Normalized Radon Cumulative Distribution Transforms for Invariance and Robustness in Optimal Transport Based Image Classification*

Matthias Beckmann†, Robert Beinert‡, and Jonas Bresch§

Abstract. The Radon cumulative distribution transform (R-CDT), is an easy-to-compute feature extractor that facilitates image classification tasks especially in the small data regime. It is closely related to the sliced Wasserstein distance and provably guaranties the linear separability of image classes that emerge from translations or scalings. In many real-world applications, like the recognition of watermarks in filigranology, however, the data is subject to general affine transformations originating from the measurement process. To overcome this issue, we recently introduced the so-called max-normalized R-CDT that only requires elementary operations and guaranties the separability under arbitrary affine transformations. The aim of this paper is to continue our study of the max-normalized R-CDT especially with respect to its robustness against non-affine image deformations. Our sensitivity analysis shows that its separability properties are stable provided the Wasserstein-infinity distance between the samples can be controlled. Since the Wasserstein-infinity distance only allows small local image deformations, we moreover introduce a mean-normalized version of the R-CDT. In this case, robustness relates to the Wasserstein-2 distance and also covers image deformations caused by impulsive noise for instance. Our theoretical results are supported by numerical experiments showing the effectiveness of our novel feature extractors as well as their robustness against local non-affine deformations and impulsive noise.

Key words. Radon-CDT, sliced Wasserstein distance, feature representation, invariance, image classification, pattern recognition, small data regime

AMS subject classifications. 94A08, 65D18, 68T10

1. Introduction.

Automatic pattern recognition and data classification play a crucial role in various scientific disciplines and applications, like medical imaging, biometrics, computer vision or document analysis, to name just a few. As of today, end-to-end deep neural networks provide the state of the art if sufficient training data is available. In the small data regime, however, or, if performance guarantees are important, hand-crafted feature extractors and classifiers are still the first choice. Ideally, the feature representation is designed to transform the different classes to linearly separable subsets. This can, for instance, be achieved by applying the Radon cumulative distribution transform (R-CDT) introduced in [15], which is based on one-dimensional optimal transport maps, also called cumulative distribution transform [21, 1], that are generalized to two-dimensional data by applying the Radon transform [25, 14], known from computerized tomography [26, 19]. This approach shows great potential in many applications [16, 27, 2, 11, 30] and is closely related to the sliced Wasserstein distance [8, 28, 17, 20, 22]. A similar approach for data on the sphere is studied in [23, 24], for multi-dimensional optimal transport maps in [18, 9], and for optimal Gromov–Wasserstein transport maps in [5, 6].

*Preliminary and exploratory ideas have been presented in our conference paper [3].

†Center for Industrial Mathematics, University of Bremen, Germany & Department of Electrical and Electronic Engineering, Imperial College London, UK (research@mbeckmann.de).

‡Institut für Mathematik, Technische Universität Berlin, Germany (beinert@math.tu-berlin.de).

§Institut für Mathematik, Technische Universität Berlin, Germany (bresch@math.tu-berlin.de).

In our recent work [3], we introduced a novel normalization of the R-CDT, we referred to as max-normalized R-CDT ($_{\mathrm{m}}$NR-CDT), to enhance linear separability in the context of affine transformations. This was inspired by the special needs for applying pattern recognition techniques in filigranology—the study of watermarks. These play a central role in provenance research like dating of historical manuscripts, scribe identification and paper mill attribution. For automatic classification, the main issues are the enormous number of classes with only few members per class, see WZIS[1], as well as the uncertainty with respect to the position, size, orientation and slight distortion of the watermark in the digitized image. A first end-to-end processing pipeline for thermograms of watermarks including an R-CDT-based classification is proposed in [13], where classification invariance with respect to translation and dilation of the watermark is achieved, but other affine transformations are not included. In contrast to this, our $_{\mathrm{m}}$NR-CDT ensures the linear separability of affinely transformed classes without any restrictions on the affine transformations, as theoretically shown in [3, Theorem 1] and numerically validated by proof-of-concept experiments in [3, § 4].

In this work, we go one step further and analyse the linear separability in $_{\mathrm{m}}$NR-CDT space when allowing for small perturbations measured in Wasserstein-∞ space in addition to affine transformations. Recall that our definition of the $_{\mathrm{m}}$NR-CDT assumes a compact support of the considered measure. To weaken this assumption, we introduce a new normalization of the R-CDT, which we call mean-normalized R-CDT ($_{\mathrm{a}}$NR-CDT). We study the linear separability in $_{\mathrm{a}}$NR-CDT space for measure classes constructed by affinely transforming distinguishable template measures. We observe that, in contrast to $_{\mathrm{m}}$NR-CDT, our new normalization step poses restrictions on the affine transformations in order to guarantee separability. However, when considering perturbations of the templates, our new normalization allows for distortions measured in Wasserstein-2 distance instead of the more restrictive Wasserstein-∞ metric.

This manuscript is organized as follows. In Section 2, we introduce basic concepts and fix our notation. Section 3 is devoted to the R-CDT for bivariate measures, where we start with explaining the CDT for probability measures on $\mathbb{R}$ and, then, extend it to the R-CDT for probability measures on $\mathbb{R}^2$ by means of the Radon transform. In Section 4 we first recall our definition of the normalized R-CDT from [3] and show elementary properties. The $_{\mathrm{m}}$NR-CDT is explained in Section 4.1, where we also recall the linear separability result from [3] and extend it by considering perturbations in Wasserstein-∞ space. Thereon, our novel $_{\mathrm{a}}$NR-CDT is introduced in Section 4.2 and we show linear separability in $_{\mathrm{a}}$NR-CDT space under affine transformations and, additionally, perturbations in Wasserstein-2 space. Our theoretical findings are illustrated by numerical experiments in Section 5, showing the effectiveness of our approach. Section 6 concludes with a discussion of our results and future research direction.

2. Preliminaries. Throughout the paper, we restrict our attention to functions and measures on the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$ and the infinite cylinder $\mathbb{R} \times \mathbb{S}_1$ with $\mathbb{S}_1 \coloneqq \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| = 1\}$. Compact subsets of these are indicated using the symbol $\subset\subset$. For domain X, we denote the Lebesgue spaces by $L^p(X)$; the space of continuous, bounded functions by $C_b(X)$; the space of continuous functions vanishing at infinity by $C_0(X)$; and the space of finite, signed, regular (Borel) measures by $\mathcal{M}(X)$. Recall that $\mathcal{M}(X)$ is the continuous dual of $C_0(X)$. For $\mu \in \mathcal{M}(X)$, its *support* $\operatorname{supp}(\mu)$ is the minimal closed subset $Y \subset X$ such that

[1] Wasserzeichen-Informationssystem: www.wasserzeichen-online.de.

$\mu(X \setminus Y) = 0$. For $\mu \in \mathcal{M}(\mathbb{R}^d)$, the *dimension* of the affine hull of $\operatorname{supp}(\mu)$ is denoted by $\dim(\mu)$, and the *diameter* is given by $\operatorname{diam}(\mu) \coloneqq \sup_{\mathbf{x},\mathbf{y}\in\operatorname{supp}(\mu)} \|\mathbf{x}-\mathbf{y}\|$.

For two domains X and Y, the *push-forward* of a Borel probability measure $\mu \in \mathcal{P}(X)$ via a mapping $T\colon X \to Y$ is defined by $T_{\#}\mu \coloneqq \mu \circ T^{-1}$. To compare two measures, we use the so-called Wasserstein or Kantorovich–Rubinstein metric, which is defined on spaces of *probability measures with finite pth moment* given by

$$\mathcal{P}_p(\mathbb{R}^d) \coloneqq \Big\{\mu \in \mathcal{P}(\mathbb{R}^d) \;\Big|\; \int_{\mathbb{R}^d} \|\mathbf{x}\|^p \,\mathrm{d}\mathbf{x} < \infty\Big\}, \qquad p \in [1,\infty),$$
$$\mathcal{P}_\infty(\mathbb{R}^d) \coloneqq \Big\{\mu \in \mathcal{P}(\mathbb{R}^d) \;\Big|\; \sup\nolimits_{\mathbf{x}\in\operatorname{supp}(\mu)} \|\mathbf{x}\| < \infty\Big\}.$$

Moreover, we introduce the canonical projections $P_1(\mathbf{x},\mathbf{y}) \coloneqq \mathbf{x}$ and $P_2(\mathbf{x},\mathbf{y}) \coloneqq \mathbf{y}$ as well as the *set of transport plans*

$$\Pi(\mu,\nu) \coloneqq \big\{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \;\big|\; (P_1)_{\#}\pi = \mu,\ (P_2)_{\#}\pi = \nu\big\}, \qquad \mu,\nu \in \mathcal{P}(\mathbb{R}^d).$$

Then, the *Wasserstein-p distance* between $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$W_p(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|^p \,\mathrm{d}\pi(\mathbf{x},\mathbf{y})\Big)^{\frac{1}{p}}, \qquad p \in [1,\infty) \tag{2.1a}$$

$$W_\infty(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \sup_{(\mathbf{x},\mathbf{y})\in\operatorname{supp}(\pi)} \|\mathbf{x}-\mathbf{y}\|. \tag{2.1b}$$

The Wasserstein-p space $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is a metric space, and the infima in (2.1) are attained by an optimal transport plan $\pi \in \Pi(\mu,\nu)$; see [12, Prop. 1 & 2].

3. Radon Cumulative Distribution Transform.

Following the approach in [21], for a probability measure $\mu \in \mathcal{P}(\mathbb{R})$, we consider its cumulative distribution function $F_\mu\colon \mathbb{R} \to [0,1]$ given by

$$F_\mu(t) = \mu((-\infty,t]), \quad t \in \mathbb{R},$$

and define the *cumulative distribution transform* $\widehat{\mu}\colon \mathbb{R} \to \mathbb{R}$, in short CDT, via

$$\widehat{\mu} = F_\mu^{[-1]} \circ F_\rho$$

with the generalized inverse, known as quantile function,

$$F_\mu^{[-1]}(t) = \inf\{s \in \mathbb{R} \mid F_\mu(s) > t\}, \quad t \in \mathbb{R}$$

and a reference $\rho \in \mathcal{P}_2(\mathbb{R})$ that does not give mass to atoms, e.g., $\rho = \chi_{[0,1]}\lambda_{\mathbb{R}}$, where $\lambda_{\mathbb{R}}$ denotes the standard Lebesgue measure on $\mathbb{R}$. Note that, if $\mu \in \mathcal{P}_2(\mathbb{R})$ has finite second moment, we have

$$\widehat{\mu} = \operatorname*{arg\,min}_{T_{\#}\rho=\mu} \int_{\mathbb{R}} |s - T(s)|^2 \,\mathrm{d}\rho(s),$$

i.e., $\widehat{\mu}$ is the unique map $T\colon \mathbb{R} \to \mathbb{R}$ transporting ρ to μ while minimizing the cost, cf. [29]. Moreover, $\widehat{\mu}$ is square integrable with respect to ρ, i.e., $\widehat{\mu} \in L^2_\rho(\mathbb{R})$, and, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, we have

$$\|\widehat{\mu} - \widehat{\nu}\|_\rho \coloneqq \Big(\int_{\mathbb{R}} |\mu(t) - \nu(t)|^2 \,\mathrm{d}\rho(t)\Big)^{\frac{1}{2}} = W_2(\mu, \nu).$$

To deal with a probability measure $\mu \in \mathcal{P}(\mathbb{R}^2)$, we adapt the approach in [15] and apply the so-called Radon transform to obtain a family of probability measures on $\mathbb{R}$. For a bivariate function $f \in L^1(\mathbb{R}^2)$, its *Radon transform* $\mathcal{R}[f]\colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$ is defined as the line integral

$$\mathcal{R}[f](t, \boldsymbol{\theta}) \coloneqq \int_{\ell_{t,\boldsymbol{\theta}}} f(s) \,\mathrm{d}s, \quad (t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}_1,$$

where $\mathrm{d}s$ denotes the arc length element of the straight line $\ell_{t,\boldsymbol{\theta}}$ with signed distance $t \in \mathbb{R}$ to the origin and normal direction $\boldsymbol{\theta} \in \mathbb{S}_1 \coloneqq \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| = 1\}$, i.e.,

$$\ell_{t,\boldsymbol{\theta}} \coloneqq \{t\boldsymbol{\theta} + \tau\boldsymbol{\theta}^\perp \mid \tau \in \mathbb{R}\} = S_{\boldsymbol{\theta}}^{-1}(t) \subset \mathbb{R}^2$$

with *slicing operator* $S_{\boldsymbol{\theta}}\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$S_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq \langle \mathbf{x}, \boldsymbol{\theta} \rangle, \quad \mathbf{x} \in \mathbb{R}^2.$$

This defines the *Radon operator* $\mathcal{R}\colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R} \times \mathbb{S}_1)$ and, for fixed $\boldsymbol{\theta} \in \mathbb{S}_1$, we set $\mathcal{R}_{\boldsymbol{\theta}} \coloneqq \mathcal{R}(\cdot, \boldsymbol{\theta})$, which is referred to as the *restricted Radon operator* $\mathcal{R}_{\boldsymbol{\theta}}\colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R})$. It is well known that $\mathcal{R}$ preserves mass, cf. [19], in the sense that, for any $f \in L^1(\mathbb{R}^2)$, we have

$$\int_{\mathbb{R}} \mathcal{R}_{\boldsymbol{\theta}}[f](t) \,\mathrm{d}t = \int_{\mathbb{R}^2} f(\mathbf{x}) \,\mathrm{d}\mathbf{x}, \qquad \int_{\mathbb{S}_1} \int_{\mathbb{R}} \mathcal{R}[f](t, \boldsymbol{\theta}) \,\mathrm{d}t \,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) = \int_{\mathbb{R}^2} f(\mathbf{x}) \,\mathrm{d}\mathbf{x}, \tag{3.1}$$

where $u_{\mathbb{S}_1} = \frac{\sigma_{\mathbb{S}_1}}{2\pi}$ with surface measure $\sigma_{\mathbb{S}_1}$ on $\mathbb{S}_1$. The adjoint $\mathcal{R}^*\colon L^\infty(\mathbb{R} \times \mathbb{S}_1) \to L^\infty(\mathbb{R}^2)$ of Radon operator $\mathcal{R}\colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R} \times \mathbb{S}_1)$, called *back projection operator*, is given by

$$\mathcal{R}^*[g](\mathbf{x}) \coloneqq \int_{\mathbb{S}_1} g(S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}) \,\mathrm{d}\sigma_{\mathbb{S}_1}(\boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{R}^2,$$

and, for fixed $\boldsymbol{\theta} \in \mathbb{S}_1$, the adjoint $\mathcal{R}^*_{\boldsymbol{\theta}}\colon L^\infty(\mathbb{R}) \to L^\infty(\mathbb{R}^2)$ of the restricted Radon operator $\mathcal{R}_{\boldsymbol{\theta}}\colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R})$ is given by

$$\mathcal{R}^*_{\boldsymbol{\theta}}[h](\mathbf{x}) = h(S_{\boldsymbol{\theta}}(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^2.$$

Furthermore, $\mathcal{R}^*\colon C_b(\mathbb{R} \times \mathbb{S}_1) \to C_b(\mathbb{R}^2)$ and $\mathcal{R}^*_{\boldsymbol{\theta}}\colon C_b(\mathbb{R}) \to C_b(\mathbb{R}^2)$. This allows us to translate the concept of the Radon transform to signed, regular, finite measures $\mu \in \mathcal{M}(\mathbb{R}^2)$. For a fixed direction $\boldsymbol{\theta} \in \mathbb{S}_1$, we generalize the *restricted Radon transform* $\mathcal{R}_{\boldsymbol{\theta}}$ to measures by setting

$$\mathcal{R}_{\boldsymbol{\theta}}\colon \mathcal{M}(\mathbb{R}^2) \to \mathcal{M}(\mathbb{R}), \quad \mu \mapsto (S_{\boldsymbol{\theta}})_{\#}\mu = \mu \circ S_{\boldsymbol{\theta}}^{-1},$$

which corresponds to the integration along the lines $\ell_{t,\boldsymbol{\theta}}$. As for functions in (3.1), we have

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu](\mathbb{R}) = \mu(\mathbb{R}^2) \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1,$$

thus, mass is preserved by $\mathcal{R}_{\boldsymbol{\theta}}$. In measure theory, $\mathcal{R}_{\boldsymbol{\theta}}$ can be considered as a disintegration family and, heuristically, we generalize the Radon transform by integrating $\mathcal{R}_{\boldsymbol{\theta}}$ along $\boldsymbol{\theta} \in \mathbb{S}_1$. Therefore, we define the *Radon transform* $\mathcal{R}\colon \mathcal{M}(\mathbb{R}^2) \to \mathcal{M}(\mathbb{R} \times \mathbb{S}_1)$ via

$$\mathcal{R}[\mu] \coloneqq \mathcal{I}_{\#}[\mu \times u_{\mathbb{S}_1}]$$

with *glueing operator* $\mathcal{I}\colon \mathbb{R}^2 \times \mathbb{S}_1 \to \mathbb{R} \times \mathbb{S}_1$ given by

$$\mathcal{I}(\mathbf{x}, \boldsymbol{\theta}) \coloneqq (S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}), \quad (\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^2 \times \mathbb{S}_1.$$

In [3, Proposition 1] we have shown that, for $\mu \in \mathcal{M}(\mathbb{R}^2)$, $\mathcal{R}[\mu]$ can indeed be disintegrated into the family $\mathcal{R}_{\boldsymbol{\theta}}[\mu]$ with respect to the uniform measure $u_{\mathbb{S}_1} = \frac{\sigma_{\mathbb{S}_1}}{2\pi}$, i.e., for all $g \in C_0(\mathbb{R}\times\mathbb{S}_1)$, we have

$$\langle \mathcal{R}[\mu], g\rangle = \int_{\mathbb{S}_1} \langle \mathcal{R}_{\boldsymbol{\theta}}[\mu], g(\cdot, \boldsymbol{\theta})\rangle \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}).$$

Moreover, in [3, Proposition 2] we have proven that the measure-valued transforms $\mathcal{R}$ and $\mathcal{R}_{\boldsymbol{\theta}}$ are the adjoints of the back projection operators $\mathcal{R}^*$ and $\mathcal{R}^*_{\boldsymbol{\theta}}$ from above. More precisely, the Radon transform of $\mu \in \mathcal{M}(\mathbb{R}^2)$ satisfies

$$\langle \mathcal{R}[\mu], g\rangle = \langle \mu, \mathcal{R}^*[g]\rangle \quad \forall\, g \in L^\infty(\mathbb{R} \times \mathbb{S}_1) \tag{3.2}$$

and

$$\langle \mathcal{R}_{\boldsymbol{\theta}}[\mu], h\rangle = \langle \mu, \mathcal{R}^*_{\boldsymbol{\theta}}[h]\rangle \quad \forall\, h \in L^\infty(\mathbb{R}) \;\forall\, \boldsymbol{\theta} \in \mathbb{S}_1.$$

This observation suggests that the Radon transform for measures can equivalently be defined through duality. However, the dual space of $\mathcal{M}(\mathbb{R})$ is $C_0(\mathbb{R})$, whereas $h \in C_0(\mathbb{R}) \setminus \{0\}$ does not imply that $\mathcal{R}^*_{\boldsymbol{\theta}}[h] \in C_0(\mathbb{R})$ for $\boldsymbol{\theta} \in \mathbb{S}_1$. But if $h \in C_0(\mathbb{R} \times \mathbb{S}_1)$, we have $\mathcal{R}^*[h] \in C_0(\mathbb{R}^2)$.

Proposition 3.1. *Let $h \in C_0(\mathbb{R} \times \mathbb{S}_1)$. Then, we have $\mathcal{R}^*[h] \in C_0(\mathbb{R}^2)$.*

Proof. Let $(\mathbf{x}_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^2$ such that $\|\mathbf{x}_n\| \to \infty$ for $n \to \infty$. For arbitrarily $\varepsilon > 0$ there exists $M_\varepsilon > 0$ such that $|h(x, \boldsymbol{\theta})| < \frac{\varepsilon}{2}$ for all $|x| \geq M_\varepsilon$ and $\boldsymbol{\theta} \in \mathbb{S}_1$. Hence, we have

$$|\mathcal{R}^*[h](\mathbf{x}_n)| = \left| \int_{\mathbb{S}_1} h(\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle, \boldsymbol{\theta}) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \right| \leq \frac{\varepsilon}{2} + \int_{\{|\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle| < M_\varepsilon\}} |h(\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle, \boldsymbol{\theta})| \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}).$$

Since $h \in C_0(\mathbb{R} \times \mathbb{S}_1)$, there exits for $\varepsilon' > 0$ a compact set $K_{\varepsilon'}$ such that $|h|_{K_{\varepsilon'}\times\mathbb{S}_1}| \geq \varepsilon'$ and hence there exists $C \coloneqq \max_{(x,\boldsymbol{\theta})\in\mathbb{R}\times\mathbb{S}_1} |h(x, \boldsymbol{\theta})| > 0$. Combining both yields

$$|\mathcal{R}^*[h](\mathbf{x}_n)| \leq \frac{\varepsilon}{2} + C\, u_{\mathbb{S}_1}(\{|\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle| < M_\varepsilon\}).$$

Since

$$|\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle| < M_\varepsilon \quad \Leftrightarrow \quad \left|\left\langle \frac{\mathbf{x}_n}{\|\mathbf{x}_n\|}, \boldsymbol{\theta}\right\rangle\right| < \frac{M_\varepsilon}{\|\mathbf{x}_n\|},$$

there exists $N \in \mathbb{N}$ such that $u_{\mathbb{S}_1}(\{|\langle \mathbf{x}_n, \boldsymbol{\theta}\rangle| < M_\varepsilon\}) < \frac{\varepsilon}{2C}$ and the assertion follows. ∎

The definition of the Radon transform for measures is compatible with the classical definition for functions in the sense that the Radon transform of an absolutely continuous measure is again absolutely continuous. To see this, we denote the surface measure by $\sigma_{\mathbb{K}}$ and the Lebesgue measure by $\lambda_{\mathbb{K}}$ for the different sets $\mathbb{K} \in \{\mathbb{R}, \mathbb{R}^2, \mathbb{S}_1, \mathbb{R} \times \mathbb{S}_1\}$.

Proposition 3.2. *Let $f \in L^1(\mathbb{R}^2)$. The Radon transform satisfies*

$$\mathcal{R}[f\lambda_{\mathbb{R}^2}] = \mathcal{R}[f]\,\sigma_{\mathbb{R}\times\mathbb{S}_1} \quad \text{and} \quad \mathcal{R}_{\boldsymbol{\theta}}[f\lambda_{\mathbb{R}^2}] = \mathcal{R}_{\boldsymbol{\theta}}[f]\,\lambda_{\mathbb{R}}.$$

Proof. We denote by $\langle\cdot,\cdot\rangle_{\mathcal{M}}$ the dual pairing between $\mathcal{M}$ and C_0 and by $\langle\cdot,\cdot\rangle_L$ the dual pairing between L^1 and L^∞. Then, the duality relation in (3.2) gives

$$\langle \mathcal{R}[f\lambda_{\mathbb{R}^2}], g\rangle_{\mathcal{M}} = \langle f\lambda_{\mathbb{R}^2}, \mathcal{R}^*[g]\rangle_{\mathcal{M}} = \langle f, \mathcal{R}^*[g]\rangle_L = \langle \mathcal{R}[f], g\rangle_L = \langle \mathcal{R}[f]\,\sigma_{\mathbb{R}\times\mathbb{S}_1}, g\rangle_{\mathcal{M}}$$

for all $g \in C_0(\mathbb{R} \times \mathbb{S}_1)$. ∎

We have the following connection between the Wasserstein distance of measures and the Wasserstein distance of the corresponding restricted Radon transforms.

Proposition 3.3. *Let $\boldsymbol{\theta} \in \mathbb{S}_1$. Then, we have*

$$W_\infty(\mathcal{R}_{\boldsymbol{\theta}}[\mu], \mathcal{R}_{\boldsymbol{\theta}}[\nu]) \le W_\infty(\mu,\nu) \quad \forall\, \mu,\nu \in \mathcal{P}_c(\mathbb{R}^2),$$
$$W_2(\mathcal{R}_{\boldsymbol{\theta}}[\mu], \mathcal{R}_{\boldsymbol{\theta}}[\nu]) \le W_2(\mu,\nu) \quad \forall\, \mu,\nu \in \mathcal{P}_2(\mathbb{R}^2).$$

Proof. For $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^2)$, let $\pi \in \Pi(\mu,\nu)$ realize $W_2(\mu,\nu)$. Since the push-forward plan satisfies $(S_{\boldsymbol{\theta}}, S_{\boldsymbol{\theta}})_{\#}\pi \in \Pi(\mathcal{R}_{\boldsymbol{\theta}}[\mu], \mathcal{R}_{\boldsymbol{\theta}}[\nu])$, the second inequality follows from

$$\begin{aligned} W_2^2(\mu,\nu) &= \int_{\mathbb{R}^2\times\mathbb{R}^2} \|\mathbf{x}-\mathbf{y}\|^2 \,\mathrm{d}\pi(\mathbf{x},\mathbf{y}) \ge \int_{\mathbb{R}^2\times\mathbb{R}^2} |\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\theta}\rangle|^2 \,\mathrm{d}\pi(\mathbf{x},\mathbf{y}) \\ &= \int_{\mathbb{R}\times\mathbb{R}} |t-s|^2 \,\mathrm{d}(S_{\boldsymbol{\theta}}, S_{\boldsymbol{\theta}})_{\#}\pi(t,s) \ge W_2^2(\mathcal{R}_{\boldsymbol{\theta}}[\mu], \mathcal{R}_{\boldsymbol{\theta}}[\nu]). \end{aligned}$$

The first inequality can be established analogously. ∎

We are now prepared to adapt the CDT to a probability measure $\mu \in \mathcal{P}(\mathbb{R}^2)$. To this end, we first consider its Radon transform $\mathcal{R}[\mu] \in \mathcal{M}(\mathbb{R} \times \mathbb{S}_1)$, which satisfies

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu] \in \mathcal{P}(\mathbb{R}) \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1,$$

and then, for each fixed $\boldsymbol{\theta} \in \mathbb{S}_1$, the CDT $\widehat{\mu}_{\boldsymbol{\theta}}$ of the Radon projection $\mu_{\boldsymbol{\theta}} = \mathcal{R}_{\boldsymbol{\theta}}[\mu]$, yielding the *Radon cumulative distribution transform* (R-CDT) $\widehat{\mathcal{R}}\mu\colon \mathbb{R}\times\mathbb{S}_1 \to \mathbb{R}$ of μ via

$$\widehat{\mathcal{R}}[\mu](t,\boldsymbol{\theta}) = \widehat{\mu}_{\boldsymbol{\theta}}(t), \quad (t,\boldsymbol{\theta}) \in \mathbb{R}\times\mathbb{S}_1.$$

In this way, any probability measure $\mu \in \mathcal{P}(\mathbb{R}^2)$ is mapped to its R-CDT $\widehat{\mathcal{R}}[\mu]\colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$. If $\mu \in \mathcal{P}_2(\mathbb{R}^2)$, then the Radon projection $\mathcal{R}_{\boldsymbol{\theta}}[\mu] \in \mathcal{P}_2(\mathbb{R})$ has finite second moment as well and we have $\widehat{\mathcal{R}}[\mu] \in L^2_{\rho \times u_{\mathbb{S}_1}}(\mathbb{R} \times \mathbb{S}_1)$. Moreover, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^2)$, the norm distance

$$\|\widehat{\mathcal{R}}[\mu] - \widehat{\mathcal{R}}[\nu]\|_{\rho \times u_{\mathbb{S}_1}} \coloneqq \Big(\int_{\mathbb{S}_1} \int_{\mathbb{R}} |\widehat{\mathcal{R}}[\mu](t, \boldsymbol{\theta}) - \widehat{\mathcal{R}}[\nu](t, \boldsymbol{\theta})|^2 \, \mathrm{d}\rho(t) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \Big)^{\frac{1}{2}}$$

agrees with the so-called sliced Wasserstein-2 distance [8].

Linear separability in R-CDT space. We now show that the above feature representation via R-CDT enhances linear separability of distinct classes that are generated from template measures by certain transformations. To be more precise, we assume that, for fixed $\boldsymbol{\theta}_0 \in \mathbb{S}_1$, we are given two classes $\mathbb{F}_{\boldsymbol{\theta}_0}, \mathbb{G}_{\boldsymbol{\theta}_0}$ that are generated by template measures $\mu_0, \ \nu_0 \in \mathcal{P}(\mathbb{R}^2)$ via

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}_0} &= \big\{ \mu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, h \in \mathbb{H}\colon \mathcal{R}_{\boldsymbol{\theta}_0}[\mu] = h_{\#} \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0] \big\}, \\ \mathbb{G}_{\boldsymbol{\theta}_0} &= \big\{ \nu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, h \in \mathbb{H}\colon \mathcal{R}_{\boldsymbol{\theta}_0}[\nu] = h_{\#} \mathcal{R}_{\boldsymbol{\theta}_0}[\nu_0] \big\}, \end{aligned}$$

where $\mathbb{H}$ is a convex set of increasing bijections $h\colon \mathbb{R} \to \mathbb{R}$. Then, we will prove that the transformed function classes in R-CDT space

$$\widehat{\mathbb{F}}_{\boldsymbol{\theta}_0} = \big\{ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu]\colon \mathbb{R} \to \mathbb{R} \mid \mu \in \mathbb{F}_{\boldsymbol{\theta}_0} \big\}, \qquad \widehat{\mathbb{G}}_{\boldsymbol{\theta}_0} = \big\{ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\nu]\colon \mathbb{R} \to \mathbb{R} \mid \nu \in \mathbb{G}_{\boldsymbol{\theta}_0} \big\}$$

are linearly separable if $\mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{F}_{\boldsymbol{\theta}_0} \cap \mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{G}_{\boldsymbol{\theta}_0} = \emptyset$, where

$$\mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{F}_{\boldsymbol{\theta}_0} = \{ \mathcal{R}_{\boldsymbol{\theta}_0}[\mu] \mid \mu \in \mathbb{F}_{\boldsymbol{\theta}_0} \}, \qquad \mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{G}_{\boldsymbol{\theta}_0} = \{ \mathcal{R}_{\boldsymbol{\theta}_0}[\nu] \mid \nu \in \mathbb{G}_{\boldsymbol{\theta}_0} \},$$

in the sense that for any two non-empty, finite subsets $\mathbb{F}_0 \subset \mathbb{F}_{\boldsymbol{\theta}_0}$ and $\mathbb{G}_0 \subset \mathbb{G}_{\boldsymbol{\theta}_0}$ there exist a continuous linear functional $\Phi\colon L^2_\rho(\mathbb{R}) \to \mathbb{R}$ and a constant $c \in \mathbb{R}$ such that

$$\Phi\Big(\widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu]\Big) < c < \Phi\Big(\widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\nu]\Big) \quad \forall\, \mu \in \mathbb{F}_0 \ \forall\, \nu \in \mathbb{G}_0.$$

Note that a slightly different version of this result has first been stated in [13, § 3.2] without a proof and we now close this gap in the literature.

Theorem 3.4. *For templates $\mu_0, \nu_0 \in \mathcal{P}_2(\mathbb{R}^2)$ and $\boldsymbol{\theta}_0 \in \mathbb{S}_1$ consider the classes*

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}_0} &= \big\{ \mu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, h \in \mathbb{H}\colon \mathcal{R}_{\boldsymbol{\theta}_0}[\mu] = h_{\#} \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0] \big\}, \\ \mathbb{G}_{\boldsymbol{\theta}_0} &= \big\{ \nu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, h \in \mathbb{H}\colon \mathcal{R}_{\boldsymbol{\theta}_0}[\nu] = h_{\#} \mathcal{R}_{\boldsymbol{\theta}_0}[\nu_0] \big\}, \end{aligned}$$

where $\mathbb{H}$ is a convex set of increasing bijections $h\colon \mathbb{R} \to \mathbb{R}$. Then, any non-empty, finite subsets $\mathbb{F}_0 \subset \mathbb{F}_{\boldsymbol{\theta}_0}$ and $\mathbb{G}_0 \subset \mathbb{G}_{\boldsymbol{\theta}_0}$ are linearly separable in R-CDT space if $\mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{F}_{\boldsymbol{\theta}_0} \cap \mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{G}_{\boldsymbol{\theta}_0} = \emptyset$ in the sense that

$$\widehat{\mathbb{F}}_0 = \big\{ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu]\colon \mathbb{R} \to \mathbb{R} \mid \mu \in \mathbb{F}_0 \big\}, \qquad \widehat{\mathbb{G}}_0 = \big\{ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\nu]\colon \mathbb{R} \to \mathbb{R} \mid \nu \in \mathbb{G}_0 \big\}$$

are linearly separable in $\mathrm{L}^2_\rho(\mathbb{R})$.

Table 1

Summary of common transformations for $\mu \in \mathcal{M}(\mathbb{R}^2)$ with $a, b > 0$ and $c, \varphi \in \mathbb{R}$. The unit circle is parametrized by $\boldsymbol{\theta}(\vartheta) \coloneqq (\cos(\vartheta), \sin(\vartheta))^\top$. The Radon transform for the left half of $\mathbb{S}_1$ follows by symmetry.

transformation	$\mathbf{A}$	$\mathbf{y}$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta)}[\mu_{\mathbf{A},\mathbf{y}}]$, $\vartheta \in (-\frac{\pi}{2}, \frac{\pi}{2})$
translation	$\mathbf{I}$	$\mathbb{R}^2$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta)}[\mu] \circ (\cdot - \langle \mathbf{y}, \boldsymbol{\theta}(\vartheta) \rangle)$
rotation	$\left(\begin{smallmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta-\varphi)}[\mu]$
reflection	$\left(\begin{smallmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\varphi-\vartheta)}[\mu]$
anisotropic scaling	$\left(\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\arctan(\frac{b}{a}\tan(\vartheta)))}[\mu] \circ ([a^2\cos^2(\vartheta) + b^2\sin^2(\vartheta)]^{-1/2}\cdot)$
vertical shear	$\left(\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\arctan(c+\tan(\vartheta)))}[\mu] \circ ([1 + c^2\cos^2(\vartheta) + c\sin(2\vartheta)]^{-1/2}\cdot)$

Proof. As $\mu_0, \nu_0 \in \mathcal{P}_2(\mathbb{R}^2)$ have finite second moments, $\mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0], \mathcal{R}_{\boldsymbol{\theta}_0}[\nu_0] \in \mathcal{P}_2(\mathbb{R})$ have finite second moments as well, and we get $\widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_0], \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\nu_0] \in \mathrm{L}^2_\rho(\mathbb{R})$ so that, in particular, $\widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}, \widehat{\mathbb{G}}_{\boldsymbol{\theta}_0} \subset \mathrm{L}^2_\rho(\mathbb{R})$. We now show that $\widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}$ and $\widehat{\mathbb{G}}_{\boldsymbol{\theta}_0}$ are convex. To this end, let $\widehat{p}_1, \widehat{p}_2 \in \widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}$ and $\alpha \in [0,1]$. Then, there exist $\mu_1, \mu_2 \in \mathbb{F}_{\boldsymbol{\theta}_0}$ such that $\widehat{p}_1 = \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_1]$ and $\widehat{p}_2 = \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_2]$. Set $p_0 = \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0]$, $p_1 = \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_1]$ and $p_2 = \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_2]$. By the definition of $\mathbb{F}_{\boldsymbol{\theta}_0}$ there are $h_1, h_2 \in \mathbb{H}$ such that $p_i = (h_i)_\# p_0$ for $i \in \{1,2\}$, where $F_{p_i} = F_{p_0} \circ h_i^{-1}$ so that

$$\widehat{p}_i = F_{p_i}^{[-1]} \circ F_\rho = h_i \circ \left(F_{p_0}^{[-1]} \circ F_\rho\right) = h_i \circ \widehat{p}_0.$$

Consequently,

$$\alpha\widehat{p}_1 + (1-\alpha)\widehat{p}_2 = (\alpha h_1 + (1-\alpha)h_2) \circ \widehat{p}_0 = h_\alpha \circ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_0]$$

with $h_\alpha = \alpha h_1 + (1-\alpha)h_2 \in \mathbb{H}$ as $\mathbb{H}$ is convex. Now, choose $\mu_\alpha \in \mathbb{F}_{\boldsymbol{\theta}_0}$ such that $\mathcal{R}_{\boldsymbol{\theta}_0}[\mu_\alpha] = (h_\alpha)_\# \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0]$, e.g., $\mu_\alpha \coloneqq (\cdot\,\boldsymbol{\theta}_0)_\#(h_\alpha)_\# \mathcal{R}_{\boldsymbol{\theta}_0}[\mu_0]$. As before, we then have $\widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_\alpha] = h_\alpha \circ \widehat{\mathcal{R}}_{\boldsymbol{\theta}_0}[\mu_0]$ and, hence, we conclude that $\alpha\widehat{p}_1 + (1-\alpha)\widehat{p}_2 \in \widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}$ so that $\widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}$ is indeed convex. Analogously, we obtain the convexity of $\widehat{\mathbb{G}}_{\boldsymbol{\theta}_0}$. Now, let $\mathbb{F}_0 \subset \mathbb{F}_{\boldsymbol{\theta}_0}$ and $\mathbb{G}_0 \subset \mathbb{G}_{\boldsymbol{\theta}_0}$ be non-empty and finite. Then, $\operatorname{conv}(\widehat{\mathbb{F}}_0) \subset \widehat{\mathbb{F}}_{\boldsymbol{\theta}_0}$ and $\operatorname{conv}(\widehat{\mathbb{G}}_0) \subset \widehat{\mathbb{G}}_{\boldsymbol{\theta}_0}$ are convex and compact. As $\mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{F}_{\boldsymbol{\theta}_0} \cap \mathcal{R}_{\boldsymbol{\theta}_0}\mathbb{G}_{\boldsymbol{\theta}_0} = \emptyset$, we also have $\widehat{\mathbb{F}}_{\boldsymbol{\theta}_0} \cap \widehat{\mathbb{G}}_{\boldsymbol{\theta}_0} = \emptyset$ and, in particular, $\operatorname{conv}(\widehat{\mathbb{F}}_0) \cap \operatorname{conv}(\widehat{\mathbb{G}}_0) = \emptyset$. Therefore, by the Hahn-Banach separation theorem $\operatorname{conv}(\widehat{\mathbb{F}}_0)$ and $\operatorname{conv}(\widehat{\mathbb{G}}_0)$ are linearly separable in $\mathrm{L}^2_\rho(\mathbb{R})$, which implies that also $\widehat{\mathbb{F}}_0$ and $\widehat{\mathbb{G}}_0$ are linearly separable in $\mathrm{L}^2_\rho(\mathbb{R})$. ∎

Remark 3.5. Note that Theorem 3.4 is similar to the result in [15]. There, however, the authors consider certain classes of functions that can be linearly separated in R-CDT space when considering *all* directions $\boldsymbol{\theta} \in \mathbb{S}_1$. As opposed to this, our result only needs *one* $\boldsymbol{\theta}_0 \in \mathbb{S}_1$ such that the corresponding restricted Radon transforms of the classes are distinguishable.

To study an example satisfying the assumptions of Theorem 3.4, we consider affinely transformed finite measure $\mu \in \mathcal{M}(\mathbb{R}^2)$. To this end, let $\mathbf{A} \in \mathrm{GL}(2)$, $\mathbf{y} \in \mathbb{R}^2$ and define $\mu_{\mathbf{A},\mathbf{y}} \in \mathcal{M}(\mathbb{R}^2)$ via

$$\mu_{\mathbf{A},\mathbf{y}} \coloneqq (\mathbf{A}\cdot + \mathbf{y})_\# \mu = \mu \circ (\mathbf{A}^{-1}(\cdot - \mathbf{y})).$$

Then, in [3, Proposition 3] we have shown that, for any $\boldsymbol{\theta} \in \mathbb{S}_1$, the restricted Radon transform of $\mu_{\mathbf{A},\mathbf{y}}$ is given by

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] = (\|A^\top \boldsymbol{\theta}\| \cdot + \langle \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#} \mathcal{R}_{\frac{A^\top \boldsymbol{\theta}}{\|A^\top \boldsymbol{\theta}\|}}[\mu] = \mathcal{R}_{\frac{A^\top \boldsymbol{\theta}}{\|A^\top \boldsymbol{\theta}\|}}[\mu] \circ \left(\tfrac{\cdot - \langle \mathbf{y}, \boldsymbol{\theta} \rangle}{\|A^\top \boldsymbol{\theta}\|} \right).$$

The effect of common affine transformations on the Radon transform is given in Table 1. In order to describe the deformation with respect to $\boldsymbol{\theta}$, we over-parametrize the unit circle $\mathbb{S}_1$ via $\boldsymbol{\theta}(\vartheta) \coloneqq (\cos(\vartheta), \sin(\vartheta))^\top$, $\vartheta \in \mathbb{R}$. We see that an affine transformation essentially causes a translation and scaling of the transformed measure together with a non-affine mapping in $\boldsymbol{\theta}$.

Note that, if $\mu = f \lambda_{\mathbb{R}^2}$ is absolutely continuous with respect to the Lebesgue measure $\lambda_{\mathbb{R}^2}$ with density function $f \in L^1(\mathbb{R}^2)$, we have

$$\mu_{\mathbf{A},\mathbf{y}} = \Big(|\det(\mathbf{A})|^{-1} \, f(\mathbf{A}^{-1}(\cdot - \mathbf{y})) \Big) \lambda_{\mathbb{R}^2}$$

and, for any $\boldsymbol{\theta} \in \mathbb{S}_1$,

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] = \left(\tfrac{1}{\|A^\top \boldsymbol{\theta}\|} \mathcal{R}_{\frac{A^\top \boldsymbol{\theta}}{\|A^\top \boldsymbol{\theta}\|}}[f] \Big(\tfrac{\cdot - \langle \mathbf{y}, \boldsymbol{\theta} \rangle}{\|A^\top \boldsymbol{\theta}\|} \Big) \right) \lambda_{\mathbb{R}},$$

i.e., $\mu_{\mathbf{A},\mathbf{y}} \in \mathcal{M}(\mathbb{R}^2)$ and $\mathcal{R}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] \in \mathcal{M}(\mathbb{R})$ are absolutely continuous as well.

Example 3.6. The set of affine linear function $\mathbb{H} = \{x \mapsto ax + b \mid a > 0,\ b \in \mathbb{R}\}$ satisfies the assumptions of Theorem 3.4 and corresponds to translation and isotropic scaling, see Table 1.

4. Normalized Radon Cumulative Distribution Transform. Inspecting Table 1 reveals that the only affine transformations that satisfy the assumptions of Theorem 3.4 are translation and isotropic scaling. To account for general affine transformations, in [3] we introduced a two-step normalization scheme for the R-CDT, which we recall here for the sake of completeness. To this end, we define the *normalized R-CDT* (NR-CDT) $\mathcal{N}[\mu] \colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$ of $\mu \in \mathcal{P}_2(\mathbb{R}^2)$ via

$$\mathcal{N}[\mu](t, \boldsymbol{\theta}) \coloneqq \frac{\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])}{\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])}, \quad (t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}_1,$$

where, for $g \in \mathrm{L}^2_\rho(\mathbb{R})$,

$$\operatorname{mean}(g) = \int_{\mathbb{R}} g(s) \, \mathrm{d}\rho(s), \qquad \operatorname{std}(g) = \sqrt{\int_{\mathbb{R}} |g(s) - \operatorname{mean}(g)|^2 \, \mathrm{d}\rho(s)}.$$

To ensure that the NR-CDT is well defined, we restrict ourselves to measures whose support is not contained in a straight line. More precisely, we consider the class

$$\mathcal{P}_2^*(\mathbb{R}^2) = \{\mu \in \mathcal{P}_2(\mathbb{R}^2) \mid \dim(\mu) > 1\},$$

and show that for these measures the standard deviation of the restricted Radon transform is bounded away from zero and cannot vanish.

Lemma 4.1. *Let $\mu \in \mathcal{P}_2^*(\mathbb{R}^2)$. Then, there exists a constant $c > 0$ such that*

$$\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \geq c \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1.$$

The proof is based on the following continuity result.

Lemma 4.2. *For fixed measure $\mu \in \mathcal{P}_2^*(\mathbb{R}^2)$, the functions $\mathbb{S}_1 \ni \boldsymbol{\theta} \mapsto \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \in \mathbb{R}$ and $\mathbb{S}_1 \ni \boldsymbol{\theta} \mapsto \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \in \mathbb{R}_{\geq 0}$ are continuous.*

Proof. We rewrite the mean as

$$\operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) = \int_{\mathbb{R}} \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t)\ \mathrm{d}\rho(t) = \int_{\mathbb{R}} t\ \mathrm{d}\mathcal{R}_{\boldsymbol{\theta}}[\mu](t) = \int_{\mathbb{R}^2} \langle \mathbf{x}, \boldsymbol{\theta} \rangle\ \mathrm{d}\mu(\mathbf{x}).$$

Since the integrand is continuous in $\boldsymbol{\theta}$ and uniformly bounded by $|\langle \cdot, \boldsymbol{\theta} \rangle| \leq \|\cdot\|$, the dominated convergence theorem yields the assertion. Analogously, we have

$$\operatorname{std}^2(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) = \int_{\mathbb{R}} |\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2\ \mathrm{d}\rho(t) = \int_{\mathbb{R}^2} |\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2\ \mathrm{d}\mu(\mathbf{x}).$$

The integrand is again continuous in $\boldsymbol{\theta}$ and uniformly bounded by

$$|\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2 \leq 2\|\cdot\|^2 + 2 \max_{\boldsymbol{\theta} \in \mathbb{S}_1} \operatorname{mean}^2(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]);$$

thus, the standard deviation is continuous by dominated convergence. ∎

Proof of Lemma 4.1. Assume the contrary, this is, $c = 0$. Then, due to the continuity of $\boldsymbol{\theta} \mapsto \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])$, there exists a minimizing and convergent sequence in $\mathbb{S}_1$ whose limit $\boldsymbol{\theta}$ is attained and satisfies $\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) = 0$, i.e.,

$$\int_{\mathbb{R}^2} |\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2\ \mathrm{d}\mu(\mathbf{x}) = 0.$$

Hence, the support of μ is contained in the line $\left\{ \mathbf{x} \in \mathbb{R}^2 \;\middle|\; \langle \mathbf{x}, \boldsymbol{\theta} \rangle = \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \right\}$ in contradiction to $\mu \in \mathcal{P}_2^*(\mathbb{R}^2)$. ∎

Lemma 4.1 implies the well-definedness and square-integrability of $\mathcal{N}[\mu]$.

Proposition 4.3. *Let $\mu \in \mathcal{P}_2^*(\mathbb{R}^2)$. Then, $\mathcal{N}[\mu] \in \mathrm{L}^2_{\rho \times u_{\mathbb{S}_1}}(\mathbb{R} \times \mathbb{S}_1)$.*

Proof. For $\mu \in \mathcal{P}_2^*(\mathbb{R}^2)$ we have $\widehat{\mathcal{R}}[\mu] \in \mathrm{L}^2_{\rho \times u_{\mathbb{S}_1}}(\mathbb{R} \times \mathbb{S}_1)$ and, due to Lemma 4.1, there exists a constant $c > 0$ such that

$$\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \geq c \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1.$$

Hence,

$$\begin{aligned} \|\mathcal{N}[\mu]\|^2_{\rho \times u_{\mathbb{S}_1}} &\leq c^{-2} \int_{\mathbb{S}_1} \int_{\mathbb{R}} |\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2\ \mathrm{d}\rho(t)\ \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \\ &\leq 4c^{-2}\, \|\widehat{\mathcal{R}}[\mu]\|^2_{\rho \times u_{\mathbb{S}_1}} < \infty \end{aligned}$$

so that $\mathcal{N}[\mu] \in \mathrm{L}^2_{\rho \times u_{\mathbb{S}_1}}(\mathbb{R} \times \mathbb{S}_1)$, as stated. ∎

4.1. Max-normalized R-CDT. As in [3], we now consider the more restricted class

$$\mathcal{P}_c^*(\mathbb{R}^2) = \{\mu \in \mathcal{P}(\mathbb{R}^2) \mid \operatorname{supp}(\mu) \subset\subset \mathbb{R}^2 \wedge \dim(\mu) > 1\}$$

and define the *max-normalized R-CDT* ($_{\mathrm{m}}$NR-CDT) $\mathcal{N}_{\mathrm{m}}[\mu]\colon \mathbb{R} \to \mathbb{R}$ via

$$\mathcal{N}_{\mathrm{m}}[\mu](t) \coloneqq \max_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}[\mu](t,\boldsymbol{\theta}) \quad \text{for } t \in \mathbb{R}.$$

In [3, Proposition 6] we have seen that $\mathcal{N}_{\mathrm{m}}[\mu] \in L^\infty_\rho(\mathbb{R})$ for all $\mu \in \mathcal{P}_c^*(\mathbb{R}^2)$ and we now focus on the linear separability of classes induced by affine transformations of template measures.

Linear separability in $_{\mathrm{m}}$NR-CDT space. Let $\mu_0 \in \mathcal{P}_2^*(\mathbb{R}^2)$ be a template measure and assume that $\mu \in \mathcal{P}_2(\mathbb{R}^2)$ satisfies

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu] = (a_{\boldsymbol{\theta}} \cdot + b_{\boldsymbol{\theta}})_{\#}\mathcal{R}_{h(\boldsymbol{\theta})}[\mu] \quad \text{with } a_{\boldsymbol{\theta}} > 0,\ b_{\boldsymbol{\theta}} \in \mathbb{R},$$

where $h\colon \mathbb{S}_1 \to \mathbb{S}_1$ is bijective. Then,

$$\widehat{\mathcal{R}}[\mu](t,\boldsymbol{\theta}) = a_{\boldsymbol{\theta}}\, \widehat{\mathcal{R}}[\mu_0](t, h(\boldsymbol{\theta})) + b_{\boldsymbol{\theta}}$$

so that

$$\operatorname{mean}(\widehat{\mathcal{R}}[\mu](\cdot,\boldsymbol{\theta})) = a_{\boldsymbol{\theta}}\, \operatorname{mean}(\widehat{\mathcal{R}}[\mu_0](\cdot, h(\boldsymbol{\theta}))) + b_{\boldsymbol{\theta}}, \quad \operatorname{std}(\widehat{\mathcal{R}}[\mu](\cdot,\boldsymbol{\theta})) = a_{\boldsymbol{\theta}}\, \operatorname{std}(\widehat{\mathcal{R}}[\mu_0](\cdot, h(\boldsymbol{\theta}))).$$

Consequently,

$$\mathcal{N}[\mu](t,\boldsymbol{\theta}) = \frac{\widehat{\mathcal{R}}[\mu_0](t, h(\boldsymbol{\theta})) - \operatorname{mean}(\widehat{\mathcal{R}}[\mu_0](\cdot, h(\boldsymbol{\theta})))}{\operatorname{std}(\widehat{\mathcal{R}}[\mu_0](\cdot, h(\boldsymbol{\theta})))} = \mathcal{N}[\mu_0](t, h(\boldsymbol{\theta}))$$

and, if $\mu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$,

$$\mathcal{N}_{\mathrm{m}}[\mu](t) = \max_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}[\mu](t,\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}[\mu_0](t, h(\boldsymbol{\theta})) = \mathcal{N}_{\mathrm{m}}[\mu_0](t).$$

This observation implies linear separability in $_{\mathrm{m}}$NR-CDT space if we consider classes in $\mathcal{P}_c^*(\mathbb{R}^2)$ generated by arbitrary affine-linear transforms, which has been shown in [3].

Theorem 4.4 (cf. [3, Theorem 1]). *For template measures $\mu_0, \nu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$ with*

$$\mathcal{N}_{\mathrm{m}}[\mu_0] \neq \mathcal{N}_{\mathrm{m}}[\nu_0]$$

consider the classes

$$\mathbb{F} = \big\{\mu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, \mathbf{A} \in \mathrm{GL}(2),\, \mathbf{y} \in \mathbb{R}^2\colon \mu = (\mathbf{A}\cdot + \mathbf{y})_{\#}\mu_0\big\},$$
$$\mathbb{G} = \big\{\nu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, \mathbf{A} \in \mathrm{GL}(2),\, \mathbf{y} \in \mathbb{R}^2\colon \nu = (\mathbf{A}\cdot + \mathbf{y})_{\#}\nu_0\big\}.$$

Then, any non-empty subsets $\mathbb{F}_0 \subseteq \mathbb{F}$ and $\mathbb{G}_0 \subseteq \mathbb{G}$ are linearly separable in $_m$NR-CDT space.

The proof of Theorem 4.4 shows that $_{\mathrm{m}}$NR-CDT maps $\mathbb{F}$ and $\mathbb{G}$ to one-point sets. More precisely,

$$\mathcal{N}_{\mathrm{m}}[\mathbb{F}] = \{\mathcal{N}_{\mathrm{m}}[\mu_0]\} \qquad \text{and} \qquad \mathcal{N}_{\mathrm{m}}[\mathbb{G}] = \{\mathcal{N}_{\mathrm{m}}[\nu_0]\}.$$

In the next step, we consider the linear separability of two generated classes when allowing for slight perturbations of the underlying template measures.

Linear separability under perturbations in Wasserstein space. To study the uncertainty of the max-normalized R-CDT under perturbations with respect to the Wasserstein-∞ distance, we first analyse how these effect the non-normalized R-CDT.

Proposition 4.5. *Let $\mu_0, \mu_\epsilon \in \mathcal{P}(\mathbb{R}^2)$ with $W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$. Then*

$$\|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty \leq \epsilon.$$

Proof. For any measure $\nu \in \mathcal{P}(\mathbb{R})$, the cumulative distribution and the quantile function fulfil

$$t \leq F_\nu\big(F_\nu^{[-1]}(t)\big) \quad \forall t \in (0,1) \qquad \text{and} \qquad F_\nu^{[-1]}\big(F_\nu(s)\big) \leq s \quad \forall s \in \mathbb{R}.$$

Exploiting $W_\infty(\mathcal{R}_{\boldsymbol{\theta}}[\mu_0], \mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]) \leq W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$ due to Proposition 3.3, and utilizing any W_∞ optimal plan $\pi_{\boldsymbol{\theta}} \in \Pi(\mathcal{R}_{\boldsymbol{\theta}}[\mu_0], \mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon])$, we observe

$$\begin{aligned} F_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_0]}(s) &= \pi_{\boldsymbol{\theta}}((-\infty, s] \times \mathbb{R}) = \pi_{\boldsymbol{\theta}}((-\infty, s] \times (-\infty, s+\epsilon]) \\ &\leq \pi_{\boldsymbol{\theta}}(\mathbb{R} \times (-\infty, s+\epsilon]) = F_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]}(s+\epsilon) \end{aligned}$$

for all $s \in \mathbb{R}$. Because of the monotonicity, we further have

$$\begin{aligned} F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]}(t) &\leq F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]}\big(F_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_0]}\big(F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_0]}(t)\big)\big) \\ &\leq F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]}\big(F_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]}\big(F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_0]}(t) + \epsilon\big)\big) \leq F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu_0]}(t) + \epsilon \end{aligned}$$

Interchanging the role of μ_0 and μ_ϵ, we obtain the lower bound, which yields the assertion. ∎

To simplify the notation, for $\mu \in \mathcal{P}_1(\mathbb{R}^2)$, we introduce the zero-mean quantile functions

$$\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu](t) \coloneqq \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1 \ \forall\, t \in \mathbb{R},$$

which corresponds to the first normalization step of the normalized R-CDT. Thus, we now transfer the estimate for R-CDT to the zero-mean quantile functions.

Lemma 4.6. *Let $\mu_0, \mu_\epsilon \in \mathcal{P}_1(\mathbb{R}^2)$ with $W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$. Then*

$$\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty \leq 2\epsilon.$$

Proof. Due to Proposition 4.5, we have

$$\begin{aligned} \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty &\leq \|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty + |\operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0]) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon])| \\ &\leq \|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty + \int_{\mathbb{R}} |\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0](t) - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)| \, \mathrm{d}\rho(t) \leq 2\epsilon. \end{aligned}$$

∎

For the next normalization step in $_{\mathrm{m}}$NR-CDT, we have to divide the zero-mean quantile functions by their standard deviations

$$\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) = \operatorname{std}(\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu]) = \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu]\|_\rho.$$

Depending on the smallest occurring standard deviation, a perturbation of the initial measure μ_0 may cause the following variation of the max-normalized R-CDT.

Proposition 4.7. *Let $\mu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$ and $\mu_\epsilon \in \mathcal{P}(\mathbb{R}^2)$ with $W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$, and define $c_0 := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0]) > 0$. If $\epsilon < c_0/2$, then*

$$\|\mathcal{N}_{\mathrm{m}}[\mu_0] - \mathcal{N}_{\mathrm{m}}[\mu_\epsilon]\|_\infty \leq \frac{\operatorname{diam}(\mu_0) + 2\epsilon}{c_0(c_0 - 2\epsilon)}\, 4\epsilon.$$

Proof. Since $W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$ and $\mu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$, the measure $\mu_\epsilon \in \mathcal{P}(\mathbb{R}^2)$ has compact support and, in particular, $\mu_\epsilon \in \mathcal{P}_1(\mathbb{R}^2)$. Employing Lemma 4.6, we have

$$\big|\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho - \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho\big| \leq \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \leq \sqrt{\int_{\mathbb{R}} 4\epsilon^2 \,\mathrm{d}\rho(t)} = 2\epsilon. \tag{4.1}$$

Moreover, $\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]) = \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho$ is bounded away from zero and, thus, $\mathcal{N}_{\mathrm{m}}[\mu_\epsilon]$ is well defined. Indeed, Lemma 4.6 in combination with $2\epsilon < c_0$ gives

$$\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \geq \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho - \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \geq c_0 - 2\epsilon > 0.$$

On the basis of (4.1), the perturbation after the second normalization step is for all $t \in \mathbb{R}$ bounded by

$$\begin{aligned}
|\mathcal{N}_{\boldsymbol{\theta}}[\mu_0](t) - \mathcal{N}_{\boldsymbol{\theta}}[\mu_\epsilon](t)| &\leq \left|\frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0](t)}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} - \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho}\right| + \left|\frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} - \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho}\right| \\
&\leq \frac{2\epsilon}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} + |\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)| \frac{2\epsilon}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho} \\
&\leq 2\epsilon \frac{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho + \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\infty}{c_0(c_0 - 2\epsilon)} \leq 2\epsilon \frac{(\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho + 2\epsilon) + (\operatorname{diam}(\mu_0) + 2\epsilon)}{c_0(c_0 - 2\epsilon)} \\
&\leq \frac{\operatorname{diam}(\mu_0) + 2\epsilon}{c_0(c_0 - 2\epsilon)}\, 4\epsilon.
\end{aligned}$$

Since this upper bound is independent of $\boldsymbol{\theta}$, for fixed $t \in \mathbb{R}$, it remains valid for the supremum over $\boldsymbol{\theta} \in \mathbb{S}_1$, which yields the assertion. ∎

An affine transformation only causes a non-linear deformation of the normalized R-CDT in the argument $\boldsymbol{\theta}$; therefore the estimate in Proposition 4.7 can be immediately generalized to small perturbations of the initial measure followed by an affine transformation.

Corollary 4.8. *Let $\mu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$ and $\mu_\epsilon \in \mathcal{P}(\mathbb{R}^2)$ with $W_\infty(\mu_0, \mu_\epsilon) \leq \epsilon$, and define $c_0 := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0]) > 0$. For $\mathbf{A} \in \mathrm{GL}(2)$, $\mathbf{y} \in \mathbb{R}^2$, let $\mu := (\mathbf{A}\cdot + \mathbf{y})_\# \mu_\epsilon$. If $2\epsilon < c_0$, then*

$$\|\mathcal{N}_{\mathrm{m}}[\mu_0] - \mathcal{N}_{\mathrm{m}}[\mu]\|_\infty \leq \frac{\operatorname{diam}(\mu_0) + 2\epsilon}{c_0(c_0 - 2\epsilon)}\, 4\epsilon.$$

The uniform bounds for the uncertainty of the max-normalized R-CDT can be incorporated into the separation guarantee in Theorem 4.4 to obtain linear separability even in the cases of slight perturbations in W_∞.

Theorem 4.9. *For template measures $\mu_0, \nu_0 \in \mathcal{P}_c^*(\mathbb{R}^2)$ with $\mathcal{N}_\mathrm{m}[\mu_0] \neq \mathcal{N}_\mathrm{m}[\nu_0]$, define $c_\mu := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0])$ and $c_\nu := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\nu_0])$. Let $\epsilon < \frac{\min\{c_\mu, c_\nu\}}{2}$ satisfy*

$$4\epsilon\left(\frac{\operatorname{diam}(\mu_0)+2\epsilon}{c_\mu(c_\mu-2\epsilon)} + \frac{\operatorname{diam}(\nu_0)+2\epsilon}{c_\nu(c_\nu-2\epsilon)}\right) < \|\mathcal{N}_\mathrm{m}[\mu_0] - \mathcal{N}_\mathrm{m}[\nu_0]\|_\infty.$$

Consider the classes

$$\mathbb{F} = \big\{(\mathbf{A}\cdot+\mathbf{y})_\#\mu \mid \mathbf{A}\in\mathrm{GL}(2),\, \mathbf{y}\in\mathbb{R}^2,\, \mu\in\mathcal{P}(\mathbb{R}^2),\, W_\infty(\mu,\mu_0)\le\epsilon\big\},$$
$$\mathbb{G} = \big\{(\mathbf{A}\cdot+\mathbf{y})_\#\nu \mid \mathbf{A}\in\mathrm{GL}(2),\, \mathbf{y}\in\mathbb{R}^2,\, \nu\in\mathcal{P}(\mathbb{R}^2),\, W_\infty(\nu,\nu_0)\le\epsilon\big\}.$$

Then, any non-empty subsets $\mathbb{F}_0 \subseteq \mathbb{F}$ and $\mathbb{G}_0 \subseteq \mathbb{G}$ are linearly separable in ${}_m$NR-CDT space.

Proof. For $\mu\in\mathbb{F}$ there exist $\mathbf{A}\in\mathrm{GL}(2)$, $\mathbf{y}\in\mathbb{R}^2$ and $\mu_\epsilon\in\mathcal{P}(\mathbb{R}^2)$ with $W_\infty(\mu_\epsilon,\mu_0)\le\epsilon$ such that $\mu = (\mathbf{A}\cdot+\mathbf{y})_\#\mu_\epsilon$. Consequently, Corollary 4.8 yields

$$\|\mathcal{N}_\mathrm{m}[\mu_0] - \mathcal{N}_\mathrm{m}[\mu]\|_\infty \le \frac{\operatorname{diam}(\mu_0)+2\epsilon}{c_\mu(c_\mu-2\epsilon)}\,4\epsilon = C_{\mu,\epsilon}$$

so that $\mathcal{N}_\mathrm{m}\mathbb{F} \subseteq B_{C_{\mu,\epsilon}}(\mathcal{N}_\mathrm{m}[\mu_0]) \subset \mathrm{L}^\infty_\rho(\mathbb{R})$. Analogously, for $\nu\in\mathbb{G}$ we obtain

$$\|\mathcal{N}_\mathrm{m}[\nu_0] - \mathcal{N}_\mathrm{m}[\nu]\|_\infty \le \frac{\operatorname{diam}(\nu_0)+2\epsilon}{c_\nu(c_\nu-2\epsilon)}\,4\epsilon = C_{\nu,\epsilon}$$

and $\mathcal{N}_\mathrm{m}\mathbb{G} \subseteq B_{C_{\nu,\epsilon}}(\mathcal{N}_\mathrm{m}[\nu_0]) \subset \mathrm{L}^\infty_\rho(\mathbb{R})$. Since $\|\mathcal{N}_\mathrm{m}[\mu_0] - \mathcal{N}_\mathrm{m}[\nu_0]\|_\infty > C_{\mu,\epsilon} + C_{\nu,\epsilon}$, the closed balls $B_{C_{\mu,\epsilon}}(\mathcal{N}_\mathrm{m}[\mu_0])$ and $B_{C_{\nu,\epsilon}}(\mathcal{N}_\mathrm{m}[\nu_0])$ are linearly separable in $\mathrm{L}^\infty_\rho(\mathbb{R})$. This implies the linear separability of $\mathcal{N}_\mathrm{m}\mathbb{F}_0$ and $\mathcal{N}_\mathrm{m}\mathbb{G}_0$ in $\mathrm{L}^\infty_\rho(\mathbb{R})$ for any non-empty subsets $\mathbb{F}_0\subseteq\mathbb{F}$ and $\mathbb{G}_0\subseteq\mathbb{G}$. ∎

4.2. Mean-normalized R-CDT. To deal with a more general measure $\mu\in\mathcal{P}_2^*(\mathbb{R}^2)$ and perturbations in W_2, we define the *mean-normalized R-CDT* (aNR-CDT) $\mathcal{N}_\mathrm{a}[\mu]\colon\mathbb{R}\to\mathbb{R}$ via

$$\mathcal{N}_\mathrm{a}[\mu](t) = \int_{\mathbb{S}_1} \mathcal{N}[\mu](t,\boldsymbol{\theta})\,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \quad \text{for } t\in\mathbb{R}.$$

Proposition 4.3 implies the well-definedness and square-integrability of $\mathcal{N}_\mathrm{a}[\mu]$.

Lemma 4.10. *Let $\mu\in\mathcal{P}_2^*(\mathbb{R}^2)$. Then, $\mathcal{N}_\mathrm{a}[\mu]\in L^2_\rho(\mathbb{R})$.*

Proof. For $\mu\in\mathcal{P}_2^*(\mathbb{R}^2)$ we have $\mathcal{N}[\mu]\in\mathrm{L}^2_{\rho\times u_{\mathbb{S}_1}}(\mathbb{R}\times\mathbb{S}_1)$ according to Proposition 4.3. Consequently, Jensen's inequality gives

$$\|\mathcal{N}_\mathrm{a}[\mu]\|_\rho^2 \le \int_{\mathbb{R}}\int_{\mathbb{S}_1} |\mathcal{N}[\mu](t,\boldsymbol{\theta})|^2\,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) = \|\mathcal{N}[\mu]\|^2_{\rho\times u_{\mathbb{S}_1}} < \infty$$

so that $\mathcal{N}_\mathrm{a}[\mu]\in L^2_\rho(\mathbb{R})$, as stated. ∎

We again focus on the linear separability of classes generated by template measures.

Linear separability in ${}_a$NR-CDT space. The linear separability in ${}_a$NR-CDT space of classes in $\mathcal{P}_2^*(\mathbb{R}^2)$ requires more restrictions on the admissible affine-linear transforms.

Theorem 4.11. *For template measures $\mu_0, \nu_0 \in \mathcal{P}_2^*(\mathbb{R}^2)$ with*

$$\mathcal{N}_\mathrm{a}[\mu_0] \neq \mathcal{N}_\mathrm{a}[\nu_0]$$

consider the classes

$$\mathbb{F} = \big\{\mu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, \mathbf{A} \in \mathcal{X},\, \mathbf{y} \in \mathbb{R}^2 \colon \mu = (\mathbf{A}\cdot + \mathbf{y})_\# \mu_0\big\},$$
$$\mathbb{G} = \big\{\nu \in \mathcal{P}(\mathbb{R}^2) \mid \exists\, \mathbf{A} \in \mathcal{X},\, \mathbf{y} \in \mathbb{R}^2 \colon \nu = (\mathbf{A}\cdot + \mathbf{y})_\# \nu_0\big\},$$

where, for some $c \in (0, \frac{1}{2})$,

$$\mathcal{X} = \left\{ \mathbf{A} \in \mathrm{GL}(2) \;\middle|\; \frac{\sigma_{\max}(\mathbf{A}) - \sigma_{\min}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} \leq \frac{c\|\mathcal{N}_\mathrm{a}[\mu_0] - \mathcal{N}_\mathrm{a}[\nu_0]\|_\rho}{\max\{\|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}}, \|\mathcal{N}[\nu_0]\|_{\rho\times u_{\mathbb{S}_1}}\}} \right\}.$$

Then, any non-empty subsets $\mathbb{F}_0 \subseteq \mathbb{F}$ and $\mathbb{G}_0 \subseteq \mathbb{G}$ are linearly separable in ${}_a$NR-CDT space.

Proof. For $\mu \in \mathbb{F}$ there exist $\mathbf{A} \in \mathcal{X}$ and $\mathbf{y} \in \mathbb{R}^2$ with $\mu = (\mathbf{A}\cdot + \mathbf{y})_\# \mu_0$ implying that

$$\mathcal{N}[\mu](t, \boldsymbol{\theta}) = \mathcal{N}[\mu_0](t, h(\boldsymbol{\theta})) \quad \text{with} \quad h(\mathbf{x}) = \frac{\mathbf{A}^\top \mathbf{x}}{\|\mathbf{A}^\top \mathbf{x}\|}, \quad \mathbf{x} \in \mathbb{R}^2 \setminus \{0\}.$$

Setting $\boldsymbol{\theta}(\varphi) = (\cos(\varphi), \sin(\varphi))^\top \in \mathbb{S}_1$ for $\varphi \in [0, 2\pi)$, the definition of $\mathcal{N}_\mathrm{a}$ gives, for all $t \in \mathbb{R}$,

$$\mathcal{N}_\mathrm{a}[\mu](t) = \int_{\mathbb{S}_1} \mathcal{N}[\mu](t, \boldsymbol{\theta})\; \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_0^{2\pi} \mathcal{N}[\mu_0](t, h(\boldsymbol{\theta}(\varphi)))\; \mathrm{d}\varphi.$$

Now, the parametrization $\eta \colon [0, 2\pi) \to \mathbb{S}_1,\ \varphi \mapsto h(\boldsymbol{\theta}(\varphi))$ is bijective and continuously differentiable on $(0, 2\pi)$ with

$$\dot\eta(\varphi) = \mathrm{D}h(\boldsymbol{\theta}(\varphi))\dot{\boldsymbol{\theta}}(\varphi) = \frac{1}{\|\mathbf{A}^\top \boldsymbol{\theta}(\varphi)\|} \left(\mathbf{A}^\top \dot{\boldsymbol{\theta}}(\varphi) - \frac{\mathbf{A}^\top \boldsymbol{\theta}(\varphi) \boldsymbol{\theta}(\varphi)^\top \mathbf{A}\mathbf{A}^\top \dot{\boldsymbol{\theta}}(\varphi)}{\|\mathbf{A}^\top \boldsymbol{\theta}(\varphi)\|^2} \right)$$

so that

$$\|\dot\eta(\varphi)\|^2 = \frac{\|\mathbf{A}^\top \boldsymbol{\theta}(\varphi)\|^2 \, \|\mathbf{A}^\top \dot{\boldsymbol{\theta}}(\varphi)\|^2 - |\langle \mathbf{A}^\top \boldsymbol{\theta}(\varphi), \mathbf{A}^\top \dot{\boldsymbol{\theta}}(\varphi)\rangle|^2}{\|\mathbf{A}^\top \boldsymbol{\theta}(\varphi)\|^4}.$$

Consequently,

$$\mathcal{N}_\mathrm{a}[\mu_0](t) = \int_{\mathbb{S}_1} \mathcal{N}[\mu_0](t, \boldsymbol{\theta})\; \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_0^{2\pi} \mathcal{N}[\mu_0](t, \eta(\varphi))\, \|\dot\eta(\varphi)\|\; \mathrm{d}\varphi,$$

which implies that

$$\begin{aligned} \mathcal{N}_\mathrm{a}[\mu](t) - \mathcal{N}_\mathrm{a}[\mu_0](t) &= \frac{1}{2\pi} \int_0^{2\pi} \mathcal{N}[\mu_0](t, \eta(\varphi))\, (1 - \|\dot\eta(\varphi)\|)\; \mathrm{d}\varphi \\ &= \frac{1}{2\pi} \int_0^{2\pi} \mathcal{N}[\mu_0](t, \eta(\varphi))\, \|\dot\eta(\varphi)\| \left(\|\dot\eta(\varphi)\|^{-1} - 1\right)\; \mathrm{d}\varphi \end{aligned}$$

and, hence, with $c_{\mathbf{A}} = \max_{\varphi\in[0,2\pi)}|1 - \|\dot\eta(\varphi)\|^{-1}|$ follows that

$$\big|(\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0])(t)\big| \leq \frac{c_{\mathbf{A}}}{2\pi}\int_0^{2\pi} |\mathcal{N}[\mu_0](t,\eta(\varphi))|\,\|\dot\eta(\varphi)\|\,\mathrm{d}\varphi = c_{\mathbf{A}}\int_{\mathbb{S}_1}|\mathcal{N}[\mu_0](t,\boldsymbol{\theta})|\,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}).$$

Thereon, Hölder's inequality gives

$$\big|(\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0])(t)\big|^2 \leq c_{\mathbf{A}}^2\left(\int_{\mathbb{S}_1}|\mathcal{N}[\mu_0](t,\boldsymbol{\theta})|\,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta})\right)^2 \leq c_{\mathbf{A}}^2\int_{\mathbb{S}_1}|\mathcal{N}[\mu_0](t,\boldsymbol{\theta})|^2\,\mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta})$$

so that

$$\|\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0]\|_\rho \leq \Big(\max_{\varphi\in[0,2\pi)}\big|1 - \|\dot\eta(\varphi)\|^{-1}\big|\Big)\,\|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}}.$$

Direct calculations show that

$$\|\mathbf{A}^\top\boldsymbol{\theta}(\varphi)\|^2\,\|\mathbf{A}^\top\dot{\boldsymbol{\theta}}(\varphi)\|^2 - |\langle\mathbf{A}^\top\boldsymbol{\theta}(\varphi),\mathbf{A}^\top\dot{\boldsymbol{\theta}}(\varphi)\rangle|^2 = |\det(\mathbf{A})|^2$$

and, thus,

$$\|\dot\eta(\varphi)\| = \frac{|\det(\mathbf{A})|}{\|\mathbf{A}^\top\boldsymbol{\theta}(\varphi)\|^2}$$

with $|\det(\mathbf{A})| = \sigma_{\min}(\mathbf{A})\,\sigma_{\max}(\mathbf{A})$ and $\|\mathbf{A}^\top\boldsymbol{\theta}(\varphi)\|_2 \in [\sigma_{\min}(\mathbf{A}),\sigma_{\max}(\mathbf{A})]$ for all $\varphi\in[0,2\pi)$. This gives

$$\|\dot\eta(\varphi)\|^{-1} \in \left[\frac{\sigma_{\min}(\mathbf{A})}{\sigma_{\max}(\mathbf{A})},\frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}\right] \quad \forall\,\varphi\in[0,2\pi),$$

which in turn implies

$$\max_{\varphi\in[0,2\pi)}\big|1 - \|\dot\eta(\varphi)\|^{-1}\big| \leq \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} - 1 = \frac{\sigma_{\max}(\mathbf{A}) - \sigma_{\min}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

and the assumption $\mathbf{A}\in\mathcal{X}$ guarantees that

$$\begin{aligned}\|\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0]\|_\rho &\leq \frac{\sigma_{\max}(\mathbf{A}) - \sigma_{\min}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}\,\|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}}\\ &\leq c\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho =: r_0.\end{aligned} \tag{4.2}$$

Consequently, $\mathcal{N}_{\mathrm{a}}\mathbb{F} \subseteq B_{r_0}(\mathcal{N}_{\mathrm{a}}[\mu_0]) \subset \mathrm{L}^2_\rho(\mathbb{R})$ and, analogously, $\mathcal{N}_{\mathrm{a}}\mathbb{G} \subseteq B_{r_0}(\mathcal{N}_{\mathrm{a}}[\nu_0]) \subset \mathrm{L}^2_\rho(\mathbb{R})$. Since $c\in(0,\frac{1}{2})$, $B_{r_0}(\mathcal{N}_{\mathrm{a}}[\mu_0])$ and $B_{r_0}(\mathcal{N}_{\mathrm{a}}[\nu_0])$ are linearly separable in $\mathrm{L}^2_\rho(\mathbb{R})$. This implies the linear separability of $\mathcal{N}_{\mathrm{a}}\mathbb{F}_0$ and $\mathcal{N}_{\mathrm{a}}\mathbb{G}_0$ in $\mathrm{L}^2_\rho(\mathbb{R})$ for any non-empty $\mathbb{F}_0\subseteq\mathbb{F}$ and $\mathbb{G}_0\subseteq\mathbb{G}$. ■

In the next step, we again consider the linear separability of two generated classes when allowing for slight perturbations of the underlying template measures.

Linear separability under perturbations in Wasserstein space. To study the uncertainty of the mean-normalized R-CDT under perturbations with respect to the Wasserstein-2 distance, we again start with analysing how these effect the non-normalized R-CDT.

Proposition 4.12. *Let $\mu_0, \mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^d)$ with $W_2(\mu_0, \mu_\epsilon) \leq \epsilon$. Then*

$$\|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \leq \epsilon.$$

Proof. The statement follows from Proposition 3.3 via

$$\|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho = W_2(\mathcal{R}_{\boldsymbol{\theta}}[\mu_0], \mathcal{R}_{\boldsymbol{\theta}}[\mu_\epsilon]) \leq W_2(\mu_0, \mu_\epsilon) \leq \epsilon. \qquad \blacksquare$$

Next, we transfer the estimate for the R-CDT to the zero-mean quantile functions.

Lemma 4.13. *Let $\mu_0, \mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^2)$ with $W_2(\mu_0, \mu_\epsilon) \leq \epsilon$. Then*

$$\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \leq 2\epsilon.$$

Proof. Utilizing Proposition 4.12 and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho &\leq \|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho + |\mathrm{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0]) - \mathrm{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon])| \\
&\leq \|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho + \int_{\mathbb{R}} |\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0](t) - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon](t)| \, \mathrm{d}\rho(t) \\
&\leq 2\, \|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0] - \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \leq 2\epsilon. \qquad \blacksquare
\end{aligned}$$

We can now study the effect of perturbations in W_2 on the mean-normalized R-CDT.

Proposition 4.14. *Let $\mu_0 \in \mathcal{P}_2^*(\mathbb{R}^2)$ and $\mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^2)$ with $W_2(\mu_0, \mu_\epsilon) \leq \epsilon$, and define $c_0 \coloneqq \min_{\boldsymbol{\theta} \in \mathbb{S}_1} \mathrm{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}} \mu_0)$. If $\epsilon < c_0/2$, then*

$$\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\mu_\epsilon]\|_\rho \leq \frac{4\epsilon}{c_0}.$$

Proof. Equation (4.1) remains valid under the given assumptions. Since $2\epsilon < c_0$, the standard deviation $\mathrm{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_\epsilon]) = \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho$ is uniformly bounded away from zero such that $\mu_\epsilon \in \mathcal{P}_2^*(\mathbb{R}^2)$. In the view of Lemma 4.13, we obtain

$$\begin{aligned}
\|\mathcal{N}_{\boldsymbol{\theta}}[\mu_0] - \mathcal{N}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho &= \left\| \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} - \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho} \right\|_\rho \\
&\leq \left\| \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} - \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} \right\|_\rho + \left\| \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} - \frac{\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho} \right\|_\rho \\
&= \frac{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} + \frac{\big|\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho - \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho\big|}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho} \|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho \\
&\leq 2\, \frac{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0] - \widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_\epsilon]\|_\rho}{\|\widetilde{\mathcal{N}}_{\boldsymbol{\theta}}[\mu_0]\|_\rho} \leq \frac{4\epsilon}{c_0}. \qquad (4.3)
\end{aligned}$$

Employing Jensen's inequality, we may bound the perturbation after normalization by

$$\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\mu_\epsilon]\|_\rho^2 = \int_{\mathbb{R}} \left| \int_{\mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu_0](t) - \mathcal{N}_{\boldsymbol{\theta}}[\mu_\epsilon](t) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \right|^2 \mathrm{d}\rho(t) \leq \left(\frac{4\epsilon}{c_0}\right)^2. \qquad \blacksquare$$

The squared bounds for the uncertainty of the mean-normalized R-CDT can be incorporated into the separation guarantee in Theorem 4.11 to obtain linear separability even in the cases of slight perturbations in W_2.

Theorem 4.15. *For template measures $\mu_0, \nu_0 \in \mathcal{P}_2^*(\mathbb{R}^2)$ with $\mathcal{N}_{\mathrm{a}}\mu_0 \neq \mathcal{N}_{\mathrm{a}}\nu_0$, define $c_\mu := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathrm{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_0])$, $C_\mu := \|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}}$ and $c_\nu := \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathrm{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\nu_0])$, $C_\nu := \|\mathcal{N}[\nu_0]\|_{\rho\times u_{\mathbb{S}_1}}$. Let $c \in (0, \frac{1}{2})$, $c' \in (c, \frac{1}{2})$ and let $\epsilon < \min\{c_\mu, c_\nu\}/2$ satisfy*

$$\epsilon < \frac{c'-c}{4} \min\{c_\mu, c_\nu\} \frac{\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho \max\{C_\mu, C_\nu\}}{c\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho + \max\{C_\mu, C_\nu\}}.$$

Consider the classes

$$\mathbb{F} := \left\{(\mathbf{A}\cdot + \mathbf{y})_\# \mu \mid \mathbf{A} \in \mathcal{X},\, \mathbf{y} \in \mathbb{R}^2,\, \mu \in \mathcal{P}_2(\mathbb{R}^2),\, W_2(\mu, \mu_0) \leq \epsilon\right\},$$
$$\mathbb{G} := \left\{(\mathbf{A}\cdot + \mathbf{y})_\# \nu \mid \mathbf{A} \in \mathcal{X},\, \mathbf{y} \in \mathbb{R}^2,\, \nu \in \mathcal{P}_2(\mathbb{R}^2),\, W_2(\nu, \nu_0) \leq \epsilon\right\},$$

where

$$\mathcal{X} = \left\{ \mathbf{A} \in \mathrm{GL}(2) \;\middle|\; \frac{\sigma_{\max}(\mathbf{A}) - \sigma_{\min}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} \leq \frac{c}{\max\{C_\mu, C_\nu\}} \|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho \right\}$$

Then, any non-empty subsets $\mathbb{F}_0 \subseteq \mathbb{F}$ and $\mathbb{G}_0 \subseteq \mathbb{G}$ are linearly separable in ${}_a$NR-CDT space.

Proof. For $\mu \in \mathbb{F}$, there are $\mathbf{A} \in \mathcal{X}$, $\mathbf{y} \in \mathbb{R}^2$ and $\mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^2)$ with $W_2(\mu_\epsilon, \mu_0) \leq \epsilon$ such that $\mu = (\mathbf{A}\cdot + \mathbf{y})_\# \mu_\epsilon$. Consequently, the proof of Theorem 4.11 in combination with Proposition 4.14 gives

$$\begin{aligned}\|\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0]\|_\rho &\leq \|\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_\epsilon]\|_\rho + \|\mathcal{N}_{\mathrm{a}}[\mu_\epsilon] - \mathcal{N}_{\mathrm{a}}[\mu_0]\|_\rho \\ &\leq \frac{\sigma_{\max}(\mathbf{A}) - \sigma_{\min}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} \|\mathcal{N}[\mu_\epsilon]\|_{\rho\times u_{\mathbb{S}_1}} + \frac{4\epsilon}{c_\mu}.\end{aligned}$$

Utilizing (4.3), we obtain

$$\begin{aligned}\|\mathcal{N}[\mu_\epsilon]\|_{\rho\times u_{\mathbb{S}_1}} &\leq \|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}} + \|\mathcal{N}[\mu_\epsilon] - \mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}} \\ &\leq \|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}} + \left(\int_{\mathbb{S}_1}\int_{\mathbb{R}} |\mathcal{N}_{\boldsymbol{\theta}}[\mu_\epsilon](t) - \mathcal{N}_{\boldsymbol{\theta}}[\mu_0](t)|^2 \,\mathrm{d}\rho(t)\, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta})\right)^{\frac{1}{2}} \\ &\leq \|\mathcal{N}[\mu_0]\|_{\rho\times u_{\mathbb{S}_1}} + \frac{4\epsilon}{c_\mu}.\end{aligned}$$

Consequently, the assumption $\mathbf{A} \in \mathcal{X}$ and (4.2) give

$$\|\mathcal{N}_{\mathrm{a}}[\mu] - \mathcal{N}_{\mathrm{a}}[\mu_0]\|_\rho \leq c\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho + \left(\frac{c\|\mathcal{N}_{\mathrm{a}}[\mu_0] - \mathcal{N}_{\mathrm{a}}[\nu_0]\|_\rho}{\max\{C_\mu, C_\nu\}} + 1\right)\frac{4\epsilon}{c_\mu}$$

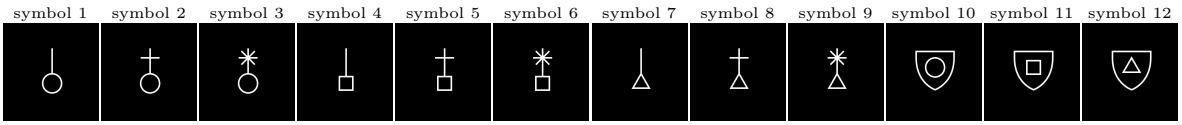


Figure 1. *Templates for academic datasets with domain* $[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^2$ *represented by* 256×256 *pixels.*

and the choice of $\epsilon > 0$ guarantees that

$$\|\mathcal{N}_\mathrm{a}[\mu] - \mathcal{N}_\mathrm{a}[\mu_0]\|_\rho < c' \, \|\mathcal{N}_\mathrm{a}[\mu_0] - \mathcal{N}_\mathrm{a}[\nu_0]\|_\rho =: r_0.$$

Consequently, $\mathcal{N}_\mathrm{a}\mathbb{F} \subseteq B_{r_0}(\mathcal{N}_\mathrm{a}[\mu_0]) \subset \mathrm{L}^2_\rho(\mathbb{R})$ and, analogously, $\mathcal{N}_\mathrm{a}\mathbb{G} \subseteq B_{r_0}(\mathcal{N}_\mathrm{a}[\nu_0]) \subset \mathrm{L}^2_\rho(\mathbb{R})$. Since $c' \in (c, \frac{1}{2})$, $B_{r_0}(\mathcal{N}_\mathrm{a}[\mu_0])$ and $B_{r_0}(\mathcal{N}_\mathrm{a}[\nu_0])$ are linearly separable in $\mathrm{L}^2_\rho(\mathbb{R})$. This implies the linear separability of $\mathcal{N}_\mathrm{a}\mathbb{F}_0$ and $\mathcal{N}_\mathrm{a}\mathbb{G}_0$ in $\mathrm{L}^2_\rho(\mathbb{R})$ for any non-empty $\mathbb{F}_0 \subseteq \mathbb{F}$ and $\mathbb{G}_0 \subseteq \mathbb{G}$. ■

5. Numerical experiments.

In order to support our theoretical findings with numerical evidence, we provide a series of proof-of-concept simulations, which focus on the (linear) separability under affine transformations established in Theorem 4.4 and 4.11 and on the influence of non-affine perturbations studied in Theorem 4.9 and 4.15. Since the original R-CDT [15] already outperforms other state-of-the-art classifiers in the small data regime [27], we restrict the experiments to comparisons with the R-CDT and the naïve Euclidean approach and omit comparisons with neural network classifiers. The new ${}_\mathrm{m}$NR-CDT and ${}_\mathrm{a}$NR-CDT methods significantly increase the classification accuracy showing their potential as feature extractor. The methods are implemented in Julia[2], and the code is publicly available at https://github.com/DrBeckmann/NR-CDT. All experiments are performed on an off-the-shelve MacBookPro 2020 with Intel Core i5 (4-Core CPU, 1.4 GHz) and 8 GB RAM.

5.1. Academic Datasets.

For the majority of the numerical simulations, we rely on academic datasets that allow us to fully control the occurring affine and non-affine perturbations. Starting from the synthetic symbols in Figure 1, whose domains are contained in the unit disc, we generate datasets with up to twelve classes by (anisotropic) scaling, rotating, shearing, and shifting the shown templates on the pixel grid using bi-quadratic interpolations. Scaling and shearing is here independently applied twice with respect to the horizontal and vertical direction. Depending on the experiment, we further apply non-affine transformations or add impulsive noise. Finally, the gray values are normalized to represent a (absolutely continuous) probability measure, where each pixel corresponds to a constant density on a square.

5.1.1. Classification under Affine Transformations.

The first numerical example deals with the ideal, unperturbed setting, where we aim to classify 10 affinely transformed versions of all twelve symbols. The theory behind ${}_\mathrm{m}$NR-CDT in Theorem 4.4 predicts that every class is transformed to a single point in ${}_\mathrm{m}$NR-CDT space. In order to observe this behaviour numerically, the underlying Radon transform and CDT have to be discretized fine enough. In this and all experiment regarding the academic datasets, we choose 850 equispaced radii in $[-1, 1]$ and 128 equispaced angles in $[0, 2\pi)$ for the Radon transform and 64 equispaced interpolation

[2]The Julia Programming Language – Version 1.9.2 (https://docs.julialang.org).

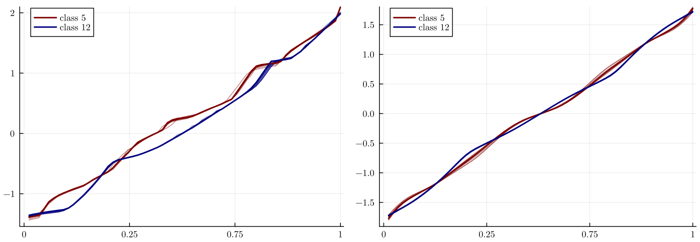


Figure 2. *$_{\mathrm{m}}$NR-CDT (left) and $_{\mathrm{a}}$NR-CDT (right) of affine classes 5 and 12. The classes are generated using horizontal/vertical scaling by a factor in $[0.5, 1.25]$, rotation by an angle in $[0^\circ, 360^\circ]$, shearing of horizontal/vertical axis by an angle in $[-45^\circ, 45^\circ]$, and horizontal/vertical shifting by a pixel number in $[-20, 20]$.*

Table 2

NT classification accuracies for academic datasets with 10 samples per class and different parameter ranges for random affine transformations. Similar to Figure 2, rotation angles are in $[0^\circ, 360^\circ]$, pixel shifts in $[-20, 20]$. The best result per dataset and angle number is highlighted.

angles	scaling in $[0.5, 1.25]$, shearing in $[-45^\circ, 45^\circ]$						scaling in $[0.75, 1.25]$, shearing in $[-35^\circ, 35^\circ]$					
	R-CDT		$_{\mathrm{m}}$NR-CDT		$_{\mathrm{a}}$NR-CDT		R-CDT		$_{\mathrm{m}}$NR-CDT		$_{\mathrm{a}}$NR-CDT	
	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$
1	0.1083	0.2333	0.1750	**0.3416**	0.1750	**0.3416**	0.1166	0.2500	0.1500	**0.3083**	0.1500	**0.3083**
2	0.1083	0.2333	0.1750	**0.3416**	0.3000	**0.3416**	0.1166	0.2500	0.1666	**0.3083**	0.2333	0.2916
4	0.1333	0.2500	0.2166	**0.5333**	0.2416	0.3583	0.1583	0.2250	0.2833	**0.5416**	0.2500	0.3833
8	0.0916	0.2000	0.4416	**0.6333**	0.4333	0.4750	0.1333	0.2083	0.3916	**0.6083**	0.4250	0.4583
16	0.1416	0.1916	0.6333	**0.8250**	0.7250	0.7750	0.1583	0.2250	0.6916	0.8583	0.8333	**0.8750**
32	0.1500	0.1916	0.8833	**0.9916**	0.8416	0.8750	0.2000	0.2333	0.9083	**1.0000**	0.9583	0.9416
64	0.1666	0.1916	0.9500	**1.0000**	0.8333	0.8583	0.2000	0.2250	0.9750	**1.0000**	0.9583	0.9416
128	0.1666	0.1916	**1.0000**	**1.0000**	0.8500	0.8583	0.2000	0.2250	**1.0000**	**1.0000**	0.9583	0.9416
256	0.1666	0.1916	0.9666	**1.0000**	0.8500	0.8666	0.1916	0.2166	**1.0000**	**1.0000**	0.9583	0.9416
Eucl.	0.0833	0.0916					0.0833	0.0833				

angles	scaling in $[0.75, 1.0]$, shearing in $[-15^\circ, 15^\circ]$						no scaling and shearing					
	R-CDT		$_{\mathrm{m}}$NR-CDT		$_{\mathrm{a}}$NR-CDT		R-CDT		$_{\mathrm{m}}$NR-CDT		$_{\mathrm{a}}$NR-CDT	
	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lVert\cdot\rVert_2$
1	0.1666	0.2666	0.1916	**0.3083**	0.1916	**0.3083**	0.2000	0.2583	0.1583	**0.3166**	0.1583	**0.3166**
2	0.1666	0.2666	0.1916	**0.3083**	0.2166	0.3000	0.2000	0.2583	0.1500	**0.3333**	0.2500	0.3083
4	0.2250	0.2083	0.3500	**0.5666**	0.1916	0.3833	0.2166	0.2083	0.3250	**0.5333**	0.1833	0.3583
8	0.1750	0.2250	0.4166	**0.6250**	0.4333	0.5000	0.2166	0.2250	0.4583	**0.6500**	0.4833	0.4750
16	0.1666	0.2250	0.7916	**0.9083**	0.8583	**0.9083**	0.2500	0.2250	0.7833	**0.9166**	0.9250	0.9000
32	0.1750	0.2083	0.9500	**1.0000**	**1.0000**	**1.0000**	0.2500	0.1266	0.9500	**1.0000**	**1.0000**	**1.0000**
64	0.1666	0.2083	0.9750	**1.0000**	**1.0000**	**1.0000**	0.2416	0.2083	0.9500	**1.0000**	**1.0000**	**1.0000**
128	0.1666	0.2083	**1.0000**	**1.0000**	**1.0000**	0.9916	0.2250	0.2083	**1.0000**	**1.0000**	**1.0000**	**1.0000**
256	0.1833	0.2083	0.9750	**1.0000**	**1.0000**	0.9833	0.2166	0.2083	0.9583	0.9916	**1.0000**	**1.0000**
Eucl.	0.0833	0.0666					0.0833	0.0666				

points in $(0, 1)$ for the CDT. For illustration, Figure 2 shows the $_{\mathrm{m}}$NR-CDT for the affine classes with respect to templates 5 and 12. The remaining numerical errors originate from bi-quadratic interpolations underlying the affine image transformations. Figure 2 also shows the $_{\mathrm{a}}$NR-CDT of both classes. Note that the $_{\mathrm{a}}$NR-CDT does not transform an affine class to a single point but to a small ball around the template whose radius depends on the eigenvalues of the affine transformations, see Theorem 4.11. The quality of both transformations is comparable but visually the $_{\mathrm{m}}$NR-CDT yields a larger distance between the classes.

To classify a given datum, we assign the label of the closest template in the considered feature spaces. Henceforth, we refer to this approach as *nearest template* (NT) *classification*. Since the quality of the $_{\mathrm{a}}$NR-CDT mainly depends of the size of the anisotropic scaling and shearing, we repeat the experiment for different parameter ranges. Our classification results are reported in Table 2, where we compare the NT performance of the $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT with the R-CDT representation from [15] and the Euclidean representation as baseline. In feature space, we use the Euclidean $\lVert\cdot\rVert_2$ and Chebyshev $\lVert\cdot\rVert_\infty$ norm to assign the labels. Moreover, we vary the number of equispaced angles for the underlying Radon transform. The $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT feature representations clearly outperform the R-CDT and the Euclidean baseline for classification under affine transformations. The results of the $_{\mathrm{m}}$NR-CDT surpass the accuracies of the $_{\mathrm{a}}$NR-CDT especially in the presence of large anisotropic scaling and shearing, which is covered by the developed theory. Notice that already small numbers of Radon angles yield high accuracies. Finally, the Euclidean norm outperforms the Chebyshev norm; for this reason, we restrict ourselves henceforth to the Euclidean norm.

5.1.2. Classification under Non-affine Deformations. The theory behind Theorem 4.9 and 4.15 guarantees the separability of affine classes in $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT space even for imperfect affine transformations. In the next experiments, we study the robustness of the proposed methods against non-affine deformations and additive impulsive noise. Both error

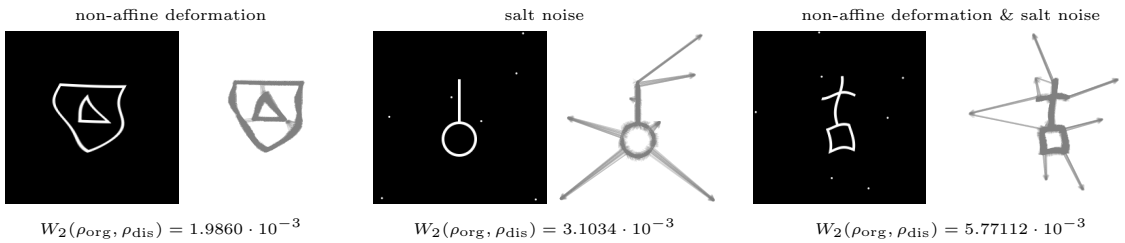


Figure 3. *Instances of corrupted data regarding non-affine deformations and impulsive/salt noise considered in the robustness analysis. The accompanying vector fields illustrate the optimal Wasserstein-2 transport between the corrupted datum ρ_{dis} and the true template ρ_{org}.*

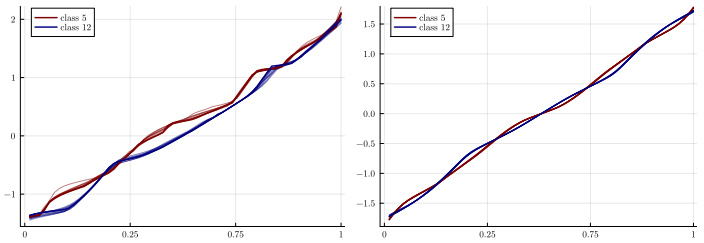


Figure 4. *Visualization of $_m$NR-CDT (left) and $_a$NR-CDT (right) for classes 5 and 12 of academic dataset, each of size 10 and generated by, first, random non-affine deformations induced by sine/cosine functions with amplitudes in $[2.5, 7.5]$ and frequencies in $[0.5, 2.0]$, and, second, random affine transformation as in Figure 2.*

Table 3

NT classification accuracies for the academic dataset with 10 samples per class and different parameter ranges for random non-affine distortions. Additionally, random affine transformations are applied with scaling in $[0.75, 1.0]$, shearing in $[-5°, 5°]$, rotation angles in $[0°, 360°]$ and pixel shifts in $[-20, 20]$. The best result per dataset and angle is highlighted.

angles	no non-affine distortion			freq. in [0.5, 2.0], amp. in [2.5, 7.5]			freq. in [0.5, 2.0], amp. in [8.0, 13.0]			freq. in [0.5, 4.0], amp. in [0.5, 2.0]			freq. in [0.5, 4.0], amp. in [0.5, 7.5]			freq. in [0.5, 4.0], amp. in [2.5, 7.5]		
	R-CDT	$_m$NR-CDT	$_a$NR-CDT	R-CDT	$_m$NR-CDT	$_a$NR-CDT	R-CDT	$_m$NR-CDT	$_a$NR-CDT	R-CDT	$_m$NR-CDT	$_a$NR-CDT	R-CDT	$_m$NR-CDT	$_a$NR-CDT	R-CDT	$_m$NR-CDT	$_a$NR-CDT
1	0.2583	**0.3083**	**0.3083**	0.2250	**0.2666**	**0.2666**	0.2250	**0.2500**	**0.2500**	0.2416	**0.2666**	**0.2666**	0.2166	**0.2250**	**0.2250**	0.2000	**0.2166**	**0.2166**
2	0.2883	**0.3083**	0.2916	0.2250	**0.2750**	**0.2750**	0.2250	**0.2416**	0.2333	0.2416	**0.2583**	0.2500	0.2166	0.2166	**0.2333**	0.2000	**0.2250**	0.2166
4	0.1833	**0.5250**	0.3666	0.1750	**0.4333**	0.2500	0.2083	**0.4000**	0.2083	0.1916	**0.4583**	0.3166	0.1666	**0.3916**	0.2333	0.1750	**0.3750**	0.2083
8	0.2083	**0.6500**	0.4916	0.1583	**0.6000**	0.4166	0.1333	**0.5583**	0.4333	0.1833	**0.6500**	0.5166	0.1966	**0.5833**	0.4416	0.1916	**0.5833**	0.4250
16	0.2000	**0.9166**	0.8750	0.1583	0.8250	**0.8666**	0.1250	0.7083	**0.7500**	0.1583	**0.9083**	0.8833	0.1916	0.7916	**0.8083**	0.2083	**0.7583**	0.7250
32	0.2000	**1.0000**	**1.0000**	0.1666	0.9916	**1.0000**	0.1250	0.8083	**0.9083**	0.1666	**1.0000**	**1.0000**	0.1916	0.9500	**0.9666**	0.2000	0.9083	**0.9250**
64	0.1916	**1.0000**	**1.0000**	0.1666	0.9833	**1.0000**	0.1250	0.8000	**0.9166**	0.1583	**1.0000**	**1.0000**	0.1916	0.9333	**0.9666**	0.2000	0.9083	**0.9166**
128	0.1916	**1.0000**	**1.0000**	0.1666	0.9916	**1.0000**	0.1250	0.8166	**0.9250**	0.1666	**1.0000**	**1.0000**	0.1916	0.9333	**0.9750**	0.2000	0.9000	**0.9250**
256	0.1916	0.9916	**1.0000**	0.1666	0.9916	**1.0000**	0.1580	0.8416	**1.0000**	0.1666	**1.0000**	**1.0000**	0.1916	0.9416	**0.9750**	0.2000	0.9000	**0.9250**
Eucl.	0.0833			0.0583			0.0666			0.0750			0.0750			0.0750		

sources are illustrated in Figure 3 together with optimal transport plans from the underlying true template. Already light impulsive noise, which is referred to as salt noise, has a similar effect on the Wasserstein-2 distance as strong non-affine perturbations. Therefore, we expect that $_m$NR-CDT and $_a$NR-CDT can manage non-affine distortions better than impulsive noise.

Non-affine Deformations. To generate non-affine deformations of an $(N \times N)$-pixel image, we assign to the (j, k)th pixel the bi-quadratically interpolated gray value at the perturbed location

$$\big(j + a_1 \sin\big(\tfrac{2\pi f_1}{N}\, k\big), k + a_2 \cos\big(\tfrac{2\pi f_2}{N}\, j\big)\big)$$

with fixed random frequencies f_1, f_2 and amplitudes a_1, a_2. Figuratively, this deformation yields local bendings of the template symbols; see Figure 3 (left). The datasets for the following experiments consist of 10 non-affinely deformed samples for each of the twelve templates in Figure 1, followed by the application of a random affine transformation. The impact on the

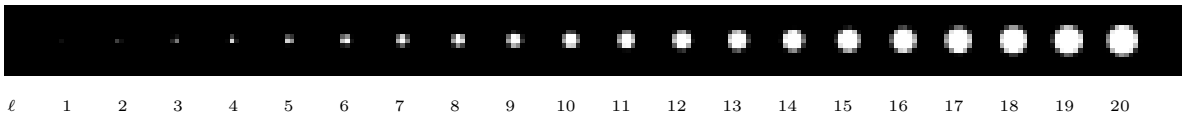
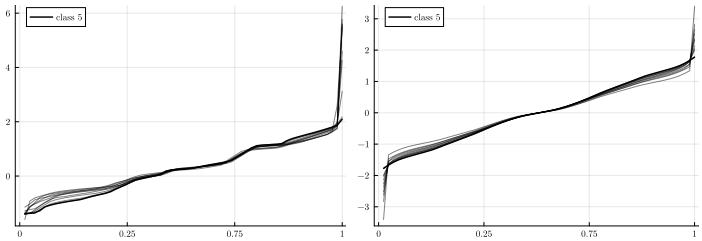


Figure 5. *Scala of exemplarily rendered salt noise of strength $\ell \in \{1, \dots, 20\}$ used in the experiments. The salt noise illustrated in Figure 3 has strength $\ell = 9$.*

Figure 6. *Visualization of $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT for class 5 of academic dataset of size 10, generated by, first, random affine transformations with scaling in $[0.75, 1.0]$, shear in $[-5°, 5°]$, rotation in $[0°, 360°]$, shift in $[-20, 20]$ and, second, adding salt noise of strength 9 according to Figure 5 at 4 to 7 locations.*



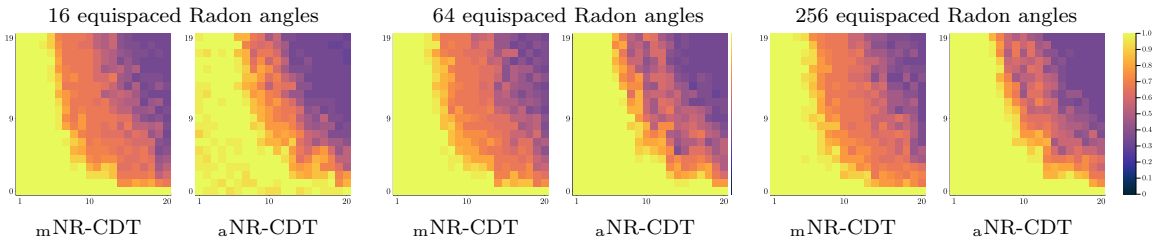


Figure 7. *Phase transition of NT classification accuracies for 3-class academic datasets with 10 samples per class and different salt noise (vertical: component numbers, horizontal: noise strengths, cf. Figure 5). The random affine transformations consist of rotations with angle in $[0°, 360°]$ and pixel shifts in $[-20, 20]$. The experiment is repeated for varying numbers of Radon angles.*

$_{\mathrm{m}}$NR-CDTs and $_{\mathrm{a}}$NR-CDTs is illustrated in Figure 4. While the non-affine deformations only have a very small impact on the $_{\mathrm{a}}$NR-CDT representation, we clearly observe within-class variations for the $_{\mathrm{m}}$NR-CDT. For classification, we again use the NT approach and repeat the entire experiment for different academic datasets whose amplitudes and frequencies of the random non-affine deformations lie in varying parameter ranges. Our results are reported in Table 3. We observe that our $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT representations clearly outperform R-CDT and Euclidean representations. In particular, for a small amount of non-affine distortions both—$_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT—yield perfect classification for sufficiently many angles. With increasing amount of distortions, we see that $_{\mathrm{a}}$NR-CDT performs better than $_{\mathrm{m}}$NR-CDT, as expected by our observations based on Figure 3.

Salt Noise. In style of salt-and-pepper noise, we use the term *salt noise* for image distortions caused by adding white discs with fixed radius to the image. Figure 5 shows a gamut of rendered salt noise used in our experiments. The impact of salt noise to the $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT is illustrated in Figure 6 and mainly consists in heavy disturbances at the end points of the domain $(0, 1)$, which also affect the $_{\mathrm{m}}$NR-CDTs and $_{\mathrm{a}}$NR-CDTs as a whole. The 3-class academic datasets in this experiment are generated as follows: first, the template images 1, 5, and 12 in Figure 1 are randomly affinely transformed (without anisotropic scaling and shearing to avoid additional disturbances of $_{\mathrm{a}}$NR-CDTs); second, the obtained images are

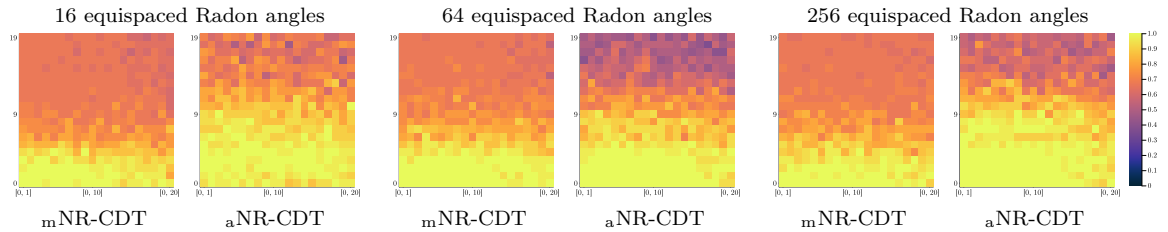


Figure 8. *Phase transition of NT classification accuracies for 3-class academic datasets with 10 samples per class, non-affine deformations with frequencies in* $[0, 2]$ *and variable amplitude ranges (horizontal), as well as salt noise of strength 9 according to Figure 5 and variable location numbers (vertical). The random affine transformations consist of rotations with angles in* $[0°, 360°]$ *and pixel shifts in* $[-20, 20]$*. The experiment is repeated for varying numbers of Radon angles.*

Table 4

NN classification accuracies (mean plus/minus standard deviation) for academic datasets with all 12 symbols from Figure 1 and 100 samples per class. In 2. and 4., random non-affine deformations with frequencies in $[0.5, 2.0]$ *and amplitudes in* $[2.5, 7.5]$ *are applied. Affine transformations consist of rotations in* $[0°, 360°]$*, shifts in* $[-20, 20]$ *and, in 1. and 2., scaling in* $[0.5, 1.25]$ *and shearing in* $[-45°, 45°]$ *as well as, in 3. and 4., scaling in* $[0.75, 1.0]$ *and shearing in* $[-5°, 5°]$*. In 3. and 4., salt noise is of strength 9, cf. Figure 5, with random location numbers in* $[4, 7]$*. The best result per dataset and training number is highlighted.*

dataset	5 training samples				10 training samples			
	Euclidean	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{\mathrm{a}}$NR-CDT	Euclidean	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{\mathrm{a}}$NR-CDT
1. affine	0.0915 ± 0.0052	0.1217 ± 0.0094	$\mathbf{0.9999 \pm 0.0001}$	0.9010 ± 0.0247	0.0958 ± 0.0083	0.1320 ± 0.0091	$\mathbf{1.0000 \pm 0.0000}$	0.9595 ± 0.0122
2. non-affine, affine	0.0903 ± 0.0081	0.1212 ± 0.0110	$\mathbf{0.9899 \pm 0.0048}$	0.9055 ± 0.0264	0.0922 ± 0.0080	0.1343 ± 0.0111	$\mathbf{0.9962 \pm 0.0031}$	0.9623 ± 0.0128
3. affine, salt	0.1003 ± 0.0072	0.1707 ± 0.0111	0.5669 ± 0.0226	$\mathbf{0.6378 \pm 0.0300}$	0.1103 ± 0.0085	0.2003 ± 0.0085	0.6542 ± 0.0209	$\mathbf{0.7236 \pm 0.0157}$
4. non-affine, affine, salt	0.1042 ± 0.0075	0.1666 ± 0.0132	0.5412 ± 0.0294	$\mathbf{0.6325 \pm 0.0271}$	0.1108 ± 0.0068	0.1904 ± 0.0080	0.6296 ± 0.0223	$\mathbf{0.7149 \pm 0.0223}$

corrupted by salt noise. The final datasets consist of 10 samples per class and are classified using the NT approach. Repeating the experiment for datasets with different noise strength and component numbers, we obtain the phase transitions in Figure 7. The $_{\mathrm{a}}$NR-CDT requires more angles but then performs better than $_{\mathrm{m}}$NR-CDT, whose phase transition is less sharp.

Non-affine Deformations and Salt Noise. In the final academic experiment, we study the combined impact of non-affine deformations and salt noise. Similar to before, we consider 3-class academic datasets based on the symbols 1, 5, and 12 from Figure 1. The datasets themselves, each consisting of 10 samples per class, are generated by, first, distorting via a non-affine deformation, second, applying an affine transformation (again without anisotropic scaling and shearing), and third, corrupting with salt noise. For classification, we employ the NT approach. The resulting phase transitions are reported in Figure 8, where the range of the random amplitudes of the non-affine deformation and the number of locations corrupted by salt noise are varied. We observe that the salt noise has more impact on the classification success and that the area of near perfect classification is larger for $_{\mathrm{a}}$NR-CDT than $_{\mathrm{m}}$NR-CDT, indicating more robustness against distortions.

5.1.3. Nearest Neighbour Classification. In the final experiments, regarding academic datasets, the NT method is replaced by the nearest neighbour (NN) method to demonstrate that the classification accuracy can be improved, especially under presence of non-affine distortions and noise. The parameter regarding the discretized Radon transform and CDT remain unchanged, i.e., we use 128 Radon angles, 850 radii, and 64 interpolation points. The gener-

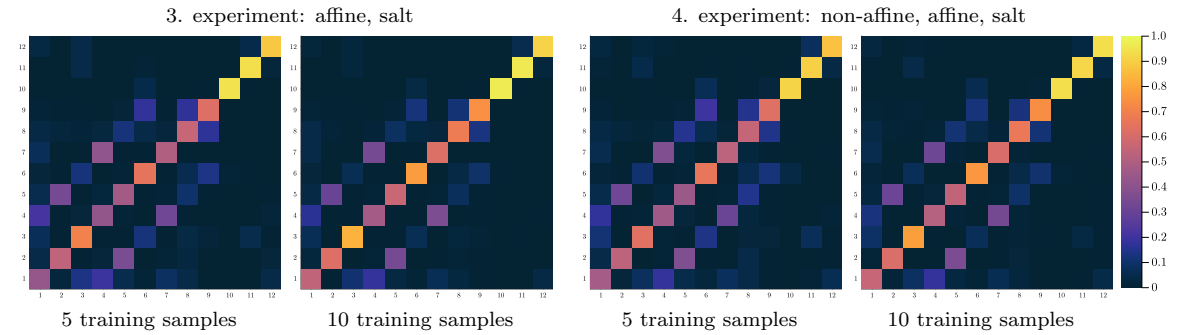


Figure 9. *Confusion maps for $_{a}$NR-CDT in the 3. and 4. experiment in Table 4. True labels are vertical and the classified labels horizontal.*

Figure 10. *Chinese character dataset. One member of each class is selected as template symbol and affinely transformed to create the dataset. Here, the templates of classes 1 to 12 are shown, each of size* 128×128 *pixels.*

ated datasets consists of 100 samples per each of the 12 classes in Figure 1. Furthermore, the employed NN classifiers rely on up to 10 random samples per class. In Table 4, the mean NN accuracies for different distortion settings are recorded, where each experiment is repeated 20 times with different training samples. In the ideal setting and under non-affine deformations, rows 1 and 2, $_{m}$NR-CDT yields perfect classifications, whereas $_{a}$NR-CDT performs slightly worse. The parameter ranges of the anisotropic scaling and shearing in these experiments are relatively large, such that the decrease of the performance of $_{a}$NR-CDT is expected. Under the presence of salt noise, rows 3 and 4, $_{a}$NR-CDT significantly outperforms $_{m}$NR-CDT. Considering the confusion maps of these experiments in Figure 9, we notice that $_{a}$NR-CDT can clearly distinguish the shield symbols and the crosses at the top but has problems to classify the basis (circle, square, triangle) of the symbols.

5.2. Semi-synthetic Chinese Character Dataset. To demonstrate the applicability of our approach to multi-class problems with a large number of classes, we consider the first 1000 classes of the Chinese character dataset [7]. For each of these, we select the first representative as template, see Figure 10 for the first 12 classes, which is then randomly scaled, rotated, sheared and shifted to form our semi-synthetic Chinese character datasets.

To start with, we restrict ourselves to the leading 100 classes with 50 samples per class and compare $_{m}$NR-CDT and $_{a}$NR-CDT with the Euclidean and R-CDT representations. To this end, we again use the NT approach as well as the NN method with 5 or 10 training samples, respectively. For the discretization of the Radon transform and CDT we use 128 Radon angles, 850 radii, and 64 interpolation points. The classification accuracies are reported in Table 5 (top). For both NT and NN, we see that $_{m}$NR-CDT and $_{a}$NR-CDT clearly outperform the Euclidean and R-CDT approaches, which only perform at the level of random guessing. As expected by our theory, we observe perfect NT and NN classification using $_{m}$NR-CDT with sufficiently many angles. For $_{a}$NR-CDT, the performance of the NT classification is worse,

Table 5

NT and NN classification accuracies for the Chinese character dataset, where each class consists of 50 samples, generated by random affine transformations with scaling in $[0.5, 1.0]$*, shearing in* $[-25°, 25°]$*, rotation angles in* $[0°, 360°]$ *and pixel shifts in* $[-20, 20]$*. Separately for NT and NN, the best result is highlighted.*

# classes	method	NT # angles 2	4	8	16	32	64	128	NN # training samples 5	10
100	Euclidean	0.0096							0.0202 ± 0.0021	0.0242 ± 0.0022
	R-CDT	0.0160	0.0180	0.0172	0.0164	0.0168	0.0170	0.0170	0.0188 ± 0.0020	0.0217 ± 0.0021
	$_{\mathrm{m}}$NR-CDT	**0.0924**	**0.1038**	**0.3124**	**0.8422**	**0.9836**	**1.0000**	**1.0000**	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$
	$_{\mathrm{a}}$NR-CDT	0.0810	0.0172	0.1860	0.4950	0.6486	0.6814	0.6852	0.8345 ± 0.0097	0.9139 ± 0.0062
1000	$_{\mathrm{m}}$NR-CDT	**0.0848**	**0.0620**	**0.2178**	**0.7368**	**0.9791**	**0.9975**	**0.9981**	$\mathbf{0.9987 \pm 0.0001}$	$\mathbf{0.9987 \pm 0.0001}$
	$_{\mathrm{a}}$NR-CDT	0.0629	0.0362	0.0751	0.3475	0.5554	0.6030	0.6104	0.7651 ± 0.0026	0.8631 ± 0.0018

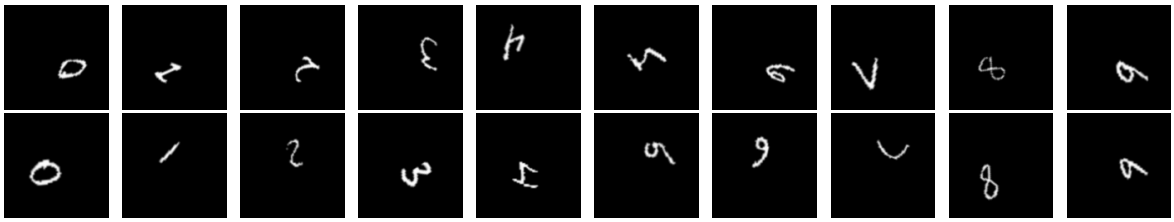


Figure 11. *Two samples of each class (zero to nine) of the LinMNIST dataset, generated by randomly rotating, anisotropically scaling and shifting a random MNIST sample of the corresponding class. Here, rotation angles are in* $[0°, 360°]$*, scalings in* $[0.75, 1.0]$ *and shifts in* $[-20, 20]$*. The image resolution is* 128×128 *pixels.*

which is also expected due to the application of rather severe scaling and shearing. However, the classification accuracy is significantly improved by NN and we achieve nearly perfect results with increasing number of training samples.

Since the Euclidean and R-CDT approach cannot successfully classify the first 100 classes, we only consider our $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT representations when dealing with 1000 Chinese characters. The results are shown in Table 5 (bottom). Again, $_{\mathrm{m}}$NR-CDT yields nearly perfect results, while the performance of $_{\mathrm{a}}$NR-CDT is improved by NN with increasing number of training samples. Hence, all in all, our numerical observations reflect our theoretical results also in the challenging case of a tremendous number of classes.

5.3. LinMNIST Dataset. For a more realistic scenario, we finally consider the LinMNIST dataset [4] consisting of affinely transformed MNIST digits [10], cf. Figure 11. More precisely, this dataset is generated by selecting the first 500 samples of each of the ten MNIST classes and, thereon, applying a random affine transformation. In this way, we combine our theoretically inspired setting of affinely transformed classes with the variety in real-world datasets. To account for this, we change the NT and NN approach to k-NN classification and vary the value of k as well as the number of training samples. Again, we compare $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT with the Euclidean and R-CDT representations. For the discretization of the Radon transform and CDT we use 128 Radon angles, 300 radii, and 64 interpolation points.

The classification accuracies are reported in Table 6. We observe that $_{\mathrm{m}}$NR-CDT clearly outperforms the other approaches followed by $_{\mathrm{a}}$NR-CDT and reaches a k-NN classification ac-

Table 6

k-NN classification accuracies (mean plus/minus standard deviation) for the LinMNIST dataset with class size 500 *generated by anisotropic scaling in* $[0.75, 1.0]$*, rotation angles in* $[0°, 360°]$ *and pixel shifts in* $[-20, 20]$*. For each number of training samples, the best result is highlighted.*

# training samples	k	Euclidean	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{\mathrm{a}}$NR-CDT
11	1	0.1222 ± 0.0078	0.1279 ± 0.0067	0.5541 ± 0.0134	0.3899 ± 0.0140
	5	0.1168 ± 0.0100	0.1105 ± 0.0037	$\mathbf{0.5591 \pm 0.0148}$	0.4005 ± 0.0135
	11	0.1135 ± 0.0097	0.1053 ± 0.0038	0.5445 ± 0.0184	0.4015 ± 0.0131
25	1	0.1387 ± 0.0067	0.1434 ± 0.0067	0.6010 ± 0.0122	0.4208 ± 0.0098
	5	0.1278 ± 0.0113	0.1260 ± 0.0060	$\mathbf{0.6147 \pm 0.0094}$	0.4402 ± 0.0111
	11	0.1229 ± 0.0121	0.1156 ± 0.0047	0.6132 ± 0.0100	0.4453 ± 0.0107
50	1	0.1597 ± 0.0052	0.1619 ± 0.0056	0.6283 ± 0.0083	0.4467 ± 0.0076
	5	0.1416 ± 0.0069	0.1457 ± 0.0047	0.6524 ± 0.0084	0.4722 ± 0.0067
	11	0.1318 ± 0.0089	0.1336 ± 0.0054	$\mathbf{0.6524 \pm 0.0075}$	0.4776 ± 0.0089

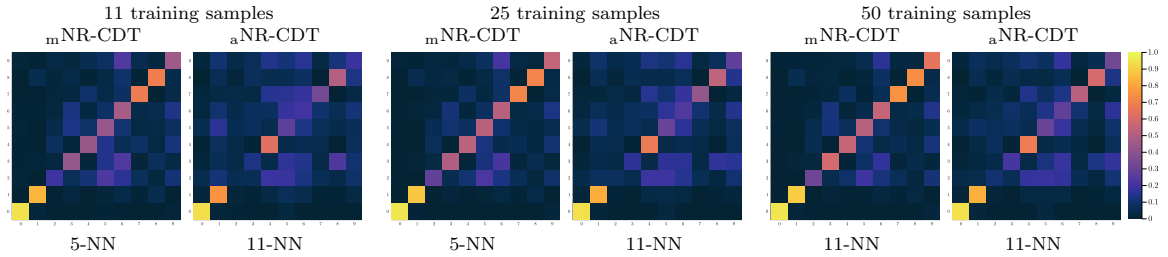


Figure 12. *Confusion maps for the best* $_{\mathrm{m}}$*NR-CDT and* $_{\mathrm{a}}$*NR-CDT results per number of training samples.*

curacy of 65% when using $k = 11$ and 50 training samples. In contrast to this, the Euclidean and R-CDT representation perform on the level of random guessing. Inspecting the confusion maps in Figure 12 reveals that both $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT nearly perfectly classify classes 0 and 1. The 4s are better classified in $_{\mathrm{a}}$NR-CDT space. For all other numbers, the representation in $_{\mathrm{m}}$NR-CDT space leads to better classifications.

6. Conclusion. In this paper, we continued our study of the $_{\mathrm{m}}$NR-CDT introduced in [3] to enhance separability by analysing its robustness with respect to non-affine perturbations. In addition, we introduced the $_{\mathrm{a}}$NR-CDT, which shows an improved numerical performance especially in the presence of impulsive noise. In future works, we wish to design refined approaches for handling more severe and realistic noise models, specifically for our motivating pattern recognition task in filigranology. In so doing, we aim to surmount the gap between mathematical theory and practice. In particular, our $_{\mathrm{m}}$NR-CDT and $_{\mathrm{a}}$NR-CDT feature representations are to be used in real-world applications like a fully automated watermark recognition and classification pipeline. Moreover, we wish to improve our estimates especially regarding the $_{\mathrm{m}}$NR-CDT as this appears to perform better than suggested by our theoretical findings. In our proof-of-concept experiments, we relied on NN classifiers to show the superior performance of the proposed feature extractors in comparison with the Euclidean and R-CDT approach. As in [27], the NN classifier may be replaced by more advanced classification meth-

ods to improve the shown results. Finally, besides classification tasks, we want to study the impact of our feature extractors on clustering problems.

REFERENCES

[1] A. Aldroubi, R. Diaz Martin, I. V. Medri, G. K. Rohde, and S. Thareja, *The signed cumulative distribution transform for 1-d signal analysis and classification*, Foundations of Data Science, 4 (2022), pp. 137–163, https://doi.org/10.3934/fods.2022001.

[2] A. Aldroubi, S. Li, and G. K. Rohde, *Partitioning signal classes using transport transforms for data analysis and machine learning*, Sampling Theory, Signal Processing, and Data Analysis, 19 (2021), p. 6, https://doi.org/10.1007/s43670-021-00009-z.

[3] M. Beckmann, R. Beinert, and J. Bresch, *Max-normalized radon cumulative distribution transform for limited data classification*, in Scale Space and Variational Methods in Computer Vision (SSVM), vol. 15667 of Lecture Notes in Computer Science, Springer, 2025, pp. 241–254, https://doi.org/10.1007/978-3-031-92366-1_19.

[4] M. Beckmann and N. Heilenkötter, *Equivariant neural networks for indirect measurements*, SIAM Journal on Mathematics of Data Science, 6 (2024), pp. 579–601, https://doi.org/10.1137/23M1582862.

[5] F. Beier, R. Beinert, and G. Steidl, *On a linear Gromov–Wasserstein distance*, IEEE Transactions on Image Processing, 31 (2022), pp. 7292–7305, https://doi.org/10.1109/TIP.2022.3221286.

[6] R. Beinert, C. Heiss, and G. Steidl, *On assignment problems related to Gromov–Wasserstein distances on the real line*, SIAM Journal on Imaging Sciences, 16 (2023), pp. 1028–1032, https://doi.org/10.1137/22M1497808.

[7] P. Bliem, *Handwritten chinese character hanzi datasets*, 2022, https://www.kaggle.com/datasets/pascalbliem/handwritten-chinese-character-hanzi-datasets. Accessed: March 26, 2025.

[8] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, *Sliced and Radon Wasserstein barycenters of measures*, Journal of Mathematical Imaging and Vision, 51 (2015), pp. 22–45, https://doi.org/10.1007/s10851-014-0506-3.

[9] A. Cloninger, K. Hamm, V. Khurana, and C. Moosmüller, *Linearized Wasserstein dimensionality reduction with approximation guarantees*, Applied and Computational Harmonic Analysis, 74 (2025), p. 101718, https://doi.org/10.1016/j.acha.2024.101718.

[10] L. Deng, *The MNIST database of handwritten digit images for machine learning research*, IEEE Signal Processing Magazine, 29 (2012), pp. 141–142, https://doi.org/10.1109/MSP.2012.2211477.

[11] R. Díaz Martín, I. V. Medri, and G. K. Rohde, *Data representation with optimal transport*, 2024, https://doi.org/10.48550/arXiv.2406.15503. arXiv:2406.15503.

[12] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Mathematical Journal, 31 (1984), pp. 231–240, https://doi.org/10.1307/mmj/1029003026.

[13] D. Hauser, M. Beckmann, G. Koliander, and H. S. Stiehl, *On image processing and pattern recognition for thermograms of watermarks in manuscripts – a first proof-of-concept*, in International Conference on Document Analysis and Recognition (ICDAR), 2024, pp. 91–107, https://doi.org/10.1007/978-3-031-70543-4_6.

[14] S. Helgason, *The Radon Transform*, Birkhäuser, 2 ed., 1999.

[15] S. Kolouri, S. R. Park, and G. K. Rohde, *The Radon cumulative distribution transform and its application to image classification*, IEEE Transactions on Image Processing, 25 (2016), pp. 920–934, https://doi.org/10.1109/TIP.2015.2509419.

[16] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, *Optimal mass transport*, IEEE Signal Processing Magazine, 34 (2017), pp. 43–59, https://doi.org/10.1109/MSP.2017.2695801.

[17] X. Liu, R. Díaz Martín, Y. Bai, A. Shahbazi, M. Thorpe, A. Aldroubi, and S. Kolouri, *Expected sliced transport plans*, 2024, https://doi.org/10.48550/arXiv.2410.12176. arXiv:2410.12176.

[18] C. Moosmüller and A. Cloninger, *Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations*, Information and Inference: A Journal of the IMA, 12 (2023), pp. 363–389, https://doi.org/10.1093/imaiai/iaac023.

[19] F. Natterer, *The Mathematics of Computerized Tomography*, SIAM, Philadelphia, 2001, https://doi.org/10.1137/1.9780898719284.

[20] S. Park and D. Slepčev, *Geometry and analytic properties of the sliced Wasserstein space*, Journal of Functional Analysis, 289 (2025), p. 110975, https://doi.org/10.1016/j.jfa.2025.110975.

[21] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde, *The cumulative distribution transform and linear pattern classification*, Applied and Computational Harmonic Analysis, 45 (2018), pp. 616–641, https://doi.org/10.1016/j.acha.2017.02.002.

[22] M. Piening and R. Beinert, *Slicing the Gaussian mixture Wasserstein distance*, 2025, https://doi.org/10.48550/arXiv.2504.08544. arXiv:2504.08544.

[23] M. Quellmalz, R. Beinert, and G. Steidl, *Sliced optimal transport on the sphere*, Inverse Problems, 39 (2023), p. 105005, https://doi.org/10.1088/1361-6420/acf156.

[24] M. Quellmalz, L. Buecher, and G. Steidl, *Parallelly sliced optimal transport on spheres and on the rotation group*, Journal of Mathematical Imaging and Vision, 66 (2024), pp. 951–976, https://doi.org/10.1007/s10851-024-01206-w.

[25] J. Radon, *Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten*, in Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, vol. 69 of Mathematisch-Physische Klasse, Königlich Sächsische Gesellschaft der Wissenschaften, Leipzig, 1917, pp. 262–277.

[26] A. G. Ramm and A. I. Katsevich, *The Radon Transform and Local Tomography*, CRC Press, 1996, https://doi.org/10.1201/9781003069331.

[27] M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, S. Li, S. Kolouri, A. Aldroubi, J. M. Nichols, and G. K. Rohde, *Radon cumulative distribution transform subspace modeling for image classification*, Journal of Mathematical Imaging and Vision, 63 (2021), pp. 1185–1203, https://doi.org/10.1007/s10851-021-01052-0.

[28] M. Shifat-E-Rabbi, Y. Zhuang, S. Li, A. H. M. Rubaiyat, X. Yin, and G. K. Rohde, *Invariance encoding in sliced-Wasserstein space for image classification with limited training data*, Pattern Recognition, 137 (2023), p. 109268, https://doi.org/10.1016/j.patcog.2022.109268.

[29] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003, https://doi.org/10.1090/gsm/058.

[30] Y. Zhuang, S. Li, M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, and G. K. Rohde, *Local sliced Wasserstein feature sets for illumination invariant face recognition*, Pattern Recognition, 162 (2025), p. 111381, https://doi.org/10.1016/j.patcog.2025.111381.