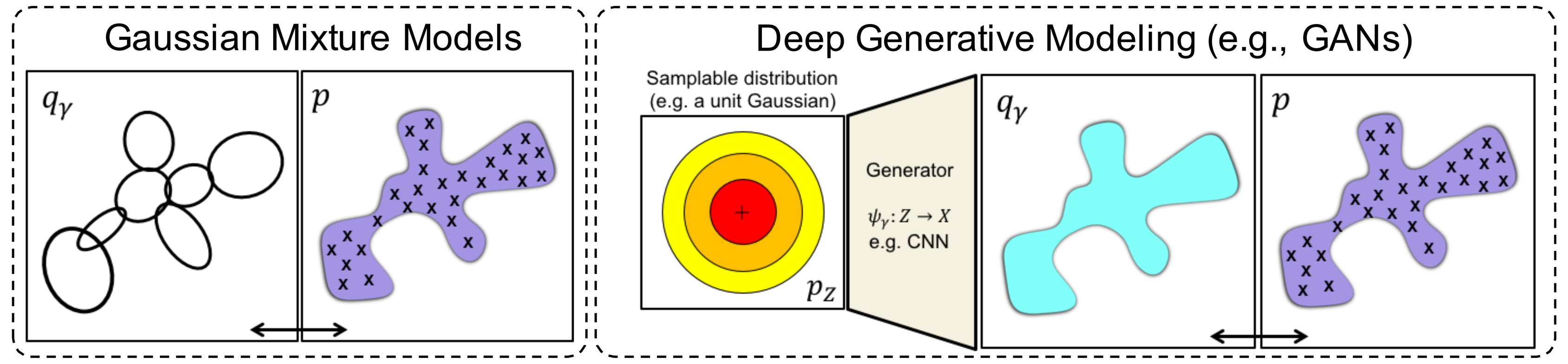
Generalized Sliced Wasserstein Distances

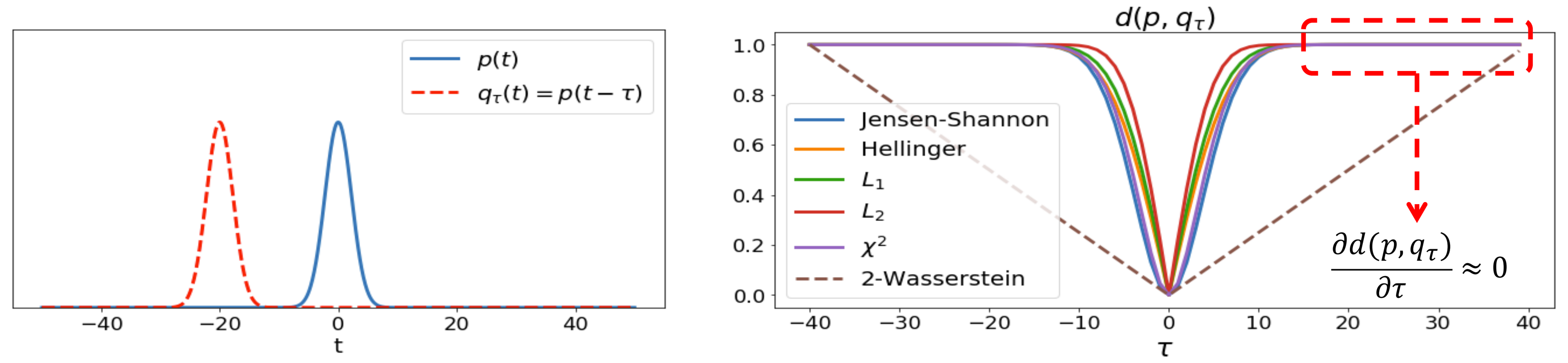
Soheil Kolouri*[1], Kimia Nadjahi*[2], Umut Şimşekli[2], Roland Badeau[2], Gustavo K. Rohde[3]
[1] HRL Laboratories, Malibu, CA. [2] Télécom Paris [3] University of Virginia

1. Why to bother about distances between probability measures?

Measuring the dissimilarity between probability distributions is at the heart of many ML tasks, including generative modeling.



Wasserstein distances have been shown to capture the underlying geometry of the space and are suitable for learning.



Calculating Wasserstein distances is computationally expensive, but for one-dimensional distributions there is a closed form.

2. Sliced Wasserstein (SW) Distances and Max-SW Distances

Formally, for probability measures μ and ν with finite p'th moment, defined on $X \subset \mathbb{R}^d$, with corresponding densities I_μ and I_ν, the SW is defined as:

$$SW_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p\left(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\right) d\theta\right)^{\frac{1}{p}}$$

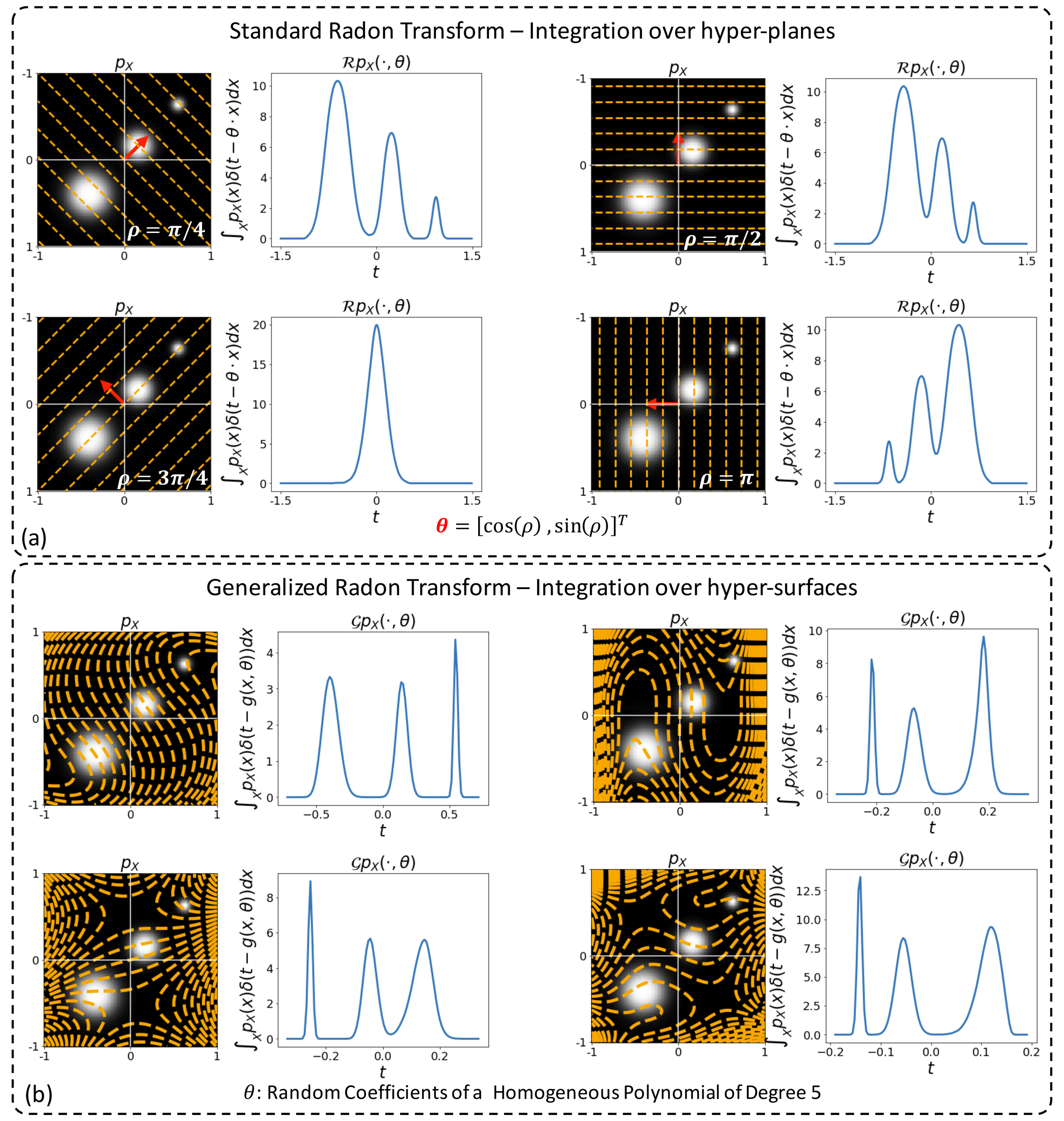
where $\mathcal{R}I$ is the Radon transform of the density I:

$$\mathcal{R}I(t,\theta) = \int_X I(x)\delta(t - \langle x,\theta\rangle)dx, \qquad \forall t \in \mathbb{R}, \theta \in \mathbb{S}^{d-1}$$

In practice, we use a Monte-Carlo approximation where samples $\{\theta_l\}_l$ are uniformly drawn on $\mathbb{S}^{d-1}$. However, in high-dimensions there is a high chance that $W_p\left(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\right) \approx 0$ for randomly drawn θs. Max-SW was proposed to alleviate this issue:

$$max\text{-}SW_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} W_p\left(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\right)$$

3. Classic Radon Transform versus Generalized Radon Transforms



The **classic Radon transform** integrates the input distribution along the **hyperplanes** of $\mathbb{R}^d$:

$$H(t,\theta) \coloneqq \{x \mid \langle x,\theta\rangle = t\}, \forall t \in \mathbb{R}, \theta \in \mathbb{S}^{d-1}$$

The **generalized Radon transform**, on the other hand, integrates the input distribution along the **hypersurfaces** of $\mathbb{R}^d$, parameterized by a defining function, $g(\cdot,\theta)$:

$$H(t,\theta) \coloneqq \{x \mid g(x,\theta) = t\}, \forall t \in \mathbb{R}, \theta \in \Omega_\theta$$

The generalized Radon transform is defined as:

$$\mathcal{G}I(t,\theta) = \int_X I(x)\delta(t - g(x,\theta))dx$$

Note that $g(x,\theta) = \langle x,\theta\rangle$ leads to the classic Radon transform.

4. Generalized Sliced Wasserstein (GSW) and Max-GSW Distances

Using the definition of generalized Radon transform, we define the GSW as follows:

$$GSW_p(\mu,\nu) = \left(\int_{\Omega_\theta} W_p^p\left(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)\right) d\theta\right)^{\frac{1}{p}}$$

and similar to the SW case, we also define the max-GSW to be:

$$max\text{-}GSW_p(\mu,\nu) = \max_{\theta\in\Omega} W_p\left(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)\right)$$

GSW_p and $max\text{-}GSW_p$ are true metrics iff the generalized Radon transform is injective. Otherwise, they provide pseudo-metrics (i.e., identity of indiscernibles would not be satisfied).

Necessary conditions on the defining function for injectivity of the generalized Radon Transform

H1. $g(x,\theta)$ is a real valued C^∞ function on $X \times \Omega_\theta$ where $\Omega_\theta \subseteq (\mathbb{R}^n \backslash \{0\})$

H2. $g(x,\theta)$ is homogeneous of degree 1 in θ:
$$g(x,\lambda\theta) = \lambda g(x,\theta), \forall\theta \in \mathbb{R}$$

H3. $g(x,\theta)$ is non-degenerate in the sense that
$$\frac{\partial g}{\partial x}(x,\theta) \neq 0, \qquad \forall (x,\theta) \in X \times \Omega_\theta$$

H4. The mixed Hessian of $g(x,\theta)$ is positive definite: $\det\left(\left(\partial^2 g/\partial x_i \partial\theta_j\right)_{i,j}\right) > 0$

5. Numerical Experiments – GSW Flows

We consider the following problem:

$$\min_\mu GSW_p(\mu,\nu)$$

where ν is a target and μ is the source probability measure, initialized with a normal density. The optimization is then solved by

$$\partial_t \mu_t = -\nabla\, GSW_p(\mu_t,\nu), \qquad I_{\mu_0} = \mathcal{N}(0,I)$$

Comparison between GSW and Max-GSW