# Information geometry of Wasserstein statistics on shapes and affine deformations

Shun-ichi Amari, Takeru Matsuda

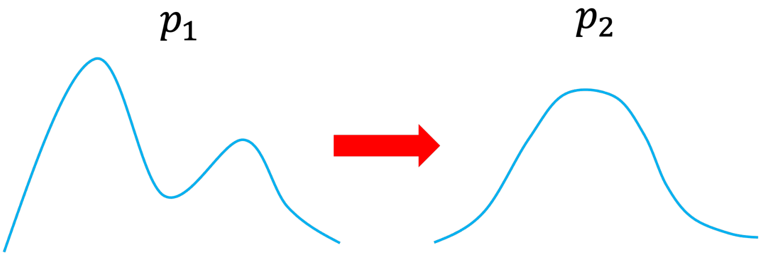
Information Geometry, 2024

# Wasserstein distance

- $L^2$ Wasserstein distance (= optimal transportation cost) between $p_1$ and $p_2$ on $\mathbb{R}^d$

$$W_2(p_1, p_2) = \inf_{X_1, X_2} \mathrm{E}[\|X_1 - X_2\|^2]^{1/2}$$

  - infimum over all joint distributions of $(X_1, X_2)$ with $X_1 \sim p_1$ and $X_2 \sim p_2$ marginally (coupling)



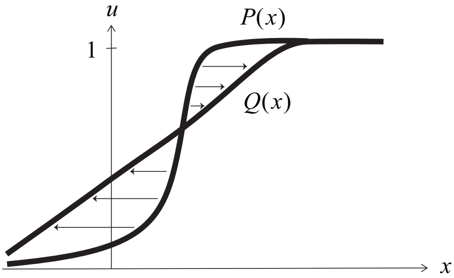
# One-dimensional case

- When $d = 1$, $W_2$ is explicitly given by the cdfs $P_1$ and $P_2$:

$$W_2(p_1, p_2) = \left( \int_0^1 (P_1^{-1}(u) - P_2^{-1}(u))^2 \mathrm{d}u \right)^{1/2}$$

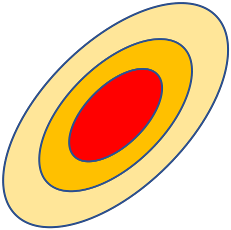
- optimal coupling = monotone map

$$X_2 = P_2^{-1}(P_1(X_1))$$



# Elliptically contoured family

- When $d \geq 2$, $W_2$ is intractable in general..
- elliptically contoured family (e.g. Gaussian)
  - $\mu$: mean, $\Sigma$: covariance, $f$: shape



$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x - \mu)\|)$$

**Proposition (Gelbrich, 1990)**

$$W_2(p(x \mid \mu_1, \Sigma_1), p(x \mid \mu_2, \Sigma_2))$$
$$= \left( \|\mu_1 - \mu_2\|^2 + \mathrm{tr}\left( \Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \right) \right)^{1/2}$$

- note: $W_2$ does not depend on the shape $f$

# Wasserstein v.s. Kullback–Leibler

- bijective variable transformation

$$y = g(x) \quad \rightarrow \quad \tilde{p}(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p(x)$$

- Kullback–Leibler divergence: invariant

$$D_{\mathrm{KL}}(p, q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

$$D_{\mathrm{KL}}(\tilde{p}, \tilde{q}) = D_{\mathrm{KL}}(p, q)$$

- Wasserstein distance: not invariant

$$W_2(\tilde{p}, \tilde{q}) \neq W_2(p, q)$$

# Li–Zhao framework

- Recently, Li and Zhao (2023) developed Wasserstein counterparts of information geometric concepts

| | |
|---|---|
| Kullback–Leibler divergence | Wasserstein distance |
| Fisher score | Wasserstein score |
| Fisher information matrix | Wasserstein information matrix |
| covariance | Wasserstein covariance |
| Cramer–Rao | Wasserstein–Cramer–Rao |
| Fisher efficiency | Wasserstein efficiency |

- We investigate their statistical meaning

# Continuity equation

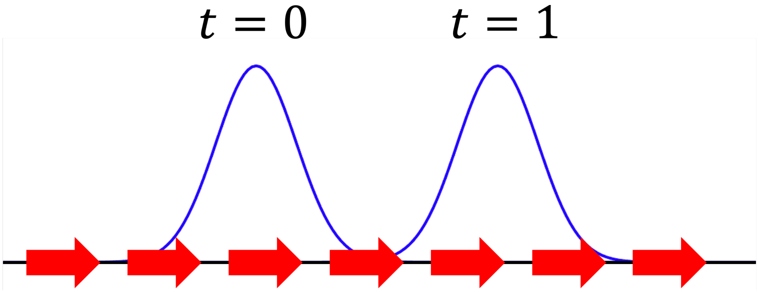
$$\frac{\partial}{\partial t} p(x,t) = -\nabla_x \cdot (p(x,t)\nabla_x \Phi(x))$$

- This PDE describes dynamics of measure transport
- intuition: Many particles are distributed with $p(x,t)$ and they move with velocity $\nabla_x \Phi(x)$

# Example: 1d linear potential

$$\frac{\partial}{\partial t} p(x,t) = -\nabla_x \cdot (p(x,t)\nabla_x \Phi(x))$$

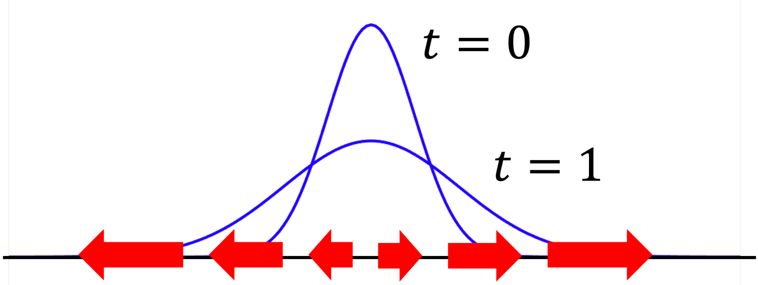
- $\Phi(x) = x \to \nabla_x \Phi(x) \equiv 1$ (const.)
- $p(x,0) = \mathrm{N}(0,1) \to p(x,t) = \mathrm{N}(t,1)$ (shift)



# Example: 1d quadratic potential

$$\frac{\partial}{\partial t} p(x,t) = -\nabla_x \cdot (p(x,t)\nabla_x \Phi(x))$$

- $\Phi(x) = x^2 \rightarrow \nabla_x \Phi(x) = 2x$
- $p(x,0) = \mathrm{N}(0,1) \rightarrow p(x,t) = \mathrm{N}(0,t+1)$ (expansion)



# Wasserstein score function

## Definition (Li and Zhao, 2023)

For $i = 1, \ldots, p$, the Wasserstein score function $\Phi_i^{\mathrm{W}}(x \mid \theta)$ is the solution of

$$-\nabla_x \cdot (p(x \mid \theta) \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta)) = \frac{\partial}{\partial \theta_i} p(x \mid \theta), \quad \mathrm{E}_\theta[\Phi_i^{\mathrm{W}}(x \mid \theta)] = 0.$$

- For infinitesimal $\delta$, the map $x \mapsto x + \delta \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta)$ is the optimal transport map from $p(x \mid \theta)$ to $p(x \mid \theta + \delta e_i)$ with transportation cost

$$W_2(p(x \mid \theta), p(x \mid \theta + \delta e_i)) = \left( \int \| \delta \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta) \|^2 p(x \mid \theta) \mathrm{d}x \right)^{1/2}$$

  - $e_i$: $i$-th standard unit vector

# Wasserstein information matrix (WIM)

## Definition (Li and Zhao, 2023)

The Wasserstein information matrix $G_{\mathrm{W}}(\theta)$ is the $p \times p$ matrix given by

$$G_{\mathrm{W}}(\theta) = \left( \int \frac{\partial}{\partial \theta_i} p(x \mid \theta) \cdot \Phi_j^{\mathrm{W}}(x \mid \theta) \mathrm{d}x \right)_{ij}$$

- cf. Fisher information matrix

$$G_{\mathrm{F}}(\theta) = \left( \int \frac{\partial}{\partial \theta_i} p(x \mid \theta) \cdot \Phi_j^{\mathrm{F}}(x \mid \theta) \mathrm{d}x \right)_{ij}$$

$$\Phi_j^{\mathrm{F}}(x \mid \theta) = \frac{\partial}{\partial \theta_j} \log p(x \mid \theta)$$

- inner product = pairing of tangent vector and cotangent vector
  - information geometry: m-representation and e-representation

# Wasserstein information matrix (WIM)

## Proposition (Li and Zhao, 2023)

$$G_{\mathrm{W}}(\theta)_{ij} = \mathrm{E}_\theta[(\nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta))^\top (\nabla_x \Phi_j^{\mathrm{W}}(x \mid \theta))]$$

## Proposition (Li and Zhao, 2023)

$$W_2(p(x \mid \theta), p(x \mid \theta + \delta))^2 = \delta^\top G_W(\theta)\delta + o(\|\delta\|^2)$$

- WIM = Hessian of Wasserstein distance
  - cf. Fisher information matrix = Hessian of Kullback–Leibler divergence
- WIM appears in Otto calculus and Wasserstein gradient flow

# Example: 1d Gaussian

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \theta = (\mu, \sigma)$$

- Wasserstein distance

$$W_2(p(x \mid \theta_1), p(x \mid \theta_2))^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2$$

- Wasserstein score function

$$\Phi_\mu^{\mathrm{W}}(x \mid \theta) = x - \mu, \quad \Phi_\sigma^{\mathrm{W}}(x \mid \theta) = \frac{(x-\mu)^2 - \sigma^2}{2\sigma}$$

- Wasserstein information matrix

$$G_{\mathrm{W}}(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- More generally, 1d location-scale model is Euclidean (totally geodesic) in $L^2$-Wasserstein geometry

# Wasserstein estimator

**Definition (Li and Zhao, 2023)**

The Wasserstein estimator $\hat{\theta}_{\mathrm{W}}(x)$ is the zero of the Wasserstein score function:

$$\Phi_i^{\mathrm{W}}(x \mid \hat{\theta}_{\mathrm{W}}(x)) = 0, \quad i = 1, \ldots, p$$

- cf. MLE = zero of the Fisher score function = projection w.r.t. Kullback–Leibler divergence
- What does it mean??
  - It is different from the projection w.r.t. Wasserstein distance studied in Amari and M. (2022)

# Example: elliptically contoured family

$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x-\mu)\|)$$

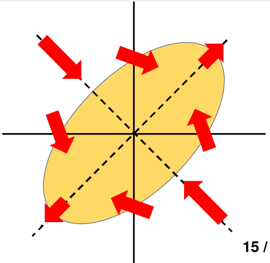
**Theorem (Amari and M., 2024)**

- Wasserstein score functions are quadratic
- Wasserstein estimator = 2nd-order moment estimator

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^\top$$

e.g. 2d Gaussian $\mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \theta\\ \theta & 1\end{pmatrix}\right)$

$$\Phi^{\mathrm{W}}(x \mid \theta) = \frac{1}{4}\begin{pmatrix}x_1\\x_2\end{pmatrix}^\top \begin{pmatrix}-\theta & 1\\ 1 & -\theta\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$$



# Wasserstein covariance & Wasserstein–Cramer–Rao

## Definition (Li and Zhao, 2023)

The Wasserstein covariance $\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}]$ of an estimator $\hat{\theta}$ is the $p \times p$ positive semidefinite matrix given by

$$\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}] = (\mathrm{E}_\theta[(\nabla_x \hat{\theta}_i)^\top (\nabla_x \hat{\theta}_j)])_{ij}$$

## Theorem (Li and Zhao, 2023)

When $\hat{\theta}$ is unbiased ($\mathrm{E}_\theta[\hat{\theta}] = \theta$),

$$\mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta}) \succeq G_{\mathrm{W}}(\theta)^{-1}$$

- What does it mean??
  - cf. usual Cramer–Rao = lower bound of mean squared error

# Wasserstein covariance and robustness

$$X \sim p(x \mid \theta), \quad Z \sim q(z)$$

- We consider estimation of $\theta$ from noisy observation $X + Z$
  - $\mathrm{E}[Z] = 0$, $\mathrm{Var}[Z] = \sigma^2 I$

**Theorem (Amari and M., 2024)**

$$\begin{aligned}\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}] = \lim_{\sigma^2 \to 0} & \frac{\mathrm{Var}_\theta[\hat{\theta}(X+Z)] - \mathrm{Var}_\theta[\hat{\theta}(X)]}{\sigma^2} \\ & - \frac{1}{2}\left(\mathrm{Cov}_\theta[\hat{\theta}_a(X), \Delta\hat{\theta}_b(X)] + \mathrm{Cov}_\theta[\hat{\theta}_b(X), \Delta\hat{\theta}_a(X)]\right)\end{aligned}$$

# Wasserstein covariance and robustness

## Corollary (Amari and M., 2024)

If $\hat{\theta}$ is quadratic,

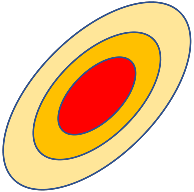
$$\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}] = \lim_{\sigma^2 \to 0} \frac{\mathrm{Var}_\theta[\hat{\theta}(X+Z)] - \mathrm{Var}_\theta[\hat{\theta}(X)]}{\sigma^2}$$

- Thus, Wasserstein covariance quantifies the robustness against additive noise of quadratic estimators.

- e.g. Wasserstein estimator for elliptically contoured family

$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x-\mu)\|)$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$$



- "additive noise": not invariant w.r.t. transformation of $x$
  - noise contamination $\approx$ (random) transportation