A Stochastic Control Model for Strategic Misdirection via Sequential Hypothesis Testing

Haosheng Zhou

University of California, Santa Barbara

July 23, 2025

Joint work with Daniel Ralston, Xu Yang and Ruimeng Hu

Previous Literature

Strategic interactions between players:

- Deceive adversaries about objectives.

Existing approach:

- Robust control[1] & RL (passive).
- Antagonistic control[2] (deterministic, constrained).
- POMDP & POSG[3] (zero-sum).
- IRL[4] (ill-posed).

---

[1] B. Taskesen et al., *Distributionally Robust Linear Quadratic Control*, NeurIPS, 2024.

[2] T. Lipp, S. Boyd, *Antagonistic Control*, Systems & Control Letters, 2016.

[3] O. Ma et al., *SUB-PLAY: Adversarial Policies against Partially Observed MARL Systems*, Proceedings of ACM SIGSAC, 2024.

[4] A. Y. Ng, S. Russell, *Algorithms for Inverse Reinforcement Learning*, ICML, 2000.

Primary Task Model

Blue team has a primary task, state dynamics:

$$\begin{aligned} \mathrm{d}V_t &= \alpha_t \mathrm{d}t + \sigma_B \mathrm{d}B_t, \\ \mathrm{d}Y_t &= (V_t + \beta_t)\mathrm{d}t + \sigma_W \mathrm{d}W_t, \end{aligned}$$

where $\alpha_t = \phi^{\alpha}(t, V_t, Y_t)$ and $\beta_t = \phi^{\beta}(t, V_t, Y_t)$. Cost functionals:

$$\begin{aligned} r(t, v, y, \alpha, \beta) &= \frac{r_\alpha}{2}\alpha^2 + \frac{r_\beta}{2}\beta^2 + \frac{r_v}{2}(v - \overline{v}(t))^2, \\ g(v, y) &= \frac{t_v}{2}(v - \overline{v}_T)^2. \end{aligned}$$

Optimal control: $\hat{\beta} \equiv 0$ (baseline).

Introduction of Adversary

Red team only observes blue's state dynamics and sample paths of $\{Y_t\}$.

Blue team's bi-objective optimization:

- Fulfill the primary task.
- Actively perturb $\{Y_t\}$ to hide its intentions.
- Act based on its belief in red (envisioned rather than actual).

Strategic Misdirection

Difficulty: modeling the misdirection.

Maximize $D(\mathcal{L}(X_T), \mathcal{L}(Y_T))$, where $\mathrm{d}X_t = (V_t + 0)\mathrm{d}t + \sigma_W \mathrm{d}W_t$?

- Ill-posed without constraints.
- Non-Markovian.

Why it fails?—Information divergences correspond to a family of "test functions" $\mathcal{F}$, e.g.,

- Levy-Prokhorov metric — $\mathcal{F} = C_b(\mathbb{R})$.
- Wasserstein-1 metric — $\mathcal{F} = \{f : \|f\|_L \leq 1\}$.

Idea: simple *v.s.* simple sequential hypothesis testing (SHT).

SHT Formulation

Assumption (Linearity of Control)

Assume the feedback functions are linear in the state variables:

$$\phi^{\alpha}(t, v, y) = b_{\alpha}(t)v + c_{\alpha}(t)y + d_{\alpha}(t),$$
$$\phi^{\beta}(t, v, y) = b_{\beta}(t)v + c_{\beta}(t)y + d_{\beta}(t),$$

Red's hypothesis testing:

$$\begin{cases} H_0 : b_{\beta} \equiv 0, c_{\beta} \equiv 0, d_{\beta} \equiv 0 \\ H_1 : b_{\beta} \equiv 0, c_{\beta} = f_c, d_{\beta} = f_d \end{cases}.$$

Remarks:

- Blue team specifies f_c, f_d.
- Reject H_0 — presence of misdirection.

SHT Formulation

The sequential probability ratio test (most powerful) with statistics $L_T := \frac{\mathrm{d}\mathcal{L}^{H_1}(V,Y)}{\mathrm{d}\mathcal{L}^{H_0}(V,Y)}$, where $\mathcal{L}^{H_1}(V,Y)(\cdot) := \mathbb{P}_{H_1}((V,Y) \in \cdot)$, a measure on $C([0,T];\mathbb{R}^2)$.

Proposition (Z.-Ralston-Yang-Hu 2025)

The expected log-likelihood ratio, evaluated at the empirically observed trajectories (V,Y), is given by:

$$\mathbb{E}\log L_T = \frac{1}{\sigma_W^2}\mathbb{E}\int_0^T \left[(f_c(t)Y_t + f_d(t))\beta_t - \frac{1}{2}(f_c(t)Y_t + f_d(t))^2\right]\mathrm{d}t.$$

Blue maximizes $\mathbb{E}\log L_T$ besides completing the primary task.

Strategic Misdirection Model

Blue's objective:

$$\inf_{\alpha,\beta} J(\alpha,\beta) := \mathbb{E}\left[\int_0^T r(t, V_t, Y_t, \alpha_t, \beta_t)\,\mathrm{d}t + g(V_T, Y_T)\right] - \lambda \mathbb{E}\log L_T.$$

A Markovian control with a modified running cost:

$$h := r - \frac{\lambda}{\sigma_W^2}(f_c(t)y + f_d(t))\beta + \frac{\lambda}{2\sigma_W^2}(f_c(t)y + f_d(t))^2.$$

Solve the Markovian control problem:

- Hamilton-Jacobi-Bellman (HJB) equation:

$$\partial_t V + \inf_{\alpha,\beta}\left\{\alpha\partial_v V + (v+\beta)\partial_y V + \frac{1}{2}\sigma_B^2\partial_{vv}V + \frac{1}{2}\sigma_W^2\partial_{yy}V + h\right\} = 0.$$

Semi-Explicit Solution

- Quadratic ansatz of value function:

$$V(t,v,y)=\frac{\mu_t}{2}v^2+\eta_t vy+\frac{\rho_t}{2}y^2+\gamma_t v+\theta_t y+\xi_t.$$

- A system of Riccati ODEs:

$$\begin{cases}
\dot{\mu}_t=\frac{1}{r_\alpha}\mu_t^2+\frac{1}{r_\beta}\eta_t^2-2\eta_t-r_v,\\
\dot{\eta}_t=\frac{1}{r_\alpha}\mu_t\eta_t+\frac{1}{r_\beta}\rho_t\eta_t-\rho_t-\frac{\lambda}{r_\beta\sigma_W^2}\eta_t f_c(t),\\
\dot{\rho}_t=\frac{1}{r_\alpha}\eta_t^2+\frac{1}{r_\beta}\rho_t^2-\frac{2\lambda}{r_\beta\sigma_W^2}\rho_t f_c(t)+(\frac{\lambda^2}{r_\beta\sigma_W^4}-\frac{\lambda}{\sigma_W^2})f_c^2(t),\\
\dot{\gamma}_t=\frac{1}{r_\alpha}\mu_t\gamma_t+\frac{1}{r_\beta}\eta_t\theta_t-\theta_t+r_v\overline{v}(t)-\frac{\lambda}{r_\beta\sigma_W^2}\eta_t f_d(t),\\
\dot{\theta}_t=\frac{1}{r_\alpha}\eta_t\gamma_t+\frac{1}{r_\beta}\rho_t\theta_t-\frac{\lambda}{r_\beta\sigma_W^2}\theta_t f_c(t)-\frac{\lambda}{r_\beta\sigma_W^2}f_d(t)\rho_t\\
\qquad+(\frac{\lambda^2}{r_\beta\sigma_W^4}-\frac{\lambda}{\sigma_W^2})f_c(t)f_d(t),\\
\dot{\xi}_t=\frac{1}{2r_\alpha}\gamma_t^2+\frac{1}{2r_\beta}\theta_t^2-\frac{1}{2}\sigma_B^2\mu_t-\frac{1}{2}\sigma_W^2\rho_t-\frac{\lambda}{r_\beta\sigma_W^2}f_d(t)\theta_t\\
\qquad-\frac{r_v[\overline{v}(t)]^2}{2}+(\frac{\lambda^2}{2r_\beta\sigma_W^4}-\frac{\lambda}{2\sigma_W^2})f_d^2(t).
\end{cases}$$

Semi-Explicit Solution

Theorem (Z.-Ralston-Yang-Hu 2025)

If $0 \leq \lambda \leq r_\beta \sigma_W^2$, *the system of ODEs satisfied by* $(\mu_t, \eta_t, \rho_t, \gamma_t, \theta_t, \xi_t)$ *has a unique solution on* $[0, T]$ *for any* $T > 0$.

Optimal control:

$$\hat{\alpha}(t, v, y) = -\frac{\mu_t}{r_\alpha} v - \frac{\eta_t}{r_\alpha} y - \frac{\gamma_t}{r_\alpha},$$
$$\hat{\beta}(t, v, y) = -\frac{\eta_t}{r_\beta} v + \Big(\frac{\lambda}{r_\beta \sigma_W^2} f_c(t) - \frac{\rho_t}{r_\beta}\Big) y + \Big(\frac{\lambda}{r_\beta \sigma_W^2} f_d(t) - \frac{\theta_t}{r_\beta}\Big).$$

Remarks:

- Justify the assumption.
- Compare with H_1.

Numerical Experiments

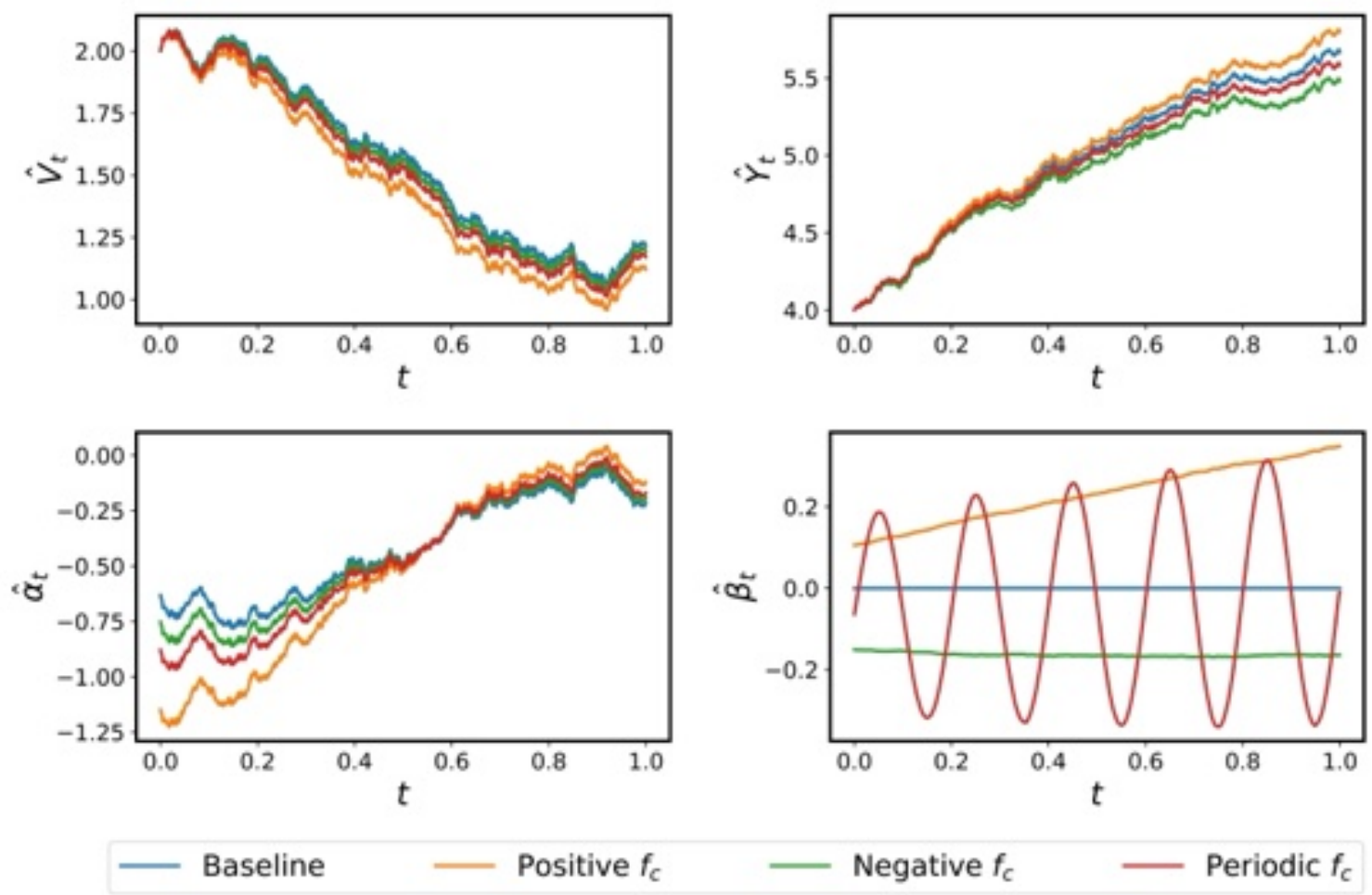


Figure: Comparisons of optimal state (upper panels) and control (lower panels) trajectories for different f_c, fixing other model parameters.

Numerical Experiments

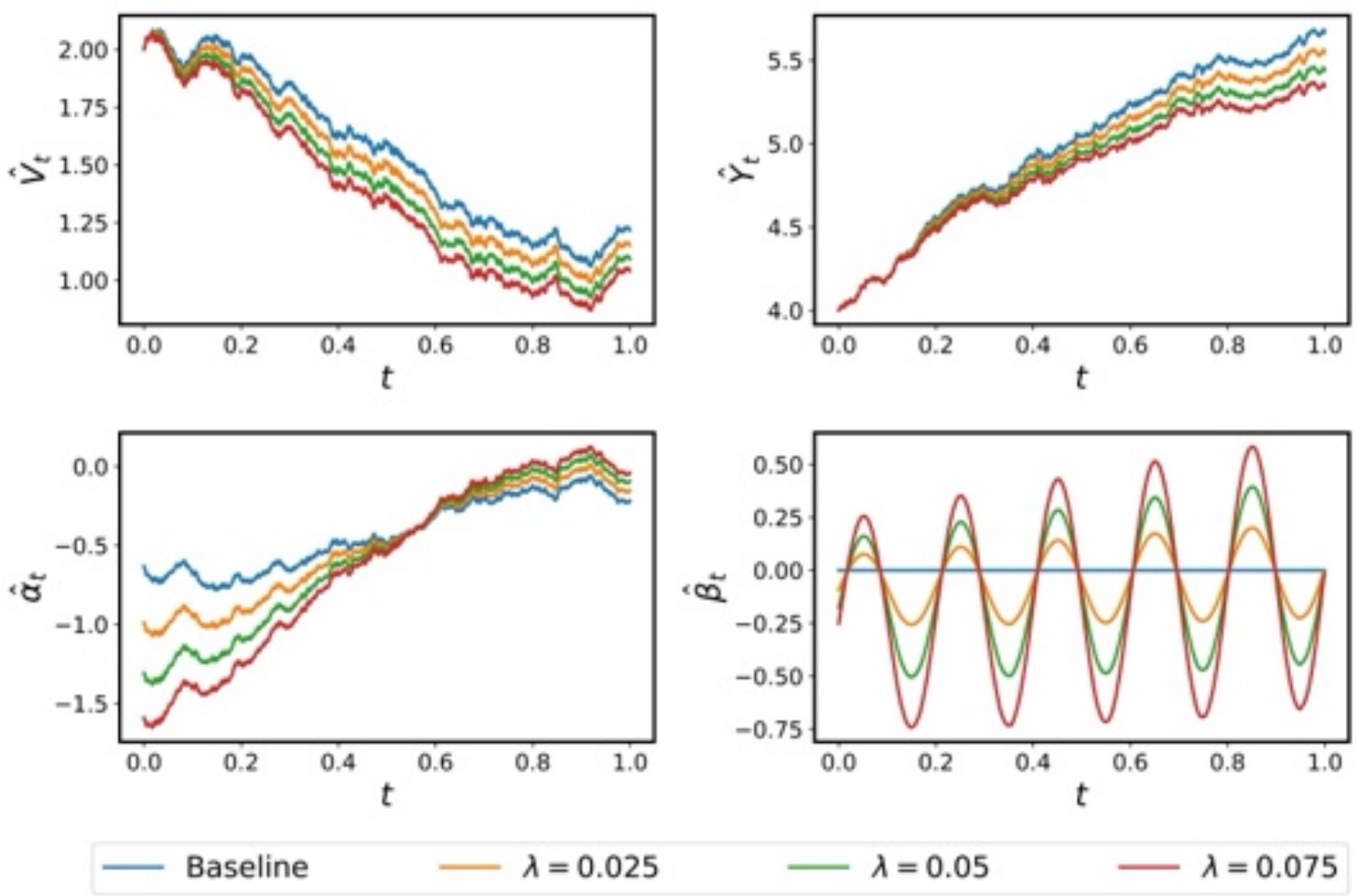


Figure: Comparisons of optimal state (upper panels) and control (lower panels) trajectories for different λ, fixing other model parameters.

Main References


R. Hu, D. Ralston, X. Yang, H. Zhou: Integrating Sequential Hypothesis Testing into Adversarial Games: A Sunzi-Inspired Framework. In Preparation (2025).


R. S. Liptser, A. N. Shiryaev: Statistics of Random Processes: I. General Theory. Springer Science & Business Media (2013).

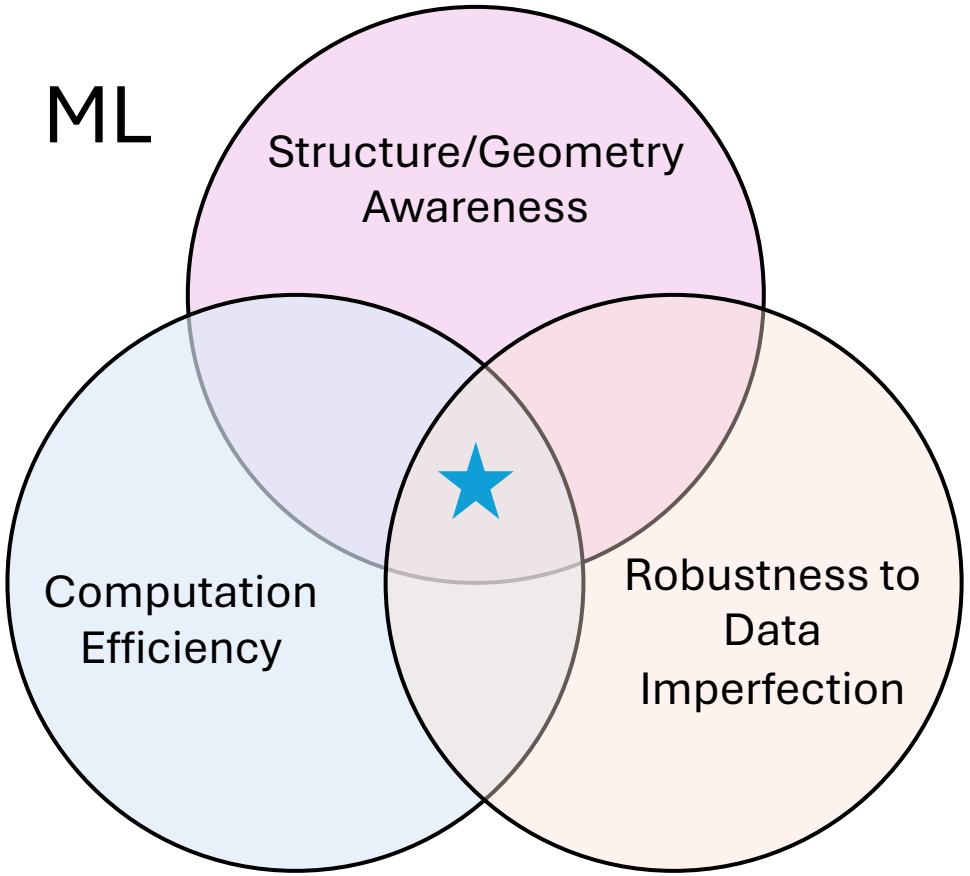
Thank you!

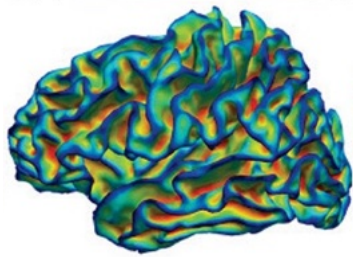
Efficient Optimal Transport in Machine Learning Applications

Xinran (Eva) Liu

Vanderbilt University

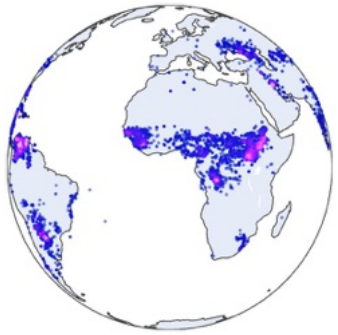
Research Overview

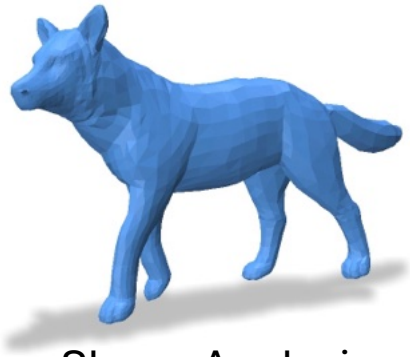




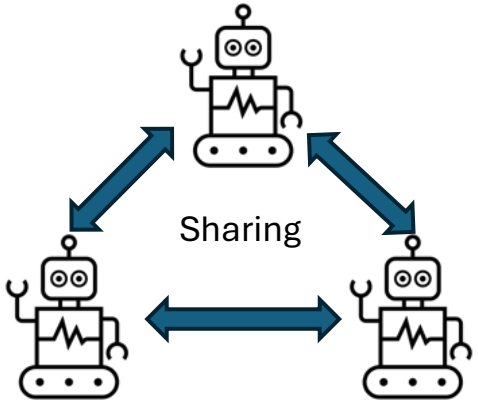
Neuroscience

Geoscience





Shape Analysis



Knowledge sharing in Multi-Agent learning

Overview

Background

Efficient Optimal Transport : Sliced Optimal Transport and Linear Optimal Transport

Linear Spherical Sliced Optimal Transport (LSSOT)

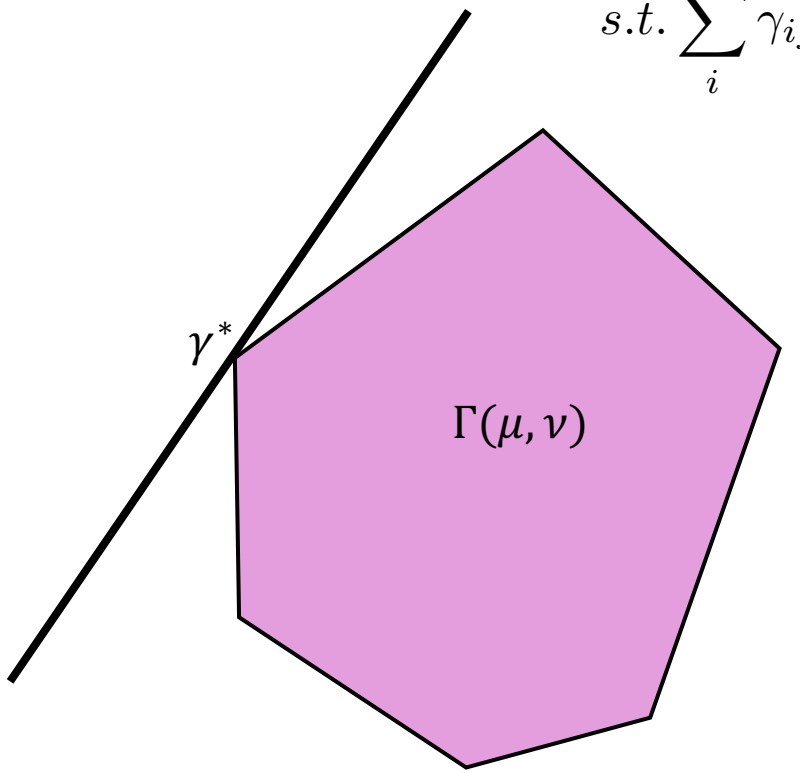
A fast metric for comparing spherical data

Solving Optimal Transport as a Linear Program

Distributions: $\mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}, \nu = \sum_{j=1}^{M} \beta_j \delta_{y_j}$, with $x_i, y_j \in \mathbb{R}^d$,

Optimization Problem: $\min_{\gamma \in \mathbb{R}_+^{N \times M}} c(x_i, y_j)\gamma_{i,j}$

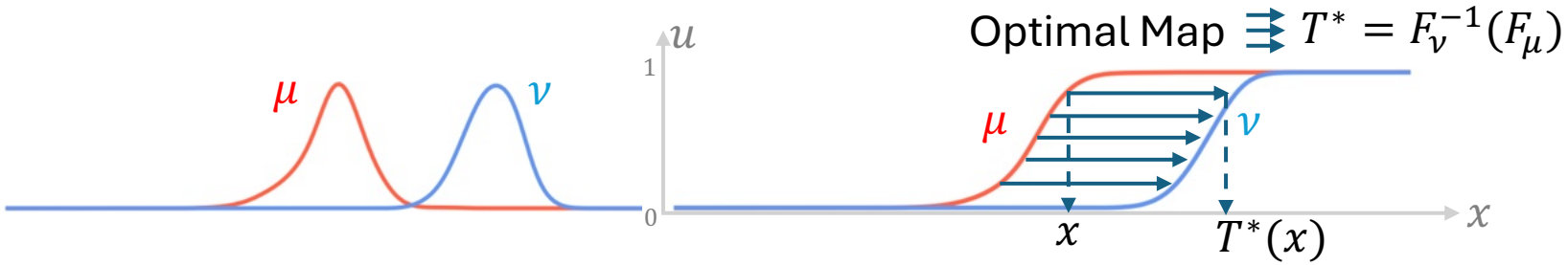
$$s.t. \sum_i \gamma_{ij} = \alpha_i, \sum_j \gamma_{ij} = \beta_j \text{ for all } i, j$$

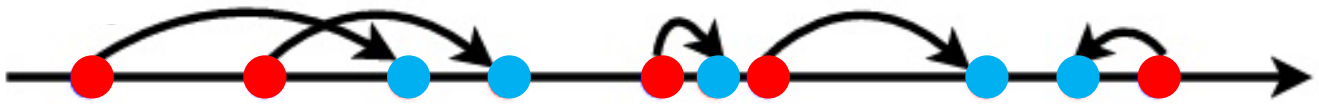


$-$ *e. g. Interior Point Method*
As expensive as $O(N^3 \log N)$

1D Optimal Transport : An Easy Problem

$$W_p^p(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^p \mathrm{d}u$$





Can be solved by Sorting in $O(N \log N)$ time!

Radon Transform and Sliced Wasserstein Distances

Radon Transform

$$\mathcal{R}_\theta[I_\mu](t) = \int_{\mathbb{R}^d} I_\mu(x)\delta(t - \langle x, \theta\rangle)\mathrm{d}x$$

Sliced Wasserstein Distance

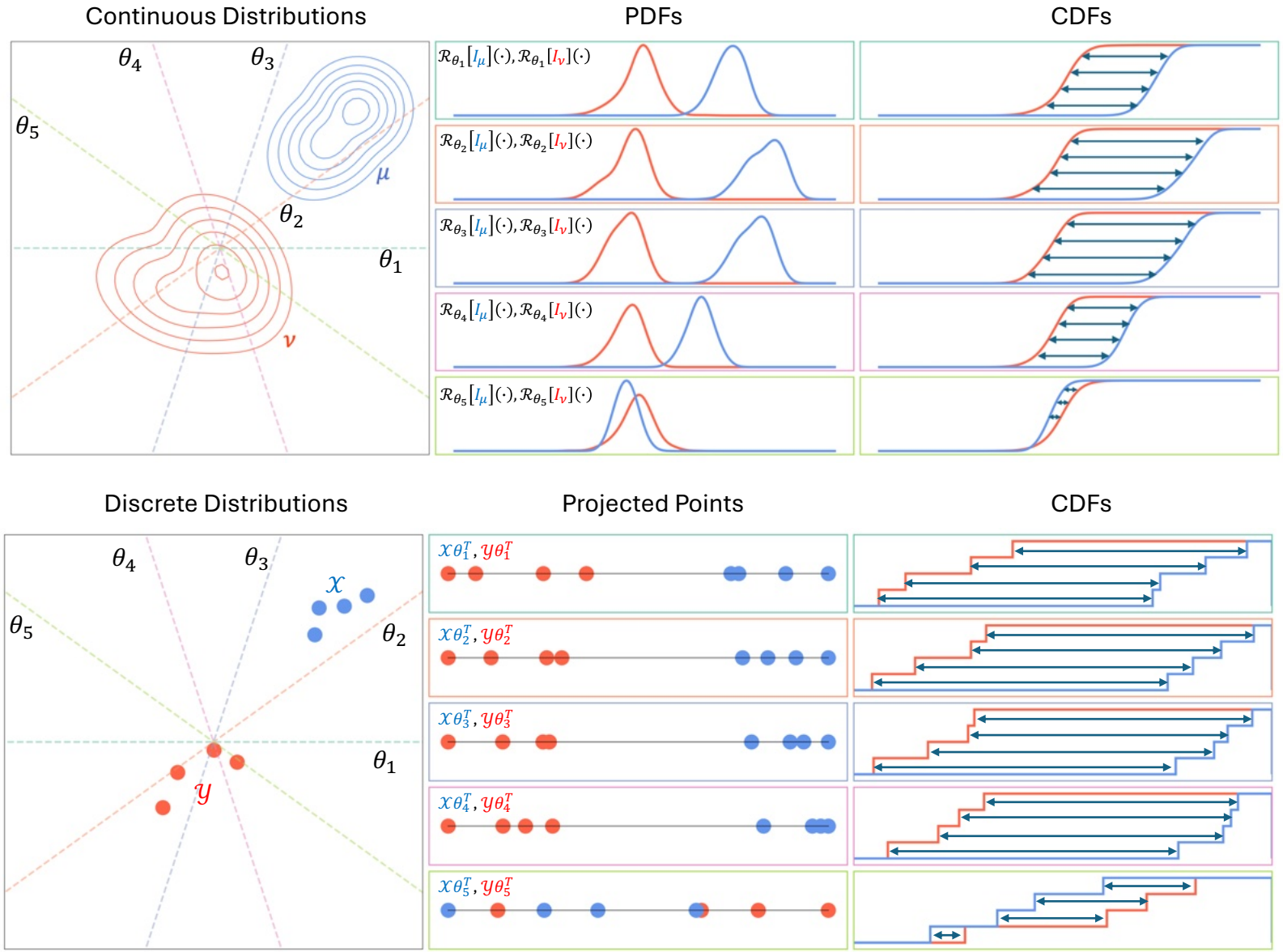
$$SW_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}_{\theta\#}\mu, \mathcal{R}_{\theta\#}\nu)\mathrm{d}\theta\right)^{\frac{1}{p}}$$

Monte Carlo Approximation

$$SW_p(\mu,\nu) \approx \left(\frac{1}{L}\sum_{l=1}^{L} W_p^p(\mathcal{R}_{\theta_l\#}\mu, \mathcal{R}_{\theta_l\#}\nu)\right)^{\frac{1}{p}}$$

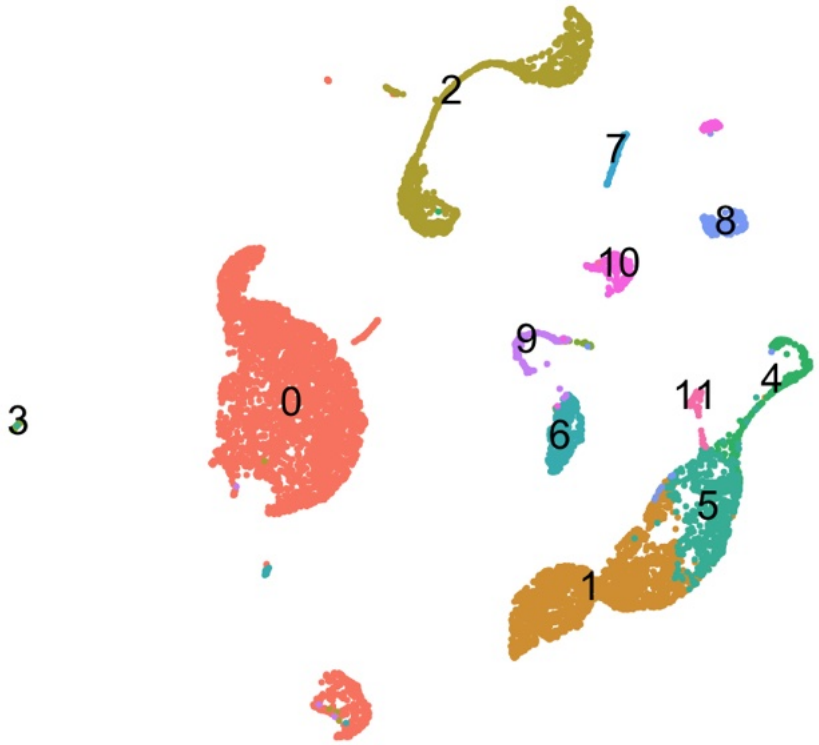
– Can be solved in $O(LN\log N)$ time,
– $O(N\log N)$ if parallelizing over slices

- Defines a metric
- Desirable statistical properties

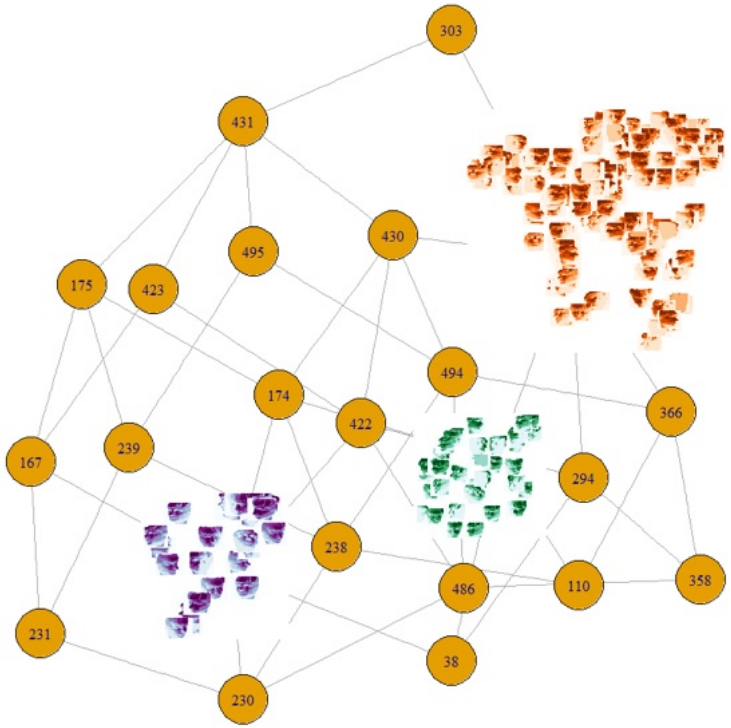


Groupwise Analysis with Pairwise Distance Calculations

Spectral Clustering



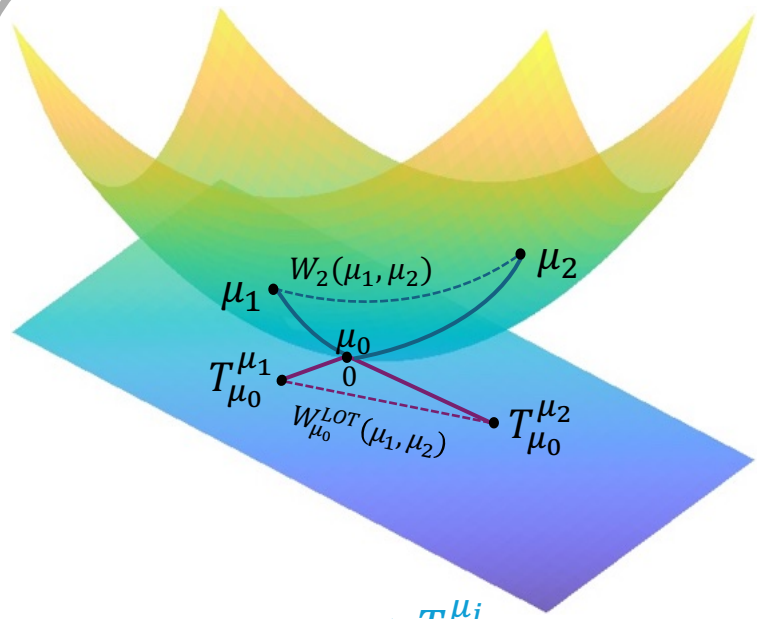
Nearest Neighbor Graph



Zhang, Shuyi, et al. "Spectral clustering of single-cell multi-omics data on multilayer graphs." *Bioinformatics* 38.14 (2022): 3600-3608

Source: LSH for Image Retrieval

Accelerating Groupwise Analysis : Linear Optimal Transport

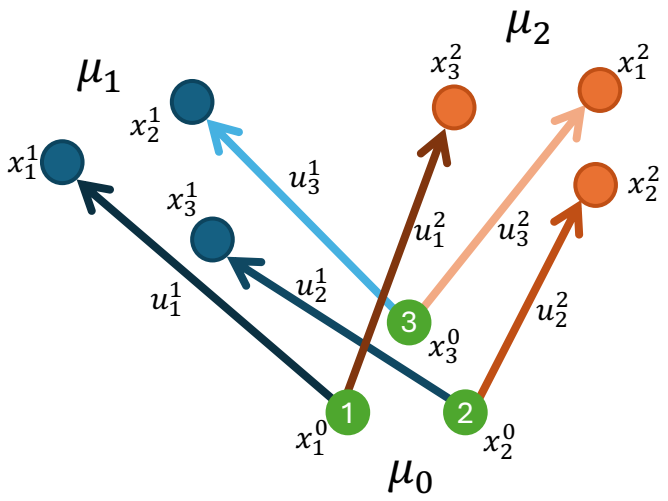


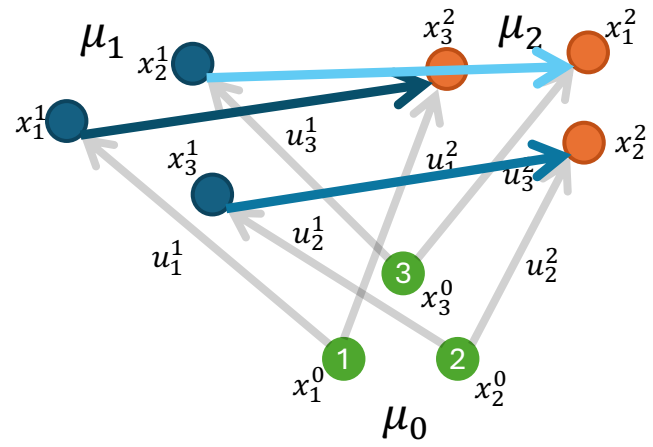
$$\mu_i \to T_{\mu_0}^{\mu_i}$$

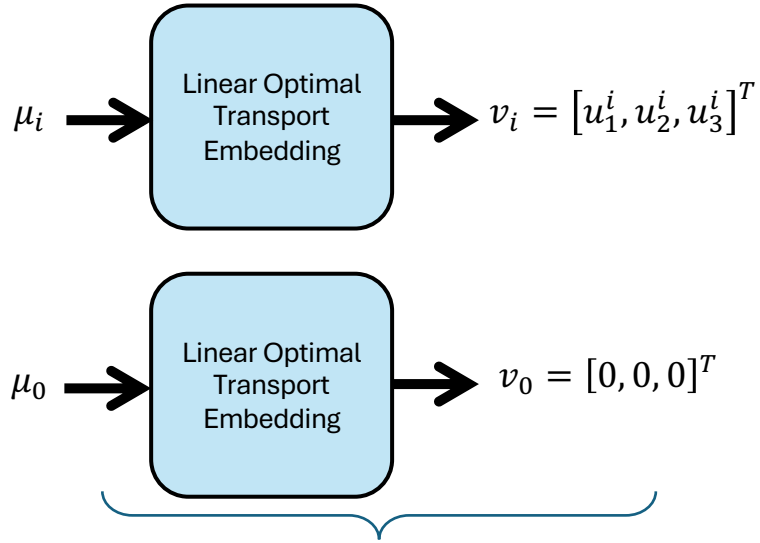
$$W_{\mu_0}^{LOT}(\mu_1,\mu_2) = \left|\left|T_{\mu_0}^{\mu_1} - T_{\mu_0}^{\mu_2}\right|\right|_2$$

$$W_{\mu_0}^{LOT}(\mu_0,\mu_i) = W_2(\mu_0,\mu_i)$$

$$W_{\mu_0}^{LOT}(\mu_1,\mu_2) \approx W_2(\mu_1,\mu_2)$$





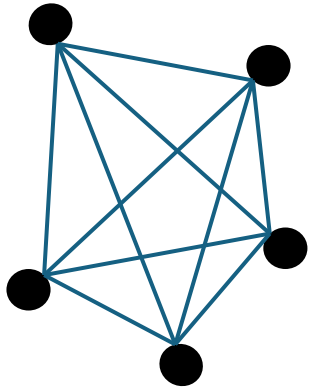


Measure to vector (Permutation Invariant)

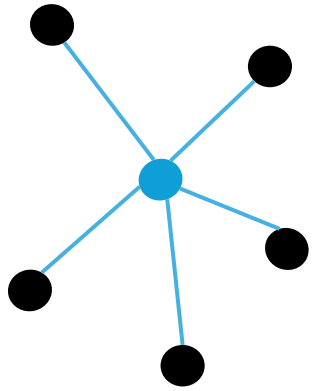
$$\|v_i - v_0\|^2 = \|v_i\|^2 = W_2^2(\mu_i,\mu_0)$$

$$\|v_i - v_j\|^2 = \sum_{n=1}^{N_0} \|u_n^i - u_n^j\|^2 \approx W_2^2(\mu_i,\mu_j)$$

Accelerating Groupwise Analysis : Linear Optimal Transport



$\frac{M(M-1)}{2}$
calculations of OT distances!



M OT distances,
with $\frac{M(M-1)}{2}$ L^2 distances

Much cheaper than
OT distances!

Overview

Background

Efficient Optimal Transport : Sliced Optimal Transport and Linear Optimal Transport

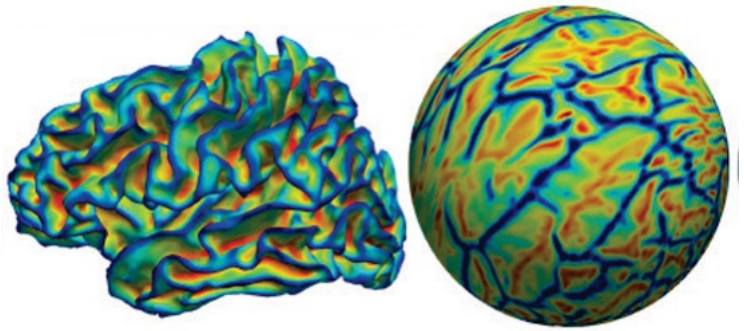
Linear Spherical Sliced Optimal Transport (LSSOT)

A fast metric for comparing spherical data

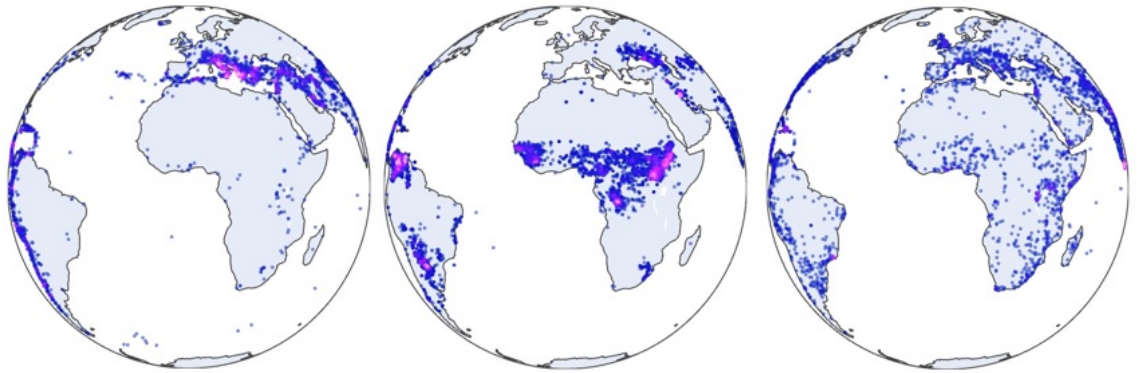
Liu, X., Bai, Y., Martin, R.D., Shi, K., Shahbazi, A., Landman, B.A., Chang, C. and Kolouri, S., Linear Spherical Sliced Optimal Transport: A Fast Metric for Comparing Spherical Data. In *The Thirteenth International Conference on Learning Representations*.

Motivation : Fast Optimal Transport on Spheres

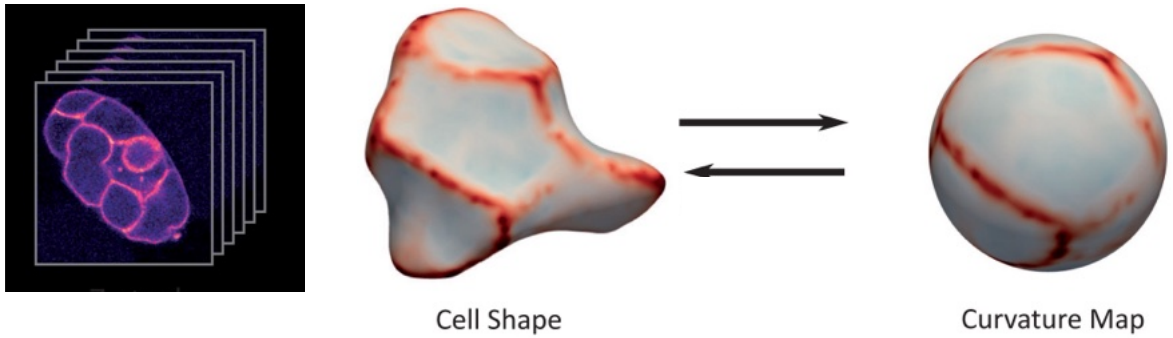
Cortical Surface



Earth Data

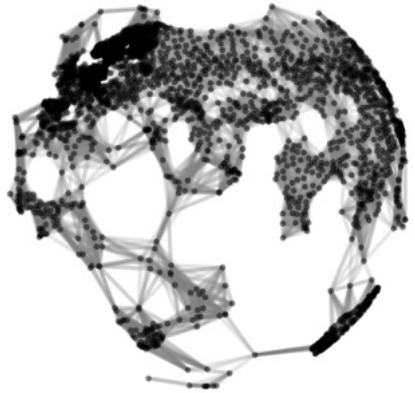


Cell Shape



van Bavel et. al. "Cell shape characterization, alignment, and comparison using FlowShape."

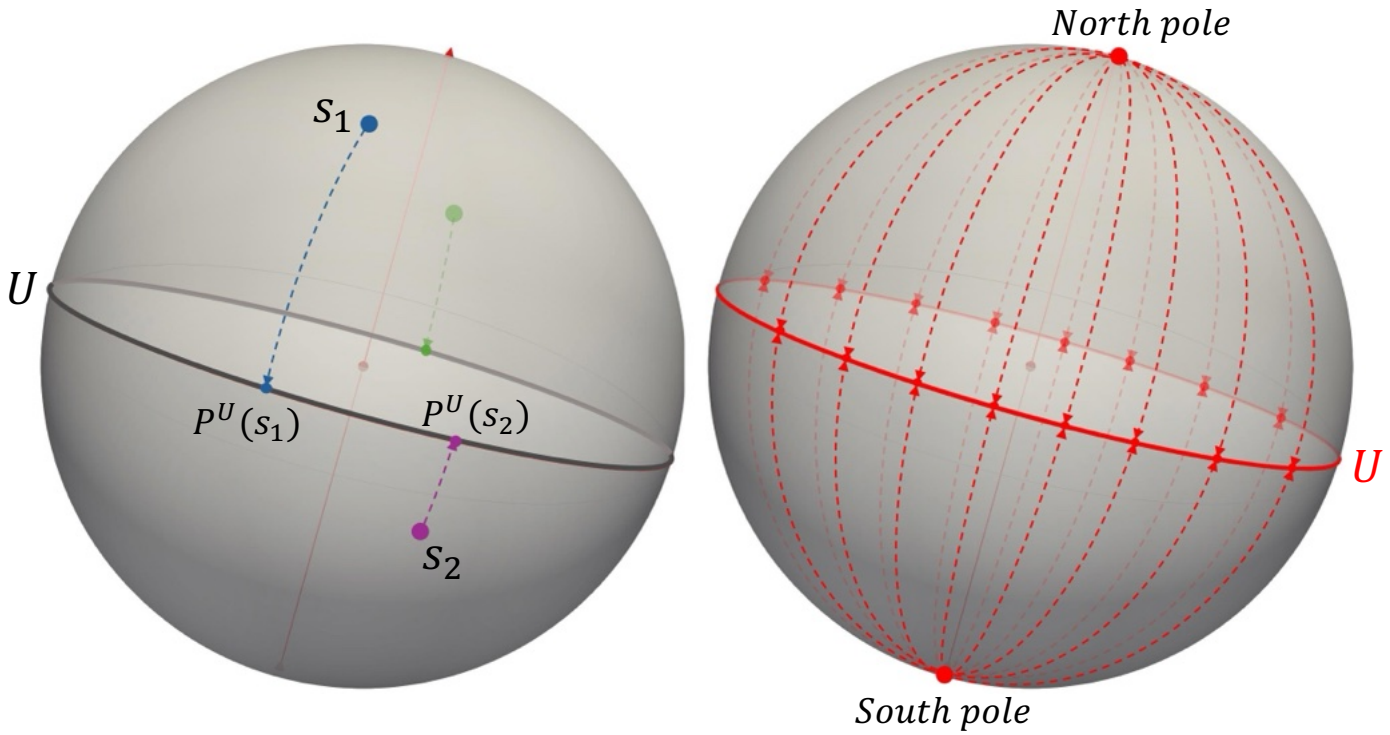
Global Historical Climatology Network



Defferrard, Michaël, et al. "Deepsphere: towards an equivariant graph-based spherical cnn."

Spherical Slicing via Spherical Radon Transform [Bonet et al. (2023)]

$$\mathcal{R}f(U,z) := \int_{\mathbb{S}^{d-1}} \delta(P^U(s) = z) f(s) ds$$



Circular Optimal Transport [Hundrieser et al. (2021)]

$$COT_h(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{S}^1\times\mathbb{S}^1} c(x,y)d\gamma(x,y)$$

General OT Formulation

$$= \inf_{x_0\in[0,1)} \int_0^1 h(|F^{-1}_{\mu,x_0}(x) - F^{-1}_{\nu,x_0}(x)|_{\mathbb{R}})\,dx$$

Optimal cutting position

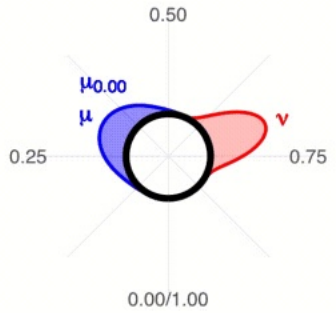
$$= \inf_{\alpha\in\mathbb{R}} \int_0^1 h(|F^{-1}_{\mu}(x) - F^{-1}_{\nu}(x-\alpha)|_{\mathbb{R}})\,dx$$

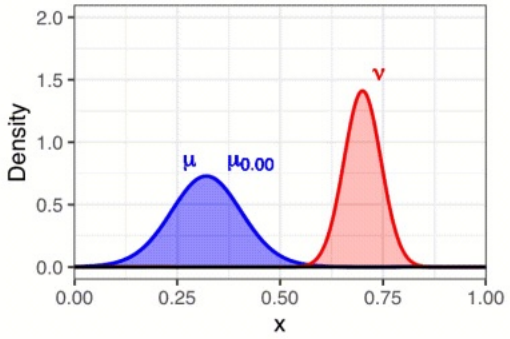
Equivalent definition with $\alpha^* = F_\mu(x_0^*) - F_\nu(x_0^*)$

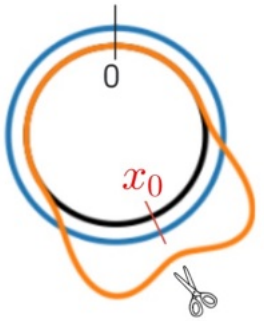
* $c(x,y) := h(|x-y|_{\mathbb{S}^1}), h : \mathbb{R}\to\mathbb{R}_+$ convex and increasing

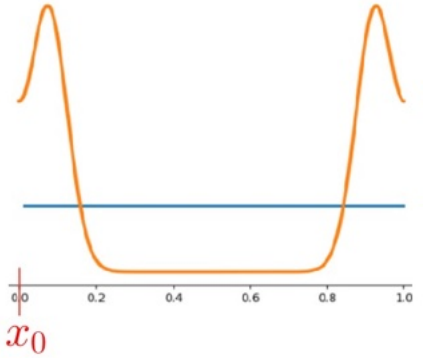
** if $\mu = Unif(\mathbb{S}^1)$ and $h(x) = |x|^2$, $\alpha^* = x_0^* - F_\nu(x_0^*) = \mathbb{E}(\nu) - 1/2$
(Computation Benefits!)

Source: https://stochastik.math.uni-goettingen.de/cot/









Linear Circular Optimal Transport

Rocío Díaz Martín, Ivan Medri, Yikun Bai, Xinran Liu, Kangbai Yan, Gustavo K. Rohde, and Soheil Kolouri.
LCOT: Linear circular optimal transport. ICLR 2024.

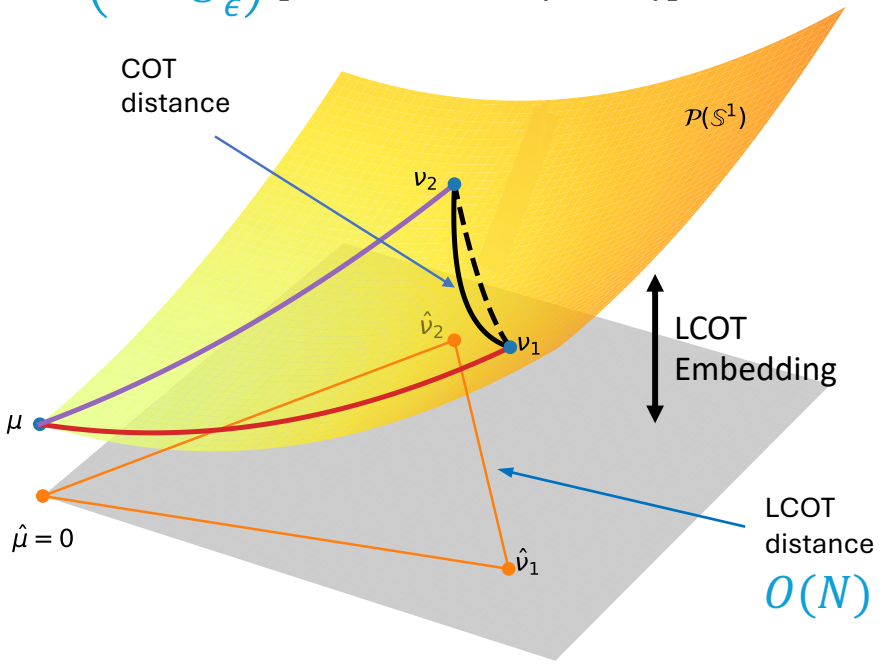
- $\mu = Unif(\mathbb{S}^1)$ –reference measure
- $h(x) = |x|^2$, – quadratic cost
- Linear Circular Optimal Transport (LCOT) Embedding

$$\widehat{\nu}(x) := F_\nu^{-1}\left(x - \mathbb{E}(\nu) + \frac{1}{2}\right) - x, \qquad x \in [0,1) \quad O(N)$$

- Linear Circular Optimal Transport (LCOT) distance

$$LCOT_2(\nu_1, \nu_2) = \|\widehat{\nu_1} - \widehat{\nu_2}\|_{L^2(\mathbb{S}^1)}$$

$O\left(N \log \frac{1}{\epsilon}\right)$ [Delon et al. (2010)]



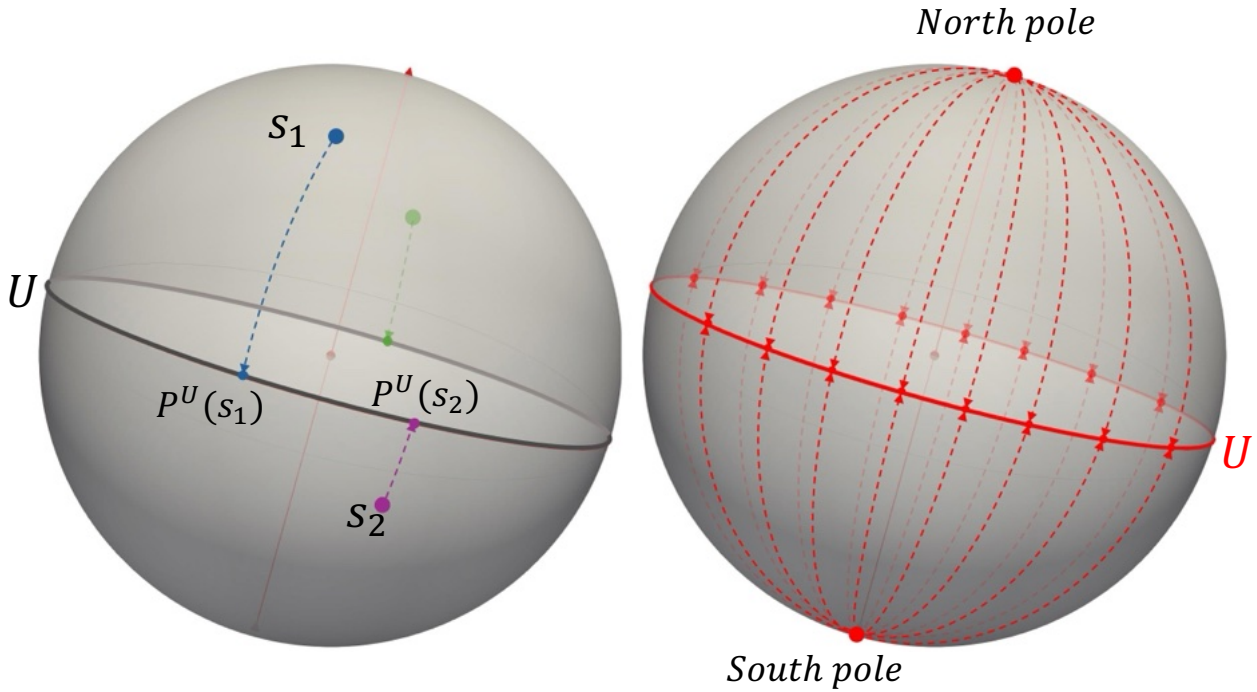
Method: Linear Spherical Sliced Optimal Transport (LSSOT)

- LSSOT Embedding

$$\widehat{\nu_i}^S(x,U) := F^{-1}_{P^U_\# \nu_i}\left(P^U(x) - \mathbb{E}(P^U_\# \nu_i) + \frac{1}{2}\right), \quad x \in \mathbb{S}^{d-1}$$

- LSSOT distance

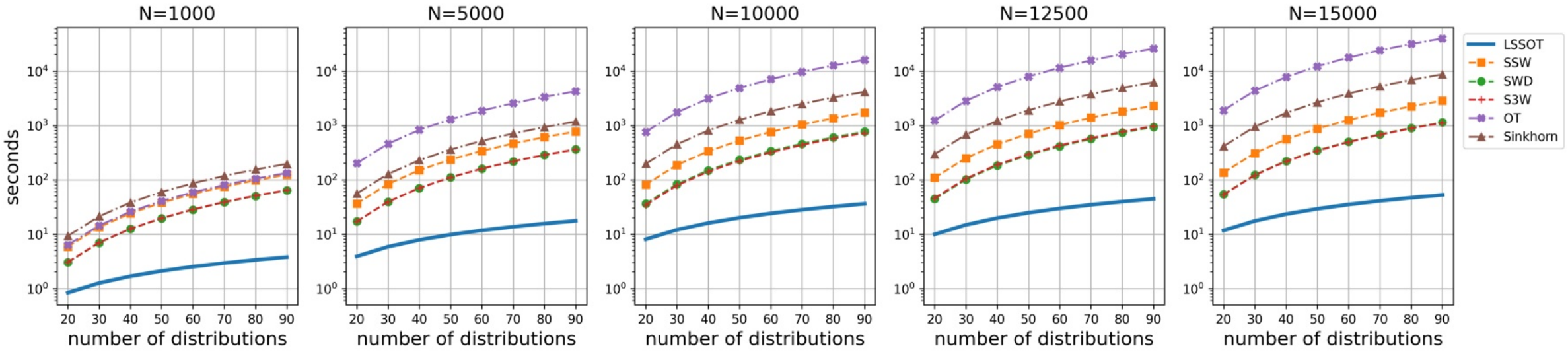
$$\begin{aligned}(LSSOT_2(\nu_1,\nu_2))^2 &:= \int_{V_2(\mathbb{R}^d)} \left(LCOT_2(P^U_\#\nu_1, P^U_\#\nu_2)\right)^2 d\sigma(U) \\ &= \int_{V_2(\mathbb{R}^d)} \|\widehat{\nu_1}^S(\cdot,U) - \widehat{\nu_2}^S(\cdot,U)\|^2_{L^2(\mathbb{S}^1)}\, d\sigma(U)\end{aligned}$$



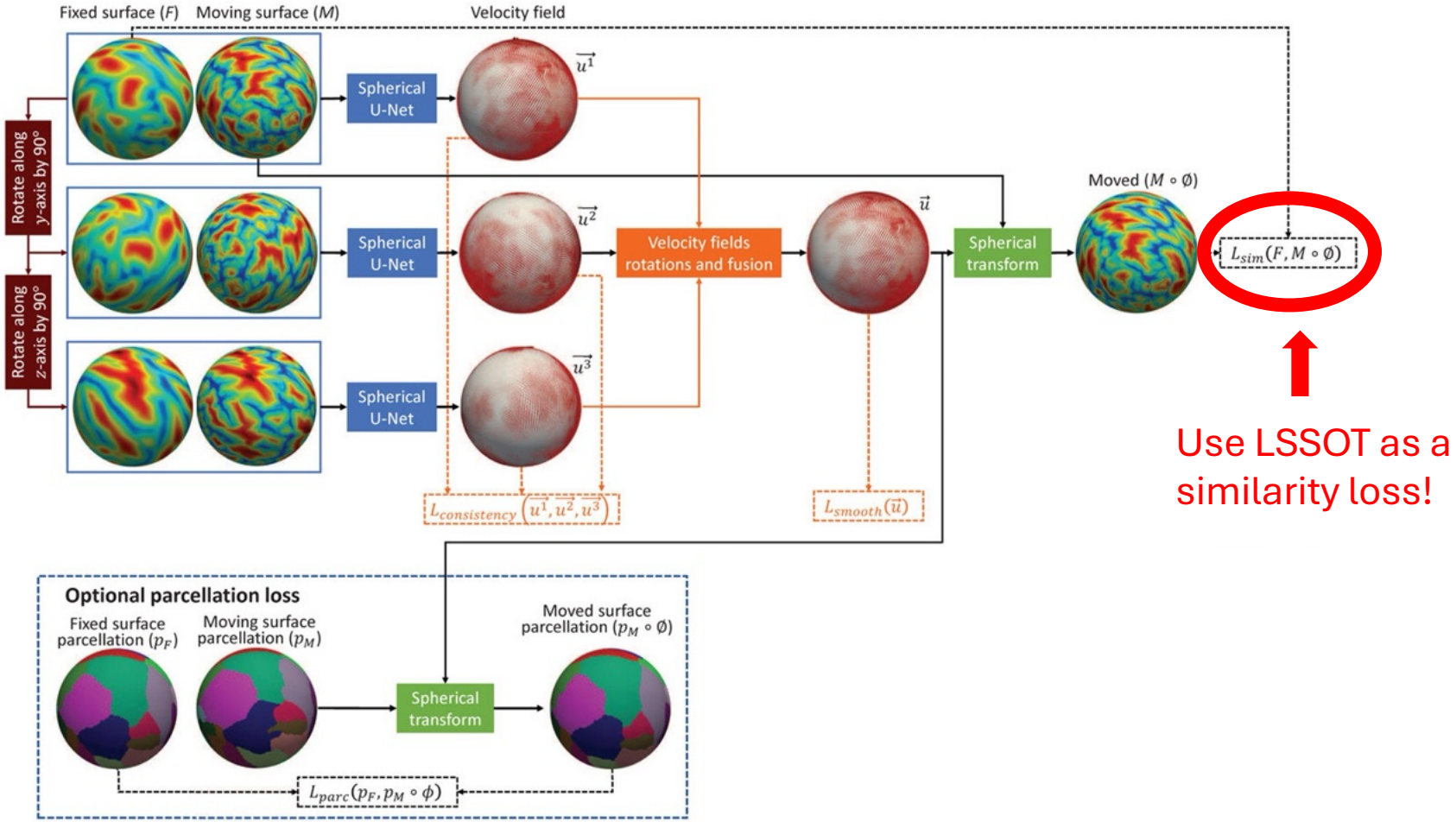
Computation Efficiency

Method	Complexity	Pairwise Distance Calculation Complexity
LSSOT	$\mathcal{O}(LN(d+\log N+1))$	$\mathcal{O}(KLN(d+\log N+1+K))$
SSW	$\mathcal{O}(LN(d+\log N+\log(\frac{1}{\epsilon})))$	$\mathcal{O}(K^2LN(d+\log N+\log(\frac{1}{\epsilon})))$
OT	$\mathcal{O}(N^2(N\log N+d))$	$\mathcal{O}(K^2N^2(N\log N+d)$
Sinkhorn	$\mathcal{O}(N^2(\frac{\log N}{\epsilon^2}+d))$	$\mathcal{O}(K^2N^2(\frac{\log N}{\epsilon^2}+d))$
S3W	$\mathcal{O}(LN(d+\log N))$	$\mathcal{O}(K^2LN(d+\log N))$
SWD	$\mathcal{O}(LN(d+\log N))$	$\mathcal{O}(K^2LN(d+\log N))$

L: number of slices; N: number of samples; K: number of distributions



Application. Cortical Surface Registration

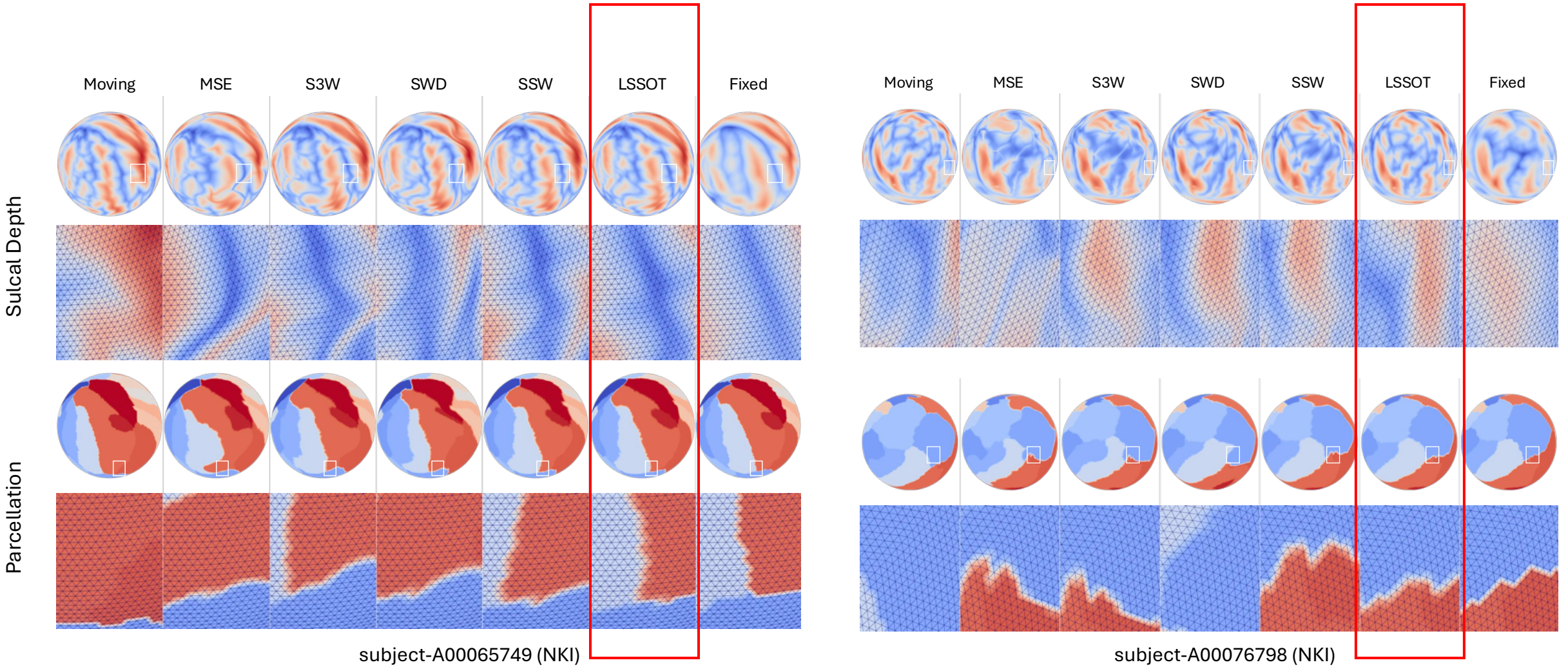


Zhao, Fenqiang, Zhengwang Wu, Fan Wang, Weili Lin, Shunren Xia, Dinggang Shen, Li Wang, and Gang Li. "S3Reg: superfast spherical surface registration based on deep learning." *IEEE transactions on medical imaging* 40, no. 8 (2021): 1964-1976.

Application. Cortical Surface Registration

		Freesurfer	MSE	S3W	SWD	SSW	LSSOT (Ours)
NKI (Left Hemisphere)	LSSOT(↓)	0.2877±0.0392	0.2754±0.0611	0.2298±0.0346	0.2411± 0.0366	**0.2079±0.0369**	**0.1890±0.0361**
	SWD(↓)	0.0060±0.0020	**0.0051±0.0017**	0.0053±0.0010	**0.0027±0.0011**	0.0059±0.0011	0.0052±0.0011
	MAE(↓)	**0.1053±0.0214**	**0.1129±0.0471**	0.2278±0.0372	0.2658±0.0410	0.2145±0.0266	0.2516±0.0397
	CC(↑)	0.8190±0.0264	0.8269±0.0425	**0.9216±0.0174**	**0.8722±0.0302**	0.8671±0.0314	0.8649±0.0295
	Dice (↑)	0.7541±0.0651**	0.7746±0.0861**	**0.8498±0.0670****	0.7984±0.0539**	0.8429±0.0692*	**0.8462±0.0548**
	Edge Dist.(↓)	0.3207±0.0436**	0.3060±0.0404**	0.3922±0.0647**	0.3442±0.0346**	**0.2476±0.0348****	**0.2365±0.0343**
	Area Dist.(↓)	0.4652±0.0698**	0.4305±0.0488**	0.4048±0.0752**	0.4073±0.0368**	**0.2733±0.0402****	**0.2897±0.0396**
	Time(seconds)(↓)	—	**73.07**	121.00	118.96	1350.96	**101.01**

Application. Cortical Surface Registration




Questions and Comments
Thank you!

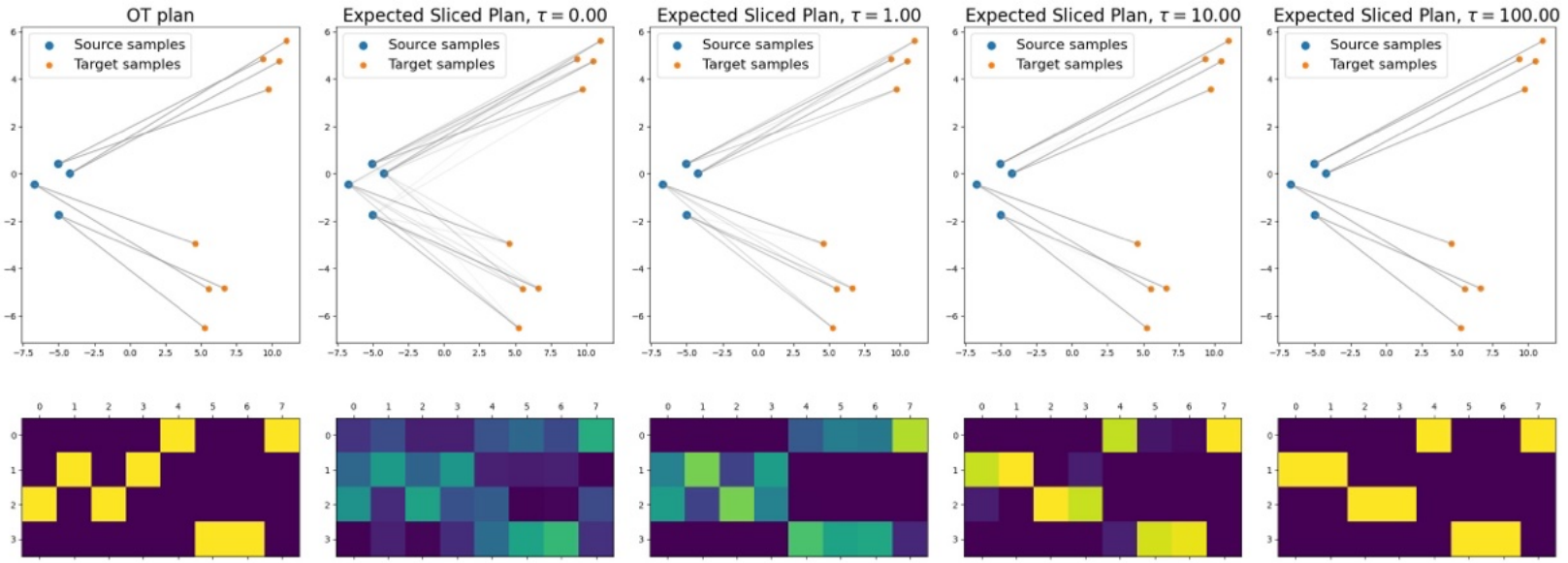
Choosing the most informative slices : Temperature Approach

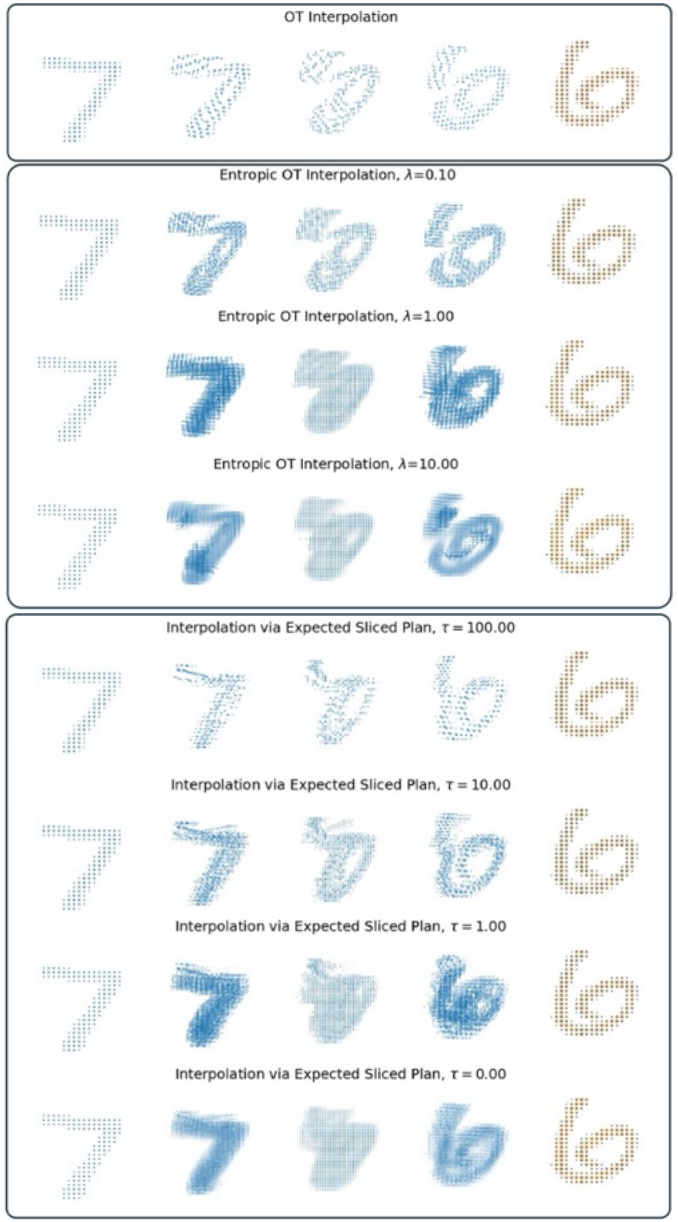
$\mathcal{D}_p^p(\mu, \nu; \theta)$: transport cost by the lifted plan from direction θ.

$$W_p(\mu, \nu) \leq \mathcal{D}_p(\mu, \nu; \theta), \forall \theta \in \mathbb{S}^{d-1}$$

$$\mathrm{d}\sigma_\tau(\theta) = \frac{e^{-\tau \mathcal{D}_p^p(\mu,\nu;\theta)}}{\int_{\mathbb{S}^{d-1}} e^{-\tau \mathcal{D}_p^p(\mu,\nu;\theta')} d\theta'} \mathrm{d}\theta$$

– The closer to the Optimal Plan, the better

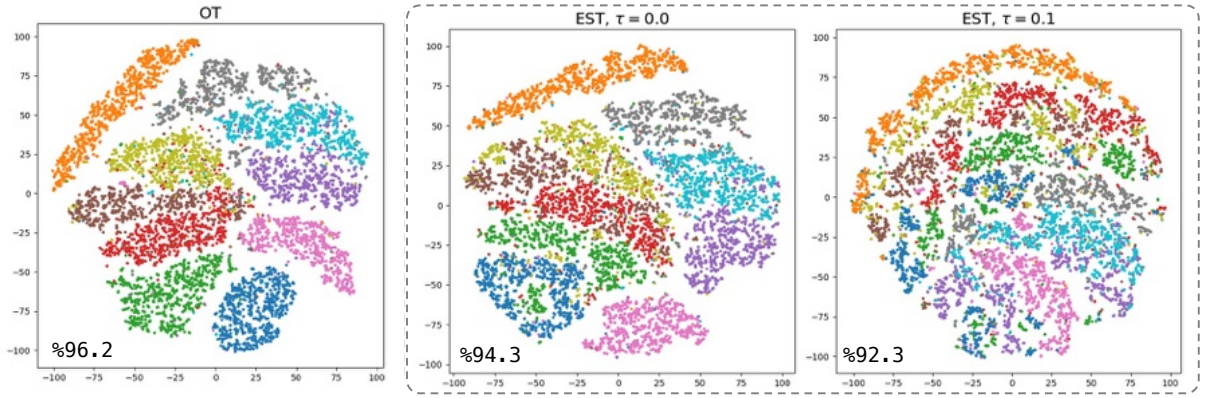




Experiments : Transport-based Embedding

- Transport Plan
- Transport Map (Linear Embedding)
- Logistic Regression / 1Nearest Neighbor Classifications

Point Cloud MNIST 2D

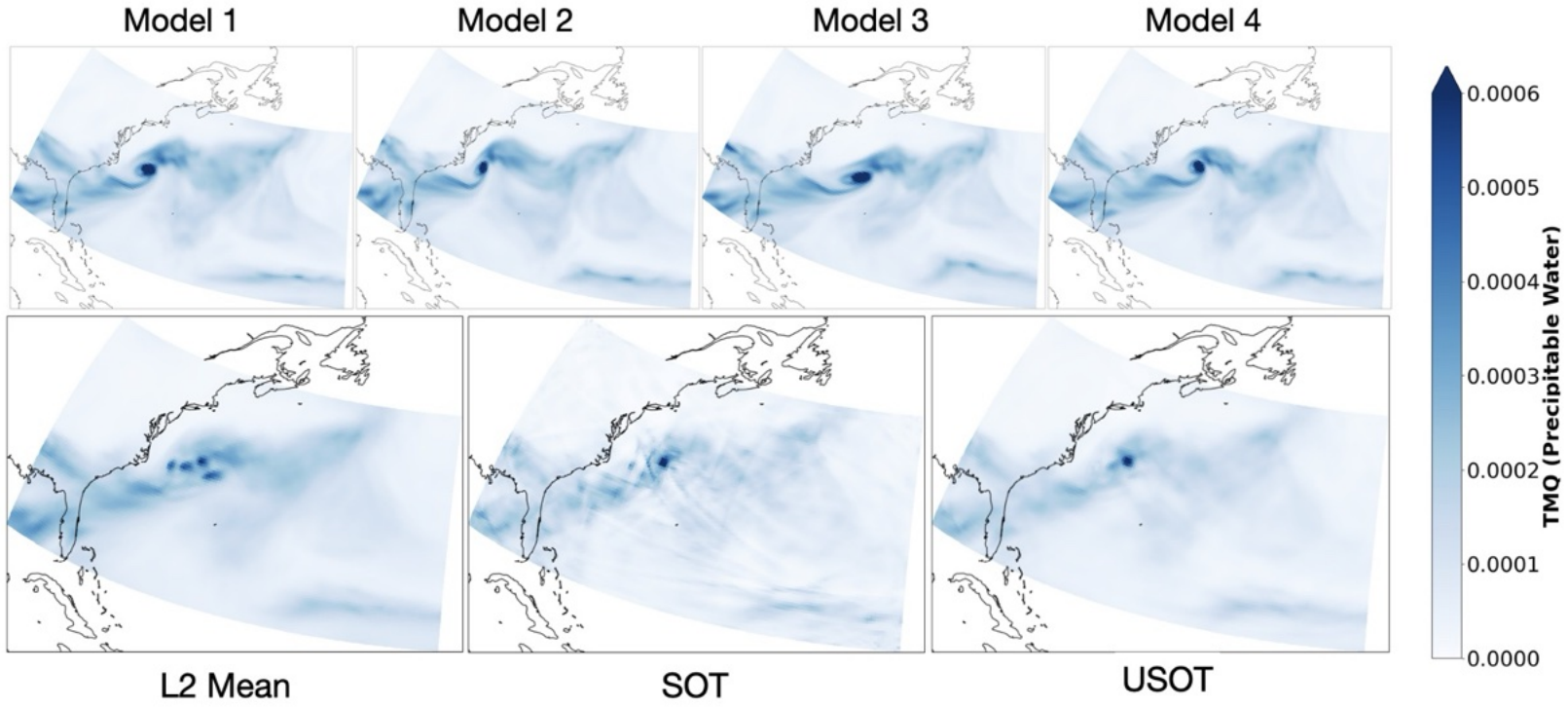


ModelNet40

1NN Classification	Sinkhorn $\lambda = 10$	Sinkhorn $\lambda = 1$	ESP $L = 128$	ESP $L = 1024$	OT (LP)
Accuracy ↑	65.96%	78.93%	77.30%	79.45%	82.09%
Time per distance (Sec) ↓	0.423	0.594	0.185	0.368	0.883

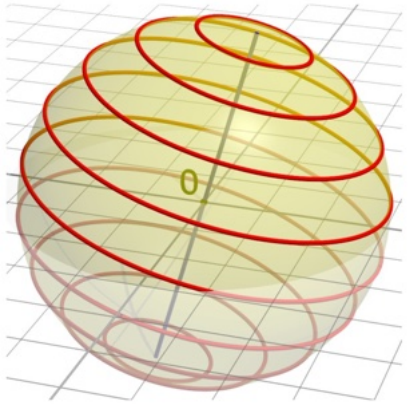
Future Direction II : Sliced Spherical Barycenters

Inspiration : OT barycenters improves ensembles of Climate Models

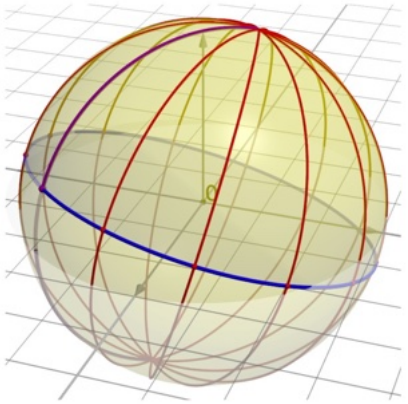


Séjourné, T., Bonet, C., Fatras, K., Nadjahi, K., & Courty, N. (2023).
Unbalanced optimal transport meets sliced-Wasserstein

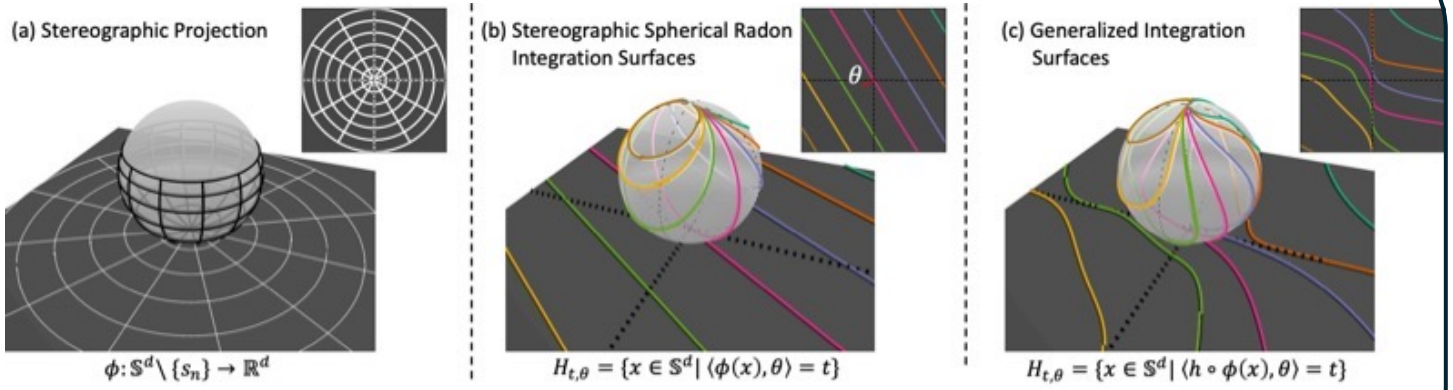
Future Direction II : Sliced Spherical Barycenters



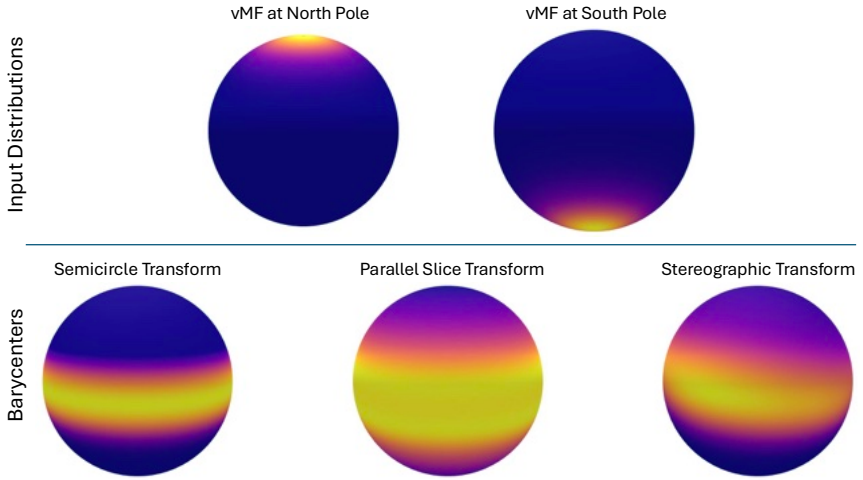
Parallel Slice Transform



Semicircle Transform



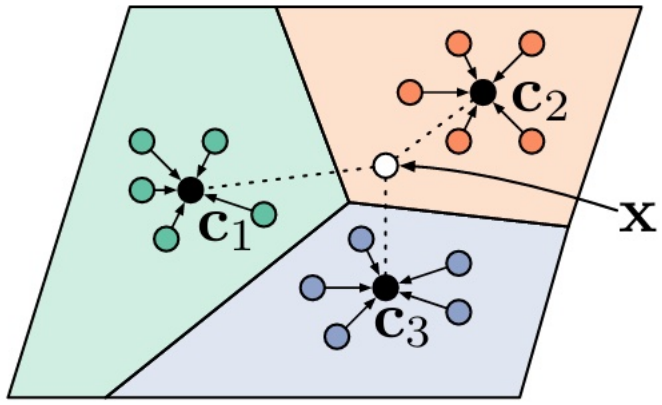
Stereographic Spherical Radon Transform



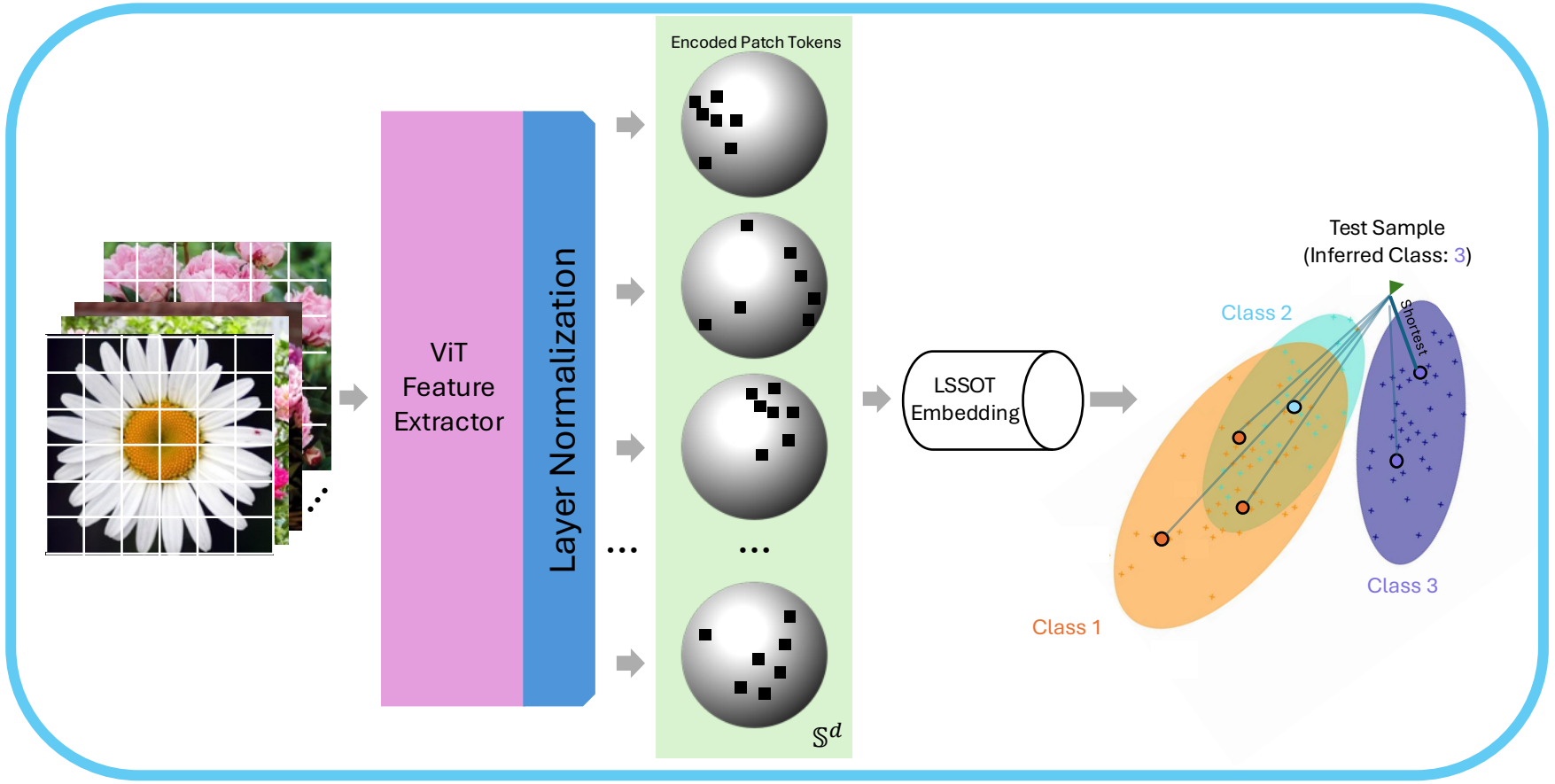
Future Direction III : Prototypical Networks

Inspirations:

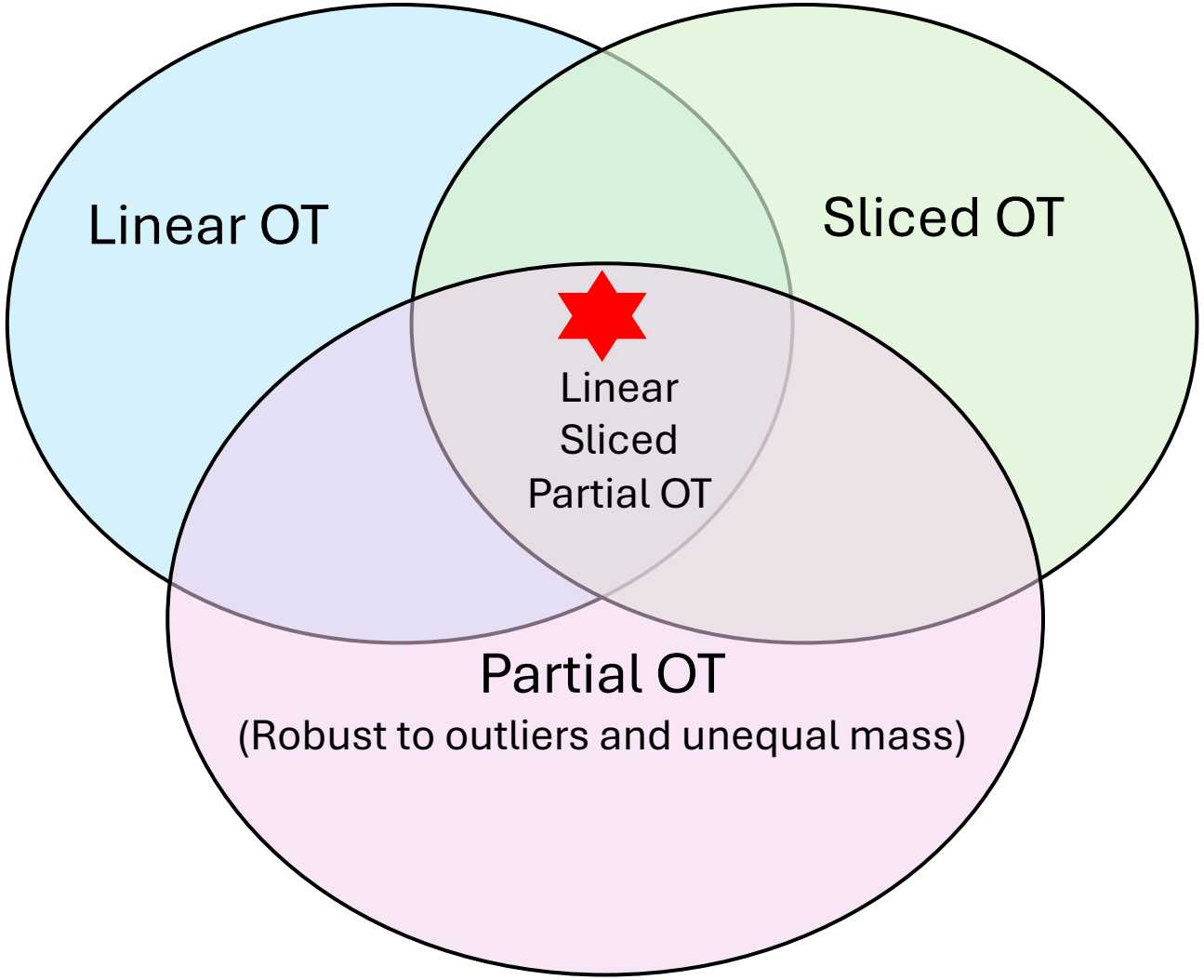
- Pooling strategy might yield better results than using only the classifier token
- Layer Normalization maps features on hyperspheres



Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." *Advances in neural information processing systems* 30 (2017)



Future Direction IV : Linear Sliced Partial Transport Embedding



LOT Embedding Approximation Error

Moosmüller and Cloninger, 2021

How well are we approximating the 2-Wasserstein distance?

If μ_i and μ_j are pushforwards of a fixed measure, μ, under shifts and scalings then the embedding is isometric.

Let $\boldsymbol{\mathcal{E}}$ identify the set of all shifts and scalings then:

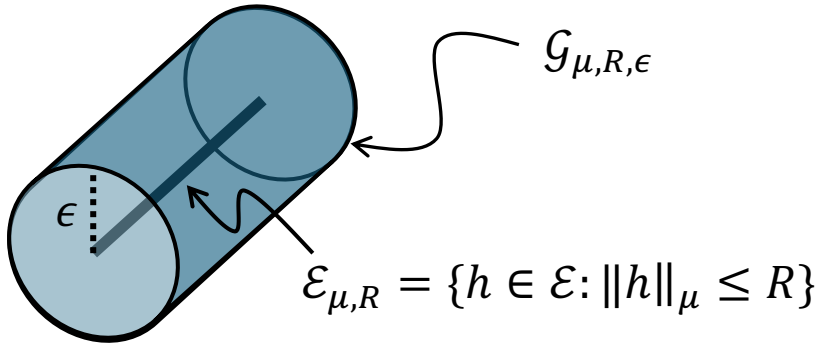
$$\|\phi(g_{1\#}\mu) - \phi(g_{2\#}\mu)\|_{\boldsymbol{\mu}} = \mathcal{W}_{\mathbf{2}}(g_{1\#}\mu, g_{2\#}\mu)\,, \quad \forall g_i \in \boldsymbol{\mathcal{E}}$$

For $R, \epsilon > 0$, define $\mathcal{E}_{\mu,R} = \{h \in \mathcal{E}: \|h\|_\mu \leq R\}$ and let define the ϵ −tube around the set of shifts and scalings as:

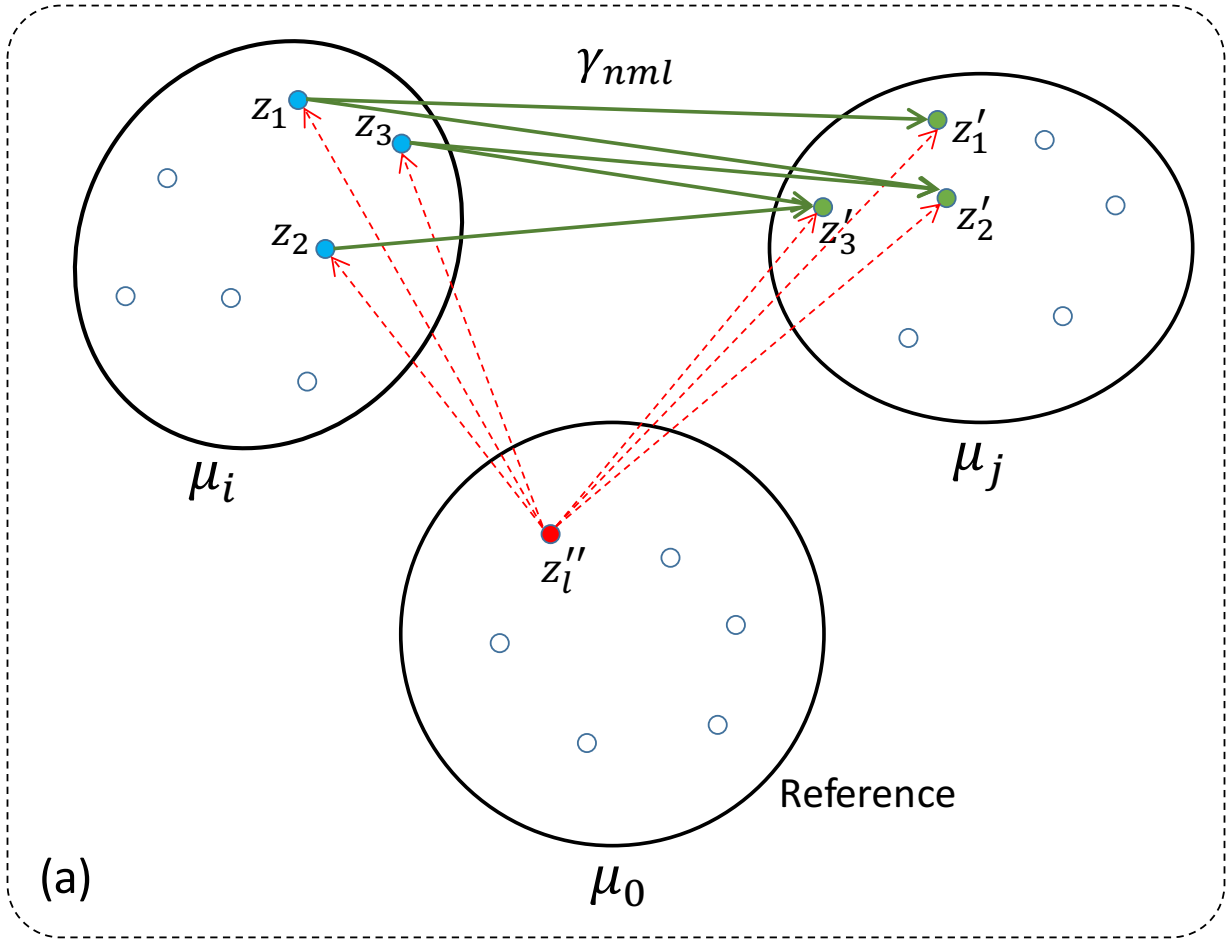
$$\mathcal{G}_{\mu,R,\epsilon} = \{g \in L^2(\mathcal{Z}, \mu): \exists h \in \mathcal{E}_{\mu,R}: \|g - h\|_\mu \leq \epsilon\}$$

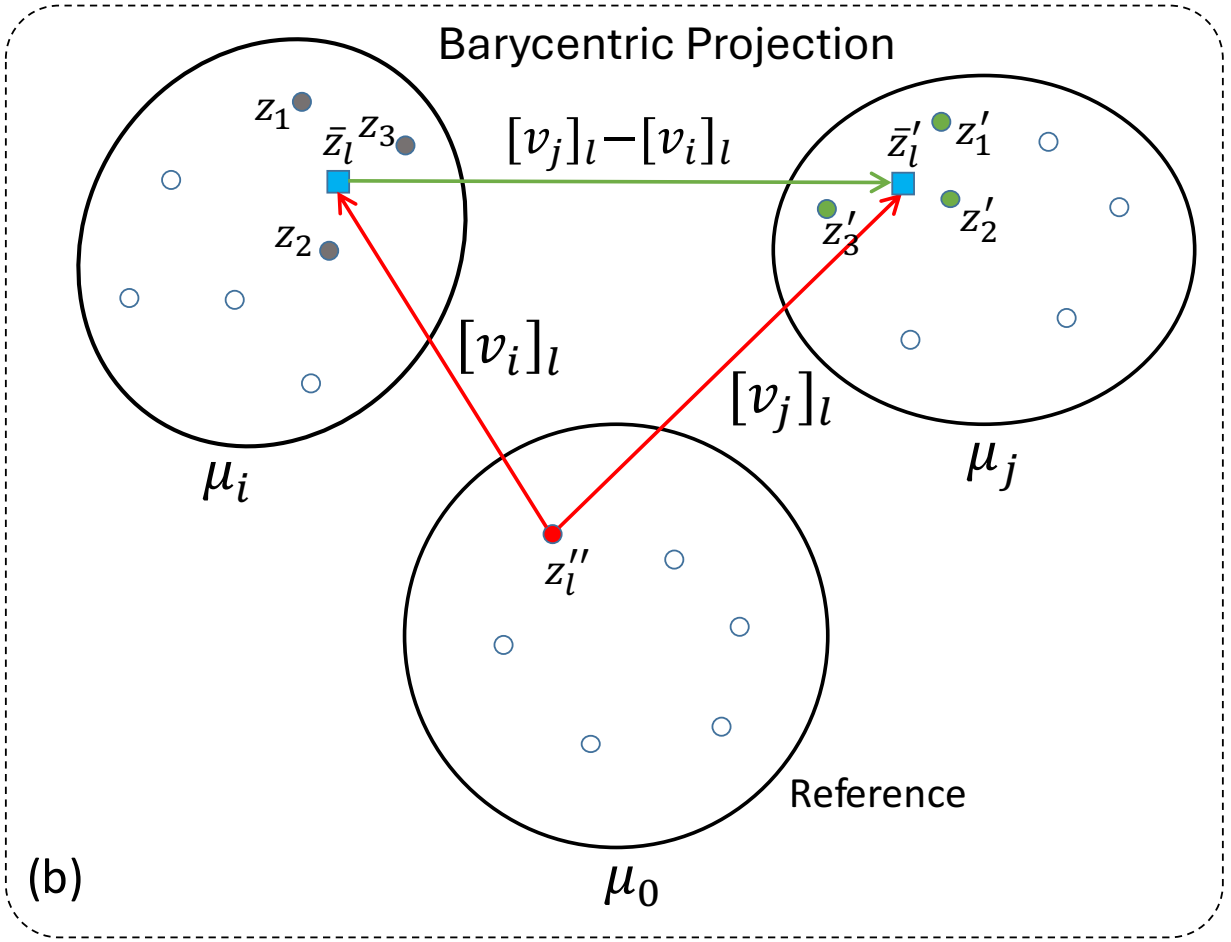
Then for $g_1, g_2 \in \mathcal{G}_{\mu,R,\epsilon}$:

$$0 \leq \|\phi(g_{1\#}\mu) - \phi(g_{2\#}\mu)\|_{\boldsymbol{\mu}} - \mathcal{W}_{\mathbf{2}}(g_{1\#}\mu, g_{2\#}\mu) \leq C_1\epsilon + C_2\epsilon^2$$



Mass Splitting is Resolved with Barycentric Projection





Why Optimal Transport : From an Optimization Perspective

- L_p-metrics ($p \geq 1$):

$$\left(\int_X |p(x) - q(x)|^p dx \right)^{\frac{1}{p}}$$

- Symmetric chi-squared distances:

$$\int_X \frac{2\big(p(x) - q(x)\big)^2}{q(x) + p(x)} dx$$

- Hellinger distance:

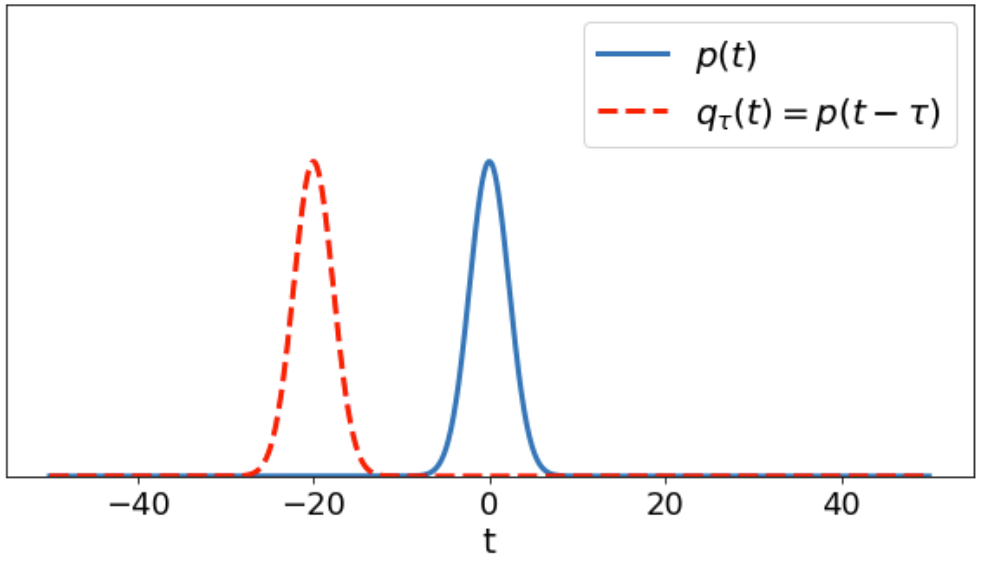
$$\left(\frac{1}{2} \int_X \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \right)^{\frac{1}{2}}$$

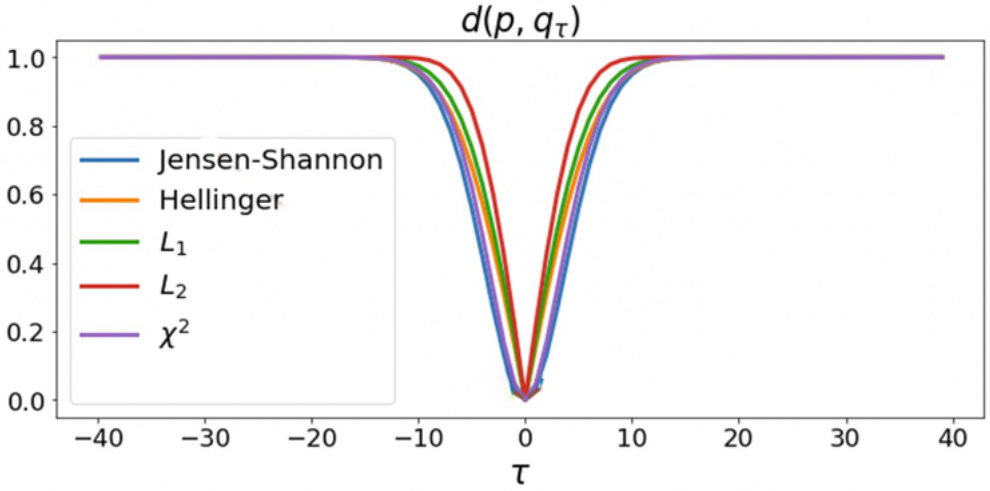
- Jensen-Shannon's divergence:

$$\left(-\frac{1}{2} \int_X \left(p(x) \log\left(\frac{m(x)}{p(x)} \right) + q(x) \log\left(\frac{m(x)}{q(x)} \right) \right) dx \right)^{\frac{1}{2}}$$

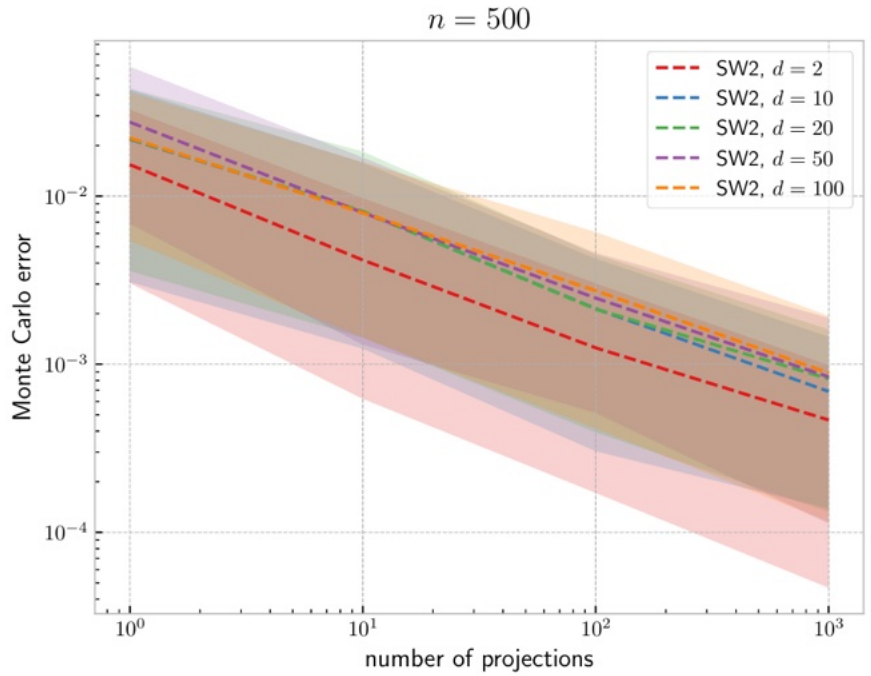
for $m(x) = \frac{p(x)+q(x)}{2}$

- ...

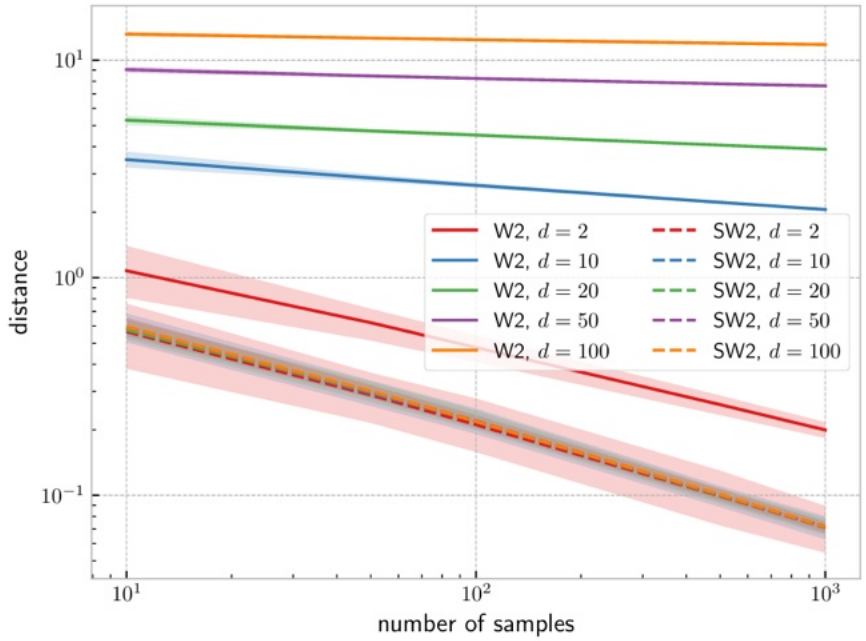




Statistical Benefits of Sliced Wasserstein



Projection complexity



Sample complexity

Nadjahi, Kimia, et al. "Statistical and topological properties of sliced probability divergences." *Advances in Neural Information Processing Systems* 33 (2020)

SWGG and min-SWGG

Mahey, Guillaume, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. "Fast optimal transport through sliced wasserstein generalized geodesics." (NeurIPS 2023).

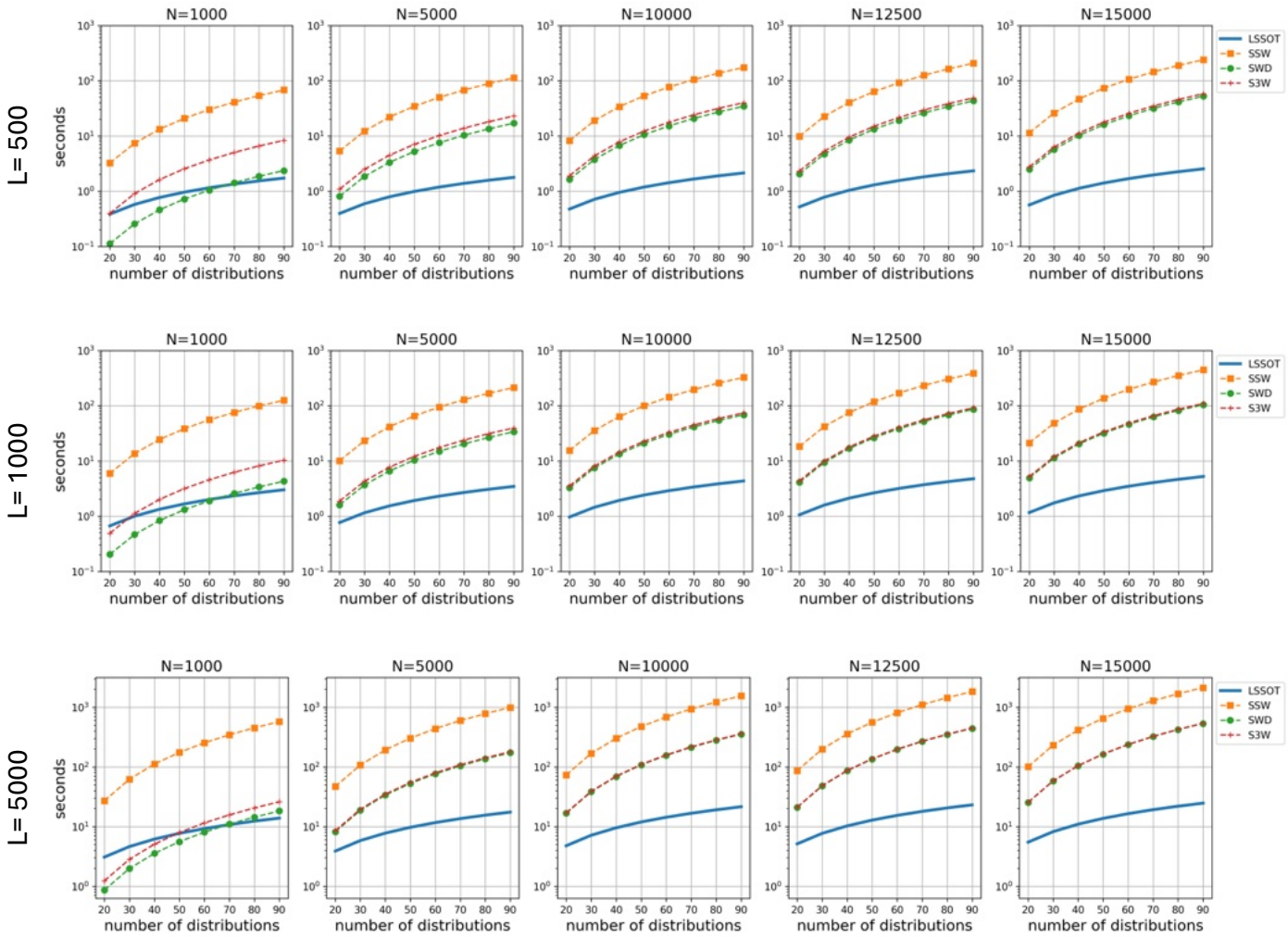
Definition 3.1 (SWGG and min-SWGG)**.** Let $\mu_1, \mu_2 \in \mathcal{P}_2^n(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$. Denote by σ_θ and τ_θ the permutations obtained by sorting the 1D projections $P^\theta_\# \mu_1$ and $P^\theta_\# \mu_2$. We define respectively SWGG and min-SWGG as:

$$\text{SWGG}_2^2(\mu_1, \mu_2, \theta) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_{\sigma_\theta(i)} - \boldsymbol{y}_{\tau_\theta(i)}\|_2^2, \tag{8}$$

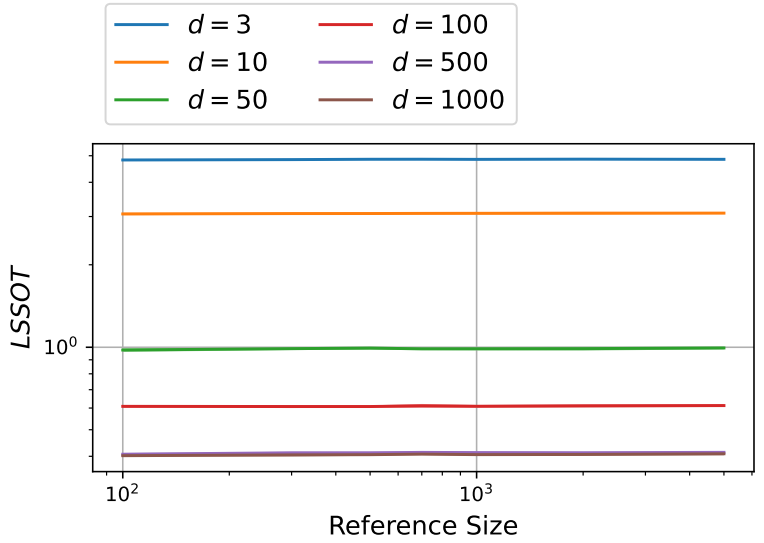
$$\text{min-SWGG}_2^2(\mu_1, \mu_2) \overset{\text{def}}{=} \min_{\theta \in \mathbb{S}^{d-1}} \text{SWGG}_2^2(\mu_1, \mu_2, \theta). \tag{9}$$

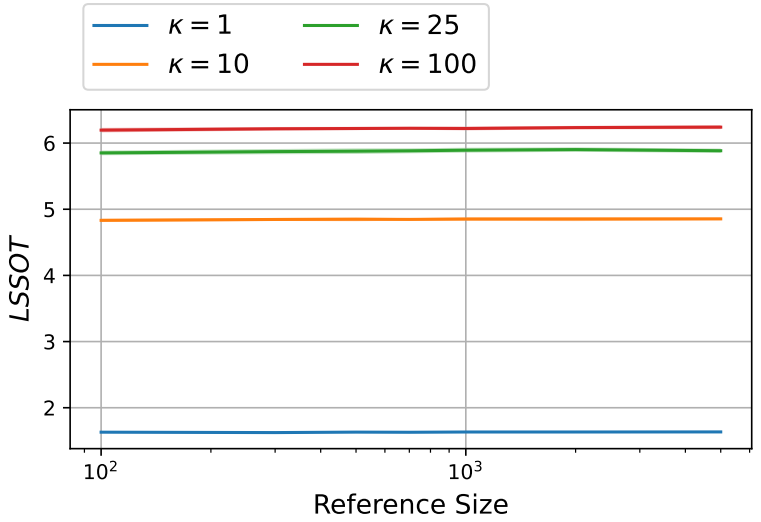
Computation Efficiency

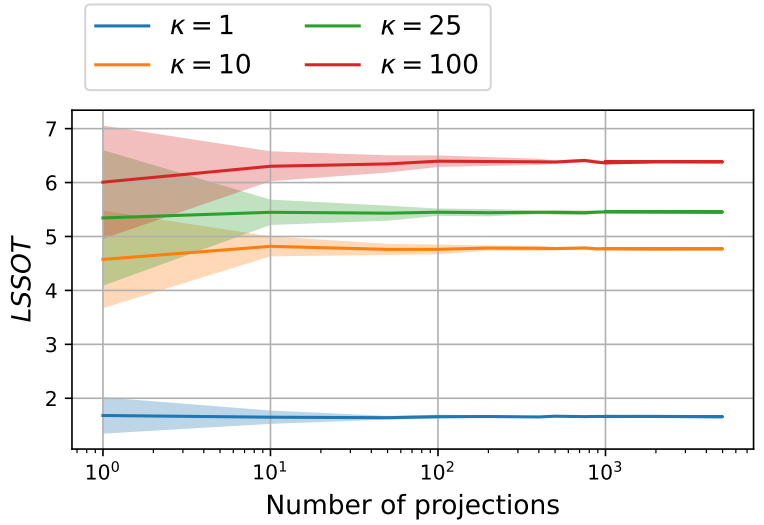
- L: number of slices; N: number of samples; K: number of distributions

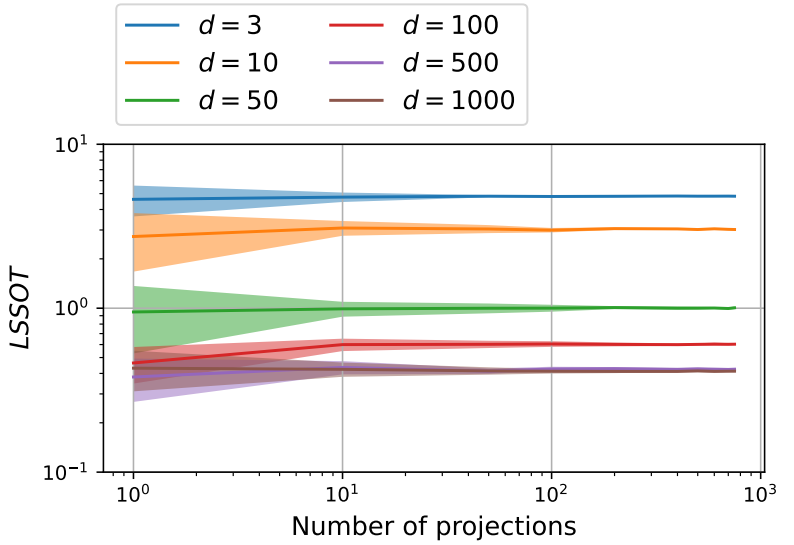


Sensitivity analysis

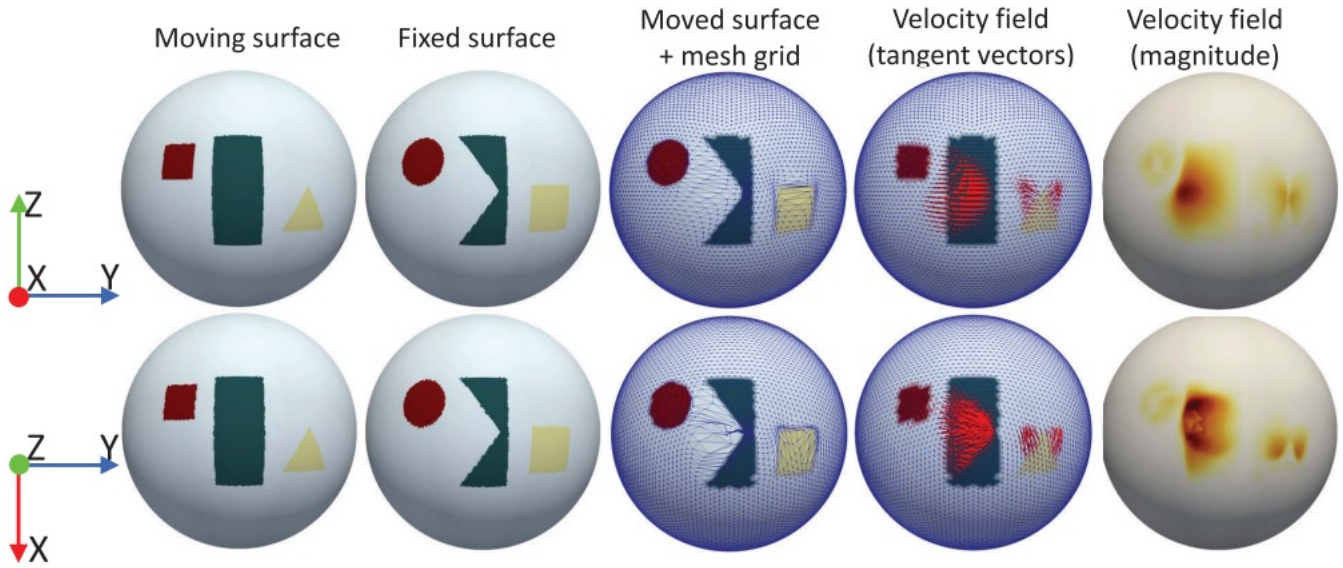








S3Reg



Distortion in poles

LSSOT Algorithm

Algorithm 1 Linear Spherical Sliced Optimal Transport (LSSOT)

Require: Spherical distributions $\mu = \sum_{i=1}^{N_\mu} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{N_\nu} b_j \delta_{y_j}$; number of slices L; reference size M, threshold ϵ

for $l = 1$ to L **do**

Construct a matrix $Z_l \in \mathbb{R}^{d\times 2}$ with entries randomly drawn from $\mathcal{N}(0,1)$

Apply QR decomposition on Z_l to get an orthogonal Q matrix U_l

Project $\{x_i\}_{i=1}^{N_\mu}$ and $\{y_j\}_{j=1}^{N_\nu}$ on $\mathbb{R}^2$ to get $\hat{x}_i^l = U_l^T x_i$ and $\hat{y}_j^l = U_l^T y_j, \forall i, j$

Find $I_l^* = \{i^*\}, J_l^* = \{j^*\}$ such that $\|\hat{x}_{i^*}^l\| \leq \epsilon$ or $\|\hat{y}_{j^*}^l\| \leq \epsilon$

$a_l^* = \sum_{i^*\in I^*} a_{i^*}, b_l^* = \sum_{j^*\in J^*} b_{j^*}; \tilde{N}_\mu = N_\mu - |I_l^*|, \tilde{N}_\nu = N_\nu - |J_l^*|$

} Apply ϵ-cap around poles

$\tilde{\mu}_l = \sum_{\tilde{i}=1}^{\tilde{N}_\mu} (a_{\tilde{i}} + \frac{a^*}{\tilde{N}_\mu}) \delta_{\hat{x}_{\tilde{i}}^l}, \tilde{\nu}_l = \sum_{\tilde{j}=1}^{\tilde{N}_\nu} (b_{\tilde{j}} + \frac{b^*}{\tilde{N}_\nu}) \delta_{\hat{y}_{\tilde{j}}^l}$

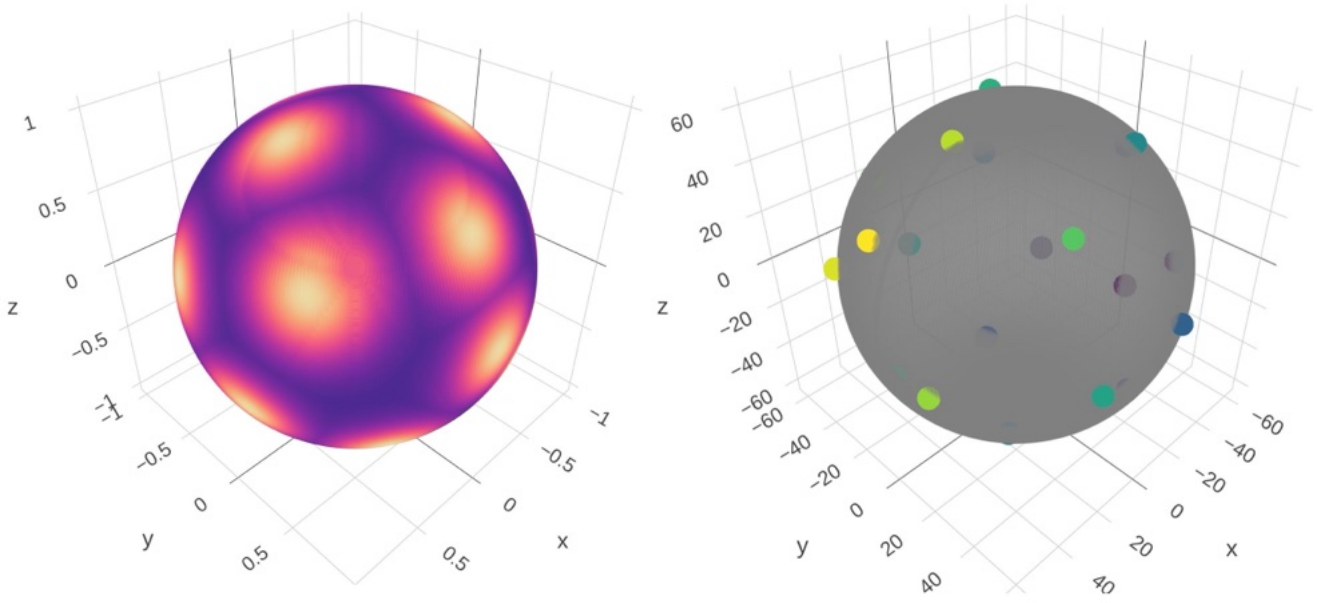
Project all $\hat{x}_{\tilde{i}}^l$ and $\hat{y}_{\tilde{j}}^l$ on $\mathbb{S}^1$: $P(\hat{x}_{\tilde{i}}^l) = \frac{\hat{x}_{\tilde{i}}^l}{\|\hat{x}_{\tilde{i}}^l\|}, P(\hat{y}_{\tilde{j}}^l) = \frac{\hat{y}_{\tilde{j}}^l}{\|\hat{y}_{\tilde{j}}^l\|}$

Calculate $LCOT_{\bar{\mu},2}(\tilde{\mu}_l^{proj}, \tilde{\mu}_l^{proj})$ by (3), where $\tilde{\mu}_l^{proj} = \sum_{\tilde{i}=1}^{\tilde{N}_\mu} (a_{\tilde{i}} + \frac{a^*}{\tilde{N}_\mu}) \delta_{P(\hat{x}_{\tilde{i}}^l)}$, $\tilde{\nu}_l^{proj} = \sum_{\tilde{j}=1}^{\tilde{N}_\nu} (b_{\tilde{j}} + \frac{b^*}{\tilde{N}_\nu}) \delta_{P(\hat{y}_{\tilde{j}}^l)}$, $\bar{\mu}$ is the discrete uniform reference measure of size M

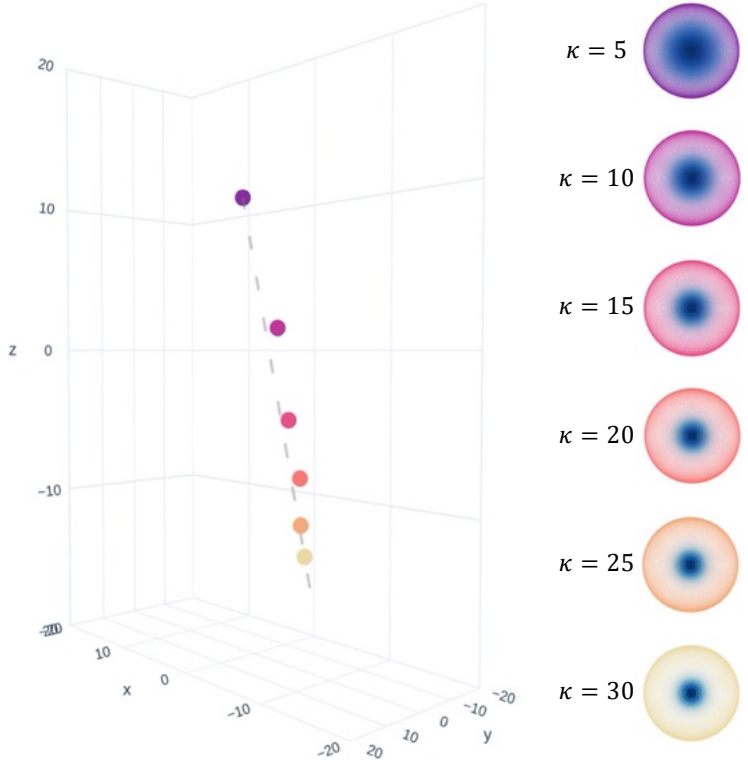
end for

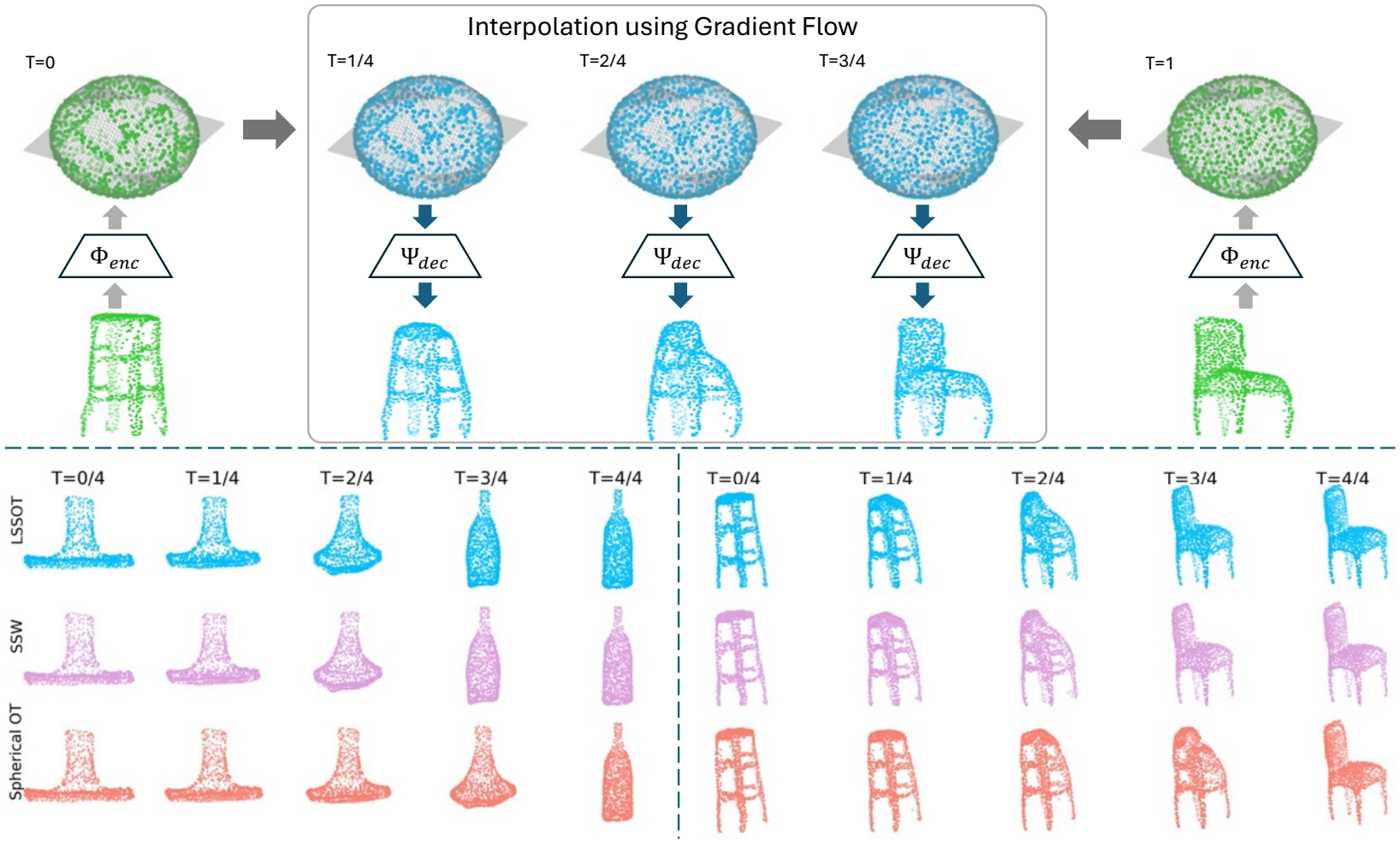
return $LSSOT_2(\mu,\nu) \approx \left(\frac{1}{L}\sum_{l=1}^{L} LCOT_2^2(\tilde{\mu}_l^{proj}, \tilde{\mu}_l^{proj})\right)^{\frac{1}{2}}$

LSSOT for Rotating VMFs



LSSOT for Scaling VMFs

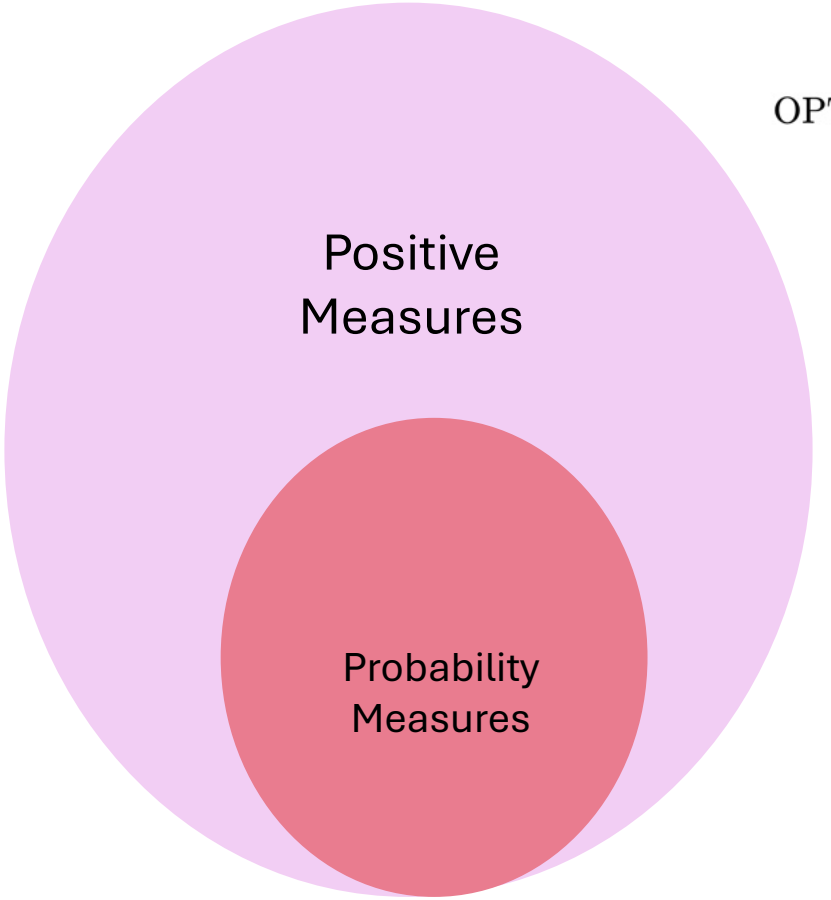



Interpolation using Gradient Flow
T=0
T=1/4
T=2/4
T=3/4
T=1
Φ_{enc}
Ψ_{dec}
Ψ_{dec}
Ψ_{dec}
Φ_{enc}
T=0/4
T=1/4
T=2/4
T=3/4
T=4/4
T=0/4
T=1/4
T=2/4
T=3/4
T=4/4
LSSOT
SSW
Spherical OT

Complexity of EST

- For each one of the L directions, $\{\theta_l\}_{l=1}^{L}$, we have to project the locations $\{x_i\}_{i=1}^{N}$, $\{y_j\}_{j=1}^{N}$ to obtain the new projected measures concentrated at $\{\theta_l \cdot x_i\}_{i=1}^{N}$ and $\{\theta_l \cdot y_j\}_{j=1}^{N}$, respectively. This requires $\mathcal{O}(LdN)$ operations.
- To compute the one-dimensional plan $\Lambda_{\theta_l}^{\mu,\nu}$, for each $L \leq \ell \leq L$, we essentially need to solve a one-dimensional optimal transport problem (i.e., a sorting problem), which is of order $\mathcal{O}(N \log(N))$. Thus, when considering all the slices, this step is of order $\mathcal{O}(LN \log(N))$.
- The lifting process that gives rise to $\gamma_{\theta_l}^{\mu,\nu}$ does not require additional operations to be taken into account (it is performed as an assignment or correspondence).
- The plan $\bar{\gamma}^{\mu,\nu}$ can be represented as an $N \times N$ matrix. The (i,j)-entry is given by $\sum_{l=1}^{L} \gamma_{\theta_l}^{\mu,\nu}(\{(x_i, y_j)\})$, requiring L operations. Thus, the complexity of this step is $\mathcal{O}(LN^2)$.
- Finally, once we have $\bar{\gamma}^{\mu,\nu}$, for computing the EST-distance we require another $\mathcal{O}(N^2 d)$ operations.

Optimal Partial Transport



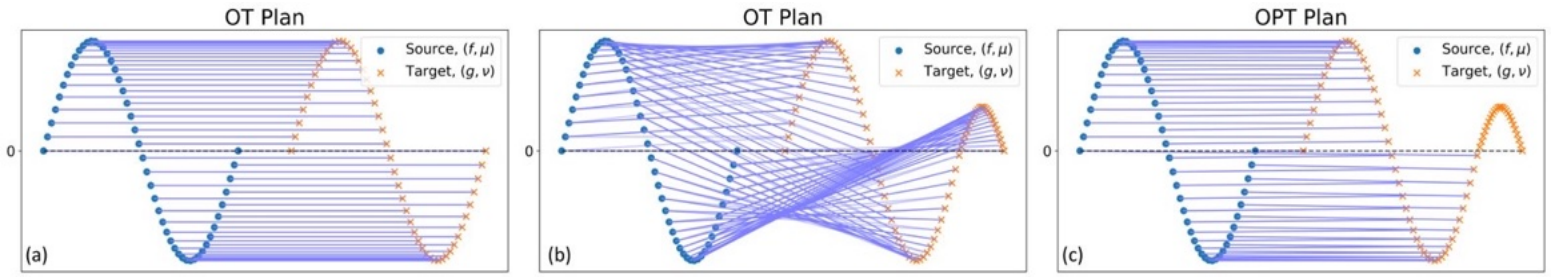
Penalty for mass destruction at the source

Penalty for mass creation at the target

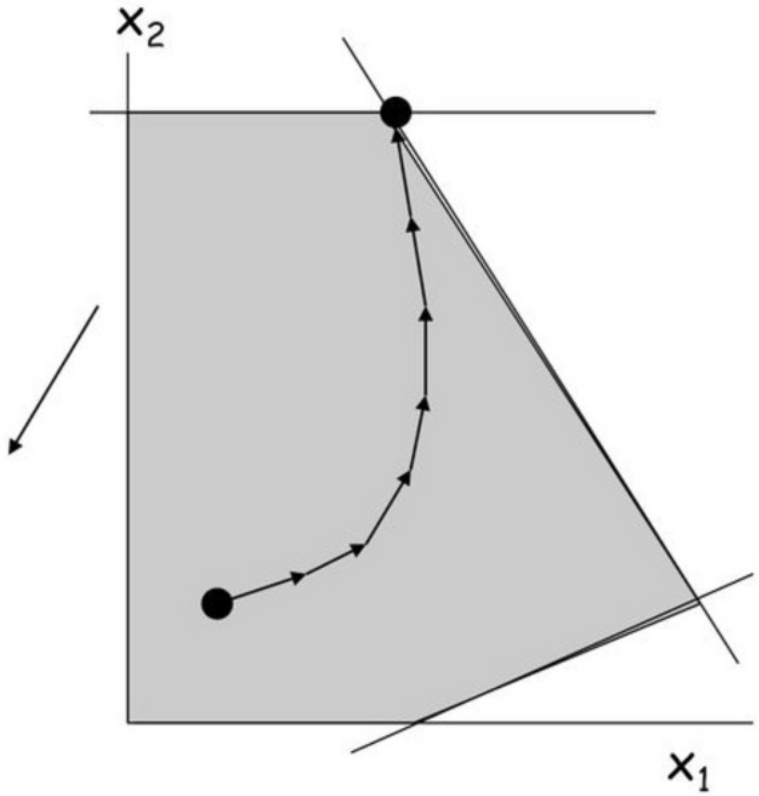
$$\mathrm{OPT}_{\lambda_1,\lambda_2,c}(\mu,\nu) := \inf_{\gamma\in\Pi_{\leq}(\mu,\nu)} \int_{\Omega^2} c(x,y)\,\mathrm{d}\gamma(x,y) + \lambda_1(\|\mu\|_{\mathrm{TV}} - \|\pi_{1\#}\gamma\|_{\mathrm{TV}}) + \lambda_2(\|\nu\|_{\mathrm{TV}} - \|\pi_{2\#}\gamma\|_{\mathrm{TV}})$$

where $\Pi_{\leq}(\mu,\nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : \pi_{1\#}\gamma \leq \mu, \pi_{2\#}\gamma \leq \nu\}$

When $\lambda_1 = \lambda_2$, and the transportation cost $c(x,y)$ is a metric, $\mathrm{OPT}_{\lambda,c}(\cdot,\cdot)$ defines a metric on $\mathcal{M}_+(\Omega)$.



Interior Point Methods



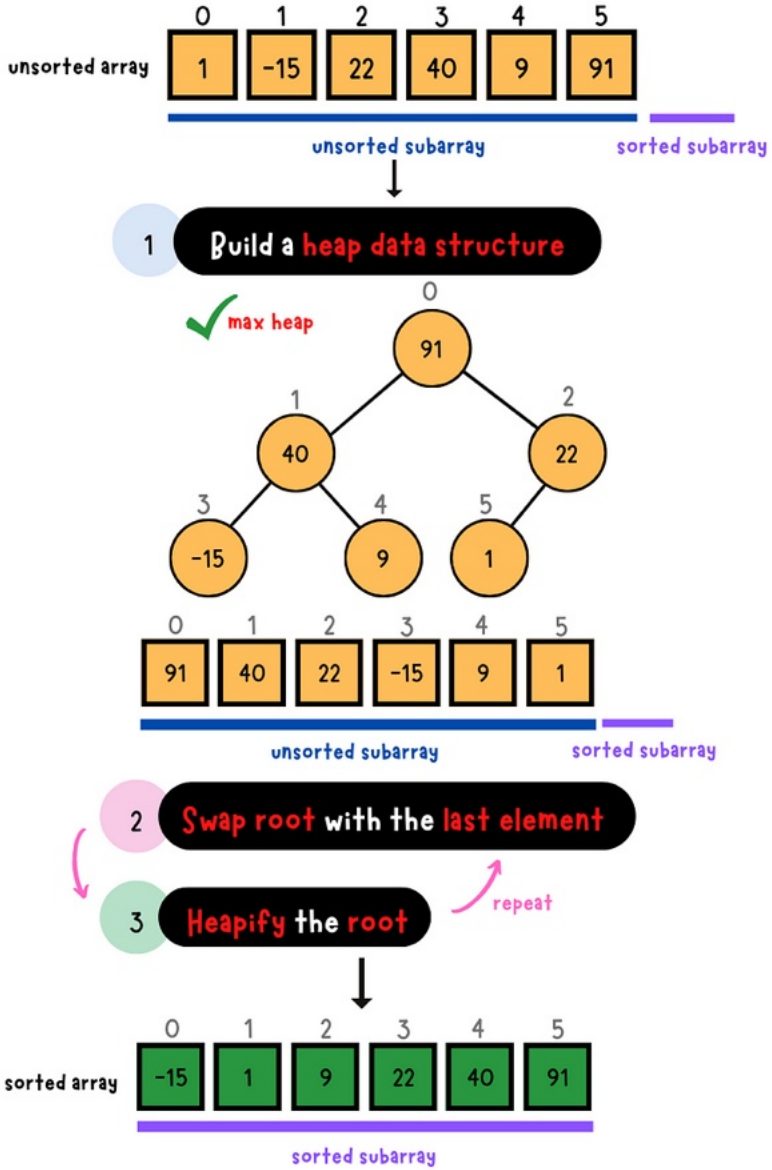
Travel through the interior with a **combination** of

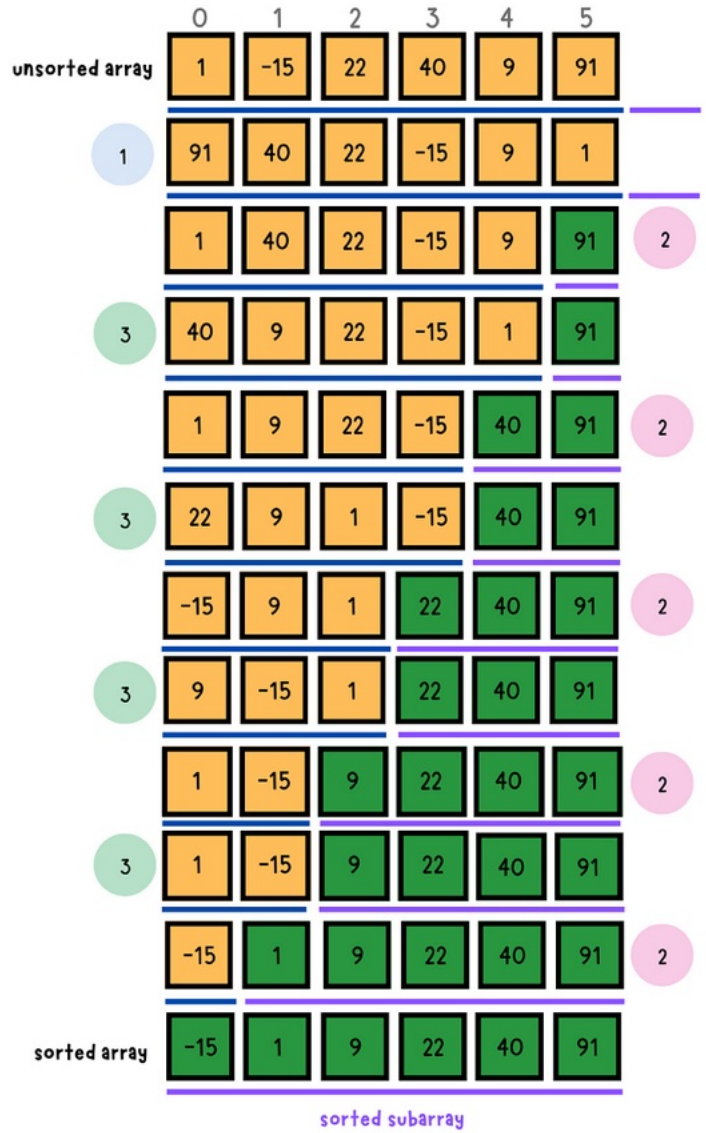
1. An **optimization** term (moves toward objective)
2. A **centering** term or constraint (keeps away from boundary)

Used since 50s for nonlinear programming.

Karmakar proved a variant is **polynomial time** in 1984

Heap Sort Algorithm

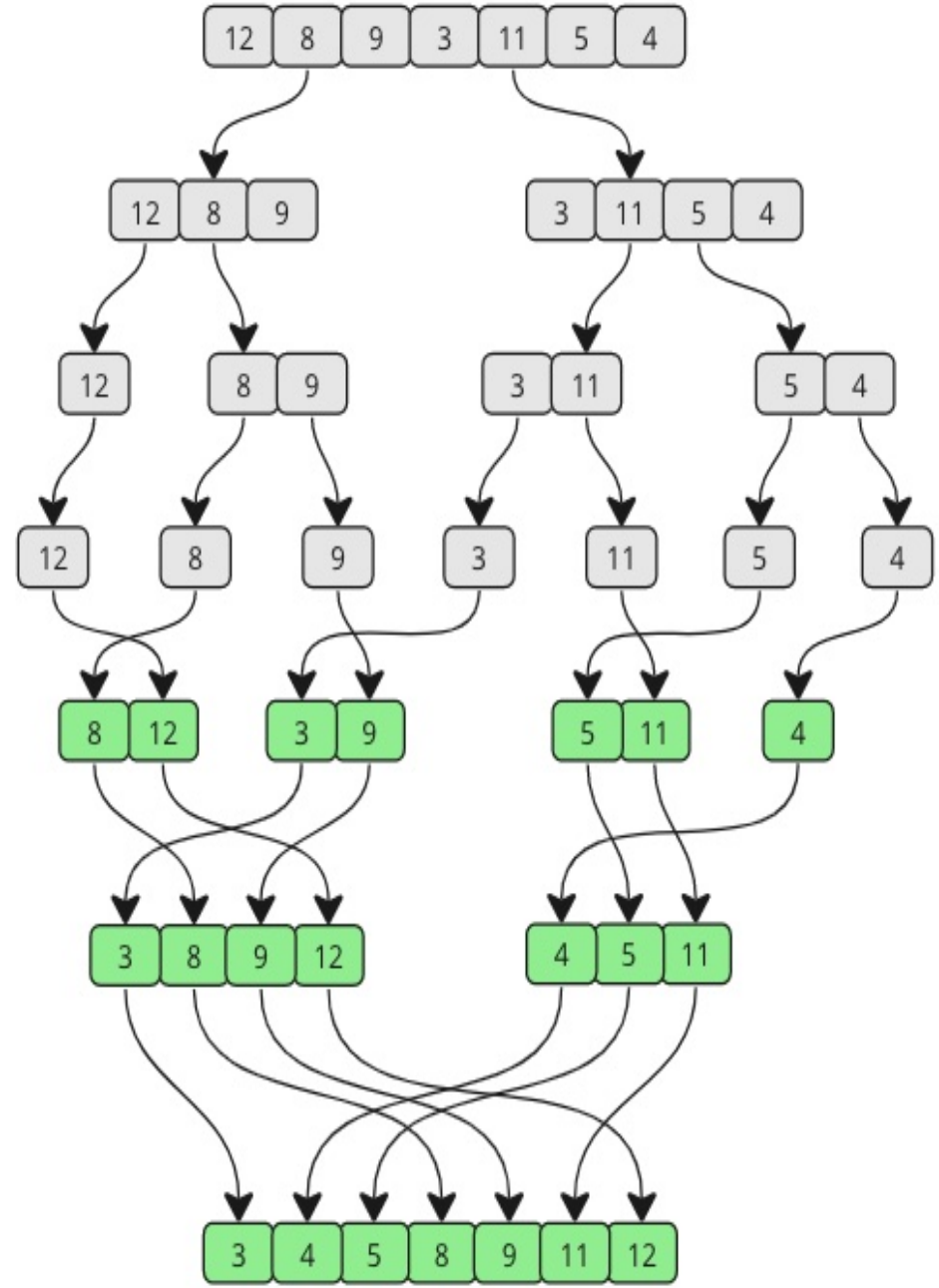




time complexity: O(nlog(n))

n: the total number of elements in the input array.

Mergesort



Conformal Prediction for Surrogate Models under Distribution Shifts

Julie Zhu

Joint work with Youssef Marzouk, Oliver Wang (LIDS, MIT)
Liviu Aolaritei, Michael Jordan (UC Berkeley)

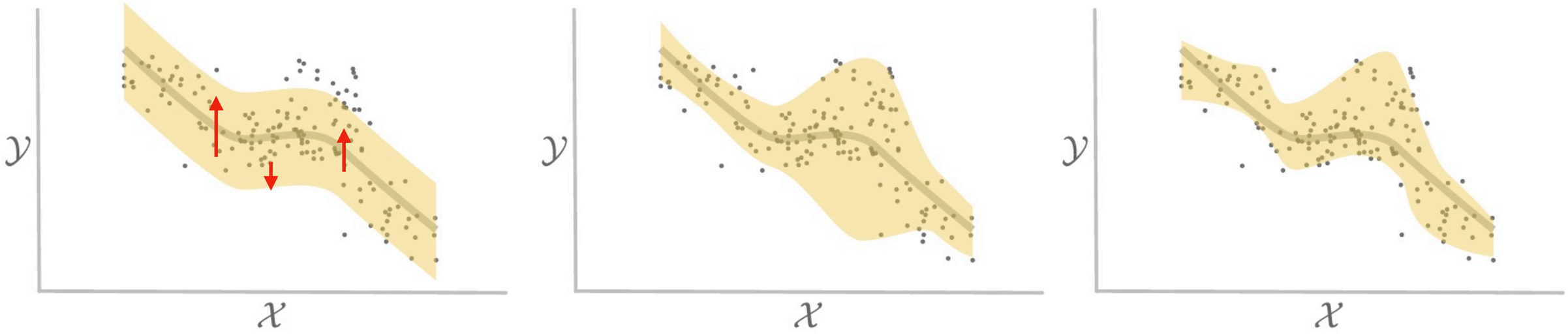
Point Prediction

- Given training data $\left\{(X_i,\ Y_i)\right\}_{i=1}^{K}$ from a distribution $\mathbb{P} \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$
- Fit a surrogate model $\hat{f}$ and get $\hat{f}(X_{n+1})$

Point Prediction is not enough

- Given training data $\{(X_i, Y_i)\}_{i=1}^{K}$ from a distribution $\mathbb{P} \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$
- Fit a surrogate model $\hat{f}$ and get $\hat{f}(X_{n+1})$

- We also want to know how close $\hat{f}(X_{n+1})$ and Y_{n+1} are
- Goal: quantify uncertainty using sets or intervals that "likely contain the true prediction"



Angelopoulos et al., Theoretical Foundations of Conformal Prediction, arXiv:2411.11824 (2024)

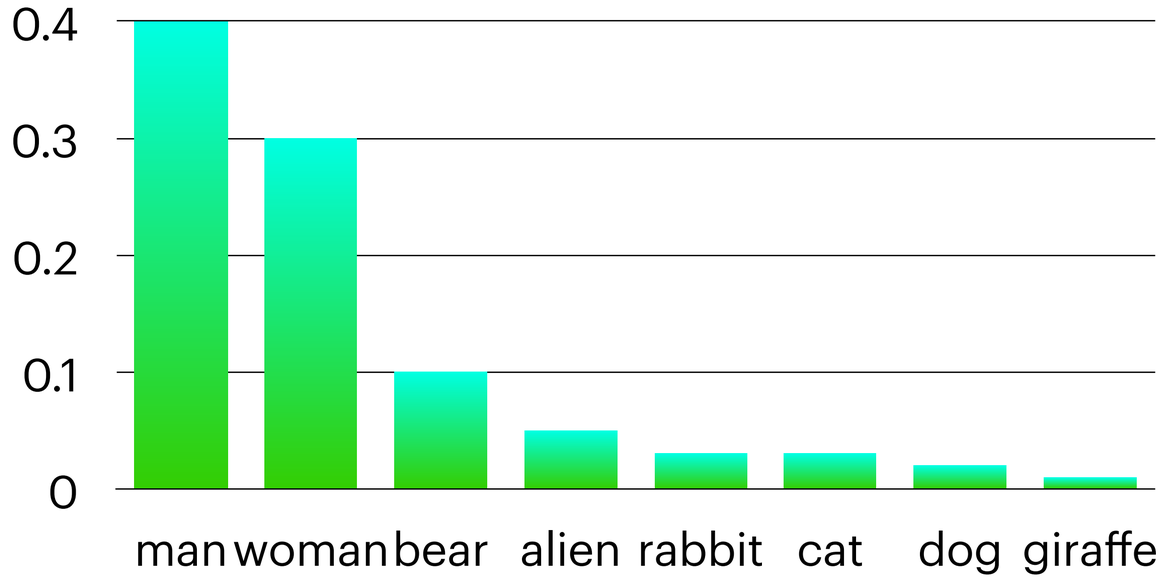
Confidence Sets
for classification

$X \in \mathcal{X} = \mathbb{R}^{m \times n}$: images

$\hat{f}$: classification model

$\hat{f}(X) \in \mathcal{Y} = \mathbb{R}^{K}$: softmax prediction

$Y \in \mathcal{Y}$: true class



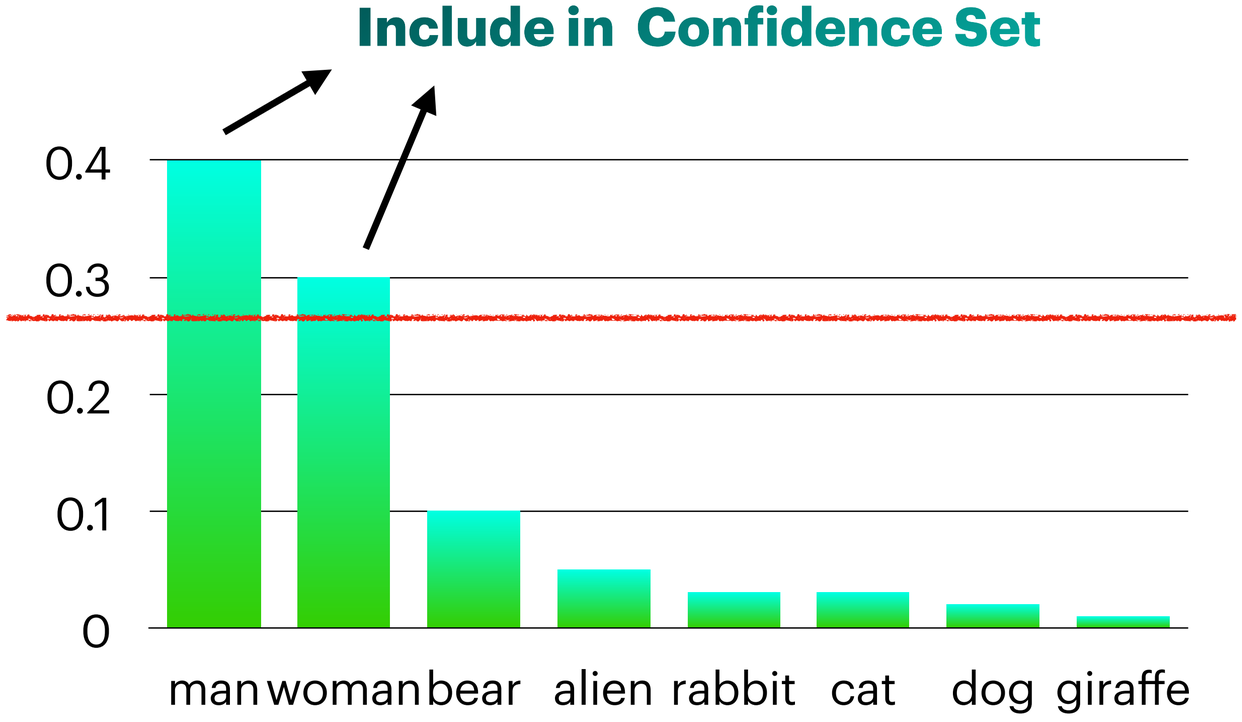
Confidence Sets
for classification

$X \in \mathcal{X} = \mathbb{R}^{m \times n}$: images

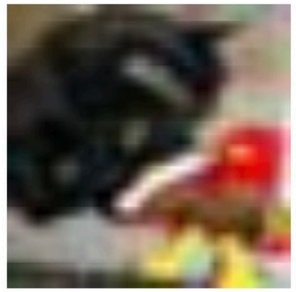
$\hat{f}$: classification model

$\hat{f}(X) \in \mathcal{Y} = \mathbb{R}^{K}$: softmax prediction

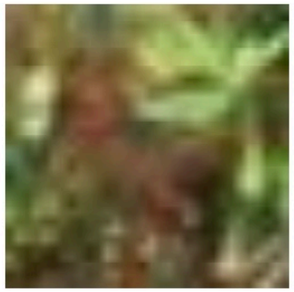
$Y \in \mathcal{Y}$: true class



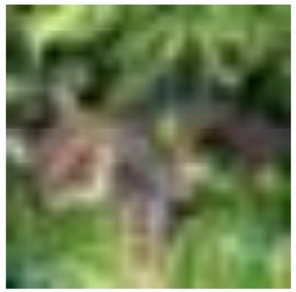
{airplane}



{cat}



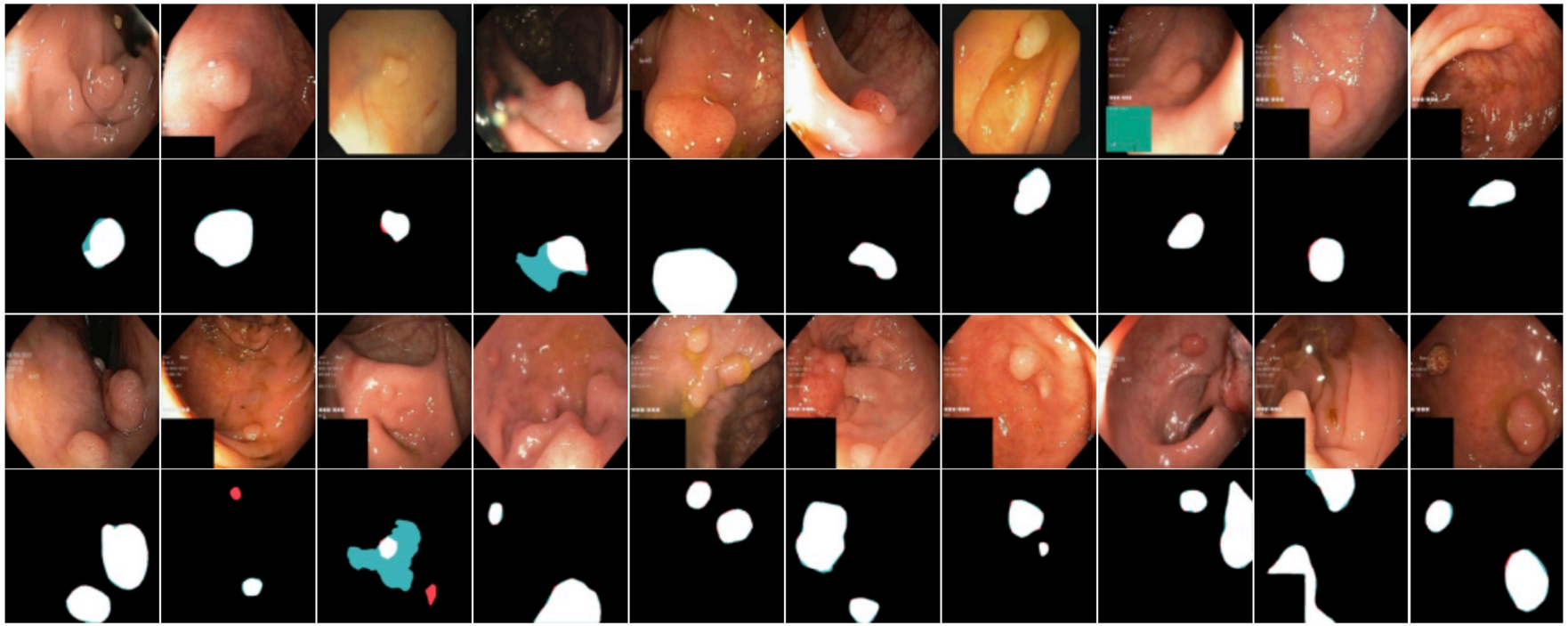
{cat,horse,dog}



{cat,frog}

true class

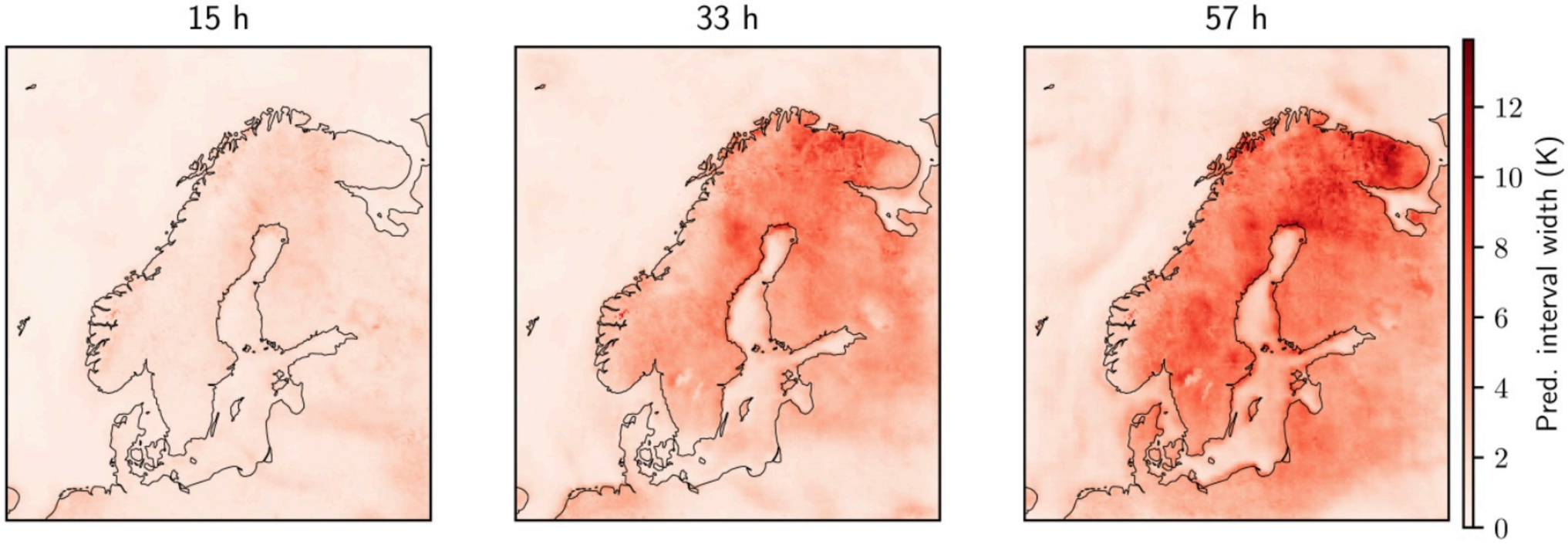
Confidence Sets
for tumor segmentation



White: Tumor Blue: False positive Red: False negative!!!

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. *Conformal Risk Control*. arXiv:2208.02814, 2023.

Confidence Sets
for weather prediction



Pixel-wise confidence interval of error for

predicting the temperature above ground

Gopakumar et al., Valid Error Bars for Neural Weather Models using Conformal Prediction, arXiv:2406.14483 (2024)

Split Conformal Prediction: 90% confidence set

- Partition dataset into

Training $D_1 = \left\{(X_i,\ Y_i)\right\}_{i=1}^{m}$

Calibration $D_2 = \left\{(X_i,\ Y_i)\right\}_{i=1}^{n}$

Split Conformal Prediction: 90% confidence set

- Partition dataset into

$$\text{Training } D_1 = \{(X_i, Y_i)\}_{i=1}^{m}$$

$$\text{Calibration } D_2 = \{(X_i, Y_i)\}_{i=1}^{n}$$

- Train a predictor $\hat{f}$ using D_1

Split Conformal Prediction: 90% confidence set

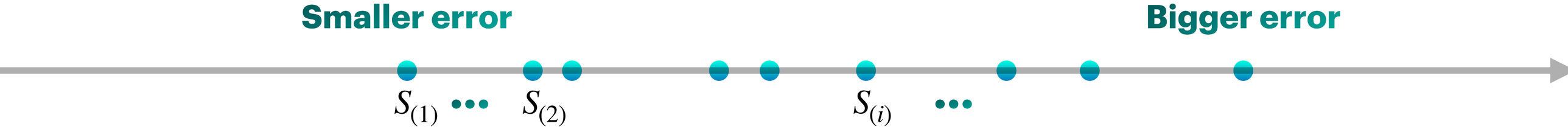
- Partition dataset into

  Training $D_1 = \{(X_i, Y_i)\}_{i=1}^{m}$

  Calibration $D_2 = \{(X_i, Y_i)\}_{i=1}^{n}$

- Train a predictor $\hat{f}$ using D_1

- Define **non-conformity score** $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, evaluate scores $\{S_i\}_{i=1}^{n}$ using D_2
  - $s(X, Y) = |Y - f(X)|_{norm}$
  - $s(X, Y) = -\langle \log f(X), Y \rangle$



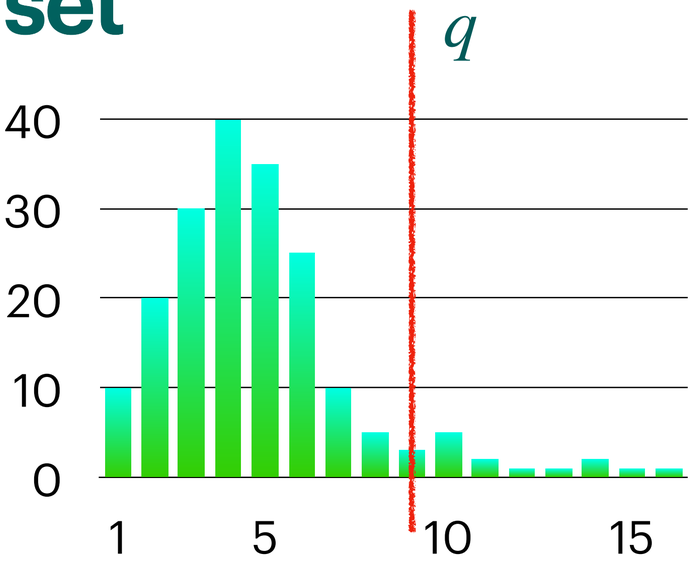
Split Conformal Prediction: 90% confidence set

- Partition dataset into

  Training $D_1 = \{(X_i, Y_i)\}_{i=1}^{m}$

  Calibration $D_2 = \{(X_i, Y_i)\}_{i=1}^{n}$

- Train a predictor $\hat{f}$ using D_1

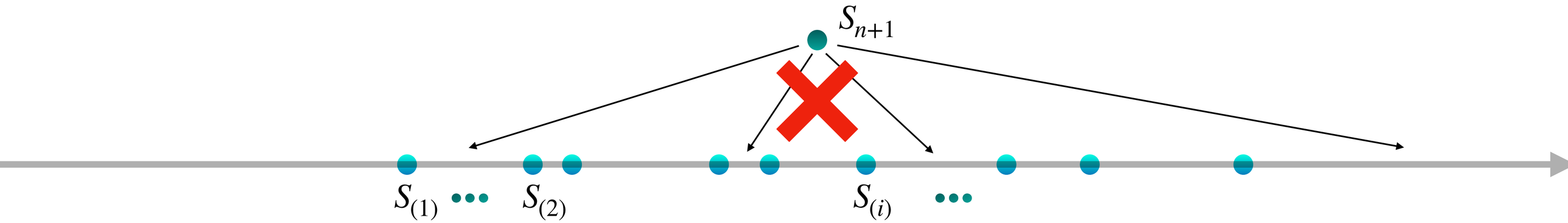


- Define **non-conformity score** $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, evaluate scores $\{S_i\}_{i=1}^{n}$ using D_2

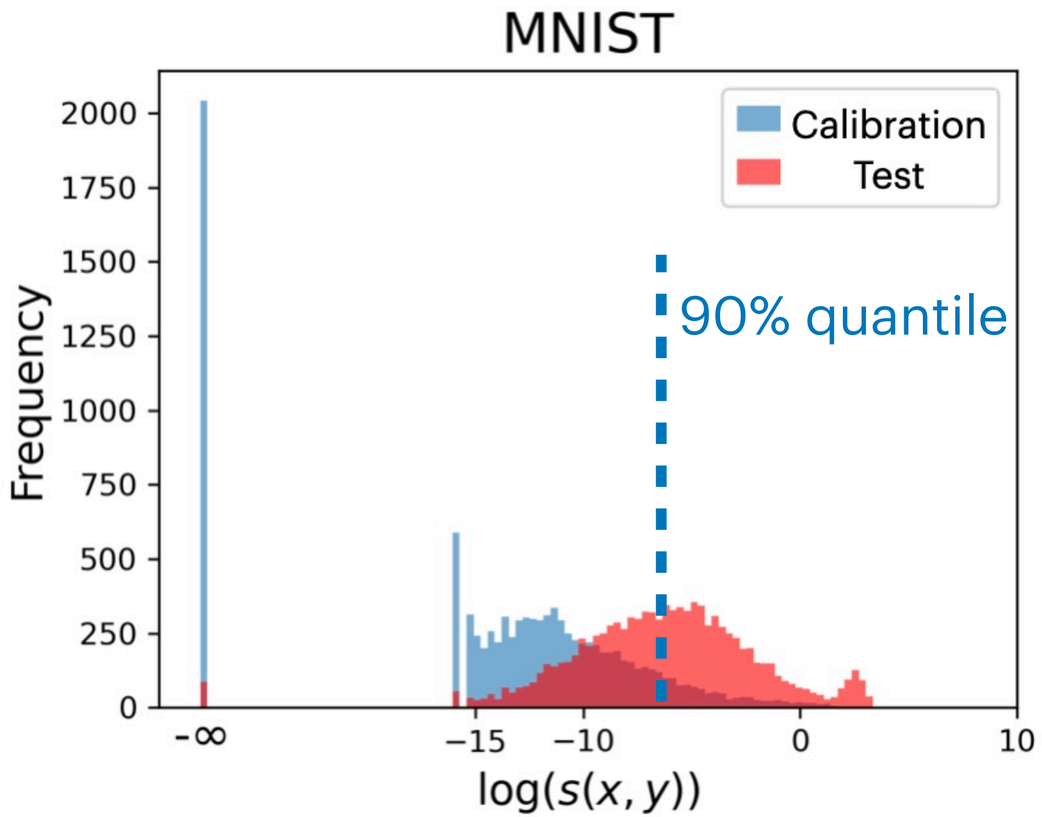
- Construct the prediction set to be

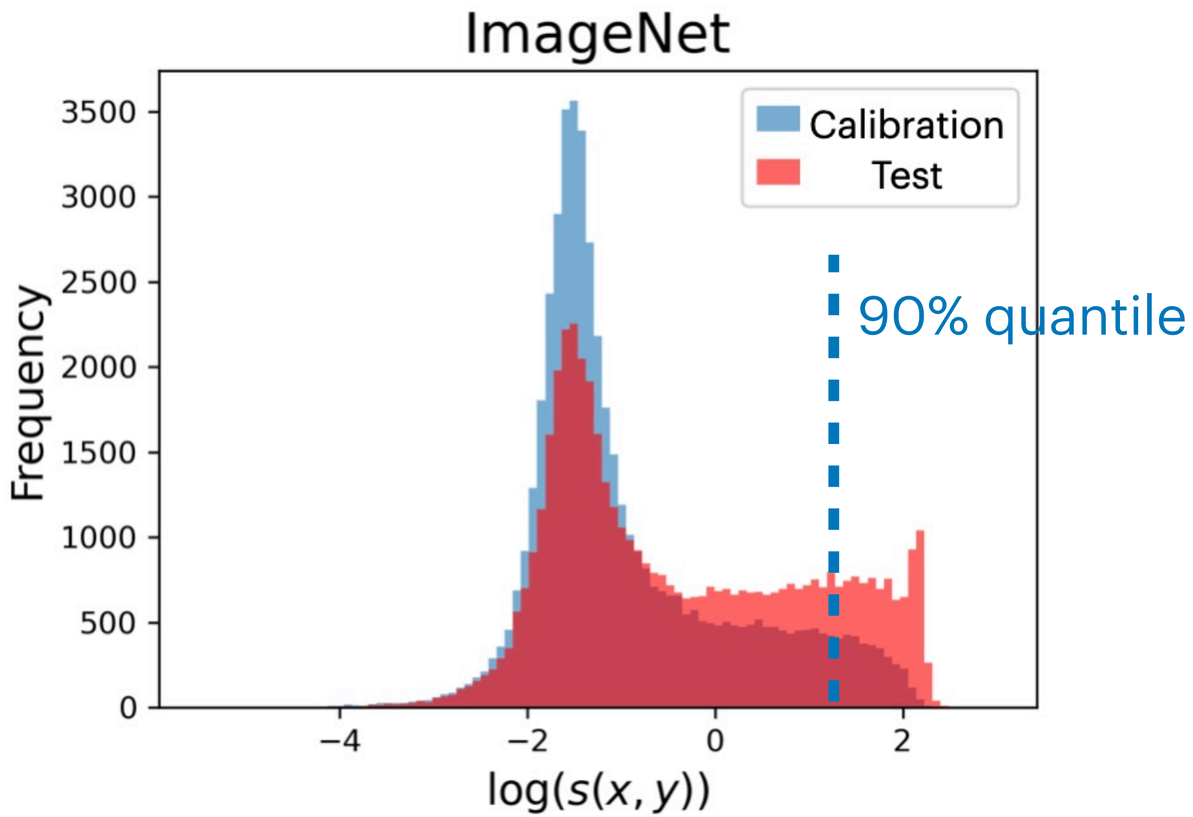
$$\hat{C}_n(x) := \{y : s(x, y) \le q\} \quad \text{where} \quad q := \text{Quant}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}; \frac{1}{n}\sum_{i=1}^{n} \delta_{S_i}\right)$$

No Exchangeability ?

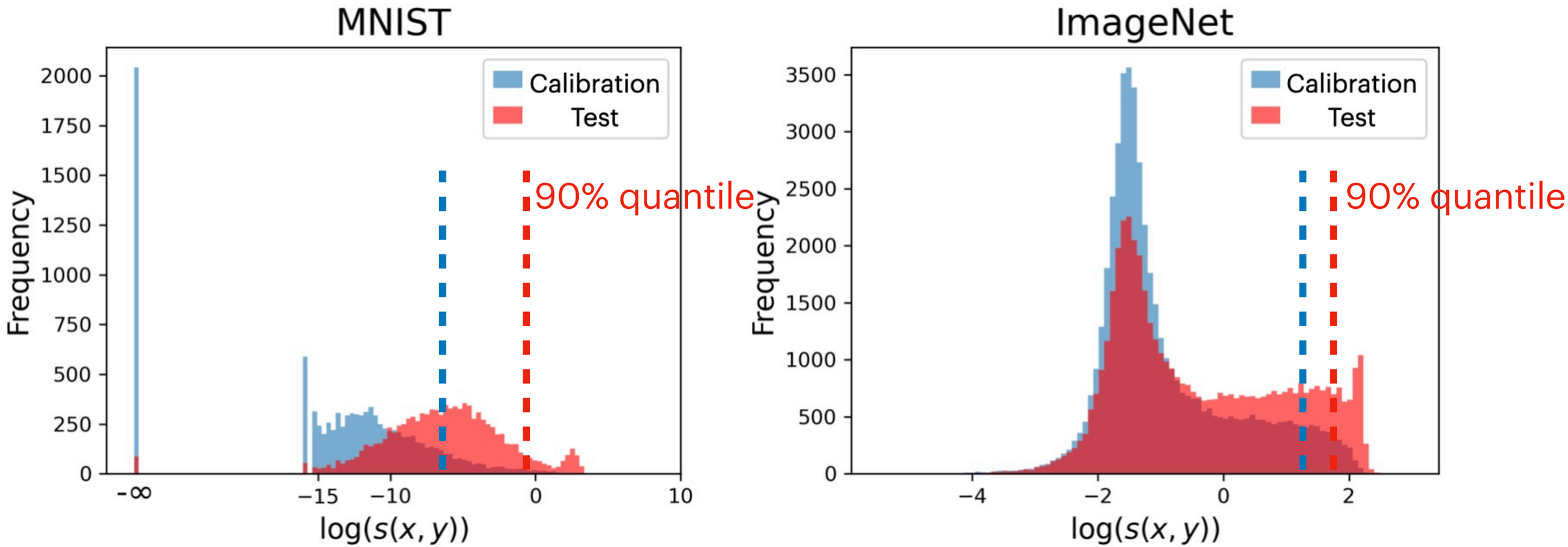


No Exchangeability ?





No Exchangeability ?



90% quantile of the test score distribution changes,

We want to calculate the new one.

Conformal under Distribution Shifts

Labeling distribution shift $p_{Y|X}(y \mid x)$

Covariate shift $p_X(x)$

Noise

Misspecification

Finite sample error

Conformal under Distribution Shifts

Labeling distribution shift $p_{Y|X}(y \mid x)$

Covariate shift $p_X(x)$

Noise
Local

Misspecification
Global

Finite sample error

Conformal under Distribution Shifts

Labeling distribution shift $p_{Y|X}(y \mid x)$

Covariate shift $p_X(x)$

Noise
Local

Misspecification
Global

Ambiguity
Balls

Finite sample error

Conformal Prediction under Levy-Prokhorov Ambiguity

Define a pseudo divergence

$$LP_N(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int 1_{\{z_1 - z_2 \notin N\}} \gamma(dz_1, dz_2)$$

Conformal Prediction under Levy-Prokhorov Ambiguity

Define a pseudo divergence

$$LP_N(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int 1_{\{z_1 - z_2 \notin N\}} \gamma(dz_1, dz_2)$$

Noise

$$Z_1 \sim \mathbb{P},$$
$$Z_{3/2} = Z_1 + \xi, \quad \xi \in N,$$

Conformal Prediction under Levy-Prokhorov Ambiguity

Define a pseudo divergence

$$LP_N(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int 1_{\{z_1 - z_2 \notin N\}} \gamma(dz_1, dz_2)$$

Noise

$$Z_1 \sim \mathbb{P},$$
$$Z_{3/2} = Z_1 + \xi, \quad \xi \in N,$$

+

Misspecification

$$Z_2 = Z_{3/2} \cdot 1_{(B=0)} + \eta \cdot 1_{(B=1)},$$
$$\text{Prob}(B = 1) \leq \rho, \quad \text{arbitrary } \eta$$

Conformal Prediction under Levy-Prokhorov Ambiguity

Define a pseudo divergence

$$LP_N(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int 1_{\{z_1 - z_2 \notin N\}} \gamma(dz_1, dz_2)$$

Noise

$$Z_1 \sim \mathbb{P},$$
$$Z_{3/2} = Z_1 + \xi, \quad \xi \in N,$$

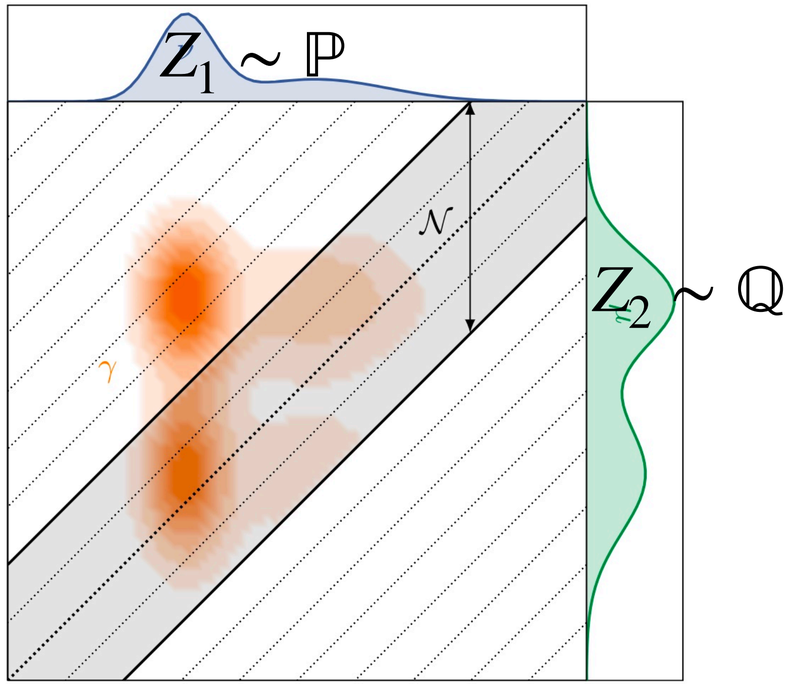
+

Misspecification

$$Z_2 = Z_{3/2} \cdot 1_{(B=0)} + \eta \cdot 1_{(B=1)},$$
$$\text{Prob}(B = 1) \leq \rho, \quad \text{arbitrary } \eta$$

$$Z_1 \sim \mathbb{P} \longrightarrow Z_2 \sim \mathbb{Q}\,.$$

We choose $N = \mathbb{B}_\epsilon(0)$.



Conformal Prediction under Levy-Prokhorov Ambiguity

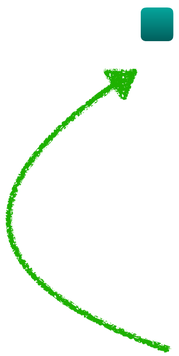
Interpolation between Wasserstein-∞ and Total Variation $\mathcal{N} = \mathbb{B}_\epsilon(0), \;\; LP_{\mathcal{N}} = LP_\epsilon$

$$\mathrm{LP}_\varepsilon(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{Z} \times \mathcal{Z}} 1_{\{\|z_1 - z_2\| > \varepsilon\}} \mathrm{d}\gamma(z_1, z_2)$$

Conformal Prediction under Levy-Prokhorov Ambiguity

Interpolation between Wasserstein-∞ and Total Variation $\mathcal{N} = \mathbb{B}_\epsilon(0),\ LP_{\mathcal{N}} = LP_\epsilon$

$LP_\epsilon = \epsilon$
"Local"



$$\mathsf{W}_\infty(\mathbb{P},\mathbb{Q}) := \inf\left\{\epsilon \geq 0 : \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} 1_{\{\|z_1-z_2\|>\epsilon\}}\, \mathrm{d}\gamma(z_1,z_2) \leq 0\right\}$$

$$\mathsf{LP}_\varepsilon(\mathbb{P},\mathbb{Q}) := \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} 1_{\{\|z_1-z_2\|>\varepsilon\}} \mathrm{d}\gamma(z_1,z_2)$$

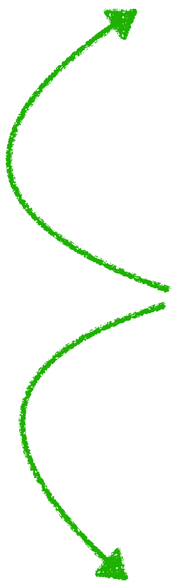
Amine Bennouna, Bart Van Parys, and Ryan Lucas. *Holistic Robust Data-Driven Decisions*. arXiv:2207.09560, 2025.

Conformal Prediction under Levy-Prokhorov Ambiguity

Interpolation between Wasserstein-∞ and Total Variation $\mathcal{N} = \mathbb{B}_\epsilon(0),\ LP_{\mathcal{N}} = LP_\epsilon$

$LP_\epsilon = \epsilon$
"Local"

$$\mathsf{W}_\infty(\mathbb{P},\mathbb{Q}) := \inf\left\{\epsilon \geq 0 : \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} 1_{\{\|z_1-z_2\|>\epsilon\}}\, \mathrm{d}\gamma(z_1,z_2) \leq 0\right\}$$



$$\mathsf{LP}_\varepsilon(\mathbb{P},\mathbb{Q}) := \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} 1_{\{\|z_1-z_2\|>\varepsilon\}} \mathrm{d}\gamma(z_1,z_2)$$

$\epsilon = 0$
"Global"

$$\mathsf{TV}(\mathbb{P},\mathbb{Q}) := \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} 1_{\{\|z_1-z_2\|>0\}} \mathrm{d}\gamma(z_1,z_2)$$

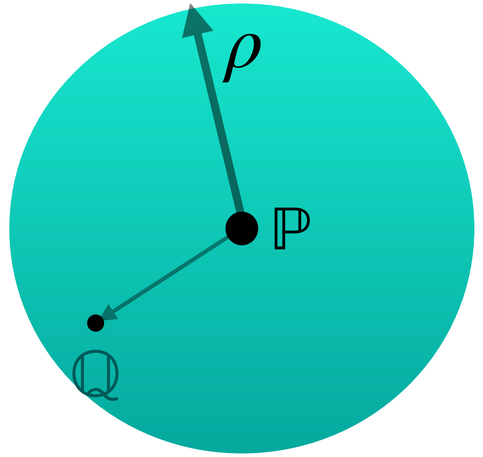
Amine Bennouna, Bart Van Parys, and Ryan Lucas. *Holistic Robust Data-Driven Decisions*. arXiv:2207.09560, 2025.

Levy-Prokhorov Ambiguity Ball

- The **LP ambiguity ball** in data space

$$\mathbb{B}_{\epsilon,\,\rho}(\mathbb{P}) \;:=\; \{\mathbb{Q} : LP_{\epsilon}(\mathbb{P},\mathbb{Q}) \le \rho\} \subset \mathbb{R}^{data}$$

ϵ: local noise level; ρ: global noise level



$\mathbb{B}_{\epsilon,\,\rho}(\mathbb{P}) \subset \mathbb{R}^{data}$

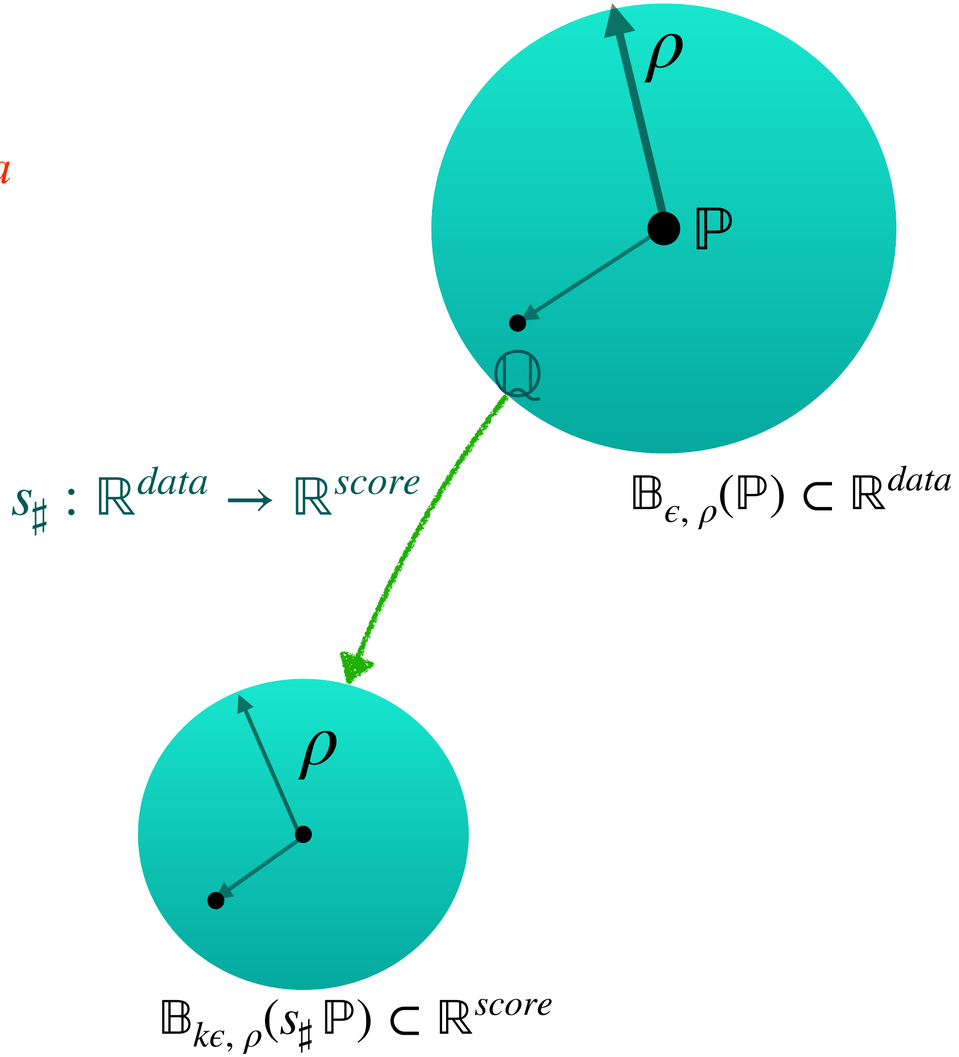
Levy-Prokhorov Ambiguity Ball

- The **LP ambiguity ball** in data space

$$\mathbb{B}_{\epsilon,\,\rho}(\mathbb{P}) \;:=\; \{\mathbb{Q} : LP_{\epsilon}(\mathbb{P},\mathbb{Q}) \le \rho\} \subset \mathbb{R}^{data}$$

ϵ: local noise level; ρ: global noise level

- Score function is k-Lipschitz



Levy-Prokhorov Ambiguity Ball

- The **LP ambiguity ball** in data space

$$\mathbb{B}_{\epsilon,\,\rho}(\mathbb{P}) \;:=\; \{\mathbb{Q} : LP_{\epsilon}(\mathbb{P},\mathbb{Q}) \le \rho\} \subset \mathbb{R}^{data}$$

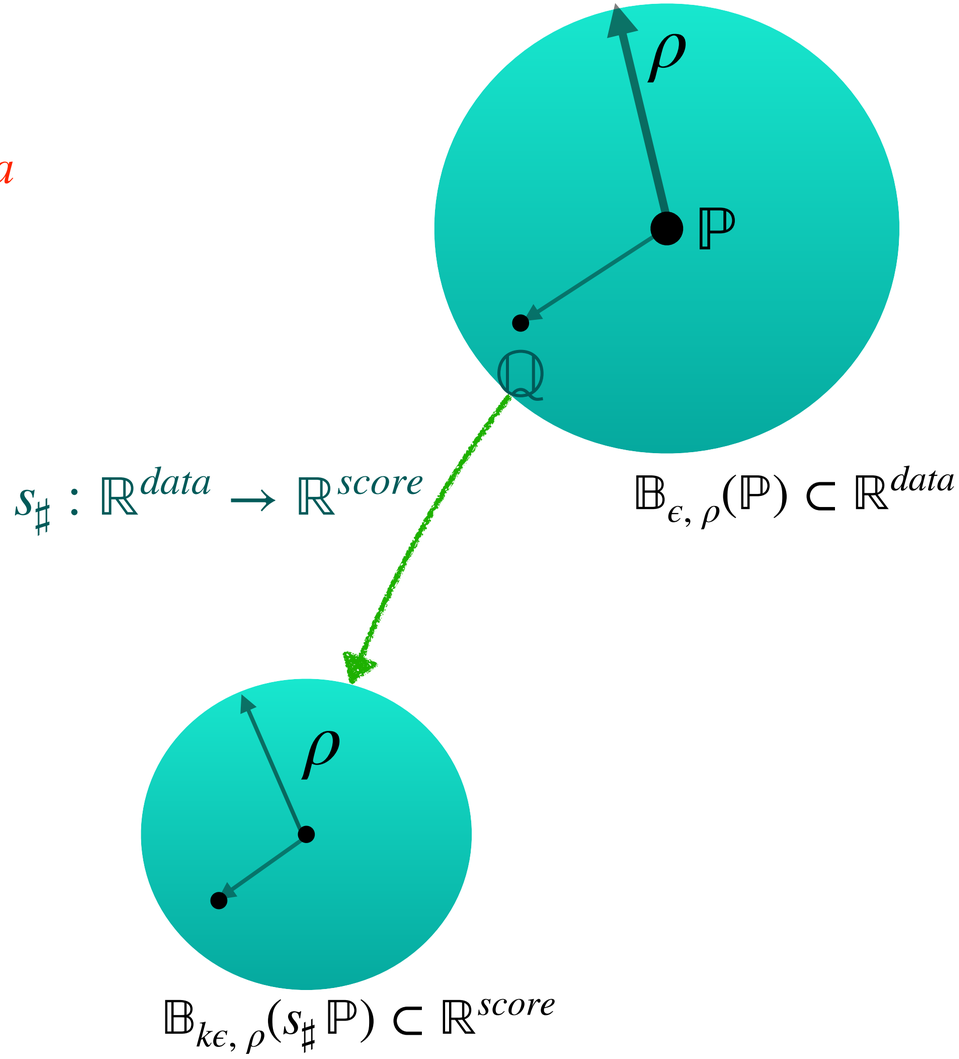
ϵ: local noise level; ρ: global noise level

- Score function is k-Lipschitz

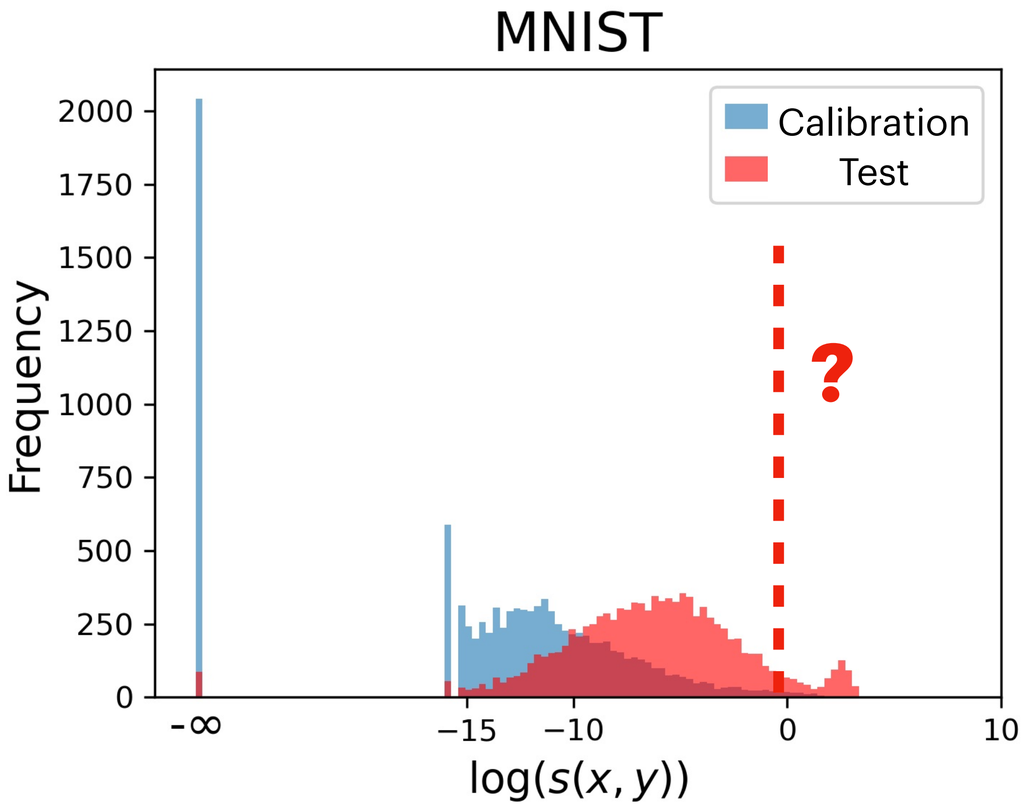
- **Propagation** under score preserves LP ball

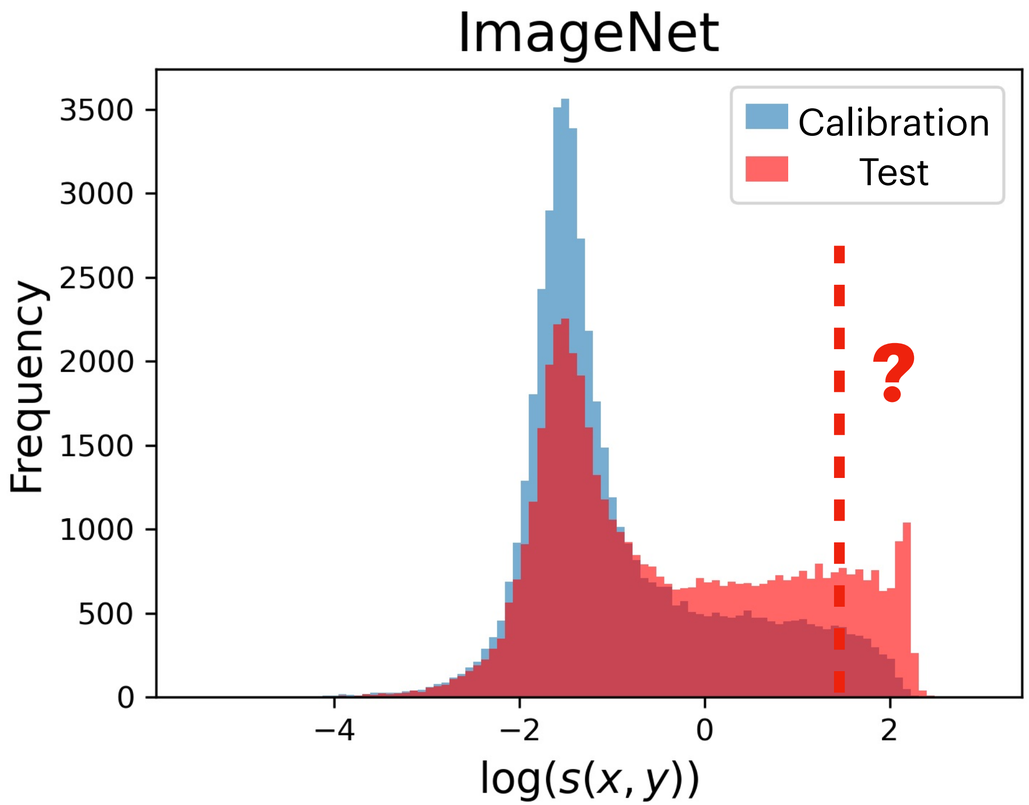
$$s_\sharp\, \mathbb{B}_{\epsilon,\,\rho}(\mathbb{P}) \;\subset\; \mathbb{B}_{k\epsilon,\,\rho}(s_\sharp\, \mathbb{P}) \subset \mathbb{R}^{score}$$

$k\epsilon$: local noise level; ρ: global noise level



Worst-case Quantile and Coverage





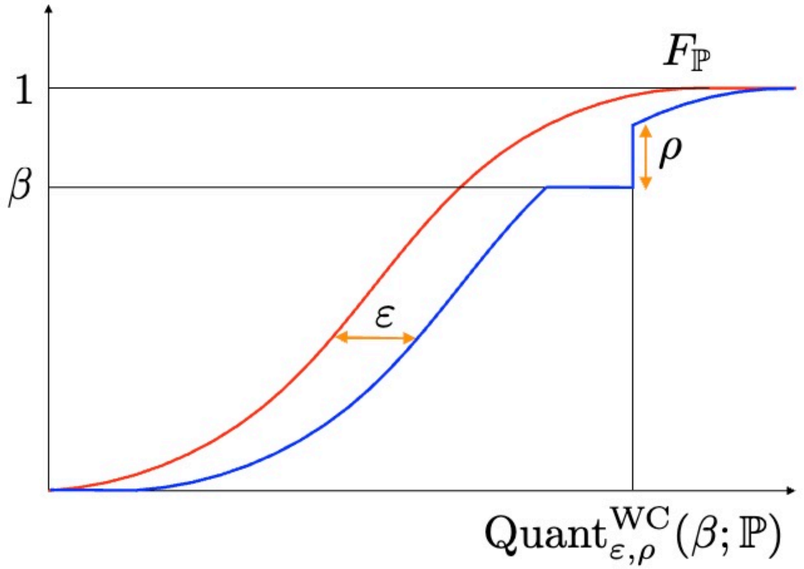
- $LP_{k\epsilon}(\mathbb{P}, P_{test}) \leq \rho$; Know empirical $\hat{\mathbb{P}}_n$
- What is the worst case 90% quantile for P_{test}?

Worst-case Quantile and Coverage

Proposition 1

smallest quantile s.t. all P achieve β coverage?

$$\begin{aligned}\text{Quant}^{\text{WC}}_{\varepsilon,\rho}(\beta;\mathbb{P}) &:= \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,\rho}(\mathbb{P})} \text{Quant}(\beta;\mathbb{Q}) \\ &= \text{Quant}(\beta+\rho;\mathbb{P})+\varepsilon\end{aligned}$$

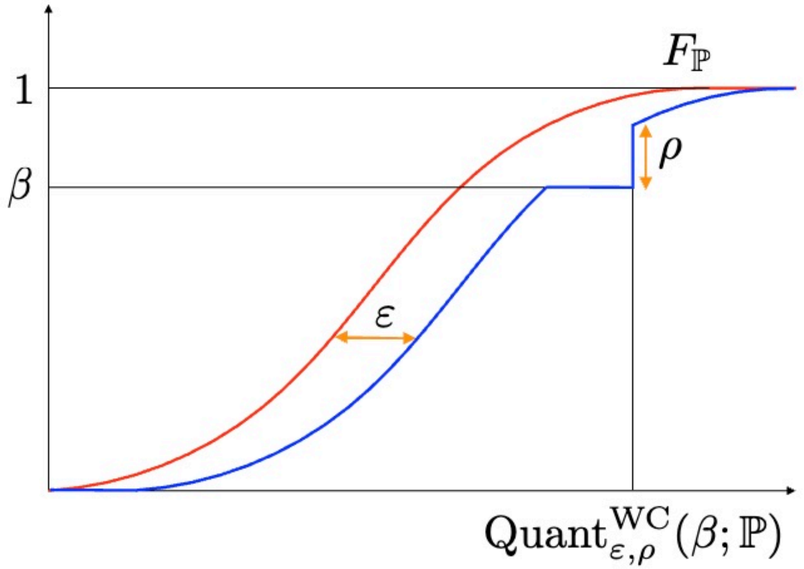


Worst-case Quantile and Coverage

Proposition 1

smallest quantile s.t. all P achieve β coverage?

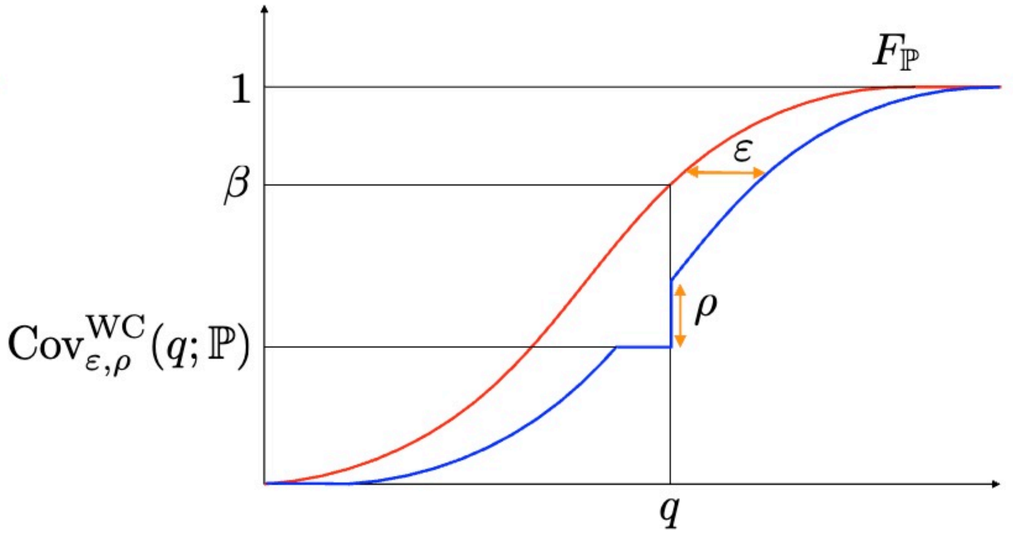
$$\begin{aligned}\mathrm{Quant}^{\mathrm{WC}}_{\varepsilon,\rho}(\beta;\mathbb{P}) &:= \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,\rho}(\mathbb{P})} \mathrm{Quant}(\beta;\mathbb{Q}) \\ &= \mathrm{Quant}(\beta+\rho;\mathbb{P})+\varepsilon\end{aligned}$$



Proposition 2

smallest coverage at given quantile?

$$\begin{aligned}\mathrm{Cov}^{\mathrm{WC}}_{\varepsilon,\rho}(q;\mathbb{P}) &:= \inf_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,\rho}(\mathbb{P})} F_{\mathbb{Q}}(q) \\ &= F_{\mathbb{P}}(q-\varepsilon)-\rho\end{aligned}$$



Empirical Coverage Theorem

Using **proposition 1**, we construct a prediction set valid under distribution shift:

$$C(x;\ \mathbb{P}) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\alpha;\ \mathbb{P}) \right\}$$

Empirical Coverage Theorem

Using **proposition 1**, we construct a prediction set valid under distribution shift:

$$C(x;\ \mathbb{P}) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\alpha;\ \mathbb{P}) \right\}$$

But $\mathbb{P}$ is unknown to us. Instead, we have $\hat{\mathbb{P}}_n$ from calibration data

$$C(x;\ \hat{\mathbb{P}}_n) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\alpha;\ \hat{\mathbb{P}}_n) \right\}$$

Which yields smaller coverage:

$$P\left(Y_{n+1} \in C\left(X_{n+1}; \hat{\mathbb{P}}_n\right) \right) \geq \frac{\lceil n(1-\alpha+\rho) \rceil}{n+1} - \rho$$

Empirical Coverage Theorem

Using **proposition 1**, we construct a prediction set valid under distribution shift:

$$C(x;\ \mathbb{P}) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\alpha;\ \mathbb{P}) \right\}$$

But $\mathbb{P}$ is unknown to us. Instead, we have $\hat{\mathbb{P}}_n$ from calibration data

$$C(x;\ \hat{\mathbb{P}}_n) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\alpha;\ \hat{\mathbb{P}}_n) \right\}$$

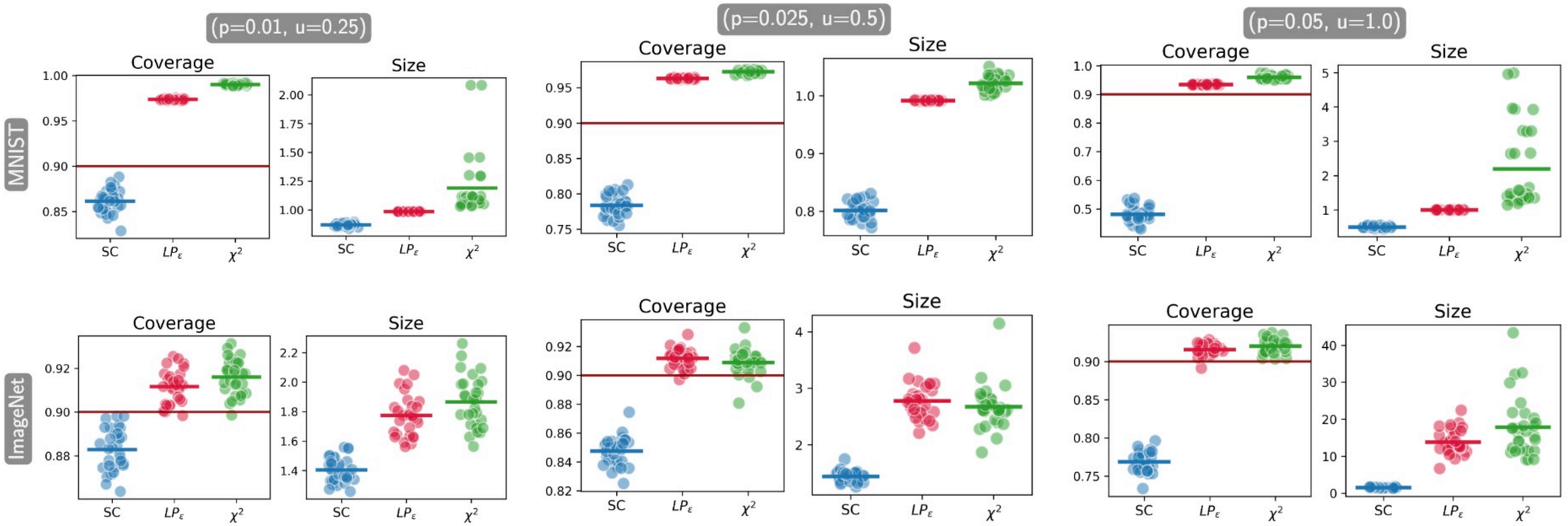
Which yields smaller coverage:

$$P\left(Y_{n+1} \in C\left(X_{n+1}; \hat{\mathbb{P}}_n\right) \right) \geq \frac{\lceil n(1-\alpha+\rho) \rceil}{n+1} - \rho$$

Set $\hat{\alpha} = \alpha + (\alpha - \rho - 2)/n$, we recover a valid prediction set

$$C(x;\ \hat{P}_n) := \left\{ y \mid s(x,y) \leq \text{Quant}^{\text{WC}}_{LP_\epsilon}(1-\hat{\alpha};\ \hat{\mathbb{P}}_n) \right\}$$

Experiments: MNIST and ImageNet

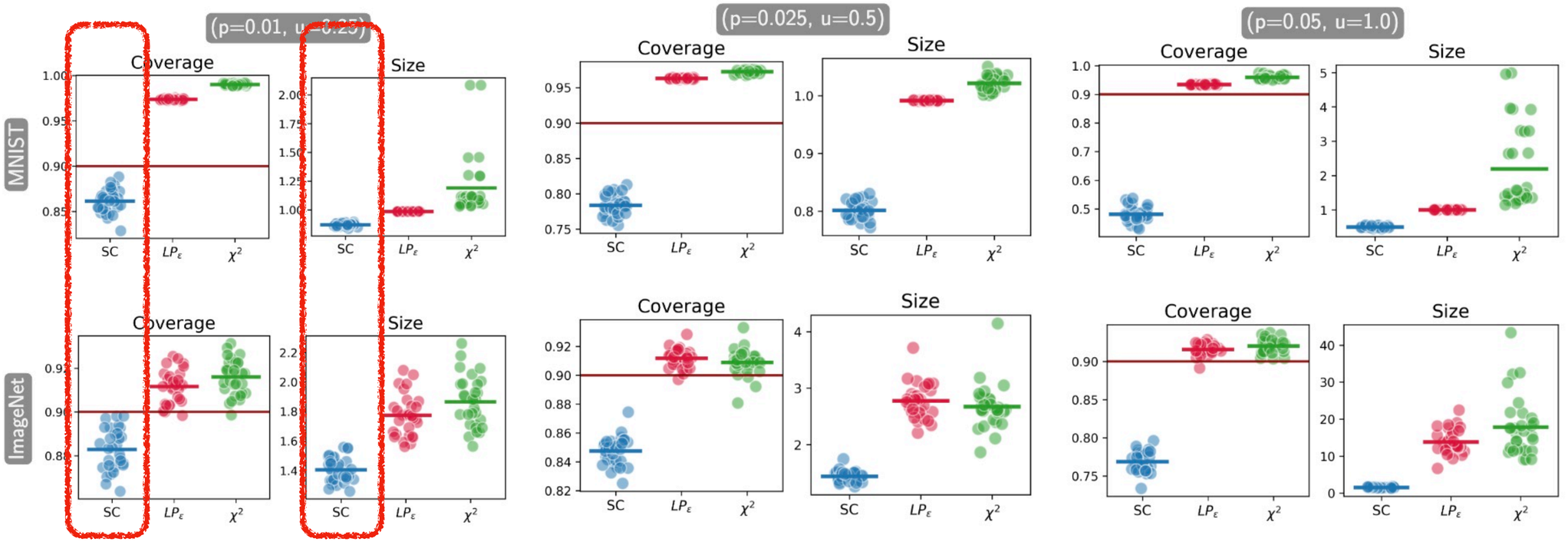


Data-space distribution shift validity and efficiency. Desired $1 - \alpha$ coverage (long dark red line);

empirical coverage and prediction set size for each split (scattered points);

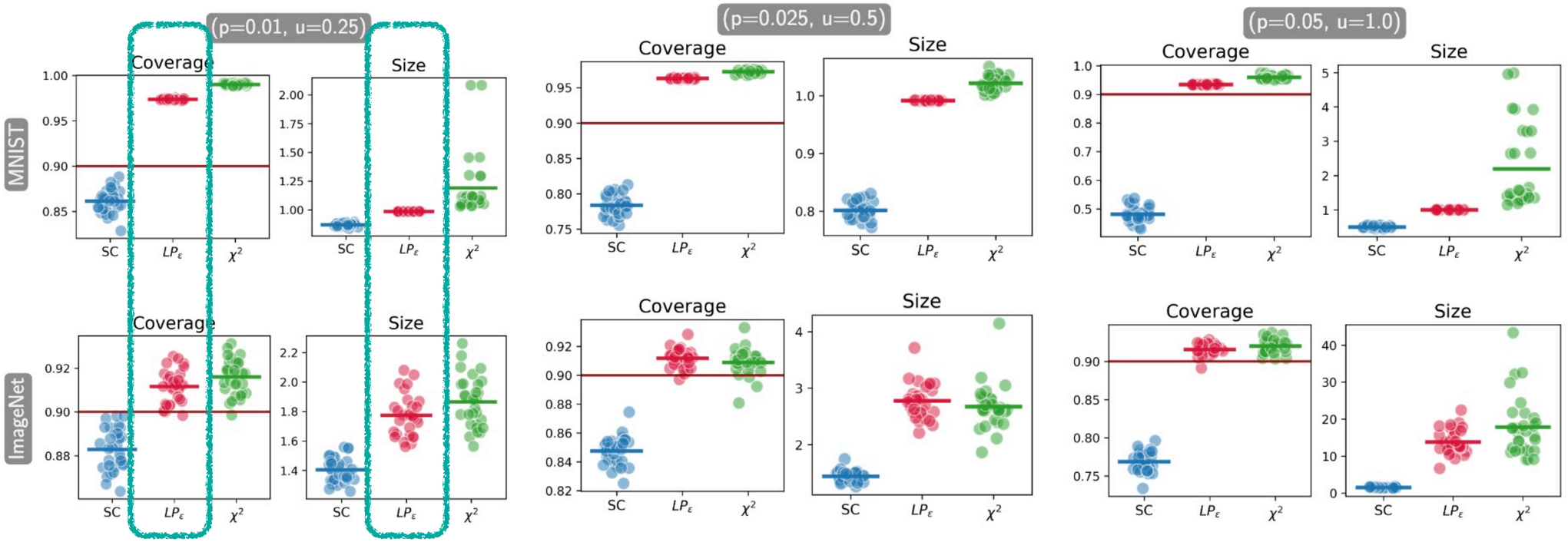
and mean coverage and prediction set size across 30 calibration-test splits (short horizontal lines).

Experiments: MNIST and ImageNet



Data-space distribution shift validity and efficiency. Desired $1 - \alpha$ coverage (long dark red line); empirical coverage and prediction set size for each split (scattered points); and mean coverage and prediction set size across 30 calibration-test splits (short horizontal lines).

Experiments: MNIST and ImageNet



Data-space distribution shift validity and efficiency. Desired $1 - \alpha$ coverage (long dark red line);

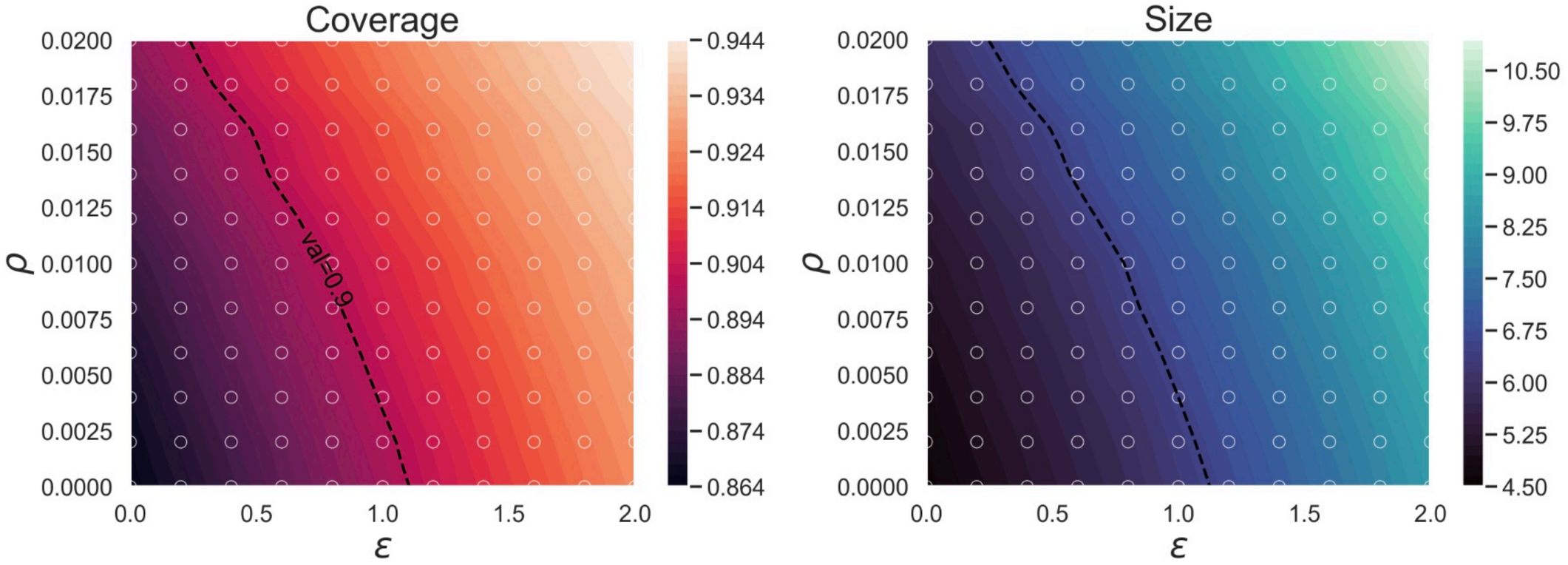
empirical coverage and prediction set size for each split (scattered points);

and mean coverage and prediction set size across 30 calibration-test splits (short horizontal lines).

Experiments: iWildCam

https://github.com/visipedia/iwildcam_comp

Experiments: iWildCam



Coverage (left) and prediction set size (right) over a range of (ε, ρ), illustrated with color maps.

Desired 90% coverage is the black dotted line; points above and to the right of this line achieve valid coverage.

The point $(0,0)$ represents standard conformal prediction.

Key Takeaways:

1. Conformal distribution under distribution shift
2. Levy-Prokhorov ambiguity set: capture local/global distribution shift
3. Closed form result for the worst-case scenario in ambiguity set

Key Takeaways:

1. Conformal distribution under distribution shift
2. Levy-Prokhorov ambiguity set: capture local/global distribution shift
3. Closed form result for the worst-case scenario in ambiguity set

Things we have not finished...

1. Develop methods to quantify ϵ and ρ;
2. Design score function for risk control, high-dim cases and generative models;
3. Application on climate models and material science

Key Takeaways:

1. Conformal distribution under distribution shift
2. Levy-Prokhorov ambiguity set: capture local/global distribution shift
3. Closed form result for the worst-case scenario in ambiguity set

Scan to see the paper:
Conformal Prediction under LP Distribution Shifts
arXiv:2502.14105

Things we have not finished...

1. Develop methods to quantify ϵ and ρ;
2. Design score function for risk control, high-dim cases and generative models
3. Application on climate models and material science

Thank you!

References:

[1] L., Aolaritei, M. I., Jordan, Y., Marzouk, Z. O., Wang, & J., Zhu. Conformal Prediction under L\'evy-Prokhorov Distribution Shifts: Robustness to Local and Global Perturbations. arXiv preprint arXiv:2502.14105, 2025.

[2] A. Bennouna and B. Van Parys. Holistic robust data-driven decisions. arXiv preprint arXiv:2207.09560, 2022.

[3] L. Aolaritei, N. Lanzetti, H. Chen, and F. Dörfler. Distributional uncertainty propagation via optimal transport. IEEE Transactions on Automatic Cntrol (Forthcoming), 2025.

[4] . R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2):816–845, 2023.

[5] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi. Robust validation: Confident predictions even when distributions shift. Journal of the American Statistical Association, pages 1–66, 2024.