Master Thesis Presentation

Optimal Transport in Linear Independent Component Analysis

Ashutosh Jha

University of Tuebingen

October 2025

Outline

Introduction (1/3)

- Linear Independent Component Analysis (ICA) seeks to recover a set of statistically independent components from their linear mixtures.
- Unlike Principal Component Analysis (PCA), which only ensures uncorrelated components, ICA imposes the stronger condition of full statistical independence.
- This stronger criterion allows ICA to disentangle latent variables that PCA cannot fully separate.

Introduction (2/3)

- The approach builds upon intuition from the Central Limit Theorem: mixtures of independent random variables tend towards Gaussian distributions regardless of original distributions.
- Consequently, any mixture of independent **non-Gaussian** sources appears more Gaussian than the original sources themselves.
- After whitening the mixed signal to remove correlations, the problem reduces to finding a rotation on the unit sphere that maximizes non-Gaussianity.

Introduction (3/3)

- The goal is to identify directions where projected data are least Gaussian, corresponding to candidate statistically independent components.
- To formalize this as an optimization problem, we measure Gaussianity using the Wasserstein-2 (W2) distance between the empirical distribution of projections and a Gaussian distribution.
- This distance comes from Optimal Transport theory, which quantifies the minimal "cost" of morphing one probability distribution into another.

ICA Model

The linear ICA model assumes that the observed signals are linear mixtures of statistically independent sources:

$$\mathbf{X} = \mathbf{AS}$$

where

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

The sources s_i are assumed to be statistically independent, and at most one s_i can be Gaussian, since a mixture of Gaussian signals remains Gaussian and cannot be separated by ICA [**?**].

Centering

- Observed mixtures: $\mathbf{X} \in \mathbb{R}^d$.
- Center before further processing: $\mathbb{E}[\mathbf{X}] = \mathbf{0}$.
- Covariance formula:

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top$$

- If $\mathbb{E}[\mathbf{X}] = 0$ then

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$$

Whitening: Decorrelating with Eigenvalue Decomposition

- Start with centered data $\mathbf{X} \in \mathbb{R}^d$.
- Compute the covariance matrix:

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$$

- Eigenvalue decomposition of covariance:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

  where $\mathbf{U}$: eigenvectors, $\mathbf{\Lambda}$: diagonal matrix of eigenvalues.
- Apply decorrelation transform:

$$\mathbf{W} = \mathbf{V}^\top$$

- $\mathbf{WX}$ yields a diagonal covariance matrix with variance entries λ_j, but not necessarily unit variance.

Whitening: Scaling to Identity Covariance

- To "whiten" data, rescale each decorrelated dimension to have unit variance.
- Whitening transformation:

$$\mathbf{W} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top$$

  Here, $\mathbf{\Lambda}^{-1/2}$ scales diagonal entries by $1/\sqrt{\lambda_j}$.
- Whitened data:

$$\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$$

- Covariance after whitening:

$$\mathrm{Cov}(\widetilde{\mathbf{X}}) = \mathbf{W}\mathrm{Cov}(\mathbf{X})\mathbf{W}^\top = \mathbf{I}_d$$

- Eigenvalue decomposition decorrelates data; scaling by $\mathbf{\Lambda}^{-1/2}$ gives unit variance.

Whitening and Scaling

- Scaling the data so all dimensions have unit variance and are uncorrelated.
- Found (with ED/SVD) a linear $\mathbf{W}$ that gives $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$ s.t.

$$\mathrm{Cov}(\widetilde{\mathbf{X}}) = \mathbb{E}[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top] = \mathbb{E}[\mathbf{W}\mathbf{X}\mathbf{X}^\top\mathbf{W}^\top] = \mathbf{W}\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{W}^\top$$

$$\mathrm{Cov}(\widetilde{\mathbf{X}}) = \mathbf{W}\mathrm{Cov}(\mathbf{X})\mathbf{W}^\top$$

- Diagonalized covariance:

$$\mathrm{Cov}(\widetilde{\mathbf{X}}) = \mathbf{W}\mathrm{Cov}(\mathbf{X})\mathbf{W}^\top = \mathbf{D}$$

  but we choose scaling ($\mathbf{\Lambda}^{-1/2}$) such that $\mathbf{D} = \mathbf{I}_d$.
- **Why this scaling?**
  - Ensures all dimensions are comparable.
  - Reduces ICA problem to a rotation (orthogonal search, next slide).

Unmixing

- After whitening, seek a matrix $\mathbf{U}$ to "unmix" such that:

$$\widetilde{\mathbf{Z}} = \mathbf{U}\widetilde{\mathbf{X}}$$

  where $\widetilde{\mathbf{Z}}$ is an estimated candidate vector for independent components.
- Covariance after unmixing:

$$\mathrm{Cov}(\widetilde{\mathbf{Z}}) = \mathbf{U}\,\mathrm{Cov}(\widetilde{\mathbf{X}})\,\mathbf{U}^\top = \mathbf{U}\mathbf{I}_d\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top$$

- When $\mathbf{U}$ is orthogonal ($\mathbf{U}\mathbf{U}^\top = \mathbf{I}_d$), recovered components are uncorrelated and ready for independence optimization.
- ICA then finds the rotation that makes elements of $\widetilde{\mathbf{Z}}$ as independent as possible.

Why PCA is not sufficient?

1. Make sample decorrelated and whitened, i.e. $E[\mathbf{X}] = 0$, $E[\mathbf{X}\mathbf{X}^T] = \mathbf{I}$. At this point, any orientation is good enough for PCA, so it is not sufficient for ICA, when the data are non gaussian. For multivariate gaussian data decorrelation implies independence, so there is not much more for ICA to do.
2. For ICA, find local optima of the function $f(\mathbf{u}) = E[(\mathbf{u}^T\mathbf{X})^4]$ where $\mathbf{u} \in S_n$ (unit sphere). Gradient descent can be used to find candidate components.

Information Geometry Perspective on ICA

- ICA aims to find a linear transform B such that $Y = BX$ has maximally independent components.
- Standard decorrelation (whitening) gives uncorrelated axes, but not full independence unless data are Gaussian.
- Information geometry relates independence and non-Gaussianity: ICA projects distributions onto product and Gaussian manifolds to quantify both phenomena [**?**].

Manifolds of Product and Gaussian Distributions

- $\mathcal{P}$: Product manifold – distributions where all components are independent.
- $\mathcal{G}$: Gaussian manifold – all multivariate Gaussians (zero mean, fixed covariance).
- ICA seeks a projection of the empirical distribution onto $\mathcal{P}$: the closest independent representation.
- After whitening, ICA operates on the intersection where covariance is identity and independence must be achieved via some optimization on rotated candidates.

KL Pythagorean Identity on Product Manifold

- For any random vector Y with density P_Y, mutual information is denoted by $L(Y)$:

$$I(Y) = D_{\mathrm{KL}}(P_Y \parallel \prod_i P_{Y_i})$$

- More generally, for any product candidate $Q = \prod_i q_i$:

$$\begin{aligned} D_{\mathrm{KL}}\left(P_Y \| Q\right) &= D_{\mathrm{KL}}\left(P_Y \| \prod_i P_{Y_i}\right) + D_{\mathrm{KL}}\left(\prod_i P_{Y_i} \| \prod_i q_i\right) \\ &= D_{\mathrm{KL}}\left(P_Y \| \prod_i P_{Y_i}\right) + \sum_i D_{\mathrm{KL}}\left(P_{Y_i} \| q_i\right) \\ &= I(Y) + \sum_i D_{\mathrm{KL}}\left(P_{Y_i} \| q_i\right) \end{aligned}$$

- Minimizing KL to a product model first projects onto marginals; the leftover is mutual information.

KL Pythagorean Identity on Gaussian Manifold

$$D_{\mathrm{KL}}\big(P_Y \| N(\Sigma)\big) = D_{\mathrm{KL}}\big(P_Y \| N(\mathrm{Cov}\, Y)\big) + D_{\mathrm{KL}}\big(N(\mathrm{Cov}\, Y) \| N(\Sigma)\big)$$

- $N(\Sigma)$ denotes the multivariate Gaussian distribution with covariance Σ and zero mean.
- $N(\mathrm{Cov}\, Y)$ is the Gaussian with the empirical mean and covariance matrix of Y.
- This formula expresses that KL divergence from P_Y to any Gaussian splits into the divergence to the best Gaussian fit (same covariance as Y) and the divergence between two Gaussian distributions (with covariances Cov Y and Σ).

Combining Pythagorean Theorems

Let us combine the Pythagorean theorem for independence with the one for Gaussianity, interpreting ICA as successive geometric approximations:

- For an N-vector with complicated joint distribution P_Y, two simplifying approximations are:
  - **Gaussian**: $P_Y^G \equiv \mathcal{N}(\text{Cov } Y)$ – best Gaussian fit (project onto $\mathcal{G}$).
  - **Product**: $P_Y^P \equiv \prod_i P_{Y_i}$ – distribution with independent marginals (project onto $\mathcal{P}$).
- Applying both approximations yields $P_Y^{PG} \equiv \mathcal{N}(\text{diag Cov } Y)$, a Gaussian with independent marginals matching those of Y.
- These yield two triangles through distribution space: $[P_Y \to P_Y^G \to P_Y^{PG}]$ and $[P_Y \to P_Y^P \to P_Y^{PG}]$, sharing the hypotenuse $[P_Y \to P_Y^{PG}]$.

Information Geometry in ICA

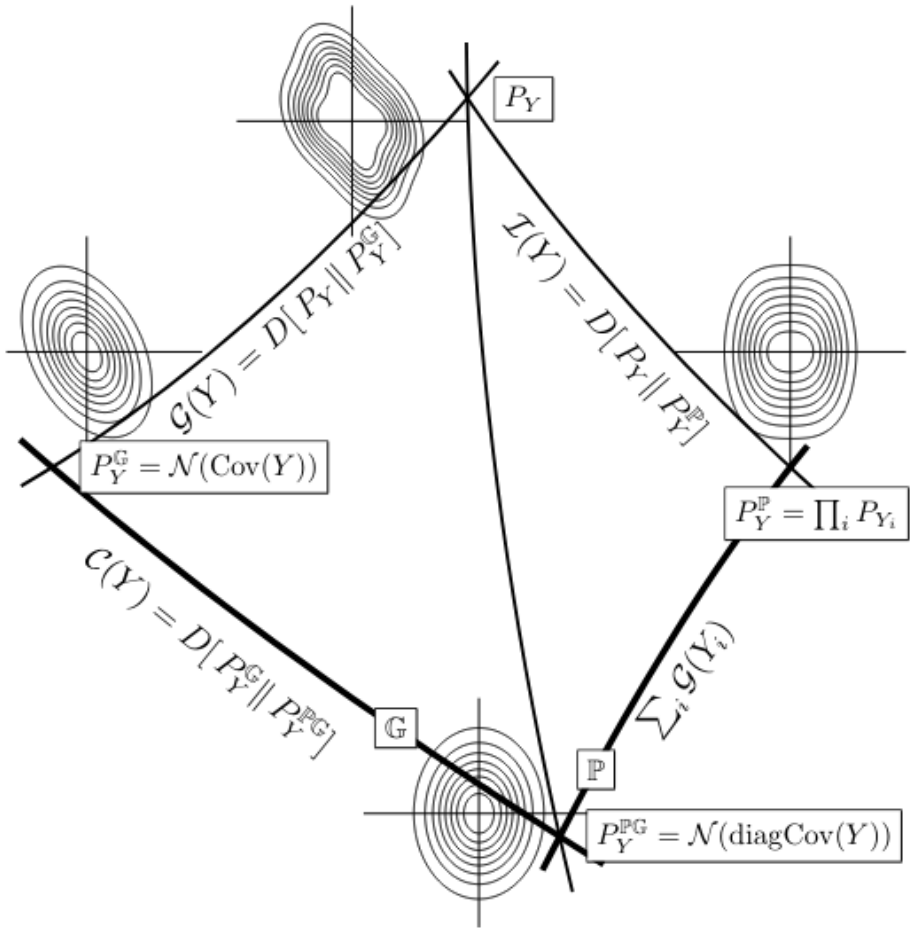


Figure 2: A probability density P_Y for a vector Y can be approximated as having independent components (approximation $P_Y^{\mathbb{P}}$) or as being Gaussian (approximation $P_Y^{\mathbb{G}}$) or both as $P_Y^{\mathbb{PG}}$. These approximations correspond to projections onto exponential manifolds. Those densities form two 'right triangles', each giving rise to a Pythagorean theorem, and sharing a common hypotenuse, thus relating the 'lengths' of the other sides, leading to Eq. (12) The lengths of all sides have a clear and simple statistical meaning, allowing to connect independence, correlation and non Gaussianity in a single information-geometric picture.

Combining Pythagorean Theorems

The KL divergence to the product-Gaussian approximation admits two complementary decompositions:

$$D\left[P_Y \| P_Y^{PG}\right] = D\left[P_Y \| P_Y^{P}\right] + D\left[P_Y^{P} \| P_Y^{PG}\right] = D\left[P_Y \| P_Y^{G}\right] + D\left[P_Y^{G} \| P_Y^{PG}\right]$$

- **Mutual Information:** $I(Y) = D[P_Y \| P_Y^P]$
- **Non-Gaussianity:** $G(Y) = D[P_Y \| P_Y^G]$
- **Correlation Scalar:**

$$C(Y) = D\left[P_Y^G \| P_Y^{PG}\right] = D[\mathcal{N}(\mathrm{Cov}\, Y) \| \mathcal{N}(\mathrm{diag}\, \mathrm{Cov}\, Y)]$$

Measures how far the Cov Y matrix is from just it's diagonal part and appears as the natural scalar measure of the correlation between entries of Y.

Non-Gaussianity, correlation and independence

- Define non-Gaussianity of Y:

$$G(Y) = D_{\mathrm{KL}}(P_Y \parallel N(\mathrm{Cov}(Y)))$$

- Cardoso's pythagorean theorem relates independence, correlation, and non-Gaussianity:

$$L(Y) + \sum_i G(Y_i) = C(Y) + G(Y)$$

- Where
  - $C(Y) = D_{\mathrm{KL}}(N(\mathrm{Cov}(Y)) \parallel N(\mathrm{diag}\,\mathrm{Cov}(Y)))$ measures how correlated the Gaussian approximation of Y and decorrelated? Y is.
  - $G(Y)$ is joint non-Gaussianity; $G(Y_i)$ is marginal non-Gaussianity.

Invariant KL Divergence of Gaussian Projection

If Y undergoes an invertible linear transformation, then its Gaussian projection $N(\operatorname{Cov} Y)$ undergoes the same transformation. The Kullback-Leibler divergence to the Gaussian projection is invariant:

$$D_{\mathrm{KL}}\left(P_Y \,\|\, N(\operatorname{Cov} Y)\right) = D_{\mathrm{KL}}\left(P_{A^{-1}Y} \,\|\, N(A^{-1} \operatorname{Cov} Y (A^{-1})^\top)\right)$$

That is, $G(Y)$ does not change under invertible linear transformations.

Whitening Reduces ICA to Maximizing Non-Gaussianity

- After whitening, $\mathrm{Cov}(Y) = I$, so $C(Y) = 0$: Pythagorean identity reduces to

$$L(Y) = 0 - \sum_i G(Y_i) + \text{const}$$

- ICA in whitened space therefore reduces to maximizing non-Gaussianity of the marginals, subject to orthogonality constraints.
- Contrast functions (kurtosis, negentropy, Wasserstein distances) measure this non-Gaussianity and solve ICA.
- Thus, Centering and whitening constrain data to the space where independence can be achieved solely by maximizing non-Gaussianity [**?**].

Motivations from Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables drawn from a non-Gaussian distribution with mean μ and variance σ^2.
- Define the sample mean as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- The CLT states that as $n \to \infty$, the distribution of

$$\sqrt{n} \left(\bar{X}_n - \mu \right) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2)$$

where the arrow $\xrightarrow{d}$ denotes convergence in distribution.

Weighted Sum Interpretation of CLT

- The sample mean can be seen as a weighted sum:

$$\bar{X}_n = \sum_{i=1}^{n} \frac{1}{n} X_i$$

  where all weights $\frac{1}{n}$ are equal.
- If viewed as an expectation, this weighted sum is

$$\mathbb{E}[X] = \sum_{i=1}^{n} w_i X_i, \quad \text{where} \quad w_i = \frac{1}{n}, \quad \sum_{i=1}^{n} w_i = 1.$$

- **Conclusion:** Any weighted mixture of non-Gaussian sources tends to be closer to a Gaussian distribution than the original sources themselves.

Why Independent Components Cannot be Gaussian

- The fundamental assumption in Independent Component Analysis (ICA) is that the independent components must be non-Gaussian for the model to be identifiable and ICA to work.
- Gaussian components cannot be separated since any linear combination of Gaussian variables is also Gaussian. This means the mixing matrix cannot be uniquely estimated.
- For Gaussian Multivariate distributions decorrelation implies independence, so after PCA, decorrelation, there is not much more ICA can achieve.
- Motivation from the CLT: weighted sums of independent non-Gaussian variables tend toward Gaussianity, so we seek sources that are maximally non-Gaussian.[**?**].

Introduction to Optimal Transport

- Optimal Transport (OT) is about moving "mass" in the most efficient way, akin to how one might move sand with minimal effort.
- The earth mover distance (EMD) quantifies this cost, measuring how much "work" it takes to transform one distribution into another.
- The problem was first formulated by Gaspard Monge in 1781, conceptualizing the minimal cost of moving soil [**?**].

Optimal Transport for Probability Distributions

- The core idea is to measure distances between probability measures using the cost of transporting one distribution into another.
- This led to metrics such as the Wasserstein distance, which quantify meaningful differences between distributions.
- The Kantorovich relaxation generalized Monge's formulation using transport plans, allowing greater mathematical and computational tractability [**?**, **?**, **?**].

Introducing Distance W_1

- W_1 distance (Earth Mover's Distance) measures the minimal total "effort" to move one distribution into another.
- For one-dimensional distributions with CDFs F and G, it can be written as:

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |F(x) - G(x)| \, dx$$

- Intuitively, W_1 sums the absolute vertical differences between the cumulative distribution functions.

Introducing Distance W_2 and Comparison

- W_2 distance uses squared differences and is given by:

$$W_2(\mu, \nu) = \left(\int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 \, dt \right)^{1/2}$$

  where F^{-1}, G^{-1} are the quantile functions (inverse CDFs).
- W_2 is more sensitive to differences in spread (variance) and central location (mean).
- When comparing distributions with $\mathcal{N}(0, 1)$, W_2 better captures deviations than W_1, making it more reliable for Gaussianity measures.
- W_1 can be more robust to outliers but might underestimate such deviations.

Motivation: Bounding Mixture Non-Gaussianity

- **Goal:** We want to justify why maximizing W_2 recovers independent sources.
- **Claim:** The non-Gaussianity of a mixture $w \cdot Z$ is upper bounded by the weighted sum of the non-Gaussianity of its independent parts.
- $$W_2^2(w \cdot Z, \mathcal{N}) \leq \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \mathcal{N})$$
- **Interpretation?:** By the Central Limit Theorem, mixtures are "more Gaussian" (have lower W_2 distance to $\mathcal{N}$) than the sources. Maximizing the LHS forces the weight vector w to align with a single source Z_i.

Proof Construction: The Common Source

To prove the bound, we construct a specific coupling using a common source of randomness.

- Let $\mathbf{N} = (N_1, \ldots, N_d)$ be a vector of d independent Standard Normal variables.
- Let T_i be the Optimal Transport map such that $(T_i)_{\#}\mathcal{N} = Z_i$.
- **Constructing the Mixture Variable X:**

$$X = \sum_{i=1}^{d} w_i T_i(N_i)$$

  Since $T_i(N_i) \sim Z_i$, it follows that $X \sim w \cdot Z$.
- **Constructing the Gaussian Target Y:** Using the *same* noise vector $\mathbf{N}$:

$$Y = \sum_{i=1}^{d} w_i N_i$$

  Since $\sum w_i^2 = 1$, it follows that $Y \sim \mathcal{N}(0, 1)$.

Proof Step: The Coupling Inequality

- The pair (X, Y) creates a valid joint distribution (coupling) $\pi^{\mathbf{N}}_{X,Y}$.
- **Definition of Wasserstein-2:** The infimum of cost over *all possible* couplings.

$$W_2^2(w \cdot Z, \mathcal{N}) = \inf_{\pi} \mathbb{E}_{\pi}[|x - y|^2]$$

- **The Inequality:** Since the infimum is the cost of the *best* plan, it must be less than or equal to the cost of *our specific* plan.

$$\underbrace{W_2^2(w \cdot Z, \mathcal{N})}_{\text{Cost of Optimal Plan}} \leq \underbrace{\mathbb{E}[|X - Y|^2]}_{\text{Cost of our constructed plan}}$$

Proof Step: Expanding the Square

Let us expand the expected squared difference $\mathbb{E}[|X - Y|^2]$:

$$|X - Y|^2 = \left| \sum_{i=1}^{d} w_i T_i(N_i) - \sum_{i=1}^{d} w_i N_i \right|^2$$

$$= \left(\sum_{i=1}^{d} w_i (T_i(N_i) - N_i) \right)^2$$

Squaring the sum yields diagonal terms ($i = j$) and cross terms ($i \neq j$):

$$= \sum_{i=1}^{d} w_i^2 (T_i(N_i) - N_i)^2 + \sum_{i \neq j} w_i w_j (T_i(N_i) - N_i)(T_j(N_j) - N_j)$$

Proof Conclusion: Vanishing Cross Terms

Taking the expectation $\mathbb{E}$:

- **Cross Terms ($i \neq j$):** vanish due to independence and zero mean.

$$\mathbb{E}[\text{Cross}] \propto \mathbb{E}[T_i(N_i) - N_i] \cdot \mathbb{E}[T_j(N_j) - N_j] = 0 \cdot 0 = 0$$

- **Diagonal Terms:**

$$\mathbb{E}[|X - Y|^2] = \sum_{i=1}^{d} w_i^2 \mathbb{E}[|T_i(N_i) - N_i|^2]$$

- Recall that $\mathbb{E}[|T_i(N_i) - N_i|^2]$ is exactly $W_2^2(Z_i, \mathcal{N})$.
- **Result:**

$$W_2^2(w \cdot Z, \mathcal{N}) \leq \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \mathcal{N})$$

This convexity adjacent property confirms that mixtures reduce the Wasserstein distance to Gaussianity.

Condition for Equality (1D Case)

Brenier's Theorem and Comonotonicity:

- We compare two scalar variables, X and Y, constructed from the random vector $\mathbf{N} \in \mathbb{R}^d$.
- In the 1D case, the Wasserstein-2 distance satisfies $W_2^2(X, Y) = \mathbb{E}[|X - Y|^2]$ **if and only if** X and Y are **comonotonic**.
- This means there exists a strictly increasing function f such that $X = f(Y)$.
- **Implication:** For this to hold, the gradients of X and Y with respect to the source noise $\mathbf{N}$ must be parallel at every point $\mathbf{N}$:

$$\nabla X(\mathbf{N}) = \lambda(\mathbf{N}) \nabla Y(\mathbf{N})$$

Note: The scaling factor $\lambda(\mathbf{N})$ *is a scalar function of* $\mathbf{N}$.

Definitions and Regularity

Assumption (Regularity): We assume the source distributions possess smooth, strictly positive densities. This guarantees that the optimal transport maps T_i are **differentiable** functions (Luis Caffarelli regularity).

Definitions:

- Noise Vector: $\mathbf{N} = [N_1, N_2, \ldots, N_d]^\top$
- Projection Vector: $\mathbf{w} = [w_1, w_2, \ldots, w_d]^\top$

1. The Gaussian Target Gradient:

$$Y(\mathbf{N}) = \sum w_i N_i \implies \nabla Y = \mathbf{w}$$

2. The Source Mixture Gradient:

$$X(\mathbf{N}) = \sum w_i T_i(N_i) \implies \nabla X = [w_1 T_1'(N_1), \ldots, w_d T_d'(N_d)]^\top$$

Applying the Product Rule

The Equality Condition: We require $\nabla X = \lambda(\mathbf{N})\nabla Y(\mathbf{N}) = \lambda(\mathbf{N})\mathbf{w}$. To test if this holds, we differentiate both sides with respect to $\mathbf{N}$ to compute the Hessian matrices.

- **LHS (Hessian of X):** Since X separates into independent terms $w_i T_i(N_i)$, the cross-derivatives $\frac{\partial^2 X}{\partial N_i \partial N_j}$ are zero. This yields a **Diagonal Matrix**.
- **RHS (Product Rule):** We differentiate the product $\lambda(\mathbf{N})\mathbf{w}$. Since $\mathbf{w}$ is constant:

$$\frac{\partial}{\partial \mathbf{N}} \left(\lambda(\mathbf{N})\mathbf{w}\right) = \mathbf{w}(\nabla\lambda)^\top$$

This yields a **Rank-1 Matrix** (Outer Product).

Visualizing the Matrix Equation

We equate the LHS (Diagonal) and the RHS (Outer Product):

$$\underbrace{\begin{bmatrix} w_1 T_1'' & 0 & \cdots & 0 \\ 0 & w_2 T_2'' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_d T_d'' \end{bmatrix}}_{\text{Hessian of } X \text{ (Diagonal)}} = \underbrace{\begin{bmatrix} w_1 \frac{\partial \lambda}{\partial N_1} & w_1 \frac{\partial \lambda}{\partial N_2} & \cdots & w_1 \frac{\partial \lambda}{\partial N_d} \\ w_2 \frac{\partial \lambda}{\partial N_1} & w_2 \frac{\partial \lambda}{\partial N_2} & \cdots & w_2 \frac{\partial \lambda}{\partial N_d} \\ \vdots & \vdots & \ddots & \vdots \\ w_d \frac{\partial \lambda}{\partial N_1} & w_d \frac{\partial \lambda}{\partial N_2} & \cdots & w_d \frac{\partial \lambda}{\partial N_d} \end{bmatrix}}_{\text{Outer Product } \mathbf{w}(\nabla \lambda)^\top \text{ (Rank-1)}}$$

The Contradiction

Let us look at any **off-diagonal** term (row i, column j where $i \neq j$) from the matrix equation on the previous slide.

- **LHS says:** The entry is **0**.
- **RHS says:** The entry is $w_i \cdot \frac{\partial \lambda}{\partial N_j}$.

The Implication:

$$w_i \cdot \frac{\partial \lambda}{\partial N_j} = 0$$

Since we are in a mixture, the weights are non-zero ($w_i \neq 0$). Therefore, it must be that:

$$\frac{\partial \lambda}{\partial N_j} = 0 \quad \text{for all } j \neq i$$

Conclusion: $\nabla \lambda = \mathbf{0}$. The function $\lambda(\mathbf{N})$ is actually a **scalar**.

Conclusion: Proof of Strict Inequality

We have proven that equality holds **if and only if** λ is constant.

$$T_i'(N_i) = \lambda \implies T_i(x) = \lambda x + c$$

1. **The Gaussian Case (Identity Map):**
   - Sources are Gaussian $\rightarrow$ $T_i(x) = x$. This is linear ($\lambda = 1$).
   - The condition is met. **Equality holds.**
2. **The Non-Gaussian Case (Curved Map):**
   - Sources are non-Gaussian $\rightarrow$ T_i is non-linear (curved).
   - The derivative T_i' is not constant.
   - The condition fails.

Strict Inequality Result

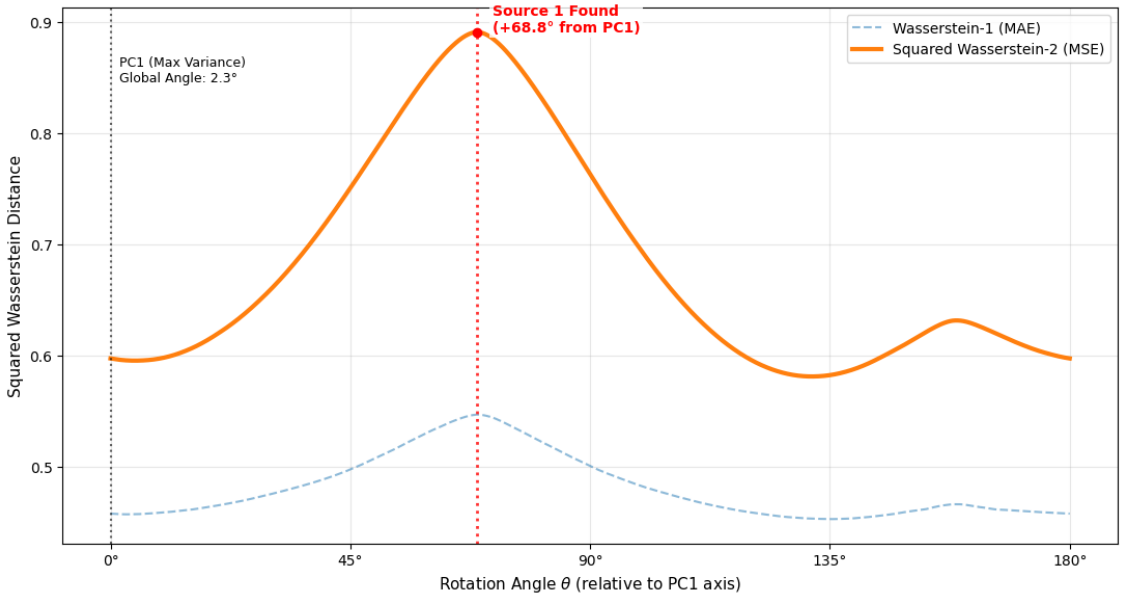
For any mixture of non-Gaussian independent sources, the bound is strict:

$$W_2^2(w \cdot Z, \mathcal{N}) < \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \mathcal{N})$$

Empirical Comparison: Experimental Setup

- To select the optimal contrast function, we empirically compared the optimization landscape of W_1 (Mean Absolute Error) versus squared W_2 (Mean Squared Error).
- **Setup:**
  - Synthetic data generated from Student-t distributions.
  - We varied the degrees of freedom (ν) to control "non-Gaussianity":
    - $\nu = 2$: Heavy tails (Strong non-Gaussian signal).
    - $\nu = 10$: Approaching Normal (Weak non-Gaussian signal).
  - We plotted the distance metric over the rotation angle θ relative to the first Principal Component.

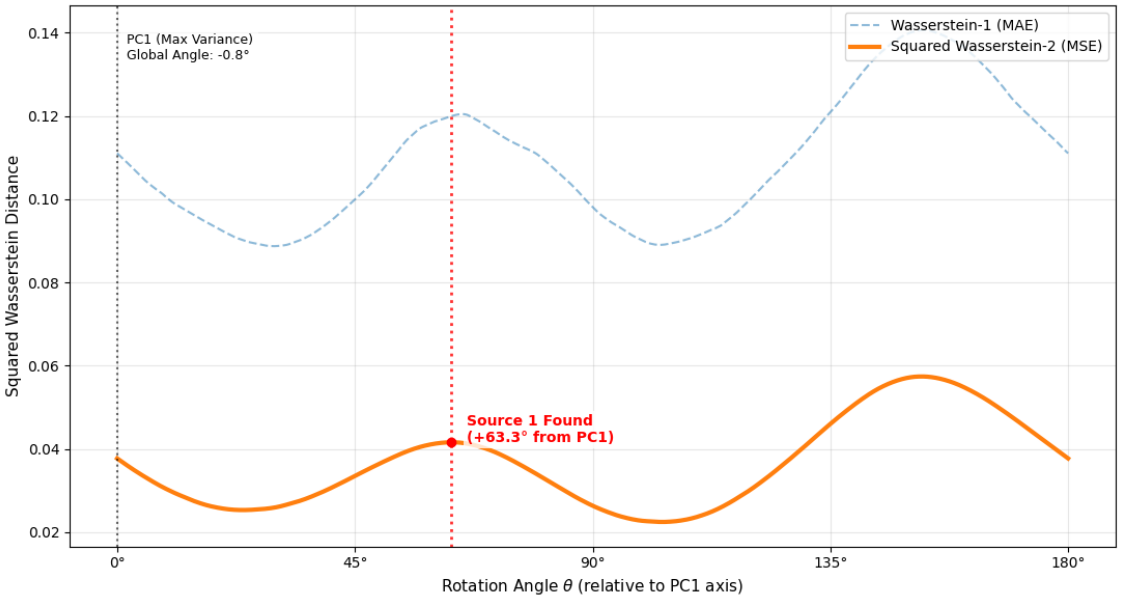
Case 1: Strong Non-Gaussianity ($\nu = 2$)



Observation:

- Both W_1 and W_2 successfully identify the source (peaks align).
- W_2 (orange) exhibits a significantly sharper peak and higher curvature around the optimum.
- This suggests W_2 provides stronger gradients for optimization when the signal is distinct.

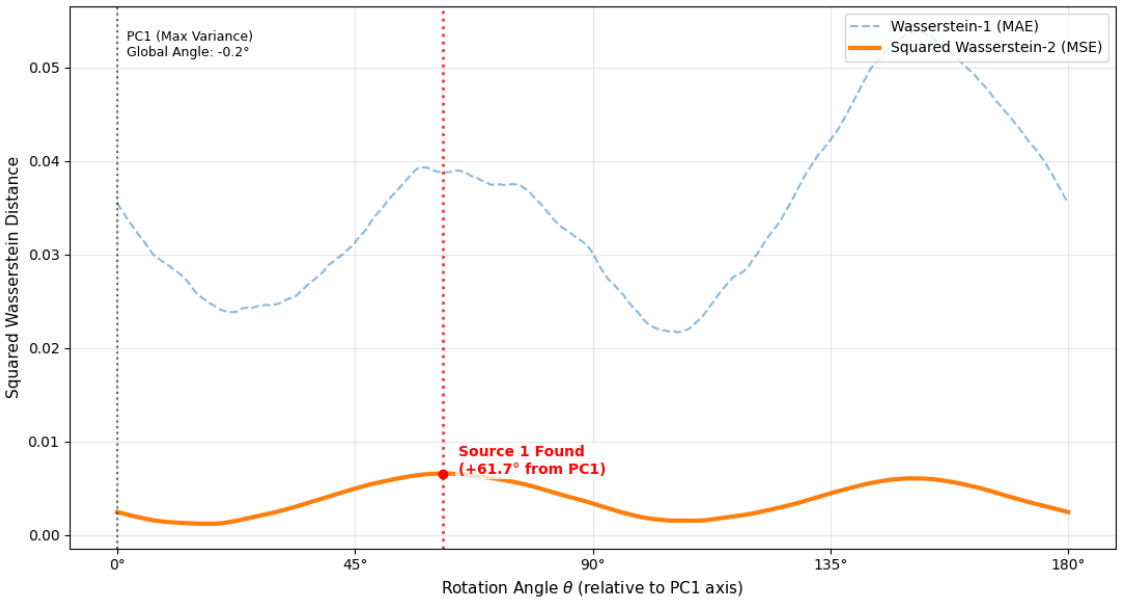
Case 2: Medium Non-Gaussianity ($\nu = 4$)



Observation:

- As tails become lighter, the signal weakens.
- W_2 maintains a smooth, convex-like profile around the optimum.
- W_1 begins to show irregularities in the landscape, though the global maximum is still recoverable.

Case 3: Weak Non-Gaussianity ($\nu = 10$)



Crucial Difference:

- The signal is now very close to Gaussian noise.
- W_1 **(blue dashed):** Becomes jagged and noisy. This "rough" landscape traps gradient-based solvers in spurious local optima.
- W_2 **(orange):** Remains smooth and unimodal.
- **Conclusion:** W_2 is slightly robust for weak signals, enabling recovery where W_1 fails.

Core Idea: Optimization Problem with W2

- Traditional ICA formulations optimize kurtosis or negentropy contrast functions.
- Our approach formulates ICA as finding the rotation $\mathbf{u} \in S_n$ that maximizes the W2 distance (W_2) from the standard Gaussian $\mathcal{N}(0,1)$:

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u} \in S_n} W_2\left(\mathbb{P}_{\mathbf{u}^\top \mathbf{X}}, \mathcal{N}(0,1)\right)$$

- This perspective leverages the optimal transport metric's ability to capture deviations from Gaussianity in a geometrically meaningful way.

Core Idea: Algorithm

Algorithm 1 ICA with W2 Distance Optimization

1: Obtain observed mixture data $\mathbf{X}$
2: Perform whitening/decorrelation on $\mathbf{X}$
3: Initialize empty set of recovered components $\mathcal{U} = \emptyset$
4: **while** number of components not found **do**
5: Find local optimum (Gradient Descent) $\mathbf{u}^* \in S_n$ maximizing

$$W_2\left(\mathbb{P}_{\mathbf{u}^\top \mathbf{X}}, \mathcal{N}(0,1)\right)$$

6: Store candidate: $\mathcal{U} \leftarrow \mathcal{U} \cup \{\mathbf{u}^*\}$
7: Restrict search space by enforcing orthogonality:

$$\mathbf{u} \perp \operatorname{span}(\mathcal{U})$$

8: **end while**
9: **return** $\mathcal{U}$ as estimated independent components

Why W2-ICA? A Comparison with FastICA

- **The FastICA Approach:** Relies on proxy contrast functions (e.g., Kurtosis or Negentropy approximations). These capture specific moments (fat tails, peaks) but ignore the full geometric structure of the distribution.
- **The W2-ICA Advantage:** The Wasserstein-2 metric quantifies the true global geometric cost of transforming the empirical distribution into a Gaussian.
- **Robustness to Weak Signals:** For highly complex, multimodal, or weakly non-Gaussian signals where polynomial proxies fail or create jagged optimization landscapes, W_2 provides a smooth, globally aware gradient.
- **Trade-off:** W_2 involves sorting operations $\mathcal{O}(N \log N)$, making it computationally heavier. We introduce several mathematical optimizations to close this performance gap.

Optimization 1: Exact Analytical Gaussian Target

- **Previous Bottleneck:** Approximating the target Gaussian involved sampling or using the standard empirical quantile function: $q_i = \Phi^{-1}((i - 0.5)/N)$. This introduces approximation noise into the gradient.
- **Analytical Solution:** We compute the *exact* expected value of a Standard Normal variable within each discrete quantile bin using integral calculus.
- Let the bin edges be $p_i = \frac{i}{N}$, with corresponding Gaussian boundaries $z_i = \Phi^{-1}(p_i)$.
- The exact target value T_i for the i-th sorted sample is:

$$T_i = N \int_{z_{i-1}}^{z_i} x\phi(x)\, dx = N\left(\phi(z_{i-1}) - \phi(z_i)\right)$$

  where $\phi(x)$ is the standard normal PDF.
- **Result:** Eliminates sampling noise.

Optimization 2: Batched Vectorization

- **The Challenge:** Finding the global optimum requires multiple random initializations (restarts). Evaluating these sequentially via standard iterative loops is highly inefficient.
- **Vectorized Approach:** We aggregate B random restarts into a single tensor $\mathbf{W}_{\text{batch}} \in \mathbb{R}^{B \times d}$.
- **Parallel Execution:**
  1. **Projection:** $\mathbf{Y} = \mathbf{W}_{\text{batch}}\mathbf{X}$ is computed for all B restarts simultaneously.
  2. **Sorting:** PyTorch sorts all B rows of $\mathbf{Y}$ in parallel.
  3. **Broadcasting:** The exact analytical target $\mathbf{T}$ is broadcast-subtracted across the batch to compute distances and gradients in one step.
- **Result:** Fully saturates CPU/GPU hardware, calculating dozens of trajectories in the time it previously took to compute one.

Optimization 3: Riemannian Gradient on the Stiefel Manifold

Formal Math:

- Unmixing matrices must be orthogonal: $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$. The space of such matrices forms the **Stiefel Manifold** $\mathcal{S}$.
- The standard Euclidean gradient $\mathbf{G} = \nabla_{\mathbf{W}} W_2^2$ points into flat space, violating orthogonality constraints.
- We project $\mathbf{G}$ directly onto the tangent space of the manifold at $\mathbf{W}$:

$$\nabla_{\mathcal{S}}\mathbf{W} = \mathbf{G} - \frac{1}{2}\left(\mathbf{G}\mathbf{W}^\top + \mathbf{W}\mathbf{G}^\top\right)\mathbf{W}$$

- **Intuitive Vector View:** The term $(\mathbf{G}\mathbf{W}^\top + \mathbf{W}\mathbf{G}^\top)$ measures the exact symmetric violation of orthogonality. We subtract this violation from $\mathbf{G}$, ensuring our update step perfectly hugs the curvature of the manifold.

Optimization 3: Retraction via Symmetric Decorrelation

- After stepping along the flat tangent plane ($\mathbf{W}_{\text{step}} = \mathbf{W} + \eta \nabla_{\mathcal{S}} \mathbf{W}$), the matrix slightly lifts off the curved manifold. We must "retract" it.
- **Formal Retraction:**

$$\mathbf{W}_{\text{new}} = (\mathbf{W}_{\text{step}} \mathbf{W}_{\text{step}}^{\top})^{-1/2} \mathbf{W}_{\text{step}}$$

- **Why not Gram-Schmidt?**
  - Gram-Schmidt is asymmetric; it permanently fixes the first vector and modifies the rest to fit, which inherently biases simultaneous optimization.
  - Symmetric decorrelation computes the overlap covariance ($\mathbf{W}\mathbf{W}^{\top}$) and applies a uniform, symmetric push ($-1/2$ power) to all vectors simultaneously, treating all independent components equally.

Optimization 4: Novel OT Fixed-Point Iteration

- We formulated a gradient-free-like update rule that exploits the Optimal Transport mapping directly, acting as a Wasserstein-specific fixed-point step.
- **Step 1 (The Ideal Target):** Sort current projections $\mathbf{Y} = \mathbf{WX}$ to find the permutation matrix $\mathbf{P}$. The ideal Gaussian target mapped back to the original data order is:

$$\mathbf{Y}_{\text{ideal}} = \mathbf{P}^{-1}\mathbf{T}$$

- **Step 2 (The Pull):** The cross-covariance computes the exact linear transformation required to push our data toward pure Gaussian noise:

$$\mathbf{G}_{\text{OT}} = \frac{1}{N-1}\mathbf{Y}_{\text{ideal}}\mathbf{X}^{\top}$$

- **Step 3 (The Anti-Gaussian Step):** ICA maximizes non-Gaussianity. We subtract this transformation to step directly *away* from the Gaussian valley:

$$\mathbf{W}_{\text{new}} = \text{Retract}(\mathbf{W} - \eta\mathbf{G}_{\text{OT}})$$

The Blind Spot of FastICA

- FastICA is the industry standard due to its fast, scale-invariant Newton-step optimizer.
- It maximizes Negentropy using proxy contrast functions, typically the `logcosh` function: $G(x) = \frac{1}{a} \log \cosh(ax)$.
- **The Mathematical Trap:** FastICA relies on **Assumption A.5** for its Newton step to converge. The denominator of its update rule is proportional to:

$$\mathbb{E}[x \cdot g(x) - g'(x)]$$

  where $g(x) = \tanh(x)$ and $g'(x) = 1 - \tanh^2(x)$.
- If a non-Gaussian source distribution naturally causes this expectation to equal zero, the algorithmic curvature vanishes. FastICA is mathematically blinded and will fail to converge.

Engineering the A.5 Trap: The Trimodal Gaussian

- To prove this limitation, we engineered a continuous distribution that forces the A.5 expectation to exactly zero.
- The target function $h(x) = x\tanh(x) - 1 + \tanh^2(x)$ is negative near zero ($|x| < 0.82$) and positive elsewhere.
- We constructed a **Trimodal Gaussian Mixture** parameterized by peak locations $\{-b, 0, b\}$ and variance σ^2.
- By solving $\int h(x)\text{PDF}(x)dx = 0$ subject to the unit variance constraint ($E[X^2] = 1$), we found the exact probability masses that perfectly balance the negative center against the positive tails.

Experimental Results: Breaking FastICA

- **Setup:** We mixed the engineered Trimodal Gaussians in 10, 20, and 30 dimensions with 50,000 samples and compared FastICA to W-ICA.
- **FastICA Results:** As mathematically predicted, FastICA hit its maximum iteration limit without converging, yielding unmixing matrices with catastrophically high Amari Errors (> 1.0).
- **W-ICA Results:** W-ICA successfully unmixed the 10D and 20D signals (Amari Error ~ 0.1).
- **Why W-ICA Succeeds:** W-ICA maps the *entire global geometry* of the empirical CDF to a Gaussian. It uses first-order optimization on the Stiefel manifold, completely bypassing the local zero-curvature traps that destroy Newton-step proxy methods.

The Dimensionality Ceiling of W-ICA

- While W-ICA bypassed the mathematical trap, its performance degraded significantly at 30 dimensions with 50,000 samples.
- **The Cause:** The Central Limit Theorem.
- A linear mixture of 30 independent trimodal sources creates a highly complex, overwhelmingly Gaussian-looking geometry.
- **Sample Starvation:** To distinguish the microscopic "steps" of this 30D discrete-like mixture from a true continuous Gaussian, the empirical CDF requires a massive sample size. Without it, the Wasserstein gradients become overwhelmed by sample noise.
- W-ICA's limitation is empirical (resolution), whereas FastICA's is structural (mathematical).

The Outlier Problem in Exact Wasserstein

- We established that W_2 provides superior, smooth gradients for weak signals compared to W_1.
- However, because W_2 uses an L2 ground cost $c(x, y) = |x - y|^2$, it is highly sensitive to outliers (e.g., sensor artifacts).
- A single massive outlier forces the optimal transport plan to expend immense gradient effort to "pull" the outlier, destroying the unmixing matrix for the normal data.
- The fact that W_2 is a strict mathematical metric guarantees geometric stability, but **not** statistical robustness.

W-Huber: Robust Geometry via Ground Cost Modification

- To combine the smooth gradients of W_2 with the outlier resistance of W_1, we modify the Optimal Transport ground cost.
- We replace the L2 cost with a stable `logcosh` formulation:

$$c(x, y) = \log(\cosh(x - y)) \approx |x - y| + \log(1 + e^{-2|x-y|}) - \log(2)$$

- **Result:** For small distances (normal data), it behaves like an L2 penalty, preserving the smooth Stiefel gradients. For large distances (outliers), it behaves like an L1 penalty, ignoring extreme noise.
- Unlike FastICA, using `logcosh` as a ground cost inside W-ICA *does not* trigger the A.5 trap, as W-ICA relies on first-order global matching rather than second-order local curvature.

The Discrete Case: A Pathological Landscape

- Real-world signals can be discrete (e.g., binary states, digital communication). To test this, we evaluated a mixture of independent Bernoulli sources ($\{-1, 1\}$).
- **FastICA Dominates:** The Bernoulli distribution has an excess kurtosis of -2 (the absolute theoretical minimum). FastICA's proxy metrics act as heat-seeking missiles for this extreme, sharp "hypercube" geometry.
- **W-ICA Fails:** Optimal Transport matches the empirical CDF to a continuous Gaussian. For binary data, any 1D projection creates a step-like, rigid "staircase" CDF, which fundamentally breaks the continuous W_2 geometry.

The Shattered Optimization Landscape

- **The Sorting Problem:** W_2 relies on sorting projected points. As the optimizer rotates the unmixing matrix, clumps of discrete points slide past each other, causing the sorting order to abruptly jump.
- **Spurious Maxima:** Every sorting jump instantly changes the gradient direction. This shatters the smooth W_2 landscape into a jagged mountain range of combinatorial local optima.
- **The Hypercube Trap:** W-ICA climbs the nearest peak (often a diagonal projection of the hypercube that vaguely approximates a binomial step-curve) and becomes permanently trapped, yielding catastrophic Amari errors.

Algorithmic Fix 1: Gaussian Dithering

- **The Goal:** Restore the smooth gradients of the Stiefel manifold without resorting to computationally expensive $O(N^2)$ Entropic Regularization (Sinkhorn distances).
- **The Method:** We inject a microscopic amount of continuous, zero-mean Gaussian noise ($\sigma \approx 0.01$) to the projections immediately before sorting.
- **Geometric Effect:** This mathematically convolves the discrete Dirac spikes with a Gaussian kernel, "melting" the rigid staircase into a strictly increasing, smooth continuous curve.
- **Result:** Eliminates non-differentiable sorting ties and restores stable gradients while maintaining $O(N \log N)$ computational efficiency.

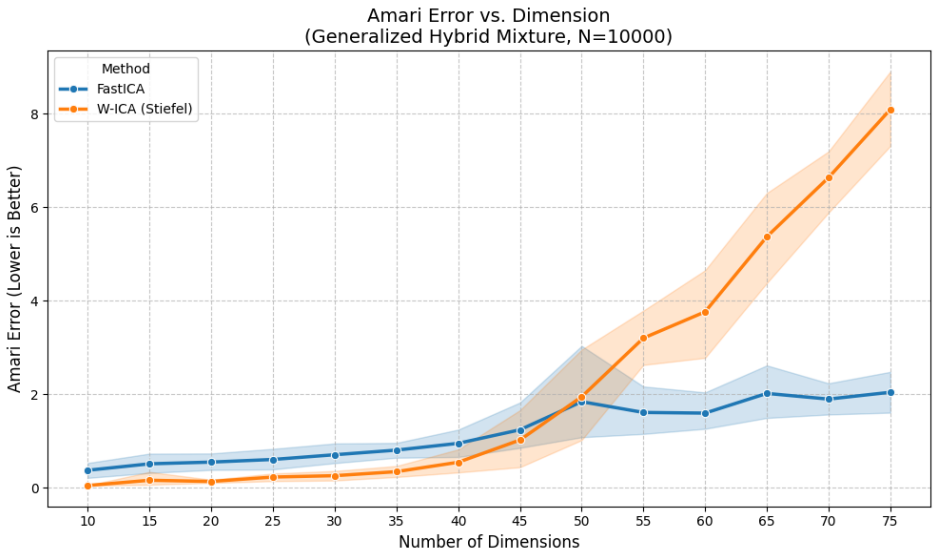
Algorithmic Fix 2: Stochastic Mini-Batching

- **The Limitation of Full-Batch:** Evaluating the exact W_2 distance on all N samples guarantees convergence to the *nearest* local maximum, which is fatal in a landscape riddled with shallow discrete traps.
- **The Method:** We replace full-batch gradient ascent with Stochastic Mini-Batching (evaluating random subsets of 512 or 1024 samples per step).
- **Simulated Annealing:** The random subsampling introduces stochastic gradient noise. A trap for the full dataset may not be a trap for a random subset, allowing the optimizer to "bounce" out of local hypercube diagonals.
- **Conclusion:** Combining dithering (smoothing the steps) with stochastic batching (escaping the traps) allows W-ICA to successfully separate discrete mixtures up to strict high-dimensional sample limits.

The Ultimate Stress Test: Generalized Mixtures

- Real-world signals rarely belong to a single statistical family. To test the true robustness of W-ICA, we designed a highly chaotic, generalized mixture.
- **The Setup:** We evaluated mixtures across 10 to 75 dimensions with $10,000$ samples.
- **Source Composition:**
  - Exactly **one** Standard Gaussian source $\mathcal{N}(0, 1)$.
  - The remaining sources were randomly drawn from a diverse pool of 8 distinct distributions: Laplace, Bernoulli, Uniform, Student-t, Poisson, Binomial, Chi-squared, and Exponential.
- **W-ICA Configuration:** We maintained the robust settings engineered for the discrete case ($\sigma_{\text{dither}} = 0.01$, batch size $= 1024$) to handle the unpredictable presence of discrete and light-tailed components.

Results: W-ICA Stability vs. FastICA Degradation



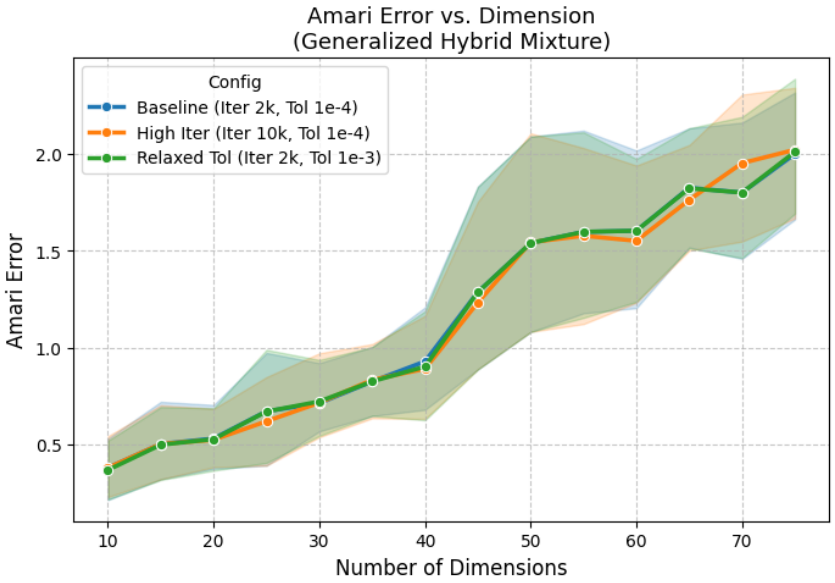
Key Observations:

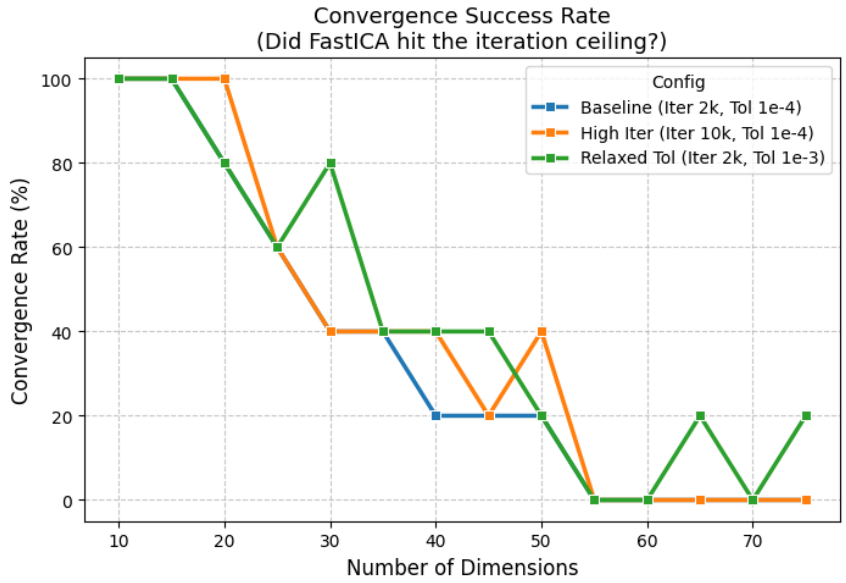
- **W-ICA (Orange):** Successfully maintains a low Amari error (< 0.25) up to 35 dimensions, proving the geometric W_2 metric generalizes across highly diverse topological landscapes.
- **FastICA (Blue):** Performance degrades rapidly, crossing the critical 0.5 Amari error threshold immediately after 20 dimensions.
- At higher dimensions, the hyper-complex mixture completely blinds FastICA's proxy contrast functions.

Diagnosing FastICA: The Oscillation Problem

- During the generalized experiment, FastICA consistently triggered non-convergence warnings.
- **The Hypothesis:** FastICA's Newton-Raphson solver assumes a relatively uniform curvature. When forced to simultaneously evaluate highly kurtotic (Laplace) and negatively kurtotic (Uniform, Bernoulli) signals, the gradient updates contradict each other, causing the solver to oscillate indefinitely.
- **The Deep-Dive Test:** To prove this was a structural failure and not merely a lack of compute time, we evaluated FastICA under extreme optimization parameters:
  1. Baseline (`max_iter=2000, tol=1e-4`)
  2. High Iterations (`max_iter=10000, tol=1e-4`)
  3. Relaxed Tolerance (`max_iter=2000, tol=1e-3`)

Convergence Failure in Complex Landscapes: Results





Convergence Failure in Complex Landscapes: Conclusions

Conclusion:

- Even when granted 5x the normal iteration limit (10,000 steps), FastICA's convergence rate plummets to 0% past 30 dimensions.
- Relaxing the tolerance only marginally delays this collapse.
- This definitively proves that FastICA is structurally oscillating.
- In unpredictable, hybrid environments, proxy-based Newton solvers become trapped in contradictory gradient loops, whereas W-ICA's first-order global CDF matching remains stable.

Why Does FastICA Fail in General setting?

- **Top Performers:** *Continuous Only* and *Strictly Super-Gaussian* are the definitive best, consistently achieving 100% convergence and maintaining the lowest Amari Errors across both dimensions.
- **Complete Failure:** The *Discrete Only* pool type failed entirely, with 0% convergence and exceptionally high errors at both dimensions.
- **Struggling Performers:** *Full Hybrid* and *Zero Gaussian* showed inconsistent convergence rates (varying between 40% and 60%) and significantly higher errors than the top performers.
- **Dimensionality Impact:** Increasing the dimension from 30 to 40 degraded performance across the board, resulting in higher Amari Errors for all methods and dropping the convergence rate for the Full Hybrid model.

Performance Visualizations

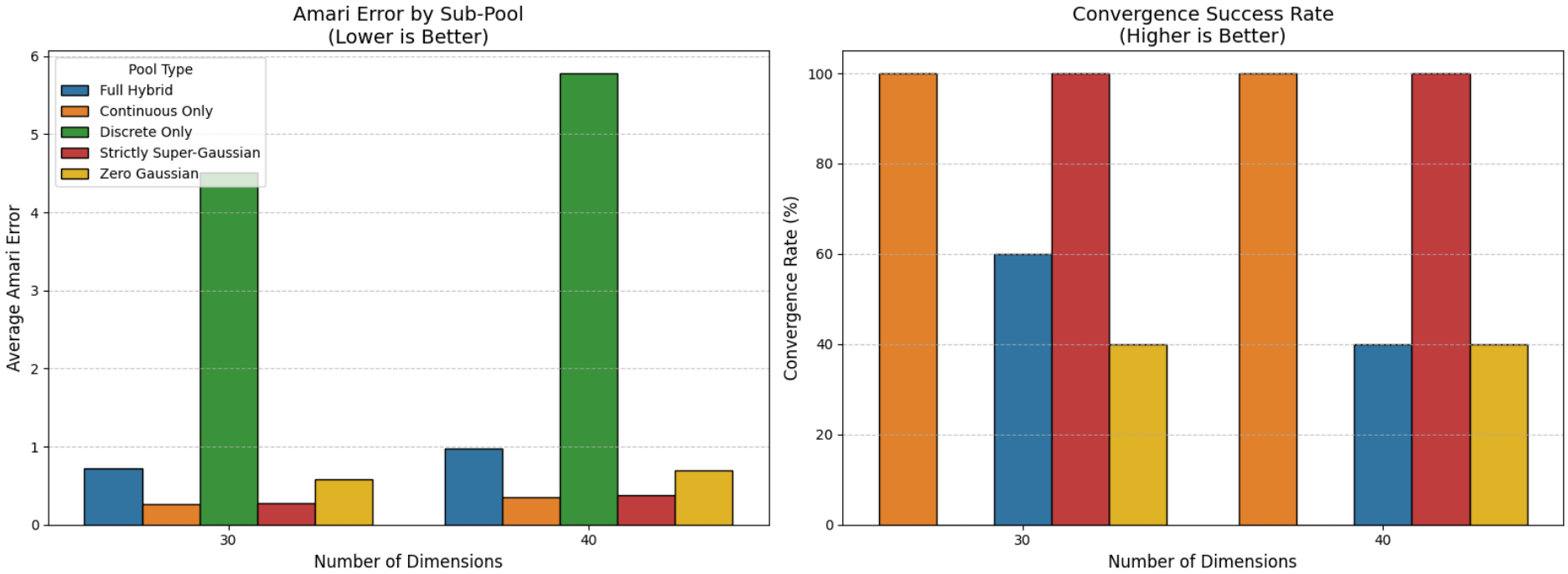


Figure: Amari Error (left) and Convergence Success Rate (right)

Deconstructing the Discrete Failure

- We established that pure discrete pools cause catastrophic failure. To understand *why*, we isolated the three discrete distributions (Bernoulli, Poisson, Binomial) in independent tests.
- **Setup:** We generated mixtures containing 1 Gaussian source and $(N-1)$ sources of a *single* discrete type at Dimensions 30 and 40 (10,000 samples).
- **The Goal:** Determine if the failure is driven by binary state-flips (Bernoulli) or by count-based discreteness (Poisson/Binomial).

FastICA: The Bernoulli Catastrophe

- **Bernoulli (Binary Flips):** FastICA fails catastrophically. It achieved 0% convergence at both Dim 30 and Dim 40, with massive Amari Errors (~ 1.0 and ~ 1.3).
- **Poisson & Binomial (Count Data):** FastICA performed significantly better, achieving 100% convergence at Dim 30 and 80% at Dim 40, with relatively low errors (~ 0.2).
- **The Diagnosis:** FastICA's proxy contrast functions break down specifically when faced with the extreme hypercube geometry and minimal kurtosis (-2) of purely binary Bernoulli signals.

Performance Visualization: FastICA Discrete Isolation

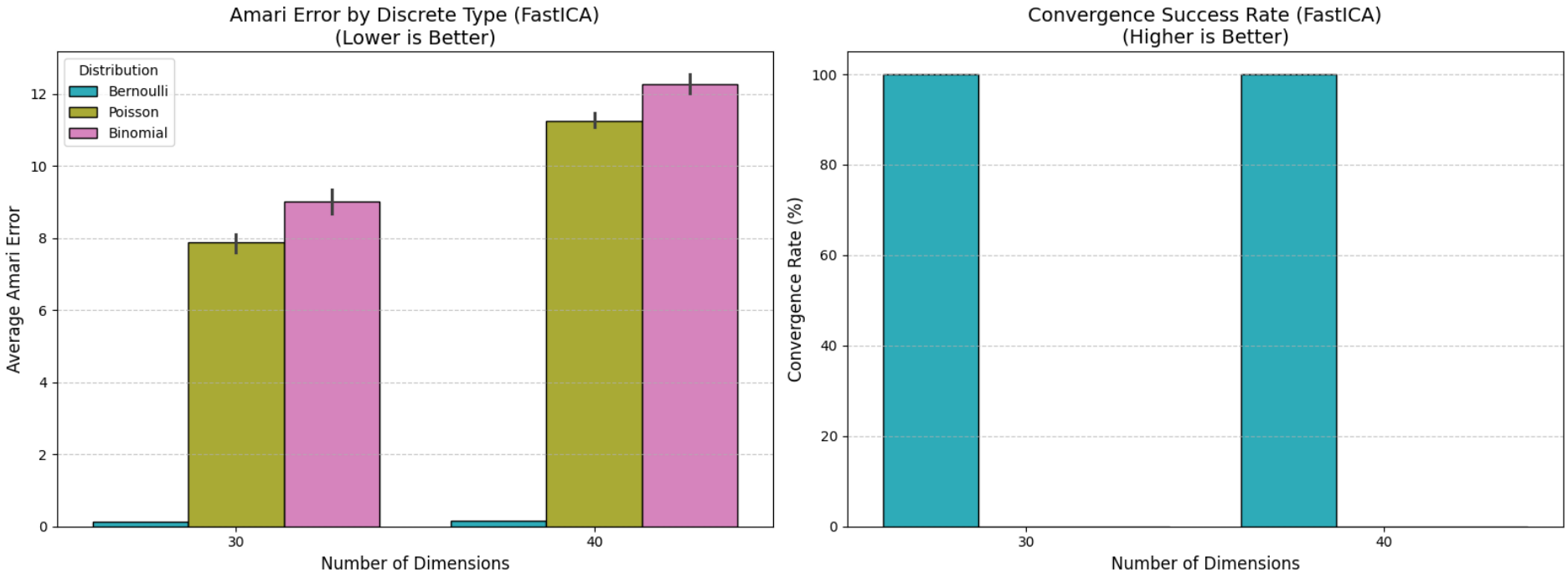


Figure: FastICA: Amari Error and Convergence across isolated discrete types.

OT-ICA: The Fundamental Sorting Trap

- We ran the exact same isolation study using OT-ICA (W-ICA).
- **The Result:** OT-ICA failed universally across *all* discrete types (Bernoulli, Poisson, Binomial).
- While it often reports algorithmic "convergence" (finding a local maximum), it consistently converges to spurious, incorrect unmixing matrices, yielding catastrophic Amari Errors (> 1.2) across the board.
- **The Diagnosis:** This confirms our hypothesis regarding the shattered W_2 optimization landscape. Any form of discreteness creates step-like empirical CDFs. The continuous Wasserstein metric cannot establish meaningful gradients across these rigid "staircases," trapping the optimizer regardless of the specific discrete shape.

Performance Visualization: OT-ICA Discrete Isolation

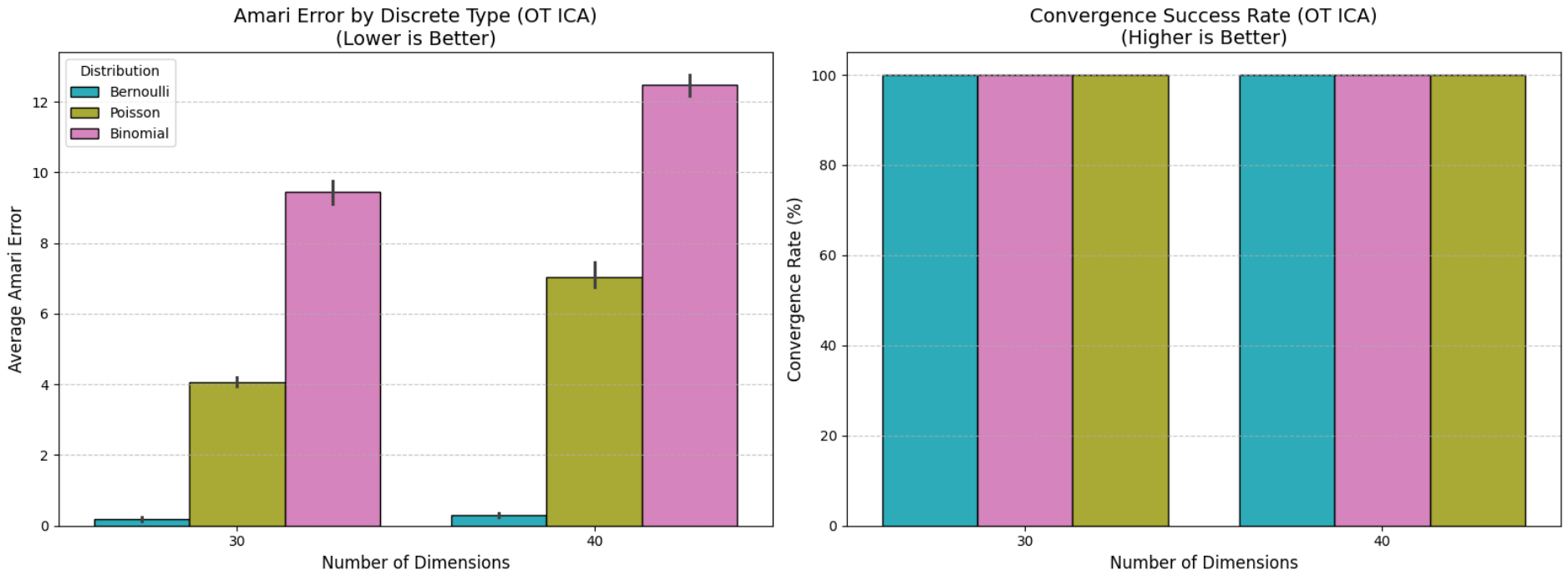


Figure: OT-ICA: High Amari Errors demonstrate failure across all isolated discrete types.

The OT-ICA Advantage: Hybrid Robustness

- If OT-ICA fails on purely discrete pools, why use it over FastICA?
- **The Answer: Real-World Hybridization.** Real signals are rarely purely discrete; they are noisy, continuous, or complex mixtures.
- As demonstrated in our Sub-Pool ablation studies:
  - **FastICA breaks** when faced with a "Full Hybrid" pool (dropping to 40% convergence with high error).
  - **OT-ICA thrives** in a "Full Hybrid" pool, maintaining perfect stability and low Amari Errors.
- **Conclusion:** While pure discreteness breaks the optimal transport sorting mechanism, the presence of continuous distributions in a hybrid mix provides enough smooth geometric structure to guide the W_2 metric past the discrete traps—an environment where FastICA's static proxies become overwhelmed.

Intractability of an Exact Newton Step

- FastICA achieves rapid convergence using a Newton (second-order) fixed-point step because its contrast function $g(\cdot)$ is static.
- To build a "Fast-Wasserstein-ICA", we would need the Hessian of the Wasserstein distance: $\mathbf{H} = \nabla^2_{\mathbf{w}} W_2^2$.
- The first derivative (Gradient) uses the Optimal Transport map $T_{\mathbf{w}}$:

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} W_2^2 \right) = \mathbb{E} \left[\mathbf{X} \left(\mathbf{w}^\top \mathbf{X} - T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}) \right) \right]$$

- The Hessian requires differentiating $T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X})$ with respect to $\mathbf{w}$. Applying the total derivative yields two terms:

$$\nabla_{\mathbf{w}}[T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X})] = \underbrace{T'_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X})\mathbf{X}}_{\text{Argument Derivative}} + \underbrace{(\nabla_{\mathbf{w}} T_{\mathbf{w}})(\mathbf{w}^\top \mathbf{X})}_{\text{Map Derivative}}$$

- Caffarelli's regularity guarantees the first term exists. The second term, however, is structurally problematic.

The Density Estimation Bottleneck

Let us expand the problematic Map Derivative. The 1D transport map to a standard Gaussian relies on the empirical CDF $F_{\mathbf{w}}$ of the projection:

$$(\nabla_{\mathbf{w}} T_{\mathbf{w}})(y) = \nabla_{\mathbf{w}} \left[\Phi^{-1}(F_{\mathbf{w}}(y))\right] = \frac{1}{\phi(\Phi^{-1}(F_{\mathbf{w}}(y)))} \nabla_{\mathbf{w}} F_{\mathbf{w}}(y)$$

The Mathematical Trap:

- The term $\nabla_{\mathbf{w}} F_{\mathbf{w}}(y)$ asks: *"How does the cumulative mass change as we rotate the projection plane?"*
- Mathematically, the derivative of a CDF boundary relies strictly on the Probability Density Function (PDF) evaluated exactly at that boundary.
- **The W-ICA Advantage Defeated:** The primary computational advantage of 1D Optimal Transport is that it uses sorting (empirical CDFs) to **completely avoid** continuous density estimation (PDFs).
- Requiring the PDF for the Hessian forces us to use Kernel Density Estimation (KDE), which introduces massive statistical noise and bandwidth-tuning fragility.

Conclusion: The Role of Quasi-Newton Methods

- Because the exact Hessian requires unstable density estimation, a true Newton-based fixed-point algorithm for exact W-ICA is mathematically intractable.
- **Our Solution:** This justifies our reliance on **Quasi-Newton methods** (like L-BFGS on the Stiefel manifold).
- L-BFGS intelligently approximates the inverse Hessian curvature strictly by observing the history of the stable, first-order Wasserstein gradients.
- This allows us to achieve super-linear convergence rates while strictly preserving the CDF-only, sorting-based elegance of the Optimal Transport formulation.

References I


Jean-Franois Cardoso.
Independent component analysis in the light of information geometry.
Entropy, 24(3):377, 2022.


Aapo Hyvärinen, Juha Karhunen, and Erkki Oja.
Independent Component Analysis.
John Wiley & Sons, Inc., New York, 2001.
Final version of 7 March 2001, Chapter 1.


Christian Jutten and Jean Herault.
Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture.
Signal Processing, 24(1):1–10, 1985.

References II

Leonid V. Kantorovich.
On the translocation of masses.
C. R. Doklady Acad. Sci. URSS (N.S.), 37:199–201, 1942.
Relaxed formulation via transport plans.

Gaspard Monge.
Mémoire sur la théorie des déblais et des remblais.
Histoire de l'Académie royale des sciences, 1781.

Cédric Villani.
Topics in Optimal Transportation, volume 58 of *Graduate Studies in Mathematics*.
American Mathematical Society, 2003.

References III


Cédric Villani.
Optimal Transport: Old and New, volume 338 of *Grundlehren der mathematischen Wissenschaften*.
Springer-Verlag, 2006.