1 Problem Statement and Algorithm Overview

1.1 Diffusion Correction in Spatial Transcriptomics

In spatial transcriptomics, mRNA molecules undergo diffusion during tissue processing and sequencing, resulting in observed expression patterns $\boldsymbol{X}_g \sim P(x_{i,g})$ for each gene g across N spots that deviate from the true undiffused states $\boldsymbol{Z}_g \sim P(z_{j,g})$. Our goal is to infer the undiffused expression field $\boldsymbol{Z}_g$ by modeling the diffusion process as an optimal transport problem for every gene $\boldsymbol{g}$.

1.2 Optimization Framework

We formulate this as a differentiable optimal transport problem with:

1.2.1 Parameters to Optimize

- **Cost Matrix Weights** $\boldsymbol{W} \in \mathbb{R}^{N \times N}$: Learns spatial relationships between spots, incorporating:

$$W_{ij} = f_\phi(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|, \text{tissue\_mask}) \tag{1}$$

- Initialization:

$$\mathbf{1}_{N \times N}$$

- **Gene-specific Thresholds** $\boldsymbol{q} \in [0,1]^G$: Adaptive quantiles to reflect diffusion levels. Essentially assuming spots with counts above quantile cutoffs as true sources.

- Initialization: For each gene, the cutoff value that maximizes Moran's I after filtering

- **Regularization Strengths** $\boldsymbol{r} \in \mathbb{R}_+^G$: Balances transport cost and entropy in the Sinkhorn algorithm.

- Initialization:

$$\mathbf{1}_G$$

1.2.2 Differentiable Components

The entire pipeline is end-to-end differentiable through:

- **Soft Thresholding**: Implements a differentiable quantile-based filter when applying gene-specific thresholds:

$$\boldsymbol{b}_g = \sigma(\beta \frac{\boldsymbol{X}_g - Q_g(q)}{\text{range}(\boldsymbol{X}_g)}) \odot \boldsymbol{X}_g \tag{2}$$

- **POT Integration**: Uses PyTorch-compatible optimal transport solvers (Sinkhorn, EMD) that preserve gradients

- **Spatial Statistics**: Moran's I and image alignment metrics are computed using differentiable Kornia operations

2 Problem Formulation

2.1 Objective Function

We minimize the composite loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{image}} + \lambda_2 \mathcal{L}_{\text{out-tissue}} + \lambda_3 \mathcal{L}_{\text{Moran}} \tag{3}$$

where $\lambda_1 = \lambda_2 = \lambda_3 = 1$ are weighting hyperparameters.

2.2 Component Losses

2.2.1 1. Image Alignment Loss

$$\mathcal{L}_{\text{image}} = \underbrace{\text{SSIM}(\boldsymbol{E}, \boldsymbol{I})}_{\text{Structural similarity}} + \underbrace{\|\nabla \boldsymbol{E} - \nabla \boldsymbol{I}\|_1}_{\text{Gradient matching}} + \underbrace{|\text{TV}(\boldsymbol{E}) - \text{TV}(\boldsymbol{I})|}_{\text{Total variation}} \tag{4}$$

where:

- $\boldsymbol{E} \in \mathbb{R}^{H \times W}$ is the gene expression grid
- $\boldsymbol{I} \in \mathbb{R}^{H \times W}$ is the H&E image
- SSIM is the Structural Similarity Index Measure
- TV is the Total Variation norm

2.2.2 2. Out-of-Tissue Expression Loss

$$\mathcal{L}_{\text{out-tissue}} = \sum_{i \notin \Omega} \boldsymbol{x}_i \tag{5}$$

where Ω is the set of in-tissue spots and $\boldsymbol{x}_i$ is the expression vector for spot i.

2.2.3 3. Spatial Autocorrelation Loss (Moran's I)

$$\mathcal{L}_{\text{Moran}} = -\frac{1}{G} \sum_{g=1}^{G} I_g \tag{6}$$

with Moran's I for gene g calculated as:

$$I_g = \frac{N}{S_0} \frac{(\boldsymbol{z}_g - \bar{z}_g)^\top \boldsymbol{W} (\boldsymbol{z}_g - \bar{z}_g)}{(\boldsymbol{z}_g - \bar{z}_g)^\top (\boldsymbol{z}_g - \bar{z}_g)} \tag{7}$$

where:

- $\boldsymbol{z}_g \in \mathbb{R}^N$ is the expression of gene g
- $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ is the spatial weight matrix
- $S_0 = \sum_{ij} W_{ij}$ is the normalization constant

2.3 Optimal Transport Framework

For each gene g, we solve the following with *ot.bregman.sinkhorn_stabilized*:

$$\begin{aligned} \gamma = \arg\min_{\gamma} \quad & \langle \gamma, \mathbf{M} \rangle_F + \text{reg} \cdot \Omega(\gamma) \\ & s.t.\ \gamma \mathbf{1} = \mathbf{a} \\ & \gamma^T \mathbf{1} = \mathbf{b} \\ & \gamma \geq 0 \end{aligned} \tag{8}$$

where:

- $\mathbf{M}$ is the (dim_a, dim_b) metric cost matrix
- Ω is the entropic regularization term $\Omega(\gamma) = \sum_{i,j} \gamma_{i,j} \log(\gamma_{i,j})$
- $\mathbf{a}$ and $\mathbf{b}$ are source and target weights (histograms, both sum to 1)

2.4 Soft Thresholding Operation

$$\boldsymbol{b}_g = \sigma \left(\beta \frac{\boldsymbol{x}_g - Q_g(q)}{\text{range}(\boldsymbol{x}_g)} \right) \odot \boldsymbol{x}_g \tag{9}$$

where:

- $Q_g(q)$ is the q-th quantile cutoff
- σ is the sigmoid function with $\beta = 50$
- $\odot$ denotes element-wise multiplication

3 Current Limitations and Proposed Improvements

3.1 Identified Challenges

The current implementation faces several technical and theoretical challenges:

- **High memory usage**:
  - The $N \times N$ cost matrix $\mathbf{C}_g$ becomes prohibitive for large datasets ($N > 10^4$ spots)
  - Current RAM usage: $\mathcal{O}(GN^2)$ where G is the number of genes
- **Gene interactions**:
  - Treats each gene independently (no cross-gene constraints)
  - Fails to capture biological correlations between genes
- **Quantile values receive gradients as 0**:
  - Even with soft-thresholding quantile values only receive 0 as gradients

3.2 Possible improvements

- **OT plan highly correlated**:
  - Computed OT plans are highly correlated for genes with similar expressions
  - May leverage this to speed up calculation
- **Incorporate spatial constraints or true physical constraints**:
  - Initialize cost weights (N*N) to reflect maximum diffusion distance
  - Need a good estimation of maximum diffusion distance / steps
- **Incorporation of reference if present**:
  - If user provides a reference spatial / single-cell dataset with less / no lateral diffusion, the information can be incorporated into loss terms
  - 1. Impute true sources based on reference (integrate and co-cluster)
    2. Encourage statistics of corrected counts to be close to reference