Low-Rank Optimal Transport through Factor Relaxation with Latent Coupling (*FRLC*)

Peter Halmos*, Xinhao Liu*, Julian Gold*, Benjamin J. Raphael

NEURAL INFORMATION PROCESSING SYSTEMS

PRINCETON UNIVERSITY

Background

- Optimal transport (OT) finds a least cost coupling $\boldsymbol{P}$ between probability distributions $\boldsymbol{a}$ and $\boldsymbol{b}$.
- Sinkhorn's Algorithm (Cuturi '13) greatly reduced the time-complexity of computing optimal couplings.
- Scaling OT to massive datasets is intractable as the coupling matrix scales quadratically.
- Low-rank OT methods (Forrow et al. '19, Scetbon et al. '21, Lin et al. '21) solve for factored couplings with linear space complexity.

Overview of FRLC

Factor Relaxation with Latent Coupling (FRLC, mnemonic "frolic") is an algorithm minimizing OT objectives subject to a low-rank constraint on coupling matrix $\boldsymbol{P}$. For example, the rank-r Wasserstein objective is:

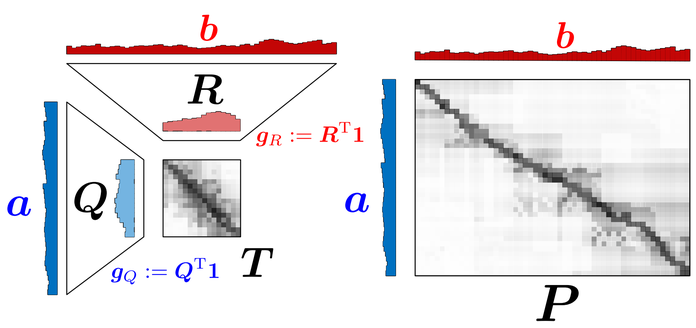
$$\mathrm{W}_r(\boldsymbol{a},\boldsymbol{b}) = \min_{\boldsymbol{P}\in\Pi_r(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{C},\boldsymbol{P}\rangle_F$$

We parameterize the set $\Pi_r(\boldsymbol{a},\boldsymbol{b})$ of rank-r couplings by components $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$ of an *LC (latent coupling) factorization,* decomposing $\boldsymbol{P}$ in terms of these components (depicted below) as:

$$\boldsymbol{P} = \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$$

FRLC minimizes the OT primal cost with a fast and scalable coordinate mirror-descent algorithm on the variables $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$.
Advantages over prior work:

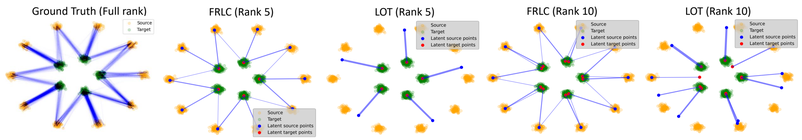
- FRLC is the only low-rank OT method solving low-rank OT with OT sub-problems, rather than Dykstra projections.
- FRLC coarse-grains optimal transport plans $\boldsymbol{P}$ and marginals $\boldsymbol{a},\boldsymbol{b}$ for general costs $\boldsymbol{C}$: $(\boldsymbol{P},\boldsymbol{a},\boldsymbol{b}) \to (\boldsymbol{T},\boldsymbol{g}_Q,\boldsymbol{g}_R)$
- It handles all OT problems (W, GW, FGW) and all relaxations (unbalanced, semi-relaxed, balanced)
- Has the best OT-cost value of any low-rank OT optimizer
- Has fewer optimization variables than other low-rank OT methods



A schematic of the LC-factorization used in FRLC, compared to a full-rank plan $\boldsymbol{P}$.

https://github.com/raphael-group/FRLC

Experimental Evaluation and Comparisons

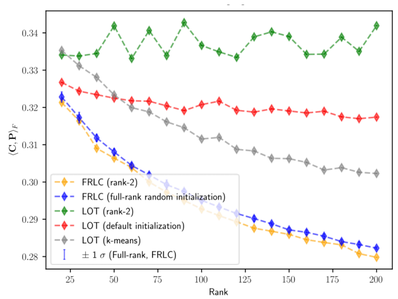


Comparison against (Scetbon et al. '21, LOT) on low-rank alignment of the n-th roots of unity — highlighting the value of a *non-square* latent coupling.

FRLC outperforms the previous state of the art low-rank optimal transport method, LOT (Scetbon et al. '21) on minimizing the primal OT cost:

Objective	Algorithm	Spearman ρ	ARI	AMI
W	LOT-U	0.394	0.332	0.397
	FRLC-U	**0.465**	**0.466**	**0.484**
	LOT-SR	0.391	0.319	0.396
	FRLC-SR	**0.467**	**0.475**	**0.492**
GW	LOT-U	0.005	0.0	0.0
	FRLC-U	**0.266**	**0.299**	**0.364**
	LOT-SR	0.004	0.0	0.0
	FRLC-SR	**0.275**	**0.318**	**0.381**
FGW	LOT-U	**0.391**	0.330	0.400
	FRLC-U	0.368	**0.391**	**0.420**
	LOT-SR	0.396	0.337	0.401
	FRLC-SR	**0.465**	**0.469**	**0.499**

Table comparing performance of gene expression prediction (Spearman) and cell type prediction tasks (ARI, AMI) on spatial transcriptomics data from (Chen et al. '22).



Plotting final OT cost as a function of rank on Gaussian synthetic data.

FRLC also outperforms other methods related to low-rank OT such as (Lin et al. '21).

Theoretical Results

We prove that the non-asymptotic stationary convergence guarantee of (Ghadimi et al. '14) for mirror-descent on smooth objectives can be extended to coordinate mirror descent (Prop 3.3). We then prove smoothness as a guarantee of the algorithm, from which non-asymptotic stationary convergence follows:

Proposition 3.4. *The FRLC algorithm with step-sizes $\gamma_k = \alpha/L$ and iterates $\boldsymbol{x}_k = (\boldsymbol{Q}_k,\boldsymbol{R}_k,\boldsymbol{T}_k)$ has non-asymptotic stationary convergence in the criterion $\Delta(\cdot,\cdot)$ with:*

$$\min_{k\in 1,..,N-1} \Delta_k(\boldsymbol{x}_k,\boldsymbol{x}_{k+1}) \le 2D^2L^2/N\alpha^2$$

With the convergence criterion:

$$\Delta_k(\boldsymbol{x}_k,\boldsymbol{x}_{k+1}) := \frac{1}{\gamma_k^2}\left[\|\boldsymbol{Q}_{k+1}-\boldsymbol{Q}_k\|_F^2 + \|\boldsymbol{R}_{k+1}-\boldsymbol{R}_k\|_F^2 + \|\boldsymbol{T}_{k+1}-\boldsymbol{T}_k\|_F^2\right]$$

In addition to these results, in our Appendix, we:

- derive the FRLC algorithm,
- we improve the bound on the rank-r approximation error of the optimal low-rank OT plan of (Scetbon et al. '22, Liu et al. '21) and show one can analytically solve for the block-optimal weights of the previous low-rank OT factorization of (Scetbon et al. '21).

FRLC algorithm

Algorithm 1 Balanced FRLC

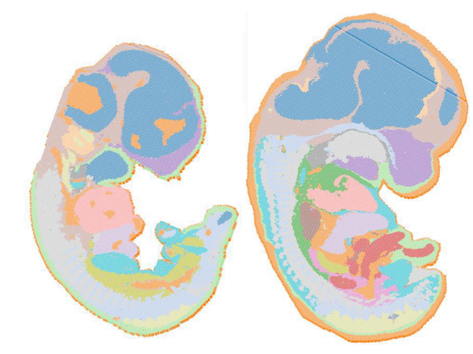
Input $\boldsymbol{C}, r, \boldsymbol{a}, \boldsymbol{b}, \tau, \gamma_1, \delta, \varepsilon$
Initialize $\boldsymbol{g}_Q, \boldsymbol{g}_R = \frac{1}{r}\mathbf{1}_r$
$\boldsymbol{Q}_0, \boldsymbol{R}_0, \boldsymbol{T}_0 \leftarrow$ *Initialize-Couplings*$(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{g}_Q, \boldsymbol{g}_R)$ # Alg. 6
$\boldsymbol{X}_0 \leftarrow \mathrm{diag}(1/\boldsymbol{Q}_0^{\mathrm{T}}\mathbf{1}_n)\boldsymbol{T}_0\,\mathrm{diag}(1/\boldsymbol{R}_0^{\mathrm{T}}\mathbf{1}_m)$
while $\Delta((\boldsymbol{Q}_k,\boldsymbol{R}_k,\boldsymbol{T}_k),(\boldsymbol{Q}_{k-1},\boldsymbol{R}_{k-1},\boldsymbol{T}_{k-1})) > \varepsilon$ **do** # Δ as in (10)
  $\nabla_{\boldsymbol{Q}} \leftarrow \boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}} - \mathbf{1}_n\mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}$
  $\nabla_{\boldsymbol{R}} \leftarrow \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k)^{\mathrm{T}}$
  $\gamma_k \leftarrow \gamma/\max\{\|\nabla_{\boldsymbol{Q}}\|_\infty, \|\nabla_{\boldsymbol{R}}\|_\infty\}$ # ℓ^∞-normalization of Scetbon & Cuturi (2022)
  $\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \boldsymbol{K}_{\boldsymbol{R}}^{(k)} \leftarrow \boldsymbol{Q}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{Q}}), \boldsymbol{R}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{R}})$
  $\boldsymbol{Q}_k \leftarrow$ SR^{R}-*projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Semi-relaxed OT, Alg. 2
  $\boldsymbol{R}_k \leftarrow$ SR^{R}-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Semi-relaxed OT
  $\boldsymbol{g}_Q, \boldsymbol{g}_R \leftarrow \boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m$
  $\nabla_{\boldsymbol{T}} = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}_k\,\mathrm{diag}(1/\boldsymbol{g}_R)$
  $\gamma_{\boldsymbol{T}} = \gamma/\|\nabla_{\boldsymbol{T}}\|_\infty$ # ℓ^∞-normalization
  $\boldsymbol{K}_{\boldsymbol{T}}^{(k)} \leftarrow \boldsymbol{T}_k \odot \exp(-\gamma_{\boldsymbol{T}}\nabla_{\boldsymbol{T}})$
  $\boldsymbol{T}_k \leftarrow$ *Sinkhorn*$(\boldsymbol{K}_{\boldsymbol{T}}^{(k)}, \boldsymbol{g}_R, \boldsymbol{g}_Q, \delta)$ # Balanced OT, Alg. 5
  $\boldsymbol{X}_k \leftarrow \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$
end while
Return $\boldsymbol{P}_r = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$

The Factor Relaxation with Latent Coupling Coordinate-MD Algorithm

Applications and Examples

FRLC, as a means of coarsening an OT plan, is readily interpretable and can perform both co-clustering and alignment of the co-clusters. One can also find the barycentric (LC) projection of the latent clusters.

Large-scale computation of growth rates (Halmos & Liu et al. '24) using semi-relaxed FRLC on mouse embryogenesis dataset (Chen et al. '22)



Co-clustering of massive-scale (100k+ points) mouse embryogenesis dataset (Chen et al. '22)

References

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022). "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays." In: *Cell*, **185**(10), pp. 1777–1792.

Cuturi, M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In: *Advances in Neural Information Processing Systems*, **26**, pp. 2292–2300. PMLR, 2013.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., & Weed, J. (2019). "Statistical Optimal Transport via Factored Couplings." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, **89** of Proceedings of Machine Learning Research, pp. 2454–2465. PMLR, 2019.

Ghadimi, S., Lan, G., & Zhang, H. (2014). "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization." In: *Mathematical Programming*, **155**(1–2), pp. 267–305.

Halmos, P., Liu, X., Gold, J., Chen, F., Ding, L., & Raphael, B. J. (2024). "DeST-OT: Alignment of Spatiotemporal Transcriptomics Data." In: *International Conference on Research in Computational Molecular Biology*, pp. 434–437.

Lin, C.-H., Azabou, M., & Dyer, E. L. (2021). "Making transport more robust and interpretable by moving data through a small number of anchor points." In: *Proceedings of Machine Learning Research*, **139**, pp. 6631–6641.

Liu, W., Zhang, C., Zheng, N., & Qian, H. (2021). "Approximating optimal transport via low-rank and sparse factorization." *arXiv preprint arXiv:2111.06546*.

Scetbon, M., Cuturi, M., & Peyré, G. (2021). "Low-Rank Sinkhorn Factorization." In: *Proceedings of the 38th International Conference on Machine Learning*, **139**, pp. 9344–9354. PMLR, 2021.

Scetbon, M., & Cuturi, M. (2022). "Low-rank Optimal Transport: Approximation, Statistics and Debiasing." In: *Advances in Neural Information Processing Systems*, **35**, pp. 6802–6814. PMLR, 2022.

Scetbon, M., Peyré, G., & Cuturi, M. (2022). "Linear-time Gromov Wasserstein Distances using Low Rank Couplings and Costs." In: *Proceedings of the 39th International Conference on Machine Learning*, **162**, pp. 19347–19365. PMLR, 2022.

Scetbon, M., Klein, M., Palla, G., & Cuturi, M. (2023). "Unbalanced Low-rank Optimal Transport Solvers." In: *Advances in Neural Information Processing Systems*, **36**, pp. 52312–52325. PMLR, 2023.

SCHMIDT FUTURES

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research