Variational Inference for Uncertain Optimal Transport via Sinkhorn Parametrization

Ananyapam De[1] Linus Bleistein[2] Anton Frederik Thielmann[3] Benjamin Säfken[1]

[1]Technische Universität Clausthal [2]EPFL [3]Amazon Music

Question

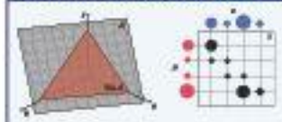
Classical OT maps one cost matrix to one deterministic coupling:

$$\Gamma^\star = \arg\min_{\Gamma \in \Pi(\mu,\nu)} \langle C, \Gamma \rangle, \qquad \Pi(\mu,\nu) = \{\Gamma \geq 0 : \Gamma \mathbf{1}_m = \mu, \Gamma^\top \mathbf{1}_n = \nu\}.$$

With noisy or conflicting costs $C^{(1)}, \ldots, C^{(N)}$, the goal is a posterior over feasible couplings, not just one MAP plan.

How can we learn $q(\Gamma \mid C^{(1:N)})$, a distribution that samples plausible future transport plans from multiple noisy cost observations?

Background: discrete optimal transport



Given source masses $\mu \in \Delta^{n-1}$, target masses $\nu \in \Delta^{m-1}$, and a cost matrix C, OT searches over couplings

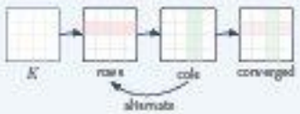
$$\Pi(\mu,\nu) = \{\Gamma \in \mathbb{R}_+^{n \times m} : \Gamma \mathbf{1}_m = \mu,\ \Gamma^\top \mathbf{1}_n = \nu\}.$$

Each entry Γ_{ij} is mass moved from source i to target j. The classical problem is

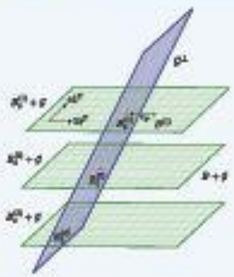
$$OT(\mu,\nu;C) = \min_{\Gamma \in \Pi(\mu,\nu)} \langle C, \Gamma \rangle.$$

The feasible set $\Pi(\mu,\nu)$ is a transport polytope of dimension $r = (n-1)(m-1)$. A deterministic OT solution can be brittle when C is uncertain; SPVI keeps the marginal constraints but learns a distribution over feasible plans:

$$p(\Gamma \mid C^{(1:N)}) \propto \exp\!\big[-R(\Gamma)\big] \sum_{k=1}^{N} \exp\!\big[-\beta \langle C^{(k)}, \Gamma \rangle\big] \mathbb{1}_{\Pi(\mu,\nu)}(\Gamma)$$



Gauge fixing



Sinkhorn maps a positive kernel to a feasible coupling:

$$S(K;\mu,\nu) = \operatorname{diag}(u) K \operatorname{diag}(v), \quad u \leftarrow \frac{\mu}{Kv}, \ v \leftarrow \frac{\nu}{K^\top u}$$

The Sinkhorn map is invariant to row/column shifts in log space:

$$B \sim B + a\mathbf{1}_m^\top + \mathbf{1}_n b^\top.$$

We remove this redundancy with the orthogonal slice

$$\mathcal{G}^\perp = \{B : B\mathbf{1}_m = 0,\ B^\top \mathbf{1}_n = 0\}.$$

- each gauge orbit has one representative in $\mathcal{G}^\perp$;
- the latent dimension becomes $r = (n-1)(m-1)$;
- with $Q \in \mathbb{R}^{nm \times r}$, the identifiable chart is

$$\Psi(z) = S(\exp(\operatorname{mat}(Qz)); \mu, \nu),$$

mapping $\mathbb{R}^r$ smoothly to $\mathcal{M} = \{\Gamma \in \Pi(\mu,\nu) : \Gamma_{ij} > 0\}$.

SPVI sampler

$\xi \sim \mathcal{N}(0, I_r)$ noise → f_ϕ flow → $x(\cdot)$ squash → Qz gauge logits → $S(\exp(\cdot))$ Sinkhorn → $\Gamma \in \Pi(\mu,\nu)$ plan

The squashing map $x(z) = c\tanh(z/c)$ stabilizes the latent logits, while Sinkhorn enforces

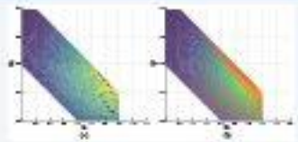
$$\Gamma(\xi)\mathbf{1}_m = \mu, \qquad \Gamma(\xi)^\top \mathbf{1}_n = \nu, \qquad \Gamma(\xi) > 0.$$

Manifold density and ELBO

For $\hat{\Gamma}(z) = S(\exp(\operatorname{mat}(Qz)); \mu, \nu)$, the area formula gives

$$\log q_\phi(\Gamma) = \log p(\xi) - \log\left|\det J_{g_\phi}(\xi)\right| - \frac{1}{2}\log\det G(z),$$

$$G(z) = J_{\hat{\Gamma}}(z)^\top J_{\hat{\Gamma}}(z).$$



HMC samples and SPVI ambient on a 2×3 polytope.

$$E(\Gamma) = R(\Gamma) - \log \sum_{k=1}^{N} \exp\!\big[-\beta \langle C^{(k)}, \Gamma \rangle\big]$$

$$\mathcal{L}(\phi) = \mathbb{E}_\xi\left[-E(\hat{\Gamma}(\xi)) + \log\left|\det J_{g_\phi}(\xi)\right| + \frac{1}{2}\log\det G(z)\right].$$

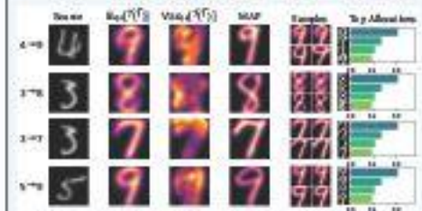
Estimate the log-volume term matrix-free using Hutchinson:

$$\log\det G(z) = \operatorname{Tr}(\log G(z)) \approx \frac{1}{M}\sum_{m=1}^{M} \omega_m^\top \log(G(z))\,\omega_m,$$

Approximated by Lanczos:

$$\omega^\top \log(G)\omega \approx \|\omega\|^2 e_1^\top \log(T_k) e_1.$$

Domain adaptation and multimodality



Setup: MNIST → USPS; 32×32, $n = m = 256$; 3-layer RealNVP + MLP

Mixed bargain: curve 32×32, $n = m = 256$; 3-layer RealNVP; setup

Barycentric projection: $T(i) = \mathbb{E}_\Gamma[Y|i]$

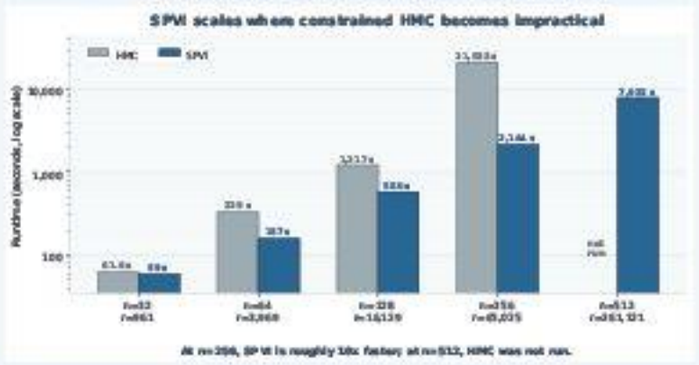
Pairs: $(4 \to 9), (3 \to 8), (3 \to 7), (5 \to 9)$

Marginals: $(0) \to (3,4), (6) \to (1,2,6), (2) \to (4,5)$

Histogram: posterior size over Barycentric sampling for pairs in $\mathcal{T}(\Gamma)$

MAP (mean $\mathbb{E}[y]$ single mode)

Scalability



At $n = 256$, SPVI is roughly 16× faster; at $n = 512$, HMC was not run.

Joint representation and transport learning

Office-31 DA: labeled source, unlabeled target, $K = 31$

$$C_{ij}^{(k)}(\psi) = \alpha D_{ij}^{(k)} + \eta(t) L_{ij}(\psi), \qquad k = 1, \ldots, 4$$

Frozen views: ResNet-50, DaT, DINOv2, OpenCLIP
View cost: median-scaled, ℓ_2-normalized feature distance
Label consistency:

$$L_{ij}(\psi) = \sum_{c=1}^{K} \ell_{ic}(Y_n^\star, g_{\psi,c}(x_j')), \qquad \ell_{ic}(y,s) = \max(0, 1 - ys)^2$$

Domain	Base	SinK	OT-MM	JDOT	SPVI
amazon → webcam	79.81	81.95	85.96	84.85	84.83
amazon → dslr	86.21	87.67	78.21	87.48	86.81
webcam → amazon	70.79	70.48	90.57	88.21	90.81
webcam → dslr	96.55	95.32	99.39	98.55	97.43
dslr → amazon	83.81	82.27	90.59	88.59	88.47
dslr → webcam	96.74	92.16	96.16	96.17	96.88
Mean	**87.15**	**86.27**	**90.48**	**90.04**	**90.72**

Office-31 accuracy (%); held-out; unlabeled source/domain pairs from SPVI posterior. Paired metric: MAP plan from SPVI posterior.

Block coordinate updates:

- ϕ-step: $\max_\phi \mathcal{L}(\phi; C(\psi^{(k)}))$
- ψ-step: posterior-predictive hinge risk, $\Gamma \sim q_\phi$

$$\mathcal{J}(\psi) = \mathbb{E}_{\Gamma \sim q_\phi} \sum_{ij} C_{\text{target}}(i,c,\psi,\Gamma) + \lambda_c C_{\text{consensus}}(\psi) + \lambda_{\text{reg}}\|\psi\|_2^2$$

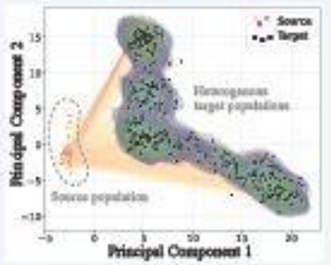
$$C_{\text{target}} = \hat{P}_{ic}(\Gamma) L_{ic}(+1, g_{\psi,c}(x_j')) + (1 - \hat{P}_{ic}(\Gamma)) L_{ic}(-1, g_{\psi,c}(x_j'))$$

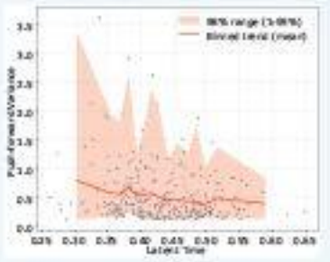
Soft target labels:

$$\hat{P}(\Gamma) = \Gamma^\top P^\star$$

$\eta(t) : 0 \to 1$ ramp; delayed label pressure

Single-cell fate uncertainty





SPVI gives uncertainty-aware couplings between developmental cell states: posterior fate uncertainty identifies ambiguous downstream assignments, while plasticity over latent time highlights transition regions and competing biological trajectories.

Theorem: Finite Sinkhorn guarantee

Bounded gauge logits $\kappa \in [-\kappa, \kappa]^r$

$$\|\Phi(\kappa) - \Phi(\kappa')\|_\infty \leq C_\Phi \varepsilon^\alpha, \qquad \|F(\kappa) - F(\kappa')\| \leq C_F \varepsilon^\alpha, \quad \alpha, \beta < 1$$

Implications: finite T valid for sampling Γ and ELBO log-density. Depth reflects scale since Sinkhorn step divides error geometrically.

References

1. Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. NeurIPS.
2. Mallasto, A., Heinonen, M., Kaski, S. (2021). Bayesian inference for optimal transport with stochastic cost. AISTATS.
3. Hallman, M. O., Bleistein, L., Wang, C., Paisley, J. (2022). Stochastic variational inference. JMLR.
4. Rangapuram, S., Gerstek, S., Ritz, O. (2024). Block low-rank variational inference. AISTATS.
5. Ruiverda, D., Mohamed, S. (2020). Variational inference with normalizing flows. ICML.
6. Kingma, D. P., Welling, M. (2014). Auto-encoding variational Bayes. ICLR.
7. Dinh, L., Sohl-Dickstein, J., Bengio, S. (2017). Density estimation using Real NVP. ICLR.

Takeaway

normalizing flow + gauge-fixed Sinkhorn map + manifold ELBO

SPVI turns Bayesian optimal transport from constrained MCMC into scalable differentiable inference while preserving feasibility at every sample.

ananyapam.de@tu-clausthal.de