Sinkhorn Networks: Using Optimal Transport Techniques to Learn Permutations.

Gonzalo E. Mena[1], David Belanger [2], Gonzalo Muñoz[3], Jasper Snoek [2].,
1. Department of Statistics, Columbia University, New York, NY, USA 2. Google Brain, Cambridge, MA. 3. Polytechnique Montréal, Montréal, Québec, Canada.,

In Brief

- Entropy regularization [3] has shown to be effective in *optimal transportation* (OT).
- We focus on permutations, a first-class citizen in the most elementary discrete OT.
- Using entropy regularization we conceive the choice of a matching (a transference plan in the discrete case) as a limit involving the **differentiable** *Sinkhorn operator*.
- We define Sinkhorn Networks: i)we parameterize a matching using neural networks, and ii)using Sinkhorn we enable automatic differentiation.
- We beat competitive baselines on problems involving permutations, and show an original application to reconstructions of objects from pieces.

Introduction

OT is relevant in machine learning as it provides richer description on the discrepancy between distributions than e.g., KL divergence. Entropy regularization of OT has also proven to be helpful [3]. More recently, this technique has enabled automatic differentiation (AD) for learning generative models based on OT [4]. Here we slightly deviate from current uses of entropy regularization, and focus on the most elementary discrete OT, where permutations arise naturally.

From the softmax to its permutation analog, the Sinkhorn operator

Softmax to approximate a category

With a temperature dependent softmax, one can approximate categories:

$$\text{softmax}_\tau(x)_i = \exp(x_i/\tau)/\sum_{j=1}\exp(x_j/\tau).$$

For $\tau > 0$, $\text{softmax}_\tau(x)_i$ is simplex-valued, and in the limit $\tau \to 0$, $\text{softmax}_\tau(x)_i$ converges to a point in the set $\mathcal{V}$ of vertices of the simplex $\mathcal{S}$, a one-hot vector. This approximation is key in [5, 7].

Sinkhorn operator is the analog in permutations

For a N dimensional square matrix $X > 0$ (elementwise) we define the Sinkhorn operator $S(\cdot)$ [1] as

$$\begin{aligned} S^0(X) &= \exp(X), \\ S^l(X) &= \mathcal{T}_c\left(\mathcal{T}_r(S^{l-1}(X))\right), \\ S(X) &= \lim_{l\to\infty} S^l(X). \end{aligned}$$

where $\mathcal{T}_r(X), \mathcal{T}_c(X)$ as the row and column-wise normalization operators of a matrix. Sinkhorn [9] proved $S(X)$ must belong to the Birkhoff polytope $\mathcal{B}_N$, the set of doubly stochastic matrices.

Parameterizing permutations, the matching operator $M(X)$

Choosing a category is a maximization problem parameterized by a vector x. The choice $\arg\max_i x_i$ given by

$$v^* = \arg\max_i x_i = \arg\max_{v\in\mathcal{S}}\langle x, v\rangle = \arg\max_{v\in\mathcal{S}}\langle x, v\rangle.$$

Likewise, one parameterizes the choice of a permutation P through a matrix X, as the solution to the linear assignment problem [6]

$$M(X) = \arg\max_{P\in\mathcal{P}_N}\langle P, X\rangle_F = \arg\max_{P\in\mathcal{B}_N}\langle P, X\rangle_F.$$

$\mathcal{P}_N$ the set of permutation matrices and $\langle A, B\rangle_F = \text{trace}(A^\top B)$. $M(\cdot)$ is the matching operator, through which we parameterize permutations (Figure 1a).

Theorem

For a doubly-stochastic matrix P, define its entropy as $h(P) = -\sum_{i,j} P_{i,j}\log\left(P_{i,j}\right)$.

$$S(X/\tau) = \arg\max_{P\in\mathcal{B}_N}\langle P, X\rangle_F + \tau h(P).$$

If $X_{i,j} \sim F_{i,j}$ are ind and $F_{i,j}$ is a.c. w.r. to the Lebesgue measure in $\mathbb{R}$, then, a.s.

$$M(X) = \lim_{\tau\to 0^+} S(X/\tau).$$

Matching and Sinkhorn operators illustrated

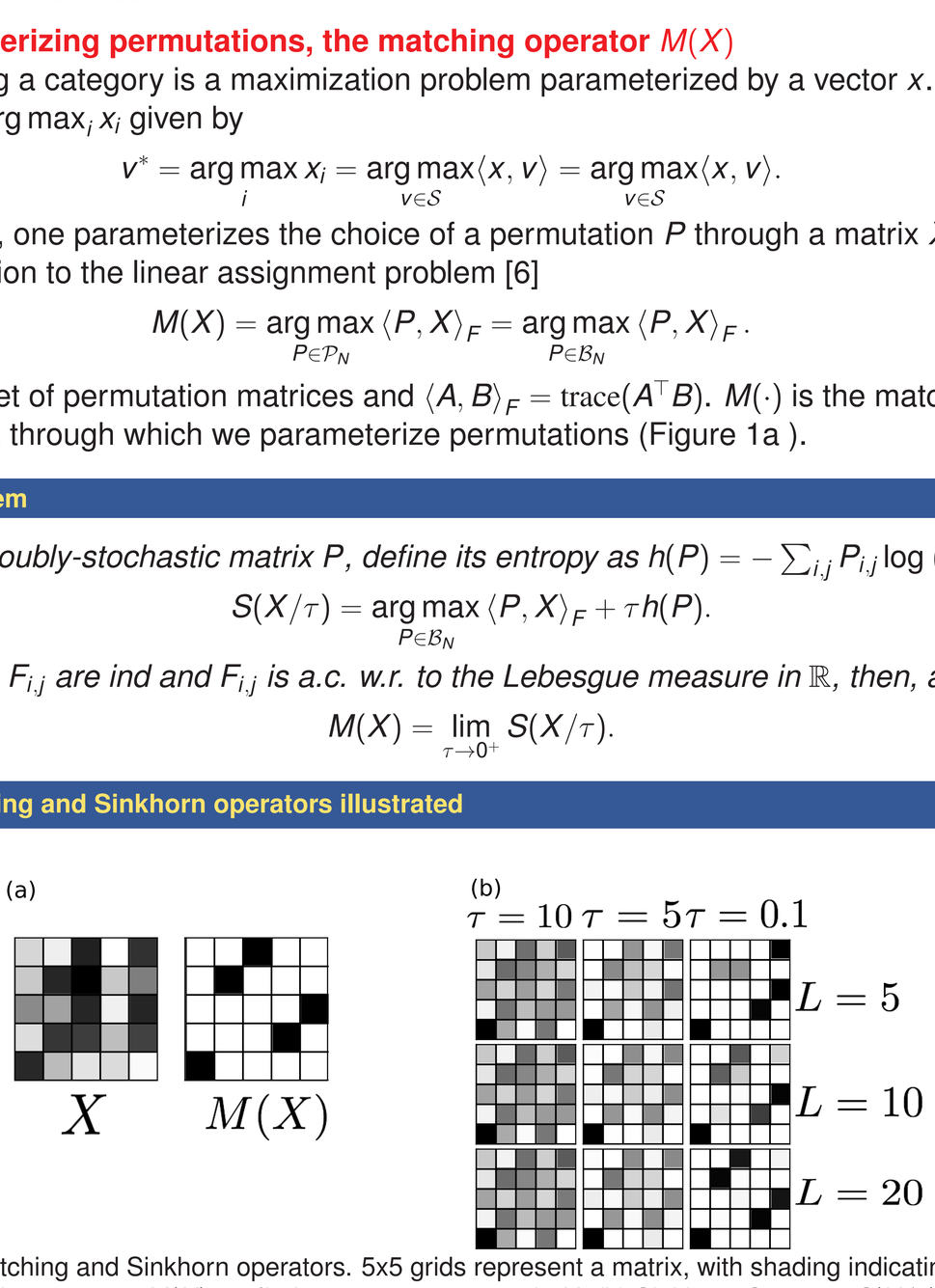

Figure: Matching and Sinkhorn operators. 5x5 grids represent a matrix, with shading indicating value (a) Matching operator $M(X)$ applied to a parameter matrix X. (b) Sinkhorn Operator $S(X/\tau)$ approximating $M(X)$ for different temperature τ and number of Sinkhorn iterations, L.

Optimal transportation perspective

Permutations naturally arise in discrete OT, as the transportation problem between the discrete measures $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ for a cost $c(\cdot,)$ is [10]:

$$\inf_{P\in\mathcal{P}_N}\frac{1}{n}\left\langle C^{X,Y}, P\right\rangle_F, \quad C^{X,Y}_{i,j} = c(x_i, y_i).$$

Our matrix X may be seen as an OT (negative) cost in the discrete transportation problem, i.e. we collapse c and x_i, y_i into a single $X_{i,j} = -c(x_i, y_j)$.

Sinkhorn Networks

Objective Consider the supervised task of learning a mapping from scrambled objects $\tilde{X}$ to non-scrambled X. Data are pairs $(X_i, \tilde{X}_i)$ with $\tilde{X}_i$ are random permutations of X_i. We minimize reconstruction error using the permutation $P_{\theta,\tilde{X}}$.

$$f(\theta, X, \tilde{X}) = \sum_{i=1}^M ||X_i - P^{-1}_{\theta,\tilde{X}_i}\tilde{X}_i||^2.$$

$P_{\theta,\tilde{X}}$ is parameterized as the solution of the assignment problem; $P_{\theta,\tilde{X}} = M(g(\tilde{X}, \theta))$, where $g(\tilde{X}, \theta)$ is a (possibly deep) transformation of the input. By Theorem 1 we replace $M(g(\tilde{X}, \theta))$ by the differentiable $S(g(\tilde{X}, \theta)/\tau)$.

Permutation equivariance

$$P_{\theta,P'(\tilde{X})}\left(P'\tilde{X}\right) = P'\left(P_{\theta,\tilde{X}}\tilde{X}\right)$$

Reconstructions of objects should not depend on how pieces were scrambled, but only on the pieces themselves.

Solving Jigsaw Puzzles

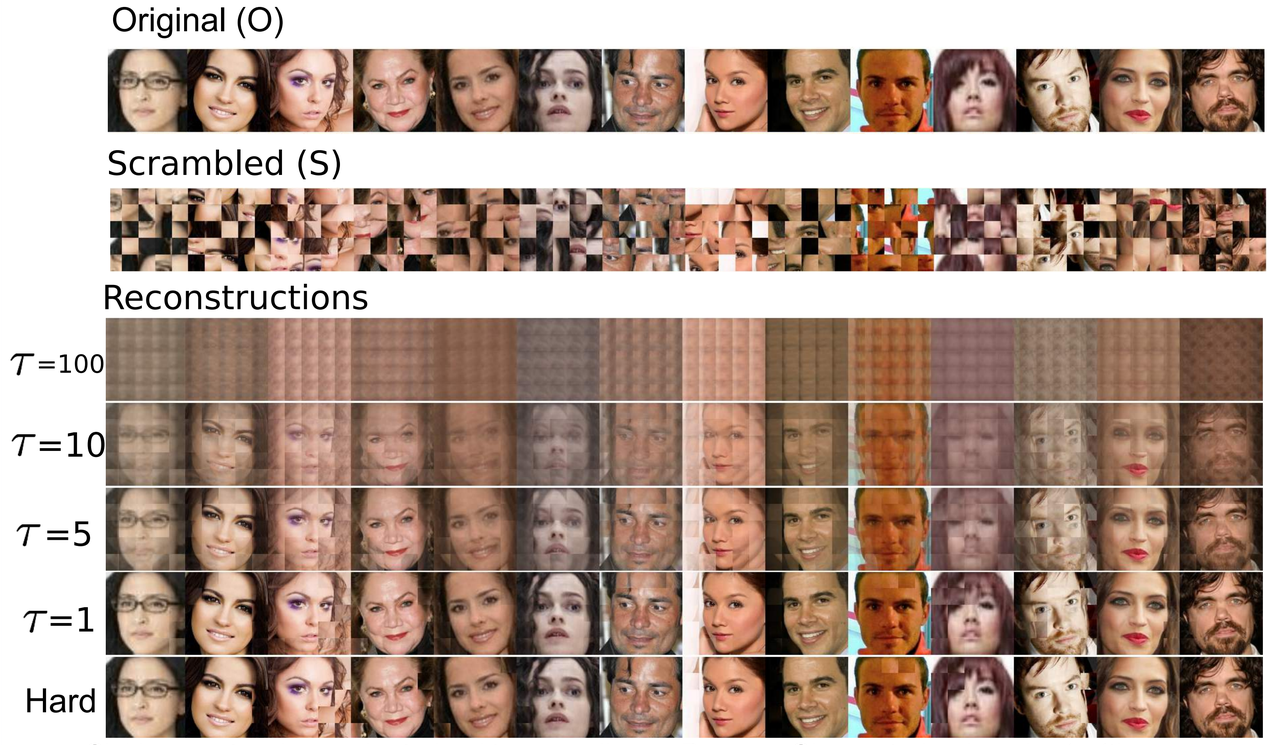

Figure: Sinkhorn networks can be trained to solve Jigsaw Puzzles. Given a trained model, 'soft' reconstructions are shown at different τ using $S(X/\tau)$. We also show hard reconstructions, made by computing $M(X)$ with the Hungarian algorithm [8]. We tie with [2] with a much simpler architecture.

Reconstruction of arbitrary MNIST digits from pieces

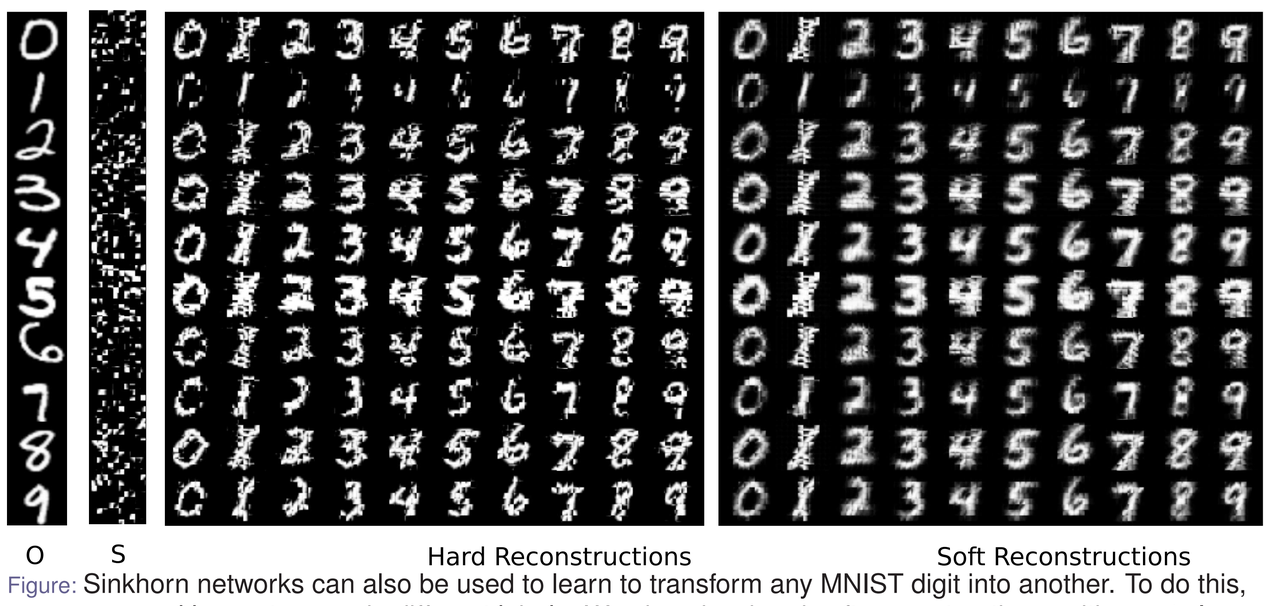

Figure: Sinkhorn networks can also be used to learn to transform any MNIST digit into another. To do this, we use several layers to encode different labels. We show hard and soft reconstructions, with $\tau = 1$. In 85% of cases a classifier is 'fooled'.

References

[1] R. P. Adams and R. S. Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.
[2] R. S. Cruz, B. Fernando, A. Cherian, and S. Gould. Deeppermnet: Visual permutation learning. *arXiv preprint arXiv:1704.02729*, 2017.
[3] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.
[4] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.
[5] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
[6] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.
[7] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
[8] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.
[9] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.
[10] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Email: gem2131@columbia.edu