Learning with a Wasserstein Loss

Charlie Frogner* Chiyuan Zhang* Hossein Mobahi Tomaso Poggio Mauricio Araya-Polo

Massachusetts Institute of Technology, Shell International E & P, Inc.

Presented by: Xi Fang
Date: 01/16/2019

Introduction

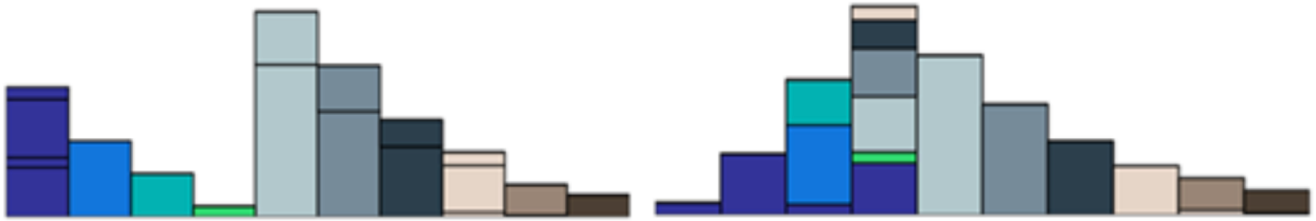
- Task: predict a non-negative measure over a finite set
- Motivation: Wasserstein loss encourages predictions that are similar to ground truth, the outputs have semantic relationships
- The similarity structure on the label space can be called the *ground metric* or *semantic similarity*

Siberian husky Eskimo dog

Figure 1: Semantically near-equivalent classes in ILSVRC

Introduction

- Wasserstein distance: the cost of the optimal transport plan for moving the mass in the predicted measure to match that in the target
  - The distance between two histogram



Distance: K*|i-j|

- Contribution:
  - first use of the Wasserstein distance as a loss for supervised learning

Problem setup

learning a map from $X \subset R^D$ into the space $Y = R^K_+$ of measures over a finite set K of size $|K| = K$

K possesses a metric $d_K(\cdot,\cdot)$, which is called the *ground metric*.

the predictor $h_\theta \in H$ that minimizes the expected risk E [E(h_θ(x),y)]

$$\min_{h_\theta \in \mathcal{H}} \left\{ \hat{\mathbb{E}}_S[\ell(h_\theta(x), y) = \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \right\}$$

Wasserstein loss

$$W_c(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{K} \times \mathcal{K}} c(\kappa_1, \kappa_2) \gamma(d\kappa_1, d\kappa_2)$$

$\Pi(\mu_1, \mu_2)$ is the set of joint probability measures on K×K having μ_1 and μ_2 as marginals

Exact Wasserstein Loss

For any $h_\theta \in H$, $h_\theta : X \rightarrow \Delta^K$, *let* $h_\theta(\varkappa|x) = h_\theta(x)_\varkappa$ *be the predicted value at element* $\varkappa \in K$, *given input* $x \in X$

$y(\varkappa)$ *be the ground truth value for* $\varkappa$ *given by the corresponding label* y

$$W_p^p(h(\cdot|x), y(\cdot)) = \inf_{T \in \Pi(h(x), y)} \langle T, M \rangle$$

$M \in R^{K \times K}$ *is the distance matrix*

the set of valid transport plans is

$$\Pi(h(x),y)=\{T \in R^{K\times K} : T1=h(x),\ T^T1=y\}$$

Algorithm 1 Gradient of the Wasserstein loss

Given $h(x)$, y, λ, $\mathbf{K}$. (γ_a, γ_b if $h(x)$, y unnormalized.)

$u \leftarrow \mathbf{1}$

while u has not converged **do**

$$u \leftarrow \begin{cases} h(x) \oslash \left(\mathbf{K}\left(y \oslash \mathbf{K}^\top u\right)\right) & \text{if } h(x), y \text{ normalized} \\ h(x)^{\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}} \oslash \left(\mathbf{K}\left(y \oslash \mathbf{K}^\top u\right)^{\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}}\right)^{\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}} & \text{if } h(x), y \text{ unnormalized} \end{cases}$$

end while

If $h(x)$, y unnormalized: $v \leftarrow y^{\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}} \oslash \left(\mathbf{K}^\top u\right)^{\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}}$

$$\partial W_p^p / \partial h(x) \leftarrow \begin{cases} \frac{\log u}{\lambda} - \frac{\log u^\top \mathbf{1}}{\lambda K}\mathbf{1} & \text{if } h(x), y \text{ normalized} \\ \gamma_a \left(\mathbf{1} - (\mathsf{diag}(u)\mathbf{K}v) \oslash h(x)\right) & \text{if } h(x), y \text{ unnormalized} \end{cases}$$

Efficient optimization via entropic regularization

- Proposition

Proposition 4.1. *The transport matrix T^* optimizing (8) satisfies $T^* = \text{diag}(u)\mathbf{K}\text{diag}(v)$, where $u = (h(x) \oslash T^*\mathbf{1})^{\gamma_a \lambda}$, $v = \left(y \oslash (T^*)^\top \mathbf{1}\right)^{\gamma_b \lambda}$, and $\mathbf{K} = e^{-\lambda M - 1}$.*

And the optimal transport matrix is a fixed point for a Sinkhorn-like iteration. [2]

Proposition 4.2. *$T^* = \text{diag}(u)\mathbf{K}\text{diag}(v)$ optimizing (8) satisfies: i) $u = h(x)^{\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}} \odot (\mathbf{K}v)^{-\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}}$, and ii) $v = y^{\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}} \odot \left(\mathbf{K}^\top u\right)^{-\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}}$, where $\odot$ represents element-wise multiplication.*

Efficient optimization via entropic regularization

Cuturi [18] proposes a smoothed transport objective that enables efficient approximation of both the transport matrix in (3) and the subgradient of the loss. [18] introduces an entropic regularization term that results in a strictly convex problem:

$$ {}^{\lambda}W_p^p(h(\cdot|x), y(\cdot)) = \inf_{T \in \Pi(h(x),y)} \langle T, M \rangle - \frac{1}{\lambda} H(T), \quad H(T) = -\sum_{\kappa,\kappa'} T_{\kappa,\kappa'} \log T_{\kappa,\kappa'}. \tag{6} $$

Importantly, the transport matrix that solves (6) is a *diagonal scaling* of a matrix $\mathbf{K} = e^{-\lambda M - 1}$:

$$ T^* = \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v) \tag{7} $$

for $u = e^{\lambda\alpha}$ and $v = e^{\lambda\beta}$, where α and β are the Lagrange dual variables for (6).

Identifying such a matrix subject to equality constraints on the row and column sums is exactly a *matrix balancing* problem, which is well-studied in numerical linear algebra and for which efficient iterative algorithms exist [19]. [18] and [3] use the well-known Sinkhorn-Knopp algorithm.

Statistical Properties of the Wasserstein loss

Let $S = ((x_1, y_1), \ldots, (x_N, y_N))$ be i.i.d. samples and $h_{\hat{\theta}}$ be the empirical risk minimizer

$$h_{\hat{\theta}} = \underset{h_\theta \in \mathcal{H}}{\operatorname{argmin}} \left\{ \hat{\mathbb{E}}_S \left[W_p^p(h_\theta(\cdot|x), y) \right] = \frac{1}{N} \sum_{i=1}^{N} W_p^p(h_x\theta(\cdot|x_i), y_i) \right\}.$$

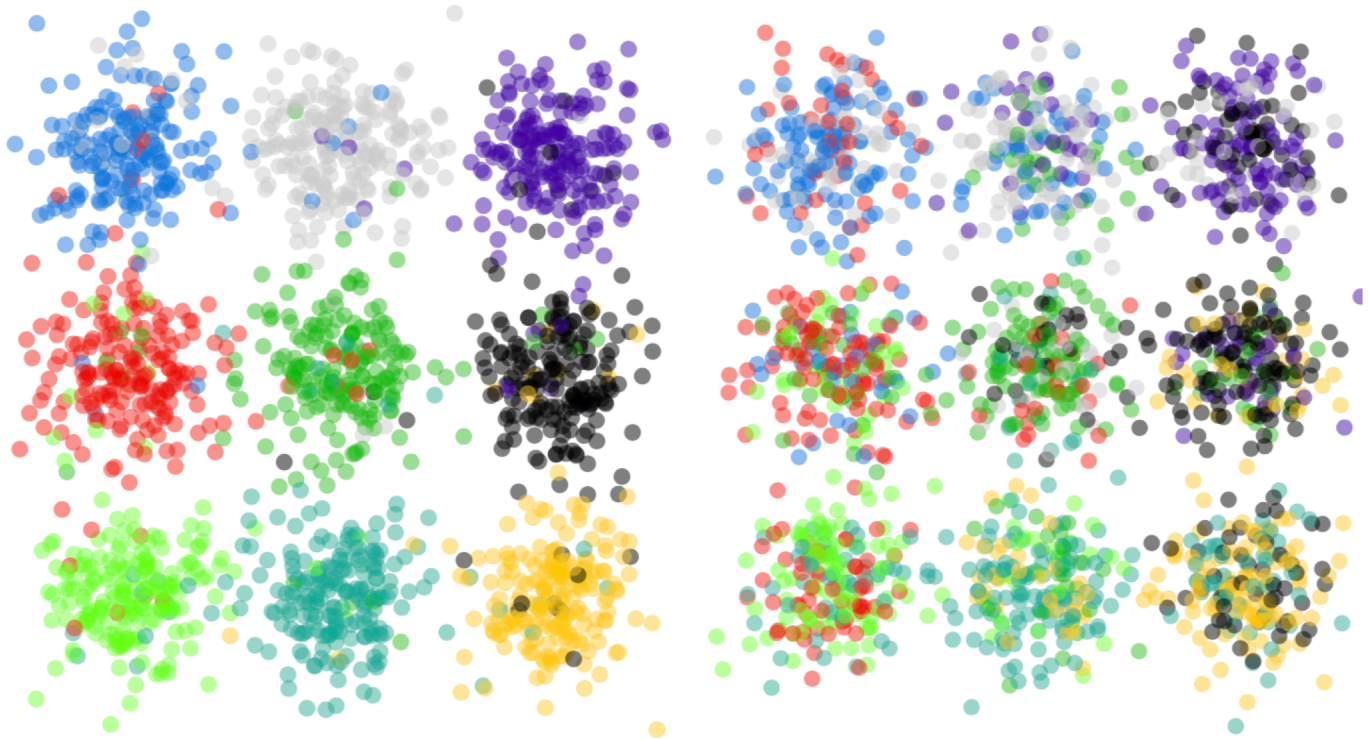
Theorem 5.1. *For $p = 1$, and any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\mathbb{E}\left[W_1^1(h_{\hat{\theta}}(\cdot|x), y)\right] \leq \inf_{h_\theta \in \mathcal{H}} \mathbb{E}\left[W_1^1(h_\theta(\cdot|x), y)\right] + 32KC_M\mathfrak{R}_N(\mathcal{H}^o) + 2C_M\sqrt{\frac{\log(1/\delta)}{2N}} \quad (9)$$

with the constant $C_M = \max_{\kappa,\kappa'} M_{\kappa,\kappa'}$. $\mathfrak{R}_N(\mathcal{H}^o)$ is the Rademacher complexity *[22] measuring the complexity of the hypothesis space $\mathcal{H}^o$.*

Experiment

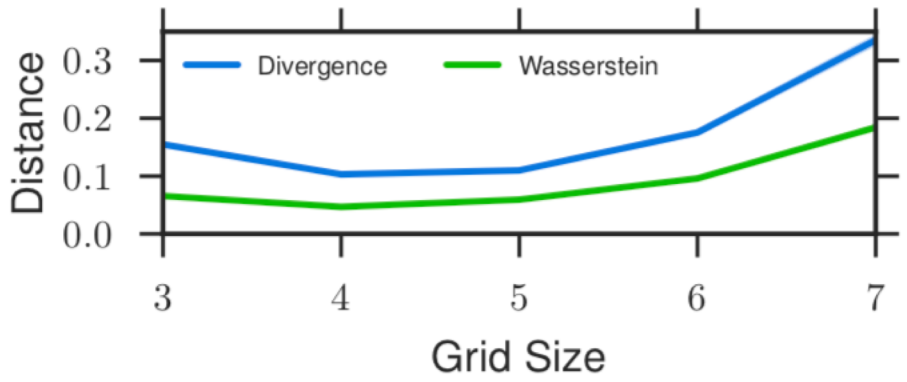
- Noise Dataset



(a) Noise level 0.1 (b) Noise level 0.5

Figure 8: Illustration of training samples on a 3x3 lattice with different noise levels.

Experiment



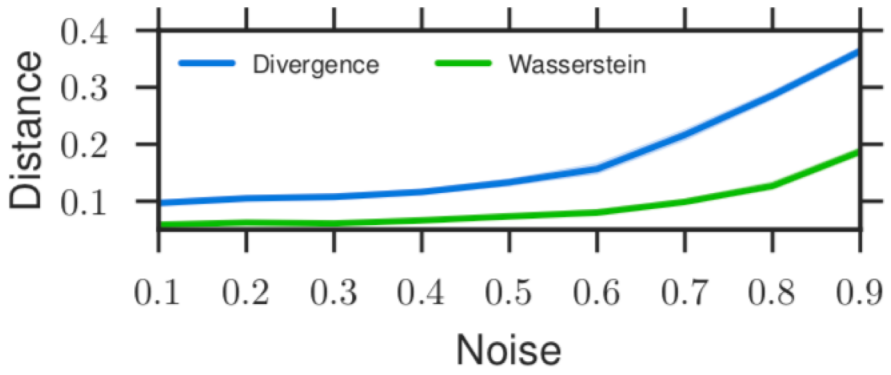
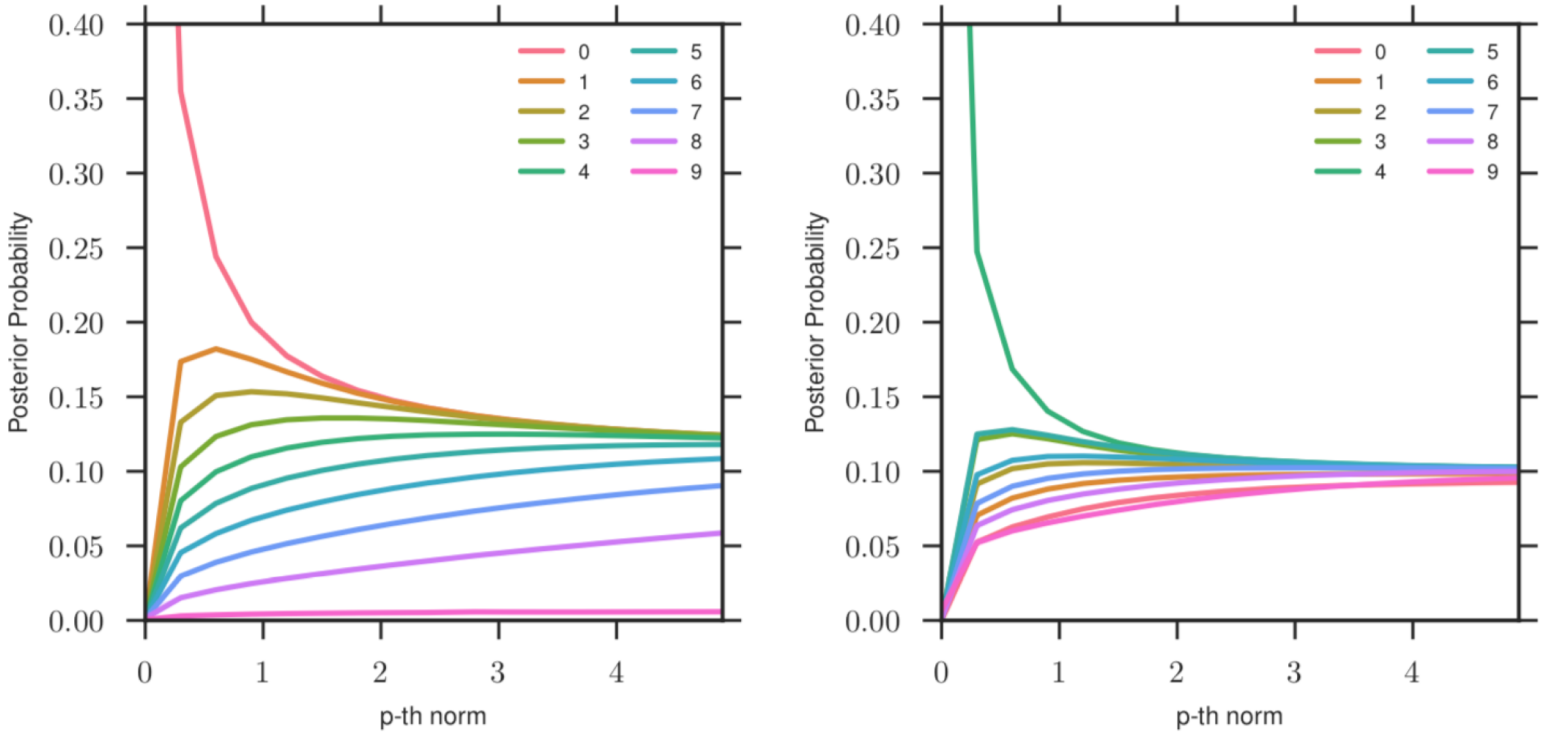


Figure 2: The Wasserstein loss encourages predictions that are similar to ground truth, robustly to incorrect labeling of similar classes (see Appendix E.1). Shown is Euclidean distance between prediction and ground truth vs. (left) number of classes, averaged over different noise levels and (right) noise level, averaged over number of classes. Baseline is the multiclass logistic loss.

Experiment



(a) Posterior prediction for images of digit 0.

(b) Posterior prediction for images of digit 4.

Figure 9: Each curve is the predicted probability for a target digit from models trained with different p values for the ground metric.

Experiment

- Derive a distance metric between tags by using word2vec [24] to embed the tags as unit vectors, then taking their Euclidean distances
- Extract image features using MatConvNet
- measure the prediction performance by the *top-K cost*, defined as CK = 1/K sum(Kk=1) min(j) dK(κˆk,κj), where {κj} is the set of ground truth tags

Experiment

Examples of images in the Flickr dataset

(a) **Flickr user tags**: street, parade, dragon; **our proposals**: people, protest, parade; **baseline proposals**: music, car, band.

(b) **Flickr user tags**: water, boat, reflection, sunshine; **our proposals**: water, river, lake, summer; **baseline proposals**: river, water, club, nature.

Conclusion

- A loss function for learning to predict a non-negative measure over a finite set, based on the Wasserstein distance
- Optimizing with respect to the exact Wasserstein loss is computationally costly, an approximation based on entropic regularization is efficiently computed
- A learning algorithm based on this regularization and they proposed a novel extension of the regularized loss to unnormalized measures that preserves its efficiency
- Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space

Reference

- A Wasserstein Distance https://www.nowcoder.com/acm/contest/91/A

- SINKHORN DISTANCES: LIGHTSPEED COMPUTATION OF OPTIMAL TRANSPORTATION DISTANCES, MARCO CUTURI, https://arxiv.org/pdf/1306.0895.pdf

- Learning with a Wasserstein Loss, https://arxiv.org/pdf/1506.05439.pdf