A Closed-form Gradient for the 1D Earth Mover's Distance for Spectral Deep Learning on Biological Data

Manuel Martinez, Makarand Tapaswi, and Rainer Stiefelhagen
Karlsruhe Institute of Technology, Karlsruhe, Germany
Email: firstname.lastname@kit.edu

1 Introduction

Spectral analysis is performed in several domains of physiological monitoring (*e.g.* respiratory analysis [1], EEG [2], ECG [3]). Regression models in the spectral domain enable several applications, often through the use of Power Spectral Density (PSD). Within machine learning frameworks, PSD is commonly treated as a probability distribution and learned using the Küllback-Leibler (KL) divergence. However, KL compares each bin independently.

The Earth Mover's Distance (EMD) is a natural metric to compare distributions, but has seen limited use due to its computational cost. Nevertheless, for one dimensional distributions (*e.g.* PSD) the EMD can be computed efficiently, and we derive a closed-form solution for its gradient. We enforce the gradient to preserve the ℓ_1 norm of the original distribution. We evaluate on a data set of 81 sleep laboratory patients, predict breathing rate, and compare EMD as a loss against KL divergence and Mean Squared Error.

2 The Earth Mover's Distance

The Earth Mover's Distance, also known as the Wasserstein distance, is a family of metrics used to compare distributions based on the optimal transport problem. The name EMD is derived from the effort required to move dirt to make the distributions equal. In the typical case, the two distributions are non-negative and are of the same size (same total amount of dirt).

The EMD metric has seen significant use to compare histograms [4, 5, 6, 7] or even entire images [8]. Recently there have been efforts to integrate EMD as a loss criterion for deep learning [9, 10]. However, as compared to other

criteria such as Mean Squared Error (MSE) or KL divergence, the perceived inefficiency in EMD computation has hindered progress.

In the general case, calculating the EMD requires to solve the optimal transport problem that turns the source distribution to the target one. As this calculation is expensive, much effort has been invested in relaxed definitions of the EMD that allow more efficient computation [11, 12, 13, 14, 15].

Recently, Frogner *et al.* [10] suggest a method to incorporate EMD in a deep learning framework using the entropic regularization proposed by Cuturi [14]. We consider that this approach to compute EMD is unnecessarily complex for the common case of one dimensional distributions for which there exists a closed-form solution.

For two discrete distributions, represented as non-negative vectors $\mathbf{a}$ and $\mathbf{b}$ of length N, the one-dimensional EMD is defined as:

$$EMD(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{N} |\varphi_i|, \text{ where } \varphi_i = \sum_{j=1}^{i} \left(\frac{a_j}{\|\mathbf{a}\|_1} - \frac{b_j}{\|\mathbf{b}\|_1} \right) . \tag{1}$$

This is rewritten using the sign function as $EMD(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{N} \operatorname{sgn}(\varphi_i) \cdot \varphi_i$.

2.1 Gradient of Earth Mover's Distance

To integrate EMD with deep learning, we now compute the analytical form of its gradient.

Let $\mathbf{e}_k$ be a unit vector of length N whose value at dimension k is 1, and 0 elsewhere. For a small perturbation h, we compute the distance between the modified distribution $\mathbf{a} + h\mathbf{e}_k$ and $\mathbf{b}$ as:

$$EMD(\mathbf{a} + h\mathbf{e}_k, \mathbf{b}) \simeq \sum_{i=1}^{N} \operatorname{sgn}(\varphi_i) \sum_{j=1}^{i} \left(\frac{a_j + h\delta_{jk}}{\|\mathbf{a}\|_1 + h} - \frac{b_j}{\|\mathbf{b}\|_1} \right) , \tag{2}$$

where $\delta_{jk} = 1$ when $j = k$. Note that by choosing h small enough, $\operatorname{sgn}(\varphi_i)$ is unchanged.

We now compute the partial derivative for EMD as follows. As we define EMD on normalized vectors (Eq. 1), without the loss of generality, we assume $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1 = 1$, and operate on unit ℓ_1 norm vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$.

$$
\begin{aligned}
gEMD &= \frac{\partial EMD(\hat{\mathbf{a}}, \hat{\mathbf{b}})}{\partial a_k} = \lim_{h \to 0} \frac{1}{h} \left(EMD(\hat{\mathbf{a}} + h\mathbf{e}_k, \hat{\mathbf{b}}) - EMD(\hat{\mathbf{a}}, \hat{\mathbf{b}}) \right) \quad (3) \\
&\simeq \lim_{h \to 0} \frac{1}{h} \sum_{i=1}^{N} \text{sgn}(\varphi_i) \left(\sum_{j=1}^{i} \left(\frac{\hat{a}_j + h\delta_{jk}}{1+h} - \hat{b}_j \right) - \sum_{j=1}^{i} \left(\hat{a}_j - \hat{b}_j \right) \right) \\
&= \lim_{h \to 0} \frac{1}{h} \sum_{i=1}^{N} \text{sgn}(\varphi_i) \sum_{j=1}^{i} \frac{h\delta_{jk} - h\hat{a}_j}{1+h} \\
&= \sum_{i=1}^{N} \text{sgn}(\varphi_i) \sum_{j=1}^{i} (\delta_{jk} - \hat{a}_j) \; . \quad (4)
\end{aligned}
$$

2.2 ℓ_1 preserving gradient

The above gradient disobeys the law of dirt conservation and creates new or destroys existing dirt, rendering it unsuitable for use (*i.e.* the gradient sum is not 0). To solve this problem, and in contrast to $\mathbf{e}_k$, we redefine our unit vector such that its sum is 0. We propose a set of vectors $\tilde{\mathbf{e}} \in \{-1, N-1\}^N$, where $\tilde{\mathbf{e}}_k$ takes the value $N-1$ at dimension k, and -1 elsewhere.

The partial derivatives for such a setting are

$$
\begin{aligned}
gEMD_{L1} &= \frac{\partial EMD(\hat{\mathbf{a}}, \hat{\mathbf{b}})}{\partial a_k} = \lim_{h \to 0} \frac{1}{h} \left(EMD(\hat{\mathbf{a}} + h\tilde{\mathbf{e}}_k, \hat{\mathbf{b}}) - EMD(\hat{\mathbf{a}}, \hat{\mathbf{b}}) \right) \quad (5) \\
&\simeq \lim_{h \to 0} \frac{1}{h} \sum_{i=1}^{N} \text{sgn}(\varphi_i) \left(\sum_{j=1}^{i} \left(\hat{a}_j + h(N\delta_{jk} - 1) - \hat{b}_j \right) - \sum_{j=1}^{i} \left(\hat{a}_j - \hat{b}_j \right) \right) \\
&= \lim_{h \to 0} \frac{1}{h} \sum_{i=1}^{N} \text{sgn}(\varphi_i) \sum_{j=1}^{i} h \left(N\delta_{jk} - 1 \right) \\
&= \sum_{i=1}^{N} \text{sgn}(\varphi_i) \sum_{j=1}^{i} (N\delta_{jk} - 1) \, . \quad (6)
\end{aligned}
$$

2.3 Implementation details

Efficiency. We implement our regression model with the EMD criterion using Torch [16]. Note that φ_i can be computed very fast through the use of

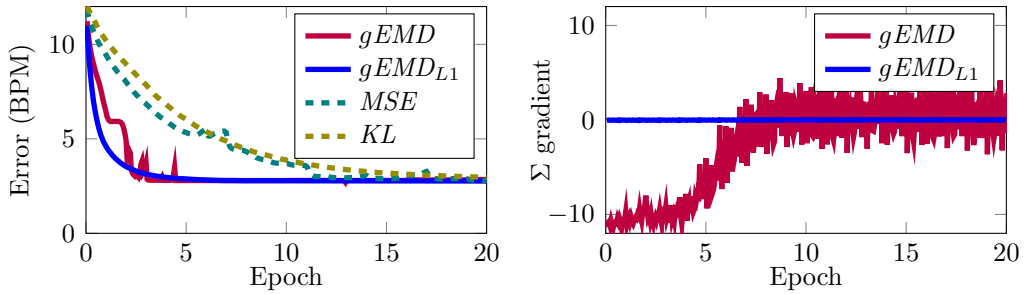

Figure 1: We train a regressor to estimate the PSD of the breathing signal. The validation error is the distance between the largest PSD peak (predicted vs. reference) in Breaths Per Minute (BPM). Left: the $gEMD_{\text{L1}}$ (ℓ_1 preserving variant of EMD) converges faster than all other losses: *gEMD*, MSE and KL divergence. Right: the sum of gradient $gEMD_{\text{L1}}$ is 0, while it fluctuates for *gEMD*.

a cumulative sum operator (`cumsum`). On a GPU, both forward and backward passes of EMD are faster than KL divergence (tested for $N = 30$).

Absolute magnitude. The original EMD is not defined for $\|\mathbf{a}\|_1 \neq \|\mathbf{b}\|_1$. Although there are several ways to modify the EMD for unnormalized distributions [13, 17] we consider that this goes against the spirit of the metric. Therefore, prior to computing the EMD, we ℓ_1-normalize the input distributions (see Eq. (1)). Such a normalization can lead to a mismatch between the absolute magnitudes, which may result in numerical instability during optimization. We address this by defining our loss as the weighted combination of EMD and MSE criterion ($w_{\text{EMD}} = 0.9, w_{\text{MSE}} = 0.1$).

Non-negative distributions. We square the output of the last layer of our model to ensure non-negative values for the distribution.

3 Experimental results

We train a two layer Multilayer Perceptron regressor to predict the breathing signal PSD from the chest excursion of 81 sleep laboratory patients with different degrees of apnoea, and use the PSD from a nose thermistor as reference. As we see in Fig. 1, the PSD obtained using $gEMD_{\text{L1}}$ loss converges faster and provides a better estimation of the breathing rate (the largest peak of the PSD). We also see that the gradient sums to 0 for $gEMD_{\text{L1}}$.

References

[1] M. Martinez and R. Stiefelhagen, "Breath Rate Monitoring During Sleep using Near-IR Imagery and PCA," in *ICPR*, 2012.

[2] A. Sano and R. W. Picard, "Comparison of sleep-wake classification using electroencephalogram and wrist-worn multi-modal sensor data," in *EMBC*, 2014.

[3] G. D. Clifford, F. Azuaje, and P. McSharry, *Advanced methods and tools for ECG data analysis*. Artech House, Inc., 2006.

[4] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[5] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *PAMI*, vol. 29, no. 5, pp. 840–853, 2007.

[6] J. Rabin, J. Delon, and Y. Gou, "Circular Earth Mover's Distance for the comparison of local features," in *ICPR*, 2008.

[7] S. Marinai, B. Miotti, and G. Soda, "Using earth mover's distance in the bag-of-visual-words model for mathematical symbol retrieval," in *ICDAR*, 2011.

[8] S. Peleg, M. Werman, and H. Rom, "A unified approach to the change of resolution: Space and gray-level," *PAMI*, vol. 11, no. 7, pp. 739–742, 1989.

[9] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher, "Wasserstein propagation for semi-supervised learning," in *ICML*, 2014.

[10] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a Wasserstein Loss," in *NIPS*, 2015.

[11] M. Cuturi, G. Peyré, and A. Rolet, "A smoothed dual approach for variational wasserstein problems," *SIAM JIS*, vol. 9, no. 1, pp. 320–343, 2016.

[12] S. Shirdhonkar and D. W. Jacobs, "Approximate earth mover's distance in linear time," in *ICPR*, 2008.

[13] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *ICCV*, 2009.

[14] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NIPS*, 2013.

[15] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," in *ICML*, 2014.

[16] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[17] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "Unbalanced optimal transport: geometry and Kantorovich formulation," *arXiv preprint arXiv:1508.05216*, 2015.