A Study of Optimal Transport in Deep Learning

Divyansh Khanna
Department of Computer Science
New York University
New York City, NY 100012
dk3399@nyu.edu

Hans Krupakar
Department of Computer Science
New York University
New York City, NY 100012
hansk@nyu.edu

Sachin Shastri
Department of Computer Science
New York University
New York City, NY 100012
sds662@nyu.edu

1 Abstract

In this report we study the theory of Optimal Transport (OT) and its application to deep learning. OT is the study of allocation of resources between two distributions such that the overall cost is minimized. This finds use in deep learning as a way to model a distribution closer to the data distribution. Though the research on OT can be dated to early 1980s, we focus in particular on the recent wave of efficient algorithms that have helped OT find relevance in deep learning ([1], [2], [3]).

OT is most commonly used as a way to compute a distance metric between distributions. Recently, Earth-Mover distance (EM) or Wasserstein-1 has brought the interest of the research community to OT based distance metrics for efficiently training generative models [1]. As a natural extension of the idea, OT has also been used for matching distributions. Optimal transport provides the tools to transform one distribution into another. This is particularly helpful in domain adaptation. A regularized optimal transportation model can be used to perform the alignment of the representations in the source and target domains [4] [5].

We touch upon gradient flows for the Wasserstein metric on the space of probability measures. This has been relevant in varied problems, like studying the global convergence of gradient descent [6] and reinforcement learning [7]. OT has been applied in NLP recently, with Word-Mover distance (WMD) [8] providing a distance metric for measuring the similarity between word documents. The report also covers the latest research on applications of OT based gradient flow techniques for natural language understanding [2].

Despite suffering from biased gradients ([9], [10]), OT provides a unique geometrical perspective along with optimum mass displacement which is usually lacking in other distance measures like Euclidean and Kullback–Leibler. Along with its applications, we plan to explore how OT based measures compare with other measures. We conclude by studying further research directions in the field and possible extensions to existing methods to further advance the application of OT in deep learning.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning

via optimal transport. *CoRR*, abs/1901.06283, 2019.

[3] Tim Salimans, Han Zhang, Alec Radford, and Dimitris N. Metaxas. Improving gans using optimal transport. *CoRR*, abs/1803.05573, 2018.

[4] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *CoRR*, abs/1507.00504, 2015.

[5] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3730–3739. Curran Associates, Inc., 2017.

[6] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3040–3050, 2018.

[7] Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as wasserstein gradient flows. *CoRR*, abs/1808.03030, 2018.

[8] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015.

[9] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *CoRR*, abs/1705.10743, 2017.

[10] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *CoRR*, abs/1707.06887, 2017.