CONNECTING CONVEX ENERGY-BASED INFERENCE AND OPTIMAL TRANSPORT FOR DOMAIN ADAPTATION

Arip Asadulaev
ITMO University
aripasadulaev@itmo.ru

ABSTRACT

The connection of optimal transport and neural networks finds a rich application in machine learning problems. In this paper, we propose a simple algorithm for the mutual improvement of optimal transport and energy-based models for the semi-supervised domain adaptation. Having target and source domain samples we use convex energy-based inference to create a new domain that is class-wise cyclical monotone to the target domain and its samples contains features from the source domain examples. Mapping from target to such domain can be solved by optimal transport much more successfully. We study the performance of our approach by benchmarking it on a range of optimal transport methods and showed that in our settings optimal transport can achieve much higher results.

1 INTRODUCTION

Optimal transport provides a theoretical framework to solve mass moving problems for continuous and discrete distributions (Villani, 2008; Peyré & Cuturi, 2019) and it seems to be a natural solution to the domain adaptation problem (Courty et al., 2015). But in practice, computational optimal transport techniques are very sensitive to regularization terms (Courty et al., 2015) and without special scaling often cannot achieve acceptable results in a high dimensional space. Furthermore, optimal transport is a cyclical monotone structured mapping, which makes it not applicable to many empirical distributions.

When we solve the domain adaptation problem with optimal transport, we trying to find a map that can make the target distribution closer to the source distribution. In our approach, we do not use the source distribution during domain adaptation. The core of our method is the idea of creating an ”easy” distribution that is class-wise cyclical monotone to the target distribution and still contains features of the real source dataset. Having a such distribution that is cyclical monotone to target one we can apply optimal transport for domain adaptation much more efficiently.

According to Rockfellar (Rockafellar, 1966) and Breirer's theorems (Theorem 1.22 (Santambrogio, 2015)), we know that the gradient of a convex function is cyclical monotone and solves a Monge problem (Villani, 2008). Owing to this theoretical insight we connect optimal transport and input convex energy-based model to build an efficient domain adaptation framework.

Our method is based on Input Convex Neural Networks (ICNNs) (Amos et al., 2017), which are energy-based models that were used for multi-label classification, structured predictions, and reinforcement learning. Recently, there has been a strong push to further incorporate ICNNs in optimal transport problems (Villani, 2008; Peyré & Cuturi, 2019; Taghvaei & Jalali, 2019; Makkuva et al., 2020). Further development of this approach enabled the construction of the non-minimax Wasserstein-2 generative framework (Korotin et al., 2019). ICNNs were also used for Wasserstein-2 Barycenters estimation (Fan et al., 2020; Korotin et al., 2021). These approaches explored ICNNs as parameterized convex potentials for optimal transport but did not explore their abilities as energy-based models.

Our contribution is two-fold because we improve both optimal transport and energy-based learning method for semi-supervised domain adaptation problems. Usually, the energy-based classifiers are not used in domain adaptation because have significant challenges in inference. On the contrary, we show that convex energy-based inference with a connection to optimal transport can achieve greater results and improve optimal transport performance in comparison to standard settings.

2 Background and Related Work

2.1 Optimal Transport

Optimal Transport aims at finding a solution to transfer mass from one distribution to another with the least effort. Monge's problem was the first example of the optimal transport problem (Villani, 2008) and can be formally expressed as follows:

$$\inf_{T\#\mu_s=\nu_t} \int_{\Omega_{\mu_s}} c(\mathbf{x}, T(\mathbf{x}))\mu(\mathbf{x})d\mathbf{x} \tag{1}$$

The Monge's formulation of optimal transport aims at finding a mapping $T : \Omega_\mu \to \Omega_\nu$ of the two probability measures μ and ν and a cost function $c : \Omega \times \Omega \to [0, +\infty]$, where $T\#\mu_s = \nu_t$ represents the mass preserving push forward operator. In Monge's formulation, T cannot split the mass from a single point. The problem is that such constraints on the mapping T may not even exist and the solution for that mapping, respectively.

To solve this problem, Kantorovitch proposed a relaxation (Villani, 2008). Instead of obtaining a mapping, the goal is to seek a joint distribution over the source and the target that determines how the mass is allocated. For a given symmetric cost function $c : \Omega \times \Omega \to [0, +\infty]$, the primal Kantorovitch formulation can be expressed as the following problem:

$$\min_{\gamma\in\Pi(\mu_s,\nu_t)} \left\{ \int_{\Omega\times\Omega} c d\gamma = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[c(\mathbf{x}, \mathbf{y})] \right\} \tag{2}$$

The primal Kantorovitch formulation has linear objective and linear constraints (Villani, 2008). In this notation, we look for the joint distribution γ with μ_s and μ_t as marginals that minimize the expected transportation cost. If the independent distribution $\gamma(\mathbf{x}, \mathbf{y}) = \mu(\mathbf{x})\nu(\mathbf{y})$ respects the constraints, linear program is convex and always has a solution for a semi lower continuous c:

$$\Pi(\mu_s, \nu_t) = \left\{ \gamma \in P(\Omega, \Omega) : \int \gamma(\mathbf{x}, \mathbf{y})d\mathbf{y} = \mu_s(\mathbf{x}), \int \gamma(\mathbf{x}, \mathbf{y})d\mathbf{x} = \nu_t(\mathbf{y}) \right\} \tag{3}$$

The primal Kantorovitch formulation can be also presented in dual form as stated by the Rockafellar-Fenchel theorem (Villani, 2008):

$$\max_{\phi\in\mathcal{C}(\Omega_s),\psi\in\mathcal{C}(\Omega_t)} \left\{ \int \phi d\mu_s + \int \psi d\nu_t \mid \phi(\mathbf{x}) + \psi(\mathbf{y}) \le c(\mathbf{x}, \mathbf{y}) \right\} \tag{4}$$

After finding a solution to the transport problem, OT provides a measure of similarity between the two distributions in the form of the optimal transport rate. This similarity is also called Wasserstein distance (Villani, 2008):

$$W_p(\mu_s, \nu_t) = \min_{\gamma\in\Pi(\mu_s,\nu_t)} \left\{ \int_{\Omega_s\times\Omega_t} c(\mathbf{x}, \mathbf{y})d\gamma(\mathbf{x}, \mathbf{y}) \right\}^{\frac{1}{p}} \tag{5}$$

where $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$ and $p > 1$. The Wasserstein distance encodes the geometry of the space through the optimization problem and can be used on any distribution of mass.

2.2 Energy-Based Learning

Energy-based learning provides a unified framework for many probabilistic and non-probabilistic approaches, particularly for non-probabilistic training of graphical models including discriminative and generative approaches, as well as conditional random fields, graph-transformer networks, maximum margin Markov networks, and several manifold learning methods (LeCun et al., 2006). In energy-based settings for some given fixed x and possibly some fixed elements of y we can perform inference by:

$$\arg\min_{y} f(x, y; \theta) \tag{6}$$

Energy-based learning approaches can be considered as an alternative to probabilistic estimation for prediction, classification, or decision-making tasks. The energy-based representation must capture both, the discriminative interactions between x and y, and also allow for efficient combinatorial optimization over y.

2.3 Domain Adaptation

Domain adaptation generalizes a learner across different domains of different distributions. The importance of the divergence between the data probability distribution functions of the different domains was theoretically researched in (Ben-David et al., 2010a;b; Germain et al., 2013). This works proposed a primary way of solving the domain adaptation problem by transforming data to make different domain distributions "closer".

Globally, we can divide domain adaptation problems into three categories: (i) unsupervised domain adaptation, when the set of labels in the target domain is unavailable; (ii) semi-supervised domain adaptation, when we have a "small" set of labeled target examples; (iii) supervised domain adaptation, when labels are available for all considered target domain examples. In our paper, we research an semi-supervised settings when we have source domain samples $X_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$ with labels $Y_s = \{\mathbf{y}_i^s\}_{i=1}^{N_s}$ and target domain samples $X_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ with small number of labels $Y_t = \{\mathbf{y}_i^t\}_{i=1}^{N_t}$.

3 Approach

3.1 Easy domain

Current domain adaptation methods estimate both target and source distributions and find a function that can make these distributions closer and more similar. In our approach instead of making target distribution closer to the source, we try to make it closer to the distribution which we call "easy".

It was shown that even after the first epoch, there exist samples that can be easily classified (Arpit et al., 2017). Our idea is that by using gradient descent in the classifier input space, we can find a transformation over the target distribution, that makes it "easy" for the classifier. This procedure is similar to adversarial attacks (Szegedy et al., 2014). We can say that in our case we apply targeted adversarial attack but using the true class as an aim for the targeted attack. As the result, we will have a target samples that is similar to the source domain samples with respect to the inductive biases of the classifier.

A problem can arise when we try to find an optimal transport map from target to the "easy" domain. The resulting "easy" samples can be less similar to the target than the source. In settings with standard feed-forward neural architectures, we have no control over the transformations from target to the "easy" samples and can't make them monotone. To have control over these transformations, we turn to ICNNs energy-based classifier, while for this model, "easy" samples generation can be a convex optimization problem.

3.2 Inversed convex energy-based inference

Here is our approach in more detail. First of all, we build a model f_θ that is Partly Input Convex Neural Network (PICNNs) (Amos et al., 2017) over x instead of y (Amos et al., 2017), it's make possible to apply convex inference over x. Secondary we train our model f_θ on the source domain (X_s, Y_s) in setup equal to Structured Prediction Energy Networks (SPEN) (Belanger & McCallum, 2016), but instead of multi-label classification, we simply perform multi-class classification.

In the next step we apply the core idea of our approach, we use a model trained on (X_s, Y_s), to create an "easy" domain which is cyclical monotone to the target domain (X_t, Y_t) and contain features from the source domain examples. To make it, we simply input $\mathbf{x}_i^t$ samples to the model f_θ and apply inference over it with the fixed $\mathbf{y}_i^t$, in general it can be presented as follows:

$$\arg\min_x f(x, y; \theta) \tag{7}$$

As stated before, Rockafellar's Theorem(Rockafellar, 1966) says that every cyclical monotone mapping g is contained in a sub-gradient of some convex function $f : \mathcal{X} \to \mathbb{R}$. Furthermore, according to Brenier's Theorem, these gradients uniquely solve the Monge problem:

Brenier's Theorem, Theorem 1.22 of (Santambrogio, 2015)). Let μ, ν be probability measures with a finite second moment, and assume μ has a Lebesgue density p_X. Then there exists a convex potential G such that the gradient map $g := \nabla G$ (defined up to a null set) uniquely solves the Monge problem eq. (1) with the quadratic cost function $c(x, y) = \|x - y\|^2$.

We used the gradient descent method to inference in (7) and while our f_θ is input convex over the x, applying this procedure over the target samples we can collect a new dataset (X_e, Y_e) where X_e samples are class-wise cyclical monotone to X_t. And according to the theorems presented before, there exists optimal transport that can solve transportation from X_t to X_e inside each class. The procedure of solving the semi-supervised domain adaptation problem by our method is presented in Algorithm 1.

Algorithm 1: Inversed energy-based inference for semi-supervised domain adpatation

Input: Input convex network f_θ, optimal transport algorithm OT, source dataset (X_s, Y_s), target dataset (X_t, Y_t) with a little Y_t and empty (X_e, Y_e).
Train EBM f_θ on (X_s, Y_s).
for x_t, y_t in (X_t, Y_t) **do**
  $x_e = \arg\min_{x_t} f(x_t, y_t; \theta)$
  $y_e = y_t$
  Append x_e and y_e to "easy" dataset (X_e, Y_e)
end for
Use OT to find a map between X_t and X_e.

4 EXPERIMENTAL EVALUATION

We tested our model on MNIST (LeCun & Cortes, 2010) and USPS datasets (Hull, 1994). In both experiments, we train PICNNs with a one hidden layer size of 100, in the SPEN settings using SGD (Ruder, 2016) optimizer with a learning rate equal to 1e-3 and momentum equal to 0.9.

Table 1: Results on MNIST and USPS dataset in semi-supervised settings. Source denote accuracy of the model on the target domain without domain adaptation. U is USPS, M is MNIST, EU is "easy" USPS, and EM is "easy" MNIST. In the top table presented results for the settings with the 10 known labels for each class in the target domain, and the bottom table present result with the 100 known labels for each class

METHOD	SOURCE	EMD	SINKH	SINKH L1LP	SINKH L1L2	SINKH SS.	MAPT
U→M	28.84	39.33	37.76	31.14	11.8	43.29	28.15
U→EU	-	**76.13**	**76.13**	**76.13**	**76.13**	**76.13**	**68.21**
M→U	28.56	34.43	30.42	25.75	14.29	44.80	28.68
M→EM	-	**68.72**	**62.74**	**57.92**	**16.02**	**63.27**	**56.11**

U→M	28.84	45.54	35.52	29.19	9.41	45.98	33.58
U→EU	-	**86.39**	**86.39**	**86.34**	**86.34**	**86.39**	**72.09**
M→U	28.56	45.01	29.9	26.09	9.68	52.92	48.12
M→EM	-	**85.80**	**77.60**	**69.31**	**12.59**	**79.56**	**61.35**

After training, we obtain an MNIST classifier with 95.70% test set accuracy and USPS trained classifier with an accuracy equal to 90.24%. The accuracies are lower compared to the standard machine learning classifiers because energy-based learning has significant challenges in inference. Then when we have a trained model, we can collect "easy" domain examples. For example: for the network trained on the MNIST dataset our aim is to improve its accuracy on the USPS dataset, so the "easy" dataset for the MNIST model will be the USPS dataset transformed by the input gradients of the network trained on MNIST.

Collected "easy" examples can contain very strong source domain features. The accuracy of the model can be higher on "easy" examples than on the source. For example, the accuracy of the model trained on MNIST is 95.70% and the accuracy of the same model on USPS samples transformed to the "easy" USPS is equal to 96%. For the model trained on USPS, the accuracy on the MNIST samples transformed to "easy" MNIST is even equal to 99%. When we have such datasets, it remains only to fit the optimal transport on this data and then transport samples without a label to this distribution to solve domain adaptation.

Based on the POT library (Flamary & Courty, 2017) we tested the different variations of optimal transport in our task. First of all, we tested and basic Earth Moving Distance (EMD) and Sinkhorn (Sinkh) (Cuturi, 2013) algorithms. Then we tested regularized versions of the Sinkhorn algorithm with a group lasso regularization (L1L2) and Laplacian regularization (L1LP) (Courty et al., 2015). Finally, we benchmarked the semi-supervised (SS) Sinkhorn algorithm and Linear OT mapping estimators (MapT) (Perrot et al., 2016) in our problem. We provide experiments in semi-supervised settings with 10 and 100 known labels per class for each dataset, the results presented in Table 1. In the tables, we can see that mapping to "easy" representation greatly improves the accuracy of the optimal transport algorithms. All these scores were computed on the test sets of the MNIST and USPS.

5 Conclusion and future work

We propose a simple algorithm for the mutual improvement of optimal transport and energy-based learning for the semi-supervised domain adaptation problem. Using input gradients of the energy-based model, we transform our target examples and get a new source distribution that can be accurately classified by the energy-based model. These transformations are class-vise cyclical monotone and can be approximated by optimal transport approaches much more successful than a standard target to the source domain transformation. We tested our algorithm over the set of optimal transport methods on the MNIST and USPS datasets. This paper is a work in progress. Using Langevin dynamics for sampling from the underlying distribution defined by EBM (instead of finding min) can improve the performance of the proposed methods. In the future, we plan to test our method over the more complicated datasets and optimal transport methods. Also, it is important to test convolution ICNNs and develop new input convex architectures to get not class-wise but full cyclical monotonicity between target and "easy" domains.

References

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017. URL `http://proceedings.mlr.press/v70/amos17b.html`.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL `http://proceedings.mlr.press/v70/arpit17a.html`.

David Belanger and Andrew McCallum. Structured prediction energy networks. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Con-*

ference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 983–992. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/belanger16.html.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010a. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In Yee Whye Teh and D. Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 129–136. JMLR.org, 2010b. URL http://proceedings.mlr.press/v9/david10a.html.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *CoRR*, abs/1507.00504, 2015. URL http://arxiv.org/abs/1507.00504.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. *CoRR*, abs/2007.04462, 2020. URL https://arxiv.org/abs/2007.04462.

R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017. URL https://pythonot.github.io/.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 738–746. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/germain13.html.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, and Evgeny Burnaev. Wasserstein-2 generative networks. *CoRR*, abs/1909.13082, 2019. URL http://arxiv.org/abs/1909.13082.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. *CoRR*, abs/2102.01752, 2021. URL https://arxiv.org/abs/2102.01752.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 2020. URL http://proceedings.mlr.press/v119/makkuva20a.html.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4197–4205, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/26f5bd4aa64fdadf96152ca6e6408068-Abstract.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

Filippo Santambrogio. Optimal transport for applied mathematicians. calculus of variations, pdes and modeling. 2015. URL https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *CoRR*, abs/1902.07197, 2019. URL http://arxiv.org/abs/1902.07197.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.ru/books?id=hV8o5R7_5tkC.