2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–8

Optimal Transport for Distribution Adaptation in Bayesian Hilbert Maps

Anthony Tompkins ANTHONY.TOMPKINS@SYDNEY.EDU.AU
The University of Sydney, Australia

Ransalu Senanayake RANSALU@STANFORD.EDU
Stanford University, USA

Fabio Ramos FABIO.RAMOS@SYDNEY.EDU
The University of Sydney, Australia and NVIDIA, USA

Abstract

Parameters are one of the most critical components of machine learning models. As datasets and learning domains change, it is often necessary and time-consuming to re-learn entire models. Rather than re-learning the parameters from scratch, replacing learning with optimization, we propose a framework building upon the theory of *optimal transport* to adapt model parameters by discovering correspondences between models and data, significantly amortizing the training cost. We demonstrate our idea on the challenging problem of creating probabilistic spatial representations for autonomous robots. Although recent mapping techniques have facilitated robust occupancy mapping, learning all spatially-diverse parameters in such approximate Bayesian models demand considerable computational time, discouraging them to be used in real-world robotic mapping. Considering the fact that the geometric features a robot would observe with its sensors are similar across various environments, in this paper, we demonstrate how to re-use parameters and hyperparameters learned in different domains. This adaptation is computationally more efficient than variational inference and Monte Carlo techniques. A series of experiments conducted on realistic settings verified the possibility of transferring thousands of such parameters with a negligible time and memory cost, enabling large-scale mapping in urban environments.

1. Introduction

The quintessential paradigm in the machine learning pipeline consists of the stages of data acquisition and inference of the given data. As data become plentiful, or as ones problem set become more diverse over time, it is common to learn new models tailored to the new data or problem. Contrasting this conventional modeling archetype, we argue that it is often redundant to perform inference and re-learn parameters from scratch. Such model adaptation procedures are indispensable in application domains such as robotics in which the operating environments change continuously. For instance, if the model is represented as a Bayesian model, its distribution should be redetermined regularly to adjust for changes in new data.

In this paper, we focus on significantly improving the training time of building Bayesian *occupancy maps* such as automorphing Bayesian Hilbert maps (ABHMs) Senanayake et al. (2018) by transferring model parameters associated with a set of source datasets to a target dataset in a zero-shot fashion Isele et al. (2016). Despite having attractive theoretical properties and being robust, the main reason that hinders models such as ABHM being used in real-world settings is the run-time cost of learning thousands of parameters (main parameters and hyperparameters). Moreover, these parameters not only vary across different places in the same environment, but also change over time.

We demonstrate domain adaptation of "geometry-dependent spatial features" of the ABHM model from a pool of source domains to the current target domain. This is efficiently done using the theory of Optimal Transport Arjovsky et al. (2017). Since the proposed approach completely bypasses explicitly learning parameters of the Bayesian model using domain adaptation, this process can be thought of as "replacing parameter learning with domain adapatation." The notation given in Table 1 will be used throughout the rest of the paper.

Table 1: Table of notations and terminology

Notation	Description
$\bar{\ }$ and $\breve{\ }$	Mean and variance
$\mathbf{x}, y$	LIDAR data positions and labels
N and M	Number of data points and parameters
$\bar{\mathbf{h}}$	Kernel hinge positions
θ	Parameter set except $\bar{\mathbf{h}}$
$(\mathcal{S})$ and $(\mathcal{T})$	Source and target
P	Coupling matrix
$a \to b$	transport = transfer = domain adaptation

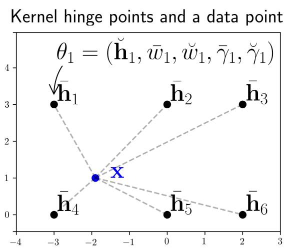

Figure 1: The distances between each datapoint $\mathbf{x}$ and kernels hinged in different locations $\{\bar{\mathbf{h}}_m\}_{m=1}^M$ are evaluated.

2. Nonstationary Mapping using Bayesian Logistic Regression

An occupancy model is typically a parameterized function which gives the probability of a given point in the environment being occupied. For instance, having learned a function with parameters θ, it is possible to query $y_* = p(\text{occupied}|\mathbf{x}_*, \theta) \in [0, 1]$ for anywhere in the space $\mathbf{x}_* = (\text{longitude}, \text{latitude}) \in \mathbb{R}^2$. The parameters θ must be estimated from data gathered using a LIDAR sensor with labels $y = \{0, 1\} = \{\text{free}, \text{hit}\}$. The high level idea of ABHM is projecting LIDAR data into the reproducing kernel Hilbert space (RKHS)—a rich high dimensional feature space—and performing Bayesian logistic regression. The occupancy probability of a point $\mathbf{x}$ is given by $p(y|\mathbf{x}) = \text{sigmoid}(\sum_{m=1}^{M} w_m \exp{(-\gamma_m \|\mathbf{x} - \mathbf{h}_m\|_2^2)})$ with weights $w \in \mathbb{R}$, kernel hinged at spatial locations $\mathbf{h} \in \mathbb{R}^2$, and width of the squared-exponential (SE) kernel $\gamma \in \mathbb{R}_+$. As shown in Figure 1, here, M SE kernels positioned at M sparial locations $\{\mathbf{h}_m\}_{m=1}^M$ are used to project 2D data into a M dimensional vector such that each kernel has more effect from data in its locality.

The distributions $\{w_m \sim \mathcal{N}(\bar{w}_m, \breve{w}_m), \gamma_m \sim \text{Gamma}(\bar{\gamma}_m, \breve{\gamma}_m), \mathbf{h}_m \sim \mathcal{N}(\bar{\mathbf{h}}_m, \breve{\mathbf{h}}_m)\}_{m=1}^M$ must be learned from LIDAR data. Slightly abusing standard notations, in this paper, $\bar{\ }$ and $\breve{\ }$ symbols are used to represent the mean and dispersion parameters, respectively. One of the most important parameters for later discussions is the location parameter $\bar{\mathbf{h}}_m \in \mathbb{R}^2$. Because of the intractable posterior, the parameters of the model are learned Senanayake et al. (2018) using variational inference through probabilistic programming Tran et al. (2017).

3. Parameter Optimal Transport (POT)

In this section, we propose a framework for swiftly adapting thousands of parameter and hyperparameters of the Bayesian mapping model. To adapt to domains, we require accurately pre-trained maps from which we can extract spatially relevant features. In the context of our problem we must extract LIDAR scans (hits and free) with their corresponding model parameters $\{(\bar{\mathbf{h}}, \theta)\}$. To simplify further discussions, as in Figure 1, θ is defined as all parameters except the mean location parameter $\bar{\mathbf{h}}$. We define source LIDAR data $\{(\mathbf{x}_n^{(\mathcal{S})}, y_n^{(\mathcal{S})})\}_{n=1}^{N^{(\mathcal{S})}}$ with corresponding parameters learned from ABHM $\{\theta_m^{(\mathcal{S})}\}_{m=1}^{M^{(\mathcal{S})}}$ as the source *atom*. The source is an environment small enough to be trainable with ABHM.

Having determined the source atom, our **objective** is to determine the new set of parameters $\{\theta^{(\mathcal{T})}\}_{m=1}^{M^{(\mathcal{T})}}$ for a new LIDAR dataset (target) $\{(\mathbf{x}_n^{(\mathcal{T})}, y_n^{(\mathcal{T}))}\}_{n=1}^{N^{(\mathcal{T})}}$. As illustrated in Figure 7, we are looking for a nonlinear mapping technique to convert a source ($\mathcal{S}$) to a target ($\mathcal{T}$). We recognize this as an *optimal transport (OT) problem*. In occupancy mapping, the probability measures are from LIDAR data. For a new target dataset, we attempt to obtain the optimal coupling,

$$P_* = \min_{P \in \Gamma(\mathbf{x}^{(\mathcal{S})}, \mathbf{x}^{(\mathcal{T})})} \sum_{ij} P_{ij} D_{ij} - \lambda^{-1} r(P) \tag{1}$$

for a given $D \in \mathbb{R}^{N^{(\mathcal{S})} \times N^{(\mathcal{T})}}$ distance matrix (e.g. Euclidean distance between source-target pairs) with the information entropy of P, $r(P) = -\sum_{ij} P_{ij} \log P_{ij}$. This entropic regularization, commonly known as the Sinkhorn distance Cuturi (2013); Genevay et al. (2017), enables solving the otherwise hard integer programming problem using an efficient iterative algorithm Sinkhorn and Knopp (1967). Here, λ controls the amount of regularization.

Having obtained the optimal coupling between source and target LIDAR, as illustrated in Figures 7 (b)-(c), now it is possible to transport corresponding source parameters $\theta^{(\mathcal{S})}$ to the target domain. This is done by associating the parameter positions with source samples $\bar{\mathbf{h}}^{(\mathcal{S})}$ as a linear map $\mathbf{x}^{(\mathcal{S})}$ Perrot et al. (2016). Note that all other $\theta^{(\mathcal{S})}$ parameters associated with $\bar{\mathbf{h}}^{(\mathcal{S})}$ will also be transported. This implicit transfer process is depicted in Figure 5. Since ABHM can only be executed in small areas due to the high computational cost, we learn individual ABHM maps for different areas and construct a dictionary of source atoms which we call a *dictionary of atoms* $\mathcal{X}^{(\mathcal{S})}$. As a result, as depicted in Figure 2, atoms from various domains will be transferred to the target. The entire algorithm is given in Algorithm 1.

4. Experiments

We used the Carla simulator Dosovitskiy et al. (2017) and KITTI benchmark dataset Geiger et al. (2013) for experiments. A summary of datasets is listed in Table 5. We compared against vanilla variational inference Senanayake and Ramos (2017); Jaakkola and Jordan (1997) and variational inference with reparameterization trick Senanayake et al. (2018); Kingma and Welling (2013).

Intra-domain and inter-domain adaptation Here we consider two paradigms: intra-domain and inter-domain transfer. In intra-domain transfer, the source atoms are generated from the first 10 frames of a particular dataset and parameters are transferred within the

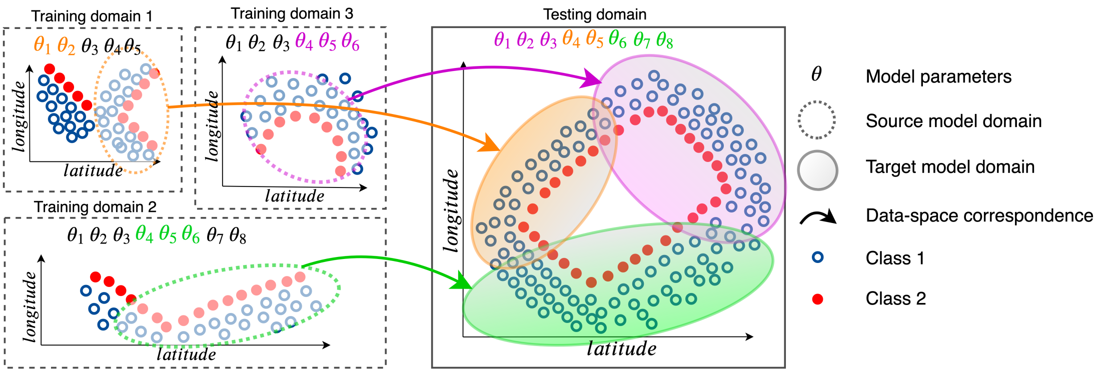

Figure 2: A high-level overview of our method: Parameter Optimal Transport. *Training domains* correspond to potentially independent, data-intensive, expensive, yet small-scale pre-learned models. After storing in a dictionary of atoms, representative data-space and model-parameter tuples from the pre-learned set of models, we find data-space correspondences using optimal transport maps via the ranking procedure. These maps are then used to transport pre-learned parameters to out-of-sample *test domains*. Our method is largely insensitive to data-space invariances between source training domains and test domains reducing knowledge loss during the transfer process.

same dataset. In inter-domain transfer they are transferred to a completely new town. Results are in Table 4 with 20% randomly sampled test LIDAR beams. We consider two paradigms: intra-domain and inter-domain transfer. In intra-domain transfer, the source atoms are generated from the first 10 frames of a particular dataset and parameters are transferred within the same dataset. In inter-domain transfer they are transferred to a new town. Results are in Table 4 with 20% randomly sampled test LIDAR beams from each town. Parameters are aggregated over time to map entire environments visualized in Figure 3.

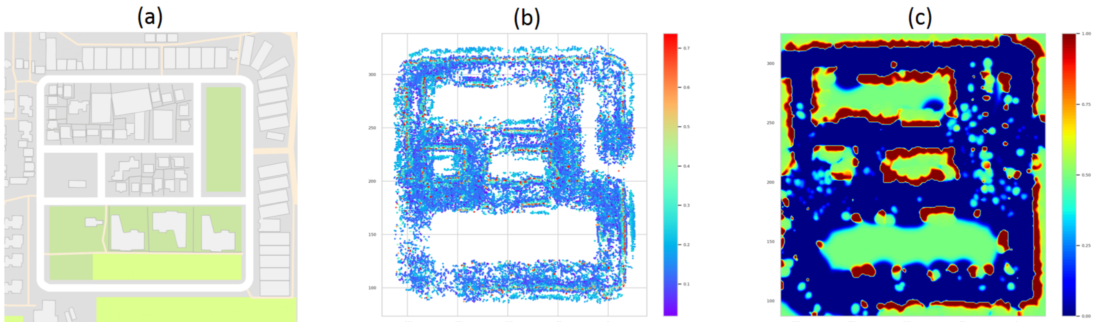

Figure 3: (a) Street map of Carla Town 2 domain. (b) POT Transported kernels for the entire domain. (c) POT probabilistic occupancy map.

Building instantaneous maps This experiment demonstrates performance of building instantaneous maps. For this purpose, we use the two dynamic datasets: SimCarla and RealKITTI. The source dictionary of atoms was prepared similar to the intra/inter-domain

Table 2: Instantaneous map building in dynamic environments. Mean and SD are given. We evaluated the test performance of our model using accuracy (ACC), area under ROC curve (AUC), and negative log-likelihood (NLL) Bishop (2006). The higher the ACC and AUC or lower the NLL, the better.

			Target	
			SimCarla	RealKITTI
Source	ACC	Town 1	0.74 ± 0.10	0.69 ± 0.06
		Town 2	0.70 ± 0.10	0.58 ± 0.06
	AUC	Town 1	0.81 ±0.11	0.77 ±0.06
		Town 2	0.77 ±0.12	0.73 ±0.06
	NLL	Town 1	1.06 ± 0.56	1.42 ± 0.38
		Town 2	1.90 ± 0.79	3.63 ± 1.04

Table 3: Performance per time unit.

		Target	
	Method	Town1	Town2
ACC	RePOT	0.95	0.93
	POT	0.85	0.83
	ABHM	0.77	0.59
	BHM	0.66	0.61
AUC	RePOT	0.99	0.98
	POT	0.92	0.92
	ABHM	0.95	0.96
	BHM	0.94	0.92
NLL	RePOT	0.71	1.41
	POT	1.64	1.69
	ABHM	0.58	0.71
	BHM	0.63	0.69

adaptation experiment. Such a map is shown in Figure 6. Table 2 shows the performance of transferring features extracted from each town to the dynamic datasets.

Performance comparison We evaluate various occupancy mapping algorithms in terms of accuracy and speed, noting that ABHM with a high accuracy cannot be run for large environments. We measure the time for POT per LIDAR scan then select the number of kernels to match the same time to run BHM and ABHM; results are in in Table 3. To further improve the map quality, we propose to use transported parameters as prior distributions of the BHM and merely update the weights w in using Senanayake and Ramos (2017). This improved method termed refined POT (RePOT) maps results in further performance gains.

5. Conclusion

This paper introduced optimal transport theory as a framework for distribution (parameter and hyperparameter) adaptation of a Bayesian model with applications to occupancy mapping. We demonstrated quickly (1 second) adapting thousands of parameters and benchmarked the performance with realistic datasets.

References

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

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems (NIPS)*, pages 2292–2300, 2013.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

David Isele, Mohammad Rostami, and Eric Eaton. Using task features for zero-shot knowledge transfer in lifelong learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1620–1626, 2016.

Tommi Jaakkola and Michael Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, volume 82, 1997.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4197–4205, 2016.

Ransalu Senanayake and Fabio Ramos. Bayesian hilbert maps for dynamic continuous occupancy mapping. In *Conference on Robot Learning (CoRL)*, pages 458–471, 2017.

Ransalu Senanayake, Anthony Tompkins, and Fabio Ramos. Automorphing kernels for nonstationarity in mapping unstructured environments. In *Conference on Robot Learning (CoRL)*, pages 443–455, 2018.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Dustin Tran, Matthew D. Hoffman, Rif A. Saurous, Eugene Brevdo, Kevin Murphy, and David M. Blei. Deep probabilistic programming. In *International Conference on Learning Representations (ICLR)*, 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.