Optimal Transport for Distribution Adaptation in Bayesian Hilbert Maps

Anthony Tompkins
The University of Sydney, Australia
anthony.tompkins@sydney.edu.au

Ransalu Senanayake
Stanford University, USA
ransalu@stanford.edu

Fabio Ramos
The University of Sydney, Australia and NVIDIA, USA
fabio.ramos@sydney.edu

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>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\gamma}$ and $\hat{\gamma}$</td>
<td>Mean and variance</td>
</tr>
<tr>
<td>$x, y$</td>
<td>LIDAR data positions and labels</td>
</tr>
<tr>
<td>$N$ and $M$</td>
<td>Number of data points and parameters</td>
</tr>
<tr>
<td>$\bar{h}$</td>
<td>Kernel hinge positions</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Parameter set except $\bar{h}$</td>
</tr>
<tr>
<td>$(S)$ and $(T)$</td>
<td>Source and target</td>
</tr>
<tr>
<td>$P$</td>
<td>Coupling matrix</td>
</tr>
</tbody>
</table>

Figure 1: The distances between each datapoint $x$ and kernels hinged in different locations $\{\bar{h}_m\}_{m=1}^M$ 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 $\theta$, it is possible to query $y_{\ast} = p(\text{occupied}|x_{\ast}, \theta) \in [0, 1]$ for anywhere in the space $x_{\ast} = (\text{longitude}, \text{latitude}) \in \mathbb{R}^2$. The parameters $\theta$ 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 $x$ is given by $p(y|x) = \text{sigmoid}(\sum_{m=1}^M w_m \exp (-\gamma_m \|x - h_m\|^2_2))$ with weights $w \in \mathbb{R}$, kernel hinged at spatial locations $h \in \mathbb{R}^2$, and width of the squared-exponential (SE) kernel $\gamma \in \mathbb{R}_+$. As shown in Figure 1, here, $M$ SE kernels positioned at $M$ spatial locations $\{h_m\}_{m=1}^M$ 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 $\{w_m \sim \mathcal{N}(\bar{w}_m, \hat{w}_m), \gamma_m \sim \Gamma(\bar{\gamma}_m, \hat{\gamma}_m), h_m \sim \mathcal{N}(\bar{h}_m, \hat{h}_m)\}_{m=1}^M$ must be learned from LIDAR data. Slightly abusing standard notations, in this paper, $\bar{\cdot}$ and $\hat{\cdot}$ symbols are used to represent the mean and dispersion parameters, respectively. One of the most important parameters for later discussions is the location parameter $\hat{h}_m \in \mathbb{R}^2$. 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 \(\{(\vec{h}, \theta)\}\). To simplify further discussions, as in Figure 1, \(\theta\) is defined as all parameters except the mean location parameter \(\vec{h}\). We define source LIDAR data \(\{(x^{(S)}_n, y^{(S)}_n)\}_{n=1}^{N^{(S)}}\) with corresponding parameters learned from ABHM \(\{\theta^{(S)}_m\}_{m=1}^{M^{(S)}}\) 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 \(\{\theta^{(T)}_m\}_{m=1}^{M^{(T)}}\) for a new LIDAR dataset (target) \(\{(x^{(T)}_n, y^{(T)}_n)\}_{n=1}^{N^{(T)}}\). As illustrated in Figure 7, we are looking for a nonlinear mapping technique to convert a source \(S\) to a target \(T\). 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,

\[
P^*_s = \min_{P \in \Gamma(x^{(S)}, x^{(T)})} \sum_{ij} P_{ij} D_{ij} - \lambda^{-1} r(P)
\]

for a given \(D \in \mathbb{R}^{N^{(S)} \times N^{(T)}}\) 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, \(\lambda\) 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 \(\theta^{(S)}\) to the target domain. This is done by associating the parameter positions with source samples \(\vec{h}^{(S)}\) Perrot et al. (2016). Note that all other \(\theta^{(S)}\) parameters associated with \(\vec{h}^{(S)}\) 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 \(\mathcal{X}^{(S)}\). 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
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 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**


Figure 4: (a) 10 red and 10 brown dots indicate samples from the source and target distributions, respectively. Gray lines indicate high probable couplings (matches) obtained after solving equation 1. (b) Pairwise cost matrix \((10 \times 10)\) between the positions of samples. (c) Coupling matrix \((10 \times 10)\) indicates the coupling probability of source points and all other target points. Determining this matrix is the goal of optimal transport.

Intuitively, as illustrated in Figures 4 and 7 (a), the OT problem attempts to determine the optimal way to move one probability distribution to another. If \(\mu^{(S)}\) and \(\mu^{(T)}\) constitute of two datasets of size \(N^{(S)}\) and \(N^{(T)}\), respectively, there always exists an optimal probabilistic coupling \(P^* \in \mathbb{R}^{N^{(S)} \times N^{(T)}}\) between the two datasets Courty et al. (2017). Here, as shown in Figures 4 in which the source and target samples are assumed to separately follow bivariate distributions, \(P^*\) is a doubly stochastic matrix—each row and column sums to one—that indicates the probability of a sample in the source match with all other points in the target.

**Theorem 1 (Monge-Kantorovich) Villani (2008)** Let \(\Omega^{(S)}\) and \(\Omega^{(T)}\) be two separable metric spaces such that probability measures \(\mu^{(S)}\) and \(\mu^{(T)}\) on \(\Omega^{(S)}\) and \(\Omega^{(T)}\), respectively, are Radon measures. The optimal coupling,

\[
P^* = \inf_{P \in \Gamma(\mu^{(S)}, \mu^{(T)})} \int_{\Omega^{(S)} \times \Omega^{(T)}} D(\mu^{(S)}, \mu^{(T)}) dP(\mu^{(S)}, \mu^{(T)}),
\]

always exists for a distance function \(D : \Omega^{(S)} \times \Omega^{(T)} \to [0, \infty)\), where \(\Gamma\) is the set of all couplings (probability measures) on \(\Omega^{(S)}\) and \(\Omega^{(T)}\) with marginals \(\mu^{(S)}\) and \(\mu^{(T)}\), respectively.
learn $P_\star$ for $x^{(S)} \xrightarrow{\text{explicit transport}} x^{(T)}$

predict $\bar{h}^{(S)} \xrightarrow{\text{explicit transport}} \bar{h}^{(T)}$ using $P_\star$

$\vdots$

$\theta^{(S)} \xrightarrow{\text{implicit transport}} \theta^{(T)}$

Figure 5: Parameter optimal transport. Known in blue and unknown in red. We learn an optimal coupling matrix $P_\star$ using training LIDAR. This coupling is then used to predict target kernel hinge positions $\bar{h}^{(T)}$ corresponding to the source kernel hinge positions $\bar{h}^{(S)}$. Parameters connected to each kernel are implicitly transported $\theta^{(S)} \rightarrow \theta^{(T)}$.

Input: LIDAR hits
Train a small-scale environment with ABHM
Create a source dictionary with the trained $\mathcal{X}^{(S)}$

while New scan do
  for Each source atom $x^{(S)} \in \mathcal{X}^{(S)}$ do
    for Each rotation of the atom $Rx^{(S)}$ do
      Compute candidate coupling matrix $P$
    end
  end

Determine the best coupling matrix $P_\star$ (Equation 1)
Transfer $\bar{h}^{(S)}$ and $\theta^{(S)}$ to targets with $P_\star$ (Figure 5)
end

Output: Parameters $\theta$

Algorithm 1: Parameter Optimal Transport

Figure 6: (a) Car camera view (b) transferred optimal parameters (lengthscales) based on LIDAR data (c) Occupancy probability map built from transferred parameters
Figure 7: Optimal transport from a square to an arc. (a) There are $N_S$ and $N_T$ number of data points in the source (red) and target (brown) datasets, respectively. With reference to Figure 4 and Theorem 1, the doubly stochastic coupling matrix $\gamma$ is size $N_S \times N_T$. A given row in $\gamma$ indicates the probabilities of the sample associated with that row could be coupled to all samples in the target dataset. Probabilities associated with one such source point to target matches are shown in white-black color scale. Note that only the 10 highest matches are shown for clarity (b) for a given set of LIDAR hits (red) spatial parameters can be learned using ABHM. For instance, the lengthscales spread all over the environment are shown here. However, for another set of LIDAR hits (brown) we do not like to learn the parameters again because it is expensive. (c) Based on the coupling matrix between the source and the target, we transport (move from the target area to the source area) the parameters around each point. Note that how the small lengthscales (cyan) stays close to the LIDAR hits and larger lengthscales (magenta) move away from the LIDAR.

Table 4: Performance of intra-domain (diagonal entries of the table) and inter-domain (off-diagonal entries of the table) transfer.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Town1</th>
<th>Town2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>Town1</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Town2</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td>AUC</td>
<td>Town1</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Town2</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>NLL</td>
<td>Town1</td>
<td>1.14</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Town2</td>
<td>3.30</td>
<td>3.23</td>
</tr>
</tbody>
</table>

Table 5: Description of domains

<table>
<thead>
<tr>
<th>Domains</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Town 1</td>
<td>2D Town 1 in Carla.</td>
</tr>
<tr>
<td>Town 2</td>
<td>2D Town 2 in Carla.</td>
</tr>
<tr>
<td>SimCarla</td>
<td>2D Town 1 with 120 vehicles.</td>
</tr>
<tr>
<td>RealKITTI</td>
<td>2D — middle LIDAR channel.</td>
</tr>
</tbody>
</table>