2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Optimal Transport for Correctional Learning

Rebecka Winqvist, Inês Lourenço, Francesco Quinzan, Cristian R. Rojas and Bo Wahlberg

Abstract— **The contribution of this paper is a generalized formulation of correctional learning using optimal transport, which is about how to optimally transport one mass distribution to another. Correctional learning is a framework developed to enhance the accuracy of parameter estimation processes by means of a teacher-student approach. In this framework, an expert agent, referred to as the teacher, modifies the data used by a learning agent, known as the student, to improve its estimation process. The objective of the teacher is to alter the data such that the student's estimation error is minimized, subject to a fixed intervention budget. Compared to existing formulations of correctional learning, our novel optimal transport approach provides several benefits. It allows for the estimation of more complex characteristics as well as the consideration of multiple intervention policies for the teacher. We evaluate our approach on two theoretical examples, and on a human-robot interaction application in which the teacher's role is to improve the robots performance in an inverse reinforcement learning setting.**

I. Introduction

Parameter estimation refers to the process of determining a model's parameter values based on measured data. The values of these parameters have a direct impact on the distribution of the data generated by the modeled system. Estimation theory is a well-researched topic with several established methods, see e.g. [1]. The interest in the subject is widespread with applications to be found in, among others, the process industries, control applications, as well as in the research of biological functions and systems [2]. Popular estimation methods range from the conventional maximum likelihood and Bayesian inference methods, to more recent learning-based methods.

Common to most of these estimators is their data-driven nature. In many real-world applications, however, the available data often does not accurately reflect the underlying distribution or behavior of the system being studied. This can be the result of limited sample sizes, biased sampling methods, measurement errors, or outliers [3]. Relying on such data when using data-driven estimators can result in inaccurate parameter estimation and poor model performance. For example, a recent study found that commercially available facial analysis algorithms showed higher error rates for darker-skinned individuals and women [4], which could be linked to unrepresentative training data.

Correctional learning is a recently developed framework that may be used to address this issue [5], [6]. The framework arises from the idea of cooperative (learning) problems, i.e., settings in which two or more agents work together towards a common goal. The framework is structured around a teacher-student model, where an expert (teacher) agent seeks to assist a learner (student) agent in its estimation process. More specifically, the teacher's goal is to modify (or correct) the collected observations, based on which the student forms its estimation. See Figure 1 for an illustration of this. Correctional learning can thus be viewed as a means of finding an optimal mapping from the original observations to a modified sequence that minimizes the student's estimation error.

Recent works on correctional learning have shown promising results in both offline and online settings [5], [6]. Nevertheless, the framework still suffers from some limitations. First of all, the derived performance guarantees hold only for simple systems. To describe real-world phenomena, however, one typically requires more complex distributions. For example, a Gaussian distribution can be used to describe biological data such as the heights of people. To model the probability of failure of an appliance, we can use the Weibull distribution. Another disadvantage is that the teacher's policy follows explicitly from the solution, leaving no room for alternative intervention strategies to be considered.

Our contribution is an alternative approach to correctional learning using tools from optimal transport [7]. Optimal transport is a mathematical framework concerned with finding the most efficient way to transport mass from one location to another, according to some cost function. Historically, optimal transport has been widely used in finance and logistics [8], but recent advances have made it an increasingly popular tool in fields such as systems, control and estimation [9], [10]. In machine learning, optimal transport has found use in a number of applications, including shape reconstruction [11], multi-label classification [12], and brain decoding [13]. Moreover, recent work in robotics demonstrates how optimal transport can be applied to mapping problems to enable robots to operate in new environments [14], and for policy fusion in reinforcement learning to speed up the process of a robot learning a new task [15].

In the context of correctional learning, we note that the optimal corrections can be viewed as a transportation of probability mass from an initial distribution into a target distribution. Furthermore, by assuming that the estimator depends on the samples only through their empirical measure, we can pose the correctional learning problem as an optimization program in terms of distribution functions – i.e., as an optimal transport problem. In contrast to [5] and [6], this novel formulation considers the samples implicitly

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), the Swedish Research Council Research Environment NewLEADS under contract 2016-06079, and the Digital Futures project EXTREMUM.

The authors are with the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden {rebwin, ineslo, quinzan, crro, bo}@kth.se

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

through their distribution, which not only enables the estimation of more complex parameters, but also allows for the consideration of alternative intervention strategies.

The main contributions of this paper are:

- *A generalized correctional learning framework:* we leverage the principles of optimal transport and propose a novel formulation of correctional learning. With this new framework, we can expand the range of applications to consider more sophisticated tasks that involve complex systems.
- *Multiple teacher policies:* in standard learning settings, a teacher agent may exhibit several intervention strategies. We show how our new formulation allows for the consideration of multiple teacher policies to fit different tasks.
- *Evaluation of performance:* we demonstrate the benefits of our optimal-transport approach by applying the framework on three different test cases. Specifically, we show how the framework can be used to estimate the parameters of more complex distributions such as the Gaussian and the Weibull. We also apply the framework to update a robot's reward function in an inverse reinforcement learning setting.

A. Related Work

Learning from experts is a widely studied problem in machine learning. Learning from demonstrations [16] and imitation learning [17] are two closely related paradigms where a robot learns from observing the behavior of an expert. In corrective feedback [18], on the other hand, the expert provides corrections to the robot's actions to improve its learning process. This is opposed to correctional learning, where the corrections are made to the data that the robot learns from.

By interpreting correctional learning as a means for customizing a data set to better suit a specific learning task, we find similarities with other techniques in machine learning and statistics. Feature selection [19] is one such example, where the aim is to find the most informative and relevant features to improve learning. Other examples include active learning [20], where the learner can interactively ask the expert to label new data, and input for system identification [21] [22], where the aim is to design input data to optimize for a model's accuracy.

In the context of system identification and estimation, our framework can also be placed around other works that also use tools from optimal transport. For example, in [23], the authors use optimal transport for state tracking of linear ensembles. More specifically, they propose an optimal-transport approach for estimating the states of multiple subsystems based on their joint output. Other related works include [24], where they propose an optimal transport formulation of the ensemble Kalman filter, and [25], where the authors study the use of optimal transport distances as objective functions for parameter estimation in dynamical systems.

II. Preliminaries

In this section, we first define the notation used throughout the paper. We then give a brief introduction to correctional learning and optimal transport.

A. Notation

We use $(\mathcal{Y}, \mathcal{B})$ to denote a measurable space, in which $\mathcal{Y}$ is a set, and $\mathcal{B}$ is a σ-algebra of subsets of $\mathcal{Y}$. We denote the set of positive measures on $(\mathcal{Y}, \mathcal{B})$ by $\mathcal{M}_+(\mathcal{Y})$. For $N \in \mathbb{N}$, we define $[N] := \{1, 2, \dots, N\}$. For a set $\mathcal{A}$, we use $|\mathcal{A}|$ to denote its cardinality. We use $\mathbb{I}$ to denote the indicator function, $\mathbb{1}$ to denote a vector of ones, and I to denote the identity matrix. Inequalities between vectors and matrices are considered element-wise. We use the words samples and observations interchangeably throughout the paper.

B. Correctional Learning

Consider a model of some data-generating system parameterized by the unknown parameter $\theta \in \Theta$. Let the true system correspond to the value θ_0. In a standard parameter estimation setting, a learner (student) agent aims to estimate θ_0 as $\hat{\theta}_N$, based on a sequence of observations sampled from the system, $\mathcal{O}_N = \{y_1, \dots, y_N\}$, distributed according to $p_0^N \in \mathcal{M}_+(\mathcal{Y}^N)$, where $y_i \in \mathcal{Y} \subseteq \mathbb{R}^d$ for all i, and $(\mathcal{Y}, \mathcal{B})$ is a measurable space. That is,

$$\hat{\theta}_N = f_N(\mathcal{O}_N) = f_N(y_1, \dots, y_N), \tag{1}$$

where $f_N \colon \mathcal{Y}^N \to \Theta$ is some estimator function. In the rest of the paper, we will omit the dependence on N if its value is clear from the context.

In the correctional learning framework, an expert (teacher) agent is introduced to help the the student in its estimation process. The teacher may do so by modifying the original observation sequence, $\mathcal{O}$, into a sequence, $\tilde{\mathcal{O}}$, that better represents the true characteristics of the system. The modified sequence is then passed on to the student, who forms the altered estimate $\tilde{\theta}$.

However, utilizing expert knowledge might be expensive or limited. The number of allowed interventions might also be restricted for privacy preserving reasons – the more observations the teacher changes, the more likely it is to be discovered. To account for this, the teacher is constrained to not exceed a certain intervention budget b. If $C_N \colon \mathcal{Y}^N \times \mathcal{Y}^N \to \mathbb{R}_0^+$ denotes a distance measure between two sequences of N elements, then the teacher must satisfy

$$C_N(\mathcal{O}_N, \tilde{\mathcal{O}}_N) \leq b. \tag{2}$$

The cost may be chosen to be any distance metric, e.g. the ℓ_1-norm for discrete observations.

The goal of the teacher agent is to find the optimal modified sequence that minimizes the student's estimation error

$$V(\theta_0, \tilde{\theta}) = \|\theta_0 - \tilde{\theta}\|. \tag{3}$$

where $\|\cdot\|$ is a norm on Θ.

Depending on the setting, this problem can be posed and solved in different ways:

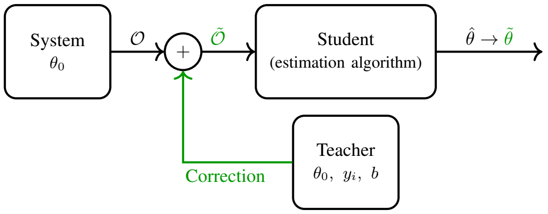

Fig. 1. A schematic view of the correctional learning framework. The teacher knows the true parameter value θ_0 and the original samples y_i. The teacher modifies the original sequence of observations, $\mathcal{O}$, into $\tilde{\mathcal{O}}$, while adhering to the budget constraint b.

1) Batch setting: In the offline (batch) setting, the observations are made available to the teacher in batches. By having access to all samples at once, it is shown in [5] that the offline problem can be cast as the optimization program

$$\begin{aligned} \min_{\tilde{\mathcal{O}}} \quad & V(\theta_0, \tilde{\theta}) \\ \text{s.t.} \quad & \tilde{y}_i \in \mathcal{Y}, \text{ for all } \tilde{y}_i \in \tilde{\mathcal{O}}, \\ & C(\mathcal{O}, \tilde{\mathcal{O}}) \leq b. \end{aligned} \tag{4}$$

2) Online setting: In the online setting, the observations are made available sequentially (one at a time). This means that the teacher has to decide, at each time step, whether or not to change the new incoming observation. In [6], the authors show how the online problem can be formulated as a Markov decision process, and solved using dynamic programming.

A schematic view of a general correctional learning framework is provided in Figure 1.

C. Optimal Transport

Assume that we are given a probability measure $\mu \in \mathcal{M}_+(\mathcal{X})$, which can be interpreted as, say, a distribution of sand in $\mathcal{X}$ of total mass 1. Assume further that we wish to transform μ into another probability measure $\nu \in \mathcal{M}_+(\tilde{\mathcal{X}})$, corresponding to a different distribution of sand, by "moving" the grains of sand with minimal transportation cost. The cost of transporting one unit of probability mass from location x to location $\tilde{x}$ is quantified by a metric on $\mathcal{X}$, $c(x, \tilde{x})$. To compute the total transportation cost, we must also define a transportation map, which is modelled by a probability measure $\pi \in \mathcal{M}_+(\mathcal{X} \times \tilde{\mathcal{X}})$, with $d\pi(x, \tilde{x})$ denoting the amount of mass transferred from x to $\tilde{x}$. Since we cannot move more mass than what we originally have, it must hold that the mass moved from one point in μ must be received by ν and vice versa. In mathematical terms, we express those conditions by

$$\int_{\tilde{\mathcal{X}}} d\pi(x, \tilde{x}) = d\mu(x) \quad \text{and} \quad \int_{\mathcal{X}} d\pi(x, \tilde{x}) = d\nu(\tilde{x}). \tag{5}$$

The problem can now be posed as the optimization program

$$\begin{aligned} \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \tilde{\mathcal{X}})} \quad & \int_{\mathcal{X} \times \tilde{\mathcal{X}}} c(x, \tilde{x}) d\pi(x, \tilde{x}) \\ \text{s.t.} \quad & \int_{\tilde{x} \in \tilde{\mathcal{X}}} d\pi(x, \tilde{x}) = d\mu(x), \\ & \int_{x \in \mathcal{X}} d\pi(x, \tilde{x}) = d\nu(\tilde{x}), \end{aligned} \tag{6}$$

where the cost function $\mathcal{I}(\pi) = \int_{\mathcal{X} \times \tilde{\mathcal{X}}} c(x, \tilde{x}) d\pi(x, \tilde{x})$ is the total transportation cost under the transport plan π. This form is known as Kantorovich's optimal transportation problem, and is a relaxation of the original formulation by Monge. The interested reader is referred to [7], [26] and [10] for more details.

III. Optimal Transport for Correctional Learning

In this section we present the main contribution of this paper: an optimal transport formulation of the *batch* correctional learning problem. The main motivation behind using optimal transport is to create a general framework that allows for the estimation of more complex parameters.

A. General Problem Formulation

Recall the problem setup of correctional learning in Section II-B. We now put it into a more general setting to make its connections to optimal transport more clear.

Assume that the data-generating system is permutation-invariant, or *exchangeable*, in the sense that the distribution p_0^N of the samples $\mathcal{O}_N$ it generates does not change if the samples in $\mathcal{O}_N$ are permuted (in a deterministic manner). It is then natural to consider estimators that are also permutation-invariant, i.e., that can be described as some function of the samples' empirical measure:

$$\hat{\theta}_N(y_1, \ldots, y_N) = J(\hat{p}_N(y_1, \ldots, y_N)), \tag{7}$$

where $\hat{p}_N : \mathcal{Y}^N \to \mathcal{M}_+(\mathcal{Y})$ is the empirical measure of the samples, defined as

$$(\hat{p}_N(y_1, \ldots, y_N))(A) := \sum_{i=1}^{N} \frac{1}{N} \mathbb{I}\{y_i = A\}, \quad A \in \mathcal{B}. \tag{8}$$

The function $J \colon \mathcal{M}_+(\mathcal{Y}) \to \Theta$ is a fixed function (i.e., independent of N), which we will later assume to be Fréchet-differentiable.

Recall that the samples can be perturbed by the teacher before they reach the student. In the batch setting, the teacher has access to all of the original samples, $\mathcal{O}$, before perturbing them into $\tilde{\mathcal{O}} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$, where $\tilde{y}_i \in \tilde{\mathcal{Y}} \subseteq \mathbb{R}^d$ for all i. The teacher is subject to a budget constraint, namely

$$\sum_{i=1}^{N} c(y_i, \tilde{y}_i) \leq b. \tag{9}$$

The goal of the teacher is still to modify the original sequence, $\mathcal{O}$, subject to (9), in order to minimize the estimation error

$$\|\theta_0 - \tilde{\theta}\|, \tag{10}$$

where $\tilde{\theta}$ is the altered estimate based on $\tilde{\mathcal{O}}$.

Since the estimator in (7) depends on the samples only through their empirical distribution, it makes sense to pose the optimization problem to be solved by the teacher in terms of distribution functions, i.e., as an optimal transport problem

$$\min_{p} \left\| \theta_0 - J\left(\int_{y\in\mathcal{Y}} dp(y,\cdot) \right) \right\|^2 \tag{11a}$$

$$\text{s.t.} \int_{(y,\tilde{y})\in\mathcal{Y}\times\tilde{\mathcal{Y}}} c(y,\tilde{y})dp(y,\tilde{y}) \leq \frac{b}{N} \tag{11b}$$

$$\int_{\tilde{y}\in\tilde{\mathcal{Y}}} \int_{y\in A} dp(y,\tilde{y}) = (\hat{p}_N(y_1,\dots,y_N))(A), \quad \forall A \in \mathcal{B}, \tag{11c}$$

where $p \in \mathcal{M}_+(\mathcal{Y}\times\tilde{\mathcal{Y}})$ is a transportation map representing the joint measure of the original and modified samples, and $\hat{p}_N$ the empirical distriubtion of the original samples. We note that this is in general an infinite-dimensional problem with linear constraints. In general, however, J is not necessarily linear. In the case that J is Fréchet-differentiable, and we assume that the budget b is "small" (in the sense that most of the original samples will not be modified), one can use the Taylor approximation of the cost of the modified samples,

$$\begin{aligned} J\left(\int_{y\in\mathcal{Y}} dp(y,\cdot) \right) &\approx J(\hat{p}_N) \\ &+ \int_{y\in\mathcal{Y}} \int_{\tilde{y}\in\tilde{\mathcal{Y}}} (\nabla J(\hat{p}_N))(\tilde{y}) dp(y,\tilde{y}) \\ &- \int_{y\in\mathcal{Y}} (\nabla J(\hat{p}_N))(y) d\hat{p}_N(y), \end{aligned} \tag{12}$$

where $\hat{p}_N$ is the empirical distribution of the original sample sequence $\mathcal{O}$. We let $J_{\text{TA}}(\mathcal{O})$ denote the Taylor approximation. Substitution into (11a) then yields the objective

$$\min_{p} \|\theta_0 - J_{\text{TA}}(\mathcal{O})\|^2. \tag{13}$$

Remark 1: Note how the constraints in (11) now consider the distribution of the samples. This is in contrast to the original formulation in (4), in which each observation is considered individually.

B. Discretization of the Continuous Case

One of the most common approaches to solve the optimal transport problem in (11) is to discretize it [10]. For simplicity, we will assume that the original samples are independent and identically distributed, with distribution p_0. We start by defining a discretized sample space. Recall that our observation sequence is given by the *multiset*[1]

$$\mathcal{O} = \{y_1, \dots, y_N\}. \tag{14}$$

We can let the *set* of unique values of $\mathcal{O}$ constitute our discretized sample space as

$$\mathcal{S} = \bigcup_{o\subseteq\mathcal{O}} o = \{s_1, \dots, s_n\} \subseteq \mathcal{O}. \tag{15}$$

For the continuous case, we note that with probability one,

$$\mathcal{S} = \mathcal{O} \quad \text{and} \quad n = |\mathcal{S}| = |\mathcal{O}| = N, \tag{16}$$

since all the samples in $\mathcal{O}$ are distinct, with probability one. The elements in $\mathcal{S}$ will be called *states*.

Remark 2: We note that there are other methods to determine the states. For instance, they may be fixed to belong to some pre-determined set of values. However, with regards to the nature of the framework of modifying an *observed* sequence, we believe the suggested approach is reasonable.

Furthermore, the teacher will be allowed to change the observations in $\mathcal{O}$ into samples from the set

$$\tilde{\mathcal{S}} = \{\tilde{\boldsymbol{s}}_1, \dots, \tilde{\boldsymbol{s}}_m\}, \tag{17}$$

which may or may not coincide with $\mathcal{S}$. Note that $|\tilde{\mathcal{S}}| = m$.

We now continue by discretizing (11). We note that both the objective in (12) as well as our constraints in (11) include our decision variable $dp(y,\tilde{y})$ in the integrals, which we cannot sample from. Thus, to approximate the integrals, we use techniques from importance sampling [27]. We consider the proposal distribution

$$d\mu(y,\tilde{y}) = dq(y)dr(\tilde{y}), \tag{18}$$

where $dq(y)$ and $dr(\tilde{y})$ are probability measures defined on $\mathcal{S}$ and $\tilde{\mathcal{S}}$, respectively. Note that, for simplicity, μ has been chosen in terms of independent proposal distributions for y and $\tilde{y}$. With this distribution, and restricting p to $\mathcal{S}\times\tilde{\mathcal{S}}$, we rewrite the estimator in (12) as

$$\begin{aligned} J_{\text{TA}}(\mathcal{O}) &= J(\hat{p}_N) \\ &+ \int_{y\in\mathcal{S}} \int_{\tilde{y}\in\tilde{\mathcal{S}}} (\nabla J(\hat{p}_N))(\tilde{y}) \frac{dp(y,\tilde{y})}{dq(y)dr(\tilde{y})} dq(y)dr(\tilde{y}) \\ &- \int_{y\in\mathcal{S}} (\nabla J(\hat{p}_N))(y) d\hat{p}_N(y). \end{aligned} \tag{19}$$

Using importance sampling, we can discretize the expression in (19) as

$$\begin{aligned} J_{\text{TAD}}(\mathcal{O}) &= J(\hat{p}_N) \\ &+ \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\partial J}{\partial p_{\tilde{y}_j}}(\hat{p}_N) \frac{dp(y_i,\tilde{y}_j)}{dq(y_i)dr(\tilde{y}_j)} \\ &- \frac{1}{m} \sum_{j=1}^{m} \frac{\partial J}{\partial \tilde{y}_j}(\hat{p}_N), \end{aligned} \tag{20}$$

where we use numerical differentiation to approximate the gradient ∇J. To simplify the notation, we define

$$\alpha \in \mathbb{R}^{n\times m} : \ \alpha_{ij} = \frac{dp(y_i,\tilde{y}_j)}{dq(y_i)dr(\tilde{y}_j)} \tag{21}$$

to be our new decision variable. Our objective in (13) can then be written as

$$\min_{\alpha} \|\theta_0 - J_{\text{TAD}}(\mathcal{O})\|^2. \tag{22}$$

We discretize the budget constraint (11b) as

$$\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} c(y_i,\tilde{y}_j)\alpha_{ij} \leq \frac{b}{N}. \tag{23}$$

[1] A sample may occur multiple times in the sequence.

To discretize the constraint in (11c), we first utilize the same trick as before and rewrite it as

$$\int_{y \in \mathcal{S}} \frac{dp(y, \tilde{y})}{dq(y) dr(\tilde{y})} dr(\tilde{y}) = \frac{d\hat{p}_N(y)}{dq(y)}. \tag{24}$$

We discretize this as

$$\frac{1}{m} \sum_{j=1}^{m} \frac{dp(y, \tilde{y}_j)}{dq(y_i) dr(\tilde{y}_j)} = \frac{d\hat{p}_N(y)}{dq(y_i)}. \tag{25}$$

Again, since we cannot sample from $p(y, \tilde{y})$, we simply say that the above relation must hold for all values of y, i.e.,

$$\frac{1}{m} \sum_{j=1}^{m} \underbrace{\frac{dp(y_i, \tilde{y}_j)}{dq(y_i) dr(\tilde{y}_j)}}_{\alpha_{ij}} = \frac{d\hat{p}_N(y_i)}{dq(y_i)}, \quad \forall i. \tag{26}$$

All constraints are now written in terms of our new decision variables α_{ij}.

C. Importance Sampling: Different Approaches

Consider the case when $\tilde{\mathcal{S}} = \mathcal{S}$, and $dq = dr = \hat{p}_N$. This means that we we will work directly with the observed samples, and the constraint in (26) then simplifies to

$$\frac{1}{m} \sum_{j=1}^{m} \frac{dp(y_i, \tilde{y}_j)}{dq(y_i) dr(\tilde{y}_j)} = \frac{d\hat{p}_N}{dq}(y_i) = \frac{d\hat{p}_N}{d\hat{p}_N}(y_i) = 1 \quad \forall i. \tag{27}$$

We note that this case is very similar to a discrete setting in the sense that we are limiting the teacher to change the observations into values that have already been seen or encountered. This approach is similar to other resampling techniques and can be viewed as a way of re-weighting the samples to change their importance for the estimation.

We can also sample from dq and dr independently, with $dq \neq dr$. Using this approach, we can impose some prior knowledge on dr, either by defining it to be the true distribution, or some distribution that will yield a more accurate estimate of the parameter we are interested in. However, using this approach, we would not be working directly with the empirical distribution, which means that we would have to perform an interpolation step to figure out how to best change the actual observations. We would also have to use a density estimation technique to enforce the constraint in (11c).

D. Modifying the Sequence

Next we describe how the teacher modifies the sequence based on the α obtained from solving (22) w.r.t. the constraints (23) and (26). Recall the definition of α in (21), and that $dp(y, \tilde{y})$ denotes the amount of probability mass transferred from y to $\tilde{y}$. Then, by applying Bayes' theorem, we compute the conditional probability mass function (pmf) as

$$p(\tilde{y} \mid y) = \alpha \hat{p}_N(\tilde{y}), \tag{28}$$

which will give us the probability of changing an observation y into $\tilde{y}$.

The teacher's intervention procedure is then as follows. For each sample in $y_i \in \mathcal{O}$, the teacher modifies it according to the pmf in (28). That is,

$$y_i \ \rightarrow \ \tilde{y}_i, \ \tilde{y}_i \sim p(\tilde{y} \mid y_i). \tag{29}$$

To ensure that the intervention budget is not exceeded, the teacher generates M_s new sequences. For each new generated sequence that satisfy the budget constraint (i.e., for which the number of corrections is less than or equal to b), an updated estimate is computed. Out of these sequences, the one yielding the lowest estimation error is then chosen to be the optimal one. Should the teacher fail to find a sequence that both improves the estimate and satisfies the budget constraint, it will keep the original sequence.

Naturally, the teacher may follow different intervention policies. Alternative approaches may include changing one sample at a time and then re-solve for a new α following each update. This policy is similar to receding horizon control strategies where we may interpret the budget to be the horizon [28].

Another possible strategy would be for the teacher to always make the change with the highest probability. This would make for a greedy approach [29].

IV. Numerical results

In this section, we evaluate our framework in three different settings; two theoretical and one applied. For simplicity, we consider $\mathcal{Y} \in \mathbb{R}$ and $\theta_0 \in \mathbb{R}$ in all settings. We evaluate the performance in terms of the absolute error, i.e., $e = |\theta_0 - \tilde{\theta}|$.

A. Variance Estimation of a Gaussian Distribution

In the first experiment, we use the framework to estimate the variance of a Gaussian distribution. We consider the observations to be sampled from the distribution $\mathcal{N}(0, 1)$, so $\theta_0 = \sigma^2 = 1$.

We perform the estimation on three sample sizes, $N = \{10, 20, 50\}$, subject to four different intervention budgets, $b = \{0, 1, 5, 10\}$. In this example, we use a uniform transportation cost, i.e., $c(y, \tilde{y}) = \mathbb{1}\mathbb{1}^T - I$. This means that all changes made are equally expensive. For each sample size and budget, we perform the experiment 100 times and compute the average absolute error. For all configurations, we use $M_s = 1000$. The results are shown in Figure 2. As expected, the plot shows a decrease in the estimation error as the sample size increase. It also shows that the error is further decreased as the budget increases.

B. Scale Estimation of a Weibull Distribution

Next we apply the framework to estimate the scale parameter of a Weibull distribution. The probability density function is given by

$$f(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-(x-\lambda)^k}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{30}$$

where $k > 0$ is called the shape parameter, and $\lambda > 0$ the scale parameter. In this example, we will consider the

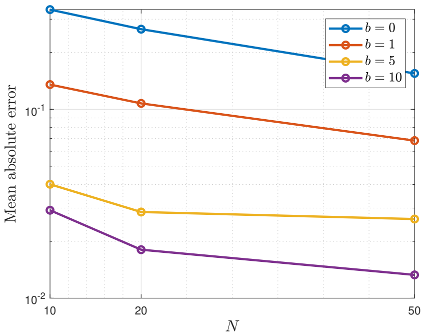

Fig. 2. The absolute estimation error averaged over 100 Monte Carlo simulations for increasing sample sizes and budgets. Note that $b = 0$ corresponds to the case with no teacher intervention.

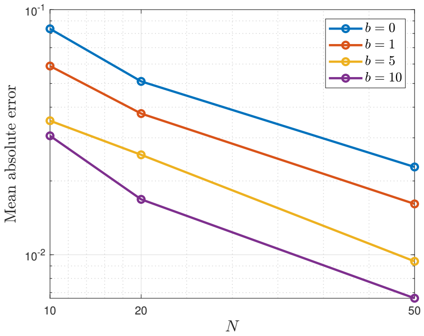

Fig. 3. The absolute estimation error averaged over 100 Monte Carlo simulations for increasing sample sizes and budgets. Note that $b = 0$ corresponds to the case with no teacher intervention.

estimation of λ of a Weibull distribution with $\theta_0 = \lambda_0 = 2$ and $k = 8$.

There are different approaches available for estimating λ, see e.g. [30]. In this experiment, we use the Bayesian two-stage approach derived in [31]. We use the proportional cost

$$c(y, \tilde{y}) = 10 \times \lceil |\tilde{y} - y| \rceil, \tag{31}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. As in the previous example, we run the estimation process on the sample sizes $N = \{10, 20, 50\}$ and for the intervention budgets $b = \{0, 1, 5, 10\}$. We use $M_s = 2000$ for all configurations. The averaged estimation errors are shown in Figure 3. The results are similar to what we observed in the previous experiment, with an improved estimation error for increasing sample sizes and budgets.

C. Reward Estimation in Inverse Reinforcement Learning

As a final example, we apply our framework to update a robot's reward function in an inverse reinforcement learning setting. Recent work on learning from human interaction shows how physical corrections made by a human (e.g. in the form of applied torque) can improve a robot's learning process [32]. Inspired by their problem setup, we apply our framework in a similar setting.

Consider a robot arm being tasked with moving a coffee cup from one side of a table to the other. To learn the task, the robot gets to observe a set of N trajectories, $\{\xi_1, \dots, \xi_N\}$, demonstrated by a human. Figure 4 illustrates some examples of trajectories demonstrated on a robotic arm with seven degrees of freedom implemented in PyBullet. Each trajectory is associated with a total feature count for each feature $i \in [n]$

$$\Phi_i(\xi) = \sum_{x \in \xi} \phi_i(x), \tag{32}$$

where $\phi_i(x)$ is the local feature value in a point x along the trajectory ξ. A high feature value corresponds to a good position in space. The features represent different subgoals in performing the task, such as "stay nearby the top of the table" and "avoid the laptop". Based on these features, the robot learns a reward function

$$R = \Theta^T \Phi = \theta_1 \Phi_1 + \dots \theta_n \Phi_n, \tag{33}$$

where the weights Θ represent the importance of each feature to the human.

Assume now that the robot has learned a reward function based on the following observations collected over $N = 5$ trajectories with $n = 3$ features

$$\begin{cases} \mathbf{\Phi_1} = \{\Phi_1(\xi_i)\}_{i=1}^{5} = \{100,\ 75,\ 50,\ 20,\ 5\} \\ \mathbf{\Phi_2} = \{\Phi_2(\xi_i)\}_{i=1}^{5} = \{90,\ 200,\ 10,\ 2,\ 30\} \\ \mathbf{\Phi_3} = \{\Phi_3(\xi_i)\}_{i=1}^{5} = \{50,\ 20,\ 3,\ 5,\ 10\} \end{cases} . \tag{34}$$

An expert may then apply our framework to improve the robot's learned θ_i's, by modifying the sets $\mathbf{\Phi}_i$ in (34) into $\tilde{\mathbf{\Phi}}_i$. As estimator, we consider a slightly modified version of the weight update in [32]:

$$\tilde{\theta}_i = \hat{\theta}_i + \beta \left(\sum_{\Phi \in \mathbf{\Phi}} \Phi - \sum_{\tilde{\Phi} \in \tilde{\mathbf{\Phi}}} \tilde{\Phi} \right), \tag{35}$$

where $\beta < 0$ is a step/scaling parameter. Here, we consider $\beta = -0.001$. Note how the feature weights are updated based on the direction of change of the feature values between the original and the modified trajectories. If the altered corrections pass further away from, say, the laptop, the θ_i corresponding to the distance-to-laptop feature will increase.

For this experiment we used $c(y, \tilde{y}) = \mathbb{1}\mathbb{1}^T - I$, $b = 1$, and $M_s = 1000$. The corrected feature values together with their corresponding updated weight estimate are shown in Table I. For reference, we also we also present the true weights and the initial estimates in the same table. The results show that the updated estimates are closer to the true values, compared to the initial estimates.

V. CONCLUSION AND FUTURE WORK

In this work, we presented a generalized formulation of the correctional learning framework using optimal transport. We demonstrated that by expressing the correctional learning problem as an optimization program in terms of distribution functions, we obtain a more general and flexible framework better suited for estimation of more complex characteristics. We successfully applied the framework on three estimation processes; for the variance estimation of a Gaussian, the scale

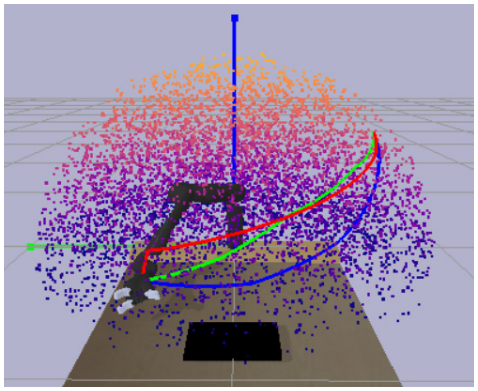

Fig. 4. The robot observes $N = 3$ trajectories with different feature values. The expert may alter some of them to the one that is closer to its preferences. For example, if the robot should avoid the laptop, the expert may change the blue trajectory into the red one, to reflect this.

TABLE I

THE CORRECTED FEATURE VALUES AND THEIR CORRESPONDING ESTIMATES.

True weight	Old est.	New est.	Corrected feature values
$\theta_1 = 0.1$	$\hat{\theta}_1 = 0.5$	$\tilde{\theta}_1 = 0.05$	$\tilde{\mathbf{\Phi}}_1 = \{100,\ 75,\ 50,\ 20,\ \mathbf{50}\}$
$\theta_2 = 1$	$\hat{\theta}_2 = 0.5$	$\tilde{\theta}_2 = 1.1$	$\tilde{\mathbf{\Phi}}_2 = \{\mathbf{30},\ 200,\ 10,\ 2,\ 30\}$
$\theta_3 = 0.8$	$\hat{\theta}_1 = 0.5$	$\tilde{\theta}_3 = 0.8$	$\tilde{\mathbf{\Phi}}_3 = \{\mathbf{20},\ 20,\ 3,\ 5,\ 10\}$

estimation of a Weibull, and, finally, in an inverse reinforcement setting where we improved a robot's reward function. This novel optimal-transport formulation opens up for several interesting extensions, including using correctional learning for differential privacy, considering time-series data, and for balancing biased or skewed learning datasets. Addressing the curse of dimensionality of the discretized problem will be an important step along the way.

ACKNOWLEDGEMENT

The authors would like to thank Isabel Haasler for valuable discussions on Optimal Transport theory.

REFERENCES

[1] L. Ljung, *System Identification: Theory for the User, 2nd Ed.* Prentice Hall, 1999.

[2] K. Åström and P. Eykhoff, "System identification—a survey," *Automatica*, vol. 7, pp. 123–162, 3 1971.

[3] F. Liu and P. Demosthenes, "Real-world data: A brief review of the methods, applications, challenges and opportunities," *BMC Medical Research Methodology*, vol. 22, p. 287, 11 2022.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91.

[5] I. Lourenço, R. Mattila, C. R. Rojas, and B. Wahlberg, "Cooperative system identification via correctional learning," *19th IFAC Symposium on System Identification*, vol. 54, no. 7, pp. 19–24, 2021.

[6] I. Lourenço, R. Winqvist, C. R. Rojas, and B. Wahlberg, "A teacher-student markov decision process-based framework for online correctional learning," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 3456–3461.

[7] C. Villani, *Topics in Optimal Transport.* American Mathematical Society, 2003.

[8] A. Galichon, *Optimal Transport Methods in Economics.* Princeton University Press, 2016.

[9] I. Haasler, J. Karlsson, and A. Ringh, "Control and estimation of ensembles via structured optimal transport," *IEEE Control Systems Magazine*, vol. 41, pp. 50–69, 8 2021.

[10] Y. Chen, J. Karlsson, and A. Ringh, "Optimal transport for applications in control and estimation," *IEEE Control Systems Magazine*, vol. 41, pp. 28–33, 8 2021.

[11] J. Digne, D. Cohen-Steiner, P. Alliez, F. de Goes, and M. Desbrun, "Feature-preserving surface reconstruction and simplification from defect-laden point sets," *Journal of Mathematical Imaging and Vision*, vol. 48, pp. 369–382, 2 2014.

[12] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a Wasserstein Loss," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[13] A. Gramfort, G. Peyré, and M. Cuturi, "Fast optimal transport averaging of neuroimaging data," in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds. Cham: Springer International Publishing, 2015, pp. 261–272.

[14] A. Tompkins, R. Senanayake, and F. Ramos, "Online domain adaptation for occupancy mapping," in *Robotics: Science and Systems*, 07 2020.

[15] J. Tan, R. Senanayake, and F. Ramos, "Renaissance robot: Optimal transport policy fusion for learning diverse skills," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7052–7059.

[16] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[17] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys*, vol. 50, no. 2, apr 2017.

[18] A. Najar and M. Chetouani, "Reinforcement learning with human advice: A survey," *Frontiers in Robotics and AI*, vol. 8, 6 2021.

[19] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*, D. Sleeman and P. Edwards, Eds. San Francisco (CA): Morgan Kaufmann, 1992, pp. 249–256.

[20] C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. Yu, *Active learning: A survey.* CRC Press, Jan. 2014, pp. 571–605, publisher Copyright: © 2015 by Taylor & Francis Group, LLC.

[21] H. Hjalmarsson, "System identification of complex and structured systems," *European Journal of Control*, vol. 15, no. 3, pp. 275–310, 2009.

[22] L. Pronzato, "Optimal experimental design and some related control problems," *Automatica*, vol. 44, no. 2, pp. 303–325, 2008.

[23] Y. Chen and J. Karlsson, "State tracking of linear ensembles via optimal mass transport," *IEEE Control Systems Letters*, vol. 2, no. 2, pp. 260–265, 2018.

[24] A. Taghvaei and P. G. Mehta, "An optimal transport formulation of the ensemble kalman filter," *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 3052–3067, 2021.

[25] Y. Yang, L. Nurbekyan, E. Negrini, R. Martin, and M. Pasha, "Optimal transport for parameter identification of chaotic dynamics via invariant measures," *SIAM Journal on Applied Dynamical Systems*, vol. 22, no. 1, pp. 269–310, 2023.

[26] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, pp. 355–607, 2019.

[27] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective.* Elsevier Science, 2015.

[28] F. Borrelli, A. Bemporad, and M. Morari, *Predictive Control for Linear and Hybrid Systems*, 1st ed. USA: Cambridge University Press, 2017.

[29] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[30] M. Teimouri, S. Hoseini, and S. Nadarajah, "Comparison of estimation methods for the weibull distribution," *Statistics: A Journal of and Applied Statistics*, 03 2011.

[31] B. Lakshminarayanan and C. R. Rojas, "A statistical decision-theoretical perspective on the two-stage approach to parameter estimation," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 5369–5374.

[32] A. Bajcsy, D. P. Losey, M. K. O'malley, and A. D. Dragan, "Learning robot objectives from physical human interaction," in *Conference on Robot Learning.* PMLR, 2017, pp. 217–226.