A TEACHER-STUDENT FRAMEWORK FOR ONLINE CORRECTIONAL LEARNING

Inês Lourenço, Rebecka Winqvist, Cristian R. Rojas, Bo Wahlberg

KTH Royal Institute of Technology, Division of Decision and Control Systems, Stockholm, Sweden. 
{ineslo, rebwin, crro, bo}@kth.se

ABSTRACT

A classical learning setting is one in which a student collects data, or observations, about a system, and estimates a certain quantity of interest about it. Correctional learning is a type of cooperative teacher-student framework where a teacher, who has knowledge about the system, has the possibility to observe and alter (correct) the observations received by the student in order to improve its estimation. In this paper, we show that the variance of the estimate of the student is reduced with the help of the teacher. We further formulate the online problem – where the teacher has to decide at each time instant whether or not to change the observations – as a Markov decision process, from which the optimal policy is derived using dynamic programming. We validate the framework in numerical experiments, and compare the optimal online policy with the one from the batch setting.

Index Terms— Correctional learning, teacher-student framework, assisted learning, Markov decision processes

1. INTRODUCTION

With the rapid growth of smart systems and IoT, we are able to collect amounts of data like never before. These data may constitute anything from medical images captured by camera sensors to distance measures from e.g. lidars and radars. Using this data, agents can learn to perform tasks such as cancer detection and prognosis [1][2], and autonomous driving [3].

In the Oxford dictionary [4], the term learning is defined as the “acquisition of knowledge or skills through study, experience, or being taught”. In this work we consider a combination of the latter two; “experience” by using dynamic programming to train a teacher, and “being taught” by letting the trained teacher transfer its knowledge to a student agent. Setups in which there are aiding expert agents are commonly denoted cooperative learning problems.

Cooperative problems play an important role in our lives, as we encounter them daily. Indeed, most tasks we perform require some sort of collaboration; acquiring a new skill, such as learning how to drive, social learning, search and rescue operations, and much more. Solving these tasks is, however, not trivial. In the literature, two famous paradigms of cooperative learning that use a teacher-student framework are learning from demonstration [5] and imitation learning [6], in which the role of the teacher is to accelerate the learning of the student by showing it the optimal policy. In the field of image classification, teacher networks have been trained to learn unlabeled data and teach a student network [7].

In this work, we consider a different teacher-student paradigm. We consider that the way the teacher can assist the student is by intercepting, and altering, the data collected by the student. This novel approach, denoted correctional learning, was recently proposed in [8] in a cooperative system identification scenario in which batches of data were altered by the teacher to maximize what the student learns about the underlying system.

In this paper, we extend the correctional learning framework to a more realistic (online) setting, where the teacher has to decide, at each time step, whether or not to alter the corresponding observation. This online teacher-student correctional learning framework can be used in a variety of settings: helping an agent learning a policy, the parameters of the system, or the state of the world. These problems are called reinforcement learning, system identification, and filtering, respectively. To the best of the authors’ knowledge, little work besides [9, Section XII] and [10] has been done to incorporate prior, or external, information on decision-making. In the former, the entropy of a message was used to measure and correct the errors in the transmission of data between a source and a receiver. In the latter, which is the area of controlled sensing, a decision-maker can choose at each time instant which sensor to use to provide the next measurement.

The main question posed in the work is then:

How can a teacher alter, at each time instant, the data received by a student, in order to assist its learning process?

In summary, the main contributions of this paper are:

- Computation of a theoretical bound for how much the teacher can improve the estimation of the student in the case of discrete systems;
- Formulation of a Markov decision process for the correctional learning framework performed in an online setting;

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), and the Swedish Research Council Research Environment NewLEADS under contract 2016-06079.
• Demonstration of the results in numerical experiments, in particular the optimal policy of the teacher obtained using dynamic programming;
• Comparison of the online and batch correctional learning frameworks.

The rest of the paper is organised as follows. In Section 2, the correctional learning problem is formulated. In Section 3 we derive bounds for how much the teacher can help the student, and, in Section 4, the proposed algorithm for performing online correctional learning is derived. Finally, Section 5 validates the presented methods in numerical experiments and Section 6 concludes the paper.

2. PROBLEM FORMULATION

In this section, we define our notation and formally introduce the correctional learning problem: a teacher, who has knowledge about a system of interest, aims to help a student in its learning process by altering the data it receives.

2.1. Notation

All vectors are column vectors and inequalities between vectors are considered element-wise. The i-th element of a vector \( v \) is \( [v]_i \). A generic probability density (or mass) function is denoted as \( p(\cdot) \) and \( P(\cdot) \) is the probability of event \( \cdot \). The vector of ones is denoted as \( \mathbb{1} \), the set of natural numbers including zero as \( \mathbb{N}_0 \), and the indicator function as \( I(\cdot) \). The \( \ell_1 \) norm of a vector \( v \) is denoted \( ||v||_1 \) and \( \text{card}(S) \) is the cardinality of set \( S \).

2.2. Introduction to Correctional Learning

An agent (student) is sequentially collecting information from a source in the form of observations \( D = \{ y_k \}_{k=1}^N \), \( y_k \in \mathcal{Y} \), throughout \( N \) time steps, and estimating characteristics of interest \( \theta \) about that system. An expert agent (teacher) has the ability to intercept, and alter, the observations collected by the student to \( \tilde{D} = \{ \tilde{y}_k \}_{k=1}^N \). A schematic representation is shown in Figure 1. The goal of the teacher is to assist the estimation process of the student. Improving the estimation might mean obtaining an altered estimate \( \tilde{\theta} \) closer to the true estimate \( ||\tilde{\theta} - \theta|| \leq ||\hat{\theta} - \theta|| \), or which converges to the true one faster \( \text{var}(\tilde{\theta}) < \text{var}(\hat{\theta}) \).

Additionally, access to the expert teacher might be expensive, or dangerous for privacy concerns since the more observations it changes, the more likely it is to be discovered. Therefore, correctional learning includes the constraint

\[
B(D, \tilde{D}) \leq c_b, \tag{1}
\]

where \( B \) is a distance measure between two sets and which represents the budget, \( c_b \), that the teacher has on how many, or by how much, it can interfere with the observations obtained from the system. If the observations are discrete, \( B \) can be defined as the \( l_1 \)-norm \( \frac{1}{N} \sum_{k=1}^N ||D - \tilde{D}||_1 \leq c_b \).

2.3. Batch Correctional Learning

In batch correctional learning multiple observations are intercepted at once by the expert. This can be, for example, the case in a communication channel, where multiple bits can be delayed on the way from the source to the receiver. In [8], the batch problem was solved by minimizing the distance between the empirical estimate computed by the student and the true one, according to the following optimization problem:

\[
\min_{\tilde{D}} V(\theta_0, \tilde{\theta})
\]

\[
s.t. \; \tilde{y}_k \in \mathcal{Y}, \text{ for all } \tilde{y}_k \in \tilde{D},
\]

\[
B(D, \tilde{D}) \leq c_b, \tag{2}
\]

where \( V(\theta_0, \tilde{\theta}) \) is a distance measure between estimates, and \( B \) is the distance measure from \( \{1\} \) between two sets. In [8] it was shown that the resulting set \( \tilde{D} \) of corrected observations was the optimal one for the case of binomial systems, and by how much the variance of the corrected estimate is decreased compared to the original one. In this paper, we extend this for multinomial systems, and in an online setting.

3. CORRECTIONAL LEARNING BOUNDS FOR DISCRETE SYSTEMS

In this section we theoretically study by how much the teacher can effectively help the student, for the case when the observations are discrete.

The following theorem compares the estimates of the mean values of two sequences of observations – the original, \( Y/N \), and the corrected, \( \tilde{Y} \). Knowing how much the variance of the corrected estimate is decreased is a measure of how much the teacher can help reducing the error of the estimation of the student.

**Theorem 1** (Variance decrease of the altered estimate). Let \( X_1, \ldots, X_N \) be i.i.d. random variables in \( \{0, 1, \ldots, M\} \) with mean \( \mu \), and \( Y = X_1 + \ldots + X_N \). Let \( B \in \{0, 1, \ldots, N\} \), and

\[
\tilde{Y} = \arg \min_{\{Z \in \{0,\ldots,N\} : |Y - Z| \leq B\}} |Z - N\mu|. \tag{3}
\]
Then, \( \text{var} \frac{\hat{Y}}{N} \leq M^2 \exp \left[ -\frac{2B^2}{NM^2} \right] \). Let us further assume that \( X_i \sim \text{Unif}([0, \ldots, M]) \) (this assumption is not crucial but provides a special case that is easier to understand). Then,

\[
\frac{\text{var} \frac{\hat{Y}}{N}}{\text{var} \frac{Y}{N}} \leq \frac{6M}{5M + 1} \exp \left[ -\frac{2B^2}{NM^2} \right].
\]

(4)

The proof of Theorem 1 is supplemented in Appendix A. This theorem derives an upper bound for the decrease in variance of the estimate computed by the student due to the help of the teacher, according to its budget \( B \). We have that:

- The teacher’s ability to improve the learning of the student increases as its budget increases;
- For a given budget, the improvement becomes less important as \( N \) grows. This is reasonable, since the average deviation of \( Y/N \) around \( \mu \) is of order \( O(1/\sqrt{N}) \), while the improvement due to the teacher can be at most \( B/N \);
- For a fixed budget and \( N \), the improvement degrades as \( M \) increases, since the variance of \( Y/N \) increases with \( M \), which makes it increasingly harder for a teacher to compensate for “bad” samples.

In the next section we propose a framework for how the teacher can alter the observations in an online setting.

4. ONLINE CORRECTIONAL LEARNING

After computing by how much the teacher can improve the estimation process of the student in a discrete setting, we propose a framework for implementing this online. Unlike the batch case, the online setting is a more realistic scenario in which the observations are obtained sequentially and the expert has to make, at each time instant, the decision of whether not to change the current observation. Consider, for example, the case when the student is updating its localization based on consecutive GPS measurements.

4.1. Formulation of the Markov Decision Process

The following is the second main result of our paper, which is the formulation of the Markov decision process (MDP) for the online correctional learning problem when sampling from a multinomial distribution \( y_k \in \{0, 1, \ldots, M\} = \mathcal{Y} \) with parameter \( \theta_0 \):

- **States**: \( s = (x_{1:k}, b_k, y_k) \)
- **Actions**: \( a = \{\text{keep } y_k, \text{change to } \tilde{y}_k\} \)
- **Time-horizon**: \( N \)
- **Reward function**: \( M - ||\hat{\theta}_N - \theta_0||_1 \)
- **Constraint**: number of times the action “change to \( \tilde{y}_k \)” is taken \( \leq c_b \)
- **Transition probabilities**: see (5)

In more details:

- **States**: The states of the MDP are tuples containing: \( x_{1:k} \) – a \( M \times 1 \) vector with the number times each observation has been seen until time \( k \); \( b \in \mathbb{N}_0 \) – the current budget left to use at time \( k \); \( y_k \) – the observation received at time \( k \). The number of states is finite and upper bounded by \( \text{card}(S) \leq NM^{c_B+1} \). However, the constraint \( \sum_{k=1}^M |x_k| \leq N \) renders many of these states invalid, which results in a much smaller and tight upper bound \( \text{card}(S) \leq \text{card}(x)c_BN \). Here, \( \text{card}(x) \) can be computed using multiset coefficients as \( \text{card}(x) = \sum_{n=1}^N \binom{M}{n} = \sum_{n=1}^N (M(M+1)\ldots(M+n-1))/n! \), which are the \( N \)-permutations of \( M \) with repetitions and which satisfy the previous constraint.
- **Terminal states**: Those where all the observations have been received, that is, \( \sum_{k=1}^M |x_k| = N \).
- **Actions**: The possible actions are to keep the last observation \( y_k \) or change it to a certain value \( \tilde{y}_k \in \mathcal{Y} \). The number of actions is \( \text{card}(A) = M \).
- **Reward function**: The reward is zero in all states except in the terminal states, where it is equal to the maximum error minus the error of the final altered estimate.
- **Transition probabilities**: If the action is “keep \( y_k \)”, the next state depends, with probability \( p(y_{k+1}) \), on the next observation \( y_{k+1} \) received. The value of the next state is obtained by simply replacing the last value of the previous state \( y_k \) by the new observation received, and adding one to that entry of the vector \( x \). \[
[x'_{y_{k+1}}] = [x]_{y_{k+1}} + 1.
\]

If the action is “change to \( \tilde{y}_k \)”, the next state will have the same probability as in the previous case, where one is added to \( [x]_{y_{k+1}} \). However, it will now have one subtracted from the previous observation in \( [x]_{y_k} \) and one added in \( [x]_{\tilde{y}_k} \) (since \( y_k \) was altered to \( \tilde{y}_k \)), as well as a budget of \( b_{k+1} = b_k - 1 \). We can write them mathematically using a simplified notation as:

\[
\mathbb{P}\{x', b, y_{k+1}) | s = (x, b, y_k), a = \text{“keep } y_k\text{”}\} = p(y_{k+1}),
\]

where \([x']_{y_{k+1}} = +1, \]

\[
\mathbb{P}\{x', b - 1, y_{k+1}) | s = (x, b, y_k), a = \text{“change to } \tilde{y}_k\text{”}\} = p(y_{k+1}) - 1,
\]

where \([x']_{y_{k+1}} = +1, [x']_{y_k} = -1, \text{ and } [x']_{\tilde{y}_k} = +1, \]

and \( \mathbb{P}\{\text{others}\} = 0. \)

(5)

Note that the chosen formulation of the states and actions satisfies the Markovian property.

- **Constraint**: The constraint was enforced by attributing an infinitely negative reward to transitions to states where the budget would be \( b_{k+1} < 0 \).

Remark 1. This framework can also be used when the observations are continuous, by discretizing the observation space and changing the constraint to the total amount of correction

\[
\sum_{k=1}^N |y_k - \tilde{y}_k| \leq c_b.
\]
In this work we considered that an expert agent, a teacher, can observe the observations collected by a student agent from a certain system of interest. We used correctional learning to study how the teacher can alter these observations in order to improve the learning process of the student. We bounded by how much the teacher can help the estimation of the student. We bounded by reducing the variance of its estimate, and derived an optimal policy for the online correctional learning problem represented as the previous MDP can be obtained using dynamic programming.

**Theorem 2.** The optimal policy for the online correctional learning problem represented as the previous MDP can be obtained using dynamic programming.

**Proof.** The resulting policy of the finite MDP using dynamic programming is optimal by construction. 

## 4.2. Solving the MDP using Dynamic Programming

**Theorem 2.** The optimal policy for the online correctional learning problem represented as the previous MDP can be obtained using dynamic programming.

**Proof.** The resulting policy of the finite MDP using dynamic programming is optimal by construction.

## 5. NUMERICAL RESULTS

Let us now demonstrate the framework proposed in numerical experiments, using Python 3.7 and 1.90 GHz CPU.

Figure 2 presents the results of the MDP proposed in Section 4 for performing correctional learning in an online way. The figure shows the error of the estimate of the student with, in blue, the original sequence of observations – that is, without the help of the teacher – and, in orange, the corrected sequence. The estimates $\hat{\theta}$ are computed as the mean of the observations, $[\hat{\theta}]_i = \frac{\sum_{k=1}^{M} I(y_k = i)}{N}$, and we consider that an observation is randomly sampled $N = 5$ times from a multinomial distribution with parameter $\theta_0 = [0.4; 0.3; 0.3]$, over 50 experiments. Unlike in the binomial case, where we can compute a closed form solution for the minimum attainable error of experiment (see Appendix B), in the multinomial this error is given by the batch error, which we computed using (8) from B and with the $l1$-norm in the objective function for consistency. The minimum error, independent of $\theta$ and $c_b$, can, however, be computed as

$$e_{\text{min}}(N, \theta_0) = \frac{\|\theta_0 - \frac{\theta_0 N}{N}\|_1 = 0.2,}$$

which is achieved by $\theta^* = [0.4; 0.4; 0.2]$ or [0.4; 0.2; 0.4].

Intuitively, one expects the teacher’s optimal policy to be to delay spending its budget as much as possible. In the binomial case we prove that the online policy learned, $\mu^* = \begin{cases} a_k = \text{keep } y_k, & \text{if } [x] y_k \leq \frac{\theta_0}{N}, N, \\ a_k = \text{change to } \tilde{y}_k = 1 - y_k, & \text{otherwise}, \end{cases}$

is optimal since it coincides with the policy computed in the batch case, as can be seen in Appendix B. In the multinomial case, both differ only in a limited amount of scenarios, when a little expected sample that has a small reward is obtained. Note that in experiment 23 the altered estimate $\tilde{\theta}$ is even worse than the original one, $\hat{\theta}$. The teacher chose to alter the fourth observation of [1, 2, 0, 2, 2] to a 0 since the expected value was larger (receiving a 1 or a 2 at $k = 5$ had a big probability and maximum reward), but the less likely observation, 0, was received instead.

Figure 3 shows that, as expected, the variance of the estimate decreases as the number of observations increases. However, as the budget of the teacher increases, the variance is further decreased. This results illustrates the conclusions from Theorem 1.

## 6. CONCLUSION

In this work we considered that an expert agent, a teacher, can observe the observations collected by a student agent from a certain system of interest. We used correctional learning to study how the teacher can alter these observations in order to improve the learning process of the student. We bounded by how much the teacher can help the estimation of the student by reducing the variance of its estimate, and derived an optimal policy to perform correctional learning in an online setting using dynamic programming.

In the future, we would like to apply correctional learning to the settings of filtering and, especially, reinforcement learning. Furthermore, the correctional learning problem can easily be translated to an adversarial setting, for which the student needs to estimate which observations were corrupted.
7. REFERENCES

[1] Li Shen, Laurie R. Margolies, Joseph Rothstein, Eugene Fluder, Russell B. McBride, and Weiva Sieh. “Deep learning to improve breast cancer detection on screening mammography,” Scientific Reports, vol. 9, 2019.

[2] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fiotiadis, “Machine learning applications in cancer prognosis and prediction,” Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun, “Are we ready for autonomous driving? the kit vision benchmark suite,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[4] “learn, v.,” in OED Online. Oxford University Press, Oct. 2021. Available at https://www.oed.com/view/Entry/106716?rskey=kOj1Kj&result=1&isAdvanced=false#eid (accessed Oct 06, 2021).

[5] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning, “A survey of robot learning from demonstration,” Robotics and Autonomous Systems, vol. 57, no. 5, pp. 469–483, 2009.

[6] Ahmed Hussein, Mohamed Gaber, Eyad Elyan, and Chrisina Jayne, “Imitation learning: A survey of learning methods,” ACM Computing Surveys, vol. 50, 04 2017.

[7] Henrik L. Gjestang, Steven A. Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A. Riegler, “A self-learning teacher-student framework for gastrointestinal image classification,” in 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), 2021, pp. 539–544.

[8] Inês Lourenço, Robert Mattila, Cristian R. Rojas, and Bo Wahlberg, “Cooperative system identification via correctional learning,” 19th IFAC Symposium on System Identification, vol. 54, no. 7, pp. 19–24, 2021, URL: https://arxiv.org/pdf/2012.05161.pdf.

[9] Claude E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[10] Vikram Krishnamurthy, Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing, Cambridge University Press, 2016.

[11] Martin L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.

A. BOUNDING THE DECREASE IN VARIANCE

Let us first present two known lemmas used to solve Theorem 1.

Lemma 3. Let $X$ be a nonnegative random variable (R.V.). Then,

$$E[X] = \int_{0}^{\infty} P(X \geq \tau)d\tau$$

Proof. Let $F$ be the CDF of $X$, i.e., $F(\tau) = P(X \leq \tau)$. Then, by integration by parts

$$\begin{align*}
\mathbb{E}[X] &= \int_{0}^{\infty} \tau dF(\tau) = -\left[ \int_{0}^{\infty} \tau d[1 - F(\tau)] \right]_{= P(X > \tau)}^{= P(X \geq \tau)} \\
&= -\tau[1 - F(\tau)]_{\tau=0}^{\infty} + \int_{0}^{\infty} \frac{d[1 - F(\tau)]}{= P(X > \tau)} \\
&= \int_{0}^{\infty} P(X > \tau)d\tau
\end{align*}$$

Note that $P(X = \tau) > 0$ for at most a ... number of values of $\tau$, so

$$\int_{0}^{\infty} P(X = \tau)d\tau = 0$$

and

$$\mathbb{E}[X] = \int_{0}^{\infty} P(X \geq \tau)d\tau.$$ 

Lemma 4. Let $X$ be a R.V., and $\lambda$ a constant. Then,

$$\mathbb{E}[(X - \lambda)^2] = \int_{0}^{\infty} P(|X - \lambda| \geq \sqrt{\tau})d\tau.$$ 

Proof. Use Lemma 3 with $X$ replaced by $(X - \lambda)^2$. 

Let us now restate Theorem 1 and prove it in three parts.

Theorem 1. Let $X_1, \ldots, X_N$ be i.i.d. R.V.’s in $\{0, 1, \ldots, M\}$ with mean $\mu$, and $Y = X_1 + \ldots + X_N$. Let $B \in \{0, 1, \ldots, N\}$, and

$$\tilde{Y} = \arg\min_{\{Z \in \{0, \ldots, N\}:Y - Z| \leq B\}} |Z - N\mu|.$$ 

Then, $\text{var} [\tilde{Y}/N] \leq M^2 \exp \left[ -\frac{2N}{M} \right]$. Let us further assume that $X_i \sim \text{Unif}([0, \ldots, M])$ (this assumption is not crucial but provides a special case that is easier to understand). Then,

$$\frac{\text{var}[\tilde{Y}/N]}{\text{var}[Y/N]} \leq \frac{6M}{5M + 1} \exp \left[ -\frac{2B^2}{NM^2} \right].$$
Finally, we conclude that
\[
\frac{\text{var}[\tilde{Y}/N]}{\text{var}[Y/N]} \leq \frac{M^2 \exp \left[ -\frac{2B^2}{NM^2} \right]}{\frac{2M^2 + M}{6}} = \frac{6M}{5M + 1} \exp \left[ -\frac{2B^2}{NM^2} \right].
\]

**B. RESULTS FOR A BINOMIAL DISTRIBUTION**

We exemplify now the case in which the observations are randomly sampled from a binomial distribution with parameter \( \theta_0 = 0.5 \), for comparison with the batch case from \([8]\). We define \( N = 10 \) and repeat the algorithm over 50 experiments.

Figure 4 shows the results of the MDP proposed in Section 4 for performing online correction learning now for a binomial distribution instead of a multinomial as in Section 5.

A main difference in this case is that we can prove that the policy learned by the teacher is optimal, as stated in Theorem 2 since the error of the estimate of the student with the corrected sequence of observations (in orange) – that is, without the help of the teacher – is always exactly the minimum error attainable for the original sequence (in green). The exact expression for this error is:

\[
e(N, \theta_0, c_b, \tilde{\theta}) = \max \left\{ \|\theta_0 - \tilde{\theta}\|_1 - \frac{2c_b}{N}, e_{\text{min}} \right\}.
\]

where the second term is the \( e_{\text{min}} \) from \([6]\). This error is now always never larger than the error with the original sequence of observations (in blue).

Here, an analysis of the policy values obtained using the dynamic programming algorithm shows that indeed the teacher chooses the optimal policy, defined in \([7]\), of delaying spending its budget as much as possible and doing so only once too many of one of the outcomes is sampled. In this case switching to the other value does not represent a risk, unlike in the multinomial case where a specific outcome to change to has to be chosen. An example from Figure 4 is the sequence 110111001, where the teacher alters the orange value that was a 1 to a 0 without caring if the next sample is 0 or a 1.
This figure "Online_N.png" is available in "png" format from:

http://arxiv.org/ps/2111.07818v1
This figure "Online_errors.png" is available in "png" format from:

http://arxiv.org/ps/2111.07818v1