KNOW THY STUDENT: INTERACTIVE LEARNING WITH GAUSSIAN PROCESSES

Rose E. Wang, Mike Wu, Noah Goodman
Stanford University
{rewang, wumike, ngoodman}@stanford.edu

ABSTRACT

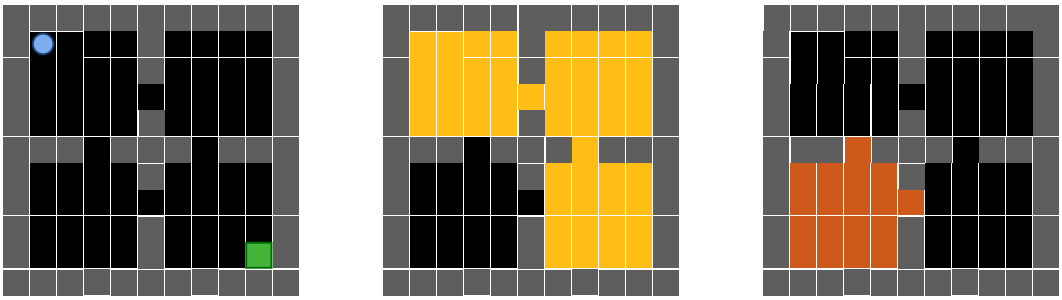
Learning often involves interaction between multiple agents. Human teacher-student settings best illustrate how interactions result in efficient knowledge passing where the teacher constructs a curriculum based on their students' abilities. Prior work in machine teaching studies how the teacher should construct optimal teaching datasets assuming the teacher knows everything about the student. However, in the real world, the teacher doesn't have complete information about the student. The teacher must interact and *diagnose* the student, before teaching. Our work proposes a simple diagnosis algorithm which uses Gaussian processes for inferring student-related information, before constructing a teaching dataset. We apply this to two settings. One is where the student learns from scratch and the teacher must figure out the student's learning algorithm parameters, eg. the regularization parameters in ridge regression or support vector machines. Two is where the student has partially explored the environment and the teacher must figure out the important areas the student has not explored; we study this in the offline reinforcement learning setting where the teacher must provide demonstrations to the student and avoid sending redundant trajectories. Our experiments highlight the importance of diagosing before teaching and demonstrate how students can learn more efficiently with the help of an interactive teacher. We conclude by outlining where diagnosing combined with teaching would be more desirable than passive learning.

1 INTRODUCTION

In natural systems, learning often involves interaction between multiple agents. Teacher-student interactions best illustrate how interactions result in efficient knowledge passing. One way teachers interact with students is through concept tests—short, targeted *tests* administered at the start of a class to help teachers gauge which concepts students have understood (Caceffo et al., 2016; Treagust, 1988; Adams & Wieman, 2011). By first understanding what their students struggle on, teachers can adapt the curriculum to best address the needs of their class and avoid repeating topics their students have already mastered. In other words, teachers construct diagnostic tests to inform them how they should later construct "training sets" for their students.

We take inspiration from this form of teacher-student interactions in reframing how we'd like to efficiently train and test machine learning models. As ML practitioners and researchers, we implicitly interact with our model learners in this way to diagnose the learner. For example, we evaluate our models by analyzing their performance on a test set. The model's test performance can inform us about the model's training dataset, for example data-centric properties like class imbalance (Horn et al., 2018; Rahman & Davis, 2013). This feedback can then suggest remedies for how to retrain our model, for example adding more examples of minority classes to balance the dataset (Hernandez et al., 2013). The collective learning that takes place between a teacher and a student inspires our work to design principled methods for constructing first diagnostic datasets, then training datasets.

We focus on the problem setup where the teacher must first diagnose the student, then construct a teaching dataset. The performance of the teacher is based on how well the student performs on a held-out test dataset. We build on prior work in machine teaching (Zhu et al., 2018; Liu et al., 2016; Cakmak & Lopes, 2012) for constructing optimal teaching datasets. We use Gaussian



(a) (b) (c)

Figure 1: Teacher needs to figure out what the student knows vs. doesn't know. (a) depicts an example of a navigation environment where the student (blue, upper left) has to navigate to a goal location (green, lower right). (b) illustrates that in some cases, the student has partial knowledge of the environment; from any of the yellow-highlighted cells, the student knows how to navigate to the green cell. (c) highlights the cells from which the student does not know how to navigate to the goal. The ideal teacher is one that can identify these states and construct examples from these states.

Processes (Williams & Rasmussen, 2006) as a tractable means for inferring student parameters when diagnosing. We combine machine teaching with diagnosing, and apply this framework to linear models for ridge regression, support vector machines (SVMs) and offline reinforcement learning settings.

We apply this to two settings. One is where the student learns from scratch and the teacher must figure out the student's learning algorithm parameters, eg. the regularization parameters in ridge regression or support vector machines. Two is where the student has partially explored the environment and the teacher must figure out the important states the student has not learn the optimal value for; we study this in the offline reinforcement learning setting where the teacher provides demonstrations to the student and should avoid sending redundant trajectories.

In summary, our work's contributions are the following:

- We show that the teacher must diagnose the student, otherwise the student can have arbitrarily bad performance.
- We show that if the teacher does first diagnose, the learner can learn from much fewer training examples than the teacher. In the offline RL setting, this results in 2-$16,000\times$ fewer training steps.
- We show that in certain cases, diagnosing with machine teaching leads to recovering optimal model parameters, where passive learning or naive machine teaching fail to do so.

2 Related works

Machine teaching The machine teaching problem is to determine an optimal training set, given knowledge about the student algorithm (eg. ridge regression) and the optimal model parameters θ^* (Zhu et al., 2018; Zhu, 2015; Simard et al., 2017). An *optimal* dataset is defined by the cardinality of the dataset, also known as the teaching dimension (Goldman et al.). Machine teaching has been broadly applied to settings like inverse reinforcement learning (Cakmak & Lopes, 2012; Brown & Niekum, 2019; Hadfield-Menell et al., 2016) where the goal is to construct a small demonstration dataset for the learner to most confidently learn the underlying reward function. Our work builds on results from Liu et al. (2016) which derive the teaching dimension of linear learners in settings including ridge regression and SVMs.

Imitation learning The imitation learning problem looks at how a student should learn a policy from an expert demonstration dataset, either by inferring the underlying reward that produced the behavior (inverse reinforcement learning, IRL) (Abbeel & Ng, 2004) or directly learning the policy (behavioral cloning, BC) (Torabi et al., 2018). Our work is close to the interactive imitation learning

Algorithm 1 Diagnosing student model

1: Initialize priors $p(z|H = \varnothing)$.
2: **for** $t = 1...T$ probing iterations **do**
3: Sample a student latent $z_t \sim p(z|H_t)$
4: *Optional*: Construct a train dataset $D_{\text{train}} = g(z_t)$ and pass to learner, $\theta = A(D_{\text{train}})$.
5: Construct a test probe x_{test} and get feedback from learner, $f(\theta, x_{\text{test}})$.
6: Use feedback to update beliefs $p(z|H)$
7: **Output** $p(z|H)$

setting Ross et al. (2011) where we rely on a teacher to provide action labels to states. The key difference is that we rely on a teacher to first probe and determine which states it should provide labels to; these might be states the student has never visited before.

Active learning The active learning problem looks at how a student should query a teacher for labels on unlabelled examples (Settles, 2009). Active learning is different from machine teaching in that the learners choose the datapoints they want to be trained on, rather than being guided by a teacher. However, the active learning framework might not be the right framework in settings where the learner doesn't have access to examples its never seen before. For example, DAgger (Ross et al., 2011) limits interaction between the learner and the teacher to the states the learner visits. This doesn't fully leverage what the learner *could* learn from the teacher, such as the teacher's privileged access to states that are important for a task of interest.

3 Interactive learning with Gaussian Processes

We study the problem of the teacher needing to infer properties of the student which will impact how the teacher teaches. In our setting, the teacher *first* diagnoses, *then* teaches the student. Extensions of this work could explore interweaving both diagnosis and teaching.

We use a Gaussian Process with a radial-basis function kernel for inference; alternative inference methods can be used as well. In our settings, the teacher must either infer something about the student's learning algorithm (eg. a hyperparameter) or the student's knowledge base (eg. its original training dataset).

Notation We denote $\mathcal{X}$ as the input space and $\mathcal{Y}$ as the output space. $f_\theta : \mathcal{X} \to \mathcal{Y}$ is the student model parameterized by model weights $\theta \in \Theta$. We refer to D as a dataset, and $\mathcal{D}$ as the space of datasets. A student runs an algorithm on a dataset to recover parameters that minimize the algorithm's learning objective, $A : \mathcal{D} \to \Theta$.

Our method for inference is simple:

1. **Sample hypothesis**: We sample a hypothesis z from our posterior distribution $p(z|H)$ according to expected improvement (EI) (Frazier, 2018). H is the history set of inputs and values given to the process.
2. **Train the learner**: We construct a train dataset based on the sampled hypothesis, $D_{\text{train}}(z) = \{(x_{\text{train}}, y_{\text{train}})\}$. We pass the dataset to the learner for it to run its learning algorithm A and learn initial model parameters, $\theta = A(D_{\text{train}})$. *This is only applicable in the setting where the learner trains from scratch.*
3. **Probe for feedback**: We construct a test probe $g(z) = x_{\text{test}}$ and send it to the learner. We receive feedback from the learner on that probe, $\texttt{feedback}(f_\theta(x_{\text{test}}))$.
4. **Update beliefs**: We use the feedback to update our beliefs, $p(z|H, x_{\text{test}}, \texttt{feedback})$.

This is summarized in Algorithm 1. After a few iterations, the teacher uses $p(z|H)$ to then construct a training dataset for the student. We apply this framework to the ridge regression, SVMs, and offline RL domains. What the teacher needs to infer in each domain varies, however the underlying mechanism is the same as outlined above.

After probing the student, the teacher constructs the optimal teaching dataset with the maximum a posteriori (MAP) estimate of $p(z|H)$. The optimal teaching dataset and teaching dimension lower bound have been studied in Liu et al. (2016) for ridge regression and SVM with linear models. This means that if the teacher *did* have complete information about its student (ie. knowing the student's regularization parameter λ), the teacher can construct the optimal teaching dataset D^*_{teach}. In later sections, we discuss how to exactly construct the optimal dataset in both settings. These domains illustrate how teaching and probing co-occur in a realistic setting where agents only have partial information about each other. Below we discuss each of the domains in more detail and what the teacher needs to infer in each domain.

3.1 Ridge regression

A linear model student running ridge regression (without a bias) minimizes the following objective:

$$l(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i \mathbf{w})^2 + \lambda \frac{\|\mathbf{w}\|^2}{2}, \tag{1}$$

where $\mathbf{w}$ are the linear model weights, $\mathbf{x}$ is the input, y is the target, and λ is the regularizer. We denote the optimal model weights to be $\mathbf{w}^*$. From Liu et al. (2016), the optimal machine teaching dataset consists of a single datapoint and is a function of the learner's regularizer λ,

$$D(\lambda) = \{(x_{\text{teach}}, y_{\text{teach}})\} = \{(a\mathbf{w}^*, \frac{\lambda + \|\mathbf{x}_{\text{t}}\|^2}{a})\}, \tag{2}$$

where a is any nonzero real number. The inhomogeneous ridge regression case (with a bias) is also discussed in Liu et al. (2016). We omit that setting as we use ridge regression as an illustrative example for optimal teaching with probing. However, our analysis is easily extendable to that setting.

We denote the learner's true regularization parameter as λ^*, and the teacher's MAP estimate of the regularization parameter as $\hat{\lambda}^*$. To simulate a teacher not having complete information about the student, we assume that the teacher does not know the learner's true λ^*; the teacher does not have complete information about *how* a student learns from data.

Concretely, the inference steps outlined in Section 3 are:

1. We sample a hypothesis for $\lambda \sim p(\lambda|H)$.
2. We construct a training dataset $D(\lambda) = (x_{\text{train}}, y_{\text{train}})$ based on the optimal teaching dataset outlined in Equation 2.
3. We construct a test probe $g(\lambda) = x_{\text{test}}$ for the learner and the real learner sends back its prediction, $f_{A(D(\lambda^*))}(x_{\text{test}})$. [1] We simulate how a learner with λ would have responded, $f_{A(D(\lambda))}(x_{\text{test}})$. We calculate the response difference, $(f_{A(D(\lambda))}(x_{\text{test}}) - f_{A(D(\lambda^*))}(x_{\text{test}}))^2$.
4. We use the difference as the target function for the Gaussian Process. It is the feedback used to update our beliefs over λ.

Our setting uses ordinary least squares linear regression for determining $\mathbf{w}$.

3.2 Support Vector Machines (SVM)

A linear SVM student minimizes the following objective:

$$l(\mathbf{w}) = \sum_{i=1}^{N} \max(1 - y_i \mathbf{x}_i \mathbf{w}, 0) + \lambda \frac{\|\mathbf{w}\|^2}{2} \tag{3}$$

[1] Empirically, we found that both randomly sampled probes or a training point from Step 2 allow for fast inference.

From Liu et al. (2016), the optimal teaching dataset consists of N identical training items where $N = \lceil \lambda \|\mathbf{w}^*\|^2 \rceil$. The teaching dataset is

$$D(\lambda) = \{(x_i, y_i)\}_{i=1}^N = \{(\frac{\lambda \mathbf{w}^*}{\lceil \lambda \|\mathbf{w}^*\|^2 \rceil}, 1)\}_{i=1}^N \tag{4}$$

Similar to ridge regression, the teacher does not know the learner's regularization parameter λ^* and must infer this through interaction. Concretely, the inference steps are:

1. We sample $\lambda \sim p(\lambda|H)$.
2. We construct a training dataset $D(\lambda) = (x_{\text{train}}, y_{\text{train}})$ based on the optimal teaching dataset outlined in Equation 4 for SVMs.
3. We construct a random test probe and pass it to the learner. The SVM learner sends back the distance of that point from its separating hyperplane. [2] We simulate what a learner with $\lambda = \hat{\lambda}$ would have responded and calculate the differences in their responses, ie. $(f_{A(D(\lambda))}(x_{\text{test}}) - f_{A(D(\lambda^*))}(x_{\text{test}}))^2$.
4. We use the difference as feedback to update our beliefs over λ.

Our setting uses linear support vector classification for determining $\mathbf{w}$.

3.3 Offline reinforcement learning

An offline RL student updates its Q-values given a dataset of trajectories $D = \{\tau_i\}_{i=1}^N$ where $\tau = [s_0, a_0, r_0, ..., s_{T-1}, a_{T-1}, r_{T-1}, s_T]$ is a trajectory of states and actions: $Q(s_t, a_t) = r_t + \gamma * V(s_{t+1}), \quad \forall t = 0...T-1$ [3]

In this domain, we assume the student can be tabula rasa or partially trained. An example is given in Figure 1. To the best of our knowledge, there isn't prior work in studying the teaching dimension problem in the offline RL setting. Nonetheless, intuitively we would like the teacher to propose trajectories that maximize coverage over the environment and minimizes overlaps with states the learner has already seen. The teacher is tasked to infer which states the learner has seen in order to provide demonstrations that are novel to the learner.

The inference steps follow as,

1. We sample a state $s \sim p(s|H)$.
2. We do *not* construct a training dataset in the behavioral cloning setting.
3. The sampled state is the test probe. We pass the sampled state s to the learner and the learner sends back the action they would take, $a_{\text{learner}} = \arg\max_a Q_{\text{learner}}(s, a)$. The feedback is a binary function of whether the learner takes the same action as the teacher would, $\mathbb{1}[a_{\text{learner}} = a_{\text{teacher}}]$.
4. We use the difference as feedback to update our beliefs over $\mathcal{S}$.

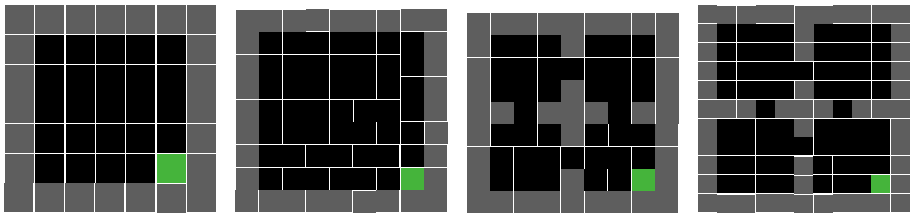
Our method uses ϵ-greedy Q-learning. Note that in the RL setting, states are used for diagnosing the student but the teacher ultimately teaches in the form of demonstrated trajectories. An open question that we hope to later pursue is the form of probe and teaching, and how this impacts the choice of target function (Step 3).

4 Experiments

We now evaluate the utility of understanding the learner before teaching and compare different teaching methods against passive learning (learning without a teacher and just from the original training dataset). In particular, we try to answer the following research questions (**RQ**s):

[2] We also tried receiving class predictions from the learner but found this to be an extremely sparse feedback signal. There might be smart ways to go about this, such as considering approaching for optimal probing. Related works in this direction include optimal experiment design (Pronzato, 2008; Foster et al., 2019).

[3] If function approximators are used, then this turns into fitting a parametric model π_θ where we represent the policy π in terms of a soft Q function Q_θ similar to Reddy et al. (2019).



(a) EmptyRoom 7x7 (b) EmptyRoom 9x9 (c) 4 Rooms 9x9 (d) 4 Rooms 11x11

Figure 2: Environments used for the RL setting. The learner can be initialized anywhere in the environment. Its goal is to navigate to the green grid cell.

RQ1: How important is it to understand the learner when teaching?

RQ2: When would you prefer teaching with diagnosing over passive learning?

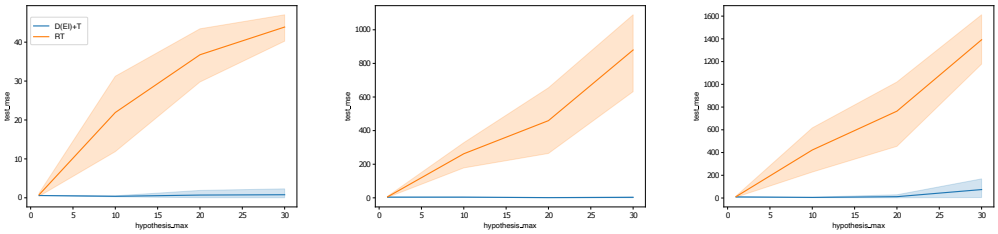
We examine this in three settings: Ridge regression, classification with support vector machines and offline RL in sequential decision making settings. The ridge regression and classification settings are meant to be simple, illustrative examples of how probing would be combined with optimal teaching, when the optimal teaching dataset can be determined with full knowledge of the student Liu et al. (2016). The regression and two-way separable classification data are generated from Pedregosa et al. (2011) without added noise. We use the MiniGrid environments (Chevalier-Boisvert et al., 2018) for the RL setting; examples of the environments are shown in Figure 2.

Methods Our main method first diagnoses the student with expected improvement (EI) as the acquisition function, then teaches the student. We denote our method as **D(EI)+T**. We compare three alternatives to our method. One is to diagnose with random sampling as the acquisition function, then teach the student; this is denoted as **D(R)+T**. Two is to teach randomly by sampling from the prior $p(z)$ and immediately teach the student; we denote this as **RT**. For instance, this would mean that in ridge regression, the teacher samples a random λ and constructs a teaching dataset according to Equation 2 without interacting with the student. Three is to passive learn without a teacher, **PL**. In ridge regression and SVMs, this means to passively learn over a fixed training dataset, and in the offline RL setting, this means the student collects and learns from its own data. For both methods **D(EI)+T** and **D(R)+T** where diagnosis is run, we allow the teacher to probe the student 5 times, ie. $T = 5$ in Algorithm 1.

4.1 Ridge regression

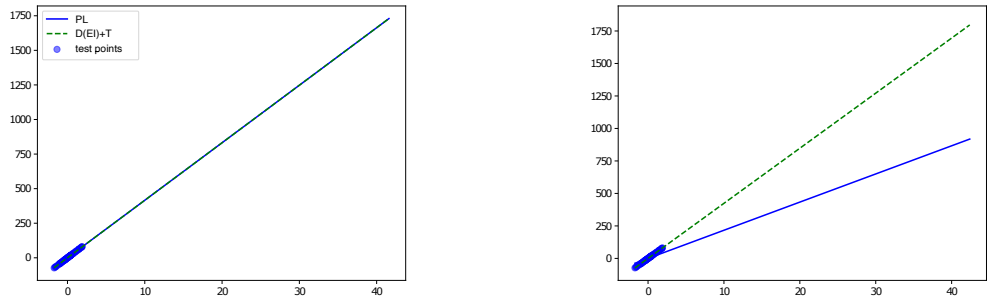
We compare the accuracy and sample efficiency (number of training examples) across the methods. Figure 3 compares **RT** and **D(EI)+T** to answer **RQ1**: How important is it for the teacher to learn about the student before teaching? We compare a teacher that randomly samples across an increasing hypothesis space for λ compared to a teacher that probes the student. We see that randomly guessing information about the student can lead to arbitrarily bad performance for the learner; these observations hold also across increasing feature dimension. With only 5 probes, **D(EI)+T** allows the student to predict with minimal error. Thus, to answer **RQ1**: It's important for the teacher to learn about the student, otherwise the student can perform poorly under larger belief spaces and feature spaces.

Figure 4 compares **PL** and **D(EI)+T** to answer **RQ2**: When would you prefer teaching with diagnosing over passive learning? With a fixed training dataset size $N = 100$, we vary the learner's regularization parameter from $\lambda = 1.0$ to $\lambda = 50.0$. We see that with particular learners (ie. low regularization), both methods do comparably in Figure 4c. However, with other learners (ie. high regularization), **PL** is not able to recover the best model. A teacher who diagnoses and then teaches, on the other hand, does enable the student to recover the best model. To answer **RQ2**: These results suggest teaching is preferable in low-data settings where sampling from the training distribution is expensive; the setting doesn't easily allow for more data to remedy the effects of high regularization or peculiarities in the student's algorithm.

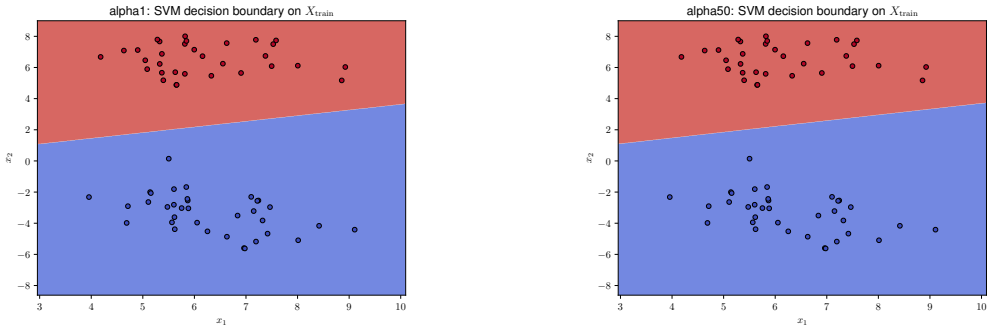


(a) MSE (1D) (b) MSE (2D) (c) MSE (3D)

Figure 3: Mean squared error on held-out regression dataset over 10 runs. Proper diagnosis of the student matters as the hypothesis space (x-axis) over possible student models grows and when the number of feature dimensions grows (a-c).



(a) $\lambda = 1$ (b) $\lambda = 50$

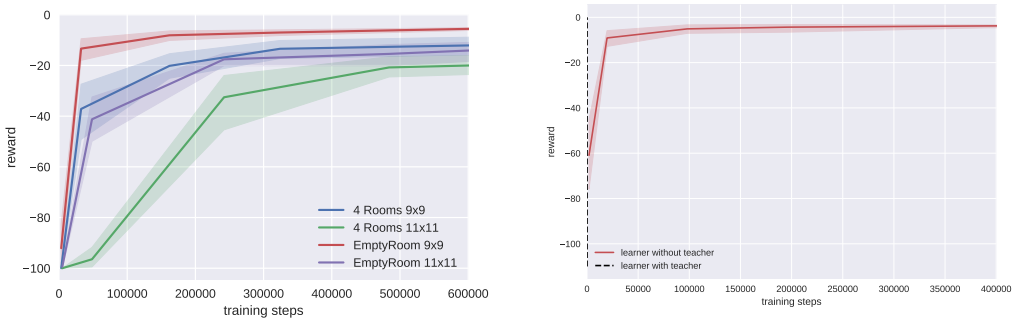


(c) $\lambda = 1$ (d) $\lambda = 50$

Figure 4: (a-b) illustrate a 2D regression task where the x- and y-axes correspond to the two coordinates. The first image is where the learners have a low regularization parameter ($\lambda = 1$), and the second image is where the learners have a high regularization parameter ($\lambda = 50$). The weights from two learners are plotted as lines. One is a learner trained on a teaching dataset (dashed, green line) by **D(EI)+T**. Another is the **PL** learner (full, blue line) trained on a *fixed* training regression dataset. The plotted points in blue are points from the test dataset. The **D(EI)+T** learner always finds the best fit model (ie. recovers θ^*) whereas the **PL** learner does not in the high regularization region. (c-d) however show that this is not the case for the SVM learners: both learners learn to separate the classes accurately.

4.2 Support Vector Machine

We conducted the same experiments as in ridge regression for the two-way separable classification with SVMs. However, we found separable classification settings with SVMs unable to yield differences among **PL**, **RT**, **D(EI)+T**: all three methods learn to separate the data within a reasonable



(a) (b)

Figure 5: Comparing the number of training samples needed for a learner with vs. without a teacher. (a) shows the number of samples needed a learner without a teacher across environments. (b) compares the number of samples needed by a learner without a teacher (red curve) and with a (black horizontal line $x \approx 50$) on the EmptyRoom 7x7 environment: with the help of a teacher sending demonstrations, the learner can learn much more quickly. All graphs were generated on 10 runs.

range over λ. Figure 4d provides intuition into why the three methods perform similarly on the classification domain. The separating hyperplane is not sensitive to increasing λ values. Therefore, when the classes are easy to separate, learners from each method will classify the points the same way. We extended these experiments to non-separable classes, however results were inconclusive as they required hand-designing the degree of overlap between classes.

4.3 Offline reinforcement learning

There are four environments we use for the RL setting; examples of the environments are shown in Figure 2. The learner can be initialized anywhere in the environment. The learner's goal is to navigate to the green cell. The reward for this setting is sparse: The learner receives -1 with every step except for when it reaches the goal where it receives 0 as its reward. The learner is evaluated based on how well it's able to navigate to the goal from any initial state. The teacher is trained until convergence by running ϵ-greedy Q-learning with $\epsilon = 0.3$. for 5000 episodes, each episode with 100 steps. The teacher learns the optimal policy at convergence. The learning curves of the teacher are shown in Figure 5a.

We assume different levels of student knowledge: Either the student has no experience (it has never interacted with the environment), it knows one optimal path, it knows a few optimal paths, or it knows all optimal paths to the goal. An example is illustrated in Table 1 on the left-hand column. The goal of the teacher is to then infer which states the student can perform well in, and ideally construct a teaching dataset of trajectories where the student would perform poorly.

We compare four different methods. One is **PL** where the passive learner runs standard ϵ-greedy Q-learning without the teacher. Two is **RT** where the teacher randomly samples a state from the environment and executes its policy until it reaches the goal; this is the trajectory that is sent over per sample. Three is **D(R)+T** where the teacher randomly samples a state for probing when approximating the target function for the Gaussian process. It then picks states that maximizes the target function as the initial states for running its policy and collecting trajectories. Four is **D(EI)+T** where the teacher samples a state with the highest expected improvement. Similarly to **D(R)+T**, **D(EI)+T** picks the states that maximize the target function as the initial states for collecting trajectories.

To answer **RQ1**, we compare against **RT** in Table 1 for EmptyRoom 7x7; results for the other environments can be found in the appendix. The first takeaway is that probing methods like **D(R)+T** and **D(EI)+T** both teach better than **RT** (middle column). Additionally, teachers that probe students avoid sending repeated states which the learner has already seen before (right-hand column). Interestingly, **D(R)+T** tends to perform better or comparably to **D(EI)+T**. This suggests that that the

target function we use (action matching) might not be ideal for determining states to probe a student with.

To answer **RQ2**, we compare **PL** with **D(EI)+T** in Figure 5. Noticeably, the student can benefit a lot from having a teacher provide demonstrations as it's able to learn from much fewer samples and still achieve high performance. Thus, a benefit to having a teacher is avoiding inefficient exploration.

5 Conclusion and future work

Our work showcases three simple settings which highlight the utility of having a teacher first learn about its student before teaching. We believe this form of collective learning—algorithmically constructing diagnostics for the learner in order to construct model-tailored training—can pave principled ways for diagnosing models and proposing remedies.

There are several interesting questions which we hope to later explore. These include,

- What is an optimal target function (form of feedback) to use for the Gaussian Process? The offline RL experiments suggest that we should revisit this choice. They also suggest that if a reasonably generic target function can be found in the RL setting, training RL policies can be made in a more sample-efficient and targeted way. A sensible, alternative target function could be the difference in value function between the student and teacher. However, one key limitation is that the teacher can access the (near-)optimal value function which is challenging to obtain in the first place.
- Under what conditions would you prefer machine teaching over passive learning? We'd like to formalize this as a function of the training set and problem difficulty. Under which domains (or types of domains) would machine teaching be useful for?
- How should we sample from a desirable data distribution given what we know about the student?

Acknowledgments

REW is supported by the National Science Foundation Graduate Research Fellowship. The authors would give special thanks to Dorsa Sadigh, Willie Neiswanger, Emma Brunskill, Tatsunori Hashimoto, Gregory Valiant, Andy Shih, Gabriel Poesia, and Rohith Kuditipudi for their helpful discussions. The authors would also like to thank Sidd Karamcheti, Jenn Grannen, and the anonymous reviewers for their feedback on the paper.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Twenty-first international conference on Machine learning - ICML '04*, pp. 1, Banff, Alberta, Canada, 2004. ACM Press. doi: 10.1145/1015330.1015430. URL http://portal.acm.org/citation.cfm?doid=1015330.1015430.

Wendy K. Adams and Carl E. Wieman. Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33:1289 – 1312, 2011.

Daniel S. Brown and Scott Niekum. Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications. *arXiv:1805.07687 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1805.07687. arXiv: 1805.07687.

Ricardo Caceffo, Steve Wolfman, Kellogg S. Booth, and Rodolfo Azevedo. Developing a computer science concept inventory for introductory programming. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pp. 364–369, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336857. doi: 10.1145/2839509.2844559. URL https://doi.org/10.1145/2839509.2844559.

Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *arXiv preprint arXiv:1903.05480*, 2019.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Sally A Goldman, Michael J Kearns, T Bell Laboratories, and Murray Hill. On the Complexity of Teaching. pp. 29.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative Inverse Reinforcement Learning. *arXiv:1606.03137 [cs]*, November 2016. URL http://arxiv.org/abs/1606.03137. arXiv: 1606.03137.

Julio Hernandez, Jesús Ariel Carrasco-Ochoa, and José Francisco Martínez-Trinidad. An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In *Iberoamerican Congress on Pattern Recognition*, pp. 262–269. Springer, 2013.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018.

Ji Liu, Xiaojin Zhu, and Hrag Ohannessian. The teaching dimension of linear learners. In *International Conference on Machine Learning*, pp. 117–126. PMLR, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Luc Pronzato. Optimal experimental design and some related control problems. *Automatica*, 44(2): 303–325, 2008.

M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*, 2019.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *arXiv:1011.0686 [cs, stat]*, March 2011. URL http://arxiv.org/abs/1011.0686. arXiv: 1011.0686.

Burr Settles. Active learning literature survey. 2009.

Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *arXiv:1707.06742 [cs, stat]*, August 2017. URL http://arxiv.org/abs/1707.06742. arXiv: 1707.06742.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

David F. Treagust. Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, 10:159–169, 1988.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

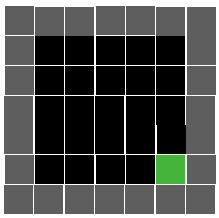
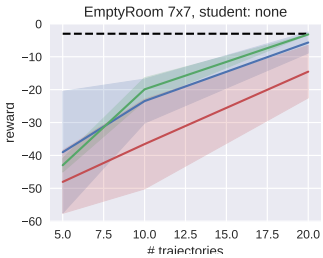
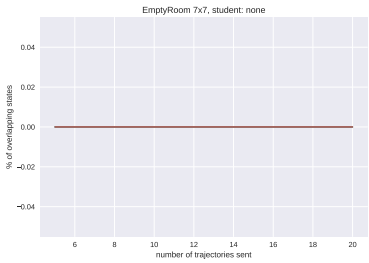
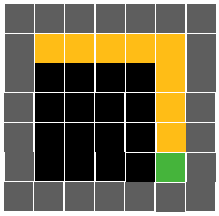
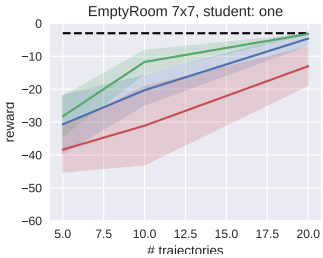
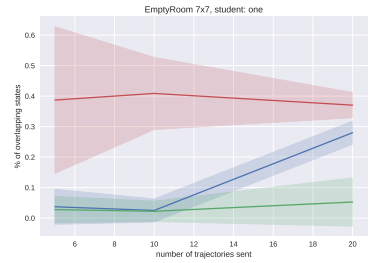
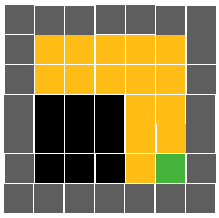
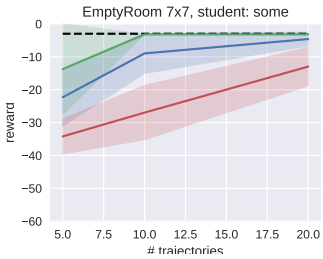
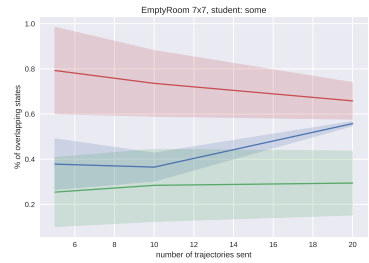
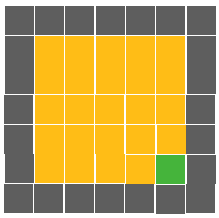
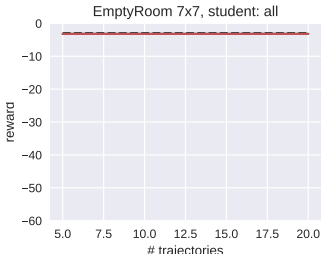
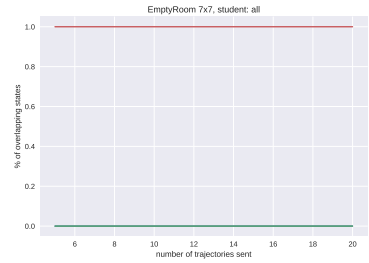
Student experience	Performance	% overlap
D(EI)+T D(R)+T RT 		
		
		
		

Table 1: Student experience with how different teachers teach the student, EmptyRoom 7x7. The dashed lines in the middle column denote the optimal policy's performance.

A APPENDIX

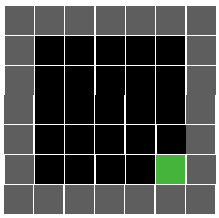
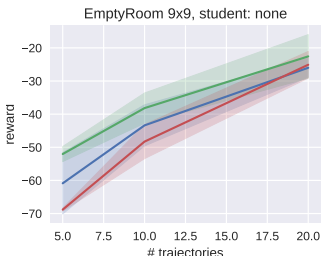
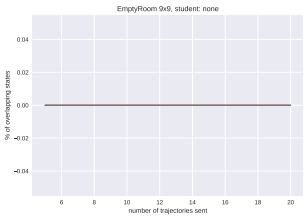
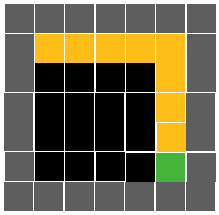
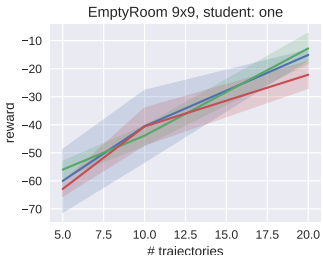
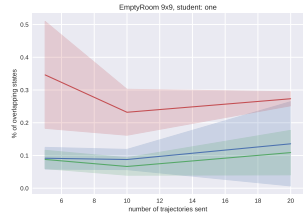
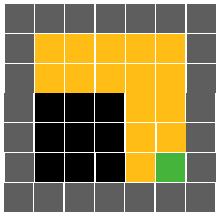
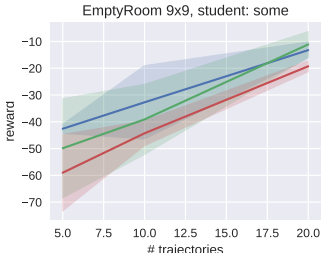
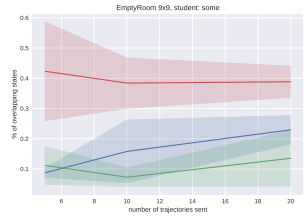
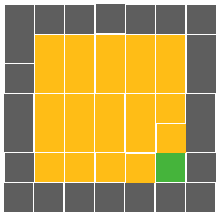
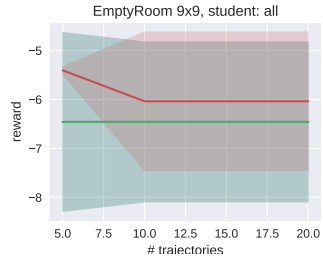
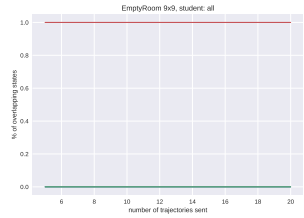
Student experience	Performance	% overlap
D(EI)+T D(R)+T RT 		
		
		
		

Table 2: Student experience with how different teachers teach the student, EmptyRoom 9x9. The dashed lines in the middle column denote the optimal policy's performance.

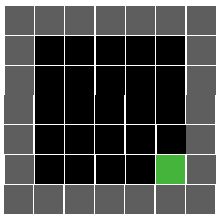
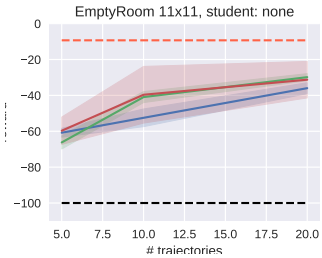
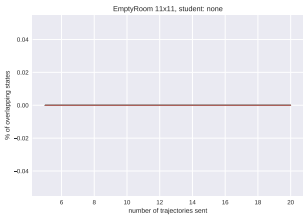
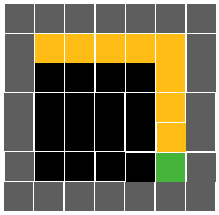
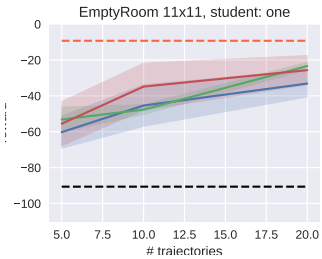
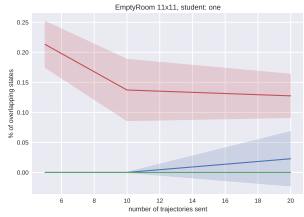
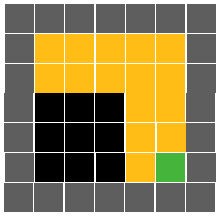
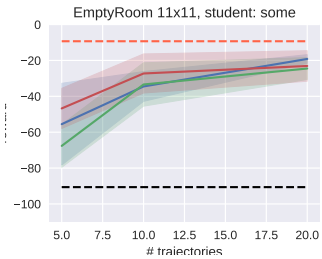
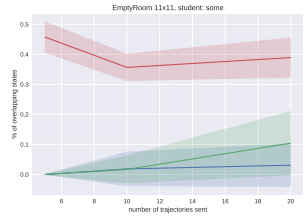
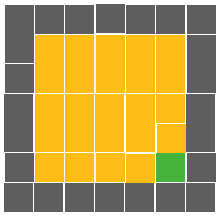
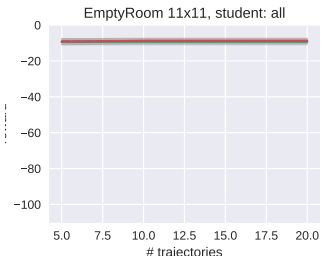
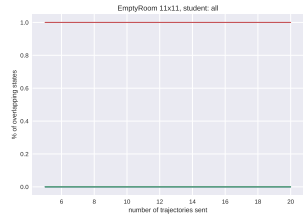
Student experience	Performance	% overlap
D(EI)+T D(R)+T RT 		
		
		
		

Table 3: Student experience with how different teachers teach the student, EmptyRoom 11x11. The dashed lines in the middle column denote the optimal policy's performance.

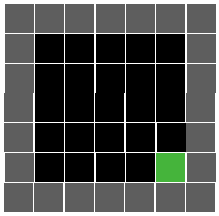
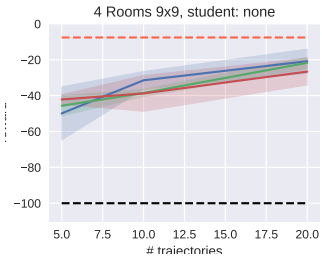
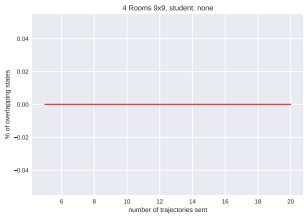
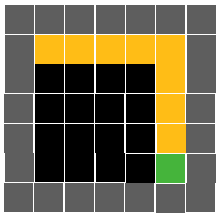
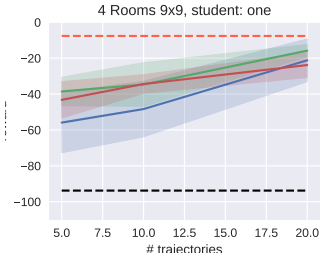
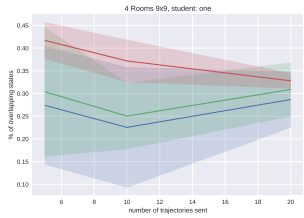
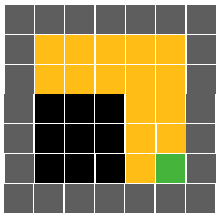
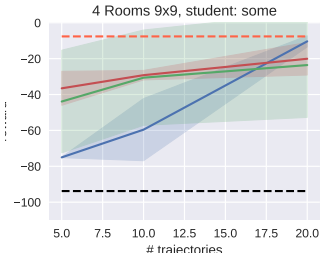
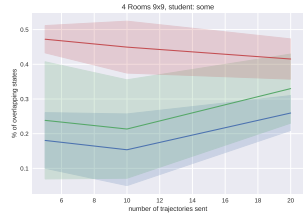
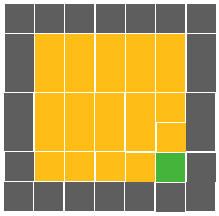
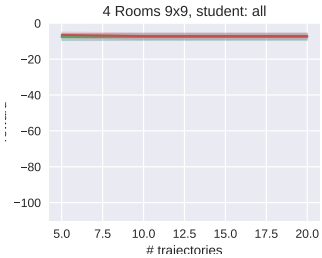
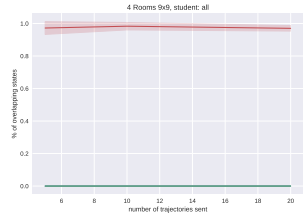
Student experience	Performance	% overlap
D(EI)+T D(R)+T RT 		
		
		
		

Table 4: Student experience with how different teachers teach the student, 4 rooms 9. The dashed lines in the middle column denote the optimal policy's performance.

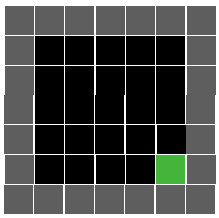
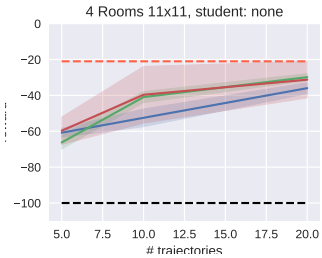
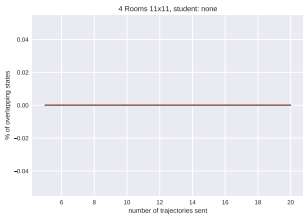
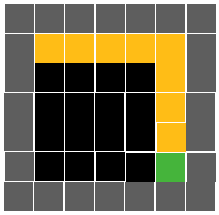
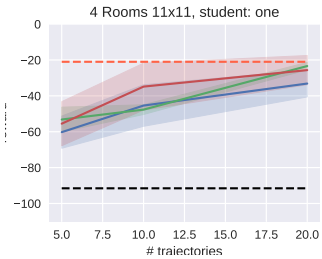
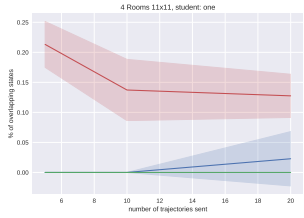
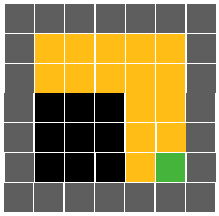
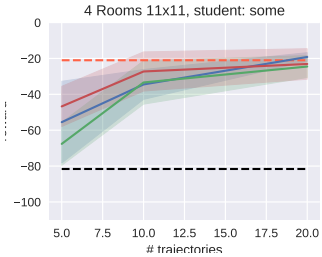
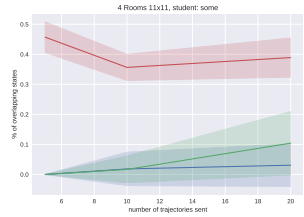
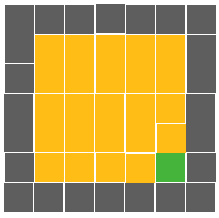
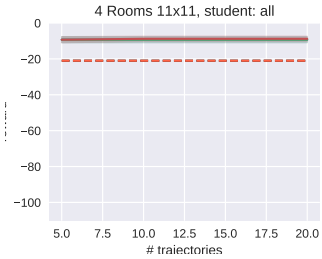
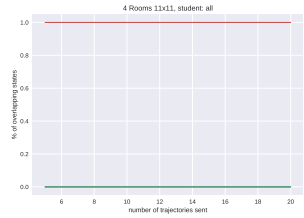
Student experience	Performance	% overlap
D(EI)+T D(R)+T RT 		
		
		
		

Table 5: Student experience with how different teachers teach the student, 4 rooms 11. The dashed lines in the middle column denote the optimal policy's performance.