

Scalable Planning with Deep Neural Network Learned Transition Models

Ga Wu

WUGA@MIE.UTORONTO.CA

*Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, ON, Canada
Vector Institute for Artificial Intelligence, Toronto, ON, Canada*

Buser Say

BUSER.SAY@MONASH.EDU

*Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, ON, Canada
Vector Institute for Artificial Intelligence, Toronto, ON, Canada
Faculty of Information Technology
Monash University, Melbourne, VIC, Australia*

Scott Sanner

SSANNER@MIE.UTORONTO.CA

*Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, ON, Canada
Vector Institute for Artificial Intelligence, Toronto, ON, Canada*

Abstract

In many complex planning problems with factored, continuous state and action spaces such as Reservoir Control, Heating Ventilation and Air Conditioning (HVAC), and Navigation domains, it is difficult to obtain a model of the complex nonlinear dynamics that govern state evolution. However, the ubiquity of modern sensors allows us to collect large quantities of data from each of these complex systems and build accurate, nonlinear deep neural network models of their state transitions. But there remains one major problem for the task of control – how can we plan with deep network learned transition models without resorting to Monte Carlo Tree Search and other black-box transition model techniques that ignore model structure and do not easily extend to continuous domains? In this paper, we introduce two types of planning methods that can leverage deep neural network learned transition models: Hybrid Deep MILP Planner (HD-MILP-Plan) and Tensorflow Planner (TF-Plan). In HD-MILP-Plan, we make the critical observation that the Rectified Linear Unit (ReLU) transfer function for deep networks not only allows faster convergence of model learning, but also permits a direct compilation of the deep network transition model to a Mixed-Integer Linear Program (MILP) encoding. Further, we identify deep network specific optimizations for HD-MILP-Plan that improve performance over a base encoding and show that we can plan optimally with respect to the learned deep networks. In TF-Plan, we take advantage of the efficiency of auto-differentiation tools and GPU-based computation where we encode a subclass of purely continuous planning problems as Recurrent Neural Networks and directly optimize the actions through backpropagation. We compare both planners and show that TF-Plan is able to approximate the optimal plans found by HD-MILP-Plan in less computation time. Hence this article offers two novel planners for continuous state and action domains with learned deep neural net transition models: one optimal method (HD-MILP-Plan) and a scalable alternative for large-scale problems (TF-Plan).

1. Introduction

In many complex planning problems with factored (Boutilier et al., 1999) and continuous state and action spaces such as Reservoir Control (Yeh, 1985), Heating, Ventilation and Air Conditioning (HVAC) (Agarwal et al., 2010), and Navigation (Faulwasser & Findeisen, 2009), it is difficult to obtain a model of the complex nonlinear dynamics that govern state evolution. For example, in Reservoir Control, evaporation and other sources of water loss are a complex function of volume, bathymetry, and environmental conditions; in HVAC domains, thermal conductance between walls and convection properties of rooms are nearly impossible to derive from architectural layouts; and in Navigation problems, nonlinear interactions between surfaces and traction devices make it hard to accurately predict odometry.

A natural answer to these modeling difficulties is to instead learn the transition model from sampled data; fortunately, the presence of vast sensor networks often make such data inexpensive and abundant. While learning nonlinear models with *a priori* unknown model structure can be very difficult in practice, recent progress in Deep Learning and the availability of off-the-shelf tools such as Tensorflow (Abadi et al., 2015) and Pytorch (Paszke et al., 2017) make it possible to learn highly accurate nonlinear deep neural networks with little prior knowledge of model structure.

However, the modeling of a nonlinear transition model as a deep neural network poses non-trivial difficulties for the optimal control task. Existing planners with nonlinear transition structure are either are not compatible with nonlinear deep network transition models and continuous (i.e., real-valued) actions¹, or only optimize goal-oriented objectives². Monte Carlo Tree Search (MCTS) methods (Coulom, 2006; Kocsis & Szepesvári, 2006; Keller & Helmert, 2013) including AlphaGo (Silver et al., 2016) that *could* exploit a deep network learned black box model of transition dynamics do not inherently work with continuous action spaces due to the infinite branching factor. While MCTS with continuous action extensions such as HOOT (Weinstein & Littman, 2012) have been proposed, their continuous partitioning methods do not scale to high-dimensional concurrent and continuous action spaces. Finally, offline model-free reinforcement learning with function approximation (Sutton & Barto, 1998; Szepesvári, 2010) and deep extensions (Mnih et al., 2013) do not directly apply to domains with high-dimensional continuous action spaces. That is, offline learning methods like Q-learning require action maximization for every update, but in high-dimensional continuous action spaces such nonlinear function maximization is non-convex and computationally intractable at the scale of millions or billions of updates.

Despite these limitations of existing methods, all is not lost. First, we remark that our deep network is not a black-box but rather a gray-box; while the learned parameters often lack human interpretability, there is still a uniform layered symbolic structure in the deep neural network models. Second, we make the critical observation that the popular Rectified Linear Unit (ReLU) (Nair & Hinton, 2010) transfer function for deep networks enables effective *nonlinear* deep neural network model learning and permits a direct compilation to a Mixed-Integer Linear Program (MILP) encoding. Given other components such as a human-specified objective function and a horizon, this permits direct optimization in a method we call Hybrid Deep MILP Planner (HD-MILP-Plan).

While arguably an important step forward, we remark that planners with optimality guarantees such as HD-MILP-Plan can only scale up to moderate-sized planning problems. Hence in an ef-

1. (Penna et al., 2009; Löhr et al., 2012; Coles et al., 2013; Ivankovic et al., 2014; Piotrowski et al., 2016; Scala et al., 2016a)

2. (Bryce et al., 2015; Scala et al., 2016b; Cashmore et al., 2016)

fort to scale to substantially larger control problems, we focus on a general subclass of planning problems with purely continuous state and action spaces in order to take advantage of the efficiency of auto-differentiation tools and GPU-based computation. Specifically, we propose to extend work using the Tensorflow tool for planning (Wu et al., 2017) in deterministic continuous RDDL (Sanner, 2010) domains to the case of learned neural network transition models investigated in this article. Specifically, we show that we can embed both a reward function and a deep-learned transition function into a Recurrent Neural Network (RNN) cell, chain multiple of these RNN cells together for a fixed horizon, and produce a plan in the resulting RNN encoding through end-to-end backpropagation in a method we call Tensorflow Planner (TF-Plan).

In brief, we can summarize the high-level procedure of the two approaches we contribute in this article by the following two steps:

1. Train a neural network to learn a transition model from sample trajectory data that predicts the next state given the current state and action.
2. Chain these (fixed) learned transition models together for a fixed planning horizon and optimize the action choices given a fixed initial state and an overall planning objective. We contribute and evaluate two different encodings and optimizers for this task:
 - HD-MILP-Plan, which compiles the learned ReLU-based neural network transition model and overall objective into a MILP and leverage the CPLEX optimizer³.
 - TF-Plan, which encodes the learned transition model in an RNN-based deep neural network unrolled for a fixed planning horizon and leverages Tensorflow (Abadi et al., 2015)⁴ for fast and efficient gradient descent through backpropagation to optimize the overall objective w.r.t. this (fixed) learned transition model.

Experimentally, we compare HD-MILP-Plan and TF-Plan versus manually specified domain-specific policies on Reservoir Control, HVAC, and Navigation domains. Our primary objectives are to comparatively evaluate the ability of HD-MILP-Plan and TF-Plan to produce high quality plans with limited computational resources in an online planning setting, and to assess their performance against carefully designed manual policies. For HD-MILP-Plan, we show that our strengthened MILP encoding improves the quality of plans produced in less computational time over the base encoding. For TF-Plan, we show the scalability and the efficiency of our planner on large-scale problems and its ability to approximate the optimal plans found by HD-MILP-Plan on moderate-sized problems. Overall, this article contributes and evaluates two novel approaches for planning in continuous state and action domains with learned deep neural net transition models: one optimal method (HD-MILP-Plan) and a scalable alternative for large-scale problems (TF-Plan).

2. Deterministic Factored Planning Problem Specification

Before we proceed to discuss deep network transition learning, we review the general planning problem that motivates this work. A deterministic factored planning problem is a tuple $\Pi = \langle S, A, C, T, I, G, Q \rangle$ where $S = \{s^1, \dots, s^{n_1}\}$ and $A = \{a^1, \dots, a^{n_2}\}$ are sets of state and ac-

3. <http://www.cplex.com/>

4. <https://www.tensorflow.org/>

tion variables with continuous domains⁵, $C : \mathbb{R}^{|S|} \times \mathbb{R}^{|A|} \rightarrow \{true, false\}$ is a function that returns true if values of action $\bar{A}_t = \langle \bar{a}_t^1, \dots, \bar{a}_t^{|A|} \rangle \in \mathbb{R}^{|A|}$ and state $\bar{S}_t = \langle \bar{s}_t^1, \dots, \bar{s}_t^{|S|} \rangle \in \mathbb{R}^{|S|}$ variables at time t satisfy global constraints, $T : \mathbb{R}^{|S|} \times \mathbb{R}^{|A|} \rightarrow \mathbb{R}^{|S|}$ denotes the transition function between time steps t and $t + 1$, $I : \mathbb{R}^{|S|} \rightarrow \{true, false\}$ is the initial state constraint indicating which assignment to state variables S is the initial state, and $G : \mathbb{R}^{|S|} \rightarrow \{true, false\}$ represents goal state constraints. Finally, $Q : \mathbb{R}^{|S|} \times \mathbb{R}^{|A|} \rightarrow \mathbb{R}$ denotes the reward function⁶. Given a planning horizon H , an optimal solution (i.e., an optimal plan) to Π is a value assignment to the action variables with values \bar{A}^t for all time steps $t \in \{1, \dots, H\}$ (and state variables with values \bar{S}^t for all time steps $t \in \{1, \dots, H + 1\}$) such that the state variables are updated by the transition function $T(\langle \bar{s}_t^1, \dots, \bar{s}_t^{|S|}, \bar{a}_t^1, \dots, \bar{a}_t^{|A|} \rangle) = \bar{S}_{t+1}$ for all time steps $t \in \{1, \dots, H\}$, initial and goal state constraints are satisfied such that $I(\bar{S}_1) = true$ and $G(\bar{S}_{H+1}) = true$, and the total reward function over horizon H , i.e., $\sum_{t=1}^H Q(\langle \bar{s}_{t+1}^1, \dots, \bar{s}_{t+1}^{|S|}, \bar{a}_t^1, \dots, \bar{a}_t^{|A|} \rangle)$, is maximized.

In many complex problems, it is difficult to model the exact dynamics of the complex nonlinear transition function T that governs the evolution of states S over the horizon H . Therefore in this paper, we do not assume a-priori knowledge of T , but rather we learn it from data. We limit our model knowledge to a human-specified reward function Q , horizon H and global constraint function C that specifies whether actions A are applicable in state S at time t , or not, e.g., the outflow from a reservoir must not exceed the present water level, and goal state constraints G . Given these known components S, A, C, I, G, Q of our deterministic factored planning problem, we next discuss how the final component T can be learned as a deep neural network given data.

3. Neural Network Transition Learning

A neural network is a layered, acyclic, directed computational network structure commonly used for supervised learning tasks (Goodfellow et al., 2016). A modern (deep) neural network typically has one (or more) hidden layers, with each hidden layer typically consisting of a linear transformation of its input followed by a nonlinear activation function to produce its output. Each successive hidden layer provides the ability to learn complex feature structure building on the previous layer(s) that allow it to model nearly arbitrary nonlinear functions. While most traditional nonlinear activation functions are bounded (e.g., sigmoid or tangent function), simple piecewise linear activation functions such as the rectified linear unit (ReLU) (Nair & Hinton, 2010) have become popular due to their computational efficiency and robustness to the so-called vanishing gradient problem.

3.1 Network Structure

In this article, we model the transition function through a deep neural network, where the network takes the current state and current action as input and produces the prediction of the next state as

5. In this article we focus on purely continuous state and action planning problems. While the extension to mixed (i.e., continuous and discrete domains) is an interesting and important problem for future work, it poses a number of significant additional challenges for both deep network model learning as well as Tensorflow-based planning through gradient descent. We discuss these challenges and possible extensions in more detail in our concluding future work discussion.

6. Q should not be confused with the Q-function in MDPs. We use Q here since the more standard notation R for immediate reward is used elsewhere for other purposes in this article.

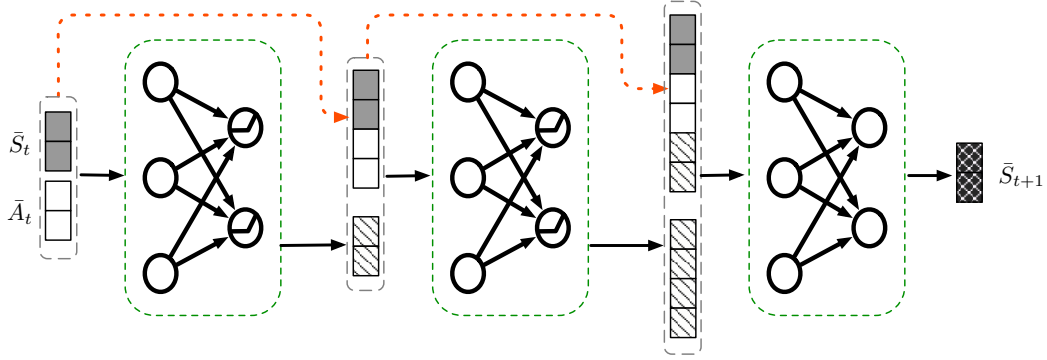


Figure 1: An example two-layer neural network transition model. At left are the input values of state \bar{S}_t and action \bar{A}_t variables at time t , and at far right are the output value state variables \bar{S}_{t+1} at time $t + 1$. The layers of the neural network between this input and output represent the transition function $T(\langle \bar{s}_t^1, \dots, \bar{s}_t^{|S|}, \bar{a}_t^1, \dots, \bar{a}_t^{|A|} \rangle) = \bar{S}_{t+1}$. In this neural network, black solid arrows indicate direct connections into each neural network layer, while red dashed lines indicate skip connections that bypass neural network layers and copy the input of the arrow to the output. Observe that the red dashed skip connections accumulate the concatenation of all previous layers. Green dashed rounded rectangles represent fully connected layers with an output nonlinearity – in this paper we use a ReLU nonlinearity.

output. In particular, we use the modified *densely-connected* deep network (Huang et al., 2016)⁷ as shown in Figure 1, which in comparison to a standard fully connected network, allows direct connections of each layer to the output via skip connections. This can be advantageous when a transition function has differing levels of nonlinearity (i.e., requiring differing network depths to model), allowing linear components of transitions to pass directly from the input to the output layer, while nonlinear components of transitions may pass through one or more of the hidden layers. We found that it is critical to allow transition function learning to bypass hidden layers when appropriate since, as we will show in Section 6.3, (a) deeper networks do not necessarily reduce error in transition learning and (b) deeper networks are significantly more time-consuming to optimize with MILP-based solvers. Thus, having shorter paths enabled by skip connections in dense deep networks can facilitate better learning and faster optimization.

Given a deep neural network configuration with K hidden layers, a hyperparameter λ , and a dataset D consisting of vectors of current state $\bar{S}_n = (\bar{s}_n^1, \dots, \bar{s}_n^{|S|})$, next state $\bar{S}'_n = (\bar{s}'_n^1, \dots, \bar{s}'_n^{|S|})$ and action $\bar{A}_n = (\bar{a}_n^1, \dots, \bar{a}_n^{|A|})$ variable values denoted as $D = \{\bar{S}_1 \hat{\ } \bar{A}_1 \hat{\ } \bar{S}'_1, \dots, \bar{S}_N \hat{\ } \bar{A}_N \hat{\ } \bar{S}'_N\}$ where $\hat{\ }$ denotes the concatenation operation between two vectors, the optimal weights \mathbf{W}_k for all layers $k \in \{1, \dots, K\}$ can be found by minimizing the following prediction error objective:

$$\underset{\mathbf{W}_k, \mathbf{b}_k, k \in \{1, \dots, K\}}{\text{minimize}} \sum_{n \in \{1, \dots, N\}} \|\gamma \odot [\bar{S}'_n - \mathbf{S}'_n]\|_2^2 + \lambda \sum_{k \in \{1, \dots, K\}} \|\mathbf{W}_k\|_2^2 \quad (1)$$

subject to

$$\mathbf{Z}_l = g([\bar{S}_n \hat{\ } \bar{A}_n \hat{\ } \mathbf{Z}_1 \hat{\ } \dots \hat{\ } \mathbf{Z}_{k-1}] \mathbf{W}_k^T + \mathbf{b}_k) \quad \forall k \in \{1, \dots, K-1\}, n \in \{1, \dots, N\} \quad (2)$$

$$\mathbf{S}'_n = [\bar{S}_n \hat{\ } \bar{A}_n \hat{\ } \mathbf{Z}_1 \hat{\ } \dots \hat{\ } \mathbf{Z}_{K-1}] \mathbf{W}_K^T + \mathbf{b}_K \quad \forall n \in \{1, \dots, N\} \quad (3)$$

7. The densely connected network (Huang et al., 2016) was first proposed for Convolutional Neural Networks but has clear analogues with the fully connected networks we use in this article.

Here, weight matrix \mathbf{W}_k and vector of biases \mathbf{b}_k for each neural network hidden layer $k \in \{1, \dots, K\}$ are the key parameters being optimized. We use the vector \mathbf{Z}_l to represent the intermediate output from the hidden layer l .

The objective (1) minimizes squared reconstruction error of transition predictions for the next state in the data, where γ denotes a vector of dimensional rescaling weights (defined in the *loss normalization* section below) elementwise multiplied \odot with the vector of residual prediction errors $[\bar{\mathbf{S}}'_n - \mathbf{S}'_n]$. The objective also contains an L_2 regularization term $\|\mathbf{W}_k\|_2^2$ with hyperparameter $\lambda \in \mathbb{R}$ to prevent overfitting⁸. Constraints (2)-(3) define the nonlinear activation of \mathbf{Z}_k for hidden layers $k \in \{1, \dots, K-1\}$ and outputs \mathbf{S}'_n , where g denotes a nonlinear activation function (specified below). To reduce notational clutter, we omit subscript n of the hidden layers \mathbf{Z}_k though we remark that each unique state $\bar{\mathbf{S}}_n$ and action $\bar{\mathbf{A}}_n$ variable values naturally lead to a different activation pattern for the \mathbf{Z}_k .

We use Rectified Linear Units (ReLU) (Nair & Hinton, 2010) of the form $\text{ReLU}(x) = \max(x, 0)$ as the activation function in this paper. ReLUs offer a threefold benefit for our planning tasks:

1. In comparison to other activation functions, such as the sigmoid or hyperbolic tangent, the ReLU can be trained efficiently and numerically stably since its derivative is either 1 or 0.
2. Since ReLUs produce a piecewise-linear approximation of the transition function, they permit direct compilation to a set of linear constraints (containing both continuous and integer variables) in Mixed-Integer Linear Programming (MILP) as we will discuss in Section 4. Since the number of ReLUs and their constraint-based encoding significantly affects solution time, we do encounter scaling issues with the MILP approach, which we address with our contribution of an alternate gradient-based optimization approach in TF-Plan as we will discuss in Section 5.
3. ReLU activations are robust to the vanishing gradient problem which is advantageous in the context of planning through backpropagation over long horizons, which we will discuss in Section 5.2.

3.2 Input Standardization and Loss Normalization

Input Standardization: Standardizing input data to have zero mean and unit variance facilitates gradient-based learning used in backpropagation for deep neural networks (Goodfellow et al., 2016). Hence, we standardize inputs before feeding them into the neural network, which we found to significantly improve training quality. Specifically, denoting the i th neural network input as \bar{x}^i , we denote the normalized input as $\hat{x}^i = \frac{\bar{x}^i - \bar{\mu}^i}{\bar{\sigma}^i}$, where $\bar{\mathbf{S}}_n \cap \bar{\mathbf{A}}_n = (\bar{x}_n^1, \dots, \bar{x}_n^{|\mathbf{A}|+|\mathbf{S}|})$, and $\bar{\mu}^i$ and $\bar{\sigma}^i$ are respectively the empirical mean and standard deviation of \bar{x}^i in the data.

After learning, we can modify the weights and biases of the learned neural network to accept unnormalized inputs instead, thus bypassing the need to normalize states and actions during online planning in subsequent sections. To do this, we define modified weights \hat{w}_{ij} and biases \hat{b}_j , respectively for each entry w_{ij} (connecting input i to hidden layer unit j) in matrix \mathbf{W}_k and bias b_j (for

8. The L_2 regularization term on weights \mathbf{W}_k is a standard method in the deep learning literature for limiting model complexity to prevent overfitting by penalizing weights with large values, cf. Section 7.1 of (Goodfellow et al., 2016).

each hidden layer unit j) in vector \mathbf{b}_k as follows:

$$\hat{w}_{ij} = \frac{w_{ij}}{\bar{\sigma}^i} \quad \text{and} \quad \hat{b}_j = - \sum_{i=1}^{|A|+|S|} \frac{\bar{\mu}^i w_{ij}}{\bar{\sigma}^i} + b_j.$$

The full derivation of this result is provided in Appendix A.

Loss Normalization: While it is critical to standardize the input to improve the learning process, it is also important to normalize the loss applied to each dimension of the output prediction.

To make the reasons for this more clear, we first begin by observing that it is not uncommon for a next state vector \mathbf{S}'_n to contain state variables of highly varying magnitude. For example, in the Reservoir control problem that we formally define in Section 6.1, state variables correspond to water volumes of different reservoirs that may have vastly different capacities. In this case, a uniformly weighted elementwise loss in the objective (1) would lead to most learning effort expended on the largest capacity reservoir at the expense of (relative) accuracy for the smaller reservoirs since the largest reservoir would dominate the squared error.

To address this loss imbalance, we employ an elementwise loss-weighting vector γ on each dimension of the squared loss in the objective (1). Specifically, letting \bar{S}_{max} be defined as the elementwise maximum of $\bar{s}'_1, \dots, \bar{s}'_N$ of each output dimension $i \in \{1, \dots, |S|\}$ in the data, each component of the loss weighting rescales the corresponding error residual by its maximum value observed in the data:

$$\gamma = \frac{1}{\bar{S}_{max}}.$$

Without the loss normalizing effect of γ in the objective (1), we found it is extremely difficult to learn an accurate predictive model over *all* next state dimensions.

3.3 Network Regularization and Complexity Reduction

Learning complex functions on limited training data naturally runs the risk of overfitting, which would lead to poor planning performance where planners explore actions and states that may differ substantially from those seen in the training data. Hence generalization is important and thus in order to combat model overfitting, we deployed *Dropout* (Srivastava et al., 2014), a regularization technique for structured deep networks that randomly “drops out” neurons with probability p in order to prevent neurons from co-adapting to each other and memorizing aspects of the training data — a key factor in deep network overfitting. More precisely, we use a slight variant known as *Inverted Dropout* (Li, Johnson, & Yeung, 2018) that proves to be more stable in post-training generalization and is the default Dropout implementation in many deep learning toolboxes such as Tensorflow.

In addition to Inverted Dropout for regularizing our deep network, it is critical to find the network structure and size that permits good generalization to future state and action scenarios not seen in the training data. The reasons for this are twofold. First, minimizing network structure reduces model representation capacity and is by itself a technique to prevent overfitting. Second, with an eye towards the ultimate planning motivation of this article, we remark that a smaller network can dramatically reduce the planning computation time. This is especially true for HD-MILP-Plan, where higher hidden layer width and additional hidden layers result in more *big-M* constraints that substantially increase the computational cost of computing a plan as we discuss and experimentally

show in Sections 4 and 6, respectively. Structural tuning typically consists of searching over the number of hidden layers as well as the width of these hidden layers and is detailed in Section 6.

4. Hybrid Deep MILP Planner (HD-MILP-Plan)

In the previous section, we have defined the underlying planning problem addressed in this paper as well as the methodology and structure of deep neural networks used to learn the transition function component. We now proceed to propose our first planner, HD-MILP-Plan, which can exploit this problem structure and specifically the ReLU structure of our deep net learned transition models by leveraging a compilation to a Mixed-Integer Linear Program (MILP).

Hybrid⁹ Deep MILP Planner (HD-MILP-Plan) is a two-stage framework for learning and optimizing planning problems with piecewise linear transition functions. The first stage of HD-MILP-Plan learns the unknown transition function T with densely-connected network as discussed previously in Section 3. The learned transition function \tilde{T} is then used to construct the learned planning problem $\tilde{\Pi} = \langle S, A, C, \tilde{T}, I, G, Q \rangle$. Given a planning horizon H , HD-MILP-Plan compiles the learned planning problem $\tilde{\Pi}$ into a MILP and finds an optimal plan to $\tilde{\Pi}$ using an off-the-shelf MILP solver. HD-MILP-Plan operates as an online planner where actions are optimized over the remaining planning horizon in response to sequential state observations from the environment.

We now describe the base MILP encoding of HD-MILP-Plan. Then, we strengthen the linear relaxation of our base MILP encoding for solver efficiency.

4.1 Base MILP Encoding

We begin with all notation necessary for the HD-MILP-Plan specification:

4.1.1 PARAMETERS

- V_s is the value of the initial state variable $s \in S$.
- R is the set of ReLUs in the neural network.
- B is the set of learned bias units in the neural network.
- O is the set of output units in the neural network.
- \tilde{w}_{ij} denotes the learned weight in the neural network between units i and j .
- A_f is the set of action variables connected to unit $f \in R \cup O$.
- S_f is the set of state variables connected to unit $f \in R \cup O$.
- U_f is the set of units connected to unit $f \in R \cup O$.
- O_s specifies the output unit with linear function that predicts the value of state variable $s \in S$.
- M is a large constant used in the big-M constraints.

4.1.2 DECISION VARIABLES

- X_t^a denotes the value of action variable $a \in A$ at time t .
- Y_t^s denotes the value of state variable $s \in S$ at time t .
- P_t^f denotes the output of ReLU $f \in R$ at time t .
- $P_t^f = 1$ if ReLU $f \in R$ is activated at time t , 0 otherwise (i.e., P_t^f is a Boolean variable).

9. The term *hybrid* refers to mixed (i.e., discrete and continuous) action and state spaces as used in MDP literature (Kveton et al., 2006).

4.1.3 THE MILP COMPILATION

Next we define the MILP formulation of our planning optimization problem that encodes the learned transition model.

$$\text{maximize } \sum_{t=1}^H Q(\langle Y_{t+1}^{s_1}, \dots, Y_{t+1}^{s_{|S|}}, X_t^{a_1}, \dots, X_t^{a_{|A|}} \rangle) \quad (4)$$

subject to

$$Y_1^s = V_s \quad \forall s \in S \quad (5)$$

$$C(\langle Y_t^{s_1}, \dots, Y_t^{s_{|S|}}, X_t^{a_1}, \dots, X_t^{a_{|A|}} \rangle) \quad (6)$$

$$G(\langle Y_{H+1}^{s_1}, \dots, Y_{H+1}^{s_{|S|}} \rangle) \quad (7)$$

$$P_t^f = 1 \quad \forall f \in B \quad (8)$$

$$P_t^f \leq M P_t'^f \quad \forall f \in R \quad (9)$$

$$P_t^g \leq M(1 - P_t'^g) + \hat{P}_t^g \quad \forall g \in R \quad (10)$$

$$P_t^g \geq \hat{P}_t^g \quad \forall g \in R \quad (11)$$

$$Y_{t+1}^s = \hat{P}_t^g \quad \forall g \in O_s, s \in S \quad (12)$$

$$\text{where expression } \hat{P}_t^g = \sum_{f \in U_g} \tilde{w}_{fg} P_t^f + \sum_{s \in S_g} \tilde{w}_{sg} Y_t^s + \sum_{a \in A_g} \tilde{w}_{ag} X_t^a \quad \forall g \in R$$

for all time steps $t \in \{1, \dots, H\}$ except constraints (5)-(7).

In the above MILP, the objective function (4) maximizes the sum of rewards over a given horizon H . Constraint (5) connects input units of the neural network to the initial state of the planning problem at time $t = 1$. Constraint (6) ensures that global constraints are satisfied at every time t . Constraint (7) ensures output units of the neural network satisfy goal state constraints of the planning problem at time $t = H + 1$. Constraint (8) sets all neurons that represent biases equal to 1. Constraint (9) ensures that a ReLU $f \in R$ is activated if the total weighted input flow into f is positive. Constraints (10)-(11) together ensure that if a ReLU $f \in R$ is active, the outflow from f is equal to the total weighted input flow. Constraint (12) predicts the values of state variables at time $t + 1$ given the values of state and action variables, and ReLUs at time t using linear activation functions.

4.2 Strengthened MILP Encoding

In the previous encoding, constraints (9)-(11) sufficiently encode the piecewise linear activation function of the ReLUs. However, the positive unbounded nature of the ReLUs leads to a poor linear relaxation of the big-M constraints, that is, when all boolean variables are relaxed to continuous $[0, 1]$ in constraints (9)-(10); this can significantly hinder the overall performance of standard branch and bound MILP solvers that rely on the linear relaxation of the MILP for guidance. Consequently, in this section, we strengthen our base MILP encoding by preprocessing bounds on state and action variables, and with the addition of auxiliary decision variables and linear constraints, to improve its LP relaxation.

In our base MILP encoding, constraints (9)-(11) encode the piecewise linear activation function, $\text{relu}(x) = \max(x, 0)$, using the big-M constraints for each ReLU $f \in R$. We strengthen the linear relaxation of constraints (9)-(10) by first finding tighter bounds on the input units of the

neural network, then separating the input x into its positive x^+ and negative x^- components. Using these auxiliary variables, we augment our base MILP encoding with an additional linear inequality in the form of $x^+ \geq \text{relu}(x)$. This inequality is valid since the constraints $x = x^+ + x^-$ and $\text{relu}(x) = \max(x, 0) \leq \max(x^+, 0) = x^+$ hold for all $x^+ \geq 0$ and $x^- \leq 0$.

4.2.1 PREPROCESSING BOUNDS

The optimization problems solved to find the tightest bounds on the input units of the neural network are as follows. The tightest lower bounds on action variables can be obtained by solving the following optimization problem:

$$\begin{aligned} & \text{minimize } X_t^a && (13) \\ & \text{subject to} \\ & \text{Constraints (5)-(12)} \end{aligned}$$

Similarly, the tightest lower bounds on state variables, upper bounds on action and state variables can be obtained by simply replacing the expression in the objective function (13) with Y_t^s , $-X_t^a$, and $-Y_t^s$, respectively. Given the preprocessing optimization problems have the same theoretical complexity as the original learned planning optimization problem (i.e., NP-hard), we limit the computational budget allocated to each preprocessing optimization problem to a fixed amount, and set the lower and upper bounds on the domains of action $a \in A$ and state $s \in S$ decision variables X_t^a , Y_t^s to the best dual bounds found in each respective problem.

4.2.2 ADDITIONAL DECISION VARIABLES

The additional decision variables required to implement our strengthened MILP are as follows:

- $X_t^{+,a}$ and $X_t^{-,a}$ denote the positive and negative values of action variable $a \in A$ at time t , respectively.
- $Y_t^{+,s}$ and $Y_t^{-,s}$ denote the positive and negative values of state variable $s \in S$ at time t , respectively.
- $X_t'^a = 1$ if X_t^a is positive at time t , 0 otherwise.
- $Y_t'^s = 1$ if Y_t^s is positive at time t , 0 otherwise.

4.2.3 ADDITIONAL CONSTRAINTS

The additional constraints in the strengthened MILP are as follows:

$$X_t^a = X_t^{+,a} + X_t^{-,a} \quad (14)$$

$$X_t^a \leq U^a X_t'^a \quad (15)$$

$$X_t^a \geq L^a(1 - X_t'^a) \quad (16)$$

$$X_t^{+,a} \leq U^a X_t'^a \quad (17)$$

$$X_t^{-,a} \geq L^a(1 - X_t'^a) \quad (18)$$

$$\text{for all action variables } a \in A \text{ where } L^a < 0 \text{ and } U^a > 0, \text{ time steps } t \in \{1, \dots, H\} \quad (19)$$

$$Y_t^s = Y_t^{+,s} + Y_t^{-,s} \quad (20)$$

$$Y_t^s \leq U^s Y_t'^s \quad (21)$$

$$Y_t^s \geq L^s(1 - Y_t'^s) \quad (22)$$

$$Y_t^{+,s} \leq U^s Y_t'^s \quad (23)$$

$$Y_t^{-,s} \geq L^s(1 - Y_t'^s) \quad (24)$$

$$\text{for all state variables } s \in S \text{ where } L^s < 0 \text{ and } U^s > 0, \text{ time steps } t \in \{1, \dots, H + 1\}$$

Here, the ranges $[L^a, U^a]$ and $[L^s, U^s]$ are the lower and upper bounds on the domains of action $a \in A$ and state $s \in S$ decision variables X_t^a, Y_t^s , respectively found by solving the preprocessing optimization problems. Given constraints (14)-(24), constraint (25) implements our strengthening constraint which provides a valid upper bound on each ReLU $f \in R$.

$$\begin{aligned} & \sum_{s \in S_g, \tilde{w}_{sg} > 0, L^s \geq 0} \tilde{w}_{sg} Y_t^s + \sum_{s \in S_g, \tilde{w}_{sg} < 0, U^s \leq 0} \tilde{w}_{sg} Y_t^s + \sum_{s \in S_g, \tilde{w}_{sg} > 0, L^s < 0} \tilde{w}_{sg} Y_t^{+,s} \\ & + \sum_{s \in S_g, \tilde{w}_{sg} < 0, L^s < 0} \tilde{w}_{sg} Y_t^{-,s} + \sum_{a \in A_g, \tilde{w}_{ag} > 0, L^a \geq 0} \tilde{w}_{ag} X_t^a + \sum_{a \in A_g, \tilde{w}_{ag} < 0, U^a \leq 0} \tilde{w}_{ag} X_t^a \\ & + \sum_{a \in A_g, \tilde{w}_{ag} > 0, L^a < 0} \tilde{w}_{ag} X_t^{+,a} + \sum_{a \in A_g, \tilde{w}_{ag} < 0, L^a < 0} \tilde{w}_{ag} X_t^{-,a} + \sum_{f \in U_g \cap R, \tilde{w}_{fg} > 0} \tilde{w}_{fg} P_t^f \\ & + \sum_{f \in U_g \cap B} \tilde{w}_{fg} P_t^f \geq P_t^g \end{aligned} \quad (25)$$

$$\text{for all ReLU } g \in R, \text{ time steps } t \in \{1, \dots, H\}$$

With this, we conclude our definition of HD-MILP-Plan — both the base encoding in Section 4.1 and the enhanced encoding in this section for that strengthened the LP relaxation of the big-M constraints for the ReLU activation functions. We will observe the performance of HD-MILP-Plan and the improvement offered by the strengthened encoding in Section 6, but we first define an alternative to HD-MILP-Plan for continuous action planning problems in the next section that offers improved scalability but sacrifices provable optimality.

5. Tensorflow-based Planner (TF-Plan)

While HD-MILP-Plan offers the advantage of provably optimal plans for ReLU deep network learned transition models and discrete and continuous action spaces, we will see in Section 6 that

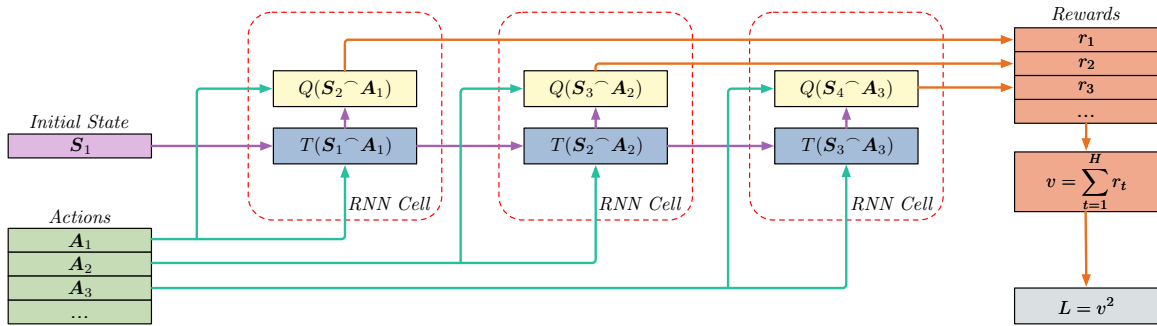


Figure 2: A recurrent neural network (RNN) encoding of a planning problem with continuous state and action parameters: A *known* next-state dependent immediate reward (Q) and *learned* transition function (T) of a discrete time decision-process are embedded in a *custom* RNN cell. RNN cell inputs correspond to the *fixed* starting state and *free variable* actions to be optimized. RNN cell outputs correspond to immediate reward and next state. Rewards for each planning instance are additively accumulated in v . Since the entire specification of objective L is a symbolic representation in Tensorflow with free variable action parameters as inputs, the sequential action plan can be directly optimized via gradient descent using the auto-differentiated representation of L . In practice, we will typically instantiate this RNN with many instances i to support parallel non-convex optimization over multiple initial seeds; here we omit the subscript for the instance i index for notational clarity.

it does face computational constraints that limit its ability to solve large hybrid planning problems. Hence as an alternative to HD-MILP-Plan, in this section we present a planner based on end-to-end auto-differentiation that we call Tensorflow Planner (TF-Plan). TF-Plan represents the planning task as a symbolic recurrent neural network (RNN) architecture with action parameter inputs directly amenable to optimization with GPU-based symbolic toolkits such as Tensorflow and Pytorch.

The TF-Plan we describe here is an extension of our previous Tensorflow-based tool for planning (Wu et al., 2017) with the following key differences: (1) Instead of compiling the hard-coded transition model of the planning domain into a Tensorflow graph, we learn the transition function through a Tensorflow-based neural network as previously described in Section 3 and then freeze the weights of this neural network to obtain a Tensorflow-based model of state transitions. (2) We describe a methodology for handling action constraints used in this work. (3) We evaluate TF-Plan and compare to HD-MILP-Plan on ReLU-based learned neural network transition models that contrast with previous experimental evaluation of TF-Plan on much simpler, manually encoded (non-learned) domain models.

We outline the TF-Plan approach in detail in the following subsections.

5.1 End-to-end Planning through Backpropagation

Backpropagation (Rumelhart et al., 1988) is a standard method for optimizing parameters of deep neural networks via gradient descent. With the chain rule of derivatives, backpropagation propagates the derivative of the output error of a neural network back to each of its parameters in a single linear-time pass in the size of the network using what is known as reverse-mode automatic differentiation (Linnainmaa, 1970). Despite its theoretical efficiency, backpropagation in large-scale deep neural networks in practice is still computationally expensive, and it is only with the advent of recent GPU-based symbolic toolkits that efficient backpropagation at a scale and efficiency required to implement end-to-end planning by gradient descent has become possible.

In planning through backpropagation, we reverse the idea of training the parameters of the network to minimize a loss function over sampled data. Instead, assuming that we have pretrained our deep network transition model, we *freeze* these trained network weights and optimize the inputs (i.e., actions) subject to the fixed transition parameters and an initial state. This end-to-end symbolic planning framework is demonstrated in the form of a recurrent neural network (RNN) as shown in Figure 2; specifically, given learned transition \tilde{T} and reward Q are piecewise differentiable vector-valued functions, we want to optimize the action input vector \mathbf{A}_t for all $t \in \{1, \dots, H\}$ to maximize the accumulated reward value $v = \sum_{t=1}^H r_t$ where $r_t = Q(\mathcal{S}_{t+1} \frown \mathbf{A}_t)$. We remark here that we want to optimize *all* actions $\mathbf{A} = \mathbf{A}_1 \frown \dots \frown \mathbf{A}_H$ with respect to a planning loss L (defined shortly as a function of v) that we minimize via the following gradient update schema

$$\mathbf{A} \leftarrow \mathbf{A} - \eta \frac{\partial L}{\partial \mathbf{A}}, \quad (26)$$

where η is the optimization rate and the partial derivatives comprising the gradient based optimization in problem instance i , where we define the total loss $L = \sum_i L_i$ over multiple planning instances i . To explain why we have multiple planning instances i , we remark that since both transition and reward functions are generally non-convex, optimization on a domain with such dynamics could result in a local minimum or saddle point. To mitigate this problem, we randomly initialize actions for a batch of instances and optimize multiple mutually independent planning instances i in parallel (leveraging the GPU implementation of Tensorflow), and eventually return the best-performing action sequence over all instances.

While there are multiple choices of loss function in Auto-differentiation toolkits, we minimize $L_i = v_i^2$ since the cumulative rewards we test in this paper are at most piecewise linear. We remark that the derivative of a linear function yields a constant value which is not informative in updating actions using the gradient update schema (26). Hence, the optimization of the sum of the squared loss functions L_i has dual effects: (1) it optimizes each problem instance i independently (implemented on the GPU in parallel) and (2) provides fast convergence (i.e., empirically faster than optimizing v_i directly). We remark that simply defining the objective L and the definition of all state variables in terms of predecessor state and action variables via the transition dynamics is enough for auto-differentiation toolkits such as Tensorflow to build the symbolic directed acyclic graph (DAG) of Figure 2 representing the objective L and take its gradient with respect to all free action parameters using reverse-mode automatic differentiation.

5.2 Planning over Long Horizons

The TF-Plan compilation of a planning problem with a neural network transition model shown in Figure 2 reflects the same abstract sequential structure as an RNN that is commonly used in deep learning. However, we critically remark that we have defined *custom* RNN cells based on the known reward model Q and learned transition model T as opposed to using more well-known RNN cells such as the LSTM (Hochreiter & Schmidhuber, 1997) that are intended for general sequential learning purposes and do not reflect the specific planning purpose of the TF-Plan architecture.

Nonetheless, the connection between our RNN representation of hybrid planning in TF-Plan and deep learning with RNNs is not entirely superficial. A longstanding difficulty with training RNNs lies in the vanishing gradient problem, that is, multiplying long sequences of gradients via the chain rule usually renders the gradients extremely small and negligible for weight updates, especially

when using nonlinear activation functions that can be saturated such as a sigmoid. Because TF-Plan must differentiate through all time steps to optimize action plans, it is similarly subject to the vanishing gradient problem.

We mitigate the vanishing gradient issue of the RNN in TF-Plan by training the transition function T of each RNN cell in Figure 2 with ReLU piecewise linear activations, which has a gradient of either 0 (ReLU is inactive) or 1 (ReLU is active) and thus guarantees that the gradient through a path of active ReLUs does not vanish due to the ReLU activations. We note that reward function Q of each RNN cell does not trigger the vanishing gradient problem since the output r_t of each time step t is directly (additively) connected to the loss function.

5.3 Handling Action Bound Constraints

Bounds on actions are common in many planning tasks. For example in the Navigation domain, the distance that the agent can move at each time step is bounded by constant minimum and maximum values. To handle actions with such range constraints in this planning by backpropagation framework, we use projected stochastic gradient descent. Projected gradient descent (PGD) (Calamai & Moré, 1987) is a method that can handle constrained optimization problems by projecting the parameters (action variables) into a feasible range after each gradient update. Precisely, we clip the values of all action variables $a \in A$ to their feasible range $[L^a, U^a]$ after each epoch of gradient descent:

$$\bar{a} \leftarrow \min(\max(\bar{a}, L^a), U^a)$$

In an online planning setting, TF-Plan only ensures the feasibility of actions with bound constraints using PSGD at time step $t = 1$. TF-Plan does not enforce the remaining global (action) constraints C and goal constraints G during planning. In general, effectively handling arbitrary constraints in the Tensorflow framework is an open research question beyond the scope of this paper. Nonetheless, action bounds are among the most common constraints (i.e., the ones we use experimentally) and TF-Plan’s initial time step handling of these constraints via PSGD is sufficient to guarantee that only feasible actions are taken during online planning with TF-Plan.

6. Experimental Results

In this section, we present experimental results that empirically test the performance of both HD-MILP-Plan and TF-Plan on multiple planning domains with learned neural network transition models. These experiments focus on continuous action domains since the intent of the paper is to compare the performance of HD-MILP-Plan to TF-Plan on domains where they are both applicable. To accomplish this task we first present three nonlinear continuous action benchmark domains, namely: Reservoir Control, Heating, Ventilation and Air Conditioning, and Navigation. Then, we validate the transition learning performance of our proposed ReLU-based densely-connected neural networks with different network configurations in each domain. Finally we evaluate the efficacy of both proposed planning frameworks based on the learned model by comparing them to strong baseline manually coded policies¹⁰ in an online planning setting. For HD-MILP-Plan, we test the effect of preprocessing to strengthened MILP encoding on run time and solution quality. For TF-Plan, we investigate the impact of the number of epochs on planning quality. Finally, we test the scalability

10. As noted in the Introduction, MCTS and model-free reinforcement learning are not applicable as baselines given our multi-dimensional concurrent *continuous* action spaces.

of both planners on large scale domains and show that TF-Plan can scale much more gracefully compared to HD-MILP-Plan.

6.1 Illustrative Domains

Full RDDDL (Sanner, 2010) specifications of all domains and instances defined below and used for data generation and plan evaluation in the experimentation are listed in Appendix B.

Reservoir Control has a single state variable $l^r \in \mathbb{R}$ for each reservoir, which denotes the water level of the reservoir r and a corresponding action variable for each r to permit a flow $f^r \in [0, r^{\max}]$ from reservoir r (with maximum allowable flow r^{\max}) to the next downstream reservoir. The transition is a nonlinear function due to the evaporation e^r from each reservoir r , which is defined by the formula

$$e_t^r = (1.0/2.0) \cdot \sin((1.0/2.0) \cdot l_t^r) \cdot 0.1,$$

and the water level transition function is

$$l_{t+1}^r = l_t^r + \sum_{r_{up}} f^{r_{up}} - f_t^r - e_t^r,$$

where $f^{r_{up}}$ ranges over all upstream reservoirs of r with bounds $0 \leq f^r \leq l^r$. The reward function minimizes the total absolute deviation from a desired water level, plus a constant penalty for having water level outside of a safe range (close to empty or overflowing), which is defined for each time step t by the expression

$$- \sum_r (0.1 \cdot \left| ((m^r + n^r)/2.0) - l_{t+1}^r \right| + 100 \cdot \max(m^r - l_{t+1}^r, 0) + 5 \cdot \max(l_{t+1}^r - n^r, 0)),$$

where m^r and n^r define the upper and lower desired ranges for each reservoir r . We report the results on small instances with 3 and 4 reservoirs over planning horizons $H = 10, 20$, and large instances with 10 reservoirs over planning horizons $H = 10, 20$.

Heating, Ventilation and Air Conditioning (Agarwal et al., 2010) has a state variable $p^r \in \mathbb{R}$ denoting the temperature of each room r and an action variable $b^r \in [0, b^{\max}]$ for sending heated air to each room r (with maximum allowable volume b^{\max}) via vent actuation. The bilinear transition function is then

$$p_{t+1}^r = p_t^r + (\Delta t / C^r)(b^r + \sum_{r'} (p_t^{r'} - p_t^r) / R^{rr'}),$$

where C^r is the heat capacity of rooms, r' represents an adjacency predicate with respect to room r and $R^{rr'}$ represents a thermal conductance between rooms. The reward function minimizes the total absolute deviation from a desired temperature for all rooms plus a linear penalty for having temperatures outside of a range plus a linear penalty for heating air with cost k , and is defined for each time step t by the expression

$$- \sum_r (10.0 \cdot \left| ((m^r + n^r)/2.0) - p_t^r \right| + k b^r + 0.1 \cdot (\max(p_t^r - n^r, 0) + \max(m^r - p_t^r, 0)).$$

We report the results on small instances with 3 and 6 rooms over planning horizons $H = 10, 20$, and a large instance with 60 rooms over planning horizon $H = 2$.

Navigation is designed to test learning of a highly nonlinear transition function and has a vector of state variables for the 2D location of an agent $\mathbf{P}_{t+1} = (p_t^x, p_t^y)$ and a 2D action variable vector intended nominally to move the agent $\Delta\mathbf{P}_t = (\Delta p_t^x, \Delta p_t^y)$ (with minimum and maximum movement boundaries $[\Delta p^{\min}, \Delta p^{\max}]$). The new location \mathbf{P}_{t+1} is a nonlinear function of the current location \mathbf{P}_t (with minimum and maximum maze boundaries $[p^{\min}, p^{\max}]$) due to higher slippage in the center of the domain where the transition function is

$$\mathbf{P}_{t+1} = \mathbf{P}_t + \Delta\mathbf{P}_t \cdot 2.0 / (1.0 + \exp(-2 \cdot \Delta d_t^p)) - 0.99.$$

where Δd_t^p is the Euclidean distance from \mathbf{P}_t to the center of the domain. The reward function minimizes the total Manhattan distance from the goal location, which is defined for each time step t by the expression

$$- \sum_{d \in \{x, y\}} |g^d - p_t^d|,$$

where g^d defines the goal location for dimension d . We report the results on small instances with maze sizes 8-by-8 (i.e., $p^{\min} = 0, p^{\max} = 8$) and 10-by-10 (i.e., $p^{\min} = 0, p^{\max} = 10$) over planning horizons $H = 8, 10$, and a large instance with minimum and maximum movement boundaries $[\Delta p^{\min} = -0.5, \Delta p^{\max} = 0.5]$ over planning horizon $H = 20$.

6.2 Transition Learning Performance

In Table 1, we show the held-out test data mean squared error (MSE) of the best configuration of neural network architectures for three instances of the previously defined planning domains. We train all neural networks using 10^5 data samples from simulation using a simple stochastic exploration policy. As standard for deep network training, we randomly permute the data order for each epoch of gradient descent. 80% of the sampled data was used for training — with hyperparameters tuned on a subset of 20% validation data of the training data — and 20% of the sampled data was held out for the test evaluation. We applied the RMSProp (Hinton et al., 2012) optimizer over 200 epochs. Since densely-connected networks (Huang et al., 2016) discussed and illustrated in Section 3.1 strictly dominated the performance of non-densely-connected networks, we only report the results of the densely-connected network. Throughout all experiments, we fixed the dropout parameter $p = 0.1$ (cf. Section 3.3) at *all* hidden layers since deviations from this value rarely improved generalization performance by a significant amount. We tuned the number of hidden layers in the set of $\{0$ (linear), $1, 2\}$ and the number of neurons for each layer in the set of $\{8, 16, 32, 64, 128\}$; for a given layer size, we chose the minimal number of neurons that performed best on validation data within statistical significance.

Overall, we see that Reservoir and HVAC can be accurately learned with one hidden layer since an additional layer hurt generalization performance on the test data. Navigation benefits from having two layers owing to the complexity of its nonlinear transition and approximation with ReLU activations. The network with the lowest MSE is used as the deep neural network model for each domain in the subsequent planning experiments.

Figure 3 visualizes the training performance of different neural network configurations over three domain instances. Figures 3 (a)-(c) visualize the loss curves over training epochs for three

Table 1: Mean Squared Error (MSE) Table for all domains and network configurations with 95% confidence intervals; three significant digits are shown for MSE.

Domain	Linear	1 Hidden	2 Hidden
Reservoir (instance with 4 reservoirs)	46500000 \pm 487000	343000 \pm 7210	653000 \pm 85700
HVAC (instance with 3 rooms)	710 \pm 2.3	520 \pm 54	75200 \pm 7100
Navigation (instance with 10 by 10 maze)	30400 \pm 9.8	9420 \pm 29	1940 \pm 50

domain instances. In Reservoir, we observe that while both 1 and 2 hidden layer networks have similar MSE values, the former has much smaller variance. In HVAC and Navigation instances, we observe that 1 and 2 hidden layer networks have the smallest MSE values, respectively. Figures 3 (d)-(f) visualize the performance of learning transition functions with different number of hidden layers. We observe that Reservoir needs at least one hidden layer to overlap with ground truth whereas HVAC is learned well with all networks (all dashed lines overlap) and Navigation only shows complete overlap (especially near the center nonlinearity) for a two layered neural network. All of these results mirror the MSE comparisons in Table 1 thus providing both empirical and intuitive evidence justifying the best neural network structure selected for each domain.

6.3 Planning Performance

In this section, we investigate the effectiveness of planning with HD-MILP-Plan and TF-Plan to plan for the original planning problem Π through optimizing the learned planning problem $\tilde{\Pi}$ in an online planning setting. We optimized the MILP encodings using IBM ILOG CPLEX 12.7.1 with eight threads and a 1-hour total time limit per problem instance on a MacBookPro with 2.8 GHz Intel Core i7 16 GB memory. We optimize TF-Plan through Tensorflow 1.9 with an Nvidia GTX 1080 GPU with CUDA 9.0 on a Linux system with 16 GB memory.¹¹ We connected both planners with the RDDLSim (Sanner, 2010) domain simulator and interactively solved multiple problem instances with different sizes and horizon lengths (as described in Figure 4). In order to assess the approximate solution quality of TF-Plan, we also report HD-MILP-Plan with 20% duality gap. The results reported for TF-Plan, unless otherwise stated, are based on fixed number of epochs for each domain where TF-Plan used 1000 epochs for Reservoir and HVAC, and 300 epochs for Navigation.

6.3.1 COMPARISON OF PLANNING QUALITY

In Figures 4 (a)-(c), we compare the planning qualities of the domain-specific policies (blue), the base MILP model (gray), MILP model with preprocessing and strengthening constraints solved optimally (orange), MILP model with preprocessing and strengthening constraints solved up to 20% duality gap (green) and TF-Plan.

In Figure 4 (a), we compare HD-MILP-Plan and TF-Plan to a rule-based local Reservoir planner, which measures the water level in reservoirs, and sets outflows to release water above a pre-specified median level of reservoir capacity. In this domain, we observe an average of 15% increase in the total reward obtained by the plans generated by HD-MILP-Plan in comparison to that of the rule-based local Reservoir planner. Similarly, we find that TF-Plan outperforms the rule-based local

11. Due to the fact that CPLEX and Tensorflow leverage different hardware components (GPUs in the case of Tensorflow), the run-times reported are machine-specific and may vary with other machine architecture configurations and components.

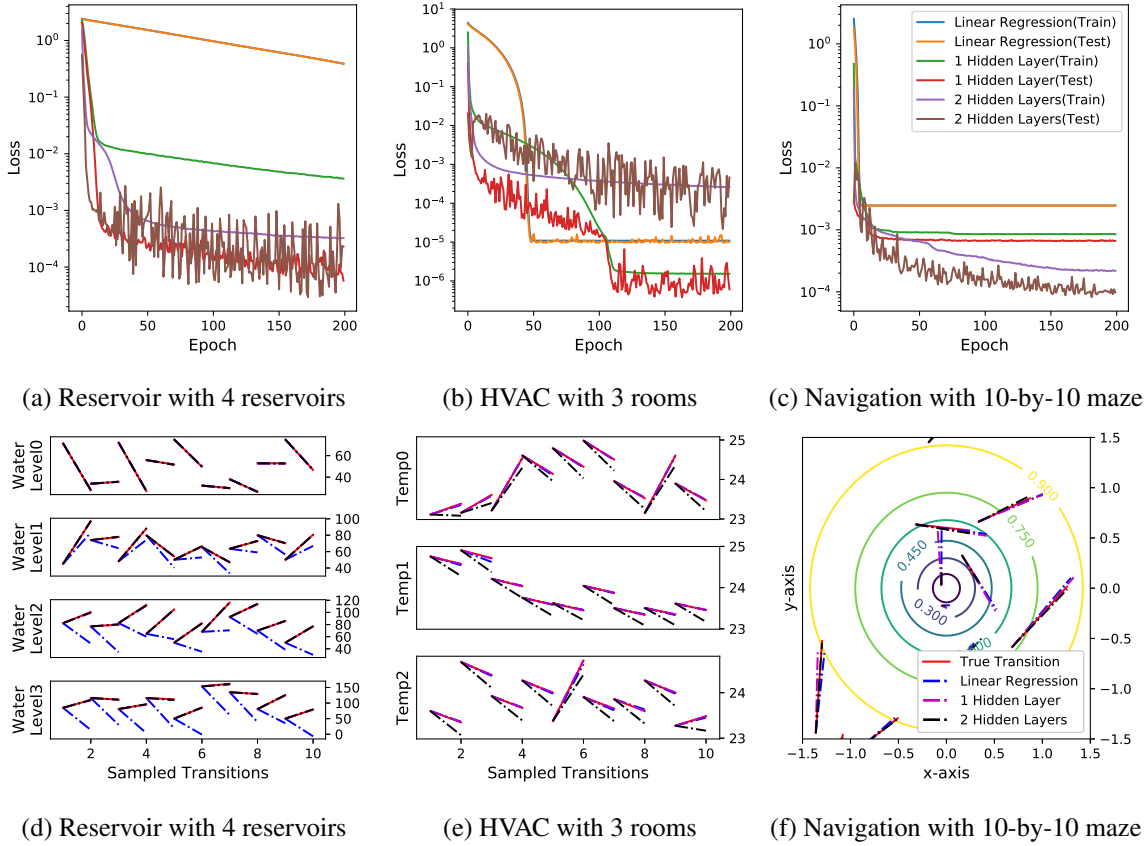


Figure 3: Visualization of training performance of different neural network configurations over three domain instances. Figures (a)-(c) represent the loss curves over training epochs for the three domains we introduced in the paper. The y-axis is in logarithmic scale. In figure (a), we observe that while MSE of 1 hidden layer and 2 hidden layer networks look similar, the former has smaller variance. In figures (b)-(c), we observe that 1 hidden layer network and 2 hidden layer network have smallest MSEs, respectively. Figures (d)-(f) represent the learning quality comparisons of linear, 1 hidden layer and 2 hidden layer networks for 3 domains. For different domain variables, we show different starting points and true transitions after a period of time. The dashed lines are the different neural network learned approximations.

Reservoir planner with a similar percentage though does not always beat HD-MILP-Plan. However, we observe that TF-Plan outperforms both the rule-based policy and HD-MILP-Plan on the larger Reservoir 4 domains. We investigate this outcome further in Figure 5 (a). We find that the plan returned by HD-MILP-Plan incurs more penalty due to the noise in the learned transition model, where the plan attempts to distribute water to multiple reservoirs and obtain higher reward. As a result, the actions returned by HD-MILP-Plan break the safety threshold and receive additional penalty. Thus, HD-MILP-Plan incurs more cost than TF-Plan.

In Figure 4 (b), we compare HD-MILP-Plan and TF-Plan to a rule-based local HVAC policy, which turns on the air conditioner anytime the room temperature is below the median value of a given range of comfortable temperatures [20,25] and turns off otherwise. While the reward (i.e., electricity cost) of the proposed models on HVAC 3 rooms are almost identical to that of the locally optimal HVAC policy, we observe significant performance improvement on HVAC 6 rooms settings,

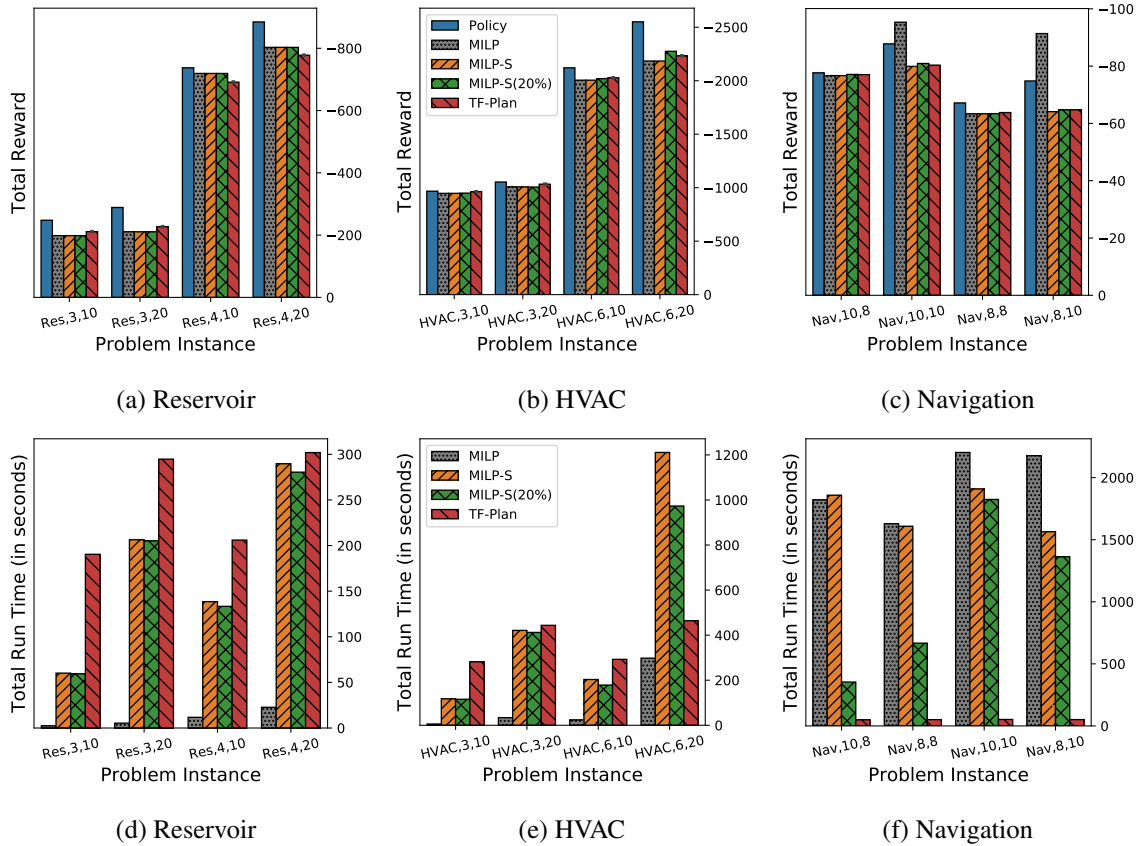


Figure 4: Overall planning performance comparison. The domain notation shown in the bar labels of each figure correspond to (DOMAIN NAME, SIZE, HORIZON). (a)-(c) visualizes the total *negative* reward comparison (lower is better) between the proposed methods and domain-specific rule-based planning. In addition, we show results for HD-MILP-Plan solved up to 20% duality gap to provide a reference to the solution quality of TF-Plan. Note that rewards in this paper represent costs to minimize and therefore *lower* total negative reward indicates *better* performance. The handcoded policies were strong baselines intended to be near-optimal, but we see greater performance separation as the domains become more nonlinear (most notably Reservoir and Navigation) and the optimal policies become harder to manually encode. Figures (d)-(f) visualizes the timing comparison among base MILP algorithm (gray bar), the strengthened MILP algorithm (orange bar), strengthened MILP with 20% gap (green bar), and the TF-Plan (red bar). As the domains become more nonlinear (i.e., as measured by the learning quality of each domain as presented in Table 1) and the deep network depth increases, the strengthened MILP encoding begins to dominate the base encoding. Deep network depth impacts performance of the base MILP more than problem size.

which suggests the advantage of the proposed models on complex planning problems, where the manual policy fails to track the temperature interaction among the rooms. Figure 5 (b) further demonstrates the advantage of our planners where the room temperatures controlled by the proposed models are identical to the locally optimal policy with 15% less power usage.

Figure 4 (c) compares HD-MILP-Plan and TF-Plan to a greedy search policy, which uses a Manhattan distance-to-goal function to guide the agent towards the direction of the goal (as visualized by Figure 5 (c)). The pairwise comparison of the total rewards obtained for each problem

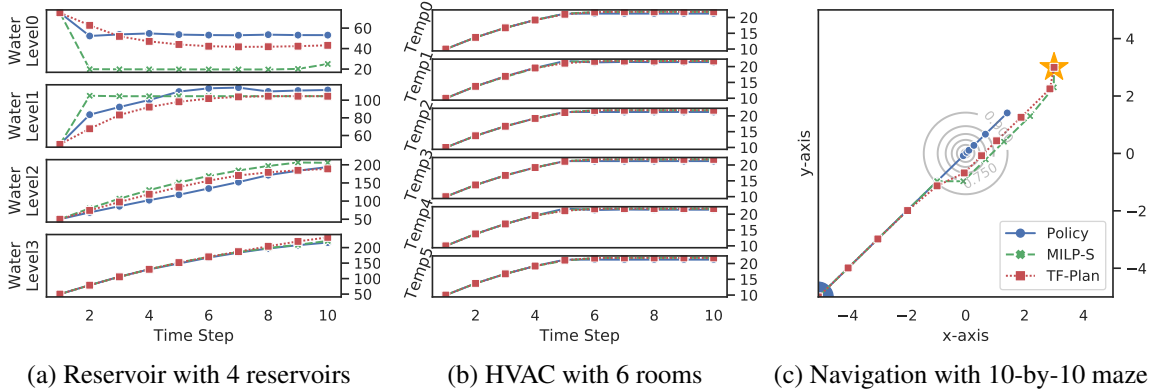


Figure 5: Planning behavior comparison between the manually encoded rule-based policy (Blue), HD-MILP-Plan (Green) and TF-Plan (Red). Compared to the strong manually coded policies, both HD-MILP-Plan and TF-Plan make more subtle nonlinear deviations in comparison to the manual policy (Blue) to better optimize the overall objective as shown in Figures 4 (a)-(c).

instance per plan shows that the proposed models can outperform the manual policy up to 15%, as observed in the problem instance Navigation,10,8 in Figure 4 (c). The investigation of the actual plans, as visualized by Figure 5 (c), shows that the local policy ignores the nonlinear region in the middle, and tries to reach the goal directly, which cause the plan to be not able to reach the goal position with given step budget. In contrast, both HD-MILP-Plan and TF-Plan can find plans that move around the nonlinearity and successfully reach the goal state, which shows their ability to model the nonlinearity and find plans that are near-optimal with respect to the learned model over the complete horizon H .

Overall we observe that in 10 out of 12 problem instances, the solution quality of the plans generated by HD-MILP-Plan and TF-Plan are significantly better than the total reward obtained by the plans generated by the respective domain-specific human-designed policies. Further we find that the quality of the plans generated by TF-Plan are between the plans generated by i) HD-MILP-Plan solved to optimality, and ii) HD-MILP-Plan solved to 20% duality gap, thus indicating the overall strong approximate performance of TF-Plan.

6.3.2 COMPARISON OF RUN TIME PERFORMANCE

In Figures 4 (d)-(f), we compare the run time performances of the base MILP model (gray), MILP model with preprocessing and strengthening constraints solved optimally (orange), MILP model with preprocessing and strengthening constraints solved upto 20% duality gap (green) and TF-Plan. Figure 4 (f) shows significant run time improvement for the strengthened encoding over the base MILP encoding, while Figures 4 (d)-(e) show otherwise. Together with the results presented in Figures 4 (a)-(c), we find that domains that utilize neural networks with only 1 hidden layer (e.g., HVAC and Reservoir) do not benefit from the additional fixed computational expense of preprocessing. In contrast, domains that require deeper neural networks (e.g., Navigation) benefit from the additional computational expense of preprocessing and strengthening. Over three domains, we find that TF-Plan significantly outperforms HD-MILP-Plan in all Navigation instances, performs slightly worse in all Reservoir instances, and performs comparable in HVAC instances.

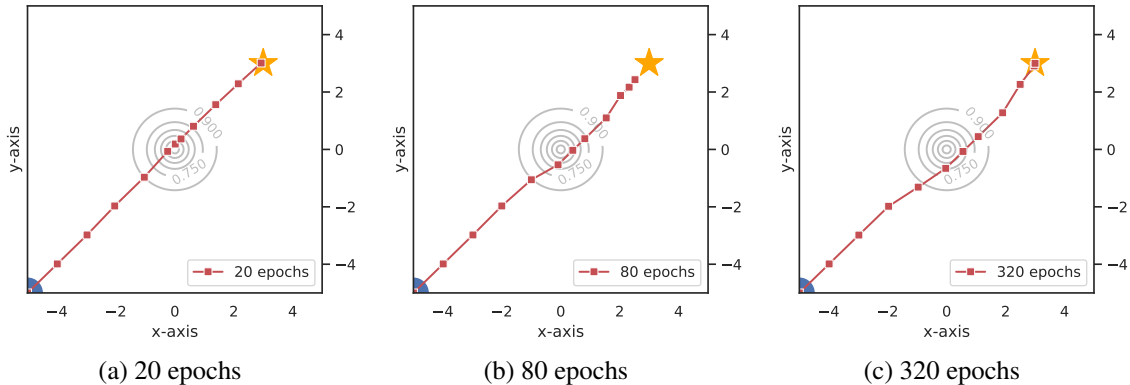


Figure 6: Performance comparison using differing numbers of epochs when optimizing with TF-Plan on the Navigation domain. (a) 20 epochs allows the model to find a brute force solution that finds a direct, but highly suboptimal path to the goal that passes through the heart of the central deceleration zone. (b) 80 epochs allows the model to avoid the deceleration zone, though it takes suboptimally large or small steps. (c) 320 epochs shows a near-optimal result where the agent skirts the deceleration zone and spaces its steps fairly equally. Furthermore, 320 epochs reaches the goal within 10 steps while 20 epochs requires 11 steps.

6.3.3 EFFECT OF TRAINING EPOCHS FOR TF-PLAN ON PLANNING QUALITY

To test the effect of the number of optimization epochs on the solution quality, we present results on 10-by-10 Navigation domain for a horizon of 10 with different epochs. Figure 6 visualizes the increase in solution quality as the number of epochs increases where Figure 6 (a) presents a low quality plan found similar to that of the manual-policy with 20 epochs, Figure 6 (b) presents a medium quality plan with 80 epochs, and Figure 6 (c) presents a high quality plan similar to that of HD-MILP-Plan with 320 epochs.

6.3.4 SCALABILITY ANALYSIS ON LARGE PROBLEM INSTANCES

To test the scalability of the proposed planning models, we create three additional domain instances that simulate more realistic planning instances. For the Reservoir domain, we create a system with 10 reservoirs with complex reservoir formations, where a reservoir may receive water from more than one upstream reservoirs. For the HVAC domain, we simulate a building of 6 floors and 60 rooms with complex adjacencies between neighboring rooms and across building levels. Moreover, in order to fully capture the complex mutual temperature impact of the rooms, we train a *large* transition function with one hidden layer of width 256. For the Navigation domain, we reduce the feasible action range from $[-1, 1]$ to $[-0.5, 0.5]$ and increase the planning horizon to 20 time steps.

In Figures 7 (a)-(c), we compare the total rewards obtained by the domain-specific rule-based policy (blue), HD-MILP-Plan (orange) and TF-Plan (red) on larger problem instances. The analysis of Figures 7 (a)-(c) shows that TF-Plan scales better compared to HD-MILP-Plan by consistently outperforming the policy, whereas HD-MILP-Plan outperforms the other two planners in two out of three domains (i.e., Reservoir and HVAC) while suffering from scalability issues in one domain (i.e., Navigation). Particularly, we find that in Navigation domain, HD-MILP-Plan sometimes does not find feasible plans with respect to the learned model and therefore returns default no-op action values leading to its poor observed performance in this domain.

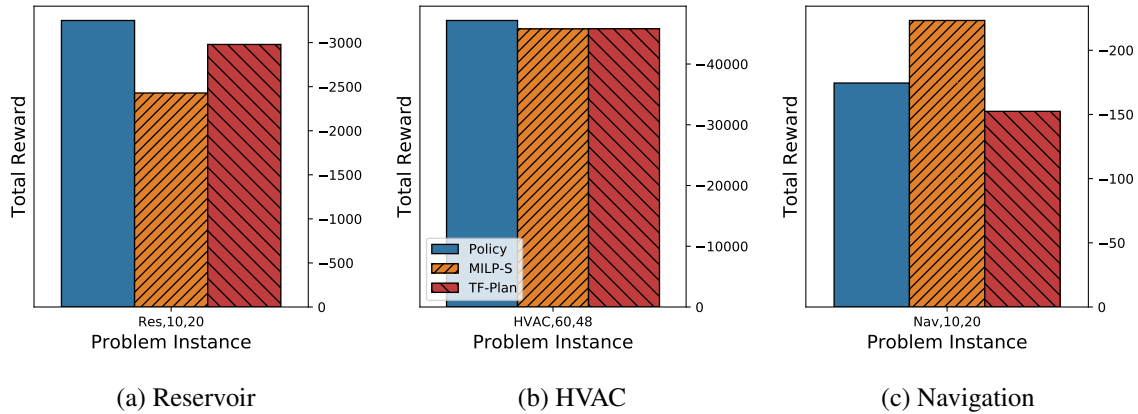


Figure 7: The total reward comparison between the proposed methods and the (strong baseline) domain-specific rule-based planning on larger problem instances. Since the y-axis is a negative cost, lower (less cost) is better. In particular, we show TF-Plan (red) scales better by consistently finding plans better than the rule-based policy (blue). Further, we show HD-MILP-Plan (orange) outperforms the other planners in two out of three domains (i.e., Reservoir and HVAC), but performs poorly in the Navigation domain due to its deeper neural network structure and longer planning horizon that requires a large MILP encoding.

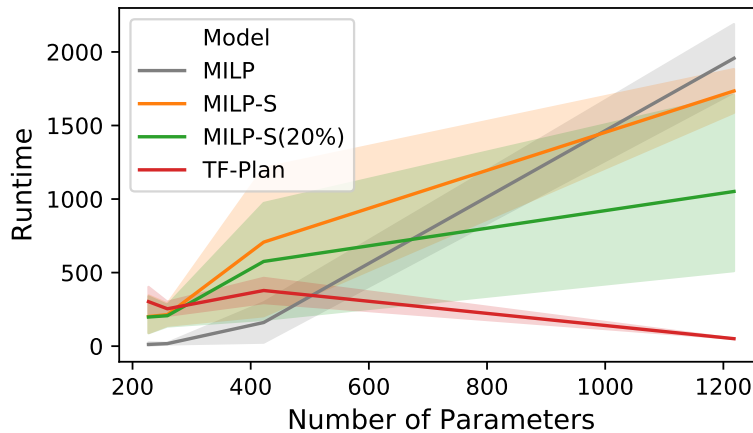


Figure 8: Visualization of run times as a function of problem sizes (specifically the number of variables in the encoding), where problem size is a function of horizon H , number of parameters in the learned model, and number of neural network layers. 95% confidence intervals on running times are shown.

In Figure 8, we compare the run-time performance of the proposed planning systems over all problem instances vs. problem size. We measure the problem size as the number of variables used in the encoding, which is a function of horizon H , number of parameters in the learned model, and number of neural network layers. We observe that as the problem size gets larger, HD-MILP-Plan takes more computational effort to solve due to the additional overhead of proving optimality, which can be mitigated by relaxing the optimality guarantee to a bound (e.g., the 20% duality gap of MILP-S(20%)). We also observe that as the problem sizes get larger, the effect of preprocessing bounds and strengthening constraints of the MILP-S variants pay off. Finally, in observing that the MILP-based running times appear to increase roughly linearly in the problem size, it is critical to note that

in contrast to classical discrete planning problems where running time often scales exponentially as a function of the encoding size, the scalability of MILP-based optimization can often be highly dependent on specific aspects of the problem encoding that lead to subexponential growth rates as a function of the number of parameters. In general, it can be difficult to predict MILP optimization time as a simple function of the number of parameters being optimized.

If we examine TF-Plan in Figure 8, we observe that it scales gracefully as the problem size gets larger — while these results surprisingly show that TF-Plan is fastest on the largest problem, we note that due to its highly parallel GPU-based implementation, adding more parameters does not necessarily linearly increase the running time of TF-Plan. Furthermore, we remark that (like MILPs) continuous optimization running times can be hard to predict as a function of problem size — small continuous problems may sometimes be very difficult to optimize while other larger continuous problems may have simpler optimization surfaces that can be optimized much more efficiently. Together, these two reasons can help explain the somewhat surprisingly low running time of TF-Plan for the largest problem sizes. In summary, if we also revisit Figures 7 (a)-(c), we observe that TF-Plan can provide a highly efficient alternative to HD-MILP-Plan in large-scale planning problems.

7. Related Work

In this section, we discuss the existing automated planning literature in relation to our data-driven planners. In this work, we have focused on planning with ReLU-based DNN learned state transition models subject to the optimization of general piecewise linear reward functions over concurrent instantaneous actions with parameters that have real (continuous) domains. Many existing PDDL-based (Penna et al., 2009; Coles et al., 2013; Ivankovic et al., 2014; Bryce et al., 2015; Piotrowski et al., 2016; Scala et al., 2016a, 2016b; Cashmore et al., 2016) and RDDDL-based (Keller & Helmert, 2013) planners focus on models with action parameters that are strictly finite (discrete). In order to represent more general problems, the PDDL formalism has been extended to handle action variables that have real domains (i.e., real-valued control parameters) with the focus of synthesizing partially ordered plans (Sava et al., 2016). While the extension of the PDDL formalism to real-valued action parameters is an important step forward towards the domain expressiveness we handle in this article, more work would be needed to show how a multilayer ReLU-based neural network transition model with piecewise nonlinearities and feedforward computations at hidden layers could be compactly and efficiently encoded in PDDL to apply such a PDDL-based planner.

In a different vein, hybrid automaton-based Domain Predictive Control can model state transition functions with linear dynamics over action parameters with real domains (Löhr et al., 2012). While Domain Predictive Control could technically be used to model ReLU-based state transition functions using linear state transitions, the number of mode switches required to represent each activation pattern of the learned DNN would be exponential in the number of ReLUs per time step which makes this approach ill-suited for solving the neural net transition planning problems considered in this work. As a further expressivity extension, hybrid planning problems with continuous (time) state transition models can be solved using a bounding technique known as *flow tubes* but this approach requires the state transition function to be modeled as ordinary differential equations (Li & Williams, 2008), which are not well-suited to our learned neural network transition models. Because techniques from robotics are largely specialized for the physics and geometry of those particular problems, we are not aware of techniques in robotics that can plan for arbitrary ReLU-based

deep neural network state transition models as we contribute here. As such, we conclude that our data-driven planners for learned neural net models uniquely address a complex problem related to existing literature in the field of domain-independent automated planning methods.

8. Concluding Remarks and Future Work

In this paper, we have tackled the question of how we can plan with expressive and accurate deep network learned transition models that are not amenable to existing solution techniques. We started by focusing on how to learn accurate learned transition functions by using densely-connected deep neural networks trained with a reweighted mean square error (MSE) loss. We then leveraged the insight that ReLU-based deep networks offer strong learning performance and permit a direct compilation of the neural network transition model to a Mixed-Integer Linear Program (MILP) encoding in a planner we called Hybrid Deep MILP Planner (HD-MILP-Plan). To further enhance planning efficiency, we strengthened the linear relaxation of the base MILP encoding. Finally, as a more efficient but not provably optimal alternative to MILP-based optimization, we proposed the end-to-end gradient optimization-based Tensorflow Planner (TF-Plan) that encodes the problem in the form of a Recurrent Neural Network (RNN) and directly optimizes plans via backpropagation.

We evaluated run-time performance and solution quality of the plans generated by both proposed planners over multiple problem instances from three diverse continuous state and action planning domains. It would be hard to definitively characterize the general comparative performance behavior of HD-MILP-Plan vs. TF-Plan for arbitrary problems; however, on these particular domains, we have shown that HD-MILP-Plan can find optimal plans with respect to the learned models, and TF-Plan can approximate the optimal plans with little computational cost. We have shown that the plans generated by both HD-MILP-Plan and TF-Plan yield better solution qualities compared to strong domain-specific human-designed policies and that TF-Plan performance generally improves with the number of optimization epochs. Also, we have shown that our strengthening constraints improved the solution quality and the run-time performance of HD-MILP-Plan as problem instances got larger. Finally, we have shown that TF-Plan can handle large-scale planning problems with very little relative computational cost compared to the MILP-based optimization approach of HD-MILP Plan and its variants defined in Section 4.

In terms of future work, while TF-Plan showed strong performance and scalability for large problems evaluated in this article, it does have two important weaknesses compared to HD-MILP-Plan that provide avenues for additional research. First, while HD-MILP-Plan in principle has no problem working with discrete actions — learning with discrete actions requires no special encodings over that previously given and MILP solvers directly support discrete variables — the only way to currently handle discrete actions in TF-Plan would be through a continuous relaxation. Second, while goal and state constraints are straightforward to handle in the HD-MILP-Plan framework since these are literally just additional constraints in the encoding, Tensorflow is generally not intended for constrained optimization. To this end, future work should consider how TF-Plan can be effectively extended to handle discrete actions and general constraints. Finally, we remark that this article did not address discrete states, with a key caveat simply being that effective learning of deterministic deep neural networks with discrete output nodes remains an open area of research investigation in deep neural networks. That said, we note that some recent work on planning with learned binarized neural network (BNN) transition models using MILP, SAT and pseudo-Boolean

encodings (Say & Sanner, 2018, 2020; Say et al., 2020) does pose one direction for future research with discrete states in combination with the contributions of this work focused on continuous states.

In conclusion, both HD-MILP-Plan and TF-Plan represent *a new class of data-driven planning methods* that can accurately learn complex state transitions of high-dimensional nonlinear continuous state and action planning domains, and provide high-quality plans with respect to these learned models. Further, we believe that this work paves the way for future research that extends HD-MILP-Plan and TF-Plan to learning and optimizing in transition models with discrete states as well as further extensions to TF-Plan for handling discrete actions and general constraints in a highly general and highly scalable Tensorflow-based planner.

Acknowledgements

Preliminary versions of this article appeared in conference papers (Say et al., 2017) and (Wu et al., 2017). Both Ga Wu and Buser Say contributed equally to this work. For Buser Say, the majority of the work was done while he was at the University of Toronto and affiliated with the Vector Institute, and only the final revision of the article was done while the he was at the Monash University. This work was supported by an NSERC Discovery Grant and an Ontario Early Researcher Award.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems..
- Agarwal, Y., Balaji, B., Gupta, R., Lyles, J., Wei, M., & Weng, T. (2010). Occupancy-driven energy management for smart building automation. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, pp. 1–6.
- Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *JAIR*, 11(1), 1–94.
- Bryce, D., Gao, S., Musliner, D., & Goldman, R. (2015). SMT-based nonlinear PDDL+ planning. In *29th AAAI*, pp. 3247–3253.
- Calamai, P. H., & Moré, J. J. (1987). Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1), 93–116.
- Cashmore, M., Fox, M., Long, D., & Magazzeni, D. (2016). A compilation of the full PDDL+ language into SMT. In *ICAPS*, pp. 79–87.
- Coles, A., Fox, M., & Long, D. (2013). A hybrid lp-rpg heuristic for modelling numeric resource flows in planning. *Journal of Artificial Intelligence Research*, 46, 343–412.
- Coulom, R. (2006). Efficient selectivity and backup operators in monte-carlo tree search. In *International Conference on Computers and Games*, pp. 72–83. Springer Berlin Heidelberg.
- Faulwasser, T., & Findeisen, R. (2009). Nonlinear Model Predictive Path-Following Control. In *Nonlinear Model Predictive Control - Towards New Challenging Applications*, pp. 335–343. Springer, Berlin, Heidelberg.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*, Vol. 1.

- Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a: overview of mini-batch gradient descent..
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks.. In *CVPR*.
- Ivankovic, F., Haslum, P., Thiebaux, S., Shivashankar, V., & Nau, D. (2014). Optimal planning with global numerical state constraints. In *ICAPS*, pp. 145–153.
- Keller, T., & Helmert, M. (2013). Trial-based heuristic tree search for finite horizon MDPs. In *ICAPS*, pp. 135–143.
- Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. In *ECML*, pp. 282–293.
- Kveton, B., Hauskrecht, M., & Guestrin, C. (2006). Solving factored mdps with hybrid state and action variables. *Journal of Artificial Intelligence Research*, 27, 153–201.
- Li, F.-F., Johnson, J., & Yeung, S. (2018). Cs231n: Convolutional neural networks for visual recognition lecture note..
- Li, H. X., & Williams, B. C. (2008). Generative planning for hybrid systems based on flow tubes. In *Proceedings of the Eighteenth International Conference on Automated Planning and Scheduling, ICAPS 2008, Sydney, Australia, September 14-18, 2008*, pp. 206–213.
- Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Masters Thesis (in Finnish), Univ. Helsinki*, 6–7.
- Löhr, J., Eyerich, P., Keller, T., & Nebel, B. (2012). A planning based framework for controlling hybrid systems. In *ICAPS*, pp. 164–171.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*.
- Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*, pp. 807–814.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch..
- Penna, G. D., Magazzeni, D., Mercurio, F., & Intrigila, B. (2009). UPMurphi: A tool for universal planning on PDDL+ problems. In *ICAPS*, pp. 106–113.
- Piotrowski, W. M., Fox, M., Long, D., Magazzeni, D., & Mercurio, F. (2016). Heuristic planning for hybrid systems. In *AAAI*, pp. 4254–4255.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5, 3.
- Sanner, S. (2010). Relational dynamic influence diagram language (rddl): Language description..
- Sava, E., Fox, M., Long, D., & Magazzeni, D. (2016). *Planning using actions with control parameters*, Vol. 285 of *Frontiers in Artificial Intelligence and Applications*, pp. 1185–1193. IOS Press.

- Say, B., Devriendt, J., Nordström, J., & Stuckey, P. (2020). Theoretical and experimental results for planning with learned binarized neural network transition models. In *Proceedings of the Twenty-Sixth International Conference on Principles and Practice of Constraint Programming*.
- Say, B., & Sanner, S. (2018). Planning in factored state and action spaces with learned binarized neural network transition models. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI'18*, pp. 4815–4821.
- Say, B., & Sanner, S. (2020). Compact and efficient encodings for planning in factored state and action spaces with learned binarized neural network transition models. *Artificial Intelligence*, 285, 103291.
- Say, B., Wu, G., Zhou, Y. Q., & Sanner, S. (2017). Nonlinear hybrid planning with deep net learned transition models and mixed-integer linear programming. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI'17*, pp. 750–756.
- Scala, E., Haslum, P., Thiebaux, S., & Ramirez, M. (2016a). Interval-based relaxation for general numeric planning. In *IOS Press*.
- Scala, E., Ramirez, M., Haslum, P., & Thiebaux, S. (2016b). Numeric planning with disjunctive global constraints via SMT. In *ICAPS*, pp. 276–284.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587), 484.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 1929–1958.
- Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning* (1st edition). MIT Press.
- Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool.
- Weinstein, A., & Littman, M. L. (2012). Bandit-based planning and learning in continuous-action markov decision processes.. In *ICAPS*.
- Wu, G., Say, B., & Sanner, S. (2017). Scalable planning with tensorflow for hybrid nonlinear domains. In *Proceedings of the Thirty First Annual Conference on Advances in Neural Information Processing Systems (NIPS-17)*, Long Beach, CA.
- Yeh, W. G. (1985). Reservoir management and operations models: A state-of-the-art review. *Water Resources research*, 21,12, 17971818.

Appendix A. Post-learning Neural Network Weight Modification

As outlined in Section 3.2, we standardize input $\bar{\mathbf{X}} = \bar{\mathbf{S}}_n \wedge \bar{\mathbf{A}}_n$ before feeding it to the neural network. Let $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\sigma}}$ respectively denote the vector of means and standard deviations for $\bar{\mathbf{X}}$ empirically calculated from the training data. Then the normalized inputs $\hat{\bar{\mathbf{X}}}$ used to train the deep neural network are defined as

$$\hat{\bar{\mathbf{X}}} = (\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}) \cdot \bar{\boldsymbol{\sigma}}^{-1}.$$

Now, let \mathbf{W} denote the vector of weights connected to a hidden unit of a neural network, let b indicate the bias term for this unit, and let \bar{z} denote the value of the hidden unit before the nonlinear activation is applied:

$$\bar{z} = \hat{\bar{\mathbf{X}}}^T \mathbf{W} + b.$$

Our objective now is to express \bar{z} in terms of $\bar{\mathbf{X}}$ and determine a modified weight vector \mathbf{W}' and bias b' that can be used to replace \mathbf{W} and b in order to allow the network to accept unnormalized inputs. This derivation requires some simple algebraic manipulation, which we show below:

$$\begin{aligned} \bar{z} &= \hat{\bar{\mathbf{X}}}^T \mathbf{W} + b \\ &= ((\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}) \cdot \bar{\boldsymbol{\sigma}}^{-1})^T \mathbf{W} + b \\ &= (\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}})^T (\mathbf{W} \cdot \bar{\boldsymbol{\sigma}}^{-1}) + b \\ &= \bar{\mathbf{X}}^T \underbrace{(\mathbf{W} \cdot \bar{\boldsymbol{\sigma}}^{-1})}_{\mathbf{W}'} + \underbrace{b - \bar{\boldsymbol{\mu}}^T (\mathbf{W} \cdot \bar{\boldsymbol{\sigma}}^{-1})}_{b'} \end{aligned} \tag{27}$$

We remark that this transformation need only apply to the weights and biases of units that connect directly to the input layer.

Appendix B. RDDDL Domain Descriptions

In this section, we list the RDDDL domain and instance files that we experimented with in this paper.

Reservoir

DOMAIN FILE

```
domain Reservoir_Problem{

requirements = {
reward-deterministic
};

types {
id: object;
};

pvariables {

// Constant
MAXCAP(id): { non-fluent, real, default = 100.0 };
HIGH_BOUND(id): { non-fluent, real, default = 80.0 };
LOW_BOUND(id): { non-fluent, real, default = 20.0 };
RAIN(id): { non-fluent, real, default = 5.0 };
DOWNSTREAM(id,id): {non-fluent, bool, default = false };
DOWNTOSEA(id): {non-fluent, bool, default = false };
}
```

```

BIGGESTMAXCAP: {non-fluent, real, default = 1000};

//Interm
vaporated(id): {interm-fluent, real};

//State
rlevel(id): {state-fluent, real, default = 50.0 };

//Action
flow(id): { action-fluent, real, default = 0.0 };
};

cpfs {
vaporated(?r) = (1.0/2.0)*sin[rlevel(?r)/BIGGESTMAXCAP]*rlevel(?r);
rlevel'(?r) = rlevel(?r) + RAIN(?r)- vaporated(?r) - flow(?r)
              + sum_{?r2: id}[DOWNSTREAM(?r2,?r)*flow(?r2)];
};

reward = sum_{?r: id} [if (rlevel'(?r)>=LOW_BOUND(?r) ^ (rlevel'(?r)<=HIGH_BOUND(?r)))
                      then 0
                      else if (rlevel'(?r)<=LOW_BOUND(?r))
                          then (-5)*(LOW_BOUND(?r)-rlevel'(?r))
                          else (-100)*(rlevel'(?r)-HIGH_BOUND(?r))]
      +sum_{?r2:id}[abs(((HIGH_BOUND(?r2)+LOW_BOUND(?r2))/2.0)-rlevel'(?r2))*(-0.1)];

state-action-constraints {

forall_{?r:id} flow(?r)<=rlevel(?r);
forall_{?r:id} rlevel(?r)<=MAXCAP(?r);
forall_{?r:id} flow(?r)>=0;
};
}

```

INSTANCE FILES

Reservoir 3

```

non-fluents Reservoir_non {
  domain = Reservoir_Problem;
  objects{
    id: {t1,t2,t3};
  };
  non-fluents {
    RAIN(t1) = 5.0;RAIN(t2) = 10.0;RAIN(t3) = 20.0;
    MAXCAP(t2) = 200.0;LOW_BOUND(t2) = 30.0;HIGH_BOUND(t2) = 180.0;
    MAXCAP(t3) = 400.0;LOW_BOUND(t3) = 40.0;HIGH_BOUND(t3) = 380.0;
    DOWNSTREAM(t1,t2);DOWNSTREAM(t2,t3);DOWNTOSEA(t3);
  };
}

instance isl{
  domain = Reservoir_Problem;
  non-fluents = Reservoir_non;
  init-state{
    rlevel(t1) = 75.0;
  };
  max-nondef-actions = 3;
  horizon = 10;
  discount = 1.0;
}

```

Reservoir 4

```

non-fluents Reservoir_non {

```

```

domain = Reservoir_Problem;
objects{
    id: {t1,t2,t3,t4};
};
non-fluents {
    RAIN(t1) = 5.0;RAIN(t2) = 10.0;RAIN(t3) = 20.0;RAIN(t4) = 30.0;
    MAXCAP(t2) = 200.0;LOW_BOUND(t2) = 30.0;HIGH_BOUND(t2) = 180.0;
    MAXCAP(t3) = 400.0;LOW_BOUND(t3) = 40.0;HIGH_BOUND(t3) = 380.0;
    MAXCAP(t4) = 500.0;LOW_BOUND(t4) = 60.0;HIGH_BOUND(t4) = 480.0;
    DOWNSTREAM(t1,t2);DOWNSTREAM(t2,t3);DOWNSTREAM(t3,t4);DOWNTOSEA(t4);
};
}

instance isl{
    domain = Reservoir_Problem;
    non-fluents = Reservoir_non;
    init-state{
        rlevel(t1) = 75.0;
    };
    max-nondef-actions = 4;
    horizon = 10;
    discount = 1.0;
}

```

Reservoir 10

```

non-fluents Reservoir_non {
    domain = Reservoir_Problem;
    objects{
        id: {t1,t2,t3,t4,t5,t6,t7,t8,t9,t10};
    };
    non-fluents {
        RAIN(t1) = 15.0;RAIN(t2) = 10.0;RAIN(t3) = 20.0;RAIN(t4) = 30.0;RAIN(t5) = 20.0;
        RAIN(t6) = 10.0;RAIN(t7) = 35.0;RAIN(t8) = 15.0;RAIN(t9) = 25.0;RAIN(t10) = 20.0;
        MAXCAP(t2) = 200.0;LOW_BOUND(t2) = 30.0;HIGH_BOUND(t2) = 180.0;
        MAXCAP(t3) = 400.0;LOW_BOUND(t3) = 40.0;HIGH_BOUND(t3) = 380.0;
        MAXCAP(t4) = 500.0;LOW_BOUND(t4) = 60.0;HIGH_BOUND(t4) = 480.0;
        MAXCAP(t5) = 750.0;LOW_BOUND(t5) = 20.0;HIGH_BOUND(t5) = 630.0;
        MAXCAP(t6) = 300.0;LOW_BOUND(t6) = 30.0;HIGH_BOUND(t6) = 250.0;
        MAXCAP(t7) = 300.0;LOW_BOUND(t7) = 10.0;HIGH_BOUND(t7) = 180.0;
        MAXCAP(t8) = 300.0;LOW_BOUND(t8) = 40.0;HIGH_BOUND(t8) = 240.0;
        MAXCAP(t9) = 400.0;LOW_BOUND(t9) = 40.0;HIGH_BOUND(t9) = 340.0;
        MAXCAP(t10) = 800.0;LOW_BOUND(t10) = 20.0;HIGH_BOUND(t10) = 650.0;
        DOWNSTREAM(t1,t2);DOWNSTREAM(t2,t3);DOWNSTREAM(t3,t4);DOWNSTREAM(t4,t5);
        DOWNSTREAM(t6,t7);DOWNSTREAM(t7,t8);DOWNSTREAM(t8,t5);
        DOWNSTREAM(t5,t6);DOWNSTREAM(t6,t10);
        DOWNSTREAM(t5,t9);DOWNSTREAM(t9,t10);
        DOWNTOSEA(t10);
    };
}

instance isl{
    domain = Reservoir_Problem;
    non-fluents = Reservoir_non;
    init-state{
        rlevel(t1) = 175.0;
    };
    max-nondef-actions = 10;
    horizon = 10;
    discount = 1.0;
}

```

HVAC

DOMAIN FILE

```

domain hvac_vav_fix{
  types {
    space : object;
  };

  pvariables {
    //Constants
    ADJ(space, space)      : { non-fluent, bool, default = false };
    ADJ_OUTSIDE(space)    : { non-fluent, bool, default = false };
    ADJ_HALL(space)       : { non-fluent, bool, default = false };
    R_OUTSIDE(space)      : { non-fluent, real, default = 4};
    R_HALL(space)         : { non-fluent, real, default = 2};
    R_WALL(space, space)  : { non-fluent, real, default = 1.5 };
    IS_ROOM(space)       : { non-fluent, bool, default = false };
    CAP(space)           : { non-fluent, real, default = 80 };
    CAP_AIR              : { non-fluent, real, default = 1.006 };
    COST_AIR             : { non-fluent, real, default = 1 };
    TIME_DELTA           : { non-fluent, real, default = 1 };
    TEMP_AIR             : { non-fluent, real, default = 40 };
    TEMP_UP(space)       : { non-fluent, real, default = 23.5 };
    TEMP_LOW(space)      : { non-fluent, real, default = 20.0 };
    TEMP_OUTSIDE(space)  : { non-fluent, real, default = 6.0 };
    TEMP_HALL(space)     : { non-fluent, real, default = 10.0 };
    PENALTY              : { non-fluent, real, default = 20000 };
    AIR_MAX(space)       : { non-fluent, real, default = 10.0 };
    TEMP(space)          : { state-fluent, real, default = 10.0 };
    AIR(space)           : { action-fluent, real, default = 0.0 };
  };

  cpfs {
    //State
    TEMP'(?s) = TEMP(?s) + TIME_DELTA/CAP(?s) *
    (AIR(?s) * CAP_AIR * (TEMP_AIR - TEMP(?s)) * IS_ROOM(?s)
    + sum_{?p : space} ((ADJ(?s, ?p) | ADJ(?p, ?s)) * (TEMP(?p) - TEMP(?s)) / R_WALL(?s, ?p))
    + ADJ_OUTSIDE(?s)*(TEMP_OUTSIDE(?s) - TEMP(?s))/ R_OUTSIDE(?s)
    + ADJ_HALL(?s)*(TEMP_HALL(?s)-TEMP(?s))/R_HALL(?s));
  };

  reward = - (sum_{?s : space} IS_ROOM(?s)*(AIR(?s) * COST_AIR
    + ((TEMP(?s) < TEMP_LOW(?s)) | (TEMP(?s) > TEMP_UP(?s))) * PENALTY)
    + 10.0*abs(((TEMP_UP(?s) + TEMP_LOW(?s))/2.0) - TEMP(?s)));

  action-preconditions{
    forall_{?s : space} [ AIR(?s) >= 0 ];
    forall_{?s : space} [ AIR(?s) <= AIR_MAX(?s)];
  };
}

```

INSTANCE FILES

HVAC 3 Rooms

```

non-fluents nf_hvac_vav_fix{
  domain = hvac_vav_fix;

  objects{
    space : { r1, r2, r3};
  };
}

```

```

non-fluents {
  //Define rooms
  IS_ROOM(r1) = true;IS_ROOM(r2) = true;IS_ROOM(r3) = true;

  //Define the adjacency
  ADJ(r1, r2) = true;ADJ(r1, r3) = true;ADJ(r2, r3) = true;

  ADJ_OUTSIDE(r1) = true;ADJ_OUTSIDE(r2) = true;
  ADJ_HALL(r1) = true;ADJ_HALL(r3) = true;
};
}

instance inst_hvac_vav_fix{
  domain = hvac_vav_fix;
  non-fluents = nf_hvac_vav_fix;

  horizon = 20;
  discount = 1.0;
}

```

HVAC 6 Rooms

```

non-fluents nf_hvac_vav_fix{
  domain = hvac_vav_fix;

  objects{
    space : { r1, r2, r3, r4, r5, r6 };
  };

  non-fluents {
    //Define rooms
    IS_ROOM(r1) = true;IS_ROOM(r2) = true;IS_ROOM(r3) = true;
    IS_ROOM(r4) = true;IS_ROOM(r5) = true;IS_ROOM(r6) = true;

    //Define the adjacency
    ADJ(r1, r2) = true;ADJ(r1, r4) = true;ADJ(r2, r3) = true;
    ADJ(r2, r5) = true;ADJ(r3, r6) = true;ADJ(r4, r5) = true;
    ADJ(r5, r6) = true;

    ADJ_OUTSIDE(r1) = true;ADJ_OUTSIDE(r3) = true;
    ADJ_OUTSIDE(r4) = true;ADJ_OUTSIDE(r6) = true;
    ADJ_HALL(r1) = true;ADJ_HALL(r2) = true;ADJ_HALL(r3) = true;
    ADJ_HALL(r4) = true;ADJ_HALL(r5) = true;ADJ_HALL(r6) = true;
  };
}

instance inst_hvac_vav_fix{
  domain = hvac_vav_fix;
  non-fluents = nf_hvac_vav_fix;

  horizon = 20;
  discount = 1.0;
}

```

HVAC 60 Rooms

```

non-fluents nf_hvac_vav_fix{
  domain = hvac_vav_fix;

  objects{
    space : { r101, r102, r103, r104, r105, r106, r107, r108, r109, r110, r111, r112,
              r201, r202, r203, r204, r205, r206, r207, r208, r209, r210, r211, r212,
              r301, r302, r303, r304, r305, r306, r307, r308, r309, r310, r311, r312,
              r401, r402, r403, r404, r405, r406, r407, r408, r409, r410, r411, r412,
              r501, r502, r503, r504, r505, r506, r507, r508, r509, r510, r511, r512
    };
  };
}

```



```

    }; //Three rooms, one hallway, and the outside world
};

non-fluents {
    //Define rooms
    //Level1
    IS_ROOM(r101) = true; IS_ROOM(r102) = true; IS_ROOM(r103) = true; IS_ROOM(r104) = true;
    IS_ROOM(r105) = true; IS_ROOM(r106) = true; IS_ROOM(r107) = true; IS_ROOM(r108) = true;
    IS_ROOM(r109) = true; IS_ROOM(r110) = true; IS_ROOM(r111) = true; IS_ROOM(r112) = true;
    //Level2
    IS_ROOM(r201) = true; IS_ROOM(r202) = true; IS_ROOM(r203) = true; IS_ROOM(r204) = true;
    IS_ROOM(r205) = true; IS_ROOM(r206) = true; IS_ROOM(r207) = true; IS_ROOM(r208) = true;
    IS_ROOM(r209) = true; IS_ROOM(r210) = true; IS_ROOM(r211) = true; IS_ROOM(r212) = true;
    //Level3
    IS_ROOM(r301) = true; IS_ROOM(r302) = true; IS_ROOM(r303) = true; IS_ROOM(r304) = true;
    IS_ROOM(r305) = true; IS_ROOM(r306) = true; IS_ROOM(r307) = true; IS_ROOM(r308) = true;
    IS_ROOM(r309) = true; IS_ROOM(r310) = true; IS_ROOM(r311) = true; IS_ROOM(r312) = true;
    //Level4
    IS_ROOM(r401) = true; IS_ROOM(r402) = true; IS_ROOM(r403) = true; IS_ROOM(r404) = true;
    IS_ROOM(r405) = true; IS_ROOM(r406) = true; IS_ROOM(r407) = true; IS_ROOM(r408) = true;
    IS_ROOM(r409) = true; IS_ROOM(r410) = true; IS_ROOM(r411) = true; IS_ROOM(r412) = true;
    //Level5
    IS_ROOM(r501) = true; IS_ROOM(r502) = true; IS_ROOM(r503) = true; IS_ROOM(r504) = true;
    IS_ROOM(r505) = true; IS_ROOM(r506) = true; IS_ROOM(r507) = true; IS_ROOM(r508) = true;
    IS_ROOM(r509) = true; IS_ROOM(r510) = true; IS_ROOM(r511) = true; IS_ROOM(r512) = true;

    //Define the adjacency
    //Level1
    ADJ(r101, r102) = true; ADJ(r102, r103) = true; ADJ(r103, r104) = true;
    ADJ(r104, r105) = true; ADJ(r106, r107) = true; ADJ(r107, r108) = true;
    ADJ(r107, r109) = true; ADJ(r108, r109) = true; ADJ(r110, r111) = true;
    ADJ(r111, r112) = true;
    //Level2
    ADJ(r201, r202) = true; ADJ(r202, r203) = true; ADJ(r203, r204) = true;
    ADJ(r204, r205) = true; ADJ(r206, r207) = true; ADJ(r207, r208) = true;
    ADJ(r207, r209) = true; ADJ(r208, r209) = true; ADJ(r210, r211) = true;
    ADJ(r211, r212) = true;
    //Level3
    ADJ(r301, r302) = true; ADJ(r302, r303) = true; ADJ(r303, r304) = true;
    ADJ(r304, r305) = true; ADJ(r306, r307) = true; ADJ(r307, r308) = true;
    ADJ(r307, r309) = true; ADJ(r308, r309) = true; ADJ(r310, r311) = true;
    ADJ(r311, r312) = true;
    //Level4
    ADJ(r401, r402) = true; ADJ(r402, r403) = true; ADJ(r403, r404) = true;
    ADJ(r404, r405) = true; ADJ(r406, r407) = true; ADJ(r407, r408) = true;
    ADJ(r407, r409) = true; ADJ(r408, r409) = true; ADJ(r410, r411) = true;
    ADJ(r411, r412) = true;
    //Level5
    ADJ(r501, r502) = true; ADJ(r502, r503) = true; ADJ(r503, r504) = true;
    ADJ(r504, r505) = true; ADJ(r506, r507) = true; ADJ(r507, r508) = true;
    ADJ(r507, r509) = true; ADJ(r508, r509) = true; ADJ(r510, r511) = true;
    ADJ(r511, r512) = true;
    //InterLevel 1-2
    ADJ(r101, r201) = true; ADJ(r102, r202) = true; ADJ(r103, r203) = true;
    ADJ(r104, r204) = true; ADJ(r105, r205) = true; ADJ(r106, r206) = true;
    ADJ(r107, r207) = true; ADJ(r108, r208) = true; ADJ(r109, r209) = true;
    ADJ(r110, r210) = true; ADJ(r111, r211) = true; ADJ(r112, r212) = true;
    //InterLevel 2-3
    ADJ(r201, r301) = true; ADJ(r202, r302) = true; ADJ(r203, r303) = true;
    ADJ(r204, r304) = true; ADJ(r205, r305) = true; ADJ(r206, r306) = true;
    ADJ(r207, r307) = true; ADJ(r208, r308) = true; ADJ(r209, r309) = true;
    ADJ(r210, r310) = true; ADJ(r211, r311) = true; ADJ(r212, r312) = true;
    //InterLevel 3-4

```

```

ADJ(r301, r401) = true;ADJ(r302, r402) = true;ADJ(r303, r403) = true;
ADJ(r304, r404) = true;ADJ(r305, r405) = true;ADJ(r306, r406) = true;
ADJ(r307, r407) = true;ADJ(r308, r408) = true;ADJ(r309, r409) = true;
ADJ(r310, r410) = true;ADJ(r311, r411) = true;ADJ(r312, r412) = true;
//InterLevel 4-5
ADJ(r401, r501) = true;ADJ(r402, r502) = true;ADJ(r403, r503) = true;
ADJ(r404, r504) = true;ADJ(r405, r505) = true;ADJ(r406, r506) = true;
ADJ(r407, r507) = true;ADJ(r408, r508) = true;ADJ(r409, r509) = true;
ADJ(r410, r510) = true;ADJ(r411, r511) = true;ADJ(r412, r512) = true;

//Outside
//Level1
ADJ_OUTSIDE(r101) = true;ADJ_OUTSIDE(r102) = true;ADJ_OUTSIDE(r103) = true;
ADJ_OUTSIDE(r104) = true;ADJ_OUTSIDE(r105) = true;ADJ_OUTSIDE(r106) = true;
ADJ_OUTSIDE(r108) = true;ADJ_OUTSIDE(r110) = true;ADJ_OUTSIDE(r111) = true;
ADJ_OUTSIDE(r112) = true;
//Level2
ADJ_OUTSIDE(r201) = true;ADJ_OUTSIDE(r202) = true;ADJ_OUTSIDE(r203) = true;
ADJ_OUTSIDE(r204) = true;ADJ_OUTSIDE(r205) = true;ADJ_OUTSIDE(r206) = true;
ADJ_OUTSIDE(r208) = true;ADJ_OUTSIDE(r210) = true;ADJ_OUTSIDE(r211) = true;
ADJ_OUTSIDE(r212) = true;
//Level3
ADJ_OUTSIDE(r301) = true;ADJ_OUTSIDE(r302) = true;ADJ_OUTSIDE(r303) = true;
ADJ_OUTSIDE(r304) = true;ADJ_OUTSIDE(r305) = true;ADJ_OUTSIDE(r306) = true;
ADJ_OUTSIDE(r308) = true;ADJ_OUTSIDE(r310) = true;ADJ_OUTSIDE(r311) = true;
ADJ_OUTSIDE(r312) = true;
//Level4
ADJ_OUTSIDE(r401) = true;ADJ_OUTSIDE(r402) = true;ADJ_OUTSIDE(r403) = true;
ADJ_OUTSIDE(r404) = true;ADJ_OUTSIDE(r405) = true;ADJ_OUTSIDE(r406) = true;
ADJ_OUTSIDE(r408) = true;ADJ_OUTSIDE(r410) = true;ADJ_OUTSIDE(r411) = true;
ADJ_OUTSIDE(r412) = true;
//Level5
ADJ_OUTSIDE(r501) = true;ADJ_OUTSIDE(r502) = true;ADJ_OUTSIDE(r503) = true;
ADJ_OUTSIDE(r504) = true;ADJ_OUTSIDE(r505) = true;ADJ_OUTSIDE(r506) = true;
ADJ_OUTSIDE(r508) = true;ADJ_OUTSIDE(r510) = true;ADJ_OUTSIDE(r511) = true;
ADJ_OUTSIDE(r512) = true;

//Hallway
//Level1
ADJ_HALL(r101) = true;ADJ_HALL(r102) = true;ADJ_HALL(r103) = true;
ADJ_HALL(r106) = true;ADJ_HALL(r107) = true;ADJ_HALL(r109) = true;
ADJ_HALL(r110) = true;
//Level2
ADJ_HALL(r201) = true;ADJ_HALL(r202) = true;ADJ_HALL(r203) = true;
ADJ_HALL(r206) = true;ADJ_HALL(r207) = true;ADJ_HALL(r209) = true;
ADJ_HALL(r210) = true;
//Level3
ADJ_HALL(r301) = true;ADJ_HALL(r302) = true;ADJ_HALL(r303) = true;
ADJ_HALL(r306) = true;ADJ_HALL(r307) = true;ADJ_HALL(r309) = true;
ADJ_HALL(r310) = true;
//Level4
ADJ_HALL(r401) = true;ADJ_HALL(r402) = true;ADJ_HALL(r403) = true;
ADJ_HALL(r406) = true;ADJ_HALL(r407) = true;ADJ_HALL(r409) = true;
ADJ_HALL(r410) = true;
//Level5
ADJ_HALL(r501) = true;ADJ_HALL(r502) = true;ADJ_HALL(r503) = true;
ADJ_HALL(r506) = true;ADJ_HALL(r507) = true;ADJ_HALL(r509) = true;
ADJ_HALL(r510) = true;
};
}

instance inst_hvac_vav_fix{
    domain = hvac_vav_fix;

```

```

non-fluents = nf_hvac_vav_fix;
//init-state{
//};
max-nonddef-actions = 60;
horizon = 12;
discount = 1.0;
}

```

Navigation

DOMAIN FILE

```

domain Navigation_Problem{

  requirements = {
    reward-deterministic
  };

  types {
    dim: object;
  };

  pvariables {

    // Constant
    MINMAZEBOUND(dim): { non-fluent, real, default = -4.0 }; //-5.0 for 10x10 instance
    MAXMAZEBOUND(dim): { non-fluent, real, default = 4.0 }; //5.0 for 10x10 instance
    MINACTIONBOUND(dim): { non-fluent, real, default = -1.0 }; //-0.5 for large scale instance
    MAXACTIONBOUND(dim): { non-fluent, real, default = 1.0 }; //0.5 for large scale instance
    GOAL(dim): { non-fluent, real, default = 3.0 };
    PENALTY: {non-fluent, real, default = 1000000.0 };
    CENTER(dim): {non-fluent, real, default = 0.0};

    // Interm
    distance: {interm-fluent,real,level=1 };
    scalefactor: {interm-fluent,real,level=2 };
    proposedLoc(dim):{interm-fluent, real, level=3};

    //State
    location(dim): {state-fluent, real, default = -4.0 }; //-5.0 for 10x10 instance

    //Action
    move(dim): { action-fluent, real, default = 0.0 };
  };

  cpfs {

    distance = sqrt[sum_{?1:dim}[pow[(location(?1)-CENTER(?1)),2]]];
    scalefactor = 2.0/(1.0+exp[-2*distance])-0.99;
    proposedLoc(?1) = location(?1) + move(?1)*scalefactor;
    location'(?1)= if(proposedLoc(?1)<=MAXMAZEBOUND(?1) ^ proposedLoc(?1)>=MINMAZEBOUND(?1))
                    then proposedLoc(?1)
                    else (if(proposedLoc(?1)>MAXMAZEBOUND(?1))
                          then MAXMAZEBOUND(?1) else MINMAZEBOUND(?1));

  };

  reward = - sum_{?1: dim}[abs[GOAL(?1) - location(?1)]];

  state-action-constraints {
    forall_{?1:dim} move(?1)<=MAXACTIONBOUND(?1);
    forall_{?1:dim} move(?1)>=MINACTIONBOUND(?1);
  }
}

```

```

        forall_{?l:dim} location(?l) <= MAXMAZEBOUND(?l);
        forall_{?l:dim} location(?l) >= MINMAZEBOUND(?l);
    };
}

```

INSTANCE FILES

Navigation 8 by 8 instance

```

non-fluents Navigation_non {
    domain = Navigation_Problem;
    objects{
        dim: {x,y};
    };
    non-fluents {
        MINMAZEBOUND(x) = -4.0;
    };
}

instance isl{
    domain = Navigation_Problem;
    non-fluents = Navigation_non;
    init-state{
        location(x) = -4.0; location(y) = -4.0;
    };
    max-nondef-actions = 2;
    horizon = 10;
    discount = 1.0;
}

```

Navigation 10 by 10 instance

```

non-fluents Navigation_non {
    domain = Navigation_Problem;
    objects{
        dim: {x,y};
    };
    non-fluents {
        MINMAZEBOUND(x) = -5.0;
    };
}

instance isl{
    domain = Navigation_Problem;
    non-fluents = Navigation_non;
    init-state{
        location(x) = -5.0; location(y) = -5.0;
    };
    max-nondef-actions = 2;
    horizon = 10; //20 for large scale instance
    discount = 1.0;
}

```