Factoring Exogenous State for Model Free Monte Carlo

Sean McGregor[0], **Rachel Houtman**[1], **Claire Montgomery**[1], **Ronald Metoyer**[2], **Thomas G. Dietterich**[0]
Oregon State University, School of Electrical Engineering and Computer Science[0]
Oregon State University, College of Forestry[1]
University of Notre Dame, College of Engineering[2]
aiorsocgood17@seanbmcgregor.com

Abstract

Policy analysts wish to visualize a range of policies for large simulator-defined MDPs. One visualization approach is to invoke the simulator to generate on-policy trajectories and then visualize those trajectories. When the simulator is expensive, this is not practical, and some method is required for generating trajectories for new policies without invoking the simulator. The method of Model-Free Monte Carlo (MFMC) can do this by stitching together state transitions for a new policy based on previously-sampled trajectories from other policies. This "off-policy Monte Carlo simulation" method works well when the state space has low dimension but fails as the dimension grows. This paper describes a method for factoring out some of the state and action variables so that MFMC can work in high-dimensional MDPs. The new method, MFMCi, is evaluated on a very challenging wildfire management MDP.

1 Introduction

One important way that AI can serve the social good is to support policy analysts in their efforts to develop good policies for managing complex systems. For example, we work on the problem of wildfire management in which land managers must decide when and where to fight fires on public lands. Our goal is to create an interactive visualization environment in which policy analysts can define various fire management polices and evaluate them through comparative visualizations. The transition dynamics of our fire management MDP are defined by a simulator that takes as input a detailed map of the landscape, an ignition location, a stream of weather conditions, and a fire fighting decision (i.e., suppress the fire vs. letting it burn), and produces as output the resulting landscape map and associated variables (fire duration, area burned, timber value lost, fire fighting cost, etc.). The simulator also models the year-to-year growth of the trees and accumulation of fuels. Unfortunately, this simulator is extremely expensive. It can take up to 7 hours to simulate a single 100-year trajectory of fire ignitions and resulting landscapes. How can we support interactive policy analysis when the simulator is so expensive?

Our approach is to develop a surrogate model that can substitute for the simulator. We start by designing a small set of "seed policies" and invoking the slow simulator to generate several 100-year trajectories for each policy. This gives us a database of state transitions of the form (s_t, a_t, r_t, s_{t+1}), where s_t is the state at time t, a_t is the selected action, r_t is the resulting reward, and s_{t+1} is the resulting state. Given a new policy π to visualize, we apply the method of Model-Free Monte Carlo (MFMC) developed by Fonteneau et al. (2013) to simulate trajectories for π by stitching together state transitions according to a given distance metric Δ. Given a current state s and desired action $a = \pi(s)$, MFMC searches the database to find a tuple $(\tilde{s}, \tilde{a}, r, s')$ that minimizes the distance $\Delta((s, a), (\tilde{s}, \tilde{a}))$. It then uses s' as the resulting state and r as the corresponding one-step reward. MFMC is guaranteed to give reasonable simulated trajectories under assumptions about the smoothness of the transition dynamics and reward function and provided that each matched tuple is removed from the database when it is used.

In high-dimensional spaces (i.e., where the states and actions are described by many features), MFMC breaks because of two related problems. First, distances become less informative in high-dimensional spaces. Second, the required number of seed-policy trajectories grows exponentially in the dimensionality of the space. The main technical contribution of this paper is to introduce a modified algorithm, MFMCi, that reduces the dimensionality of the distance matching process by factoring out certain exogenous state variables and removing the features describing the action. In many applications, this can very substantially reduce the dimensionality of the matching process to the point that MFMC is again practical.

This paper is organized as follows. First, we briefly review previous research in surrogate modeling. Then we introduce our method for factoring out exogenous variables. The method requires a modification to the way that trajectories are generated from the seed policies. With this modification, we prove that MFMCi gives sound results and that it has lower bias and variance than MFMC. Finally, we conduct an experimental evaluation of MFMCi on our fire management problem. We show that MFMCi gives good performance for three different classes of policies and that for a fixed database size, it gives much more accurate visualizations.

2 Related Work

Surrogate modeling is the construction of a fast simulator that can substitute for a slow simulator. When designing a surrogate model for our wildfire suppression problem, we can consider several possible approaches.

First, we could develop a special-purpose simulator for fire spread, timber harvest, weather, and vegetative growth that computes the state transitions more efficiently. For instance, Arca et al. (2013) use a custom built model running on GPUs to calculate fire risk maps and mitigation strategies. However, developing a special-purpose simulator requires additional work to design, implement, and validate the simulator. This cost may exceed the resulting time savings.

A second approach would be to learn a parametric surrogate model from data generated by the slow simulator. For instance, Abbeel et al. (2005) learn helicopter dynamics by updating the parameters of a function designed specifically for helicopter flight. Designing a suitable parametric model that can capture weather, vegetation, fire spread, and the effect of fire suppression would require a major modeling effort.

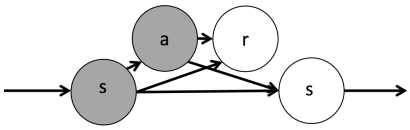
Instead of pursuing these two approaches, we adopted the method of Model-Free Monte Carlo (MFMC). In MFMC, the model is replaced by the database of transitions computed from the slow simulator. MFMC is "model free" in the sense that it does not need to model the transition because it uses the Monte Carlo transition from the database. Error accumulates from dissimilarity between the stitched states, but for many MDPs in computational sustainability the dimensionality required to model similarity may be small even for very large state spaces. We achieve the required dimensionality reduction by factoring the state space.

3 Factoring State for MFMC

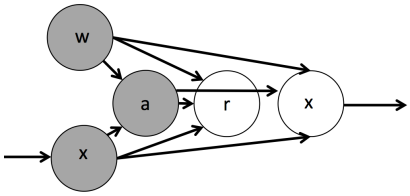
We work with the standard finite horizon undiscounted MDP (Bellman 1957; Puterman 1994), denoted by the tuple $\mathcal{M} = \langle S, A, P, R, P_0, h\rangle$. S is a finite set of states of the world; A is a finite set of possible actions that can be taken in each state; $P : S \times A \times S \mapsto [0, 1]$ is the conditional probability of entering state s' when action a is executed in state s; $R(s, a)$ is the finite reward received after performing action a in state s; P_0 is the distribution over starting states; and $\pi : S \mapsto A$ is the policy function that selects actions until reaching a terminal state or the horizon (h). We additionally define $M : S \times A \mapsto S$ as the model constructed from a database of trajectories D. MFMC selects transitions minimizing the distance between the current state and the candidate state from D.

In Factored MDPs (FMPDs), the goal is to leverage the probabilistic structure of state transitions to efficiently learn state transition dynamics (Guestrin et al. 2003; Hallak et al. 2015). Since MFMC doesn't model transition probabilities, the factors allow us to remove certain variables from a distance metric (Δ) that determines which states are stitched.

State variables can be divided into Markovian and Time-Independent random variables. A time-independent random variable x_{t+1} is independent of the values of (s_t, a_t), whereas in a Markovian random variable depends on the values of (s_t, a_t). Variables can also be classified as endogenous and exogenous. A variable is exogenous if it does not depend on a_t and it is endogenous otherwise. If a variable is time-independent and exogenous, then we can remove it from the MFMC stitching calculation.



(a) The standard MDP transition.



(b) MDP transition with *exogenous* (w) and *Markovian* variables (x).

Figure 1: MDP probabilistic graphical models.

Formally, the tuple (x, w) represents a complete state s, where w is time-independent and exogenous. See Figure 1 for a graphical depiction of the dependencies among variables in MDP state transitions. In our wildfire suppression domain, the state of the trees from one time step to another is Markovian, but our policy decisions also depend on exogenous weather events such as rain, wind, and lightning.

Factoring Exogenous State (W). We specify exogenous state independencies as

Definition 3.1. A Factored Exogenous MDP is an MDP such that the state (x, w) and next state (x', w') are related as

$$Pr(x', w'|x, w, a) = Pr(w')Pr(x'|x, w, a). \quad (1)$$

This factorization allows us to avoid modeling similarity of the complete state space s, and instead only consider the similarity of the Markov state. More formally, MFMC stitches (s, a) to the $(\tilde{s}, \tilde{a})$ in D that minimizes a distance metric Δ, so Δ has the form $\Delta((s, a), (\tilde{s}, \tilde{a})) \mapsto \mathbb{R}^+$. The key advantage of factoring out $w \in W$ is that the distance function only needs to match the Markov state x, so it has the form $\Delta_i(x, \tilde{x}) \mapsto \mathbb{R}^+$. The reduced dimensionality of Δ_i leads to reductions in the bias and the variance as we will show in Section 3.2. In the wildfire domain, this means we can compare landscapes without needing to compare specific wildfires.

In order for this to work correctly, two conditions must hold. First, the values of w in each state must be independent and identically distributed. Second, the database D must contain a separate tuple for each possible action that can be taken on a Markov state. The second condition fixes a subtle problem shown in Figure 1. Since w and x are both parents of a, it is possible to bias the exogenous variables by observing the action and matching on x. We formalize these

conditions in Section 3.1.

3.1 MFMC with independencies (MFMCi)

We introduce additional notation before formally stating MFMCi. In any MDP, we can rewrite the transition function $P(s'|s,a)$ as the combination of an exogenous random variable w and a function f such that the transition is generated by drawing $w \sim P(w)$ and then computing $s' := f(s,a,w)$. Similarly, a function o can be defined for the rewards such that $r := o(s,a,w)$. Furthermore, in many problems, the state variable s can be replaced by a pair (x, w_o) where w_o influences the policy, transition, and rewards, but $P(w_o) = P(w_o|x,a)$, i.e. it is exogenous. In these cases, we let w_u be the unobserved exogenous variables, w_o be the observed exogenous variables, and $w = (w_u, w_o)$. A state transition is then generated as follows: $w \sim P(w)$, $x' := f(x, \pi(x, w_o), w)$, and $r := o(x, \pi(x, w_o), w)$.

As noted in the prior section, factoring the exogenous state can introduce bias unless we sample each action for every $x \in D$. We formalize the additional action sampling by populating the database with *transition sets* B, where $B^r_{x',a}$ refers to the reward obtained by executing action a in state x', $B^{x'_{result}}_{x',a}$ refers to the result state, and $B^w_{x'}$ refers to the exogenous state. We use Algorithm 1 to populate the database with transition sets.

Algorithm 1: Populating D for MFMCi by sampling whole trajectories.

// Policy, horizon, trajectory count, simulator
Input Parameters: π, h, n, f
Returns: nh **transition sets** B
$D \leftarrow \emptyset$
while $|D| < nh$ **do**
  $x' = f_{x'}(\cdot,\cdot,\cdot)$ // Draw initial Markov state
  $w = f_w(\cdot,\cdot,\cdot)$ // Draw initial exogenous state
  $l = 0$
  while $l < h$ **do**
    $B \leftarrow \emptyset$
    $B^w_{x'} = w$
    for $a \in A$ **do**
      $B^r_{x',a} = f_r(x', a, w)$
      $B^{x'_{result}}_{x',a} = f_{x'_{result}}(x', a, w)$
    end
    append(D,B)
    $x' = B^{x'_{result}}_{x',\pi(x')}$
    $w = f_w(x', \cdot, w)$
    $l = l + 1$
  end
end
return(D)

Now we present MFMCi in Algorithm 2. When generating multiple trajectories with Algorithm 2, the database does not reuse transition sets for the same set of trajectories. We prove MFMCi's bias and variance bounds in the following section.

Algorithm 2: MFMCi

Input Parameters: $\pi, h, x_0, \Delta(.,.), D$ // Policy, horizon, starting state, distance metric, database
Returns: $(x_0, w, a, r)_1, ..., (x', w', a', r')_h$
$t \leftarrow \emptyset$
$x \leftarrow x_0$
while *length(t)*$< h$ **do**
  $B \leftarrow \text{argmin}_{B_{x'} \in D} \Delta(x, B_{x'})$
  $a \leftarrow \pi(x)$
  $r \leftarrow B^r_a$
  $w \leftarrow B^{w'}$
  append$(t, (x, a, r, w))$
  $D \leftarrow D \setminus B$
  $x \leftarrow B^{x'_{result}}_a$
end
return(t)

3.2 MFMCi Bias and Variance Bound

To maintain consistency with Fonteneau et al. (2013; 2014; 2010), we focus our theoretical analysis on the bias and variance of the estimator of the total reward.

Fonteneau et al.'s bounds depend on Lipschitz constants, the number of generated trajectories, and the sparsity of the database of transitions. In this work, we employ a similar set of assumptions, but by reducing the dimensionality of Δ, we shrink the bounds derived from these assumptions.

We define MFMC's three Lipschitz constants L_f, L_R, and L_π as follows:

$$||f(s,a,w_u) - f(s',a',w_u)||_S \leq L_f(||s-s'||_S + ||a-a'||_A) \quad (2)$$

$$|o(s,a,w_u) - o(s',a',w_u)| \leq L_R(||s-s'||_S + ||a-a'||_A) \quad (3)$$

$$||\pi(s) - \pi(s')||_A \leq L_\pi(||s-s'||_S + ||a-a'||_A). \quad (4)$$

Let E^π_r be the true h-horizon expected return of policy π when starting in state s_0, and let $MFMC^\pi_r$ denote the estimate of the expected return computed by the MFMC algorithm. To characterize the database's coverage of the state-action space, let $\alpha_k(D)$ be the maximum distance from any state-action pair (s,a) to its k-th nearest neighbor in database D. This is known as the k-sparsity of the database. Fonteneau et al. prove the bias bound given in Equation 5, and the variance bound given in Equation 6.

Theorem 1. *(Fonteneau et al. 2010) Under the Lipschitz continuity assumptions of Equations 2, 3, and 4 the bias and variance of $MFMC^\pi_r$ for any policy π are bounded as follows:*

$$|MFMC^\pi_r(s_0) - E^\pi_r(s_0)| \leq C\alpha_{nh}(D) \quad (5)$$

$$Var^\pi_{n,d}(s_0) \leq \left(\frac{\sigma_{V^\pi}(s_0)}{\sqrt{n}} + 2C\alpha_{nh}(D)\right)^2 \quad (6)$$

where $\sigma_{V^\pi}(s_0)$ is the true variance of the total reward and C is defined in terms of the Lipschitz constants as

$$C = L_R \sum_{i=0}^{h-1} \sum_{j=0}^{h-i-1} [L_f(1 + L_\pi)]^j. \quad (7)$$

Now we derive analogous bias and variance bounds for MFMCi. To this end, define $MFMC_i^\pi$ to be the estimate of the expected h-horizon return of policy π, and define two Lipschitz constants L_{F_i} and L_{R_i} such that the following conditions hold for the MDP:

$$||f(x, a, w) - f(x', a, w)||_X \leq L_{f_i}(||x - x'||_X) \quad (8)$$

$$|o(x, a, w) - o(x', a, w)| \leq L_{R_i}(||x - x'||_X). \quad (9)$$

Let $\alpha_{i,k}(D)$ be the maximum distance from any Markov state x to its k-th nearest neighbor in database D for the distance metric Δ_i. Then we have

Corollary 1. *Under the Lipschitz and k-sparsity conditions, the bias and variance of MFMCi are bounded as follows*

$$|MFMCi^\pi(s_0) - E_r^\pi(s_0)| \leq C_i \alpha_{i,nh}(D) \quad (10)$$

$$Var_{i,n,d}^\pi(s_0) \leq \left(\frac{\sigma_{V^\pi}(s_0)}{\sqrt{n}} + 2C_i\alpha_{i,nh}(D)\right)^2 \quad (11)$$

where C_i is defined as

$$C_i = L_{R_i} \sum_{b=0}^{h-1} \sum_{j=0}^{h-b-1} [L_{f_i}]^j. \quad (12)$$

Proof. (Sketch) The result follows by observing that because there is always a matching action for each transition set, a will equal a' and $||a - a'||_A$ will be zero, so we can eliminate L_π. Similarly, because we can factor out w_o, we only match on x, so we can replace L_f with L_{f_i} and replace the norms with respect to S by the norms with respect to X. Finally, as we argued above, by using transition sets we do not introduce any added bias by adopting w_o instead of matching against it. Formally, we can view this as converting w_o from an observable exogenous variable to being part of the unobserved exogenous source of stochasticity w_u. With these changes, the proof of Fonteneau, et al., holds. □

We believe similar proof techniques will prove the validity of MFMCi for visualizations, such as the quantile plots we demonstrate in our experiments. We leave this to future work.

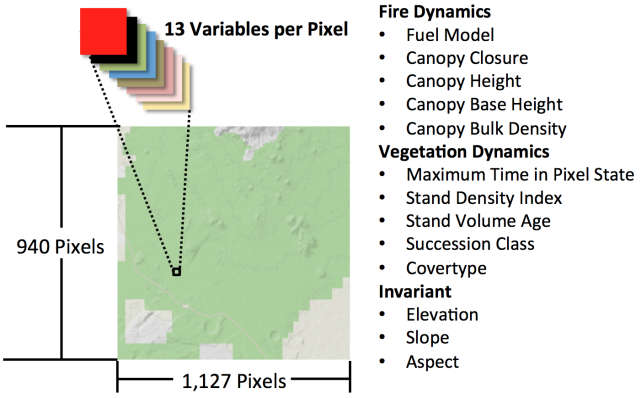


Figure 2: The landscape totals approximately one million pixels, each of which has 13 state variables that influence the spread of wildfire on the landscape. We use summary statistics of the dynamic state variables in MFMC's distance metric. (Map is copyright of OpenStreetMap contributors)

4 Experimental Evaluation

In our experiments we test whether we can generate accurate trajectory visualizations for a wildfire, timber, vegetation, and weather simulator. The aim of the wildfire management simulator is to inform wildfire suppression policies that determine whether the US government will suppress a wildfire. Each trajectory takes up to 7 hours to generate (Houtman et al. 2013).

Figure 2 shows a snapshot of the landscape as simulated by the Houtman simulator. The landscape is comprised of approximately one million pixels, each with 13 state variables. At the time of the ignition there are two actions: *Suppress* (fight the fire) and *Let Burn* (do nothing). Hence, $|A| = 2$.

The fire simulator spreads fire spatially from the ignition point according to the surrounding pixel variables (X) and the hourly weather. Weather variables include *hourly wind speed, hourly wind direction, hourly cloud cover, daily maximum/minimum temperature, daily maximum/minimum humidity, daily hour of maximum/minimum temperature, daily precipitation, and daily precipitation duration.* These are generated by resampling from 26 years of observed weather. We use MFMCi to synthesize trajectories by modeling the weather time series and ignition locations as w_o. The weather is exogenous because, to a first approximation, neither the selected actions nor the landscape influence the weather. Ignition location is exogenous to the landscape because tree cover does not affect the spatial probability of lightning strikes.

We constructed three policy classes that map fires to fire suppression decisions. We label these policies *intensity*, *fuel*, and *location*. The intensity policy suppresses fires based on the weather conditions at the time of the ignition, and the number of days remaining in the fire season. The fuel policy suppresses fires when the landscape accumulates sufficient high-fuel pixels. The location policy suppresses fires starting on one half of the landscape, and allows fires on the other half of the landscape to burn.

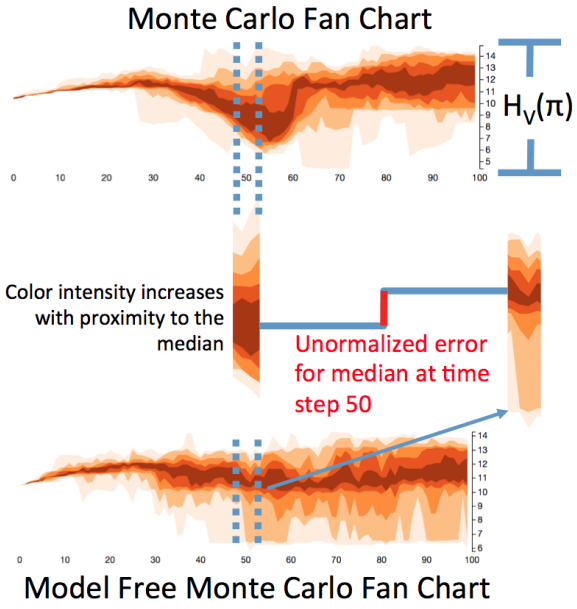


Figure 3: Here we enlarge and annotate fan charts where the x axis is the time step and the y axis is the value of the state variable at that time step. Each change in color shows a quantile boundary for a set of trajectories generated under policy π. The distance between the Monte Carlo median value and the MFMC median value is the error measure we use in the quantitative evaluation of MFMCi. We normalize the error across fan charts with $H_v(\pi)$, which is the Monte Carlo fan chart height for policy π and variable v. By measuring error in pixel units, we automatically incorporate the range normalization of the visualization procedure.

Fonteneau et al.'s (2013) theoretical analysis assumes the database is populated with state-action transitions covering the entire state-action space. The dimensionality of the wildfire state space makes this assumption impractical. We bias sampling towards states likely to be entered by future policy queries by seeding the database with one trajectory for 360 policies grid sampled over the intensity policy space.

All three policy classes are independent of each other, i.e. knowing the action selected by one policy class does not provide information on what action would be selected by one of the other policy classes. Thus demonstrating the ability for the state transitions from the intensity policy to generate transitions from the fuel and location policies would be strong evidence MFMC's generality. We evaluate MFMCi by generating 30 trajectories for a policy from each of these policy classes.

We use summary statistics of *Canopy Closure, Canopy Height, Canopy Base Height, Canopy Bulk Density, Stand Density Index, High Fuel Count, and Stand Volume Age* in the distance metric. We normalize these variables in the distance metric by their observed variance in the database. We also give arbitrarily large weight to the time step so states will only stitch if they were generated after a consistent number of years of growth.

We visualize trajectories using the visualization tool MDPVIS (McGregor et al. 2015). The key visualization in MDPVIS is the fan chart, which depicts various quantiles of the set of trajectories as a function of time (see Figure 3).

To evaluate the quality of the fan charts generated using surrogate trajectories, we define visual fidelity error in terms of the difference in vertical position between the true median and its position under the surrogate. Specifically, we define error(v,t) as the distance the median shifts for state variable v in time step t. We weight the error by the height of the fan chart for the rendered policy ($H_v(\pi)$). The re-weighted error is thus $\sum_{v \in S} \sum_{t=0}^{h} \frac{\text{error}(v,t)}{H_v(\pi)}$.

We visualize 20 variables related to the counts of burnt pixels, fire suppression expenses, timber loss, timber harvest, and landscape ecology.

4.1 Experimental Results

We evaluated the visual fidelity for MFMCi when we excluded exogenous variables from Δ_i, included exogenous variables, and excluded the bias correction sample from D. We constructed the biased database by removing all the transitions not consistent with the database policy (the innermost loop of Algorithm 1). We also compare against two baseline methods. First, we bootstrap resample the MC trajectories and report the average of the performance measure on the resampled trajectories. The bootstrap measure is an estimate of the achievable performance for 30 MC trajectories. We also report the performance of sampling 30 whole database trajectories and view its performance as a lower bound of the performance achievable with the database. See Figure 4 for the quantitative results.

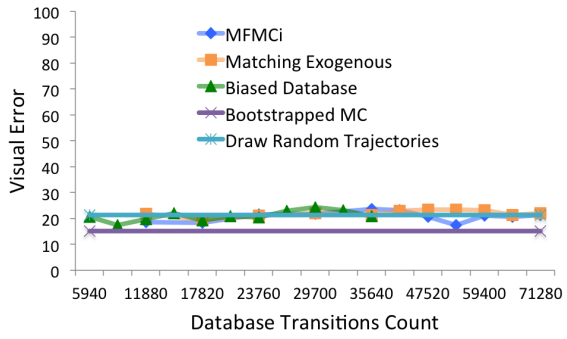
5 Discussion

Each of the target policies we selected illustrate a different strength or fault of MFMC for the wildfire problem.

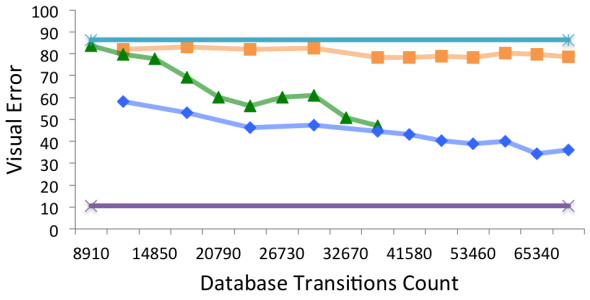
Intensity Policy. There are many policies in the database that agree with the target policy on the majority of fires. Thus, for the intensity policy it is sufficient to find a policy with a high-level of agreement, then sample the entire trajectory. Figure 4a shows that the median value across all the visualized variables is captured well by the median of a randomly sampled trajectories from the database. Unsurprisingly, we can approximate policies from the database policy class with very few samples.

Location Policy. We would expect our distance metric to be insufficient for representing spatial policies like the location policy. Every time a fire is allowed to burn on the left side of the landscape, the imbalance in fuels is smoothed over through stitching. However, our non-spatial metric shows that the visualization improves with additional samples when we don't include exogenous variables. Since fires are typically much smaller than half the landscape, the spatial effects are minimal. With additional samples, matching on exogenous variables will be capable of generalizing, but at these small sample counts it is essentially a random sample of trajectories in D.

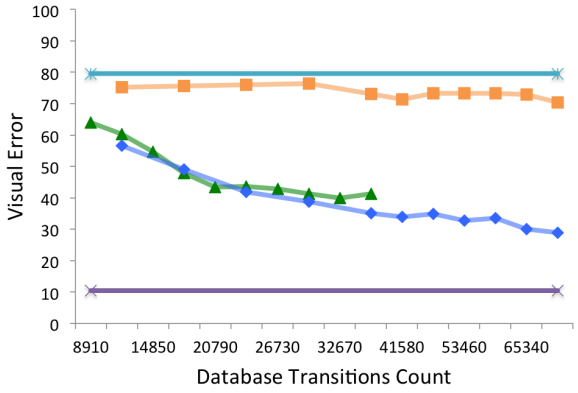
Fuel Policy. The fuel policy shows that including exogenous variables is very damaging for generalization. The bi-



(a) Visual fidelity errors for a weather *intensity* policy class. Fires are suppressed based on a combination of the weather and how much time is left in the fire season.



(b) Visual fidelity errors for a ignition *location* policy class. Fires are always suppressed if they start on the left half of the landscape, otherwise they are always allowed to burn.



(c) Visual fidelity errors for a *fuel* accumulation policy class. Fires are always suppressed if the landscape is at least 30 percent in high fuels, otherwise the fire is allowed to burn.

Figure 4: Policy classes for the wildfire domain under a variety of distance metrics and sampling procedures.

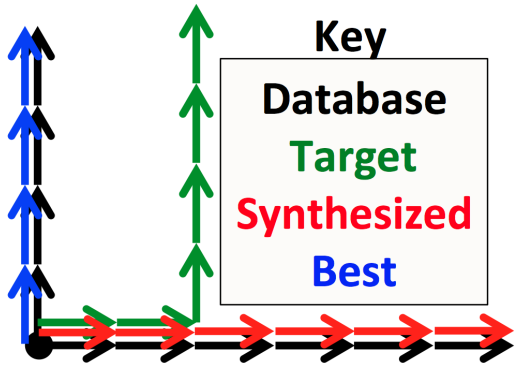


Figure 5: Example of MFMC's autoregressive tendency. Here the black arrows represent state transitions for two trajectories sampled into the database. The Green trajectory is a trajectory we want to generate for a new target policy. The trajectory that maintains minimal Euclidean distance from the target trajectory is the blue arrows, but distance metrics will force the synthesized (red) trajectory to follow the horizontal database trajectory if it has a bias correction sample. In some instances it is better to bias the exogenous variables than repeatedly stitch to the same trajectories.

ased database's close performance to MFMCi illustrates an important property of how we built the database. Within 7 time steps, fuel accumulation causes the policy action to switch from let-burn-all to suppress-all. MFMC will ideally follow a database trajectory that allows all wildfires to burn regardless of intensity and then switch to trajectories that suppress all wildfires. The "policy switching" behavior is implicit in the biased database, because the stitching process will jump to trajectories consistent with the target policy. If we select a policy that changes from let-burn-all to suppress-all after many time steps, it is possible that the unbiased database will trap the stitching operation in an unrepresentative region of the state space. We illustrate the problem in Figure 5 for a grid world example.

Despite these theoretical limitations, these experiments show that MFMCi is able to generalize across policy classes and that it requires only a small number of database trajectories to accurately reproduce the median of each state variable at each future time step.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 1331932.

References

Abbeel, P.; Ganapathi, V.; and Ng, A. Y. 2005. Learning Vehicular Dynamics, with Application to Modeling Helicopters. *Advances in Neural Information Processing Systems (NIPS)* 1–8.

Arca, B.; Ghisu, T.; Spataro, W.; and Trunfio, G. a. 2013. GPU-accelerated Optimization of Fuel Treatments for Mitigating Wildfire Hazard. *Procedia Computer Science* 18:966–975.

Bellman, R. 1957. *Dynamic Programming*. New Jersey: Princeton University Press.

Fonteneau, R., and Prashanth, L. 2014. Simultaneous Perturbation Algorithms for Batch Off-Policy Search. In *53rd IEEE Conference on Conference on Decision and Control*.

Fonteneau, R.; Murphy, S. A.; Wehenkel, L.; and Ernst, D. 2010. Model-Free Monte Carlo-like Policy Evaluation. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)* 217–224.

Fonteneau, R.; Murphy, S. a.; Wehenkel, L.; and Ernst, D. 2013. Batch Mode Reinforcement Learning based on the Synthesis of Artificial Trajectories. *Annals of Operations Research* 208(1):383–416.

Guestrin, C.; Koller, D.; Parr, R.; and Venkataraman, S. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research* 19(c):399–468.

Hallak, A.; Schnitzler, F.; Mann, T.; and Mannor, S. 2015. Off-policy Model-based Learning under Unknown Factored Dynamics. *Proceedings of the 32nd International Conference on Machine Learning* 37.

Houtman, R. M.; Montgomery, C. A.; Gagnon, A. R.; Calkin, D. E.; Dietterich, T. G.; McGregor, S.; and Crowley, M. 2013. Allowing a Wildfire to Burn: Estimating the Effect on Future Fire Suppression Costs. *International Journal of Wildland Fire* 22(7):871–882.

McGregor, S.; Buckingham, H.; Dietterich, T. G.; Houtman, R.; Montgomery, C.; and Metoyer, R. 2015. Facilitating Testing and Debugging of Markov Decision Processes with Interactive Visualization. In *IEEE Symposium on Visual Languages and Human-Centric Computing*.

Puterman, M. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1st edition.