Magical Policy Search: Data Efficient Reinforcement Learning with Guarantees of Global Optimality

Philip S. Thomas PHILIPT@CS.CMU.EDU
Carnegie Mellon University

Emma Brunskill EBRUN@CS.CMU.EDU
Carnegie Mellon University

Abstract

We present a batch policy search algorithm that has several desirable properties: it has few parameters that require expert tuning, it can leverage approximate models of the environment, it can seamlessly handle continuous states and actions, (informally speaking) it is guaranteed to converge to a globally optimal policy even in partially observable environments, and in our simulations it outperforms a state-of-the-art baseline. The primary limitation of our algorithm is its high computational complexity—each policy improvement step involves the optimization of a known (not necessarily convex) function.

1. Introduction

We present a batch policy search algorithm for episodic *Markov decision processes* (MDPs) or *partially observable Markov decision processes* (POMDPs). Given that one or more policies have been deployed in an environment that can be modeled as an MDP or POMDP, batch policy search algorithms take data collected from running the currently deployed policies and use it to search for an improved policy. Being able to use past experience to improve future performance in this way is a critical capability for intelligent agents (Shortreed et al., 2011; Pietquin et al., 2011).

One strong batch reinforcement learning algorithm is *fitted q-iteration* (Ernst et al., 2005, FQI). Although existing batch policy search algorithms like FQI have achieved significant success, most prior methods have several drawbacks: **1)** they are not guaranteed to converge to a globally optimal policy when using function approximation or in the presence of partial observability, **2)** if an approximate model of the MDP or POMDP is available, it is not always obvious how such information can or should be incorporated to improve data efficiency, and **3)** value-function based methods like FQI cannot leverage knowledge about a policy class that is known to perform well (e.g., PID controllers for control problems).

In this paper we present *magical policy search* (MPS), a new batch policy search algorithm that addresses the limitations just described. MPS is the result of the straightforward combination of two ideas. The first idea is that a batch policy search algorithm can simply return the policy that maximizes a data-based prediction of how good each policy is (Levine and Koltun, 2013, Section 3). We combine this idea with the recently proposed MAGIC estimator, which uses historical data to make predictions about how good each policy is (Thomas and Brunskill, 2016). The synthesis of these two ideas results in a batch policy search algorithm that can drastically improve data efficiency relative to existing model-free methods like FQI when an approximate model of the (PO)MDP is available.

2. Notation and Problem Statement

We assume that the reader is familiar with reinforcement learning (Sutton and Barto, 1998) and, although our results extend directly to POMDPs, for simplicity we adopt notational standard MDPNv1 for MDPs (Thomas, 2015a). For all policies, π, let $v(\pi) \coloneqq \mathbf{E}[\sum_{t=0}^{L} \gamma^t R_t | \pi]$ denote the *expected discounted return* if policy π were to be used, where $\gamma^t \in [0,1]$ is a discounting parameter, L is the finite horizon, and R_t is the uniformly bounded reward at time t. We do not assume that the transition and reward functions of the MDP are known, and so v is not known. Instead, we must make inferences about it from *historical data*, D. This historical data contains the observed states (observations), actions, and rewards, produced from the deployment of the current and past policies called *behavior policies*. We assume that D contains data from $n \in \mathbb{N}_{>0}$ episodes. *Batch policy search* algorithms are algorithms that take the historical data, D, an approximate model, and a set of policies, Π, called the *feasible set*, as input. They produce as output a policy $\pi \in \Pi$ with the goal of maximizing $v(\pi)$.

3. Background: Off-Policy Policy Evaluation (OPE)

In this section we review *off-policy policy evaluation* (OPE) algorithms that lie at the heart of our proposed method. An OPE algorithm, $\hat{v}$, takes historical data, D, as input as well as a policy, π_e, called the *evaluation policy*. It produces as output an estimate, $\hat{v}(\pi_e|D)$, of the performance, $v(\pi_e)$, of the evaluation policy. That is, $\hat{v}(\pi_e|D) \approx v(\pi_e)$. One of the earliest and most well-known OPE estimators is the *importance sampling* (IS) estimator, $\hat{v}_{\text{IS}}$, introduced by Precup et al. (2000). Given mild assumptions, the IS estimator is unbiased, i.e., $\mathbf{E}[\hat{v}_{\text{IS}}(\pi_e|D)] = v(\pi_e)$ and strongly consistent, i.e., $\mathbf{E}[\hat{v}_{\text{IS}}(\pi_e|D)] \xrightarrow{\text{a.s.}} v(\pi_e)$, (Precup et al., 2000). The primary drawback of the IS estimator is that it tends to have impractically high variance when the behavior and evaluation policies are not similar.

The *approximate model* (AM) estimator, $\hat{v}_{\text{AM}}$, is an OPE algorithm that uses historical data to build an approximation of the MDP or POMDP. It then uses the performance of the evaluation policy on this approximate model as an estimate of the evaluation policy's actual performance. Although the AM estimator tends to have much lower variance than the IS estimator, it is often not unbiased nor asymptotically correct. That is, if the true MDP cannot be represented by the approximate model (e.g., when using function approximation) or if there is partial observability, then as the amount of historical data goes to infinity, the most common maximum-likelihood approximate models will cause $\hat{v}_{\text{AM}}(\pi_e|D)$ to converge to values that may be quite different from $v(\pi_e)$.

The MAGIC estimator, $\hat{v}_{\text{MAGIC}}$, is a recently proposed estimator that combines the desirable properties of the IS and AM estimators. Experiments show that it tends to perform like whichever estimator is better, $\hat{v}_{\text{AM}}$ or a variant of $\hat{v}_{\text{IS}}$, and sometimes performs orders of magnitude better than both (in terms of mean squared error) (Thomas and Brunskill, 2016). The variant of $\hat{v}_{\text{IS}}$ used by MAGIC is called the *weighted doubly robust* (WDR) estimator, which we denote by $\hat{v}_{\text{WDR}}$. The MAGIC estimator is derived by defining a spectrum of estimators between the WDR and AM estimators: $\hat{v}^{(j)}$ for $j \in \{-2, -1, 0, 1, \ldots, L-1\}$,

where $\hat{v}^{(-2)} = \hat{v}_{\text{AM}}$ and $\hat{v}^{(L-1)} = \hat{v}_{\text{WDR}}$.[1] Rather than select a single estimator from this spectrum of estimators, the MAGIC estimator acts more like a *product of experts* system (Hinton, 2002)—it uses a weighted combination of all of the different estimators, $\hat{v}_{\text{MAGIC}}(D) \coloneqq \sum_{j=-2}^{L-1} w_j \, \hat{v}^{(j)}(D)$, where the weights, w_i, are selected to minimize a prediction of the resulting mean squared error.

4. Background: Batch Policy Search

Some recent works have used a batch policy search algorithm that we call *surrogate objective policy search* (SOPS).[2] SOPS creates a surrogate objective function, $\hat{v}$, that estimates the true objective function, v, and for which $\hat{v}(\pi)$ is known for all $\pi \in \Pi$. SOPS then selects the policy that maximizes $\hat{v}$. Specifically, SOPS selects the policy that maximizes an OPE estimator: $\arg\max_{\pi \in \Pi} \hat{v}(\pi|D) - \lambda c(\pi)$, where $c(\pi)$ is some notion of the complexity of the policy π and $\lambda \in \mathbb{R}_{\geq 0}$ is a parameter that scales regularization. Levine and Koltun (2013) suggest using a weighted importance sampling estimator (Precup et al., 2000) and a definition of c that leverages properties that are particular to the importance sampling estimator. Thomas et al. (2015) also propose using a weighted form of importance sampling, but regularize based on how close to deterministic a policy is. However, both Levine and Koltun (2013) and Thomas et al. (2015) propose more than just SOPS. Levine and Koltun (2013) propose guiding the policy search using the *iterative linear-quadratic regulator* control theoretic algorithm, and it is not clear how this could be extended to non-control applications like the digital marketing experiment that we propose. Thomas et al. (2015) propose using SOPS as a subroutine in an algorithm that guarantees improvement with high confidence, which requires significantly more data than merely finding a policy that tends to be better than the behavior policies.

To the best of our knowledge, using SOPS (without augmentations) as a batch policy search algorithm has not been proposed. One reason for this might be that the search over $\pi \in \Pi$ requires the OPE algorithm to be run on a wide distribution of policies, some of which can be very different from the behavior policies. In these cases importance sampling based methods tend not to perform well, and so the policy search algorithm must be restricted to proposing policies that are too similar to the behavior policies to be of practical use. In the next section we argue that this limitation can be mitigated by using MAGIC as the OPE algorithm in SOPS.

5. Magical Policy Search (MPS)

We propose using the MAGIC estimator as the OPE subroutine in SOPS, and therefore refer to our algorithm as *magical policy search* (MPS). MPS is defined by the equation:

$$\pi \in \arg\max_{\pi \in \Pi} \hat{v}_{\text{MAGIC}}(\pi|D). \tag{1}$$

1. Although Thomas and Brunskill (2016) focus on the case where $j \neq -2$, they point out in footnote 8 that when an approximate model is available *a priori*, using $j = -2$ can be important.
2. Levine and Koltun (2013, Section 3) call this method *importance sampled policy search*. We do not adopt this name because we consider using OPE methods like $\hat{v}_{\text{AM}}$ that do not use importance sampling.

In practice, we find approximate solutions to this arg max using CMA-ES, a general-purpose black-box optimization algorithm (Hansen, 2006).[3] Using MAGIC as a subroutine in SOPS has several benefits over using conventional importance sampling methods. Most importantly, its ability to leverage approximate models means that the estimates produced by MAGIC can be accurate even for policies that are quite different from the behavior policies.

That is, if using SOPS with an importance sampling variant, the OPE predictions will have extremely high variance for policies that are very different from the behavior policies. As a result, SOPS tends to find policies that importance sampling predicts will be near optimal, but which in fact are quite bad. To stop SOPS from selecting these policies, past work used regularization to limit the set of policies to those that are in some way "close" to the behavior policies.

Unlike purely importance sampling methods, for policies that are very different from the behavior policy, MAGIC automatically relies on the approximate model. So, when there is little data, the MAGIC estimator is similar to the AM estimator, and so MPS proposes a policy that makes sense given the prior knowledge provided in the approximate model. As more data arrives, the MAGIC estimator begins to look more like an importance sampling estimator for policies close to the behavior policy, but still resembles the AM estimator for distant policies. MPS will therefore continue to return the policy that makes sense based on the prior knowledge provided in the approximate model until the importance sampling estimator becomes more accurate than the model for a region of policy space that includes better policies. As a result of this behavior, in our experiments we found that MPS does not require the regularization term in SOPS. Although we therefore omit the regularizing term from MPS, it can easily be included by adding $-\lambda c(\pi)$ to the right side of (1), using a model selection method like cross-validation to tune λ, and one of the proposed definitions of $c(\pi)$ (Levine and Koltun, 2013; Thomas et al., 2015).

Recall that MAGIC uses an approximate model. For some applications an approximate model might be available *a priori*. For example, approximate models of some medical reinforcement learning applications have been constructed from expert knowledge and used to find policies that were then deployed for patients (Jagodnik and van den Bogert, 2007). Although the models resulted in decent policies, there remained room for improvement due to discrepancies between the expert models and the real world (Thomas et al., 2009). In other cases, a model might not be available *a priori*. In these cases, the approximate model can be constructed using some or all of the historical data.

6. Theoretical Analysis of MPS

In the supplemental document we present our primary theoretical result formally—here we provide an overview. We show that the performance of the policy produced by SOPS converges almost surely to the performance of an optimal policy given remarkably weak

3. The only parameter of CMA-ES that we tuned was the initial covariance matrix, which we set so that it covered Π. The low number of hyperparameters that must be tuned for the MAGIC estimator and CMA-ES means that MPS does not require much expert tuning. Specifically, the hyperparameters of MPS that might require tuning are the initial covariance matrix for CMA-ES, the number of iterations of CMA-ES to run (as many as possible) and the number of bootstrap samples to use in MAGIC (also as many as possible). Additionally, some terms that we view as part of the problem statement could be viewed as tunable hyperparameters: the approximate model used and the feasible set, Π.

assumptions. Essentially, as long as the OPE method used is strongly consistent—like all of the importance sampling estimators and the MAGIC estimator (Thomas, 2015b; Thomas and Brunskill, 2016))—then even if the objective function, v, is discontinuous (even point discontinuities are allowed), the value, $v(\pi)$, of the policy, π, returned by MPS will converge almost surely to the value, $v(\pi^\star)$ of an optimal policy, $\pi^\star \in \Pi$, as the amount of historical data goes to infinity (as $n \to \infty$). Intuitively, SOPS and MPS are able to achieve global optimality because every trajectory can be used to improve the estimate of the performance of every policy in Π simultaneously (via the OPE method). Recall that MPS is a variant of SOPS, and so this result applies to MPS.

Notice, however, that the convergence guarantee requires the assumption that[4] the implementation of MPS finds the $\pi \in \Pi$ that maximizes the magic estimator, which in itself is a challenging optimization problem. However, the convergence guarantee that we provide is still important for several reasons. First, our guarantee is interesting because it informs the user about how the method should behave. If a gradient method for a smooth optimization problem fails to converge to a solution (global or local) using a large fixed step size, we know that we can decrease the step size and, once the step size is small enough, it should converge. Similarly here, if our method does not converge to a sufficiently good policy, the proof shows that we should increase the power of the internal search algorithm. In our implementation, this would mean running more iterations of CMA-ES and including random restarts. Second, our convergence guarantee allows for partial observability (since the MAGIC estimator is strongly consistent even if the environment is a POMDP). To the best of our knowledge, the convergence guarantees for FQI and KBRL do not hold for POMDPs—they work for continuous states, but not when there is partial observability (which is common for real applications).

7. Empirical Evaluation of MPS

We applied MPS, SOPS (using the WDR estimator rather than MAGIC), and FQI to a simulated digital marketing problem. In this problem, the agent is given a description of the user visiting a web page and must select which advertisement to show to the user. The agent gets a reward of one if the user clicks on the advertisement and a reward of zero otherwise. The description of the user is a real vector in $[0, 1]^m$, each element of which encodes the agent's current belief about the user's interest in a particular topic. After each advertisement is shown, the belief about the user's interests is updated (based on which advertisement was shown and whether or not it was clicked) and the agent must select another advertisement to show.

Digital marketing is a good application for showcasing the benefits of MPS because it is reasonable to assume that an approximate model is available *a priori*. Specifically the mechanism that updates the belief about a users interests is known, so an approximate model can perfectly model these dynamics. However, there may be uncertainty about exactly how likely a user is to click based on the current belief state. Therefore in our simulation we use an approximate model with only a rough hand-crafted estimate of the true dynamics for computing whether or not a user will click.

4. Several other assumptions are required, like that real numbers are perfectly represented and that the real environment is perfectly modeled as a POMDP, which often may be false (Thomas et al., 2017).

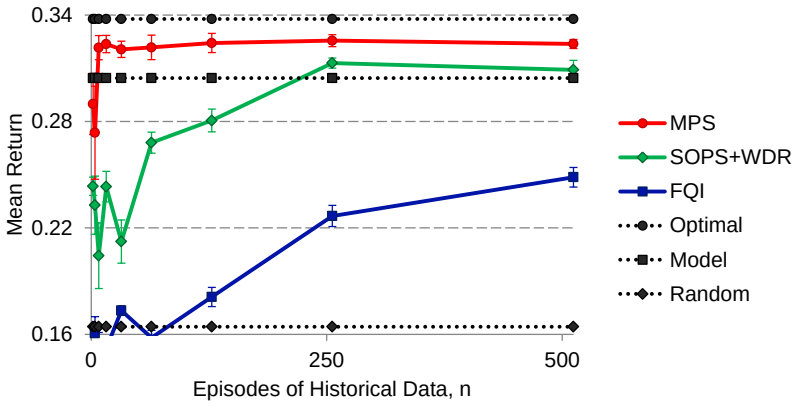

Figure 1: Performances on the digital marketing benchmark. "Optimal", "Model", and "Random" correspond to the mean returns of an optimal policy, and optimal policy for the approximate model, and the policy that selects actions uniformly randomly.

Figure 1 shows the mean returns of the policies produced by the methods that we considered. Notice that this is not a sequential plot—the point at $n = 256$ corresponds to selecting a policy using 256 trajectories of historical data generated by the random policy (e.g., the policy produced from $n = 128$ trajectories was not used to create the policy shown for $n = 256$ trajectories). We averaged the results over 20 trials and include standard error bars.

Given just $n = 8$ trajectories of historical data, MPS is able to improve significantly upon the model policy. The fact that MPS performs better than both the model policy and SOPS+WDR suggests that this is a setting where the MAGIC estimator outperforms both WDR and AM (the two estimators that it blends between). Notice that SOPS+WDR improves faster than FQI—we suspect that this is because it is "guided" by the approximate model (Thomas and Brunskill, 2016), while FQI ignores the approximate model. As anticipated, in trials using far more data, we found that eventually the performance of FQI reached levels similar to that of MPS and SOPS+WDR.

Given extremely limited data, $n = 2$ or $n = 4$, MPS does better than FQI and SOPS+WDR, but dips below the performance of the model policy. We suspect that this is because given so little data, MAGIC is unable to determine whether it should trust the approximate model or the importance sampling estimates. Still, MPS's use of the model likely gives it an advantage over model-free FQI and SOPS+WDR, which uses the approximate model only as a control variate.

8. Conclusion

We have presented a new policy search algorithm, *magical policy search* (MPS), which searches for the policy that maximizes an estimate of the true objective function, created from historical data. The estimate of the true objective function is generated by a recently

proposed estimator called the MAGIC estimator, which can make efficient use of approximate models. We proved that the expected returns of the policies produced by MPS, and a more general form of MPS that we call *surrogate objective policy search* (SOPS), converges with probability one to the expected return of a globally optimal policy as the amount of historical data goes to infinity. We also argued that the assumptions required to ensure convergence are remarkably weak relative to the assumptions required by other reinforcement learning algorithms. Finally, we evaluated MPS on a simulated digital marketing benchmark wherein it is reasonable to expect that an approximate model of some environmental dynamics would be available. We found that, by leveraging the approximate model, MPS significantly outperformed fitted q-iteration and another related SOPS algorithm.

References

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

N. Hansen. The CMA evolution strategy: A comparing review. In J.A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a New Evolutionary Computation. Advances On Estimation of Distribution Algorithms*, pages 75–102. Springer, 2006.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

K. Jagodnik and A. van den Bogert. A proportional derivative FES controller for planar arm movement. In *12th Annual Conference International FES Society*, Philadelphia, PA, 2007.

S. Levine and V. Koltun. Guided policy search. In *ICML*, pages 1–9, 2013.

Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Trans. Speech and Language Processing (TSLP)*, 7(3):7, 2011.

D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine Learning*, 84(1-2):109–136, 2011.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

P. S. Thomas. A notation for Markov decision processes. *ArXiv*, arXiv:1512.09075v1, 2015a.

P. S. Thomas. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015b.

P. S. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

P. S. Thomas, M. S. Branicky, A. J. van den Bogert, and K. M. Jagodnik. Application of the actor-critic architecture to functional electrical stimulation control of a human arm. In *IAAI*, pages 165–172, 2009.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015.

P. S. Thomas, G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Twenty-Ninth Conference on Innovative Applications of Artificial Intelligence*, 2017.

Appendix A. Theoretical Results

In §A.1 we present formal clarifications, definitions and assumptions. In §A.2 we discuss the assumptions and how they can be satisfied. Finally, in §A.3 we present our main result as Theorem 1.

A.1 A Clarification, Definitions, and Assumptions

Recall that the historical data, D, is a random variable. To further formalize this notion, let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $D_n : \Omega \to \mathcal{D}$ be a random variable. That is, $D_n(\omega)$ is a particular sample of the entire set of historical data with n trajectories, where $\omega \in \Omega$.

Definition 1 (Piecewise Lipschitz continuity)**.** *We say that a function $f : M \to \mathbb{R}$ on a metric space (M, d) is piecewise Lipschitz continuous with Lipschitz constant K and with respect to a countable partition, $\{M_1, M_2, \dots\}$, of M if f is Lipschitz continuous with Lipschitz constant K on all metric spaces in $\{(M_i, d)\}_{i=1}^{\infty}$.*

Definition 2 (δ-covering)**.** *If (M, d) is a metric space, a set $X \subseteq M$ is a δ-covering of (M, d) if and only if $\max_{y \in M} \min_{x \in X} d(x, y) \leq \delta$.*

Assumption 1 (Consistent OPE)**.** *For all $\pi \in \Pi$, $\hat{v}(\pi | D_n(\omega)) \xrightarrow{a.s.} v(\pi)$, and $\lambda = 0$.*

Assumption 2 (Piecewise Lipschitz objectives)**.** *The feasible set of policies, Π, is equipped with a metric, d_Π, such that for all $D_n(\omega)$ there exist countable partition of Π, $\Pi^v := \{\Pi_1^v, \Pi_2^v, \dots\}$ and $\Pi^{\hat{v}} := \{\Pi_1^{\hat{v}}, \Pi_2^{\hat{v}}, \dots\}$, where v and $\hat{v}(\cdot | D_n(\omega))$ are piecewise Lipschitz continuous with respect to Π^v and $\Pi^{\hat{v}}$ respectively with Lipschitz constants K and $\hat{K}$. Furthermore, for all $i \in \mathbb{N}_{>0}$ and all $\delta > 0$ there exist countable δ-covers of Π_i^v and $\Pi_i^{\hat{v}}$.*

A.2 Discussion of Definitions and Assumptions

Assumption 1 will be useful to show that the surrogate objective function converges almost surely to the true objective function, v. Thomas and Brunskill (2016, Theorem 3) showed that Assumption 1 holds for $\hat{v}_{\text{MAGIC}}$ given reasonable assumptions. Furthermore, the assumption that $\lambda = 0$ is satisfied by MPS since it does not use a regularizing term.

Assumption 2 ensures that v and $\hat{v}$ are piecewise Lipschitz continuous and that Π is a reasonable set (i.e., notice that the requirement that δ-covers exist is satisfied if Π is countable or $\Pi \subseteq \mathbb{R}^{n_{\boldsymbol{\theta}}}$ for any integer $n_{\boldsymbol{\theta}} \in \mathbb{N}_{>0}$). Consider the two most common settings considered by proofs for reinforcement learning methods: the setting where the number of policies is countable (e.g., deterministic policies for an MDP with countable state and action sets), and the setting where a stochastic policy is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^{n_{\boldsymbol{\theta}}}$. In the former setting, Π can be partitioned into $|\Pi|$ singletons (sets with one element) to satisfy Assumption 2. In the latter setting, a common assumption is that $\partial \pi(a|s, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ exists and is bounded for all s, a, and $\boldsymbol{\theta}$. In this setting, v is Lipschitz continuous over Π (this is evident from the policy gradient theorem (Sutton and Barto, 1998) and the fact that the q-function is bounded given our assumption that rewards are bounded and the horizon, L, is finite). If the importance weights are bounded (which Thomas and Brunskill (2016) already require to ensure Assumption 1) then $\hat{v}_{\text{MAGIC}}$ is also piecewise Lipschitz continuous in this setting.

Although these two settings are the most commonly considered, Assumption 2 holds for a significantly more general class of problems. For example, consider parameterized

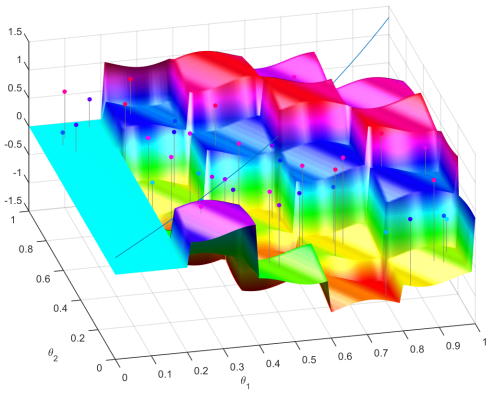

Figure 2: Example piecewise Lipschitz function, v. Here v is zero for small values of θ_1 and elsewhere it is a grid of functions that are each Lipschitz, but where there are discontinuities between each cell of the grid. Furthermore, there may be a countable number of point discontinuities (and the maximum may be at one of these points, which would have a probability of zero of being sampled during a random search of policy space). Here, the partition of Π would include separate regions for each point discontinuity, each cell of the grid (minus the point discontinuities), and a separate region for the line of discontinuities occurring at $\theta_1 = \theta_2$.

policies for an MDP where $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$. The surface in Figure 2 gives an example of a piecewise Lipschitz v (as a function of $\boldsymbol{\theta}$) that showcases how general Assumption 2 is. Discontinuous objective functions like this can occur, for example, when the stochastic policy is parameterized using a neural network with a step activation function. The fact that SOPS and MPS can almost surely converge to an optimal policy for v with flat regions and discontinuities like those of Figure 2 stems from its use of an OPE estimator. When a trajectory is generated using a single behavior policy (which may or may not even be in Π), the resulting information is used by the OPE algorithm (e.g., MAGIC) to update the surface of the entire surrogate objective function—to improve the predictions of the value of *every* policy in Π simultaneously. As a result, even if a unique optimal policy is at a point discontinuity which the gradient might never point toward and which has zero probability of being sampled by any continuous distribution over the feasible set, SOPS and MPS will find it with probability one.

A.3 Primary Theoretical Result

We are now ready to present our primary result: given the aforementioned assumptions, the performance of the policy produced by SOPS (e.g., MPS) is guaranteed to converge almost surely to the performance of an optimal policy as the amount of historical data goes to infinity.

Theorem 1 (Global optimality)**.** *If Assumptions* 1 *and* 2 *hold, then* $v(\text{SOPS}(D)) \xrightarrow{a.s.} \max_{\pi \in \Pi} v(\pi)$.

Proof By Assumption 1 and one of the common definitions of almost sure convergence,

$$\forall \pi \in \Pi, \forall \epsilon > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon\}\right) = 1.$$

Notice that, because Π may not be countable, this does not immediately imply that the probabilistic statement happens for all $\pi \in \Pi$ simultaneously, i.e., it does not immediately follow that:

$$\forall \epsilon > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : \forall \pi \in \Pi, |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon\}\right) = 1.$$

Let $C(\delta)$ denote the union of all of the points in the δ-covers of the countable partitions of Π assumed to exist by Assumption 2. Since the s are countable and the δ-covers for each region are assumed to be countable, we have that $C(\delta)$ is countable for all δ. So, we have that, for all δ, the probabilistic statement in question does hold for all $\pi \in C(\delta)$ simultaneously. That is:

$$\forall \delta > 0, \forall \epsilon > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : \forall \pi \in C(\delta), |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon\}\right) = 1. \tag{2}$$

Consider a $\pi \notin C(\delta)$. By the definition of a δ-cover and Assumption 2, we have that $\exists \pi' \in \Pi_i^v, d(\pi, \pi') \leq \delta$. Furthermore, since Assumption 2 requires v to be Lipschitz continuous on Π_i^v, we have that $|v(\pi) - v(\pi')| \leq K\delta$. By the same argument for $\hat{v}$ rather than v, we have that $|\hat{v}(\pi) - \hat{v}(\pi')| \leq \hat{K}\delta$. So, $|\hat{v}(\pi|D_n(\omega)) - v(\pi)| \leq |\hat{v}(\pi|D_n(\omega)) - v(\pi')| + K\delta \leq |\hat{v}(\pi'|D_n(\omega)) - v(\pi')| + \delta(K + \hat{K})$. This means that for all $\delta > 0$:

$$\left(\forall \pi \in C(\delta), |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon\right) \implies \left(\forall \pi \in \Pi, |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon + \delta(K + \hat{K})\right).$$

Substituting this into (2), we have that

$$\forall \delta > 0, \forall \epsilon > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : \forall \pi \in \Pi, |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon + \delta(K + \hat{K})\}\right) = 1.$$

Consider the specific choice of $\delta \coloneqq \epsilon/(K + \hat{K})$. We then have the following, where $\epsilon' = 2\epsilon$:

$$\forall \epsilon' > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : \forall \pi \in \Pi, |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon'\}\right) = 1. \tag{3}$$

Let $\pi^\star \in \arg\max_{\pi\in\Pi} v(\pi)$ and $\hat{\pi}^\star \in \arg\max_{\pi\in\Pi} \hat{v}(\pi|D_n(\omega))$ (we suppress the dependency of $\hat{\pi}^\star$ on ω). If $\forall \pi \in \Pi, |\hat{v}(\pi|D_n(\omega)) - v(\pi)| < \epsilon'$, then we have (by considering $\pi = \pi^\star$ and $\pi = \hat{\pi}^\star$):

$$|\hat{v}(\pi^\star|D_n(\omega)) - v(\pi^\star)| < \epsilon' \tag{4}$$
$$|\hat{v}(\hat{\pi}^\star|D_n(\omega)) - v(\hat{\pi}^\star)| < \epsilon', \tag{5}$$

and so:

$$\begin{aligned} v(\hat{\pi}^\star) &\overset{(a)}{\leq} v(\pi^\star) && (6)\\ &\overset{(b)}{<} \hat{v}(\pi^\star|D_n(\omega)) + \epsilon' \overset{(c)}{\leq} \hat{v}(\hat{\pi}^\star|D_n(\omega)) + \epsilon' \\ &\overset{(d)}{\leq} v(\hat{\pi}^\star) + 2\epsilon', && (7) \end{aligned}$$

where **(a)** comes from the definition of $\pi^\star$ as the maximizer of v, **(b)** comes from (4), **(c)** comes from the definition of $\hat{\pi}^\star$ as the maximizer of $\hat{v}$, and **(d)** comes from (5). So, considering (6) and (7), it follows that $|v(\hat{\pi}^\star) - v(\pi^\star)| \leq 2\epsilon'$. So, (3) implies that:

$$\forall \epsilon' > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : |v(\hat{\pi}^\star) - v(\pi^\star)| < 2\epsilon'\}\right) = 1.$$

Equivalently, using $\epsilon'' = 2\epsilon'$ and $\text{SOPS}(D) \coloneqq \hat{\pi}^\star$, we have that

$$\forall \epsilon'' > 0, \Pr\left(\liminf_{n\to\infty} \{\omega \in \Omega : |v(\text{SOPS}(D) - v(\pi^\star)| < \epsilon''\}\right) = 1,$$

which, by the definition of almost sure convergence, means that $v(\text{SOPS}(D)) \xrightarrow{\text{a.s.}} \max_{\pi\in\Pi} v(\pi)$.