Bad-Policy Density: A Measure of Reinforcement Learning Hardness

David Abel [1]   Cameron Allen [2]   Dilip Arumugam [3]
D. Ellis Hershkowitz [4]   Michael L. Littman [2]   Lawson L.S. Wong [5]

Abstract

Reinforcement learning is hard in general. Yet, in many specific environments, learning is easy. What makes learning easy in one environment, but difficult in another? We address this question by proposing a simple measure of reinforcement-learning hardness called the *bad-policy density*. This quantity measures the fraction of the deterministic stationary policy space that is below a desired threshold in value. We prove that this simple quantity has many properties one would expect of a measure of learning hardness. Further, we prove it is NP-hard to compute the measure in general, but there are paths to polynomial-time approximation. We conclude by summarizing potential directions and uses for this measure.

1. Introduction

Markov Decision Processes (MDPs) have long stood as a central model for characterizing the environments that reinforcement-learning agents inhabit. Indeed, the generality of the MDP (and its kin) allows for the description of small and simple environments such as grid worlds, but also large sophisticated ones such as the problem facing a robot organizing books on a shelf or even writing one of the books. A typical objective of research in reinforcement learning (RL) is to develop algorithms that can learn efficiently across the entire space of MDPs. For this reason, it is typically desirable to determine the worst case performance of an RL algorithm across all MDPs of a chosen size, and perhaps, horizon (Strehl et al., 2009; Azar et al., 2013).

However, understanding an algorithm's learning efficiency with respect to the size of the MDP misses out on the potentially crucial presence of structure in a given learning problem that can alter the nature and difficulty of learning. Moreover, it is likely that not all MDPs are of interest—oftentimes, relevant subsets of finite MDP space are isolated as being of particular use, such as those with objects (Diuk et al., 2008), features (Guestrin et al., 2001), or deterministic transition dynamics (Wen & Van Roy, 2013), to name a few. In this sense, it is likely the case that forcing algorithms to perform well on *all* MDPs misses out on important insights that ensure algorithms are well behaved on the MDPs that matter by forcing them to be competent on chaotic environments in which the rapid acquisition of competence *should* be impossible. Indeed, this insight is well established in the literature, with many notable examples establishing extreme efficiency in the presence of structure (Mersereau et al., 2009; Lattimore & Munos, 2014; Van Roy & Dong, 2019; Lattimore et al., 2020; Tirinzoni et al., 2020).

In this paper, we introduce Bad-Policy Density (BPD) as a simple new measure of RL hardness in finite MDPs. For a given MDP, the BPD measures the fraction of policies in the deterministic policy space that are below a desired threshold (τ) in start-state value. We argue that this measure picks up on interesting structure of MDPs, and can help understand the settings in which RL algorithms might be able to achieve extreme degrees of sample efficiency. We also advocate for this measure as a potential mechanism for determining the difficulty of MDPs in practice—those with higher BPD might be considered more difficult. Further, we suggest that this measure can be useful as a diagnostic tool to assess the impact that priors and structures have on learning hardness, and more generally to identify what characteristics of a task give rise to harder learning.

Previous RL Hardness Measures. Our proposal builds on the insights established by prior hardness measures for RL in finite MDPs (such as the mixing time (Kearns & Singh, 2002)), which we now briefly summarize. First, Maillard et al. (2014) address the question "How hard is my MDP?", with the environmental norm, measuring the maximum next-state variance of value throughout the MDP. This measure enables strong theoretical guarantees (Zanette & Brunskill, 2019) and picks up on an appealing notion of the kinds of MDPs that make RL more difficult—the more costly a mistake might be, the harder time a learning

---

[1]DeepMind [2]Department of Computer Science, Brown University [3]Department of Computer Science, Stanford University [4]Department of Computer Science, Carnegie Mellon University [5]Khoury College of Computer Sciences, Northeastern University. Correspondence to: David Abel <dmabel@deepmind.com>.

algorithm may have in learning effective behavior in the MDP. However, this measure is precisely zero for all deterministic MDPs, assigning all of them the lowest difficulty achievable under the measure. In this sense, there is room to sharpen our understanding of what constitutes a difficult MDP for RL. Farahmand (2011) and Bellemare et al. (2016) measure hardness through the gap between the $Q^\star$-values of the best and second-best action across state–action pairs; small action gaps induce more challenging problems as an agent requires more samples to reduce estimation error and identify the optimal action. The eluder dimension (Russo & Van Roy, 2013; Osband & Van Roy, 2014; Wang et al., 2020) of a value-function class is a worst-case measure of the maximal number of state–action pairs that must be observed before being able to extrapolate to unseen inputs. Intuitively, if each state–action pair of an MDP yields no information about any others, an agent has no capacity for generalization and is forced into prolonged exploration (Du et al., 2019; Van Roy & Dong, 2019).

Jiang et al. (2017) propose the Bellman Rank as a suitable measure for the difficulty of Contextual Decision Processes, a generalization of MDPs. Intuitively, the Bellman rank considers the matrix of Bellman residuals induced by a value-function class and creates a link between the behavior policy used to arrive at a particular state–action pair and the value function whose greedy policy defines behavior from that state–action pair. Sun et al. (2019) introduce the witness rank as an analogue to Bellman rank for model-based RL whereas Jin et al. (2021) introduce a generalization, the Bellman eluder dimension, as the eluder dimension on the function class of Bellman residuals. One important note on all three measures is their dependence on not only the MDP but also a particular function class; in this way, these measures address the difficulty inherent to learning in the MDP in the specific function class containing the solution. In contrast, BPD focuses on the former source of hardness.

Jaksch et al. (2010) propose the *diameter*—closely related to the span (Bartlett & Tewari, 2009; Fruit et al., 2018)— as a measure of MDP hardness, denoting the max number of steps between any two states in the environment. Naturally, a small diameter is suggestive of easier exploration, as the agent may acquire most information about the problem in few steps. Orthogonally, Arumugam et al. (2021) offer information sparsity as an information-theoretic measure of the difficulty of credit assignment within a MDP. With the exception of the environmental norm, all of the aforementioned hardness measures are concerned with characterizing the difficulty of generalization, exploration, or credit assignment. In contrast, BPD is agnostic to any one particular obstacle to efficient RL and instead simply asks what fraction of the solution space must be eliminated by any agent to solve an MDP, without regard for how efficiently individual agents may prune away sub-optimal solutions.

2. Bad-Policy Density

We now introduce and motivate the BPD measure. For a MDP $M = \langle \mathcal{S}, \mathcal{A}, R, T, \gamma \rangle$, we call M finite when $|\mathcal{S}|, |\mathcal{A}| < \infty$ and assume that M has an initial state $s_0 \in \mathcal{S}$.

Definition 1. *The **Bad-Policy Density** (BPD) of a finite MDP M, for a chosen $\tau \in \mathbb{R}$, is given by,*

$$\text{BPD}_\tau(M) := \frac{1}{|\Pi_M|} \sum_{\pi \in \Pi_M} \mathbb{1}\{V^\pi(s_0) \leq \tau\}, \quad (1)$$

for Π_M the set of all deterministic mappings from $\mathcal{S}$ to $\mathcal{A}$.

This measure answers a simple question about an MDP: What fraction of deterministic stationary policies are below τ in start-state value? The BPD has several straightforward properties, which we summarize in the following proposition.

Proposition 1. *The BPD satisfies the following properties:*

(i) *For any real τ and MDP M, $0 \leq \text{BPD}_\tau(M) \leq 1$.*

(ii) *For any rational number $q \in [0,1]$, there exists a choice of τ and deterministic M_d in which $\text{BPD}_\tau(M_d) = q$.*

(iii) *For any MDP M, if $\tau_1 < \tau_2$, then $\text{BPD}_{\tau_1}(M) \leq \text{BPD}_{\tau_2}(M)$.*

To summarize, this measure is applicable to both deterministic and stochastic MDPs; provides a universal, bounded scale of difficulty (the interval $[0,1]$) thereby allowing normalization-free comparison across MDPs; and ensures monotonic increase as τ increases. We explore more significant aspects of the measure shortly.

Weaknesses. There are clear shortcomings to the BPD. First, it is dependent on a potentially arbitrary choice of τ. Across MDPs, it is unclear what the right choice might be. In this sense, a potentially more informative view of the hardness of an MDP is given by the cumulative graph of $\text{BPD}_\tau(M)$, for $\tau \in [\text{VMIN}, \text{VMAX}]$; such a graph will illustrate what region of value space most policies lie in. Second, this measure is not immediately suitable to infinite MDPs. A natural consideration is to replace the enumeration of Equation 1 with a probability distribution—perhaps, in the assumption-free case, with the uniform distribution. Such a prospect is enticing, but is beyond the scope of this work. Third, the measure fails to capture the impact reward has on the learning dynamics of different RL algorithms. For instance, a shaped reward function that induces the same $\text{BPD}_\tau(M)$ may very well lead to a dramatically easier learning problem for many algorithms. Fourth, the measure makes a commitment as to the significance of the value in s_0 (or, more generally, in expectation under a start

state distribution). In long horizon tasks, this is not clearly the right perspective to take. Finally, the measure excludes stochastic policies from consideration in favor of focusing on deterministic policies.

In spite of these shortcomings, we take the BPD to serve as a useful and simple measure of RL difficulty in finite MDPs. A natural extension to infinite MDPs measures the probability of sampling a policy below a particular threshold, for some choice of probability distribution over the policy space. We next further establish the usefulness of the measure.

3. Analysis: Properties of the BPD

We next present three properties of the BPD: 1) there exists a simple RL algorithm whose episodic sample complexity depends only on the BPD and γ, and not on $|\mathcal{S}|$ or $|\mathcal{A}|$; 2) computing the BPD is NP-hard; 3) there is hope that a polynomial-time approximation is achievable.

3.1. BPD-Dependent Sample Complexity

Consider the following algorithmic structure, letting E denote the number of episodes and H denote the horizon:

$$\pi_t = \texttt{choose}(\Pi_t) \tag{2}$$

$$\hat{V}^\pi(s_0) = \texttt{eval}(M, \pi_t, E, H) \tag{3}$$

$$\Pi_{t+1} = \texttt{prune}(\Pi_t, \pi, \tau). \tag{4}$$

Suppose `choose` samples a policy uniformly at random, `eval` evaluates the sampled policy in the MDP, and `prune` removes π if $\hat{V}^\pi(s_0) \leq \tau$ and returns the pruned version of Π. Let us refer to this simple strategy as the POLICYSAMPLING algorithm. We note that the sample complexity of this approach will depend on the BPD as follows, where τ inversely defines the magnitude of the mistake bound.

Proposition 2. *Let $\delta \in (0, 1)$ be the desired confidence parameter, M the given MDP, $\eta \geq \text{VMAX} - \tau$ the mistake threshold, and $\rho = 1 - \text{BPD}_\tau(M)$. Then, the episodic POLICYSAMPLING algorithm has, with probability $1 - \delta$, sample complexity upper bounded by*

$$\frac{\log_\rho(\delta)\,\text{RMAX}^2 \ln\left(\frac{2}{\delta}\right)}{2\eta^2(1-\gamma)^3}. \tag{5}$$

Clearly, this algorithm is entirely impractical in many contexts. However, it illustrates a sense in which the sample complexity of learning *may* depend on the BPD, rather than quantities such as $|\mathcal{S}|$ and $|\mathcal{A}|$. Identifying structure-dependent guarantees of this form for existing algorithms is a natural direction for future work. We also note that the POLICYSAMPLING algorithm bears a resemblance to sparse sampling (Kearns et al., 2002), the PAC bandit approach by Goschin et al. (2013), and to the OLIVE algorithm

by Jiang et al. (2017), which iteratively prunes candidate value-function approximators based on the inconsistency of Bellman residuals.

3.2. Computing BPD is NP-hard

Naturally, the usefulness of a hardness measure is likely to depend on the practicality in applying it. We first note that computing the number of *optimal* policies is poly-time solvable by assessing the number of actions in each state that yield $Q^*(s, a)$ as follows,

$$\prod_{s \in \mathcal{S}} \left| \{a \in \mathcal{A} : Q^*(s, a) = V^*(s)\} \right|. \tag{6}$$

Unfortunately, when we move to computing the BPD, it is #P-hard in general. Concretely, we define the problem of computing $\text{BPD}_\tau(M)$ as follows.

Definition 2. *The* BPD-PROBLEM *is:* `given` *a finite MDP M and a $\tau \in \mathbb{R}$,* `return` $\text{BPD}_\tau(M)$.

To analyze the difficulty of this problem, we inspect its decision counterpart,

Definition 3. *The* BPD-DECISION-PROBLEM *is defined as follows:* `given` *a finite MDP M, $\tau \in \mathbb{R}$, and a proposed level of hardness κ,* `return` *true iff $\text{BPD}_\tau(M) = \kappa$.*

Theorem 1. BPD-DECISION-PROBLEM *is NP-hard.*

All proofs are presented in the appendix. As an immediate corollary of the theorem, we note that the counting variant, BPD-PROBLEM, is #P-hard.

3.3. Approximating the BPD

In light of the computational intractability of computing the BPD, we instead seek to approximate it. In this section, we give methods for efficiently approximating the good-policy density (GPD), where $\text{GPD}_\tau(M) = 1 - \text{BPD}_\tau(M)$; all of the proofs and algorithms given in this section also apply directly to the bad-policy density but it will be more convenient to describe our results in terms of good-policy density. There are two kinds of approximations for which one might aim: 1) multiplicative approximations; 2) additive approximations. In the former, we can obtain estimates that are within an arbitrary fraction of the true quantity (and thus, are more desirable), while in the latter, we can obtain estimates that are within a chosen ϵ radius of the true value (which is problematic when the true quantity is near zero or one, as the BPD is likely to be). The additive form is obtained through application of Chernoff inequalities.

Proposition 3. *There is a poly-time algorithm that, given a finite MDP M, $\tau \in \mathbb{R}$ and constant $\varepsilon > 0$ computes a value $\widehat{\text{GPD}}_\tau(M)$ satisfying $|\widehat{\text{GPD}}_\tau(M) - \text{GPD}_\tau(M)| \leq \varepsilon$ with high probability.[1]*

---

[1] At least $1 - \frac{1}{\text{poly}(n)}$ where n is the size of the input MDP.

(a) Difficult N-Chain   (b) Easy N-Chain   (c) BPD vs. $|\mathcal{S}|$
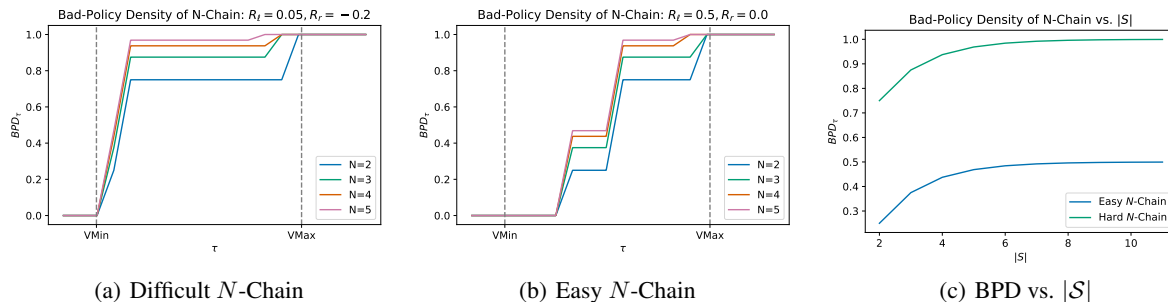
*Figure 1.* The BPD in N-Chain problems.

However, achieving the more useful form of multiplicative approximations is more involved. Our next result gives a nearly-complete story as to a path for multiplicative approximation. To actually achieve the result, we first require sampling access to a uniform distribution over "policies" which closely resemble policies with at least τ value on s_0. In particular, given some $i \geq 0$ we require access to uniform samples from what we call τ-quality i-controlled policies: These are "policies" where the agent achieves at least τ value on s_0 but only controls the first i states (for an arbitrary ordering of the states) and is forced to behave optimally on the remaining states. We defer a formal definition of these "policies" to the appendix.

Theorem 2. *There is a poly-time algorithm which, given constant $\varepsilon > 0$ and sample access to a uniform distribution over τ-quality i-controlled policies for each $i \geq 0$, returns a value $\widehat{\mathrm{GPD}}_\tau(M)$ satisfying $(1 - \varepsilon) \cdot \mathrm{GPD}_\tau(M) \leq \widehat{\mathrm{GPD}}_\tau(M) \leq (1 + \varepsilon) \cdot \mathrm{GPD}_\tau(M)$ with high probability.*

We suspect such a distribution can be sampled from in polynomial time, as this procedure is closely related to other polynomial-time constructions. See Jerrum & Sinclair (1989) for a similar result on approximating the number of matchings in a graph or Jerrum (2003) for a comprehensive overview of such approximate counting results.

4. Discussion

We conclude with a simple case study of the BPD in small MDPs, and by suggesting directions for future work.
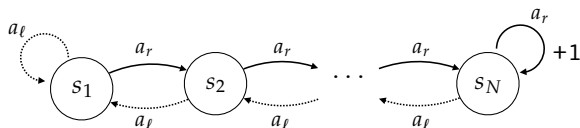
Figure 2. The structure of the N-Chain MDP.

A Small Example. First, we examine the BPD of a simple set of MDPs. Because the approximation algorithm described by Theorem 2 requires sampling a particular distribution, we here only examine hardness of small MDPs. Specifically, we consider an N-Chain graph MDP with N states and 2 actions, pictured in Figure 2. The a_r action moves the agent right, and the a_ℓ action moves the agent left along the chain. In the rightmost state of the chain, the reward for a_r is $+1$. Otherwise, rewards are determined by constants R_r and R_ℓ. As shown in Figure 1, we calculate the BPD of two variants of N-Chain for different settings of τ. In the left figure, we inspect a variation of N-Chain where $R_\ell = 0.05$ and $R_r = -0.2$, ensuring that the locally greedy action will take a_ℓ rather than a_r. In contrast, in the center figure, we inspect N-Chain where $R_\ell = 0.5$ and $R_r = 0$. Here, many more policies will be considered good for most values of τ. Indeed, as expected, we find that as we vary τ from VMIN to VMAX, the hardness of the case on the left sharply increases, whereas this increase is considerably more gradual in (b). In the right figure, we inspect the BPD of the two N-chain instances as we vary N for a particular choice of τ. As expected, we find that the harder variant of N-chain has considerably higher BPD, and that the impact of increased N is most dramatic early on. Further details and an additional experiment are presented in the appendix.

Future Work. We foresee many avenues for further research extending the BPD. First, we might remove the BPD's dependence on choice of τ by instead measuring the *cumulative* BPD across all choices of τ ($\int_{\mathrm{VMIN}}^{\mathrm{VMAX}} \mathrm{BPD}_\tau(M) d\tau$), rather than with respect to a fixed τ. Analysis and computation of this cumulative quantity are a clear direction for future work. Next, the BPD might be useful to assess the contribution made by different kinds of structures and priors in RL—which ones most dramatically reduce the BPD? The BPD could be used both as an objective (find the structure that minimizes BPD), and as an evaluation (does structure X or Y most reduce BPD?). Lastly, it is natural to extend the BPD to infinite MDPs, stochastic policies, and perhaps, learning algorithms.

Acknowledgements

References

Arumugam, D., Henderson, P., and Bacon, P.-L. An information-theoretic perspective on credit assignment in reinforcement learning. *arXiv preprint arXiv:2103.06224*, 2021.

Azar, M. G., Munos, R., and Kappen, H. J. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3), 2013.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2009.

Bellemare, M. G., Ostrovski, G., Guez, A., Thomas, P. S., and Munos, R. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Diuk, C., Cohen, A., and Littman, M. L. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2008.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *Proceedings of the International Conference on Learning Representations*, 2019.

Farahmand, A.-m. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2011.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

Goschin, S., Weinstein, A., Littman, M. L., and Chastain, E. Planning in reward-rich domains via PAC bandits. In *European Workshop on Reinforcement Learning*, 2013.

Guestrin, C., Koller, D., and Parr, R. Max-norm projections for factored MDPs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jerrum, M. *Counting, sampling and integrating: Algorithms and complexity*. Springer Science & Business Media, 2003.

Jerrum, M. and Sinclair, A. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are PAC-learnable. In *Proceedings of the International Conference on Machine Learning*, 2017.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Kearns, M., Mansour, Y., and Ng, A. Y. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3), 2002.

Lattimore, T. and Munos, R. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, 2014.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the International Conference on Machine Learning*, 2020.

Maillard, O.-A., Mann, T. A., and Mannor, S. "How hard is my MDP?" the distribution-norm to the rescue. In *Advances in Neural Information Processing Systems*, 2014.

Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12), 2009.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, 2014.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, 2013.

Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Proceedings of the Conference on Learning Theory*, 2019.

Tirinzoni, A., Lazaric, A., and Restelli, M. A novel confidence-based algorithm for structured bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

Van Roy, B. and Dong, S. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 2020.

Wen, Z. and Van Roy, B. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, 2013.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the International Conference on Machine Learning*, 2019.

A. Proofs

We first provide proofs of central results.

Proof of Proposition 1.

First, simply note that by definition $\mathrm{BPD}_\tau(M)$ is minimal when *all* policies are good (so $V^\pi(s_0) > \tau, \forall_{\pi \in \Pi_M}$, yielding $\mathrm{BPD}_\tau(M) = 0$. By the same reasoning, when the value of every policy is greater than τ, we find $\mathrm{BPD}_\tau(M)$ is maximized at 1.

Second, note that for any choice of rational $q \in [0, 1]$, there must exist integers z_1 and z_2 such that $q = \frac{z_1}{z_2}$. Since q can be zero, but non-negative, we note that z_1 and z_2 must take on non-negative quantities, too. Note that for any q, there exists an x and a finite MDP with $|\mathcal{S}|$ and $|\mathcal{A}|$ such that $\frac{x}{|\Pi_M|} = \frac{x}{|\mathcal{A}|^{|\mathcal{S}|}}$ (in the trivial case, $|\mathcal{S}| = 1$, and $|\mathcal{A}|$ can be any natural number). Then, choose T, R, and γ such that $|\{\pi \in \Pi_M : V^\pi(s_0) \geq \tau\}| = x$. Note again that this may achieved in the trivial case when

$|\mathcal{S}| = 1$, and $R(s_0, \pi_i(s_0)) > \tau(1 - \gamma)$, for each π_i in the good set, and $R(s_0, \pi_i(s_0)) \leq \tau(1 - \gamma)$ for each π_i in the bad set.

Third, note that as τ increases, the number of policies that are considered bad cannot decrease, as any policy where $V^\pi(s_0) \leq \tau$ will also adhere to $V^\pi(s_0) \leq \tau + \epsilon$, for arbitrarily small positive ϵ. □

Proof of Proposition 2.

By Hoeffding's inequality, we bound the number of episodes of a given policy π needed to obtain an $\varepsilon \in \mathbb{R}$ accurate estimate of $V^\pi(s_0)$ as follows,

$$\Pr\{|\hat{V}^\pi(s_0) - V^\pi(s_0)| \leq \varepsilon\} \geq 1 - \delta_1, \quad (7)$$

with $\delta_1 = 2 \exp\left(-\frac{2m^2\varepsilon^2}{m\mathrm{VMAX}^2}\right)$. We let $\varepsilon = \eta \geq \mathrm{VMAX} - \tau$, for the chosen τ. Moreover we run each policy for $h = \frac{1}{1-\gamma}$ steps per episode to yield the estimate $\hat{V}$.

Next, note that BPD induces a geometric distribution on the number of sampled policies needed to sample at least one good policy. Specifically we can bound the error probability as:

$$(1 - \mathrm{BPD}_\tau(M))^k \leq \delta_2, \quad (8)$$

for k the number of evaluated policies, and $\delta_2 \in (0, 1)$ a confidence parameter. Then, for

$$k = \log_{1-\mathrm{BPD}_\tau}(\delta_2) \quad (9)$$

Thus, after kmh time-steps, we will find a policy with $V^\pi(s_0) \geq \tau$ with high probability, where:

- k is the number of policies we need to sample.

$$k \leq \log_{1-\mathrm{BPD}_\tau(M)}(\delta_2).$$

- m is the number of episodes we need to evaluate each policy.

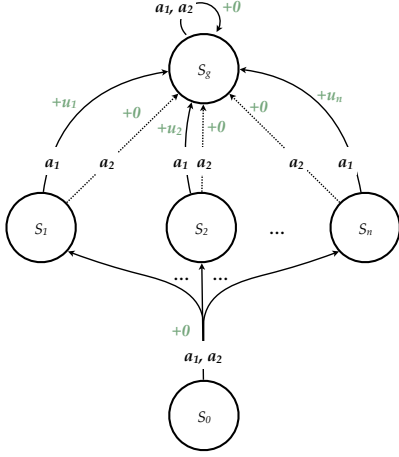$$m \leq \frac{\mathrm{VMAX}^2 \ln\left(\frac{2}{\delta_1}\right)}{2\eta^2}$$

Figure 3. Given an instance of subset sum with $\mathcal{U} = \{u_1, \dots, u_n\}$, we construct the above MDP and use a BPD solver to solve the subset sum instance.

- h is the number of steps per-episode we let each policy run,

$$h = \frac{1}{1 - \gamma}. \tag{10}$$

Hence, letting the mistake bound $\eta \geq V^*(s_0) - \tau$,

$$kmh \leq \log_{1-\mathrm{BPD}_\tau(M)} (\delta_2) \frac{\mathrm{VMAX}^2 \ln\left(\frac{2}{\delta_1}\right)}{2\eta^2(1 - \gamma)}, \tag{11}$$

$$= \log_{1-\mathrm{BPD}_\tau(M)} (\delta_2) \frac{\mathrm{RMAX}^2 \ln\left(\frac{2}{\delta_1}\right)}{2\eta^2(1 - \gamma)^3}. \tag{12}$$

$\square$

Proof of Theorem 1.

Recall that an instance of the SUBSETSUM problem is given by a finite, non-empty set of N non-negative integers $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ and a target value t. In the decision version of the SUBSETSUM problem, we ask if there exists a subset $X \subseteq \mathcal{U}$ such that $\sum_{x \in X} x = t$. Recall that this problem is NP-complete.

Given an instance of SUBSETSUM as defined above, we construct an MDP $M = \langle \mathcal{S}, \mathcal{A}, R, T, \gamma \rangle$ as follows: add a single state s_i to $\mathcal{S}$ for each element

$u_i \in \mathcal{U}$. Also add two more states, an initial state s_0 and terminal state s_T. There are two discrete actions $\mathcal{A} = \{a_1, a_2\}$ and, recalling the correspondence between states and elements of $\mathcal{U}$, we define rewards as $R(s_i, a_1) = u_i$ and $R(s_i, a_2) = 0$, $\forall i \in [N]$. All rewards for the initial and terminal states are 0. We define the transition function as $T(s_i, a, s_T) = 1, \forall a \in \mathcal{A}, i \in [N]$ such that taking either action from a state corresponding to an element of $\mathcal{U}$ leads directly to the terminal state. Moreover, $T(s_0, a, s_i) = \frac{1}{N}, \forall a \in \mathcal{A}, i \in [N]$ which, from the initial state, transitions to a state matching an element of $\mathcal{U}$ uniformly at random, under either action. Finally, let the discount factor $\gamma = 1$. This MDP is pictured in Figure 3.

Consider any deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ for MDP M and construct a subset $X = \{u_i : u_i \in \mathcal{U}, s_i \in \mathcal{S} \setminus \{s_0, s_T\}, \pi(s_i) = a_1\} \subseteq \mathcal{U}$ consisting of all states in $\mathcal{S}$ corresponding to elements of $\mathcal{U}$ where policy π takes action a_1. Examining the Bellman equation for the value function induced by policy π, we see that:

$$V^\pi(s_0) = R(s_0, \pi(s_0))$$
$$+ \gamma \sum_{s' \in \mathcal{S}} T(s' \mid s_0, \pi(s_0)) V^\pi(s')$$
$$= \frac{1}{N} \sum_{i=1}^{N} V^\pi(s_i)$$
$$= \frac{1}{N} \sum_{i=1}^{N} R(s_i, \pi(s_i))$$
$$= \frac{1}{N} \sum_{i=1}^{N} u_i \mathbb{1} \{\pi(s_i) = a_1\}$$
$$= \frac{1}{N} \sum_{x \in X} x$$

From this, we see that computing $N \cdot V^\pi(s_0)$ for any policy π yields the sum of the subset induced by π. Let $\Pi_M = \{\pi : \mathcal{S} \to \mathcal{A}\}$ be the policy class for MDP M and note that $|\Pi_M| = |\mathcal{A}|^{|\mathcal{S}|} = 2^{N+2} = 4|\mathcal{P}(\mathcal{U})|$ where $\mathcal{P}(\mathcal{U})$ denotes the power set of $\mathcal{U}$. We see that the action choices made by each policy π at states $\{s_1, \dots, s_N\}$ encode a unique subset of $\mathcal{U}$. Assume we have access to a polynomial-time algorithm for computing $\mathrm{BPD}_\tau(M)$. For a small constant $\varepsilon > 0$, consider computing $(\mathrm{BPD}_{\frac{1}{N}(t+\varepsilon)}(M) - \mathrm{BPD}_{\frac{1}{N}(t-\varepsilon)}(M))$

where

$$\mathrm{BPD}_{\frac{1}{N}(t+\varepsilon)}(M)$$

$$= \frac{1}{|\Pi_M|} \sum_{\pi \in \Pi_M} \mathbb{1}\left\{V^\pi(s_0) \le \frac{1}{N}(t+\varepsilon)\right\}$$

$$= \frac{1}{2^{N+2}} \sum_{\pi \in \Pi_M} \mathbb{1}\left\{NV^\pi(s_0) \le (t+\varepsilon)\right\}$$

$$= \frac{1}{4|\mathcal{P}(\mathcal{U})|} \sum_{\pi \in \Pi_M} \mathbb{1}\left\{\sum_{x \in X_\pi} x \le (t+\varepsilon)\right\}$$

$$= \frac{4}{4|\mathcal{P}(\mathcal{U})|} \sum_{X_\pi \in \mathcal{P}(\mathcal{U})} \mathbb{1}\left\{\sum_{x \in X_\pi} x \le (t+\varepsilon)\right\}$$

$$= \frac{1}{|\mathcal{P}(\mathcal{U})|} \sum_{X_\pi \in \mathcal{P}(\mathcal{U})} \mathbb{1}\left\{\sum_{x \in X_\pi} x \le (t+\varepsilon)\right\}$$

where the set X_π is constructed for each policy $\pi \in \Pi_M$ as described above. Consequently, we've shown that $\mathrm{BPD}_{\frac{1}{N}(t+\varepsilon)}(M)$ equals the fraction of subsets of $\mathcal{U}$ whose total element-wise sum is upper bounded by $(t+\varepsilon)$; a similar statement follows for $\mathrm{BPD}_{\frac{1}{N}(t-\varepsilon)}(M)$. With an arbitrarily small constant $\varepsilon > 0$, we can use our polynomial-time algorithm for the BPD problem to compute $\mathbb{1}\left\{(\mathrm{BPD}_{\frac{1}{N}(t+\varepsilon)}(M) - \mathrm{BPD}_{\frac{1}{N}(t-\varepsilon)}(M)) > 0\right\}$ and determine the existence of a subset $X \subseteq \mathcal{U}$ such that $\sum_{x \in X} x = t$, with arbitrarily high accuracy. Thus, we arrive at a polynomial-time algorithm for the decision version of the SUBSETSUM problem and the BPD problem must be NP-hard. $\qquad \square$

Proof of Proposition 3.

Our algorithm is as follows. We sample $k = \Theta\left(\frac{1}{\varepsilon}\log n\right)$ policies uniformly at random for some sufficiently large hidden constant. We return as our estimate of the good policy density $\widehat{\mathrm{GPD}}_\tau(M)$ the fraction of these policies that achieve value at least τ on s_0. Recall the Chernoff-Hoeffding bound which states that given $X := \sum_{i=1}^k X_i$ where each X_i is an i.i.d. Bernoulli with $a_i \le X_i \le b_i$ with probability 1 we have

$$\Pr[|X - E[X]| \ge \varepsilon] \le 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^k (b_i - a_i)}\right).$$

We apply this bound where each X_i corresponds to one of our samples and is $1/k$ if in the sampled policy we have value at least τ on s_0 and 0 otherwise. Notice that $a_i = 0$ and $b_i = 1/k$ for every i. Thus, we have $\mathrm{GPD}_\tau(M) = E[X]$ and $\widehat{\mathrm{GPD}}_\tau(M) = X$ and so

$$\Pr[|\widehat{\mathrm{GPD}}_\tau(M) - \mathrm{GPD}_\tau(M)| \ge \varepsilon]$$

$$\le 2 \exp\left(\frac{-2\varepsilon^2}{1/k^2}\right)$$

$$= 2 \exp(-\Theta(\log n))$$

$$= \frac{1}{\mathrm{poly}(n)}.$$

as desired. $\qquad \square$

Proof of Theorem 2.

For this proof we will assume that every state has exactly two actions a_1 and a_2. The proof easily generalizes to general action spaces.

In many ways the principle challenge in accurately estimating the good policy density is overcoming the "needle in the haystack" situation. In particular, if a constant fraction of all policies have value at least τ on the initial state s_0 then we can simply repeatedly sample a policy uniformly at random and then use the fraction of policies with value at least τ on s_0 among all sampled policies as our estimate of the good policy density. Standard Chernoff-bound-type arguments will show that such an estimate is very accurate with only polynomially-many samples.

However, that's assuming a whole lot of needles! If, on the other hand, a very small number of policies have value at least τ on s_0 then no such sampling strategy could possibly work efficiently as one would never sample a policy with value at least τ on s_0.

As an alternative we will utilize a general strategy first proposed by (Jerrum & Sinclair, 1989) to estimate the permanent of a matrix (which is computationally equivalent to estimating the number of matchings in a bipartite graph).

We will sketch this strategy in terms of MDPs. We will construct a sequence of MDP-like things $(M_i)_{i=1}^k$ where M_k is the original MDP in which we would like to estimate the good policy density. Let T_i be the policies in M_i with value at least τ

on s_0 and let $N_i := |T_i|$. We can estimate the good policy density by dividing N_k by the total number of policies. To estimate N_k we observe by a simple telescoping multiplication that

$$N_k = \frac{N_k}{N_{k-1}} \cdot \frac{N_{k-1}}{N_{k-2}} \cdot \ldots \cdot \frac{N_1}{N_0} \cdot N_0.$$

Thus, to estimate the good policy density it suffices to estimate every $\frac{N_{i+1}}{N_i}$ and N_0.

Let us suppose that N_0 is trivial to estimate given the way we constructed our sequence. Why should we expect that estimating the ratio $\frac{N_{i+1}}{N_i}$ is any easier than just estimating N_k? Well suppose that our sequence satisfied the following two smoothness properties

1. $T_{i+1} \subseteq T_i$

2. $\frac{1}{2} \leq \frac{N_{i+1}}{N_i}$

The existence of such smoothness properties suggests the following algorithm for estimating $\frac{N_{i+1}}{N_i}$: sample polynomially many policies independently and uniformly at random from T_{i+1} and estimate $\frac{N_{i+1}}{N_i}$ as the proportion of sampled policies that are also in T_i. Since $\frac{1}{2} \leq \frac{N_{i+1}}{N_i}$ and $T_{i+1} \subseteq T_i$ we know that a constant fraction of the time when we take a sample and perform our check the sampled policy will indeed be in T_{i+1}; by standard Chernoff bound-type arguments mentioned above we will attain a good estimate of our ratio. Thus, our smoothness property has greatly increased the proportion of needles to hay and in this way allows us to use a Chernoff bound.

Thus, to realize this strategy we must construct a sequence $(M_i)_{i=1}^k$ satisfying the above smoothness property where additionally N_0 is efficiently estimable.

We proceed to discuss how to construct the aforementioned sequence. Our sequence will be $M_0, M_1, \ldots, M_{|S|}$ where $M_{|S|}$ is the MDP in which we would like to estimate the number of policies whose value on s_0 is at least τ, i.e. $N_{|S|}$.

Our M_i for $i < |S|$ won't exactly be an MDP per se. Rather, M_i can be intuitively thought of as an MDP in which the agent only has control over the first i states (for some canonical ordering over states) and in the remaining states the agent is forced to behave optimally.

More formally, let $\Pi_i^\pi := \{\pi' : \pi'(s_j) = \pi(s_j) \text{ for } j \leq i\}$ be all policies consistent with π on the first i states. Since the Bellman equations have a fixed point, by viewing Π_i^π as all policies in an MDP in which every state s_j for $j \leq i$ has only the action taken by π available in it, we know that there is a policy in Π_i^π whose value is greater than or equal to that of all other policies in Π_i^π on all states. Let $(\pi_i^* \mid \pi) \in \Pi_i^\pi$ be any such optimal completion of π. We can now define the "value" of policy π in a state in M_i as follows.

Definition 4 ($V_i^\pi(s)$). $V_i^\pi(s) := V^{(\pi_i^* \mid \pi)}(s)$. *That is, the value of π in state s in M_i is the value of the optimal completion of π, i.e. $(\pi_i^* \mid \pi)$, in this state.*

Notice that under this definition $V_{|S|}^\pi(s) = V^\pi(s)$ and so indeed $M_{|S|}$ just is our original MDP.

We can now define the above set T_i.

Definition 5 (T_i). *Let* $T_i := \{\pi : \tau \leq V_i^\pi(s_0)\}$ *be all policies that in M_i have value at least τ on s_0. We call T_i the set of all τ-quality i-controlled policies.*

We begin by noting that the number of τ-quality i-controlled policies is trivial to compute in N_0 since no matter what the policy of the agent, the agent is forced to behave optimally on all states. Also, we may assume that $V^*(s_0) \geq \tau$ since otherwise the good policy density is always 0 and can be trivially computed as such.

Lemma 1. $N_0 = |A|^{|S|}$ *if* $V^*(s_0) \geq \tau$.

We now prove the above two smoothness properties.

Lemma 2. $T_{i+1} \subseteq T_i$.

Proof. Consider a $\pi \in T_{i+1}$. We will show that $\pi \in T_i$. That is, we know that $V_{i+1}^\pi(s) = (\pi_{i+1}^* \mid \pi) \geq \tau$ and we must argue that $V_i^\pi(s) = (\pi_i^* \mid \pi) \geq \tau$ for every state s. However, we need only notice that $\Pi_{i+1}^\pi \subseteq \Pi_i^\pi$ since the former is all policies consistent with π on the first $i+1$ states and the latter is all policies consistent with π on the first i states. Thus, we know that $(\pi_{i+1}^* \mid \pi) \in \Pi_i^\pi$ and so an optimal policy in Π_i^π—namely $(\pi_i^* \mid \pi)$—must achieve at least the same value as $(\pi_{i+1}^* \mid \pi)$ on every state—namely at least τ on s_0. $\square$

Lemma 3. $\frac{1}{2} \leq \frac{N_{i+1}}{N_i}$

Proof. To show that $\frac{1}{2} \leq \frac{N_{i+1}}{N_i}$, we will construct an injective function $f : T_i \setminus T_{i+1} \to T_{i+1}$. Letting $\neg a_1 = a_2$ and $\neg a_2 = a_1$, our function is $f(\pi) = \pi'$ where π' is defined as follows:

$$\pi' := \begin{cases} \neg\pi(s_j) & \text{if } j = i+1 \\ \pi(s_j) & \text{o/w} \end{cases}$$

That is, f outputs a policy that is identical to π but which takes the opposite action as π on state $i + 1$. To complete our proof we argue that if $\pi \in T_i \setminus T_{i+1}$ then $f(\pi) \in T_{i+1}$.

To do so, it suffices to argue that $(\pi_i^* \mid \pi) \in \Pi_{i+1}^{f(\pi)}$. Since $(\pi_i^* \mid \pi)$ matches π on the first i states and $f(\pi)$ matches π on the first i states, we know that $f(\pi)$ also matches $(\pi_i^* \mid \pi)$ on the first i states. However, recall that $(\pi_i^* \mid \pi)$ has value at least τ on s_0, but $(\pi_{i+1}^* \mid \pi)$ does not and so we know that $(\pi_i^* \mid \pi) \notin \Pi_{i+1}^\pi$. It must therefore be the case that $\pi(s_{i+1}) \neq (\pi_i^* \mid \pi)(s_{i+1})$ which is to say that $f(\pi)(s_{i+1}) = (\pi_i^* \mid \pi)$; that is $f(\pi)$ matches $(\pi_i^* \mid \pi)$ on the $(i+1)$th state. But then $f(\pi)$ matches $(\pi_i^* \mid \pi)$ on all of the first $(i+1)$th states and so we know that $(\pi_i^* \mid \pi) \in \Pi_i^{f(\pi)}$. $\square$

Thus, we have constructed our sequence satisfying the aforementioned smoothness properties and we can easily compute N_0. Assuming uniform sampling access to the τ-quality i-controlled policies, standard Chernoff bound arguments show we can estimate each $\frac{N_{i+1}}{N_i}$ up to a multiplicative $(1 \pm \varepsilon)$ using the previously-mentioned strategy in poly-time. By our earlier arguments this suffices to estimate the good policy density up to a multiplicative $(1 \pm \varepsilon)$ in poly-time. $\square$

B. Experimental Details

We next provide additional details about the N-chain experiment in Figure 1, and an additional experiment of the same form in the 4×3 Russell & Norvig grid world (Russell & Norvig, 2009).

N-Chain Details. As discussed, the results pictured in Figure 1 highlight the change of $\mathrm{BPD}_\tau(M)$ for different choices of τ and M. We compute the $\mathrm{BPD}_\tau(M)$ for each M by exactly computing the start-state value of every deterministic policy, which is feasible given the size of the MDPs. The start state is the left-most state in each chain, and γ was set to 0.95. For each MDP, we compute $\mathrm{BPD}_\tau(M)$ for 12 choices of τ, starting from VMIN and incrementing by
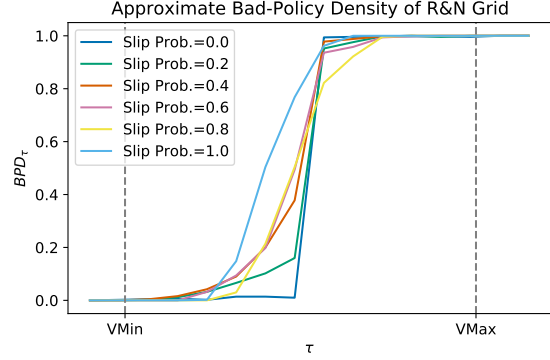
(a) With Lava
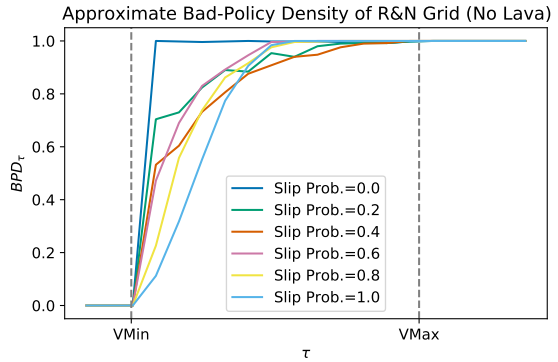
(b) No Lava

Figure 4. The BPD of variants of the Russell & Norvig grid world. On the top plot, we visualize the BPD of the traditional grid that includes a terminal lava state at $(4, 2)$, in which the agent receives -1 reward. On the bottom, we visualize the BPD of this grid world without the lava state—instead, $(4, 2)$ is a non-terminal empty cell that yields 0 reward. We vary the slip probability in each problem from 0.0 up to 1.0.

$\frac{\text{VMAX} - \text{VMIN}}{12}$ up to VMAX. In general, we anticipate that most of the interesting action will happen relatively close to $\mathrm{BPD}_\tau(M) \approx 1$, as on most MDPs of relevance the policy space likely only contains a few good policies.

Grid World. We next experiment with the Russell & Norvig grid world (Russell & Norvig, 2009), a 4×3 grid world containing a wall at $(2, 2)$, a terminal goal at $(4, 3)$ that awards $+1$ upon entering, and a terminal lava pit at $(4, 2)$ that awards a -1 upon entering. The agent starts at $(1, 1)$ and can move in each of the four cardinal directions. We introduce a slip probability ϵ in which, for each (s, a) pair there is an ϵ probability that the agent will execute an action uniformly at random with probability each time step. In our experiment, we inspect the $\mathrm{BPD}_\tau(M)$ for different settings of ϵ from 0.0 up to 1.0. Given the size of the policy space of even this small grid world, we here use the additive

approximation method to estimate $\text{BPD}_\tau(M)$ for each τ and M combination. We sample 500 policies at random and evaluate them to form our estimate of $\text{BPD}_\tau(M)$. Note that the additive approximation method is only suitable for MDPs where $\text{BPD}_\tau(M)$ is non-negligibly distant from one or zero—otherwise it is likely that estimate will always yield zero or one, despite the existence of several good or bad policies. For this reason, we emphasize the significance and important of Theorem 2, though note that there is still a remaining step to make this approach practical.

Result are presented in Figure 4. On the top, we show the estimated BPD of the grid that contains the lava cell for each value of τ. On the bottom, we show the estimated BPD of the grid with the lava cell replaced by an empty cell (so the agent receives +0 and does not terminate at $(4, 2)$). We again vary τ from VMIN up to VMAX in increments of $\frac{\text{VMAX}-\text{VMIN}}{12}$.

First note that we lose monotonicity because we are *estimating* $\text{BPD}_\tau(M)$ rather than computing it exactly. Next, observe that for any middling choice of τ (less than roughly $\frac{\text{VMAX}-\text{VMIN}}{2}$), we find the problem to be trivially easy for lower choices of slip probability when lava is present. This is because only a few policies make their way to the lava cell—thus, because value range includes the costly policies that go directly into the lava, the bad policies are only those that move directly to the lava. Even the uniform random policy only has a small probability of arriving at the lava, and thus even the light blue curve (when the slip probability is 1.0) decays toward 0 for lower values of τ. Conversely, once the τ threshold requires the agent actually reaches the goal in a timely fashion, nearly all policies are considered below optimal, and thus the problem becomes more difficult. While these dramatic swings are interesting and anticipated, we suspect that for MDPs of practical interest, $\log_{\text{BPD}_\tau(M)}$ may be the more interesting property to inspect as it will help distinguish when there are a handful of good policies as opposed to several handfuls.

In the case with no lava, we first note that the range $[\text{VMIN}, \text{VMAX}]$ is in fact different, as there is no longer a negative reward in the MDP. Hence, we find that when the threshold increases past $\tau = 0$, many policies are quickly taken to be bad, as they do not reach the goal. The higher the slip probability, the more policies that can be considered good, as the stochasticity of the environment pushes more policies toward *eventually* finding the goal. In contrast, we note that when the slip probability is zero, there are only a select few (out of $|\mathcal{A}|^{|\mathcal{S}|} = 4,194,304$) policies that ever reach the goal, which is why we see the darker blue curve rise quickly around $\tau = \text{VMIN}$.

As a final note, we observe that comparing the value *intervals* between the lava and no-lava cases gives the impression that the MDP without lava is harder, in general. This is pri-

marily due to the fact that there are few values of τ for which any policy might be considered good. In contrast, when lava is present, VMIN is considerably lower. We suggest these facts may be important when considering generalizations of the BPD that avoid explicit dependence on τ: Such generalizations may need to account for the *range* of values policies can take on. One route to incorporating such considerations is to move to a probabilistic view of the $\text{BPD}_\tau(M)$, in which we assess the *probability* of sampling a policy below a particular value threshold.