The Utility of Sparse Representations for Control in Reinforcement Learning*

Vincent Liu[1]†, **Raksha Kumaraswamy**[1]†, **Lei Le**[2], **Martha White**[1]
[1]Department of Computing Science, University of Alberta
[2]Department of Computer Science, Indiana University Bloomington

Abstract

We investigate sparse representations for control in reinforcement learning. We begin by demonstrating that learning a control policy incrementally with a representation from a standard neural network fails in classic control domains, whereas learning with a representation obtained from a neural network that has sparsity properties enforced is effective. We provide evidence that the reason for this is that the sparse representation provides locality, and so avoids catastrophic interference.

1 Introduction

Learning performance in artificial intelligence systems is highly dependent on the data representation—the features. An effective representation captures important attributes of the state, as well as simplifies the estimation of predictors. Consider a reinforcement learning agent. A local representation enables the agent to more feasibly make accurate predictions for that local region, because the local dynamics are likely to be a simpler function than learning global dynamics. Additionally, such a representation can help prevent forgetting or interference [McCloskey and Cohen, 1989, French, 1991], by only updating local weights, as opposed to dense representations where any update would modify many weights. At the same time, it is important to have a distributed representation [Bengio, 2009, Bengio et al., 2013], where the representation for an input is distributed across multiple features or attributes, promoting generalization and a more compact representation.

Such properties can be well captured by sparse representations: those for which only a few features are active for a given input. Enforcing sparsity promotes identifying key attributes, because it encourages the input to be well-described by a small subset of attributes. Sparsity then promotes locality, because local inputs are likely to share similar attributes (similar activation patterns) with less overlap to non-local inputs. In fact, many hand-crafted features are sparse representations, including tile coding [Sutton, 1996, Sutton and Barto, 1998], radial basis functions and sparse distributed memory [Kanerva, 1988, Ratitch and Precup, 2004].

Traditionally, sparse representations have been common for control in reinforcement learning, such as tile coding and radial basis functions [Sutton and Barto, 1998]. They are effective for incremental learning, but can be difficult to scale to high-dimensional inputs because they grow exponentially with input dimension. Neural networks much more feasibly enable scaling to high-dimensional inputs, such as images, but can be problematic when used with incremental training. Instead, techniques like target networks, inspired by batch methods such as fitted Q-iteration [Riedmiller, 2005], have been necessary for many of the successes of control with neural networks. We provide some evidence in this paper that this modification is necessary with dense, but not sparse networks because the reinforcement learning agent bootstraps off its own estimates. If the value in other states are overwritten, the agent will bootstrap off inaccurate estimate. Local representations, however, are much less likely to suffer from interference and these issues with bootstrapping. Learned sparse representations, then, are a promising strategy to obtain the benefits of previously common, fixed sparse representations with the scaling of neural networks.

Learning sparse representations, however, does remain a challenge. There have been some approaches developed to learning sparse representations incrementally, particularly through factorization ap-

*The full paper has been submitted to AAAI 2019.

†The authors contribute equally to this paper.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

proaches for dictionary learning [Mairal et al., 2009, 2010, Le et al., 2017] or for general sparse distributions [Olshausen and Field, 1997, Olshausen, 2002, Teh et al., 2003, Ranzato et al., 2006, 2007, Lee et al., 2008], like Boltzmann machines. Many of the methods for general sparse distribution, however, are expensive or complex to train and those based on sparse coding have been found to have serious out-of-sample issues [Mairal et al., 2009, Lemme et al., 2012, Le et al., 2017]. Moreover, there are some methods using feedforward neural network architectures, including k-sparse auto-encoders [Makhzani and Frey, 2013] and Winnner-Take-All auto-encoders [Makhzani and Frey, 2015]. However, in our experiments with these methods, we found they can be problematic because they tend to truncate non-negligible values or produce insufficiently sparse representations.

In this work, we highlight that learned sparse representations can significantly improve control performance, under an incremental learning setting, compared to a dense representation. We visualize the activation of the hidden nodes for the sparse representation as well as the action-values for particular states. These provide evidence that locality helps avoid catastrophic interference and improves accuracy of action-values for bootstrapping. In the Appendix, we discuss an effective strategy for learning sparse representations and compare it to other approaches which aim to do so.

2 Background

In reinforcement learning (RL), an agent interacts with its environment, receiving observations and selecting actions to maximize a reward signal. The environment is formalized by a Markov decision process (MDP), with states $\mathcal{S}$, actions $\mathcal{A}$, transition probabilities $\Pr : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, rewards $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and discount function $\gamma : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ [White, 2017].

One algorithm for on-policy control is Sarsa, where the agent updates its action-values for its current policy and acts near-greedily according to these action-values. The action-values for a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are the expected return for that policy, starting from state s and action a:

$$Q^\pi(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a], \text{ where } G_t = R_{t+1} + \gamma_{t+1} G_{t+1}.$$

These action-values can be estimated with function approximation, such as with neural network. Because the expected return is a real-value target, such a neural network typically uses a linear activation on the last layer:

$$Q^\pi(s, a) \approx \hat{Q}_{\mathbf{w},\boldsymbol{\theta}}(s, a) := \boldsymbol{\phi}_{\boldsymbol{\theta}}(s, a)^\top \mathbf{w} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weights in the last layer and $\boldsymbol{\phi}_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is the *representation* learned by the network with weights $\boldsymbol{\theta}$, composed of all the hidden layers in the network. The function $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s, a)$ corresponds to the last layer in the network, with $\boldsymbol{\theta}$ the weights of the network. The efficacy of the action-value approximation, therefore, relies on this representation $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s, a)$.

3 The Utility of Sparsity for Control

We highlight the utility of sparsity for control via experiments in four benchmark RL domains: Mountain Car, Puddle World, Acrobot and Catcher. The experimental set-up is two-stage where a representation is learned with a mean-squared temporal difference error (MSTDE) objective and the applicable regularization strategies. This learned representation is then fixed, and used by a (fully incremental) Sarsa(0) agent for learning a control policy, where only the weights $\mathbf{w}$ on the last layer are updated. We provide details about the domains, and experimental setup in the Appendix. We choose this two-stage training regime to remove confounding factors in training neural networks incrementally. Our goal here is to identify if a sparse representation can improve control performance, and if so, why. The networks are trained with an objective for learning values, on a large batch of data generated by a policy that covers the space; the learned representations are capable of representing the optimal policy. We investigate their utility for fully incremental learning. Outside of this carefully controlled experiment, we advocate for learning the representation incrementally, for the task faced by the agent.

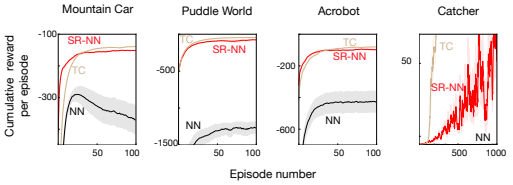

Figure 1: Learning curves for Sarsa(0) comparing SR-NN, Tile Coding and vanilla NN in the four domains. Both neural nets are of two layers with size 32 and 256, with ReLU activations.

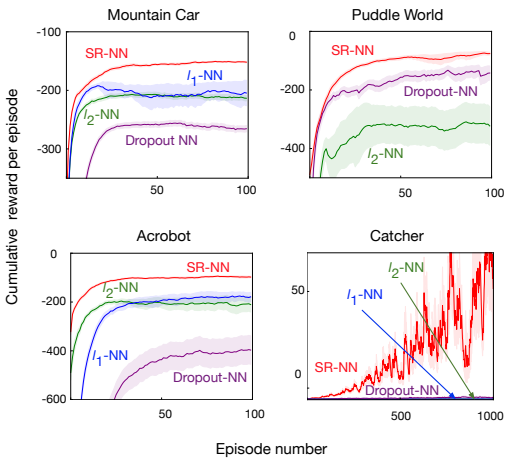

Figure 2: Learning curves for Sarsa(0) comparing SR-NN to the regularized representations. ℓ_1 does poorly in Puddle World, and is not visible. In Figure 2, we can see that regularization is unlikely to account for the improvements in control. ℓ_1-NN and ℓ_2-NN perform well in Mountain Car during early learning, but fail in other domains. Dropout-NN performs poorly in all domains except Puddle World. Interestingly, in this one domain, Dropout-NN appears to have learned a sparse representation, based on the heatmap shown in Figure 3. It has been observed that Dropout can at times learn sparse representations [Banino et al., 2018], but not consistently, as corroborated by our experiments.

We compare a learned sparse representation (SR-NN) to tile coding (TC) - a static representation, that is known to perform very well in these benchmark domains [Sutton and Barto, 1998], and dense representation (NN) in Figure 1. The NNs perform surprisingly poorly, in some case increasing and then decreasing in performance (Mountain Car), and in others failing altogether (Catcher). In all the benchmark RL domains, TC performs well, as expected. Specifically in Catcher, TC learns a close-to-optimal policy as the representation is powerful. The learned SR-NN performs as well in all domains, and is effective for learning in Catcher. Although, SR-NN and NN representations were trained in the same regime, the sparsity of SR-NN enables the Sarsa(0) agent to learn, whereas the regular feed-forward NN does not. We investigate this effect further in the next sets of experiments, to better understand the phenomenon.

To determine if the main impact of the sparse representation is simply from regularization, preventing overfitting, we tested several regularization strategies for the neural network - ℓ_2 and ℓ_1 on the weights of the network (ℓ_2-NN and ℓ_1-NN respectively) and dropout on the activation [Srivastava et al., 2014] (Dropout-NN). In Figure 2 we see SR-NN performs well across all domains, whereas none of the regularization strategies consistently perform well.

We next investigate the hypothesis that locality is preventing catastrophic interference by visualizing the locality of the representations, as well as examining the bootstrap values over time. We show results for Puddle World here, as it is an interpretable two-dimensional domain. Figure 3(a) shows the activation map of randomly selected hidden neurons with the different networks. We can see that each hidden neuron in SR-NN only responds to a local region of the input space, while some hidden neurons in NN respond to a large part of the space. Consequently, when one state is updated in a part of the space with the NN representation, it is more likely to significantly shift the values in other parts of the space, as compared to the more local SR-NN. The ℓ_2-NN, and ℓ_1-NN representations do not exhibit any discernible locality properties. Dropout-NN does achieve some degree of locality in this domain, as mentioned earlier. To show the stability (or lack of stability) of bootstrap targets used during control, we select five states and evaluate their action-values for the optimal action over the course of learning. These states are distributed across the observation space, depicted in Figure 3(b). The bootstrap estimates, that correspond to the algorithm settings for the learning curves, are plotted in Figure 3(c). We can see that the relative ordering of the value estimates is maintained with SR-NN and Dropout-NN, which were the two NNs effective for on-policy control, and that their values converge to near the true values (given in Figure 3(d)). The other representations, on the other hand, have very poor estimates. Moreover, these estimates seem to decrease together, suggesting interference is causing overgeneralization to reduce values in other states.

Finally, we report additional measures of locality, to determine if the successful methods are indeed sparse. The heatmaps provide some evidence of locality, but are more qualitative than quantitative. We report two qualitative measures: *instance sparsity* and *activation overlap*.

Instance sparsity corresponds to the percentage of active units for each input. A sparse representation should be instance sparse, where most inputs produce relatively low percentage activation, as shown in Figure 4. SR-NNs have consistently low instance sparsity across all four domains, with slightly higher level in Catcher, potentially explaining the noisy behaviour in that domain. The NNs representation, which has no regularization, has some instance sparsity, likely due to simply using ReLU activation. Interestingly, ℓ_1-NN and ℓ_2-NN actually produced less instance sparsity.

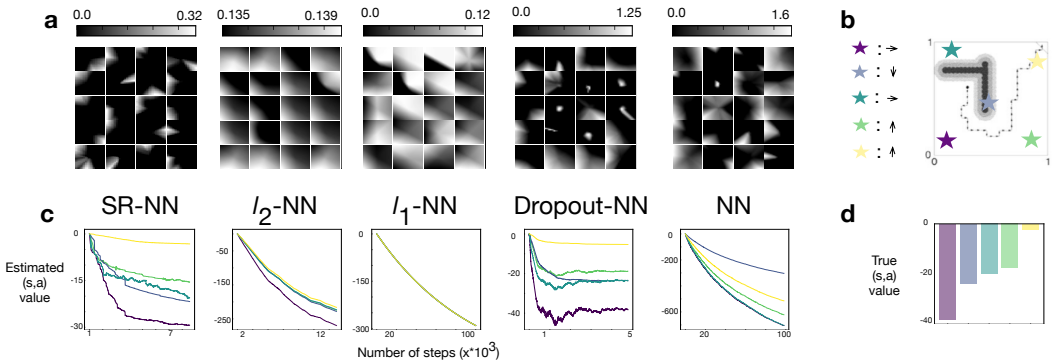

Figure 3: A study in Puddle World to investigate the effect of locality during on-policy control. (**a**) The activation maps for 20 randomly chosen neurons for different representations - each cell in the heatmap corresponds to the complete Puddle World state-space. (**b**) A visualization of the domain, denoting the selected state-action pairs used in the analysis. (**c**) The estimated state-action values for the selected configurations during on-policy control with Sarsa(0) ($\epsilon = 0.1$), while utilizing the specific representation of interest. (**d**) The true state-action values for the selected configurations with an $\epsilon = 0.1$-optimal policy, estimated from 100k Monte Carlo rollouts.

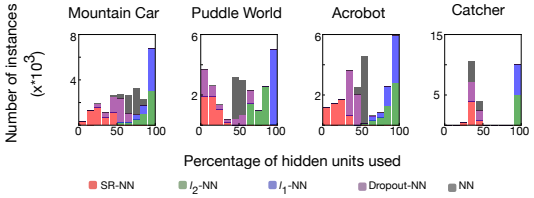

Figure 4: Instance sparsity comparing SR-NN to the regularized variants and vanilla NN. The percentage evaluation is designed to disregard units that are never active across all samples in the batch (dead units).

Activation overlap, introduced by French [1991], reflects the amount of shared activation between any two inputs. We consider a variant of activation overlap that measures the number of shared activation between two representations, $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$, for two samples, $\mathbf{x}_1$, and $\mathbf{x}_2$: $\text{overlap}(\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)) = \sum_j \mathbb{1}[(\phi_j(\mathbf{x}_1) > 0) \wedge (\phi_j(\mathbf{x}_2) > 0)]$. We measure the activation overlap of the five chosen states, distributed across Puddle World. If the overlap between two representations is zero, the interference would be zero. Updating the value function with respect to one state, therefore, would not affect the other state's value. Figure 5 shows the average overlap for these five states, and again SR-NN has significantly lesser overlap than other approaches.

Overall, these results provide some evidence that (a) sparse representations can improve control performance in an incremental learning setting, (b) these sparse representations appear to provide locality, (c) this locality reduces interference and improves accuracy of bootstrap values in Sarsa(0. These results are a first step, and warrant further investigation. They do nonetheless motivate that learning sparse representations could be a promising direction for control in reinforcement learning.

4 Conclusion

This work highlights an important phenomenon that arises in control, beyond the typical issues with catastrophic interference. Interference is typically considered for sequential multi-task learning, where previous functions are forgotten by training on a new task. Interference could occur even in a single-task setting, if an agent remains in a particular area of the space for a long time. In reinforcement learning, however, this problem is magnified by the fact that the agent uses its own estimates as targets. If estimates change incorrectly due to interference, there could be a cascading effect. This work provides some first empirical steps, in a carefully controlled set of experiments, to identify that this could be an issue, and that sparse representations could be a promising direction to alleviate the problem. We hope for this work to spur further empirical investigation into how widespread this issue is, and further algorithmic development into learning sparse representations for reinforcement learning.

SR-NN	ℓ_2-NN	ℓ_1-NN	Dropout-NN	NN
8.8	111.5	142.5	31.2	54.0

Figure 5: Activation overlap in Puddle World. The numbers are the average overlap over all pairs of selected states. For example, SR-NN has an average of 8.8 shared activation over all pairs of 5 selected states defined in Figure 3 (a).

References

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705–1749, 2005.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 2018.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2009.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Robert M French. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Annual Cognitive Science Society Conference*, 1991.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, 2015.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. 2012.

Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Lei Le, Raksha Kumaraswamy, and Martha White. Learning sparse representations in reinforcement learning with sparse coding. *arXiv:1707.08316*, 2017.

Honglak Lee, C Ekanadham, AY Ng Advances in neural information, and 2008. Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems*, 2008.

Andre Lemme, René Felix Reinhart, and Jochen Jakob Steil. Online learning and generalization of parts-based image representations by non-negative sparse autoencoders. *Neural Networks*, 2012.

Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, 2009.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2010.

A Makhzani and B Frey. Winner-take-all autoencoders. In *Adv. in Neural Information Processing Systems*, 2015.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Michael McCloskey and Neal J Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 1989.

Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 2011.

Bruno A Olshausen. Sparse Codes and Spikes. In *Probabilistic Models of the Brain*. 2002.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 1997.

Marc'Aurelio Ranzato, Christopher S Poultney, Sumit Chopra, and Yann LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model. In *Adv. in Neural Info. Process. Sys.*, 2006.

Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun. Sparse Feature Learning for Deep Belief Networks. In *Advances in Neural Information Processing Systems*, 2007.

B Ratitch and D Precup. Sparse distributed memories for on-line value-based reinforcement learning. In *Machine Learning: ECML PKDD*, 2004.

Martin Riedmiller. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, 2005.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems*, 1996.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-Based Models for Sparse Overcomplete Representations. *Journal of Machine Learning Research*, 2003.

Martha White. Unifying Task Specification in Reinforcement Learning. In *Inter. Conf. on Machine Learning*, 2017.

A Distributional Regularizers for Sparsity

In this section, we describe how to use Distributional Regularizers to learn sparse representations with neural networks.[3] We introduce a Set Distributional Regularizer, which when paired with ReLU activations enables sparse representations to be learned, as we demonstrate in the next section. We first describe how to define Distributional Regularizers on neural networks, and then discuss the extension to a Set Distributional Regularizer, and motivation for doing so.

The goal of using Distributional Regularizers is to encourage the distribution of each hidden node—across samples—to match a desired target distribution. In a neural network, we can view the hidden nodes, $Y_1, \ldots, Y_d$, as random variables, with randomness due to random inputs. Each of these random variables Y_j has a distribution $p_{\hat{\beta}_j(\theta)}$, where the parameters $\hat{\beta}_j(\theta)$ of this distribution are induced by the weights $\boldsymbol{\theta}$ of the neural network:

$$p_{\hat{\beta}_j(\theta)}(y) = \int_{s \in \mathcal{S}} p(s) p(\boldsymbol{\phi}_{j,\boldsymbol{\theta}}(s) = y) ds.$$

This provides a distribution over the values for the feature $\boldsymbol{\phi}_{j,\boldsymbol{\theta}}(s)$, across inputs s. A Distributional Regularizer is a KL divergence $KL(p_\beta || p_{\hat{\beta}_j(\boldsymbol{\theta})})$ that encourages this distribution to match a desired target distribution p_β with parameter β.

Such a regularizer can be used to encourage sparsity, by selecting a target distribution that has high mass or density at zero. Consider a Bernoulli distribution for activations, with $Y_j \in \{0, 1\}$. Using a Bernoulli target distribution with $\beta = 0.1$, giving $p_\beta(Y = 1) = 0.1$, encodes a desired activation of 10%. As another example, for continuous nonnegative Y_j, the target distribution can be set to an exponential distribution $p_\beta(y) = \beta^{-1} \exp(-y/\beta)$, which has highest density at zero with expected value β. Setting $\beta = 0.1$ encourages the average activation to be 0.1 and increases density on $y = 0$.

The efficacy of this regularizer, however, is tied to the parameterization of the network, which should match the target distribution. For a ReLU activation, for example, which has a range $[0, \infty)$, a Bernoulli target distribution is not appropriate. Rather, for the range $[0, \infty)$, an exponential distribution is more suitable. For a Sigmoid activation, giving values between $[0, 1]$, a Bernoulli is reasonably appropriate. Additionally, the parametrization should be able to set activations to zero. The ReLU activation naturally enables zero values [Glorot et al., 2011], by pushing activations to negative values. The addition of a Distributional Regularizer simply encourages this natural tendency, and is more likely to provide sparse representations. Activations under Sigmoid and tanh, on the other hand, are more difficult to encourage to zero, because they require highly negative input values or input values exactly equal to 0.5, respectively, to set the hidden node to zero. For these reasons, we advocate for ReLU for the sparse layer, with an exponential target distribution.

Finally, we modify this regularizer to provide a Set Distributional Regularizer, which does not require an exact level of sparsity to be achieved. It can be difficult to choose a precise level of sparsity, making the Distributional Regularizer prone to misspecification. Rather, the actual goal is typically to obtain *at least* some level of sparsity, where some nodes can be even more sparse. For this modification, we specify that the distribution should match any of a set of target distributions Q_β, giving a *Set KL*: $\min_{p \in Q_\beta} KL(p || p_{\hat{\beta}_j(\boldsymbol{\theta})})$. Generally, this Set KL can be hard to evaluate. However, as we show below, it corresponds to a simple clipped KL-divergence for certain choices of Q_β, importantly including for exponential distributions where $Q_\beta = \{p_{\tilde{\beta}} \,|\tilde{\beta} \leq \beta\}$.

Theorem 1 (Set KL as a Clipped-KL)**.** *Let p_η be a one-dimensional exponential family distribution with the natural parameter η, $B = [\eta_1, \eta_2]$ be a convex set in the natural parameter space and $Q_B = \{p_\eta : \eta \in B\}$. Then the Set KL divergence*

$$SKL(Q_B || p_\eta) := \min_{p \in Q_B} KL(p || p_\eta) \tag{2}$$

[3]The idea was originally introduced for neural networks with Sigmoid activations in an unpublished set of notes [Ng, 2011], and as yet has not been systematically explored. When used out-of-the-box, we found important limitations in the learned representations, including from using Sigmoid activations instead of ReLU and from using the KL to a specific distribution. We explore the idea in-depth here, to make it a practical option for learning sparse representations.

is (a) non-negative (b) convex in η and (c) corresponds to a simple clipped form

$$SKL(Q_B||p_\eta) = \begin{cases} KL(p_{\eta_2}||p_\eta) & \text{if } \eta > \eta_2 \\ KL(p_{\eta_1}||p_\eta) & \text{if } \eta < \eta_1 \\ 0 & \text{else} \end{cases} \tag{3}$$

Proof. For exponential families, the KL divergence correspond to a Bregman divergence [Banerjee et al., 2005]:

$$KL(p_{\eta_1}||p_\eta) = D_F(\eta||\eta_1)$$

for a convex potential function F that depends on the exponential family. Hence, we have

$$SKL(Q_B||p_\eta) = \arg\min_{\tilde{\eta}\in B} D_F(\eta||\tilde{\eta})$$

If $\eta \in B$, this minimum over Bregman divergences is clearly zero. If $\eta < \eta_1$ and $\eta > \eta_2$, we have to consider the minimization. The Bregman divergence is not necessarily convex in the second argument. Instead, we can rely on convexity of the set B. Taking the derivative of $D_F(\eta||\tilde{\eta})$ wrt $\tilde{\eta}$, we get

$$\begin{aligned} \frac{d}{d\tilde{\eta}} D_F(\eta||\tilde{\eta}) &= \frac{d}{d\tilde{\eta}} \left[F(\eta) - F(\tilde{\eta}) - (\eta - \tilde{\eta}) \frac{d}{d\tilde{\eta}} F(\tilde{\eta}) \right] \\ &= -\frac{d}{d\tilde{\eta}} F(\tilde{\eta}) + \frac{d}{d\tilde{\eta}} F(\tilde{\eta}) - (\eta - \tilde{\eta}) \frac{d^2}{d\tilde{\eta}^2} F(\tilde{\eta}) \\ &= -\frac{d^2}{d\tilde{\eta}^2} F(\tilde{\eta})(\eta - \tilde{\eta}) \end{aligned}$$

Now because F is convex, $-\frac{d^2}{d\tilde{\eta}^2} F(\tilde{\eta})$ is always negative. The derivative, then, is negative when $\tilde{\eta} < \eta$, indicating $\tilde{\eta}$ should be increased to decrease $D_F(\eta||\tilde{\eta})$. Similarly, when $\tilde{\eta} > \eta$, the derivative is positive, indicating $\tilde{\eta}$ should be decreased to decrease $D_F(\eta||\tilde{\eta})$. This derivative, then, points $\tilde{\eta}$ to the boundaries when $\eta \notin B$, respectively to the boundary points closest to η. □

Corollary 1 (SKL for Exponential Distributions)**.** *For p_β an exponential distribution, with natural parameter $\eta = \beta^{-1}$, and $B = (0, \beta]$, then*

$$SKL(Q_B||p_{\hat{\beta}}) = \begin{cases} \log\hat{\beta} + \frac{\beta}{\hat{\beta}} - \log\beta - 1 & \text{if } \hat{\beta} > \beta \\ 0 & \text{else} \end{cases} \tag{4}$$

We use the SKL in Corollary 1, to encode a sparsity level of at least β—rather than exactly β—for the last layer in a two-layer neural network with ReLU activations. This regularizer was used to for SR-NN in the preceding section. We include pseudocode for optimizing the regularized objective with the SKL in Algorithm 1.

B Additional algorithmic details

In general, we advocate for learning the representation incrementally, for the task faced by the agent. However, for our experiments, we learned the representations first to remove confounding factors. We detail that learning regime here.

The problem of learning a good representation $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s, a)$ in the case of finite actions can be transformed to learning a good representation of the form $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s)$, and using that to represent the action-value function from Equation (1) as:

$$\hat{Q}_{\mathbf{w},\boldsymbol{\theta}}(s, a) := \boldsymbol{\phi}_{\boldsymbol{\theta}}(s)^\top \mathbf{w}_a \tag{5}$$

Here, $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s)$ is the linear representation of the state s, which is used in conjunction with the linear predictor $\mathbf{w}_a$ to estimate action-values for action a across the state space. Under a given policy, like the action-values $Q^\pi(s, a)$, corresponding state-values, $V^\pi(s)$, are defined as:

$$V^\pi(s) := \mathbb{E}[G_t|S_t = s] \text{ ,where } G_t = R_{t+1} + \gamma_{t+1} G_{t+1}.$$

Algorithm 1 Optimizing the regularized objective

1: Initialize neural networks weights based on He initialization He et al. [2015]: for each layer l and each element ij of the weight matrix $\mathbf{W}_{ij}^{(l)} \sim \mathcal{N}(0, \frac{2}{n_l})$ and $\mathbf{b}^{(l)} = \mathbf{0}$ where n_l the number of input nodes for layer l.
2: **while** not converge to a minimum **do**
3: Draw m i.i.d. samples $\{y_1, ..., y_m\}$ from the true data distribution
4: For $j = 1, ..., k$, compute $\hat{\beta}_j = \sum_{i=1}^{m} y_{ij}/m$ and the gradient:

$$\frac{\partial KL(p_\beta || p_{\hat{\beta}_j})}{\partial \hat{\beta}_j} = (\frac{1}{\hat{\beta}_j} - \frac{\beta}{\hat{\beta}_j^2}) \mathbb{1}[\hat{\beta}_j > \beta]$$

5: Update each weight $\boldsymbol{\theta} \in \{\forall l, \mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}$ with the gradient:

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \lambda_{KL} \sum_{j=1}^{k} \frac{\partial KL(p_\beta || p_{\hat{\beta}_j})}{\partial \hat{\beta}_j} \frac{\partial \hat{\beta}_j}{\partial \boldsymbol{\theta}}$$

6: **end while**

An easy objective to train connectionist networks with simple backpropagation is the Mean Squared Temporal Difference Error (MSTDE) Sutton [1988]. For a given policy, the MSTDE is defined as:

$$\sum_{s \in \mathcal{S}} \mathbf{d}(s) \mathbb{E}[\delta_t^2 | S_t = s] \tag{6}$$

$$\text{where, } \delta_t := R_{t+1} + \gamma_{t+1} \boldsymbol{\phi}_{\boldsymbol{\theta}}(S_{t+1})^\top \mathbf{w}_v - \boldsymbol{\phi}_{\boldsymbol{\theta}}(S_t)^\top \mathbf{w}_v$$

Here, $\mathbf{d}$ denotes the stationary distribution over the states induced by the given policy, and $\boldsymbol{\theta}$ and $\mathbf{w}_v$ are parameters that can be estimated with stochastic gradient descent. Therefore, given experience generated by a policy that explores sufficiently in an environment, a strong function approximator (a dense neural network) can be trained to estimate useful features, $\boldsymbol{\phi}_{\boldsymbol{\theta}}(s)$. These features can then be used for estimating action-values in on-policy control for learning the (close-to) optimal behaviour policy in the environment.

C Evaluation of Distributional Regularizers

In this section, we investigate the efficacy of Distributional Regularizers for obtaining sparsity. There are a variety of possible choices with Distributional Regularizers, including activation function and corresponding target distribution and using a KL versus a Set KL. In this section, we investigate some of these combinations, particularly focusing on the difference in sparsity and performance when using (a) KL versus SKL; (b) Sigmoid (with a Bernoulli target distribution) versus ReLU (with an Exponential target distribution); and (c) previous strategies to obtain sparse representations versus the proposed variant of the Distributional Regularizer.

In the first set of experiments, we compare the instance sparsity of KL to Set KL, with ReLU activations and Exponential Distributions (ReLU+KL and ReLU+SKL). Figure 6 shows the instance sparsity for these, and for the NN without regularization. Interestingly, ReLU+KL actually reduces sparsity in several domains, because the optimization encouraging an exact level of sparsity is quite finicky. ReLU+SKL, on the other hand, significantly improves instance sparsity over the NN. This

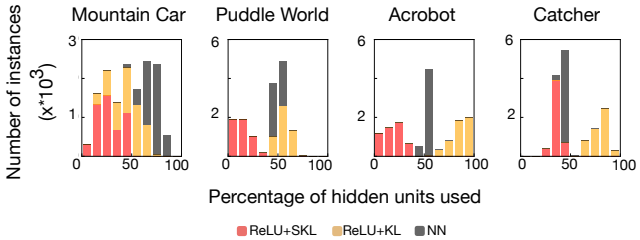

Figure 6: Instance sparsity as evaluated on a batch of test data comparing ReLU+KL and ReLU+SKL to NN. While ReLU+KL can make representations denser than just NN, ReLU+SKL always results in sparser representations.

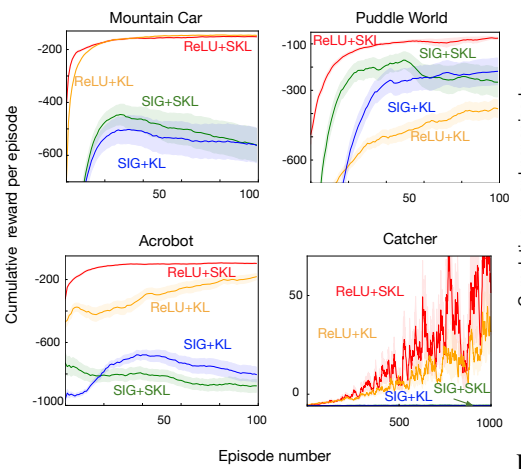

Figure 7: Learning curves for Sarsa(0) with different Distributional Regularizers.

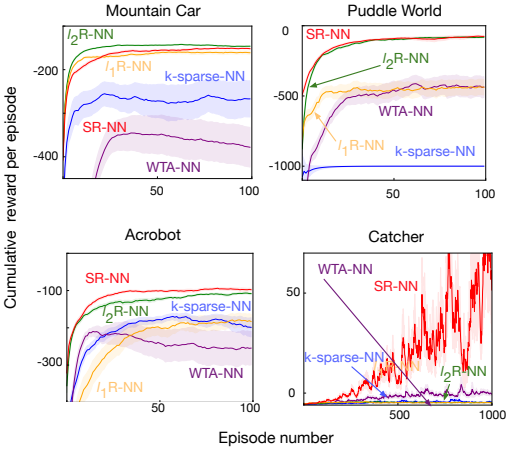

Figure 8: Learning curves for Sarsa(0) comparing SR-NN to previous proposed sparse representations learning strategies.

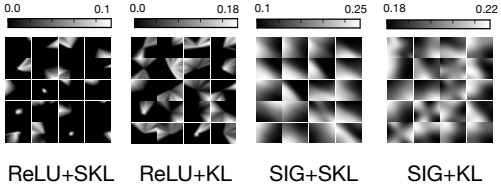

Figure 9: Heatmaps of activations with different Distributional Regularizers in Puddle World.

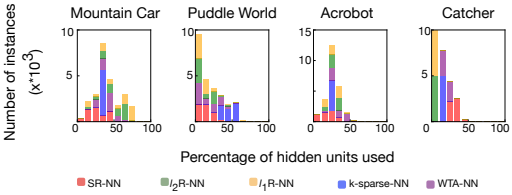

Figure 10: Instance sparsity comparing SR-NN to previous proposed sparse representations learning strategies.

instance sparsity again translates into control performance, where ReLU+KL does noticeably worse than ReLU+SKL across the four domains in Figure 7. Despite the poor instance sparsity, ReLU+KL does actually seem to provide some useful regularity, that does allow some learning across all four domains. This contrasts the previous regularization strategies, ℓ_2, ℓ_1 and Dropout, which all failed to learn on at least one domain, particularly Catcher.

In the next set of experiments, we compare Sigmoid (with a Bernoulli target distribution) versus ReLU (with an Exponential target distribution). We included both KL and Set KL, giving the combinations ReLU+KL, ReLU+SKL, SIG+KL, and SIG+SKL. We expect Sigmoid with Bernoulli to perform significantly worse—in terms of sparsity levels, locality and performance—because the Sigmoid activation makes it difficult to truly get sparse representations. This hypothesis is validated in the learning curves in Figure 7 and the heatmaps for Puddle World in Figure 9. SIG+KL and SIG+SKL perform poorly across domains, even in Puddle World, where they achieved their best performance. Unlike ReLU with Exponential, here the Set KL seems to provide little benefit. The heatmaps in Figure 9 show that both versions, SIG+KL and SIG+SKL, cover large porions of the space, and do not have local activations for hidden nodes. In fact, SIG+KL and SIG+SKL use all the hidden nodes for all the samples across domains, resulting in no instance sparsity.

Next, we compare to previously proposed strategies for learning sparse representations with neural networks. These include using ℓ_1 and ℓ_2 regularization on the activation (denoted by ℓ_1R-NN and ℓ_2R-NN respectively); k-sparse NNs, where all but the top k activations are zeroed [Makhzani and Frey, 2013]; and Winner-Take-All (WTA) NNs that keep the top k% of the activations per node across instances, to promote sparse activations of nodes over time [Makhzani and Frey, 2015].[4]

We include learning curves and instance sparsity for these methods, for a ReLU activation, in Figures 8 and 10. Neither WTA-NN nor k-sparse NN are effective. We found the k-sparse NN was prone to dead units, and often truncates non-negligible value. Surprisingly, ℓ_2R-NN performs comparably to SR-NN in all domains but Catcher, whereas ℓ_1R-NN is effective only during early learning in

[4]Both k-sparse NNs and WTA NNs were introduced for auto-encoders, though the idea can be applied more generally to NNs. We additionally tested these methods with autoencoders, but performance was significantly worse.

Mountain Car. From the instance sparsity plots in Catcher, we see that ℓ_1R-NN and ℓ_2R-NN produce highly sparse (2%-3% instance sparsity), potentially explaining its poor performance. While similar instance sparsity was effective in Puddle World, this is unlikely to be true in general. This was with considerable parameter optimization for the regularization parameter.

D Experimental Details

D.1 Policies to generate training and testing data

In Mountain Car, we use the standard energy pumping policy with 10% randomness. In Puddle World, by a policy that chooses to go North with 50% probability, and East with 50% probability on each step. The data in Acrobot is generated by a near-optimal policy. In Catcher, the agent chooses to move toward the apple with 50% probability, and selects a random action with 50% probability on each step; and gets only 1 life in the environment.

D.2 Tile Coding

For Tile Coding (TC), we experiment with several configurations for the fixed representation, particularly with grid-sizes(N) in $\{4, 8, 16\}$ and number of tilings (D) in $\{8, 16, 32\}$. We use a hash size of 8192, which is significantly larger than the largest feature size of 256, as used in the other learned representation models we compare to. The results shown in Figure 2 are for the best configuration of the static tile-coder after a sweep.

D.3 Neural networks

Architecture and optimizer: We used neural networks with two hidden layers. The first layer 32 hidden units. The second layer, which is the representation layer used for prediction, has 256 units. We optimized the neural network weights using Adam optimization Kingma and Ba [2014] with a batch size of 64. The neural network weights are initialized based on He initialization He et al. [2015]. That is, the neural networks weights are initialized with zero-mean Gaussian distribution with variance equals to $2/n_l$, where n_l is the number of input nodes for layer l. We train each neural network until convergence. That is, in our experiments, most algorithms converges within 50 epochs in Mountain Car, Puddle World and Catcher, and 100 epochs in Acrobot.

Representation hyperparameters: The range of grid search for the representation hyperparameters are as follows:

$$\lambda_{KL} \in \{0.1, 0.01, 0.001\}$$
$$\beta \in \{0.05, 0.1, 0.2\}$$
$$\lambda_{NN} \text{ for } \ell_1 \text{ and } \ell_2 \in \{0.1, 0.01, 0.001, 0.0001\}$$
$$\text{dropout probability } p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$$
$$k \text{ for k-sparse} \in \{16, 32, 64, 128\}$$
$$k \text{ for WTA} \in \{6.25\%, 12.5\%, 25\%, 50\%\}$$

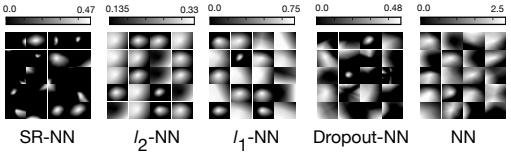

Figure 11: Heatmaps of activations comparing SR-NN to different regularization strategies and NN in Mountain Car.

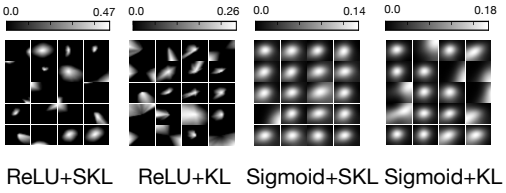

Figure 12: Heatmaps of activations with different Distributional Regularizers in Mountain Car.

Algorithmic choices: For k-sparse networks, only the top-k hidden units in the representation layer are activated. We also use scheduling of sparsity level described in the original paper Makhzani and Frey [2013]. If used in conjunction with a distributional regularizer, the top-k nodes are chosen before application of the distributional regularizer. For dropout, given the form of the supervision goal (MSTDE), the same dropout mask is chosen to generate the representation for both states S_{t+1} and S_t[5] – this preserves dropouts role as regularizer w.r.t. the target, and promotes diversity in learning.

[5]We have experimented with different dropout masks for S_{t+1} and S_t, and the result suggests that it is not able to learn good representations even for prediction across all domains.

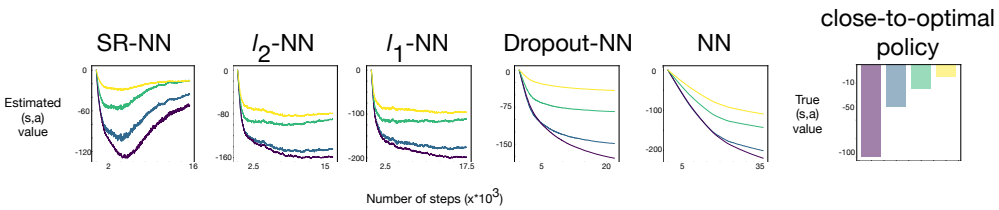

Figure 14: This plot compares the bootstrap estimates of SR-NN to various regularization stratergies during on-policy control for the chosen 4 state-action pairs denoted in the following format in the legend: <car-position,car-velocity>:action. Again, we see that the relative ordering of bootstrap values is maintained with SR-NN, and it tends towards the true values of the ($\epsilon = 0.1$)-optimal policy. The optimal policy estimates (currently) use 10k Monte Carlo rollouts with a powerful close-to-optimal tile-coder policy.

Grid-search evaluation metric: The learned representations are then used for on-policy control in Sarsa(0) with fixed $\epsilon = 0.1$. The value function for Sarsa is initialized with zero-mean Gaussian distribution with small variance. For sparse representations, we use semi-gradient Sarsa with step decay learnining rate. For dense representatinos, we use adaptive learnining rate method RMSprop Hinton et al. [2012]. The initial learning rate for Sarsa(0) is swept in the set: $\alpha_0 \in \{0.1, 0.04, 0.01, 0.004, 0.001, 0.0004, 0.0001\}$. All the sweeps for selecting the representation learning hyperparameters across domains use 50 epochs and 10 runs.

E More results

E.1 Activation heatmaps

The activation heatmaps for randomly selected neurons (excluding dead neurons) in Mountain Car with different regularization stratergies are shown in Figure 11, and with differnt Distributional Regularization designs are shown in Figure 12. Heatmaps for sparsity inducing networks with ReLU activations and Sigmoid activation, for Mountain Car and Puddle World are shown in Figure 13.

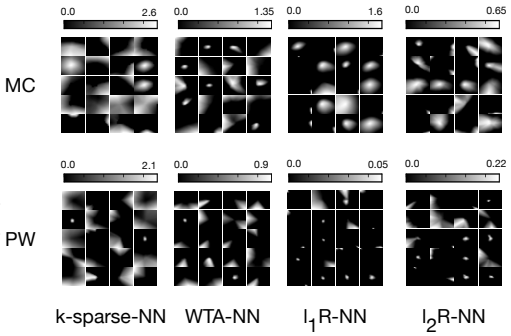

Figure 13: Heatmaps of activations for nodes from other networks which aim to generate sparse representations.

E.2 Bootstrap values

The bootstrap values comparing SR-NN to different regularization strategies, and NN are shown in Figure 14. Since it is not easy to visualize 4-dimensional space, we only include the bootstrap value result of Mountain Car here.

E.3 Activation overlap

We show the overlap of representations learned by different networks in Table 1 for Mountain Car and Puddle World. ℓ_2R-NN and ℓ_1R-NN have low overlap values. However, the regularizers tend to push many neurons to be activated for a really small region to reduce penalty as shown in Figure 13. SR-NN, on the other hand, learns a more distributed representation.

	Mountain Car	Puddle World
SR-NN	16.8	8.8
ℓ_2-NN	112.3	111.5
ℓ_1-NN	109.5	142.5
Dropout-NN	72.5	31.2
NN	106.5	54.0
ReLU+KL	36.8	71.4
SIG+SKL	256.0	256.0
SIG+KL	256.0	256.0
k-sparse-NN	36.6	61.8
WTA-NN	24.8	6.5
ℓ_2R-NN	30.0	3.8
ℓ_1R-NN	10.5	0.4

Table 1: Activation overlap in Mountain Car and Puddle World. For Mountain Car, the numbers are the average overlap over all pairs of selected states defined in Figure 14.