LESS SUBOPTIMAL LEARNING AND CONTROL IN VARIATIONAL POMDPS

Baris Kayalibay **Atanas Mirchev** **Patrick van der Smagt** **Justin Bayer**

Machine Learning Research Lab, Volkswagen Group, Munich, Germany
{bkayalibay,atanas.mirchev,bayerj}@argmax.ai

ABSTRACT

A recently uncovered pitfall in learning generative models with amortised variational inference, the conditioning gap, questions common practices in model-based reinforcement learning. Withholding a part of the quantities that the true posterior depends on from the inference network leads to a biased generative model and an approximate posterior that underestimates uncertainty. We examine the effect of the conditioning gap on model-based reinforcement learning with variational world models. We study the effect in three settings with known dynamics, which enables us to compare to a near-optimal policy. Our finding is that the impact of the conditioning gap becomes severe in systems where the state is hard to estimate.

1 INTRODUCTION

Variational state-space models (VSSMs) based on deep neural networks (Karl et al., 2017a; Fraccaro et al., 2016) have become a practical choice for learning world models for control. Their flexibility allows scaling to real-world scenarios and aids solving a broad variety of challenging problems such as robotic control, computer games and meta reinforcement learning (Becker-Ehmck et al., 2020; Hafner et al., 2020a; Zhao et al., 2020).

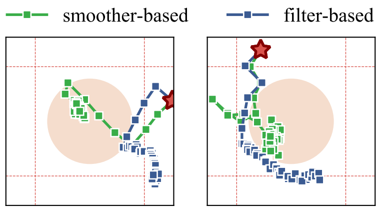

Figure 1: Effect of the conditioning gap on variational world models in the Dark Room environment, where the agent has to reach the circle in the middle. The agent is blind in the center square and can reason about its location only close to the walls. A filter-based variational posterior leads to a degenerate world model which yields policies unable to solve the problem. The smoother-based one recovers the true model and deduces an optimal policy.

Moreover, VSSMs are a natural fit for problems involving imperfect state information, commonly formalised as partially-observable Markov decision processes (POMDPs, Åström (1965)). Here, the agent perceives the world only in parts and has to estimate the system state. Still, the recent literature mostly deals with problems where the state of an agent is only observable through images or noisy measurements. In such settings, a concatenation of observations is often sufficient to transform the partially-observed problem into a fully-observed one, as shown by Srinivas et al. (2020). Studies on problems involving more difficult state estimation tasks such as the "Heaven and Hell" system described by Thrun (1999) are missing. In such "hard" imperfect state information problems the agent has to actively reason about its uncertainty in the state and incorporate it into its decision making. Whether VSSMs are up to the task remains an open question.

VSSMs rely on an approximation to the true posterior of the latent variables given the observed ones. Successful learning of model parameters hinges on minimising the KL-divergence to the approximate posterior from the true one. Part of this divergence is the *conditioning gap*, a suboptimality present when the approximate posterior does not depend on the same conditions as the true posterior (Bayer et al., 2021). This case is quite common in VSSMs. For example, if the approximate posterior over a system state $\mathbf{z}_t$ is predicted by a neural net that only looks at the past and present observations, while the true posterior also depends on the future. This is often done to use the approximate posterior as a state estimator in model-based control.

We hypothesise that this approach is a fallacy: applying VSSMs to optimal control in imperfect state problems requires addressing the conditioning gap. Our contributions are:

- We identify two common instances of the conditioning gap in the context of POMDPs.
- We introduce two new continuous POMDPs, Dark Room and Meta Pendulum, which require more challenging state estimation.
- We demonstrate the negative effect of the conditioning gap in three low-dimensional continuous POMDPs, which we implement in a fully-differentiable fashion. This allows comparing model-based policies with near-optimal policies trained on the real system.

Related Work Using variational inference for state-space models in conjunction with an optimal control objective was first done by Raiko & Tornio (2009). It was not until the advent of amortised inference (Kingma & Welling, 2014; Rezende et al., 2014) however, that variational sequence models scaled to high-dimensional data (Bayer & Osendorfer, 2014; Chung et al., 2015; Fabius & van Amersfoort, 2015; Archer et al., 2015; Krishnan et al., 2015; Fraccaro et al., 2016; 2017; Becker-Ehmck et al., 2019; Karl et al., 2017a; Doerr et al., 2018). Henceforth, many authors have applied VSSMs to decision-making problems (Karl et al., 2017b; Hafner et al., 2019; 2020a;b; Lee et al., 2019; Becker-Ehmck et al., 2020; Buesing et al., 2018; Zhao et al., 2020). POMDPs have been focused on by Igl et al. (2018); Han et al. (2020). We are not aware of any work using a fully-conditioned variational posterior in learning variational models for control.

2 Background and Methods

2.1 Learning Variational State-Space Models

We define state-space models as a latent Markov chain over latent states $\mathbf{z}_{1:T} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$, observations $\mathbf{x}_{1:T} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ and controls $\mathbf{u}_{1:T-1} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{T-1})$:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T} \mid \mathbf{u}_{1:T-1}) = p(\mathbf{z}_1) \prod_{t=1}^{T} p(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{u}_t) p(\mathbf{x}_t \mid \mathbf{z}_t).$$

We refer to $p(\mathbf{z}_1)$, $p(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{u}_t)$ and $p(\mathbf{x}_t \mid \mathbf{z}_t)$ as the initial state, transition and emission distributions. Given appropriate approximation architectures for each we can represent any model through a set of parameters θ. Maximum likelihood learning from data can be done through the gradient-based maximisation of the ELBO w. r. t. both the model parameters θ, as well as a helper distribution $q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}, \mathbf{u}_{1:T-1}) \approx p(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T})$, which is parameterised by the neural network parameters ϕ:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\mathbf{z}_{1:T} \sim q}\left[\log \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T} \mid \mathbf{u}_{1:T-1})}{q(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}, \mathbf{u}_{1:T-1})}\right].$$

Monte Carlo estimates in conjunction with the reparameterisation trick and amortised inference (Kingma & Welling, 2014; Rezende et al., 2014) yield an efficient learning algorithm. Here, the parameters of a distribution given the conditions are e. g. $q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}, \mathbf{u}_{1:T-1}) = \mathcal{N}\left(\mathbf{z}_{1:T} \mid \mu_\phi(\mathbf{x}_{1:T}, \mathbf{u}_{1:T-1}), \sigma^2_\phi(\mathbf{x}_{1:T}, \mathbf{u}_{1:T-1})\right)$, where both μ_ϕ and σ^2_ϕ are neural networks.

2.1.1 The Conditioning Gap

As noted by Bayer et al. (2021), amortised inference gives way to a previously unknown suboptimality. For a thorough explanation we refer the reader to that paper, however we will re-state the most important results here.

If the conditions of the true posterior $p\left(\mathbf{z} \;\middle|\; \tilde{C}\right)$ are partitioned into two disjoint sets $\tilde{C} := C \coprod \bar{C}$, where the amortised variational posterior is only conditioned on C and has no access to $\bar{C}$, then:

1. The optimal amortised variational posterior matches neither $p(\mathbf{z} \mid C)$ nor $p\big(\mathbf{z} \bigm| C, \bar{C}\big)$,
2. The ELBO-optimal generative model p_θ under the optimal amortised variational posterior does not match the maximum-likelihood model,

as long as $p\big(\mathbf{z} \bigm| C, \bar{C}\big) \neq p(\mathbf{z} \mid C)$.

2.2 Learning POMDPs

We define a partially-observable Markov decision process (POMDP) as a state-space model (cf. section 2.1), where the emissions are partitioned into observations $\mathbf{o}_{1:T}$ and instantaneous costs $\mathbf{c}_{1:T}$, i. e. $\mathbf{x}_t := [\mathbf{o}_t, \mathbf{c}_t]$. Solving a POMDP refers to minimising the total cost w. r. t. the control sequence $\mathbf{u}_{1:T}$. For example, an infinite discounted sum $\sum_{t=1}^{\infty} \gamma^t \mathbf{c}_t$ or a finite sum $\sum_{t=1}^{T} \mathbf{c}_t$. In this work, we focus exclusively on the latter.

A trend we observe in most VSSM-based approaches to learning POMDPs is to use approximate posteriors of the form $q(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}) = \prod_t q(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{o}_{t+1})$ (Karl et al., 2017a; Hafner et al., 2019; 2020b; Lee et al., 2019; Zhao et al., 2020), which differ from the true posterior $p(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{o}_{t+1:T}, \mathbf{c}_{1:T})$ by not considering future observations and past and future costs. This is a design choice made for practicality, because the inference network is often used as a state estimator at test time, approximating the filtering distribution $p(\mathbf{z}_t \mid \mathbf{x}_{1:t})$. Due to Section 2.1.1, however, this is not the case. Borrowing notation from Section 2.1.1, the approximate posterior is given access to the conditions $C_t = [\mathbf{z}_t, \mathbf{o}_{t+1}]$ and does not consider the set $\bar{C}_t := [\mathbf{o}_{t+2:T}, \mathbf{c}_{1:T}]$.

We examine the effect of omitting $\mathbf{o}_{t+2:T}$ and $\mathbf{c}_{1:T}$ by treating the condition set of the inference network as a hyper-parameter. Thus our inference networks vary in two ways, either by looking or not looking at future observations $\mathbf{o}_{t+2:T}$ and by looking or not looking at the instantaneous costs $\mathbf{c}_{1:T}$. We work around the need for an explicit state estimator by training policies represented through recurrent networks that internalise the state estimation problem. The inference network is hence discarded during policy optimisation.

Our implementation of VSSMs follows that of Bayer et al. (2021) and we omit details here.

2.3 Policy Optimisation

We consider finite-time stochastic optimal control problems with imperfect state information. The state of the system can only be estimated through the information vector $\mathbf{I}_t = (\mathbf{o}_1, \mathbf{u}_1, \mathbf{o}_2, \mathbf{u}_2, \ldots, \mathbf{o}_t)$, and the optimal control signal is hence a function thereof. We implement a policy $\mathbf{u}_t = \pi_\nu(\mathbf{I}_t)$ with parameters ν using recurrent networks. Given a fixed set of model parameters θ, an optimal policy can be found by minimization of the expected total cost w. r. t. the policy's parameters ν:

$$\mathcal{L}(\nu) = \mathbb{E}_{\substack{\mathbf{o}_t, \mathbf{c}_t \sim p_\theta \\ \mathbf{u}_t = \pi_\nu(\mathbf{I}_t)}} \left[\sum_{t=1}^{T} \mathbf{c}_t \right].$$

Many methods for approximately doing so exist (Bertsekas, 2005). We employ gradient-based optimisation using Monte-Carlo rollouts.

2.4 Model-Based Reinforcement Learning

We use an alternating scheme of data acquisition, system identification (see section 2.2) and policy optimisation (see section 2.3). For the first iteration, we acquire N_{init} episodes of length T using random actions. For each later iteration, N episodes are collected by the current policy. All data gathered so far is used to learn the parameters θ for a fixed number of updates L_θ. The parameters of the policy ν are then optimised on the most recent model for a fixed number of updates L_ν. This approach has been shown to work well in deep reinforcement learning settings (Kaiser et al., 2020).

3 Experiments

We conduct a series of studies that illustrate the suboptimality of variational filtering posteriors with and without access to the instantaneous cost compared to fully-conditioned variational smoothers.

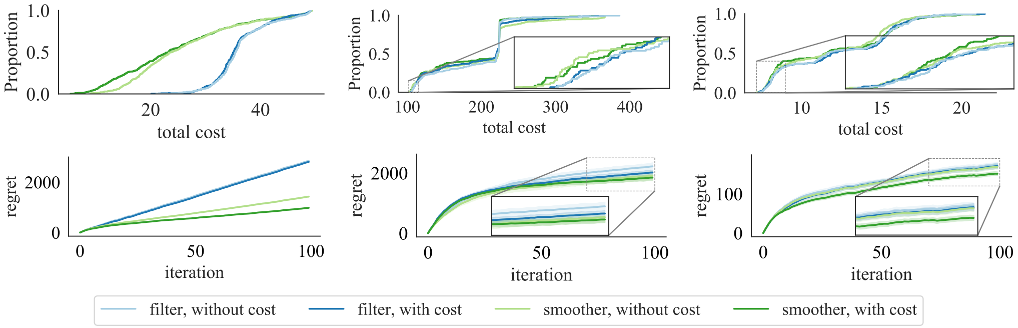

Figure 2: Top row: Cumulative distribution function of the total cost averaged over the final ten episodes. Bottom row: Average regret curves of the best-performing hyper-parameter configurations. Left to right: Dark Room, Mountain Hike and Meta Pendulum. For Dark Room, the average regret is not sublinear on average, yet, for individual smoothing experiments it is (not shown).

While most works in model-based reinforcement learning concentrate on domains such as computer games or simulators of complex dynamics, we explicitly choose settings that are simple in their dynamics, but where inference is challenging and the conditioning gap therefore plays a role.

Further, our environments are fully differentiable w. r. t. the control inputs. This allows us to train near-optimal policies using the true POMDP during policy optimisation for evaluation purposes. More specifically, this allows us to calculate the *regret* of each method, which is the total cost some policy achieves minus the total cost of executing an optimal policy.

We first introduce the *Dark Room* environment. An agent is placed in an empty unit square that it can traverse up to its surrounding walls. Its state is its location only, i. e. $\mathbf{z}_t \in \mathbb{R}^2$. The agent can sense the walls only with its distance sensors that have a maximum range of 0.2 facing north, east, south and west ($\mathbf{x}_t \in \mathbb{R}^4$); out of range, the sensor yields the maximum value. The cost is 1 everywhere, except at the goal, where it is 0. The goal is a circle of radius 0.5 in the center of the room. The agent's controls $\mathbf{u}_t \in [-0.2, 0.2]^2$ are added to its current location, after which it is projected back into the unit square: $\mathbf{z}_{t+1} = \rho\left(\mathbf{z}_t + \mathbf{u}_t\right)$. For optimising the near-optimal policy, we use straight-through estimation of ρ and the cost.

The *Mountain Hike* environment has been introduced by Igl et al. (2018) and features a two-dimensional linear Gaussian system with a non-linear cost function. We use the *medium* variant that has a moderate amount of observation and transition noise.

We further introduce a version of the classic pendulum swingup task where the mass of the pendulum is randomised at the beginning of each episode, *Meta Pendulum*. It is distributed uniformly in the range $[0.8, 1.2]$.

For each experiment, we conducted a study of 300 different hyper-parameter configurations per each conditioning set. We used $N_{\text{init}} = 25$ and $N = 5$ with varying lengths of $T = 50$ for Dark Room, $T = 75$ for Mountain Hike and $T = 50$ for Meta Pendulum. The algorithm did a total of 100 iterations in each case.

In all studies, the smoothing posterior with access to the instantaneous costs achieves the smallest regret w. r. t. the near-optimal policy on average. The picture is clearest in the dark room environment, where the agent has to actively localise itself before going to the goal. In the other environments, the uncertainty about the state diminishes even if the agent does not perform a strategy of active uncertainty reduction. Consequently, the smoothing posterior variant is not as dominant, but still clearly superior. For more details, see the illustrations in fig. 2.

4 Conclusion

We have presented simple settings where under-conditioning inference networks leads to inferior model-based reinforcement learning. In all cases we see that world models learned from smoothing posteriors lead to superior performance. Our findings suggest that scaling up VSSM-based reinforcement learning to hard imperfect state information problems will require careful consideration of the true posterior's condition set.

References

Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*, 2015.

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Justin Bayer and Christian Osendorfer. Learning stochastic recurrent networks. *CoRR*, abs/1411.7610, 2014. URL http://arxiv.org/abs/1411.7610.

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models, 2021.

Philip Becker-Ehmck, Jan Peters, and Patrick Van Der Smagt. Switching linear dynamics for variational Bayes filtering. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 553–562. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/becker-ehmck19a.html.

Philip Becker-Ehmck, Maximilian Karl, Jan Peters, and Patrick van der Smagt. Learning to fly via deep model-based reinforcement learning. *CoRR*, abs/2003.08876, 2020. URL https://arxiv.org/abs/2003.08876.

Dimitri P. Bertsekas. *Dynamic programming and optimal control, 3rd Edition*. Athena Scientific, 2005. ISBN 1886529264. URL http://www.worldcat.org/oclc/314894080.

Lars Buesing, Theophane Weber, Sebastien Racaniere, S. M. Ali Eslami, Danilo Rezende, David P. Reichert, Fabio Viola, Frederic Besse, Karol Gregor, Demis Hassabis, and Daan Wierstra. Learning and querying fast generative models for reinforcement learning, 2018.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2980–2988, 2015. URL http://papers.nips.cc/paper/5653-a-recurrent-latent-variable-model-for-sequential-data.

Andreas Doerr, Christian Daniel, Martin Schiegg, Nguyen-Tuong Duy, Stefan Schaal, Marc Toussaint, and Trimpe Sebastian. Probabilistic recurrent state-space models. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1280–1289, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/doerr18a.html.

Otto Fabius and Joost R. van Amersfoort. Variational recurrent auto-encoders, 2015.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2199–2207, 2016. URL http://papers.nips.cc/paper/6039-sequential-neural-models-with-stochastic-layers.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 2555–2565, 2019. URL `http://proceedings.mlr.press/v97/hafner19a.html`.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a. URL `https://openreview.net/forum?id=S1lOTC4tDS`.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *CoRR*, abs/2010.02193, 2020b. URL `https://arxiv.org/abs/2010.02193`.

Dongqi Han, Kenji Doya, and Jun Tani. Variational recurrent models for solving partially observable control tasks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=r1lL4a4tDB`.

Maximilian Igl, Luisa M. Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2122–2131. PMLR, 2018. URL `http://proceedings.mlr.press/v80/igl18a.html`.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=S1xCPJHtDB`.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017a. URL `https://openreview.net/forum?id=HyTqHL5xg`.

Maximilian Karl, Maximilian Soelch, Philip Becker-Ehmck, Djalel Benbouzid, Patrick van der Smagt, and Justin Bayer. Unsupervised real-time control through variational empowerment, 2017b.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *CoRR*, abs/1511.05121, 2015. URL `http://arxiv.org/abs/1511.05121`.

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

Tapani Raiko and Matti Tornio. Variational bayesian learning of nonlinear hidden state-space models for model predictive control. *Neurocomputing*, 72(16-18):3704–3712, 2009. doi: 10.1016/j.neucom.2009.06.009. URL `https://doi.org/10.1016/j.neucom.2009.06.009`.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1278–1286, 2014. URL `http://proceedings.mlr.press/v32/rezende14.html`.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Sebastian Thrun. Monte carlo pomdps. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 10641070, Cambridge, MA, USA, 1999. MIT Press.

Tony Z. Zhao, Anusha Nagabandi, Kate Rakelly, Chelsea Finn, and Sergey Levine. MELD: meta-reinforcement learning from images via latent state models. *CoRR*, abs/2010.13957, 2020. URL `https://arxiv.org/abs/2010.13957`.