Disentangling Independently Controllable Factors in Reinforcement Learning

Rafael Rodriguez-Sanchez
Brown University
rrs@brown.edu

Cameron Allen
University of California, Berkeley

George Konidaris
Brown University

Abstract

Leveraging the factored structure of the world leads to efficient algorithms for reinforcement learning that allows agents to abstract states, explore the world and discover skills. However, all these methods require access to a factored representation a priori. Typically, these representations are hand-specified and it remains an open problem how this representation can be learned directly from data. Therefore, applying these methods to problems with high-dimensional observations is not yet practical. In this work, we take a step toward factored representation in reinforcement learning. We introduce Action Controllable Factorization (ACF), a contrastive learning approach that focuses on disentangling *independently controllable* latent variables. These are variables the agent can affect directly without affecting others. The core idea of ACF is to leverage action sparsity: actions typically affect only a subset of variables, while the rest evolve under the environment's dynamics, yielding informative data for contrastive training. ACF recovers the ground-truth controllable factors directly from pixel observations on three benchmarks with known factored structure—TAXI and MINIGRID-DOORKEY—consistently outperforming baseline disentanglement algorithms.

Keywords: contrastive learning, representation learning, disentanglement

1 Motivation

Classical work in factored RL shows that, if the underlying Markov decision process (MDP) can be decomposed into state-variable factors with sparse dependencies, one can achieve exponential gains in both model learning and planning [Boutilier et al., 1995, Guestrin et al., 2003]. Indeed, factored variants of PAC-RL algorithms such as factored E^3 [Kearns and Koller, 1999] and Factored RMax [Guestrin et al., 2002, Brafman and Tennenholtz, 2002], provably exploit these structures for faster convergence, and subsequent methods even learn the dependency graph online [Strehl et al., 2007, Diuk et al., 2009]. More recently, factored representations have proven useful for world modeling [Wang et al., 2022, Pitis et al., 2020, 2022], exploration [Wang et al., 2023, Seitzer et al., 2021], and skill discovery [Vigorito and Barto, 2010, Wang et al., 2024, Chuck et al., 2024, 2025]. Crucially, all these gains depend on having access to a hand-specified factored representation. In this work, we introduce ACF, a contrastive learning algorithm that address this representation gap by leveraging an agent's actions to disentangle independently controllable factors directly from pixels.

2 ACF: Action Controllable Factorization

Setting We assume that the agent does not have access to the ground truth factored state space S. Instead, it gets high-dimensional observations that are generated by an unknown decoder $o : S \to X \subseteq \mathbb{R}^{d_x}$. Hence, we are concerned with learning from the observed samples of $T(x' \mid x, a)$ an encoder $f_\phi : X \to Z$, where Z factorizes as $Z = Z_1 \times \cdots \times Z_K$, that identifies the underlying factors. Moreover, in many problems, the agent's actions have sparse effects on the environment: just a few factors are controlled, while others just follow their natural transition, unaffected by the agent. To help the agent understand its environment, we assume that the agent has a *special action* a_0 that corresponds to a *no-op* (or observe) action that allows the agent to observe the natural evolution of the environment without intervening.

Transition Dynamics Let $\Psi(s, a) = S \times A \to \mathcal{P}([1, 2, \ldots, K])$ be the set of variables affected by action a in state s. We assume the transition dynamics factorize as $T(s' \mid s, a) = \prod_{i \in \Psi(s,a)} T(s'_i \mid s, a) \prod_{j \notin \Psi(s,a)} T(s'_j \mid s, a_0)$, where $T(s'_i \mid s, a_0)$ represents the natural (or observational) dynamics.

Algorithm We parameterize the encoder by $f_\phi(x) \mapsto z$, with parameters ϕ, and, more importantly, we parameterize the transition function as the sum of energy functions (unnormalized probability densities) such that $T(z' \mid z, a) \propto \exp\left(\sum_{i=1}^{K} E_\theta(z'_i, a, z)\right)$, with $i \in [K]$ and parameters θ. This sum of energies reflects the factorized structure where each energy represent the transition dynamics of latent variable z_i.

In order to estimate these energy functions from data and learn a Markov representation suitable for RL [Allen et al., 2021]. Hence, we estimate the inverse dynamics model and forward dynamics model by training a multiclass classifier and InfoNCE [Oord et al., 2018], respectively.

$$\mathcal{L}_{\text{inv}}(\phi, \theta) = -\log I^\pi(a \mid z, z') = -\log \frac{\exp\left(\sum_i E_\theta(z'_i, a, z)\right) \pi(a \mid z)}{\sum_{a' \in A} \exp\left(\sum_i E_\theta(z'_i, a', z)\right) \pi(a' \mid z)}; \tag{1}$$

$$\mathcal{L}_{\text{fwd}}(\phi, \theta) = -\log \frac{\exp\left(\sum_i E_\theta(z'_i, a, z)\right)}{\sum_{z^j \in B} \exp\left(\sum_i E_\theta(z^j_i, a, z)\right)}. \tag{2}$$

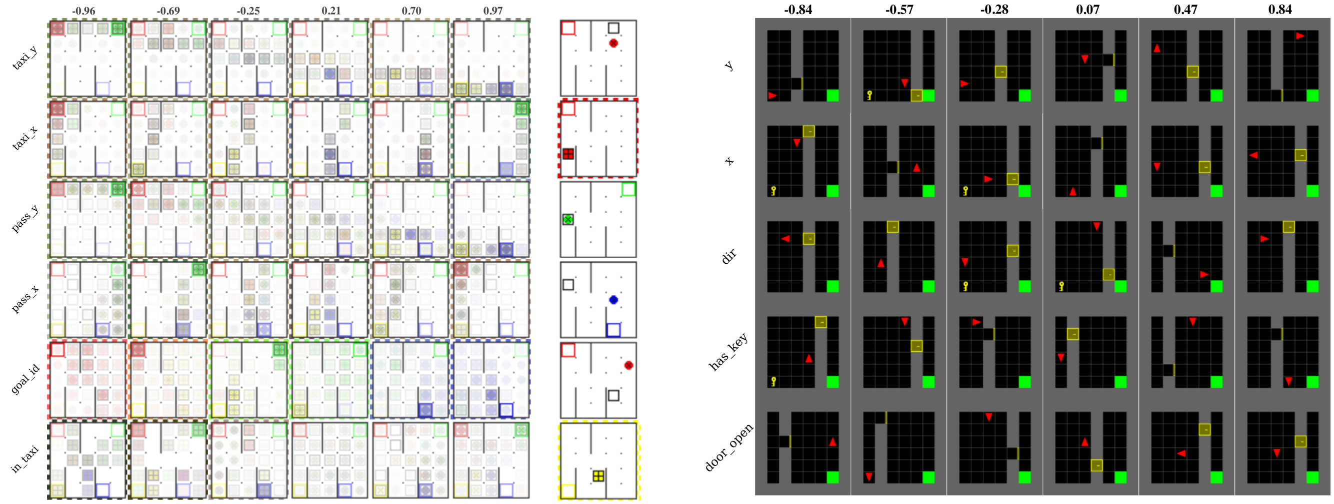
However, these two alone do not ensure that the representation will align with the controllable factors.

Factorizing the Controllable Variables We formalize our intuition and exploit the sparsity of the actions' effects to learn a latent representation Z that identifies the controllable factors. The core idea is to contrast the effect of an action, the distribution $T(x' \mid x, a)$, against the natural dynamics $T(x' \mid x, a_0)$, where a_0 is the no-op action. We leverage the fact that, $\log r_a(x', x) = \log \frac{T(x'|x,a)}{T(x'|x,a_0)} = \log \frac{T(s'_j|s,a)}{T(s'_j|s,a_0)} = \log r_a(s', s)$, where s_j is the factor affected by a when executed in s. Therefore, this ratio is invariant to the representation and provides a signal to *separate* a controlled factor from the rest. In practice, we estimate these ratios from observed transitions using Noise Contrastive Estimation (NCE; Gutmann and Hyvärinen [2010], Hyvärinen et al. [2019]) and leveraging our energy parameterization: $\log r_a(z', z) := \log r_a(f_\phi(x'), f_\phi(x)) := \sum_i E_\theta(z'_i, a, z) - E_\theta(z'_i, a_0, z)$.

Therefore, we train our energy functions to match the observed ratios by training $|A| - 1$ binary classifiers computed by $\sigma(\log r_a(z', z))$ where σ is the sigmoid function. We use the transitions of other actions as negative samples and minimize the following binary cross-entropy loss:

$$\mathcal{L}_r(\theta, \phi) = \sum_{a' \in A} [a' = a] \log \sigma\left(\log r_a + \zeta_a\right) + [a' \neq a] \log\left(1 - \sigma\left(\log r_a + \zeta_a\right)\right),$$

with $\zeta_a := \log \frac{\pi(a|z)}{\pi(a_0|z)}$ and $[\cdot]$ is indicator functions that is 1 when the condition holds, and ζ_a are correction weights to account for the policy used to collect the data. In practice, we estimate the policy from the dataset and use the estimate



(a) Taxi Environment (b) Minigrid DoorKey Environment

Figure 1: Latent traversals: By traversing the values of each latent variable we can observe the disentanglement effect on the observations.

to compute the loss. The core assumption of ACF is that variables are independently controllable, that is, for every state variable s_i, there exists a context $s \in S$ and action $a \in A$, where the action effect is sufficiently different from the natural dynamics of the variable (a_0 effect). In the following section, we will show empirically cases where this might not hold but our algorithm still manages to identify some of these variables.

3 Results and Discussion

We empirically evaluate ACF in classical RL test domains. We consider a visual variation of the classical Taxi domain [Dietterich, 2000] and visual Minigrid DoorKey [Chevalier-Boisvert et al., 2023][1]. We compare ACF with GCL (Generalized Contrastive Learning; Hyvärinen et al. [2019]) that can be seen as a vanilla contrastive-based disentanglement algorithm, and DMS (Disentanglement via Mechanism Sparsity; Lachapelle et al. [2022]), a VAE-based [Kingma and Welling, 2014] method that explicitly maximizes sparsity in state dependencies and action effects to drive disentanglement, and MSA (Markov State Abstractions; Allen et al. [2021]), a contrastive-based algorithm that leverages both forward and inverse dynamics to learn Markovian representations but does not explicitly optimize for disentanglement.

To measure disentanglement, we consider test datasets of pairs of $\{(s^i, z^i)\}_i$ where s is the ground truth representation and z is the corresponding learned latent representation. Then, we fit factor-wise regressors (parameterized by feed-forward networks), $h_{ij}(z_i) \mapsto s_j$. The performance of h_{ij} is limited by the amount of information z_i contains about s_j, therefore we measure the quality of the learned regressor using the coefficient of determination R^2. Table 1 show the results of the mean diagonal of the R^2 matrices and the maximum off-diagonal value. Moreover, Figure 1 show qualitative latent factor traversals.

Method	**Doorkey**		**Taxi**	
	Mean diagonal ↑	Max off-diagonal ↓	Mean diagonal ↑	Max off-diagonal ↓
acf	**0.565±0.042**	0.250±0.021	**0.698±0.084**	0.251±0.044
dms	0.301±0.029	0.250±0.028	0.309±0.058	**0.153±0.027**
gcl	0.459±0.067	0.227±0.076	0.606±0.054	0.281±0.034
markov	0.169±0.116	**0.115±0.085**	0.373±0.049	0.178±0.058

Table 1: Qualitative Results: Perfect disentanglement would be 1 for the mean diagonal value and minimal off-diagonal.

While these domains are simple from an RL perspective, we can see that they are challenging for the factorization task. Moreover, we see that ACF outperforms all the baselines disentangling the intended factors. However, ACF works for disentangling one-step controllable factors. Future work must focus in improving factor discovery by considering longer-term effects and control. Moreover, not all factors are always controllable but they might be relevant for a task, hence, including factors affecting the reward signal is also an important direction.

[1] We use Minigrid JAX [Bradbury et al., 2018] re-implementation [Pignatelli et al., 2024].

References

Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning Markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8229–8241, 2021.

Craig Boutilier, Richard Dearden, Moisés Goldszmidt, et al. Exploiting structure in policy construction. In *IJCAI*, volume 14, pages 1104–1113, 1995.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Caleb Chuck, Kevin Black, Aditya Arjun, Yuke Zhu, and Scott Niekum. Granger causal interaction skill chains. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=iA2KQyoun1`.

Caleb Chuck, Fan Feng, Carl Qi, Chang Shi, Siddhant Agarwal, Amy Zhang, and Scott Niekum. Null counterfactual factor interactions for goal-conditioned reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Carlos Diuk, Lihong Li, and Bethany R Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 249–256, 2009.

Carlos Guestrin, Relu Patrascu, and Dale Schuurmans. Algorithm-directed exploration for model-based reinforcement learning in factored mdps. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 235–242, 2002.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, pages 740–747, 1999.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Eduardo Pignatelli, Jarek Liesen, Robert Tjarko Lange, Chris Lu, Pablo Samuel Castro, and Laura Toni. Navix: Scaling minigrid environments with jax. *arXiv preprint arXiv:2407.19396*, 2024.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. Mocoda: Model-based counterfactual data augmentation. *Advances in Neural Information Processing Systems*, 35:18143–18156, 2022.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021.

Alexander L. Strehl, Carlos Diuk, and Michael L. Littman. Efficient structure learning in factored-state mdps. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, page 645–650. AAAI Press, 2007. ISBN 9781577353232.

Christopher M. Vigorito and Andrew G. Barto. Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Transactions on Autonomous Mental Development*, 2(2):132–143, 2010. doi: 10.1109/TAMD.2010.2050205.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*, 2022.

Zizhao Wang, Jiaheng Hu, Peter Stone, and Roberto Martín-Martín. Elden: Exploration via local dependencies. *Advances in Neural Information Processing Systems*, 36:15456–15474, 2023.

Zizhao Wang, Jiaheng Hu, Caleb Chuck, Stephen Chen, Roberto Martín-Martín, Amy Zhang, Scott Niekum, and Peter Stone. Skild: Unsupervised skill discovery guided by factor interactions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.