Pareto Value-Conditioned Networks for MORL in Stochastic Environments

Liam P. H. Mertens[1], Ann Nowé[1], and Diederik M. Roijers[1,2]

[1] AI Lab, Vrije Universiteit Brussel (VUB), Belgium
[2] Innovation Dept., City of Amsterdam, The Netherlands
[liam.phi.h.mertens,ann.nowe,diederik.roijers]@vub.be

Keywords: Multiple Objectives · Reinforcement Learning

Many key sequential decision problems, such as climate change mitigation [1] or epidemic mitigation [6], have multiple conflicting objectives. Multi-objective reinforcement learning (MORL) algorithms can handle such problems [4].

When the user utility is unknown and can be non-linear, as is often the case with human decision makers, a Pareto front is often the desired solution concept. MORL algorithms that compute the Pareto front do exist [5, 9], but are often tailored towards and/or tested deterministic environments. This is undoubtedly in part due to the specific difficulties that the combination of multiple objectives and stochastic environments entail. In contrast to the single-objective case, it becomes highly non-trivial to execute the a value-matching policy from a value function. In fact, this can involve solving a combinatorial optimization problem at every timestep during execution [7]. However, many environments just are intrinsically stochastic in nature [3, 8].

In this paper, we propose *Pareto Value Conditioned Networks (PVCN)*, a new method that builds on Pareto Conditioned Networks (PCN) [5] and Pareto-optimal policy following (POPF) networks [7]. PVCN effectively discovers Pareto-optimal policies in stochastic environments with accurate value estimates.

1 The Pareto Value Conditioned Networks Algorithm

We propose Pareto Value Conditioned Networks (PVCN). PVCN maintains two networks: a policy network $\pi_\theta(s|\mathbf{V}, h)$, that outputs a distribution over actions, and is conditioned on a remaining time horizon h, and a *desired value* $\mathbf{V}$ that the agent aims to achieve from state s and horizon h; and a POPF-network $\phi_\kappa(\mathbf{N}, s, a, s')$ that helps us track the next desired value.

To see why a POPF-network is necessary, we note that we explicitly assume stochastic environments. This means that once we take an action a from state s, the subsequent states and their associated achievable values are not always the same. For example, assume that an agent desires a 2-objective value of $(5, 5)$ and performs an action a in s, leading to a $(0, 0)$ reward. There are two subsequent states s' and s'' where the agent can end up with equal probability. From neither of these states $(5, 5)$ is an achievable value, but $(10, 0)$ is achievable from s', and $(0, 10)$ from s''. By taking the average of these value vectors, we can still achieve a value of $(5, 5)$ by taking a in s, as long as the agent knows

to either chase $(0, 10)$ in s'' and $(10, 0)$ in s'. This is what the POPF network, $\phi_\kappa(\mathbf{N}, s, a, s')$, does; it is conditioned on the transition, and $\mathbf{N}$, which is the previous desired value minus the immediate reward associated with the transition.

During training PVCN collects samples by selecting a desired value $\hat{\mathbf{V}}$ for the initial state s (and appropriate time horizon h), and then repeatedly selects an action a using the policy network π_θ, observing the transition $(s, a, h, \mathbf{r}, s')$, and then determines the next desired value $\hat{\mathbf{V}}'$. An experience replay buffer is kept with $(s, a, h, \mathbf{r}, s')$ in sequences making it easier to extract Monte-Carlo estimates of the values during training. When selecting the initial desired value vector during training, we randomly pick a vector from the approximated Pareto coverage set (PCS) for the initial state and add a small bonus vector.

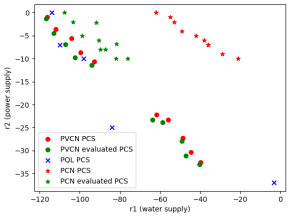

Fig. 1. The value estimates, evaluated policy values, and buffer returns of PVCN in discrete watershed, with the value estimates, evaluated policy values of PCN. And value estimates of PQL [9].

Targets for training π_θ are computed by selecting a (s, a, h)-tuple and sampling batches $(s, a, h, \mathbf{r}, s')$ with identical (s, a, h), along with their returns. Then the Monte-Carlo value estimate $\hat{\mathbf{V}}$ is used as the desired value. The loss function is NLL with an added entropy term. For the POPF networks, the $\mathbf{N}$ values are computed using $\hat{\mathbf{V}}$, and the subsequent desired values are the Monte-Carlo estimates of the subsequent values for s' from the same batch of $(s, a, h, \mathbf{r}, s')$-tuples. The POPF network's loss function is MSE.

2 Preliminary Experiments & Discussion

To test PVCN on stochastic environments, we use a discretised version of the Watershed problem by [2]. As we can see in Figure 1, PVCN learns to accurately estimate the values of the policies and learns a broad PCS. PCN [5] on the hand cannot handle the stochasticity of the environment well, and is overly optimistic with its value estimates. Pareto Q-learning learns a slightly worse PCS, but its policies cannot be executed faithfully. In addition, we compared PVCN and PCN on Deep Sea Treasure (DST), a deterministic environment with a known Pareto coverage set. For DST, PVCN and PCN both learn the entire PCS, but PVCN is slightly slower as expected, as it is not tailored to deterministic environments.

Conclusion and Discussion In this paper, we proposed PCVN, a new MORL algorithm for learning Pareto-optimal policies stochastic environments. We have shown that we can learn the entire Pareto coverage set in Deep Sea Treasure, which is deterministic. But more importantly, we have shown that PCVN can learn Pareto-optimal policies in highly stochastic environments using Random MOMDPs. In future work, we aim to expand this to high-dimensional state spaces, and perform more thorough experimentation.

Acknowledgements This research was in part supported by EU's Horizon Europe Research and Innovation Programme, under Grant Agreement number 101120406 (PEER).

References

1. Biswas, P., Osika, Z., Tamassia, I., Whorra, A., Zatarain-Salazar, J., Kwakkel, J., Oliehoek, F.A., Murukannaiah, P.K.: Exploring equity of climate policies using multi-agent multi-objective reinforcement learning. arXiv preprint arXiv:2505.01115 (2025)
2. Castelletti, A., Pianosi, F., Restelli, M.: A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. Water Resources Research **49**(6), 3476–3486 (2013). https://doi.org/https://doi.org/10.1002/wrcr.20295, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/wrcr.20295
3. Felten, F., Alegre, L.N., Nowe, A., Bazzan, A., Talbi, E.G., Danoy, G., C da Silva, B.: A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. Advances in Neural Information Processing Systems **36**, 23671–23700 (2023)
4. Hayes, C.F., Radulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A.A., Mannion, P., Nowé, A., de Oliveira Ramos, G., Restelli, M., Vamplew, P., Roijers, D.M.: A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems **36**, 26 (2022)
5. Reymond, M., Bargiacchi, E., Nowé, A.: Pareto conditioned networks. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS'22). p. 1110–1118 (2022)
6. Reymond, M., Hayes, C.F., Willem, L., Rădulescu, R., Abrams, S., Roijers, D.M., Howley, E., Mannion, P., Hens, N., Nowé, A., et al.: Exploring the Pareto front of multi-objective covid-19 mitigation policies using reinforcement learning. Expert Systems with Applications **249**, 123686 (2024)
7. Roijers, D.M., Röpke, W., Nowe, A., Radulescu, R.: On following Pareto-optimal policies in multi-objective planning and reinforcement learning. In: Multi-Objective Decision Making Workshop 2021 (2021)
8. Teoh, J., Varakantham, P., Vamplew, P.: On generalization across environments in multi-objective reinforcement learning. arXiv preprint arXiv:2503.00799 (2025)
9. Van Moffaert, K., Nowé, A.: Multi-objective reinforcement learning using sets of Pareto dominating policies. The Journal of Machine Learning Research **15**(1), 3483–3512 (2014)