Diversity Progress for Goal Selection in Discriminability-Motivated RL

Erik M. Lintunen, Nadia M. Ady*, Christian Guckelsberger*
Department of Computer Science, Aalto University, Finland
{erik.lintunen,nadia.ady,christian.guckelsberger}@aalto.fi
*Joint senior authors, order determined alphabetically

Abstract

Non-uniform goal selection has the potential to improve the reinforcement learning (RL) of skills over uniform-random selection. In this paper, we introduce a method for learning a goal-selection policy in intrinsically-motivated goal-conditioned RL: "Diversity Progress" (DP). The learner forms a curriculum based on observed improvement in discriminability over its set of goals. Our proposed method is applicable to the class of discriminability-motivated agents, where the intrinsic reward is computed as a function of the agent's certainty of following the true goal being pursued. This reward can motivate the agent to learn a set of diverse skills without extrinsic rewards. We demonstrate empirically that a DP-motivated agent can learn a set of distinguishable skills faster than previous approaches, and do so without suffering from a collapse of the goal distribution—a known issue with some prior approaches. We end with plans to take this proof-of-concept forward.

1 Introduction

Intrinsically-motivated learning has been studied extensively in the reinforcement learning (RL) literature (see recent reviews by Colas et al., 2022; Aubret et al., 2023; Lidayan et al., 2024). Here, we focus on intrinsically-motivated skill acquisition, often referred to as competence-based intrinsic motivations (CB-IMs), an area of control problems requiring multiple skills. How to learn a diverse set of skills is a key subproblem of CB-IMs (Colas et al., 2022, p. 1161). In answer, one class of CB-IMs uses rewards computed as functions of the agent's certainty of following the goal it chose to pursue, that is, the goal's discriminability. It is postulated in existing work that the more discriminable a set of goals is, the more diverse we expect the skills to be in terms of observed behaviour (see Section 2.1). Many such discriminator-based models select goals uniformly during training; yet, learning a distribution over goals has the potential to speed up learning (see Section 3).

Our contribution is threefold. (1) We present Diversity Progress (DP) (Section 4), a method for learning a goal-selection policy prioritising goals based on the observed learning progress over a set of goals. (2) We complement the formalism with empirical findings (Section 5), suggesting that a DP-motivated agent can improve existing discriminator-based CB-IMs by speeding up the learning of a diverse set of skills. (3) We detail plans for taking this proof-of-concept forward (Section 6).

2 Background

2.1 Discriminability-motivated RL

Our method falls within the framework of goal-conditioned reinforcement learning (GCRL), in which a *skill* is a policy paired with a goal, which can be identified with a goal-conditioned reward function (see Appendix A). We focus on a class of intrinsic rewards used to maximise behavioural

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

mutual information (BMI) (termed by Hansen et al., 2020, p. 2; e.g., Gregor et al., 2016; Warde-Farley et al., 2019; Eysenbach et al., 2019; Baumli et al., 2021; Laskin et al., 2022), though our method could be used with any discriminator-based CB-IM. These rewards approximate the mutual information between the goal-defining variable, $\mathbf{g}$, and some function of the trajectory drawn from the corresponding skill, $f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}})$. Typically, f maps a trajectory to a single state, but some formulations map to multiple states (e.g., Gregor et al., 2016, use the initial and final states). Following Hansen et al. (2020, p. 4), the BMI objective can be defined as the maximisation of:

$$I(\mathbf{g}; f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}})) := H(\mathbf{g}) - H(\mathbf{g} \mid f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}})), \tag{1}$$

where H represents Shannon entropy. However, mutual information can be notoriously hard to compute (Hjelm et al., 2019, p. 1), and a popular approach is to maximise a variational lower bound:

$$\tilde{I}(\mathbf{g}; f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}})) \geq H(\mathbf{g}) - \mathbb{E}_{\mathbf{g}\sim p(\mathbf{g}), \boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}}\sim\pi(\mathbf{g})} \left[\log q(\mathbf{g} \mid f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}}))\right], \tag{2}$$

where q is an arbitrary variational distribution (Barber and Agakov, 2003, p. 2). Given $f(\boldsymbol{\mathcal{T}}_{\pi_{\mathbf{g}}})$, q defines a probability distribution over goals. That is, a probability model, q, predicts, for each skill, the probability that the skill has induced the observations—thus the model is canonically known as a *discriminator*—and is typically used to compute the reward. Successfully discriminating skills in the observation space requires the agent to observe distinct regions of the state space, encouraging the agent to learn a set of diverse behaviours. If a goal is not discriminable based on observations, two or more skills are producing overlapping behaviours (and therefore, the skills lack diversity); conversely, if a goal is discriminable, then the corresponding skill is inducing trajectories unique to that skill.

2.2 Learning Progress

Learning Progress (LP) (e.g., Oudeyer et al., 2007; Schmidhuber, 2010) is designed "to select goals that are of learning-optimal difficulty with respect to [the agent's] current capabilities" (Lintunen et al., 2024, p. 39). Oudeyer et al. (2007, pp. 270–271) formulate LP as an intrinsic reward measuring how much the agent has improved in some prediction over a window of time. For any decision that affects learning, the LP at time $t+1$ can be computed for each option, indexed by n:

$$\mathrm{LP}_n(t+1) := \overline{e}_n(t+1-\tau) - \overline{e}_n(t+1), \tag{3}$$

given an offset, τ, of the comparison, in time steps. A smoothing hyperparameter η adds robustness to stochastic fluctuations, so the average error $\overline{e}_n$ is defined as the mean over the smoothing window:

$$\overline{e}_n(t+1-\tau) := \frac{1}{\eta+1} \sum_{i=0}^{\eta} e_n(t+1-\tau-i), \tag{4}$$

where $e_n(t+1)$ is the difference between some prediction and the true observation made due to choosing option n. Originally, predictions were of state observations (Oudeyer et al., 2007, p. 271). LP has since been extended to measures of competence beyond prediction error (Colas et al., 2019).

3 Related work

BMI-maximising CB-IMs In VIC (Gregor et al., 2016), one of the earliest BMI-related approaches, the agent learns a categorical distribution over goals, $\mathbf{g} \sim p(\mathbf{g})$—like our method. VIC reinforces $p(\mathbf{g})$ to maximise the same BMI-based intrinsic reward as is used for learning the skills (p. 4). However, Eysenbach et al. (2019, p. 6) showed that in VIC the probability mass collapses to a handful of skills. In response, Eysenbach et al. fixed $p(\mathbf{g})$ as uniform in DIAYN, so all skills receive, in expectation, equal training signal. A side effect is inefficient use of training samples: the curriculum is not optimised for improving discriminability. DP aims to train more efficiently, via a learned goal-selection policy that could be used with any BMI-maximising CB-IMs. Laskin et al. (2022, Table 2, p. 10) list several discriminator-based approaches, noting key differences in reward formulations.

LP-based CB-IMs LP has been used to motivate goal selection in a variety of ways (Colas et al., 2022, p. 1185). Learning $p(\mathbf{g})$ can significantly improve the speed of learning a set of skills (Stout and Barto, 2010, pp. 260–261). Our method formulates goal selection as a bandit problem rewarded using LP, much like the approach used by CURIOUS (Colas et al., 2019), one such LP-based CB-IM. This formulation is in contrast with VIC, which sets an RL problem, conditioning its goal-selection policy

on a state. However, while DP selects between single goals in the full sensory space, CURIOUS supports the provision of predefined subspaces, which the authors call modules. The CURIOUS agent learns a distribution over the modules, learning to prioritise subspaces of goals it is increasingly or decreasingly successful in reaching (success is determined by some binary function).

4 Diversity Progress

Our method motivates the agent to prioritise goals that provide most progress in the learning of the discriminator, over all skills. That is, if pursuing one goal improves the discriminability of multiple skills, it is reflected positively in the goal-selection probabilities. We call this Diversity Progress.

At a given time, $t+1$, the agent's prediction error for a given goal, $\mathbf{g} \sim p(\mathbf{g})$, is defined by:

$$e_{\mathbf{g}}(t+1) := \begin{cases} 1 - q(\mathbf{g} \mid \mathbf{s}_{t+1}), & \text{if } \mathbf{g} \text{ corresponds to the skill being followed} \\ q(\mathbf{g} \mid \mathbf{s}_{t+1}), & \text{otherwise,} \end{cases} \tag{5}$$

where $e_{\mathbf{g}}(t+1)$ is an element of $\mathbf{e}(t+1) \in \mathbb{R}^{|\mathcal{G}|}$, a vector composed of the error, at $t+1$, for each goal, and q is a discriminator. Then, the errors over an epoch of fixed time length T form a $T \times |\mathcal{G}|$ matrix. Given hyperparameters *offset*, τ, and *smoothing*, η, (see Section 2.2), we use Equation 4 to compute the most recent average errors, $\overline{\mathbf{e}}(t+1)$, and the average errors from τ time steps earlier, $\overline{\mathbf{e}}(t+1-\tau)$. Following Equation 3, we compute the LP for each goal, with goal $\mathbf{g}$ corresponding to option n. Then, the LP values are averaged over goals to compute the Diversity Progress:

$$\overline{DP}(t+1) := \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} \overline{\mathbf{e}}_{\mathbf{g}}(t+1-\tau) - \overline{\mathbf{e}}_{\mathbf{g}}(t+1). \tag{6}$$

In this way, the mean progress *over all goals* is attributed to the goal being pursued, since that goal determines the actions taken. For each goal, $\mathbf{g}$, $\overline{DP}$ is updated after executing the corresponding skill. The goal-selection probabilities, $p(\mathbf{g})$, are computed via softmax. For the algorithm see Appendix B.

5 Experiments

For details of implementation and evaluation see Appendices C and D, respectively.

RQ1. Does the probability mass of the goal distribution collapse?

To learn a diverse set of skills, it is material to train a substantial number of skills. Eysenbach et al. (2019, p. 6) showed that, with VIC's goal-selection method, the effective number of skills (see Appendix D.2) collapses to a handful. The method by Eysenbach et al. (2019), DIAYN, therefore selects goals uniformly. Does DP avoid a collapse like VIC's? Figure 1 shows the effective number of skills over training time for all three methods in three environments (see Appendix D.1). With the fixed $p(\mathbf{g})$ in DIAYN, its effective number of skills is constant—equal to $|\mathcal{G}|$. DP does not collapse, and, additionally, we have control over the effective number of skills via the softmax temperature.

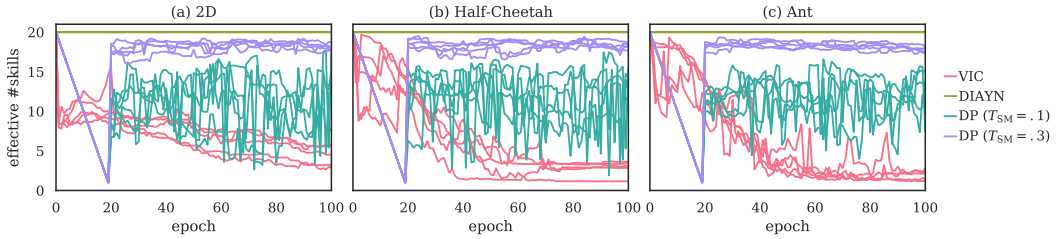

Figure 1: **The effective number of skills** over training time in three environments. The agent is learning 20 skills. We compare VIC, DIAYN, and DP with two different softmax temperatures (0.1 and 0.3) determining how greedy the policy is. The linear decline of the effective number of skills for epochs up to the number of skills is due to DP's initialisation, that is, randomly selecting goals without replacement (see Algorithm 1, ll. 6–10). Results from five random seeds; each line is a seed.

RQ2. What are the effects of DP on the dynamics of goal selection?

Figure 2 shows DP values, goal-selection probabilities, and cumulative frequencies (counts of how many times each goal has been selected thus far) for a small number of skills (five). Initially, without evidence on all goals, each goal is selected once. Then, the agent should focus on the "easiest-to-discriminate" skills, up until repeated attempts generate less progress than other goals. Goals that are too "hard" generate little progress, so should be avoided until other goals are sufficiently mastered.

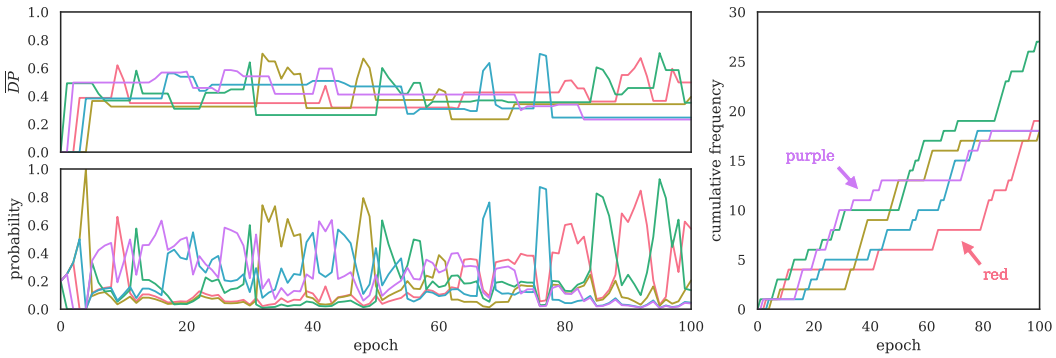

Figure 2: **The effects of DP on goal selection** over training time in the Half-Cheetah environment. The agent is learning five skills, each shown in a different colour. *Upper left:* DP values, $\overline{DP}$; updated for the current skill at the end of an epoch. *Lower left:* goal-selection probabilities after the softmax transformation. *Right:* cumulative frequencies of goal selection. The initial trend for epochs up to the number of skills is due to DP's initialisation, where goals are selected randomly without replacement (see Algorithm 1, ll. 6–10). The plotted traces represent data from running a single random seed.

Consider the purple and red goals. Red initially provided low DP, so was largely ignored until it was sufficiently easy (the policy is not completely greedy, so red is still updated occasionally despite its difficulty) or others provide less DP in comparison (e.g., purple declining). From the cumulative frequencies of selection, we see the red goal eventually caught up—explained by its increase in DP.

RQ3. Does DP-guided selection of goals speed up learning distinguishable skills?

LP should prioritise goals that are providing the most improvement to the discriminator. Therefore, we hypothesise that a DP-motivated agent learns a set of diverse skills faster than a learner sampling goals uniformly. We probe this by visualising trajectories in a dimension-reduced feature space to find out whether trajectories from early training are distinguishable based on their mean observations (Figure 3; details in Appendix D.3). Compared to random skills (no learning) and skills learned with uniform goal selection (DIAYN), the set of skills learned with DP is distinguishable earlier.

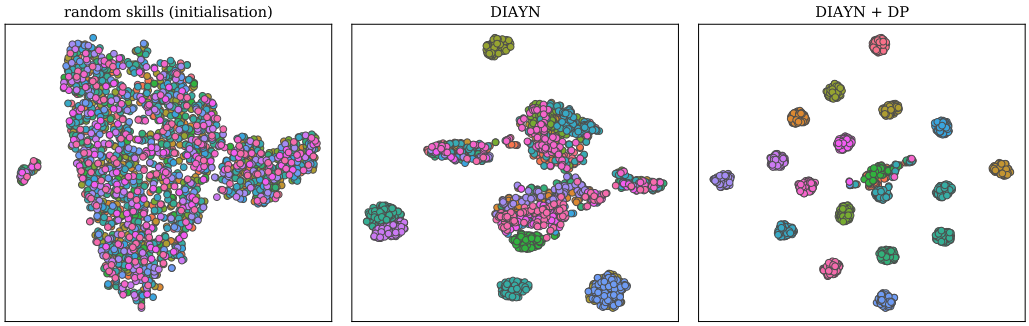

Figure 3: **A dimension-reduced feature space (t-SNE) for 20 skills** in the Half-Cheetah environment. For each skill (one colour), the 100 data points represent i.i.d. draws of trajectories. *Left:* randomly initialised skills with no training. *Middle, Right:* trajectories sampled after 100 epochs of training. Note both DP and DIAYN eventually learn distinguishable skills. Experiment details in Appendix D.3.

6 Conclusion and future work

We propose DP as a method for learning a goal-selection policy in discriminability-motivated RL, prioritising goals based on overall improvements in discriminability. We have shown in three experiments, that: (1) a DP-motivated agent learns a distribution over goals without the probability mass collapsing; (2) the DP values motivate goal selection with respect to observed LP; and (3) with DP, an agent can learn a diverse set of skills in less training time than with uniform-random selection.

In future work, we aim to better understand how different factors affect goal selection. We plan to test other intrinsic rewards combining discriminability with LP, including absolute LP, where the agent also attends to goals that it is forgetting (i.e., skills decreasing discriminability). Different entropy regularisation regimes may benefit diversity in terms of increased state-space coverage but make discrimination of skills harder. Following Eysenbach et al. (2019), the utility of DP can be tested on transfer learning and hierarchical tasks (pp. 6, 7–8), comparing with other discriminator-based CB-IMs, and on a range of environments including non-episodic and stochastic ones. However, evaluating open-ended learning is notoriously difficult. When the aim is to learn a diverse set of "interesting" skills, what is interesting is typically subjectively defined by the researchers and the community, and not clearly designed to maximise a well-defined objective (Colas et al., 2022, p. 1168). We plan to improve the evaluation of skill diversity by considering different diversity metrics, and use our findings to quantitatively evaluate DP against the current state-of-the-art BMI-maxisiming CB-IMs.

Author contributions

We follow the Contributor Role Taxonomy (CRediT) introduced by Brand et al. (2015). Conceptualisation: initial equals EML, NMA; evolution equals EML, NMA, CG. Data curation: EML. Formal analysis: EML. Funding acquisition: CG. Investigation: EML. Design of methodology: lead EML; supporting NMA, CG. Creation of models: leads EML, NMA; supporting CG. Software: EML. Resources: CG. Oversight: equals NMA, CG. Leadership responsibility: CG. Visualisation: lead EML; supporting NMA, CG. Writing: original draft EML; review & editing equals EML, NMA, CG.

Acknowledgements and disclosure of funding

EML and CG received financial support from the Research Council of Finland (NEXT-IM, grant 349036) and NMA from the Helsinki Institute for Information Technology. We acknowledge the computational resources provided by the Aalto Science-IT project. We thank the Aalto Scientific Computing team, particularly Simo Tuomisto, Mira Salmensaari, and Hossein Firooz, for their support. For their time, ideas, and feedback on the topic of measuring skill diversity, we thank Perttu Hämäläinen, Marlos Machado, Sebastian Berns, Nam Hee Kim, and Luigi Acerbi. We thank the reviewers for their service and for providing useful future-looking feedback.

References

Arthur Aubret, Laetitia Matignon, and Salima Hassas. An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, 25(2), 2023. URL `https://doi.org/10.3390/e25020327`.

David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press, 2003. URL `https://dl.acm.org/doi/10.5555/2981345.2981371`.

Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. Relative variational intrinsic control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(8), pages 6732–6740. AAAI Press, 2021. URL `https://doi.org/10.1609/aaai.v35i8.16832`.

Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, 2015. URL `http://dx.doi.org/10.1087/20150211`.

Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer. CURIOUS: Intrinsically motivated modular multi-goal reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1331–1340. PMLR, 2019. URL https://proceedings.mlr.press/v97/colas19a.html.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022. URL https://doi.org/10.1613/jair.1.13554.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016. URL https://doi.org/10.48550/arXiv.1611.07507.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Steven Hansen, Will Dabney, André Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *8th International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJeAHkrYDS.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Leslie Pack Kaelbling. Learning to achieve goals. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1094–8. International Joint Conferences on Artificial Intelligence Organization, 1993. URL https://api.semanticscholar.org/CorpusID:5538688.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. In *Advances in Neural Information Processing Systems*, volume 35, pages 34478–34491. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/debf482a7dbdc401f9052dbe15702837-Paper-Conference.pdf.

Aly Lidayan, Michael Dennis, and Stuart Russell. BAMDP shaping: a unified theoretical framework for intrinsic motivation and reward shaping. *arXiv preprint arXiv:2409.05358*, 2024. URL https://doi.org/10.48550/arXiv.2409.05358.

Erik M Lintunen, Nadia M Ady, Christian Guckelsberger, and Sebastian Deterding. Advancing self-determination theory with computational intrinsic motivation: The case of competence. *PsyArXiv preprint 10.31234/osf.io/7qy35*, 2024. URL https://doi.org/10.31234/osf.io/7qy35.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5502–5511. International Joint Conferences on Artificial Intelligence Organization, 2022. URL https://doi.org/10.24963/ijcai.2022/770.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, volume 31, pages 9209–9220. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/7ec69dd44416c46745f6edd947b470cd-Paper.pdf.

Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007. URL https://doi.org/10.1109/TEVC.2006.890271.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*, 2023. URL https://doi.org/10.48550/arXiv.2304.01315.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1312–1320. PMLR, 2015. URL https://proceedings.mlr.press/v37/schaul15.html.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. URL https://doi.org/10.1109/TAMD.2010.2056368.

Andrew Stout and Andrew G. Barto. Competence progress intrinsic motivation. In *2010 IEEE 9th International Conference on Development and Learning*, pages 257–262. IEEE, 2010. URL https://doi.org/10.1109/DEVLRN.2010.5578835.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. URL https://doi.org/10.1109/IROS.2012.6386109.

David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, and Volodymyr Mnih. Unsupervised control through non-parametric discriminative rewards. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eVMnA9K7.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-SNE effectively. *Distill*, 2016. URL https://doi.org/10.23915/distill.00002.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. URL https://doi.org/10.1007/BF00992696.

A Goal-conditioned RL

Goal-conditioned reinforcement learning (for a review see Liu et al., 2022) extends the standard definition of a reward function (Kaelbling, 1993, p. 1) to be conditioned on goals:

$$R_G : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{G} \to \mathbb{R}, \tag{7}$$

where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of actions available to the agent, and $\mathbf{g} \in \mathcal{G}$, known as a *goal*, or a *goal-defining variable*, is a parameter to the reward function (cf., Colas et al., 2022, p. 1165; Aubret et al., 2023, p. 6). Note that $R_G(\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{g})$ is a random variable, even for fixed $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, $\mathbf{a} \in \mathcal{A}$, and $\mathbf{g} \in \mathcal{G}$. Then, a *skill* is a policy paired with a goal, optimising for the return according to the reward conditioned on that goal. For example, the goal-defining variable, $\mathbf{g}$, can be an index (e.g., Eysenbach et al., 2019) or an element drawn from a learned distribution (e.g., Nair et al., 2018). Its role is simply to indicate which reward function the agent is aiming to maximise.

Goals can be viewed as "a set of *constraints* ... that the agent seeks to respect" (Colas et al., 2022, p. 1165, emphasis in original). While the most immediate intuition of a goal is often as a desired state for the agent to reach (e.g., Kaelbling, 1993, p. 1; Schaul et al., 2015, p. 2), the formalism allows for a more general set of constraints on behaviour (Lintunen et al., 2024, pp. 22–23). In effect, any behaviour that can be defined by attempting to maximise some reward function on the environment can be formulated as a goal–skill pairing.

B Algorithm

Algorithm 1: Diversity Progress for goal selection

1 Set hyperparameters: smoothing η, offset τ, softmax temperature T_{SM}, number of goals $|\mathcal{G}|$.
2 Initialise skills $\pi_{\boldsymbol{\theta}}$, with parameters $\boldsymbol{\theta}$, and discriminator $q_{\boldsymbol{\phi}}$, with parameters $\boldsymbol{\phi}$.
3 Initialise uniform, categorical distribution over goals: $p(\mathbf{g}) = \mathcal{U}\{1, |\mathcal{G}|\}$.
4 Initialise zero vector for Diversity Progress: $\mathbf{dp} = \mathbf{0}$, where $\mathbf{dp} \in \mathbb{R}^{|\mathcal{G}|}$.
5 Initialise environment with initial state $\mathbf{s}_0$ according to the environment.
6 **for** *epoch* $\leftarrow 1$ **to** $|\mathcal{G}|$ **do** `/* Select each goal once to set` $p(\mathbf{g})$ `*/`
7 | Sample goal $\mathbf{g} \sim p(\mathbf{g})$ without replacement.
8 | `UpdateDP(`$\mathbf{g}$`)`
9 Normalise $\mathbf{dp}$ to a probability mass function via softmax: $\exp(\mathbf{dp}/T_{\mathrm{SM}}) / \sum_j \exp(\mathrm{dp}_j/T_{\mathrm{SM}})$.
10 Update goal distribution: $p(\mathbf{g}) \leftarrow \mathbf{dp}$.
11 **while** *training* **do**
12 | Sample goal $\mathbf{g} \sim p(\mathbf{g})$.
13 | `UpdateDP(`$\mathbf{g}$`)`
14 | Normalise $\mathbf{dp}$ and update $p(\mathbf{g})$ as in lines 9–10.

15 **function** `UpdateDP(`$\mathbf{g}$`):`
16 | **for** $t \leftarrow 1$ **to** *steps_per_epoch* **do**
17 | | Sample action $\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t, \mathbf{g})$ from skill associated with $\mathbf{g}$.
18 | | Step environment: $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$.
19 | | Compute prediction errors: $1 - q_{\boldsymbol{\phi}}(\mathbf{g} \mid \mathbf{s}_{t+1})$ if $\mathbf{g}$ is goal being pursued else $q_{\boldsymbol{\phi}}(\mathbf{g} \mid \mathbf{s}_{t+1})$.
20 | | Update discriminator ($\boldsymbol{\phi}$) with optimiser of choice.
21 | | Use reward / algorithm / optimiser of choice to learn goal-conditioned policy.
22 | Compute most recent mean errors: $\overline{\mathbf{e}}(t+1) = \frac{1}{\eta+1} \sum_{i=0}^{\eta} \mathbf{e}(t+1-i)$.
23 | Compute mean errors from τ time steps earlier: $\overline{\mathbf{e}}(t+1-\tau) = \frac{1}{\eta+1} \sum_{i=0}^{\eta} \mathbf{e}(t+1-\tau-i)$.
24 | Compute Diversity Progress: $\overline{DP}(t+1) = \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} \overline{\mathbf{e}}(t+1-\tau) - \overline{\mathbf{e}}(t+1)$.
25 | Update Diversity Progress for current goal $\mathbf{g}$: $\mathrm{dp}_{\mathbf{g}} \leftarrow \overline{DP}(t+1)$.

In our implementation (see Appendix C), we follow DIAYN (Eysenbach et al., 2019) on lines 20–21.

C Implementation

C.1 Learning objective

Our implementation is based on DIAYN, so following Eysenbach et al. (2019, p. 3) we wish to ensure: (1) that states—not actions—are used to distinguish skills; and (2) that the agent is maximising the entropy of its policies. This leads to the following instantiation of the BMI objective (see Section 2.1 for BMI; see Eysenbach et al., 2019, Equations 1–2, p. 4, for the full derivation):

$$\mathcal{F}(\boldsymbol{\theta}) := I(\mathcal{S};\mathcal{G}) + H(\mathcal{A} \mid \mathcal{S}) - I(\mathcal{A};\mathcal{G} \mid \mathcal{S}) = H(\mathcal{A} \mid \mathcal{S},\mathcal{G}) - H(\mathcal{G} \mid \mathcal{S}) + H(\mathcal{G}). \tag{8}$$

The objective function is a variational lower bound on Equation 8 (see Eysenbach et al., 2019, p. 4):

$$\mathcal{F}(\boldsymbol{\theta},\boldsymbol{\phi}) := \underbrace{H(\mathcal{A} \mid \mathcal{S},\mathcal{G})}_{\text{(a)}} + \underbrace{\mathbb{E}_{\mathbf{g}\sim p(\mathbf{g}),\mathbf{s}\sim\pi_{\boldsymbol{\theta}}(\mathbf{g})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{g} \mid \mathbf{s}) - \log p(\mathbf{g})\right]}_{\text{(b)}}, \tag{9}$$

where $q_{\boldsymbol{\phi}}$ is a discriminator, specifically a multi-layer perceptron (MLP), parametrised by $\boldsymbol{\phi}$.

C.2 Optimiser and reward function

In optimising its policy to maximise Equation 9, the agent relies on soft actor-critic (SAC) (Haarnoja et al., 2018), an off-policy maximum entropy actor-critic algorithm. In maximum entropy RL, where learning a stochastic policy is desirable, the standard RL objective of maximising expected return is augmented with an entropy maximisation term, so the optimiser takes care of the entropy term (a) (cf., Eysenbach et al., 2019, p. 4). The expectation (b) is maximised with the intrinsic reward function

$$R_G^i := \log q_{\boldsymbol{\phi}}(\mathbf{g} \mid \mathbf{s}) - \log p(\mathbf{g}). \tag{10}$$

This expresses that the agent is rewarded for its ability to discriminate the skill being followed (see Section 2.1). The distribution over goals, $p(\mathbf{g})$, is: (1) categorical; and (2) learned using DP (we describe the method for learning a goal-selection policy in Section 4). At the beginning of an epoch, the agent samples a skill, $\pi_{\mathbf{g}\sim p(\mathbf{g})}$, and follows it until termination. That is when $p(\mathbf{g})$ is updated.

C.3 Action-selection policy

At the action-selection level, the agent learns a goal-conditioned policy, $\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s},\mathbf{g})$. Following DIAYN (mujoco_all_diayn.py, ll. 210–216), the policy is a Gaussian mixture model (GMM) with K components, where an MLP, parametrised by $\boldsymbol{\theta}$, maps from state–goal pairs to the (log) weight w_k, mean vector $\boldsymbol{\mu}_k$, and vector $\boldsymbol{\sigma}_k$ of (log) standard deviations, i.e., of the diagonal entries of the covariance matrix $\boldsymbol{\Sigma}_k$, of each Gaussian component:

$$\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s},\mathbf{g}) := \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{a};\boldsymbol{\mu}_k,\mathrm{diag}(\boldsymbol{\sigma}_k)). \tag{11}$$

The weights $w_1,\cdots,w_K$ are transformed into probabilities for selecting the component (let the parameters of the choice be indexed by $*$) from which to sample an action: $\mathbf{a} \sim \mathcal{N}_*(\mathbf{a};\boldsymbol{\mu}_*,\mathrm{diag}(\boldsymbol{\sigma}_*))$.

C.4 Hyperparameters

We tested various candidates for the DP hyperparameters: smoothing, η, offset, τ, and softmax temperature T_{SM}. However, since we currently lack a good diversity metric to evaluate differences in performance, the hyperparameter choices in our experiments were mainly based on qualitative observations. Smoothing ranged 100–250 time steps, offset 250–900, and softmax temperature 0.1–0.75. The hyperparameter values used in producing the figures are listed in Table 1. The most noticeable difference was given by varying T_{SM}. To achieve stability in the level of greediness of the goal-selection policy, we normalised the mean LP before computing the mean progress over goals.

For reference, in the 2D environment, episodes last for 100 time steps, so a goal is sampled every 10 episodes. In the Half-Cheetah environment, episodes last for 1000 time steps, so a new goal is sampled for every episode. In the Ant environment, episodes last for a maximum of 1000 time steps (there is a potential termination condition before the maximum episode length), so a sampled goal

Table 1: The hyperparameters used to generate the figures.

		Fig. 1(a)	Fig. 1(b)	Fig. 1(c)	Fig. 2	Fig. 3
smoothing	η	250	250	250	250	100
offset	τ	250	250	750	750	900
temperature	T_{SM}	0.1/0.3	0.1/0.3	0.1/0.3	0.1	0.75
number of goals	$\lvert\mathcal{G}\rvert$	20	20	20	5	20
number of Gaussian components	K	4	4	4	4	4
entropy regularisation scaling	α	0.1	0.1	0.1	0.1	0.1
MLP number of hidden units		32	300	300	300	300

may be used starting from partway through an episode and into another episode. For environment details see Appendix D.1.

For clarification, the same K and α are used for DIAYN and DIAYN+DP across all our experiments. In Figure 3, the DP-specific hyperparameters (η, τ, and T_{SM}) are not needed for DIAYN without DP.

D Evaluation

D.1 Environments

We carried out experimental comparisons of DP against DIAYN, our baseline method, in environments where DIAYN has been shown to perform well, following the advice of Patterson et al. (2023, p. 26). This choice eases comparison with related work. Specifically, we used three environments tested by Eysenbach et al. (2019): a slightly modified version of their *2D Navigation* environment and two MuJoCo environments (Todorov et al., 2012), namely *Half-Cheetah* and *Ant*. The 2D Navigation environment affords diagnostics into the algorithms' function, and with the MuJoCo environments, we represent tasks of increasing complexity in terms of both observation space and degrees of freedom.

2D Navigation Our 2D environment (Figure 4) is a slightly modified version of the one constructed by Eysenbach et al. (2019). An agent starts in the center of the unit box $\boldsymbol{s}_0 = (0.5, 0.5)$ and observes states $\mathbf{s} \in [0, 1]^2$. We modified the action space to $\mathbf{a} \in [-0.05, 0.05]^2$ (whereas the original used larger actions, in $[-0.1, 0.1]^2$). The environment is bounded, in that if an action takes the learner outside of the box, they are projected back to the closest point in the support of the observation space.

MuJoCo HalfCheetah-v1 and Ant-v1 with no modifications (Todorov et al., 2012).

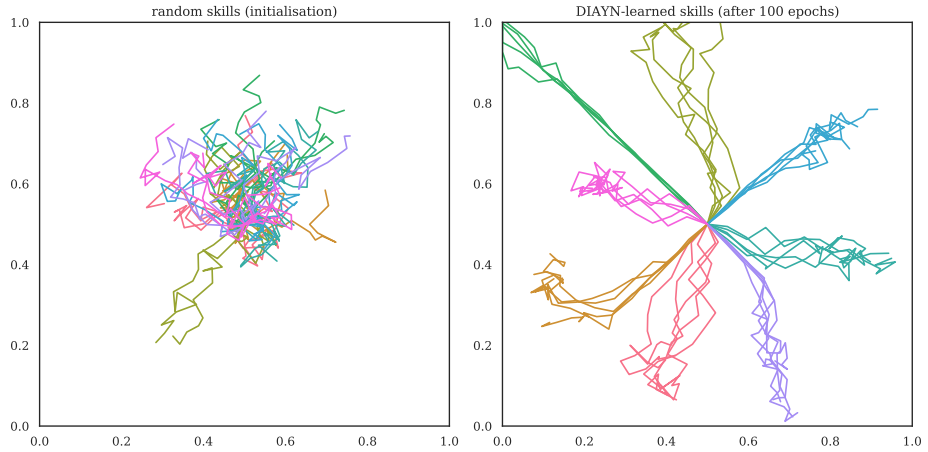

Figure 4: Trajectories drawn from eight stochastic skills in **our modified version of the 2D Navigation environment** constructed by Eysenbach et al. (2019). *Left:* random skills with no training. *Right:* DIAYN-learned skills. Note, these visualisations are provided for intuition, showing trajectories only 15 steps long. In Figure 1(a), we included 20 skills and trajectories were 100 steps long.

D.2 Effective number of skills

Following Eysenbach et al. (2019, Appendix E.2, p. 18), we compute the effective number of skills by exponentiating the entropy of the goal-selection policy, quantifying the number of skills that the policy is effectively sampling from at the time of selection. For intuition, consider an example of an occasion in which one goal is providing significantly more DP than all other goals. This will decrease the effective number of skills, as the learned distribution over goals is less uniform (favouring that one goal). If the DP is similar for all goals, then the distribution is close to uniform and the number of effective skills is close to the total number of skills being learned. With DP, we have some control over the effective number of skills via the softmax temperature: higher temperatures lead to less emphasis on the differences in DP and thus a more uniform sampling distribution; whereas, with low temperatures, we observe higher variance in the effective number of skills over time. High variance in the effective number of skills is likely desirable, since we wish to favour goals only if they provide more DP in comparison to others, and wish for the learner to sample more uniformly in cases of uncertainty (in terms of small differences between the DP values).

Like Eysenbach et al. (2019, p. 6), we believe that VIC's goal-selection policy collapses due to the choice of optimiser (REINFORCE; Williams, 1992). But our results, as well as those from the wider LP literature, demonstrate their claim that learning $p(\mathbf{g})$ "results in learning fewer diverse skills" (Appendix E.2, p 18) to be overgeneralised.

D.3 Evaluating skill diversity with dimensionality reduction

To generate Figure 3, we first draw 100 trajectories from each skill, for each: (1) a set of random skills (no learning), (2) skills learned using DIAYN (uniform goal selection), and (3) skills learned using DIAYN with DP motivating goal selection. Then, we compute the multivariate means over each trajectory, and dimension-reduce the feature space to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) with the following hyperparemeters: a perplexity of 30, a learning rate of 10, and running the optimisation for a maximum of 5000 iterations. In the plot, the distances between clusters may mean nothing (Wattenberg et al., 2016).

The same random seeds were used for the training of DIAYN and DIAYN+DP. Similarly, the seeds for the random number generators of both the policy simulator and t-SNE algorithm were fixed for the processes of simulating trajectories and dimensionality reduction across all three methods.