HOW THE LEVEL SAMPLING PROCESS IMPACTS ZERO-SHOT GENERALISATION IN DEEP REINFORCEMENT LEARNING

Samuel Garcin
University of Edinburgh
s.garcin@ed.ac.uk

James Doran
University of Edinburgh

Shangmin Guo
University of Edinburgh

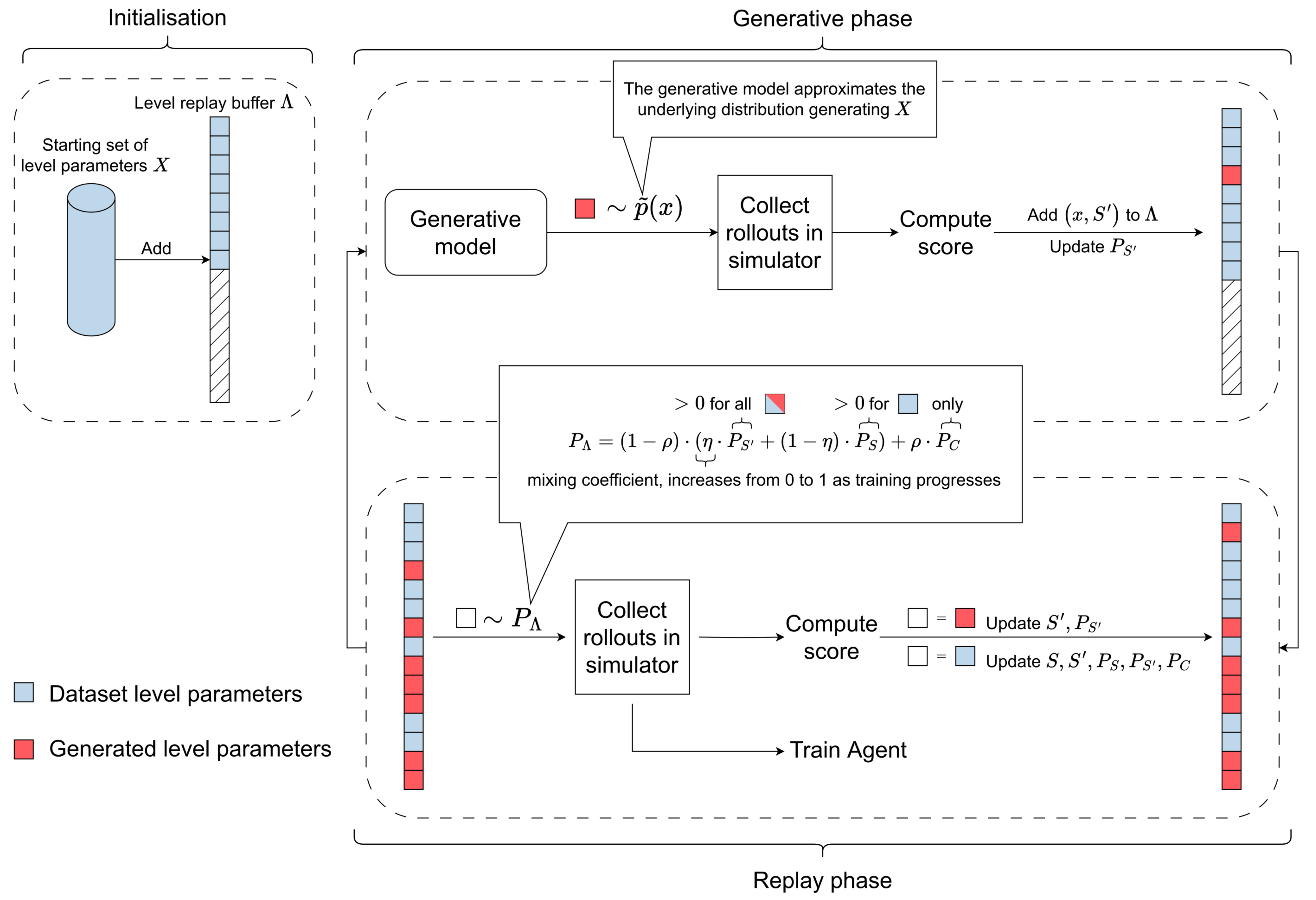
Christopher G. Lucas
University of Edinburgh

Stefano V. Albrecht
University of Edinburgh

Autonomous Agents Research Group
THE UNIVERSITY of EDINBURGH
EDINBURGH CENTRE FOR ROBOTICS

SELF-SUPERVISED ENVIRONMENT DESIGN IMPROVES ZERO-SHOT GENERALISATION…

- RL agents generalise poorly without access to a **large set of training levels**.
- This is a problem as the specification or collection of large sets of level parameters is often **costly**.
- **Self-Supervised Environment Design** (SSED) maximises the **generalisation potential** achievable from a **limited starting set of level parameters**.
- SSED learns a **generative model** of their underlying distribution to **augment** the training set with **synthetic levels**.
- By **adaptively sampling** over this augmented set, SSED can further improve generalisation.
- De-prioritising levels with low value loss lowers the **mutual information** between the agent's model and the training set, minimising an **upper bound on the generalisation gap**.
- SSED's generative model and the gradual inclusion of augmented levels limit **distributional shift**, preventing **overgeneralisation**.



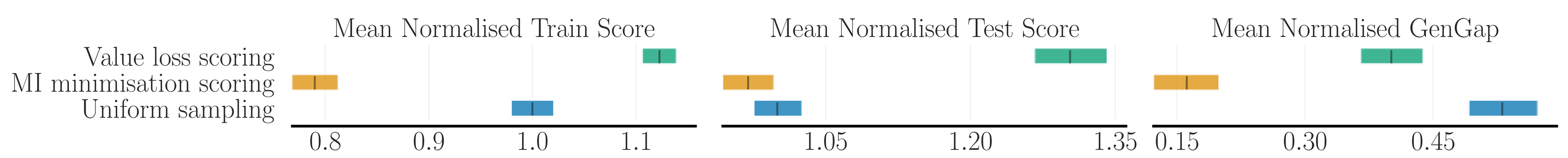
…BY MINIMISING AN UPPER BOUND ON THE GENERALISATION GAP…

The **generalisation gap** measures the gap in agent performance that exists between levels encountered during training and those that were never seen. We target its **upper bound**,

$$\mathrm{GenGap}(\pi) \leq \sqrt{\frac{\mathrm{const}}{|L|} \times \mathrm{MI}(L;\pi)}$$

The number of training levels is increased via **level set augmentation**.

The mutual information between the learned policy and the training levels is minimised by an **adaptive sampling strategy**.

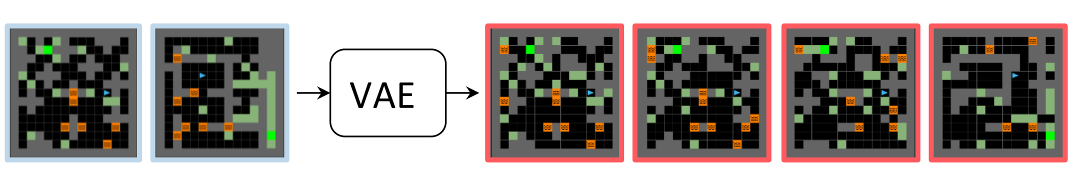


- The impact the level sampling process makes on the **generalisation gap** depends on how well it **regularises the training data** to have **low levels of mutual information**.
- However, directly sampling levels according to a mutual information estimate should be avoided as it impacts **training efficiency** more than it reduces the generalisation gap.

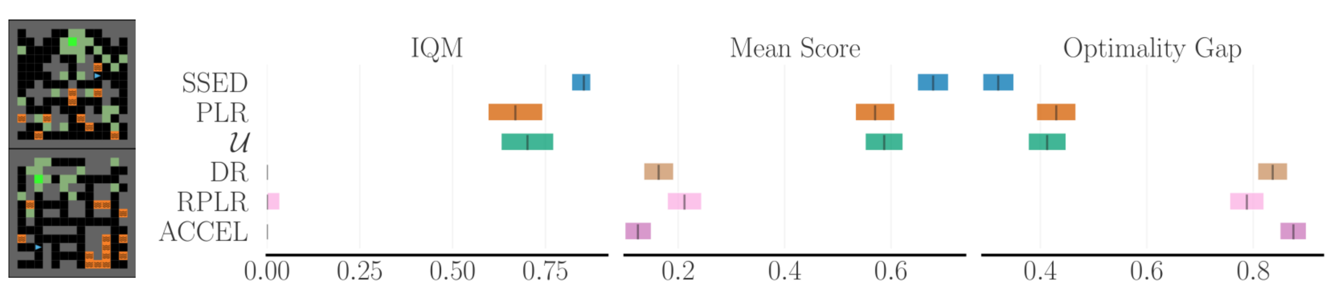
- When the value function contains **level-specific** components, accurate prediction (i.e. a small value loss) is only possible when the **critic's internal representation** is informative of the **level identity**.
- We find that sampling levels according to their **value loss** minimises mutual information (while also slightly improving training efficiency). Our findings help explain the effectiveness of this class of adaptive sampling strategies in reducing the **generalisation gap**.

…WHILE PREVENTING OVERGENERALISATION INDUCED BY DISTRIBUTIONAL SHIFT.

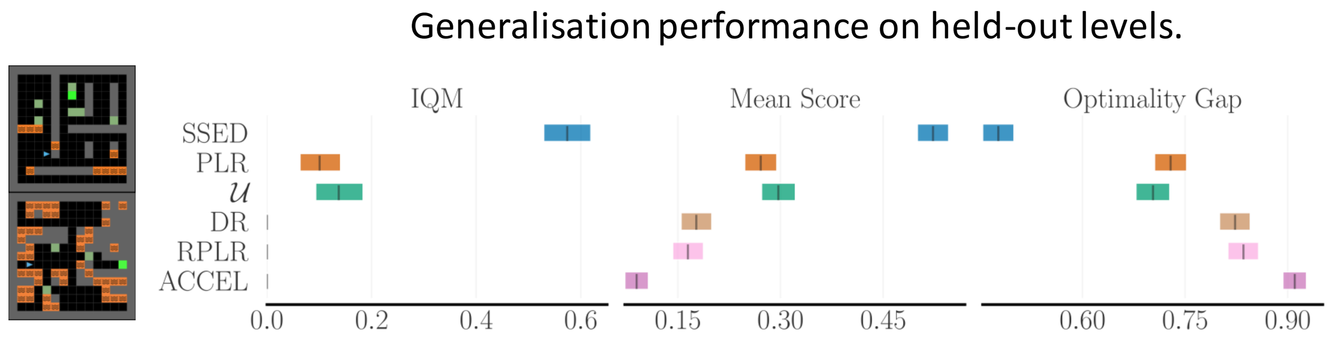
- Employing a generative process that ensures the augmented set of training levels remains consistent with the **ground-truth distribution** is a central component of SSED.
- SSED employs a **VAE** that approximates this distribution by being pre-trained on the starting set of level parameters.
- To **generate** new level parameters, we first compute the **latent encodings** of the parameters in the starting set. We **interpolate** between pairs of latent encodings, exploiting the latent space's smoothness, and decode interpolated points to obtain synthetic level parameters.



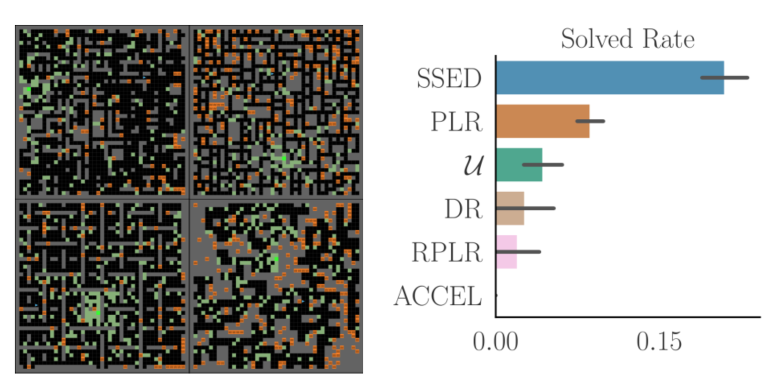
- **Unsupervised** level generation processes risk shifting the learning problem towards undesirable or ineffective policies. These techniques will **overgeneralise** and perform poorly when levels inconsistent with the **task semantics** can be generated.
- In contrast, methods **restricted to the starting set** will **overfit** and are not robust to **edge cases**.
- SSED strikes the right balance between **minimising overfitting** and **preventing overgeneralisation**.



Generalisation performance on held-out levels.



Generalisation performance on edge cases out of distribution w.r.t. the starting set but respecting the task semantics.



Generalisation performance on harder versions of the task.