Discovering General Reinforcement Learning Algorithms with Adversarial Environment Design — Supplementary Materials

1 Hyperparameters

1.1 GROOVE

Hyperparameters shared between GROOVE and LPG were tuned using LPG on Grid-World, then transferred to GROOVE without further tuning. The additional GROOVE hyperparameters (regarding the level buffer) were then tuned separately on Grid-World.

Table 1: GROOVE/LPG hyperparameters

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.0001
Discount factor	0.99
Policy entropy coefficient (β_0)	0.05
Bootstrap entropy coefficient (β_1)	0.001
L2 regularization coefficient for $\hat{\pi}$ (β_2)	0.005
L2 regularization coefficient for $\hat{y}$ (β_3)	0.001
Level buffer size	4000
Replay probability	0.5
Number of interactions per agent update	20
Number of agent updates per optimizer update	5
Number of parallel lifetimes	512
Number of parallel environments per lifetime	64
Algorithmic regret baseline algorithm	A2C

1.2 Agents

Agent hyperparameters were based on tuned A2C agents, before being fine-tuned with LPG. Since we meta-train on a continuous distribution of Grid-World environments, we do not use the agent hyperparameter bandit proposed by Oh et al. [2020] for meta-training.

Table 2: Agent hyperparameters—architecture descriptions $D(N)$ and $C(N)$ respectively refer to dense and convolutional layers of size N; ReLU activations are used throughout.

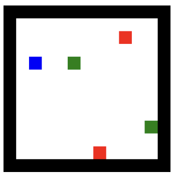
Hyperparameter	Environment		
	Grid-World	Min-Atar	Atari
Architecture	Tabular	D(64)-D(64)	C(32)-C(64)-C(64)-D(512)
Optimizer	SGD	Adam	Adam
Learning rate	40	0.0005	0.0005
Bootstrap KL coefficient (α_y)	0.5	0.5	0.5
Train steps	2500	100,000	100,000
Agent seeds per LPG seed	64	16	1

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

2 Handcrafted Environments

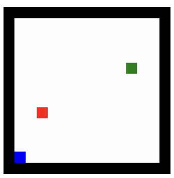
For our handcrafted environment set, we use the set of five tabular Grid-World configurations from Oh et al. [2020]. Grid-World objects are defined by $[r, \epsilon_{\text{term}}, \epsilon_{\text{respawn}}]$, where r represents the reward when collected, ϵ_{term} is the episode-termination probability and $\epsilon_{\text{respawn}}$ is the probability of the object respawning each step after collection.

2.1 Dense



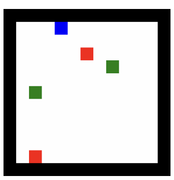
Property	Value
Size	11×11
Objects	$2 \times [1, 0, 0.05], [-1, 0.5, 0.1], [-1, 0, 0.5]$
Maximum episode length	500

2.2 Sparse



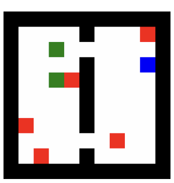
Property	Value
Size	13×13
Objects	$[1, 1, 0], [-1, 1, 0]$
Maximum episode length	50

2.3 Long Horizon



Property	Value
Size	11×11
Objects	$2 \times [1, 0, 0.01], 2 \times [-1, 0.5, 1]$
Maximum episode length	1000

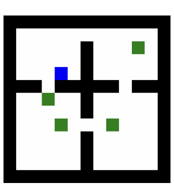
2.4 Longer Horizon



Property	Value
Size	9×9
Objects	$2 \times [1, 0.1, 0.01], 5 \times [-1, 0.8, 1]$
Maximum episode length	2000

Note: size is increased from 7×9 for consistency with our generalized Grid-World distribution.

2.5 Long Dense



Property	Value
Size	11×11
Objects	$4 \times [1, 0, 0.005]$
Maximum episode length	2000

3 Atari Training Curves

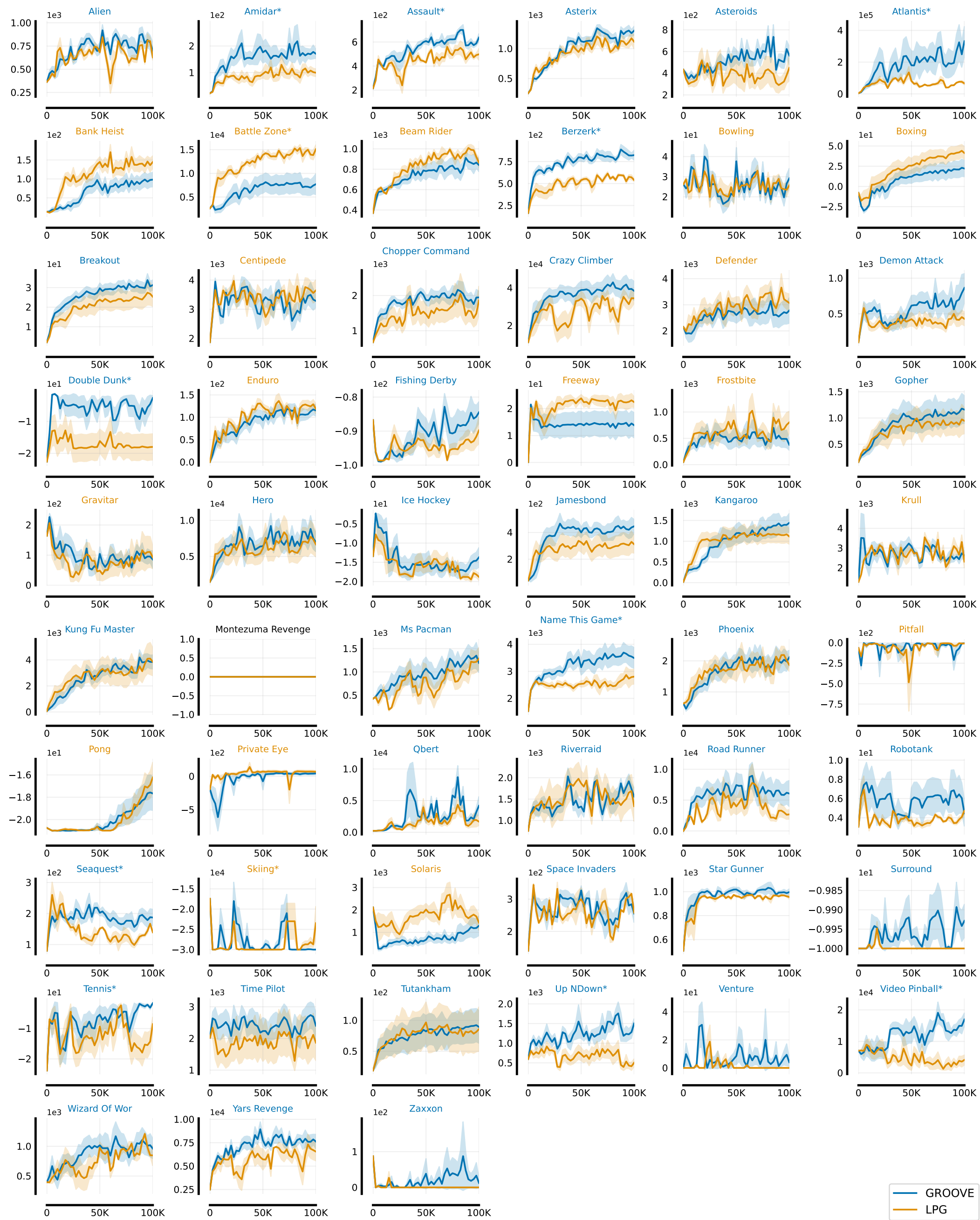


Figure 7: Atari training curves—environment names are highlighted according to highest evaluation return, asterisks (*) denote significant differences in evaluation return (5 seeds, $p < 0.05$).

4 Min-Atar Per-Task Performance

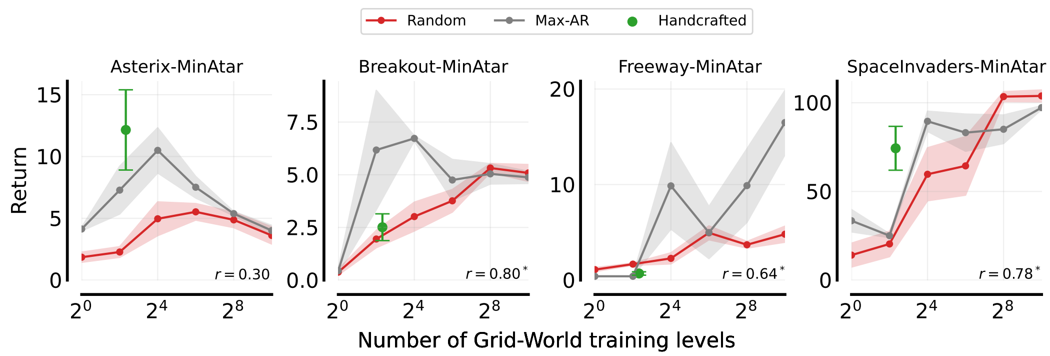


Figure 8: Generalization performance on Min-Atar, after meta-training LPG on variable-sized sets of Grid-World levels (5 seeds)—levels are selected through uniform-random sampling of all levels ("Random"), from the highest-regret levels of a previous LPG instance ("Max-AR"), or from a set of five handcrafted levels ("Handcrafted"). Pearson correlation coefficient is given for Random levels; significant positive correlations are marked with an asterisk (*).

As expected, we observe increased noise when breaking down performance by individual Min-Atar tasks, however, the results from the majority of tasks support our earlier conclusions. Firstly, we observe a significant positive correlation between the number of random training levels and return on three of the four Min-Atar tasks, again demonstrating the impact of task diversity on generalization. When controlling for the number of levels, we observe improved performance after training on handcrafted, rather than random, levels on three of the four Min-Atar tasks. Furthermore, on Asterix, training on handcrafted levels results in higher performance than the largest set of $2^{10} = 1024$ random levels, supporting our conclusion about level informativeness.

After training on high-AR levels, we observe an improvement against random levels on at least three of the four Min-Atar tasks for all sizes of training environment set up to $2^6 = 64$ levels. Beyond this, random and high-AR levels outperform each other on an equal number of tasks, however the dilution in mean AR for larger training sets makes this convergence unsurprising. Furthermore, high-AR levels are competitive with handcrafted levels at the same training set size and quickly outperform the fixed handcrafted set as more high-AR levels are added, demonstrating the effectiveness of AR at identifying informative curricula.