

---

# All You Need Is Supervised Learning: From Imitation Learning to Meta-RL With Upside Down RL

---

Kai Arulkumaran<sup>\*,1,2</sup>, Dylan R. Ashley<sup>3,4,5</sup>, Jürgen Schmidhuber<sup>3,4,5,6,7</sup>, and Rupesh K. Srivastava<sup>7</sup>

<sup>1</sup> ARAYA, Inc., Tokyo, Japan

<sup>2</sup> Imperial College London, London, UK

<sup>3</sup> The Swiss AI Lab IDSIA, Lugano, Switzerland

<sup>4</sup> Università della Svizzera Italiana (USI), Lugano, Switzerland

<sup>5</sup> Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Lugano, Switzerland

<sup>6</sup> King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

<sup>7</sup> NNAISENSE, Lugano, Switzerland

## Abstract

Upside down reinforcement learning (UDRL) flips the conventional use of the return in the objective function in RL upside down, by taking returns as input and predicting actions. UDRL is based purely on supervised learning, and bypasses some prominent issues in RL: bootstrapping, off-policy corrections, and discount factors. While previous work with UDRL demonstrated it in a traditional online RL setting, here we show that this single algorithm can also work in the imitation learning and offline RL settings, be extended to the goal-conditioned RL setting, and even the meta-RL setting. With a general agent architecture, a single UDRL agent can learn across all paradigms.

**Keywords:** supervised learning, imitation learning, offline reinforcement learning, goal-conditioned reinforcement learning, meta-reinforcement learning

## Acknowledgements

This work was partially supported by the European Research Council (ERC, Advanced Grant Number 742870).

---

\*Correspondence to [kai\\_arulkumaran@araya.org](mailto:kai_arulkumaran@araya.org). Work partially done while author was at NNAISENSE.

# 1 Introduction

Over time, the field of AI has exhibited significant convergence, with the increasing popularity of artificial neural networks (NNs) trained on large amounts of data resulting in improved performance across many benchmarks. While many earlier successes were based on supervised learning (SL), NNs have also revolutionised reinforcement learning (RL).

We begin by noting two further significant convergences in recent years. The first pertains to NN architectures. While convolutional NNs (CNNs) and recurrent NNs (RNNs) have for decades been applied to images and text, respectively, a new general-purpose architecture has emerged [16]. **Originally applied to text, the Transformer architecture [34] has been successfully extended to images [6], as well as other structured input/outputs [16]. The advantage of such models is the ability to operate over sets, allowing a single, modular model to reuse knowledge over a dynamic set of inputs.**

The second development is the rise of self-supervised learning (SSL) across different domains [36]: the use of “pretext tasks”, constructing labels from unlabelled data in order to imitate SL training, has resulted in models that rival those trained with SL. SSL benefits from rich “supervisory signals”, with models learning to *predict* transformations,  $f$ , of their inputs,  $x$ ; loosely speaking,  $p(f(x)|x)$ , versus SL’s  $p(y|x)$ , where  $y$  is a label. In a single domain, a variety of tasks can enable a model to gain complementary knowledge over different facets of the world.

We propose uniting these directions to create general learning agents, with the position that RL can itself be framed as an SL problem. This is not a novel proposition [25, 31, 32, 20, 12, 3, 17, 8, 11], but in contrast to prior works, we provide a general framework that includes online RL, goal-conditioned RL (GCRL) [28], imitation learning (IL) [26], offline RL [9], and meta-RL [29], as well as other paradigms contained within partially observed Markov decision processes (POMDPs) [23]. We build upon the proposal of upside down RL (UDRL) by Schmidhuber [31], the implementation of Srivastava et al. [32], and sequence modelling via Decision Transformers [3, 17]. We examine current implementations, discuss a generalisation of the framework, and then demonstrate a *single algorithm and architecture* on a standard control problem.

## 2 Upside Down RL

The core of UDRL is the policy,  $\pi$ , conditioned on “commands”,  $c$  [31]. Given a dataset  $D$  of trajectories (states,  $s$ , actions,  $a$ , and rewards,  $r$ ), the policy, parameterised by  $\theta$ , is trained using an SL loss,  $\mathcal{L}$ , to map states and commands to actions:

$$\arg \min_{\theta} \mathbb{E}_{s,a,r \sim D} [\mathcal{L}(a, \pi(a|s, c; \theta))]. \quad (1)$$

In the initial implementation [32],  $c = [d^H, d^R]$ , where  $d^H$  is the desired (future) time horizon,  $t_2 - t_1$ , and  $d^R$  is the (time-bounded) return-to-go,  $\sum_{t=t_1}^{t_2} r_t$ . In a similar fashion to hindsight experience replay<sup>1</sup> [19, 1], the agent can be trained on data sampled from  $D$ , calculating  $d^H$  and  $d^R$  directly from the data. Trained thus, the agent can then achieve a desired return by sampling actions from its stochastic command-conditioned policy. Unlike typical RL agents, UDRL agents are capable of achieving low returns, medium returns, or high returns, based on the choice of  $d^R$  [32].

There are two key elements to the efficacy of UDRL—the first being the dataset (and its use) [27]. In the online RL setting, a UDRL agent can learn to perform reward maximisation in a manner akin to expectation maximisation: collect data using the policy, then train on the most rewarding data. This can be achieved by weighting the data [25], or controlling data storage/sampling [32]. Given an existing dataset of demonstrations, the agent can be trained in the offline RL setting [20, 17, 3].  $D$  can be used for auxiliary tasks to improve learning/generalisation, such as via learning a world model [17].

The second key element is the use of commands.  $c$  can be any computable predicate that is consistent with the data [31]. In the initial implementation, the agent is trained to map observed actions  $a_t$  to the corresponding states  $s_t$  and  $[d^H, d^R]$ , where the latter is calculated from  $t$  to the terminal timestep  $T$ , allowing the agent to perform (undiscounted) credit assignment across the entire episode. During exploration in the environment (which generates more data for the agent),  $d^H$  is set to the mean of the most rewarding episodes in  $D$ , and  $d^R$  is sampled uniformly from values between the mean of the most rewarding episodes’ returns, and the mean plus one standard deviation (encouraging optimism). After every environment interaction,  $d^H$  is decremented by 1 and  $d^R$  is decremented by  $r$ . If  $c$  is simplified to only contain the desired return, we recover reward-conditioned policies [20], and if  $c = \emptyset$ , we recover behavioural cloning (BC) [26], which is the simplest IL algorithm. Conversely, the trivial augmentation of  $c$  with a goal vector  $g$  extends UDRL to the GCRL setting [31, 12, 17], with zero-shot generalisation enabled via appropriate goal spaces (e.g., language [18]).

As the agent is trained using SL on observed data, UDRL bypasses several common issues in RL: bootstrapping (temporal-difference updates), off-policy corrections, and discount factors. This allows us to more easily focus on standard points in ML: the agent’s ability to generalise to novel states and commands is based on the data available, the class of policies, and the optimisation process [8]. Srivastava et al. [32] trained fully-connected-/CNN-based UDRL agents using stochastic gradient descent on the cross-entropy loss, but replacing the architecture (e.g., with Transformers [17, 3])

<sup>1</sup>But without specific environment settings, e.g., symbolic state spaces or a distance-based goal-conditional reward functions.

or the optimisation (e.g., with evolutionary algorithms) are valid alternatives. While the move to sequence modelling allows the agent to benefit from greater contexts [17, 3], in general, a *stateful* agent is needed in order to deal with POMDPs, as well as more general computable predicates [31].

### 3 Generalised Upside Down RL in POMDPs

While many realistic problems cannot be captured by MDPs, they can be reasonably modelled by POMDPs, in which the agent only receives a partial observation  $o$  of the true state  $s$ . A principled approach to solving POMDPs is to keep a “belief” over the state, which can be achieved implicitly by training an RNN agent, which updates a hidden state vector  $h$  [37]. POMDPs encompass many other problems, such as hidden parameter MDPs [5], which consider related tasks, and even the general meta-RL setting (where  $h$  is typically augmented with previous action, reward and terminal indicator,  $\mathbb{1}_{\text{terminal}}$  [7, 35]). POMDPs can also be related to generalisation in RL and the robust RL [22] setting, which considers worst-case performance. While various specialised algorithms exist for these different problem settings, recent results from Ni et al. [23] have shown that simple recurrent model-free RL agents can perform well across the board. Motivated by this (and further related arguments by Schmidhuber [31]), we proceed by treating every environment as a POMDP, in which an agent attempts to learn using a general algorithmic framework.

A final ingredient enables a *single agent* to deal with all RL problem settings (plus IL) with *one model* — an architecture that can deal with arbitrary structured inputs and outputs (e.g., Perceiver IO [16]). Such a model allows  $c$  to be dynamic as needed: being null in the pure IL setting (where UDRL reduces to BC), being  $[d^H, d^R, g]$  in the GCRL setting, and extending further beyond to incorporate SSL tasks. This alleviates the problem of using unlabelled demonstrations to bootstrap an RL agent, as the agent does not need to model the rewards achieved. Furthermore, the ability to adapt to different observation and action spaces allows such an agent to use third-person data for representation learning.<sup>2</sup>

Given this, we present a generalised algorithm for UDRL (Algorithm 1). Although nearly all RL problems consider the episodic setting (in which a terminal indicator is given before the environment resets), the following algorithm is applicable in both episodic and non-episodic MDPs, making it capable of continual/lifelong learning.

---

#### Algorithm 1 Generalised Upside Down RL

---

**Require:**  $E$  ▷ POMDP environment  
**Require:**  $\pi(a|o, c, h)$  ▷ Command-conditioned recurrent policy  
**Require:**  $D$  ▷ Experience replay memory; existing data optional unless performing IL/offline RL

**function** RESET( $E, \pi, D$ )  
  Reset environment  $E$  and  $\pi$ 's hidden state  $h$  ▷  $h$  also contains the previous action, reward, and terminal indicator  
  Get initial observation and goal  $(o, g)$  from  $E$   
  Sample  $c$  based on  $D$  and  $(o, g)$  ▷ Requires a procedure for sampling an initial command  $c$  based on observed data

Train  $\pi$  on batches from  $D$  ▷ Data can be non-uniformly/adaptively sampled from  $D$ ; auxiliary objectives can be used if performing IL or offline RL without environment interaction **then return**

RESET( $E, \pi, D$ )  
**while** true **do**  
  Act with  $a, h \sim \pi(a|o, c, h)$   
  Observe  $(o', r, g', \mathbb{1}_{\text{terminal}})$  from environment transition ▷  $g'$  given for goal-conditioned RL,  $\mathbb{1}_{\text{terminal}}$  for an episodic MDP  
  Update  $D$  with  $(o, a, r, g, \mathbb{1}_{\text{terminal}})$  ▷  $D$  may prioritise updates/remove old data  
  Update  $h$  (to contain  $a$  and  $r$ ) and  $c$  ▷ Requires a procedure for updating  $c$ , e.g., include  $g$ , decrement  $d^H$   
  Train  $\pi$  on batches from  $D$   
  **if**  $\mathbb{1}_{\text{terminal}}$  **then** RESET( $E, \pi, D$ )

---

### 4 Experiments

For our experiments, we use the classic CartPole control problem [2], where the agent receives a +1 reward for every timestep the pole is balanced, with a time limit of 500 timesteps. We adapt this environment to demonstrate how a single architecture can be used in the following settings: online RL, IL, offline RL, GCRL, and meta-RL. For the IL setting, we train the agent on 5 episodes with the maximum return of 500, collected from an online RL agent, and disable the reward and desired return inputs. For the offline setting, we train the agent on the worst 1000 trajectories from the online agent, with an average return of  $162 \pm 195$ . In the GCRL setting, at the beginning of each episode the agent is given an  $x$  goal position uniformly sampled from  $[-1, 1]$ , with the reward function set to  $e^{-|x-g|}$ . In the meta-RL setting, several environment parameters are randomly sampled at the beginning of each episode [21]. In the GCRL and meta-RL settings, during testing the agent is evaluated on a cross product of a uniform spread over goals/environment parameters.

<sup>2</sup>Correspondences between the agent and third party's observation/action spaces would be needed for true third-person IL [33]. To maintain such flexibility, the previous action and reward can be included in a dynamic  $c$  structure, instead of within  $h$ .

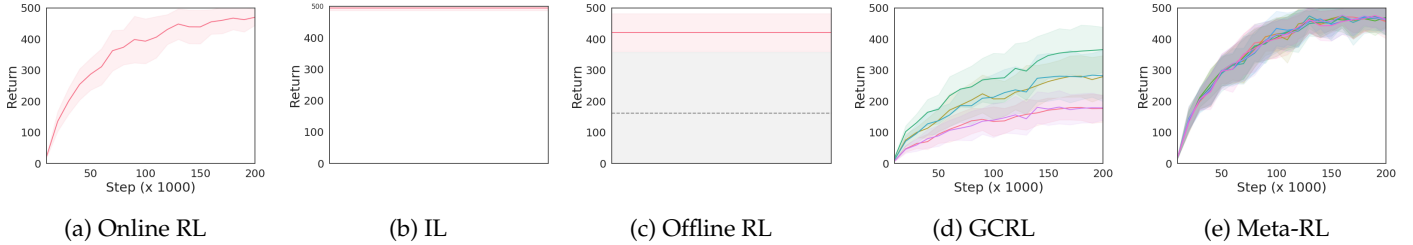


Figure 1: Generalised UDRL agent trained under the (a) online RL (b) IL (c) offline RL (d) GCRL and (e) meta-RL settings in the CartPole environment. Results averaged over 10 test episodes per evaluation  $\times$  20 random seeds. The dashed line in (c) represents the distribution of returns in  $D$ . The different colours in (d) and (e) represent different goals/environment parameters, respectively.

As the observation and action spaces are constant, we use a simple base policy that consists of a linear layer, a sigmoid-gated linear layer, an LSTM [14], and a final linear layer to predict the logits of a categorical distribution. Each command ( $d^H$  and  $d^R$ ), the goal, previous action, reward, and terminal indicator, are all embedded and concatenated with a learnable encoding vector, before being processed by a Transformer encoder layer [34]; the resulting vectors are aggregated using the max function and then used in the gated linear layer before the LSTM. For simplicity, all settings use the same architecture and hyperparameters; the code is available at <https://github.com/Kaixhin/GUDRL>.

As shown in Figure 1, the generalised UDRL agent is able to learn under all settings. While its performance may not match more specialised RL algorithms, it demonstrates that a single SL objective is sufficient for a variety of sequential decision making problems. Whilst it is possible to tune hyperparameters for individual tasks, a more compelling avenue for improving the performance of such agents is to incorporate further commands/tasks that relate to the environment’s structure. This can include learning world models, SSL tasks [36], and more general computable predicates [31].

## 5 Discussion

Given the increased interest in the RL-as-SL paradigm, this work aims to construct a more general purpose agent/learning algorithm, but with more concrete implementation details and links to existing RL concepts than prior work [31]. A major question is whether such an idea can scale? As mentioned previously, optimisation and function approximation are key limiters. Other experiments with tabular representations have yielded UDRL agents that can learn more complex commands [31]; and in further experiments, the act of resetting weights has been useful for more complex agents. With less of the confounding problems of other RL algorithms, UDRL lays bare the problem of continual learning (and proactive interference, in particular) [10, 15].

Another discussion point UDRL introduces is the method by which (value) credit assignment can occur. While other agents typically consider the (discounted) episodic return, UDRL agents can incorporate a desired horizon. As such, UDRL might lend itself better to hierarchical RL [13]. In the current formulation, desired returns and goals are almost interchangeable in the command structure, but a more powerful formulation is to consider these as being variables that can be inferred themselves, i.e., a joint model  $(o, a, d^R, g)$ . In the same way that inverse models can complement forward models, the joint distribution can interchange  $Q$ -functions,  $Q(d^R|o, a, g)$  and return-conditioned policies,  $\pi(a|o, d^R, g)$ , utilising whichever is better for the situation (acting, or learning, in a data-dependent manner). Being able to infer the desired goal becomes particularly important when it comes to generalisation in the IL setting [4], and could even allow the use of “suboptimal” data [8]. Looking even further forward, with both hierarchy and a more intelligent command selection strategy, UDRL agents could be powerful vessels for implementing open-ended, goal-driven curiosity [24, 30].

## References

- [1] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba. Hindsight Experience Replay. In *NeurIPS*, 2017.
- [2] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Trans. SMC*, 13(5):834–846, 1983.
- [3] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision Transformer: Reinforcement Learning via Sequence Modeling. *arXiv:2106.01345*, 2021.
- [4] P. De Haan, D. Jayaraman, and S. Levine. Causal Confusion in Imitation Learning. *NeurIPS*, 2019.
- [5] F. Doshi-Velez and G. Konidaris. Hidden Parameter Markov Decision Processes: A Semiparametric Regression Approach for Discovering Latent Task Parametrizations. In *IJCAI*, 2016.

- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.
- [7] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel.  $RL^2$ : Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779*, 2016.
- [8] S. Emmons, B. Eysenbach, I. Kostrikov, and S. Levine. RvS: What is Essential for Offline RL via Supervised Learning? *arXiv:2112.10751*, 2021.
- [9] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based Batch Mode Reinforcement Learning. *JMLR*, 6:503–556, 2005.
- [10] W. Fedus, D. Ghosh, J. D. Martin, M. G. Bellemare, Y. Bengio, and H. Larochelle. On Catastrophic Interference in Atari 2600 Games. *arXiv:2002.12499*, 2020.
- [11] H. Furuta, Y. Matsuo, and S. S. Gu. Generalized Decision Transformer for Offline Hindsight Information Matching. In *ICLR*, 2022.
- [12] D. Ghosh, A. Gupta, A. Reddy, J. Fu, C. M. Devin, B. Eysenbach, and S. Levine. Learning to Reach Goals via Iterated Supervised Learning. In *ICLR*, 2021.
- [13] N. Gürtler, D. Büchler, and G. Martius. Hierarchical Reinforcement Learning with Timed Subgoals. In *NeurIPS*, 2021.
- [14] S. Hochreiter and J. Schmidhuber. Long Short-term Memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [15] M. Igl, G. Farquhar, J. Luketina, W. Boehmer, and S. Whiteson. Transient Non-stationarity and Generalisation in Deep Reinforcement Learning. In *ICLR*, 2021.
- [16] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv:2107.14795*, 2021.
- [17] M. Janner, Q. Li, and S. Levine. Offline Reinforcement Learning as One Big Sequence Modeling Problem. In *NeurIPS*, 2021.
- [18] Y. Jiang, S. S. Gu, K. P. Murphy, and C. Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. *NeurIPS*, 2019.
- [19] L. P. Kaelbling. Learning to Achieve Goals. In *IJCAI*, 1993.
- [20] A. Kumar, X. B. Peng, and S. Levine. Reward-conditioned Policies. *arXiv:1912.13465*, 2019.
- [21] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Learning to Generalize: Meta-learning for Domain Generalization. In *AAAI*, 2018.
- [22] J. Morimoto and K. Doya. Robust Reinforcement Learning. In *NeurIPS*, 2000.
- [23] T. Ni, B. Eysenbach, and R. Salakhutdinov. Recurrent Model-free RL is a Strong Baseline for Many POMDPs. *arXiv:2110.05038*, 2021.
- [24] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Trans. Evol. Comput.*, 11(2):265–286, 2007.
- [25] J. Peters and S. Schaal. Reinforcement Learning by Reward-weighted Regression for Operational Space Control. In *ICML*, pages 745–750, 2007.
- [26] D. A. Pomerleau. ALVINN: An Autonomous Land Vehicle in a Neural Network. In *NeurIPS*, 1988.
- [27] M. Riedmiller, J. T. Springenberg, R. Hafner, and N. Heess. Collect & Infer-A Fresh Look at Data-efficient Reinforcement Learning. In *CoRL*, 2022.
- [28] J. Schmidhuber. Learning Algorithms for Networks with Internal and External Feedback. In *Connectionist Models*, pages 52–61. Elsevier, 1990.
- [29] J. Schmidhuber. On Learning How to Learn Learning Strategies. Technical Report FKI-198-94, Technische Universität München, 1994.
- [30] J. Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE TAMC*, 2(3):230–247, 2010.
- [31] J. Schmidhuber. Reinforcement Learning Upside Down: Don’t Predict Rewards—Just Map Them to Actions. *arXiv:1912.02875*, 2019.
- [32] R. K. Srivastava, P. Shyam, F. Mutz, W. Jaśkowski, and J. Schmidhuber. Training Agents Using Upside-down Reinforcement Learning. In *NeurIPS Deep RL Workshop*, 2019.
- [33] B. C. Stadie, P. Abbeel, and I. Sutskever. Third-person Imitation Learning. In *ICLR*, 2017.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017.
- [35] J. X. Wang, Z. Kurth-Nelson, H. Soyer, J. Z. Leibo, D. Tirumala, R. Munos, C. Blundell, D. Kumaran, and M. M. Botvinick. Learning to Reinforcement Learn. In *CogSci*, 2017.
- [36] L. Weng. Self-Supervised Representation Learning. *lilianweng.github.io/lil-log*, 2019.
- [37] D. Wierstra, A. Foerster, J. Peters, and J. Schmidhuber. Solving Deep Memory POMDPs with Recurrent Policy Gradients. In *ICANN*, 2007.