Reinforcement learning for legged robots

Stéphane Caron

November 28, 2025

Département d'informatique de l'ENS (CNRS, ENS-PSL, Inria)

Quick history of robot learning

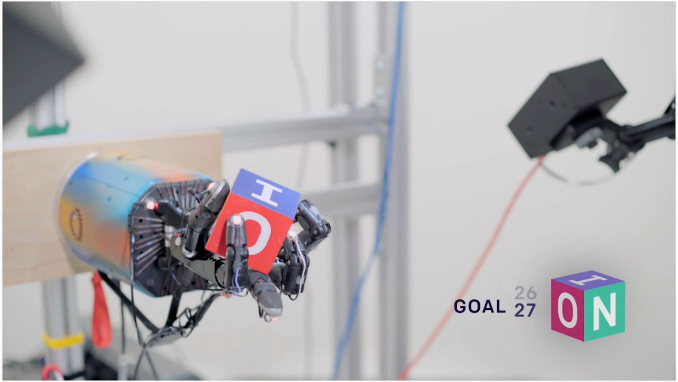
Policy trained with current RL techniques [Lia+24]

Video: `https://youtu.be/8pR1HE-wMHw`

2020: QUADRUPEDAL LOCOMOTION

Teacher-student residual reinforcement learning [Lee+20]

Video: `https://youtu.be/oPNkeoGMvAE`



LSTM policy with domain randomization [And+20]

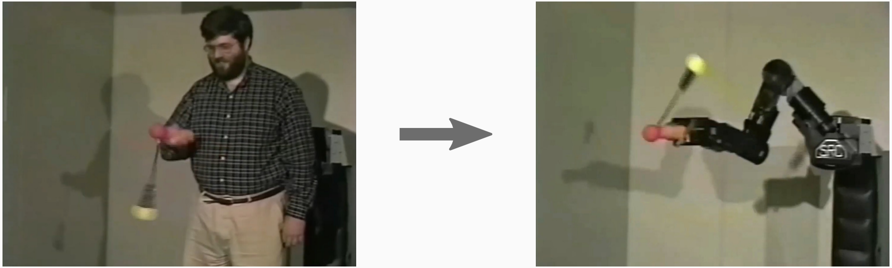
Video: `https://youtu.be/jwSbzNHGflM`

2010: HELICOPTER STUNTS

Helicopter aerobatics through apprenticeship learning [ACN10]

Video: `https://youtu.be/M-QUkgk3HyE`

1997: PENDULUM SWING UP

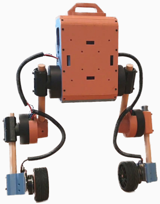


Swinging up an inverted pendulum from human demonstrations [AS97]

Video: `https://youtu.be/g3I2VjeSQUM?t=294`

Intro to reinforcement learning

Agent



action $a \in \mathcal{A}$

observation $o \in \mathcal{O}$

reward $r \in \mathbb{R}$

Environment

REWARDS

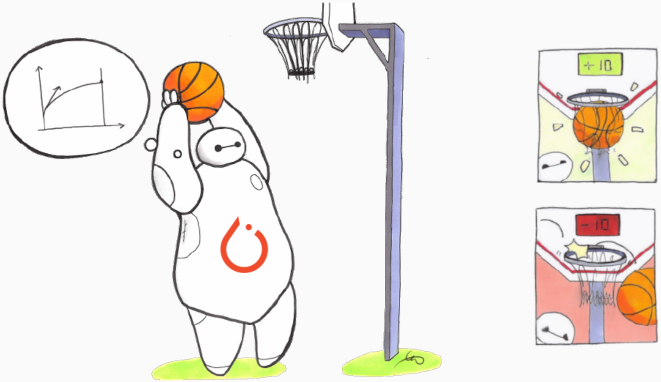
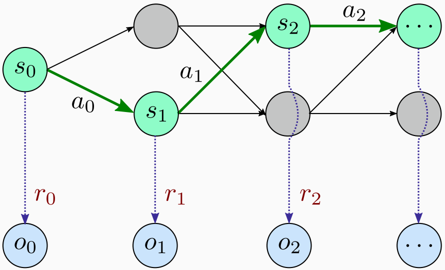


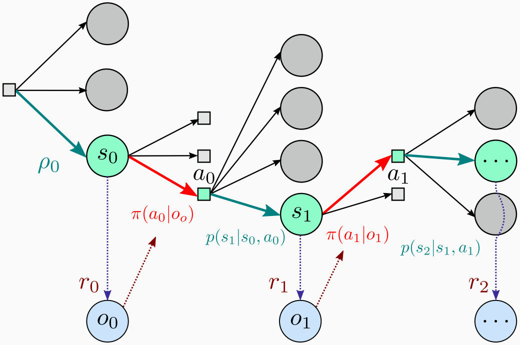
Image credit: L. M. Tenkes, source: `https://araffin.github.io/post/sb3/`

PARTIALLY OBSERVABLE MARKOV DECISION PROCESS (1/2)

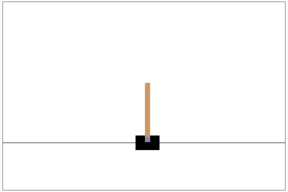


- **State:** s_t, ground truth of the environment
- **Action:** a_t, decision of the agent (discrete or continuous)
- **Transitions:** only depend on current state and action (Markov property)
- **Observation:** o_t, *partial* estimation of the state from sensors
- **Reward:** $r_t \in \mathbb{R}$, scalar feedback, often $r_t = r(s_t, a_t)$ or $r(s_t, a_t, s_{t+1})$

PARTIALLY OBSERVABLE MARKOV DECISION PROCESS (2/2)



	Deterministic	*Stochastic*	
Transitions:	$s_{t+1} = f(s_t, a_t)$	$s_{t+1} \sim p(\cdot|s_t, a_t)$	how the environment evolves
Initial state:	s_0	$s_0 \sim \rho_0(\cdot)$	where we start from
Observation:	$o_t = h(s_t)$	$o_t \sim z(\cdot|s_t)$	how sensors measure the world
Policy:	$a_t = g(s_t)$	$a_t \sim \pi(\cdot|o_t)$	what the agent decides



```python
import gymnasium as gym

with gym.make("CartPole-v1", render_mode="human") as env:
    observation, _ = env.reset()
    action = env.action_space.sample()
    for step in range(1_000_000):
        observation, reward, terminated, truncated, _ = env.step(action)
        if terminated or truncated:
            observation, _ = env.reset()
        position = observation[0]
        action = 0 if position > 0.0 else 1
```

SAME API FOR SIMULATION AND REAL ROBOTS

```python
import gymnasium as gym

with gym.make("Upkie-PyBullet-Pendulum", frequency=200.0) as env:
    observation, _ = env.reset()
    action = env.action_space.sample()
    for step in range(1_000_000):
        observation, reward, terminated, truncated, _ = env.step(action)
        if terminated or truncated:
            observation, _ = env.reset()
        pitch = observation[0]
        action[0] = 10.0 * pitch  # action is [ground_velocity]
```

Two last missing pieces:

- **Episode:** $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ truncated or infinite[1]
- **Return:** $R(\tau) = \sum_{t \in \tau} r_t$ or with discount $\gamma \in]0, 1[$: $R(\tau) = \sum_{t \in \tau} \gamma^t r_t$

We can now state what reinforcement learning is about:

Goal of reinforcement learning

The goal of reinforcement learning is to *find a policy that maximizes returns*.

[1] In practice episodes contain o_t rather than s_t. In RL, we implicitly assume that observations contain enough information to be in bijection with their corresponding states. See also *Augmenting observations* thereafter.

In the stochastic setting, the goal of reinforcement learning is:

$$
\begin{aligned}
\max_{\pi}\ & \mathbb{E}_\tau[R(\tau)] \\
\text{s.t. }\ & R(\tau) = \textstyle\sum_t r_t \\
& \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots) \\
& s_0 \sim \rho_0(\cdot) \\
& o_0 \sim z(\cdot|s_0) \\
& a_0 \sim \pi(\cdot|o_0) \\
& s_1 \sim p(\cdot|s_0, a_0) \\
& \vdots
\end{aligned}
$$

The value $V(s) \in \mathbb{R}$ of a state s is the expected return achieved when starting from s.

On-policy value function

Expected return from s following a given policy:

$$V^{\pi}(s) = \mathbb{E}_{\tau \sim \pi}(R(\tau) | s_0 = s)$$

Optimal value function

Best return we can expect from s:

$$V^{*}(s) = \max_{\pi} \mathbb{E}_{\tau \sim \pi}(R(\tau) | s_0 = s)$$

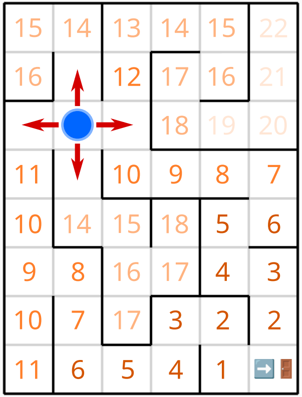


Figure 1: Labyrinth with discrete NSEW actions. Orange: distance to the exit.

Value functions satisfy the Bellman equation:

Bellman equation

For the on-policy value function, we have recursively:

$$V^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s),(r,s') \sim p(s'|s,a)}[r + \gamma V^{\pi}(s')]$$

And similarly for the optimal value function:

$$V^{*}(s) = \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s),(r,s') \sim p(s'|s,a)}[r + \gamma V^{*}(s')]$$

An optimal policy π^* can thus be derived from an optimal value function V^*.

This gives rise to algorithms like Q-learning that search for optimal value functions.

COMPONENTS OF AN RL ALGORITHM

A reinforcement-learning algorithm may include any of the following:

- **Policy:** compute the agent's action from an observation
- **Value function:** estimate the policy/best return from a state
- **Model:** an internal function approximating the (unknown) environment dynamics

An algorithm with a policy (actor) and a value function (critic) is called *actor-critic*.

An algorithm with an explicit model is called *model-based* (without: *model-free*).

A TAXONOMY OF RL ALGORITHMS

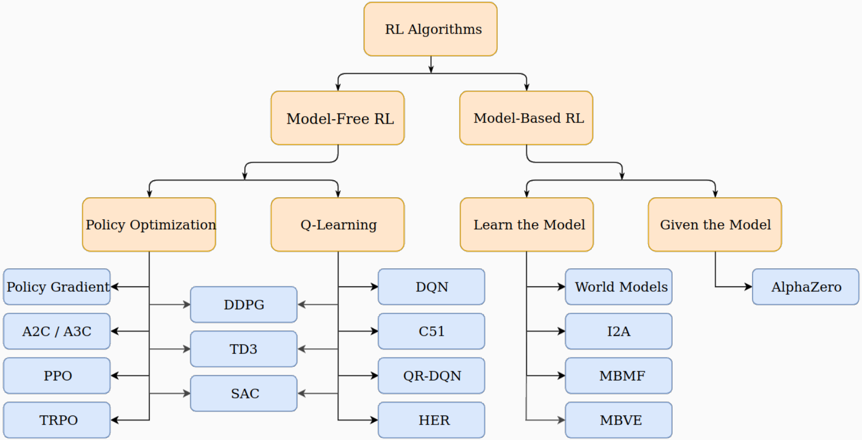


Figure 2: There are several taxonomies, none of them fully works. This one is from [Ach18].

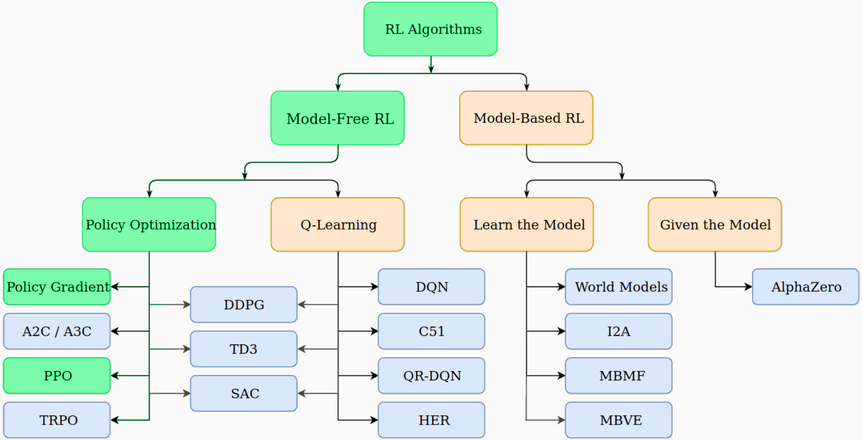


Figure 3: We will focus on this branch in what follows.

Policy optimization

PARAMETERIZED POLICY

We parameterize our policy π_θ by a vector $\theta \in \mathbb{R}^n$.

For continuous actions, it is common to use a *diagonal Gaussian policy*:

$$a \sim \pi_\theta(\cdot|s) \iff a = \mu_\theta(s) + \mathrm{diag}(\sigma_\theta(s))z,\ \ z \sim \mathcal{N}(0, I_m)$$

where μ_θ and σ_θ are neural networks mapping states to means and standard deviations.[2]

[2] In practice, σ often does not depend on s and we store $\log \sigma \in \mathbb{R}^m$ rather than $\sigma \in \mathbb{R}_+^m$ in θ.

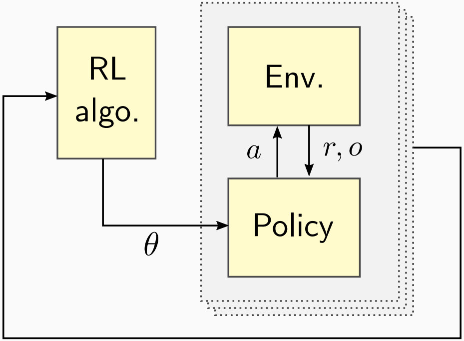
POLICY-BASED ALGORITHMS

At each iteration k:

- **Collect** a *batch* of episodes $\mathcal{D}_k = \{\tau\}$
- **Update** policy parameters

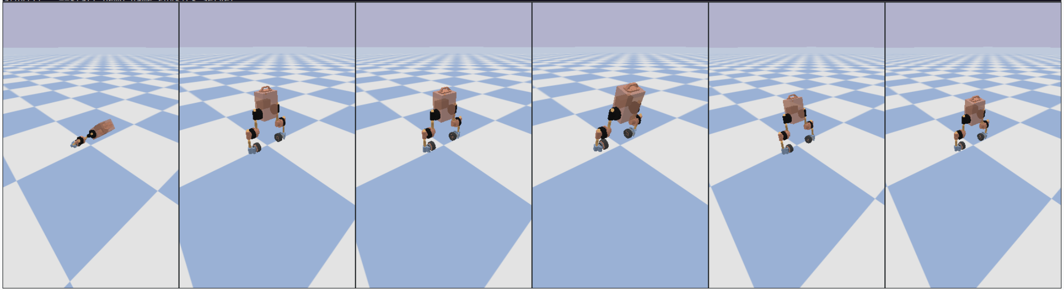
$$\theta_{k+1} = \text{update}(\theta_k, \mathcal{D}_k)$$

to get a new policy $\pi_{\theta_{k+1}}$



Episodes $\mathcal{D} = \{\tau_i\}$

```
~/src/upkie > playground > ./tools/bazel run //agents/ppo_balancer:train -- --nb-envs 6 --show
INFO: Analyzed target //agents/ppo_balancer:train (108 packages loaded, 17832 targets configured).
INFO: Found 1 target...
Target //agents/ppo_balancer:train up-to-date:
  bazel-bin/agents/ppo_balancer/train
INFO: Elapsed time: 4.220s, Critical Path: 0.16s
INFO: 1 process: 1 internal.
INFO: Build completed successfully, 1 total action
INFO: Running command line: bazel-bin/agents/ppo_balancer/train --nb-envs 6 --show
[2023-11-14 11:31:04,519] [info] Logging training data in /home/scaron/src/upkie/training/2023-11-14 (train.py:365)
[2023-11-14 11:31:04,519] [info] To track in TensorBoard, run `tensorboard --logdir /home/scaron/src/upkie/training/2023-11-14` (train.py:366)
[2023-11-14 11:31:04,524] [info] New policy name is "marshiest" (train.py:236)
[2023-11-14 11:31:04,524] [info] Training data will be logged to /home/scaron/src/upkie/training/2023-11-14/marshiest_1 (train.py:237)
[2023-11-14 11:31:04,550] [info] Waiting for spine /monogamous to start (trial 1 / 10)... (spine_interface.py:46)
[2023-11-14 11:31:04,552] [info] Waiting for spine /rundown to start (trial 1 / 10)... (spine_interface.py:46)
[2023-11-14 11:31:04.554] [info] Command line: shm_name = /monogamous
[2023-11-14 11:31:04.554] [info] Command line: nb_substeps = 5
[2023-11-14 11:31:04.554] [info] Command line: spine_frequency = 1000 Hz
[2023-11-14 11:31:04.554] [warning] [Joystick] Observer disabled: no joystick found at /dev/input/js0
startThreads creating 1 threads.
starting thread 0
started thread 0
argc=2
argv[0] = --unused
argv[1] = --start_demo_name=Physics_Server
```



The goal of RL is to find a policy that maximizes the expected return. In terms of θ:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$$

In policy optimization, we seek an optimum by gradient ascent:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta_k)$$

The gradient $\nabla_\theta J$ with respect to policy parameters θ is called the *policy gradient*.

Policy gradient theorem

The policy gradient can be computed from returns and the log-policy gradient $\nabla_\theta \log \pi_\theta$ as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left(R(\tau) \sum_{s_t, a_t \in \tau} \nabla_\theta \log \pi_\theta(a_t | s_t) \right)$$

LHS: the graal. RHS: things we observe ($R(\tau)$) or know by design ($\nabla_\theta \log \pi_\theta$).

POLICY GRADIENT THEOREM: PROOF SKETCH

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}(R(\tau)) && \text{definition} \\
&= \nabla_\theta \int_\tau R(\tau)\mathbb{P}(\tau|\theta)\mathrm{d}\tau && \text{expectation as integral} \\
&= \int_\tau R(\tau)\nabla_\theta \mathbb{P}(\tau|\theta)\mathrm{d}\tau && \text{Leibniz integral rule} \\
&= \int_\tau R(\tau)\mathbb{P}(\tau|\theta)\nabla_\theta \log \mathbb{P}(\tau|\theta)\mathrm{d}\tau && \text{log-derivative trick} \\
&= \int_\tau R(\tau)\sum_{s_t,a_t\in\tau} \nabla_\theta \log \pi_\theta(a_t|s_t)\mathbb{P}(\tau|\theta)\mathrm{d}\tau && \text{expand } \mathbb{P}(\tau|\theta) \text{ as product} \\
&= \mathbb{E}_{\tau\sim\pi_\theta}\left(R(\tau)\sum_{s_t,a_t\in\tau} \nabla_\theta \log \pi_\theta(a_t|s_t)\right) && \text{integral as expectation}
\end{aligned}
$$

With a diagonal Gaussian policy $\mu_\theta(s), \sigma_\theta$:

$$\pi_\theta(a|s) = \prod_{i=1}^{\dim(a)} \frac{1}{\sqrt{2\pi\sigma_{\theta,i}^2}} \exp\left(\frac{-(a_i - \mu_{\theta,i}(s))^2}{\sigma_{\theta,i}^2}\right)$$

$$\log \pi_\theta(a|s) = -\frac{1}{2} \sum_{i=1}^{\dim(a)} \left[\frac{(a_i - \mu_{\theta,i}(s))^2}{\sigma_{\theta,i}^2} + 2\log\sigma_{\theta,i} + \log 2\pi\right]$$

$$\nabla_\theta \log \pi_\theta(a|s) = \sum_{i=1}^{\dim(a)} \left[\frac{a_i - \mu_{\theta,i}(s)}{\sigma_{\theta,i}^2} \nabla_\theta \mu_{\theta,i}(s) + \left(\frac{(a_i - \mu_{\theta,i}(s))^2}{\sigma_{\theta,i}^3} - \frac{1}{\sigma_{\theta,i}}\right) \nabla_\theta \sigma_{\theta,i}\right]$$

where $s \mapsto \mu_\theta(s)$ is typically a neural network, from which we get $\nabla_\theta \mu_\theta(s)$.

REINFORCE (1/2)

REINFORCE algorithm [SB18]

Data: initial policy parameters θ_0, learning rate α

Initialize policy parameters θ (e.g. to 0);

for $k = 0, 1, 2, \ldots$ **do**

  Roll out an episode $\tau = (o_0, a_0, \ldots, o_N, a_N)$ following π_{θ_k};

  for *each step* $t \in \tau$ **do**

    $R \leftarrow \sum_{t'=t+1}^{N} \gamma^{t'-t-1} r_{t'}$;

    $\theta \leftarrow \theta + \alpha \gamma^t R \nabla_\theta \log \pi_\theta(a_t | s_t)$

  end

end

REINFORCE (2/2)

Gradient ascent:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta_k)$$

From the policy gradient theorem, this is equivalent to:

$$\theta_{k+1} = \theta_k + \alpha \mathbb{E}_{\tau \sim \pi_\theta} \left(R(\tau) \sum_{s_t, a_t \in \tau} \nabla_\theta \log \pi_\theta(a_t | s_t) \right)$$

REINFORCE drops the expectation:

$$\theta_{k+1} = \theta_k + \alpha R(\tau_k) \sum_{s_t, a_t \in \tau_k} \nabla_\theta \log \pi_\theta(a_t | s_t)$$

Vanilla policy gradient [Ach18]

Data: initial policy parameters θ_0, initial value function parameters ϕ_0, learning rate α

for $k = 0, 1, 2, \ldots$ **do**

Collect episodes $\mathcal{D}_k = \{\tau_i\}$ by running $\pi_\theta = \pi(\theta_k)$;

Compute returns $\hat{R}_t$ and advantage estimates $\hat{A}_t$ based on V_{ϕ_k};

Estimate the policy gradient as

$$\hat{g}_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_k} \hat{A}_t$$

Update policy parameters by *e.g.* gradient ascent, $\theta_{k+1} = \theta_k + \alpha \hat{g}_k$;

Fit value function by regression on mean-square error:

$$\phi_{k+1} = \arg\min_\phi \frac{1}{T|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(\hat{R}_t - V_\phi(s_t)\right)^2$$

end

Proximal policy optimization [Sch+17]

Data: initial policy parameters θ_0, initial value function parameters ϕ_0

for $k = 0, 1, 2, \dots$ **do**

Collect episodes $\mathcal{D}_k = \{\tau_i\}$ by running $\pi_\theta = \pi(\theta_k)$;

Compute returns $\hat{R}_t$ and advantage estimates $\hat{A}_t$ based on V_{ϕ_k};

Clipping: Update policy parameters by maximizing the clipping objective:

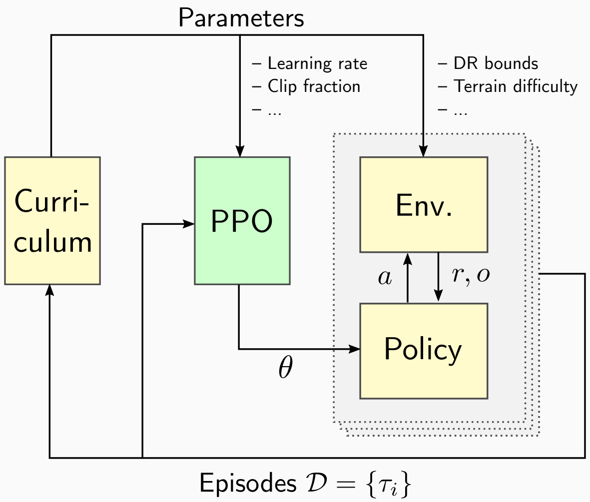
$$\theta_{k+1} = \arg\max_\theta \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), \mathrm{clip}(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right)$$

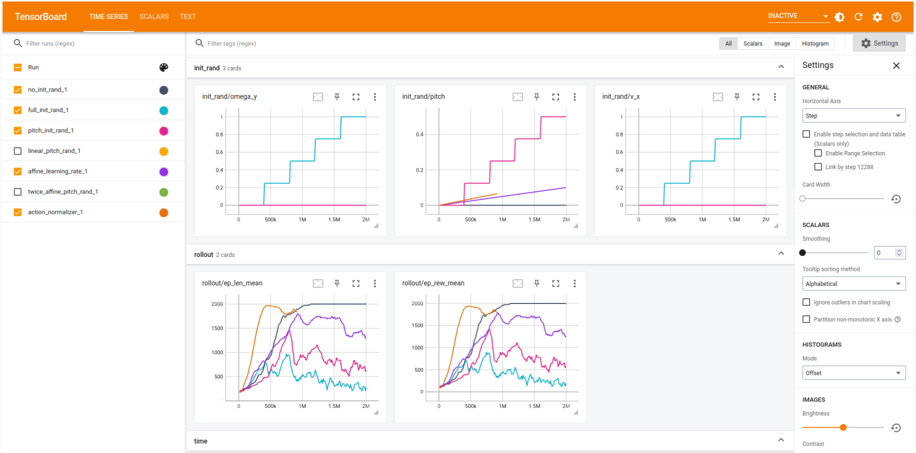
where $\mathrm{clip}(\epsilon, A) = (1+\epsilon)A$ if $A \geq 0$ else $(1-\epsilon)A$

Fit value function by regression on mean-square error:

$$\phi_{k+1} = \arg\min_\phi \frac{1}{T|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(\hat{R}_t - V_\phi(s_t) \right)^2$$

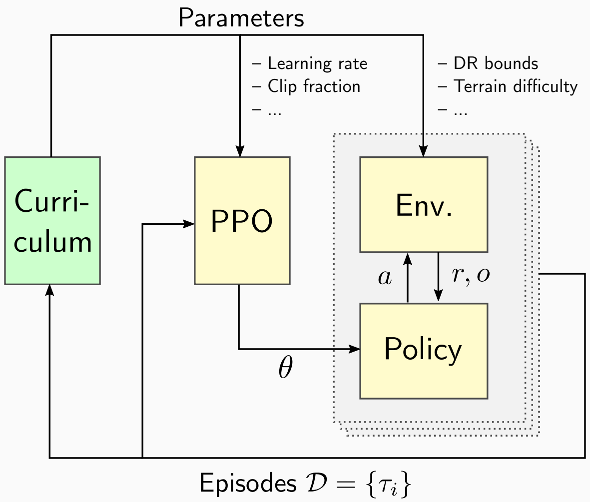
end


Parameters
– Learning rate
– Clip fraction
– ...
– DR bounds
– Terrain difficulty
– ...
Curri-
culum
PPO
Env.
Policy
a
r, o
θ
Episodes $\mathcal{D} = \{\tau_i\}$



Monitor the average return `ep_rew_mean` and length `ep_rew_len` of episodes.

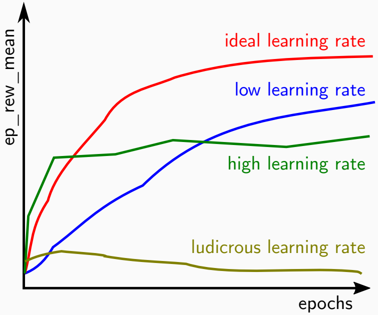
If training goes well, both eventually plateau at their maximum values.


Parameters
– Learning rate
– Clip fraction
– ...
– DR bounds
– Terrain difficulty
– ...
Curri-
culum
PPO
Env.
Policy
a
r, o
θ
Episodes $\mathcal{D} = \{\tau_i\}$

OPTIMIZER PARAMETERS: STEPS, EPOCHS, MINI-BATCHING

The optimizer behind PPO, usually Adam [KB14], comes with parameters:

- `learning_rate`: step size parameter, typically decreasing with a linear schedule.
- `n_epochs`: number of uses of the rollout buffer while optimizing the surrogate loss.
- `batch_size`: mini-batch size, same as in stochastic gradient descent.



Application to robotics

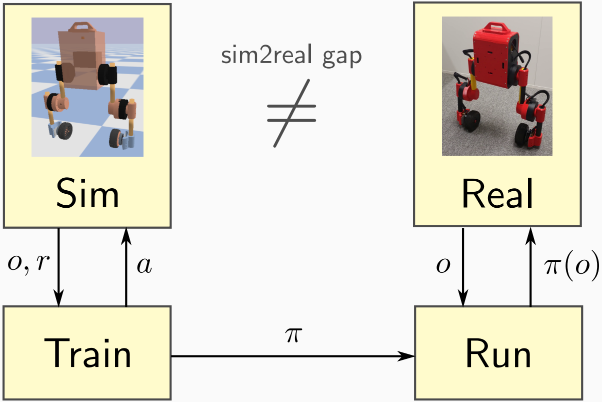


Figure 4: The "sim-to-real gap" is a metaphor for model mismatch.

General reinforcement learning techniques:

- Normalize observations and actions
- Augment observations with history
- Curriculum learning
- Reward shaping

For the sim-to-real gap in robotics:

- Domain randomization
- Data-based simulation
- Teacher-student distillation

OBSERVATION-ACTION NORMALIZATION

Unnormalized actions don't work well on actors with Gaussian policies:

- Bounds too large $\Rightarrow$ sampled actions cluster around zero.
- Bounds too small $\Rightarrow$ sampled actions saturate all the time, *bang-bang* behavior.

Good practice: bound observations/states, rescale actions to $[-1, 1]$.

AUGMENTING OBSERVATIONS WITH HISTORY

We assumed a Markovian system, but real systems have lag:

Definition

The *lag of a system* is the number of observations required to estimate its state.

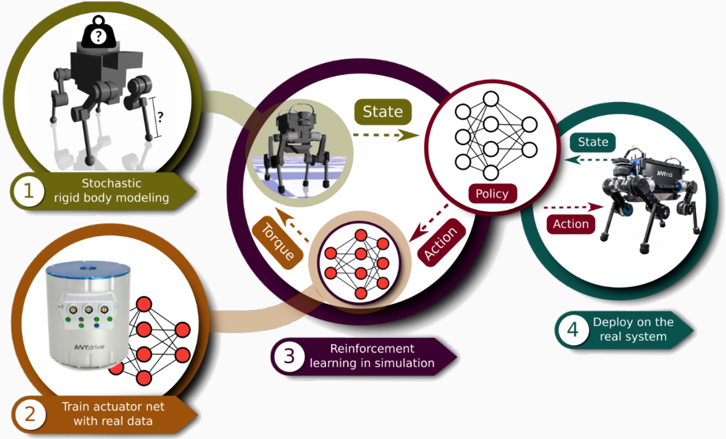
Counter-measure: augment observations with history to restore the Markov property.

Randomize selected environment parameters:

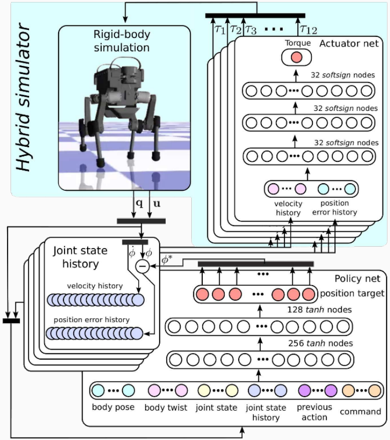
- **Robot geometry:** limb lengths, wheel diameters, ...
- **Inertias:** masses, mass distributions
- **Initial state:** $s_0 \sim \rho_0(\cdot)$
- **Actuation models:** delays, bandwidth, ...
- **Perturbations:** send $(1 \pm \epsilon)\tau$ torques...

There is a tradeoff to this:

- Pro: closer/may include real-robot distribution.
- Con: makes policies **more conservative**.

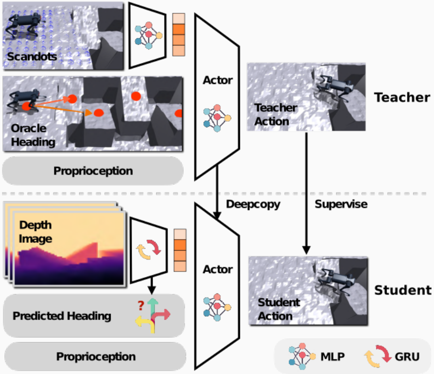


[3] Jemin Hwangbo et al. **"Learning agile and dynamic motor skills for legged robots".** In: *Science Robotics* 4.26 (2019).



[4]Jemin Hwangbo et al. **“Learning agile and dynamic motor skills for legged robots”.** In: *Science Robotics* 4.26 (2019).

- Train a **teacher policy** in simulation with privileged information
- Train a **student policy** in simulation with observations and teacher action

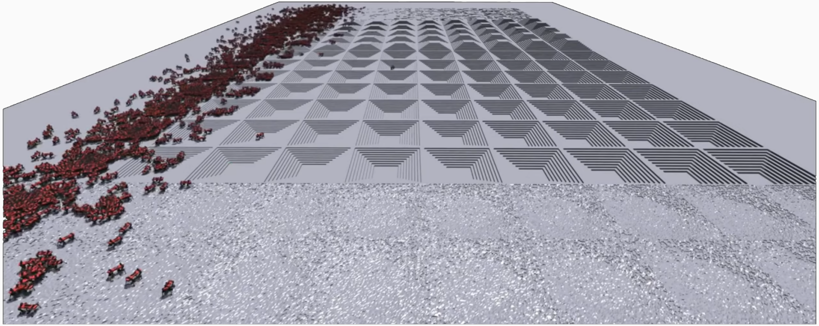


[5] Xuxin Cheng et al. **"Extreme parkour with legged robots".** In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 11443–11450.

Randomization and task difficulty vary based on policy performance.

Example: terrain curriculum for quadrupedal locomotion [Lee+20]:

easier → harder



Let r_e denote the reward associated with an error function e:

Motivation:

- Exponential: $r_e = \exp(-e^2)$

Penalization:

- Absolute value $r_e = -|e|$
- Squared value: $r_e = -e^2$

Making an RL pipeline work can lead to complex rewards, e.g. in [Lee+20]:

- Linear velocity tracking: $r_{lv} = \exp(-2.0(v_{pr} - 0.6)^2)$, or 1, or 0
- Angular velocity tracking: $r_{av} = \exp(-1.5(\omega_{pr} - 0.6)^2)$, or 1
- Base motion tracking: $r_b = \exp(-1.5 v_o^2) + \exp(-1.5 \|({}_{IB}^{B}\omega)_{xy}\|^2)$
- Foot clearance: $r_{fc} = \sum_{i \in I_{swing}} \mathbf{1}_{fclear}(i)/|I_{swing}|$
- Body-terrain collisions: $r_{bc} = -|I_{c,body} \setminus I_{c,foot}|$
- Foot acceleration smoothness: $r_s = -\|(r_{f,d})_t - 2(r_{f,d})_{t-1} + (r_{f,d})_{t-2}\|$
- Torque penalty: $r_\tau = -\sum_i |\tau_i|$

Final reward: $r = 0.05 r_{lv} + 0.05 r_{av} + 0.04 r_b + 0.01 r_{fc} + 0.02 r_{bc} + 0.025 r_s + 2 \cdot 10^{-5} r_\tau$

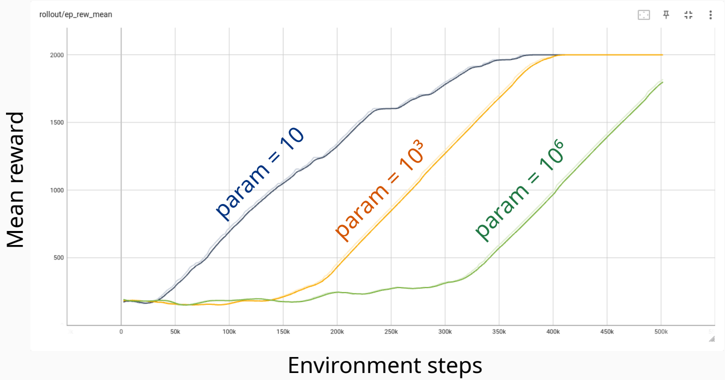


Figure 5: We may be observing the effect of our parameter.

Figure 5: Or we may be observing the variance of the training process.

What did we see?

WE SAW

Introduction to policy optimization:

- Partially-observable Markov decision process (POMDP)
- The goal of reinforcement learning
- Model, policy and value function
- Policy optimization: REINFORCE, policy gradient, PPO

Application to robotics:

- Sim-to-real gap: domain randomization, hybrid simulation
- Techniques: curriculum, distillation, history, “RewArt”

RL is not magic: great results, possibly going to great lengths!

Thank you for your attention![6]

[6]These slides have received feedback from E. Chane-Sane, T. Flayols, N. Perrin-Gilbert, R. P. Singh, P. Souères, and the 2023-2025 classes at ENS–PSL, Mines Paris–PSL and Master MVA. Thank you all!

Bibliography

References I

[Ach18] Josh Achiam. ***Spinning Up in Deep Reinforcement Learning.*** `https://spinningup.openai.com/`. 2018.

[ACN10] Pieter Abbeel, Adam Coates, and Andrew Y Ng. **“Autonomous helicopter aerobatics through apprenticeship learning”.** In: *The International Journal of Robotics Research* 29.13 (2010), pp. 1608–1639.

[And+20] OpenAI: Marcin Andrychowicz et al. **“Learning dexterous in-hand manipulation”.** In: *The International Journal of Robotics Research* 39.1 (2020), pp. 3–20.

[AS97] Christopher G Atkeson and Stefan Schaal. **“Robot learning from demonstration”.** In: *ICML*. Vol. 97. 1997, pp. 12–20.

[Che+24] Xuxin Cheng et al. **“Extreme parkour with legged robots”.** In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 11443–11450.

[Hwa+19] Jemin Hwangbo et al. **“Learning agile and dynamic motor skills for legged robots”.** In: *Science Robotics* 4.26 (2019).

[KB14] Diederik P Kingma and Jimmy Ba. **“Adam: A method for stochastic optimization”.** In: *arXiv preprint arXiv:1412.6980* (2014).

[Lee+20] Joonho Lee et al. **“Learning quadrupedal locomotion over challenging terrain”.** In: *Science robotics* 5.47 (2020).

REFERENCES II

[Lia+24] Qiayuan Liao et al. **“Berkeley humanoid: A research platform for learning-based control”.** In: *arXiv preprint arXiv:2407.21781* (2024).

[SB18] Richard S Sutton and Andrew G Barto. ***Reinforcement learning: An introduction.*** MIT press, 2018.

[Sch+17] John Schulman et al. **“Proximal policy optimization algorithms”.** In: *arXiv preprint arXiv:1707.06347* (2017).

Bonus slides on PPO

INTUITION BEHIND CLIPPING IN PPO

When the advantage is positive:

$$L(s, a, \theta_k, \theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, (1+\epsilon)\right) A^{\pi_{\theta_k}}(s, a)$$

The objective increases if the action becomes more likely $\pi_\theta(a|s) > \pi_{\theta_k}(a|s)$, but no extra benefit as soon as $\pi_\theta(a|s) > (1+\epsilon)\pi_{\theta_k}(a|s)$.

When the advantage is negative: *idem mutatis mutandis*.

PPO LOSS FUNCTION

Surrogate loss of PPO

```
loss = policy_gradient_loss + ent_coef * entropy_loss + vf_coef * value_loss
```

- `policy_gradient_loss`: regular loss resulting from episode returns.
- `entropy_loss`: negative of the average policy entropy. It should increase to zero over training as the policy becomes more deterministic.
- `value_loss`: value function estimation loss, *i.e.* error between the output of the function estimator and Monte-Carlo or TD(GAE lambda) estimates.

PPO HYPERPARAMETERS

The PPO implementation in Stable Baselines3 has > 25 parameters, including:

- `clip_range` : clipping factor in policy loss.
- `ent_coef` : weight of entropy term in the surrogate loss.
- `gae_lambda` : parameter of Generalized Advantage Estimation.
- `net_arch_pi` : policy network architecture.
- `net_arch_vf` : value network architecture.
- `normalize_advantage` : use advantage normalization?
- `vf_coef` : weight of value-function term in the surrogate loss.

Some metrics indicate whether training is going well:

- `approx_kl` : approximate KL divergence between the old policy and the new one.
- `clip_fraction` : mean fraction of policy ratios that were clipped.
- `clip_range` : value of the clipping factor for policy ratios.
- `explained_variance` : ≈ 1 when the value function is a good predictor for returns.