Distilling Policy Distillation

W. M. Czarnecki, R. Pascanu, S. Osindero,
S. Jayakumar, G. Swirszcz, M. Jaderberg

PMLR, 2019

Presentation by Anthony Coache
STA4273 Minimizing Expectations ⋆ March 18, 2021

Motivations

Distillation

Knowledge transfer; learn an optimal behavior from expert (e.g. a pre-trained model or a human) interactions with an environment.

- Speed up the learning process
- Achieve model compression

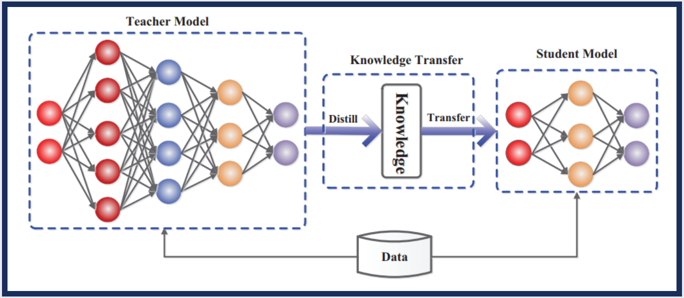


Figure 1 from (Gou et al., 2020)

Markov Decision Process

MDP

A **Markov decision process** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \pi, \gamma)$, where

- $\mathcal{S}$ – Finite state space
- $\mathcal{A}$ – Finite action space
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ – State-action dependent reward function
- $P : \mathcal{S} \times \mathcal{A} \to \Delta^{|\mathcal{S}|}$ – Transition probability distribution
- $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ – Policy
- $\gamma \in [0, 1]$ – Discount factor

- One trajectory from $\mathcal{M}$ is denoted by

$$\tau = (s_1, a_1, r_1, \ldots, s_{|\tau|}, a_{|\tau|}, r_{|\tau|}).$$

- Typical goal of reinforcement learning: find a policy

$$\pi^* = \operatorname{argmax}_\pi \left\{ \mathbb{E}_\pi \left[\sum_{t=1}^{|\tau|} \gamma^{t-1} r_t \right] \right\}.$$

General Problem

Policy Distillation

Goal: extract knowledge from a teacher policy, and transfer it to a student policy using trajectories sampled from interactions between a control policy and the environment.

- π – Teacher policy
- π_θ – Student policy
- q_θ – Control policy

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

General Problem

Policy Distillation

Goal: extract knowledge from a teacher policy, and transfer it to a student policy using trajectories sampled from interactions between a control policy and the environment.

- π – Teacher policy
- π_θ – Student policy
- q_θ – Control policy

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

General Problem

Policy Distillation

Goal: extract knowledge from a teacher policy, and transfer it to a student policy using trajectories sampled from interactions between a control policy and the environment.

- π – Teacher policy
- π_θ – Student policy
- q_θ – Control policy

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

Update Rules

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

- q_θ
  - Control policy, i.e. measure of state space exploration
  - Teacher $q_\theta = \pi$; Student $q_\theta = \pi_\theta$
  - Could also be a uniform distribution or a mixture of policies
- $\widehat{R}_t = \sum_{i=t}^{|\tau|} \widehat{r}_i = \sum_{i=t}^{|\tau|} \widehat{r}(\pi_\theta, V_{\pi_\theta}, s_i, a_i, s_{i+1}, a_{i+1}, r_i)$
  - Reward term, i.e. long-term alignment
  - $\widehat{r}_i = \log(\pi(a_i|s_i))$
  - $\widehat{r}_i = r_i + V_\pi(s_{i+1}) - V_{\pi_\theta}(s_i)$
- $\ell_t = \ell(\pi_\theta, V_{\pi_\theta}, s_t)$
  - Loss term, i.e. policy alignment with the teacher
  - Cross-entropy $\ell_t = -\mathbb{E}_{a\sim\pi(s_t)}[\log(\pi_\theta(a|s_t))]$

Update Rules

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

- q_θ
  - Control policy, i.e. measure of state space exploration
  - Teacher $q_\theta = \pi$; Student $q_\theta = \pi_\theta$
  - Could also be a uniform distribution or a mixture of policies
- $\widehat{R}_t = \sum_{i=t}^{|\tau|} \widehat{r}_i = \sum_{i=t}^{|\tau|} \widehat{r}(\pi_\theta, V_{\pi_\theta}, s_i, a_i, s_{i+1}, a_{i+1}, r_i)$
  - Reward term, i.e. long-term alignment
  - $\widehat{r}_i = \log(\pi(a_i|s_i))$
  - $\widehat{r}_i = r_i + V_\pi(s_{i+1}) - V_{\pi_\theta}(s_i)$
- $\ell_t = \ell(\pi_\theta, V_{\pi_\theta}, s_t)$
  - Loss term, i.e. policy alignment with the teacher
  - Cross-entropy $\ell_t = -\mathbb{E}_{a\sim\pi(s_t)}[\log(\pi_\theta(a|s_t))]$

Update Rules

Update rules for the parameters of π_θ are proportional to

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

- q_θ
  - Control policy, i.e. measure of state space exploration
  - Teacher $q_\theta = \pi$; Student $q_\theta = \pi_\theta$
  - Could also be a uniform distribution or a mixture of policies
- $\widehat{R}_t = \sum_{i=t}^{|\tau|} \widehat{r}_i = \sum_{i=t}^{|\tau|} \widehat{r}(\pi_\theta, V_{\pi_\theta}, s_i, a_i, s_{i+1}, a_{i+1}, r_i)$
  - Reward term, i.e. long-term alignment
  - $\widehat{r}_i = \log(\pi(a_i|s_i))$
  - $\widehat{r}_i = r_i + V_\pi(s_{i+1}) - V_{\pi_\theta}(s_i)$
- $\ell_t = \ell(\pi_\theta, V_{\pi_\theta}, s_t)$
  - Loss term, i.e. policy alignment with the teacher
  - Cross-entropy $\ell_t = -\mathbb{E}_{a\sim\pi(s_t)}[\log(\pi_\theta(a|s_t))]$

Offline vs Online RL

Offline RL:

- Learning without interacting with the environment, only observing transitions from some policies
- Turning (large) datasets of transitions into decision making engines

Online RL:

- Learning while interacting with the environment, working with data as it is made available
- Improving policies with the latest collected experience

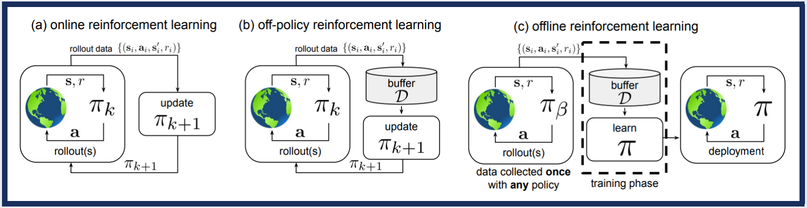


Figure 1 from (Levine et al., 2020)

Offline Policy Distillation

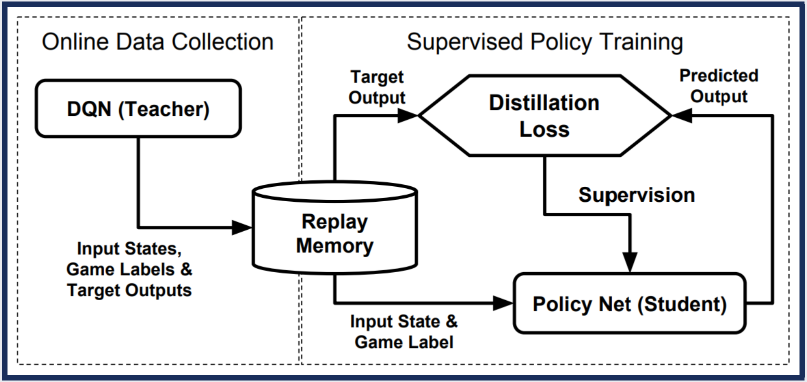


Figure 2 (a) from (Rusu et al., 2015)

Online Policy Distillation

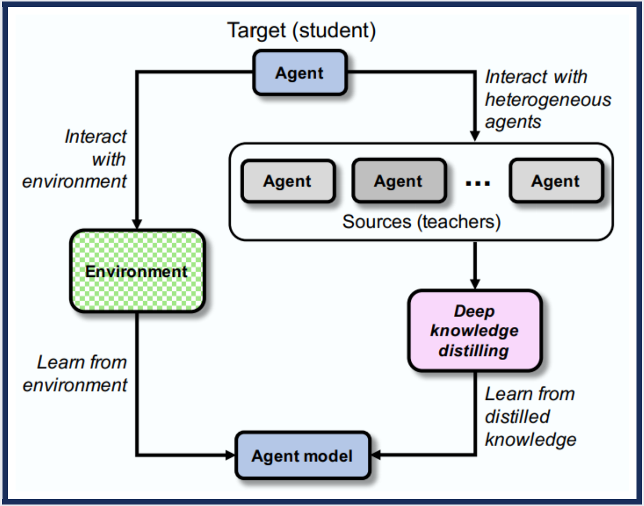


Figure 1 from (Lin et al., 2017)

Contributions

- Exploration of multiple policy distillation approaches
  - Naive student distillation does not form a gradient vector field
  - That property can be recovered adding an additional reward term
  - Student distillation has convergence guarantees in simple tabular cases
- Proposition of new algorithm variations
  - Modifications of known distillation techniques to address some issues
  - Methods using the value function
  - Methods splitting the loss between the reward and loss terms
- Empirical evaluation of different distillation techniques
  - Performance comparison with different control policies
  - Performance comparison with different update rules
  - Policy distillation method selection diagram

Contributions

- Exploration of multiple policy distillation approaches
  - Naive student distillation does not form a gradient vector field
  - That property can be recovered adding an additional reward term
  - Student distillation has convergence guarantees in simple tabular cases
- Proposition of new algorithm variations
  - Modifications of known distillation techniques to address some issues
  - Methods using the value function
  - Methods splitting the loss between the reward and loss terms
- Empirical evaluation of different distillation techniques
  - Performance comparison with different control policies
  - Performance comparison with different update rules
  - Policy distillation method selection diagram

Contributions

- Exploration of multiple policy distillation approaches
  - Naive student distillation does not form a gradient vector field
  - That property can be recovered adding an additional reward term
  - Student distillation has convergence guarantees in simple tabular cases
- Proposition of new algorithm variations
  - Modifications of known distillation techniques to address some issues
  - Methods using the value function
  - Methods splitting the loss between the reward and loss terms
- Empirical evaluation of different distillation techniques
  - Performance comparison with different control policies
  - Performance comparison with different update rules
  - Policy distillation method selection diagram

Update Rules with the Student as Control

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

Gradient is under an expectation wrt. the same θ it operates upon

Theorem 1

If $g(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t]$ is differentiable and there exists $\alpha_\tau \in \mathbb{R}$ such that $\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t = \alpha_\tau \nabla_\theta \pi_\theta(\tau)$, then $g(\theta)$ is not a gradient vector field of any function.

Unclear if distillation with student-generated trajectories will converge...

Theorem 2

The gradient vector field property can be recovered adding an appropriate extra reward term.

Update Rules with the Student as Control

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

Gradient is under an expectation wrt. the same θ it operates upon

Theorem 1

If $g(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t]$ is differentiable and there exists $\alpha_\tau \in \mathbb{R}$ such that $\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t = \alpha_\tau \nabla_\theta \pi_\theta(\tau)$, then $g(\theta)$ is not a gradient vector field of any function.

Unclear if distillation with student-generated trajectories will converge...

Theorem 2

The gradient vector field property can be recovered adding an appropriate extra reward term.

Update Rules with the Student as Control

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t))\widehat{R}_t + \nabla_\theta \ell_t\right].$$

Gradient is under an expectation wrt. the same θ it operates upon

Theorem 1

If $g(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t]$ is differentiable and there exists $\alpha_\tau \in \mathbb{R}$ such that $\nabla_\theta \sum_{t=1}^{|\tau|} \ell_t = \alpha_\tau \nabla_\theta \pi_\theta(\tau)$, then $g(\theta)$ is not a gradient vector field of any function.

Unclear if distillation with student-generated trajectories will converge...

Theorem 2

The gradient vector field property can be recovered adding an appropriate extra reward term.

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t)||\pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta)\right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau)\ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau)\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau))\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau))\ell(\tau, \theta)\right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta)\right].
\end{aligned}$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1})||\pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t \right].$$

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t)||\pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$
\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta)\right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau) \ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau)\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau))\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau)) \ell(\tau, \theta)\right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta)\right].
\end{aligned}
$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1})||\pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$
\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t\right].
$$

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t)||\pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta)\right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau) \ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau)\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau))\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau)) \ell(\tau, \theta)\right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta)\right].
\end{aligned}$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1})||\pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t \right].$$

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t)||\pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$
\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta)\right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau)\ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau)\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau))\right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta)\right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau))\ell(\tau, \theta)\right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta)\right].
\end{aligned}
$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1})||\pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$
\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t \right].
$$

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t)||\pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta) \right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau) \ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau) \right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta) \right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau)) \right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta) \right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau)) \ell(\tau, \theta) \right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta) \right].
\end{aligned}$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1})||\pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t|s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t \right].$$

Gradient Vector Field

Let $\ell(\tau, \theta) = \sum_{t=1}^{|\tau|} \ell(\pi(s_t) || \pi_\theta(s_t))$, for a certain loss, e.g. cross-entropy. Then computing the gradient of this loss function yields

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \, \mathbb{E}_{\pi_\theta(\tau)} \left[\ell(\tau, \theta) \right] \\
&= \nabla_\theta \int_\tau \pi_\theta(\tau) \ell(\tau, \theta) d\tau \\
&= \int_\tau \left(\nabla_\theta \pi_\theta(\tau) \right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta) \right) d\tau \\
&= \int_\tau \left(\pi_\theta(\tau) \nabla_\theta \log(\pi_\theta(\tau)) \right) \ell(\tau, \theta) + \pi_\theta(\tau) \left(\nabla_\theta \ell(\tau, \theta) \right) d\tau \\
&= \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \log(\pi_\theta(\tau)) \ell(\tau, \theta) \right] + \mathbb{E}_{\pi_\theta(\tau)} \left[\nabla_\theta \ell(\tau, \theta) \right].
\end{aligned}$$

Setting $\widehat{r}_i = -\ell(\pi(s_{i+1}) || \pi_\theta(s_{i+1}))$, we recover the gradient vector field property

$$\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{|\tau|} -\nabla_\theta \log(\pi_\theta(a_t | s_t)) \sum_{i=t}^{|\tau|} \widehat{r}_i + \nabla_\theta \ell_t \right].$$

Experimental Setting

- Thousand randomly sampled 20×20 grid worlds MDPs with rewards and terminal states

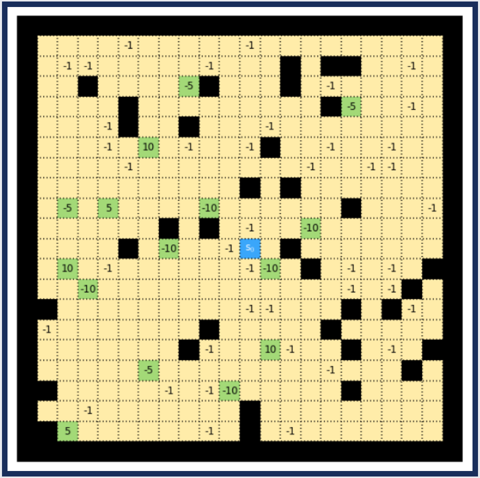


Figure 7 from (Czarnecki et al., 2019)

Experimental Setting

- Teacher trained with standard Q-learning and ϵ-greedy policies

$$Q(a_t, s_t) = (1 - \lambda) Q(a_t, s_t) + \lambda \left(r_t + \gamma \max_a Q(a, s_{t+1}) \right)$$

- Distillation with different control policies, 30,000 optimization steps
  - Teacher, student or uniform driven distillation

- Minimizing the per-step cross-entropy:

$$\mathbb{E}_{q_\theta} \left[\sum_{t=1}^{|\tau|} \nabla_\theta H^\times(\pi(s_t) || \pi_\theta(s_t)) \right].$$

$$H^\times(p_1(s) || p_2(s)) = -\mathbb{E}_{a \sim p_1(s)} [\log p_2(a|s)]$$

Experimental Setting

- Teacher trained with standard Q-learning and ϵ-greedy policies

$$Q(a_t, s_t) = (1-\lambda)Q(a_t, s_t) + \lambda\left(r_t + \gamma \max_a Q(a, s_{t+1})\right)$$

- Distillation with different control policies, 30,000 optimization steps
  - Teacher, student or uniform driven distillation

- Minimizing the per-step cross-entropy:

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} \nabla_\theta H^\times(\pi(s_t)||\pi_\theta(s_t))\right].$$

$$H^\times(p_1(s)||p_2(s)) = -\mathbb{E}_{a\sim p_1(s)}[\log p_2(a|s)]$$

Experimental Setting

- Teacher trained with standard Q-learning and ϵ-greedy policies

$$Q(a_t, s_t) = (1-\lambda)Q(a_t, s_t) + \lambda\left(r_t + \gamma \max_a Q(a, s_{t+1})\right)$$

- Distillation with different control policies, 30,000 optimization steps
  - Teacher, student or uniform driven distillation

- Minimizing the per-step cross-entropy:

$$\mathbb{E}_{q_\theta}\left[\sum_{t=1}^{|\tau|} \nabla_\theta H^\times(\pi(s_t)||\pi_\theta(s_t))\right].$$

$$H^\times(p_1(s)||p_2(s)) = -\mathbb{E}_{a\sim p_1(s)}\left[\log p_2(a|s)\right]$$

Results of Experiments - Control Policy

- Student-driven distillation needs 3x less iterations than teacher-driven distillation to recover the full teacher performance
- Student-driven distillation explores more the state space; visits states that would be less visited with a teacher-driven distillation
- Student-driven distillation leads to less of a distribution-shift between the training phase and the testing deployment
- Student-driven distillation provides in general more robust policies than teacher-driven distillation

Results of Experiments - Control Policy

- Student-driven distillation needs 3x less iterations than teacher-driven distillation to recover the full teacher performance
- Student-driven distillation explores more the state space; visits states that would be less visited with a teacher-driven distillation
- Student-driven distillation leads to less of a distribution-shift between the training phase and the testing deployment
- Student-driven distillation provides in general more robust policies than teacher-driven distillation

Results of Experiments - Control Policy

- Student-driven distillation needs 3x less iterations than teacher-driven distillation to recover the full teacher performance
- Student-driven distillation explores more the state space; visits states that would be less visited with a teacher-driven distillation
- Student-driven distillation leads to less of a distribution-shift between the training phase and the testing deployment
- Student-driven distillation provides in general more robust policies than teacher-driven distillation

Results of Experiments - Control Policy

Notebook experiments!

Results of Experiments - Control Policy

- KL over various sampling distributions and returns when following a teacher policy

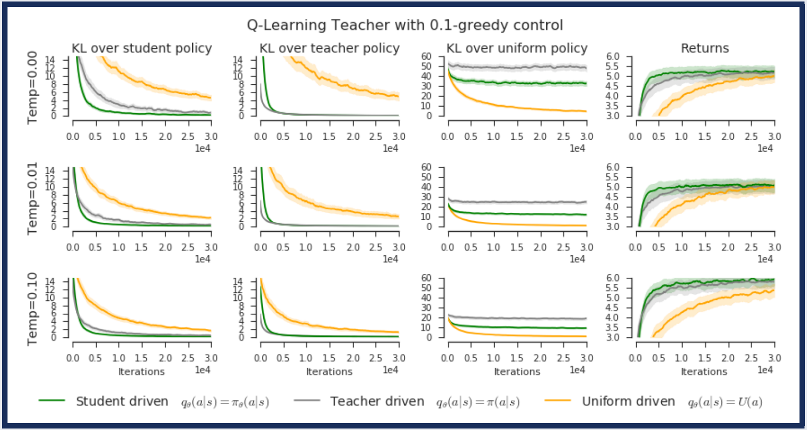


Figure 3 from (Czarnecki et al., 2019)

Results of Experiments - Loss Term

- Common approaches with $H^{\times}(p_1||p_2) = -\mathbb{E}_{a\sim p_1(s)}[\log p_2(a|s)]$:
  - Using $H^{\times}(\pi||\pi_\theta)$: trying to replicate the π
  - Using $H^{\times}(\pi_\theta||\pi)$: trying to find the most probable action of π
- Since $H^{\times}(\pi||\pi_\theta) = H(\pi) + \mathrm{KL}(\pi||\pi_\theta)$, the minimum is given by π
- When optimizing $H^{\times}(\pi_\theta||\pi)$, the minimum is the dirac delta distribution of the most probable action a^* of π

$$\begin{aligned} H^{\times}(\pi_\theta||\pi) &= -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a|s)] \\ &> -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a^*|s)] \end{aligned}$$

- Better empirical results with $H^{\times}(\pi_\theta||\pi)$, i.e. when directly maximizing the probability of the student produced trajectories under the teacher policy

Results of Experiments - Loss Term

- Common approaches with $H^{\times}(p_1||p_2) = -\mathbb{E}_{a\sim p_1(s)}[\log p_2(a|s)]$:
  - Using $H^{\times}(\pi||\pi_\theta)$: trying to replicate the π
  - Using $H^{\times}(\pi_\theta||\pi)$: trying to find the most probable action of π
- Since $H^{\times}(\pi||\pi_\theta) = H(\pi) + \mathrm{KL}(\pi||\pi_\theta)$, the minimum is given by π
- When optimizing $H^{\times}(\pi_\theta||\pi)$, the minimum is the dirac delta distribution of the most probable action a^* of π

$$\begin{aligned} H^{\times}(\pi_\theta||\pi) &= -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a|s)] \\ &> -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a^*|s)] \end{aligned}$$

- Better empirical results with $H^{\times}(\pi_\theta||\pi)$, i.e. when directly maximizing the probability of the student produced trajectories under the teacher policy

Results of Experiments - Loss Term

- Common approaches with $H^{\times}(p_1||p_2) = -\mathbb{E}_{a\sim p_1(s)}[\log p_2(a|s)]$:
  - Using $H^{\times}(\pi||\pi_\theta)$: trying to replicate the π
  - Using $H^{\times}(\pi_\theta||\pi)$: trying to find the most probable action of π
- Since $H^{\times}(\pi||\pi_\theta) = H(\pi) + \mathrm{KL}(\pi||\pi_\theta)$, the minimum is given by π
- When optimizing $H^{\times}(\pi_\theta||\pi)$, the minimum is the dirac delta distribution of the most probable action a^* of π

$$\begin{aligned} H^{\times}(\pi_\theta||\pi) &= -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a|s)] \\ &> -\mathbb{E}_{a\sim\pi_\theta(s)}[\log \pi(a^*|s)] \end{aligned}$$

- Better empirical results with $H^{\times}(\pi_\theta||\pi)$, i.e. when directly maximizing the probability of the student produced trajectories under the teacher policy

Results of Experiments - Loss Term

- Common approaches with $H^{\times}(p_1||p_2) = -\mathbb{E}_{a\sim p_1(s)}[\log p_2(a|s)]$:
  - Using $H^{\times}(\pi||\pi_\theta)$: trying to replicate the π
  - Using $H^{\times}(\pi_\theta||\pi)$: trying to find the most probable action of π
- Since $H^{\times}(\pi||\pi_\theta) = H(\pi) + \mathrm{KL}(\pi||\pi_\theta)$, the minimum is given by π
- When optimizing $H^{\times}(\pi_\theta||\pi)$, the minimum is the dirac delta distribution of the most probable action a^* of π

$$\begin{aligned} H^{\times}(\pi_\theta||\pi) &= -\mathbb{E}_{a\sim\pi_\theta(s)}[\log\pi(a|s)] \\ &> -\mathbb{E}_{a\sim\pi_\theta(s)}[\log\pi(a^*|s)] \end{aligned}$$

- Better empirical results with $H^{\times}(\pi_\theta||\pi)$, i.e. when directly maximizing the probability of the student produced trajectories under the teacher policy

Proposed Distillation Methods

- Best empirical results obtained with their expected entropy regularized distillation algorithm
  - It creates a gradient vector field
  - It reduces the variance by splitting the entropy between the reward and loss term

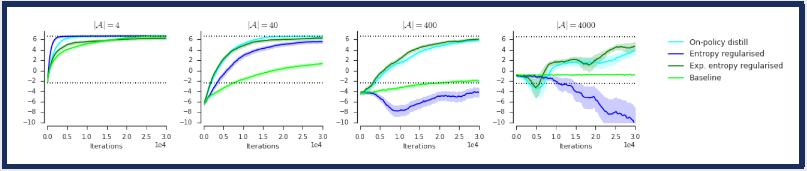


Figure 4 from (Czarnecki et al., 2019)

Proposed Distillation Methods

- Their teacher V reward distillation algorithm uses the value function $V_\pi(s) = \mathbb{E}_\pi[\sum_t r_t]$ in the distillation process
  - It can estimate how much we trust the teacher
  - It allows the student to learn with imperfect teachers

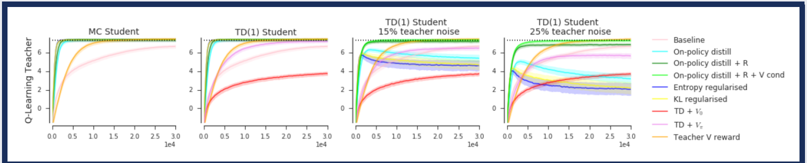


Figure 5 from (Czarnecki et al., 2019)

Policy Distillation Method Selection

Despite the multiple other factors affecting performance of policy distillation in practice, this provides a method suggestion based on different settings:

- Do we want convergence guarantees?
- Do we prefer improvement over speed?
- Is the teacher relatively strong?

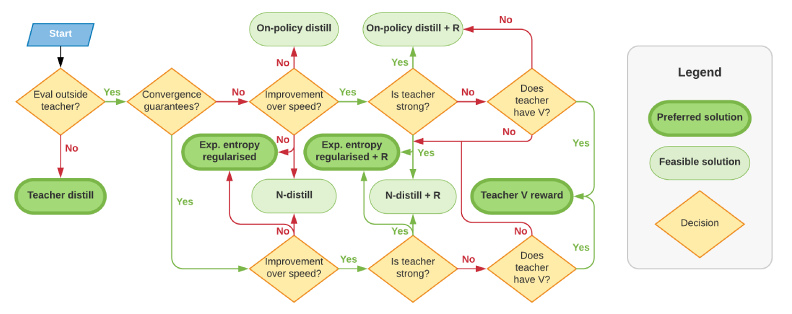


Figure 1 from (Czarnecki et al., 2019)

Summary and Limitations

- Student-driven policy distillation provides better empirical results over teacher-driven distillation
- Their proposed expected entropy regularized and teacher V reward distillation algorithms combine benefits of various methods:
  - Creates a gradient vector field
  - Reduces the variance
  - Allows the agent to learn from imperfect teachers
- Their distillation method selection diagram gives a general rule of thumb when choosing the most suitable algorithms in practice
- Some open questions:
  - Same behaviors on real-world problems, e.g. continuous spaces?
  - Similar results using functions approximators, e.g. neural nets?
  - Convergence guarantees?
  - Infinite horizon problems instead of episodic?

Summary and Limitations

- Student-driven policy distillation provides better empirical results over teacher-driven distillation
- Their proposed expected entropy regularized and teacher V reward distillation algorithms combine benefits of various methods:
  - Creates a gradient vector field
  - Reduces the variance
  - Allows the agent to learn from imperfect teachers
- Their distillation method selection diagram gives a general rule of thumb when choosing the most suitable algorithms in practice
- Some open questions:
  - Same behaviors on real-world problems, e.g. continuous spaces?
  - Similar results using functions approximators, e.g. neural nets?
  - Convergence guarantees?
  - Infinite horizon problems instead of episodic?

Summary and Limitations

- Student-driven policy distillation provides better empirical results over teacher-driven distillation
- Their proposed expected entropy regularized and teacher V reward distillation algorithms combine benefits of various methods:
  - Creates a gradient vector field
  - Reduces the variance
  - Allows the agent to learn from imperfect teachers
- Their distillation method selection diagram gives a general rule of thumb when choosing the most suitable algorithms in practice
- Some open questions:
  - Same behaviors on real-world problems, e.g. continuous spaces?
  - Similar results using functions approximators, e.g. neural nets?
  - Convergence guarantees?
  - Infinite horizon problems instead of episodic?

Thank you!

Czarnecki, W. M., Pascanu, R., Osindero, S., Jayakumar, S., Swirszcz, G., and Jaderberg, M. (2019). Distilling policy distillation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1331–1340. PMLR.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2020). Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Lin, K., Wang, S., and Zhou, J. (2017). Collaborative deep reinforcement learning. *arXiv preprint arXiv:1702.05796*.

Ross, S., Gordon, G., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.

Rusu, A. A., Colmenarejo, S. G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., and Hadsell, R. (2015). Policy distillation. *arXiv preprint arXiv:1511.06295*.

Related Work

Distillation

Form of **knowledge** transfer; learn an optimal behavior from expert (e.g. a pre-trained model or a human) interactions with an environment.

- Compress the knowledge of an ensemble of large neural networks into a single model (Hinton et al., 2015)
- Train a new network based on a already trained RL agent (Rusu et al., 2015)
  - The smaller network achieves expert level performance
  - Procedure can be used for multi-task policy distillation
- Mimic an expert on a dataset of trajectories, imitation learning (Ross et al., 2011)
  - DAGGER algorithm
  - Similar to the Follow-The-Leader approach, best policy over the iterations

Related Work

Distillation

Form of **knowledge** transfer; learn an optimal behavior from expert (e.g. a pre-trained model or a human) interactions with an environment.

- Compress the knowledge of an ensemble of large neural networks into a single model (Hinton et al., 2015)
- Train a new network based on a already trained RL agent (Rusu et al., 2015)
  - The smaller network achieves expert level performance
  - Procedure can be used for multi-task policy distillation
- Mimic an expert on a dataset of trajectories, imitation learning (Ross et al., 2011)
  - DAGGER algorithm
  - Similar to the Follow-The-Leader approach, best policy over the iterations

Related Work

Distillation

Form of **knowledge** transfer; learn an optimal behavior from expert (e.g. a pre-trained model or a human) interactions with an environment.

- Compress the knowledge of an ensemble of large neural networks into a single model (Hinton et al., 2015)
- Train a new network based on a already trained RL agent (Rusu et al., 2015)
  - The smaller network achieves expert level performance
  - Procedure can be used for multi-task policy distillation
- Mimic an expert on a dataset of trajectories, imitation learning (Ross et al., 2011)
  - DAGGER algorithm
  - Similar to the Follow-The-Leader approach, best policy over the iterations

Related Work

Distillation

Form of **knowledge** transfer; learn an optimal behavior from expert (e.g. a pre-trained model or a human) interactions with an environment.

- Compress the knowledge of an ensemble of large neural networks into a single model (Hinton et al., 2015)
- Train a new network based on a already trained RL agent (Rusu et al., 2015)
  - The smaller network achieves expert level performance
  - Procedure can be used for multi-task policy distillation
- Mimic an expert on a dataset of trajectories, imitation learning (Ross et al., 2011)
  - DAGGER algorithm
  - Similar to the Follow-The-Leader approach, best policy over the iterations

Policy Distillation Algorithms

name	q_θ	$\ell(\pi_\theta, V_{\pi_\theta}, \tau_t)$	$\widehat{r}_i$	is ∇?	Loss
Teacher distill	π	$\mathrm{H}^\times(\pi(\tau_t)\|\pi_\theta(\tau_t))$	0	yes [1]	$\mathbb{E}_\pi[\sum_t \mathrm{H}^\times(\pi(\tau_t)\|\pi_\theta(\tau_t))]$
On-policy distill	π_θ	$\mathrm{H}^\times(\pi(\tau_t)\|\pi_\theta(\tau_t))$	0	no*	does not exist*
Entropy regularised	π_θ	0	$\log\pi(a_i\|\tau_i)$	yes [4]	$\mathbb{E}_{\pi_\theta}[\sum_t -\log\pi(a_t\|\tau_t)]$
N-distill	π_θ	$\mathrm{H}^\times(\pi(\tau_t)\|\pi_\theta(\tau_t))$	$-\mathrm{H}^\times(\pi(\tau_{i+1})\|\pi_\theta(\tau_{i+1}))$	yes**	$\mathbb{E}_{\pi_\theta}[\sum_t \mathrm{H}^\times(\pi(\tau_t)\|\pi_\theta(\tau_t))]$
Exp. entropy regularised	π_θ	$\mathrm{H}^\times(\pi_\theta(\tau_t)\|\pi(\tau_t))$	$\log\pi(a_{i+1}\|\tau_{i+1})$	yes**	$\mathbb{E}_{\pi_\theta}[\sum_t -\log\pi(a_t\|\tau_t)]$
Teacher V reward	π_θ	0	$r_i + V_\pi(\tau_{i+1}) - V_{\pi_\theta}(\tau_i)$	yes**	$\mathbb{E}_{\pi_\theta}[\sum_t r_t]$

Table 1 from (Czarnecki et al., 2019)

- Different control policies, loss terms and reward terms
- Methods below the mid line are introduced in this paper
- Usually modifications of known techniques to address specific issues

Oscillation Example

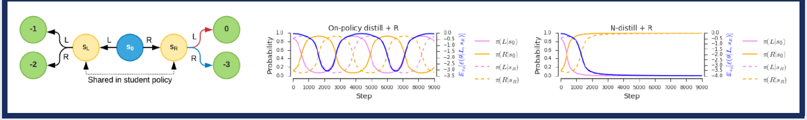


Figure 2 from (Czarnecki et al., 2019)

- Student policy is parameterized with sigmoids and shares parameter for both yellow states
- Teacher policy prefers to go right when in s_R with $\ell(\theta|s_R) = -4\pi_\theta(R|s_R)$
- On-policy distillation diverges

$$\ell_t = H^\times(\pi(s_t)||\pi_\theta(s_t)), \quad \hat{r}_t = 0$$

- N-distillation converges

$$\ell_t = H^\times(\pi(s_t)||\pi_\theta(s_t)), \quad \hat{r}_t = -H^\times(\pi(s_{t+1})||\pi_\theta(s_{t+1}))$$

Oscillation Example

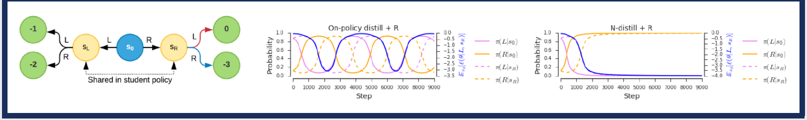


Figure 2 from (Czarnecki et al., 2019)

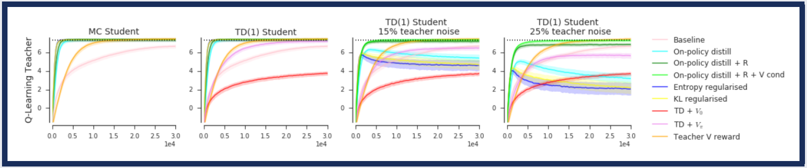
- Student policy is parameterized with sigmoids and shares parameter for both yellow states
- Teacher policy prefers to go right when in s_R with $\ell(\theta|s_R) = -4\pi_\theta(R|s_R)$
- On-policy distillation diverges

$$\ell_t = H^\times(\pi(s_t)||\pi_\theta(s_t)), \quad \widehat{r}_t = 0$$

- N-distillation converges

$$\ell_t = H^\times(\pi(s_t)||\pi_\theta(s_t)), \quad \widehat{r}_t = -H^\times(\pi(s_{t+1})||\pi_\theta(s_{t+1}))$$

Results of Experiments - Cont'd



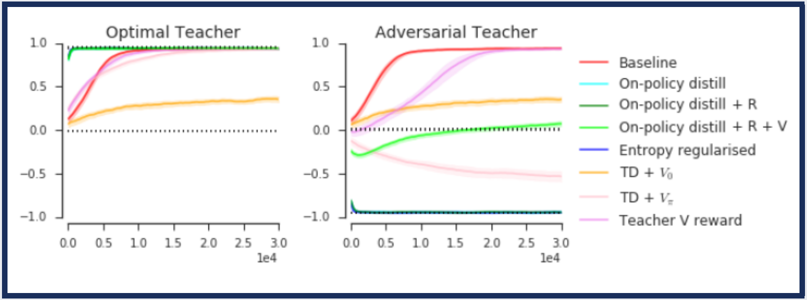


Figure 4 & 6 from (Czarnecki et al., 2019)