A GRADIENTS BETWEEN POLICY IMPROVEMENT AND POLICY EVALUATION

Theorem 1 (Sutton et al. (1999)). *If Q_θ satisfies $\mathbb{E}_\pi[(Q^\pi - Q_\theta)\nabla_\theta Q_\theta] = 0$ and $\nabla_\theta Q_\theta = \nabla_\theta \log \pi_\theta$, then we have*

$$\nabla_\theta \mathcal{J} = \mathbb{E}_\pi[Q_\theta \nabla_\theta \log \pi_\theta].$$

	Function Approximation	Train Gradients	Interested Angles
PPO	$(V, logit) = (V_\theta, logit_\theta)$ $\pi = \text{softmax}(logit)$	$0.5\nabla L_V + \nabla \mathcal{J}$	$< \nabla L_V, \nabla \mathcal{J} >$
PPO ver.1	$(Q, logit) = (Q_\theta, logit_\theta)$, $\pi = \text{softmax}(logit)$ $V = sg(\pi) \cdot Q$	$0.5\nabla L_V + \nabla \mathcal{J}$	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$ $< \nabla Q, \nabla \log \pi >$
PPO ver.2	$(Q, logit) = (Q_\theta, logit_\theta)$, $pi = \text{softmax}(logit)$ $V = sg(\pi) \cdot Q$	$0.5\nabla L_V + \nabla L_Q + \nabla \mathcal{J}$	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$ $< \nabla Q, \nabla \log \pi >$
PPO+CASA	$(V, A) = (V_\theta, A_\theta)$, $\pi = \text{softmax}(A/\tau)$, $\bar{A} = A - sg(\pi) \cdot A$ $Q = \bar{A} + sg(V)$	$0.5\nabla L_V + \nabla L_Q + \nabla \mathcal{J}$	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$ $< \nabla Q, \nabla \log \pi >$

Table 5: PPO is the original PPO. PPO ver.1 and PPO ver.2 are adapted versions to calculate ∇L_Q. PPO+CASA is applying CASA on PPO, which is described in Sec. 4.2.

	Function Approximation	Train Gradients	Interested Angles
R2D2	$(V, A) = (V_\theta, A_\theta)$ $Q = A + V$ $\pi = \text{softmax}(A/\tau)$	∇L_Q	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$
R2D2 ver.1	$(V, A) = (V_\theta, A_\theta)$ $Q = A + V$ $\pi = \text{softmax}(A/\tau)$	$0.5\nabla L_V + \nabla L_Q$	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$
R2D2+CASA	$(V, A) = (V_\theta, A_\theta)$, $\pi = \text{softmax}(A/\tau)$, $\bar{A} = A - sg(\pi) \cdot A$ $Q = \bar{A} + sg(V)$	$0.5\nabla L_V + \nabla L_Q + \nabla \mathcal{J}$	$< \nabla L_V, \nabla \mathcal{J} >$ $< \nabla L_Q, \nabla \mathcal{J} >$ $< \nabla L_V, \nabla L_Q >$

Table 6: R2D2 is the original R2D2. R2D2 ver.1 is adapted version to include ∇L_V for training. R2D2+CASA is applying CASA on R2D2, which is described in Sec. 4.2.

To understand the behavior of

$$\beta =< (Q^\pi - Q_\theta)\nabla_\theta Q_\theta,\ (Q^\pi - V_\theta)\nabla_\theta \log \pi_\theta >$$

in reinforcement learning algorithms, we choose PPO as a representative as policy-based methods and R2D2 as a representative as value-based algorithms.

Define

$$L_V(\theta) = \mathbb{E}_\pi[(V^\pi - V_\theta)^2],\ L_Q(\theta) = \mathbb{E}_\pi[(Q^\pi - Q_\theta)^2],$$

and

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_\pi \left[(Q^\pi - V_\theta)\nabla_\theta \log \pi\right].$$

We usually have above three kinds of loss functions in reinforcement learning, which aim to estimate the state values, state-action values and the policy. We do not talk about the estimations of V^π and Q^π as they are estimated as their usual way of PPO's and R2D2's. All hyperparameters are listed in Appendix D.

The fact that PPO only has $\nabla_\theta L_V$ and $\nabla_\theta \mathcal{J}$ and R2D2 only has $\nabla_\theta L_Q$ is the main difficulty to track both the gradients of policy improvement and policy evaluation. To solve the problem, we adjust PPO and R2D2 with different versions.

For PPO, we displace the estimation of V_θ by $sg(\pi) \cdot Q_\theta$, where Q_θ is estimated by function approximation and V is estimated by taking the expectation of Q_θ. All versions of PPO are listed in Table 5.

For R2D2, we point out that though we apply ϵ-greedy to interact with environments, ϵ is only used for exploration and the final target policy of value-based methods is simply $\arg\max Q_\theta$. Because $\arg\max Q_\theta$ breaks the gradient, we use a surrogate policy to approximate the gradient of policy improvement. Since R2D2 uses dueling structure and $\text{softmax}(A_\theta/\tau) = \text{softmax}(Q_\theta/\tau) \xrightarrow{\tau \to 0+} \arg\max Q_\theta$, we use $\pi_{surrogate} = \text{softmax}(A_\theta/\tau)$ to calculate the policy gradient. We only use $\pi_{surrogate}$ on learner to calculate the gradient, the policy that interacts with environments is still ϵ-greedy. All versions of R2D2 are listed in Table 6.

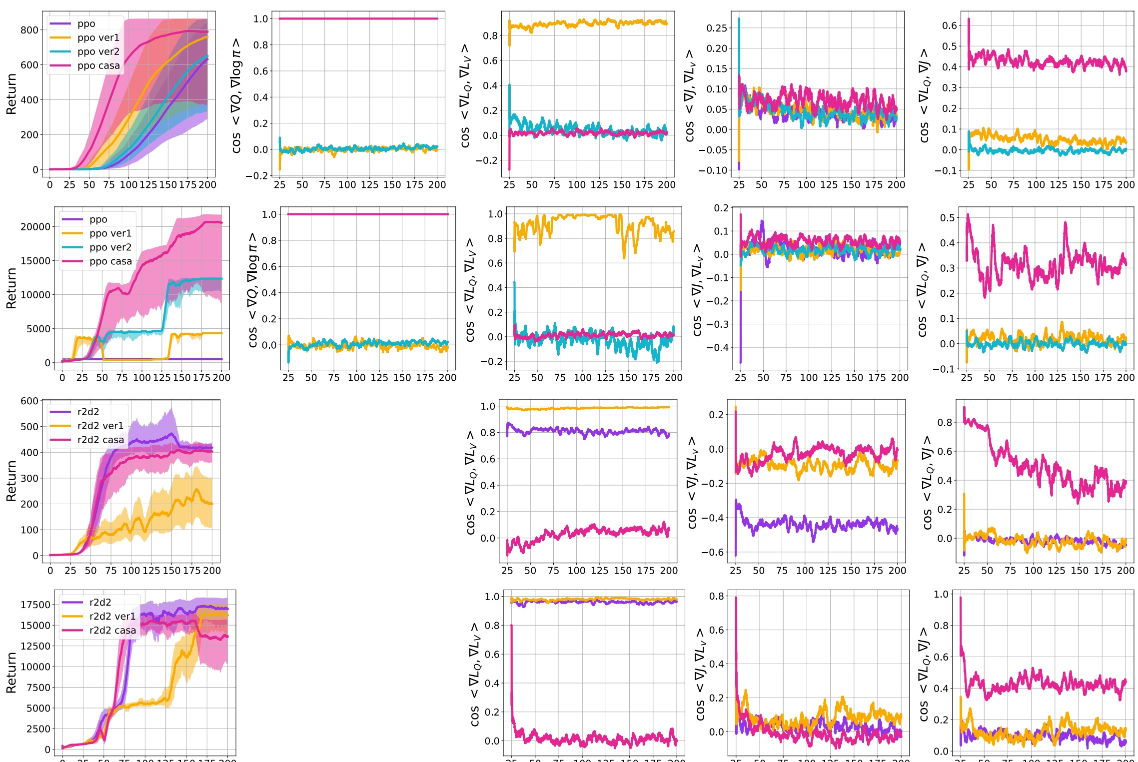


Figure 5: Angles of Gradients and Returns of versions of PPO and R2D2 defined in Table 5 and Table 6.

B DR-TRACE

As CASA estimates (V, Q, π), we would ask **i)** how to guarantee that $\tilde{\pi}_{VTrace} = \tilde{\pi}_{ReTrace}$, **ii)** how to exploit (V, Q, π) to make a better estimation. Though we can apply V-Trace to estimate V and ReTrace to estimate Q with proper hyperparameters to guarantee $\tilde{\pi}_{VTrace} = \tilde{\pi}_{ReTrace}$, it's more reasonable to estimate (V, Q) together. Inspired by Doubly Robust, which is shown to maximally reduce the variance, we introduce DR-Trace, which estimates V by

$$V_{DR}^{\tilde{\pi}}(s_t) \overset{def}{=} \mathbb{E}_{\mu}[V(s_t) + \sum_{k\geq 0} \gamma^k c_{[t:t+k-1]} \rho_{t+k} \delta_{t+k}^{DR}],$$

where $\delta_t^{DR} \overset{def}{=} r_t + \gamma V(s_{t+1}) - Q(s_t, a_t)$ is one-step Doubly Robust error, $\rho_t \overset{def}{=} \min\{\frac{\pi_t}{\mu_t}, \bar{\rho}\}$ and $c_t \overset{def}{=} \min\{\frac{\pi_t}{\mu_t}, \bar{c}\}$ are clipped per-step importance sampling, $c_{[t:t+k]} \overset{def}{=} \prod_{i=0}^{k} c_{t+i}$.

With one step Bellman equation, we estimate Q by

$$\begin{aligned} Q_{DR}^{\tilde{\pi}}(s_t, a_t) &\overset{def}{=} \mathbb{E}_{s_{t+1}, r_t \sim p(\cdot,\cdot|s_t,a_t)}[r_t + \gamma V_{DR}^{\tilde{\pi}}(s_{t+1})] \\ &= \mathbb{E}_{\mu}[Q(s_t, a_t) + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} \delta_{t+k}^{DR}], \end{aligned}$$

where $\tilde{\rho}_{t,k} = 1_{\{k=0\}} + 1_{\{k>0\}} \rho_{t+k}$.

Theorem 2. *Define* $\bar{A} = A - \mathbb{E}_\pi[A]$, $Q = \bar{A} + sg(V)$,

$$\begin{aligned} &\mathscr{T}(Q) \overset{def}{=} \mathbb{E}_{\mu}[Q(s_t, a_t) + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} \delta_{t+k}^{DR}], \\ &\mathscr{S}(V) \overset{def}{=} \mathbb{E}_{\mu}[V(s_t) + \sum_{k\geq 0} \gamma^k c_{[t:t+k-1]} \rho_{t,k} \delta_{t+k}^{DR}], \\ &\mathscr{U}(Q, V) = (\mathscr{T}(Q) - \mathbb{E}_\pi[Q] + \mathscr{S}(V), \mathscr{S}(V)), \\ &\mathscr{U}^{(n)}(Q, V) = \mathscr{U}(\mathscr{U}^{(n-1)}(Q, V)), \end{aligned}$$

then $\mathscr{U}^{(n)}(Q, V) \to (Q^{\tilde{\pi}}, V^{\tilde{\pi}})$ *that corresponds to*

$$\tilde{\pi}(a|s) = \frac{\min\{\bar{\rho}\mu(a|s), \pi(a|s)\}}{\sum_{b\in\mathcal{A}} \min\{\bar{\rho}\mu(b|s), \pi(b|s)\}}.$$

as $n \to +\infty$.

Proof. See Appendix C, Theorem C.1 □

Theorem 2 shows that DR-Trace is a contraction mapping and (V, Q) converges to $(V^{\tilde{\pi}}, Q^{\tilde{\pi}})$ that corresponds to

$$\tilde{\pi}(a|s) = \frac{\min\{\bar{\rho}\mu(a|s), \pi(a|s)\}}{\sum_{b\in\mathcal{A}} \min\{\bar{\rho}\mu(b|s), \pi(b|s)\}}.$$

According to our proof, DR-Trace should work similar to V-Trace and ReTrace, as the convergence rate and the limitation are same. We compare DR-Trace with V-Trace+ReTrace in Figure 6, where we replace estimation of state values by V-Trace and estimation of state-action values by ReTrace. We call V-Trace+ReTrace as No-DR-Trace for brevity. No-DR-Trace performs better on Breakout and ChopperCommand, but fails to make a breakthrough on Krull. Recalling the fact that Doubly Robust can maximally reduce the variance of Bellman error, No-DR-Trace is less stable but also potential to achieve a better performance. A conclusion cannot be made about No-DR-Trace, as this phenomenon means that No-DR-Trace is less stable than DR-Trace, but it also holds the potential to achieve a better performance.

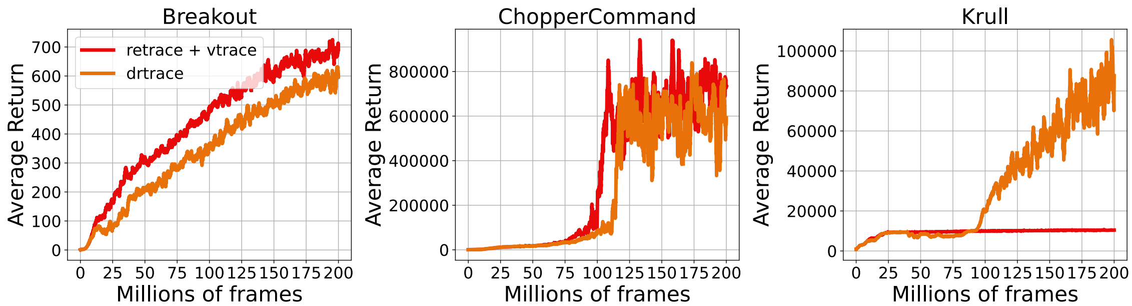


Figure 6: Ablation study for w/wo DR-Trace on Breakout, ChopperCommand and Krull.

C PROOFS

Lemma C.1. *(i) Define* $\pi = softmax(A/\tau)$*, then* $\nabla \log \pi = (\mathbf{1} - \pi)\frac{\nabla A}{\tau}$*. (ii) Denote* sg *to be stop gradient and define* $\bar{A} = A - \mathbb{E}_\pi[A]$*,* $Q = \bar{A} + sg(V)$*, then* $\nabla Q = (\mathbf{1} - \pi)\nabla A$*.*

Proof. As $Q = \bar{A} + sg(V) = A - sg(\pi) \cdot A + sg(V)$, it's obvious that $\nabla Q = (\mathbf{1} - \pi)\nabla A$.

For $\log \pi$, it's a standard derivative of cross entropy, so we have $\nabla \log \pi = (\mathbf{1} - \pi)\nabla(A/\tau) = (\mathbf{1} - \pi)\frac{\nabla A}{\tau}$. □

Lemma C.2. *Define* $\bar{A} = A - \mathbb{E}_\pi[A]$*,* $Q = \bar{A} + sg(V), \pi = softmax(A/\tau)$*, then*

$$\mathbb{E}_\pi\left[(Q - V)\nabla \log \pi\right] = -\tau \nabla \boldsymbol{H}[\pi].$$

Proof. Since

$$\pi = \exp(A/\tau)/Z,\ Z = \int_{\mathcal{A}} \exp(A/\tau),$$

we have

$$A = \tau \log \pi + \tau \log Z.$$

Based on the observation that $\mathbb{E}_\pi\left[f(s)\nabla \log \pi(\cdot|s)\right] = 0$, we have

$$\mathbb{E}_\pi\left[\mathbb{E}_\pi[A] \cdot \nabla \log \pi\right] = 0,$$

$$\mathbb{E}_\pi\left[\log Z \cdot \nabla \log \pi\right] = 0.$$

On the one hand,

$$\begin{aligned}\mathbb{E}_\pi\left[(Q - V)\nabla \log \pi\right] &= \mathbb{E}_\pi\left[A\nabla \log \pi\right] - \mathbb{E}_\pi\left[\mathbb{E}_\pi[A] \cdot \nabla \log \pi\right] \\ &= \tau\mathbb{E}_\pi\left[\log \pi \nabla \log \pi\right] + \tau\mathbb{E}_\pi\left[\log Z \cdot \nabla \log \pi\right] \\ &= \tau\mathbb{E}_\pi\left[\log \pi \nabla \log \pi\right].\end{aligned}$$

On the other hand,

$$\begin{aligned}\nabla \mathbf{H}[\pi] &= -\nabla \int_{\mathcal{A}} \pi_i \log \pi_i \\ &= -\int_{\mathcal{A}} \nabla \pi_i \cdot \log \pi_i - \int_{\mathcal{A}} \pi_i \nabla \log \pi_i \\ &= -\int_{\mathcal{A}} \pi_i \nabla \log \pi_i \cdot \log \pi_i - \int_{\mathcal{A}} \pi_i \frac{\nabla \pi_i}{\pi_i} \\ &= -\mathbb{E}_\pi \left[\log \pi \nabla \log \pi\right].\end{aligned}$$

Hence, $\mathbb{E}_\pi\left[(Q - V)\nabla \log \pi\right] = -\tau \nabla \mathbf{H}[\pi]$. □

Theorem C.1. *Define* $\bar{A} = A - \mathbb{E}_\pi[A]$, $Q = \bar{A} + sg(V)$. *Define*

$$\begin{aligned}&\mathscr{T}(Q) \overset{def}{=} \mathbb{E}_\mu[Q(s_t, a_t) + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} \delta^{DR}_{t+k}], \\ &\mathscr{S}(V) \overset{def}{=} \mathbb{E}_\mu[V(s_t) + \sum_{k\geq 0} \gamma^k c_{[t:t+k-1]} \rho_{t,k} \delta^{DR}_{t+k}], \\ &\mathscr{U}(Q, V) = (\mathscr{T}(Q) - \mathbb{E}_\pi[Q] + \mathscr{S}(V), \mathscr{S}(V)), \\ &\mathscr{U}^{(n)}(Q, V) = \mathscr{U}(\mathscr{U}^{(n-1)}(Q, V)),\end{aligned}$$

then $\mathscr{U}^{(n)}(Q, V) \to (Q^{\tilde{\pi}}, V^{\tilde{\pi}})$ *that corresponds to*

$$\tilde{\pi}(a|s) = \frac{\min\left\{\bar{\rho}\mu(a|s), \pi(a|s)\right\}}{\sum_{b\in\mathcal{A}} \min\left\{\bar{\rho}\mu(b|s), \pi(b|s)\right\}}.$$

as $n \to +\infty$.

Remark. $\mathscr{T}(Q) - \mathbb{E}_\pi[Q] + \mathscr{S}(V)$ is **exactly** how Q is updated at training time. Since $Q = \bar{A} + sg(V)$, if we apply gradient ascent on Q and V in directions $\nabla L_Q(\theta)$ and $\nabla L_V(\theta)$ respectively, change of Q comes from two aspects. One comes from $\nabla L_Q(\theta)$, which changes A, the other comes from $\nabla L_V(\theta)$, which changes V. Because the gradient of V is stopped when estimating Q, the latter is captured by "minus old baseline, add new baseline", which is $-\mathbb{E}_\pi[Q] + \mathscr{S}(V)$ in Theorem C.1.

Proof. Define

$$\begin{aligned}\widetilde{\mathscr{T}}(Q) &= -\mathbb{E}_\pi[Q] + \mathscr{T}(Q), \\ \widetilde{\mathscr{U}}(Q, V) &= (\widetilde{\mathscr{T}}(Q), \mathscr{S}(V)), \\ \widetilde{\mathscr{U}}^{(n)}(Q, V) &= \widetilde{\mathscr{U}}(\widetilde{\mathscr{U}}^{(n-1)}(Q, V)).\end{aligned}$$

By Lemma C.3, $\widetilde{\mathscr{T}}^{(n)}(Q)$ converges to some A^* as $n \to \infty$. This process will not influence the estimation of V as the gradient of V is stopped when estimating Q. According to the proof, A^* does not depend on V.
By Lemma C.4, $\mathscr{S}^{(n)}(V)$ converges to some V^* as $n \to \infty$.
Hence, we have

$$\widetilde{\mathscr{U}}^{(n)}(Q, V) \to (A^*, V^*) \;\; as \;\; n \to +\infty.$$

By definition,

$$\mathscr{U}(Q, V) = (\widetilde{\mathscr{T}}(Q) + \mathscr{S}(V), \mathscr{S}(V)),$$

we can regard $\widetilde{\mathscr{T}}(Q) + \mathscr{S}(V)$ as Q and regard $\mathscr{S}(V)$ as V, then

$$\begin{aligned}\mathscr{U}^{(2)}(Q,V) &= \mathscr{U}(\widetilde{\mathscr{T}}(Q) + \mathscr{S}(V), \mathscr{S}(V)) \\ &= (\mathscr{T}(\widetilde{\mathscr{T}}(Q) + \mathscr{S}(V)) - \mathscr{S}(V) + \mathscr{S}^{(2)}(V), \mathscr{S}^{(2)}(V)) \\ &= (\widetilde{\mathscr{T}}^{(2)}(Q) + \mathscr{S}^{(2)}(V), \mathscr{S}^{(2)}(V)).\end{aligned}$$

By induction,

$$\begin{aligned}\mathscr{U}^{(n)}(Q,V) &= (\widetilde{\mathscr{T}}^{(n)}(Q) + \mathscr{S}^{(n)}(V), \mathscr{S}^{(n)}(V)) \\ &\to (A^* + V^*, V^*) \ \ as \ \ n \to +\infty.\end{aligned}$$

Same as (Espeholt et al., 2018),

$$\tilde{\pi}(a|s) = \frac{\min\{\bar{\rho}\mu(a|s), \pi(a|s)\}}{\sum_{b\in\mathcal{A}} \min\{\bar{\rho}\mu(b|s), \pi(b|s)\}}.$$

is the policy s.t. the Bellman equation holds, which is

$$\mathbb{E}_\mu[\rho_t(r_t + \gamma V_{t+1} - V_t)|\mathscr{F}_t] = 0,$$

and $\mathscr{U}(Q^{\tilde{\pi}}, V^{\tilde{\pi}}) = (Q^{\tilde{\pi}}, V^{\tilde{\pi}})$.
So we have $(A^* + V^*, V^*) = (Q^{\tilde{\pi}}, V^{\tilde{\pi}})$. □

Lemma C.3. *Define* $\bar{A} = A - \mathbb{E}_\pi[A]$, $Q = \bar{A} + sg(V)$, *then operator*

$$\mathscr{T}(Q) \stackrel{def}{=} \mathbb{E}_\mu[Q(s_t, a_t) + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} \delta^{DR}_{t+k}]$$

is a contraction mapping w.r.t. Q.

Remark. Note that $\mathscr{T}(Q)$ is exactly equation B.

Since $Q = A + sg(V)$, the gradient of V is stopped when estimating Q, updating Q will not change V, which is equivalent to updating A. Without loss of generality, we assume V is fixed as V^* in the proof.

Proof. $\bar{A} = A - \mathbb{E}_\pi[A]$ shows $\mathbb{E}_\pi[\bar{A}] = 0$, which guarantees that no matter how we update A, we always have $\mathbb{E}_\pi[Q] = V^*$.

Based on above observations, define

$$\widetilde{\mathscr{T}}(Q) \stackrel{def}{=} -\mathbb{E}_\pi[Q] + \mathscr{T}(Q).$$

It's obvious that we only need to prove $\widetilde{\mathscr{T}}(Q)$ is a contraction mapping.

For brevity, we denote

$$Q_t = Q(s_t, a_t), A_t = A(s_t, a_t), V_t^* = V^*(s_t).$$

Noticing that $\tilde{\rho}_{t,0} = 1$, let $\mathscr{F}$ represent filtration, we can rewrite $\widetilde{\mathscr{T}}$ as

$$\begin{aligned}\widetilde{\mathscr{T}}(Q) &= \mathbb{E}_\mu[A_t + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} \delta^{DR}_{t+k}] \\ &= \mathbb{E}_\mu[-V_t^* + \sum_{k\geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k} r_{t+k} + \sum_{k\geq 0} \gamma^{k+1} c_{[t+1:t+k-1]} \Delta_k],\end{aligned} \tag{9}$$

where

$$\Delta_k = \mathbb{E}_\mu \left[\tilde{\rho}_{t,k} V^*_{t+k+1} - c_{t+k} \tilde{\rho}_{t,k+1} Q_{t+k+1} | \mathscr{F}_{t+k} \right] . \tag{10}$$

By definition of Q,

$$\mathbb{E}_\mu [V^*_{t+k+1} | \mathscr{F}_{t+k}] = \mathbb{E}_\mu [\mathbb{E}_\pi [Q_{t+k+1} | \mathscr{F}_{t+k+1}] | \mathscr{F}_{t+k}],$$

we can rewrite equation 10 as

$$\Delta_k = \mathbb{E}_\mu [(\tilde{\rho}_{t,k} \frac{\pi_{t+k+1}}{\mu_{t+k+1}} - c_{t+k} \tilde{\rho}_{t,k+1}) Q_{t+k+1} | \mathscr{F}_{t+k}]. \tag{11}$$

For any $Q_1 = A_1 + sg(V^*)$, $Q_2 = A_2 + sg(V^*)$, since

$$\mathbb{E}_\mu [(\tilde{\rho}_{t,k} \frac{\pi_{t+k+1}}{\mu_{t+k+1}} - c_{t+k} \tilde{\rho}_{t,k+1}) | \mathscr{F}_{t+k}] \geq 0,$$

by equation 9 equation 11, we have

$$||\widetilde{\mathscr{T}}(Q_1) - \widetilde{\mathscr{T}}(Q_2)|| \leq \mathcal{C} ||Q_1 - Q_2||,$$

where

$$\begin{aligned}
\mathcal{C} &= \mathbb{E}_\mu [\sum_{k \geq 0} \gamma^{k+1} c_{[t+1:t+k-1]} (\tilde{\rho}_{t,k} \frac{\pi_{t+k+1}}{\mu_{t+k+1}} - c_{t+k} \tilde{\rho}_{t,k+1})] \\
&= \mathbb{E}_\mu [1 - 1 + \sum_{k \geq 0} \gamma^{k+1} c_{[t+1:t+k-1]} \left(\tilde{\rho}_{t,k} - c_{t+k} \tilde{\rho}_{t,k+1} \right)] \\
&= 1 - (1 - \gamma) \mathbb{E}_\mu [\sum_{k \geq 0} \gamma^k c_{[t+1:t+k-1]} \tilde{\rho}_{t,k}] \\
&\leq 1 - (1 - \gamma) < 1.
\end{aligned}$$

Hence, $\widetilde{\mathscr{T}}(Q)$ is a contraction mapping and converges to some fixed function, which we denote as A^*. So $\mathscr{T}(Q)$ is also a contraction mapping and converges to $A^* + V^*$. □

Lemma C.4. *Define $Q = A + sg(V)$ with $\mathbb{E}_\pi[A] = 0$, then operator*

$$\mathscr{S}(V) \stackrel{def}{=} \mathbb{E}_\mu [V(s_t) + \sum_{k \geq 0} \gamma^k c_{[t:t+k-1]} \rho_{t,k} \delta^{DR}_{t+k}]$$

is a contraction mapping w.r.t. V.

Remark. Note that $\mathscr{S}(V)$ is exactly equation B.

Proof. Same as Lemma C.3, we can get

$$\Delta_k = \mathbb{E}_\mu \left[\left(\rho_{t+k} - c_{t+k} \rho_{t+k+1} \right) V_{t+k+1} - c_{t+k} \rho_{t+k+1} A^*_{t+k+1} | \mathscr{F}_{t+k} \right] ,$$

so we have

$$\Delta^1_k - \Delta^2_k = \mathbb{E}_\mu \left[\left(\rho_{t+k} - c_{t+k} \rho_{t+k+1} \right) \cdot (V^1_{t+k+1} - V^2_{t+k+1}) | \mathscr{F}_{t+k} \right] .$$

The remaining proof is identical to (Espeholt et al., 2018)'s. □

D HYPERPARAMETERS

Our python packages are shown in Table 7

Package	Version
ale-py	0.6.0.dev20200207
gym	0.19.0
tensorflow	1.15.2
opencv-python	4.1.2.30
opencv-contrib-python	4.4.0.46

Table 7: Versions for python packages among all experiments.

All experiments follow the shared hyperparameters as in Table 8. The specific hyperparameters for PPO, R2D2 and CASA+DR-Trace are shown in Table 9, Table 10 and Table 11. The only exceptions are V-loss scaling, Q-loss scaling and π-loss scaling, which may be zero depending on some specific ablation settings. We will state these three hyperparameters every time in all experiments.

Parameter	Value
Atari Version	NoFrameskip-v4
Atari Wrapper	gym.wrappers.atari_preprocessing
Image Size	(84, 84)
Grayscale	Yes
Num. Action Repeats	4
Num. Frame Stacks	4
Action Space	Full
End of Episode When Life Lost	No
Num. States	200M
Num. Environments	160
Random No-ops	30
Burn-in	40
Seq-length	80
Burn-in Stored Recurrent State	Yes
Bootstrap	Yes
Batch size	64
Backbone	IMPALA,deep
LSTM Units	256
Optimizer	Adam Weight Decay
Weight Decay Rate	0.01
Weight Decay Schedule	Anneal linearly to 0
Learning Rate	5e-4
Warmup Steps	4000
Learning Rate Schedule	Anneal linearly to 0
AdamW β_1	0.9
AdamW β_2	0.98
AdamW ϵ	1e-6
AdamW Clip Norm	50.0
Learner Push Model Every n Steps	25
Actor Pull Model Every n Steps	64

Table 8: Configurations for shared hyperparameters among all experiments.

Parameter	Value
Sample Reuse	1
Reward Shape	$\text{clip}(r, 0, 1)$
Discount (γ)	0.995
V-loss Scaling (α_1)	0.5
Q-loss Scaling (α_2)	1.0
π-loss Scaling (α_3)	1.0
PPO clip ϵ	0.2
GAE λ	0.8
Temperature (τ)	0.1

Table 9: Hyperparameter configurations for PPO.

Parameter	Value
Sample Reuse	2
Target Shape	$Q_t^{\tilde{\pi}} = h(\sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n h^{-1}(\text{Double}(Q_{t+n})))$
Target Shape Function h	$h(x) = \text{sign}(x) \cdot (\sqrt{\lvert x \rvert + 1} - 1) + 10^{-3} x$
Bootstrap Length n	5
ϵ-greedy	$\epsilon \sim 0.4^{\text{uniform}(1,8)}$
PER Sample Temperature α	0.9
PER Buffer Size	400000
Discount (γ)	0.997
V-loss Scaling (α_1)	0.5
Q-loss Scaling (α_2)	1.0
π-loss Scaling (α_3)	1.0
Temperature (τ)	0.1

Table 10: Hyperparameter configurations for R2D2.

Parameter	Value
Sample Reuse	2
Reward Shape	$\log(\lvert r \rvert + 1.0) \cdot (2 \cdot 1_{\{r \geq 0\}} - 1_{\{r<0\}})$
Discount (γ)	0.997
V-loss Scaling (α_1)	1.0
Q-loss Scaling (α_2)	10.0
π-loss Scaling (α_3)	10.0
Temperature (τ)	1.0
Importance Sampling Clip $\bar{c}$	1.05
Importance Sampling Clip $\bar{\rho}$	1.05

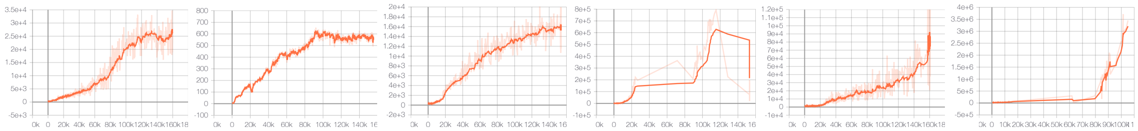
Table 11: Hyperparameter configurations for CASA + DR-Trace.

E EVALUATION OF CASA ON ATARI GAMES

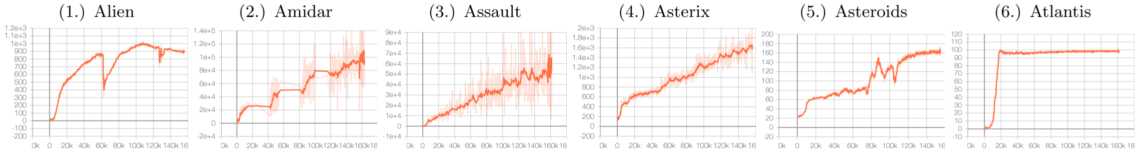
Random scores and average human's scores are from (Badia et al., 2020). Human World Records (HWR) are from (Toromanoff et al., 2019). Rainbow's scores are from (Hessel et al., 2017). IMPALA's scores are from (Espeholt et al., 2018). LASER's scores are from (Schmitt et al., 2020), no sweep at 200M.

Games	RND	HUMAN	RAINBOW	HNS(%)	IMPALA	HNS(%)	LASER	HNS(%)	CASA	HNS(%)
Scale			200M		200M		200M		200M	
alien	227.8	7127.8	9491.7	134.26	15962.1	228.03	**35565.9**	**512.15**	26137	375.50
amidar	5.8	1719.5	**5131.2**	**299.08**	1554.79	90.39	1829.2	106.4	560	32.34
assault	222.4	742	14198.5	2689.78	19148.47	3642.43	**21560.4**	**4106.62**	16228	3080.37
asterix	210	8503.3	**428200**	**5160.67**	300732	3623.67	240090	2892.46	213580	2572.80
asteroids	719	47388.7	2712.8	4.27	108590.05	231.14	**213025**	**454.91**	80339	170.60
atlantis	12850	29028.1	826660	5030.32	849967.5	5174.39	841200	5120.19	**3211600**	**19772.10**
bank heist	14.2	753.1	**1358**	**181.86**	1223.15	163.61	569.4	75.14	895.3	119.24
battle zone	236	37187.5	62010	167.18	20885	55.88	64953.3	175.14	**91269**	**246.36**
beam rider	363.9	16926.5	16850.2	99.54	32463.47	193.81	**90881.6**	**546.52**	57456	344.70
berzerk	123.7	2630.4	2545.6	96.62	1852.7	68.98	**25579.5**	**1015.51**	1648	60.81
bowling	23.1	160.7	30	5.01	59.92	26.76	48.3	18.31	**162.4**	**101.24**
boxing	0.1	12.1	99.6	829.17	99.96	832.17	**100**	**832.5**	98.3	818.33
breakout	1.7	30.5	417.5	1443.75	**787.34**	**2727.92**	747.9	2590.97	624.3	2161.81
centipede	2090.9	12017	8167.3	61.22	11049.75	90.26	**292792**	**2928.65**	102600	1012.57
chopper command	811	7387.8	16654	240.89	28255	417.29	**761699**	**11569.27**	616690	9364.42
crazy climber	10780.5	36829.4	**168788.5**	**630.80**	136950	503.69	167820	626.93	161250	600.70
defender	2874.5	18688.9	55105	330.27	185203	1152.93	336953	2112.50	**421600**	**2647.75**
demon attack	152.1	1971	111185	6104.40	132826.98	7294.24	133530	7332.89	**291590**	**16022.76**
double dunk	-18.6	-16.4	-0.3	831.82	-0.33	830.45	14	1481.82	**20.25**	**1765.91**
enduro	0	860.5	2125.9	247.05	0	0.00	0	0.00	**10019**	**1164.32**
fishing derby	-91.7	-38.8	31.3	232.51	44.85	258.13	45.2	258.79	**53.24**	**273.99**
freeway	0	29.6	**34**	**114.86**	0	0.00	0	0.00	3.46	11.69
frostbite	65.2	4334.7	**9590.5**	**223.10**	317.75	5.92	5083.5	117.54	1583	35.55
gopher	257.6	2412.5	70354.6	3252.91	66782.3	3087.14	114820.7	5316.40	**188680**	**8743.90**
gravitar	173	3351.4	1419.3	39.21	359.5	5.87	1106.2	29.36	**4311**	**130.19**
hero	1027	30826.4	**55887.4**	**184.10**	33730.55	109.75	31628.7	102.69	24236	77.88
ice hockey	-11.2	0.9	1.1	101.65	3.48	121.32	**17.4**	**236.36**	1.56	105.45
jamesbond	29	302.8	19809	72.24	601.5	209.09	**37999.8**	**13868.08**	12468	4543.10
kangaroo	52	3035	**14637.5**	**488.05**	1632	52.97	14308	477.91	5399	179.25
krull	1598	2665.5	8741.5	669.18	8147.4	613.53	9387.5	729.70	**64347**	**5878.13**
kung fu master	258.5	22736.3	52181	230.99	43375.5	191.82	**607443**	**2701.26**	124630.1	553.31
montezuma revenge	0	**4753.3**	384	8.08	0	0.00	0.3	0.01	2488.4	52.35
ms pacman	307.3	6951.6	5380.4	76.35	7342.32	105.88	6565.5	94.19	**7579**	**109.44**
name this game	2292.3	8049	13136	188.37	21537.2	334.30	26219.5	415.64	**32098**	**517.76**
phoenix	761.5	7242.6	108529	1662.80	210996.45	3243.82	**519304**	**8000.84**	498590	7681.23
pitfall	-229.4	**6463.7**	0	3.43	-1.66	3.40	-0.6	3.42	-17.8	3.16
pong	-20.7	14.6	20.9	117.85	20.98	118.07	**21**	**118.13**	20.39	116.40
private eye	24.9	**69571.3**	4234	6.05	98.5	0.11	96.3	0.10	134.1	0.16
qbert	163.9	13455.0	33817.5	253.20	**351200.12**	**2641.14**	21449.6	160.15	27371	204.70
riverraid	1338.5	17118.0	22920.8	136.77	29608.05	179.15	**40362.7**	**247.31**	11182	62.38
road runner	11.5	7845	62041	791.85	57121	729.04	45289	578.00	**251360**	**3208.64**
robotank	2.2	11.9	61.4	610.31	12.96	110.93	**62.1**	**617.53**	10.44	84.95
seaquest	68.4	**42054.7**	15898.9	37.70	1753.2	4.01	2890.3	6.72	11862	28.09
skiing	-17098	**-4336.9**	-12957.8	32.44	-10180.38	54.21	-29968.4	-100.86	-12730	34.23
solaris	1236.3	**12326.7**	3560.3	20.96	2365	10.18	2273.5	9.35	2319	9.76
space invaders	148	1668.7	18789	1225.82	43595.78	2857.09	**51037.4**	**3346.45**	3031	189.58
star gunner	664	10250	127029	1318.22	200625	2085.97	321528	3347.21	**337150**	**3510.18**
surround	-10	6.5	**9.7**	**119.39**	7.56	106.42	8.4	111.52	-10	0.00
tennis	-23.8	-8.3	0	153.55	0.55	157.10	**12.2**	**232.26**	-21.05	17.74
time pilot	3568	5229.2	12926	563.36	48481.5	2703.84	**105316**	**6125.34**	84341	4862.62
tutankham	11.4	167.6	241	146.99	292.11	179.71	278.9	171.25	**381**	**236.62**
up n down	533.4	11693.2	125755	1122.08	332546.75	2975.08	345727	3093.19	**416020**	**3723.06**
venture	0	**1187.5**	5.5	0.46	0	0.00	0	0.00	0	0.00
video pinball	0	17667.9	533936.5	3022.07	**572898.27**	**3242.59**	511835	2896.98	297920	1686.22
wizard of wor	563.5	4756.5	17862.5	412.57	9157.5	204.96	**29059.3**	**679.60**	26008	606.83
yars revenge	3092.9	54576.9	102557	193.19	84231.14	157.60	**166292.3**	**316.99**	118730	224.61
zaxxon	32.5	9173.3	22209.5	242.62	32935.5	359.96	41118	449.47	**46070.8**	**503.66**
MEAN HNS(%)	0.00	100.00		873.97		957.34		1741.36		1941.08
MEDIAN HNS(%)	0.00	100.00		230.99		191.82		454.91		246.36

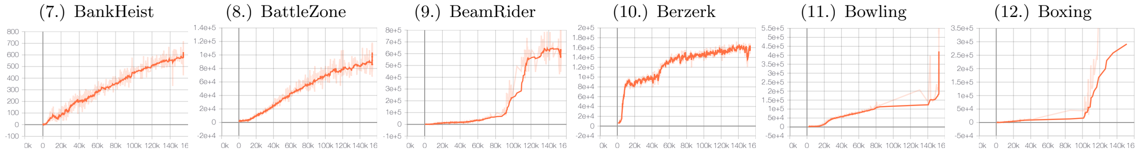
Games	RND	HWR	RAINBOW	SABER(%)	IMPALA	SABER(%)	LASER	SABER(%)	CASA	SABER(%)
Scale			200M		200M		200M		200M	
alien	227.8	**251916**	9491.7	3.68	15962.1	6.25	976.51	14.04	26137	10.29
amidar	5.8	**104159**	5131.2	4.92	1554.79	1.49	1829.2	1.75	560	0.53
assault	222.4	8647	14198.5	165.90	19148.47	200.00	**21560.4**	**200.00**	16228	189.99
asterix	210	**1000000**	428200	42.81	300732	30.06	240090	23.99	213580	21.34
asteroids	719	**10506650**	2712.8	0.02	108590.05	1.03	213025	2.02	80339	0.76
atlantis	12850	**10604840**	826660	7.68	849967.5	7.90	841200	7.82	3211600	30.20
bank heist	14.2	**82058**	1358	1.64	1223.15	1.47	569.4	0.68	895.3	1.07
battle zone	236	**801000**	62010	7.71	20885	2.58	64953.3	8.08	91269	11.37
beam rider	363.9	**999999**	16850.2	1.65	32463.47	3.21	90881.6	9.06	57456	5.71
berzerk	123.7	**1057940**	2545.6	0.23	1852.7	0.16	25579.5	2.41	1648	0.14
bowling	23.1	**300**	30	2.49	59.92	13.30	48.3	9.10	162.4	50.31
boxing	0.1	**100**	99.6	99.60	99.96	99.96	**100**	**100.00**	98.3	98.3
breakout	1.7	**864**	417.5	48.22	787.34	91.11	747.9	86.54	624.3	72.20
centipede	2090.9	**1301709**	8167.3	0.47	11049.75	0.69	292792	22.37	102600	7.73
chopper command	811	**999999**	16654	1.59	28255	2.75	761699	76.15	616690	61.64
crazy climber	10780.5	**219900**	168788.5	75.56	136950	60.33	167820	75.10	161250	71.95
defender	2874.5	**6010500**	55105	0.87	185203	3.03	336953	5.56	421600	6.97
demon attack	152.1	**1556345**	111185	7.13	132826.98	8.53	133530	8.57	291590	18.73
double dunk	-18.6	**21**	-0.3	46.21	-0.33	46.14	14	82.32	20.25	98.11
enduro	0	9500	2125.9	22.38	0	0.00	0	0.00	**10019**	**105.46**
fishing derby	-91.7	**71**	31.3	75.60	44.85	83.93	45.2	84.14	53.24	89.08
freeway	0	**38**	34	89.47	0	0.00	0	0.00	3.46	9.11
frostbite	65.2	**454830**	9590.5	2.09	317.75	0.06	5083.5	1.10	1583	0.33
gopher	257.6	**355040**	70354.6	19.76	66782.3	18.75	114820.7	32.29	188680	53.11
gravitar	173	**162850**	1419.3	0.77	359.5	0.11	1106.2	0.57	4311	2.54
hero	1027	**1000000**	55887.4	5.49	33730.55	3.27	31628.7	3.06	24236	2.32
ice hockey	-11.2	**36**	1.1	26.06	3.48	31.10	17.4	60.59	1.56	27.03
jamesbond	29	**45550**	19809	43.45	601.5	1.26	37999.8	83.41	12468	27.33
kangaroo	52	**1424600**	14637.5	1.02	1632	0.11	14308	1.00	5399	0.38
krull	1598	**104100**	8741.5	6.97	8147.4	6.39	9387.5	7.60	64347	61.22
kung fu master	258.5	**1000000**	52181	5.19	43375.5	4.31	607443	60.73	124630.1	12.44
montezuma revenge	0	**1219200**	384	0.03	0	0.00	0.3	0.00	2488.4	0.20
ms pacman	307.3	**290090**	5380.4	1.75	7342.32	2.43	6565.5	2.16	7579	2.51
name this game	2292.3	25220	13136	47.30	21537.2	83.94	26219.5	104.36	**32098**	**130.00**
phoenix	761.5	**4014440**	108529	2.69	210996.45	5.24	519304	12.92	498590	12.40
pitfall	-229.4	**114000**	0	0.20	-1.66	0.20	-0.6	0.20	-17.8	0.19
pong	-20.7	**21**	20.9	99.76	20.98	99.95	**21**	**100.00**	20.39	98.54
private eye	24.9	**101800**	4234	4.14	98.5	0.07	96.3	0.07	134.1	0.11
qbert	163.9	**2400000**	33817.5	1.40	351200.12	14.63	21449.6	0.89	27371	1.13
riverraid	1338.5	**1000000**	22920.8	2.16	29608.05	2.83	40362.7	3.91	11182	0.99
road runner	11.5	**2038100**	62041	3.04	57121	2.80	45289	2.22	251360	12.33
robotank	2.2	**76**	61.4	80.22	12.96	14.58	62.1	81.17	10.44	11.17
seaquest	68.4	**999999**	15898.9	1.58	1753.2	0.17	2890.3	0.28	11862	1.18
skiing	-17098	**-3272**	-12957.8	29.95	-10180.38	50.03	-29968.4	-93.09	-12730	31.59
solaris	1236.3	**111420**	3560.3	2.11	2365	1.02	2273.5	0.94	2319	0.98
space invaders	148	**621535**	18789	3.00	43595.78	6.99	51037.4	8.19	3031	0.46
star gunner	664	77400	127029	164.67	200625	200.00	321528	200.00	**337150**	**200.00**
surround	-10	9.6	**9.7**	**100.51**	7.56	89.59	8.4	93.88	-10	0.00
tennis	-23.8	**21**	0	53.13	0.55	54.35	12.2	80.36	-21.05	6.14
time pilot	3568	65300	12926	15.16	48481.5	72.76	**105316**	**164.82**	84341	130.84
tutankham	11.4	**5384**	241	4.27	292.11	5.22	278.9	4.98	381	6.88
up n down	533.4	82840	125755	152.14	332546.75	200.00	345727	200.00	**416020**	**200.00**
venture	0	**38900**	5.5	0.01	0	0.00	0	0.00	0	0.00
video pinball	0	**89218328**	533936.5	0.60	572898.27	0.64	511835	0.57	297920	0.33
wizard of wor	563.5	**395300**	17862.5	4.38	9157.5	2.18	29059.3	7.22	26008	6.45
yars revenge	3092.9	**15000105**	102557	0.66	84231.14	0.54	166292.3	1.09	118730	0.77
zaxxon	32.5	**83700**	22209.5	26.51	32935.5	39.33	41118	49.11	46070.8	55.03
MEAN SABER(%)	0.00	100.00		28.39		29.45		36.78		36.10
MEDIAN SABER(%)	0.00	100.00		4.92		4.31		8.08		10.29



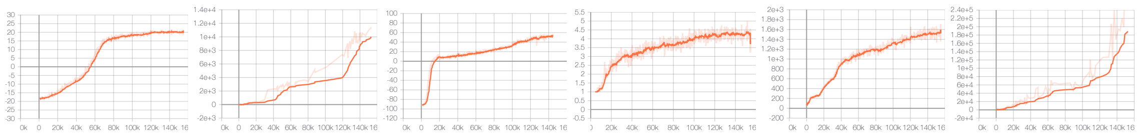
(1.) Alien (2.) Amidar (3.) Assault (4.) Asterix (5.) Asteroids (6.) Atlantis



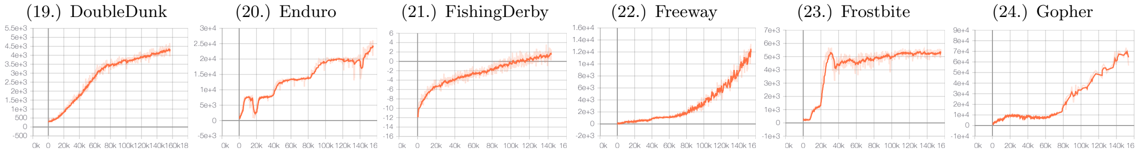
(7.) BankHeist (8.) BattleZone (9.) BeamRider (10.) Berzerk (11.) Bowling (12.) Boxing



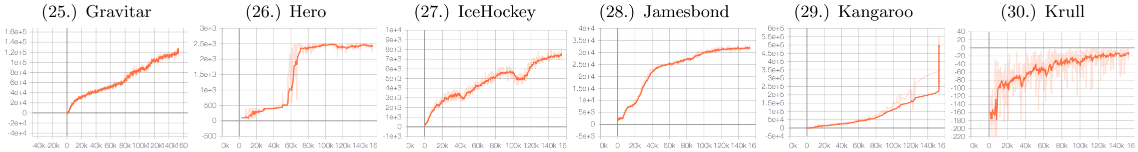
(13.) Breakout (14.) Centipede (15.) ChopperCommand (16.) CrazyClimber (17.) Defender (18.) DemonAttack



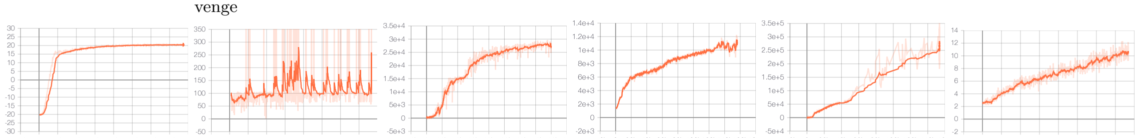
(19.) DoubleDunk (20.) Enduro (21.) FishingDerby (22.) Freeway (23.) Frostbite (24.) Gopher



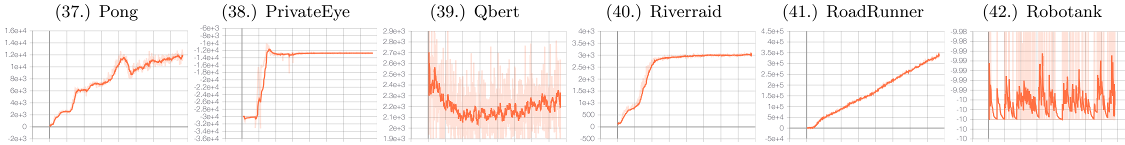
(25.) Gravitar (26.) Hero (27.) IceHockey (28.) Jamesbond (29.) Kangaroo (30.) Krull



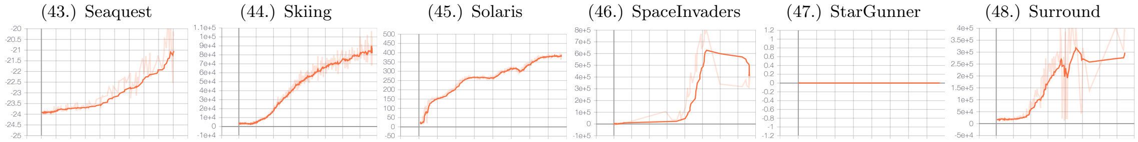
(31.) KungFuMaster (32.) MontezumaRevenge (33.) MsPacman (34.) NameThisGame (35.) Phoenix (36.) Pitfall



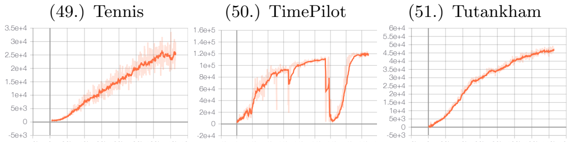
(37.) Pong (38.) PrivateEye (39.) Qbert (40.) Riverraid (41.) RoadRunner (42.) Robotank



(43.) Seaquest (44.) Skiing (45.) Solaris (46.) SpaceInvaders (47.) StarGunner (48.) Surround



(49.) Tennis (50.) TimePilot (51.) Tutankham (52.) UpNDown (53.) Venture (54.) VideoPinball



(55.) WizardOfWor (56.) YarsRevenge (57.) Zaxxon