

---

# Return-based Scaling: Yet Another Normalisation Trick for Deep RL

Tom Schaul<sup>1</sup> Georg Ostrovski<sup>1</sup> Iurii Kemaev<sup>1</sup> Diana Borsa<sup>1</sup>

발표자 김봉석

# 1. Abstract

1. Scaling issues are mundane yet irritating for practitioners of reinforcement learning
2. Error scales vary across domains, tasks, and stages of learning; sometimes by many orders of magnitude
3. This can be detrimental to learning speed and stability, create interference between learning tasks, and necessitate substantial tuning

We revisit this topic for agents based on temporal-difference learning,

**The mechanism we propose requires neither tuning, clipping, nor adaptation.  
We validate its effectiveness and robustness on the suite of Atari gam**

## 2. Introduction

unlike in supervised learning,

- there is **no standard preprocessing step** (such as whitening) that adjusts the scales of the learning targets.
- multiple sources of non-stationarity can cause scales to vary during learning

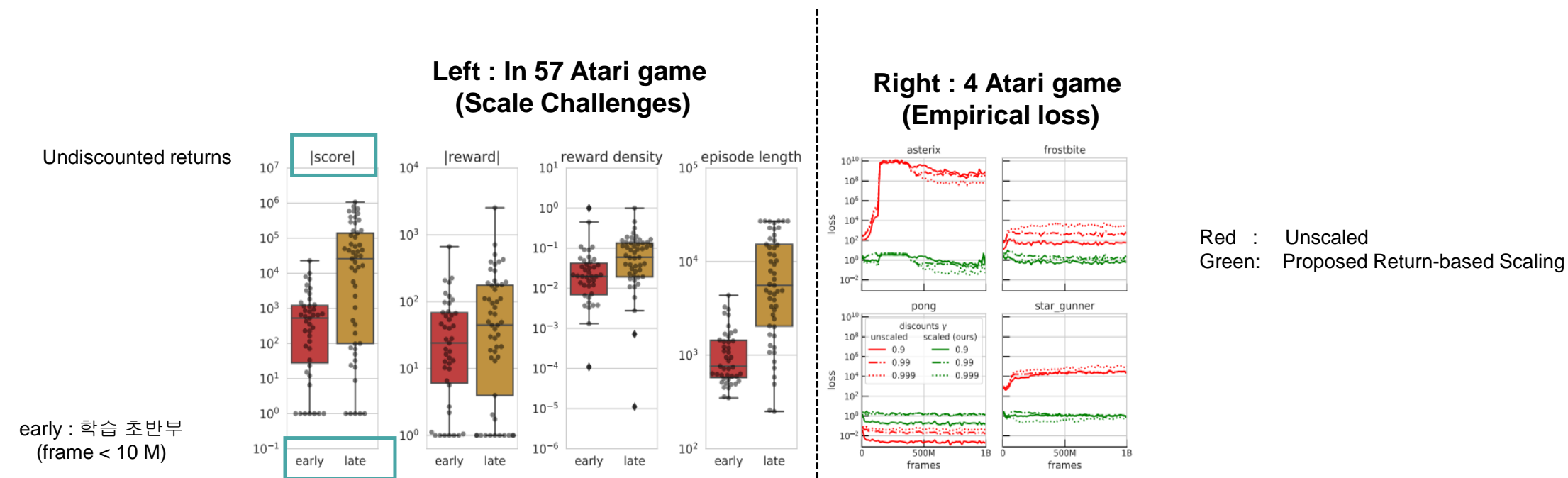
The resulting error scales depend on,

- the **reward scales** (and density), which can vary widely across tasks or domains, Reward
- cumulative quantities, **Return**
- **discount factor**

$$G_t := R_t + \gamma G_{t+1} = \sum_{t'=t}^T \gamma^{t'-t} R_{t'},$$

## 2. Introduction

- In practice, these variations can span many orders of magnitude



**Figure 1. Left:** Scale challenges in Atari. Each subplot shows the variability of scales across 57 Atari games (one point per game), for both the early phase of learning (frames < 10M, in red) and the late phase (800M < frames < 1G, in orange). Note how scores (undiscounted returns) vary by many orders of magnitude, which is a compound effect of changing reward scales, reward densities, and episode lengths. Also note how much these statistics can change over the course of learning. **Right:** Illustration of empirical loss scales on a few individual Atari games. They can vary by 10 orders of magnitude across domains, and can increase or decrease substantially over the course of training (red curves), or both. Each line-style corresponds to a different discount factor  $\gamma$ . Green curves show how our proposed scaling maps the corresponding scales to a much narrower range.

## 2. Introduction

- This seemingly mundane phenomenon is nevertheless a frequent contributor to thorny practical issues,
- strategies abound in the field of deep RL;
  1. reward and gradient clipping (Mnih et al., 2015; Espeholt et al., 2018)
  2. discount factors, non-linear reward or value transforms (Hester et al., 2017; Pohlen et al., 2018; van Hasselt et al., 2019)
  3. separations between value and advantage (Wang et al., 2016)
  4. separate networks instead of a shared torso (Badia et al., 2020a).

**can cause undesirable side-effects**

### **< return-based scaling >**

- sidesteps undesirable side-effects
- algorithm-agnostic as it requires no access to agent internals
- low computational and implementation complexity
- does not introduce any new hyperparameters

## 2. Introduction

- consider several scenarios, varying a different aspect of a reference reward sequence in each, illustrate how that affects the scales of rewards and returns

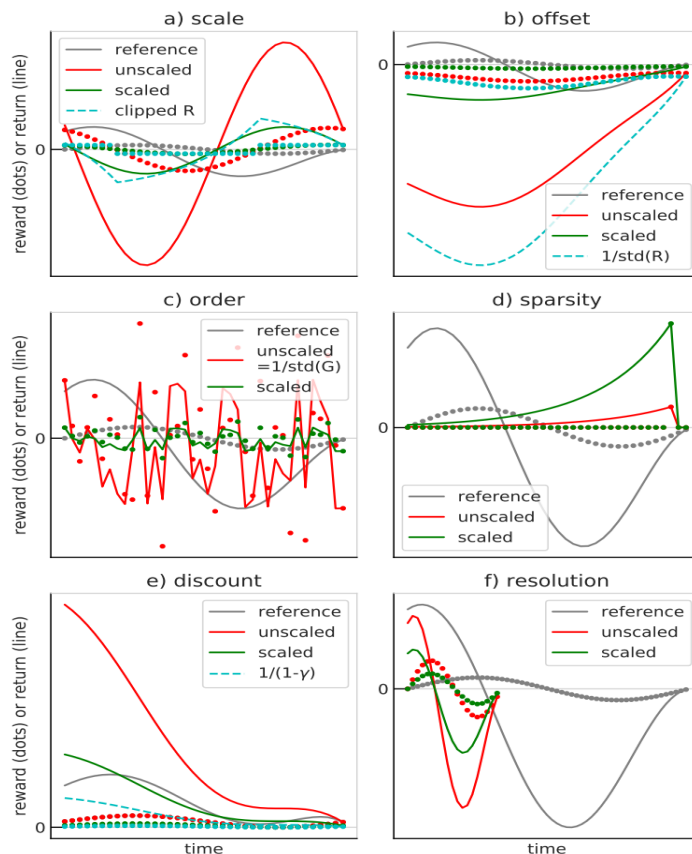


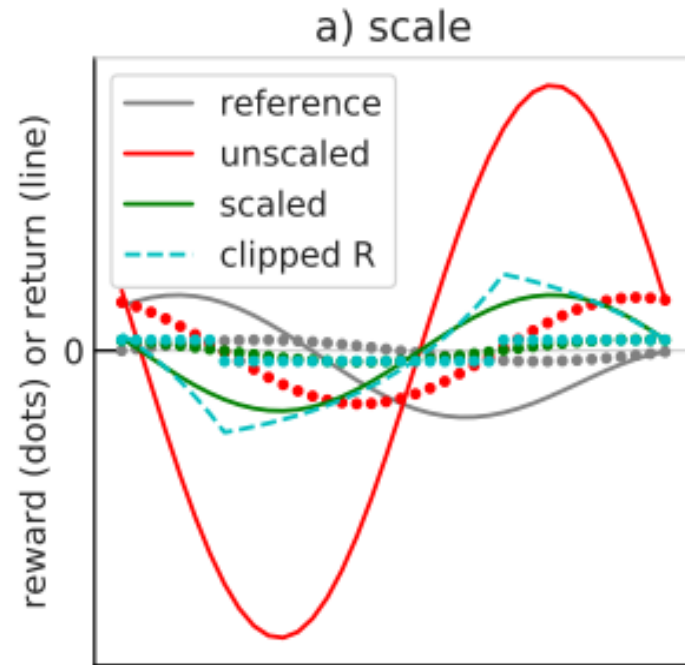
Figure 2. Six scenarios for establishing scale intuitions. Rewards sequences are shown as dots, return curves as full lines (of matching color). Each panel juxtaposes a reference sequence (in gray) at canonical scale, a second unscaled sequence (in red) and how our method would linearly adjust its scale (in green). We encourage the reader to consider whether the match of scales between the green line and the gray line is satisfying, or at least better than the red line. In addition, scenarios (a), (b), (c) and (e) highlight a failure mode of some ‘false friend’ (see Section 2.1), shown in cyan. In scenario (c), cyan and red lines overlap exactly, and in scenario (e) red and gray dots overlap.

✓ Reward (dot)  
✓ Return (lines)

✓ Red (unscaled)  
✓ Green (scaled: proposed)  
✓ Gray (reference)  
✓ Cyan (false friend) : minimalist ideas that quickly come to mind often look tempting, but fail in others

We encourage the reader to consider  
**whether the match of scales between the green line and the gray line is satisfying**

### 3. Intuition

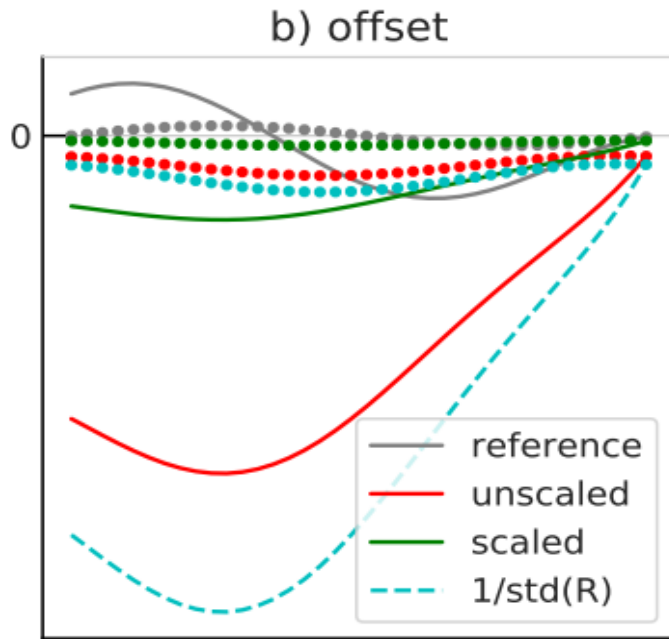


- The purest scenario is when all rewards (and thus returns) are linearly scaled up or down by some factor:
- we expect a reasonable scaling mechanism to correct by the same factor

✓ Reward (dot)  
✓ Return (lines)

✓ Red (unscaled)  
✓ Green (scaled: proposed)  
✓ Gray (reference)  
✓ Cyan (false friend) : minimalist ideas that quickly come to mind often look tempting, but fail in others

### 3. Intuition



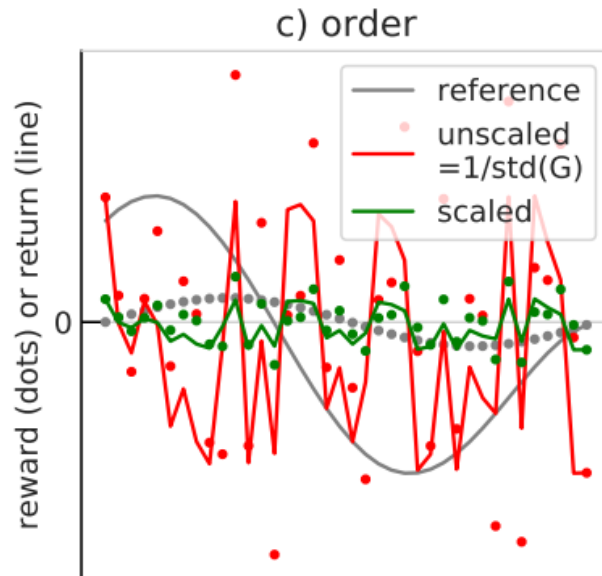
- when rewards are offset additively instead of multiplicatively
- their variance is identical, but the scale of returns can vary significantly

✓ Reward (dot)  
✓ Return (lines)

✓ Red (unscaled)  
✓ Green (scaled: proposed)  
✓ Gray (reference)  
✓ Cyan (false friend) : minimalist ideas that quickly come to mind often look tempting, but fail in others



### 3. Intuition



- The scales **of returns on their own** are not a sufficient characteristic of the desirable scale, **the sequential structure** matters as well.
- the comparison sequence has identical returns to the (smooth) reference, but in shuffled order (adjusting rewards to produce such a sequence):

✓ Reward (dot)  
✓ Return (lines)

✓ Red (unscaled)  
✓ Green (scaled: proposed)  
✓ Gray (reference)  
✓ Cyan (false friend) : minimalist ideas that quickly come to mind often look tempting, but fail in others

### 3. Intuition

종합해서 말하자면..

1. a reasonable scaling mechanism is needed to correct by the same factor
2. reward scales on their own are insufficient.
3. returns on their own are not a sufficient characteristic of the desirable scale, the sequential structure matters as well
4. Another dimension to take into account is the discount  $\gamma$ , which influences how reward scales relate to return scales

**we only consider simple linear rescaling methods that do not have any hyper-parameters (so that the tuning effort is not just shifted) and can be quickly implemented in a wide range of agents**

## 4. Return-based scaling

- we only consider simple linear rescaling methods that do not have any hyper-parameters (so that the tuning effort is not just shifted) and can be quickly implemented in a wide range of agents
- Our starting point is the temporal-difference (TD) error  $\delta$ , which takes the form

$$\delta_t = R_t + \gamma_t V'_{t+1} - V_t, \quad V_t := v(S_t) := \mathbb{E}[G_t | S_t]$$

- We propose to replace raw TD-errors  $\delta_t$  by a scaled version where  $\sigma \in \mathbb{R}^+$  is an adaptive scale factor

$$\bar{\delta}_t := \frac{\delta_t}{\sigma},$$

## 4. Return-based scaling

- For determining overall error scales, the regime of interest is the transient regime, long before convergence (what happens in early learning)
- the errors in the transient regime are the ones that characterise the problem
- We can write the variance of TD-errors in the following way

$$\begin{aligned}\mathbb{V}[\delta] &= \mathbb{V}[R + \gamma V' - V] \\ &= \mathbb{V}[R + \gamma(V' - V) - (1 - \gamma)V]\end{aligned}$$

**We assume the (approximate) independence relation, in early learning (i.e., the transient regime)**

$$R \perp\!\!\!\perp \gamma(V' - V) \perp\!\!\!\perp (1 - \gamma)V$$

$$\mathbb{V}[\delta] \approx \mathbb{V}[R] + \mathbb{V}[\gamma(V' - V)] + \mathbb{V}[(1 - \gamma)V]$$

## 4. Return-based scaling

- The variance of a product of independent variables obeys

$$\begin{aligned}
 \mathbb{V}[XY] &= \mathbb{E}[X]^2\mathbb{V}[Y] + \mathbb{V}[X]\mathbb{E}[Y]^2 + \mathbb{V}[X]\mathbb{V}[Y] \\
 &= \mathbb{E}[X]^2\mathbb{V}[Y] + \mathbb{V}[X](\mathbb{E}[Y]^2 + \mathbb{V}[Y]) \\
 &= \mathbb{E}[X]^2\mathbb{V}[Y] + \mathbb{V}[X]\mathbb{E}[Y^2]
 \end{aligned}$$

$$\bar{\gamma} := \mathbb{E}[\gamma]$$

$$\mathbb{V}[\delta] \approx \mathbb{V}[R] + \mathbb{V}[\gamma(V' - V)] + \mathbb{V}[(1 - \gamma)V]$$

$$\begin{aligned}
 \mathbb{V}[\delta] \approx & \mathbb{V}[R] + \bar{\gamma}^2\mathbb{V}[V' - V] + \mathbb{V}[\gamma]\mathbb{E}[(V' - V)^2] \\
 & + (1 - \bar{\gamma})^2\mathbb{V}[V] + \mathbb{V}[\gamma]\mathbb{E}[V^2], \quad (4)
 \end{aligned}$$

(Note that, perhaps unconventionally, we treat  $\gamma$  as a random variable here)

## 4. Return-based scaling

- It is reasonable to assume that values take on similar overall scales to returns, very early in learning. So to a first approximation, return statistics can take the place of value statistics

$$\mathbb{V}[V] \approx \mathbb{V}[G]$$

$$\mathbb{E}[V^2] \approx \mathbb{E}[G^2]$$

$$\begin{aligned} \mathbb{V}[\delta] \approx & \mathbb{V}[R] + \bar{\gamma}^2 \mathbb{V}[V' - V] + \mathbb{V}[\gamma] \mathbb{E}[(V' - V)^2] \\ & + (1 - \bar{\gamma})^2 \mathbb{V}[V] + \mathbb{V}[\gamma] \mathbb{E}[V^2], \end{aligned} \quad (4)$$



$$\begin{aligned} \mathbb{V}[\delta] \approx & \mathbb{V}[R] + \bar{\gamma}^2 \mathbb{V}[G' - G] + \mathbb{V}[\gamma] \mathbb{E}[(G' - G)^2] \\ & + (1 - \bar{\gamma})^2 \mathbb{V}[G] + \mathbb{V}[\gamma] \mathbb{E}[G^2] \end{aligned} \quad (5)$$

## 4. Return-based scaling

- One way to approximate the statistics of one-step differences is to use analogous (approximate) independence

$$\begin{aligned} \mathbb{V}[\delta] \approx & \mathbb{V}[R] + \bar{\gamma}^2 \mathbb{V}[G' - G] + \mathbb{V}[\gamma] \mathbb{E}[(G' - G)^2] \\ & + (1 - \bar{\gamma})^2 \mathbb{V}[G] + \mathbb{V}[\gamma] \mathbb{E}[G^2] \end{aligned} \quad (5)$$

$$G' - G = R - (1 - \gamma)G$$

$$\begin{aligned} \mathbb{E}[(G' - G)^2] &= \mathbb{E}[(R - (1 - \gamma)G)^2] \\ &\approx \mathbb{E}[R^2] + (1 - \bar{\gamma})^2 \mathbb{E}[G^2] \\ &\quad - 2(1 - \bar{\gamma}) \mathbb{E}[R] \mathbb{E}[G] \\ &\approx \mathbb{E}[R^2] + (1 - \bar{\gamma})^2 \mathbb{E}[G^2] \\ &\quad - (1 - \bar{\gamma})^2 \mathbb{E}[G]^2 - \mathbb{E}[R]^2 \\ &= \mathbb{V}[R] + (1 - \bar{\gamma})^2 \mathbb{V}[G] \end{aligned}$$

## 4. Return-based scaling

- last approximation uses  $\mathbb{V}[G' - G] \approx \mathbb{V}[R] + (1 - \bar{\gamma})^2 \mathbb{V}[G] + \mathbb{V}[\gamma] \mathbb{E}[G^2]$ .

$$\begin{aligned}
 \mathbb{V}[\delta] &\approx \mathbb{V}[R] + \bar{\gamma}^2 \mathbb{V}[R] + \bar{\gamma}^2 (1 - \bar{\gamma})^2 \mathbb{V}[G] \\
 &\quad + \bar{\gamma}^2 \mathbb{V}[\gamma] \mathbb{E}[G^2] + \mathbb{V}[\gamma] \mathbb{V}[R] \\
 &\quad + \mathbb{V}[\gamma] (1 - \bar{\gamma})^2 \mathbb{V}[G] \\
 &\quad + (1 - \bar{\gamma})^2 \mathbb{V}[G] + \mathbb{V}[\gamma] \mathbb{E}[G^2] \\
 &= (1 + \bar{\gamma}^2 + \mathbb{V}[\gamma]) \mathbb{V}[R] \\
 &\quad + (1 + \bar{\gamma}^2 + \mathbb{V}[\gamma]) (1 - \bar{\gamma})^2 \mathbb{V}[G] \\
 &\quad + (1 + \bar{\gamma}^2) \mathbb{V}[\gamma] \mathbb{E}[G^2] \\
 &\approx \mathbb{V}[R] + (1 - \bar{\gamma})^2 \mathbb{V}[G] + \mathbb{V}[\gamma] \mathbb{E}[G^2] \\
 &\approx \mathbb{V}[R] + \mathbb{V}[\gamma] \mathbb{E}[G^2]
 \end{aligned}$$



## 4. Return-based scaling

Return-based scaling

- raw TD-errors  $\delta_t$  by a scaled version,  $\sigma \in \mathbb{R}^+$  is an adaptive scale factor

$$\bar{\delta}_t := \frac{\delta_t}{\sigma}, \quad \sigma^2 := \mathbb{V}[R] + \mathbb{V}[\gamma] \mathbb{E}[G^2] \approx \mathbb{V}[\delta].$$

- this is sufficient to satisfactorily address all of the scenarios discussed  
its three components have sufficient information
  1. reward scale
  2. discounting
  3. offset
  4. etc.

## 4. Return-based scaling

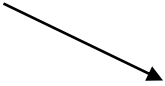
### Return-based scaling Implementation

- In keeping with our aim of not adding any hyper-parameters, we propose to estimate the statistics based on all data the agent has ever seen  $\mathbb{V}[R]$ ,  $\mathbb{V}[\gamma]$  and  $\mathbb{E}[G^2]$
- This is a conservative approach that we prefer for robustness and simplicity

$$\bar{\delta}_t := \frac{\delta_t}{\sigma}, \quad \sigma^2 := \mathbb{V}[R] + \mathbb{V}[\gamma]\mathbb{E}[G^2] \approx \mathbb{V}[\delta].$$

- Specifically, we use  $\bar{\delta}_t := \delta_t / \max(\sigma, \sigma_V, \sigma_{\text{batch}})$ :

it resolves the stability issue for this edge case

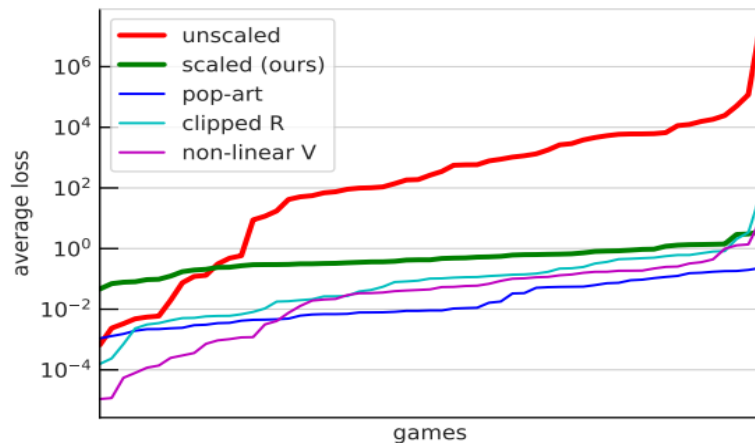


the moment a reward arrives that is much larger than all previously encountered ones

## 4 Return-based scaling

- Note how return-based scaling brings all loss scales into a narrow band

### In 57 Atari game averaged loss



- ✓ Red (unscaled)
- ✓ Green (scaled: proposed)
- ✓ Else: alternative scaling

alternative scaling methods,  
each of which spans a wider range

*Figure 3.* Loss scales across 57 Atari games, when using return-based scaling (green) or not (red). Data is averaged across training, and the 57 per-game averages are sorted before plotting, i.e., a handful of games have (unscaled) average losses below  $10^{-2}$ , as well as a handful above  $10^4$ . Note how return-based scaling brings all loss scales into a narrow band. Thin lines show the corresponding results for three commonly used alternative scaling methods, each of which spans a wider range.

## 4. Proposal: Return-based scaling

- We find that return-based scaling has a massive benefit for the 10-head setup for any metric considered
- but is on par with the unscaled baseline in the 1-head setup

Head: neural net output 개수

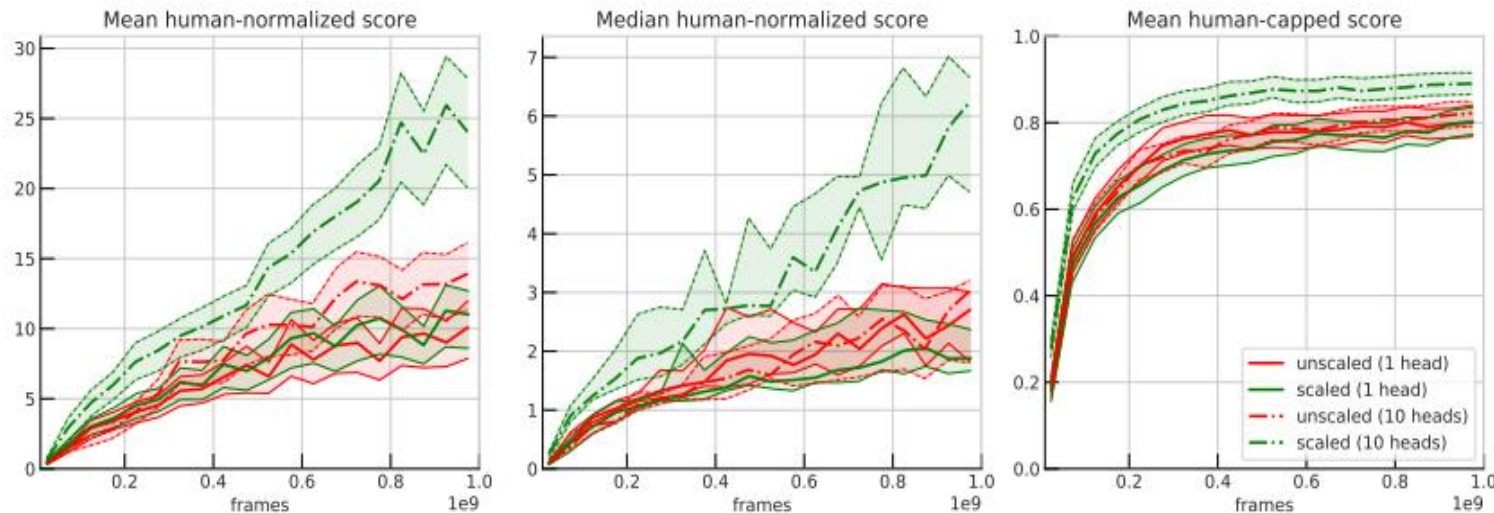


Figure 4. Aggregate performance results across 57 Atari games (1 seed), see Figure 15 (appendix) for per-game details. The four variants show unscaled (red) and scaled (green) results, for both the single head (solid lines) and 10-head (dash-dotted lines) scenarios. Shaded areas indicate inter-quartile ranges computed via bootstrap sampling the games (indicating sensitivity to scores in individual games). We find that return-based scaling has a massive benefit for the 10-head setup, for any metric considered, but is on par with the unscaled baseline in the 1-head setup. Also, the 10-head setup demonstrates its benefit over the single-head one only when the losses of the different heads are appropriately balanced via return-based scaling, but collapses to essentially 1-head performance otherwise.

## 4. Return-based scaling

