# Notes on Reward Scaling Laws

Faraz Rahman

January 3, 2026

## Contents

## 1 Intro

The goal of these notes are to provide some insight on generating / analyzing scaling curves of deep neural networks in environments (i.e. multistep tasks) where we are interested in “S”-shaped metrics like reward/value/returns which have upper and lower bounds.

Some common challanges in this setting are:

- Rewards vs. ________ follow an “S”-curve which is harder to paremetrize than e.g. $\log(\ell_{\text{val}})$ which is linear
- Rollouts are expensive to run, and various scales may require differing number of trials to acheive statistical certainty. Optimal compute allocation can accelerate iteration speed.
- Error is in reward is non symmetric - if average reward is near zero, then your lower-uncertainty should be small.

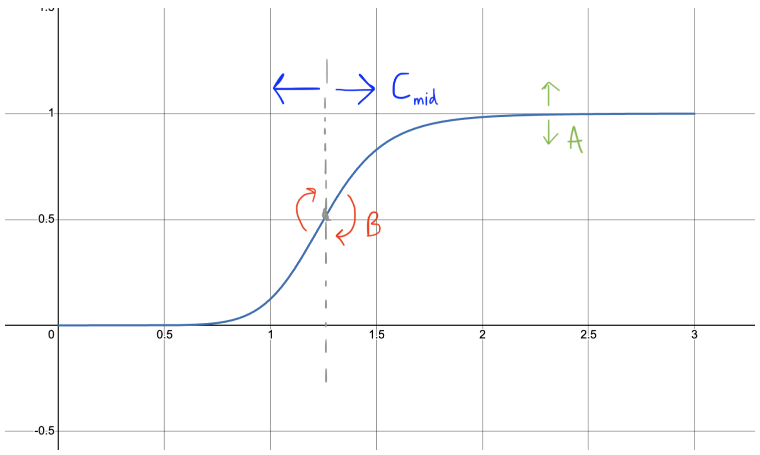
These notes will primarily centralize tools and code as well as statistical justification from existing work such as Agarwal et al. (2022), Khatri et al. (2025), and Snyder et al. (2025).

## 2 Sigmoidal Scaling Curves

Rewards are typically bounded $[0, 1]$. As a result, we should not expect scaling laws to be linear on a log scale. Khatri et al. (2025) advise using a sigmoidal curve with the following parametrization.

$$R(C) = A \cdot \frac{1}{1 + (C_{\text{mid}}/C)^B} \tag{1}$$

Where $R(C)$ is the gain in reward (i.e. $R(0) = 0$). There are three parameters: $A$ is the asymptotic reward gain, $C_{\text{mid}}$ is the compute budget that defines the midpoint of the performance curve and $B$ is the scaling exponent. Below is a visualization, an interactive Desmos plot can be found https://www.desmos.com/calculator/rctmkfophe.



Khatri et al. (2025) find fitting all three parameters to be hard in general so they do a grid search over $A, C_{\text{mid}}$ while $B$ is fit and select $A, C_{\text{mid}}$ that minimizes the $\ell_2$ error. They provide code to help at www.devvrit.com/scalerl_curve_fitting.

Sigmoidal Scaling Curves look like a power law for large $C$. This is discussed in appendix section A.4 from Khatri et al. (2025).

Applying the inverse sigmoid transform should also allow you to linearize the reward scaling curve. I have not seend this applied in the literature yet though and will update this note when I get a chance to try it. The inverse sigmoid transform is

$$\sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right)$$

Applying this to our functional form when $A = 1$ we get

$$\sigma^{-1}(R(C)) = \underbrace{-B \log(C_{\text{mid}})}_{\text{constant}} + B \log(C)$$

which is linear when the $C$ axis is plotted on a log-scale.

## 3 Error Bars

Good error-bars can depdend on the number of seeds available for some point on the x-axis. In the 3-5 seed range, Agarwal et al. (2022) reccommend using the Interquartile Mean (IQM) as measure of the center and using the Interquartile Range (IQR) to provide a rough sense of range. When 10-20 seeds are available other methods such as bootstrap-refit-take-quantiles can be useful to provide confidence bands arounda parametric fit curve. Both of these measure of uncertainty are assymetric and will not fall into the issue of predicting "negative" accuracy.

## 4 Identifying if two configs provide approximately the same value

This is useful if you have two scaling curves and want to know if both reach approximately the same final reward or if one configuration is actually better. Check out this https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Paired_T-Test_for_Equivalence.pdf

## References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A., and Bellemare, M. G. (2022). Deep reinforcement learning at the edge of the statistical precipice.

Khatri, D., Madaan, L., Tiwari, R., Bansal, R., Duvvuri, S. S., Zaheer, M., Dhillon, I. S., Brandfonbrener, D., and Agarwal, R. (2025). The art of scaling reinforcement learning compute for llms.

Snyder, D., Hancock, A. J., Badithela, A., Dixon, E., Miller, P., Ambrus, R. A., Majumdar, A., Itkina, M., and Nishimura, H. (2025). Is your imitation learning policy better than mine? policy comparison with near-optimal stopping.