Replicating the Adam optimizer paper

Graham Burgess Thomas Newman Tamás P. Papp Drupad Parmar
Jack Trainer

24–25 April 2023

1 Introduction

We aimed to replicate the experiments in the Adam (stochastic optimization algorithm) paper Kingma and Ba (2015). The Adam algorithm is clearly explained and the empirical results in Kingma and Ba (2015) show that it clearly outperforms its competitors. However, a number of design choices in the paper are not made explicit, which make the experiments in the paper difficult to replicate.

2 Challenges in replicating the paper

1. There is undefined notation in the paper – although this can be inferred.
   - In the reformulation of the optimization algorithm, the parameter ε (which guards against numerical errors introduced by extremely small estimates of the second moment) is replaced by a parameter $\hat{\varepsilon}$ which is undefined in the paper. The choice which makes the two formulations equivalent is $\hat{\varepsilon} = \varepsilon(1 - \beta_2^t)$.
2. There is no author source code.
3. The hyperparameters used in the experiments are not explicit – both for the optimizer and for the models. The authors state that these have been hand-optimized for each algorithm, but do not report their values.
   - An additional source of confusion is as follows. The authors specify fixed default values for the parameters in their Algorithm 1, e.g. the learning rate is defaulted to $\alpha = 0.001$. However, in both the theory and the experiments in Kingma and Ba (2015, Section 6.1), the learning rate is time-varying $\alpha_t = \alpha/\sqrt{t}$.
4. There is no description of the initialization for the optimization, other than it being consistent across all methods.
5. The method and experiments are inherently stochastic, so the results can only be reproduced up to Monte Carlo error.

3 Attempted experiments

We made an attempt to replicate the logistic regression example on the MNIST (digit classification) dataset (Kingma and Ba, 2015, Section 6.1). We used built-in implementations of the model and the Adam algorithm in three software packages, hereafter referred to as "frameworks":

1. `Keras` (with a `Tensorflow` backend), in Python.

2. `Keras` (with a `Tensorflow` backend), in R.
3. `Jax` (with `Flax` models and `Optax` optimizers), in Python.

Experiments were run in all frameworks. Due to lack of time, ambiguity of documentation, and our general unfamiliarity with the aforementioned machine learning frameworks, it is possible that our coded models and optimizer, do not exactly match the textual description in (Kingma and Ba, 2015, Section 1). Code is nonetheless available on the workshop Github, which attempts to replicate parts of Kingma and Ba (2015, Figure 1 left).

Discrepancy with Kingma and Ba (2015, Figure 1) In frameworks (1,3), we found a discrepancy between our training loss and the "training cost" used in Kingma and Ba (2015, Figure 1 left). Ours was a factor of ≈ 10 larger – which happens to coincide with the number of class labels (digits 1-10) – although we found both loss curves to be similar in shape. The ambiguous notion of "training cost" in Kingma and Ba (2015, Figure 1) therefore makes the figure difficult to replicate.

References

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*