

Are DeepBigWideNets provably better than linear predictors?

author names withheld

Under Review for OPT 2025

Abstract

Recently, a transformer network (DeepBigWideNets-1) has been shown to outperform linear predictors, in the sense that all its local minima are at least as good as the best linear predictor. We take a step towards extending this result to deep DeepBigWideNets-2.

1. Main

Main text limited to 6 pages, excluding references and appendices.

1.1. Citations (`natbib` package)

?] did preliminary studies in the field, their results were reproduced in [? , Theorem 1], see also [?] or [see also ?].

1.2. Algorithms (`algorithm2e` package)

For how to use this package, check out the `doc`. You are free to choose another package, add line numbering, etc.

Algorithm 1: Example Algorithm

Data: this text

Result: how to write algorithm with $\text{\LaTeX}2\text{\epsilon}$ initialization;

```
while not at end of this document do
  read current;
  if understand then
    go to next section;
    current section becomes this one;
  else
    go back to the beginning of current section;
  end
end
```

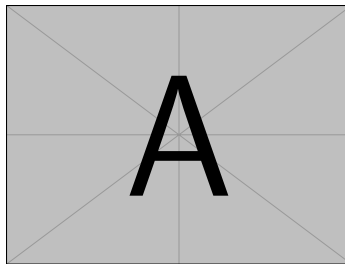


Figure 1: This is a figure.