NLU Course Project – Language Model Experiments

Name Surname (mat. number)

University of Trento

name.surname@studenti.unitn.it

1. Introduction

I conducted two complementary experiments on Penn Treebank Dataset (PTB) to evaluate methods for reducing test perplexity (PPL). In **Part A**, I began with a single-layer LSTM (embedding size = 650, hidden size = 650) trained with vanilla SGD, then added dropout layers and finally switched from SGD to AdamW for the optimizer. Each modification was applied incrementally, retraining to convergence with early stopping and hyperparameter tuning. In **Part B**, I built on that LSTM backbone by integrating three techniques from Merity [1]: (i) *weight tying* between embedding and output projections; (ii) *variational dropout* at the embedding and output stages; and (iii) *non-monotonically triggered Averaged SGD* (NT-AvSGD). I compared four configurations to measure their impact on final test PPL: two where I initially trained with SGD and switched to AvSGD when the non-monotonic trigger activated, and two where I started with AdamW and then switched to AvSGD when triggered. Each model was evaluated under the same PTB splits, employing early stopping on validation PPL with a patience of 5 epochs.

2. Implementation details

2.1. Part A

My baseline model is a single-layer LSTM with an embedding dimension of 650 and hidden state size of 650. I trained it using SGD (learning rate = 5, gradient clipping = 0.25). The vocabulary was built from PTB training text, and I appended an end-of-sentence token (`<eos>`). After verifying a base test PPL of approximately 131.4, I inserted dropout layers: a 0.2 dropout on the embedding outputs and a 0.3 dropout before the final linear projection. Introducing dropout reduced overfitting significantly, yielding a test PPL of about 110.61.

Next, I replaced SGD with AdamW, using a learning rate of 0.001. I found that AdamW stabilized training earlier: validation PPL plateaued around epoch 25 (versus epoch 30 for SGD). Early stopping was triggered when validation PPL did not improve for five consecutive epochs. Under AdamW, the final test PPL decreased slightly to 110.19, demonstrating a modest gain from the adaptive optimizer.

2.2. Part B

All experiments in Part B share a single-layer LSTM backbone (no bidirection), with weight tying applied only when embedding and hidden sizes match. I used embedding/hidden dimensions of 800 or 950. Variational dropout was implemented (one dropout mask per sequence), following the method proposed in [2], at two points: immediately after the embedding lookup and immediately before the linear classifier. The linear layer's weights are tied to the embedding matrix when embed = hidden, reducing parameter count and following Inan et al.

Training schedules varied as follows:

- **SGD-800**: Embedding = 800, hidden = 800; I started training with SGD (lr = 5, clip = 5), and switched to AvSGD (lr = 5) when the non-monotonic trigger was activated.
- **SGD-950**: Same as SGD-800 but with embedding/hidden size = 950.
- **AdamW-1**: Embedding/hidden = 950; initial training with AdamW (lr = 0.005, clip = 5); then switch to AvSGD (lr = 1) when the non-monotonic trigger (NT threshold = 5) was activated, following the triggering strategy described in Merity et al. [1].
- **AdamW-5**: Identical to AdamW-1 but the AvSGD phase uses lr = 5 instead of 1.

For all four runs, I used a mini-batch size of 64 for training and 128 for dev and test sets, padding/tokenization as per PTB conventions, and applied gradient clipping at a norm of 5. Early stopping monitored validation PPL with a patience of 5 epochs.

3. Results

Table 1 reports Part A test PPLs. The baseline LSTM trained with SGD (no dropout) reached 131.41. Introducing dropout reduced PPL to 110.61, confirming dropout's strong regularization effect. Switching to AdamW gave a further modest gain (final PPL 110.19), as AdamW stabilized weights during later epochs, though its effect was smaller than that of dropout alone.

Table 1: *Part A: Test Perplexities*

Experiment	**Test PPL**
LSTM base	131.41
LSTM + dropout layers	110.61
LSTM + dropout layers + AdamW optimizer	110.19

Table 2 shows Part B results. Both SGD-800 and SGD-950 runs achieved lower test PPL than any configuration evaluated in Part A, at 102.20 and 99.85, respectively. The **AdamW-1** configuration—leveraging AdamW then NT-AvSGD (lr = 1)—yielded the best PPL of 95.91, a further improvement over SGD-950. The **AdamW-5** run (AvSGD lr = 5) achieved 98.84, showing that a lower AvSGD learning rate (lr = 1) is preferable under NT triggering. Training for this configuration required nearly 100 epochs to converge, reaching the maximum number of allowed epochs rather than stopping early due to the patience criterion. This is in contrast to the other configurations, which all converged within 50–80 epochs. The slower convergence is illustrated by the gradual flattening of the validation perplexity curve in Figure 1. These results confirm that weight tying and variational dropout reduce model complexity and overfitting, while the AdamW+NT-AvSGD schedule offers additional gains through stabilization of the final weights.

Table 2: *Part B: Test Perplexities with Regularization*

Experiment	**Test PPL**
SGD-800	102.20
SGD-950	99.85
AdamW-1	95.91
AdamW-5	98.84

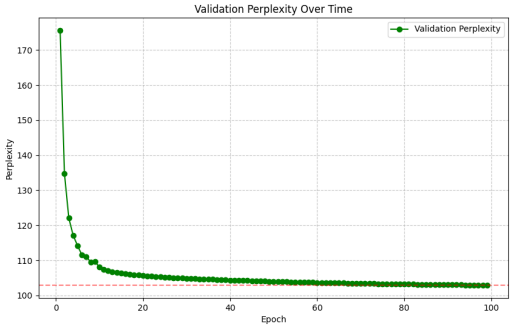

Figure 1: *AdamW-1: Validation perplexity over training epochs.*

4. References

[1] N. S. K. . R. S. Stephen Merity, "Regularizing and optimizing lstm language models," 2018.

[2] Z. G. Yarin Gal, "A theoretically grounded application of dropout in recurrent neural networks," 2016.