Supplementary Material for Adaptively Truncating BPTT to Control Gradient Bias

Christopher Aicher[1] Nicholas J. Foti[2] Emily B. Fox[1,2]

[1] Department of Statistics, University of Washington
[2] Paul G. Allen School of Computer Science and Engineering, University of Washington
{aicherc, nfoti, ebfox}@uw.edu

We start the supplement from 'B' to avoid confusion between equation numbering and assumption numbering in the main text for ($A\#$).

B Proofs for Section 3

B.1 Proof of Theorem 1

The proof of Theorem 1 consists of two part. First we bound the absolute bias by $\mathcal{E}(K,\theta)$ using assumptions (A-1) and (A-2). Then we bound the relative bias using the triangle inequality

Proof of Theorem 1. The bias of $\hat{g}_K$ is bounded by the expected error between $\hat{g}_K$ and $\hat{g}_T$

$$\|\mathbb{E}\left[\hat{g}_K(\theta)\right] - g(\theta)\| = \|\mathbb{E}\left[\hat{g}_K(\theta) - \hat{g}_T(\theta)\right]\| \leq \mathbb{E}\left[\|\hat{g}_K(\theta) - \hat{g}_T(\theta)\|\right] \quad . \tag{B.1}$$

Applying the triangle-inequality to the difference between $\hat{g}_K$ and $\hat{g}_T$ gives

$$\|\hat{g}_K(\theta) - \hat{g}_T(\theta)\| = \left\| \sum_{k=K+1}^{s} \frac{\partial \mathcal{L}_s}{\partial h_{s-k}} \cdot \frac{\partial h_{s-k}}{\partial \theta} \right\| \leq \sum_{k=K+1}^{s} \left\| \frac{\partial \mathcal{L}_s}{\partial h_{s-k}} \right\| \cdot \left\| \frac{\partial h_{s-k}}{\partial \theta} \right\| \leq \sum_{k=K+1}^{s} \left\| \frac{\partial \mathcal{L}_s}{\partial h_{s-k}} \right\| \cdot M \;\; , \tag{B.2}$$

where in the last inequality we apply the assumption (A-2), $\|\partial h_t / \partial \theta\| < M$ for all t. Taking the expectation with respect to s of both sides of Eq. (B.2) gives

$$\mathbb{E}\left[\|\hat{g}_K(\theta) - \hat{g}_T(\theta)\|\right] \leq \sum_{k=K+1}^{s} \mathbb{E} \left\| \frac{\partial \mathcal{L}_s}{\partial h_{s-k}} \right\| \cdot M = \sum_{k=K+1}^{s} \mathbb{E}\left[\phi_k\right] \cdot M \;\; , \tag{B.3}$$

where we recall that $\phi_k = \|\partial \mathcal{L}_s / \partial h_{s-k}\|$.

Recursively applying assumption (A-2) to $\mathbb{E}\left[\phi_{\tau+t}\right]$ gives

$$\mathbb{E}\left[\phi_{\tau+t}\right] \leq \beta \cdot \mathbb{E}\left[\phi_{\tau+t-1}\right] \leq \ldots \leq \beta^t \cdot \mathbb{E}\left[\phi_\tau\right] \;\; . \tag{B.4}$$

Combining Eqs. (B.1), (B.3), and (B.4) gives us the first half of the result

$$\|\mathbb{E}\left[\hat{g}_K(\theta)\right] - g(\theta)\| \leq \mathbb{E}\left[\|\hat{g}_K(\theta) - \hat{g}_T(\theta)\|\right] \leq \sum_{k=K+1}^{s} \mathbb{E}\left[\phi_k\right] \cdot M = \mathcal{E}(K,\theta) \;\; . \tag{B.5}$$

To bound the relative error, we apply the reverse triangle inequality to $\|g(\theta)\|$

$$\|g(\theta)\| \geq \|\mathbb{E}\left[\hat{g}_K(\theta)\right]\| - \|\mathbb{E}\left[\hat{g}_K(\theta)\right] - g(\theta)\| \geq \|\mathbb{E}\left[\hat{g}_K(\theta)\right]\| - \mathcal{E}(K,\theta) \;\; , \tag{B.6}$$

when $\|\mathbb{E}\left[\hat{g}_K(\theta)\right]\| - \mathcal{E}(K,\theta) > 0$.

Since $\mathcal{E}(K,\theta)$ is an upper bound for the numerator and $\|\mathbb{E}\left[\hat{g}_K(\theta)\right]\| - \mathcal{E}(K,\theta)$ is a lower bound for the denominator, we obtain the result

$$\frac{\|\mathbb{E}\left[\hat{g}_K(\theta)\right] - g(\theta)\|}{\|g(\theta)\|} \leq \frac{\mathcal{E}(K,\theta)}{\|\mathbb{E}\left[\hat{g}_K(\theta)\right]\| - \mathcal{E}(K,\theta)} = \delta(K,\theta) \ . \tag{B.7}$$

□

B.2 Proof of Theorem 2

Let $\langle x_1, x_2 \rangle$ denote the inner-product between two vectors.

We first presents some Lemmas involving $\hat{g}(\theta)$ and $g(\theta)$ when the gradient has bounded relative bias δ.

Lemma 1. *If $\hat{g}(\theta)$ has bounded relative bias of δ then*

$$\mathbb{E}\left\langle g(\theta), \hat{g}(\theta) - g(\theta)\right\rangle \leq \delta\|g(\theta)\|^2 \quad \textit{and} \quad \mathbb{E}\left\langle g(\theta), \hat{g}(\theta)\right\rangle \geq (1-\delta)\|g(\theta)\|^2 \tag{B.8}$$

Proof of Lemma 1. The first inequality follows from the Cauchy-Schwartz inequality and bound on relative bias

$$\mathbb{E}\left\langle g(\theta), \hat{g}(\theta) - g(\theta)\right\rangle = \left\langle g(\theta), \mathbb{E}\left[\hat{g}(\theta)\right] - g(\theta)\right\rangle \leq \|g(\theta)\|\|\mathbb{E}\left[\hat{g}(\theta) - g(\theta)\right\| \leq \delta\|g(\theta)\|^2 \ . \tag{B.9}$$

The second inequality follows immediately from the first

$$\langle g(\theta), \hat{g}(\theta)\langle = \langle g(\theta), g(\theta)\rangle + \langle g(\theta), \hat{g}(\theta) - g(\theta)\rangle \leq \|g(\theta)\|^2 - \delta\|g(\theta)\|^2 = (1-\delta)\|g(\theta)\|^2 \ . \tag{B.10}$$

□

The next lemma bounds the second moment of $\|\hat{g}(\theta)\|$.

Lemma 2. *If $\hat{g}$ has bounded relative bias δ and bounded variance σ^2 for all θ (assumption (A-4)), then*

$$\mathbb{E}\left[\|\hat{g}\|^2\right] \leq (1+\delta)^2\|g\|^2 + \sigma^2 \ . \tag{B.11}$$

Proof of Lemma 2.

$$\|\hat{g}\|^2 = \|g\|^2 + 2\langle g, \hat{g} - g\rangle + \|\hat{g} - g\|^2 \tag{B.12}$$

Take the expectation, we obtain the result

$$\mathbb{E}\|\hat{g}\|^2 = \|g\|^2 + 2\mathbb{E}\left\langle g, \hat{g} - g\right\rangle + \mathbb{E}\|\hat{g} - g\|^2 \ , \tag{B.13}$$

where expand the mean-squared error into the bias squared plus variance

$$\mathbb{E}\|\hat{g} - g\|^2 = \|\mathbb{E}\,\hat{g} - g\|^2 + \mathbb{E}\|\hat{g} - \mathbb{E}\,\hat{g}\|^2 \leq \delta^2\|g\|^2 = \sigma^2 \ . \tag{B.14}$$

Therefore

$$\mathbb{E}\|\hat{g}\|^2 \leq \|g\|^2 + 2\delta\|g\|^2 + (\delta^2\|g\|^2 + \sigma^2) = (1+\delta)^2\|g\|^2 + \sigma^2 \tag{B.15}$$

□

We now begin the proof of Theorem 2 which builds off the proof in [Ghadimi and Lan, 2013].

Proof of Theorem 2. From the L-smoothness of $\mathcal{L}$, assumption (A-3), we have

$$\mathcal{L}(\theta) - \mathcal{L}(\theta') - |\langle g(\theta), \theta - \theta'\rangle| \leq \frac{L}{2}\|\theta' - \theta\|^2, \quad \forall\theta,\theta' \ . \tag{B.16}$$

Substituting $\theta = \theta_{n+1}$ and $\theta' = \theta_n$, where θ_{n+1} and θ_n are connected through SGD Eq. (10), we obtain

$$\mathcal{L}(\theta_{n+1}) \leq \mathcal{L}(\theta_n) + \langle g(\theta_n), \theta_{n+1} - \theta_n \rangle + \frac{L}{2}\|\theta_{n+1} - \theta_n\|^2 \tag{B.17}$$

$$= \mathcal{L}(\theta_n) - \gamma_n \langle g(\theta_n), \hat{g}(\theta_n) \rangle + \frac{L}{2}\gamma_n^2 \|\hat{g}(\theta_n)\|^2 \ . \tag{B.18}$$

Taking the expectation with respect to $\hat{g}(\theta_n)$ on both sides and using Lemmas 1 and 2 gives us

$$\mathbb{E}\,\mathcal{L}(\theta_{n+1}) = \mathcal{L}(\theta_n) - \gamma_n \mathbb{E}\,\langle g(\theta_n), \hat{g}(\theta_n) \rangle + \frac{L}{2}\gamma_n^2 \mathbb{E}\,\|\hat{g}(\theta_n)\|^2 \tag{B.19}$$

$$\leq \mathcal{L}(\theta_n) - \gamma_n (1-\delta)\|g(\theta_n)\|^2 + \frac{L}{2}\gamma_n^2((1+\delta)^2\|g(\theta_n)\|^2 + \sigma^2) \ . \tag{B.20}$$

Rearranging terms with γ_n gives

$$\frac{\gamma_n(1-\delta)}{2}\left(2 - \gamma_n \frac{L(1+\delta)^2}{(1-\delta)}\right) \cdot \|g(\theta_n)\|^2 \leq \mathcal{L}(\theta_n) - \mathbb{E}\,\mathcal{L}(\theta_{n+1}) + \gamma_n^2 \frac{L\sigma^2}{2} \ . \tag{B.21}$$

As we assume the stepsizes are $\gamma_n < \frac{1-\delta}{L(1+\delta)^2}$, therefore $(2 + \gamma_n \frac{L(1+\delta)^2}{(1-\delta)}) < 1$ and we can drop these terms. Taking the summation over n and taking the expectation with respect to $\hat{g}(\theta_n)$ for $n = 1, \ldots, N$ we obtain

$$\sum_{n=1}^{N} \gamma_n \frac{(1-\delta)}{2} \cdot \min_{n \in [1, N+1]} \|g(\theta_n)\|^2 \leq \mathcal{L}(\theta_1) - \mathbb{E}\,\mathcal{L}(\theta_{N+1}) + \sum_{n=1}^{N} \gamma_n^2 \frac{L\sigma^2}{2} \ . \tag{B.22}$$

Finally, we divide both sides by $\sum_n \gamma_n \frac{1-\delta}{2}$ and apply $\mathbb{E}\,\mathcal{L}(\theta_{N+1}) \geq \min_{\theta^*} \mathcal{L}(\theta^*)$ to obtain the result

$$\min_{n \in [1, N+1]} \|g(\theta_n)\|^2 \leq \frac{2D_{\mathcal{L}} + L\sigma^2 \sum_{n=1}^{N} \gamma_n^2}{(1-\delta)\sum_{n=1}^{N} \gamma_n} \ , \tag{B.23}$$

where $D_{\mathcal{L}} = \mathcal{L}(\theta_1) - \min_{\theta^*} \mathcal{L}(\theta^*)$.

If we use a constant stepsize $\gamma_n = \gamma$ for all $n \in [1, N]$, then the optimal stepsize for N steps of SGD is

$$\gamma = \sqrt{\frac{2D_{\mathcal{L}}}{NL\sigma^2}} \quad \text{which achieves} \quad \min_{n \in [1, N+1]} \|g(\theta_n)\|^2 \leq \frac{1}{1-\delta} \cdot \sqrt{\frac{8D_{\mathcal{L}}L\sigma^2}{N}} \ . \tag{B.24}$$

If instead a decaying $\mathcal{O}(n^{-1/2})$ stepsize is used, then the numerator of Eq. (B.23) grows as a harmonic series $\mathcal{O}(\sum_n n^{-1}) = \mathcal{O}(\log n)$, while the denominator grows $\mathcal{O}(\sum_n n^{-1/2}) = \mathcal{O}(n^{1/2})$. Therefore the overall rate is $\mathcal{O}(n^{-1/2} \log n)$. □

B.3 Comparison of Bounds to [Chen and Luss, 2018]

In Section 3.3 for Theorem 2, we assume the *relative bias* is bounded, that is $\|\mathbb{E}\,[\hat{g}(\theta)] - g(\theta)\| \leq \delta\|g(\theta)\|$ for all θ (Eq. (11)). Chen and Luss [2018] prove similar results to Theorem 2, where they assume the relative error of each gradient is bounded in high probability, that is there exists $\delta, \epsilon > 0$ such that

$$\Pr(\|\hat{g}(\theta) - g(\theta)\| \leq \delta\|g(\theta)\|) > 1 - \epsilon \ , \text{ for all } \theta \ . \tag{B.25}$$

Although Markov's inequality implies that if the relative bias is bounded by $\delta \cdot \epsilon$, when Eq. (B.25) holds for δ, ϵ, their non-convex optimization results only hold in high probability rather than uniformly. A key drawback of their results, is that the relative error must be bounded in high probability for all steps of SGD ($\hat{g}_{1:N}$); therefore the required ϵ for each step depends on the total number of SGD steps during training [see Chen and Luss, 2018, Eq.(7) and Theorem 5]. Specifically, Chen and Luss [2018] observe that the probability the relative error is controlled for all N steps is bounded by $1 - \epsilon_{\text{total}} \leq (1-\epsilon)^N$ under the additional assumption that the noise in $\hat{g}(\theta)$ is independent. For their results to hold with probability $1 - \epsilon_{\text{total}}$ after N steps, each gradient must have a relative error bound with $\epsilon \leq 1 - (1 - \epsilon_{\text{total}})^{1/N}$. Chen and Luss [2018] achieve this by restricting $\epsilon \leq \epsilon_{\text{total}}/N$. Our result assumes the relative error is bounded in expectation, which sides steps this issue. However our results are not as robust in the sense that they do not hold if the noise in $\hat{g}(\theta)$ does not have an expected value (e.g. if $\hat{g}(\theta) - g(\theta)$ is Cauchy).

C Additional Experiments

This section provides additional tables and figures for the experiments in Section 5 as well as results on time series prediction with temporal point processes.

In our experiments, we selected the stepsize γ for SGD by performing a grid search over powers of 10 and selected the largest stepsize that did not diverge for fixed TBPTT (with $K = 15$ for the synthetic tasks, $K = 100$ for the language modeling tasks, and $K = 6$ for the temporal point process tasks). We also consider adaptive and decaying stepsizes (specifically ADADELTA, SGD with Momentum, and epoch-wise stepsize decay); however, we did not see a significant difference in results.

C.1 Additional Figures and Tables

C.1.1 Synthetic 'Copy' Experiment

Figure C.1 shows the validation PPL for the two experiments in Section 5.1. The left pair of figures show the validation PPL while the right pair shows the cumulative minimum (i.e. the 'best') validation PPL. The test PPL plots in Figure 1 are piecewise constant evaluated using these 'best' validation PPL parameters. The top row corresponds to the fixed-memory $m = 10$ copy experiment, and we see the loss decays relatively smoothly. The bottom row corresponds to the variable-memory $m \in [5, 10]$ copy experiment, and we see heavy oscillation in the validation error as it decays.

Table C.1 is a table of the test PPL results evaluated at the 'best' validation PPL. This table provides the numeric values of the 'best' PPL values for Figures 1 and C.1. We see that the adaptive TBPTT perform as well as or outperform the best fixed K TBPTT.

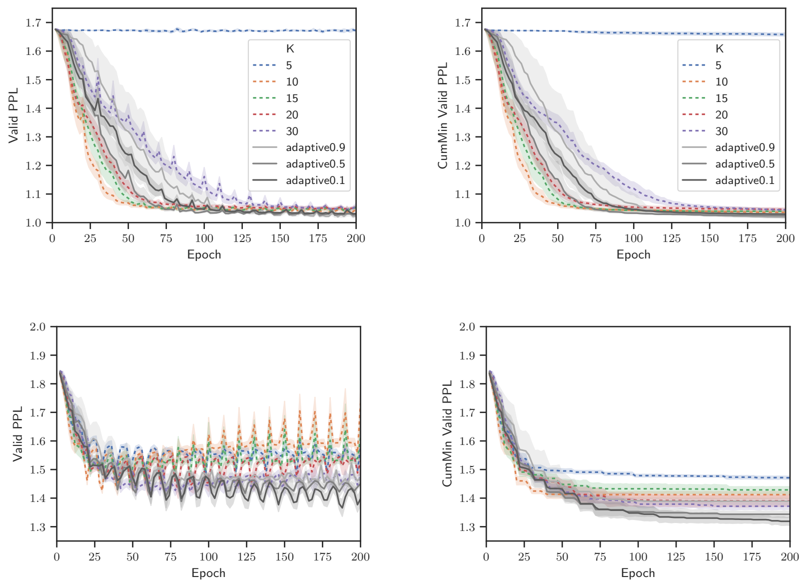

Figure C.1: Synthetic Copy Supplement: (left) Valid PPL vs epoch, (right) 'Best' Valid PPL vs epoch (Top) fixed $m = 10$, (bottom) variable $m \in [5, 10]$. Solid dark lines are our adaptive TBPTT methods, dashed colored lines are fixed TBPTT baselines.

Table C.1: Table of PPL for Synthetic Copy Experiments: (left) fixed $m = 10$, (right) variable $m \in [5, 10]$. 'Valid PPL' is the best validation set PPL. 'Test PPL' is the test set PPL at parameters of the best validation set PPL. Standard deviation over multiple initializations are in parentheses.

Fixed Copy $m = 10$

K	Valid PPL	Test PPL
5	1.655 (0.012)	1.646 (0.012)
10	1.035 (0.007)	1.036 (0.005)
15	1.038 (0.005)	1.039 (0.003)
20	1.045 (0.009)	1.040 (0.006)
30	1.044 (0.007)	1.043 (0.004)
$\delta = 0.9$	**1.018 (0.005)**	**1.022 (0.006)**
$\delta = 0.5$	**1.024 (0.003)**	**1.027 (0.002)**
$\delta = 0.1$	**1.029 (0.004)**	**1.030 (0.005)**

Variable Copy $m \in [5, 10]$

K	Valid PPL	Test PPL
5	1.46 (0.01)	1.47 (0.01)
10	1.41 (0.02)	1.39 (0.02)
15	1.39 (0.03)	1.37 (0.03)
20	1.39 (0.03)	1.35 (0.03)
30	**1.33 (0.02)**	**1.31 (0.01)**
$\delta = 0.9$	1.37 (0.02)	1.35 (0.02)
$\delta = 0.5$	**1.33 (0.01)**	**1.32 (0.02)**
$\delta = 0.1$	**1.31 (0.01)**	**1.29 (0.01)**

C.1.2 Language Modeling Experiment

Figure C.2 shows the validation PPL for the two language modeling experiments in Section 5.2. The left pair of figures show the validation PPL while the right pair shows the cumulative minimum (i.e. the 'best') validation PPL. The top row corresponds to the PTB experiment. We see that fixed TBPTT with small K quickly begins to over-fit (as the validation PPL increases). With larger K, fixed TBPTT achieves lower validation (and test) PPL, but requires more epochs. We see that the adaptive TBPTT with $\delta = 0.1$, achieves a better PPL much more rapidly. The bottom row corresponds to Wiki2 experiment, where we see that the adaptive TBPTT and best fixed TBPTT method perform similarly.

Table C.2 is a table of the test PPL results evaluated at the 'best' validation PPL. This table provides the numeric values of the 'best' PPL values for Figures 2 and C.2. We see that the adaptive TBPTT perform as well as or outperform the best fixed K TBPTT.

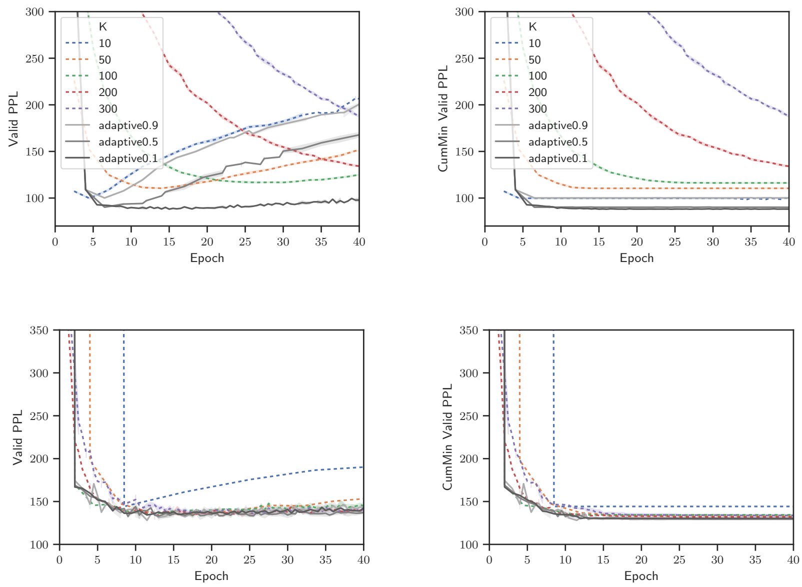

Figure C.2: Language Modeling Supplement: (left) Valid PPL vs epoch, (right) 'Best' Valid PPL vs epoch. (Top) PTB, (bottom) Wiki2. Solid dark lines are our adaptive TBPTT methods, dashed colored lines are fixed TBPTT baselines.

Table C.2: Table of PPL for Language Modeling experiments: (left) PTB, (right) Wiki2. 'Valid PPL' is the best validation set PPL. 'Test PPL' is the test set PPL at parameters of the best validation set PPL. Standard deviation over multiple initializations are in parentheses.

PTB			Wiki2		
K	Valid PPL	Test PPL	K	Valid PPL	Test PPL
10	99.7 (0.6)	99.9 (0.8)	10	144.2 (0.4)	136.5 (1.3)
50	110.4 (0.4)	110.8 (0.8)	50	133.4 (2.9)	127.2 (2.8)
100	116.2 (0.5)	116.9 (0.5)	100	134.4 (0.3)	127.8 (0.5)
200	125.2 (1.2)	126.1 (0.9)	200	130.3 (1.1)	124.6 (0.7)
300	161.5 (0.5)	161.2 (0.3)	300	**129.6 (1.4)**	**124.0 (2.2)**
$\delta = 0.9$	100.1 (0.5)	99.0 (0.5)	$\delta = 0.9$	**130.0 (1.3)**	**124.1 (2.2)**
$\delta = 0.5$	90.1 (0.4)	89.5 (0.3)	$\delta = 0.5$	**127.2 (0.7)**	**121.7 (0.6)**
$\delta = 0.1$	**88.1 (0.2)**	**87.2 (0.2)**	$\delta = 0.1$	**127.5 (0.6)**	**121.9 (1.2)**

C.2 Temporal Point Process Estimation

We now consider applying our adaptive TBPTT scheme to optimizing neural networks for temporal point prediction as in [Du et al., 2016]. Given a sequence $\{(y_i, t_i)_{i=1}^N\}$ of categorical observations $y_i \in \mathcal{Y}$ and observation times $t_i \in \mathbb{R}$, the task consider by [Du et al., 2016] is to predict (y_i, t_i) given $(y_j, t_j)_{j<i}$. Following [Du et al., 2016], we model the sequence using an RNN, with input embedding layers for y_{i-1} and t_{i-1}, and two output prediction layers: one for y_i and another for $\lambda(t)$ the *conditional temporal point process intensity*. The loss now consists of two terms, which define the negative log-likelihood (NLL) for a temporal point process: (i) cross entropy loss for y_i and (ii) a temporal point process loss for $\lambda(t_i)$ given by Eq.(12) in [Du et al., 2016]. [Du et al., 2016] also evaluate the neural network model by measuring the zero-one loss of the predicted observation $\hat{y}_i$ to y_i and the root mean-squared error (RMSE) of the mean predicted observation time $\hat{t}_i = \mathbb{E}[t_i|\lambda(t)]$ to t_i.

We fit such a model using a two-layer LSTM to the 'Book Order' financial data used in [Du et al., 2016]. For the input layers, we use an embedding of size 128 for the two state categorical observations y and a two dimensional encoding of t_i (i.e. $[t_i - t_{i-1}, t_i]$). For the two-layer LSTM, we use a hidden and cell state dimension of size 128. And the output layer dimensions follow [Du et al., 2016]. For training, we use a batchsize of $S = 64$ and a fixed learning rate of $\gamma = 0.1$ for SGD. We compare gradients from fixed TBPTT $K \in [3, 6, 9, 15, 21]$ and our adaptive TBPTT method $\delta \in [0.9, 0.5, 0.1]$. We set $W = 200$, $K_0 = 6$ and $[K_{\min}, K_{\max}] = [1, 100]$ for Algorithm 1.

The 'Book Order' dataset consists of the high-frequency financial transactions from the NYSE for a stock in one day. It consists of 0.7 million transactions records (in milliseconds) and the possible actions $\mathcal{Y}$ are 'to buy' or 'to sell'. We use the train-test split of [Du et al., 2016] and split their test set in half to form a validation set.

The results of our experiment in Figures C.3 and C.4 and Table C.3. From Figure C.3(bottom center-right and bottom right) we see that the adaptive methods control for bias, by slowly increasing K. From Figure C.3(top right) and Table C.3, we find that adaptive TBPTT methods achieve the best test set NLL. We also see from Figure C.3(bottom left and bottom center-left) and Table C.3 that fixed TBPTT $K = 3$ performs better at predicting y_i at the cost of increased error in predicting t_i. Similarly, fixed TBPTT $K = 15$ and $K = 21$ are better at predicting t_i, but poorer at predicting y_i.

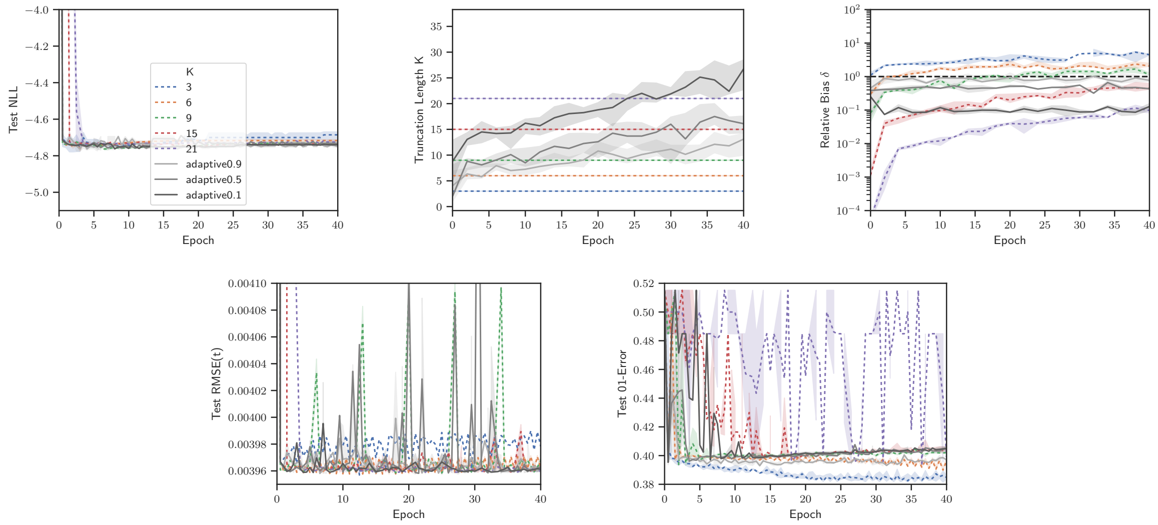

Figure C.3: Book Order Experiment. Top row: (left) Test NLL. (center) truncation length $\hat{\kappa}(\delta, \theta_n)$, (right) relative bias $\hat{\delta}(K)$. Bottom row: (left) Test RMSE for t, (right) Test 01-Error for y. Solid dark lines are our adaptive TBPTT methods, dashed colored lines are fixed TBPTT baselines.

Table C.3: Table of metrics for Book Order experiment. Test metrics are evaluated at the parameters of the best valdiation set NLL. Standard deviation over multiple initializations are in parentheses.

K	Valid NLL	Test NLL	RMSE(t) 10^{-3}	01-Loss(y)
3	-4.983 (0.013)	-4.694 (0.015)	3.9705 (0.0016)	**0.3827 (0.0003)**
6	-4.905 (0.006)	-4.716 (0.006)	3.9691 (0.0007)	0.3959 (0.0009)
9	-4.898 (0.005)	**-4.732 (0.005)**	3.9634 (0.0003)	0.3944 (0.0011)
15	-4.875 (0.007)	**-4.734 (0.004)**	**3.9619 (0.0011)**	0.3971 (0.0010)
21	**-4.831 (0.026)**	-4.719 (0.019)	**3.9622 (0.0001)**	0.4316 (0.0336)
$\delta = 0.9$	-4.930 (0.016)	**-4.745 (0.009)**	3.9641 (0.0007)	0.3932 (0.0006)
$\delta = 0.5$	-4.890 (0.003)	**-4.733 (0.013)**	**3.9662 (0.0043)**	0.3953 (0.0002)
$\delta = 0.1$	**-4.867 (0.001)**	**-4.739 (0.002)**	3.9634 (0.0003)	0.3954 (0.0001)

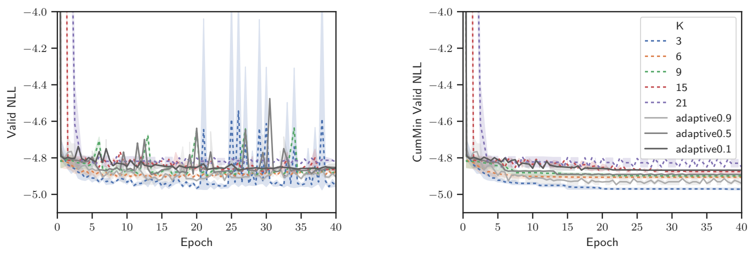

Figure C.4: Book Order Experiment: (left) Valid NLL, (right) 'Best' Valid NLL. Solid dark lines are our adaptive TBPTT methods, dashed colored lines are fixed TBPTT baselines.

References

J. Chen and R. Luss. Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*, 2018.

N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent marked temporal point processes: Embedding event history to vector. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564. ACM, 2016.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.