A. Training Details

A.1. Machine Translation.

Dataset The training/validation/test sets for the IWSLT14 dataset contain about 153K/7K/7K sentence pairs, respectively. We use a vocabulary of 10K tokens based on a joint source and target byte pair encoding (BPE) (Sennrich et al., 2016). For the WMT14 dataset, we follow the setup of (Vaswani et al., 2017), which contains 4.5M training parallel sentence pairs. Newstest2014 is used as the test set, and Newstest2013 is used as the validation set. The 37K vocabulary for WMT14 is based on a joint source and target BPE factorization.

Hyperparameter Given the unstable gradient issues of decoders in NMT (Zhang et al., 2019a), we only change all the normalization layers in the 6 encoder layers from LN to BN/PN, and we keep all the 6 decoder layers to use LN. For Transformer$_{\text{PN-V}}$ `big` and Transformer$_{\text{BN}}$ `big` (not Transformer$_{\text{PN}}$ `big`), we use the synchronized version, where each FP and BP will synchronize the mean/variance/quadratic mean of different batches at different nodes. For PN, we set the α in the forward and backward steps differently, and we tune the best setting over 0.9/0.95/0.99 on the validation set. To control the scale of the activation, we also involve a layer-scale layer (Zhang & Sennrich, 2019) in each model setting before the normalization layer. The warmup scheme for accumulating ψ is also employed, as suggested in (Yan et al., 2020) . Specifically, we do not tune the warmup steps, but we set it identical to the warmup steps for the learning rate schedule in the optimizer (Vaswani et al., 2017). We set dropout as 0.3/0.0 for Transformer `big`/`small` model, respectively. We use the Adam optimizer and follow the optimizer setting and learning rate schedule in (Wang et al., 2019). We set the maximum number of updates following (Ott et al., 2018) to be 300k for WMT and 100k for IWSLT. We used early stopping to stop the experiments by showing no improvement over the last 10/5 epochs. For the `big` model, we enlarge the batch size and learning rate, as suggested in (Ott et al., 2019), to accelerate training. We employ label smoothing of value $\epsilon_{\text{ls}} = 0.1$ in all experiments. We implement our code for MT using *fairseq-py* (Ott et al., 2019).

Evaluation We use BLEU[4] (Papineni et al., 2002) as the evaluation metric for MT. Following standard practice, we measure tokenized case-sensitive BLEU and case-insensitive BLEU for WMT14 En-De and IWSLT14 De-En, respectively. For a fair comparison, we do not include other external datasets. For inference, we average the last 10 checkpoints, and we set the length penalty to 0.6/1.0 and beam size to 4/5 for WMT/IWSLT, following (Ott et al., 2019).

A.2. Language Modeling.

Dataset PTB (Mikolov et al., 2011) has 0.93M training tokens, 0.073M validation words, and 0.082M test word. Wikitext-103 (Merity et al., 2017) contains 0.27M unique tokens, and 100M training tokens from 28K articles, with an average length of 3.6K tokens per article. We use the same evaluation scheme that was provided in (Dai et al., 2019).

Hyperparameter We use three layers tensorized transformer core-1 for PTB and six layers tensorized transformer core-1 for Wikitext-103, following (Ma et al., 2019). This means there exists only one linear projection in multi-linear attention. We replace every LN layer with a PN layer. For PN, we set the α in forward and backward differently, and we tune the best setting over 0.9/0.95/0.99 on the validation set. The warmup scheme and layer-scale are also the same as the hyperparameter setting introduced for machine translation. We set the dropout as 0.3 in all the datasets. The model is trained using 30 epochs for both PTB and WikiText-103. We use the Adam optimizer, and we follow the learning rate setting in (Ma et al., 2019). We set the warmup steps to be 4000 and label smoothing to be $\epsilon_{\text{ls}} = 0.1$ in all experiments.

B. Extra Results

B.1. Empirical Results for Lemma 2.

Under Assumption 9, mentioned in Section 4.1 and discussed in Appendix C.1, we show

$$\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}_{:,i}}\|^2 = \frac{\gamma_i^2}{(\psi_B)_i^2}\big(\|\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{X}_{:,i}}\|^2 - \langle \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{X}_{:,i}}, \frac{\widehat{\boldsymbol{X}}_{:,i}}{\sqrt{B}}\rangle^2\big).$$

Given that $\langle \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{X}_{:,i}}, \frac{\widehat{\boldsymbol{X}}_{:,i}}{\sqrt{B}}\rangle^2$ is non-negative, the Lipschitz constant of $\mathcal{L}$ is smaller than that of $\widehat{\mathcal{L}}$ if $\gamma_i \leqslant (\psi_B)_i$. Here, we

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

report the empirical results to show that $\gamma_i \leqslant (\psi_B)_i$ holds for each $i \in \{1, 2, ..., d\}$ on IWSLT14; see Figure 7. Observe that the Lipschitz constant of $\mathcal{L}$ is smaller than that of $\widehat{\mathcal{L}}$ empirically in our setting.

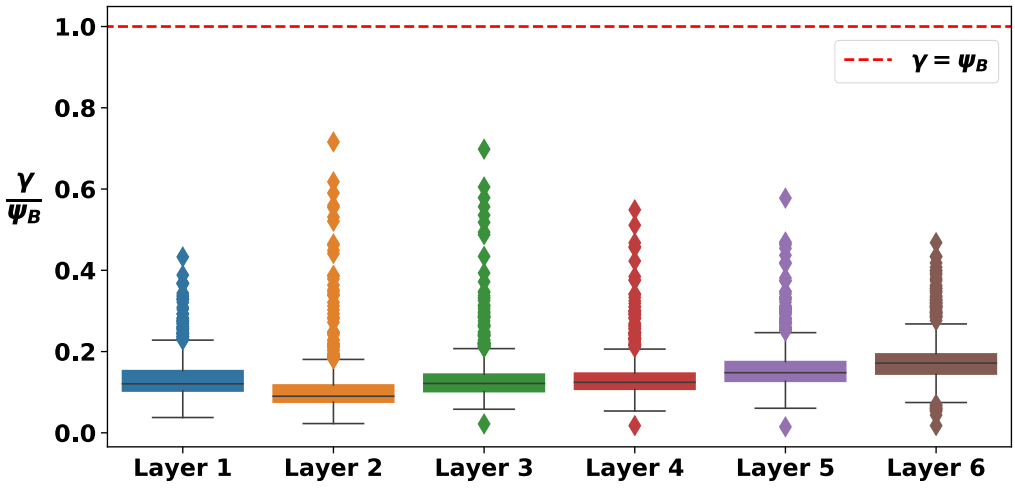

Figure 7. The empirical results of the distribution of $\frac{\gamma}{(\psi_B)} \in \mathbb{R}^d$ in different layers of Transformer$_{\text{PN-V}}$ on IWSLT14. Given that $\gamma_i \leqslant (\psi_B)_i$ holds for each $i \in \{1, 2, ..., d\}$, **Lemma 2** holds as well.

B.2. Validation Results on Language Modeling.

Model	PTB		WikiText-103	
	Val PPL	Test PPL	Val PPL	Test PPL
Tied-LSTM (Inan et al., 2017)	75.7	48.7	–	48.7
AWD-LSTM-MoS (Yang et al., 2018)	58.1	56.0	29.0	29.2
Adaptive Input (Baevski & Auli, 2019)	59.1	57.0	19.8	20.5
Transformer-XL$_{\texttt{base}}$ (Dai et al., 2019)	56.7	54.5	23.1	24.0
Transformer-XL$_{\texttt{large}}$ (Dai et al., 2019)	–	–	–	18.3
Tensor-Transformer$_{1\text{core}}$ (Ma et al., 2019)	55.4	57.9	23.6	20.9
Tensor-Transformer$_{2\text{core}}$ (Ma et al., 2019)	54.3	49.8	19.7	18.9
Tensor-Transformer$_{1\text{core}}$ + LN	58.0*	53.2*	22.7*	20.9*
Tensor-Transformer$_{1\text{core}}$ + BN	71.7	60.7	28.4	27.2
Tensor-Transformer$_{1\text{core}}$ + PN-V	59.7	55.3	23.6	21.3
Tensor-Transformer$_{1\text{core}}$ + PN	**51.6**	**47.6**	**18.3**	**17.9**

Table 3. Additional Validation and Test results with state-of-the-art results on PTB and WikiText-103. '-' indicates no reported results in that setting, '$*$' indicates that the results are from our own implementation. PN achieves 5.6/3.0 points lower testing PPL on PTB and WikiTest-103, respectively, as compared to LN.

B.3. More Comparisons.

As shown in B.3, we present more comparison results for different normalization method including Moving Average Batch Normalizatio (MABN) (Yan et al., 2020), Batch Renormalization (BRN) (Ioffe, 2017), and Group Normalization (GN) (Wu & He, 2018). We can observe that BRN/MABN/PN-V is better than BN but worse than LN, which suggests the small batch-size setting (main focus of (Yan et al., 2020; Ioffe, 2017; Wu & He, 2018)) may have similar characteristic of the setting in NLP, where there exists large variance across batches. Obviously, GN performs the best among the previous proposed methods given LN can be viewed as the special case of GN (group number as 1). [5] Throughout the comparisons,

[5]We empirically found that setting group number the same as head number leads to the best performance.

Model	IWSLT14	PTB
$\text{Transformer}_{\text{BN}}$	34.4	60.7
$\text{Transformer}_{\text{BRN}}$	34.7	58.3
$\text{Transformer}_{\text{MABN}}$	34.9	57.2
$\text{Transformer}_{\text{LN}}$	35.5	53.2
$\text{Transformer}_{\text{GN}}$	35.7	51.7
$\text{Transformer}_{\text{PN-V}}$	35.5	55.3
$\text{Transformer}_{\text{PN}}$	**35.9**	**47.6**

Table 4. (Left) NMT performance (BLEU) on IWSLT14 De-En. (Right) LM performance (Test PPL) on PTB.

PN still performs the best in the two tasks, which may validate the effectiveness of our method.

C. Theoretical Results

In this section, we discuss the theoretical results on BN and PN. We assume γ and β to be constants for our analysis on BN, PN-V and PN.

Since the derivative of loss $\mathcal{L}$ w.r.t. $\boldsymbol{Y}$ is known as $\frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}$, trivially, we will have $\gamma \odot \frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}$. Also, it is not hard to get the following fact.

Fact 5. *The derivatives of μ_B and σ_B^2 w.r.t. $\boldsymbol{x}_i$ are*

$$\frac{\partial \mu_B}{\partial \boldsymbol{x}_i} = \frac{1}{B} \quad \textit{and} \quad \frac{\partial \sigma^2}{\partial \boldsymbol{x}_i} = \frac{2}{B}(\boldsymbol{x}_i - \mu_B). \tag{14}$$

We are now ready to show the derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{x}_i$ under BN.

Lemma 6 (Derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{x}_i$ in BN)**.** *Based on the Fact 5, it holds that*

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} = \frac{1}{\sigma_B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_i} - \frac{1}{\sigma_B B}\sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}(1 + \check{\mathbf{x}}_j\check{\mathbf{x}}_i). \tag{15}$$

Proof. Based on chain rule, we will have

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} &= \frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_i}\frac{\partial \check{\mathbf{x}}_i}{\partial \boldsymbol{x}_i} + \sum_{j\in B}\Big(\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}\frac{\partial \check{\mathbf{x}}_j}{\partial \mu_B}\frac{\partial \mu_B}{\partial \boldsymbol{x}_i} + \frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}\frac{\partial \check{\mathbf{x}}_j}{\partial \sigma_B}\frac{\partial \sigma_B}{\partial \boldsymbol{x}_i}\Big) \\
&= \frac{1}{\sigma_B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_i} + \sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}\Big(\frac{\partial \check{\mathbf{x}}_j}{\partial \mu_B}\frac{1}{B} + \frac{\partial \check{\mathbf{x}}_j}{\partial \sigma_B^2}\frac{2}{B}(\boldsymbol{x}_i - \mu_B)\Big) \\
&= \frac{1}{\sigma_B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_i} - \frac{1}{\sigma_B B}\sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}\Big(1 + \frac{\boldsymbol{x}_i - \mu_B}{\sigma_B}\frac{\boldsymbol{x}_j - \mu_B}{\sigma_B}\Big) \\
&= \frac{1}{\sigma_B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_i} - \frac{1}{\sigma_B B}\sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \check{\mathbf{x}}_j}(1 + \check{\mathbf{x}}_j\check{\mathbf{x}}_i).
\end{aligned} \tag{16}$$

□

Replacing $\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}$ by $\gamma \odot \frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}$, we can get Eq. 3.

In the following, we will first discuss the theoretical properties of PN-V in Appendix C.1; and then we discuss how to use running statistics in the forward propagation and how to modify the corresponding backward propagation in Appendix C.2.

C.1. Proof of PN-V

Before showing the gradient of $\mathcal{L}$ w.r.t. $\boldsymbol{x}_i$ under PN-V, we note the following fact, which is not hard to establish.

Fact 7. *The derivatives of ψ_B w.r.t. $\boldsymbol{x}_i$ are,*

$$\frac{\partial {\psi_B}^2}{\partial \boldsymbol{x}_i} = \frac{2}{B}\boldsymbol{x}_i. \tag{17}$$

With the help of Fact 7, we can prove the following lemma

Lemma 8 (Derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{x}_i$ in PN-V)**.** *Based on the Fact 7, it holds that that*

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} = \frac{1}{\psi_B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i} - \frac{1}{B\psi_B}\sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_j}\hat{\mathbf{x}}_j\hat{\mathbf{x}}_i. \tag{18}$$

Proof. Based on chain rule, we will have

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} &= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i}\frac{\partial \hat{\mathbf{x}}_i}{\partial \boldsymbol{x}_i} + \sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_j}\frac{\partial \hat{\mathbf{x}}_j}{\partial {\psi_B}^2}\frac{\partial {\psi_B}^2}{\partial \boldsymbol{x}_i} \\
&= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i}\frac{\partial \hat{\mathbf{x}}_i}{\partial \boldsymbol{x}_i} + \sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_j}(-\frac{1}{2}\frac{\boldsymbol{x}_j}{{\psi_B}^3})\frac{2\boldsymbol{x}_i}{B} \\
&= \frac{1}{\psi_B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i} - \frac{1}{B\psi_B}\sum_{j\in B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_j}\hat{\mathbf{x}}_j\hat{\mathbf{x}}_i.
\end{aligned} \tag{19}$$

□

Replacing $\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}$ by $\gamma \odot \frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}$, we can get Eq. 6.

In order to show the effect of PN-V on the Lipschitz constant of the loss, we make the following standard assumption, as in (Santurkar et al., 2018).

Assumption 9. *Denote the loss of the non-normalized neural network, which has the same architecture as the* PN-V *normalized neural network, as $\widehat{\mathcal{L}}$. We assume that*

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_i} = \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{x}_i}, \tag{20}$$

where $\boldsymbol{y}_i$ is the i-th row of $\boldsymbol{Y}$.

Based on these results, we have the following proof of Lemma 2, which was stated in Section 4.

Proof of Lemma 2. Since all the computational operator of the derivative is element-wise, here we consider $d = 1$ for notational simplicity[6]. When $d = 1$, Lemma 8 can be written as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} = \frac{1}{\psi_B}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i} - \frac{1}{B\psi_B}\langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \widehat{\boldsymbol{X}}\rangle\hat{\mathbf{x}}_i. \tag{21}$$

Therefore, we have

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}} = \frac{1}{\psi_B}\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}} - \frac{1}{B\psi_B}\langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \widehat{\boldsymbol{X}}\rangle\widehat{\boldsymbol{X}}. \tag{22}$$

Since

$$\|\widehat{\boldsymbol{X}}\|^2 = \frac{\sum_{i\in B}\hat{\mathbf{x}}_i}{\frac{1}{B}\sum_{i\in B}\hat{\mathbf{x}}_i} = B, \tag{23}$$

[6]For $d \geqslant 2$, we just need to separate the entry and prove them individually.

the following equation can be obtained

$$\begin{aligned}
\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}}\|^2 &= \frac{1}{{\psi_B}^2}\|\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}} - \langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle\frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\|^2 \\
&= \frac{1}{{\psi_B}^2}\big(\|\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}\|^2 - 2\langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle\frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle + \|\langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle\frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle\|^2\big) \\
&= \frac{1}{{\psi_B}^2}\big(\|\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}\|^2 - \langle\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{X}}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle^2\big) \\
&= \frac{\gamma^2}{{\psi_B}^2}\big(\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}\|^2 - \langle\frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle^2\big) \\
&= \frac{\gamma^2}{{\psi_B}^2}\big(\|\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{X}}\|^2 - \langle\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{X}}, \frac{\widehat{\boldsymbol{X}}}{\sqrt{B}}\rangle^2\big).
\end{aligned} \tag{24}$$

□

C.2. Proof of PN

In order to prove that after the replacement of $\frac{\partial \mathcal{L}}{\partial (\boldsymbol{X}^{(t)})}$ with Eq. 12, the gradient of the input is bounded, we need the following assumptions.

Assumption 10. *We assume that*

$$\|\hat{\mathbf{x}}_i\| \leqslant C_1 \quad \textit{and} \quad \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i}\| \leqslant C_2, \tag{25}$$

for all input datum point and all iterations. We also assume that the exponentially decaying average of each element of $\hat{\mathbf{x}}_i$ is bounded away from zero,

$$(1-\alpha)\sum_{j=0}^{t}\alpha^{t-j}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i > C_3 > 0, \ \ \forall t, \tag{26}$$

where we denote α as the decay factor for the backward pass. In addition, we assume that α satisfies

$$(C_1)^2 < \frac{1}{1-\alpha}. \tag{27}$$

W.l.o.g., we further assume that every entry of $\psi^{(t)}$ is bounded below, i.e.

$$C_0 < \psi^{(t)}, \quad \forall t. \tag{28}$$

If we can prove or $\nu^{(t)}$ is bounded by some constant C_4 (the official proof is in Lemma 11), then it is obvious to prove the each datum point of $\widetilde{\boldsymbol{X}}'$ is bounded.

Based on these results, we have the following proof of Theorem 4, which was stated in Section 4.

Proof of Theorem 4. It is easy to see that

$$\begin{aligned}
\|\widetilde{\boldsymbol{X}}'_{i,:}\|^2 &= \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}} - \nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)}\|^2 \\
&= \langle \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}} - \nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)}, \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}} - \nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)} \rangle \\
&= \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|^2 + \|\nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)}\|^2 - 2\langle \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}, \nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)} \rangle \\
&\leqslant \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|^2 + \|\nu^{(t-1)}\|^2\|\hat{\mathbf{x}}_i^{(t)}\|^2 - 2\langle \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}, \nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)} \rangle \\
&\leqslant \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|^2 + \|\nu^{(t-1)}\|^2\|\hat{\mathbf{x}}_i^{(t)}\|^2 + 2\|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|\|\nu^{(t-1)}\hat{\mathbf{x}}_i^{(t)}\| \\
&\leqslant \|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|^2 + \|\nu^{(t-1)}\|^2\|\hat{\mathbf{x}}_i^{(t)}\|^2 + 2\|\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i^{(t)}}\|\|\nu^{(t-1)}\|\|\hat{\mathbf{x}}_i^{(t)}\| \\
&\leqslant (C_2)^2 + (C_1)^2(C_4)^3 + C_1C_2C_4
\end{aligned}$$

All these inequalities come from Cauchy-Schwarz inequity and the fact that

$$(a_1b_1)^2 + ... + (a_db_d)^2 \leqslant (a_1^2 + ... + a_d^2)(b_1^2 + ... + b_d^2).$$

□

In the final step of Theorem 4, we directly use that $\nu^{(t)}$ is uniformly bounded (each element of $\nu^{(t)}$ is bounded) by C_4. The exact proof is shown in below.

Lemma 11. *Under Assumption 10, $\nu^{(t)}$ is uniformly bounded.*

Proof. For simplicity, denote $\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}_i}$ as $\hat{\mathbf{x}}'_i$. It is not hard to see,

$$\begin{aligned}
\|\Gamma^{(t)}\|^2 &= \frac{1}{B^2}\|\sum_{i=1}^{B}\hat{\mathbf{x}}_i^{(t)}\hat{\mathbf{x}}_i^{(t)}\|^2 \\
&= \frac{1}{B^2}\langle\sum_{i=1}^{B}\hat{\mathbf{x}}_i^{(t)}\hat{\mathbf{x}}_i^{(t)}, \sum_{i=1}^{B}\hat{\mathbf{x}}_i^{(t)}\hat{\mathbf{x}}_i^{(t)}\rangle \\
&\leqslant \frac{1}{B^2}(B^2 \max_j\{\langle\hat{\mathbf{x}}_j^{(t)}\hat{\mathbf{x}}_j^{(t)}, \hat{\mathbf{x}}_i^{(t)}\hat{\mathbf{x}}_i^{(t)}\rangle\}) \\
&\leqslant (C_1)^2.
\end{aligned}$$

Similarly, we will have $\|\Lambda^{(t)}\| \leqslant C_1C_2$ as well as $(1-\alpha)\sum_{j=0}^{t}\alpha^{t-j}\Gamma^{(j)}\Gamma^{(j)} > C_3$. We have

$$\begin{aligned}
\nu^{(t)} &= (1-(1-\alpha)\Gamma^{(t)})\nu^{(t-1)} + (1-\alpha)\Lambda^{(t)} \\
&= (1-(1-\alpha)\Gamma^{(t)})((1-(1-\alpha)\Gamma^{(t-1)})\nu^{(t-2)} + (1-\alpha)\Lambda^{(t-1)}) + (1-\alpha)\Lambda^{(t)} \\
&\vdots \\
&= (1-\alpha)\sum_{j=0}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}.
\end{aligned}$$

Then,

$$\frac{1}{(1-\alpha)^2}\|\nu^{(t)}\|^2 = \langle\sum_{j=0}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}, \sum_{j=0}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}\rangle.$$

Notice that with the definition,

$$\Gamma^{(m)} = \frac{1}{B}\sum_{i=1}^{B} \hat{\mathbf{x}}_i^{(m)} \hat{\mathbf{x}}_i^{(m)}, \tag{29}$$

we will have that all entries of $\Gamma^{(m)}$ are positive, for $m \in \{0, 1, ..., t\}$. It is clear that when all entries of $\Lambda^{(m)}$, for $m \in \{0, 1, ..., t\}$, have the same sign (positive or negative), the above equation achieves its upper bound. W.l.o.g., we assume they are all positive.

Since $0 < \alpha < 1$, it is easy to see that, when $K = \lceil (\log(\frac{(1-\alpha)C_3}{2C_1C_1})/\log(\alpha)) \rceil$, then the following inequality holds,

$$(1-\alpha)\sum_{j=K}^{\infty} \alpha^j < \frac{C_3}{2C_1C_1}. \tag{30}$$

Since $\|\Gamma^{(k)}\| \leqslant C_1$, the value of any entry of $\Gamma^{(k)}$ is also bounded by C_1. Therefore, based on this and Eq. 30, when $t > K$, we will have

$$\begin{aligned}
(1-\alpha)\sum_{k=t-K+1}^{t} \alpha^{t-k}\,\Gamma^{(k)}\Gamma^{(k)} &= (1-\alpha)\sum_{k=0}^{t} \alpha^{t-k}\,\Gamma^{(k)}\Gamma^{(k)} - (1-\alpha)\sum_{k=0}^{t-K} \alpha^{t-k}\,\Gamma^{(k)}\Gamma^{(k)} \\
&> C_3\vec{1} - (1-\alpha)\sum_{k=0}^{t-K} \alpha^{t-k}\,\|\Gamma^{(k)}\|\|\Gamma^{(k)}\| \\
&> C_3\vec{1} - (1-\alpha)C_1C_1\vec{1}\sum_{k=0}^{t-K} \alpha^{t-k} \\
&= C_3\vec{1} - (1-\alpha)C_1C_1\vec{1}\sum_{k=K}^{t} \alpha^{k} \\
&> C_3\vec{1} - (1-\alpha)C_1C_1\vec{1}\sum_{k=K}^{\infty} \alpha^{k} \\
&> C_3\vec{1} - \frac{C_3\vec{1}}{2} \\
&= \frac{C_3\vec{1}}{2},
\end{aligned} \tag{31}$$

where $\vec{1}$ is the unit vector. Then, for $t > K$, we can bound from below the arithmetic average of the K corresponding items of Γ,

$$\begin{aligned}
(C_1)^2 > \frac{1}{K}\sum_{k=0}^{K-1} \Gamma^{(t-k)}\Gamma^{(t-k)} &> \frac{1}{\alpha^{K-1}}\sum_{k=0}^{K-1} \alpha^k\Gamma^{(t-k)}\Gamma^{(t-k)} \\
&= \frac{1}{\alpha^{K-1}}\sum_{k=t-K+1}^{t} \alpha^{t-1}\Gamma^{(k)}\Gamma^{(k)} \\
&> \frac{C_3}{2(1-\alpha)\alpha^{K-1}} = C_5 > 0.
\end{aligned} \tag{32}$$

This inequality shows that after the first K items, for any K consecutive $\Gamma^{(k)}$, the average of them will exceeds a constant number, C_5. Therefore, for any $t > T > K$, we will have

$$\frac{1}{T-K}\sum_{k=0}^{T-K} \Gamma^{(t-k)}\Gamma^{(t-k)} > \lfloor \frac{T-K}{K} \rfloor (K\frac{1}{T-K})C_5 > \frac{C_5}{2}. \tag{33}$$

Let us split $\sum_{j=0}^{t} \big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}$ into two parts: (i) $\sum_{j=K}^{t} \big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}$, and (ii) $\sum_{j=0}^{K-1} \big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}$. From so on, we will discuss how we deal with these two parts respectively.

Case 1: $\sum_{j=K}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}$ Notice that for $0<a_j<1$, the following inequality can be proven with simply induction,

$$\prod_{j=0}^{k-1}(1-a_j)\leqslant(1-\frac{1}{k}\sum_{j=0}^{k-1}\alpha_j)^k. \tag{34}$$

Replacing a_j with $(1-\alpha)\Gamma^{(t-j+1)}$, we will have

$$\begin{aligned}\sum_{j=K}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}&\leqslant\sum_{j=K}^{t}\big((1-\frac{(1-\alpha)}{j}\sum_{k=0}^{j-1}\Gamma^{(t-k+1)})\big)^j\Lambda^{(t-j)}\\&\leqslant\sum_{j=K}^{t}\big((1-(1-\alpha)\frac{C_5}{2})\big)^j\Lambda^{(t-j)}\\&\leqslant\sum_{j=K}^{t}\big((1-(1-\alpha)\frac{C_5}{2})\big)^jC_1C_2\\&\leqslant\frac{2}{(1-\alpha)C_5}C_1C_2=C_6.\end{aligned} \tag{35}$$

Here the second inequality comes from Eq. 33, and the third inequality comes form the the fact each entry of $\Lambda^{(m)}$ is smaller than C_1C_2, given $\|\Lambda^{(m)}\|\leqslant C_1C_2$. The final inequality comes from Eq. 31, where $0<C_5<(C_1)^2<1/(1-\alpha)$, then we can have $0<\big(1-(1-\alpha)C_5/2\big)<1$.

Case 2: $\sum_{j=0}^{K-1}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}$ It is easy to see

$$\begin{aligned}\sum_{j=0}^{K-1}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}&\leqslant\sum_{j=0}^{K-1}\big(\prod_{k=0}^{j-1}(\vec{1})\big)\Lambda^{(t-j)}\\&\leqslant KC_1C_2.\end{aligned} \tag{36}$$

Combining Case 1 and 2, we have

$$\begin{aligned}\frac{1}{(1-\alpha)^2}\|\nu^{(t)}\|^2&=\langle\sum_{j=0}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)},\sum_{j=0}^{t}\big(\prod_{k=0}^{j-1}(1-(1-\alpha)\Gamma^{(t-k+1)})\big)\Lambda^{(t-j)}\rangle\\&\leqslant\langle C_6\vec{1}+KC_1C_2\vec{1},C_6\vec{1}+KC_1C_2\vec{1}\rangle<C_7,\end{aligned} \tag{37}$$

which indicates $\|\nu^{(t)}\|$ is bounded and $C_4=(1-\alpha)\sqrt{C_7}$.

□