Appendix: Reviving and Improving Recurrent Back-Propagation

Renjie Liao [1 2 3] **Yuwen Xiong** [1 2] **Ethan Fetaya** [1 3] **Lisa Zhang** [1 3] **KiJung Yoon** [4] **Xaq Pitkow** [4 5]
Raquel Urtasun [1 2 3] **Richard Zemel** [1 3 6]

A similar technique to RBP was discovered in physics by Richard Feynman (Feynman, 1939) in modeling molecular forces back in the 1930's. When the energy of molecules are in steady state, the forces on the molecules are defined as the gradient of energy w.r.t. the position parameters of molecules.

1. Assumptions of RBP

In this section, we will further discuss the assumptions imposed by RBP.

1.1. Contraction Map

Contraction map is often adopted for constructing a convergent dynamic system. But it also largely restricts the model capacity and is also hard to satisfy for general neural networks. Moreover, as pointed out by (Li et al., 2016), on a special cycle graph, contraction map will make the impact of one node on the other decay exponentially with their distance.

1.2. Local Regularization At Convergence

Recall that in order to apply implicit function theorem, we just to need to make sure that no singular value of the Jacobian is zero. In particular, note that $|\det(I - J_{F,h^*})| > 0$ is equivalent to $|\det(I - J_{F,h^*})|^2 > 0$, one can equivalently rewrite the condition II as,

$$|\det(I - J_{F,h^*})|^2 = \prod_i |\sigma_i(I - J_{F,h^*})|^2 > 0. \quad (1)$$

Note that for any square matrix A, we have,

$$\det(A^\top A) = \det(A^\top)\det(A) = \det(A)^2 \quad (2)$$

Therefore, we can instead focus on the positive semi-definite matrix $(I - J_{F,h^*})^\top (I - J_{F,h^*})$. The condition can be equivalently stated as below,

$$\lambda_{\min}\left((I - J_{F,h^*})^\top (I - J_{F,h^*})\right) > 0, \quad (3)$$

where $\lambda_{\min}$ is the smallest eigenvalue. We now briefly discuss two ways of maximizing the smallest eigenvalue.

Maximizing Lower Bound One way to achieve this is to enforce the lower bound of $\lambda_{\min}$ is larger than zero. Specifically, according to Gershgorin Circle Theorem, if A is positive definite, we have,

$$\lambda_{\min}(A) \geq 1 - \|A - I\|_\infty \geq 1 - \sqrt{n}\|A - I\|_F. \quad (4)$$

We can instead maximize the lower bound by adding the term $\max(0, \sqrt{n}\|A - I\|_F - 1)$ to the loss function. One may need to add a small constant to A if A is only positive semi-definite rather than positive definite. Note that the RHS term is not necessarily larger than zero.

Direct Maximizing By Differentiating Through Lanczos Another possible solution is to treat Lanczos algorithm as a fix computational graph to compute the smallest eigenvalue of $(I - J_{F,h^*})^\top (I - J_{F,h^*})$. The most expansive operator in one step Lanczos is the matrix-vector product $(I - J_{F,h^*})^\top (I - J_{F,h^*})v$ which has doubled complexity as back-propagation. Differentiating through Lanczos via BPTT is even more expansive which also provides rooms for applying RBP. We can add a term $\max(0, -\lambda_{\min})$ to the loss function. Note that the computational complexity of this method is generally high which seems to be impractical for large scale problems.

2. Recurrent Back-Propagation based on Neumann Series

In this section, we restate the propositions and prove them.

Proposition 1. *Assume that there exists some step K where $0 < K \leq T$ such that for the convergent sequence of hidden states $h^0, h^1, \ldots, h^T$ of an RNN, we have $h^* = h^T = h^{T-1} = \cdots = h^{T-K}$ where h^* is the fixed point. If the Neumann series $\sum_{t=0}^{\infty} J_{F,h^*}^t$ converges, then the full and K-step Neumann-RBP are equivalent to BPTT and K-step TBPTT respectively.*

Proof. Since Neumann series $\sum_{t=0}^{\infty} J_{F,h^*}^t$ converges, we have $(I - J_{F,h^*})^{-1} = \sum_{t=0}^{\infty} J_{F,h^*}^t$. By substituting it into

Eq. (8), we have,

$$\begin{aligned}\frac{\partial L}{\partial w_F} &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\left(I - J_{F,h^*}\right)^{-1}\frac{\partial F(x, w_F, h^*)}{\partial w_F} \\ &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\left(I + J_{F,h^*} + J_{F,h^*}^2 + \dots\right)\frac{\partial F(x, w_F, h^*)}{\partial w_F} \\ &= \frac{\partial L}{\partial y}\sum_{k=0}^{\infty}\frac{\partial y}{\partial h^*}J_{F,h^*}^k\frac{\partial F(x, w_F, h^*)}{\partial w_F}. \end{aligned} \tag{5}$$

Therefore, the full Neumann-RBP is equivalent to original RBP which is further equivalent to BPTT due the implicit function theorem. If we truncate K steps from the end, then the gradient of TBPTT is

$$\begin{aligned}\frac{\partial L}{\partial w_F} &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\sum_{k=0}^{K}\left(\prod_{i=T}^{T-k} J_{F,h^i}\right)\frac{\partial F(x, w_F, h^{T-k})}{\partial w_F}. \\ &= \frac{\partial L}{\partial y}\sum_{k=0}^{K}\frac{\partial y}{\partial h^*}J_{F,h^*}^k\frac{\partial F(w_F, h^*)}{\partial w_F}, \end{aligned} \tag{6}$$

where the second row uses the fact that $h^* = h^T = h^{T-1} = \cdots = h^{T-K}$. Comparing Eq. (5) and Eq. (6), it is clear that we exactly recover the K-step Neumann-RBP. □

Proposition 2. *If the Neumann series $\sum_{t=0}^{\infty} J_{F,h^*}^t$ converges, then the error between K-step and full Neumann series is as following,*

$$\left\|\sum_{t=0}^{K} J_{F,h^*}^t - \left(I - J_{F,h^*}\right)^{-1}\right\| \leq \left\|\left(I - J_{F,h^*}\right)^{-1}\right\| \left\|J_{F,h^*}\right\|^{K+1}$$

Proof. First note that,

$$\begin{aligned}\left(I - J_{F,h^*}\right)\left(\sum_{t=0}^{K} J_{F,h^*}^t\right) &= I\left(\sum_{t=0}^{K} J_{F,h^*}^t\right) - J_{F,h^*}\left(\sum_{t=0}^{K} J_{F,h^*}^t\right) \\ &= I - J_{F,h^*}^{K+1}. \end{aligned} \tag{7}$$

Multiplying $\left(I - J_{F,h^*}\right)^{-1}$ on both sides, we get,

$$\left(\sum_{t=0}^{K} J_{F,h^*}^t\right) = \left(I - J_{F,h^*}\right)^{-1}\left(I - J_{F,h^*}^{K+1}\right). \tag{8}$$

With a bit rearrange, we have,

$$\left(\sum_{t=0}^{K} J_{F,h^*}^t\right) - \left(I - J_{F,h^*}\right)^{-1} = -\left(I - J_{F,h^*}\right)^{-1} J_{F,h^*}^{K+1}. \tag{9}$$

The result is then straightforward by using Cauchy-Schwarz inequality. □

We further make the following proposition regarding to the relationship between Neumann-RBP and the original RBP algorithm.

Proposition 3. *$K+1$-step RBP algorithm returned the same gradient with K-step Neumann-RBP if z_0 in original RBP is initialized as a zero vector.*

Truncate Step	TBPTT	RBP	CG-RBP	Neumann-RBP
10	100%	1%	100%	100%
20	100%	4%	100%	100%
30	100%	99%	100%	100%

Table A1. Success rate of different methods with different truncation steps. RBP is unstable until the truncation step reaches 30.

Proof. To prove this proposition, we only need to compare the vector z_{K+1} and g_K returned by two algorithms respectively. For original RBP, recall in Algorithm 1, we have the following recursion,

$$z_t = J_{F,h^*}^{\top} z_{t-1} + \left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\right)^{\top}. \tag{10}$$

Therefore, after $K+1$ step, we have,

$$z_{K+1} = \left(J_{F,h^*}^{\top}\right)^{K+1} z_0 + \sum_{t=0}^{K}\left(J_{F,h^*}^{\top}\right)^t\left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\right)^{\top}. \tag{11}$$

For Neumann-RBP, we have the following recursion from Algorithm 2,

$$\begin{aligned} v_t &= J^{\top} v_{t-1} \\ g_t &= g_{t-1} + v_t \end{aligned} \tag{12}$$

with $v_0 = g_0 = \left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\right)^{\top}$. Therefore, after K step, we have the following expansion,

$$g_K = \sum_{t=0}^{K}\left(J_{F,h^*}^{\top}\right)^t\left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial h^*}\right)^{\top}. \tag{13}$$

The relationship is now obvious. □

3. Experiments

3.1. Example Code

Our Neumann-RBP is very simple to implement as long as the auto-differentiation function is provided. Here we show an example code based on PyTorch in Listing 1. The effective number of lines is less than 10.

3.2. Continuous Hopfield Network

The success rates out of 100 experiments with different random corruptions and initialization are counted in Table A1. We consider one trial as successful if its final training loss is less than 50% of the initial loss. From the table, we can see that original RBP almost always fails to converge until the truncation step increases to 30 whereas both CG-RBP and Neumann-RBP have no issues to converge.

Figure A1 shows full results of visualization of associative memory.

3.3. Citation Networks

Table A2 shows the statistics of datasets we used in our experiments.

Dataset	#Nodes	#Edges	#Classes	#Features
Cora	2,708	5,429	7	1,433
Pubmed	19,717	44,338	3	500

Table A2. Dataset statistics of citation networks.

We also inlcude the comparsion with the recent work ARTBP (Tallec & Ollivier, 2017). The experiment setting is exactly the same as described in the paper. Since the underlying RNN has the loss defined at the last time step, i.e., 100th step, we adapt the ARTBP as follows: instead of randomly truncating at multiple locations, we randomly choose one time step to truncate. Similar analysis can be derived to compensate the truncated gradient such that it is unbiased. Due to the limited time, we only tried uniform and truncated Poisson distribution (expected truncation point is roughly at the 95th time step which is where TBPTT stops) over the truncation location. We use SGD with momentum as the optimizer for all methods. The average validation accuracy over 10 runs are in the table below. We can see that both ARTBP variants do not perform as well as Neumann-RBP in this setting. ARTBP with truncated Poisson is better than the one with uniform which matches the other observation that TBPTT is better than full BPTT.

Test Acc.	Cora
Baseline	39.96 ± 3.4
BPTT	24.48 ± 6.6
TBPTT	46.55 ± 6.4
Uniform-ARTBP	27.88 ± 3.2
TPoisson-ARTBP	42.22 ± 7.1
RBP	29.25 ± 3.3
CG-RBP	39.26 ± 6.5
Neumann-RBP	$\mathbf{46.63} \pm 8.3$

Table A3. Test accuracy of different methods on citation networks.

References

Feynman, R. P. Forces in molecules. *Physical Review*, 56 (4):340, 1939.

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. Gated graph sequence neural networks. *ICLR*, 2016.

Tallec, C. and Ollivier, Y. Unbiasing truncated backpropagation through time. *arXiv preprint arXiv:1705.08209*, 2017.

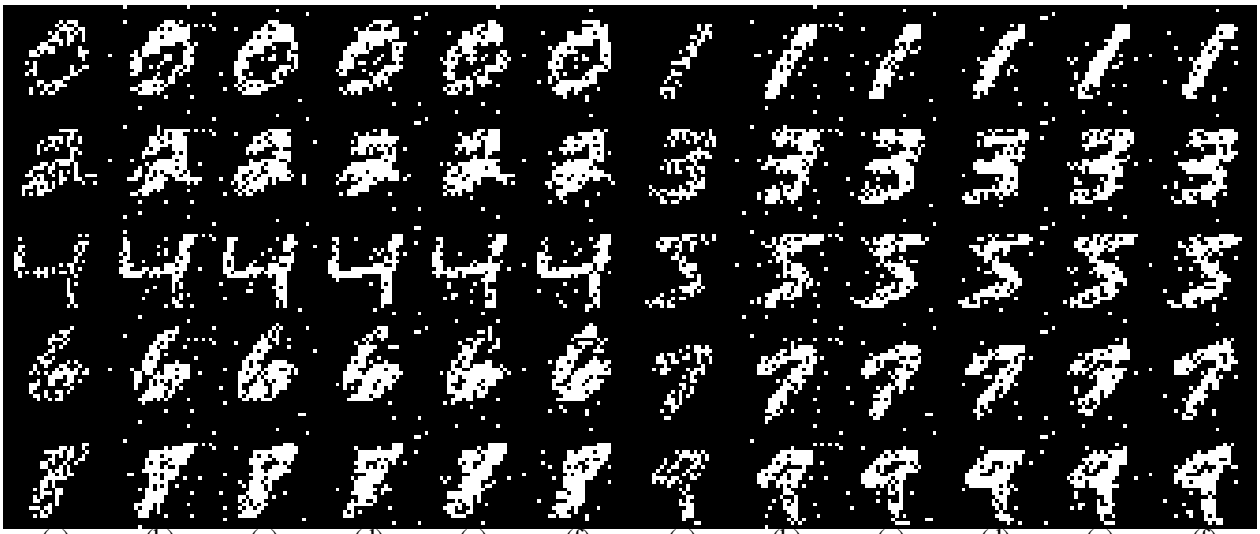

Figure A1. Visualization of associative memory. (a) Corrupted input image; (b)-(f) are retrieved images by BPTT, TBPTT, RBP, CG-RBP, Neumann-RBP respectively.

```
def neumann_rbp(weight, hidden_state, loss, rbp_step)
  # get the gradient of last hidden state
  grad_h = autograd.grad(loss, hidden_state[-1], retain_graph=True)

  # set v, g to grad_h
  neumann_v = grad_h.clone()
  neumann_g = grad_h.clone()

  for i in range(rbp_step):
    # set last hidden_state's gradient to neumann_v[prev]
    # and get the gradient of last second hidden state
    neumann_v = autograd.grad(
                           hidden_state[-1], hidden_state[-2],
                           grad_outputs=neumann_v,
                           retain_graph=True)

    neumann_g += neumann_v

  # set last hidden_state's gradient to neumann_g
  # and return the gradient of weight
  return autograd.grad(hidden_state[-1], weight, grad_outputs=neumann_g)
```

Listing 1. PyTorch example code