Luck Matters: Understanding Training Dynamics of Deep ReLU Networks

Abstract

We analyze the dynamics of training deep ReLU networks and their implications on generalization capability. Using a teacher-student setting, we discovered a novel relationship between the gradient received by hidden student nodes and the activations of teacher nodes for deep ReLU networks. With this relationship and the assumption of small overlapping teacher node activations, we prove that (1) student nodes whose weights are initialized to be close to teacher nodes converge to them at a faster rate, and (2) in over-parameterized regimes and 2-layer case, while a small set of lucky nodes do converge to the teacher nodes, the fan-out weights of other nodes converge to zero. This framework provides insight into multiple puzzling phenomena in deep learning like over-parameterization, implicit regularization, lottery tickets, etc. We verify our assumption by showing that the majority of BatchNorm biases of pre-trained VGG11/16 models are negative. Experiments on (1) random deep teacher networks with Gaussian inputs, (2) teacher network pre-trained on CIFAR-10 and (3) extensive ablation studies validate our multiple theoretical predictions.

1. Introduction

Although neural networks have made strong empirical progress in a diverse set of domains (e.g., computer vision (16; 32; 10), speech recognition (11; 1), natural language processing (22; 3), and games (30; 31; 35; 23)), a number of fundamental questions still remain unsolved. How can Stochastic Gradient Descent (SGD) find good solutions to a complicated non-convex optimization problem? Why do neural networks generalize? How can networks trained with SGD fit both random noise and structured data (38; 17; 24), but prioritize structured models, even in the presence of massive noise (27)? Why are flat minima related to good generalization? Why does over-parameterization lead to better generalization (25; 39; 33; 26; 19)? Why do lottery tickets exist (6; 7)?

In this paper, we propose a theoretical framework for multilayered ReLU networks. Based on this framework, we try to explain these puzzling empirical phenomena with a unified view. We adopt a teacher-student setting where the label provided to an over-parameterized deep student ReLU network is the output of a fixed teacher ReLU network of the same depth and unknown weights (Fig. 1(a)). In this perspective, hidden student nodes are randomly initialized with different activation regions. (Fig. 2(a)). During optimization, student nodes compete with each other to explain teacher nodes. Theorem 4 shows that *lucky* student nodes which have greater overlap with teacher nodes converge to those teacher nodes at a *fast rate*, resulting in *winner-take-all* behavior. Furthermore, Theorem 5 shows that if a subset of student nodes are close to the teacher nodes, they converge to them and the fan-out weights of other irrelevant nodes of the same layer vanishes.

With this framework, we can explain various neural network behaviors as follows:

Fitting both structured and random data. Under gradient descent dynamics, some student nodes, which happen to overlap substantially with teacher nodes, will move into the teacher node and cover them. This is true for both structured data that corresponds to small teacher networks with few intermediate nodes, or noisy/random data that correspond to large teachers with many intermediate nodes. This explains why the same network can fit both structured and random data (Fig. 2(a-b)).

Over-parameterization. In over-parameterization, lots of student nodes are initialized randomly at each layer. Any teacher node is more likely to have a substantial overlap with some student nodes, which leads to fast convergence (Fig. 2(a) and (c), Thm. 4), consistent with (6; 7). This also explains that training models whose capacity just fit the data (or teacher) yields worse performance (19).

Flat minima. Deep networks often converge to "flat minima" whose Hessian has a lot of small eigenvalues (28; 29; 21; 2). Furthermore, while controversial (4), flat minima seem to be associated with good generalization, while sharp minima often lead to poor generalization (12; 14; 36; 20). In our theory, when fitting with structured data, only a few lucky student nodes converge to the teacher, while for other nodes, their fan-out weights shrink towards zero, making them (and their fan-in weights) irrelevant to the final outcome (Thm. 5), yielding flat minima in which movement along most dimensions ("unlucky nodes") results in minimal change in output. On the other hand, sharp min-

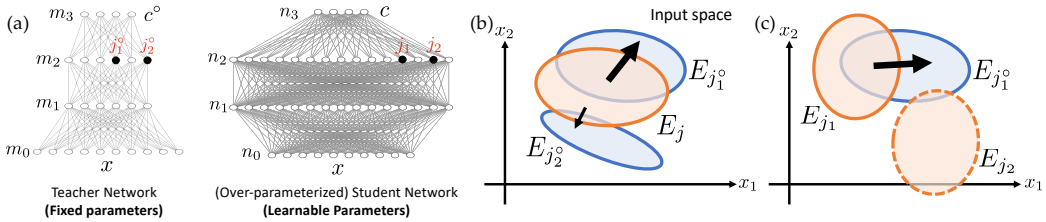

Figure 1. **(a)** Teacher-Student Setting. For each node j, the activation region is $E_j = \{x : f_j(x) > 0\}$. **(b)** node j initialized to overlap substantially with a teacher node j_1° converges faster towards j_1° (Thm. 4).. **(c)** Student nodes initialized to be close to teacher converges to them, while the fan-out weights of other irrelevant student nodes goes to zero. (Thm. 5).

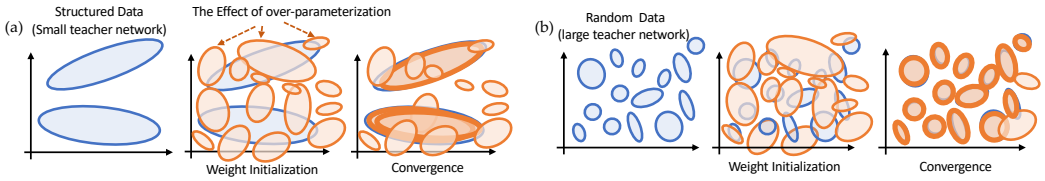

Figure 2. Explanation of implicit regularization. Blue are activation regions of teacher nodes, while orange are students'. **(a)** When the data labels are structured, the underlying teacher network is small and each layer has few nodes. Over-parameterization (lots of red regions) covers them all. Moreover, those student nodes that heavily overlap with the teacher nodes converge faster (Thm. 4), yield good generalization performance. **(b)** If a dataset contains random labels, the underlying teacher network that can fit to it has a lot of nodes. Over-parameterization can still handle them and achieves zero training error.

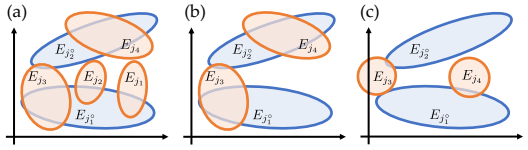

Figure 3. Explanation of lottery ticket phenomenon. **(a)** A successful training with over-parameterization (2 filters in the teacher network and 4 filters in the student network). Node j_3 and j_4 are lucky draws with strong overlap with two teacher node j_1° and j_2°, and thus converges with high weight magnitude. **(b)** Lottery ticket phenomenon: initialize node j_3 and j_4 with the same initial weight, clamp the weight of j_1 and j_2 to zero, and retrain the model, the test performance becomes better since j_3 and j_4 still converge to their teacher node, respectively. **(c)** If we reinitialize node j_3 and j_4, it is highly likely that they are not overlapping with teacher node j_1° and j_2° so the performance is not good.

ima is related to noisy data (Fig. 2(d)), in which more student nodes match with the teacher.

Implicit regularization. On the other hand, the snapping behavior enforces *winner-take-all*: after optimization, a teacher node is fully covered (explained) by a few student nodes, rather than splitting amongst student nodes due to over-parameterization. This explains why the same network, once trained with structured data, can generalize to the test set.

Lottery Tickets. Lottery Tickets (6; 7) is an interesting phenomenon: if we reset "salient weights" (trained weights with large magnitude) back to the values before optimization but after initialization, prune other weights (often $> 90\%$ of total weights) and retrain the model, the test performance is the same or better; if we reinitialize salient weights, the test performance is much worse. In our theory, the salient weights are those lucky regions (E_{j_3} and E_{j_4} in Fig. 3) that happen to overlap with some teacher nodes after initialization and converge to them in optimization. Therefore, if we reset their weights and prune others away, they can still converge to the same set of teacher nodes, and potentially achieve better performance due to less interference with other irrelevant nodes. However, if we reinitialize them, they are likely to fall into unfavorable regions which cannot cover teacher nodes, and therefore lead to poor performance (Fig. 3(c)), just like in the case of under-parameterization.

2. Overview of the Framework

The details of our proposed theory can be found in Appendix (Sec. 5). Here we list the summary. First we show that for multilayered ReLU, there exists a relationship between the gradient $g_j(x)$ of a student node j and teacher and student's activations of the same layer (Thm. 1):

$$g_j(x) = f_j'(x)\left[\sum_{j^\circ} \beta_{jj^\circ}^* f_{j^\circ}(x) - \sum_{j'} \beta_{jj'} f_{j'}(x)\right], \tag{1}$$

where f_{j° is the activation of node j° in the teacher, and j' is the node at the same layer in the student. For each node j, we don't know which teacher node corresponds to it, hence the linear combination terms. Typically the number of student nodes is much more than that of teachers'. Thm. 1 applies to arbitrarily deep ReLU networks.

Then with a mild assumption (Assumption 1), we can write the gradient update rule of each layer l in the following

concise form (also Eqn. 8 in the Appendix):

$$\dot{W}_l = L_l^* W_l^* H_{l-1}^* - L_l W_l H_{l-1} \quad (2)$$

where L and L^* are correlations matrix of activations from the bottom layers, and H and H^* are *modulation* matrix from the top layers.

We then make an assumption that different teacher nodes of the same layer have small overlap in node activations (Assumption 3 and Fig. 7), and verify it in VGG16/VGG11 by showing that the majority of their BatchNorm bias are negative (Fig. 4 and Fig. 14). With this assumption, we prove two theorems:

- When the number of student nodes is the same as the number of teacher nodes ($m_l = n_l$), and each student's weight vector $\mathbf{w}_j$ is close to a corresponding teacher $\mathbf{w}_{j^\circ}^*$, then the dynamics of Eqn. 2 yields (recovery) convergence $\mathbf{w}_j \to \mathbf{w}_{j^\circ}^*$ (Thm. 4). Furthermore, such convergence is super-linear (i.e., the convergence rate is higher when the weights are closer).
- In the over-parameterization setting ($n_l > m_l$), we show that in the 2-layer case, with the help of top-layer, the portion of weights W_u that are close to teacher W^* converge ($W_u \to W^*$). For other irrelevant weights, while their final values heavily depends on initialization, with the help of top-down modulation, their fan-out top-layer weights converge to zero, and thus have no influence on the network output.

3. Experiments

3.1. Checking Assumption 3

To make Theorem 4 and Theorem 5 work, we make Assumption 3 that the activation field of different teacher nodes should be well-separated. To justify this, we analyze the BatchNorm bias of pre-trained VGG11 and VGG16. We check the BatchNorm bias c_1 as both VGG11 and VGG16 use Linear-BatchNorm-ReLU architecture. After BatchNorm first normalizes the input data into zero mean distribution, the BatchNorm bias determines how much data pass the ReLU threshold. If the bias is negative, then a small portion of data pass ReLU gating and Assumption 3 is likely to hold. From Fig. 4, it is quite clear that the majority of BatchNorm bias parameters are negative, in particular for the top layers.

3.2. Experiment Setup

We evaluate both the fully connected (FC) and ConvNet setting. For FC, we use a ReLU teacher network of size 50-75-100-125. For ConvNet, we use a teacher with channel size 64-64-64-64. The student networks have the same depth but with 10x more nodes/channels at each layer, such that they are substnatially over-parameterized. When BatchNorm is added, it is added after ReLU.

We use random i.i.d Gaussian inputs with mean 0 and std 10 (abbreviated as `GAUS`) and `CIFAR-10` as our dataset in the experiments. `GAUS` generates infinite number of samples while `CIFAR-10` is a finite dataset. For `GAUS`, we use a random teacher network as the label provider (with 100 classes). To make sure the weights of the teacher are weakly overlapped, we sample each entry of $\mathbf{w}_j^*$ from $[-0.5, -0.25, 0, 0.25, 0.5]$, making sure they are non-zero and mutually different within the same layer, and sample biases from $U[-0.5, 0.5]$. In the FC case, the data dimension is 20 while in the ConvNet case it is 16×16. For `CIFAR-10` we use a pre-trained teacher network with BatchNorm. In the FC case, it has an accuracy of 54.95%; for ConvNet, the accuracy is 86.4%. We repeat 5 times for all experiments, with different random seed and report min/max values.

Two metrics are used to check our prediction that some lucky student nodes converge to the teacher:

Normalized correlation $\bar{\rho}$. We compute normalized correlation (or cosine similarity) ρ between teacher and student activations evaluated on a validation set. At each layer, we average the best correlation over teacher nodes: $\bar{\rho} = \text{mean}_{j^\circ} \max_j \rho_{jj^\circ}$, where ρ_{jj° is computed for each teacher and student pairs (j, j°). $\bar{\rho} \approx 1$ means that most teacher nodes are covered by at least one student.

Mean Rank $\bar{r}$. After training, each teacher node j° has the most correlated student node j. We check the correlation rank of j, normalized to $[0, 1]$ (0=rank first), back at initialization and at different epoches, and average them over teacher nodes to yield mean rank $\bar{r}$. Small $\bar{r}$ means that student nodes that initially correlate well with the teacher keeps the lead toward the end of training.

3.3. Results

Experiments are summarized in Fig. 5 and Fig. 6. $\bar{\rho}$ indeed grows during training, in particular for low layers that are closer to the input, where $\bar{\rho}$ moves towards 1. Furthermore, the final winning student nodes also have a good rank at the early stage of training. BatchNorm helps a lot, in particular for the CNN case with `GAUS` dataset. For `CIFAR-10`, the final evaluation accuracy (see Appendix) learned by the student is often $\sim 1\%$ higher than the teacher. Using BatchNorm accelerates the growth of accuracy, improves $\bar{r}$, but seems not to accelerate the growth of $\bar{\rho}$.

The theory also predicts that the top-down modulation $\boldsymbol{\beta}$ helps the convergence. For this, we plot $\beta_{jj^\circ}^*$ at different layers during optimization on `GAUS`. For better visualization, we align each student node index j with a teacher node j° according to highest ρ. Despite the fact that correlations are computed from the low-layer weights, it matches well with the top-layer modulation (identity matrix structure in Fig. 16). More ablation studies are in Sec. 8.

4. Conclusion

We propose a novel mathematical framework for multi-layered ReLU networks. This could tentatively explain many puzzling empirical phenomena in deep learning.

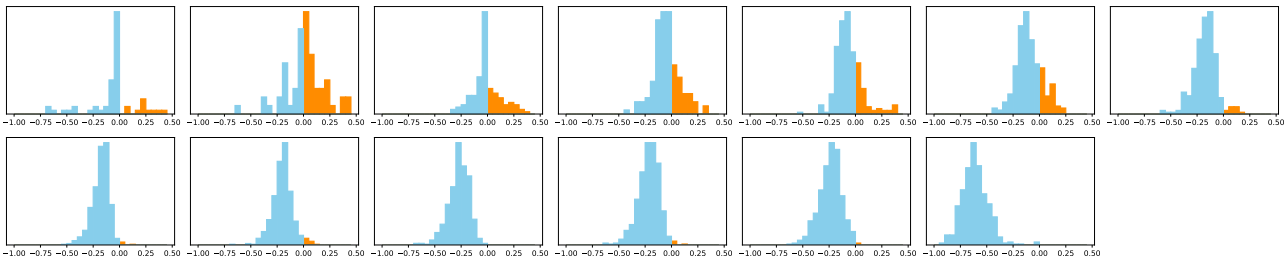

Figure 4. Distribution of BatchNorm bias in pre-trained VGG16 on ImageNet. Orange/blue are positive/negative biases. The first plot corresponds to the lowest layer (closest to the input). VGG11 in Fig. 14.

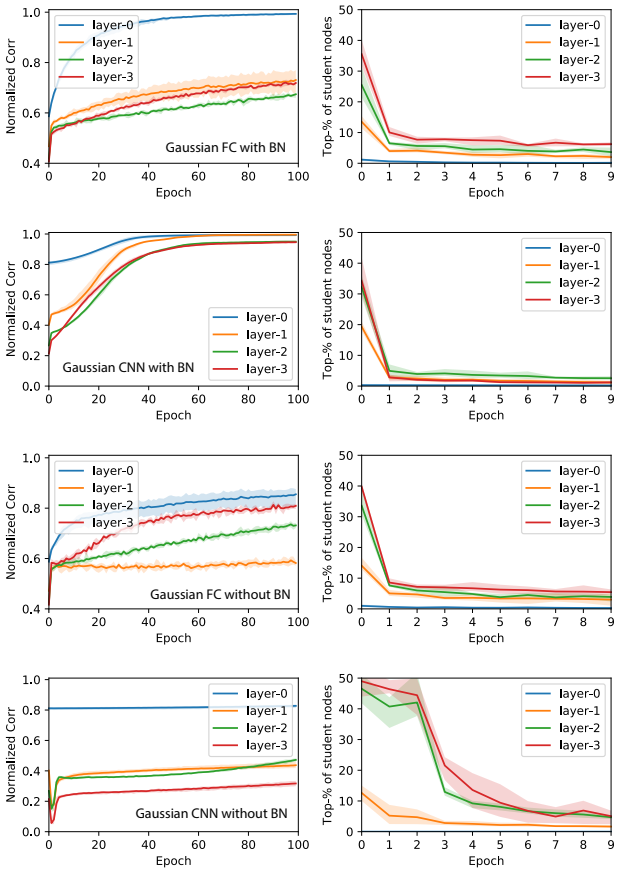

Figure 5. Correlation $\bar{\rho}$ and mean rank $\bar{r}$ over training on `GAUS`. $\bar{\rho}$ steadily grows and $\bar{r}$ quickly improves over time. `Layer-0` (the lowest layer that is closest to the input) shows best match with teacher nodes and best mean rank. BatchNorm helps achieve both better correlation and lower $\bar{r}$, in particular for the CNN case.

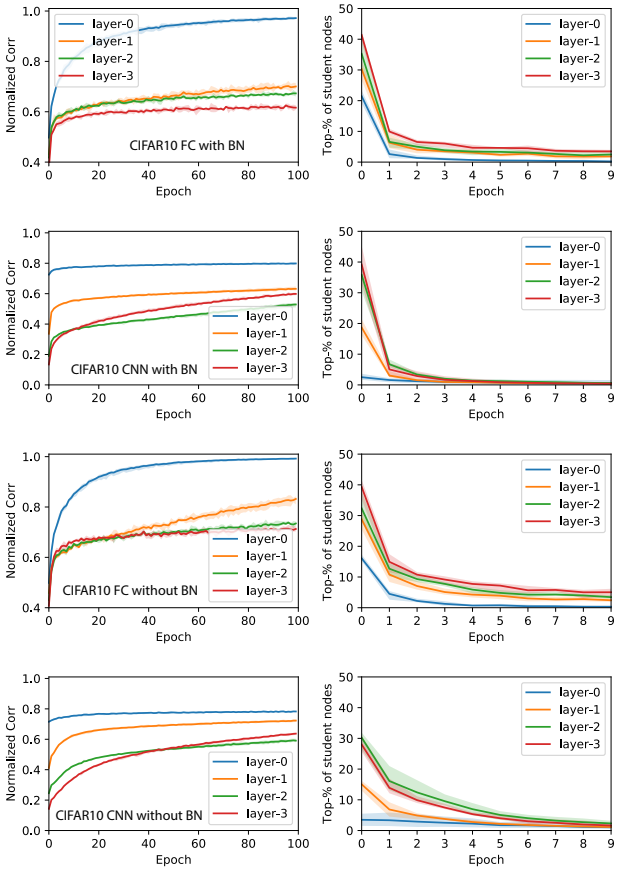

Figure 6. Same experiment setting as in Fig. 5 on `CIFAR-10`. BatchNorm helps achieve lower $\bar{r}$.

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[2] Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, G Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli. Comparing dynamics: Deep neural networks versus glassy systems. *ICML*, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.

[5] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *ICML*, 2018.

[6] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Training pruned neural networks. *ICLR*, 2019.

[7] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611*, 2019.

[8] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[9] Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *IEEE international conference on neural networks*, pages 293–299. IEEE, 1993.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

[14] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*, 2017.

[15] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Ming Zhou, Klaus Neymeyr, and Thomas Hofmann. Towards a theoretical understanding of batch normalization. *arXiv preprint arXiv:1805.10694*, 2018.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] David Krueger, Nicolas Ballas, Stanislaw Jastrzebski, Devansh Arpit, Maxinder S Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, and Aaron Courville. Deep nets don't learn via memorization. *ICLR Workshop*, 2017.

[18] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.

[19] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *ICLR*, 2018.

[20] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399, 2018.

[21] Zachary C Lipton. Stuck in a what? adventures in weight space. *arXiv preprint arXiv:1602.07320*, 2016.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[24] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[25] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[26] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *ICLR Workshop*, 2015.

[27] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[28] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *ICLR*, 2017.

[29] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *ICLR 2018 Workshop Contribution, arXiv preprint arXiv:1706.04454*, 2018.

[30] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[31] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

[34] Yuandong Tian. A theoretical framework for deep locally connected relu network. *arXiv preprint arXiv:1809.10829*, 2018.

[35] Yuandong Tian and Yan Zhu. Better computer go player with neural network and long-term prediction. *ICLR*, 2016.

[36] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[37] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. *ICLR*, 2019.

[38] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

[39] Chiyuan Zhang, Samy Bengio, Moritz Hardt, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.

5. Appendix: Mathematical Framework

Notation. Consider a student network and its associated teacher network (Fig. 1(a)). Denote the input as x. For each node j, denote $f_j(x)$ as the activation, $f'_j(x)$ as the ReLU gating, and $g_j(x)$ as the backpropagated gradient, all as functions of x. We use the superscript $^\circ$ to represent a teacher node (e.g., j°). Therefore, g_{j° never appears as teacher nodes are not updated. We use w_{jk} to represent weight between node j and k in the student network. Similarly, $w^*_{j^\circ k^\circ}$ represents the weight between node j° and k° in the teacher network.

We focus on multi-layered ReLU networks. We use the following equality extensively: $\sigma(x) = \sigma'(x)x$. For ReLU node j, we use $E_j \equiv \{x : f_j(x) > 0\}$ as the activation region of node j.

Objective. We assume that both the teacher and the student output probabilities over C classes. We use the output of teacher as the input of the student. At the top layer, each node c in the student corresponds to each node c° in the teacher. Therefore, the objective is:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2}\mathbb{E}_x\left[\|f_c(x) - f_{c^\circ}(x)\|^2\right] \tag{3}$$

By the backpropagation rule, we know that for each sample x, the (negative) gradient $g_c(x) \equiv \partial J/\partial f_c = f_{c^\circ}(x) - f_c(x)$. The gradient gets backpropagated until the first layer is reached.

Note that here, the gradient $g_c(x)$ sent to node c is *correlated* with the activation of the corresponding teacher node $f_{c^\circ}(x)$ and other student nodes at the same layer. Intuitively, this means that the gradient "pushes" the student node c to align with class c° of the teacher. If so, then the student learns the corresponding class well. A natural question arises:

Are student nodes at intermediate layers correlated with teacher nodes at the same layers?

One might wonder this is hard since the student's intermediate layer receives no *direct supervision* from the corresponding teacher layer, but relies only on backpropagated gradient. Surprisingly, the following theorem shows that it is possible for every intermediate layer:

Theorem 1 (Recursive Gradient Rule)**.** *If all nodes j at layer l satisfies Eqn. 4*

$$g_j(x) = f'_j(x)\left[\sum_{j^\circ}\beta^*_{jj^\circ}(x)f_{j^\circ}(x) - \sum_{j'}\beta_{jj'}(x)f_{j'}(x)\right], \tag{4}$$

then all nodes k at layer $l-1$ also satisfies Eqn. 4 with $\beta^*_{kk^\circ}(x)$ and $\beta_{kk'}(x)$ defined as follows:

$$\beta^*_{kk^\circ}(x) \equiv \sum_{jj^\circ} w_{jk}f'_j(x)\beta^*_{jj^\circ}(x)f'_{j^\circ}(x)w^*_{j^\circ k^\circ}, \quad \beta_{kk'}(x) \equiv \sum_{jj'} w_{jk}f'_j(x)\beta_{jj'}(x)f'_{j'}(x)w_{j'k'} \tag{5}$$

Note that this formulation allows different number of nodes for the teacher and student. In particular, we consider the *over-parameterization* setting: the number of nodes on the student side is much larger (e.g., 5-10x) than the number of nodes on the teacher side. Using Theorem 1, we discover a novel and concise form of gradient update rule:

Assumption 1 (Separation of Expectations)**.**

$$\mathbb{E}_x\left[\beta^*_{jj^\circ}(x)f'_j(x)f'_{j^\circ}(x)f_k(x)f_{k^\circ}(x)\right] = \mathbb{E}_x\left[\beta^*_{jj^\circ}(x)\right]\mathbb{E}_x\left[f'_j(x)f'_{j^\circ}(x)\right]\mathbb{E}_x\left[f_k(x)f_{k^\circ}(x)\right] \tag{6}$$

$$\mathbb{E}_x\left[\beta_{jj'}(x)f'_j(x)f'_{j'}(x)f_k(x)f_{k'}(x)\right] = \mathbb{E}_x\left[\beta_{jj'}(x)\right]\mathbb{E}_x\left[f'_j(x)f'_{j'}(x)\right]\mathbb{E}_x\left[f_k(x)f_{k'}(x)\right] \tag{7}$$

Theorem 2. *If Assumption 1 holds, the gradient dynamics of deep ReLU networks with objective (Eqn. 3) is:*

$$\dot{W}_l = L^*_l W^*_l H^*_{l+1} - L_l W_l H_{l+1} \tag{8}$$

Here we explain the notations. $W^*_l = \left[\mathbf{w}^*_1, \ldots, \mathbf{w}^*_{m_l}\right]$ is m_l teacher weights, $\boldsymbol{\beta}^*_{l+1} = \mathbb{E}_x\left[\beta_{jj^\circ}(x)\right]$, $d^*_{jj^\circ} = \mathbb{E}_x\left[f'_j(x)f'_{j^\circ}(x)\right]$ and $D^*_l = [d^*_{jj^\circ}]$, $H^*_{l+1} = [h_{jj^\circ}] = \boldsymbol{\beta}^*_{l+1} \circ D_l$, $l^*_{jj^\circ} = \mathbb{E}_x\left[f_j(x)f_{j^\circ}(x)\right]$ and $L^*_l = [l^*_{jj^\circ}]$. We can define similar notations for W (which has n_l columns/filters), $\boldsymbol{\beta}$, D, H and L (Fig. 7(c)). At the lowest layer $l = 0$, $L_0 = L^*_0$, at the highest layer $l = l_{\max} - 1$ where there is no ReLU, we have $\boldsymbol{\beta}_{l_{\max}} = \boldsymbol{\beta}^*_{l_{\max}} = H_{l_{\max}} = H^*_{l_{\max}} = I$ due to Eqn. 3. According to network structure, $\boldsymbol{\beta}_{l+1}$ and $\boldsymbol{\beta}^*_{l+1}$ only depends on weights $W_{l+1}, \ldots, W_{l_{\max}-1}$, while L_l and L^*_l only depend on $W_0, \ldots, W_{l-1}$.

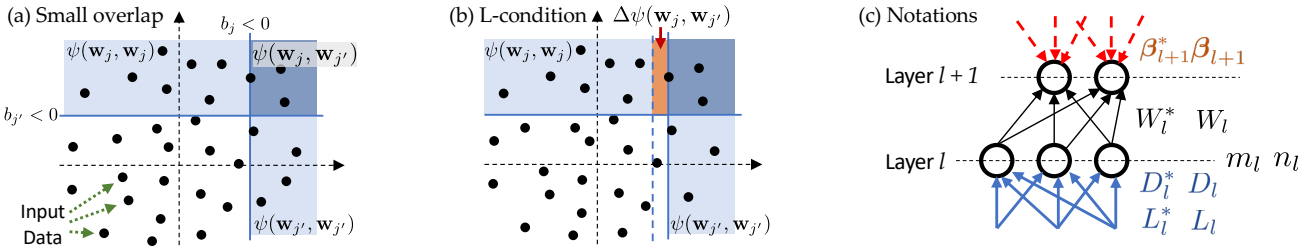

Figure 7. **(a)** Small overlaps between node activations. **(b)** Assumption 2. **(c)** Notation in Thm. 2.

6. Appendix: Analysis on the Dynamics

In the following, we will use Eqn. 8 to analyze the dynamics of the multi-layer ReLU networks. For convenience, we first define the two functions ψ_l and ψ_d (σ is the ReLU function):

$$\psi_l(\mathbf{w}, \mathbf{w}') = \mathbb{E}_x\left[\sigma(\mathbf{w}^T x)\sigma(\mathbf{w}'^T x)\right], \quad \psi_d(\mathbf{w}, \mathbf{w}') = \mathbb{E}_x\left[\mathbb{I}(\mathbf{w}^T x)\mathbb{I}(\mathbf{w}'^T x)\right]. \tag{9}$$

We assume these two functions have the following property .

Assumption 2 (Lipschitz condition)**.** *There exists K_d and K_l so that:*

$$\|\psi_i(\mathbf{w}, \mathbf{w}_1) - \psi_i(\mathbf{w}, \mathbf{w}_2)\| \le \psi_i(\mathbf{w}, \mathbf{w}_1)(1 + K_i\|\mathbf{w}_1 - \mathbf{w}_2\|), \quad i \in \{d, l\} \tag{10}$$

Using this, we know that $d_{jj'} = \psi_d(\mathbf{w}_j, \mathbf{w}_{j'})$, $d^*_{jj'} = \psi_d(\mathbf{w}_j, \mathbf{w}^*_{j'})$, and so on. For brevity, denote $d^{**}_{jj'} = \psi_d(\mathbf{w}^*_j, \mathbf{w}^*_{j'})$ (when notation j°_1 is heavy) and so on. We impose the following assumption:

Assumption 3 (Small Overlap between teacher nodes)**.** *There exists $\epsilon_l \ll 1$ and $\epsilon_d \ll 1$ so that:*

$$d^{**}_{j_1 j_2} \le \epsilon_d d^{**}_{j_1 j_1} \text{ (or } \epsilon_d d^{**}_{j_2 j_2}), \quad l^{**}_{j_1 j_2} \le \epsilon_l l^{**}_{j_1 j_1} \text{ (or } \epsilon_l l^{**}_{j_2 j_2}), \quad \text{for } j_1 \neq j_2 \tag{11}$$

Intuitively, this means that the probability of the simultaneous activation of two teacher nodes j_1 and j_2 is small. One such case is that the teacher has negative bias, which means that they *cut corners* in the input space (Fig. 7a). We have empirically verified that the majority of biases in BatchNorm layers (after the data are whitened) are negative in VGG11/16 trained on ImageNet (Sec. 3.1).

6.1. Effects of BatchNorm

Batch Normalization (13) has been extensively used to speed up the training, reduce the tuning efforts and improve the test performance of neural networks. Here we use an interesting property of BatchNorm: the total "energy" of the incoming weights of each node j is conserved over training iterations:

Theorem 3 (Conserved Quantity in Batch Normalization)**.** *For Linear $\to$ ReLU $\to$ BN or Linear $\to$ BN $\to$ ReLU configuration, $\|\mathbf{w}_j\|$ of a filter j before BN remains constant in training. (Fig. 11).*

See Appendix for the proof. This may partially explain why BN has stabilization effect: energy will not leak from one layer to nearby ones. Due to this property, in the following, for convenience we assume $\|\mathbf{w}_j\|^2 = \|\mathbf{w}^*_j\|^2 = 1$, and the gradient $\dot{\mathbf{w}}_j$ is always orthogonal to the current weight $\mathbf{w}_j$. Note that on the teacher side we can always push the magnitude component to the upper layer; on the student side, random initialization naturally leads to constant magnitude of weights.

6.2. Same number of student nodes as teacher

If $n_l = m_l$, $L^*_l = L_l = I$ (e.g., the input of layer l is whitened) and $\boldsymbol{\beta}^*_{l+1} = \boldsymbol{\beta}_{l+1} = \mathbf{1}\mathbf{1}^T$ (all β entries are 1), then the following theorem shows that weight recovery could follow (we use j' as j°).

Theorem 4. *For dynamics $\dot{\mathbf{w}}_j = P^\perp_{\mathbf{w}_j}(W^*\mathbf{h}^*_j - W\mathbf{h}_j)$, where $P^\perp_{\mathbf{w}_j} \equiv I - \mathbf{w}_j\mathbf{w}_j^T$ is a projection matrix into the orthogonal complement of $\mathbf{w}_j$. $\mathbf{h}^*_j$, $\mathbf{h}_j$ are corresponding j-th column in H^* and H. Denote $\theta_j = \angle(\mathbf{w}_j, \mathbf{w}^*_j)$ and assume $\theta_j \le \theta_0$. If $\gamma = \cos\theta_0 - (m-1)\epsilon_d M_d > 0$, then $\mathbf{w}_j \to \mathbf{w}^*_j$ with the rate $1 - \eta\bar{d}\gamma$ (η is learning rate). Here $\bar{d} = [1 + 2K_d \sin(\theta_0/2)] \min_j d^{*0}_{jj}$ and $M_d = (1 + K_d)\left[1 + 2K_d \sin(\theta_0/2)\right]^2 / \cos\frac{\theta_0}{2}$.*

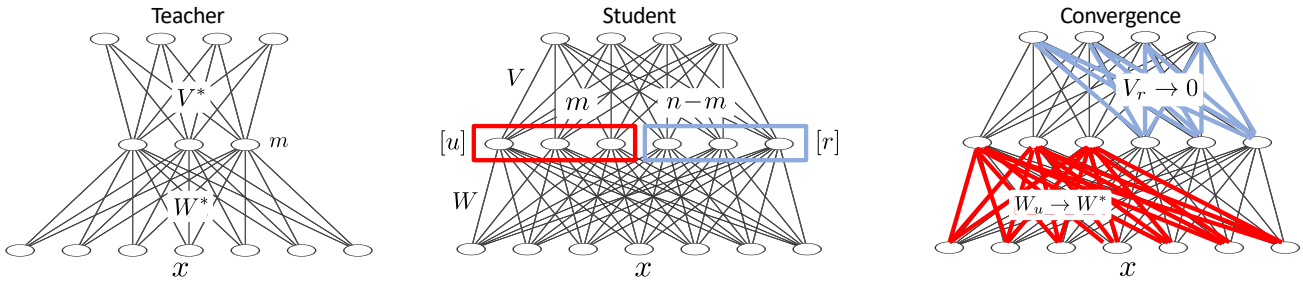

Figure 8. Over-parameterization and top-down modulation. Thm. 5 shows that under certain conditions, the relevant weights $W_u \to W^*$ and weights connecting to irrelevant student nodes $V_r \to 0$.

See Appendix for the proof. Here we list a few remarks:

Faster convergence near $\mathbf{w}_j^*$. we can see that due to the fact that h_{jj}^* in general becomes larger when $\mathbf{w}_j \to \mathbf{w}_j^*$ (since $\cos\theta_0$ can be close to 1), we expect a *super-linear* convergence near $\mathbf{w}_j^*$. This brings about an interesting *winner-take-all* mechanism: if the initial overlap between a student node j and a particular teacher node is large, then the student node will snap to it (Fig. 1(c)).

Importance of projection operator $P_{\mathbf{w}_j}^{\perp}$. Intuitively, the projection is needed to remove any ambiguity related to weight scaling, in which the output remains constant if the top-layer weights are multiplied by a constant α, while the low-layer weights are divided by α. Previous works (5) also uses similar techniques while we justify it with BN. Without $P_{\mathbf{w}_j}^{\perp}$, convergence can be harder.

6.3. Over-Parameterization and Top-down Modulation

In the over-parameterization case ($n_l > m_l$, e.g., 5-10x), we arrange the variables into two parts: $W = [W_u, W_r]$, where W_u contains m_l columns (same size as W^*), while W_r contains $n_l - m_l$ columns. We use $[u]$ (or u-set) to specify nodes $1 \le j \le m$, and $[r]$ (or r-set) for the remaining part.

In this case, if we want to show "the main component" W_u converges to W^*, we will meet with one core question: to where will W_r converge, or whether W_r will even converge at all? We need to consider not only the dynamics of the current layer, but also the dynamics of the upper layer. Using a 1-hidden layer over-parameterized ReLU network as an example, Theorem 5 shows that the upper-layer dynamics $\dot{V} = L^*V^* - LV$ automatically apply *top-down modulation* to suppress the influence of W_r, regardless of their convergence. Here $V = \begin{bmatrix} V_u \\ V_r \end{bmatrix}$, where V_u are the weight components of u-set. See Fig. 8.

Theorem 5 (Over-Parameterization and Top-down Modulation)**.** *Consider $\dot{W} = W^*H^* - WH$ with over-parameterization ($n > m$) and its upper-layer dynamics $\dot{V} = L^*V^* - LV$. Assume that initial value W_u^0 is close to W^*: $\theta_j = \angle(\mathbf{w}_j, \mathbf{w}_j^*) \le \theta_0$ for $j \in [u]$. If (1) Assumption 3 holds for all pairwise combination of columns of W^** ***and*** *W_r^0, and (2) there exists $\gamma = \gamma(\theta_0, m) > 0$ and $\bar{\lambda}$ so that Eqn. 43 and Eqn. 44 holds, then $W_u \to W^*$, $V_u \to V^*$ and $V_r \to 0$ with rate $1 - \eta\bar{\lambda}\gamma$.*

See Appendix for the proof (and definition of $\bar{\lambda}$ in Eqn. 47). The intuition is: if W_u is close to W^* and W_r are far away from them due to Assumption 3, the off-diagonal elements of L and L^* are smaller than diagonal ones. This causes V_u to move towards V^* and V_r to move towards zero. When V_r becomes small, so does $\beta_{jj'}$ for $j \in [r]$ or $j' \in [r]$. This in turn suppresses the effect of W_r and accelerates the convergence of W_u. $V_r \to 0$ exponentially so that W_r stays close to its initial locations, and Assumption 3 holds for all iterations. A few remarks:

Flat minima. Since $V_r \to 0$, W_r can be changed arbitrarily without affecting the outputs of the neural network. This could explain why there are many flat directions in trained networks, and why many eigenvalues of the Hessian are close to zero (28).

Understanding of pruning methods. Theorem 5 naturally relates two different unstructured network pruning approaches: pruning small weights in magnitude (8; 6) and pruning weights suggested by Hessian (18; 9). It also suggests a principled structured pruning method: instead of pruning a filter by checking its weight norm, pruning accordingly to its top-down modulation.

Accelerated convergence and learning rate schedule. For simplicity, the theorem uses a uniform (and conservative)

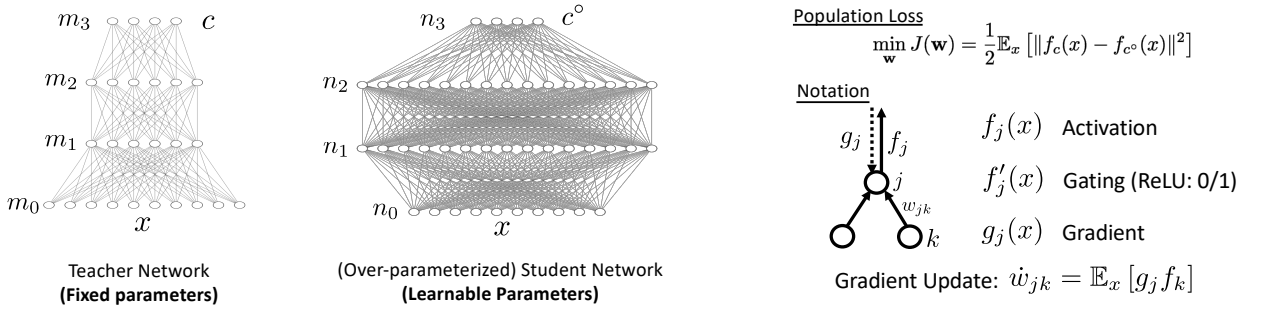

Figure 9. Teacher-Student Setting, loss function and notations.

γ throughout the iterations. In practice, γ is initially small (due to noise introduced by r-set) but will be large after a few iterations when V_r vanishes. Given the same learning rate, this leads to accelerated convergence. At some point, the learning rate η becomes too large, leading to fluctuation. In this case, η needs to be reduced.

Many-to-one mapping. Theorem 5 shows that under strict conditions, there is one-to-one correspondence between teacher and student nodes. In general this is not the case. Two students nodes can be both in the vicinity of a teacher node $\mathbf{w}_j^*$ and converge towards it, until that node is fully explained. We leave it to the future work for rigid mathematical analysis of many-to-one mappings.

Random initialization. One nice thing about Theorem 5 is that it only requires the initial $\|W_u - W^*\|$ to be small. In contrast, there is *no* requirement for small $\|V_r\|$. Therefore, we could expect that with more over-parameterization and random initialization, in each layer l, it is more likely to find the u-set (of fixed size m_l), or the *lucky weights*, so that W_u is quite close to W^*. At the same time, we don't need to worry about $\|W_r\|$ which grows with more over-parameterization. Moreover, random initialization often gives orthogonal weight vectors, which naturally leads to Assumption 3.

6.4. Extension to Multi-layer ReLU networks

Using a similar approach, we could extend this analysis to multi-layer cases. We *conjecture* that similar behaviors happen: for each layer, due to over-parameterization, the weights of some *lucky* student nodes are close to the teacher ones. While these converge to the teacher, the final values of others *irrelevant* weights are initialization-dependent. If the irrelevant nodes connect to lucky nodes at the upper-layer, then similar to Thm. 5, the corresponding fan-out weights converge to zero. On the other hand, if they connect to nodes that are also irrelevant, then these fan-out weights are not-determined and their final values depends on initialization. However, it doesn't matter since these upper-layer irrelevant nodes eventually meet with zero weights if going up recursively, since the top-most output layer has no over-parameterization. We leave a formal analysis to future work.

7. Appendix: Proofs

7.1. Theorem 1

Proof. The first part of gradient backpropagated to node j is:

$$g_j^1(x) = f'_j(x)\sum_{j^\circ}\beta^*_{jj^\circ}(x)f_{j^\circ}(x) \tag{12}$$

$$= f'_j(x)\sum_{j^\circ}\beta^*_{jj^\circ}(x)f'_{j^\circ}(x)\sum_{k^\circ}w^*_{j^\circ k^\circ}f_{k^\circ}(x) \tag{13}$$

$$= \sum_{k^\circ}\left[f'_j(x)\sum_{j^\circ}\beta^*_{jj^\circ}(x)f'_{j^\circ}(x)w^*_{j^\circ k^\circ}\right]f_{k^\circ}(x) \tag{14}$$

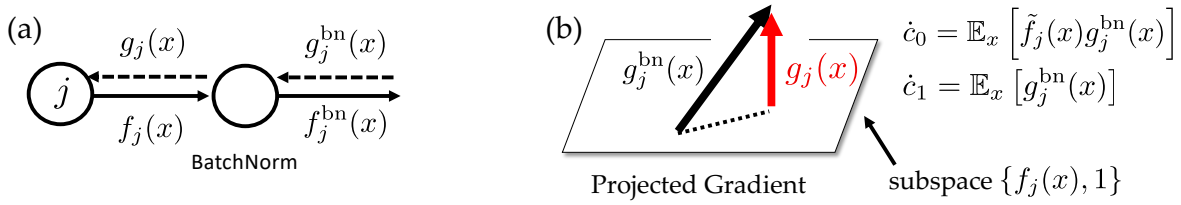

Figure 10. BatchNorm explanation

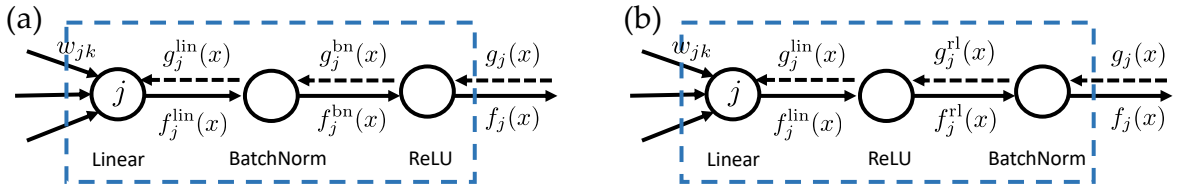

Figure 11. Different BatchNorm Configuration.

Therefore, for the gradient to node k, we have:

$$g_k^1(x) = f_k'(x) \sum_j w_{jk} g_j^1(x) \tag{15}$$

$$= f_k'(x) \sum_{k^\circ} \underbrace{\left[\sum_{jj^\circ} w_{jk} f_j'(x) \beta_{jj^\circ}^*(x) f_{j^\circ}'(x) w_{j^\circ k^\circ}^*\right]}_{\beta_{kk^\circ}^*(x)} f_{k^\circ}(x) \tag{16}$$

And similar for $\beta_{kk'}(x)$. Therefore, by mathematical induction, we know that all gradient at nodes in different layer follows the same form. □

7.2. Theorem 2

Proof. Using Thm. 1, we can write down weight update for weight w_{jk} that connects node j and node k:

$$\dot{w}_{jk} = \sum_{j^\circ,k^\circ} w_{j^\circ k^\circ}^* \underbrace{\mathbb{E}_x\left[f_j'(x)\beta_{jj^\circ}^*(x) f_{j^\circ}'(x) f_k(x) f_{k^\circ}(x)\right]}_{\beta_{jj^\circ kk^\circ}^*}$$
$$- \sum_{j',k'} w_{j'k'} \underbrace{\mathbb{E}_x\left[f_j'(x)\beta_{jj'}(x) f_{j'}'(x) f_k(x) f_{k'}(x)\right]}_{\beta_{jj'kk'}} \tag{17}$$

Note that j°, k°, j' and k' run over all parents and children nodes on the teacher side. This formulation works for over-parameterization (e.g., j° and j' can run over different nodes). Applying Assumption 1 and rearrange terms in matrix form yields Eqn. 8. □

7.3. Theorem 3

Proof. Given a batch with size N, denote pre-batchnorm activations as $\mathbf{f} = [f_j(x_1), \ldots, f_j(x_N)]^T$ and gradients as $\mathbf{g} = [g_j(x_1), \ldots, g_j(x_N)]^T$ (See Fig. 10(a)). $\tilde{\mathbf{f}} = (\mathbf{f} - \mu)/\sigma$ is its whitened version, and $c_0\tilde{\mathbf{f}} + c_1$ is the final output of BN. Here $\mu = \frac{1}{N}\sum_i f_j(x_i)$ and $\sigma^2 = \frac{1}{N}\sum_i (f_j(x_i) - \mu)^2$ and c_1, c_0 are learnable parameters. With vector notation, the gradient update in BN has a compact form with clear geometric meaning:

Lemma 1 (Backpropagation of Batch Norm (34))**.** *For a top-down gradient* $\mathbf{g}$*, BN layer gives the following gradient update (*$P_{\mathbf{f},\mathbf{1}}^\perp$ *is the orthogonal complementary projection of subspace* $\{\mathbf{f}, \mathbf{1}\}$*):*

$$\mathbf{g_f} = J^{BN}(\mathbf{f})\mathbf{g} = \frac{c_0}{\sigma} P_{\mathbf{f},\mathbf{1}}^\perp \mathbf{g}, \quad \mathbf{g_c} = S(\mathbf{f})^T \mathbf{g} \tag{18}$$

Intuitively, the back-propagated gradient $J^{BN}(\mathbf{f})\mathbf{g}$ is zero-mean and perpendicular to the input activation $\mathbf{f}$ of BN layer, as illustrated in Fig. 10. Unlike (15; 37) that analyzes BN in an approximate manner, in Thm. 1 we do not impose any assumptions.

Given Lemma 1, we can prove Thm. 3. For Fig. 11(a), using the property that $\mathbb{E}_x\left[g_j^{\text{lin}} f_j^{\text{lin}}\right] = 0$ (the expectation is taken over batch) and the weight update rule $\dot{w}_{jk} = \mathbb{E}_x\left[g_j^{\text{lin}} f_k\right]$ (over the same batch), we have:

$$\frac{1}{2}\frac{\mathrm{d}\|\mathbf{w}_j\|^2}{\mathrm{d}t} = \sum_{k\in\text{ch}(j)} w_{jk}\dot{w}_{jk} = \mathbb{E}_x\left[\sum_{k\in\text{ch}(j)} w_{jk} f_k(x) g_j^{\text{lin}}(x)\right] = \mathbb{E}_x\left[f_j^{\text{lin}}(x) g_j^{\text{lin}}(x)\right] = 0 \tag{19}$$

For Fig. 11(b), note that $\mathbb{E}_x\left[g_j^{\text{lin}} f_j^{\text{lin}}\right] = \mathbb{E}_x\left[g_j^{\text{rl}} f_j^{\text{rl}'} f_j^{\text{lin}}\right] = \mathbb{E}_x\left[g_j^{\text{rl}} f_j^{\text{rl}}\right] = 0$ and conclusion follows. □

7.4. Lemmas

For simplicity, in the following, we use $\delta\mathbf{w}_j = \mathbf{w}_j - \mathbf{w}_j^*$.

Lemma 2 (Bottom Bounds)**.** *Assume all* $\|\mathbf{w}_j\| = \|\mathbf{w}_{j'}\| = 1$. *Denote*

$$\mathbf{p}_{jj'}^* \equiv \mathbf{w}_{j'}^* d_{jj'}^*, \quad \mathbf{p}_{jj'} \equiv \mathbf{w}_{j'} d_{jj'}, \quad \Delta\mathbf{p}_{jj'} \equiv \mathbf{p}_{jj'}^* - \mathbf{p}_{jj'} \tag{20}$$

If Assumption 2 holds, we have:

$$\|\Delta\mathbf{p}_{jj'}\| \le (1 + K_d) d_{jj'}^* \|\delta\mathbf{w}_{j'}\| \tag{21}$$

If Assumption 3 also holds, then:

$$d_{jj'}^* \le \epsilon_d (1 + K_d\|\delta\mathbf{w}_{j'}\|)(1 + K_d\|\delta\mathbf{w}_j\|) d_{jj}^* \tag{22}$$

Proof. We have for $j \neq j'$:

$$\begin{aligned}
\|\Delta\mathbf{p}_{jj'}\| &= \|\mathbf{w}_{j'}^* d_{jj'}^* - \mathbf{w}_{j'} d_{jj'}\| && (23)\\
&= \|\mathbf{w}_{j'}(d_{jj'}^* - d_{jj'}) + (\mathbf{w}_{j'}^* - \mathbf{w}_{j'}) d_{jj'}^*\| && (24)\\
&\le \|\mathbf{w}_{j'}\|\|d_{jj'}^* - d_{jj'}\| + \|\mathbf{w}_{j'}^* - \mathbf{w}_{j'}\| d_{jj'}^* && (25)\\
&\le d_{jj'}^* K_d\|\delta\mathbf{w}_{j'}\| + d_{jj'}^*\|\delta\mathbf{w}_{j'}\| && (26)\\
&\le (1 + K_d) d_{jj'}^*\|\delta\mathbf{w}_{j'}\| && (27)
\end{aligned}$$

If Assumption 3 also holds, we have:

$$\begin{aligned}
d_{jj'}^* &\le d_{jj'}^{**}(1 + K_d\|\delta\mathbf{w}_{j'}\|) && (28)\\
&\le \epsilon_d d_{jj}^{**}(1 + K_d\|\delta\mathbf{w}_{j'}\| && (29)\\
&\le \epsilon_d d_{jj}^*(1 + K_d\|\delta\mathbf{w}_j\|)(1 + K_d\|\delta\mathbf{w}_{j'}\|) && (30)
\end{aligned}$$

□

Lemma 3 (Top Bounds)**.** *Denote*

$$\mathbf{q}_{jj'}^* \equiv \mathbf{v}_{j'}^* l_{jj'}^*, \quad \mathbf{q}_{jj'} \equiv \mathbf{v}_{j'} l_{jj'}, \quad \Delta\mathbf{q}_{jj'} \equiv \mathbf{q}_{jj'}^* - \mathbf{q}_{jj'} \tag{31}$$

If Assumption 2 holds, we have:

$$\|\Delta\mathbf{q}_{jj'}\| \le (1 + K_l) l_{jj'}^* \|\delta\mathbf{w}_{j'}\| \tag{32}$$

If Assumption 3 also holds, then:

$$l_{jj'}^* \le \epsilon_l (1 + K_l\|\delta\mathbf{w}_{j'}\|)(1 + K_l\|\delta\mathbf{w}_j\|) l_{jj}^* \tag{33}$$

Proof. The proof is similar to Lemma 2. □

Lemma 4 (Quadratic fall-off for diagonal elements of L)**.** *For node j, we have:*

$$\|l_{jj}^* - l_{jj}\| \le C_0 l_{jj}^* \|\delta\mathbf{w}_j\|^2 \tag{34}$$

Proof. The intuition here is that both the volume of the affected area and the weight difference are proportional to $\|\delta\mathbf{w}_j\|$. $\|l_{jj}^* - l_{jj}\|$ is their product and thus proportional to $\|\delta\mathbf{w}_j\|^2$. See Fig. 12. □

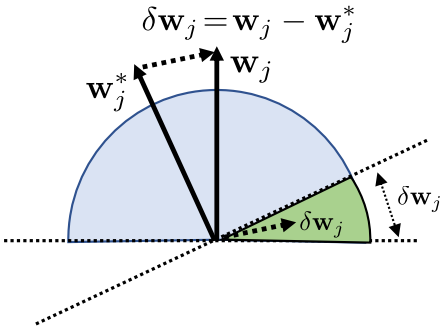

Figure 12. Explanation of Lemma. 4.

7.5. Theorem 4

Proof. First of all, note that $\|\delta \mathbf{w}_j\| = 2 \sin \frac{\theta_j}{2} \leq 2 \sin \frac{\theta_0}{2}$. So given θ_0, we also have a bound for $\|\delta \mathbf{w}_j\|$.

When $\boldsymbol{\beta} = \boldsymbol{\beta}^* = \mathbf{1}\mathbf{1}^T$, the matrix form can be written as the following:

$$\dot{\mathbf{w}}_j = P^{\perp}_{\mathbf{w}_j} \mathbf{w}_j^* h_{jj}^* + \sum_{j' \neq j} P^{\perp}_{\mathbf{w}_j} \left(\mathbf{w}_{j'}^* h_{jj'}^* - \mathbf{w}_{j'} h_{jj'} \right) = P^{\perp}_{\mathbf{w}_j} \mathbf{p}_{jj}^* + \sum_{j' \neq j} P^{\perp}_{\mathbf{w}_j} \Delta \mathbf{p}_{jj'} \tag{35}$$

by using $P^{\perp}_{\mathbf{w}_j} \mathbf{w}_j \equiv 0$ (and thus h_{jj} doesn't matter). Since $\|\mathbf{w}_j\|$ is conserved, it suffices to check whether the projected weight vector $P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j$ of $\mathbf{w}_j$ onto the complementary space of the ground truth node $\mathbf{w}_j^*$, goes to zero:

$$P^{\perp}_{\mathbf{w}_j^*} \dot{\mathbf{w}}_j = P^{\perp}_{\mathbf{w}_j^*} P^{\perp}_{\mathbf{w}_j} \mathbf{p}_{jj}^* + \sum_{j' \neq j} P^{\perp}_{\mathbf{w}_j^*} P^{\perp}_{\mathbf{w}_j} \Delta \mathbf{p}_{jj'} \tag{36}$$

Denote $\theta_j = \angle(\mathbf{w}_j, \mathbf{w}_j^*)$ and a simple calculation gives that $\sin \theta_j = \|P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j\|$. First we have:

$$P^{\perp}_{\mathbf{w}_j^*} P^{\perp}_{\mathbf{w}_j} \mathbf{w}_j^* = P^{\perp}_{\mathbf{w}_j^*} (I - \mathbf{w}_j \mathbf{w}_j^T) \mathbf{w}_j^* = -P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j \mathbf{w}_j^T \mathbf{w}_j^* = -\cos \theta_j P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j \tag{37}$$

From Lemma 2, we knows that

$$\|\Delta \mathbf{p}_{jj'}\| \leq (1 + K_d) d_{jj'}^* \|\delta \mathbf{w}_{j'}\| \leq \epsilon_d (1 + K_d) \left[1 + 2K_d \sin(\theta_0/2)\right]^2 d_{jj}^* \|\delta \mathbf{w}_{j'}\| \tag{38}$$

Note that here we have $\|\delta \mathbf{w}_{j'}\| = 2 \sin \frac{\theta_j}{2} = \sin \theta_j / \cos \frac{\theta_j}{2} \leq \sin \theta_j / \cos \frac{\theta_0}{2}$. We discuss finite step with very small learning rate $\eta > 0$:

$$\sin \theta_j^{t+1} = \|P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j^{t+1}\| = \|P^{\perp}_{\mathbf{w}_j^*} \mathbf{w}_j^t + \eta P^{\perp}_{\mathbf{w}_j^*} \dot{\mathbf{w}}_j^t\| \tag{39}$$

$$\leq (1 - \eta d_{jj}^* \cos \theta_j^t) \sin \theta_j^t + \eta \epsilon_d M_d \sum_{j' \neq j} d_{jj}^* \sin \theta_{j'}^t \tag{40}$$

since $\|P^{\perp}_{\mathbf{w}_j^*}\| = \|P^{\perp}_{\mathbf{w}_j}\| = 1$. Here

$$M_d = (1 + K_d) \left[1 + 2K_d \sin(\theta_0/2)\right]^2 / \cos \frac{\theta_0}{2} \tag{41}$$

is an iteration independent constant.

We set $\gamma = \cos \theta_0 - (m-1)\epsilon_d M_d$. If $\gamma > 0$, denote a constant $\bar{d} = [1 + 2K_d \sin(\theta_0/2)] \min_j d_{jj}^{*0}$ and from Lemma 2 we know $d_{jj}^* \geq \bar{d}$ for all j. Then given the inductive hypothesis that $\sin \theta_j^t \leq (1 - \eta \bar{d} \gamma)^{t-1} \sin \theta_0$, we have:

$$\sin \theta_j^{t+1} \leq (1 - \eta \bar{d} \gamma)^t \sin \theta_0 \tag{42}$$

Therefore, $\sin \theta_j^t \to 0$, which means that $\mathbf{w}_j \to \mathbf{w}_j^*$. □

A few remarks:

The projection operator $P^{\perp}_{\mathbf{w}_j}$. Note that $P^{\perp}_{\mathbf{w}_j}$ is important. Intuitively, without the projection, if the same proof logic worked, one could have concluded that $\mathbf{w}$ converges to any $\alpha\mathbf{w}^*$, where α is a constant scaling factor, which is obviously wrong.

Indeed, without $P^{\perp}_{\mathbf{w}_j}$, there would be a term $\mathbf{w}^*_j h^*_{jj} - \mathbf{w}_j h_{jj}$ on RHS. This term breaks into $\mathbf{w}_j(h^*_{jj} - h_{jj}) + (\mathbf{w}^*_j - \mathbf{w}_j)h^*_{jj}$. Although there could exist C so that $\|h^*_{jj} - h_{jj}\| \leq C\|\delta\mathbf{w}_j\|$, unlike Lemma 4, C may not be small, and convergence is not guaranteed.

7.6. Theorem 5

Proof. First, only for $j \in [u]$, we have their ground truth value $\mathbf{w}^*_j$. For $j \in [r]$, we assign $\mathbf{w}^*_j = \mathbf{w}^0_j$, i.e., their initial values. As we will see, this will make things easier.

From the assumption, we know that $\sin\theta_j \leq \sin\theta_0$ for $j \in [u]$. In addition, denote that $\|\delta\mathbf{v}^0_j\| \leq B_{\delta v}$ for $j \in [u]$. Denote B_v as the bound for all $\|\mathbf{v}^*_j\|$.

Now suppose we can find a $\gamma > 0$ if the following set of equations are satisfied:

$$\gamma \geq (B_v - B_{\delta v})\cos\theta_0 - \epsilon_d(B_v + B_{\delta v})\max(B_{d,u}, B_{d,r}) > 0 \tag{43}$$
$$\gamma \geq 1 - \epsilon_l \max(B_{l,u}, B_{l,r}) - \kappa > 0 \tag{44}$$

Here

$$\bar{d} = (1 - K_d C_{d,j}) \min_j d^{*0}_{jj} > 0 \tag{45}$$
$$\bar{l} = (1 - K_l C_{l,j}) \min_j l^{*0}_{jj} > 0 \tag{46}$$
$$\bar{\lambda} = \min(\bar{d}, \bar{l}) \tag{47}$$
$$\kappa = 2C_0 \sin(\theta_0/2)(1 + B_{\delta v}) \tag{48}$$
$$C_{d,u} = 2K_d \sin(\theta_0/2) \tag{49}$$
$$C_{d,r} = \epsilon_d K_d \frac{B_{d,r}(B_v + B_{\delta v})B_v}{\bar{\lambda}\gamma(2 - \eta\bar{\lambda}\gamma)} \tag{50}$$
$$M^{uu}_d = (1 + K_d)(1 + C_{d,u})^2 / \cos\frac{\theta_0}{2} \tag{51}$$
$$M^{ur}_d = (1 + K_d)(1 + C_{d,u})(1 + C_{d,r}) \tag{52}$$
$$M^{ru}_d = (1 + K_d)(1 + C_{d,u})(1 + C_{d,r}) / \cos\frac{\theta_0}{2} \tag{53}$$
$$M^{rr}_d = (1 + K_d)(1 + C_{d,r})^2 \tag{54}$$
$$B_{d,u} = (m - 1)M^{uu}_d + (n - m)M^{ur}_d \tag{55}$$
$$B_{d,r} = (m - 1)M^{ru}_d + (n - m)M^{rr}_d \tag{56}$$

and similarly we can define C_l and M_l etc. If we can find such a $\gamma > 0$ then the dynamics converges. Here all C are close to 0 and M are close to 1.

Note that if ϵ_d and ϵ_l are small, it is obvious to see there exists a feasible $\gamma > 0$ (e.g., $\gamma = 1$).

To prove it, we maintain the following induction hypothesis for iteration t :

$$d^{*t}_{jj'} \leq \epsilon_d M_{d,jj'} d^{*t}_{jj}, \;\; l^{*t}_{jj'} \leq \epsilon_l M_{l,jj'} l^{*t}_{jj}, \quad j' \neq j \tag{W-Separation}$$

$$\sin\theta^t_j \leq (1 - \eta\bar{d}\gamma)^{t-1}\sin\theta_0, \quad j \in [u] \tag{W_u-Contraction}$$

$$\|\delta\mathbf{v}^t_j\| \leq (1 - \eta\bar{l}\gamma)^{t-1}B_{\delta v}, \;\; j \in [u], \qquad \|\mathbf{v}^t_j\| \leq (1 - \eta\bar{l}\gamma)^{t-1}B_v, \;\; j \in [r] \tag{V-Contraction}$$

Besides, the following condition is involved (but it is not part of induction hypothesis):

$$\|\mathbf{w}^t_j - \mathbf{w}^0_j\| \leq C_{d,r}, \quad j \in [r] \tag{W_r-Bound}$$

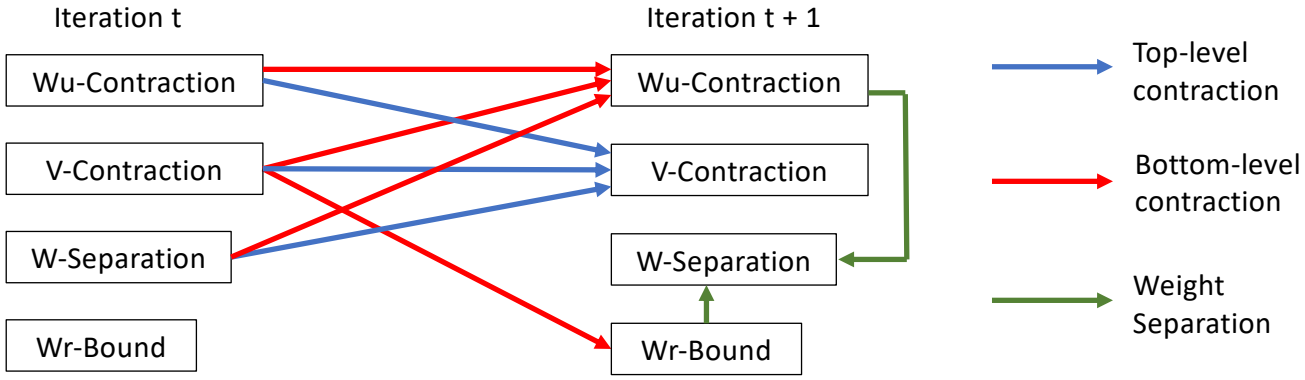

Figure 13. Proof sketch of Thm. 5.

$$d_{jj}^{*t} \geq d_{jj}^{*0}(1 - K_d C_{d,j}) \geq \bar{d} > 0, \quad l_{jj}^{*t} \geq l_{jj}^{*0}(1 - K_l C_{l,j}) \geq \bar{l} > 0 \tag{57}$$

The proof can be decomposed in the following three lemma.

Lemma 5 (Top-layer contraction)**.** *If* (W-Separation) *holds for t, then* (V-Contraction)*) holds for iteration $t+1$.*

Lemma 6 (Bottom-layer contraction)**.** *If* (V-Contraction) *holds for t, then* (W_u-Contraction) *holds for $t+1$ and* (W_r-Bound) *holds for $t+1$.*

Lemma 7 (Weight separation)**.** *If* (W-Separation) *holds for t,* (W_r-Bound) *holds for $t+1$ and* (W_u-Contraction) *holds for $t+1$, then* (W-Separation) *holds for $t+1$.*

As suggested by Fig. 13, if all the three lemmas are true then the induction hypothesis are true. □

In the following, we will prove the three lemmas.

7.6.1. Lemma 5

Proof. On the top-layer, we have $\dot{V} = L^* V^* - LV$. Denote that $V = \begin{bmatrix} \mathbf{v}_1 \\ \dots \\ \mathbf{v}_n \end{bmatrix}$, where $\mathbf{v}_j$ is the j-th row of the matrix V. For each component, we can write:

$$\dot{\mathbf{v}}_j = \mathbb{I}(j \in [u])\mathbf{q}_{jj}^* - \mathbf{q}_{jj} + \sum_{j' \neq j, j' \in [u]} \Delta \mathbf{q}_{jj'} + \sum_{j' \neq j, j' \in [r]} \mathbf{q}_{jj'} \tag{58}$$

Note that there is no projection (if there is any, the projection should be in the columns rather than the rows).

If (W-Separation) is true, we know that for $j \neq j'$,

$$\|\Delta \mathbf{q}_{jj'}\| \leq \epsilon_l M_{l,uu} l_{jj}^* \|\delta \mathbf{v}_{j'}\|, \quad \|\mathbf{q}_{jj'}\| \leq \epsilon_l M_{l,ur} l_{jj}^* \|\mathbf{v}_{j'}\|, \quad j \in [u] \tag{59}$$

$$\|\Delta \mathbf{q}_{jj'}\| \leq \epsilon_l M_{l,ru} l_{jj}^* \|\delta \mathbf{v}_{j'}\|, \quad \|\mathbf{q}_{jj'}\| \leq \epsilon_l M_{l,rr} l_{jj}^* \|\mathbf{v}_{j'}\|, \quad j \in [r] \tag{60}$$

Now we discuss $j \in [u]$ and $j \in [r]$:

Relevant nodes. For $j \in [u]$, the first two terms are:

$$\Delta \mathbf{q}_{jj} = -l_{jj}^* \delta \mathbf{v}_j + (l_{jj}^* - l_{jj}) \mathbf{v}_j \tag{61}$$

From Lemma 4 we know that:

$$\|(l_{jj}^* - l_{jj})\mathbf{v}_j\| \leq C l_{jj}^* \|\delta \mathbf{w}_j\|^2 \|\mathbf{v}_j\| \leq 2C \sin(\theta_0/2)(1 + B_{\delta v}) l_{jj}^* \|\delta \mathbf{w}_j\| = \kappa l_{jj}^* \|\delta \mathbf{w}_j\| \tag{62}$$

Therefore using (V-Contraction) and (W_u-Contraction) at iteration t, we have:

$$\begin{aligned}\|\delta\mathbf{v}_j^{t+1}\| &\leq (1-\eta l_{jj}^*)\|\delta\mathbf{v}_j^t\| + \eta\kappa l_{jj}^*\|\delta\mathbf{w}_j^t\| + \eta\epsilon_l M_{l,uu} l_{jj}^* \sum_{j'\neq j, j'\in[u]} \|\delta\mathbf{v}_{j'}^t\| + \eta\epsilon_l M_{l,ur} l_{jj}^* \sum_{j'\neq j, j'\in[r]} \|\mathbf{v}_{j'}^t\| \\ &\leq (1-\eta\bar{l}\gamma)^{t+1} B_{\delta v} \end{aligned} \tag{63}$$

Since γ satisfies Eqn. 44.

Irrelevant nodes. Note that for $j\in[r]$, we don't have the term $\mathbf{q}_{jj}^*$. Therefore, we have:

$$\begin{aligned}\|\mathbf{v}_j^{t+1}\| &\leq (1-\eta l_{jj})\|\mathbf{v}_j^t\| + \eta\epsilon_l M_{l,ru} l_{jj}^* \sum_{j'\neq j, j'\in[u]} \|\delta\mathbf{v}_{j'}^t\| + \eta\epsilon_l M_{l,rr} l_{jj}^* \sum_{j'\neq j, j'\in[r]} \|\mathbf{v}_{j'}^t\| \\ &\leq (1-\eta l_{jj}^*)\|\mathbf{v}_j^t\| + \eta\kappa l_{jj}^*\|\mathbf{v}_j^t\| + \eta\epsilon_l M_{l,ru} l_{jj}^* \sum_{j'\neq j, j'\in[u]} \|\delta\mathbf{v}_{j'}^t\| + \eta\epsilon_l M_{l,rr} l_{jj}^* \sum_{j'\neq j, j'\in[r]} \|\mathbf{v}_{j'}^t\| \\ &\leq (1-\eta\bar{l}\gamma)^{t+1} B_v \end{aligned} \tag{64}$$

□

7.6.2. LEMMA 6

Proof. Similar to the proof of Thm. 4, for node j, in the lower-layer, we have:

$$\dot{\mathbf{w}}_j = \mathbb{I}(j\in[u]) P_{\mathbf{w}_j}^{\perp} \tilde{\mathbf{p}}_{jj}^* + P_{\mathbf{w}_j}^{\perp} \sum_{j'\neq j, j'\in[u]} \Delta\tilde{\mathbf{p}}_{jj'} + P_{\mathbf{w}_j}^{\perp} \sum_{j'\in[r], j'\neq j} \tilde{\mathbf{p}}_{jj'} \tag{65}$$

where $h_{jj'} = d_{jj'}\mathbf{v}_j^T\mathbf{v}_{j'}$ and $\tilde{\mathbf{p}}_{jj'} = \mathbf{p}_{jj'}\mathbf{v}_j^T\mathbf{v}_{j'} = \mathbf{w}_{j'}h_{jj'}$.

Due to (W-Separation) and $\|\mathbf{w}_{j'}\|=1$, we know that for $j\neq j'$:

$$\|\Delta\tilde{\mathbf{p}}_{jj'}\| \leq \epsilon_d M_{d,uu} d_{jj}^* \|\delta\mathbf{w}_{j'}\|\|\mathbf{v}_j\|\|\mathbf{v}_{j'}\|, \;\; \|\tilde{\mathbf{p}}_{jj'}\| \leq \epsilon_d M_{d,ur} d_{jj}^* \|\delta\mathbf{w}_{j'}\|\|\mathbf{v}_j\|\|\mathbf{v}_{j'}\|, \quad j\in[u] \tag{66}$$

$$\|\Delta\tilde{\mathbf{p}}_{jj'}\| \leq \epsilon_d M_{d,ru} d_{jj}^* \|\delta\mathbf{w}_{j'}\|\|\mathbf{v}_j\|\|\mathbf{v}_{j'}\|, \;\; \|\tilde{\mathbf{p}}_{jj'}\| \leq \epsilon_d M_{d,rr} d_{jj}^* \|\delta\mathbf{w}_{j'}\|\|\mathbf{v}_j\|\|\mathbf{v}_{j'}\|, \quad j\in[r] \tag{67}$$

Note that if $\|\mathbf{v}_{j'}\|$(for $j\in[r]$) doesn't converge to zero, then due to Eqn. 67, there is always residue and $\mathbf{w}_j$ won't converge to $\mathbf{w}_j^*$.

Now we discuss two cases:

Relevant nodes. For $j\in[u]$, similar to Eqn. 37 we have:

$$\begin{aligned}\sin\theta_j^{t+1} &\leq (1-\eta d_{jj}^*\|\mathbf{v}_j^t\|^2\cos\theta_j^t)\sin\theta_j^t + \eta\|\mathbf{v}_j^t\|\epsilon_d M_{d,uu} d_{jj}^* \sum_{j'\neq j, j\in[u]} \|\mathbf{v}_{j'}^t\|\sin\theta_{j'}^t \\ &+ \eta\|\mathbf{v}_j^t\|\epsilon_d M_{d,ur} d_{jj}^* \sum_{j'\neq j, j\in[r]} \|\mathbf{v}_{j'}^t\| \end{aligned} \tag{68}$$

Since (W_u-Contraction) and (V-Contraction) holds for time t, we know that:

$$\sin\theta_j^{t+1} \leq (1-\eta\bar{d}\gamma)^{t+1}\sin\theta_0 \tag{69}$$

since Eqn. 43 holds.

Irrelevant nodes. In this case, we cannot prove for $j\in[r]$, $\mathbf{w}_j$ converges to any determined target. Instead, we show that $\mathbf{w}_j$ won't move too much from its initial location $\mathbf{w}_j^0$, which is also set to be $\mathbf{w}_j^*$, before its corresponding $\mathbf{v}_j$ converges to zero. This is important to ensure that (W-Separation) remains correct thorough-out the iterations.

For any $j\in[u]$, using (W_u-Contraction) and (V-Contraction), we know that the distance between the current $\mathbf{w}_j$ and its

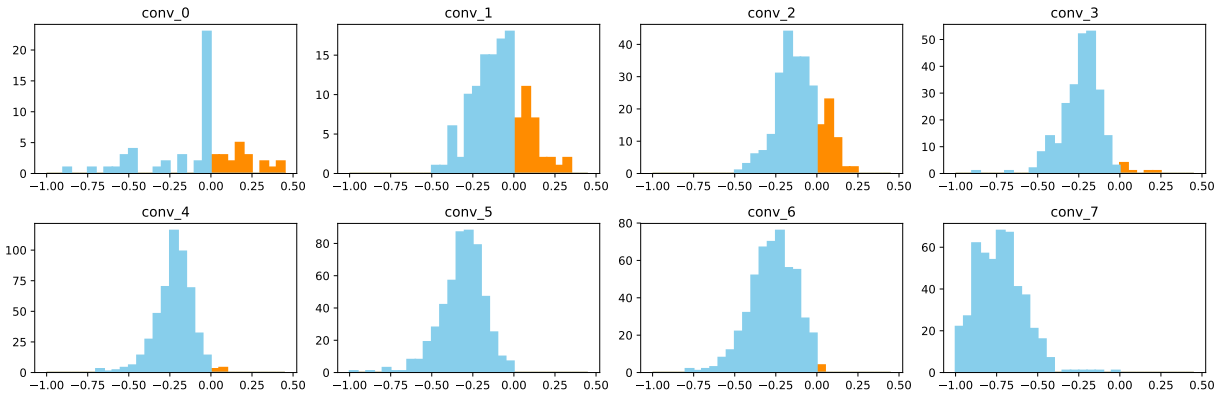

Figure 14. BatchNorm bias distribution of pre-trained VGG11 on ImageNet. Orange/blue are positive/negative biases. The first plot corresponds to the lowest layer (closest to the input).

initial value is

$$\|\mathbf{w}_j^{t+1} - \mathbf{w}_j^0\| \leq \eta \sum_{t'=0}^{t} \|\dot{\mathbf{w}}_j^{t'}\| \leq \eta \sum_{t'=0}^{t} \left\| \sum_{j' \neq j, j' \in [u]} \Delta \tilde{\mathbf{p}}_{jj'}^{t'} + \sum_{j' \in [r], j' \neq j} \tilde{\mathbf{p}}_{jj'}^{t'} \right\| \tag{70}$$

$$\leq \eta \epsilon_d B_{d,u} (B_v + B_{\delta v}) B_v \sum_{t'=0}^{t} (1 - \eta \bar{\lambda} \gamma)^{2t'} \tag{71}$$

$$= \frac{\epsilon_d B_{d,r} (B_v + B_{\delta v}) B_v}{\bar{\lambda}\gamma(2 - \eta\bar{\lambda}\gamma)} = C_{d,r} \tag{72}$$

Therefore, we prove that (W_r-Bound) holds for iteration $t+1$. □

7.7. Lemma 7

Proof. Simply followed from combining Lemma 3, Lemma 2 and weight bounds (W_u-Contraction) and (V-Contraction). □

8. Appendix: More experiments

Besides, we also perform ablation studies on `GAUS`.

Size of teacher network. As shown in Fig. 15(a), for small teacher networks (FC 10-15-20-25), the convergence is much faster and training without BatchNorm is faster than training with BatchNorm. For large teacher networks, BatchNorm definitely increases convergence speed and growth of $\bar{\rho}$.

Finite versus Infinite Dataset. We also repeat the experiments with a pre-generated finite dataset of `GAUS` in the CNN case, and find that the convergence of node similarity stalls after a few iterations. This is because some nodes receive very few data points in their activated regions, which is not a problem for infinite dataset. We suspect that this is probably the reason why `CIFAR-10`, as a finite dataset, does not show similar behavior as `GAUS`.

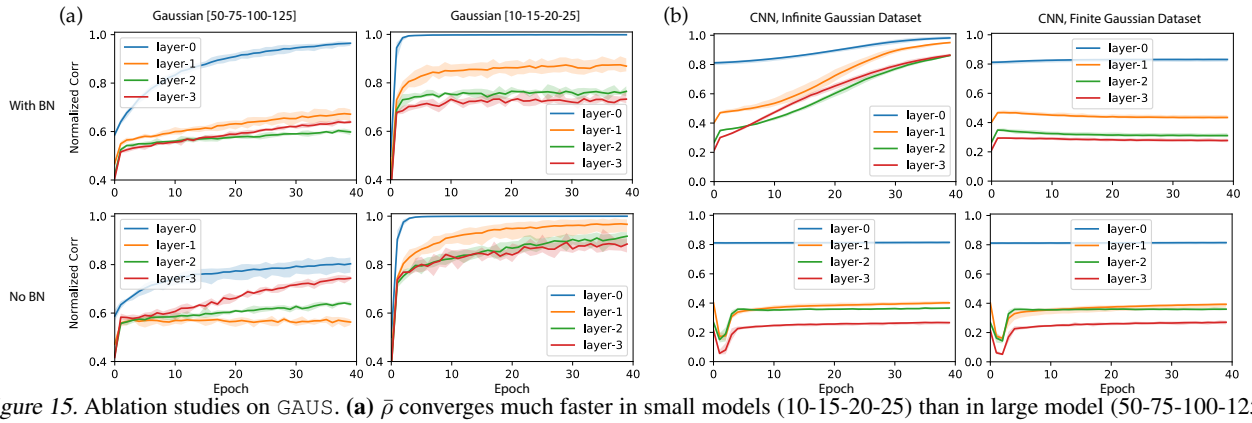

Figure 15. Ablation studies on `GAUS`. **(a)** $\bar{\rho}$ converges much faster in small models (10-15-20-25) than in large model (50-75-100-125). BatchNorm hurts in small models. **(b)** $\bar{\rho}$ stalls using finite samples.

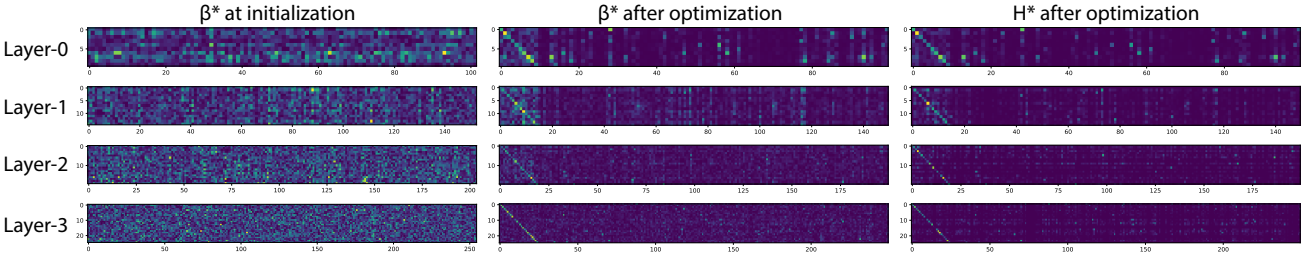

Figure 16. Visualization of (transpose of) H^* and $\boldsymbol{\beta}^*$ matrix before and after optimization (using `GAUS`). Student node indices are reordered according to teacher-student node correlations. After optimization, student node who has high correlation with the teacher node also has high β entries. Such a behavior is more prominent in H^* matrix that combines $\boldsymbol{\beta}^*$ and the activation patterns D^* of student nodes (Sec. 5).

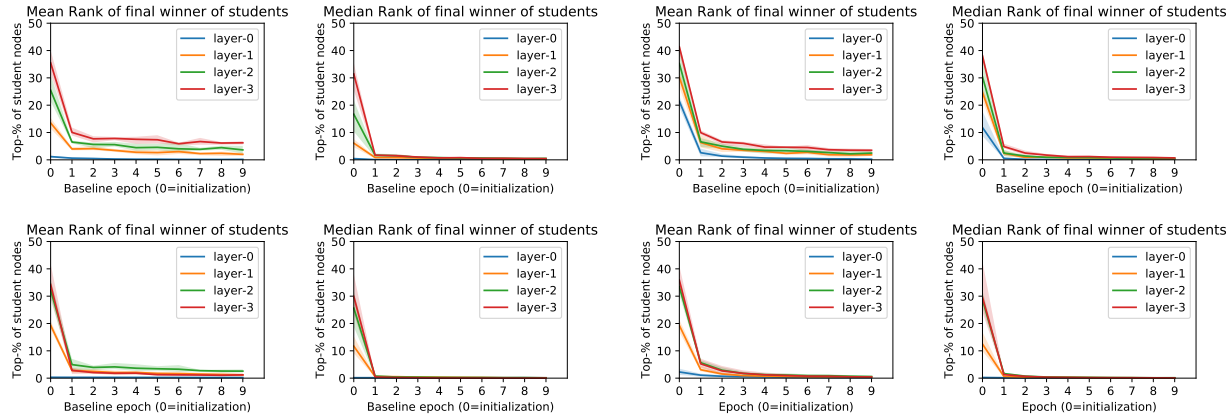

Figure 17. Mean/Median rank at different epoch of the final winning student nodes that best match the teacher nodes after the training using BatchNorm. Gaussian (left) versus CIFAR10 (right). FC (top) versus CNN (bottom).

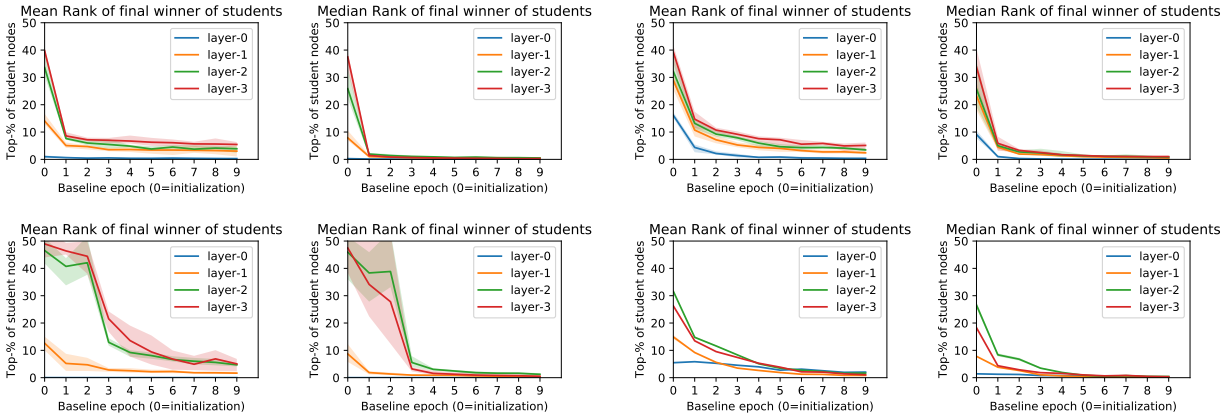

Figure 18. Mean/Median rank at different epoch of the final winning student nodes that best match the teacher nodes after the training *without* BatchNorm. Gaussian (left) versus CIFAR10 (right). FC (top) versus CNN (bottom).

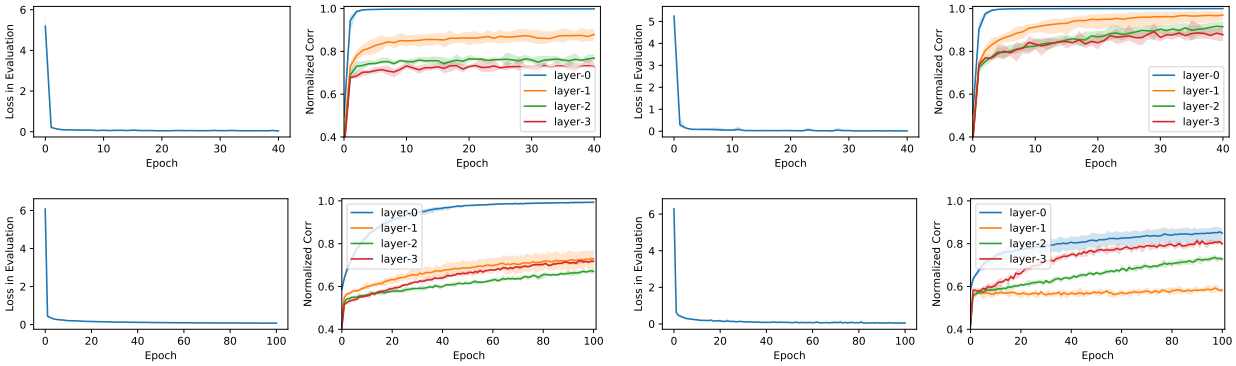

Figure 19. Gaussian data with small (10-15-20-25) and large (50-75-100-125) FC models. Small models (top) versus large models (bottom). With BN (left) versus Without BN (right).

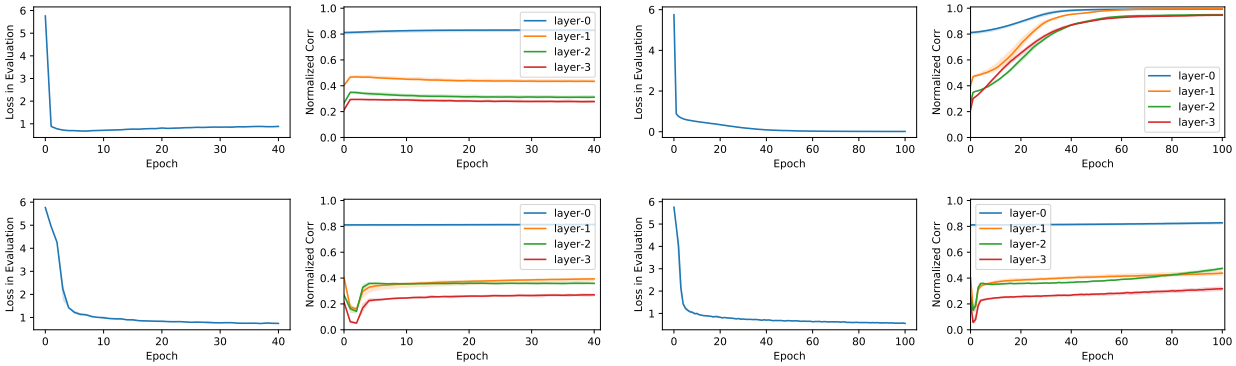

Figure 20. Gaussian CNN. With BN (top) versus Without BN (bottom). Finite Dataset (left) versus Infinite Dataset (right).