Slot Machines: Discovering Winning Combinations of Random Weights in Neural Networks

Maxwell Mbabilla Aladago [1] **Lorenzo Torresani** [1]

A. Distribution of Selected Weights and Scores

As discussed in Section 4.8 in the main paper, we observe that slot machines tend to choose increasingly large magnitude weights as learning proceeds. In Figures 1, 2, and 3 of this appendix, we provide additional plots demonstrating this phenomenon for other architectures. It may be argued that the observed behavior might be due to the Glorot Uniform distribution from which the weights are sampled. Accordingly, we performed ablations for this where we used a Glorot Normal distribution for the weights as opposed to the Glorot Uniform distribution used throughout the paper. As shown in Figure 2a, the initialization distribution do indeed contribute to observed pattern of preference for large magnitude weights. However, initialization may not be the only reason as the models continue to choose large magnitude weights even when the weights are sampled from a Glorot Normal distribution. This is shown more clearly in the third layer of Lenet which has relatively fewer weights compared to the first two layers. We also observed a similar behavior in normally distributed convolutional layers.

Different from the weights, notice that the selected scores are distributed normally as shown in Figure 1. The scores in PS move much further away from the initial values compared to those in GS. This is largely due to the large learning rates used in PS models.

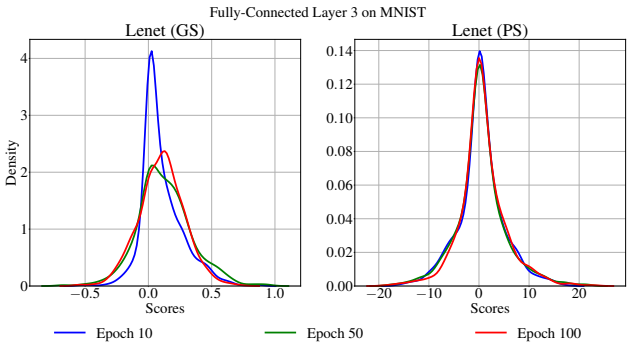
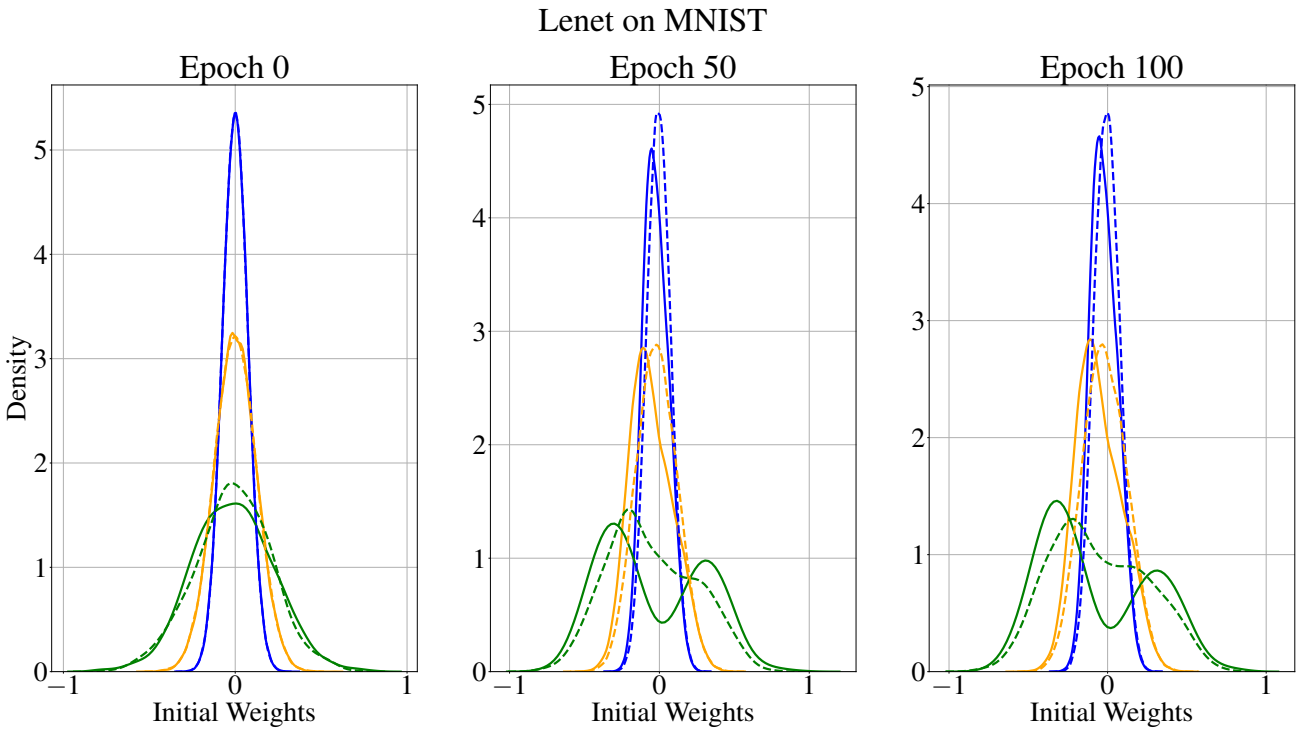


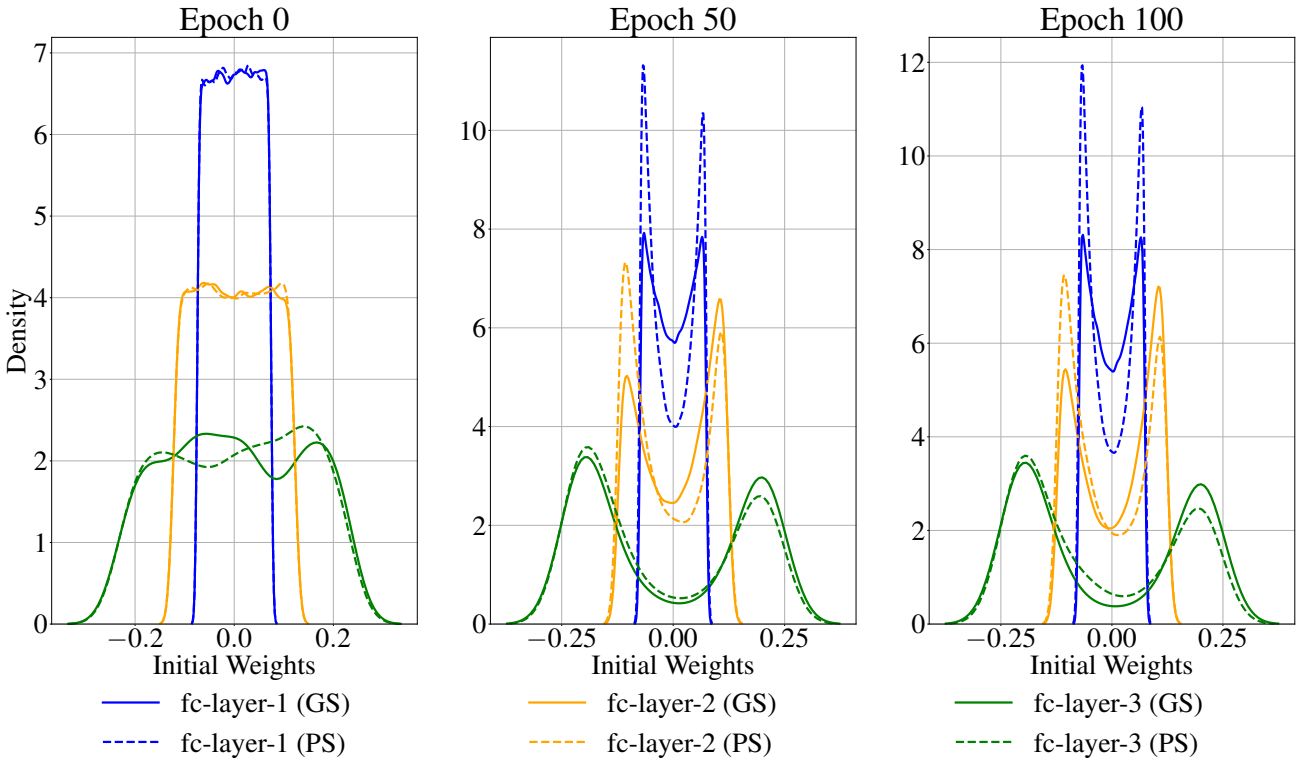
Figure 1. **Distribution of selected scores.** Different from the selected weights, the selected scores tend to be normally distributed for both GS and PS. We show only the scores for layer 3 of Lenet because it is the layer with the fewest number of weights. However, the other layers show a similar trend except that the selected scores in them have very narrow distributions which makes them uninteresting. Notice that although we sample the scores uniformly from the non-negative range $\mathbb{U}(0, 0.1 * \sigma_x)$ where σ_x is the standard deviation of the Glorot Normal distribution, gradient descent is able to drive them into the negative region. The scores in PS slot machines move much farther away from the initialization compared to those in GS due to the large learning rates used in PS models.

[1]Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA. Correspondence to: Maxwell Mbabilla Aladago <maxwell.m.aladago.gr@dartmouth.edu>, Lorenzo Torresani <LT@dartmouth.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.



(a) Glorot Normal Initialization



(b) Glort Uniform Initialization

Figure 2. **Distribution of selected weights on MNIST.** As noted above, both sampling methods tend to choose larger magnitude weights as oppose to small values. This behavior is more evident when the values are sampled from a Glorot Uniform distribution (*bottom*) as opposed to a Glorot Normal distribution (*top*). However, layer 3 which has the fewest number of weights of any layer in this work continue to select large magnitude weights even when using a normal distribution.

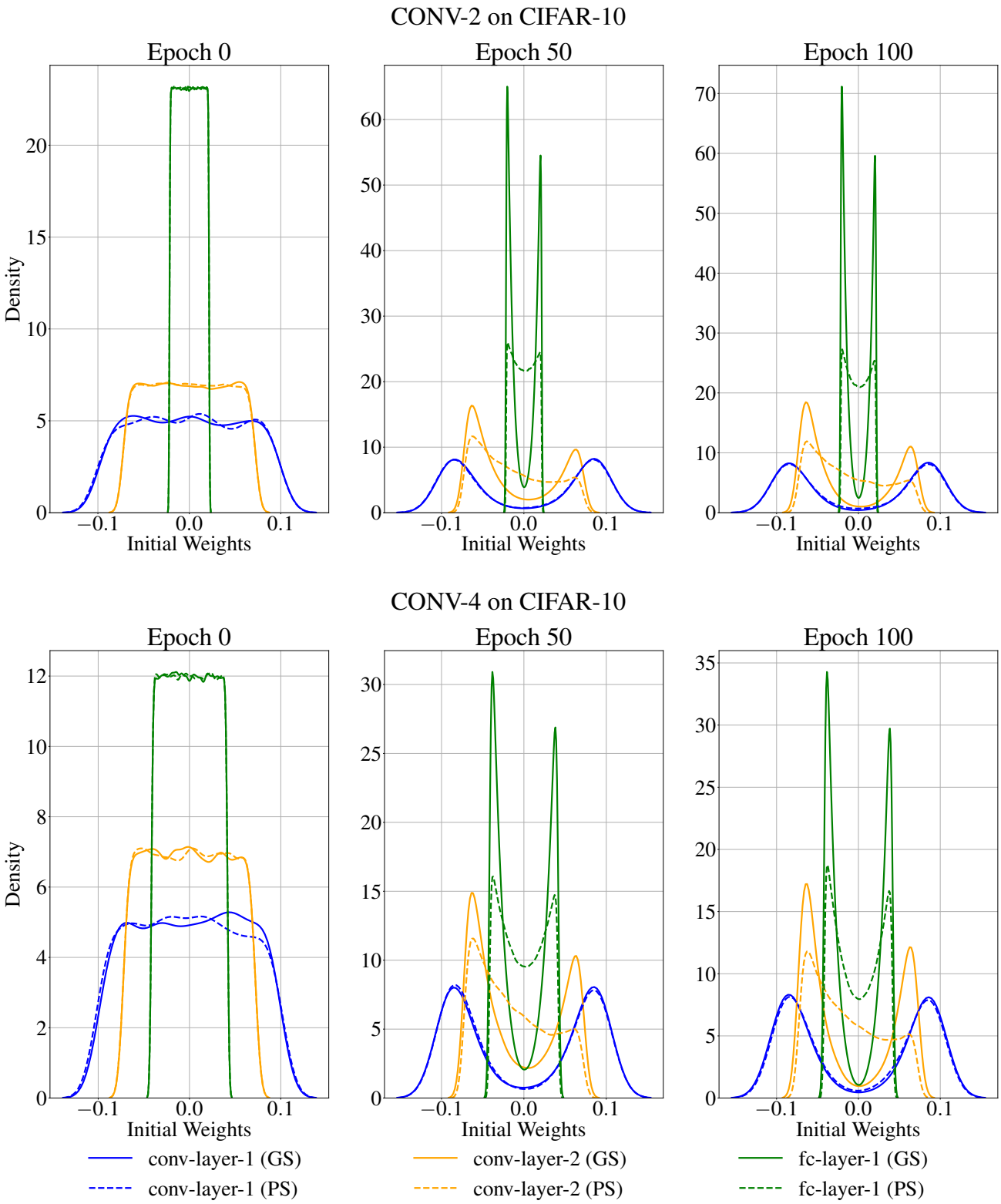


Figure 3. **Distribution of selected weights on CIFAR-10.** Similar to the plots shown in Figure 11 in the paper, both CONV-2 and CONV-4 on CIFAR-10 tend to choose bigger and bigger weights in terms of magnitude as training progresses. Here, we show the distribution of the selected networks in the first two convolutional layers and the first fully-connected layer of the above networks but all the layers in all slot machines show a similar pattern.

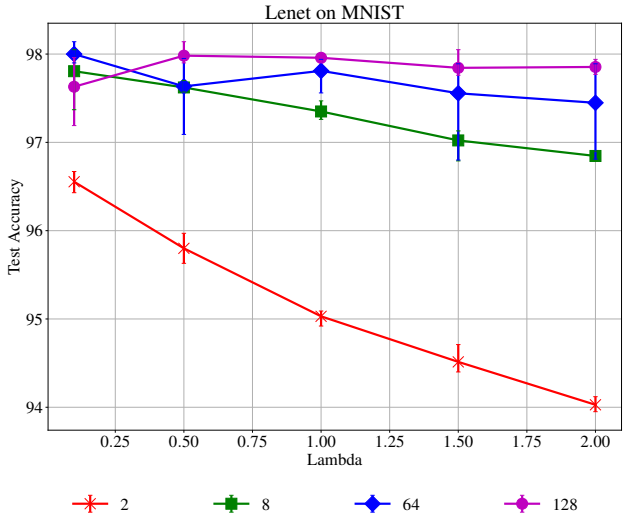


Figure 4. **Scores Initialization.** The models are sensitive to the range of the sampling distribution. As discussed in Section 4.1 of the main paper, the initial scores are sampled from the uniform distribution $\mathbb{U}(\gamma, \gamma + \lambda\sigma_x)$. The value of γ does not affect performance and so we always set it to 0. These plots are averages of 5 different random initializations of Lenet on MNIST.

B. Scores Initialization

We initialize the quality scores by sampling from a uniform distribution $\mathbb{U}(\gamma, \gamma + \lambda\sigma_x)$. As shown in Figure 4, we observe that our networks are sensitive to the range of the uniform distribution the scores are drawn from when trained using GS. However, as expected we found them to be insensitive to the position of the distribution γ. Generally, narrow uniform distributions, e.g., $\mathbb{U}(0, 0.1)$, lead to higher test set accuracy compared to wide distributions e.g., $\mathbb{U}(0, 1)$. This matches intuition since the network requires relatively little effort to drive a very small score across a small range compared to a large range. To concretize this intuition, take for example a weight $\tilde{w}$ that gives the minimum loss for connection (i, j). If its associated score $\tilde{s}$ is initialized poorly to a small value, and the range is small, the network will need little effort to push it to the top to be selected. However, if the range is large, the network will need much more effort to drive $\tilde{s}$ to the top for $\tilde{w}$. We believe that this sensitivity to the distribution range could be compensated by using higher learning rates for wider distributions of scores and vice-versa.