Strong Lottery Ticket Hypothesis with ε-Perturbation

Zheyang Xiong*, Fangshuo Liao*, Anastasios Kyrillidis
Department of Computer Science, Rice University
{zx21, Fangshuo.Liao, anastasios}@rice.edu
*Equal Contribution

Central Question

Strong Lottery Ticket Hypothesis: There exists a subnetwork in a sufficiently over-parameterized, randomly initialized neural network that approximates a target neural network.

Limitation: Strong LTH does not deal with the weight change during the pre-training of LTH.

Idea: Weight change during pre-training = Perturbation around initialization.

Central Question: By allowing an ε-perturbation on the initial weights, can we reduce the over-parameterization for the candidate network in the SLTH? If so, how can we find such a good perturbation?

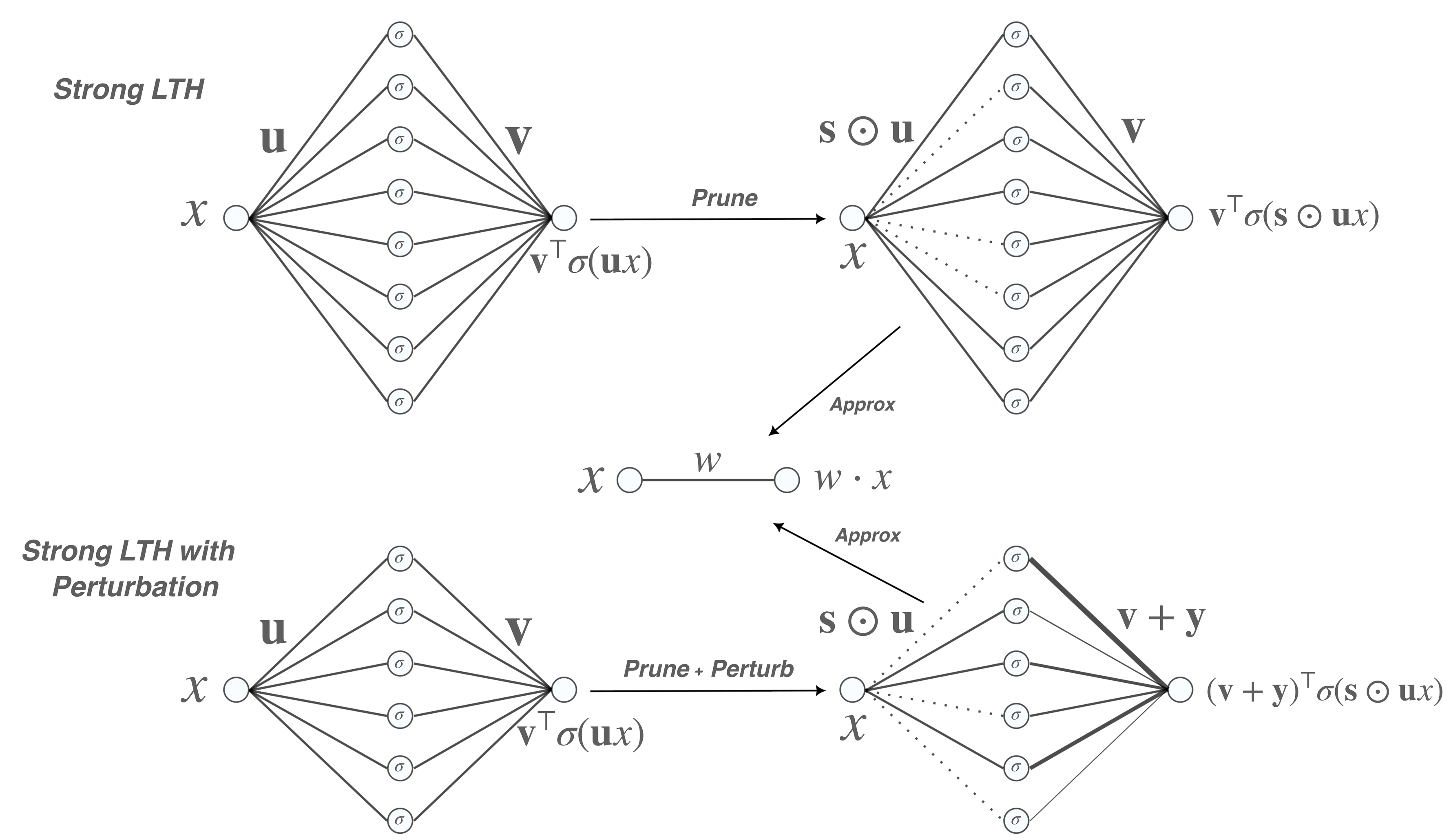
Perturbed Subset Sum Problem

Given a set of random candidates $\{x_i\}_{i=1}^n$ and a target value z, the ε-perturbed subset sum problem considers the following approximation

$$\eta^\star = \min_{\boldsymbol{\delta}\in\{0,1\}^n, \mathbf{y}\in[-\varepsilon,\varepsilon]^n} \left| \sum_{i=1}^n \delta_i (x_i + y_i) - z \right|. \tag{1}$$

Theorem 1. *For all $K \geq 0$, with probability at least $1 - \exp\left(-\frac{(n-K)(1+\varepsilon)^2}{8(3-\varepsilon)}\right) - \exp(-K)$, every $z \in [-1/2, 1/2]$ has an 2η approximation as long as the number of candidates n satisfies*

$$n = O\left(\frac{\log \eta^{-1}}{1+\varepsilon} + K\right).$$



ε-Perturbed Strong LTH

Let $\mathcal{F}$ be a target neural network with depth L, and the width of the ℓth layer is d_ℓ, and let $\mathcal{G}_\mathbf{W}$ be the candidate neural network with depth $2L$. We approximate f using $\mathcal{G}_\mathbf{W}$ by allowing *pruning* and *perturbation* on the weights of $\mathcal{G}$

$$\eta = \min_{\Delta\boldsymbol{W},\mathcal{M}} \sup_{\mathbf{x}} \|\mathcal{F}(\mathbf{x}) - (\mathcal{M}\circ\mathcal{G}_{\boldsymbol{W}+\Delta\boldsymbol{W}})(\mathbf{x})\|. \tag{2}$$

Theorem 2. *For $\mathcal{G}$, if the width of the $(2\ell-1)$th layer is d'_ℓ, the width of the 2ℓth layer is d_ℓ. As long as*

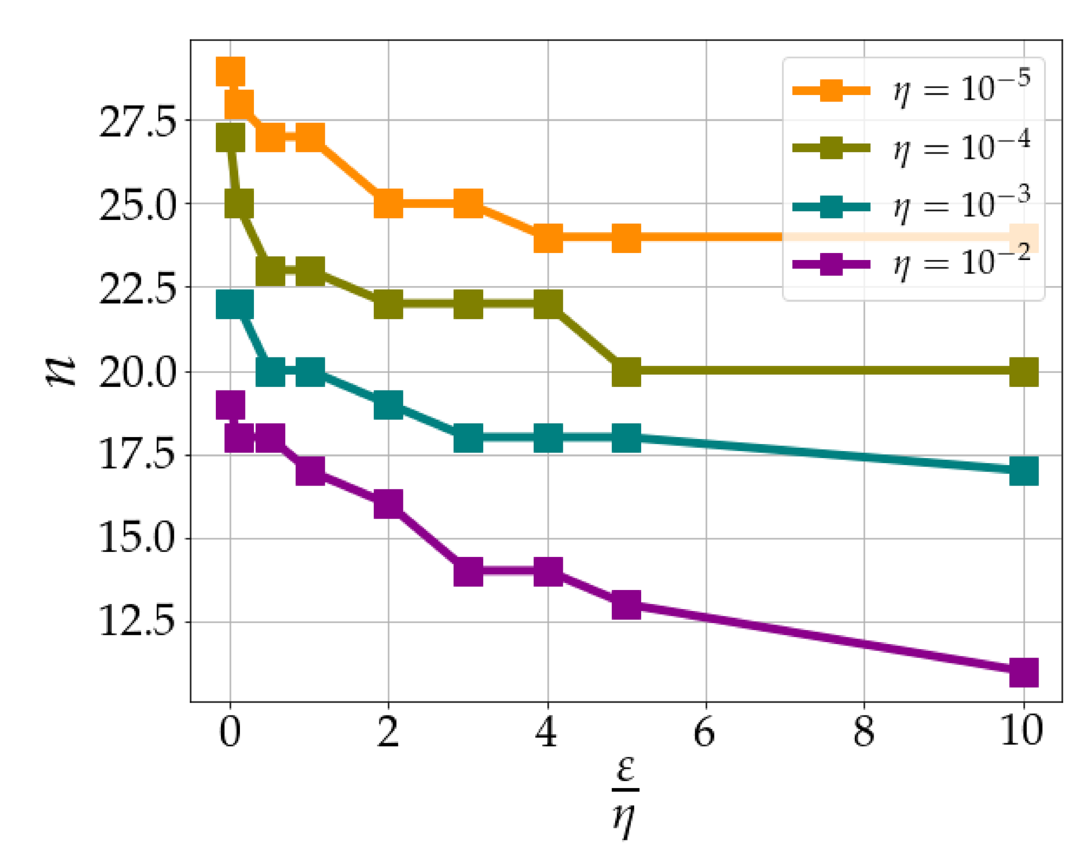
$$d'_\ell = O\left(d_{\ell-1}\frac{\log(\hat\eta^{-1}d_\ell d_{\ell-1}L)}{1+\varepsilon}\right),$$

then with high probability η defined in Equation (2) has $\eta \leq \hat\eta$

Remark: The original SLTH requires $d'_\ell = O\left(d_{\ell-1}\log(\hat\eta^{-1}d_\ell d_{\ell-1}L)\right)$. Compared with the original SLTH, our result is smaller by a factor of $\frac{1}{1+\varepsilon}$. As $\varepsilon\to\infty$, the required width of the candidate network goes to d_ℓ.

PSSP Experiments

With the goal of approximating some target value z, we search for the number n such that 90% of the randomly generated candidate sets with n elements gives an η approximation of z.



PGD+Edge-Popup

Idea: Training the neural network using SGD while bounding the max-norm of the weight change to ε. How does the pruned accuracy vary as we vary ε

Algorithm 1 PGD+StrongLTH

Input: Perturbation scale ε, neural network loss $\mathcal{L}$, initial weight $\mathbf{W}_0$, learning rate $\{\alpha_t\}_{t=0}^{T-1}$

$\Delta\mathbf{W} \leftarrow 0$
for $t \in \{0,\dots,T-1\}$ **do**
  $\hat{\mathbf{W}} \leftarrow \Delta\mathbf{W} - \alpha_t\nabla\mathcal{L}(\mathbf{W}_t)$
  $\Delta\mathbf{W} \leftarrow \texttt{sign}(\hat{\mathbf{W}})\cdot\texttt{min}\{\texttt{abs}(\hat{\mathbf{W}}),\varepsilon\}$
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_0 + \Delta\mathbf{W}$
end for
$\ell^* \leftarrow \infty$, $\mathcal{M}^* \leftarrow$ None
for pruning level $s \in \{0.1, 0.2, \dots, 0.9\}$ **do**
  $\ell, \mathcal{M} \leftarrow \texttt{Edge-Popup}(\mathcal{L}, \mathbf{W}_T, s)$
  if $\ell \leq \ell^*$ **then**
    $\ell^* \leftarrow \ell$, $\mathcal{M}^* \leftarrow \mathcal{M}$
  end if
end for
return Optimal loss ℓ^*, mask $\mathbf{M}^*$ and sparsity level s

SGD Finds a Good Weight Perturbation

Red:
Strong LTH

Blue:
Standard Training with SGD

Orange:
Pruning Dominated by SGD

Sparsity s	Perturbation Scale ε										
	0	10^{-3}	$5\cdot10^{-3}$	10^{-2}	$2\cdot10^{-2}$	$3\cdot10^{-2}$	$4\cdot10^{-2}$	$5\cdot10^{-2}$	10^{-1}	$2\cdot10^{-1}$	$3\cdot10^{-1}$
0	0.12	0.14	0.25	0.42	0.68	0.84	**0.90**	**0.93**	**0.96**	**0.97**	**0.98**
0.1	0.49	0.48	0.65	0.70	0.78	0.82	0.87	0.87	0.94	0.97	0.98
0.2	0.75	0.76	0.77	0.79	0.84	0.86	0.88	0.87	0.93	0.96	0.97
0.3	0.83	0.82	0.82	0.82	0.88	0.88	0.86	0.90	0.92	0.94	0.93
0.4	0.82	0.86	**0.88**	**0.89**	**0.90**	**0.89**	0.90	0.90	0.88	0.91	0.86
0.5	**0.85**	**0.88**	0.86	0.89	0.87	0.88	0.89	0.89	0.90	0.89	0.76
0.6	0.83	0.87	0.87	0.83	0.86	0.88	0.87	0.88	0.87	0.85	0.54
0.7	0.81	0.85	0.84	0.83	0.86	0.82	0.81	0.81	0.79	0.74	0.29
0.8	0.73	0.71	0.71	0.75	0.77	0.75	0.73	0.68	0.77	0.55	0.17

Reference

[1] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. *Optimal Lottery Tickets via SUBSETSUM: Logarithmic over-Parameterization is Sufficient*. Curran Associates Inc., Red Hook, NY, USA, 2020.

[2] George S. Lueker. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures & Algorithms*, 12(1):51–62, 1998.

[3] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[4] Arthur da Cunha, Francesco d'Amore, Frédéric Giroire, Hicham Lesfari, Emanuele Natale, and Laurent Viennot. Revisiting the random subset sum problem, 2022.