Self-Normalized Resets for Plasticity in Continual Learning

Vivek F. Farias
Sloan School of Management
Massachusetts Institute of Technology
vivekf@mit.edu

Adam D. Jozefiak[*]
Operations Research Center
Massachusetts Institute of Technology
jozefiak@mit.edu

Abstract

Plasticity Loss is an increasingly important phenomenon that refers to the empirical observation that as a neural network is continually trained on a sequence of changing tasks, its ability to adapt to a new task diminishes over time. We propose Self-Normalized Resets (SNR), which resets a neuron's weights when evidence indicates its firing rate has collapsed. Across a battery of continual learning problems and network architectures, we demonstrate that SNR consistently attains superior performance compared to its competitor algorithms. We establish the necessity of neuron-resets for mitigating plasticity loss by analyzing the task of learning a single ReLU neuron with gradient descent. Under an adversarial-target regime, an idealized SNR learns the target while regularization-based schemes can fail to learn. SNR's reset-threshold is motivated by a simple hypothesis test for detecting inactive neurons. Seen through the lens of this hypothesis test, competing reset proposals yield suboptimal error rates in correctly detecting inactive neurons.

1 Introduction

Plasticity Loss is an increasingly important phenomenon studied broadly under the rubric of continual learning [8]. This phenomenon refers to the empirical observation that as a neural network is continually trained on a sequence of changing tasks, its ability to adapt to a new task diminishes over time. While this is distinct from the problem of catastrophic forgetting (also studied under the rubric of continual learning [10, 13]), it is of significant practical importance. In the context of pre-training language models, an approach that continually trains models with newly collected data is preferable to training from scratch [11, 24]. On the other hand, the plasticity loss phenomenon demonstrates that such an approach will likely lead to models that are increasingly unable to adapt to new data. Similarly, in the context of reinforcement learning using algorithms like TD, where the learning tasks are inherently non-stationary, the plasticity loss phenomenon results in actor or critic networks that are increasingly unable to adapt to new data [19]. Example B.1 and Figure 1 (in the appendix) illustrate plasticity loss in the 'Permuted MNIST' problem introduced by [10].

One formal definition of plasticity measures the ability of a network initialized at a specific set of parameters to fit a random target function using some pre-specified optimization procedure. In this sense, random parameter initializations (eg. [21]) are known to enjoy high plasticity. This has motivated two related classes of algorithms that attempt to mitigate plasticity loss. The first explicitly 'resets' neuron's that are deemed to have low 'utility' [7, 22]. A reset re-initializes the neurons input weights and bias according to some suitable random initialization rule, and sets the output weights to zero; algorithms vary in how the utility of a neuron is defined and estimated from

[*] Code: `https://github.com/ajozefiak/SelfNormalizedResets`. An earlier version was published in ICLR 2025 [9]. This version extends Theorems 2.1 and 2.2 from $\mathbb{R}$ to $\mathbb{R}^d$ and includes additional experiments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML x OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

Algorithm 1: SNR: Self-Normalized Resets

```
Input: Reset percentile threshold $\eta$
Initialize: Initialize weights $\theta_0$ randomly. Set inter-firing time $a_i = 0$ for each neuron $i$
for each training example $x_t$ do
    Forward Pass: Evaluate $f(x_t; \theta_t)$. Get neuron activations $z_{t,i}$ for each neuron $i$
    Update inter-firing times: For each neuron $i$, $a_i \leftarrow a_i + 1$ if $z_{t,i} = 0$. Otherwise, $a_i \leftarrow 0$
    Optimize: $\theta_{t+1} \leftarrow O_t(H_t)$
    Resets: For each neuron $i$, reset if $\mathbb{P}(A_i^{\mu_t} \geq a_i) \leq \eta$.
```

online data. A second class of algorithms perform this reset procedure implicitly via regularization [1, 15]. These latter algorithms differ in their choice of what to regularize towards, with choices including the original network initialization; a new randomly drawn initialization; or even zero. The aforementioned approaches to mitigating plasticity loss attempt to adjust the training process; other research has studied the role of architectural and optimizer hyperparameter choices. Across all of the approaches to mitigating plasticity loss described above, no single approach is yet to emerge as both robust to hyperparameter choices, and simultaneously performant across benchmark problems. To address this gap, we propose *Self-Normalized Resets* (SNR) and demonstrate its efficacy theoretically and empirically.

To make ideas precise, consider a sequence of training examples $(X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$, drawn from some distribution μ_t. Denote the network by $f : \mathcal{X} \times \Theta \to \mathcal{Y}$, and let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be our loss function. Denote by $H_t \in \mathcal{H}_t$, the history of network weights and training examples up to time t, and assume access to an optimization oracle $O_t : \mathcal{H}_{t-1} \to \Theta$ that maps the history of weights and training examples to a new set of network weights. As a concrete example, O_t might correspond to stochastic gradient descent.

Let θ_t^* minimize $\mathbb{E}_{\mu_t}[l(f(X_t; \theta), Y_t)]$, denote $\Theta_t = O_t(H_{t-1})$, and consider average expected regret

$$\frac{1}{T} \sum_t \mathbb{E}_{\mu_t}[l(f(X_t; \Theta_t), Y_t)] - \mathbb{E}_{\mu_t}[l(f(X_t; \theta_t^*), Y_t)]$$

Plasticity loss describes the phenomenon where, for certain continual learning processes Θ_t, such as those corresponding to SGD or Adam, average expected regret increases over time, even for benign choices of μ_t.[2]

To motivate our algorithm, SNR, consider applying the network $f(\cdot; \theta_t^*)$ to a hypothetical sequence of training examples drawn i.i.d. from μ_t indexed by s. Let $Z_{s,i}^{\mu_t}$ indicate the sequence of activations of neuron i, and let $A_i^{\mu_t}$ be a random variable distributed as the random time between any two consecutive activations over this hypothetical sequence of examples. Now turning to the *actual* sequence of training examples, let a_i^t count the time since the last firing of neuron i prior to time t. Our (idealized) proposal is then exceedingly simple: reset neuron i at time t iff $\mathbb{P}(A_i^{\mu_t} \geq a_i^t) \leq \eta$ for some suitably small threshold η. We dub this algorithm *Self-Normalized Resets* and present it as Algorithm 1. The algorithm requires a single hyper parameter, η. In practice, the distribution of $A_i^{\mu_t}$ is unknown, and so an implementable version of Algorithm 1 simply approximates this distribution by the histogram of inter-firing times of neuron i prior to time t (over a fixed length trailing window).

In Section 2 we justify, theoretically, (1) the necessity of neuron-resets for mitigating plasticity loss and (2) the optimality of SNR relative to other reset methods. In Section 3 we demonstrate the efficacy of SNR relative to competitor algorithms across benchmark problems and network architectures.

2 Theoretical Analysis

Several correlates of plasticity loss have been identified such as neuron inactivity, feature/weight–rank collapse, increasing weight norms, and loss of curvature in the loss surface [6, 20, 22, 17, 14]. However, the literature remains largly empirical, offering little theoretical insight into *why* plasticity fades or *how* best to restore it. To address this gap, we analyze the simplest non-trivial continual learning model: gradient descent for learning a single ReLU neuron, minimizing the loss in (1).

[2] Common in the literature: divide T into Δ-length intervals and set $\mu_t = \mu_i$ on the i^{th} interval.

$$L(w) = \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2], \tag{1}$$

where $\sigma(z) = \max(0, z)$ is the ReLU activation function, v is the target neuron's weight vector, and the expectation is taken over $x \in \mathbb{R}^{d+1}$, sampled from $\mathcal{N}(0, I_d) \times \{1\}$, where I_d is the d-dimensional identity matrix. We define the *average regret* in learning the target neuron's weights as

$$R_T = \frac{1}{T} \sum_{t=0}^{T-1} (L(w_t) - L(w^*)) = \frac{1}{T} \sum_{t=0}^{T-1} L(w_t), \tag{2}$$

Existing work assumes fixed v, then samples w_0, and then runs gradient descent; we call this the *non-adversarial target regime*. In this regime, [23] prove that, under mild assumptions, gradient descent converges linearly to the unique minimizer $w^* = v$, yielding $O(1/T)$ expected average regret. We instead introduce the *adversarial target regime*: sample w_0; after observing w_0, an adversary chooses v; then run gradient descent. This abstraction captures continual learning, with the adversary's choice of v modeling task transitions.

In this adversarial-target regime, we show that gradient descent with L2 regularization incurs $\Omega(1)$ expected average regret (Theorem 2.1), formalizing neuron-level plasticity loss: weights learned for earlier tasks induce poor initializations for later ones. In contrast, granting gradient descent access to a reset oracle yields $O(d^2 \ln T/T)$ expected average regret (Theorem 2.2), demonstrating that neuron resets are an effective mechanism for identifying poor initializations and thus remedying plasticity loss. The reset oracle is $\mathcal{R}_\delta(w) = \mathbb{1}\{\mathbb{E}[\sigma(w^\top x)] \leq \delta\}$; we reset w_t iff $\mathcal{R}_\delta(w_t) = 1$.

Theorem 2.1. Let $\hat{L}(w) = L(w) + \frac{\lambda}{2}\|w\|^2$ (see Eq. (1)), and let $w_0 \in \mathbb{R}^{d+1}$ be an initialization satisfying $(w_0)_{d+1} \leq 0$. Define the target weight vector $v \in \mathbb{R}^{d+1}$ so that $v_{1:d} = -(w_0)_{1:d}$ and $v_{d+1} < -(w_0)_{d+1}$. Then, for any step size $\alpha < \frac{1}{2(d+1)}$ and any regularization parameter $\lambda < 2d$, performing gradient descent on the L2-regularized objective $\hat{L}(\cdot)$ from w_0 satisfies

$$\forall T \geq 0, \quad R_T \;\geq\; L(0) \;>\; 0.$$

Theorem 2.2. Let w_0 be sampled from $\mathcal{D}_{w_0} = \text{Unif}(l \cdot S^{d-1} \times \{0\})$ for some positive constant $l > 0$, where $l \cdot S^{d-1} \subseteq \mathbb{R}^d$ is the l-radius sphere. Let $v \in \mathbb{R}^{d+1}$ (possibly chosen adversarially with knowledge of w_0) be the target neuron's weights satisfying

$$-v_{d+1} \leq (2\sqrt{2\pi})^{-1}\|v_{1:d}\|, \quad \|v_{1:d}\| = 2\pi l, \quad \|v\| \geq C_1, \quad \|v_{1:d}\| \geq C_2\|v\|,$$

for some constants $C_1, C_2 > 0$, and suppose $\delta < \left(\frac{C_2^2}{8\pi^2\left(\sqrt{3} + \frac{2}{C_1}\right)} \right)^3$. Then, minimizing $L(\cdot)$ via gradient descent with step size $\alpha \leq \frac{1}{2(d+1)}$, initialized at w_0, and employing a reset oracle $\mathcal{R}_\delta$ yields

$$\forall T \geq 0, \quad \mathbb{E}[R_T] \leq \mathcal{O}\Big(\frac{d^2 \ln T}{T}\Big),$$

where the expectation is taken over the randomness of the the initialization of w_0 and the resets.

The assumptions of Theorem 2.2, including the initialization distribution of w_0, the bound on v_{d+1}, and the constraints on $\|v\|$ and $\|v_{1:d}\|$, are identical, up to constants, to those in the analysis of [23], the state-of-the-art work on learning a single ReLU-activated neuron in the non-adversarial setting.

We motivate SNR, in relation to other reset-schemes, by casting dead-neuron detection as an optimal hypothesis test (details in the appendix). Proposition D.1 shows the optimal algorithm uses a reset threshold proportional to a neuron's nominal firing rate—exactly SNR. Proposition D.2 shows fixed-threshold or fixed-frequency resets (e.g., Continual Backprop, ReDO [8, 22]) can be arbitrarily worse than SNR when neurons have different nominal firing rates.

3 Experiments

We evaluate the efficacy and robustness of SNR on a series of benchmark problems from the continual learning literature, measuring regret with respect to prediction accuracy $l(y, y') = \mathbb{1}\{y \neq y'\}$.

Extensive details of experimental setups and network architectures are in the appendix. We evaluate on the following problems: Permuted MNIST (PM) [10, 6, 15], Random Label MNIST (RM) [15, 20], Random Label CIFAR (RC) [15, 20], Continual Imagenet (CI) [7, 15], and Permuted Shakespeare (PS) (introduced by us). Our experimental setup, for all but PS, follows that of [15], with some minor exceptions that we outline in the appendix.

Our baseline in all problems consist simply of using SGD or Adam as the optimizer with no further intervention. We consider the following neuron-reset algorithms: SNR, Continual Backprop (CBP) [6], and ReDO [22]. We consider the following regualrization-based algorithms: L2 regularization, L2 Init [15], and Shrink and Perturb (S&P) [1]. As an architectural modification we consider Layer Normalization [3].

We utilize the following network architectures: MLP identical to that in [15] (for PM and RM), CNN identical to that in [15] (for RC and CI), transformer decoder (for PS), and ViT identical to that in [16] and [18] (for CI, which we call CI-ViT to distinguish from the CNN experiments).

We report results in Tables 1 and 2. For the transformer-based PS and CI-ViT models, we also evaluate SNR+L2 and SNR+L2*: SNR with L2-regularization on all weights or only on the attention K/Q/V matrices, respectively, since self-attention lacks explicit neurons.

Optimizer	**SGD**				**Adam**			
Algorithm	**PM**	**RM**	**RC**	**CI**	**PM**	**RM**	**RC**	**CI**
No Intv.	0.71	0.11	0.18	0.78	0.64	0.11	0.15	0.58
SNR	**0.85**	**0.97**	**0.99**	**0.89**	**0.88**	**0.98**	**0.98**	**0.85**
CBP	0.84	0.95	0.96	0.84	**0.88**	0.95	0.33	0.82
ReDO	0.83	0.72	0.98	0.87	0.85	0.67	0.74	0.80
L2 Reg.	0.82	0.80	0.95	0.83	**0.88**	0.95	0.97	0.80
L2 Init	0.83	0.91	0.97	0.83	**0.88**	0.96	**0.98**	0.83
S&P	0.83	0.92	0.97	0.85	**0.88**	0.96	0.97	0.81
Layer Norm.	0.69	0.14	0.96	0.82	0.66	0.11	0.96	0.58

Table 1: Average accuracy on the last 10% of tasks on the benchmark continual learning problems over 5 seeds. Standard deviations are provided in the extended Table 5.

	PS			**CI-ViT**
Algorithm	**All Tasks**	**First 50 Tasks**	**Last 50 Tasks**	**Train Acc. (Std.)**
L2	0.2762 (0.0309)	0.1560 (0.0236)	0.3101 (0.0425)	0.8140 (0.0081)
SNR+L2	0.2177 (0.0196)	0.1370 (0.0169)	0.2551 (0.0454)	0.8733 (0.0071)
No Intv.	2.7397 (0.0140)	1.8164 (0.0295)	3.0147 (0.0250)	0.6432 (0.0140)
L2 Init	1.5052 (0.0437)	1.2931 (0.0486)	1.5262 (0.0420)	0.9196 (0.0039)
SNR	2.6872 (0.0222)	1.7338 (0.0295)	3.0242 (0.0408)	0.6555 (0.0242)
CBP	2.4922 (0.0171)	1.3732 (0.0234)	2.9410 (0.0188)	0.7037 (0.0314)
L2*	0.1506 (0.0279)	0.1874 (0.0348)	0.1092 (0.0429)	N/A
SNR+L2*	**0.1402** (0.0246)	**0.1549** (0.0107)	**0.0909** (0.0177)	**0.9337** (0.0092)
ReDO	2.3258 (0.0356)	1.3689 (0.0686)	2.8119 (0.0373)	0.6380 (0.0189)
S&P	N/A	N/A	N/A	0.5459 (0.0007)

Table 2: Combined results. **PS**: average loss on the final epoch of each task with standard deviations over 9 seeds (Permuted Shakespeare). **CI-ViT**: average training accuracy on the last epoch of the final 50 tasks with standard deviations over 5 seeds. L2* denotes L2 regularization applied only to attention weights.

4 Conclusion

We introduce SNR, a reset-based method for mitigating plasticity loss in continual learning. Theoretically, we show (1) for a single ReLU-activated neuron resets mitigate plasticity loss, and, (2) SNR is optimal relative to fixed-threshold/frequency reset-schemes. Empirically, SNR consistently yields higher online accuracy and lower loss across benchmark problems and architectures. A limitation is layers without explicit neurons (e.g., self-attention), for which we pair SNR with L2 regularization.

References

[1] Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894, 2020.

[2] Dylan R Ashley, Sina Ghiassian, and Richard S Sutton. Does the adam optimizer exacerbate catastrophic forgetting? *arXiv preprint arXiv:2102.07686*, 2021.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

[5] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[6] Shibhansh Dohare, Richard S Sutton, and A Rupam Mahmood. Continual backprop: Stochastic gradient descent with persistent randomness. *arXiv preprint arXiv:2108.06325*, 2021.

[7] Shibhansh Dohare, J Fernando Hernandez-Garcia, Parash Rahman, Richard S Sutton, and A Rupam Mahmood. Maintaining plasticity in deep continual learning. *arXiv preprint arXiv:2306.13812*, 2023.

[8] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

[9] Vivek Farias and Adam Daniel Jozefiak. Self-normalized resets for plasticity in continual learning. In *The Thirteenth International Conference on Learning Representations*.

[10] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[11] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

[12] Seonho Kim and Kiryung Lee. Max-affine regression via first-order methods. *SIAM Journal on Mathematics of Data Science*, 6(2):534–552, 2024.

[13] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[14] Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. *arXiv preprint arXiv:2010.14498*, 2020.

[15] Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity via regenerative regularization. *arXiv preprint arXiv:2308.11958*, 2023.

[16] Hojoon Lee, Hyeonseo Cho, Hyunseung Kim, Donghu Kim, Dugki Min, Jaegul Choo, and Clare Lyle. Slow and steady wins the race: Maintaining plasticity with hare and tortoise networks. In *Forty-first International Conference on Machine Learning*, 2024.

[17] Alex Lewandowski, Haruto Tanaka, Dale Schuurmans, and Marlos C Machado. Curvature explains loss of plasticity. *arXiv preprint arXiv:2312.00246*, 2023.

[18] Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C Machado. Learning continually by spectral regularization. *arXiv preprint arXiv:2406.06811*, 2024.

[19] Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. *arXiv preprint arXiv:2204.09560*, 2022.

[20] Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, pages 23190–23211. PMLR, 2023.

[21] Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. *arXiv preprint arXiv:2402.18762*, 2024.

[22] Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pages 32145–32168. PMLR, 2023.

[23] Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, 34:28690–28700, 2021.

[24] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

[25] Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR, 2020.

A Experimental Setup

Each problem consists of tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N$, each of which contains training examples in $\mathcal{X} \times \mathcal{Y}$. A network is trained for a fixed number of epochs per task to minimize cross-entropy loss. We perform an initial hyperparameter sweep over 5 seeds to determine the optimal choice of hyperparameters (see Appendix A.2). For each algorithm and problem, we select the hyperparameters that attain the lowest average loss and repeat the experiment on 5 new random seeds. A random seed determines the network's parameter initialization, the generation of tasks, and any randomness in the algorithms evaluated. We evaluate both SGD and Adam as the base optimization algorithm, as earlier literature has argued that Adam can be less performant than SGD in some continual learning settings [7, 2]. We evaluate on the following problems:

Permuted MNIST (PM) [10, 6, 15]: A subset of 10000 image-label pairs from the MNIST dataset are sampled for an experiment. A task consists of a random permutation applied to each of the 10000 images. The network is presented with 2400 tasks appearing in consecutive order. Each task consists of a single epoch and the network receives data in batches of size 16.
Random Label MNIST (RM) [15, 20]: A subset of 1200 images from the MNIST dataset are sampled for an experiment. An experiment consists of 100 tasks, where each tasks is a random assignment of labels, consisting of 10 classes, to the 1200 images. A network is trained for 400 epochs on each task with a batch size 16.
Random Label CIFAR (RC) [15, 20]: A subset of 128 images from the CIFAR-10 dataset are sampled for an experiment. An experiment consists of 50 tasks, where each tasks is a random assignment of labels, consisting of 10 classes, to the 128 images. An agent is trained for 400 epochs on each task with a batch size 16.
Continual Imagenet (CI) [7, 15]: An experiment consists of all 1000 classes of images from the ImageNet-32 dataset [5] containing 600 images from each class. Each task is a binary classification problem between two of the 1000 classes, selected at random. The experiment consists of 500 tasks and each class occurs in exactly one task. Each task consists of 1200 images, 600 from each class, and the network is trained for 10 epochs with a batch size of 100.
Permuted Shakespeare (PS): We propose this problem to facilitate studying the transformer architecture in analogy to the MNIST experiments. An experiment consists of 32768 tokens of text from Shakespeare's Tempest. For any task, we take a random permutation of the vocabulary of the Tempest and apply it to the text. The network is presented with 500 tasks. Each task consists of 100 epochs and the network receives data in batches of size 8 with a context widow of width 128. We evaluate over 9 seeds.

This experimental setup, for all but Permuted Shakespeare, follows that of [15], with the exceptions of Permuted MNIST which has its task count increased from 500 to 2400, Random Label MNIST which has its task count increased from 50 to 100, and Random Label CIFAR which has its dataset reduced from 1200 to 128 images.

A.1 Algorithms and Architectures

Our baseline in all problems consist simply of using SGD or Adam as the optimizer with no further intervention. We then consider several interventions to mitigate plasticity loss. First, we consider algorithms that employ an explicit reset of neurons: these include SNR, Continual Backprop (CBP) [6], and ReDO [22]. Among algorithms that attempt to use regularization, we consider vanilla

L2 regularization, L2 Init [15], and Shrink and Perturb [1]. Finally, as a potential architectural modification we consider the use of Layer Normalization [3].

We utilize the following network architectures:
MLP: For Permuted MNIST and Random Label MNIST we use an MLP identical to that in Kumar et al. [15] which in turn is a slight modification to that in Dohare et al. [7].
CNN: For Random Label CIFAR and Continual ImageNet we use a CNN architectures identical to that in Kumar et al. [15] which in turn is a slight modification to that in Dohare et al. [7].
Transformer: We use a decoder model with a single layer consisting of 2 heads, dimension 16 for each head, and with 256 neurons in the feed forward layer with ReLU activations. We deploy this architecture on the Permuted Shakespeare problem using the GPT-2 BPE tokenizer (limited to the set of unique tokens present in the sampled text).
ViT: In addition, we benchmark a Vision Transformer (ViT) on the Continual ImageNet problem. Our ViT follows the specifications of Lee et al. [16] and Lewandowski et al. [18], using 4×4 input patches, three attention heads, and 12 transformer layers with an embedding dimension $d_{\text{model}} = 192$. This architecture comprises approximately 5.3 million non-embedding parameters. We refer to this variant as CI-ViT to distinguish it from the CNN-based experiments.

A.2 Hyperparameter Sweep

With SGD we train with learning rate 10^{-2} on all problems except Random Label MNIST, for which we train with learning rate 10^{-1}. With Adam we train with learning rate 10^{-3} on all problems, including Permuted Shakespeare and we use the standard parameters of $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-7}$. For Permuted Shakespeare we train our networks solely with Adam. The learning rates were selected after an initial hyperparameter sweep.

For each algorithm we vary its hyperparameter(s) by an appropriate constant over 7 choices, effectively varying the hyperparameters over a log scale. With the exception of the Permuted Shakespeare experiment, we limit over hyperparameter search to 5 choices. In Table 3 we provide the hyperparameter sweep for the 4 benchmark problems. CBP's replacement rate r is to be interpreted as one replacement per layer every r^{-1} training examples, as presented in Dohare et al. [8]. ReDO's reset frequency r determines the frequency of resets in units of tasks, as implemented and evaluated in Kumar et al. [15].

	Hyperparameter Strength						
Algorithm	0	1	2	3	4	5	6
L2 Reg. (λ)	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
L2 Init (λ)	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
S&P (1-p)	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
S&P (σ)	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
CBP (r)	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
ReDO (τ)	0	0.01	0.02	0.04	0.08	0.16	0.32
ReDO (r)	64	32	16	8	4	2	1
SNR (η)	0.08	0.04	0.02	0.01	0.005	0.0025	0.00125

Table 3: Hyperparameter sweep for the Permuted MNIST (PM), Random Label MNIST (RM), Random Label CIFAR (RC), and Continual ImageNet (CI) problems.

	Hyperparameter Strength				
Algorithm	0	1	2	3	4
L2 Reg. (λ)	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
L2 Init (λ)	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
SNR (η)	0.1	0.05	0.03	0.01	0.001
SNR + L2 Reg (η)	0.1	0.05	0.03	0.01	0.001

Table 4: Hyperparameter sweep for the Permuted Shakespeare problem. For the combination of SNR and L2 regularization we use the regularization strength of 10^{-4}, the best performing regularization strength for L2 regularization, and vary the rejection percentile threshold η

.

B Additional Experimental Results

Example B.1 (The Permuted MNIST problem)**.** Consider a sequence of 'tasks' presented sequentially to SGD, wherein each task consists of 10000 images from the MNIST dataset with the pixels permuted. SGD trains over a single epoch on each task before the subsequent task is presented. Figure 1 measures average accuracy on each task; we see that average accuracy decreases over tasks. The figure also shows a potential correlate of this phenomenon: the number of 'dead' or inactive neurons[3] in the network increases as training proceeds, diminishing the network's effective capacity.

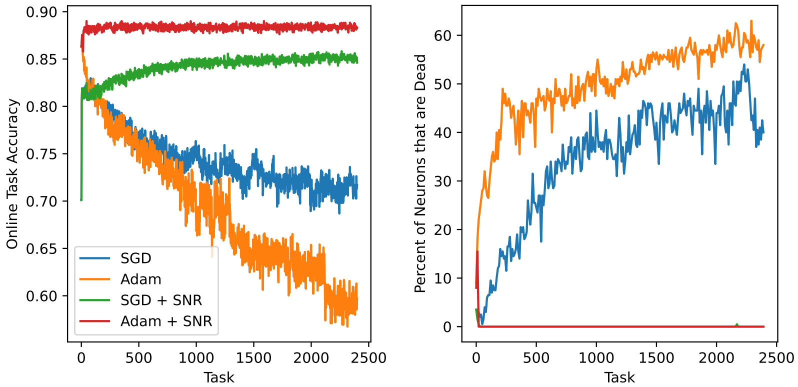

Figure 1: Illustration of plasticity loss and its mitigation by SNR during training of a multilayer perceptron on the Permuted MNIST problem for a single random seed. For this figure, a neuron is declared dead if it has not fired for the last 1000 consecutive training examples.

B.1 Complete Results for Benchmark Problems

To maintain brevity in the main body of the paper, we include here the complete results for the benchmark problems PM, RC, RM, and CI which include both the mean and standard deviation of terminal task accuracies over random seeds.

Optimizer			**SGD**	
Algorithm	**PM**	**RM**	**RC**	**CI**
No Intv.	0.710 (0.007)	0.113(0.004)	0.180 (0.011)	0.784 (0.019)
SNR	**0.851**(0.002)	**0.975** (0.001)	**0.987** (0.002)	**0.888** (0.010)
CBP	0.844(0.002)	0.951 (0.007)	0.961 (0.011)	0.840 (0.015)
ReDO	0.831 (0.013)	0.716 (0.024)	0.981 (0.003)	0.869 (0.038)
L2 Reg.	0.818 (0.001)	0.803 (0.011)	0.952 (0.006)	0.833 (0.011)
L2 Init	0.829 (0.001)	0.913 (0.001)	0.966 (0.002)	0.832 (0.010)
S&P	0.826 (0.002)	0.920 (0.009)	0.971 (0.004)	0.853 (0.006)
Layer Norm.	0.687 (0.009)	0.143 (0.015)	0.959 (0.005)	0.819 (0.009)
Optimizer			**Adam**	
Algorithm	**PM**	**RM**	**RC**	**CI**
No Intv.	0.641 (0.007)	0.114 (0.005)	0.151 (0.005)	0.581 (0.081)
SNR	**0.889**(0.001)	**0.982**(0.001)	**0.976** (0.002)	**0.847** (0.005)
CBP	0.876 (0.001)	0.948 (0.003)	0.331 (0.312)	0.818 (0.005)
ReDO	0.846 (0.002)	0.671 (0.021)	0.744 (0.131)	0.803 (0.063)
L2 Reg.	0.876(0.002)	0.948 (0.002)	0.967 (0.011)	0.803 (0.009)
L2 Init	0.883 (0.002)	0.961 (0.003)	**0.976** (0.002)	0.827 (0.008)
S&P	0.876 (0.002)	0.955 (0.006)	0.971 (0.005)	0.814 (0.005)
Layer Norm.	0.662 (0.001)	0.113 (0.005)	0.955 (0.005)	0.651 (0.053)

Table 5: Average accuracy on the last 10% of tasks on the benchmark continual learning problems with standard deviations over 5 seeds.

[3] this notion is formalized in Section D

C Learning a Single ReLU-Activated Neuron Under Adversarial Selection

We provide a two-fold theoretical analysis. We begin in Section C.1 by analyzing the problem of learning a single ReLU-activated neuron with gradient descent under an adversarial selection of the target weights. This is the simplest possible (component of a) neural network and it captures the essence of continual learning, whereby the adversarial selection of the target weights models task transitions. We begin by showing that gradient descent with L2 regularization attains $\Omega(1)$ expected average regret. This analysis characterizes plasticity loss, at the neuron-level, as being a consequence of network weights learned for previous tasks forming poor initializations for subsequent tasks. Secondly, we provide a positive result demonstrating that gradient descent with access to a reset oracle attains $O(d^2 \ln T/T)$ expected average regret, demonstrating that neuron resets are an effective mechanism for identifying poor initializations and thus remedying plasticity loss.

Having characterized plasticity loss and shown that resets are an effective remedy in Section C.1, in Section D we answer the question of what makes an effective reset oracle in practice when data is arriving sequentially. To this end, we frame the problem of detecting an inactive neuron as that of an optimal hypothesis test. We show that an optimal reset scheme must have a reset threshold that is proportional to the neuron's firing rate; this is precisely the SNR mechanism. In contrast, existing competitor reset schemes define a fixed reset threshold or frequency. Our experiments in Section 3 show that such reset schemes attain performance inferior to that of SNR. To shed intuition on this observation, we show that under this optimal hypothesis test model, one can incur arbitrarily large error relative to SNR if attempting to detect two dead neurons with distinct firing rates with a single threshold.

C.1 Characterizing Plasticity Loss in a Single ReLU and the Promise of Resets

In this section, we provide a theoretical analysis of learning a single ReLU-activated neuron with gradient descent and with access to a reset oracle, where the target neuron's weights are chosen adversarially after the network's weight initialization is fixed. This adversarial selection serves as an abstraction of task transitions in a continual learning setting, allowing for scenarios where successive tasks may be unrelated—or even deliberately chosen to be adversarial. In contrast, prior work has solely considered settings where the target neuron's weights are fixed before the network's initialization is sampled [23, 12, 25]. While such literature demonstrate the effectiveness of first-order methods for learning a single neuron and the benefits of modern weight initializations, they do not capture the challenges of task transitions, where the initialization for a new task may be influenced by prior learning.

C.1.1 Preliminaries

To study the learning dynamics of a single ReLU-activated neuron, we consider the expected squared error:

$$L(w) = \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2], \tag{3}$$

where $\sigma(z) = \max(0, z)$ is the ReLU activation function, v is the target neuron's weight vector, and the expectation is taken over $x \in \mathbb{R}^{d+1}$, sampled from $\mathcal{N}(0, I_d) \times \{1\}$, where I_d is the d-dimensional identity matrix. The gradient of the loss is then given by

$$\nabla L(w) = \mathbb{E}[2(\sigma(w^\top x) - \sigma(v^\top x))x\mathbb{1}_{\{w^\top x \geq 0\}}].$$

where we define the subgradient of σ as $\sigma'(z) = \mathbb{1}_{\{z \geq 0\}}$, which corresponds to a particular choice of subgradient at $z = 0$. Next, we introduce the L_2-regularized loss function $\hat{L}(w)$, given by

$$\hat{L}(w) = \mathbb{E}\big[(\sigma(w^\top x) - \sigma(v^\top x))^2\big] + \frac{\lambda}{2}\|w\|^2, \tag{4}$$

where $\lambda \geq 0$ is the regularization parameter. Setting $\lambda = 0$ recovers the original loss function $L(w)$. The additional L_2 regularization term contributes an additive term λw to the gradient, yielding

$$\nabla \hat{L}(w) = \nabla L(w) + \lambda w = \mathbb{E}\big[2(\sigma(w^\top x) - \sigma(v^\top x))x\mathbb{1}_{\{w^\top x \geq 0\}}\big] + \lambda w.$$

For an iterate w_t, step size α, and a general loss function $\mathcal{L}(w)$, which may represent either the standard loss $L(w)$ or the L_2-regularized loss $\hat{L}(w)$, the gradient descent update is defined as

$$w_{t+1} = w_t - \alpha \nabla \mathcal{L}(w_t).$$

Finally, we define the *average regret* in learning the target neuron's weights as

$$R_T = \frac{1}{T}\sum_{t=0}^{T-1}\left(L(w_t) - L(w^*)\right) = \frac{1}{T}\sum_{t=0}^{T-1} L(w_t), \tag{5}$$

where the second equality follows from the fact that the optimal weight vector for (3) is $w^* = v$, and thus

$$L(w^*) = L(v) = 0.$$

We also define the *expected average regret*, $\mathbb{E}[R_T]$, which accounts for randomness in the learning process. Specifically, the expectation is taken over any stochasticity introduced by the learning algorithm, including the initialization of w_0 and any reinitializations of w_t due to the reset oracle. Thus, the goal of learning a ReLU-activated neuron is to minimize (expected) average regret.

C.1.2 Non-Adversarial Target-Weights

Prior work on gradient-descent dynamics has almost exclusively considered the regime in which the target weights v are fixed before the network weights w_0 are drawn from a prescribed initialization distribution [23, 25, 12]. For clarity, Algorithm 2 formalizes gradient descent under this non-adversarial setup.

Algorithm 2: Gradient Descent for Minimizing $\mathcal{L}(w)$ with Non-Adversarial Target Weights

```
Target Selection      : Target weights $v$ are fixed
Weight Initialization : Initialize $w_0$ from some distribution
for $t \leftarrow 0, 1, 2, \dots$ do
    Compute gradient $\nabla\mathcal{L}(w_t)$;
    Update weights: $w_{t+1} \leftarrow w_t - \alpha\,\nabla\mathcal{L}(w_t)$;
end
```

Vardi et al. [23] show that gradient descent (Algorithm 2) under mild distributional conditions which capture our setting of $x \sim N(0, I_d) \times \{1\}$, converges *linearly* to the unique global minimizer $w^* = v$:

Theorem C.1 (High-Probability Convergence, Corollary 5.5 of Vardi et al. [23])**.** Under the boundedness (Assumption 5.1) and spread (Assumption 5.3) conditions on the first d coordinates, initialize

$$w_0 = (\tilde{w}_0,\, 0), \qquad \tilde{w}_0 \;\sim\; \text{Uniform}\big(\mathbb{S}^{d-1}(\rho)\big), \quad \rho = \frac{M}{c^2}.$$

Then for any step-size $\alpha \le \gamma / c^4$, there exist constants $M, c, \gamma > 0$ (depending on the distribution) such that with probability at least $1 - e^{-\Omega(d)}$,

$$\|w_t - v\|^2 \;\le\; \big(1 - \gamma\,\alpha\big)^t\,\|w_0 - v\|^2.$$

The high-probability, linear-rate convergence established in Theorem C.1 implies an $O(1/T)$ expected average regret when training a single ReLU neuron with non-adversarial targets. In contrast, we show below that if the target weights can be chosen after the network is initialized, that is adversarially, the average regret no longer vanishes, even under very mild conditions.

C.1.3 Adversarial Target-Weights

In continual learning, the weight initialization for a task can be those learned for a preceding task and can yield a poor initialization or local minimum. We formalize this notion by introducing the setting of adversarially selected target weights v. We formally define (L_2-regularized) gradient descent under adversarial selection of target weights below in Algorithm 3.

Algorithm 3: Gradient Descent for Minimizing $\mathcal{L}(w)$ with Adversarial Target Weights

```
Weight Initialization : Initialize $w_0$ from some distribution
Target Selection      : Target weights $v$ chosen adversarially with knowledge of $w_0$
for $t \leftarrow 0, 1, 2, \dots$ do
    Compute gradient $\nabla\mathcal{L}(w_t)$;
    Update weights: $w_{t+1} \leftarrow w_t - \alpha\,\nabla\mathcal{L}(w_t)$;
end
```

Below in Theorem, C.2 we show that (L_2-regularized) gradient descent can attain $\Omega(1)$ average regret under this adversarial model.

Theorem C.2. Let $L(\cdot)$ be the objective function for learning a single neuron (see Eq. (3)), and let $w_0 \in \mathbb{R}^{d+1}$ be an initialization satisfying $(w_0)_{d+1} \leq 0$. Define the target weight vector $v \in \mathbb{R}^{d+1}$ so that $v_{1:d} = -(w_0)_{1:d}$ and $v_{d+1} < -(w_0)_{d+1}$. Then, for any step size $\alpha < \frac{1}{2(d+1)}$ and any regularization parameter $\lambda < 2d$, performing gradient descent on the L2-regularized objective $\hat{L}(\cdot)$ from w_0, i.e. Algorithm 3, satisfies

$$\forall T \geq 0, \quad R_T \;\geq\; L(0) \;>\; 0.$$

We interpret this result as w_0 being the optimal solution for a preceding task and thereby the starting point for the current task at time $t = 0$. It is a priori unclear that there is an efficient way to detect when network weights w_t are in a poor local minimum, and moreover, how to remedy it. Fortunately, weight-resets in conjunction with a reset-oracle solve this issue.

C.1.4 Adversarial Target Weights with a Reset Oracle

We define the *reset oracle* $\mathcal{R}_\delta(w)$ as

$$\mathcal{R}_\delta(w) = \begin{cases} 1, & \text{if } \mathbb{E}[\sigma(w^\top x)] \leq \delta, \\ 0, & \text{if } \mathbb{E}[\sigma(w^\top x)] > \delta. \end{cases}$$

The reset oracle $\mathcal{R}_\delta$ is parameterized by a threshold δ and returns 1 (True) if and only if the expected activation of the neuron, defined by the weight vector w, falls below δ. This reset oracle is an idealization of the reset-criteria of reset-algorithms like CBP, ReDO, and SNR.

Next, we introduce the gradient descent algorithm with the reset oracle $\mathcal{R}_\delta$ for minimizing the loss function $L(w)$. The reset oracle monitors the neuron's expected activation and determines whether a reset is necessary. If the reset condition is met, the weights are reinitialized from the original distribution; otherwise, the standard gradient descent update is applied to minimize $L(w)$.

Algorithm 4: Gradient Descent with Reset Oracle for Minimizing $L(w)$

```
Weight Initialization : Initialize $w_0$ from some distribution
Target Selection      : Target weights $v$ chosen adversarially with knowledge of $w_0$
for $t \leftarrow 0, 1, 2, \dots$ do
    if $\mathcal{R}_\delta(w_t) = 0$ then
        Sample $w_{t+1}$ from the initial distribution;
    else
        Compute gradient $\nabla L(w_t)$;
        Update weights: $w_{t+1} \leftarrow w_t - \alpha \nabla L(w_t)$;
    end
end
```

Below, we present our main theoretical contributions. When the target weights v are selected adversarially with knowledge of the network initialization w_0, Theorem C.3 establishes that gradient descent with the reset oracle (Algorithm 4) achieves vanishing expected average regret.

Theorem C.3. Let w_0 be sampled from $\mathcal{D}_{w_0} = \text{Unif}(l \cdot S^{d-1} \times \{0\})$ for some positive constant $l > 0$, where $l \cdot S^{d-1} \subseteq \mathbb{R}^d$ is the l-radius sphere. Let $v \in \mathbb{R}^{d+1}$ (possibly chosen adversarially with knowledge of w_0) be the target neuron's weights satisfying

$$-v_{d+1} \leq \frac{\|v_{1:d}\|}{2\sqrt{2\pi}}, \quad \|v_{1:d}\| = 2\pi l, \quad \|v\| \geq C_1, \quad \|v_{1:d}\| \geq C_2\|v\|,$$

for some constants $C_1, C_2 > 0$, and suppose

$$\delta < \left(\frac{C_2^2}{8\pi^2\left(\sqrt{3} + \frac{2}{C_1}\right)} \right)^3.$$

Then, minimizing $L(\cdot)$ via gradient descent with step size $\alpha \leq \frac{1}{2(d+1)}$, initialized at w_0, and employing a reset oracle $\mathcal{R}_\delta$, i.e. Algorithm 4, yields

$$\forall T \geq 0, \quad \mathbb{E}[R_T] \leq \mathcal{O}\Big(\frac{d^2 \ln T}{T}\Big),$$

where the expectation is taken over the randomness of the the initialization of w_0 and the resets.

Remarks. For simplicity of exposition, Theorem C.3 assumes that $w_{1:d}$ is sampled uniformly from the sphere of radius $l = \frac{\|v_{1:d}\|}{2\pi}$. Alternatively, one may sample $w_{1:d}$ from $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = \frac{\|v_{1:d}\|^2}{4\pi^2 d}$, in which case standard concentration of measure arguments imply that

$$\mathbb{P}\Big(\|w\|^2 \in \Big[(1-\epsilon)\frac{\|v_{1:d}\|^2}{4\pi^2}, (1+\epsilon)\frac{\|v_{1:d}\|^2}{4\pi^2}\Big]\Big) \geq 1 - e^{-\Omega(d\epsilon^2)}.$$

Additionally, the assumption $-v_{d+1} \leq \frac{\|v_{1:d}\|}{2\sqrt{2\pi}}$ ensures that the bias term v_{d+1} is not arbitrarily negative relative to $\|v_{1:d}\|$. The assumptions of Theorem C.3, including the initialization distribution of w_0, the latter bound on v_{d+1}, and the constraints on $\|v\|$ and $\|v_{1:d}\|$, are identical, up to the choice of constants, to those used in the gradient descent analysis of [23], the most state-of-the-art work on learning a single ReLU-activated neuron in the non-adversarial setting. While the assumptions of Theorem C.3 align with those in prior work, our analysis of gradient descent with resets takes an independent approach.

C.2 Supporting Propositions

Proposition C.1. Let $v \in \mathbb{R}^{d+1}$ be a unit-norm vector and suppose that $x \sim \mathcal{N}(0, I_d) \times \{1\}$. Then,

$$\mathbb{E}[\sigma(v^\top x)^4] \leq 3$$

Proof. First, we note that

$$\mathbb{E}[\sigma(v^\top x)^4] \leq \mathbb{E}[(v^\top x)^4]$$

Then $v^\top x = z + \beta$ where $z \sim \mathcal{N}(0, \alpha^2)$ such that $\alpha^2 = \|v_{1:d}\|^2$ and $\alpha^2 + \beta^2 = 1$ by the fact that $\|v\|^2 = 1$. Furthermore,

$$\begin{aligned}\mathbb{E}[(v^\top x)^4] &= \mathbb{E}[(z+\beta)^4] \\ &= \mathbb{E}[z^4 + 4\beta z^3 + 6\beta^2 z^2 + 4\beta^3 z + \beta^4]\end{aligned}$$

By symmetry, $\mathbb{E}[4\beta z^3] = \mathbb{E}[4\beta^3 z] = 0$. For a zero-mean Gaussian with variance α^2, $\mathbb{E}[z^4] = 3\alpha^4$ and $\mathbb{E}[6\beta^2 z^2] = 6\beta^2\alpha^2$. Therefore,

$$\mathbb{E}[(v^\top x)^4] = 3\alpha^4 + 6\beta^2\alpha^2 + \beta^4$$

We define $a = \alpha^2$ and by $\beta^2 = 1 - \alpha^2$ we have that

$$\begin{aligned}\mathbb{E}[(v^\top x)^4] &= 3a^2 + 6a(1-a) + (1-a)^2 \\ &= -2a^2 + 4a + 1\end{aligned}$$

By standard analysis of quadratic functions, the maximum is attained at $a = 1$ with a value of 3. Therefore, $\mathbb{E}[\sigma(v^\top x)^4] \leq 3$ □

Proposition C.2. Let $v \in \mathbb{R}^{d+1}$ be a unit-norm vector and suppose that $x \sim \mathcal{N}(0, I_d) \times \{1\}$. Then,

$$\mathbb{E}[\sigma(v^\top x)^2] \leq 1$$

Proof. First, we note that

$$\mathbb{E}[\sigma(v^\top x)^2] \leq \mathbb{E}[(v^\top x)^2]$$

Then $v^\top x = z + \beta$ where $z \sim \mathcal{N}(0, \alpha^2)$ such that $\alpha^2 = \|v_{1:d}\|^2$ and $\alpha^2 + \beta^2 = 1$ by the fact that $\|v\|^2 = 1$. We then observe that,

$$\begin{aligned}\mathbb{E}[(v^\top x)^2] &= \mathbb{E}[(z+\beta)^2] \\ &= \mathbb{E}[z^2 + 2\beta z + \beta^2] \\ &= \alpha^2 + \beta^2 && \text{by } z \sim \mathcal{N}(0, \alpha^2) \\ &= 1 && \text{by } \alpha^2 + \beta^2 = 1\end{aligned}$$

□

Proposition C.3. If $x, y \in \mathbb{R}$ such that $y \leq 0$ then $\sigma(x+y) \geq \sigma(x) + y$.

Proof. We consider two cases. If $x \geq 0$ then

$$\sigma(x) + y = x + y \leq \sigma(x+y)$$

On the other hand, if $x < 0$ then

$$\sigma(x) + y = y \leq 0 \leq \sigma(x+y)$$

□

C.3 Convergence in Norm

In the following lemma, we show that L is bounded above by a quadratic function, which in turn guarantees the convergence of gradient descent when minimizing L.

Lemma C.1. For any $w, u \in \mathbb{R}^{d+1}$ then

$$L(w) \leq L(u) + \nabla L(u)^\top (w-u) + M \|w-u\|^2 \tag{6}$$

where $M = d+1$.

Proof. We let $x \in \mathbb{R}^{d+1}$ such that $x_{d+1} = 1$ be arbitrary and we define

$$L_x(w) = (\sigma(w^\top x) - \sigma(v^\top x))^2$$

and

$$\bar{L}_x(w) = (w^\top x - \sigma(v^\top x))^2$$

such that $L(w) = \mathbb{E}[L_x(w)]$ and $\nabla L(w) = \mathbb{E}[\nabla L_x(w)]$. In order to prove the desired result, (6), by the linearity of expectation it suffices to prove that

$$L_x(w) \leq L_x(u) + \nabla L_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2 \tag{7}$$

since $\mathbb{E}[\|x\|^2] = d+1$. We first prove that $\bar{L}_x$ has a $2\|x\|^2$-Lipschitz continuous gradient.

$$\begin{aligned} ||\nabla \bar{L}_x(w) - \nabla \bar{L}_x(u)|| &= 2\left\|((w^\top x - \sigma(v^\top x)) - (u^\top x - \sigma(v^\top x)))x\right\| \\ &= 2\|x\| \, |w^\top x - u^\top x| \\ &\leq 2\|x\|^2 \|w-u\| && \text{by Cauchy-Schwarz} \end{aligned}$$

Therefore, $\bar{L}_x$ has a $2\|x\|^2$-Lipschitz continuous gradient, and moreover, it follows by standard analysis that $\bar{L}_x$ has a quadratic upper bound of the form

$$\bar{L}_x(w) \leq \bar{L}_x(u) + \nabla \bar{L}_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2 \tag{8}$$

We additionally observe that $L_x(w) \leq \bar{L}_x(w)$ by the following argument. If $w^\top x \geq 0$ then clearly $L_x(w) = \bar{L}_x(w)$. Otherwise, if $w^\top x < 0$ then we have that

$$L_x(w) = (-\sigma(v^\top x))^2 \leq (w^\top x - \sigma(v^\top x))^2 = \bar{L}_x(w)$$

where the inequality follows by the fact that $w^\top x < 0$ and $-\sigma(v^\top x) \leq 0$. To establish (7), we consider several cases. Firstly, if $u^\top x \geq 0$ then

$$L_x(u) = \bar{L}_x(u) \text{ and } \nabla L_x(u) = \nabla \bar{L}_x(u) \tag{9}$$

Then it follows that

$$\begin{aligned} L_x(w) &\leq \bar{L}_x(w) && \text{as shown above} \\ &\leq \bar{L}_x(u) + \nabla \bar{L}_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2 && \text{by (8)} \\ &= L_x(u) + \nabla L_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2 && \text{by (9)} \end{aligned}$$

Next, we consider the case where $u^\top x < 0$ and $w^\top x < 0$. Then it follows that $L_x(w) = L_x(u) = \sigma(v^\top x)^2$ and $\nabla L_x(u) = 0$. Hence,

$$L_x(w) = L_x(u) \leq L_x(u) + \nabla L_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2$$

Finally, we consider the case where $u^\top x < 0$ and $w^\top x \geq 0$. Then, we let $\bar{w}$ be the vector in the convex combination of u and w such that $\bar{w}^\top x = 0$. Then we have by the choice of $\bar{w}$ and $u^\top x < 0$

$$\|w - \bar{w}\| \leq \|w - u\| \tag{10}$$

$$L_x(u) = \bar{L}_x(\bar{w}) = \sigma(v^\top x)^2 \tag{11}$$

$$\nabla L_x(u) = 0 \tag{12}$$

and

$$\nabla \bar{L}_x(\bar{w}) = 2(\sigma(\bar{w}^\top x) - \sigma(v^\top x))x\mathbb{1}_{\{\bar{w}^\top x \geq 0\}} = -2\sigma(v^\top x)x \tag{13}$$

Then we argue as follows

$$\begin{aligned}
L_x(w) &\leq \bar{L}_x(w) && \text{as shown above}\\
&\leq \bar{L}_x(\bar{w}) + \nabla \bar{L}_x(\bar{w})^\top (w-\bar{w}) + \|x\|^2 \|w-\bar{w}\|^2 && \text{by (8)}\\
&= L_x(u) + \nabla \bar{L}_x(\bar{w})^\top (w-\bar{w}) + \|x\|^2 \|w-\bar{w}\|^2 && \text{by (11)}\\
&\leq L_x(u) + \nabla \bar{L}_x(\bar{w})^\top (w-\bar{w}) + \|x\|^2 \|w-u\|^2 && \text{by (10)}\\
&= L_x(u) - 2\sigma(v^\top x)(w^\top x - \bar{w}^\top x) + \|x\|^2 \|w-u\|^2 && \text{by (13)}\\
&= L_x(u) - 2\sigma(v^\top x)(w^\top x) + \|x\|^2 \|w-u\|^2 && \text{by } \bar{w}^\top x = 0\\
&\leq L_x(u) + \|x\|^2 \|w-u\|^2 && \text{by } w^\top x, \sigma(v^\top x) \geq 0\\
&= L_x(u) + \nabla L_x(u)^\top (w-u) + \|x\|^2 \|w-u\|^2 && \text{by (12)}
\end{aligned}$$

This completes the proof. □

Lemma C.2. Let $w_0, w_1, \ldots, w_T$ be iterates of gradient descent applied to minimizing L, equation (3), with initialization at w_0 and step size $\alpha \leq \frac{1}{2(d+1)}$. Then there exists an iterate $w_t \in \{w_0, w_1, \ldots, w_{T-1}\}$ such that

$$\|\nabla L(w_t)\|^2 \leq \frac{L(w_0)}{T(\alpha - (d+1)\alpha^2)} \leq \frac{\|w_0\|^2 + \|v\|^2}{T(\alpha - (d+1)\alpha^2)} \tag{14}$$

and

$$L(w_{t+1}) \leq L(w_t), \, \forall t \geq 0 \tag{15}$$

Proof. According to Lemma C.1

$$\begin{aligned}
L(w_{t+1}) &\leq L(w_t) + \nabla L(w_t)^\top (w_{t+1} - w_t) + (d+1)\|w_{t+1} - w_t\|^2\\
&= L(w_t) - \alpha \|\nabla L(w_t)\|^2 + (d+1)\alpha^2 \|\nabla L(w_t)\|^2 && \text{by GD update}\\
&= L(w_t) + ((d+1)\alpha^2 - \alpha)\|\nabla L(w_t)\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
(\alpha - (d+1)\alpha^2)\sum_{t=0}^{T-1} \|\nabla L(w_t)\|^2 &\leq \sum_{t=0}^{T-1} L(w_t) - L(w_{t+1})\\
&= L(w_0) - L(w_T)\\
&\leq L(w_0) && \text{by } L(w_T) \geq 0
\end{aligned}$$

This implies that there exists some index $t \in \{0, 1, \ldots, T-1\}$ such that

$$\|\nabla L(w_t)\|^2 \leq \frac{L(w_0)}{T(\alpha - (d+1)\alpha^2)}$$

Then we bound $L(w_0)$ as follows.

$$\begin{aligned}
L(w_0) &= \mathbb{E}[(\sigma(w_0^\top x) - \sigma(v^\top x))^2] \\
&= \mathbb{E}[\sigma(w_0^\top x)^2 + \sigma(v^\top x)^2 - 2\sigma(w_0^\top x)\sigma(v^\top x)] \\
&\leq \mathbb{E}[\sigma(w_0^\top x)^2] + \mathbb{E}[\sigma(v^\top x)^2] \\
&= \|w_0\|^2 \, \mathbb{E}[\sigma(\frac{w_0}{\|w_0\|}^\top x)^2] \mathbb{1}_{\{w_0 \neq 0\}} + \|v\|^2 \, \mathbb{E}[\sigma(\frac{v}{\|v\|}^\top x)^2] \mathbb{1}_{\{v \neq 0\}} \\
&\leq \|w_0\|^2 + \|v\|^2 && \text{by Proposition C.2}
\end{aligned}$$

Finally, (15) follows by $(d+1)\alpha^2 - \alpha \leq 0$ for all $\alpha \in [0, \frac{1}{d+1}]$. Therefore, by the choice of step size $\alpha \leq \frac{1}{2(d+1)}$ we have that $(d+1)\alpha^2 - \alpha \leq 0$ and so

$$L(w_{t+1}) \leq L(w_t) + ((d+1)\alpha^2 - \alpha) \|\nabla L(w_t)\|^2 \leq L(w_t)$$

□

C.4 Convergence in Iterates

We define the directional derivative of $L(w)$ as

$$D_d L(w) = \nabla L(w)^\top d = 2\mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x)) x^\top d \mathbb{1}_{\{w^\top x \geq 0\}}] \tag{16}$$

For any set $A \subseteq \mathbb{R}^{d+1}$, we define the restricted "OLS" objective to be

$$L_A^r(w) = \mathbb{E}[(w^\top x - v^\top x)^2 \mathbb{1}_{\{x \in A\}}] \tag{17}$$

Lemma C.3. Let $A = \{x : w^\top x, v^\top x \geq 0\}$. Then

$$D_{w-v} L(w) = 2L_A^r(w) + c$$

where $c \geq 0$ is some nonnegative value.

Proof. Let $x \in \mathbb{R}^{d+1}$ be arbitrary. We break the proof into several cases. If $w^\top x, v^\top x \geq 0$ then

$$\begin{aligned}
2(\sigma(w^\top x) - \sigma(v^\top x)) x^\top (w - v) \mathbb{1}_{\{w^\top x \geq 0\}} &= 2(w^\top x - v^\top x)^2 \\
&= 2(w^\top x - v^\top x)^2 \mathbb{1}_{\{x \in A\}}
\end{aligned}$$

If $w^\top x \geq 0$ and $v^\top x < 0$ then

$$\begin{aligned}
&2(\sigma(w^\top x) - \sigma(v^\top x)) x^\top (w - v) \mathbb{1}_{\{w^\top x \geq 0\}} \\
&= 2(w^\top x)(w^\top x + |v^\top x|) && \text{by } w^\top x \geq 0 \text{ and } v^\top x < 0 \\
&\geq 0
\end{aligned}$$

For brevity, we define $B = \{x : w^\top x \geq 0, v^\top x < 0\}$. The preceding case analysis implies that

$$\begin{aligned}
D_{w-v} L(w) &= 2\mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x)) x^\top (w - v) \mathbb{1}_{\{w^\top x \geq 0\}}] \\
&= 2\mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x)) x^\top (w - v)(\mathbb{1}_{\{x \in A\}} + \mathbb{1}_{\{x \in B\}})] \\
&= 2\mathbb{E}[(w^\top x - v^\top x)^2 \mathbb{1}_{\{x \in A\}}] + 2\mathbb{E}[w^\top x (w^\top x + |v^\top x|) \mathbb{1}_{\{x \in C\}}] \\
&= 2L_A^r(w) + c
\end{aligned}$$

where we let $c = 2\mathbb{E}[w^\top x (w^\top x + |v^\top x|) \mathbb{1}_{\{x \in C\}}]$. Then $c \geq 0$ by the preceding cases analysis. This completes the proof. □

Proposition C.4. $D_{w-v}(w) \leq \|\nabla L(w)\| \, \|w - v\|$

Proof. By Cauchy-Schwarz we have that

$$D_{w-v}(w) = \nabla L(w)^\top (w - v) \leq \|\nabla L(w)\| \, \|w - v\|$$

□

Proposition C.5. Let $A \subseteq \mathbb{R}^{d+1}$, then

$$L_A^r(w) \geq \|w - v\|^2 \lambda_{\text{min}}(\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}])$$

where $\lambda_{\text{min}}(\cdot)$ is the minimum eigenvalue of a matrix.

Proof.

$$\begin{aligned} L_A^r(w) &= \mathbb{E}[(w^\top x - v^\top x)^2 \mathbb{1}_{\{x \in A\}}] \\ &= (w - v)^\top \mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}](w - v) \\ &\geq \|w - v\|^2 \lambda_{\text{min}}(\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}]) \end{aligned}$$

where the last line holds by the fact that $\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}]$ is a real symmetric matrix. □

Lemma C.4. Let x be sampled from $\mathcal{N}(0, I_d) \times \{1\}$. Suppose $A = \{x : w^\top x \geq 0, v^\top x \geq 0\}$ satisfies $\mathbb{P}(A) \geq p$ for some $p > 0$. Then it follows that

$$\lambda_{\min} \left(\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}] \right) \geq \frac{p^3}{4\sqrt{2\pi e}}$$

Proof. We begin by noting that

$$\lambda_{\min} \left(\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}] \right) = \min_{u : \|u\| = 1} \mathbb{E}[(u^\top x)^2 \mathbb{1}_{\{x \in A\}}]$$

We proceed by showing that for any u such that $\|u\| = 1$ then $\mathbb{E}[(u^\top x)^2 \mathbb{1}_{\{x \in A\}}] \geq \frac{p^3}{4\sqrt{2\pi e}}$. We let u be an arbitrary vector such that $\|u\| = 1$. We have that $Z_u = u^\top x \sim \mathcal{N}(u_{d+1}, \|u_{1:d}\|^2)$. Then $\mathbb{E}[Z_u^2 \mathbb{1}_A]$ is minimized when the measure of A contains as much of $u_{1:d}^\perp$ as possible, where

$$u^\perp = \{x \in \mathbb{R}^d : x^\top u_{1:d} = 0\}$$

Therefore, we suppose that

$$A \supseteq (u_{1:d}^\perp + \{Z_u : Z_u \in [a, b]\}) \times \{1\} \tag{18}$$

for some $a < b$. The set above takes its particular form since $A = \{x : w^\top x \geq 0, v^\top x \geq 0\}$ is a closed convex set, and hence, among all convex sets of measure $\mathbb{P}(A)$, the one that minimizes $\mathbb{E}[Z_u^2 \mathbb{1}_A]$ is necessarily a slab of the form (18). Then,

$$\mathbb{E}[Z_u^2 \mathbb{1}_A] \geq \mathbb{E}[Z_u^2 \mathbb{1}_{\{a \leq Z \leq b\}}]$$

We can furthermore minimize the above expectation by considering the interval $[a, b] = [-r, r]$ for any $r > 0$ such that $\mathbb{P}(-r \leq Z_u \leq r) \leq \mathbb{P}(A)$.

Hence,

$$\mathbb{E}[Z_u^2 \mathbb{1}_{\{a \leq Z_u \leq b\}}] \geq \mathbb{E}[Z_u^2 \mathbb{1}_{\{-r \leq Z_u \leq r\}}]$$

Furthermore, the above expectation is minimized when Z_u^2 is concentrated as much as possible around zero, which occurs precisely when $u_{d+1} = 0$ and $Z_u = Z \sim \mathcal{N}(0, \|u_{1:d}\|^2) = \mathcal{N}(0, 1)$, where we recall that $\|u\| = 1$. Therefore,

$$\mathbb{E}[Z_u^2 \mathbb{1}_{\{-r \leq Z_u \leq r\}}] \geq \mathbb{E}[Z^2 \mathbb{1}_{\{-r \leq Z \leq r\}}]$$

Then we make the following lower bound

$$\mathbb{E}[Z^2 \mathbb{1}_{\{-r \leq Z \leq r\}}] \geq \frac{r^2}{4} \mathbb{P}\left(|Z| \in [\frac{r}{2}, r] \right) \tag{19}$$

We return to the choice of r, recalling that r can be as large as possible provided that $\mathbb{P}(-r \leq Z \leq r) \leq \mathbb{P}(A)$. We therefore, choose $r = p$ and verify that this inequality holds.

$$\begin{aligned}
\mathbb{P}(-r \leq Z \leq r) &= \int_{-r}^{r} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&\leq \int_{-r}^{r} \frac{1}{\sqrt{2\pi}}\, dx \\
&= \sqrt{\frac{2}{\pi}} r \\
&= \sqrt{\frac{2}{\pi}} p && \text{by choice of } r \\
&\leq p && \text{by } 2 \leq \pi \\
&\leq \mathbb{P}(A) && \text{by assumption}
\end{aligned}$$

Hence, we return to (19)

$$\begin{aligned}
\frac{r^2}{4}\mathbb{P}(|Z| \in [\frac{r}{2}, r]) &= \frac{p^2}{4}\mathbb{P}(|Z| \in [\frac{p}{2}, p]) && \text{by } r = p \\
&= \frac{p^2}{2} \int_{\frac{p}{2}}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&\geq \frac{p^2}{2} \int_{\frac{p}{2}}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}\, dx && \text{by } |p| \leq 1 \\
&= \frac{p^3}{4\sqrt{2\pi e}}
\end{aligned}$$

Hence, we have our desired result

$$\lambda_{\min}\left(\mathbb{E}[xx^\top \mathbb{1}_A]\right) \geq \frac{p^3}{4\sqrt{2\pi e}}$$

□

Lemma C.5. If $\|\nabla L(w)\|^2 \leq \epsilon$, then

$$\|w - v\| \leq \sqrt{\epsilon} \frac{2\sqrt{2\pi e}}{\mathbb{P}(w^\top x \geq 0, v^\top x \geq 0)^3}$$

Proof. We let $A = \{x : w^\top x \geq 0, v^\top x \geq 0\} \times \{1\}$ then

$$\begin{aligned}
\sqrt{\epsilon}\,\|w - v\| &\geq \|\nabla L(w)\|\,\|w - v\| && \text{by } \|\nabla L(w)\|^2 \leq \epsilon \\
&\leq D_{w-v} L(w) && \text{by Proposition C.4} \\
&\geq 2L_A^r(w) && \text{by Lemma C.3} \\
&\geq 2\,\|w - v\|^2 \lambda_{\min}(\mathbb{E}[xx^\top \mathbb{1}_{\{x \in A\}}]) && \text{by Proposition C.5} \\
&\geq \|w - v\|^2 \frac{\mathbb{P}(w^\top x \geq 0, v^\top x \geq 0)^3}{2\sqrt{2\pi e}} && \text{by Lemma C.4}
\end{aligned}$$

Then the desired result follows

$$\|w - v\| \leq \sqrt{\epsilon} \frac{2\sqrt{2\pi e}}{\mathbb{P}(w^\top x \geq 0, v^\top x \geq 0)^3}$$

□

C.5 Resets and The Reset Oracle

Lemma C.6. If $\|\nabla L(w)\|^2 \leq \delta^2$, and $\mathbb{P}(w^\top x \geq 0, v^\top x \geq 0) \leq \frac{\delta^2}{4\|v\|^2}$ then

$$\mathcal{R}_\delta(w) = 1$$

Proof. By $\|\nabla L(w)\|^2 \leq \delta^2$ we have that

$$2|\mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))\mathbb{1}_{\{w^\top x \geq 0\}}]| = |\nabla L(w)_{d+1}| \leq \delta \tag{20}$$

Then,

$$\begin{aligned}
&\mathbb{E}[\sigma(w^\top x)] \\
&= \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))\mathbb{1}_{\{w^\top x \geq 0\}}] + \mathbb{E}[\sigma(v^\top x)\mathbb{1}_{\{w^\top x \geq 0\}}] \\
&= \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))\mathbb{1}_{\{w^\top x \geq 0\}}] + \mathbb{E}[\sigma(v^\top x)\mathbb{1}_{\{w^\top x, v^\top x \geq 0\}}] \\
&\leq \frac{\delta}{2} + \mathbb{E}[\sigma(v^\top x)\mathbb{1}_{\{w^\top x, v^\top x \geq 0\}}] && \text{by (20)} \\
&\leq \frac{\delta}{2} + \sqrt{\mathbb{E}[\sigma(v^\top x)^2]\mathbb{P}(w^\top x, v^\top x \geq 0)} && \text{by Cauchy-Schwarz} \\
&= \frac{\delta}{2} + \|v\| \sqrt{\mathbb{E}[\sigma(\frac{v}{\|v\|}^\top x)^2]\mathbb{P}(w^\top x, v^\top x \geq 0)}\mathbb{1}_{\{v \neq 0\}} \\
&\leq \frac{\delta}{2} + \|v\| \sqrt{\mathbb{P}(w^\top x, v^\top x \geq 0)}\mathbb{1}_{\{v \neq 0\}} && \text{by Proposition C.2} \\
&\leq \delta && \text{by assumption on} \\
& && \mathbb{P}(w^\top x \geq 0, v^\top x \geq 0)
\end{aligned}$$

Therefore, the condition of the reset oracle $\mathcal{R}_\delta(w)$ is satisfied and so $\mathcal{R}_\delta(w) = 1$. □

Lemma C.7. If $E[\sigma(w^\top x)] \leq \delta$ then for any $\epsilon > 0$

$$\mathbb{P}(\sigma(w^\top x) \geq \epsilon) \leq \frac{\delta}{\epsilon}$$

and

$$L(w) \geq L(0) - \sqrt{\frac{\delta}{\epsilon}}\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} - 2\epsilon\sqrt{L(0)}$$

Proof. The first inequality follows immediately by Markov's inequality

$$\mathbb{P}(\sigma(w^\top x) \geq \epsilon) \leq \frac{\mathbb{E}[\sigma(w^\top x)]}{\epsilon} \leq \frac{\delta}{\epsilon} \tag{21}$$

As for the second inequality, we proceed as follows.

$$\begin{aligned}
L(w) &= \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2] \\
&\geq \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2\mathbb{1}_{\{\sigma(w^\top x) \leq \epsilon\}}] \\
&\geq \mathbb{E}[\sigma(v^\top x)^2\mathbb{1}_{\{\sigma(w^\top x) \leq \epsilon\}}] - 2\mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)\mathbb{1}_{\{\sigma(w^\top x) \leq \epsilon\}}] \\
&\geq \mathbb{E}[\sigma(v^\top x)^2\mathbb{1}_{\{\sigma(w^\top x) \leq \epsilon\}}] - 2\epsilon\sqrt{\mathbb{E}[\sigma(v^\top x)^2]\mathbb{P}(\sigma(w^\top x) \leq \epsilon)} && \text{by Cauchy Schwarz} \\
&\geq \mathbb{E}[\sigma(v^\top x)^2\mathbb{1}_{\{\sigma(w^\top x) \leq \epsilon\}}] - 2\epsilon\sqrt{\mathbb{E}[\sigma(v^\top x)^2]} \\
&= \mathbb{E}[\sigma(v^\top x)^2] - \mathbb{E}[\sigma(v^\top x)^2\mathbb{1}_{\{\sigma(w^\top x) > \epsilon\}}] - 2\epsilon\sqrt{\mathbb{E}[\sigma(v^\top x)^2]} \\
&\geq \mathbb{E}[\sigma(v^\top x)^2] - \sqrt{\mathbb{E}[\sigma(v^\top x)^4]\mathbb{P}(\sigma(w^\top x) > \epsilon)} - 2\epsilon\sqrt{\mathbb{E}[\sigma(v^\top x)^2]} && \text{by Cauchy Schwarz} \\
&\geq \mathbb{E}[\sigma(v^\top x)^2] - \sqrt{\frac{\delta}{\epsilon}}\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} - 2\epsilon\sqrt{\mathbb{E}[\sigma(v^\top x)^2]} && \text{by (21)} \\
&= \mathbb{E}[\sigma(v^\top x)^2] - \sqrt{\frac{\delta}{\epsilon}}\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} - 2\epsilon\sqrt{L(0)}
\end{aligned}$$

□

Corollary C.1. If $E[\sigma(w^\top x)] \leq \delta$ then

$$\mathbb{P}(\sigma(w^\top x) \geq \delta^{\frac{1}{3}}) \leq \delta^{\frac{2}{3}}$$

and

$$L(w) \geq L(0) - \delta^{\frac{1}{3}} \left(\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} + 2\sqrt{L(0)} \right)$$

Proof. The proof follows immediately by Lemma C.7 and setting $\epsilon = \delta^{\frac{1}{3}}$. □

C.6 Convergence with High Probability

Lemma C.8. Let $v \in \mathbb{R}^{d+1}$ be the target neuron's weights such that $-v_{d+1} \leq \frac{\|v_{1:d}\|}{2\sqrt{2\pi}}$. Let w be sampled such that $w_{d+1} = 0$ almost surely and $w_{1:d}$ is sampled from rS^{d-1}, the sphere with radius $r = \frac{\|v_{1:d}\|}{2\pi}$. Then with probability at least $\frac{1}{2}$

$$L(w) \leq L(0) - \frac{\|v_{1:d}\|^2}{8\pi^2}$$

Proof.

$$\begin{aligned} L(w) &= \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2] \\ &= \mathbb{E}[\sigma(v^\top x)^2] + \mathbb{E}[\sigma(w^\top x)^2] - 2\mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)] \\ &= L(0) + \|w\|^2 \mathbb{E}[\sigma((\frac{w}{\|w\|})^\top x)^2] - 2\mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)] \end{aligned}$$

Since $w_{d+1} = 0$, then $(\frac{w}{\|w\|})^\top x = \sum_{i=1}^d \frac{w_i x_i}{\|w\|}$ and so $(\frac{w}{\|w\|})^\top x$ is a zero-mean normal random variable with variance $\sum_{i=1}^d \frac{w_i^2}{\|w\|^2} = 1$. Hence, $(\frac{w}{\|w\|})^\top x = Z \sim \mathcal{N}(0,1)$ and we have that

$$\begin{aligned} \mathbb{E}[\sigma((\frac{w}{\|w\|})^\top x)^2] &= \mathbb{E}[Z^2 \mathbb{1}_{\{Z \geq 0\}}] \\ &= \frac{1}{2}\mathbb{E}[Z^2] && \text{by symmetery} \\ &= \frac{1}{2} \end{aligned}$$

Thus, we have that

$$\begin{aligned} L(w) &= L(0) + \frac{\|w\|^2}{2} - 2\mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)] \\ &= L(0) + \frac{\|w\|^2}{2} - 2\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x_{1:d} + v_{d+1})] \\ &\leq L(0) + \frac{\|w\|^2}{2} - 2\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x_{1:d} + v_{d+1}\mathbb{1}_{\{v_{d+1} \leq 0\}})] \\ &\leq L(0) + \frac{\|w\|^2}{2} - 2\mathbb{E}[\sigma(w^\top x)(\sigma(v_{1:d}^\top x_{1:d}) + v_{d+1}\mathbb{1}_{\{v_{d+1} \leq 0\}})] \\ &= L(0) + \frac{\|w\|^2}{2} - 2\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x_{1:d})] - 2\mathbb{E}[\sigma(w^\top x)]v_{d+1}\mathbb{1}_{\{v_{d+1} \leq 0\}} \end{aligned}$$

where the second-last line above follows by Proposition C.3. Then, we bound the fourth term above as follows

$$\begin{aligned}
&- 2\mathbb{E}[\sigma(w^\top x)]v_{d+1}\mathbb{1}_{\{v_{d+1}\leq 0\}} \\
&= -2\left\|w\right\|\mathbb{E}[Z\mathbb{1}_{\{Z\geq 0\}}]v_{d+1}\mathbb{1}_{\{v_{d+1}\leq 0\}} && \text{where } Z\sim\mathcal{N}(0,1) \\
&= -\left\|w\right\|\mathbb{E}[|Z|]v_{d+1}\mathbb{1}_{\{v_{d+1}\leq 0\}} && \text{by symmetry} \\
&= -\left\|w\right\|\sqrt{\frac{2}{\pi}}v_{d+1}\mathbb{1}_{\{v_{d+1}\leq 0\}} && \text{by } \mathbb{E}[|Z|]=\sqrt{\frac{2}{\pi}} \\
&\leq \frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|\mathbb{1}_{\{v_{d+1}\leq 0\}} && \text{by assumption } -v_{d+1}\leq\frac{\left\|v_{1:d}\right\|}{2\sqrt{2\pi}} \\
&\leq \frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|
\end{aligned}$$

With this upper bound, we continue our proof

$$L(w)\leq L(0)+\frac{\left\|w\right\|^2}{2}+\frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|-2\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x)]$$

In order to bound $-2\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x)]$ we note the following closed form solution due to [4]

$$\mathbb{E}[\sigma(w^\top x)\sigma(v_{1:d}^\top x)]=\frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|(\sin\theta+(\pi-\theta)\cos\theta)$$

where $\theta=\arccos\left(\frac{w^\top v_{1:d}}{\|w\|\|v_{1:d}\|}\right)$. Defining $f(\theta)=\sin\theta+(\pi-\theta)\cos\theta$, we briefly argue that $f(\theta)\geq 1,\forall\theta\in[0,\frac{\pi}{2}]$. First we note that $f'(\theta)=-(\pi-\theta)\sin\theta$ and thus $f'(\theta)\leq 0,\forall\theta\in[0,\pi]$. Since $f(\frac{\pi}{2})=1$ then it follows that $f(\theta)\geq 1,\forall\theta\in[0,\frac{\pi}{2}]$. Then with probability $\frac{1}{2}$, $w^\top v\geq 0$ and hence $(\sin\theta+(\pi-\theta)\cos\theta)\geq 1$. Thus,

$$\begin{aligned}
L(w)&\leq L(0)+\frac{\left\|w\right\|^2}{2}+\frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|-\frac{1}{\pi}\left\|w\right\|\left\|v_{1:d}\right\|(\sin\theta+(\pi-\theta)\cos\theta) \\
&\leq L(0)+\frac{\left\|w\right\|^2}{2}+\frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|-\frac{1}{\pi}\left\|w\right\|\left\|v_{1:d}\right\| && \text{by } w^\top v\geq 0 \\
&\leq L(0)+\frac{\left\|w\right\|^2}{2}-\frac{1}{2\pi}\left\|w\right\|\left\|v_{1:d}\right\|
\end{aligned}$$

Then by the assumption that $w_{1:d}$ is sampled from rS^{d-1} where $r=\frac{\|v_{1:d}\|}{2\pi}$ we have our desired result, with probability at least $\frac{1}{2}$

$$L(w)\leq L(0)+\frac{\left\|v_{1:d}\right\|^2}{8\pi^2}-\frac{\left\|v_{1:d}\right\|^2}{4\pi^2}=L(0)-\frac{\left\|v_{1:d}\right\|^2}{8\pi^2}$$

□

Remarks. For simplicity of exposition, in Lemma C.8 we assume that $w_{1:d}$ is sampled uniformly from the sphere of radius $\frac{\|v_{1:d}\|}{2\pi}$; alternatively, one may sample $w_{1:d}$ from $\mathcal{N}(0,\sigma^2 I_d)$ with $\sigma^2=\frac{\|v_{1:d}\|^2}{4\pi^2 d}$, in which case standard concentration of measure arguments imply that

$$\mathbb{P}\Big(\|w\|^2\in\Big[(1-\epsilon)\frac{\left\|v_{1:d}\right\|^2}{4\pi^2},(1+\epsilon)\frac{\left\|v_{1:d}\right\|^2}{4\pi^2}\Big]\Big)\geq 1-e^{-\Omega(d\epsilon^2)}$$

We note that the $-v_{d+1}\leq\frac{\|v_{1:d}\|}{2\sqrt{2\pi}}$ in Lemma C.8 ensures that the bias v_{d+1} is not arbitrarily negative relative to $\|v_{1:d}\|$, an assumption nearly identical to that used in the gradient descent analysis of [23].

Lemma C.9. Suppose that $C_1\leq\left\|v\right\|$ and $C_2\left\|v\right\|\leq\left\|v_{1:d}\right\|$ for some constants $C_1,C_2>0$. Let $\delta<\left(\frac{C_2^2}{8\pi^2(\sqrt{3}+\frac{2}{C_1})}\right)^3$ and let $w\in\mathbb{R}^{d+1}$ such that $\mathbb{E}[\sigma(w^\top x)]\leq\delta$ then

$$L(w)>L(0)-\frac{\left\|v_{1:d}\right\|^2}{8\pi^2}$$

Proof. According to Corollary C.1, $\mathbb{E}[\sigma(w^\top x)] \leq \delta$ implies that

$$L(w) \geq L(0) - \delta^{\frac{1}{3}} \left(\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} + 2\sqrt{L(0)} \right)$$

Then the desired result holds if

$$\delta^{\frac{1}{3}} \left(\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} + 2\sqrt{L(0)} \right) < \frac{\left\|v_{1:d}\right\|^2}{8\pi^2}$$

By the assumption that $C_2 \left\|v\right\| \leq \left\|v_{1:d}\right\|$, the above holds if

$$\delta^{\frac{1}{3}} \left(\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} + 2\sqrt{L(0)} \right) < \frac{C_2^2}{8\pi^2} \left\|v\right\|^2$$

We then observe that

$$\begin{aligned}
&\frac{\frac{C_2^2}{8\pi^2} \left\|v\right\|^2}{\sqrt{\mathbb{E}[\sigma(v^\top x)^4]} + 2\sqrt{L(0)}} \\
&= \frac{\frac{C_2^2}{8\pi^2}}{\sqrt{\mathbb{E}[\sigma(\frac{v}{\left\|v\right\|}^\top x)^4]} + \frac{2}{\left\|v\right\|}\sqrt{\mathbb{E}[\sigma(\frac{v}{\left\|v\right\|}^\top x)^2]}} \\
&\geq \frac{C_2^2}{8\pi^2(\sqrt{3} + \frac{2}{\left\|v\right\|})} && \text{by Propositions C.1 and C.2} \\
&\geq \frac{C_2^2}{8\pi^2(\sqrt{3} + \frac{2}{C_1})} && \text{by } C_1 \leq \left\|v\right\| \\
&> \delta^{\frac{1}{3}} && \text{by choice of } \delta
\end{aligned}$$

Therefore, we have our desired result,

$$L(w) > L(0) - \frac{\left\|v_{1:d}\right\|^2}{8\pi^2}$$

□

Lemma C.10. If $\left\|\nabla L(w)\right\|^2 \leq \epsilon$ then

$$L(w) \leq \epsilon \frac{8\pi e(d+1)}{\mathbb{P}(w^\top x \geq, v^\top x \geq 0)^6}$$

Proof.

$$\begin{aligned}
L(w) &= \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2] \\
&\leq \mathbb{E}[(w^\top x - v^\top x)^2] \\
&\leq \left\|w - v\right\|^2 \mathbb{E}[\left\|x\right\|^2] && \text{by Cauchy-Schwarz} \\
&= (d+1)\left\|w - v\right\|^2 \\
&\leq \epsilon \frac{8\pi e(d+1)}{\mathbb{P}(w^\top x \geq 0, v^\top x \geq 0)^6} && \text{by Lemma C.5}
\end{aligned}$$

□

Lemma C.11. Suppose that $C_1 \leq \left\|v\right\|$ and $C_2 \left\|v\right\| \leq \left\|v_{1:d}\right\|$ for some constants $C_1, C_2 > 0$. Let $\delta < \left(\frac{C_2^2}{8\pi^2(\sqrt{3} + \frac{2}{C_1})} \right)^3$. We define

$$t_r = \max\left\{ \frac{\left\|w_0\right\|^2 + \left\|v\right\|^2}{\delta^2(\alpha - (d+1)\alpha^2)}, \frac{2 \cdot 4^7 \pi e(\left\|w_0\right\|^2 + \left\|v\right\|^2)\left\|v\right\|^{12}(d+1)}{(L(0) - \frac{\left\|v_{1:d}\right\|^2}{8\pi^2})\delta^{12}(\alpha - (d+1)\alpha^2)} \right\} \tag{22}$$

Then the maximum number of gradient descent steps, with step size $\alpha \leq \frac{1}{2(d+1)}$ and weights initialized at w_0, until a reset is triggered by the reset oracle $\mathcal{R}_\delta$ is at most t_r.

Proof. By Lemma C.2, there exists a time step $t \in \{0, 1, \dots, t_{r-1}\}$ such that

$$\|\nabla L(w_t)\|^2 \le \frac{\|w_0\|^2 + \|v\|^2}{t_r(\alpha - (d+1)\alpha^2)}$$

Then by the choice of t_r, we have that $\|\nabla L(w_t)\|^2 \le \delta^2$. If $\mathbb{P}(w_t^\top x \ge 0, v^\top x \ge 0) \le \frac{\delta^2}{4\|v\|^2}$ then according to Lemma C.6 we have that $\mathcal{R}_\delta(w_t) = 1$ and a reset of the weights is triggered at time t. As for the case where

$$\mathbb{P}(w_t^\top x \ge 0, v^\top x \ge 0) > \frac{\delta^2}{4\|v\|^2} \tag{23}$$

we argue that $\mathcal{R}_\delta(w_{t'}) = 0$ for all $t' \ge t$, or in other words, no resets are triggered by the reset oracle $\mathcal{R}_\delta$ beyond time t. By taking the contrapositive of Lemma C.9 we have that

$$L(w_t) \le L(0) - \frac{\|v_{1:d}\|^2}{8\pi^2} \tag{24}$$

implies that $\mathbb{E}[\sigma(w_t^\top x)] > \delta$, or equivalently, $\mathcal{R}_\delta(w_t) = 0$. Furthermore, Lemma C.2 ensures that $L(w_{t'}) \le L(w_t)$ for all $t' \ge t$, and so it suffices to establish (24) for iterate w_t. Then according to Lemma C.10 we have that

$$\begin{aligned}
L(w_t) &\le \frac{\|w_0\|^2 + \|v\|^2}{t_r(\alpha - (d+1)\alpha^2)} \frac{8\pi e(d+1)}{\mathbb{P}(w_t^\top x \ge 0, v^\top x \ge 0)^6} \\
&< \frac{2 \cdot 4^7 \pi e(\|w_0\|^2 + \|v\|^2)\|v\|^{12}(d+1)}{t_r \delta^{12}(\alpha - (d+1)\alpha^2)} && \text{by (23)} \\
&\le L(0) - \frac{\|v_{1:d}\|^2}{8\pi^2} && \text{by chocie of } t_r
\end{aligned}$$

□

C.7 Regret Guarantees

Theorem C.4. Let w_0 be sampled from $\mathcal{D}_{w_0} = \text{Unif}(l \cdot S^{d-1} \times \{0\})$ for some positive constant $l > 0$, where $l \cdot S^{d-1} \subseteq \mathbb{R}^d$ is the r-radius sphere. Let $v \in \mathbb{R}^{d+1}$ (possibly chosen adversarially with knowledge of w_0) be the target neuron's weights satisfying

$$-v_{d+1} \le \frac{\|v_{1:d}\|}{2\sqrt{2\pi}}, \quad \|v_{1:d}\| = 2\pi l, \quad \|v\| \ge C_1, \quad \|v_{1:d}\| \ge C_2\|v\|,$$

for some constants $C_1, C_2 > 0$, and suppose

$$\delta < \left(\frac{C_2^2}{8\pi^2\left(\sqrt{3} + \frac{2}{C_1}\right)} \right)^3.$$

Then, minimizing $L(\cdot)$ via gradient descent with step size $\alpha \le \frac{1}{2(d+1)}$, initialized at w_0, and employing a reset oracle $\mathcal{R}_\delta$, yields

$$\forall T \ge 0, \quad \mathbb{E}[R_T] \le \mathcal{O}\Big(\frac{d^2 \ln T}{T}\Big),$$

where the expectation is taken over the randomness of the the initialization of w_0 and the resets.

Proof. We let A_r be the event that exactly r resets occur while minimizing the objective $L(\cdot)$ with gradient descent using step size α and employing a reset oracle $\mathcal{R}_\delta$. Therefore, we have that

$$\mathbb{E}[R_T] = \frac{1}{T}\sum_{r=0}^{T} \mathbb{E}[R_T \mid A_r]\mathbb{P}(A_r)$$

We begin by bounding $\mathbb{P}(A_r)$. We let p be the probability that a reset does not ever occur when (re)initializing w_t from $\mathcal{D}_{w_0}$. We proceed to show that $p \ge \frac{1}{2}$. To this end, we suppose that

w_t is sampled from $\mathcal{D}_{w_0}$, and according to Lemma C.8, with probability at least $\frac{1}{2}$ we have that $L(w_t) \leq L(0) - \frac{\|v_{1:d}\|^2}{8\pi^2}$. Then for any iterate $w_{t'}$, where $t' \geq t$, we have that $L(w_{t'}) \leq L(w_t)$ due to Lemma C.2. By the contrapositive of Lemma C.9, we have that $\mathbb{E}[\sigma(w_{t'}^\top x)] > \delta$, or equivalently, $\mathcal{R}_\delta(w_{t'}) = 0$. Hence, (re)sampling w_t from $\mathcal{D}_{w_0}$ ensures that no resets occur for w_t and any future iterates with probability $p \geq \frac{1}{2}$.

We return to bounding $\mathbb{P}(A_r)$. In the worst case, if v is chosen adversarially with the knowledge of w_0 such that a reset occurs then $\mathbb{P}(A_0) = 0$ and $\forall r \geq 1, \mathbb{P}(A_r) = p(1-p)^{r-1} \leq \frac{1}{2^{r-1}}$. Hence,

$$\mathbb{E}[R_T] \leq \frac{1}{T}\sum_{r=1}^{T} \frac{1}{2^{r-1}} \mathbb{E}[R_T \mid A_r]$$

Next, we bound $\mathbb{E}[R_T \mid A_r]$. We let $\mathbb{E}[R_T \mid A_r] = M_{1,r} + M_{2,r}$, where $M_{1,r}$ is the total loss of all iterates preceding the final (r^{th}) reset, and where $M_{2,r}$ is the total loss of all iterates after the final reset. We begin by bounding $M_{1,r}$. We let w_t be an arbitrary (re)initialization. Then for any subsequent iterate $w_{t'}$ that precedes the next reset, we have that

$$\begin{aligned}
L(w_{t'}) &\leq L(w_t) && \text{by Lemma C.2} \\
&\leq \mathbb{E}[\sigma(v^\top x)^2] + \mathbb{E}[\sigma(w_t^\top x)^2] \\
&\leq \|v\|^2 + \|w_t\|^2 && \text{by Proposition C.2} \\
&= \|v\|^2 + l^2 && \text{by } w_t \sim \mathcal{D}_{w_0}
\end{aligned}$$

According to Lemma C.11

$$t_{\text{reset}} = \max\left\{ \frac{\|w_t\|^2 + \|v\|^2}{\delta^2(\alpha - (d+1)\alpha^2)}, \frac{2 \cdot 4^7 \pi e (\|w_t\|^2 + \|v\|^2) \|v\|^{12} (d+1)}{(L(0) - \frac{\|v_{1:d}\|^2}{8\pi^2})\delta^{12}(\alpha - (d+1)\alpha^2)} \right\} \tag{25}$$

is the maximum number of gradient descent updates before a reset occurs. Therefore,

$$M_{1,r} \leq r t_{\text{reset}}(\|v\|^2 + l^2) \tag{26}$$

As for $M_{2,r}$, we have the following upper bound

$$M_{2,r} \leq \sum_{t=0}^{T-1-r} L(w_t)$$

where we suppose that $w_0, w_1, \ldots, w_{T-1-r}$ are arbitrary iterates that do not trigger the reset oracle $\mathcal{R}_\delta$ and w_0 is sampled from $\mathcal{D}_{w_0}$. We proceed to bound $L(w_t)$. To this end, we let

$$\bar{t} = \lceil \frac{\|w_0\|^2 + \|v\|^2}{\delta^2(\alpha - (d+1)\alpha^2)} \rceil = \lceil \frac{l^2 + \|v\|^2}{\delta^2(\alpha - (d+1)\alpha^2)} \rceil \tag{27}$$

For $t \leq \bar{t}$ we crudely bound $L(w_t) \leq \|v\|^2 + l^2$, as argued earlier in the proof. For $t \geq \bar{t}$ we have, according to Lemma C.2 and the choice of $\bar{t}$, that for some $t' \leq t$,

$$\|\nabla L(w_{t'})\|^2 \leq \frac{\|w_0\|^2 + \|v\|^2}{t(\alpha - (d+1)\alpha^2)} \leq \frac{\|w_0\|^2 + \|v\|^2}{\bar{t}(\alpha - (d+1)\alpha^2)} \leq \delta^2 \tag{28}$$

By the fact that $\mathcal{R}_\delta(w_{t'}) = 0$ and the contrapositive of Lemma C.6, it follows that

$$\mathbb{P}(w_{t'}^\top x \geq 0, v^\top x \geq 0) > \frac{\delta^2}{4\|v\|^2}$$

Then we have that

$$\begin{aligned}
L(w_t) &\leq L(w_{t'}) && \text{by Lemma C.2} \\
&\leq \|\nabla L(w_{t'})\|^2 \frac{8\pi e(d+1)}{\mathbb{P}(w_{t'}^\top x \geq 0, v^\top x \geq 0)^6} && \text{by Lemma C.10} \\
&\leq \frac{1}{t} \cdot \frac{2 \cdot 4^7 \pi e (\|w_0\|^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)} && \text{by (27) and (28)} \\
&= \frac{1}{t} \cdot \frac{2 \cdot 4^7 \pi e (l^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)} && \text{by } w_0 \sim \mathcal{D}_{w_0}
\end{aligned}$$

Therefore,

$$M_{2,r} \leq (\bar{t}+1)(\|v\|^2 + l^2) + \sum_{t=\bar{t}+1}^{T-1-r} \frac{1}{t} \cdot \frac{2 \cdot 4^7 \pi e(l^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)}$$

$$\leq (\bar{t}+1)(\|v\|^2 + l^2) + \ln T \cdot \frac{2 \cdot 4^7 \pi e(l^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)}$$

where the last line follows by the standard upper bound on the Harmonic sum. Putting everything together,

$$\mathbb{E}[R_T] \leq \frac{1}{T} \sum_{r=1}^{T} \frac{1}{2^{r-1}} (M_{1,r} + M_{2,r})$$

and proceed by bounding each term separately.

$$\frac{1}{T} \sum_{r=1}^{T} \frac{1}{2^{r-1}} M_{1,r} \leq \frac{1}{T} \sum_{r=1}^{T} \frac{1}{2^{r-1}} (r t_{\text{reset}}(\|v\|^2 + l^2)$$

$$\leq \frac{4 t_{\text{reset}}(\|v\|^2 + l^2)}{T}$$

where the last line above follows by upper bounding $\sum_{r=1}^{T} \frac{r}{2^{r-1}} \leq 4$ by a standard analysis of the arithmetico-geometric series. Similarly, we upper bound

$$\frac{1}{T} \sum_{r=1}^{T} \frac{1}{2^{r-1}} M_{2,r} \leq \frac{2(\bar{t}+1)(\|v\|^2 + l^2)}{T} + \frac{\ln T}{T} \cdot \frac{4^8 \pi e(l^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)}$$

where we bound $\sum_{r=1}^{T-1} \frac{1}{2^{r-1}} \leq 2$ by a standard analysis of a geometric series. Then the expected average regret is

$$\mathbb{E}[R_T] \leq \frac{4 t_{\text{reset}}(\|v\|^2 + l^2)}{T} + \frac{2\bar{t}(\|v\|^2 + l^2)}{T} + \frac{\ln T}{T} \cdot \frac{4^8 \pi e(l^2 + \|v\|^2) \|v\|^{12} (d+1)}{\delta^{12}(\alpha - (d+1)\alpha^2)}$$

Then taking $\alpha = \frac{c}{2(d+1)}$ for some constant $c \in (0, 1]$ we have that

$$\alpha - (d+1)\alpha^2 = \frac{c}{2(d+1)} - \frac{c^2}{4(d+1)} \geq \frac{c}{4(d+1)}$$

Furthermore, if we take $\|v\|$, $\|w_0\| = l$, and δ as constants bounded according to the assumptions of the theorem statement, we have that $t_{\text{reset}} \leq \mathcal{O}(d^2)$ and $\bar{t} \leq \mathcal{O}(d)$ and, moreover,

$$\mathbb{E}[R_T] \leq \mathcal{O}(\frac{d^2}{T} + \frac{d}{T} + \frac{d^2 \ln T}{T}) = \mathcal{O}(\frac{d^2 \ln T}{T})$$

□

Corollary C.2. Let $L(\cdot)$ be the objective function for learning a single neuron (see Eq. (3)), and let $w_0 \in \mathbb{R}^{d+1}$ be an initialization satisfying $(w_0)_{d+1} \leq 0$. Define the target weight vector $v \in \mathbb{R}^{d+1}$ so that $v_{1:d} = -(w_0)_{1:d}$ and $v_{d+1} < -(w_0)_{d+1}$. Then, for any step size $\alpha < \frac{1}{2(d+1)}$ and any regularization parameter $\lambda < 2d$, performing gradient descent on the L2-regularized objective $\hat{L}(\cdot)$ from w_0 satisfies

$$\forall T \geq 0, \quad R_T \;\geq\; L(0) \;>\; 0.$$

Proof. The result follows immediately by Theorem C.5 and noting that

$$R_T = \frac{1}{T} \sum_{t=0}^{T-1} L(w_t) \geq \frac{1}{T} \sum_{t=0}^{T-1} L(0) = L(0)$$

□

C.8 Negative Results

Theorem C.5. Let $L(\cdot)$ be the objective function for learning a single neuron (see Eq. (3)), and let $w_0 \in \mathbb{R}^{d+1}$ be an initialization satisfying $(w_0)_{d+1} \leq 0$. Define the target weight vector $v \in \mathbb{R}^{d+1}$ such that $v_{1:d} = -(w_0)_{1:d}$ and $v_{d+1} < -(w_0)_{d+1}$. For any step size $\alpha < \frac{1}{2(d+1)}$ and any regularization parameter $\lambda < 2d$, every iterate w_t of gradient descent (starting from w_0) that minimizes the L2-regularized objective $\hat{L}(\cdot)$ satisfies

$$L(w_t) \geq L(0) > 0.$$

Proof. We first prove that for any iterate w_t,

$$\{x \in \mathbb{R}^d \times \{1\} : w_t^\top x \geq 0\} \cap \{x \in \mathbb{R}^d \times \{1\} : v^\top x \geq 0\} = \emptyset \text{ and } (w_t)_{d+1} \leq 0 \qquad (29)$$

We prove this claim by induction. For the base case of w_0, we let $x \in \mathbb{R}^d \times \{1\}$ be any vector such that $w_0^\top x \geq 0$, then

$$\begin{aligned} v^\top x &= v_{1:d}^\top x_{1:d} + v_{d+1} \\ &= -(w_t)_{1:d}^\top x_{1:d} + v_{d+1} && \text{by choice of } v \\ &< -(w_t)_{1:d}^\top x_{1:d} - w_{d+1} && \text{by choice of } v \\ &= -w_0^\top x \\ &\leq 0 && \text{by choice of } x \end{aligned}$$

By assumption $(w_0)_{d+1} \leq 0$, and therefore, (29) holds for the initialization w_0. Next, we consider an arbitrary iterate w_t and suppose that (29) holds for w_t and we proceed to show that (29) holds for w_{t+1}. We let $y \in \mathbb{R}^d \times \{1\}$ be an arbitrary vector such that $v^\top y \geq 0$. Then,

$$\begin{aligned} w_{t+1}^\top y &= w_t^\top y - \alpha \nabla L(w_t)^\top y - \alpha\lambda w_t^\top y && \text{by gradient descent update} \\ &= (1-\alpha\lambda) w_t^\top y - 2\alpha \mathbb{E}[(\sigma(w_t^\top x) - \sigma(v^\top x))\mathbb{1}_{\{w_t^\top x \geq 0\}} x]^\top y \\ &= (1-\alpha\lambda) w_t^\top y - 2\alpha \mathbb{E}[\sigma(w_t^\top x) x]^\top y \end{aligned}$$

where the last line follows by the fact (29) implies that $(\sigma(w_t^\top y) - \sigma(v^\top x))\mathbb{1}_{\{w_t^\top x \geq 0\}} = \sigma(w_t^\top x)$. We let $x = x_w + x_u$ where $x_w = (x_{\tilde{w}}, 1)$ such that $x_{\tilde{w}}$ is the projection of $x_{1:d}$ onto the subspace spanned by $(w_t)_{1:d}$ and $x_u = (x_{\tilde{u}}, 0)$ such that $x_{\tilde{u}}$ is the orthogonal complement and hence $w_t^\top x_u = 0$. Then

$$\begin{aligned} \mathbb{E}[\sigma(w_t^\top x)x] &= \mathbb{E}[\sigma(w_t^\top (x_w + x_u))(x_w + x_u)] \\ &= \mathbb{E}[\sigma(w_t^\top x_w)(x_w + x_u)] && \text{by } w_t^\top x_u = 0 \\ &= \mathbb{E}[\sigma(w_t^\top x_w)x_w] + \mathbb{E}[\sigma(w_t^\top x_w)x_u] \end{aligned}$$

Because $x_{1:d} \sim \mathcal{N}(0, I_d)$ is rotationally symmetric, its projections onto orthogonal subspaces are independent; consequently $x_{\tilde{w}}$ and $x_{\tilde{u}}$ are independent. Moreover, $w_t^\top x_w = (w_t)_{1:d}^\top x_{\tilde{w}} + w_{d+1}$ is independent of $x_u^\top y = x_{\tilde{u}}^\top y_{1:d}$. Therefore,

$$\begin{aligned} \mathbb{E}[\sigma(w_t^\top x_w)x_u]^\top y &= \mathbb{E}[\sigma(w_t^\top x_w)x_u^\top y] \\ &= \mathbb{E}[\sigma(w_t^\top x_w)]\mathbb{E}[x_u^\top y] && \text{by independence} \\ &= 0 \end{aligned}$$

where the last line follows by the fact that $x_u^\top y = x_{\tilde{u}}^\top y_{1:d}$ is a zero-mean normal random variable. Therefore,

$$\begin{aligned} w_{t+1}^\top y &= (1-\alpha\lambda) w_t^\top y - 2\alpha \mathbb{E}[\sigma(w_t^\top x_w) x_w^\top y] \\ &= (1-\alpha\lambda)(w_t)_{d+1} - 2\alpha \mathbb{E}[\sigma(w_t^\top x_w)] + (1-\alpha\lambda)(w_t)_{1:d}^\top y_{1:d} - 2\alpha \mathbb{E}[\sigma(w_t^\top x_w) x_{\tilde{w}}^\top y_{1:d}] \\ &\leq (1-\alpha\lambda)(w_t)_{d+1} + (1-\alpha\lambda)(w_t)_{1:d}^\top y_{1:d} - 2\alpha \mathbb{E}[\sigma(w_t^\top x_w) x_{\tilde{w}}^\top y_{1:d}] \end{aligned}$$

where the last line above follows by $2\alpha\mathbb{E}[\sigma(w_t^\top x_w)] \geq 0$. Since $x_{\tilde{w}}$ is the projection of $x_{1:d}$ onto the subspace spanned by $(w_t)_{1:d}$ then

$$x_{\tilde{w}}^\top y_{1:d} = \frac{(w_t)_{1:d}^\top x_{1:d}}{\|(w_t)_{1:d}\|^2}(w_t)_{1:d}^\top y_{1:d}$$

Hence,

$$2\alpha\mathbb{E}[\sigma(w_t^\top x_w)x_{\tilde{w}}^\top y_{1:d}] = 2\alpha\mathbb{E}[\sigma(w_t^\top x_w)\frac{(w_t)_{1:d}^\top x_{1:d}}{\|(w_t)_{1:d}\|^2}](w_t)_{1:d}^\top y_{1:d}$$

Next, we argue that

$$0 \leq 2\alpha\mathbb{E}[\sigma(w_t^\top x_w)\frac{(w_t)_{1:d}^\top x_{1:d}}{\|(w_t)_{1:d}\|^2}] < \frac{1}{d+1} \tag{30}$$

Since $(w_t)_{d+1} \leq 0$ then $w_t^\top x \geq 0$ implies that $(w_t)_{1:d}^\top x_{1:d} \geq 0$. Therefore the nonnegativity of (30) follows immediately. Moreover, this also implies that

$$\sigma(w_t^\top x)(w_t)_{1:d}^\top x_{1:d} \leq \sigma((w_t)_{1:d}^\top x_{1:d})(w_t)_{1:d}^\top x_{1:d} \leq ((w_t)_{1:d}^\top x_{1:d})^2$$

By the fact that $x_{1:d} \sim \mathcal{N}(0, I_d)$ we have

$$\frac{(w_t)_{1:d}^\top x_{1:d}}{\|(w_t)_{1:d}\|} = Z \sim \mathcal{N}(0,1) \tag{31}$$

Hence

$$\begin{aligned}
2\alpha\mathbb{E}[\sigma(w_t^\top x_w)\frac{(w_t)_{1:d}^\top x_{1:d}}{\|(w_t)_{1:d}\|^2}] &\leq 2\alpha\mathbb{E}[\frac{((w_t)_{1:d}^\top x_{1:d})^2}{\|(w_t)_{1:d}\|^2}] \\
&= 2\alpha\mathbb{E}[Z^2] && \text{by (31)} \\
&= 2\alpha \\
&< \frac{1}{d+1} && \text{by choice of } \alpha
\end{aligned}$$

Hence, (30) holds and therefore there exists some $D \in [0, \frac{1}{d+1})$ such that

$$(1-\alpha\lambda)(w_t)_{1:d}^\top y_{1:d} - 2\alpha\mathbb{E}[\sigma(w_t^\top x_w)x_{\tilde{w}}^\top y_{1:d}] = (1-\alpha\lambda-D)(w_t)_{1:d}^\top y_{1:d}$$

Continuing from earlier

$$\begin{aligned}
& w_{t+1}^\top y \\
&\leq (1-\alpha\lambda)(w_t)_{d+1} + (1-\alpha\lambda)(w_t)_{1:d}^\top y_{1:d} - 2\alpha\mathbb{E}[\sigma(w_t^\top x_w)x_{\tilde{w}}^\top y_{1:d}] \\
&= (1-\alpha\lambda)(w_t)_{d+1} + (1-\alpha\lambda-D)(w_t)_{1:d}^\top y_{1:d} \\
&\leq (1-\alpha\lambda-D)(w_t)_{d+1} + (1-\alpha\lambda-D)(w_t)_{1:d}^\top y_{1:d} && \text{by } (w_t)_{d+1} \leq 0 \leq D \\
&= (1-\alpha\lambda-D)w_t^\top y
\end{aligned}$$

Given that $\alpha < \frac{1}{2(d+1)}$, $\lambda < 2d$, and $D \in [0, \frac{1}{d+1})$ then

$$0 < 1-\alpha\lambda-D \leq 1$$

and therefore

$$w_{t+1}^\top y \leq (1-\alpha\lambda-D)w_t^\top y < 0$$

where the final inequality above follows by the inductive hypothesis (29) and the fact that $v^\top y \geq 0$. We additionally note that

$$(w_{t+1})_{d+1} = (1-\alpha\lambda)(w_t)_{d+1} - 2\alpha\mathbb{E}[\sigma(w_t^\top x)] \leq (1-\alpha\lambda)(w_t)_{d+1} \leq 0$$

where the finally inequality holds by $(w_t)_{d+1} \leq 0$ due to the inductive hypothesis (29) and the fact that $1 - \alpha\lambda \in (\frac{1}{d+1}, 1]$. Hence (29) holds for the iterate w_{t+1}. Finally, we show that if (29) holds then

$$L(w_t) \geq L(0) > 0$$

We first note that (29) implies

$$\forall x \in \mathbb{R}^d \times \{1\},\ \sigma(w_t^\top x)\sigma(v^\top x) = 0 \tag{32}$$

and so

$$\begin{aligned}
L(w_t) &= \mathbb{E}[(\sigma(w_t^\top x) - \sigma(v^\top x))^2] \\
&= \mathbb{E}[\sigma(w_t^\top x)^2] + \mathbb{E}[\sigma(v^\top x)^2] - 2\mathbb{E}[\sigma(w_t^\top x)\sigma(v^\top x)] \\
&= \mathbb{E}[\sigma(w_t^\top x)^2] + \mathbb{E}[\sigma(v^\top x)^2] && \text{by (32)} \\
&\geq \mathbb{E}[\sigma(v^\top x)^2] \\
&= L(0) \\
&> 0 && \text{by } v \neq 0
\end{aligned}$$

□

D Motivating SNR and Comparison to Other Reset Schemes

Here we motivate the SNR heuristic and compare it to other proposed reset schemes. Consider the following simple hypothesis test: we observe a discrete time process $Z_s \in \{0, 1\}$ which under the null hypothesis H_0 is a Bernoulli process with mean $p > 0$. The alternative hypothesis H_1 is that the mean of the process is identically zero. A hypothesis test must, at some stopping time τ, either reject ($X_\tau = 1$) or accept ($X_\tau = 0$) the null; an optimal such test would choose to minimize the sum of type-1 and type-2 errors (the 'error rate') and a penalty for delays:

$$\mathbb{P}(X_\tau = 1|H_0) + \mathbb{P}(X_\tau = 0|H_1) + \lambda\left(\mathbb{E}[\tau|H_0] + \mathbb{E}[\tau|H_1]\right)$$

Here the multiplier $\lambda > 0$ penalizes the delay in a decision. If $\lambda < p/2$, the optimal test takes a simple form: for some suitable threshold $\bar{T}$, reject the null iff $Z_s = 0$ for all times s up to $\bar{T}$:

Proposition D.1. Let $\bar{T}$ be the $1 - \lambda(p - \lambda)^{-1}$ percentile of a Geometric(p) distribution. Then the optimal hypothesis test takes the form $X_\tau = \mathbf{1}\{Z_\tau = 0\}$ where $\tau = \min(s : Z_s = 1) \wedge \bar{T}$.

Notice that if $\lambda \propto p$, the percentile threshold above is independent of p. Applying this setup to the setting where under the null, we observe the firing of neuron i under i.i.d. training examples from μ_t and a neuron is considered 'dead' or inactive if the alternate hypothesis is true, imagine that $p = \mathbb{P}(Z_{s,i}^{\mu_t} = 1)$. Further, we assume $\lambda = \alpha p$ ($\alpha < 1/2$); a reasonable assumption which models a larger penalty for late detection of neurons that are highly active. It is then optimal to declare neuron i 'inactive' if the length of time it has not fired exceed the $1 - \alpha(1 - \alpha)^{-1}$ percentile of the distribution of $A_i^{\mu_t}$. This is the underlying motivation for the SNR heuristic.

Comparison with Reset Schemes: Neuron reset heuristics such as Sokar et al. [22] define (sometimes complex) notions of neuron 'utility' to determine whether or not to re-initialize a neuron. The utility of every neuron is computed over every consecutive (say) r minibatches, and neurons with utility below a threshold are reset. To facilitate a comparison, consider the setting where neurons that do not fire at all over the course of the r mini batches are estimated to have zero utility, and that only neurons with zero utility are re-initialized.

This reveals an interesting comparison with SNR. The schemes above will re-initialize a neuron after inactivity over a period of time that is *uniform* across all neurons. On the other hand, SNR will reset a neuron after it is inactive for a period that corresponds to a fixed percentile of the inter-firing time distribution of that neuron. Whereas this percentile is fixed across neurons, the corresponding period of inactivity after which a neuron is reset will vary across neurons: shorter for neurons that tend to fire frequently, and longer for neurons that fire less frequently.

We can make this comparison precise in the context of the hypothesis testing setup above: specifically, consider two neurons with null firing rates p_1 and p_2 respectively ($p_1 < p_2$), and delay multipliers, λ, of αp_1 and αp_2 respectively. By Proposition D.1, under SNR, the first is reset if it is inactive for time at least $\log(\alpha(1-\alpha)^{-1})/\log(1-p_1)$ and the second if it is inactive for time at least $\log(\alpha(1-\alpha)^{-1})/\log(1-p_2)$. In contrast, for a fixed threshold scheme such as Sokar et al. [22], either neuron would be reset after being inactive for some fixed threshold, say r^*. Assume r^* is set to minimize the sum of the error rates of the two neurons while keeping the total delay identical to that for SNR. The proposition below compares the error rate between the two schemes:

Proposition D.2. The ratio of total error rate with a fixed threshold r^* to that under SNR scales like

$$\Omega\left(\exp\left(\log(\alpha(1-\alpha)^{-1})\left(-\frac{1}{2}+\frac{1}{2}\frac{\log(1-p_1)}{\log(1-p_2)}\right)\right)\right)$$

Now recall that $\alpha < 1/2$ and $p_1 < p_2$. The result above then shows that: (a) the error rate under an (optimal) fixed threshold can grow arbitrarily larger than the error rate under SNR as the penalty for delay shrinks to zero and (b) the rate at which this gap grows itself scales with the difference in the nominal firing rates of the neurons under consideration. This provides insight into the relative merits of using a scheme like SNR in lieu of existing reset proposals: *resets under SNR detect changes in the firing rate of a neuron faster and more accurately; this matters particularly in situations where there is wide disparity in the nominal firing rates of neurons across the network.*

D.1 Proofs of Proposition D.1 and D.2

Proposition D.3 (Restatement of Proposition D.1)**.** Let T be the $1-\frac{\lambda}{p-\lambda}$ percentile of a Geometric(p) distribution. Then the optimal hypothesis test takes the form $X_\tau = \mathbf{1}\{Z_\tau = 0\}$ where $\tau = \min(s : Z_s = 1) \wedge T$.

Proof. We begin by assuming equal priors $\mathbb{P}(H_0) = \mathbb{P}(H_1) = \frac{1}{2}$. We note that for any time s, if $Z_s = 1$ then any optimal hypothesis test must declare $X_s = 0$ as $Z_s = 1$ is impossible under H_1 and waiting to make a future declaration will incur additional cost of at least λ. Therefore, it remains for us to derive an optimal stopping time for the collection of states $\{Z_1 = \ldots = Z_s = 0 : s \in \mathbb{Z}_+\}$.

Let $V(s)$ be the expected total future cost at time s given that we have observed $Z_1 = \ldots = Z_s = 0$. We define

$$\begin{aligned}\pi_s &= \mathbb{P}(H_0|Z_1 = \ldots = Z_s = 0)\\ &= \frac{\mathbb{P}(H_0, Z_1 = \ldots = Z_s = 0)}{\mathbb{P}(Z_1 = \ldots = Z_s = 0)}\\ &= \frac{(1-p)^s\mathbb{P}(H_0)}{(1-p)^s\mathbb{P}(H_0)+\mathbb{P}(H_1)}\\ &= \frac{(1-p)^s}{(1-p)^s+1} && \text{by } \mathbb{P}(H_0)=\mathbb{P}(H_1)\end{aligned}$$

If we stop at time s and make a declaration, we choose the hypothesis with higher positive probability in order tom minimize the error probability

$$\mathbb{P}(X_s = 1|H_0) + \mathbb{P}(X_s = 0|H_1)$$

Thus, the expected cost of stopping is

$$C^{\text{stop}}(s) = \min\{\pi_s, 1-\pi_s\}$$

We can simplify this further by noting that $\pi_s \le 1-\pi_s$. We note that

$$\frac{1}{2}(1-p)^s \le \frac{1}{2}$$

by $1-p \in [0,1]$. This is equivalent to

$$(1-p)^s \le \frac{1}{2}((1-p)^s+1)$$

by adding $\frac{1}{2}(1-p)^s$, which in turn, is equivalent to

$$\pi_s = \frac{(1-p)^s}{(1-p)^s+1} \le \frac{1}{2}$$

Therefore, $\pi_s \leq \frac{1}{2} \leq 1 - \pi_s$ and so we have that

$$C^{\text{stop}}(s) = \min\{\pi_s, 1 - \pi_s\} = \pi_s$$

This also implies that if we are to stop at some state $\{Z_1 = \ldots = Z_s = 0\}$, it is optimal to declare $X_s = 1$.

If we continue at time s to $s + 1$, we incur an additional delay cost of λ, and the expected future cost depending on whether we see a $Z_{s+1} = 1$ or $Z_{s+1} = 0$.

- With probability $p\pi_s$ we obserbes Z_{s+1}, under H_0, and we stop the process with $X_{s+1} = 0$, incurring zero error cost since $Z_{s+1} = 1$ cannot occur under H_1.
- With probability $(1 - p)\pi_s + (1 - \pi) = 1 - p\pi_s$ we observe $Z_{s+1} = 0$ and the process continues.

Therefore, the expected cost of continuing at time s is

$$C^{\text{cont}}(s) = \lambda + (1 - p\pi_s)V(s + 1)$$

Then the Bellman equation for the optimal cost-to-go function is

$$V(s) = min\{C^{\text{stop}}(s), C^{\text{cont}}(s)\}$$

To determine an optimal stopping time, our goal is to find smallest T for which

$$C^{\text{stop}}(T) \leq C^{\text{cont}}(T) \tag{33}$$

Assuming we stop at time T,

$$V(T + 1) = C^{\text{stop}}(T + 1) = \pi_{T+1}$$

Therefore,

$$C^{\text{cont}}(T) = \lambda + (1 - p\pi_s)V(T + 1) = \lambda + (1 - p\pi_s)\pi_{T+1}$$

and to establish (33) it suffices to show that

$$\pi_T \leq \lambda + (1 - p\pi_T)\pi_{T+1} \tag{34}$$

First, we write π_{T+1} in terms of π_T. Under the updating rule for the posterior probability, we have that

$$\begin{aligned}
&\pi_{T+1} - \mathbb{P}(H_0|Z_1 = \ldots = Z_{T+1} = 0) \\
&\quad = \frac{\mathbb{P}(Z_{T+1} = 0|H_0)\mathbb{P}(H_0|Z_1 = \ldots = Z_T = 0)}{\mathbb{P}(Z_{T+1} = 0|Z_1 = \ldots = Z_T = 0)} \\
&\quad = \frac{(1 - p)\pi_T}{\mathbb{P}(Z_{T+1} = 0|H_0)\pi_T + \mathbb{P}(Z_{T+1} = 0|H_1)(1 - \pi_T)} \\
&\quad = \frac{(1 - p)\pi_T}{(1 - p)\pi_T + (1 - \pi_T)} \\
&\quad = \frac{(1 - p)\pi_T}{1 - p\pi_T}
\end{aligned}$$

Returning to (34), we need to show that

$$\pi_T \leq \lambda + (1 - p\pi_T)\frac{(1 - p)\pi_T}{1 - p\pi_T} = \lambda + (1 - p)\pi_T$$

Simplifying the above inequality, we have that

$$\pi_T \leq \frac{\lambda}{p}$$

Substituting in our formula for π_T, the above is equivalent to

$$\frac{(1 - p)^\top}{(1 - p)^\top + 1} \leq \frac{\lambda}{p}$$

which after simplification is equivalent to

$$(1-p)^{\top} \leq \frac{\frac{\lambda}{p}}{1-\frac{\lambda}{p}} = \frac{\lambda}{p-\lambda}$$

Let F be the CDF of the Geometric(p) distribution. Let T^* be the $1-\frac{\lambda}{p-\lambda}$ percentile of the Geometric(p) distribution. Note, since $\lambda < \frac{p}{2}$ then $1-\frac{\lambda}{p-\lambda} \in (0,1)$ and is a valid percentile. Then for any $T \geq T^*$ we have that

$$\begin{aligned} 1-(1-p)^{\top} &= F(T) \\ &\geq F(T^*) && \text{by } T \geq T^* \\ &\geq 1-\frac{\lambda}{p-\lambda} && \text{by choice of } T^* \end{aligned}$$

which is equivalent to

$$(1-p)^{\top} \leq \frac{\lambda}{p-\lambda}$$

and therefore, the optimal hypothesis test is to declare $X_T = 1$ for any $T \geq T^*$ if $Z_1 = \ldots = Z_T = 0$. Hence, the optimal hypothesis takes the form of

$$X_\tau = \mathbf{1}\{Z_\tau = 0\} \text{ where } \tau = \min(s : Z_s = 1) \wedge T^*$$

□

Proposition D.4 (Restatement of Proposition D.2)**.** The ratio of total error rate with a fixed threshold r^* to that under SNR scales like

$$\Omega\left(\exp\left(\log(\alpha(1-\alpha)^{-1})\left(-\frac{1}{2}+\frac{1}{2}\frac{\log(1-p_1)}{\log(1-p_2)}\right)\right)\right)$$

Proof. For notational convenience, define $\bar{\alpha} = \alpha(1-\alpha)^{-1}$. Notice that by Proposition C.4, under SNR, neuron i, $(I = 1, 2)$, is reset if it inactive for any longer that time $\bar{T} = \log(\bar{\alpha}))/\log(1-p_i)$. Consequently, the expected delay penalty, $\mathbb{E}[\tau|H_0] + \mathbb{E}[\tau|H_1]$ for neuron i, is simply

$$\frac{\log(\bar{\alpha})}{\log(1-p_i)} + \frac{1}{p_i}(1-\bar{\alpha})$$

Letting the optimal fixed threshold be r^*, we must have that the expected total delay across both neurons under this fixed threshold is at least $2r^*$. This total expected delay can be no larger than that under SNR. Thus,

$$2r^* \leq \log(\bar{\alpha})\left(\frac{1}{\log(1-p_1)}+\frac{1}{\log(1-p_2)}\right)+(1-\bar{\alpha})\left(\frac{1}{p_1}+\frac{1}{p_2}\right)$$

But the sum of the error rates across the two neurons with the fixed threshold r^* is at least $(1-p_1)^{r^*}+\bar{\alpha}$, while the total error rate under SNR is precisely $\bar{\alpha}$. Dividing these two quantities and employing the upper bound derived on r^* then yields the result. □