Length Independent PAC-Bayes Bounds for Simple RNNs

Volodimir Mitarchuk, Clara Lacroce, Rémi Eyraud, Rémi Emonet, Amaury Habrard, Guillaume Rabusseau

Length Independent PAC-Bayes Bounds for Simple RNNs

Volodimir Mitarchuk
UJM-Saint-Etienne-LabHC[1]

Clara Lacroce
McGill University,
Mila, Montréal, Canada

Rémi Eyraud
UJM-Saint-Etienne-LabHC[1]

Rémi Emonet
UJM-Saint-Etienne-LabHC[1]
Institut Universitaire de France

Amaury Habrard
UJM-Saint-Etienne-LabHC[1]
Institut Universitaire de France

Guillaume Rabusseau
Université de Montréal, Mila,
DIRO, CIFAR AI Chair

Abstract

While the practical interest of Recurrent Neural Networks (RNNs) is attested, much remains to be done to develop a thorough theoretical understanding of their abilities, particularly in what concerns their learning capacities. A powerful framework to tackle this question is the one of PAC-Bayes theory, which allows one to derive bounds providing guarantees on the expected performance of learning models on unseen data. In this paper, we provide an extensive study on the conditions leading to PAC-Bayes bounds for non-linear RNNs that are independent of the length of the data. The derivation of our results relies on a perturbation analysis on the weights of the network. We prove bounds that hold for *β-saturated* and *DS β-saturated* SRNs, classes of RNNs we introduce to formalize saturation regimes of RNNs. The first regime corresponds to the case where the values of the hidden state of the SRN are always close to the boundaries of the activation functions. The second one, closely related to practical observations, only requires that it happens at least once in each component of the hidden state on a sliding window of a given size.

1 INTRODUCTION

In recent years the increase in computational power allowed neural networks to achieve above human-level success in a variety of tasks. Among these models, Recurrent Neural Networks (RNNs) hold a special place. RNNs process sequential inputs recursively by using the same set of parameters to update their internal state after reading each element in an input sequence. Their recursive nature is at the core of the struggle to understand their behavior and generalization abilities.

RNNs expressiveness has been studied from various perspectives. It is known that RNNs with unbounded precision and computation time are Turing complete (Siegelmann and Sontag, 1992). This is even the case with bounded precision and growing memory (Chung and Siegelmann, 2021). Connections between RNNs and classical models have also been investigated, from finite automata (Weiss et al., 2018; Eyraud and Ayache, 2021; Li et al., 2022) to more complex models (Merrill et al., 2020; Delétang et al., 2023; Hao et al., 2022).

While these results are of crucial interest, they focus on the computational expressiveness of RNNs without investigating their learning abilities. In this regard, generalization bounds for RNNs have been derived within the PAC-learning formalism. Chen et al. (2020) derived generalz bounds depending on the spectral norms of weight matrices and the total number of parameters. Tu et al. (2020), Wang et al. (2021), and Panigrahi and Goyal (2021) have provided bounds for specifically defined learnable concept classes. Zhang et al. (2018) and Allen-Zhu and Li (2019) study the links between gradient descent and provable generalization. The main caveat of these bounds is that they depend on a arbitrary fixed maximal length of sequences.

In this work, we show how this caveat can be overcome for Simple RNNs (SRNs). Our result is based on the

[1]Université Jean Monnet Saint-Etienne, CNRS, Institut d'Optique Graduate School, Inria, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

PAC-Bayesian theory (McAllester, 2003; Guedj, 2019) which has recently proved useful to derive theoretical guarantees on the generalization ability of feed forward neural networks (Neyshabur et al., 2018; Jiang et al., 2019; Dziugaite et al., 2020).

Most of our results require the SRN to be at least partially saturated, a regime in which a RNN has its squashing activation functions fed with inputs large enough to approach their boundaries. The saturation phenomenon in RNNs has been observed empirically, particularly in Natural Language Processing (Shibata et al., 2020), and we provide further evidence with a study of the TAYSIR benchmark models (Eyraud et al., 2023). Saturation was also the subject of a theoretical study: Merrill et al. (2020) consider the more restrictive case where the limit of the squashing functions is reached, forcing the hidden states to behave binarily. This hypothesis allows the authors to establish results on the expressiveness of different RNN architectures.

Our analysis relies on bounding the stability of RNNs with respect to perturbed parameters and combining this bound with tools from PAC-Bayesian theory, extending the framework introduced by Neyshabur et al. (2018). Indeed, while this framework requires the perturbation of every parameters of the SRN, we prove that adding a noise to the bias is sufficient. The idea behind this result is that a sufficiently strong bias noise can cover the noise generated by the perturbation of all the parameters. This simplifies the proofs with the cost of loosening the derived bounds.

Using this new framework, we first demonstrate a general result: if the norm of the difference between the hidden states of a SRN and those its bias-perturbed counter part can be bounded in some way, then a length-independent PAC Bayes result can be obtained. We first apply this theorem to *stable* SRNs, which are models whose recurrent weight matrix W has a spectral norm lower than 1. This non-trivial result is clear intuitively: the update map of a stable SRN being contractive, input perturbations are not magnified through its multiple applications and can thus be controlled over sequences of arbitrary length.

We then pioneer the question of non stable RNNs by showing that the non-contractiveness of the update map can be compensated by the saturating effect of the squashing recurrent activation function. This yields to generalization bounds, without any restriction on the sequence length, for SRNs that are in a saturation regime, a notion we formalize in two ways in this work: The first where the limit of the activation function is almost reached at each time step, the second where this happens at least once on a sliding window.

After introducing the main definitions and notations in Section 2, we detail our main results in Section 3 and provide the sketch of their proofs in Section 4. Section 5 provides experimental evidence of saturation while Section 6 discusses the results and the related works. Section 7 concludes.

2 PRELIMINARIES

We introduce first our main notations, then Simple Recurrent Networks (SRNs), saturation regimes and the subclasses of SRNs under consideration. The supervised classification setting considered is then presented before an introduction to the PAC-Bayesian framework

2.1 Notations

We denote matrices and vectors by uppercase and lowercase letters, respectively, and we use $x[j]$ to indicate the j^{th} coordinate of a vector x. We denote by I_d the identity matrix of size $d \times d$. Given $v \in \mathbb{R}^d$, we denote by $\|v\|$ its Euclidean norm, and by $\|v\|_\infty$ its infinity norm. Given a matrix M, its spectral norm is denoted by $\|M\|$, and its Frobenius norm by $\|M\|_{\text{Fro}}$. Given an univariate function σ, we denote by σ' its derivative. Given two integers k, s, with $k \leq s$, we denote by $[\![k, s]\!]$ the set of integers within the range of k and s. When k and s are real numbers, we use $[k, s]$ to denote a closed interval and $]k, s[$ for an open interval.

2.2 Simple Recurrent Networks

Simple Recurrent Networks (SRNs) are a class of models designed to process sequential data (Elman, 1990). An input sequence of length T, is denoted $X^T := \{x_k\}_{k=1}^T$ with $x_k \in \mathbb{R}^u$. Let $h_0 \in \mathbb{R}^d$ be the initial hidden state. The computation of a SRN is defined recursively by the following equations:

$$\begin{cases} h_k = \sigma\left(Ux_k + Wh_{k-1} + b\right) \\ y_k = \sigma\left(Vh_k + c\right) \end{cases} \tag{1}$$

where $U \in \mathbb{R}^{d\times u}$, $W \in \mathbb{R}^{d\times d}$, $V \in \mathbb{R}^{o\times d}$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}^o$, and where we denote by σ the activation function and by $y_k \in \mathbb{R}^o$ the SRN output. We denote a SRN by $\mathcal{R}_P$, where $P = \mathbf{vec}(U, W, V, b, c)$ is the vector containing all the parameters of the SRN.

In this article, we focus on SRNs trained for classification tasks with $\sigma = \tanh$. Our results can, with some adaptation, be extended to the sigmoidal function. For a given sequence X^T, the output of $\mathcal{R}_P(X^T)$ is the last output vector y_T, i.e. $\mathcal{R}_P(X^T) = y_T$. Let $\mathcal{X} = \{X^T : T > 0\}$ be the set of all finite sequences that the SRN $\mathcal{R}_P$ can process. We assume[1] that there ex-

[1]Note that this assumption is trivial in the case of one-hot encoding, with $B = 1$.

ists a constant B such that for all $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$, $\|x_k\| \leq B$ for all $1 \leq k \leq T$.

2.2.1 β-saturation

We first formalize the saturation regime of SRNs.

Definition 2.1 (β-**Saturated SRN**). *A SRN $\mathcal{R}_P$ is* β-saturated *if there exists β, $0 < \beta \leq 1$, such that for all $X^T \in \mathcal{X}$ and all $1 \leq k \leq T$,*

$$\min_{1 \leq j \leq d} |h_k[j]| > \beta.$$

Note that when $\beta = 1$, our definition coincide with the one of Merrill et al. (2020). We then introduce a relaxation of the β-saturation constraint as follows:

Definition 2.2 (**Desynchronised Sliding β-saturation** (DS β-saturation)). *A SRN $\mathcal{R}_P$ is* DS β-saturated *if there exist $\beta > 0$ and an integer $F \geq 1$ called the window size, such that for all $X^T \in \mathcal{X}$ with $T \geq F$, and for all $1 \leq k \leq T - F$, for all $1 \leq j \leq d$:*

$$\exists\, s \in [\![k, k + F - 1]\!] \text{ such that } |h_s[j]| > \beta.$$

This defines a SRN that may not saturate at every iteration, but every coordinate will saturate at least once within a sliding window of size F. Note that a DS β-saturated SRN is a β-saturated SRN when $F = 1$.

The work presented here being of probabilistic nature, considered SRNs do not have to be β-saturated or DS β-saturated on every sequences of $\mathcal{X}$: it is required to happen on a set $\mathcal{X}_\tau \subset \mathcal{X}$ whose measure is $1 - \tau$.

We now provide an intuition about these saturation regimes from a geometrical standpoint. First, because of the tanh activation function, all the hidden state vectors lie in $[-1, 1]^d$. Let $\mathcal{R}_P$ be a β-saturated SRN for some $0 \leq \beta \leq 1$. By definition, every coordinate of each hidden state vector is greater than β in absolute value. For $\beta = 1$, all the hidden state vectors live in $\{-1, 1\}^d$. This set corresponds to the vertices of the hypercube $[-1, 1]^d$. For β smaller than 1, the hidden state vector will land next to a vertex and within a infinite norm ball of radius $1 - \beta$.

The dynamics of a DS β-saturated SRN can be intuitively represented as: within every window of size F the hidden state vector visit the neighborhood of a boundary in every dimension of the hypercube $[-1, 1]^d$ without necessarily landing near the vertices.

2.2.2 Classes of SRNs

We first give a formal definition of stable SRNs, consistent with that of Miller and Hardt (2019).

Definition 2.3 (**Stable SRN**). *Let $\mathcal{R}_P$ be a SRN as in Equation 1, we say that $\mathcal{R}_P$ is stable if $\|W\| < 1$.*

Now, we introduce the notion of perturbed SRN, which is obtained by perturbing the parameters of a SRN by a given noise vector ϑ.

Definition 2.4 (**Perturbed SRN**). *Given a SRN $\mathcal{R}_P$ as in Equation 1, and $\vartheta \in \mathbb{R}^{dim(P)}$, we define the perturbed SRN $\mathcal{R}_{P+\vartheta}$ by the following set of equations:*

$$\begin{cases} \mathring{h}_k = \sigma\Big((U + \vartheta_U)x_k + (W + \vartheta_W)\mathring{h}_{k-1} + b + \vartheta_b\Big) \\ \mathring{y}_k = \sigma\Big((V + \vartheta_V)\mathring{h}_k + c + \vartheta_c\Big) \end{cases},$$

where $\vartheta = \mathbf{vec}(\vartheta_U, \vartheta_W, \vartheta_V, \vartheta_b, \vartheta_c)$, and $\mathring{h}_0 = h_0$.

Perturbed SRNs are central to our analysis. We will show that robustness to perturbations can be translated to generalization guarantees by leveraging tools from the PAC-Bayesian theory. A particular case is the one where the matrices ϑ_U, ϑ_W and ϑ_V are zero matrices:

Definition 2.5 (**Fuzzy SRN**). *Let $\mathcal{R}_P$ be a SRN and $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$. $\mathcal{R}_{P+\vartheta^\varepsilon}$ is a fuzzy SRN if $\mathbb{R}^{dim(P)} \ni \vartheta^\varepsilon = \mathbf{vec}(\vartheta_U^\varepsilon, \vartheta_W^\varepsilon, \vartheta_V^\varepsilon, \vartheta_b^\varepsilon, \vartheta_c^\varepsilon) = \mathbf{vec}(0, 0, 0, \varepsilon_d, \varepsilon_o)$.*

In order to have a clear distinction between perturbed and fuzzy SRNs, we denote a fuzzy SRN by $\mathcal{R}_P^\varepsilon$, and refer to its hidden and output vectors with $\tilde{h}_k$ and $\tilde{y}_k$.

2.3 Supervised Classification Setting

We consider supervised classification tasks where the input space is the set of all finite sequences $\mathcal{X}$ and the output space is a set of discrete labels $\mathcal{Y} = \{1, \ldots, o\}$. Let D be a fixed unknown distribution over $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{Z}_m$ be a learning sample of size m identically and independently drawn (*i.i.d.*) from D. Lastly, $\mathcal{H}$ is a set of classifiers $f_P : \mathcal{X} \to \mathcal{Y}$ parameterized by a vector P. In a usual deep learning setting, $f_P = g \circ \mathrm{f}_P$ where $\mathrm{f}_P : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ and $g(x) = \max_{1 \leq i \leq |\mathcal{Y}|} x[i]$. Since g is generic and has no influence on the decision, we allow ourselves an abuse of notation by considering f_P as being f_P.

We now define the expected margin loss that will be used to assess the quality of a (learned) classifier.

Definition 2.6 (Expected margin loss). *Let $\gamma \geq 0$ a margin. The expected margin loss is defined by:*

$$L_\gamma(f_P) = \mathbb{P}_{(X^T, y) \sim \mathcal{D}}\Big[f_P(X^T)[y] - \max_{j \neq y} f_P(X^T)[j] \leq \gamma\Big].$$

The margin loss can be seen as a measurement of the classifier's strength toward to perturbations. Its empirical counterpart is classically defined as follows:

Definition 2.7 (Empirical margin loss)**.** *Let f_P be a classifier on $\mathcal{X} \times \mathcal{Y}$ and $\gamma > 0$. The empirical margin loss is:*

$$\widehat{L}_\gamma(f_P) = \frac{1}{m} \sum_{(X^T,y)\in\mathcal{Z}_m} \mathbf{1}_{\{f_P(X^T)[y] - \max_{j\neq y} f_P(X^T)[j] \leq \gamma\}}$$

where $\mathbf{1}$ is the indicator function.

Note that if $\gamma = 0$, $L_0(f_P)$ is the expected zero-one loss and $\widehat{L}_0(f_P)$ is the empirical zero-one loss.

2.4 PAC-Bayesian Framework

The Probably Approximately Correct (PAC) learning framework (Valiant, 1984) allows one to evaluate the quality of a hypothesis (model) by assessing to which extent the empirical loss can provide a good approximation of the expected loss with high probability. The PAC-Bayesian theory (McAllester, 2003; Alquier, 2021) is a framework to obtain generalization guarantees over a random prediction instead of a single predictor as in the classic PAC setting. In the PAC-Bayes framework, one considers a distribution $\mathcal{Q}$ over $\mathcal{H}$ which is learned from data. The performance of randomized predictors drawn from $\mathcal{Q}$ is then analyzed with respect to a prior distribution π over $\mathcal{H}$ that is independent of the training data. In order to get a bound on a single hypothesis f_P, one needs to link the expected loss over $\mathcal{Q}$ with the loss of f_P. For this purpose, we introduce the following lemma.

Lemma 2.8. *Let $\mathcal{H}$ be a set of classifiers on $\mathcal{X} \times \mathcal{Y}$. Let $\pi, \mathcal{Q}$ two distributions on $\mathcal{H}$. Let $\mathcal{Z}_m \subset \mathcal{X} \times \mathcal{Y}$ a set of m training samples assumed to be drawn iid from an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{D}_\mathcal{X}$ the marginal over $\mathcal{X}$. The distribution π is the prior and is assumed to be independent of $\mathcal{Z}_m$. Let $f_P \in \mathcal{H}$ drawn with respect to the distribution $\mathcal{Q}$. Let $\mathcal{X}_\tau \subset \mathcal{X}$ a subset of $\mathcal{X}$ such that $\mathcal{D}_\mathcal{X}(\mathcal{X}_\tau) = 1 - \tau$ for $\tau \in [0, 1[$, if :*

$$\mathbb{P}_{J\sim\mathcal{Q}}\left[\sup_{X^T\in\mathcal{X}_\tau} \left\|f_P(X^T) - f_J(X^T)\right\|_\infty < \gamma/4\right] \geq \frac{1}{2},$$

then, with probability $1 - \delta$ over $\mathcal{Z}_m$, we have

$$L_0(f_P) \leq \widehat{L}_\gamma(f_P) + \tau + 4\sqrt{\frac{2KL(\mathcal{Q}\|\pi) + \ln\left(\frac{6m}{\delta}\right)}{m-1}},$$

where $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

The proof is in the Appendix. This lemma is a relaxation of the one of Neyshabur et al. (2018, Lemma 1). While in the latter, one has to control the distance – between a fixed classifier and one randomly sampled using $\mathcal{Q}$ – over all possible data, in Lemma 2.8 it is sufficient to control this distance with probability $1-\tau$.

3 MAIN RESULTS

3.1 PAC-Bayesian Bounds

To obtain PAC-Bayesian bounds, we follow the principle of Neyshabur et al. (2018) by analyzing how perturbations of the weights affect the outputs of a network. In particular, we study the distance between a SRN $\mathcal{R}_P$ and the perturbed SRN $\mathcal{R}_{P+\vartheta}$. We define the prior $\pi = \mathcal{N}(0, \rho^2 I_{dim(P)})$ and the posterior $\mathcal{Q} = \mathcal{N}(P, \rho^2 I_{dim(P)})$ for $\rho > 0$. We assumed that the network parameters are initialized according to the prior distribution π. Note that the posterior is equivalent to the distribution $P + \vartheta$ with $\vartheta \sim \mathcal{N}(0, \rho^2 I_{dim(P)})$. In this context, it is well known that:

$$KL(\mathcal{Q}\|\pi) = KL(P + \vartheta\|\pi) = \frac{\|P\|^2_{\text{Fro}}}{2\rho^2}.$$

We now present a general result that will allow us to derive several generalization bounds for SRNs. It uses the fact that it is not necessary to perturb all the parameters of the SRN but only the bias. Combined with Lemma 2.8, we obtain a particularly convenient tool for studying the generalization power of SRNs independently of the data length.

Theorem 3.1 (Backbone theorem)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}_\mathcal{X}$ the marginal over $\mathcal{X}$. Let $\mathcal{E} > 0$, $\mathcal{R}_P$ be a SRN and $\gamma \geq 0$ a margin, such that there exists $\mathcal{C} > 0$ and $0 < \tau < 1$ verifying for all $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$, $\|\varepsilon\| \leq \mathcal{E}$:*

$$\mathbb{P}_{X^T\sim\mathcal{D}_\mathcal{X}}\left[\forall\ 1 \leq k \leq T, \left\|\tilde{h}_k - h_k\right\| \leq \mathcal{C}\|\varepsilon_d\| + \alpha\right] \geq 1 - \tau$$

where $\tilde{h}_k$ is the hidden state vector of the fuzzy SRN $\mathcal{R}^\varepsilon_P$, α is constant such that $4\|V\|\alpha < \min(\gamma, 4\mathcal{E}\|V\|(\mathcal{C}+1))$ and $X^T = \{x_k\}_{k=1}^T$. Then we have the following PAC-Bayes bound with probability at least $1 - \delta$ over the training sample Z_m:

$$L_0(\mathcal{R}_P) - \widehat{L}_\gamma(\mathcal{R}_P) \leq \tau + \tilde{\mathcal{O}}\left(\frac{D\mathcal{C}B\|V\|\|P\|_{\text{Fro}} + \ln(\frac{1}{\delta})}{(\overline{\gamma} - \alpha\|V\|)\sqrt{m}}\right)$$

with $D = \max(d, o)$, $\overline{\gamma} = \min(\gamma, 4\mathcal{E}\|V\|(\mathcal{C}+1))$ and B such that for all k, $1 \leq k \leq T$, $\|x_k\| \leq B$.

The detailed proof is given in Appendix and a sketch is presented in Section 4. The first step of the proof consists in bounding the distance between the outputs of $\mathcal{R}_P$ and $\mathcal{R}^\varepsilon_P$. We show in particular that:

$$\left\|\mathcal{R}_P(X^T) - \mathcal{R}^\varepsilon_P(X^T)\right\| \leq \|V\|\|\tilde{h}_T - h_T\| + \|\varepsilon_o\|. \quad (2)$$

It is then obvious that bounding $\|\tilde{h}_T - h_T\|$ is crucial. In order to apply Theorem 3.1, it is important to exhibit reasonable constants C and α. Depending on the type

of SRNs considered, we will derive different values for $\mathcal{C}$ and α. It is crucial to note that for SRNs exhibiting a form of saturation, the constants C and α not only act as guarantors for the application of Theorem 3.1, but also guarantee that saturation is maintained, which allows us to apply Lemma 2.8. To conclude this section, we notice that Theorem 3.1 relies on the use of the fuzzy SRN $\mathcal{R}_P^\varepsilon$: this implies that perturbing only the biases is sufficient to obtain PAC-Bayesian bounds.

3.2 PAC-Bayes Bound for Stable SRNs

We first present a length independent PAC-Bayes bound for stable SRNs.

Theorem 3.2 (Stable SRN). *Let $\mathcal{R}_P$ be a stable SRN (see Definition 2.3). Then for all $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ and all $X^T \in \mathcal{X}$ we have:*

$$\left\|\tilde{h}_T - h_T\right\| \leq (1 - \|W\|)^{-1} \|\varepsilon_d\|.$$

We then can apply Theorem 3.1 for stable SRNs with constants $\mathcal{C} = (1 - \|W\|)^{-1}$, $\alpha = 0$, and unbounded $\mathcal{E}$.

We highlight below the key points of the proof (the full proof can be found in the appendix).

First, by leveraging the fact that the tanh activation function σ is 1-Lipschitz, we can derive the bound:

$$\|\tilde{h}_T - h_T\| \leq \|W\| \|\tilde{h}_{T-1} - h_{T-1}\| + \|\varepsilon_d\|.$$

By applying this argument recursively on T, we obtain:

$$\|\tilde{h}_T - h_T\| \leq \left(\sum_{s=0}^{T-1} \|W\|^s \right) \|\varepsilon_d\|. \tag{3}$$

Now, the SRN is stable, so $\|W\| < 1$. Thus, for $T \to \infty$ the series $\|\varepsilon_d\| \sum_{s=0}^{T-1} \|W\|^s$ converges to $\|\varepsilon_d\| \frac{1}{1-\|W\|}$.

The fundamental step to determine $\mathcal{C}$ in the above derivation is the use of the Lipschitz property. In fact, for all $x \in \mathbb{R}^u$ and all $h, h' \in \mathbb{R}^d$ we have:

$$\begin{aligned} &\|\sigma(Ux + Wh' + b) - \sigma(Ux + Wh + b)\| \\ &\leq \|W\| \|h' - h\| < \|h' - h\|. \end{aligned} \tag{4}$$

In particular, $\mathcal{R}_P$ has Lipschitz constant strictly smaller than 1 with respect to the hidden state vectors. We refer to this property as *global contractiveness*, corresponding to the defining property of stable SRNs according to Miller and Hardt (2019). In the next section, we show how the requirement of global contractiveness can be relaxed by leveraging the notions of β-saturation and DS β-saturation. Specifically, these properties provide us with additional information on the dynamic of the hidden state vectors, allowing us to apply Theorem 3.1 to a setting that is only *locally contractive*, that is, where there exists an open sets of hidden states vectors for which Equation 4 holds.

3.3 PAC-Bayes Bound for β-saturated and DS β-saturated SRNs

When the considered SRN is not stable, we find that the effect of the spectral norm of W on the bound can be compensated by a saturation regime. We first state a length independent PAC-Bayes bound for β-saturation.

Theorem 3.3 (β-saturated SRN). *Let $\mathcal{R}_P$ be a β-saturated SRN with $\|W\| \geq 1$ with $\eta = \beta - z > 0$, and $z = \sqrt{1 - \frac{1}{\|W\|}}$. Let $t \in]0, 1[$, $\Delta = \tanh'(\tanh^{-1}(z) + \frac{t\eta}{1-z^2})$ and $\nabla = \frac{(1-t)\eta}{1-z^2}$, then for any $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\|\varepsilon\| \leq \nabla(1 - \Delta\|W\|)$ we have:*

$$\left\|\tilde{h}_T - h_T\right\| \leq \frac{1}{1 - \Delta\|W\|} \|\varepsilon_d\|.$$

This result allows us to apply Theorem 3.1 to β-saturated SRN with $\mathcal{C} = (1 - \Delta\|W\|)^{-1}$, $\alpha = 0$, and $\mathcal{E} = \nabla(1 - \Delta\|W\|)$.

To better understand the regime in which this theorem is relevant, we provide intuitive explanation on the different variables below and a visualization in Figure 1.

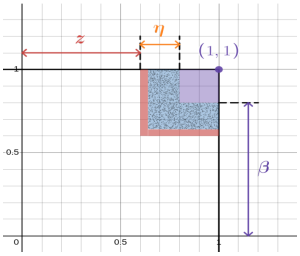

Figure 1: Detail of a corner of the hypercube $[-1, 1]^2$ with the main variables of the theorems.

We consider the setting of Theorem 3.3. Since $\|W\| \geq 1$, the SRN is not contractive everywhere, but if β is large enough, the SRN will be locally contractive in a neighborhood of the hypercube vertices, whose size depends on z. More precisely, the SRN is locally contractive in the ℓ_∞ ball of radius $\eta = \beta - z$ centered on any h in the violet region of Figure 1. In this 2D visualization, the violet square represents a region where the hidden state vectors of the SRN are guaranteed to land due to β-saturation. Figure 1 zooms on one vertex but similar figures can be drawn for each vertex of the hypercube. The red borders delimit the regions where the SRN is (locally) contractive.

Intuitively, the proof of Theorem 3.3 relies on the fact that $\|h_k - \tilde{h}_k\| \propto \|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\|$ can be bounded *independently of* T if both $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$ are locally contractive. First consider the case where noise would be injected in the SRN at only one time step: $\tilde{h}_k = h_k + N$. One can show that if $\|N\|_\infty < \eta/(1 - z^2)$, then $\tilde{h}_{k+1}$ is guaranteed to still lie in the contractive

region (details are provided in Section 4). Thus, in this case, by successively applying the contractive mapping, the noise injected at iteration k will be exponentially reduced and will not affect the local contractiveness of the following iterations. However, this only holds if the noise is injected once during the execution. In the case where a constant noise is injected at each iteration, the accumulation of the noise terms is likely to drive the fuzzy hidden state outside the locally contractive area.

A key element is that for small enough noise injected at each iteration, the fuzzy SRN remains locally contractive *at each iteration*. This constraint on the magnitude of the noise is captured by the definitions of ∇ and Δ.

From a high-level perspective, $\eta/(1-z^2)$ represents a noise budget that can be divided into two parts. This is illustrated in Figure 1. The blue pigmented zone represents the region where local contractiveness is guaranteed at each step when noise is added at each iteration. The width of this region is directly controlled by the variable $\nabla = \frac{(1-t)\eta}{1-z^2}$. As for Δ, it corresponds to a strict lower bound on $\|W\|^{-1}$ needed to ensure that all fuzzy SRN hidden states remain in the blue region. Section 4.2 provides detailed technical explanations of how these variables are obtained and used.

In the following, we show that local constractiveness is not mandatory for having a SRN robust to noise injection. DS β-saturation is a relaxation of the β-saturation in two ways: 1) β-saturation does not need to occur at every iteration, 2) not all hidden vector coordinates have to be β-saturated simultaneously. The hypothesis of DS β-saturation is sufficiently structuring to derive guarantees, and closely related to observed phenomena is practical RNNs. We discuss this last point further in the Section 5.

Theorem 3.4 (DS β-saturated SRN)**.** *Let $\mathcal{R}_P$ be a DS β-saturated SRN with $\|W\|_\infty \geq 1$, $F \geq 1$ the window size verifying $\eta = \beta - z > 0$, where $z = \sqrt{1 - \frac{1}{2\|W\|_\infty^F}}$. For any $t \in]0,1[$, we set $\Delta = \tanh'(\tanh^{-1}(z) + \frac{t\eta}{1-z^2})$ and $\nabla = \frac{(1-t)\eta}{1-z^2}$, then for any $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\|\varepsilon\| \leq \left(\frac{\nabla\Delta}{4\sum_{j=0}^{F-1}\|W\|_\infty^j}\right)$ we have:*

$$\left\|\tilde{h}_T - h_T\right\|_\infty \leq \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j + \frac{\nabla\Delta}{4}.$$

DS β-saturation cannot be interpreted as local contractiveness. From a high level perspective, if one observes a single coordinate of a DS β-saturated SRN they will find that in absolute value this coordinate will frequently go above the threshold β. The idea inherited from β-saturation is that when a neuron reaches a certain degree of saturation, it compresses the noise it might have received from the previous layer. We show that desynchronized, sparse but regular β-saturation is sufficient to contain the accumulated noise and thus extract generalization guarantees. The quantities η, ∇, Δ and $t \in]0,1[$ have a similar role to that of the β-saturation regime: they represent the partitioning of the budget available for noise accumulation across iterations.

Theorems 3.3 [resp. 3.4] is easily adaptable to the case where a SRN $\mathcal{R}_P$ is β-saturated [resp. DS β-saturated respectively] only on a subset $\mathcal{X}_\tau \subset \mathcal{X}$. Indeed, if $\mathcal{R}_P$ is β-saturated on $\mathcal{X}_\tau$ with $\mathcal{D}_\mathcal{X}(\mathcal{X}_\tau) = 1 - \tau$, this means that the results of these theorems hold with probability greater than or equal to $1 - \tau$. Theorem 3.1 can thus be used to prove the bounds.

4 SKETCH OF PROOFS

In this section we sketch the proof of Theorem 3.1 and the derivations of the constants $\mathcal{C}$ and α of the other theorems. We also provide a sketch showing why the perturbation of the biases is enough to obtain PAC-Bayes bounds. The detailed proofs are in the appendix.

4.1 Backbone Theorem

Let $\mathcal{R}_P$ be a SRN. First, we prove that it is sufficient to study the distance between $\mathcal{R}_P$ and a fuzzy SRN $\mathcal{R}_P^\varepsilon$, where $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$. Leveraging Equation 2 and the hypothesis of Theorem 3.1, we show that for an appropriately chosen ρ, and for $\varepsilon \sim \mathcal{N}(0, \rho^2 I)$, we have with high probability:

$$\begin{aligned}\left\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\right\| &\leq \|V\|\mathcal{C}\|\varepsilon_d\| + \|\varepsilon_o\| \\ &\leq \|V\|(\mathcal{C}+1)\|\varepsilon_{\max(o,d)}\|.\end{aligned}$$

Then, it is possible to set the variance of the noise ε such that the following inequality holds:

$$\mathbb{P}\left[\|V\|(\mathcal{C}+1)\|\varepsilon_{\max(o,d)}\| < \gamma/4\right] \geq \frac{1}{2}.$$

Finally, we can apply Lemma 2.8, from which we derive the bound stated in Theorem 3.1.

4.2 PAC-Bayes Bound for β-saturated SRNs

The sketch of the proof for stable SRNs being given in Section 3.2, we focus on the saturated regimes.

Let $\mathcal{R}_P$ be a β-saturated SRN with $\beta = z + \eta$, for some $\eta > 0$ and $z = \sqrt{1 - \frac{1}{\|W\|}}$. Let $X^T \in \mathcal{X}$ a sequence of length T. As explained previously, β-saturated SRNs are locally contractive around the hidden state vectors. In order to obtain a bound on $\|\tilde{h}_T - h_T\|$ independent from T, the local contractiveness must be preserved in the fuzzy version of $\mathcal{R}_P$ to prevent noise explosion.

We prove that z is a boundary of the local contractiveness region, thus for $\beta = z + \eta$, η represents the budget for accumulating the noise in the post-activation space. The reasoning can then focus only on the choice of $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\mathcal{R}_P^\varepsilon$ is z-saturated.

First, given that β is defined in the post-activation space, we need to transpose it into a pre-activation quantity, since the noise is added before the activation. This is achieved by applying $\tanh^{-1}$ and some basic calculus, leading to the following inequality:

$$\min_{1\leq j\leq d} \tanh^{-1}(|h_k[j]|) > \tanh^{-1}(z) + \frac{\eta}{1-z^2} \quad (5)$$

It gives us the pre-activation budget $\frac{\eta}{1-z^2}$ before loosing the local contractiveness. In order to ensure that the fuzzy SRN remains locally contractive, we use only a fraction of this budget. We call this fraction $\nabla := \frac{(1-t)\eta}{1-z^2}$, for $0 < t < 1$. Now, assume that we can define $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that the noise accumulated by $\mathcal{R}_P^\varepsilon$ never grows bigger than ∇. Concretely, for a given $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$, we assume that $\tilde{h}_k$ produced by $\mathcal{R}_P^\varepsilon$ never deviates from h_k further than ∇: $\tilde{h}_k = h_k + N_k$, $\|N_k\| \leq \nabla$ where N_k is the noise accumulated up to iteration k. From Equation 5, we can deduce:

$$\min_{1\leq j\leq d} \tanh^{-1}(|\tilde{h}_k[j]|) > \tanh^{-1}(z) + \frac{t\eta}{1-z^2} \quad (6)$$

Leveraging Equation 6, we prove that the Lipschitz constant of $\mathcal{R}_P^\varepsilon$ evaluated on the data from $\mathcal{X}$ is bounded by $\Delta\|W\| < 1$ where $\Delta := \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1-z^2}\right)$. From here, we prove in the appendix an expression of the accumulated noise N_T at time T:

$$N_T = \sum_{s=1}^{T-1} \Lambda(c_T) \left(\prod_{l=1}^{T-s} W\Lambda(c_{T-l}) \right) \varepsilon_d + \Lambda(c_T)\varepsilon_d \quad (7)$$

where $1 \leq k \leq T$, c_k is a d-dimensional vector with coordinates $c_k[j]$ taking values between $\tanh^{-1}(h_k[j])$ and $\tanh^{-1}(h_k[j] + N_k[j])$, and $\Lambda(c_k) = \mathrm{Diag}(\tanh'(c_k))$.

From Equations 5 and 6, it follows that all the entries of $\Lambda(c_k)$ are bounded by Δ, and thus $\|\Lambda(c_k)\| < \Delta$. With the triangle inequality we extract a bound on $\|N_T\|$ which takes the form of a convergent geometric series times $\|\varepsilon_d\|$. For $T \to \infty$ we obtain a bound on the maximum amplification of the noise. From this follows a bound on ε_d such that the initial condition $\|N_T\| < \nabla$ is fulfilled for all T: it suffices to choose ε_d such that $\|\varepsilon_d\| < \nabla(1 - \Delta\|W\|)$ where $(1 - \Delta\|W\|)^{-1}$ is the limit of the converging geometric series.

4.3 PAC-Bayes Bound for DS β-saturated SRNs

The DS β-saturation is a relaxation of the β-saturation hypothesis. However, the handling of DS β-saturation needs to be more careful. Indeed, for β-saturation we can unroll the entire recurrence of T iterations and obtain an expression for the vector $\tilde{h}_T - h_T$. It is made possible by the regularity and synchronicity of saturation resulting from the β-saturation assumption. Abandoning synchronicity in the DS β-saturation framework no longer allows us to unroll the entire recurrence, forcing us to approach the problem in a different way. We prove by induction that the coordinates of the vector $\tilde{h}_k - h_k$ do not exceed the threshold set in the theorem with $0 \leq k \leq T - F$. To do this, we first show a bound on the maximal amplification that a coordinate of the vector $\tilde{h}_k - h_k$ can experience in $F - 1$ iterations: $\|\tilde{h}_{k+F-1} - h_{k+F-1}\|_\infty$ is bounded by $\|\varepsilon_d\|_\infty \sum_{j=0}^{F-2} \|W\|_\infty^j + \|W\|_\infty^{F-1}\|\tilde{h}_k - h_k\|_\infty$.

In the induction initialization k is equal to zero, thus $\|W\|_\infty^{F-1}\|\tilde{h}_k - h_k\|_\infty = 0$. The theorem assumption on $\|\varepsilon\|$ allows us to show that the accumulated noise during the $F-1$ iterations is bounded by $\frac{\nabla\Delta}{2}$. This last bound is crucial to: 1) guarantee that $\mathcal{R}_P^\varepsilon$ will have a similar saturation pattern (refer to C.8 for a formal definition) as that of $\mathcal{R}_P$ on the first F iterations, 2) define the induction hypothesis. For the rest of the proof we assume that for $l \geq 1$ we have $\|\tilde{h}_{lF} - h_{lF}\|_\infty \leq \frac{\nabla\Delta}{2}$ and prove that $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$ will have a similar saturation pattern for the next F iterations proving the fulfilment of the induction hypothesis up to iteration $(l+1)F$.

4.4 Sufficiency of Bias Perturbation

The previous sketches of proof have all been based on the fact that it is sufficient to perturb only the bias. On that point, they actually all rely on the same result: in this sub-section we provide a sketch of it. Let $\mathcal{R}_P$ be a SRN, we recall that for $\vartheta \in \mathbb{R}^{dim(P)}$ $\mathcal{R}_{P+\vartheta}$ is a perturbed SRN producing hidden state vectors denoted with $\mathring{h}_k$. In the Appendix C.4 we prove that, in a similar way as Equation 7, it is possible to express $\mathring{h}_k$ as: $\mathring{h}_k = h_k + \mathring{N}_k$, with $\mathring{N}_k$ being defined recursively. Now, we focus on the process of generating the next hidden state $\mathring{h}_{k+1}$. Let $x_{k+1} \in \mathbb{R}^u$ such that $\|x_{k+1}\| \leq B$. By the definition of the perturbed SRN $\mathcal{R}_{P+\vartheta}$ we have:

$$\begin{aligned}
\mathring{h}_{k+1} &= \sigma\left(\mathring{U}x_{k+1} + \mathring{W}(h_k + \mathring{N}_k) + \mathring{b}\right) \\
&= \sigma\left(Ux_{k+1} + Wh_k + b + \mathring{N}_{k+1}\right) \\
&= \sigma\left(Ux_{k+1} + Wh_k + b\right) + \Lambda(c_{k+1})\mathring{N}_{k+1} \\
&= h_{k+1} + \Lambda(c_{k+1})\mathring{N}_{k+1},
\end{aligned}$$

where $\Lambda(c_{k+1})$ is defined similarly as for Equation 7, and $\mathring{U} = (U + \vartheta_U)$, $\mathring{W} = (W + \vartheta_W)$, $\mathring{b} = (b + \vartheta_b)$ $\mathring{N}_{k+1} := \vartheta_U x_{k+1} + W\mathring{N}_k + \vartheta_W(h_k + \mathring{N}_k) + \vartheta_b$.

The total noise added in the production of $\mathring{N}_{k+1}$ on top of the noise $\mathring{N}_k$, inherited from previous iterations,

is $Z := \vartheta_U x_{k+1} + \vartheta_W(h_k + \mathring{N}_k) + \vartheta_b$. We recall that $h_k + \mathring{N}_k = \mathring{h}_k \in [-1,1]^d$, therefore $\|h_k + \mathring{N}_k\| \leq \sqrt{d}$. Hence one can derive the following bound:

$$\begin{aligned}\|Z\| &\leq \|\vartheta_U\|\|x_{k+1}\| + \|\vartheta_W\|\|(h_k + \mathring{N}_k)\| + \|\vartheta_b\| \\ &\leq \|\vartheta_U\|B + \|\vartheta_W\|\sqrt{d} + \|\vartheta_b\| \\ &\leq \|\vartheta\|_{\text{Fro}}B + \|\vartheta\|_{\text{Fro}}\sqrt{d} + \|\vartheta\|_{\text{Fro}} \\ &\leq (B+1+d)\|\vartheta\|_{\text{Fro}}.\end{aligned}$$

Now if we consider a fuzzy SRN $\mathcal{R}_P^\varepsilon$, with $\|\varepsilon\| \leq r$, for some $r > 0$, whose execution deviates little from $\mathcal{R}_P$, then by choosing $\vartheta \in \mathbb{R}^{dim(P)}$ such that $\|\vartheta\| \leq \frac{r}{(B+1+d)}$, we can be sure that $\mathcal{R}_{P+\vartheta}$ will not deviate from $\mathcal{R}_P$ further than $\mathcal{R}_P^\varepsilon$ does.

In this work, the random nature of parameter perturbation is central. However, the fact that we restrict ourselves to Gaussian noise provides us with a way of guaranteeing the amplitude of the noise with arbitrarily high probability. So by adjusting the variance of the Gaussian noise we can control, with a desired probability, the deviation of a fuzzy SRN, and hence the deviation on the perturbed SRN from the SRN $\mathcal{R}_P$ that we are studying.

5 EXPERIMENTAL CLUES FOR SATURATION

The TAYSIR benchmark (Eyraud et al., 2023) provides a set of 10 small RNNs trained on a classification task. We fed the validation set to these models and kept track of the values of their hidden states. We observe a trend in half of them (all details are available in Appendix): in these models, the values tend to be either around 0 or close to the extrema $\{-1, 1\}$.

These models have one clear bias: their initial state h_0 is the null vector. We then retrained the models using the same hyper-parameters, but with values of the initial state randomly chosen in the set $\{-1, 1\}$. The noticed trend in the distribution of hidden values is then particularly significant in half of the models (see Figure 2 for 3 examples).

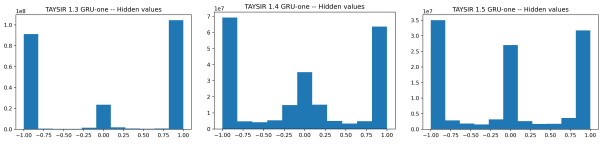

Figure 2: Examples of hidden value distributions of TAYSIR models with values of h_0 chosen in $\{-1, 1\}$.

Because of the values around the origin, these models are not saturated in the sense of our definitions. However, these distributions offer a clear argument in favor of studying saturation: practical RNNs can have the majority of their hidden values closed to the extrema.

These models being GRU Cho et al. (2014), we trained SRNs on the corresponding TAYSIR data. Since keeping the same architecture with SRN neurons instead of GRU gates did not achieved acceptable accuracy, we ran a grid search on the range of hyper-parameters used in TAYSIR. Three of the trained SRNs obtained comparable accuracy to the original models, other tasks seemingly being too complex for SRNs. Among these, two exhibit the noticeable hidden value distribution.

As DS β-saturation relies on a sliding window of size F, we then parsed the sequences one by one and we tracked, in each component of the hidden state, how many consecutive values of the hidden states were less than than a threshold (0.7) in absolute value. We kept the max for each component on each sequence, counted these maxima over all sequences of the validation set, and obtained Figure 3.

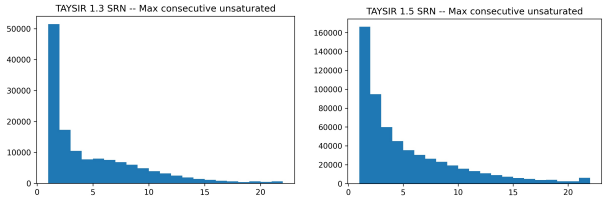

Figure 3: Max Consecutive values below the threshold in SRNs trained on TAYSIR data.

All sequences of these validation sets are of length 22. This figure shows that almost all components of the hidden state have an absolute value close to an extremum at least once every 12 to 15 time steps. Again, this does not exactly correspond to our formalisation of DS β-saturation with a potential sliding window $F = 15$ since, for instance, it may be the case than one component is never greater than the threshold. Nevertheless, it provides observations that ground this saturation regime with the reality of practical RNNs.

6 DISCUSSION AND RELATED WORKS

In this work, our main goal was to study the generalization abilities of RNNs. We first propose a theorem that provides a sufficient condition on the distance between a SRN and its fuzzy version to obtain a length independent PAC-Bayes bound. This general result offers a new potentially impactful framework to study RNNs generalisation ability. We first exemplify its use on stable SRNs , models where the hidden state dynamic is contractive, and obtained a PAC-Bayes bound independent from the length of the data. For

non-contractive SRNs, we derived PAC-Bayes bounds under the additional assumption of saturation. We think that the extension to DS β-saturated SRNs is particularly valuable, as this is arguably a much more realistic setting (Shibata et al., 2020; Chandar et al., 2019; Oliva et al., 2017). Moreover, this is the first bound independent from the length of the data for non-stable RNNs, which is of great interest for studying networks modelling long-term dependencies. Indeed, contractive RNNs can only take decision based on the last elements of the input sequence (Hammer and Tiňo, 2003), while 1-saturated SRNs are equivalent to Deterministic Finite Automata (DFAs) (Merrill et al., 2020; Eyraud and Mitarchuk, 2022), models able to handle unbounded long-term dependencies (Carroll and Long, 1989). The correspondence between these two classes of models adds another dimension to our result: it leads the way towards a first PAC-Bayes bound for DFAs.

Concerning limitations, the main challenge posed by the saturation assumption is that β-saturated SRNs can hardly be learned using gradient descent because of vanishing gradients (Chandar et al., 2019). However, this drawback is not shared with DS β-saturation since gradient at unsaturated time steps does not vanish and since this regime only requires each component to saturated asynchronously once over a sliding window. This is enforced by the experimental observation that well trained RNNs often have hidden state values close to the extrema, as exemplified for instance on the TAYSIR models or the ones of Shibata et al. (2020): training can drive the model to tend to this regime, our result showing its interest for generalisation.

We conclude this section by mentioning some related work on the topic of generalisation bounds for RNNs. The first PAC-Bayesian bound for a class of linear RNNs is proven by Eringis et al. (2021). In Zhang et al. (2018) the authors follow Neyshabur et al. (2018) approach to prove a PAC-Bayes bound for nonlinear RNNs, dependent on the input length. Moreover, the authors propose a way to control the singular values of the transition matrix W to reduce the dependency on the sequence length. Chen et al. (2018, 2020) obtain, for general SRNs, a bound on the empirical Rademacher complexity which is length independent only when $\|W\| < 1$. Tu et al. (2020) introduce a norm for RNNs reflecting the information on the data transmitted by the gradient during training. They use the Rademacher complexity and derive a bound that does not depend directly on the size of the model, but that exhibits a strong dependency on the length of the data. A totally different approach is explored by Wang et al. (2021), Allen-Zhu and Li (2019) and Panigrahi and Goyal (2021). In these works, the authors study the learnability of a class of functions defined by a complexity criterion and provide bounds depending on the length of the data. In particular, Panigrahi and Goyal (2021) define a set of concepts using a complexity measure, and a hypothesis class given by RNNs with a fixed set of hyperparameters (including the length of the string T). They prove that, within this hypothesis class, there is a RNN that will achieve perfect generalization and provide a bound depending on the set of hyperparameters.

7 CONCLUSION

In this paper, we propose PAC-Bayes bounds for a class of nonlinear RNNs working in different regimes. To the best of our knowledge, this is the first extensive theoretical study on generalizations that does not depend on the length of the sequences. We first prove a general theorem that provides a sufficient condition to obtain length independent PAC-Bayes bounds. Then, we show how this result can be used to prove a generalization bound assuming that the spectral norm of the weight matrix W is strictly smaller than one, a case known as stable RNN. Finally, we show that it is possible to obtain bounds also when $\|W\| \geq 1$, under the assumption that the network is in a saturation regime. While in the first setting the update map is contractive, and then perturbations are not magnified through the iterations, it is not the case in the latter, so obtaining length independent bounds is more challenging. This work has led to the development of tools for studying SRNs that are worth highlighting, notably the fact that it is sufficient to perturb only the bias when perturbing the SRN parameters, explained in greater detail in Section 4. Another theoretical contribution is the extension of the notion of saturation. Merrill et al. (2020) define saturation as a limit and therefore does not describe an observable phenomenon; we propose a more empirically realistic formalisation of this notion.

An important element is that our goal in this paper was to propose and study a framework for length independent bounds. We did not try to tighten these bounds as it would have complexified the demonstration. A clear future work is thus to stiffen these bounds.

Another continuation of this work is to extend the result to Deterministic Finite Automata: given their correspondence with saturated RNNs (Eyraud and Mitarchuk, 2022), it seems possible to derive a first PAC-Bayes bound for DFAs (Bengio et al., 1994).

Finally, our results pave the way for the proposal of new learning algorithms: as saturation is proven to be useful for generalisation, a well-designed normalisation of gradient descent based on it, or even a brand new approach, represent likely directions for improvement.

Acknowledgments

We are deeply grateful to the anonymous reviewers for their thoughtful remarks that helped increase the quality of this article.

This work is supported in part by the ANR TAUDoS (ANR-20-CE23-0020). It is supported for another part by the Canadian Institute for Advanced Research (CIFAR AI chair program) and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Allen-Zhu, Z. and Li, Y. (2019). Can SGD learn recurrent neural networks with provable generalization? In *NeurIPS*.

Alquier, P. (2021). User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Carroll, J. and Long, D. (1989). *Theory of Finite Automata with an Introduction to Formal Languages*. Prentice-Hall, Inc., USA.

Chandar, S., Sankar, C., Vorontsov, E., Kahou, S. E., and Bengio, Y. (2019). Towards non-saturating recurrent units for modelling long-term dependencies. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press.

Chen, M., Li, X., and Zhao, T. (2018). On generalization bounds of a family of recurrent neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Chen, M., Li, X., and Zhao, T. (2020). On generalization bounds of a family of recurrent neural networks. In Chiappa, S. and Calandra, R., editors, *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1233–1243. PMLR.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, S. and Siegelmann, H. T. (2021). Turing completeness of bounded-precision recurrent neural networks. In *NeurIPS*.

Cristinel, M. (2010). Sharp inequalities and complete monotonicity for the wallis ratio. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 17(5):929–936.

Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., and Ortega, P. A. (2023). Neural networks and the chomsky hierarchy. In *The Eleventh International Conference on Learning Representations, ICLR*. OpenReview.net.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. (2020). In search of robust measures of generalization. In *NeurIPS*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Eringis, D., Leth, J., Tan, Z., Wisniewski, R., Esfahani, A. F., and Petreczky, M. (2021). Pac-bayesian theory for stochastic LTI systems. In *60th IEEE Conference on Decision and Control (CDC)*, pages 6626–6633. IEEE.

Eyraud, R. and Ayache, S. (2021). Distillation of weighted automata from recurrent neural networks using a spectral approach. *Machine Learning*, pages 1–34.

Eyraud, R., Lambert, D., Tahri Joutei, B., Gaffarov, A., Cabanne, M., Heinz, J., and Shibata, C. (2023). Taysir competition: Transformer+RNN: Algorithms to yield simple and interpretable representations. In Coste, F., Ouardi, F., and Rabusseau, G., editors, *Proceedings of 16th edition of the International Conference on Grammatical Inference*, volume 217, pages 275–290. PMLR.

Eyraud, R. and Mitarchuk, V. (2022). On the limit of gradient decent for simple recurrent neural networks with finite precision. *LearnAut workshop*.

Guedj, B. (2019). A primer on pac-bayesian learning. *CoRR*, abs/1901.05353.

Hammer, B. and Tiňo, P. (2003). Recurrent neural networks with small weights implement definite memory machines. *Neural Computation*, 15(8):1897–1929.

Hao, Y., Angluin, D., and Frank, R. (2022). Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2019). Fantastic generalization measures and where to find them. In *ICLR*.

Li, T., Precup, D., and Rabusseau, G. (2022). Connecting weighted automata, tensor networks and recurrent neural networks through spectral learning. *Machine Learning*, pages 1–35.

McAllester, D. (2003). Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer.

Merrill, W., Weiss, G., Goldberg, Y., Schwartz, R., Smith, N. A., and Yahav, E. (2020). A formal hierarchy of RNN architectures. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 443–459. Association for Computational Linguistics.

Miller, J. and Hardt, M. (2019). Stable recurrent models. In *International Conference on Learning Representations*.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *ICLR*.

Oliva, J. B., Póczos, B., and Schneider, J. (2017). The statistical recurrent unit. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2671–2680. JMLR.org.

Panigrahi, A. and Goyal, N. (2021). Learning and generalization in rnns. In *NeurIPS*.

Shibata, C., Uchiumi, K., and Mochihashi, D. (2020). How LSTM encodes syntax: Exploring context vectors and semi-quantization on natural text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4033–4043. International Committee on Computational Linguistics.

Siegelmann, H. T. and Sontag, E. D. (1992). On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 440–449.

Tu, Z., He, F., and Tao, D. (2020). Understanding generalization in recurrent neural networks. In *ICLR*.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, L., Shen, B., Hu, B., and Cao, X. (2021). On the provable generalization of recurrent neural networks. In *NeurIPS*.

Weiss, G., Goldberg, Y., and Yahav, E. (2018). Extracting automata from recurrent neural networks using queries and counterexamples. In *International Conference on Machine Learning*, pages 5247–5256. PMLR.

Zhang, J., Lei, Q., and Dhillon, I. S. (2018). Stabilizing gradients for deep neural networks via efficient SVD parameterization. In *ICML*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDICES

This supplementary material is organized as follows. Appendix A is dedicated to the presentation of our experimental study. Appendix B presents the extension of the results of Neyshabur et al. (2018). Finally, Appendix C contains the proofs of the main theorems of the paper.

A EXPERIMENTS

To study the practicality of our theoretical work, we use the recent 2023 TAYSIR competition Eyraud et al. (2023) that provides already trained neural net models for a goal of knowledge distillation. Though extraction of surrogate models is not our aim, this benchmark is of interest since it provides a large variety of architectures, including SRNs, trained on various tasks (artificial, bio-informatics, NLP, etc.).

We first looked at the already-trained models by parsing the validation set while keeping track of the hidden state reached. The distribution of values observed in the hidden states components are given in Figure 4. For LSTMs (models 1.6 and 1.11) we report only the values of the h vector though the state of a LSTM model is the concatenation of this vector and the cell one c. Note that model 1.7 is a transformer and thus is not studied here.

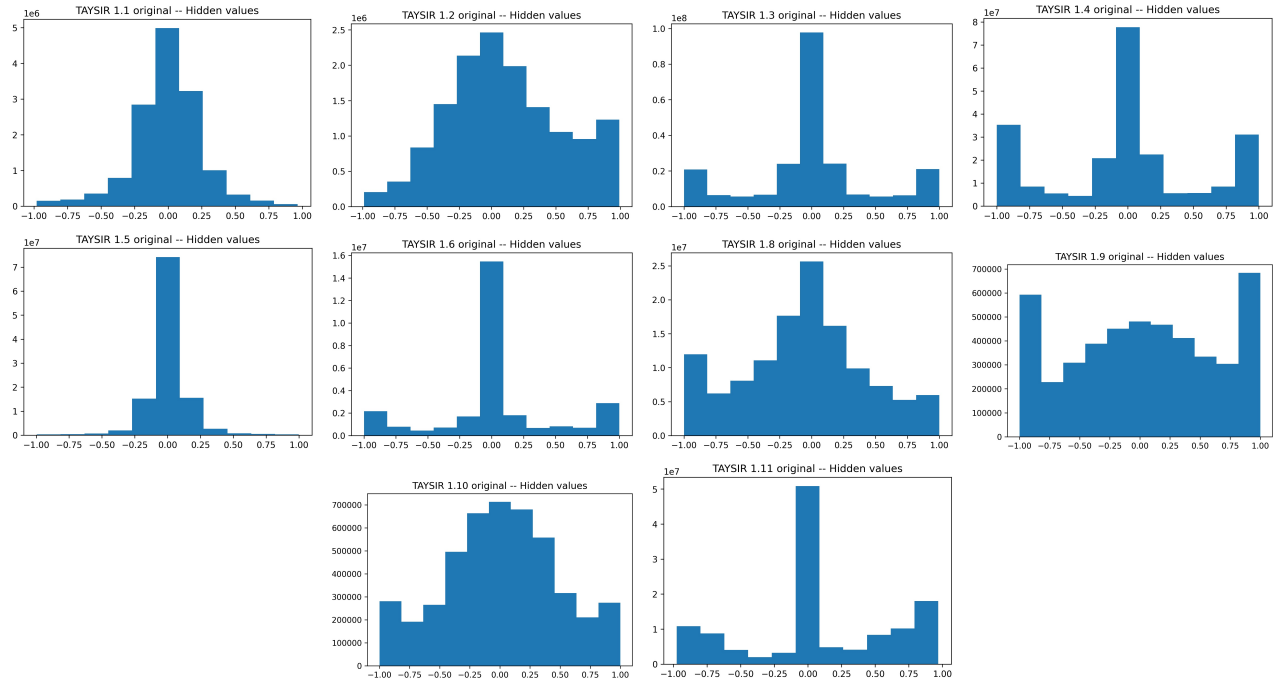

Figure 4: Distribution of hidden state values while parsing the validation set with the original TAYSIR models

We observe a trend in this figure: several distributions show values either around zero or close to the extrema -1 and 1. This is already an observation in favor of our work: it shows that asking hidden values to be close to the extrema is connected to practical RNN behavior.

However, one characteristic of the TAYSIR models may bias these plots: the initial state is always the vector made of zeros. We thus retrained the models keeping the same architecture but with an initial state made of -1's and 1's, randomly selected. The result is given in Figure 5.

While model 1.11 surprisingly does not have the noticeable distribution anymore, model 1.5 now exhibits it in a clear way. Overall, the trend is more visible, with 3 models where the large majority of hidden state values are around -1 and 1.

However, these 3 models, numbered 1.3, 1.4, and 1.5, are GRUs while our paper focuses on SRN. Since directly replacing the GRU cells by SRNs does not achieve interesting accuracy, we grid search the space of hyperparameters used for the competition to obtain competitive SRNs (with initial vectors made of 1's and -1's). The achieve

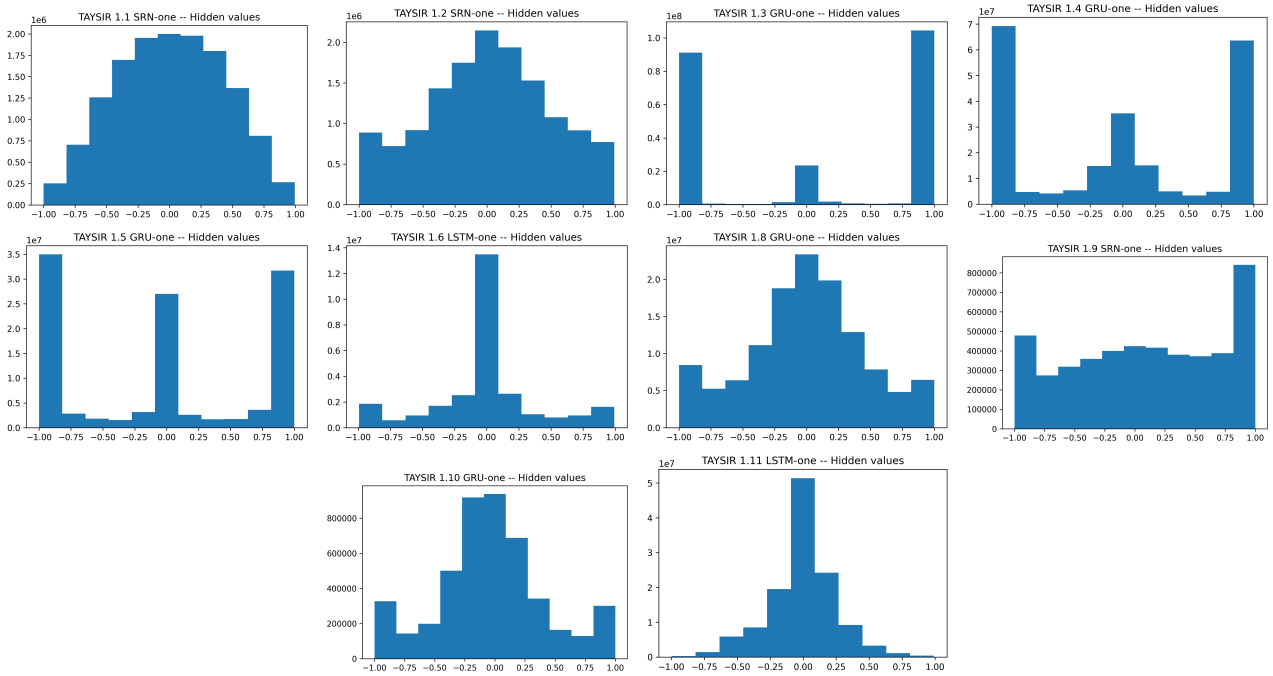

Figure 5: Distribution of hidden state values while parsing the validation set with the original TAYSIR models

accuracy is given in Table 1 and compares to the original model one.

Number	Original model	SRN
1.3	1.0	0.99983
1.4	0.99983	0.99983
1.5	1.0	0.99842

Table 1: Accuracy of original GRU models and of trained SRNs

In the left part of Figure 6 we provide the distribution of hidden state values for each of the obtained SRN: almost all these values are greater that 0.8 for model 1.3 and 1.5.

To obtain the plots on the right of the figure, we fed the models with each sentence of the validation set. For each component of the hidden states, we track how many consecutive times the value was less than 0.7. We report in these plots the maximum value reached by this number for each sequence. Notice that the length of sequences on all these datasets is 22.

For SRN models 1.3 and 1.5, on almost all sequences, almost all components of the hidden states exceed the threshold at most every 10 time steps. As discussed in the main text, this is a clear practical observation in favor of DS β-saturation.

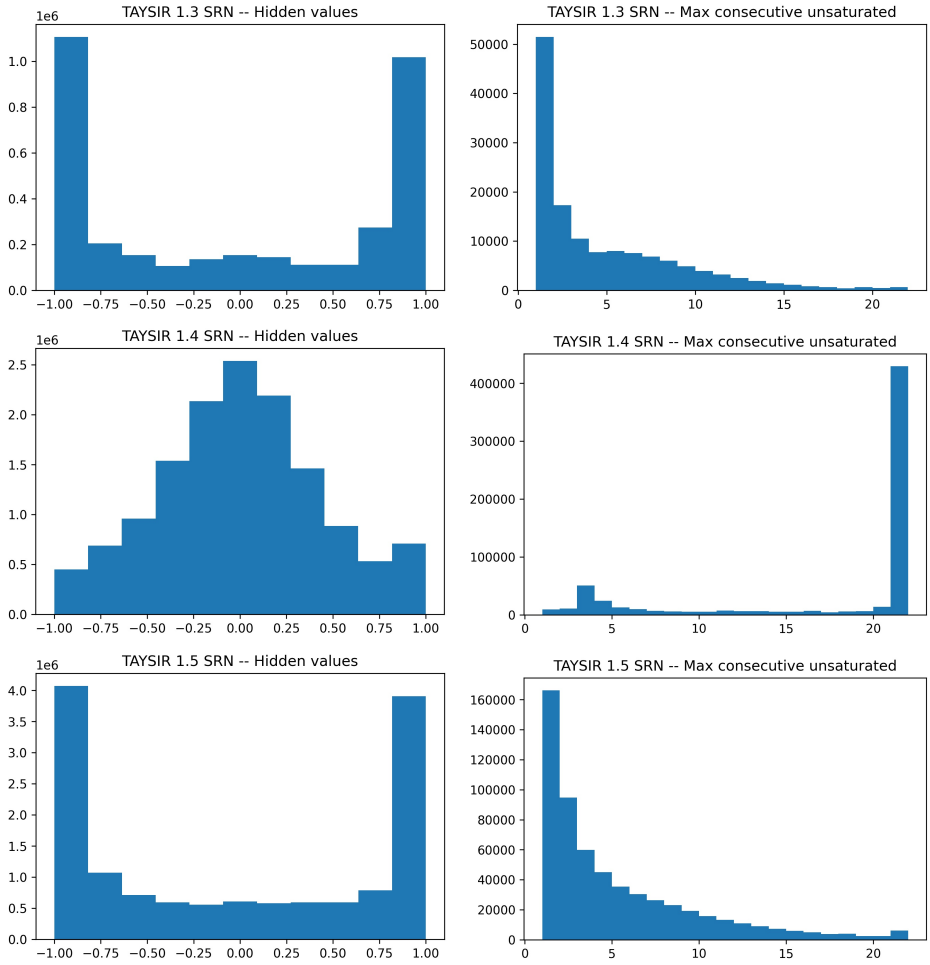

Figure 6: Accurate SRN hidden value distributions and maximum consecutive under a threshold per component.

B SUFFICIENT CONDITION FOR PAC-BAYES BOUNDS

In this section we propose a relaxation of the results in Neyshabur et al. (2018). We start by recalling some definitions and Lemma 1 in Neyshabur et al. (2018).

Definition 2.6 (Expected margin loss)**.** *Let $\gamma \geq 0$ be the margin. The expected margin loss is defined by:*

$$L_\gamma(f_P) = \mathbb{P}_{(X^T,y)\sim\mathcal{D}}\Big[f_P(X^T)[y] - \max_{j\neq y} f_P(X^T)[j] \leq \gamma\Big].$$

The margin loss can be seen as a measurement of the classifier's strength toward to perturbations. Its empirical counterpart is classically defined as follows:

Definition 2.7 (Empirical margin loss)**.** *Let f_P be a classifier on $\mathcal{X}\times\mathcal{Y}$ and $\gamma > 0$. The empirical margin loss is:*

$$\widehat{L}_\gamma(f_P) = \frac{1}{m}\sum_{(X^T,y)\in\mathcal{Z}_m} \mathbf{1}_{\{f_P(X^T)[y]-\max_{j\neq y} f_P(X^T)[j]\leq\gamma\}}$$

where $\mathbf{1}$ *is the indicator function.*

Note that if $\gamma = 0$, $L_0(f_P)$ is the expected zero-one loss and $\widehat{L}_0(f_P)$ is the empirical zero-one loss.

Lemma B.3 (Lemma 1 in Neyshabur et al. (2018))**.** *Let $\mathcal{H}$ be a set of classifiers on $\mathcal{X}\times\mathcal{Y}$. Let $\pi, \mathcal{Q}$ two distributions on $\mathcal{H}$. Let also $\mathcal{Z}_m \subset \mathcal{X}\times\mathcal{Y}$ a sample of m training instances iid from an unknown distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$. The distribution π is called the prior and is assumed to be independent of $\mathcal{Z}_m$. Let $f_P \in \mathcal{H}$ drawn with respect to the distribution $\mathcal{Q}$. If:*

$$\mathbb{P}_{J\sim\mathcal{Q}}\left[\sup_{X^T\in\mathcal{X}} \left\|f_P(X^T) - f_J(X^T)\right\|_\infty < \gamma/4\right] \geq \frac{1}{2},$$

then, with probability $1-\delta$ over the learning sample of size m,

$$L_0(f_P) \leq \widehat{L}_\gamma(f_P) + 4\sqrt{\frac{2KL\left(\mathcal{Q}\|\pi\right) + \ln\left(\frac{6m}{\delta}\right)}{m-1}}.$$

We now provide the main result of this section, that is, a relaxation of this lemma.

Lemma 2.8 (Relaxation of Lemma 1 in Neyshabur et al. (2018))**.** *Let $\mathcal{H}$ be a set of classifiers on $\mathcal{X}\times\mathcal{Y}$. Let $\pi, \mathcal{Q}$ two distributions on $\mathcal{H}$. Let $\mathcal{Z}_m \subset \mathcal{X}\times\mathcal{Y}$ a set of m training samples assumed to be drawn iid from an unknown distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$ with $\mathcal{D}_\mathcal{X}$ the marginal over $\mathcal{X}$. The distribution π is the prior and is assumed to be independent of $\mathcal{Z}_m$. Let $f_P \in \mathcal{H}$ drawn with respect to the distribution $\mathcal{Q}$. If $\exists\, \mathcal{X}_\tau \subset \mathcal{X}$ such that $\mathcal{D}_\mathcal{X}\left(\mathcal{X}_\tau\right) = 1-\tau$ for $\tau \in [0,1[$,*

$$\mathbb{P}_{J\sim\mathcal{Q}}\left[\sup_{X^T\in\mathcal{X}_\tau} \left\|f_P(X^T) - f_J(X^T)\right\|_\infty < \gamma/4\right] \geq \frac{1}{2},$$

then, with probability $1-\delta$ over $\mathcal{Z}_m$, we have

$$L_0(f_P) \leq \widehat{L}_\gamma(f_P) + \tau + 4\sqrt{\frac{2KL\left(\mathcal{Q}\|\pi\right) + \ln\left(\frac{6m}{\delta}\right)}{m-1}},$$

where $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

The proof of this statement is of similar nature of the one of Neyshabur et al. (2018).

Proof. Let $\mathcal{H}$ be a set of classifiers on $\mathcal{X}\times\mathcal{Y}$. Let $\pi, \mathcal{Q}$ two distributions on $\mathcal{H}$. Let also $\mathcal{Z}_m \subset \mathcal{X}\times\mathcal{Y}$. The distribution π is called the prior and is assumed to be independent of $\mathcal{Z}_m$. Let $f_P \in \mathcal{H}$ drawn with respect to the distribution $\mathcal{Q}$. We assume that we have:

- $\mathcal{X}_\tau \subset \mathcal{X}$ such that $\mathcal{D}_{\mathcal{X}}(\mathcal{X}_\tau) = 1 - \tau < 1$. If $\tau = 0$ we are in the case of Lemma 1 in Neyshabur et al. (2018)
- and $\mathbb{P}_{J\sim\mathcal{Q}}\left[\sup_{X^T\in\mathcal{X}_\tau} \left\|f_P(X^T) - f_J(X^T)\right\|_\infty < \gamma/4\right] \geq \frac{1}{2}$.

We start by defining the margin function:

$$\mathcal{M}_\gamma(X^T, y) = f_P(X^T)[y] - \max_{j\neq y} f_P(X^T)[j] - \gamma,$$

and the set:

$$S_P = \left\{J : \sup_{X^T\in\mathcal{X}_\tau} \left\|f_P(X^T) - f_J(X^T)\right\|_\infty < \gamma/4\right\}.$$

Then we define, given $A \in \mathcal{X}$

$$\mathcal{D}_\tau(A) := \frac{\mathcal{D}(A\cap\mathcal{X}_\tau)}{1-\tau}$$
$$\mathcal{D}^c_\tau(A) := \frac{\mathcal{D}(A\cap\mathcal{X}^c_\tau)}{\tau},$$

where $\mathcal{X}^c_\tau = \mathcal{X}\setminus\mathcal{X}_\tau$. The definitions of $\mathcal{D}_\tau$ and of $\mathcal{D}^c_\tau$ give us the expression:

$$\mathcal{D}(A) = \tau\mathcal{D}^c_\tau(A) + (1-\tau)\mathcal{D}_\tau(A).$$

We also define the distribution $\tilde{\mathcal{Q}}$ on $\mathcal{H}$ by the density q of the distribution Q:

$$\tilde{q}(J) = \frac{1}{Z}\begin{cases} q(J) & \text{if } J \in S_P \\ 0 & \text{otherwise,}\end{cases}$$

where $Z := \mathbb{P}[J \in S_P] \geq \frac{1}{2}$ is a constant. We define:

$$L^\tau_\gamma(f_P) = \mathbb{P}_{(X^T,y)\sim\mathcal{D}_\tau}\left[\mathcal{M}_\gamma(X^T, y) \leq 0\right]$$
$$\widehat{L}^\tau_\gamma(f_P) = \frac{1}{m}\sum_{(X^T,y)\in\mathcal{Z}_m} \mathbf{1}_{\{f_P(X^T)[y]-\max_{j\neq y} f_P(X^T)[j]\leq\gamma\}}\mathbf{1}_{\{X^T_i\in\mathcal{X}_\tau\}}.$$

By definition of $\mathcal{D}_\tau$ and from Neyshabur et al. (2018) we have:

$$L^\tau_0(f_P) \leq L^\tau_{\gamma/2}(f_J)$$
$$\widehat{L}^\tau_{\gamma/2}(f_{P+\vartheta}) \leq \widehat{L}^\tau_\gamma(f_P).$$

Also by definition of $\mathcal{D}_\tau$ and by what precedes we have:

$$\begin{aligned} L_0(f_P) &= (1-\tau)L^\tau_0(f_P) + \tau L^{\tau^c}_0(f_P) \\ &\leq (1-\tau)L^\tau_0(f_P) + \tau \\ &\leq L^\tau_{\gamma/2}(f_{P+\vartheta}) + \tau. \end{aligned} \tag{8}$$

Then:

$$\begin{aligned} \widehat{L}^\tau_{\gamma/2}(f_J) &\leq \widehat{L}^\tau_\gamma(f_P) \\ &= \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{\mathcal{M}_\gamma(X^T_i,y)\leq 0\}}\mathbf{1}_{\{X^T_i\in\mathcal{X}_\tau\}} \\ &\leq \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{\mathcal{M}_\gamma(X^T_i,y)\leq 0\}} \\ &= \widehat{L}_\gamma(f_P), \end{aligned} \tag{9}$$

because for all $A_1, A_2 \subset \mathcal{X}$ we have $\mathbf{1}_{A_1}\mathbf{1}_{A_2} \leq \mathbf{1}_{A_i}$ with $i = 1, 2$. Now we apply the inequalities 8, 9 and obtain:

$$\begin{aligned}
L_0(f_P) &\leq \mathbb{E}_{J\sim\tilde{\mathcal{Q}}}\left[L^{\tau}_{\gamma/2}(f_J) + \tau\right] \\
&\leq \mathbb{E}_{J\sim\tilde{\mathcal{Q}}}\left[\widehat{L}^{\tau}_{\gamma/2}(f_J)\right] + \tau + 2\sqrt{\frac{2(KL(\tilde{\mathcal{Q}}\|\pi) + \ln\left(\frac{2m}{\delta}\right))}{m-1}} \\
&\leq \widehat{L}_{\gamma}(f_P) + \tau + 2\sqrt{\frac{2(KL(\tilde{\mathcal{Q}}\|\pi) + \ln\left(\frac{2m}{\delta}\right))}{m-1}} \\
&\leq \widehat{L}_{\gamma}(f_P) + \tau + 2\sqrt{2}\sqrt{\frac{2(KL(\mathcal{Q}\|\pi) + 1) + \ln\left(\frac{2m}{\delta}\right)}{m-1}} \\
&\leq \widehat{L}_{\gamma}(f_P) + \tau + 2\sqrt{2}\sqrt{\frac{2(KL(\mathcal{Q}\|\pi) + \ln(3)) + 2\ln\left(\frac{2m}{\delta}\right)}{m-1}} \\
&\leq \widehat{L}_{\gamma}(f_P) + \tau + 4\sqrt{\frac{KL(\mathcal{Q}\|\pi) + \ln\left(\frac{6m}{\delta}\right)}{m-1}},
\end{aligned}$$

where the proof of $KL(\tilde{\mathcal{Q}}\|\pi) \leq 2(KL(\mathcal{Q}\|\pi) + 1)$, can be found in Neyshabur et al. (2018). □

C MAIN THEOREMS

In this section, we provide detailed proofs of the PAC-Bayes results stated in the paper. It is organized as follows. In Section C.1, we state and prove a series of lemmas that constitutes the building blocks of the main results. The main theorems are stated and proved in Section C.2. In Section C.3, we recall a few technical results that are needed in the rest of the document.

For clarity reasons, we first recall our main notations: if $\mathcal{R}_P$ is a SRN, then (h_k, y_k), $(\mathring{h}_k, \mathring{y}_k)$ and $(\tilde{h}_k, \tilde{y}_k)$ are the hidden state vector and the output vectors produced at step k by $\mathcal{R}_P$, $\mathcal{R}_{P+\vartheta}$, and $\mathcal{R}^{\varepsilon}_P$, respectively, where $\mathcal{R}_{P+\vartheta}$ is the perturbed SRN and $\mathcal{R}^{\varepsilon}_P$ is the fuzzy SRN.

C.1 Auxiliary Lemmas

C.1.1 Relationship Between Perturbed and Fuzzy SRNs

In this subsection, after stating some results on the distance between a SRN $\mathcal{R}_P$ and $\mathcal{R}_{P+\vartheta}$ or $\mathcal{R}^{\varepsilon}_P$, we show that a fuzzy SRN can dominate a perturbed SRN if ϑ and ε are Gaussian multivariate independent variables, and if the size of the noise ε is large enough. This simplifies the problem of bounding $\|\mathcal{R}_P(X^T) - \mathcal{R}_{P+\vartheta}(X^T)\|$ for a given X^T.

Lemma C.1 (Distance between $\mathcal{R}_P$ and $\mathcal{R}_{P+\vartheta}$)**.** *Let $\mathcal{R}_P$ be a SRN and let $\vartheta \in \mathbb{R}^{dim(P)}$. Let $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$. If the activation function σ is 1-Lipschitz, then:*

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_{P+\vartheta}(X^T)\| \leq \|V\|\|h_T - \mathring{h}_T\| + \|\vartheta_V \mathring{h}_T + \vartheta_c\|.$$

Proof. Let $\mathcal{R}_{P+\vartheta}(X^t) = \mathring{y}_T$. Following the notations and assumptions introduced in the statement of this lemma, for all $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$ we have:

$$\begin{aligned}
\|\mathcal{R}_P(X^T) - \mathcal{R}_{P+\vartheta}(X^T)\| &= \|\sigma\left(Vh_T + c\right) - \sigma\left((V+\vartheta_V)\mathring{h}_T + c + \vartheta_c\right)\| \\
&\leq \|Vh_T + c - \left((V+\vartheta_V)\mathring{h}_T + c + \vartheta_c\right)\| \\
&= \|V(h_T - \mathring{h}_T) - \vartheta_V\mathring{h}_T - \vartheta_c\| \\
&\leq \|V\|\|h_T - \mathring{h}_T\| + \|\vartheta_V\mathring{h}_T + \vartheta_c\|.
\end{aligned}$$

where the first inequality follows from the fact that σ is 1-Lipschitz. □

Corollary C.2 (Distance between $\mathcal{R}_P$ and $\mathcal{R}_P^{\varepsilon},$)**.** *Let $\mathcal{R}_P$ be a SRN, and let $\varepsilon = (\varepsilon_d, \varepsilon_o) \in \mathbb{R}^{d+o}$. For $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$ we have:*

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^{\varepsilon}(X^T)\| \leq \|V\|\|h_T - \tilde{h}_T\| + \|\varepsilon_o\|.$$

Proof. We define $\vartheta \in \mathbb{R}^{dim(P)}$ such that $\vartheta_V = 0, \vartheta_U = 0$ and $\vartheta_W = 0$ are zero matrices and $\vartheta_c = \varepsilon_o$ as well as $\vartheta_b = \varepsilon_d$. It is easy to check that $\mathcal{R}_P^{\varepsilon} = \mathcal{R}_{P+\vartheta}$. From this equality we deduce the claim of the corollary. □

Lemma C.3 (Noise regrouping)**.** *Let $\mathcal{R}_P$ be a SRN, $\vartheta \sim \mathcal{N}(0, \rho^2 I_{dim(P)})$ and $\varepsilon = (\varepsilon_d, \varepsilon_o) \sim \mathcal{N}(0, \nu^2 I_{(d+o)})$ two random vectors. Let B such that for all $\{x_k\}_{k=1}^T$ and for all k, $\|x_k\| \leq B$. We set $\nu^2 := \rho^2(B^2 + d + 1)$, then for all $t > 0$ such that*

$$\rho^2(B^2 + d + 1)(\max(o, d) - 2) \leq t^2$$

and for all $\{x_k\}_{k=1}^T$ we have:

$$\mathbb{P}\Big[\|\vartheta_U x_k + \vartheta_W \mathring{h}_{k-1} + \vartheta_b\| < t\Big] \geq \mathbb{P}\Big[\|\varepsilon_d\| < t\Big],$$
$$\mathbb{P}\Big[\|\vartheta_V \mathring{h}_k + \vartheta_c\| < t\Big] \geq \mathbb{P}\Big[\|\varepsilon_o\| < t\Big].$$

Proof. Following the notations and assumptions introduced in the statement of this lemma, we have that $\vartheta_U x_k + \vartheta_W \mathring{h}_{k-1} + \vartheta_b$ and $\vartheta_V \mathring{h}_k + \vartheta_c$ are centered Gaussian random variables of dimension d and o, respectively. By hypothesis on $\mathcal{X}$ we have $\|x_k\| \leq B$ for all $X^T = \{x_k\}_{k=1}^T$. And due to the fact that for all $\omega \in \mathbb{R}$, $|\tanh(\omega)| \leq 1$, for all hidden state vectors $\mathring{h}_k$ we have $\|\mathring{h}_k\| \leq \sqrt{d}$. Thus by Lemma C.19 and usual properties of Gaussian variables we have that the variance of the coordinates of $\vartheta_U x_k + \vartheta_W \mathring{h}_{k-1} + \vartheta_b$ is bounded by $\rho^2(B^2 + d + 1)$. By applying the same arguments we have that the variance of the coordinates of $\vartheta_V \mathring{h}_k + \vartheta_c$ is bounded by $\rho^2(d + 1) \leq \rho^2(B^2 + d + 1)$. We define the following function:

$$f : (\nu, d) \mapsto e^{-(t-(\sqrt{d-2})\nu)^2/4d\nu^2}.$$

It is easy to see that the function $f(\cdot, d)$ is strictly increasing on $0 < \nu < \frac{t}{\sqrt{d-2}}$. Thus, under the assumptions $\nu^2 = \rho^2(B^2 + d + 1)$ and $\nu^2 \leq \frac{t^2}{\max(o,d)}$, we can prove the lemma by using the concentration inequality from Lemma C.20. □

C.1.2 On Perturbed SRN

In this subsection we state a result allowing us to estimate the gap between an SRN and its perturbed version. An analogous result is proved for Fuzzy SRN . The purpose of this Lemma is to prove that it is sufficient to noise only the bias in the SRN parameters.

Lemma C.4 (Expression of the perturbed hidden state vector)**.** *Let $\mathcal{R}_P$ be a SRN and $\mathcal{R}_{P+\vartheta}$ a perturbed SRN with $\vartheta \in \mathbb{R}^{dim(P)}$. For all $\{x_k\}_{k=1}^T \in \mathcal{X}$ there exists a sequence $\{c_k\}_{k=1}^T \in \mathbb{R}^h$ such that for $1 \leq k \leq T$:*

$$\mathring{h}_k = h_k + \sum_{s=1}^{k-1} \Lambda(c_k) \left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) \right) n_s + \Lambda(c_k) n_k$$

where :

- $\Lambda(c_k) = \mathrm{Diag}\big(\sigma'(c_k)\big)$ *with c_k provided by Lemma C.18*
- $\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) = W\Lambda(c_{k-1})W\Lambda(c_{k-2}) \cdots W\Lambda(c_s)$.
- $n_k = \Big(\vartheta_U x_k + \vartheta_W \mathring{h}_{k-1} + \vartheta_b\Big)$

Proof. Following the notations introduced in the statement of the lemma, we prove the result by induction on k. We analyze the execution of the perturbed SRN $\mathcal{R}_P^{\varepsilon}$. By Lemma C.18, there exists $c_1 \in \mathbb{R}^h$ such that:

$$\begin{aligned}\mathring{h}_1 &= \sigma\left((Ux_1 + \vartheta_U x_1) + (Wh_0 + \vartheta_W h_0) + (b + \vartheta_b)\right)\\ &= \sigma\left((Ux_1 + Wh_0 + b) + (\vartheta_U x_1 + \vartheta_W h_0 + \vartheta_b)\right)\\ &= \sigma\left(Ux_1 + Wh_0 + b\right) + \sigma'(c_1) \odot (\vartheta_U x_1 + \vartheta_W h_0 + \vartheta_b).\end{aligned}$$

By definition of the perturbed FP-SRN we have $\mathring{h}_0 = h_0$ what gives us the following equality:

$$\left(\vartheta_U x_1 + \vartheta_W h_0 + \vartheta_b\right) = \left(\vartheta_U x_1 + \vartheta_W \mathring{h}_0 + \vartheta_b\right) =: n_1.$$

Hence we deduce the equality:

$$\mathring{h}_1 = h_1 + \Lambda(c_1) n_1,$$

where in the last line we set $\Lambda(c) = \mathrm{Diag}(\sigma'(c))$. We proved the initialisation step (i.e. $k = 1$) but for comprehensibility we prove the statement for $k = 2$. By reapplying the same argument we can find a vector $c_2 \in \mathbb{R}^h$ such that:

$$\begin{aligned}\mathring{h}_2 &= \sigma\left((U + \vartheta_U)x_2 + (W + \vartheta_W)(h_1 + \Lambda(c_1)n_1) + (b + \vartheta_b)\right)\\ &= \sigma\left(Ux_2 + W(h_1 + \Lambda(c_1)n_1) + b + (\vartheta_U x_1 + \vartheta_W \mathring{h}_1 + \vartheta_b)\right)\\ &= \sigma\left(Ux_2 + Wh_1 + b\right) + \sigma'(c_2) \odot (W\Lambda(c_1)n_1) + \sigma'(c_2) \odot (\vartheta_U x_1 + \vartheta_W \mathring{h}_1 + \vartheta_b)\\ &= h_2 + \Lambda(c_2)W\Lambda(c_1)n_1 + \Lambda(c_2)n_2\\ &= h_2 + \sum_{s=1}^{2-1} \Lambda(c_2)\left(\prod_{l=1}^{2-s} W\Lambda(c_{2-l})\right) n_1 + \Lambda(c_2)n_2.\end{aligned}$$

Let us assume now that for $k \geq 1$ we have the following expression:

$$\mathring{h}_k = h_k + \sum_{s=1}^{k-1} \Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right) n_s + \Lambda(c_k)n_k.$$

By Lemma C.18 there exists a vector $c_{k+1} \in \mathbb{R}^h$ such that:

$$\begin{aligned}\tilde{h}_{k+1} &= \sigma\left((U + \vartheta_U)x_{k+1} + (W + \vartheta_W)\Big(h_k + \sum_{s=1}^{k-1} \Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right) n_s + \Lambda(c_k)n_k\Big) + (b + \vartheta_b)\right)\\ &= \sigma\left(Ux_{k+1} + W\Big(h_k + \sum_{s=1}^{k-1} \Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right) n_s + \Lambda(c_k)n_k\Big) + b + n_{k+1}\right),\end{aligned}$$

where $n_{k+1} := \vartheta_U x_{k+1} + \vartheta_W \mathring{h}_k + \vartheta_b$. Now by applying Lemma C.18 we obtain:

$$\begin{aligned}\mathring{h}_{k+1} &= \sigma\left(Ux_{k+1} + Wh_k + b\right) + \sigma'(c_{k+1}) \odot \left(W \sum_{s=1}^{k-1} \Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right) n_s + W\Lambda(c_k)n_k + n_{k+1}\right)\\ &= h_{k+1} + \Lambda(c_{k+1}) \sum_{s=1}^{k-1} \left(\prod_{l=1}^{k-s+1} W\Lambda(c_{k-l+1})\right) n_s + \Lambda(c_{k+1})W\Lambda(c_k)n_k + \Lambda(c_{k+1})n_{k+1}\\ &= h_{k+1} + \sum_{s=1}^{k} \Lambda(c_{k+1})\left(\prod_{l=1}^{k-s+1} W\Lambda(c_{k-l+1})\right) n_s + \Lambda(c_{k+1})n_{k+1}.\end{aligned}$$

Thus by induction Lemma C.4 is true for all $\{x_1, \ldots, x_T\} \in \mathcal{X}$. □

C.1.3 On Fuzzy SRN

From the previous subsection we know that it is enough to cleverly choose the variance of the noise ε to simplify the problem and study fuzzy SRNs. We state here a couple of results concerning the latter. The first result is a bound on the distance between $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$, and the second result gives an expression of $\tilde{h}_k$ as a function of h_k and the noise ε_d.

Lemma C.5 (A general bound on the distance between $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$)**.** *Let $\mathcal{R}_P$ be a SRN and let $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{d+o}$. Then for all $X^T \in \mathcal{X}$ we have:*

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\| \leq \|V\| \left(\sum_{s=0}^{T-1} \|W\|^s \right) \|\varepsilon_d\| + \|\varepsilon_o\|.$$

Proof. Let $\mathcal{R}_P$ be a SRN and let $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{d+o}$. Let $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$. The proof is by induction on the time step k. We have by Lemma C.2:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\| \leq \|V\|\|h_T - \tilde{h}_T\| + \|\varepsilon_o\|.$$

We need to bound $\|h_T - \tilde{h}_T\|$. At first we will bound $\|h_1 - \tilde{h}_1\|$ by leveraging that σ is 1-Lipschitz:

$$\begin{aligned}
\|h_1 - \tilde{h}_1\| &= \|\sigma\left(Ux_1 + Wh_0 + b\right) - \sigma\left(Ux_1 + Wh_0 + b + \varepsilon_d\right)\| \\
&\leq \|Ux_1 + Wh_0 + b - Ux_1 - Wh_0 - b - \varepsilon_d\| \\
&= \|\varepsilon_d\|.
\end{aligned}$$

Now we assume that up to $k-1$, $1 \leq k-1 < T$, we have:

$$\|h_{k-1} - \tilde{h}_{k-1}\| \leq \left(\sum_{s=0}^{k-2} \|W\|^s \right) \|\varepsilon_d\|,$$

and we then prove that this property is true also for k.

$$\begin{aligned}
\|h_k - \tilde{h}_k\| &= \left\|\sigma\Big(Ux_k + Wh_{k-1} + b\Big) - \sigma\Big(Ux_k + W\tilde{h}_{k-1} + b + \varepsilon_d\Big)\right\| \\
&\leq \left\|Ux_k + Wh_{k-1} + b - Ux_k - W\tilde{h}_{k-1} - b - \varepsilon_d\right\| \\
&= \left\|W\left(h_{k-1} - \tilde{h}_{k-1}\right) - \varepsilon_d\right\| \\
&\leq \|W\| \left\|h_{k-1} - \tilde{h}_{k-1}\right\| + \|\varepsilon_d\| \\
&\leq \|W\| \left[\left(\sum_{s=0}^{k-2} \|W\|^s \right) \|\varepsilon_d\| \right] + \|\varepsilon_d\| \\
&= \left(\sum_{s=0}^{k-1} \|W\|^s \right) \|\varepsilon_d\|.
\end{aligned}$$

Finally, we combine this result for $k = T$ with the one of Lemma C.2 and we obtain:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\| \leq \|V\| \left(\sum_{s=0}^{T-1} \|W\|^s \right) \|\varepsilon_d\| + \|\varepsilon_o\|.$$

□

Lemma C.6 (Expression of the fuzzy hidden state vector)**.** *Let $\mathcal{R}_P$ be a SRN and $\mathcal{R}_P^\varepsilon$ a fuzzy SRN with $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{d+o}$. For all $\{x_k\}_{k=1}^T \in \mathcal{X}$ there exists a sequence $\{c_k\}_{k=1}^T \in \mathbb{R}^d$ such that for $1 \leq k \leq T$:*

$$\tilde{h}_k = h_k + \sum_{s=1}^{k-1} \Lambda(c_k) \left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) \right) \varepsilon_d + \Lambda(c_k)\varepsilon_d$$

where :

- $\Lambda(c_k) = \mathrm{Diag}\big(\sigma'(c_k)\big)$
- $\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) = W\Lambda(c_{k-1})W\Lambda(c_{k-2})\cdots W\Lambda(c_s)$.

Proof. Following the notations introduced in the statement of the lemma, we prove the result by induction on k. We analyze the execution of the fuzzy SRN $\mathcal{R}_p^\varepsilon$. By Lemma C.18, there exists $c_1 \in \mathbb{R}^d$ such that:

$$\begin{aligned}\tilde{h}_1 &= \sigma\left(Ux_1 + Wh_0 + b + \varepsilon_d\right)\\ &= \sigma\left(Ux_1 + Wh_0 + b\right) + \sigma'(c_1) \odot \varepsilon_d\\ &= h_1 + \Lambda(c_1)\varepsilon_d,\end{aligned}$$

where in the last line we set $\Lambda(c) = \mathrm{Diag}(c)$ and leverage the linearity of the Hadamard product denoted by $\odot$.

We proved the initialisation step (i.e. $k = 1$) but for comprehensibility we prove the statement for $k = 2$. By reapplying the same argument we can find a vector $c_2 \in \mathbb{R}^d$ such that:

$$\begin{aligned}\tilde{h}_2 &= \sigma\left(Ux_2 + W(h_1 + \Lambda(c_1)\varepsilon_d) + b + \varepsilon_d\right)\\ &= \sigma\left(Ux_2 + Wh_1 + b + W\Lambda(c_1)\varepsilon_d + \varepsilon_d\right)\\ &= \sigma\left(Ux_2 + Wh_1 + b\right) + \sigma'(c_2) \odot \left(W\Lambda(c_1)\varepsilon_d + \varepsilon_d\right)\\ &= h_2 + \Lambda(c_2)W\Lambda(c_1)\varepsilon_d + \Lambda(c_2)\varepsilon_d\\ &= h_2 + \sum_{s=1}^{2-1}\Lambda(c_2)\left(\prod_{l=1}^{2-s} W\Lambda(c_{2-l})\right)\varepsilon_d + \Lambda(c_2)\varepsilon_d.\end{aligned}$$

Let us assume now that for $k \geq 1$ we have the following expression:

$$\tilde{h}_k = h_k + \sum_{s=1}^{k-1}\Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + \Lambda(c_k)\varepsilon_d.$$

By Lemma C.18 there exists a vector $c_{k+1} \in \mathbb{R}^d$ such that:

$$\begin{aligned}\tilde{h}_{k+1} &= \sigma\left(Ux_{k+1} + W\left(h_k + \sum_{s=1}^{k-1}\Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + \Lambda(c_k)\varepsilon_d\right) + b + \varepsilon_d\right)\\ &= \sigma\left(Ux_{k+1} + Wh_k + b\right) + \sigma'(c_{k+1}) \odot \left(W\sum_{s=1}^{k-1}\Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + W\Lambda(c_k)\varepsilon_d + \varepsilon_d\right)\\ &= h_{k+1} + \Lambda(c_{k+1})\sum_{s=1}^{k-1}\left(\prod_{l=1}^{k-s+1} W\Lambda(c_{k-l+1})\right)\varepsilon_d + \Lambda(c_{k+1})W\Lambda(c_k)\varepsilon_d + \Lambda(c_{k+1})\varepsilon_d\\ &= h_{k+1} + \sum_{s=1}^{k}\Lambda(c_{k+1})\left(\prod_{l=1}^{k-s+1} W\Lambda(c_{k-l+1})\right)\varepsilon_d + \Lambda(c_{k+1})\varepsilon_d.\end{aligned}$$

Thus by induction Lemma C.6 is true for all $X^T \in \mathcal{X}$. □

C.1.4 On the Properties of DS β-saturated SRN

In this subsection, we state and prove a series of results for desynchronised sliding β-saturation (DS β saturation) with window of length F. Since DS β-saturation and β-saturation coincide for $F = 1$, all the results in this section will also apply to the β-saturation. These results focus on the conditions under which a locally contractive DS β-saturated SRN R_P has a fuzzy version R_P^ε that remains locally contractive. It tackles this question by introducing restrictions on the norm of ε. We first introduce a definition that will allow us to differentiate the different timesteps.

Definition C.7 (Saturation Iteration). *Let $\mathcal{R}_P$ be a DS β-saturated SRN with window size $F \geq 1$. Let $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$ a sequence of length $T > F$. By definition of DS β-saturation, for every coordinate $1 \leq i \leq d$ there exists $\{\varsigma_1, \dots, \varsigma_t\}_i \subset [\![1, T]\!]$ a non empty subset of indices such that for all $\varsigma \in \{\varsigma_1, \dots, \varsigma_t\}_i$ we have:*

$$\left|\tanh^{-1}(h_\varsigma[i])\right| > \beta,$$

and for all $\kappa \in [\![1, T]\!] \setminus \{\varsigma_1, \dots, \varsigma_t\}_i$ we have:

$$\left|\tanh^{-1}(h_\kappa[i])\right| \leq \beta.$$

The set $\{\varsigma_1, \dots, \varsigma_t\}_i$ will be called the set of saturation iterations for the coordinate i.

Definition C.8 (Similar Saturation Pattern). *Let $\mathcal{R}_P$ and $\mathcal{R}_{P'}$ be two DS β-saturated SRN. We say that $\mathcal{R}_{P'}$ has a similar saturation pattern $\mathcal{R}_P$ to on $X^T \in \mathcal{X}$ if for all $1 \leq i \leq d$ we have:*

$$\{\varsigma_1, \dots, \varsigma_t\}_i \subset \{\varsigma'_1, \dots, \varsigma'_t\}_i,$$

with $\{\varsigma'_1, \dots, \varsigma'_t\}_i$ the set of saturation iteration of $\mathcal{R}_{P'}$ on X^T.

Lemma C.9 (Pre-activation error margin). *Let $z, \eta \in \left]0, \frac{1}{2}\right[$ and $x \in \mathbb{R}$ such that $|\tanh(x)| \geq z + \eta$, then we have:*

$$|x| > \tanh^{-1}(z) + \frac{\eta}{1 - z^2}.$$

Proof. Let $z, \eta \in \left]0, \frac{1}{2}\right[$ and $x \in \mathbb{R}$ such that $|\tanh(x)| > z + \eta$. If $x > 0$, we have by the mean value theorem for some ω, $z < \omega < z + \eta$:

$$\begin{aligned}\tanh^{-1}(z + \eta) &= \tanh^{-1}(z) + \frac{\eta}{1 - \omega^2} \quad \text{because } \frac{d}{dt}\tanh^{-1}(\omega) = \frac{1}{1 - \omega^2} \\ &\geq \tanh^{-1}(z) + \frac{\eta}{1 - z^2}\end{aligned}$$

where the last inequality follows from the fact that $-\frac{d}{dt}\tanh^{-1}$ is an increasing function on the open interval $]0, 1[$. Thus:

$$\tanh^{-1}(x) > \tanh^{-1}(z) + \frac{\eta}{1 - z^2}.$$

If $x < 0$ we apply the same argument and obtain:

$$\begin{aligned}\tanh^{-1}(-z - \eta) &= -\tanh^{-1}(z + \eta) \\ &= -\tanh^{-1}(z) - \frac{\eta}{1 - \omega^2} \\ &\leq -\tanh^{-1}(z) - \frac{\eta}{1 - z^2}\end{aligned}$$

where the last inequality follows from the fact that $-\frac{d}{dt}\tanh^{-1}$ is an increasing function on the open interval $]-1, 0[$.

We obtain that:

$$\tanh^{-1}(x) < -\tanh^{-1}(z) - \frac{\eta}{1 - z^2}.$$

Finally we combine both inequalities to deduce:

$$\left|\tanh^{-1}(x)\right| > \tanh^{-1}(z) + \frac{\eta}{1 - z^2}.$$

□

Lemma C.10 (Margin error control). *Let $z, \eta \in \left]0, \frac{1}{2}\right[$ and $x \in \mathbb{R}$ such that $|\tanh(x)| \geq z + \eta$. For all $0 < t < 1$ and $-\frac{(1-t)\eta}{1-z^2} < g < \frac{(1-t)\eta}{1-z^2}$ we have :*

$$\tanh'\left(\tanh^{-1}(x + g)\right) < \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1 - z^2}\right)$$

Proof. Let $z, \eta \in \left]0, \frac{1}{2}\right[$ and $x \in \mathbb{R}$ such that $|\tanh(x)| > z + \eta$. Since the function $\tanh'$ is symmetric, we can assume that $x > 0$, without loss of generality. By Lemma C.9 we have:

$$\tanh^{-1}(x) > \tanh^{-1}(z) + \frac{\eta}{1-z^2} = \tanh^{-1}(z) + \frac{t\eta}{1-z^2} + \frac{(1-t)\eta}{1-z^2},$$

which is equivalent to:

$$\tanh^{-1}(x) - \frac{(1-t)\eta}{1-z^2} > \tanh^{-1}(z) + \frac{t\eta}{1-z^2}.$$

Thus, since $\tanh'$ is a strictly decreasing function on $]0, \infty[$, we have:

$$\tanh'\left(\tanh^{-1}(x) - \frac{(1-t)\eta}{1-z^2}\right) < \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1-z^2}\right).$$

If we set $-\frac{(1-t)\eta}{1-z^2} < g < \frac{(1-t)\eta}{1-z^2}$, we have:

$$\tanh^{-1}(x) - g > \tanh^{-1}(x) - \frac{(1-t)\eta}{1-z^2}.$$

Thus, by applying the same argument we obtain the desired inequality:

$$\tanh'\left(\tanh^{-1}(x) - g\right) < \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1-z^2}\right).$$

□

Lemma C.11 (Control over the noise in a fuzzy saturated SRN)**.** *Let $\mathcal{R}_P$ be a β-saturated SRN with $\|W\| > 1$, $z = \sqrt{1 - \frac{1}{\|W\|}}$, $\beta > z$ and $\eta = \beta - z$. We set:*

- $0 < t < 1$,
- $\Delta := \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1-z^2}\right)$,
- $\nabla := \frac{(1-t)\eta}{1-z^2}$,
- $\varepsilon = (\varepsilon_d, \varepsilon_o) \in \mathbb{R}^{d+o}$ *such that* $\|\varepsilon_d\| < \nabla\left(1 - \Delta\|W\|\right)$.

Let $\mathcal{R}_P^\varepsilon$ be the fuzzy SRN related to $\mathcal{R}_P$. By leveraging the expression obtained in Lemma C.6, we claim that for all $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$ and for all $1 \leq k \leq T$:

$$\left\| W \sum_{s=1}^{k-1} \Lambda(c_k) \left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) \right) \varepsilon_d + W\Lambda(c_k)\varepsilon_d + \varepsilon_d \right\| \leq \sum_{s=0}^{k-1} \left(\Delta\|W\|\right)^s \|\varepsilon_d\| < \nabla. \tag{10}$$

Proof. Under the notations and assumptions of Lemma C.11 we are going to prove this theorem by induction. Remark first that by Lemma C.22 and the definition of Δ we have:

$$1 \leq \sum_{s=0}^{\infty} \left(\Delta\|W\|\right)^s = \frac{1}{1 - \Delta\|W\|} < \infty.$$

Then by hypothesis we have:

$$\|\varepsilon_d\| \leq \frac{\|\varepsilon_d\|}{1 - \Delta\|W\|} < \nabla.$$

Thus for $k = 1$ the desired property is true. Now we assume that for all $1 \leq j \leq k-1$ the lemma holds. By Lemma C.6, we have:

$$\tilde{h}_k = h_k + \sum_{s=1}^{k-1} \Lambda(c_k) \left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l}) \right) \varepsilon_d + \Lambda(c_k)\varepsilon_d.$$

By the induction hypothesis we have:

$$\left\|\sum_{s=1}^{k-1}\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + \varepsilon_d\right\| \leq \sum_{s=0}^{k-1}\Big(\Delta\|W\|\Big)^s\|\varepsilon_d\| < \nabla.$$

This implies that every coordinate of $\left(\sum_{s=1}^{k-1}\left(\prod_{l=k-1}^{s} W\Lambda(c_l)\right)\varepsilon_d + \varepsilon_d\right)$ is in the open interval $]-\nabla, \nabla[$. Thus we can prove that:

$$\|\Lambda(c_k)\| < \Delta$$

since, for $1 \leq j \leq d$, we have:

$$\|\Lambda(c_k)\| = \|\sigma'(c_k)\|_\infty = \tanh'(\min_{1\leq j\leq d}|c_k[j]|),$$

and that by Lemma C.18 we know that for all $1 \leq j \leq d$ we have $|h_k[j]| < c_k[j] < |h_k[j]| + \nabla$ thus by Lemma C.10 and the definition of Δ, ∇ we have that $\|\Lambda(c_k)\| < \Delta$. We can now bound the following sum of vectors:

$$\begin{aligned}
&\left\|W\sum_{s=1}^{k-1}\Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + W\Lambda(c_k)\varepsilon_d + \varepsilon_d\right\| \\
&\leq \sum_{s=1}^{k-1}\|W\|\|\Lambda(c_k)\|\left(\prod_{l=1}^{k-s}\|W\|\|\Lambda(c_{k-l})\|\right)\|\varepsilon_d\| + \|W\|\|\Lambda(c_k)\|\|\varepsilon_d\| + \|\varepsilon_d\| \\
&= \sum_{s=1}^{k}\left(\prod_{l=1}^{k-s+1}\|W\|\|\Lambda(c_{k-l+1})\|\right)\|\varepsilon_d\| + \|\varepsilon_d\| \\
&\leq \sum_{s=1}^{k}\Big(\Delta\|W\|\Big)^{k-s}\|\varepsilon_d\| + \|\varepsilon_d\| \\
&\leq \sum_{s=1}^{k}\Big(\Delta\|W\|\Big)^{s}\|\varepsilon_d\| + \|\varepsilon_d\| \\
&= \sum_{s=0}^{k}\Big(\Delta\|W\|\Big)^{s}\|\varepsilon_d\| \leq \sum_{s=0}^{\infty}\Big(\Delta\|W\|\Big)^{s}\|\varepsilon_d\| < \nabla.
\end{aligned}$$

This proves, by induction, the Lemma's claim. □

C.2 Main Results

In this section we state our main results, starting by the sufficiency of disturbing only the bias in the parameters of an SRN. Then we state the Backbone theorem followed by the PAC-Bayes bounds for different setup's we talked abound in the article.

The idea of the following lemma is that for a given SRN $\mathcal{R}_P$, if one can define a fuzzy SRN $\mathcal{R}_P^\varepsilon$ that does not deviates from $\mathcal{R}_P$ further then a constant r, we can define a perturbed SRN $\mathcal{R}_{P+\vartheta}$, by controlling the magnitude of $\|\vartheta\|$, that will remain within the same range of $\mathcal{R}_P$ that $\mathcal{R}_P^\varepsilon$.

Lemma C.12 (Fuzzy is enough)**.** *Let $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$ be a SRN and its fuzzy version and let $X^T = \{x_1, \ldots, x_T\} \in \mathcal{X}$. By Lemma C.6 we have the following expression for $\|\tilde{h}_T - h_T\|$:*

$$\begin{aligned}
\|\tilde{h}_T - h_T\| &= \left\|\sum_{s=1}^{T-1}\Lambda(c_T)\left(\prod_{l=1}^{T-s} W\Lambda(c_{T-l})\right)\varepsilon_d + \Lambda(c_T)\varepsilon_d\right\| \\
&\leq \sum_{s=1}^{T-1}\|\Lambda(c_T)\|\left(\prod_{l=1}^{T-s}\|W\|\|\Lambda(c_{T-l})\|\right)\|\varepsilon_d\| + \|\Lambda(c_T)\|\|\varepsilon_d\| \\
&= \|\Lambda(c_T)\|\left[\sum_{s=1}^{T-1}\left(\prod_{l=1}^{T-s}\|W\|\|\Lambda(c_{T-l})\|\right) + 1\right]\|\varepsilon_d\| \,.
\end{aligned}$$

Proof. We set $\mathcal{X}^T = \{X^k \in \mathcal{X} \,/\, k \leq T\}$ and $\mathcal{B}(0, \|\varepsilon\|)$ the Euclidean ball centered at zero and radius $\|\varepsilon\|$ of dimension $dim(\varepsilon)$. We define the function $r(T, \|\varepsilon\|)$ as follows:

$$r(T, \|\varepsilon\|) := \sup_{\mathcal{X}^T \times \mathcal{B}(0,\|\varepsilon\|)} \left\{ \|\Lambda(c_T)\| \left[\sum_{s=1}^{T-1} \left(\prod_{l=1}^{T-s} \|W\| \|\Lambda(c_{T-l})\| \right) + 1 \right] \|\varepsilon_d\| \right\} .$$

By Lemma C.4 one can deduce that:

$$\|\mathring{h}_T - h_T\| = \left\| \sum_{s=1}^{T-1} \Lambda(c_T) \left(\prod_{l=1}^{T-s} W \Lambda(c_{T-l}) \right) n_s + \Lambda(c_T) n_T \right\|,$$

with $n_s = \left(\vartheta_U x_s + \vartheta_W \mathring{h}_{s-1} + \vartheta_b \right)$. We can establish the following bound on n_s:

$$\begin{aligned} \|n_s\| &= \|\vartheta_U x_s + \vartheta_W \mathring{h}_{s-1} + \vartheta_b\| \\ &\leq \|\vartheta_U\| \|x_s\| + \|\vartheta_W\| \|\mathring{h}_{s-1}\| + \|\vartheta_b\| \\ &\leq \|\vartheta\|_{\text{Fro}} \cdot (\|x_s\| + \|\mathring{h}_{s-1}\| + 1) \\ &\leq \|\vartheta\|_{\text{Fro}} \cdot (B + \sqrt{d} + 1) \ . \end{aligned}$$

The last inequality comes from the hypothesis that the data is bounded by B, and from the fact that $\mathring{h}_{s-1}$ is produced by the tanh therefor all the entries of $\mathring{h}_{s-1}$ are between -1 and 1. One can remark that this bound does not depend on s, therefor we can assert that:

$$\|\mathring{h}_T - h_T\| \leq \|\Lambda(c_T)\| \left[\sum_{s=1}^{T-1} \left(\prod_{l=1}^{T-s} \|W\| \|\Lambda(c_{T-l})\| \right) + 1 \right] \|\vartheta\|_{\text{Fro}} \cdot (B + \sqrt{d} + 1) \ .$$

If one sets ϑ such that $\|\vartheta\|_{\text{Fro}} \leq \frac{r(T, \|\varepsilon\|)}{B + \sqrt{d} + 1)}$ then we get:

$$\|\mathring{h}_T - h_T\| \leq r(T, \|\varepsilon\|)$$

□

In the following we consider the worst case scenario therefor Lemma C.12 is well suited.

Theorem 3.1 (Backbone theorem)**.** *We assume that there is a distribution $\mathcal{D}$ on the data set $\mathcal{X}$. Let $\mathcal{E} > 0$ and $\mathcal{R}_P$ be a SRN and $\gamma \geq 0$ a margin, such that there exists $\mathcal{C} > 0$ and $0 < \tau < 1$ verifying for all $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$, $\|\varepsilon\| \leq \mathcal{E}$:*

$$\mathbb{P}_{X^T \sim \mathcal{D}} \left[\forall \ 1 \leq k \leq T \ \left\| \tilde{h}_k - h_k \right\| \leq \mathcal{C} \|\varepsilon_d\| + \alpha \right] \geq 1 - \tau$$

where $\tilde{h}_k$ is the hidden state vector of the fuzzy SRN $\mathcal{R}_P^\varepsilon$, α is constant such that $4\|V\|\alpha < \min(\gamma, 4\mathcal{E}\|V\|(\mathcal{C}+1))$ and $X^T = \{x_k\}_{k=1}^T$. Then we have the following PAC-Bayes bound with probability at least $1-\delta$ over the training sample:

$$L_0(\mathcal{R}_P) - \widehat{L}_\gamma(\mathcal{R}_P) \leq \tau + \tilde{\mathcal{O}} \left(\frac{D\mathcal{C}B\|V\|\|P\|_{\text{Fro}} + \ln(\frac{1}{\delta})}{(\overline{\gamma} - \alpha\|V\|)\sqrt{m}} \right)$$

with $D = \max(d, o)$, $\overline{\gamma} = \min(\gamma, 4\mathcal{E}\|V\|(\mathcal{C}+1))$ and B such that for all k, $1 \leq k \leq T$, $\|x_k\| \leq B$.

Proof. Let $\mathcal{R}_P$ be a SRN. Let $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{d+o}$, $\|\varepsilon\| \leq \mathcal{E}$. Let $X^T \in \mathcal{X}$. We have, by Lemma C.2:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\| \leq \|V\| \|\tilde{h}_T - h_T\| + \|\varepsilon_o\|.$$

By hypothesis $\mathbb{P}_{X^T \sim \mathcal{D}} \left[\left\| \tilde{h}_T - h_T \right\| \leq \mathcal{C}\|\varepsilon_d\| + \alpha \right] \geq 1 - \tau$, thus there exists $\mathcal{X}_\tau \subset \mathcal{X}$ such that $\mathcal{D}(\mathcal{X}_\tau) \geq 1 - \tau$ and $\forall\, X^T \in \mathcal{X}_\tau$ we have $\left\| \tilde{h}_T - h_T \right\| \leq \mathcal{C}\|\varepsilon_d\| + \alpha$. From now we suppose that $X^T \in \mathcal{X}_\tau$, what allows us to assert:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}_P^\varepsilon(X^T)\| \leq \|V\| \Big(\mathcal{C}\|\varepsilon_d\| + \alpha \Big) + \|\varepsilon_o\|.$$

Now we assume that $\varepsilon = (\varepsilon_o, \varepsilon_d) \sim \mathcal{N}(0, \nu^2 I)$. We will show that for a well chosen ν^2 we can bound $\|\varepsilon\|$ with with arbitrary probability. In the proof of Lemma C.20 we can use the lower bound on $\mathbb{E}[\|\varepsilon\|]$ to state that:

$$\mathbb{P}[\|\varepsilon_D\| > t] \geq \mathbb{P}[\|\varepsilon_d\| > t] \text{ as well as}$$
$$\mathbb{P}[\|\varepsilon_D\| > t] \geq \mathbb{P}[\|\varepsilon_o\| > t],$$

due to the fact that $\varepsilon_D, \varepsilon_d$ and ε_o share the same variance coordinate-wise and that the norm of ε_D is impacted by the dimension. Then:

$$\mathbb{P}\left[\|V\|\Big(\mathcal{C}\|\varepsilon_d\| + \alpha\Big) + \|\varepsilon_o\| < t\right] \geq \mathbb{P}\left[\|V\|\Big(\mathcal{C}\|\varepsilon_D\| + \alpha\Big) + \|\varepsilon_D\| < t\right].$$

Thus for a well chosen ν^2 we will have:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}^\varepsilon_P(X^T)\| \leq \|V\|(\mathcal{C}+1)\|\varepsilon_D\| + \|V\|\alpha$$

with high probability. We now set:

- $\overline{\gamma} = \min(\gamma, 4\mathcal{E}\|V\|(\mathcal{C}+1))$,
- $C = \left(\frac{1-\sqrt{4\ln(2)}}{1-12\ln(2)}\right)^2$,
- $\nu^2 = \left(\frac{\overline{\gamma}-\alpha\|V\|}{4\|V\|(\mathcal{C}+1)}\right)^2 \left(\frac{C}{D}\right)$.

By defining ν^2 as above we can assert thanks to Lemma C.20 that:

$$\mathbb{P}\Big[\|\varepsilon_D\| < \frac{\overline{\gamma} - \alpha\|V\|}{4\|V\|(\mathcal{C}+1)}\Big] \geq \frac{1}{2}.$$

We have that with probability at least $\frac{1}{2}$ that for all $X^T \in \mathcal{X}$:

$$\|\mathcal{R}_P(X^T) - \mathcal{R}^\varepsilon_P(X^T)\| \leq \|V\|(\mathcal{C}+1)\|\varepsilon_D\| + \alpha\|V\| < \overline{\gamma}/4.$$

Finally, we define $\vartheta \sim \mathcal{N}(0, \rho^2 I)$ where $\vartheta \in \mathbb{R}^{dim(P)}$ and $\rho^2 = \frac{\nu^2}{(B^2+d+1)}$. In virtue of Lemma C.12 and by Lemma C.3 we can claim that with probability at least $\frac{1}{2}$ we have:

$$\sup_{X^T \in \mathcal{X}_\tau} \|\mathcal{R}_P(X^T) - \mathcal{R}_{P+\vartheta}(X^T)\| < \overline{\gamma}/4.$$

As a result of this we can apply Lemma 2.8 and state that, with probability $1-\delta$ over the training samples of size m, we have:

$$L(\mathcal{R}_P) \leq \widehat{L}_\gamma(\mathcal{R}_P) + \tau + 4\sqrt{\frac{\frac{64\ln(4)}{C}\Big(DB^2 + dD + D\Big)\left(\frac{\|V\|(\mathcal{C}+1)}{\overline{\gamma}-\alpha\|V\|}\right)^2 \|P\|^2_{\text{Fro}} + \ln\frac{6m}{\delta}}{m-1}}.$$

□

A very important remark is that this theorem is valid if the Euclidean norm $\|\cdot\|$ is replaced by the infinite norm $\|\cdot\|_\infty$ as well as respectively the subordinate matrix norm $\|\cdot\|$ is replaced by the subordinate matrix norm $\|\cdot\|_\infty$. This is due to the fact that the Euclidean norm dominates the infinite norm, and in the result from Neyshabur et al. (2018) it is sufficient to get a bound involving the infinite norm.

C.2.1 A PAC-Bayes Bound for Stable SRNs

Theorem 3.2 (Stable SRN). *Let $\mathcal{R}_P$ be a stable SRN (see Definition 2.3). Then for all $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ and all $X^T \in \mathcal{X}$ we have:*

$$\left\|\tilde{h}_T - h_T\right\| \leq \frac{1}{1 - \|W\|}\|\varepsilon_d\|.$$

Proof. Let $\mathcal{R}_P$ be a SRN with $\|W\| < 1$ Let $\varepsilon = (\varepsilon_o, \varepsilon_d) \sim \mathcal{N}(0, \rho^2 I_{d+o})$. Let $X^T \in \mathcal{X}$ we have, in Lemma C.5 we prove that:

$$\|\tilde{h}_T - h_T\| \leq \left(\sum_{s=0}^{T-1} \|W\|^s\right)\|\varepsilon_d\|.$$

By hypothesis $\|W\| < 1$, thus:

$$\left(\sum_{s=0}^{T-1} \|W\|^s\right) \leq \left(\sum_{s=0}^{\infty} \|W\|^s\right) = \frac{1}{1 - \|W\|}.$$

Therefore we can derive a bound on $\|\tilde{h}_T - h_T\|$ that is independent of T:

$$\|\tilde{h}_T - h_T\| \leq \frac{1}{1 - \|W\|}\|\varepsilon_d\|.$$

Thus, we can apply Theorem 3.1 with $\mathcal{C} = \frac{1}{1-\|W\|}$ and any $\mathcal{E} \in \mathbb{R}_+$. As a result we obtain that, with probability $1 - \delta$ over the training samples of size m, we have:

$$L(\mathcal{R}_P) \leq \widehat{L}_\gamma(\mathcal{R}_P) + 4\sqrt{\frac{\frac{64\ln(4)}{C}\left(DB^2 + dD + D\right)\left(\frac{\|V\|+1-\|W\|}{\gamma(1-\|W\|)}\right)^2 \|P\|_{\text{Fro}}^2 + \ln\frac{6m}{\delta}}{m-1}}.$$

□

C.2.2 A PAC-Bayes Bound for β-saturated SRNs

Theorem 3.3 (β-saturated SRN). *Let $\mathcal{R}_P$ be a β-saturated SRN with $\|W\| > 1$ and such that $\eta = \beta - z > 0$, where $z = \sqrt{1 - \frac{1}{\|W\|}}$. Let $t \in]0, 1[$, $\Delta = \tanh'(\tanh^{-1}(z) + \frac{t\eta}{1-z^2})$ and $\nabla = \frac{(1-t)\eta}{1-z^2}$, then for any $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\|\varepsilon\| \leq \nabla(1 - \Delta\|W\|)$ we have:*

$$\left\|\tilde{h}_T - h_T\right\| \leq \frac{1}{1 - \Delta\|W\|}\|\varepsilon_d\|.$$

Proof. Let $\mathcal{R}_P$ be a β-saturated SRN with $\|W\| > 1, z = \sqrt{1 - \frac{1}{\|W\|}}$, $\beta > z$ and $\eta = \beta - z$. We also suppose that $\|V\| \geq 1$ and we set a margin constant $\gamma > 0$ so that it is possible to choose $0 < t < 1$ such that:

- $\nabla := \frac{(1-t)\eta}{1-z^2} = \gamma/4$,
- $\Delta := \tanh'\left(\tanh^{-1}(z) + \frac{t\eta}{1-z^2}\right)$.

By Lemma C.6 we have the following expression:

$$\left\|\tilde{h}_T - h_T\right\| = \left\|\sum_{s=1}^{T-1} \Lambda(c_T)\left(\prod_{l=1}^{T-s} W\Lambda(c_{T-l})\right)\varepsilon_d + \Lambda(c_T)\varepsilon_d\right\|.$$

In Lemma C.11 we proved that for $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\|\varepsilon_d\| < \nabla(1 - \Delta\|W\|)$ we have:

$$\begin{aligned}\left\|W\sum_{s=1}^{k-1} \Lambda(c_k)\left(\prod_{l=1}^{k-s} W\Lambda(c_{k-l})\right)\varepsilon_d + W\Lambda(c_k)\varepsilon_d + \varepsilon_d\right\| &\leq \sum_{s=0}^{\infty}\left(\Delta\|W\|\right)^s\|\varepsilon_d\| \\ &\leq \frac{\|\varepsilon_d\|}{1 - \Delta\|W\|}.\end{aligned}$$

Moreover, we also showed that $\Delta\|W\| < 1$. Thus, for $k \to \infty$ the series converges. This explains the previous inequalities. We can now apply Theorem 3.1 with $\mathcal{C} = \frac{1}{1-\Delta\|W\|}$ that holds for $\|\varepsilon_d\| < \nabla(1-\Delta\|W\|)$. The way we define $\nabla = \gamma/4$ combined with the fact that $1-\Delta\|W\| = \mathcal{C}^{-1}$ we can assert that $\frac{\gamma}{4\|V\|(\mathcal{C}+1)} \leq \nabla(1-\Delta\|W\|)$. This allows us to apply Theorem 3.1 leading us to the following bound that holds with probability $1-\delta$ over the training sets of size m we have:

$$L(\mathcal{R}_P) \leq \widehat{L}_\gamma(\mathcal{R}_P) + 4\sqrt{\frac{\frac{64\ln(4)}{C}\left(DB^2+dD+D\right)\left(\frac{\|V\|+1-\Delta\|W\|}{\gamma(1-\Delta\|W\|)}\right)^2\|P\|_{\text{Fro}}^2 + \ln\frac{6m}{\delta}}{m-1}}.$$

□

C.2.3 A PAC-Bayes Bound for Desynchronised Sliding β-Saturated SRNs

We start with a result that is an upper bound on the amplification that the noise can experience in a fuzzy SRN. This result is closely related to Lemma C.5 with the difference to not consider the same norm and not unrolling the entire recurrence.

Lemma C.16 (Maximal coordinate amplification)**.** *Let $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$ respectively a SRN and its fuzzy version, $X^T \in \mathcal{X}$ a sequence and $S \geq 1$ an integer. For all $k \in \{1,\dots,T-S\}$ we have:*

$$\|\tilde{h}_{k+S} - h_{k+S}\|_\infty \leq \|W\|_\infty^S \|\tilde{h}_k - h_k\|_\infty + \|\varepsilon_d\|_\infty \sum_{j=0}^{S-1} \|W\|_\infty^j$$

Proof. Let $\mathcal{R}_P$ and $\mathcal{R}_P^\varepsilon$ be and SRN and its fuzzy version respectively, $X^T \in \mathcal{X}$ a sequence and $S \geq 1$ an integer. Let $k \in \{1,\dots,T-S\}$. The proof is a straightforward application of triangle inequality and operator norm applied recursively S. We start by writing $\tilde{h}_s$ as a function of h_s with $s := k+S$:

$$\begin{aligned}
\tilde{h}_s &= \sigma\left(Ux_s + W\tilde{h}_{s-1} + b + \varepsilon_d\right) \\
&= \sigma\left(Ux_s + Wh_{s-1} + b + W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right) \\
&= \sigma\left(Ux_s + Wh_{s-1} + b\right) + \Lambda(c_s)\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right)
\end{aligned}$$

by Lemma C.18 and the definition of c_s. We then have:

$$\tilde{h}_s = h_s + \Lambda(c_s)\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right).$$

Hence, we obtain:

$$\tilde{h}_s - h_s = \Lambda(c_s)\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right).$$

We set $i \in \{1,\dots,d\}$ (d being the dimension of the hidden state vector), and we denote W_i the i^{th} row of a matrix W. We observe:

$$\left|\left(\tilde{h}_s - h_s\right)[i]\right| = \left|\Lambda(c_s)\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right)[i]\right|.$$

The matrix $\Lambda(c_s)$ being diagonal with entries in $[0,1]$ we obtain:

$$\begin{aligned}
\left|\Lambda(c_s)\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right)[i]\right| &\leq \left|\left(W\left(\tilde{h}_{s-1} - h_{s-1}\right) + \varepsilon_d\right)[i]\right| \\
&\leq \left|\langle W_i, \tilde{h}_{s-1} - h_{s-1}\rangle + \varepsilon_d[i]\right| \\
&\leq \left|\langle W_i, \tilde{h}_{s-1} - h_{s-1}\rangle\right| + \|\varepsilon_d\|_\infty \\
&\leq \|W_i\|_1 \|\tilde{h}_{s-1} - h_{s-1}\|_\infty + \|\varepsilon_d\|_\infty \\
&\leq \|W\|_\infty \|\tilde{h}_{s-1} - h_{s-1}\|_\infty + \|\varepsilon_d\|_\infty.
\end{aligned}$$

The projection operator $\mathbb{R}^d \ni h :\to h[i] \in \mathbb{R}$ is linear and $Wh[i]$ is the scalar product between the i^{th} row of the matrix W and the vector h. We apply the Hölder's inequality and bound $\|W_i\|$ with $\max_{1\leq i\leq d}\|W_i\| = \|W\|_\infty$. One can remark that these arguments apply to all $i \in \{1,\dots,d\}$, hence:

$$\|\tilde{h}_s - h_s\|_\infty \leq \|W\|_\infty \|\tilde{h}_{s-1} - h_{s-1}\|_\infty + \|\varepsilon_d\|_\infty.$$

Moreover this reasoning does not depend on s, thus one can recursively repeat it $S-1$ more times and obtain:

$$\|\tilde{h}_{k+S} - h_{k+S}\|_\infty \leq \|W\|_\infty^S \|\tilde{h}_k - h_k\|_\infty + \|\varepsilon_d\|_\infty \sum_{j=0}^{S-1} \|W\|_\infty^j$$

□

Theorem 3.4 (desynchronized sliding β-saturated SRN)**.** *Let $\mathcal{R}_P$ be a desynchronized sliding β-saturated SRN with $\|W\|_\infty \geq 1$, $F \geq 1$ the max window verifying $\eta = \beta - z > 0$, where $z = \sqrt{1 - \frac{1}{2\|W\|_\infty^F}}$. Let $t \in]0,1[$, $\Delta = \tanh'(\tanh^{-1}(z) + \frac{t\eta}{1-z^2})$ and $\nabla = \frac{(1-t)\eta}{1-z^2}$, then for any $\varepsilon = (\varepsilon_o, \varepsilon_d) \in \mathbb{R}^{o+d}$ such that $\|\varepsilon\| \leq \left(\frac{\nabla\Delta}{4\sum_{i=0}^F \|W\|_\infty^i}\right)$ we have:*

$$\left\|\tilde{h}_T - h_T\right\|_\infty \leq \|\varepsilon_d\|_\infty \sum_{j=0}^{F} \|W\|_\infty^j + \frac{\nabla}{4}.$$

Proof. With the notations from the theorem above, we choose any $i \in \{1,\dots,d\}$ and a sequence $X^T = \{x_k\}_{k=1}^T \in \mathcal{X}$. There are two points that must be proven simultaneously: 1) That for all $X^T \in \mathcal{X}$ and for all $1 \leq s \leq T$ we have $\left\|\tilde{h}_k - h_k\right\|_\infty \leq \|\varepsilon_d\|_\infty \sum_{j=0}^F \|W\|_\infty^j + \frac{1}{4}\nabla$; 2) that $\mathcal{R}_P^\varepsilon$ has a similar saturation pattern as $\mathcal{R}_P$ (see Definition C.8). We will use a proof by induction on l that we have 1) for the any iteration $s \in \{lF,\dots,(l+1)F\}$ and we will use 1) to prove 2). This will allow us to refine the bound in 1).

For all $1 \leq k \leq T$, we have that:

$$\begin{aligned}
\tilde{h}_k &= \sigma\left(Ux_k + W\tilde{h}_{k-1} + b + \varepsilon_d\right) \\
&= \sigma\left(Ux_k + Wh_{k-1} + b + \varepsilon_d + W(\tilde{h}_{k-1} - h_{k-1})\right) \\
&= \sigma\left(Ux_k + Wh_{k-1} + b\right) + \sigma'\left(c_k\right) \odot \left(\varepsilon_d + W(\tilde{h}_{k-1} - h_{k-1})\right),
\end{aligned}$$

where the second term and $\sigma'\left(c_k\right)$ are introduced by applying Lemma C.18. We then have:

$$\tilde{h}_k = h_k + \sigma'\left(c_k\right) \odot \left(\varepsilon_d + W(\tilde{h}_{k-1} - h_{k-1})\right). \tag{11}$$

We now begin the induction by the initialization with $l = 1$. For $s \leq F$, by Lemma C.16 combined with the fact that by definition of a fuzzy SRN, $\tilde{h}_0 = h_0$, we have:

$$\begin{aligned}
\|\tilde{h}_s - h_s\|_\infty &\leq \|\varepsilon_d\|_\infty \sum_{j=0}^{s-1} \|W\|_\infty^j \\
&\leq \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j.
\end{aligned}$$

Because of Equation 11, we need a bound on $\|\varepsilon_d + W(\tilde{h}_{s-1} - h_{s-1})\|_\infty$:

$$\begin{aligned}
\|\varepsilon_d + W(\tilde{h}_{s-1} - h_{s-1})\|_\infty &\leq \|\varepsilon_d\|_\infty + \|W\|_\infty \|\tilde{h}_{s-1} - h_{s-1})\|_\infty \\
&\leq \|\varepsilon_d\|_\infty + \|W\|_\infty \left(\|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j\right) \\
&= \|\varepsilon_d\|_\infty \sum_{j=0}^{F} \|W\|_\infty^j < \frac{\nabla\Delta}{4} < \frac{\nabla\Delta}{2}.
\end{aligned}$$

We get $\|\varepsilon_d\|_\infty \sum_{j=0}^{F} \|W\|_\infty^j < \frac{\nabla}{4}$ from the hypothesis $\|\varepsilon\| \leq \left(\frac{\nabla\Delta}{4\sum_{i=0}^{F}\|W\|_\infty^i}\right)$. We proved that for all $0 \leq s \leq F$ the noise accumulated up to s^{th} iteration in the fuzzy SRN $\mathcal{R}_P^\varepsilon$ does not exceed ∇, therefore by Lemma C.10 we can assert that $\mathcal{R}_P^\varepsilon$ will have a similar saturation pattern as $\mathcal{R}_P$ up to iteration F. We have simultaneously

- $\|\tilde{h}_F - h_F\|_\infty \leq \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j + \frac{\nabla}{4}$
- $\|\tilde{h}_F - h_F\|_\infty \leq \frac{\nabla\Delta}{2}$

because $\|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j \leq \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j + \frac{\nabla}{4}$. The initialization is thus proven.

Now we suppose that for $l \geq 1$, we have $\|\tilde{h}_{lF} - h_{lF}\|_\infty \leq \frac{\nabla\Delta}{2}$. We need to show that $\|\tilde{h}_{(l+1)F} - h_{(l+1)F}\|_\infty \leq \frac{\nabla\Delta}{2}$. We start by applying Lemma C.16 to prove that $\mathcal{R}_P^\varepsilon$ will have a similar saturation pattern as $\mathcal{R}_P$ for the iterations $lF \leq s \leq (l+1)F$. By Lemma C.16:

$$\|\tilde{h}_s - h_s\| \leq \|W\|_\infty^{s-lF} \|\tilde{h}_{lF} - h_{lF}\|_\infty + \|\varepsilon_d\| \sum_{j=0}^{s-lF-1} \|W\|_\infty^j.$$

By induction hypothesis, we have $\|\tilde{h}_{lF} - h_{lF}\|_\infty \leq \frac{\nabla\Delta}{2}$. Therefore we obtain:

$$\begin{aligned}\|\tilde{h}_s - h_s\|_\infty &\leq \|W\|_\infty^{s-lF} \frac{\nabla\Delta}{2} + \|\varepsilon_d\| \sum_{j=0}^{s-lF-1} \|W\|_\infty^j \\ &\leq \|W\|_\infty^{(l+1)F-lF} \frac{\nabla\Delta}{2} + \|\varepsilon_d\| \sum_{j=0}^{F-1} \|W\|_\infty^j,\end{aligned}$$

because $\|W\| \geq 1$, and $lF \leq s \leq (l+1)F$. This gives us:

$$\|\tilde{h}_s - h_s\|_\infty \leq \underbrace{\|W\|_\infty^{F-1} \frac{\nabla\Delta}{2}}_{\leq \frac{\nabla}{4}} + \underbrace{\|\varepsilon_d\| \sum_{j=0}^{F-1} \|W\|_\infty^j}_{\frac{\nabla\Delta}{4}} \leq \frac{\nabla}{2}.$$

By the assumption on β and Lemma C.22, we know that $\|W\|^F \Delta \leq 1/2$ thus $\|W\|^F \frac{\nabla\Delta}{2} \leq \frac{\nabla}{4}$ and, since by assumption we have $\|\varepsilon\| \leq \left(\frac{\nabla\Delta}{4\sum_{i=0}^{F}\|W\|_\infty^i}\right)$. We thus have:

$$\|\tilde{h}_s - h_s\|_\infty \leq \frac{\nabla}{4} + \frac{\nabla\Delta}{4} \leq \frac{\nabla}{2},$$

because $\Delta \leq 1$ and hence $\frac{\nabla\Delta}{4} \leq \frac{\nabla}{4}$. By leveraging the previously proven identity:

$$\tilde{h}_s = h_s + \sigma'(c_s) \odot \left(\varepsilon_d + W(\tilde{h}_{s-1} - h_s)\right),$$

we can assert that $\mathcal{R}_P^\varepsilon$ will have a similar saturation pattern as $\mathcal{R}_P$ up to iteration $(l+1)F$. This is due to the fact that the perturbation never exceeds ∇.

From this, we are going to prove an upper bound on $\|\tilde{h}_{(l+1)F} - h_{(l+1)F}\|_\infty$ with the knowledge that one saturation has occurred in the window $\{lF, \ldots, (l+1)F\}$. For a coordinate $1 \leq i \leq d$, we do not know at which iteration s, $lF \leq s \leq (l+1)F$ a saturation occurs. To estimate the worst possible case, we will reason as follows: we fix a coordinate $1 \leq i \leq d$ and chose $s \in \{\varsigma_1, \ldots, \varsigma_t\}_i \cap \{lF, \ldots, (l+1)F\}$, where $\{\varsigma_1, \ldots, \varsigma_t\}_i$ is the set of saturating iterations of coordinate i. We then know that a saturation will occur on i^{th} coordinate at iteration s. Lemma C.16 is then applied for iterations between lF and $s-1$ in order to have an upper bound on the accumulated noise up to the saturating iteration. We then use Equation 11 to extract a more accurate bound for the iteration s. Finally we use that $s \leq (l+1)F$ and $\Delta \leq 1$ to extract a bound independent of s and thus independent of i.

Formally, we begin with bounding the noise accumulated between the iteration lF and $s-1$ by using Lemma C.16:

$$\begin{aligned}\|\tilde{h}_{s-1} - h_{s-1}\|_\infty &\leq \|W\|_\infty^{s-lF-1}\|\tilde{h}_{lF} - h_{lF}\| + \|\varepsilon_d\|_\infty \sum_{j=0}^{s-lF-2} \|W\|_\infty^j \\ &\leq \|W\|_\infty^{s-lF-1}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{s-lF-2} \|W\|_\infty^j \text{ by induction hypothesis.}\end{aligned}$$

Coordinate i will saturate on iteration s:

$$\begin{aligned}\left|\left(\tilde{h}_s - h_s\right)[i]\right| &= \sigma'(c_s[i]) \cdot \left|\left(\varepsilon_d + W(\tilde{h}_{s-1} - h_s)\right)[i]\right| \\ &\leq \Delta \cdot \left|\left(\varepsilon_d + W(\tilde{h}_{s-1} - h_s)\right)[i]\right| \qquad \text{by Lemma C.10} \\ &\leq \Delta \cdot \|\varepsilon_d + W(\tilde{h}_{s-1} - h_s)\|_\infty \\ &\leq \Delta \cdot \left(\|\varepsilon_d\|_\infty + \|W\|_\infty\|\tilde{h}_{s-1} - h_s\|_\infty\right).\end{aligned}$$

Now if we plug in the bound on $\|\tilde{h}_{s-1} - h_s\|_\infty$ we obtain:

$$\begin{aligned}\left|\left(\tilde{h}_s - h_s\right)[i]\right| &\leq \Delta \cdot \left(\|\varepsilon_d\|_\infty + \|W\|_\infty\left(\|W\|_\infty^{s-lF-1}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{s-lF-2} \|W\|_\infty^j\right)\right) \\ &= \Delta \cdot \left(\|W\|_\infty^{s-lF}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{s-lF-1} \|W\|_\infty^j\right).\end{aligned}$$

Since $s \leq (l+1)F$ and that $\Delta \leq 1$, we can derive a bound on $\left|\left(\tilde{h}_s - h_s\right)[i]\right|$ that does not depend on s:

$$\begin{aligned}\Delta \cdot \left(\|W\|_\infty^{s-lF}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{s-lF-1} \|W\|_\infty^j\right) &\leq \Delta \cdot \left(\|W\|_\infty^{(l+1)F-lF}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{(l+1)F-lF-1} \|W\|_\infty^j\right) \\ &= \Delta \cdot \left(\|W\|_\infty^{F}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j\right) \\ &\leq \Delta\|W\|_\infty^{F}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j.\end{aligned}$$

Consequently we have:

$$\left|\left(\tilde{h}_s - h_s\right)[i]\right| \leq \Delta\|W\|_\infty^{F}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j.$$

Note that these arguments do not depend on (i, s), all that was assumed is that a coordinate $1 \leq i \leq d$ will saturate during an iteration $s \in \{lF, \ldots, (l+1)F\}$. These assumptions are fulfilled because we showed above that $\mathcal{R}_P^\varepsilon$ has a similar saturation pattern as $\mathcal{R}_P$ on $\{lF, \ldots, (l+1)F\}$. Therefore we can assert that:

$$\|\tilde{h}_{(l+1)F} - h_{(l+1)F}\|_\infty \leq \Delta\|W\|_\infty^{F}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j.$$

Now we exploit the assumption on Δ and $\|\varepsilon\|$ to obtain:

$$\Delta\|W\|_\infty^{F}\frac{\nabla\Delta}{2} + \|\varepsilon_d\|_\infty \sum_{j=0}^{F-1} \|W\|_\infty^j \leq \frac{\nabla\Delta}{4} + \frac{\nabla\Delta}{4} = \frac{\nabla\Delta}{2},$$

obtained from Lemma C.22 that implies $\Delta\|W\|_\infty^F \leq 1/2$. Hence:

$$\|\tilde{h}_{(l+1)F} - h_{(l+1)F}\|_\infty \leq \frac{\nabla\Delta}{2}.$$

This finishes the induction and thus the proof.

□

C.3 Technical Results

This section is dedicated to a series of technical lemmas.

Lemma C.18 (Vector variant of the mean value theorem). *Let $y, \zeta \in \mathbb{R}^d$ and let $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable activation function. Then there exists $c \in \mathbb{R}^d$ such that:*

$$\sigma(y + \zeta) = \sigma(y) + \sigma'(c) \odot \zeta$$

and $\in y[j] < c[j] < (y + \zeta)[j]$ for all $1 \leq j \leq d$.

Proof. Let $y, \zeta \in \mathbb{R}^d$, and let $1 \leq j \leq d$. By the mean value theorem there exists c_j, $y[j] < c_j < (y + \zeta)[j]$ such that:

$$\sigma\big((y + \zeta)[j]\big) = \sigma(y[j]) + \sigma'(c_j) \cdot \zeta[j].$$

We can repeat this operation for any coordinate $1 \leq j \leq d$ and create a vector $c = (c_1, \ldots, c_d)$ such that:

$$\sigma(y + \zeta) = \sigma(y) + \sigma'(c) \odot \zeta.$$

□

The following two lemmas are classical results about Gaussian multivariate variables.

Lemma C.19 (A bound on the variance of the noise). *Let $\Upsilon = (\Upsilon_{k,l})$ be a random matrix of size $p \times q$ where $\Upsilon_{k,l} \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$, and let $x \in \mathbb{R}^q$ with $\|x\| \leq B$. Then for all $1 \leq j \leq p$ we have:*

$$\mathbb{V}\Big[(\Upsilon x[j])\Big] \leq \rho^2 B^2.$$

Moreover, the coordinates of the random vector Υx are independent.

Proof. Following the notations introduced in the statement of the lemma, we have:

$$\Upsilon x[j] = \Upsilon_j^T x = x^T \Upsilon_j$$

where Υ_j denotes the j^{th} row of the matrix Υ. By hypothesis the matrix Υ is centered, thus the random variable $\Upsilon_j^T x$ is also centered. Therefore we have:

$$\begin{aligned}
\mathbb{V}\Big[\Upsilon x[j]\Big] &= \mathbb{E}\Big[(\Upsilon_j^T x)^2\Big] \\
&= \mathbb{E}\Big[x^T \Upsilon_j \Upsilon_j^T x\Big] \\
&= x^T \mathbb{E}\Big[\Upsilon_j \Upsilon_j^T\Big] x \\
&= x^T (\rho^2 I_q) x = \rho^2 \|x\|^2 \leq \rho^2 B^2.
\end{aligned}$$

Since the parameters of Υ are iid, the coordinates of Υx are independent. □

Lemma C.20 (Concentration of Gaussian vectors). *Let $\varepsilon \sim \mathcal{N}(0, \nu^2 I_d)$ a Gaussian random vector with $d > 2$, and let $t \geq 0$. If*

$$\nu^2 = \frac{t^2}{d} \left(\frac{1 - \sqrt{4\ln(2)}}{1 - 12\ln(2)} \right)^2,$$

then:

$$\mathbb{P}\Big[\|\varepsilon\| > t\Big] \leq \frac{1}{2}.$$

Proof. Let $\varepsilon \sim \mathcal{N}(0, \nu^2 I_d)$ a Gaussian random vector. We have the following concentration bound from the book of Vershynin (2018):

$$\mathbb{P}\left[\|\varepsilon\| > t + \mathbb{E}\Big[\|\varepsilon\|\Big]\right] \leq e^{-t^2/4d\nu^2}.$$

Let $X \sim \mathcal{N}(0, I_d)$ the random variable $\|X\|$ follows the χ law of degree d. Therefore:

$$\mathbb{E}\Big[\|X\|\Big] = \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}$$

where Γ is the Gamma function, defined by:

$$\Gamma : \omega \mapsto \int_0^\infty t^{\omega-1}e^{-t}dt.$$

In the article from Cristinel (2010) we can find the following inequality for all $x > 0$:

$$\sqrt{x + \frac{1}{4}} < \frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})}$$

which implies the inequality:

$$\sqrt{d-2} < \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} = \mathbb{E}\Big[\|X\|\Big].$$

Since $\mathbb{E}\Big[\|\varepsilon\|\Big] = \nu\mathbb{E}\Big[\|X\|\Big]$ we obtain:

$$\mathbb{E}\Big[\|\varepsilon\|\Big] > \nu\sqrt{d-2}$$

and therefore:

$$\mathbb{P}\left[\|\varepsilon\| > t\right] \leq e^{-\left(t-(\sqrt{d-2})\nu\right)^2/4d\nu^2}.$$

We need to arrange ν such that:

$$\mathbb{P}\left[\|\varepsilon\| > t\right] \leq \frac{1}{2},$$

for that purpose we need to solve:

$$e^{-\left(t-(\sqrt{d-2})\nu\right)^2/4d\nu^2} = \frac{1}{2},$$

which is equivalent to solving the quadratic equation:

$$(d - 2 - 4\ln(2)d)\nu^2 - (2t\sqrt{d-2})\nu + t^2 = 0.$$

The above equation has a unique positive solution:

$$\nu = t\frac{\sqrt{d-2} - \sqrt{d4\ln(2)}}{(d-2) - 4\ln(2)d}.$$

In our situation a lower bound for ν is enough, thus:

$$\begin{aligned}
t\frac{\sqrt{d-2} - \sqrt{d4\ln(2)}}{(d-2) - 4\ln(2)d} &\geq t\frac{\sqrt{d-2} - \sqrt{(d-2)4\ln(2)}}{(d-2) - 4\ln(2)d} \\
&= t\left(\frac{\sqrt{d-2}}{d-2}\right)\left(\frac{1 - \sqrt{4\ln(2)}}{1 - 4\ln(2)\frac{d}{d-2}}\right) \\
&= \frac{t}{\sqrt{d-2}}\left(\frac{1 - \sqrt{4\ln(2)}}{1 - 4\ln(2)\frac{d}{d-2}}\right) \\
&\geq \frac{t}{\sqrt{d}}\left(\frac{1 - \sqrt{4\ln(2)}}{1 - 12\ln(2)}\right).
\end{aligned}$$

The last inequality is due to the fact that $d > d-2$ and $\frac{d}{d-2} \leq 3$ for $d > 2$. As a result, by setting

$$\nu = \frac{t}{\sqrt{d}} \left(\frac{1 - \sqrt{4\ln(2)}}{1 - 12\ln(2)} \right)$$

we have:

$$\mathbb{P}\Big[\|\varepsilon\| > t \Big] \leq \frac{1}{2}.$$

□

Lemma C.21. *Let $C \geq 1$, we set $z_a := \sqrt{1 - \frac{a}{C}}$, where $0 < a < 1$. Then, for all $\alpha > 0$ we have:*

$$\tanh'\big(\tanh^{-1}(z_a) + \alpha\big)C < a,$$

Proof. From the notation above one can assert:

$$\begin{aligned} \tanh'\big(\tanh^{-1}(z_a)\big) &= 1 - \tanh^2\big(\tanh^{-1}(z_a)\big) \\ &= 1 - z_a^2 \\ &= \frac{a}{C}, \end{aligned}$$

hence $\tanh'\big(\tanh^{-1}(z_a)\big)C = a$. The function $\tanh'$ being strictly decreasing on the interval $[0, +\infty[$, we can claim that for all $\alpha > 0$ we have:

$$\tanh'\big(\tanh^{-1}(z_a) + \alpha\big)C < a$$

□

Lemma C.22 (Converging geometric sum)**.** *Let $W \in \mathbb{R}^{d\times d}$ be a matrix such that $\|W\| > 1$ and $F \geq 1$, we set $z := \sqrt{1 - \frac{1}{\|W\|^F}}$. Then, for all $\alpha > 0$ we have:*

$$\tanh'\big(\tanh^{-1}(z) + \alpha\big)\|W\|^F < 1,$$

and thus

$$\sum_{s=0}^{\infty} \Big(\tanh'\big(\tanh^{-1}(z) + \alpha\big)\|W\|^F \Big)^s = \frac{1}{1 - \tanh'\big(\tanh^{-1}(z) + \alpha\big)\|W\|^F} < \infty.$$

Proof. Let $W \in \mathbb{R}^{d\times d}$ be a matrix such that $\|W\| > 1$ and $F \geq 1$. We set $z := \sqrt{1 - \frac{1}{\|W\|^F}}$. Then, for all $0 < \alpha$ we have:

$$\begin{aligned} \tanh'(\tanh^{-1}(z)) &= 1 - \tanh^2(\tanh^{-1}(z)) \\ &= 1 - \sqrt{1 - \frac{1}{\|W\|^F}}^2 \\ &= \frac{1}{\|W\|^F}. \end{aligned}$$

Secondly, the function $\tanh'$ is a strictly decreasing function on $(0, +\infty)$, thus for all $0 < \alpha < 1$ we have

$$\begin{aligned} \tanh'(\tanh^{-1}(z) + \alpha) &< \tanh'(\tanh^{-1}(z)) \\ \tanh'(\tanh^{-1}(z) + \alpha)\|W\|^F &< \tanh'(\tanh^{-1}(z))\|W\|^F = 1. \end{aligned}$$

Finally the limit of the geometric series is the usual computation. □