Appendices

A More Training Details

We collect the training hyper-parameters in the following table.

	sl-MNIST	SHD	PTB
batch size	256	256	32
weight decay	0.0	0.1	0.1
gradient norm	1.0	1.0	1.0
train/val/test	45k/5k/10k samples	8k/1k/2k samples	930k/74k/82k words
learning rate	$3.16 \cdot 10^{-4}$	$3.16 \cdot 10^{-4}$	$3.16 \cdot 10^{-5}$
layers width	128, 128	256, 256	1700, 300
label smoothing	0.1	0.1	0.1
time step repeat	2	2	2
SELT factor	0.8	0.436	4.595

The learning rates were chosen after a grid search fixing dampening and sharpness to 1. The learning rates considered are in the set $\{10^{-2}, 3.16 \cdot 10^{-3}, 10^{-3}, 3.16 \cdot 10^{-4}, 10^{-4}, 3.16 \cdot 10^{-5}, 10^{-5}\}$. The results of the grid search are reported in figure 2. The learning rate chosen for the rest of the paper was the one that made all the shapes perform reasonably well, rectangular included. This mostly resulted in a suboptimal learning rate only for the derivative of the fast sigmoid, which still out-performed the rest in the sl-MNIST and SHD, and performed comparatively on the PTB.

We train with crossentropy loss, the AdaBelief optimizer (Zhuang et al., 2020), Stochastic Weight Averaging (Izmailov et al., 2018) and Decoupled Weight Decay (Loshchilov and Hutter, 2019). For the BiGamma distribution, we choose $\alpha = 5$ and $\beta = \sqrt{\alpha(\alpha+1)} = 5.47$ to have a variance of 1. For the PTB task, the input passes through an embedding layer before passing to the first layer, and the output of the last layer is multiplied by the embedding to produce the output, removing the need for the readout (Woźniak et al., 2020; Radford et al., 2018).

Notice that we do not implement forced refractory periods that would prevent the neuron from firing too fast, as sometimes done in the neuromorphic literature, since we want to reduce the non differentiable steps in the system. Thus, $p = 1$ is possible if the inputs are strong and frequent enough.

B Neuron Model Complexity

The energy consumed per layer can be used as a metric of neuron complexity, as done in (Yin et al., 2021; Hunger, 2005).

Neural model	Energy (Complexity)
LIF	$(mnp_{l-1} + nnp_l)E_{AC} + nE_{MAC}$
ALIF	$(mnp_{l-1} + nnp_l + 2np_l)E_{AC} + 3nE_{MAC}$
LSTM	$4(mn + nn)E_{MAC} + 17nE_{MAC}$
sLSTM	$4(mnp_{l-1} + nnp_l)E_{AC} + 3np_lE_{AC}$

Table 1: **Neuron complexity.** We use the energy consumed per layer as a metric of neuron complexity (Yin et al., 2021; Hunger, 2005). We use $n = n_l$ and $m = n_{l-1}$ as the width of the layer and its input, p_l for the firing rate of the layer l. E_{MAC} is the energy cost of a multiply-accumulate operation and E_{AC} of an accumulate operation. As shown, $ALIF$ always results in a larger number of operations and energy consumption than LIF. For large networks, $n, m \gg 1$, the square terms dominates, and the $sLSTM$ results in 4 times more energy consumption.

C Interplay between SG and Initialization

We show how each initialization has a different preferred SG, and viceversa, how each SG has a different preferred initialization in figure 6. Best mean across SG is achieved by the Orthogonal Normal initialization. Best mean across initializations is achieved by the derivative of the fast-sigmoid and the exponential SG.

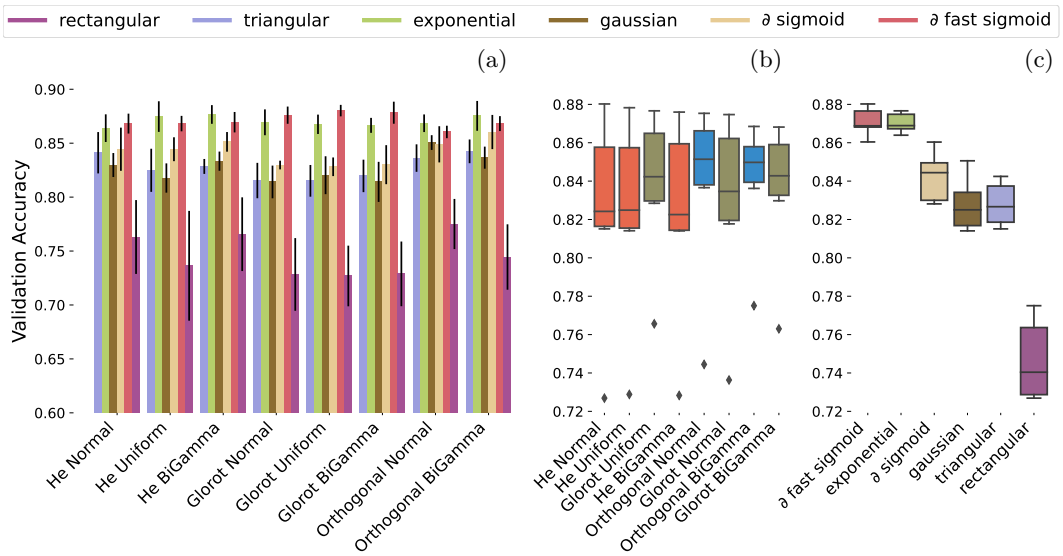

Figure 6: **Orthogonal initialization leads to higher accuracy on the LIF network.** Results for a LIF network trained on the SHD task. (a) Best accuracy when different SG trainings are initialized with different schemes. (b) Results aggregated across SG shapes. Overall, *He* initialization, in orange, achieves the best extreme values, but the *Orthogonal* achieves the best mean, in blue. The BiGamma distribution decreases the accuracy variance. (c) Results aggregated across initializations. Both, the derivative of the fast sigmoid and the exponential SG achieve the best mean accuracy.

D List of Surrogate Gradients shapes

We list here the shapes that we used in this article as surrogate gradients.

SG name	$f(v)$
triangular	$\max(1-\lvert v\rvert, 0)$
exponential	$e^{-2\lvert v\rvert}$
gaussian	$e^{-\pi v^2}$
∂ **sigmoid**	$4\ sigmoid(4\ v)\,(1-sigmoid(4\ v))$
∂ **fast-sigmoid**	$\frac{1}{(1+\lvert 2v\rvert)^2}$
rectangular	$\mathbb{1}_{\lvert v\rvert<\frac{1}{2}}$
q**-PseudoSpike** $(q>1)$	$\frac{1}{(1+\frac{2}{q-1}\lvert v\rvert)^q}$

Table 2: **Mathematical definitions of the surrogate gradients studied in this article.** Our Heaviside activation $\sigma(v) = \tilde{H}(v)$, where v is the centered voltage, has the SG $\sigma'(v) = \gamma f(\beta \cdot v)$, where β is the SG sharpness, γ the SG dampening, and f is the shape of choice. The constants, are chosen for the SG to have a maximal value of 1 and an area under the curve of 1.

E Detailed derivation of the conditions

We derive the constraints on the hyper-parameters that will lead the LIF to meet the conditions proposed at initialization. The LIF we will be using is defined by

$$\boldsymbol{y}_t = \boldsymbol{\alpha}_{decay}\boldsymbol{y}_{t-1}(1-\boldsymbol{x}_{t-1}) + \boldsymbol{i}_t \tag{1}$$

where $\boldsymbol{i}_t = W_{rec}\boldsymbol{x}_{t-1} + W_{in}\boldsymbol{z}_t + \boldsymbol{b}$, as described in the main text, and the multiplicative factor $(1-\boldsymbol{x}_{t-1})$ represents the reset mechanism.

E.1 Recurrent matrix mean sets the firing rate (I)

We show how condition (I) leads to a constraint on the mean of the recurrent connectivity with a LIF neuron model, that will lead the network to meet that condition at initialization.

Lemma 1. *Applying condition (I), which states that we want $Median[v] = 0$, to an LIF network, and further assuming $\overline{w}_{in} = 0, b = 0$, the approximation $Mean[v] \approx Median[v]$, and constant $\bar{i}_t$ over time, it results in the constraint*

$$\overline{w}_{rec} = \frac{1}{n_{rec}-1}(2-\overline{\alpha}_{decay})\overline{\vartheta} \tag{2}$$

Proof. First we show that $Median[v] = 0 \implies Mean[x] = 1/2$, where $x = \tilde{H}(v)$. In equation 4 we write the marginal distribution of $p(x) = \int p(x|v)p(v)dv$, and the double integral is represented with one integration symbol. Then, we notice that x has a deterministic dependence on v, $x = H(v)$, which proprbabilistically is described by the delta function $p(x|v) = \delta(x - H(v))$. Then, we integrate over x, and in the last equation we notice that integrating with respect to the Heaviside is equivalent to restricting the integration limits from zero to infinity.

$$
\begin{aligned}
Mean[x] &= \int xp(x)dx && (3)\\
&= \int xp(x|v)p(v)dxdv && (4)\\
&= \int xp(v)\delta(x - H(v))dxdv && (5)\\
&= \int p(v)dvH(v) && (6)\\
&= \int_0^\infty p(v)dv && (7)
\end{aligned}
$$

If $Median[v] = 0$, half of it's probability mass is on each side of 0, so the last integral is equal to $1/2$, QED.

Since working with medians is mathematically harder than working with means, we assume that $Mean[v] \approx Median[v]$, with the caveat that it will make the result approximate. To justify that they are similar, it can be shown that for a unimodal distribution $v \sim p(v)$ with the first two moments defined, we have $|Mean[v] - Median[v]| \leq \sqrt{0.6Var[v]}$

We use the notation $\overline{x} = Mean[x]$ interchangeably. We calculate how the mean of the voltage elements is propagated through time, assuming the mean input current to remain constant over time $\bar{i}_t = \bar{i}$ at initialization, to simplify the mathematical development, and assuming per condition (I), that $\overline{x} = \overline{1 - x} = 1/2$ we have

$$
\begin{aligned}
\overline{y}_t =& \overline{\alpha}_{decay}\overline{(1 - \boldsymbol{x}_{t-1})}\overline{y}_{t-1} + \bar{i} && (8)\\
=& \frac{1}{2}\overline{\alpha}_{decay}\overline{y}_{t-1} + \bar{i} && (9)\\
=& \frac{1}{2}\overline{\alpha}_{decay}\Big(\frac{1}{2}\overline{\alpha}_{decay}\overline{y}_{t-2} + \bar{i}\Big) + \bar{i} && (10)\\
=& \frac{1}{2^{t-1}}\overline{\alpha}_{decay}^{t-1}\overline{y}_1 + \Big(\sum_{t'=0}^{t-2}\frac{1}{2^{t'}}\overline{\alpha}_{decay}^{t'}\Big)\bar{i} && (11)\\
=& \frac{1}{2^{t-1}}\overline{\alpha}_{decay}^{t-1}\overline{y}_1 + \frac{1 - \frac{1}{2^{t-1}}\overline{\alpha}_{decay}^{t-1}}{1 - \frac{1}{2}\overline{\alpha}_{decay}}\bar{i} && (12)
\end{aligned}
$$

where we used the fact that the same LIF definition applies to different time steps, the geometric series formula, and the fact that for independent random variables $E[XY] = E[X]E[Y]$. For $t \to \infty$ and using $0 < \overline{\alpha}_{decay} < 1$

$$
\begin{aligned}
\overline{y}_t =& \frac{1}{1 - \frac{1}{2}\overline{\alpha}_{decay}}\bar{i} && (13)\\
\overline{y}_t - \overline{\vartheta} =& \frac{1}{1 - \frac{1}{2}\overline{\alpha}_{decay}}\bar{i} - \overline{\vartheta} && (14)
\end{aligned}
$$

Assuming we want this condition to hold independently of the dataset, we set $Mean[W_{in}] = 0$, and assuming that we do not want to promote this behavior with fixed internal currents, but with the recurrent activity instead, then $\boldsymbol{b} = 0$.

We remark that we denote $Mean[W\boldsymbol{x}]$ as the mean vector whose element i is

$$
\begin{aligned}
Mean[W\boldsymbol{x}]_i =& Mean[\sum_{\substack{j=1\\ j\neq i}} w_{ij}x_j] && (15)\\
=& \sum_{\substack{j=1\\ j\neq i}} Mean[w_{ij}x_j] && (16)\\
=& \sum_{\substack{j=1\\ j\neq i}} Mean[wx] && (17)\\
=& (n_{rec}-1)Mean[wx] && (18)
\end{aligned}
$$

where the condition $j \neq i$ in the summand reminds that neurons are not connected to themselves in our recurrent architecture. In the first equality, the index i denoting the element in the vector, is equivalent as choosing the row i of W, so it is not necessary to specify it outside the square brakets. The equality before the last one is a consequence of considering any neuron as mutually independent to any other at initialization, as done by

Then,

$$
\begin{aligned}
Mean[y_t - \vartheta] =& \frac{1}{1-\frac{1}{2}\overline{\alpha}_{decay}}\Big((n_{rec}-1)\overline{w}_{rec}\overline{x}_{t-1}\Big) - \overline{\vartheta} && (19)\\
0 =& \frac{1}{1-\frac{1}{2}\overline{\alpha}_{decay}}(n_{rec}-1)\overline{w}_{rec}\overline{x}_{t-1} - \overline{\vartheta} && (20)\\
0 =& \frac{1}{1-\frac{1}{2}\overline{\alpha}_{decay}}(n_{rec}-1)\overline{w}_{rec}\frac{1}{2} - \overline{\vartheta} && (21)\\
(n_{rec}-1)\overline{w}_{rec} =& (2-\overline{\alpha}_{decay})\ \overline{\vartheta} && (22)\\
\overline{w}_{rec} =& \frac{1}{n_{rec}-1}\Big(2-\overline{\alpha}_{decay}\Big)\overline{\vartheta} && (23)
\end{aligned}
$$

where in the second line we applied condition (I) in the form of $Mean[v_t] \approx Median[v_t] = 0$, so $Mean[y_t - \vartheta] = 0$, and in the third line we applied again condition (I), $\overline{x}_t = 1/2$. In the main text we turn $\overline{\vartheta}, \overline{\alpha}_{decay} \to \vartheta, \alpha_{decay}$, since here we consider the more general case where those are as well random variables, and we simplify it in the main text for cleanliness, assuming they are constant.

□

We therefore found a constraint on the mean of the recurrent matrix initialization, that leads the LIF network to satisfy condition I at initialization. The constraint is equation 23 with $\overline{w}_{in} = 0$, and $\boldsymbol{b} = 0$.

E.2 Recurrent matrix variance can make recurrent and input voltages comparable (II)

We apply condition (II) to the LIF network, that gives us a constraint that the recurrent matrix has to meet at initialization for the condition to be true.

Lemma 2. *Applying condition (II), which states that we want* $Var[W_{rec}x_{t-1}] = Var[W_{in}z_t]$*, to an LIF network, and further assuming* $\overline{x} = 1/2$*, and* $\overline{w}_{in} = 0$*, it results in the constraint*

$$
Var[w_{rec}] = 2(Var[z_t] + \overline{z}_t^2)\frac{n_{in}}{n_{rec}-1}Var[w_{in}] - \frac{1}{2}\overline{w}_{rec}^2 \qquad (24)
$$

Proof. The second condition, is that the recurrent and the input contribution to the variance need to match

$$Var[W_{rec}x_{t-1}] = Var[W_{in}z_t] \tag{25}$$

where the variance is computed at each element, after the matrix multiplication is performed, following the method described in

$$\begin{aligned} Var[W\boldsymbol{x}]_i =& Var[\sum_{j=1} w_{ij}x_j] && (26)\\ =& \sum_{j=1} Var[w_{ij}x_j] && (27)\\ =& \sum_{j=1} Var[wx] && (28)\\ =& n_W Var[wx] && (29) \end{aligned}$$

The second and third equality are a consequence of considering any neuron as mutually independent to any other at initialization, as done by

Therefore the vector-wise condition II is equivalent to the element-wise

$$(n_{rec} - 1)Var[w_{rec}x_{t-1}] = n_{in}Var[w_{in}z_t] \tag{30}$$

Since the time dimension is averaged out, the time axis can be randomly shuffled, and the LIF activity is indistinguishable from a Bernouilli process through the mean and variance of the activity. Therefore if $\overline{x}_t = p$, we have $Var[x_t] = p(1-p)$ when averaged over time, with p the probability of firing. Therefore it is as well true that $\overline{x_t^2} = Var[x_t] + \overline{x}_t^2 = p$.

We apply the fact that for independent w, x

$$Var[wx] = \overline{w^2}\ \overline{x^2} - \overline{w}^2\overline{x}^2 \tag{31}$$

and assuming $\overline{w}_{in} = 0$ and $p = 1/2$ we have

$$\begin{aligned} Var[w_{rec}x_{t-1}] =& (Var[w_{rec}] + \overline{w}_{rec}^2)p - \overline{w}_{rec}^2 p^2 && (32)\\ =& \frac{1}{4}(2Var[w_{rec}] + \overline{w}_{rec}^2) && (33)\\ Var[w_{in}z_t] =& (Var[z_t] + \overline{z}_t^2)Var[w_{in}] && (34) \end{aligned}$$

Substituting in equation 30 implies

$$\frac{1}{4}(2Var[w_{rec}] + \overline{w}_{rec}^2) = (Var[z_t] + \overline{z}_t^2)\frac{n_{in}}{n_{rec} - 1}Var[w_{in}] \tag{35}$$

$$Var[w_{rec}] = 2(Var[z_t] + \overline{z}_t^2)\frac{n_{in}}{n_{rec} - 1}Var[w_{in}] - \frac{1}{2}\overline{w}_{rec}^2 \tag{36}$$

□

Therefore condition (II) led us to the constraint that W_{rec} has to meet at initialization, equation 36, for the condition to be true. The final equation further assumes that $\overline{w}_{in} = 0$ and $p = 1/2$.

E.3 SG dampening controls gradient maximum (III)

We apply condition (III) to the LIF network, which gives us a constraint that the dampening has to meet at initialization for the condition to be true.

Lemma 3. *Applying condition (III), which states that we want $Max[\frac{\partial}{\partial\theta}y_t] = Max[\frac{\partial}{\partial\theta}y_{t-1}]$, to an LIF network, and assuming that (1) σ' and $\frac{\partial}{\partial\theta}y_{t-1}$ are statistically independent and (2) we do not pass the gradient through the reset, it results in the constraint*

$$\gamma = \frac{1}{(n_{rec}-1)\hat{w}_{rec}}\Big(1 - \hat{\alpha}_{decay} - \xi \cdot n_{in}\hat{w}_{in}\gamma_{in}\Big) \tag{37}$$

where ξ is zero for the first layer and it's one for the other layers in the stack.

Proof. We want the maximal value of the gradient to remain stable, without exploding, when transmitted through time and through different layers

$$Max[\frac{\partial}{\partial\theta}y_t] = Max[\frac{\partial}{\partial\theta}y_{t-1}] \tag{38}$$

where when we write $\partial/\partial\theta$, we use θ as a placeholder for any quantity that we want to propagate through gradient descent. Taking the derivative of the LIF definition and stopping the gradient from going through the reset we have

$$\frac{\partial}{\partial\theta}y_t = \alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}(1-x_{t-1}) + W_{rec}\frac{\partial}{\partial\theta}x_{t-1} + \xi W_{in}\frac{\partial}{\partial\theta}z_t \tag{39}$$

Here we introduce the symbol $\xi \in \{0,1\}$, where $\xi = 1$ is used when z_t comes from a trainable layer below, and $\xi = 0$ when z_t represents the data. We consider as well that

$$\frac{\partial}{\partial\theta}z_t = \frac{\partial}{\partial\theta}\tilde{H}_{in}(y_t^{in} - \vartheta_{in}) = \sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in} \tag{40}$$

$$\frac{\partial}{\partial\theta}x_{t-1} = \frac{\partial}{\partial\theta}\tilde{H}(y_{t-1} - \vartheta) = \sigma'\frac{\partial}{\partial\theta}y_{t-1} \tag{41}$$

where $\tilde{H}_{in}, y_t^{in}, \vartheta_{in}$ are the Heaviside, the voltage and the threshold of the layer below, $\sigma' = \frac{\partial\tilde{H}}{\partial v}$ is the surrogate gradient, and σ'_{in} is the surrogate gradient from the layer below. Substituting in equation 39, then

$$\frac{\partial}{\partial\theta}y_t = \alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}(1-x_{t-1}) + W_{rec}\sigma'\frac{\partial}{\partial\theta}y_{t-1} \tag{42}$$

$$+ \xi W_{in}\sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in} \tag{43}$$

We use Max and Min in a statistical ensemble sense, as the maximum/minimum value that a variable could take if sampled over and over again

$$Max[X] = \sup_{x\sim p(x)} x \tag{44}$$

$$Min[X] = \inf_{x\sim p(x)} x \tag{45}$$

With this definition, if X, Y are independent random variables $Max[X+Y] = Max[X] + Max[Y]$ and if they are positive $Max[XY] = Max[X]Max[Y]$. We observe, as we did before for the variance and the mean of Wx, that

$$Max[Wx] = n_W Max[wx] \tag{46}$$
$$Min[Wx] = n_W Min[wx] \tag{47}$$

We take the maximal value of $\frac{\partial}{\partial\theta}y_t$, we make the assumption that σ' and $\frac{\partial}{\partial\theta}y_{t-1}$ are statistically independent, we use the fact that the highest value that the surrogate gradient can take is given by the dampening factor $Max[\sigma'] = \gamma$, we denote as γ_{in} the dampening factor of the layer below in the stack, and we take $Max[1 - x_{t-1}] = 1$:

$$\begin{aligned} Max[\frac{\partial}{\partial\theta}y_t] =& Max[\alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ (n_{rec}-1)Max[w_{rec}\sigma'\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ \xi n_{in} Max[w_{in}\sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in}] \end{aligned} \tag{48}$$

$$\begin{aligned} =& Max[\alpha_{decay}]Max[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ (n_{rec}-1)Max[w_{rec}]Max[\sigma']Max[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ \xi n_{in} Max[w_{in}]Max[\sigma'_{in}]Max[\frac{\partial}{\partial\theta}y_t^{in}] \end{aligned} \tag{49}$$

$$\begin{aligned} =& Max[\alpha_{decay}]Max[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ (n_{rec}-1)Max[w_{rec}]\gamma Max[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ \xi n_{in} Max[w_{in}]\gamma_{in} Max[\frac{\partial}{\partial\theta}y_t^{in}] \end{aligned} \tag{50}$$

where we used the fact that σ' is positive in the second equality. We apply condition (III), which states that all maximal gradients are equivalent, and for cleanliness we use the notation $Max[x] = \hat{x}$

$$1 = \hat{\alpha}_{decay} + (n_{rec}-1)\hat{w}_{rec}\gamma + \xi n_{in}\hat{w}_{in}\gamma_{in} \tag{51}$$
$$\gamma = \frac{1}{(n_{rec}-1)\hat{w}_{rec}}\Big(1 - \hat{\alpha}_{decay} - \xi \cdot n_{in}\hat{w}_{in}\gamma_{in}\Big) \tag{52}$$

where we only had to rearrange terms.

□

We set $\xi = 0$ in the main text for readability and because we observed better performance with it. This final equation 52 gives the value that the dampening has to take to keep the maximal gradient value stable, namely, condition (III) true at initialization.

E.4 SG sharpness controls gradient variance (IV)

We apply condition (IV) to the LIF network to constrain the choice of surrogate gradient variance.

Lemma 4. *Applying condition (IV), which states that we want $Var[\frac{\partial}{\partial\theta}y_t] = Var[\frac{\partial}{\partial\theta}y_{t-1}]$, to an LIF network, and assuming that (1) we do not pass the gradient through the reset, and (2) zero mean gradients at initialization, it results in the constraint*

$$\overline{\sigma'^2} = \frac{1 - \frac{1}{2}\overline{\alpha^2}_{decay} - \xi \cdot n_{in}\overline{w_{in}^2}\ \overline{\sigma'^2_{in}}}{(n_{rec} - 1)\overline{w^2}_{rec}} \tag{53}$$

where ξ is zero for the first layer and is one for the other layers in the stack.

Proof. Condition (IV) states that we want the variance of the gradient to remain stable across time and layers. Taking the derivative of the LIF we arrive at equation 43:

$$\begin{aligned}\frac{\partial}{\partial\theta}y_t =& \alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}(1 - x_{t-1}) \\ &+ W_{rec}\sigma'\frac{\partial}{\partial\theta}y_{t-1} + \xi W_{in}\sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in}\end{aligned} \tag{54}$$

Taking the variance and assuming that the monomials in the polynomial are statistically independent, we can consider the variance of the sum to be the sum of the variances:

$$\begin{aligned}Var[\frac{\partial}{\partial\theta}y_t] =& Var[\alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}(1 - x_{t-1})] \\ &+ Var[W_{rec}\sigma'\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ Var[\xi W_{in}\sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in}]\end{aligned} \tag{55}$$

$$\begin{aligned}Var[\frac{\partial}{\partial\theta}y_t] =& Var[\alpha_{decay}\frac{\partial}{\partial\theta}y_{t-1}(1 - x_{t-1})] \\ &+ (n_{rec} - 1)Var[w_{rec}\sigma'\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ n_{in}Var[\xi w_{in}\sigma'_{in}\frac{\partial}{\partial\theta}y_t^{in}]\end{aligned} \tag{56}$$

where $\xi = 0$ if w_{in} connects to the data and $\xi = 1$ if it connects to the layer below in the stack. We denote by σ'_{in} the surrogate gradient of the layer below.

Assuming gradients g with mean zero, and weights and gradients w, g to be independent random variables at initialization:

$$Var[wg] = (Var[g] + E[g]^2)(Var[w] + E[w]^2) - E[g]^2E[w]^2 \tag{57}$$

$$= Var[g](Var[w] + E[w]^2) \tag{58}$$

$$= Var[g]E[w^2] \tag{59}$$

which gives

$$\begin{aligned}Var[\frac{\partial}{\partial\theta}y_t] =& E[\alpha_{decay}^2(1 - x_{t-1})^2]Var[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ (n_{rec} - 1)E[(w_{rec}\sigma')^2]Var[\frac{\partial}{\partial\theta}y_{t-1}] \\ &+ \xi \cdot n_{in}E[(w_{in}\sigma'_{in})^2]Var[\frac{\partial}{\partial\theta}y_t^{in}]\end{aligned} \tag{60}$$

We apply condition IV, we want gradients to have the same variance, irrespective of the time step, or the neuron in the stack, which results in

$$
\begin{aligned}
1 =& \frac{1}{2}E[\alpha_{decay}^2] + (n_{rec}-1)E[(w_{rec}\sigma')^2] \\
&+ \xi \cdot n_{in}E[(w_{in}\sigma'_{in})^2] \qquad (61)\\
1 =& \frac{1}{2}E[\alpha_{decay}^2] + (n_{rec}-1)E[w_{rec}^2]E[\sigma'^2] \\
&+ \xi \cdot n_{in}E[w_{in}^2]E[\sigma'^2_{in}] \qquad (62)
\end{aligned}
$$

where we used the fact that for independent variables X, Y we have $E[X^pY^q] = E[X^p]E[Y^q]$ in the third and fourth line. Using the notation $E[x] = \overline{x}$, the implied condition on the SG is

$$
\overline{\sigma'^2} = \frac{1 - \frac{1}{2}\overline{\alpha^2}_{decay} - \xi \cdot n_{in}\overline{w_{in}^2}\ \overline{\sigma'^2_{in}}}{(n_{rec}-1)\overline{w^2}_{rec}} \qquad (63)
$$

□

We therefore found the constraint that the second non-centered moment of the SG has to satisfy, equation 63, if we want condition IV to hold. We set $\xi = 0$ in the main text for readability and because we observed better performance with it. We show how to relate it to the sharpness of the exponential SG in Appendix E.5.

E.5 Applying Condition IV to the exponential SG

We show how we apply equation 63, to choose the sharpness of an exponential SG. For that we need to define the dependence of the variance of the SG with its sharpness. We use as equivalent notation for the surrogate gradient

$$
\sigma'(v) = \frac{\partial \tilde{H}(v)}{\partial v} = \gamma f(\beta \cdot v)
$$

We denote no dependency with the voltage in σ', when we consider it as a random variable, and we introduce the dependency $\sigma'(v)$ when we assume the voltage dependence is known. The moments of the surrogate gradient are given by

$$
\begin{aligned}
E[\sigma'^m] &= \int \sigma'^m p(\sigma')d\sigma' \qquad (64)\\
&= \iint \sigma'^m p(\sigma'|v)p(v)dvd\sigma' \qquad (65)\\
&= \iint \sigma'^m \delta\Big(\sigma' - \sigma'(v)\Big)d\sigma' p(v)dv \qquad (66)\\
&= \int \sigma'(v)^m p(v)dv \qquad (67)
\end{aligned}
$$

where we used the marginalization rule in the second equality and in the third equality we used the fact that σ is a deterministic function of v, so it inherits its randomness from v. We are going to assume as the non-informative prior a uniform distribution between the minimal and maximal values of $y_t - \vartheta$.

$$E[\sigma'(v)^m] = \int \sigma'(v)^m p(v) dv \tag{68}$$

$$= \frac{1}{y_{max} - y_{min}} \int_{y_{min}-\vartheta}^{y_{max}-\vartheta} \sigma'(v)^m dv \tag{69}$$

$$= \frac{\gamma^m}{\beta(y_{max} - y_{min})} \int_{\beta(y_{min}-\vartheta)}^{\beta(y_{max}-\vartheta)} f(v')^m dv' \tag{70}$$

where we used the non informative uniform prior assumption in the second equality and we used $\sigma' = \gamma f(\beta v)$ followed by the change of variable $v' = \beta v$ in the third equality. Considering the exponential SG we have that, calling v_i one of the integration limits above, if v_i is positive

$$\int_0^{v_i} g(|v|)^m dv = \int_0^{v_i} g(v)^m dv \tag{71}$$

and if v_i is negative

$$\int_0^{v_i} g(|v|)^m dv = \int_0^{v_i} g(-v)^m dv \tag{72}$$

$$= -\int_0^{-v_i} g(v)^m dv \tag{73}$$

$$= -\int_0^{|v_i|} g(v)^m dv \tag{74}$$

where we made the change of variable $v \to -v$ in the second equality. Therefore

$$\int_{v_-}^{v_+} g(|v|)^m dv = sign(v_+) \int_0^{|v_+|} g(v)^m dv \tag{75}$$

$$- sign(v_-) \int_0^{|v_-|} g(v)^m dv \tag{76}$$

Given that for $v_i > 0$ we have

$$\int_0^{v_i} \text{exponential}(v)^m dv = \int_0^{v_i} e^{-2m|v|} dv \tag{77}$$

$$= \int_0^{v_i} e^{-2mv} dv \tag{78}$$

$$= -\frac{1}{2m} e^{-2mv_i} + \frac{1}{2m} \tag{79}$$

$$= -\frac{1}{2m} e^{-2m|v_i|} + \frac{1}{2m} \tag{80}$$

then, for $v_+ > 0$ and $v_- < 0$

$$\int_{v_-}^{v_+} \text{exponential}(v)^m dv =$$

$$-\frac{1}{2m} e^{-2m|v_+|} - \frac{1}{2m} e^{-2m|v_-|} + \frac{2}{2m} \tag{81}$$

where $v_+ = \beta(y_{max} - \vartheta)$ and $v_- = \beta(y_{min} - \vartheta)$ and we show how to compute $y_{max} = Max[y_t]$ and $y_{min} = Min[y_t]$ in section E.7. Since the dependence with β is quite complex, we find the β that satisfies the last equation and equation 63 through random search. Notice how equation 68, shows a dependence of the SG variance proportional to the square of the dampening and inversely proportional to the sharpness, which recalls our numerical results, where a high sharpness and a low dampening were preferred. This is how condition (IV) is used to fix the sharpness of the exponential SG.

E.6 Applying Condition IV to the q-PseudoSpike SG

Instead, when using (IV) to determine the tail-fatness of the SG, we set $\beta = 1$ and use

$$\int_{v_-}^{v_+} \text{q-PseudoSpike}(|v|)^2 dv = \int_{v_-}^{v_+} \text{q-PseudoSpike}(|v|)^2 dv \tag{82}$$

$$= \int_0^{|v_+|} \text{q-PseudoSpike}(v)^2 dv + \int_0^{|v_-|} \text{q-PseudoSpike}(v)^2 dv \tag{83}$$

$$= -\frac{q + 2|v_+| - 1}{2(2q-1)} \frac{1}{\left(1 + \frac{2}{q-1}|v_+|\right)^{2q}} - \frac{q + 2|v_-| - 1}{2(2q-1)} \frac{1}{\left(1 + \frac{2}{q-1}|v_-|\right)^{2q}} + \frac{q-1}{(2q-1)} \tag{84}$$

When inserted in equation 63, we use gradient descent to optimize q and find the value that satisfies (IV).

E.7 Maximal and Minimal voltage values achievable by the network at initialization

We calculate the maximum and minimum value that the voltage y can take, to be able to complete the argument about the variance of the backward pass of section E.5. First, we use Max and Min in a statistical ensemble sense, as the maximum/minimum value that a variable could take if sampled over and over again

$$Max[X] = \sup_{x \sim p(x)} x \tag{85}$$

$$Min[X] = \inf_{x \sim p(x)} x \tag{86}$$

When applied to the definition of LIF

$$Max[y_t] = Max[\alpha_{decay} y_{t-1}(1 - x_{t-1})] + Max[W_{rec} x_{t-1}] + Max[b] + Max[W_{in} z_t] \tag{87}$$

$$= Max[\alpha_{decay}] Max[y_{t-1}] + (n_{rec} - 1) Max[w_{rec}] + Max[b] + n_{in} Max[w_{in}] \tag{88}$$

$$Max[y_t] = \frac{1}{1 - Max[\alpha_{decay}]} \Big((n_{rec} - 1) Max[w_{rec}] + Max[b] + n_{in} Max[w_{in}] \Big) \tag{89}$$

where we used the fact that if x_t, z_t were sampled over and over, the maximum value that they could take is all neurons having fired at the same time, we used the fact that

$\alpha_{decay}, \vartheta > 0$, and we assumed that the maximum is going to stay constant through time $Max[y_{t-1}] = Max[y_t]$. Notice that the maximal voltage is achieved when all neurons in the layer fired at $t-1$, equation 88, except for the neuron under study, that stayed silent at $t-1$, to have 89. Similarly for the bound to the minimal voltage:

$$\begin{aligned}
Min[y_t] =& Min[\alpha_{decay} y_{t-1}(1 - x_{t-1})] \\
&+ (n_{rec} - 1) Min[w_{rec} x_{t-1}] + Min[b] && (90)\\
&+ n_{in} Min[w_{in} z_t] && (91)\\
=& Max[\alpha_{decay}] Min[y_{t-1}] \\
&+ (n_{rec} - 1) Min[w_{rec}] + Min[b] && (92)\\
&+ n_{in} Min[w_{in}] && (93)\\
Min[y_t] =& \frac{1}{1 - Max[\alpha_{decay}]} \Big((n_{rec} - 1) Min[w_{rec}] \\
&+ Min[b] + n_{in} Min[w_{in}] \Big) && (94)
\end{aligned}$$

Second, we consider another definition of Max and Min, where we consider the maximum value achievable by the current sample from the weight distribution. The real maximum value of the voltage will be achieved when the presynaptic neurons to fire are those that are connected with positive weight, we then have that our equation turns to

$$\begin{aligned}
Max[y_t]_i =& \alpha_{decay,i} Max[y_{t-1}]_i + \sum_j ReLU[W_{rec}]_{ij} \\
&+ b_i + \sum_j ReLU[W_{in}]_{ij} && (95)\\
Max[y_t]_i =& \frac{1}{1 - \alpha_{decay,i}} \Big(\sum_j ReLU[W_{rec}]_{ij} \\
&+ b_i + \sum_j ReLU[W_{in}]_{ij} \Big) && (96)
\end{aligned}$$

where we refer as $\sum_j ReLU[W_{rec}]_{ij}$ the sum over columns, where we have typically ommitted the index i for the element of the vector for cleanliness in the rest of the article. The case for the minimum is analogous

$$\begin{aligned}
Min[y_t]_i =& \alpha_{decay,i} Min[y_{t-1}]_i + Min[W_{rec} x_{t-1}] \\
&+ b_i + Min[W_{in} z_t] && (97)\\
Min[y_t]_i =& \frac{1}{1 - \alpha_{decay,i}} \Big(- \sum_j ReLU[-W_{rec}]_{ij} \\
&+ b_i - \sum_j ReLU[-W_{in}]_{ij} \Big) && (98)
\end{aligned}$$

With this we showed how we calculated the maximal and minimal value of the voltage, to be able to use condition (IV) to define the sharpness of the SG in section E.5.

F Applying conditions I-IV to an alternative definition of reset

We want to show how the constraints on the weights initialization and on the SG choice change, when the neuron model definition changes. We will use the notation $i_t = W_{rec} x_t + W_{in} z_t + b$. The reset used by

$$y_t = (\alpha_{decay} y_{t-1} + i_t)(1 - x_{t-1}) \tag{99}$$

that we will call *post-reset*. Instead,

$$y_t = \alpha_{decay} y_{t-1}(1 - x_{t-1}) + i_t \tag{100}$$

that we will call *pre-reset*, since it resets before applying the new current. Another example is given by

$$y_t = \alpha_{decay} y_{t-1} + i_t - \vartheta x_{t-1} \tag{101}$$

and we will call it *minus-reset*.

The first definition performs as well one refractory period, while the second does not result in a y_t clamped to zero when $x_t = 1$. The factor $(1 - x_t)$ takes the voltage exactly to zero every time the neuron has fired, zero being the equilibrium voltage. What is interesting about this form of reset is that the voltage is reset exactly to $y = 0$ after firing, while with the subtractive reset it is not the case. We consider training without passing the gradient through the reset, since

Post-reset:

$$y_t = (\alpha_{decay} y_{t-1} + i_t)(1 - x_{t-1})$$

$$\overline{w}_{rec} = \frac{2}{n_{rec} - 1}\Big(1 - \alpha_{decay}\Big)\vartheta \qquad \text{I}$$

$$Var[w_{rec}] = 2(Var[z_t] + \overline{z_t}^2)\frac{n_{in}}{n_{rec} - 1}Var[w_{in}] - \frac{1}{2}\overline{w}^2_{rec} \qquad \text{II}$$

$$\gamma = \frac{1}{(n_{rec} - 1)\hat{w}_{rec}}\Big(1 - \alpha_{decay} - \xi n_{in}\hat{w}_{in}\gamma_{in}\Big) \qquad \text{III}$$

$$\overline{\sigma'^2} = \frac{2 - \alpha^2_{decay} - \xi n_{in}\overline{w^2_{in}}\ \overline{\sigma'^2_{in}}}{(n_{rec} - 1)\overline{w^2}_{rec}} \qquad \text{IV}$$

Pre-reset:

$$y_t = \alpha_{decay} y_{t-1}(1 - x_{t-1}) + i_t$$

$$\overline{w}_{rec} = \frac{1}{n_{rec} - 1}\Big(2 - \alpha_{decay}\Big)\vartheta \qquad \text{I}$$

$$Var[w_{rec}] = 2(Var[z_t] + \overline{z_t}^2)\frac{n_{in}}{n_{rec} - 1}Var[w_{in}] - \frac{1}{2}\overline{w}^2_{rec} \qquad \text{II}$$

$$\gamma = \frac{1}{(n_{rec} - 1)\hat{w}_{rec}}\Big(1 - \alpha_{decay} - \xi n_{in}\hat{w}_{in}\gamma_{in}\Big) \qquad \text{III}$$

$$\overline{\sigma'^2} = \frac{1 - \frac{1}{2}\alpha^2_{decay} - \xi n_{in}\overline{w^2_{in}}\ \overline{\sigma'^2_{in}}}{(n_{rec} - 1)\overline{w^2}_{rec}} \qquad \text{IV}$$

Minus-reset:

$$y_t = \alpha_{decay} y_{t-1} + i_t - \vartheta x_{t-1}$$

$$\overline{w}_{rec} = \frac{1}{n_{rec}-1}\Big(3 - 2\alpha_{decay}\Big)\vartheta \qquad \text{I}$$

$$Var[w_{rec}] = 2(Var[z_t] + \overline{z}_t^2)\frac{n_{in}}{n_{rec}-1}Var[w_{in}] - \frac{1}{2}\overline{w}_{rec}^2 \qquad \text{II}$$

$$\gamma = \frac{1}{(n_{rec}-1)\check{w}_{rec} - \vartheta}\frac{\check{w}_{rec}}{\hat{w}_{rec}}\Big(1 - \alpha_{decay} - \xi n_{in}\hat{w}_{in}\gamma_{in}\Big) \qquad \text{III}$$

$$\overline{\sigma'^2} = \frac{1 - \alpha_{decay}^2 - \xi n_{in}\overline{w_{in}^2}\ \overline{\sigma_{in}'^2}}{(n_{rec}-1)\overline{w^2}_{rec} + \vartheta^2} \qquad \text{IV}$$

To have the conditions when the gradient does not pass through the reset, put $\vartheta = 0$ in (III) and (IV), but not in (I).

G ALIF and sLSTM models

To study the variability of SG trining with architecture choice, we tested different SG shapes on the ALIF and sLSTM networks. We used the following ALIF implementation

$$\begin{aligned}\boldsymbol{y}_{t,l} =& \boldsymbol{\alpha}^y_{decay,l}\boldsymbol{y}_{t-1,l} \\ &+ W_{rec,l}\boldsymbol{x}_{t-1,l} + W_{in,l}\boldsymbol{x}_{t-1,l-1} + \boldsymbol{b}_l \\ &- \boldsymbol{\vartheta}_{t-1,l}\ \boldsymbol{x}_{t-1,l} \end{aligned} \tag{102}$$

$$\boldsymbol{\vartheta}_{t,l} = \boldsymbol{\alpha}^{\vartheta}_{decay,l}\boldsymbol{\vartheta}_{t-1,l} + \boldsymbol{b}^{\vartheta}_l + \boldsymbol{\beta}_l\boldsymbol{x}_{t-1,l} \tag{103}$$

where we initialized W_{rec}, W_{in} as Glorot Uniform, $b_l = 0$, $\alpha^y_{decay,l} = 4\cdot 10^{-5}$, $\alpha^{\vartheta}_{decay,l} = 0.992$ for the SHD task and $\alpha^{\vartheta}_{decay,l} = 0.98$ for the sl-MNIST task, $b^{\vartheta}_l = 0.01$, and $\beta_l = 1.8$.

The LSTM implementation that we used is the following

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{104}$$
$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{105}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{106}$$
$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{107}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{108}$$
$$h_t = o_t \circ \sigma_h(c_t) \tag{109}$$

The dynamical variables i_t, f_t, o_t represent the input, forget and output gates, that prevent representations and gradients from exploding, while c_t, h_t represent the two hidden layers of the LSTM, that work as the working memory and are maintained and updated through data time t. To construct the spiking version of the LSTM (sLSTM) we turned the activations into $\sigma_g(x) = H(x)$ and $\sigma_c = \sigma_h = 2H(x) - 1$. The matrices W_j, U_j are initialized with Glorot Uniform initialization, and the biases b_j as zeros, with $j \in \{i, f, o, c\}$.

H More on Sparsity

We investigate if the role of sparsity remains consistent across SG shapes in Fig. 7, and across tasks in Fig. 8. Notice that Fig. 3 is repeated in Fig. 7 and 8 to ease the comparison.

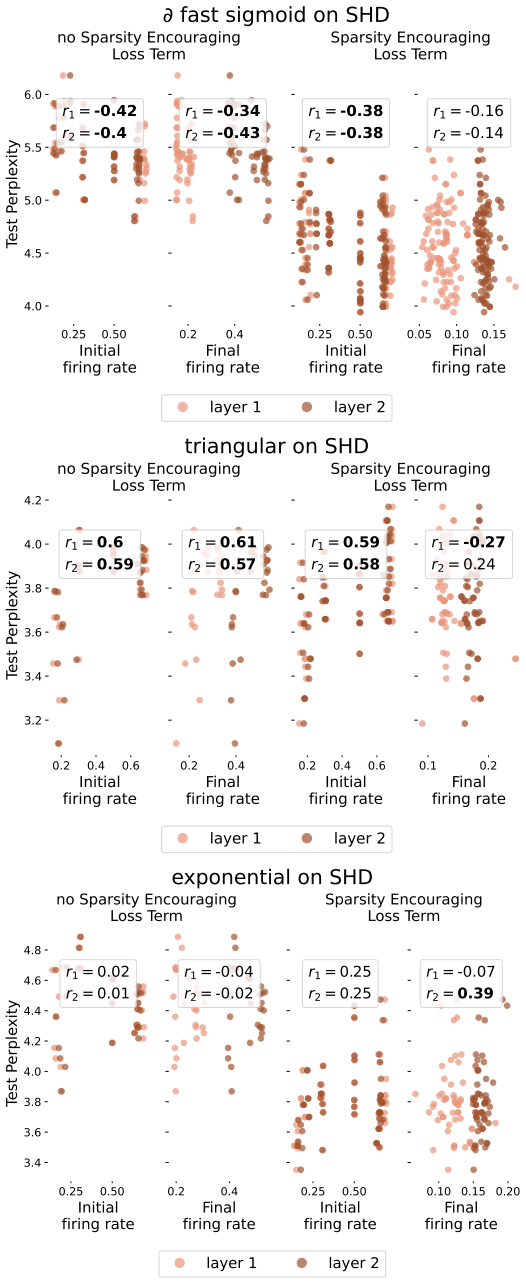

Figure 7: **Sparsity role is not consistent across SG shapes.** When we fix the task to be the SHD task, we see that the derivative of the fast sigmoid has preference for high p_i, the triangular SG has preference for low p_i, while for the exponential, p_i does not seem to correlate with final performance.

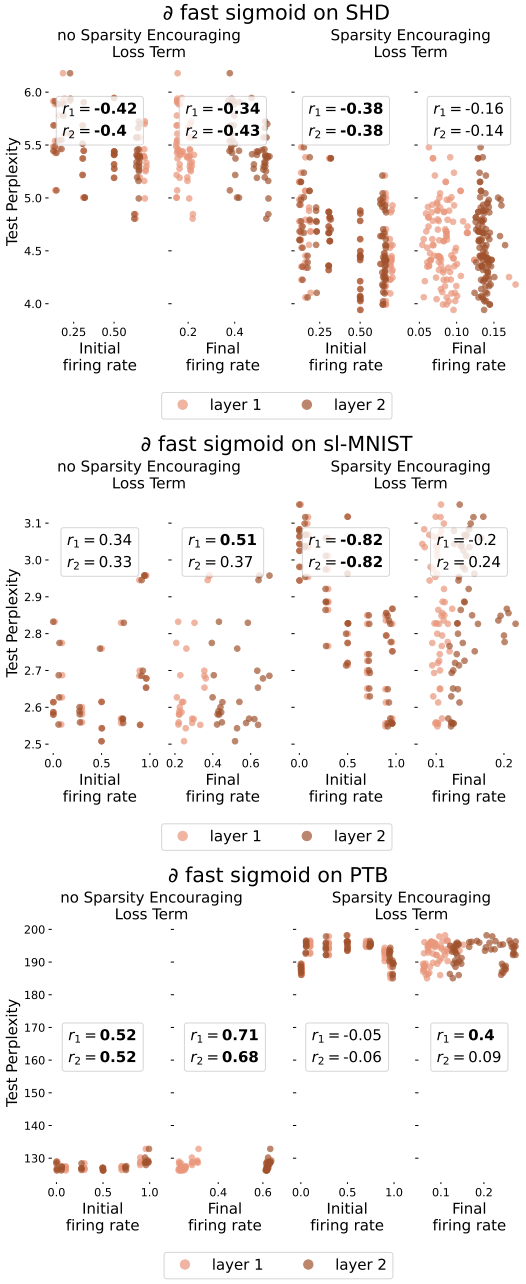

Figure 8: **Sparsity role is consistent across tasks.** Here we fix the SG shape to the derivative of the fast sigmoid and we change the task. On sl-MNIST, we see a similar trend than on SHD, where high initial firing rate is preferred for better performance when sparsity is encouraged. Encouraging sparsity has a negative effect on learning language modeling on the PTB task. However, when no sparsity is encouraged, best performance on PTB is still at $p_i = 0.5$.