Appendix

Organization of the Appendix In Section A, we introduce additional notations that will be used throughout the Appendix, then proceed to prove useful technical lemmas. We proceed in Section B to prove the results presented in the main text. Section C contains details about our experimental settings as well as some additional simulations.

A Additional notations and technical lemmas

For a vector a, we denote $\|a\|$ its ℓ^2-norm, $\|a\|_1$ its ℓ^1-norm and $\|a\|_\infty$ its ℓ^∞-norm. For matrices H, $\|H\|$ denotes the operator norm associated to the ℓ^2 norm and $\|H\|_{\mathrm{F}}$ denotes the Frobenius norm. For real-valued functions f, $\|f\|_\infty$ denotes the supremum norm.

In all of the Appendix, we denote $u_0 = -\eta/2$ and $u_{m+1} = 1 + \eta/2$. Note that $\sigma_\eta(x - u_0) = 1$ for all $x \in [0,1]$, meaning that $\sigma_\eta(\cdot - u_0)$ corresponds to the bias term. This notation allows to treat the bias term in a unified fashion with respect to the other terms of $f(x; a, u)$. Since $u_i \in (0,1)$ for $i \in \{1, \dots, m\}$, we assume in the following w.l.o.g. that the $(u_i)_{0 \leqslant i \leqslant m+1}$ are ordered in increasing order. Note that we prove in the following that the $(u_i)_{1 \leqslant i \leqslant m}$ do not cross during the dynamics, so they remain ordered throughout the dynamics.

The proofs involve comparisons of some quantities when $\eta > 0$ and when $\eta = 0$. To avoid confusion, we make explicit the dependency of L on $\eta \geqslant 0$, i.e., we let $L_\eta(a, u)$ in place of $L(a, u)$ of the main paper, and similarly, when the $\arg\min$ is well-defined and unique,

$$a_\eta^*(u) = \operatorname*{arg\,min}_{a \in \mathbb{R}^{m+1}} L_\eta(a, u)\,.$$

in place of $a^*(u)$. Similarly, we now make explicit the dependence of f on $\eta \geqslant 0$, i.e., we denote

$$f_\eta(x; a, u) = a_0 + \sum_{j=1}^m a_j \sigma_\eta(x - u_j) = \sum_{j=0}^m a_j \sigma_\eta(x - u_j)\,.$$

The Hessian of the quadratic function $L_\eta(\cdot, u)$ is denoted $H_\eta(u) \in \mathbb{R}^{(m+1)\times(m+1)}$ (in place of $H(u)$), and satisties that, for $i, j \in \{0, \dots, m\}$,

$$H_{\eta,ij}(u) = \int_0^1 \sigma_\eta(x - u_i)\sigma_\eta(x - u_j)\mathrm{d}x\,.$$

Also let, for $\eta \geqslant 0$ and $u \in \mathbb{R}^m$, $b_\eta(u) \in \mathbb{R}^{m+1}$ such that, for $j \in \{0, \dots, m\}$,

$$b_{\eta,j}(u) = \int_0^1 f^*(x)\sigma_\eta(x - u_j)\mathrm{d}x\,.$$

Finally, we let $\mathcal{U}_\eta$ in place of $\mathcal{U}$ in the paper.

With these notations, we have, for $\eta \geqslant 0$ and $a, u \in \mathbb{R}^m$,

$$\begin{aligned}\frac{\partial L_\eta}{\partial u_j}(a, u) &= \int_0^1 \frac{\partial f_\eta(x; a, u)}{\partial u_j}\left(f_\eta(x; a, u) - f^*(x)\right)\mathrm{d}x \\ &= -a_j \int_0^1 \sigma_\eta'(x - u_j)\Big(\sum_{k=0}^m a_k \sigma_\eta(x - u_k) - f^*(x)\Big)\mathrm{d}x\,. \end{aligned} \tag{10}$$

and

$$\begin{aligned}\frac{\partial L_\eta}{\partial a_j}(a, u) &= \int_0^1 \frac{\partial f_\eta(x; a, u)}{\partial a_j}\left(f_\eta(x; a, u) - f^*(x)\right)\mathrm{d}x \\ &= \int_0^1 \sigma_\eta(x - u_j)\Big(\sum_{k=0}^m a_k \sigma_\eta(x - u_k) - f^*(x)\Big)\mathrm{d}x \\ &= H_{\eta,j}(u)^\top a - b_{\eta,j}(u)\,. \end{aligned} \tag{11}$$

We now move on to a series to lemmas that will be helpful in the proofs of Appendix B.

Lemma 1. *For $\eta \geqslant 0$ and $u \in \mathbb{R}^m$, we have*

$$\|b_\eta(u) - b_0(u)\| \leqslant M\eta\sqrt{m+1} \quad \text{and} \quad \|b_\eta(u)\| \leqslant M\sqrt{m+1}\,.$$

Proof. For any $j \in \{0, \dots, m\}$,

$$\begin{aligned}
|b_{\eta,j}(u) - b_{0,j}(u)| &= \Big| \int_0^1 f^*(x)(\sigma_\eta(x-u_j) - \sigma_0(x-u_j))\mathrm{d}x \Big| \\
&\leqslant \|f^*\|_\infty \int_0^1 |\sigma_\eta(x-u_j) - \sigma_0(x-u_j)|\,\mathrm{d}x \\
&\leqslant M\eta\,,
\end{aligned}$$

where in the last step we use that $\|f^*\|_\infty \leqslant M$ and that $\sigma_\eta(x) = 0$ for $x \leqslant -\eta/2$, $\sigma_\eta(x) \in [0,1]$ for $-\eta/2 < x < \eta/2$ and $\sigma_\eta(x) = 1$ for $x \geqslant \eta/2$.

Similarly,

$$|b_{\eta,j}(u)| = \Big| \int_0^1 f^*(x)\sigma_\eta(x-u_j)\mathrm{d}x \Big| \leqslant \|f^*\|_\infty \leqslant M\,.$$

□

Lemma 2. *For $\eta \geqslant 0$ and $u \in \mathcal{U}_\eta$, $H_\eta(u) = H_0(u) + D_\eta$, where D_η is a diagonal matrix whose elements are independent of u and bounded in absolute value by $\eta/2$.*

Proof. Let $i, j \in \{0, \dots, m\}$, and denote $c = \max(u_i, u_j, 0)$. Then

$$H_{0,ij}(u) = \int_0^1 \sigma_0(x-u_i)\sigma_0(x-u_j)dx = 1 - c\,.$$

If $i = j = 0$, $\max(u_i, u_j) = -\eta/2$, and $H_{\eta,ij}(u) = 1 = H_{0,ij}(u)$. If $i = j \neq 0$,

$$\begin{aligned}
H_{\eta,ij}(u) &= \int_0^1 \sigma_\eta(x-c)^2 dx \\
&= 1 - c - \frac{\eta}{2} + \int_{c-\eta/2}^{c+\eta/2} \sigma_\eta(x-c)^2 dx \\
&= H_{0,ij}(u) - \frac{\eta}{2} + \eta \int_{-1/2}^{1/2} \sigma^2\,.
\end{aligned}$$

Note that the last integral is non-negative and less than 1, hence $|H_{\eta,ij}(u) - H_{0,ij}(u)| \leqslant \eta/2$. Finally, if $i \neq j$, since $|u_i - u_j| > \eta$,

$$H_{\eta,ij}(u) = \int_0^1 \sigma_\eta(x-u_i)\sigma_\eta(x-u_j)dx = \int_0^1 \sigma_\eta(x - \max(u_i,u_j))dx\,.$$

Furthermore, $0 < \max(u_i, u_j) < 1 - \frac{\eta}{2}$, thus

$$H_{\eta,ij}(u) = \int_0^1 \sigma_\eta(x-c)dx = 1 - c - \frac{\eta}{2} + \int_{c-\eta/2}^{c+\eta/2} \sigma_\eta(x-c)dx = 1 - c\,,$$

where the last equality comes from the oddness of $\sigma - 1/2$. □

Lemma 3. *For $\eta > 0$, let $a_\eta^* : u \in \mathcal{U}_\eta \mapsto a_\eta^*(u)$. Then a_η^* is differentiable and for any $u \in \mathcal{U}_\eta$,*

$$\Big\| \frac{\partial a_\eta^*(u)}{\partial u} \Big\| \leqslant \frac{8}{\Delta(u)} \Big(2(m+1)\|a_\eta^*(u)\| + M \Big)\,.$$

Proof. By Proposition 1 (whose proof does not rely on this lemma), for $u \in \mathcal{U}_\eta$, $L_\eta(\cdot, u)$ has a unique minimizer $a_\eta^*(u)$, which is equal to $H_\eta(u)^{-1}b_\eta(u)$ by (11). Furthermore, H_η and b_η are differentiable with respect to u, hence a_η^* is also differentiable with respect to u, and we have

$$\frac{\partial a_\eta^*(u)}{\partial u_k} = -H_\eta(u)^{-1}\frac{\partial H_\eta}{\partial u_k}(u)a_\eta^*(u) + H_\eta(u)^{-1}\frac{\partial b_\eta}{\partial u_k}(u)\,.$$

Denote $w_k(u) := \frac{\partial H_\eta}{\partial u_k}(u) a^*_\eta(u)$ and $W(u)$ the $(m+1) \times (m+1)$ matrix formed by stacking column-wise the vectors $(w_k(u))_{0 \leqslant k \leqslant m}$. Then

$$\frac{\partial a^*_\eta(u)}{\partial u} = -H_\eta(u)^{-1} W(u) + H_\eta(u)^{-1} \frac{\partial b_\eta}{\partial u}(u) \,.$$

We now estimate the Frobenius norm of the matrix $W(u)$. By Lemma 2, for $u \in \mathcal{U}_\eta$, $H_\eta(u) = H_0(u) + D_\eta$. Take $i, j \in \{0, \dots, m\}$, then

$$H_{\eta,ij}(u) = H_{0,ij}(u) + D_{\eta,ij} = \int_0^1 \sigma_0(x - u_i)\sigma_0(x - u_j)dx + D_{\eta,ij} = 1 - \max(u_i, u_j, 0) + D_{\eta,ij} \,.$$

Hence $\frac{\partial H_{\eta,ij}}{\partial u_k} = 0$ if $i, j \neq k$. Further, if $i = k$ and $j \neq k$,

$$\Big| \frac{\partial H_{\eta,ij}}{\partial u_k}(u) \Big| = \Big| \frac{\partial}{\partial u_i}(1 - \max(u_i, u_j)) \Big| \leqslant 1 \,.$$

Of course, the bound $|\frac{\partial H_{\eta,ij}}{\partial u_k}(u)| \leqslant 1$ also holds when $j = k$ and $i \neq j$. Finally, a similar bound shows that $|\frac{\partial H_{\eta,ij}}{\partial u_k}(u)| \leqslant 2$ when $i = j = k$.

As a consequence, for $k, i \in \{0, \dots, m\}$,

$$|w_{k,i}(u)| \leqslant \sum_{j=0}^{m} \Big| \frac{\partial H_{\eta,ij}}{\partial u_k}(u) \Big| |a^*_{\eta,j}(u)| \leqslant \begin{cases} |a^*_{\eta,k}(u)| & \text{if } i \neq k \,, \\ |a^*_{\eta,k}(u)| + \|a^*_\eta(u)\|_1 & \text{if } i = k \,. \end{cases}$$

Thus

$$\begin{aligned} \|W(u)\|_{\mathrm{F}} &= \Big(\sum_{i=0}^{m} \sum_{k=0}^{m} |w_{k,i}(u)|^2 \Big)^{1/2} \leqslant \Big(\sum_{i=0}^{m} \Big(\sum_{k=0}^{m} |w_{k,i}(u)| \Big)^2 \Big)^{1/2} \\ &\leqslant \Big(\sum_{i=0}^{m} \big(2\|a^*_\eta(u)\|_1 \big)^2 \Big)^{1/2} = 2\sqrt{m+1} \|a^*_\eta(u)\|_1 \,. \end{aligned}$$

With a reasoning similar to the above, note that $\frac{\partial b_\eta}{\partial u}(u)$ is a diagonal matrix with diagonal entries in $[-M, M]$. Finally, putting these elements together, using Proposition 1 and that $\|W(u)\| \leqslant \|W(u)\|_{\mathrm{F}}$, we obtain

$$\Big\| \frac{\partial a^*_\eta(u)}{\partial u} \Big\| \leqslant \|H_\eta(u)^{-1}\| \|W(u)\|_{\mathrm{F}} + \|H_\eta(u)^{-1}\| \Big\| \frac{\partial b_\eta(u)}{\partial u} \Big\| \leqslant \frac{8}{\Delta(u)} \Big(2\sqrt{m+1} \|a^*_\eta(u)\|_1 + M \Big) \,.$$

□

The following lemma gives exact formulae for the derivative of the loss L_η with respect to the positions of the neurons, evaluated for $a = a^*_0(u)$, that is the best piecewise constant approximation of f^* with subdivision $\{u_1, \dots, u_m\}$. Note that the formulae are the same as in Section 4.2, but the derivation is slightly more intricate since we consider here the loss L_η and not L_0.

Lemma 4. *Take $\eta > 0$ and $u \in \mathcal{U}_\eta$ such that there are at least two neurons on each piece $[v_i, v_{i+1}]$ of f^*. Then, if u_j does not flank a discontinuity of f^*,*

$$\frac{\partial L_\eta}{\partial u_j}(a^*_0(u), u) = 0.$$

Furthermore, for a discontinuity v_i, denote u_i^{L} is the closest neuron to its left and u_i^{R} the closest neuron to its right. If $v_i - u_i^{\mathrm{L}} \geqslant \frac{\eta}{2}$ and $u_i^{\mathrm{R}} - v_i \geqslant \frac{\eta}{2}$, then

$$\begin{aligned} \frac{\partial L_\eta}{\partial u_i^{\mathrm{L}}}(a^*_0(u), u) &= -\frac{1}{2} \frac{(u_i^{\mathrm{R}} - v_i)^2}{(u_i^{\mathrm{R}} - u_i^{\mathrm{L}})^2} (f^*_i - f^*_{i-1})^2 \,, \\ \frac{\partial L_\eta}{\partial u_i^{\mathrm{R}}}(a^*_0(u), u) &= \frac{1}{2} \frac{(v_i - u_i^{\mathrm{L}})^2}{(u_i^{\mathrm{R}} - u_i^{\mathrm{L}})^2} (f^*_i - f^*_{i-1})^2 \,. \end{aligned}$$

Proof. In this proof, let us denote for simplicity $a = a_0^*(u)$. At the condition that there is at least two neurons on each piece of f^*, Section 4.2 gives the optimal approximation $f_0(x; a, u)$ of f^* that is piecewise constant with respect to the subdivision $\{u_1, \dots, u_m\}$. As a consequence, we easily get the value of a. Namely, if u_j does not flank a discontinuity of f^*, the value of $f_0(x; a, u)$ is locally constant around u_j, thus $a_j = 0$. Plugging into (10), we obtain

$$\frac{\partial L_\eta}{\partial u_j}(a, u) = 0\,.$$

Further, for a discontinuity v_i, denote respectively a_i^{L} and a_i^{R} the coefficients associated to u_i^{L} and u_i^{R}. At u_i^{L}, the value of $f_0(x; a, u)$ jumps from f_{i-1}^* to $\frac{v_i - u_i^{\mathrm{L}}}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} f_{i-1}^* + \frac{u_i^{\mathrm{R}} - v_i}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} f_i^*$, thus

$$a_i^{\mathrm{L}} = \frac{v_i - u_i^{\mathrm{L}}}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} f_{i-1}^* + \frac{u_i^{\mathrm{R}} - v_i}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} f_i^* - f_{i-1}^* = \frac{u_i^{\mathrm{R}} - v_i}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} (f_i^* - f_{i-1}^*)\,.$$

Similarly, we have

$$a_i^{\mathrm{R}} = \frac{v_i - u_i^{\mathrm{L}}}{u_i^{\mathrm{R}} - u_i^{\mathrm{L}}} (f_i^* - f_{i-1}^*)\,.$$

We now compute, using (10),

$$\begin{aligned}
\frac{\partial L_\eta}{\partial u_i^{\mathrm{L}}}(a, u) &= -a_i^{\mathrm{L}} \int_0^1 \sigma_\eta'(x - u_i^{\mathrm{L}}) \left(f_\eta(x; a, u) - f^*(x)\right) \mathrm{d}x \\
&= -a_i^{\mathrm{L}} \int_{u_i^{\mathrm{L}} - \eta/2}^{u_i^{\mathrm{L}} + \eta/2} \sigma_\eta'(x - u_i^{\mathrm{L}}) \left(f_\eta(x; a, u) - f^*(x)\right) \mathrm{d}x\,.
\end{aligned}$$

Using that $\Delta(u) > 2\eta$ and that there are at least two neurons on each piece of f^*, we have that $u_i^{\mathrm{L}} - v_{i-1} \geqslant 2\eta$. Since, in addition, by assumption, $v_i - u_i^{\mathrm{L}} \geqslant \frac{\eta}{2}$, we get that for $x \in \left[u_i^{\mathrm{L}} - \frac{\eta}{2}, u_i^{\mathrm{L}} + \frac{\eta}{2}\right]$, $f^*(x) = f_{i-1}^*$. Moreover, using again $\Delta(u) \geqslant 2\eta$ that σ_η is equal to σ_0 on $(-\infty, -\eta/2]$ and $[\eta/2, \infty)$, we have for $x \in \left[u_i^{\mathrm{L}} - \frac{\eta}{2}, u_i^{\mathrm{L}} + \frac{\eta}{2}\right]$,

$$f_\eta(x; a, u) = \sum_{k=0}^{m} a_k \sigma_\eta(x - u_k) = f_0\left(u_i^{\mathrm{L}} - \frac{\eta}{2}; a, u\right) + a_i^{\mathrm{L}} \sigma_\eta(x - u_i^{\mathrm{L}}) = f_{i-1}^* + a_i^{\mathrm{L}} \sigma_\eta(x - u_i^{\mathrm{L}})\,.$$

Thus we obtain

$$\begin{aligned}
\frac{\partial L_\eta}{\partial u_i^{\mathrm{L}}}(a, u) &= -a_i^{\mathrm{L}} \int_{u_i^{\mathrm{L}} - \eta/2}^{u_i^{\mathrm{L}} + \eta/2} \sigma_\eta'(x - u_i^{\mathrm{L}}) a_i^{\mathrm{L}} \sigma_\eta(x - u_i^{\mathrm{L}}) \mathrm{d}x \\
&= -\frac{(a_i^{\mathrm{L}})^2}{2} \Big(\sigma_\eta\big(\frac{\eta}{2}\big)^2 - \sigma_\eta\big(-\frac{\eta}{2}\big)^2\Big) \\
&= -\frac{(a_i^{\mathrm{L}})^2}{2} \\
&= -\frac{1}{2} \frac{(u_i^{\mathrm{R}} - v_i)^2}{(u_i^{\mathrm{R}} - u_i^{\mathrm{L}})^2} (f_i^* - f_{i-1}^*)^2\,.
\end{aligned}$$

The computation of $\frac{\partial L_\eta}{\partial u_i^{\mathrm{R}}}(a, u)$ is similar.

□

Lemma 5. *Consider $\eta \geqslant 0$ and $u \in \mathcal{U}_\eta$ such that there are at least two neurons on each piece $[v_i, v_{i+1}]$ of f^*. Then, for all $x \in [0, 1]$, $|f_\eta(x; a_0^*(u), u)| \leqslant M$.*

Proof. In the case where $\eta = 0$, the result easily follows from the expressions for $f_0(x; a_0^*(u), u)$ provided in Section 4.2. We now assume $\eta > 0$.

Denote $A^*_k(u) = \sum_{j=0}^k a^*_{0,j}(u)$ (with the convention $A^*_{-1}(u) = 0$). Recall the convention $u_0 = -\eta/2$. We compute

$$\begin{aligned} f_\eta(x; a^*_0(u), u) &= \sum_{k=0}^{m} a^*_{0,k}(u)\sigma_\eta(x - u_k) \\ &= \sum_{k=0}^{m} \left(A^*_k(u) - A^*_{k-1}(u)\right) \sigma_\eta(x - u_k) \\ &= \sum_{k=0}^{m-1} A^*_k(u) \left(\sigma_\eta(x - u_k) - \sigma_\eta(x - u_{k+1})\right) + A^*_m(u)\sigma_\eta(x - u_m) \end{aligned}$$

Note that $A^*_k(u) = \lim_{x \to u_k +} f_0(x; a^*_0(u), u)$, and thus, from the case $\eta = 0$, we have $|A^*_k(u)| \leqslant M$. Moreover, σ_η is increasing and the u_k are in increasing order. We thus get

$$\begin{aligned} |f_\eta(x; a^*_0(u), u)| &\leqslant M\Big(\sum_{k=0}^{m-1} \left(\sigma_\eta(x - u_k) - \sigma_\eta(x - u_{k+1})\right) + \sigma_\eta(x - u_m)\Big) \\ &= M\sigma_\eta(x - u_0) \leqslant M\,. \end{aligned}$$

□

Lemma 6. *Consider $\eta > 0$ and $u \in \mathcal{U}_\eta$ such that there are at least two neurons on each piece $[v_i, v_{i+1}]$ of f^*. Then, for $j \in \{0, \dots, m\}$,*

$$|a^*_{0,j}(u)| \leqslant 2M$$

and, for any $a \in \mathbb{R}^{m+1}$,

$$\Big|\frac{\partial L_\eta}{\partial u_j}(a, u) - \frac{\partial L_\eta}{\partial u_j}(a^*_0(u), u)\Big| \leqslant 2M(\sqrt{m+1} + 1)\|a - a^*_0(u)\| + \sqrt{m+1}\|a - a^*_0(u)\|^2\,.$$

Proof. The first statement of the Lemma comes from the explicit formulae for $a^*_0(u)$ given in the proof of Lemma 4, namely each $a^*_{0,j}(u)$ is either zero or less in magnitude than the gap between two pieces of f^* that is less than $2M$.

By (10), we have

$$\begin{aligned} &\Big|\frac{\partial L_\eta}{\partial u_j}(a, u) - \frac{\partial L_\eta}{\partial u_j}(a^*_0(u), u)\Big| \\ &\qquad = \Big| a_j \int_0^1 \sigma'_\eta(x - u_j)\Big(f_\eta(x; a, u) - f^*(x)\Big)\mathrm{d}x \\ &\qquad\qquad - a^*_{0,j}(u) \int_0^1 \sigma'_\eta(x - u_j)\Big(f_\eta(x; a^*_0(u), u) - f^*(x)\Big)\mathrm{d}x\Big| \\ &\qquad \leqslant |a_j - a^*_{0,j}(u)| \int_0^1 \sigma'_\eta(x - u_j)\,|f_\eta(x; a^*_0(u), u) - f^*(x)|\,\mathrm{d}x \\ &\qquad\qquad + |a_j| \int_0^1 \sigma'_\eta(x - u_j)\,|f_\eta(x; a, u) - f_\eta(x; a^*_0(u), u)|\,\mathrm{d}x\,. \end{aligned}$$

We bound the two terms separately. For the first term, we use Lemma 5.

$$\begin{aligned} \int_0^1 \sigma'_\eta(x - u_j)\,|f_\eta(x; a^*_0(u), u) - f^*(x)| &\leqslant \int_0^1 \sigma'_\eta(x - u_j)\,(|f_\eta(x; a^*_0(u), u)| + |f^*(x)|) \\ &\leqslant 2M \int_0^1 \sigma'_\eta(x - u_j)\mathrm{d}x \leqslant 2M\,. \end{aligned}$$

We now continue with the second term.

$$|f_\eta(x; a, u) - f_\eta(x; a^*_0(u), u)| = \Big|\sum_{k=0}^{m}(a_k - a^*_{0,k}(u))\sigma_\eta(x - u_k)\Big| \leqslant \|a - a^*_0(u)\|_1\,,$$

and thus

$$\int_0^1 \sigma'_\eta(x-u_j)\,|f_\eta(x;a,u)-f_\eta(x;a_0^*(u),u)|\,\mathrm{d}x \leqslant \|a-a_0^*(u)\|_1 \int_0^1 \sigma'_\eta(x-u_j)\mathrm{d}x \leqslant \|a-a_0^*(u)\|_1 .$$

Returning to our initial upper bound, we obtain, using the first statement of the Lemma,

$$\begin{aligned}\Big|\frac{\partial L_\eta}{\partial u_j}(a,u)-\frac{\partial L_\eta}{\partial u_j}(a_0^*(u),u)\Big| &\leqslant 2M\|a-a_0^*(u)\| + (|a_{0,j}^*(u)|+|a_j-a_{0,j}^*(u)|)\|a-a_0^*(u)\|_1 \\ &\leqslant 2M\|a-a_0^*(u)\| + (2M+\|a-a_0^*(u)\|)\sqrt{m+1}\|a-a_0^*(u)\| \\ &= 2M(\sqrt{m+1}+1)\|a-a_0^*(u)\| + \sqrt{m+1}\|a-a_0^*(u)\|^2 .\end{aligned}$$

□

Lemma 7. *For $\eta \geqslant 0$ and $u \in \mathcal{U}_\eta$,*

$$\|a_\eta^*(u)-a_0^*(u)\| \leqslant \frac{16M\sqrt{m+1}\eta}{\Delta(u)} .$$

Proof. By (11),

$$H_\eta(u)a_\eta^*(u) = b_\eta(u)$$

and by (11) and by Lemma 2,

$$H_\eta(u)a_0^*(u) = H_0(u)a_0^*(u) + D_\eta a_0^*(u) = b_0(u) + D_\eta a_0^*(u).$$

According to Proposition 1 (whose proof does not rely on this lemma), $H_\eta(u)$ is invertible with $\|H_\eta(u)^{-1}\| \leqslant 8/\Delta(u)$. We thus have

$$\begin{aligned}\|a_\eta^*(u)-a_0^*(u)\| &= \|H_\eta(u)^{-1}(H_\eta(u)a_\eta^*(u)-H_\eta(u)a_0^*(u))\| \\ &\leqslant \frac{8}{\Delta(u)}\|b_\eta(u)-b_0(u)-D_\eta a_0^*(u)\| \\ &\leqslant \frac{8}{\Delta(u)}\big(\|b_\eta(u)-b_0(u)\|+\|D_\eta a_0^*(u)\|\big) \\ &\leqslant \frac{8}{\Delta(u)}\big(\|b_\eta(u)-b_0(u)\|+\frac{\eta}{2}\|a_0^*(u)\|\big) .\end{aligned}$$

The result then unfolds from Lemmas 1 and 6. □

Lemma 8. *Let $\eta > 0$, $u \in \mathbb{R}^m$ and $a, a' \in \mathbb{R}^{m+1}$. Then*

$$\begin{aligned}\|\nabla_u L_\eta(a,u)\| &\leqslant \sqrt{m+1}\|a\|^2 + M\|a\| , \\ \|\nabla_a L_\eta(a,u)\| &\leqslant \sqrt{m+1}(\|a\|\sqrt{m+1}+M) .\end{aligned}$$

As a consequence of the second inequality, by the fundamental theorem of calculus for line integrals,

$$|L_\eta(a,u)-L_\eta(a',u)| \leqslant \sqrt{m+1}\left(\max(\|a\|,\|a'\|)\sqrt{m+1}+M\right)\|a-a'\| .$$

Proof. Recall that, for all $j \in \{1,\dots,m\}$, and for all $a, u \in \mathbb{R}^m$,

$$\begin{aligned}\frac{\partial L_\eta}{\partial u_j}(a,u) &= -a_j \int_0^1 \sigma'_\eta(x-u_j)\Big(\sum_{k=0}^m a_k\sigma_\eta(x-u_k)-f^*(x)\Big)\mathrm{d}x , \\ \frac{\partial L_\eta}{\partial a_j}(a,u) &= \int_0^1 \sigma_\eta(x-u_j)\Big(\sum_{k=0}^m a_k\sigma_\eta(x-u_k)-f^*(x)\Big)\mathrm{d}x .\end{aligned}$$

From the first equality, we have

$$\begin{aligned}\Big|\frac{\partial L_\eta}{\partial u_j}(a,u)\Big| &\leqslant |a_j|\int_0^1 |\sigma'_\eta(x-u_j)|\Big(\sum_{k=1}^m |a_k|\sigma_\eta(x-u_k)+|f^*(x)|\Big)dx \\ &\leqslant |a_j|(\|a\|_1+M)\int_0^1 |\sigma'_\eta(x-u_j)|dx \\ &\leqslant |a_j|(\|a\|_1+M) .\end{aligned}$$

As a consequence,

$$\|\nabla_u L_\eta(a,u)\| \leqslant \|a\|(\|a\|_1 + M) \leqslant \sqrt{m+1}\|a\|^2 + M\|a\|\,.$$

Similarly, from the second equality, we have

$$\left|\frac{\partial L_\eta}{\partial a_j}(a,u)\right| \leqslant \|a\|_1 + M\,.$$

As a consequence,

$$\|\nabla_a L_\eta(a,u)\| \leqslant \sqrt{m+1}(\|a\|_1 + M) = \sqrt{m+1}(\|a\|\sqrt{m+1} + M)\,.$$

□

Lemma 9. *Consider $\eta \geqslant 0$ and $u \in \mathcal{U}_\eta$ such that there is a neuron at distance less than η from each discontinuity of f^* and $3\eta \leqslant \Delta v$. Then*

$$\int_0^1 |f_\eta(x; a_\eta^*(u), u) - f^*(x)|^2 \mathrm{d}x \leqslant 6M^2\eta n\,.$$

Proof. By definition of $a_\eta^*(u)$,

$$\int_0^1 |f_\eta(x; a_\eta^*(u), u) - f^*(x)|^2 \mathrm{d}x = \min_{a\in\mathbb{R}^{m+1}} \int_0^1 |f_\eta(x; a, u) - f^*(x)|^2 \mathrm{d}x\,.$$

Thus it is enough to exhibit some a for which the latter integral is smaller than $6M^2\eta n$ to conclude.

We construct such an a as follows: set $a_0 = f^*(0)$, and for each discontinuity v_i, set the coefficient of a neuron at distance less than η to the value $f_i^* - f_{i-1}^*$ and set all other neurons to zero. Note that the active neurons are distinct since $3\eta \leqslant \Delta v$.

Then the neural network is equal to the target function everywhere except on an interval of size $3\eta/2$ around each discontinuity, where they disagree (in infinite norm) by at most $2M$.

□

Lemma 10. *Let m be a positive integer and $u_1, \dots, u_m$ be i.i.d. uniform random variables in $[0,1]$. Assume that*

$$m \geqslant \frac{6}{\Delta v}\Big(4 + \log n + \log\frac{1}{\delta}\Big)\,.$$

Then, with probability at least $1-\delta$, the vector u is D-good with $D = \frac{\delta}{6(m+1)^2}$.

Proof. We define the following events:

(a) A is the event "there are at least 6 positions u_j in each interval $[v_i, v_{i+1}]$ for $i \in \{0, \dots, n-1\}$",

(b) B is the event "$\Delta(u) \geqslant D$",

(c) for all $i \in \{1, \dots, n-1\}$, E_i is the event "there are at least one neuron on the left and on the right of v_i" and C_i is the event "E_i holds and $|u_i^{\mathrm{R}} + u_i^{\mathrm{L}} - 2v_i| \geqslant D$".

Note that by Definition 2, u is D-good if and only if the event $A \cap B \cap (\bigcap_i C_i)$ holds. To show that this holds with high probability, we bound the probability of the complement

$$\begin{aligned}\Big(A \cap B \cap \Big(\bigcap_i C_i\Big)\Big)^c &= A^c \cup B^c \cup \Big(\bigcup_i C_i^c\Big) = A^c \cup B^c \cup \Big(\bigcup_i (C_i^c \cap A)\Big) \\ &\subset A^c \cup B^c \cup \Big(\bigcup_i (C_i^c \cap E_i)\Big) \qquad \text{(as } A \subset E_i\text{)}\,.\end{aligned}$$

Thus

$$\mathbb{P}(u \text{ is not } D\text{-good}) \leqslant \mathbb{P}(A^c) + \mathbb{P}(B^c) + \sum_{i=1}^{n-1} \mathbb{P}(C_i^c \cap E_i)\,.$$

Below, we bound separately the three terms of the right hand side.

(a) Denote $m' = \lfloor m/6 \rfloor$. For any $i \in \{0, \dots, n-1\}$, the set $\mathcal{A}_i = \{j \in \{1, \dots, m'\} \,|\, u_j \in [v_i, v_{i+1}]\}$ is empty with probability $(1 - (v_{i+1} - v_i))^{m'} \leqslant (1 - \Delta v)^{m'}$. Thus by the union bound, the probability that at least one of $\mathcal{A}_1, \dots, \mathcal{A}_n$ is empty is upper bounded by $n(1 - \Delta v)^{m'}$.

We now check that $n(1 - \Delta v)^{m'} \leqslant \delta/18$. Indeed,

$$m' = \left\lfloor \frac{m}{6} \right\rfloor \geqslant \frac{m}{6} - 1 \geqslant \frac{3 + \log n + \log \frac{1}{\delta}}{\Delta v} \geqslant \frac{\log n + \log \frac{18}{\delta}}{\Delta v} \geqslant -\frac{\log n + \log \frac{18}{\delta}}{\log(1 - \Delta v)} \,,$$

where we use $\Delta v \leqslant 1$, $3 \geqslant \log(18)$, and $\log(1 - \Delta v) \leqslant -\Delta v < 0$. This gives the desired inequality.

In other words, the probability that at least one of the intervals $[v_i, v_{i+1}]$ contains none of the $u_1, \dots, u_{m'}$ is bounded by $\delta/18$. As a consequence, by the union bound, the probability that at least one of the intervals $[v_i, v_{i+1}]$ contains strictly less than 6 of the $u_1, \dots, u_m$ is bounded by $\delta/3$, i.e., $\mathbb{P}(A^c) \leqslant \delta/3$.

(b) Recall that by convention, $u_0 = -\frac{\eta}{2}$ and $u_{m+1} = 1 + \frac{\eta}{2}$. For all $i \in \{0, \dots, m+1\}$, denote $I_i = (u_i - D, u_i + D)$. Denote F_j the event "$u_j \in I_i$ for some $i \in \{0, \dots, m+1\}$, $i \neq j$". Note that $B^c = \cup_{j=1}^m F_j$.

Fix $j = 1, \dots, m$. By conditioning on u_i for all $i \in \{0, \dots, m+1\}$, $i \neq j$, we see that $\mathbb{P}(F_j) \leqslant 2(m+1)D$. By the union bound,

$$\mathbb{P}(B^c) \leqslant 2m(m+1)D \leqslant \frac{\delta}{3}.$$

(c) Take $i \in \{1, \dots, n-1\}$. For convenience, we define the random variable u_i^{L} (resp. u_i^{R}) on the full probability space by setting $u_i^{\mathrm{L}} = 0$ (resp. $u_i^{\mathrm{R}} = 1$) when there is no neuron on the left (resp. the right) of v_i. We compute the joint cumulative distribution function of $(u_i^{\mathrm{L}}, u_i^{\mathrm{R}})$ (with a convenient change of inequality): for all $0 \leqslant y \leqslant v_i \leqslant z \leqslant 1$,

$$\mathbb{P}(u_i^{\mathrm{L}} \leqslant y, u_i^{\mathrm{R}} \geqslant z) = \mathbb{P}(\forall j \in \{1, \dots, m\},\, u_j \notin [y, z]) = (1 - (z - y))^m \,.$$

We observe that the joint cumulative distribution function of $(u_i^{\mathrm{L}}, u_i^{\mathrm{R}})$ is a smooth function of (y, z) when $(y, z) \in (0, v_i) \times (v_i, 1)$. Note that the events E_i and $\{(u_i^{\mathrm{L}}, u_i^{\mathrm{R}}) \in (0, v_i) \times (v_i, 1)\}$ are equal up to a null set. Therefore, on this event, $(u_i^{\mathrm{L}}, u_i^{\mathrm{R}})$ is an absolutely continuous random variable with density $g : (0, v_i) \times (v_i, 1) \to \mathbb{R}$,

$$g(y, z) = -\frac{\partial^2}{\partial y \partial z} \mathbb{P}(u_i^{\mathrm{L}} \leqslant y, u_i^{\mathrm{R}} \geqslant z) = m(m-1)\,(1 - (z - y))^{m-2} \,.$$

We compute

$$\begin{aligned}
\mathbb{P}(C_i^c \cap E_i) &= \mathbb{P}(\{|u_i^{\mathrm{R}} + u_i^{\mathrm{L}} - 2v_i| \leqslant D\} \cap E_i) \\
&= \int_{\{0 < y < v_i < z < 1\}} m(m-1)\,(1 - (z - y))^{m-2}\, \mathbf{1}_{\{|y + z - 2v_i| \leqslant D\}} \mathrm{d}y \mathrm{d}z \,.
\end{aligned}$$

We make the change of variables $\theta = z - y$, $\nu = z + y$.

$$\begin{aligned}
\mathbb{P}(C_i^c \cap E_i) &= \frac{m(m-1)}{2} \int_{\{0 < \frac{\nu - \theta}{2} < v_i < \frac{\nu + \theta}{2} < 1\}} (1 - \theta)^{m-2} \mathbf{1}_{|\nu - 2v_i| \leqslant D} \mathrm{d}\theta \mathrm{d}\nu \\
&\leqslant \frac{m(m-1)}{2} \Big(\int_0^1 (1 - \theta)^{m-2} \mathrm{d}\theta \Big) \Big(\int_{-\infty}^{\infty} \mathbf{1}_{|\nu - 2v_i| \leqslant D} \mathrm{d}\nu \Big) \\
&= Dm \,.
\end{aligned}$$

Using $m \geqslant 24/\Delta v \geqslant 24n$, we have

$$\sum_{i=1}^{n-1} \mathbb{P}(C_i^c \cap E_i) \leqslant (n-1)Dm \leqslant \frac{\delta}{24 \times 6} \leqslant \frac{\delta}{3} \,.$$

This concludes the proof.

□

B Proofs of the results of the main text

B.1 Proof of Proposition 1

Let us lower-bound the smallest eigenvalue of $H_\eta(u)$ which is equal to

$$\min_{\|a\|=1} a^\top H_\eta(u) a\,.$$

Now for $a \in \mathbb{R}^{m+1}$ such that $\|a\| = 1$,

$$a^\top H_\eta(u) a = \sum_{i,j=0}^{m} a_i a_j \int_0^1 \sigma_\eta(x-u_i)\sigma_\eta(x-u_j)\mathrm{d}x = \int_0^1 \left(\sum_{i=0}^{m} a_i \sigma_\eta(x-u_i)\right)^2 \mathrm{d}x\,.$$

Since $\Delta u > 2\eta$ (because $u \in \mathcal{U}$) and $u_0 = -\eta/2$, $u_{m+1} = 1+\eta/2$, the intervals $[u_i+\eta/2, u_{i+1}-\eta/2]$ for $i \in \{0,\dots,m\}$ are disjoint and included in $[0,1]$. Thus

$$a^\top H_\eta(u) a \geqslant \sum_{i=0}^{m} \int_{u_i+\eta/2}^{u_{i+1}-\eta/2} \left(\sum_{i=0}^{m} a_i \sigma_\eta(x-u_i)\right)^2 \mathrm{d}x\,.$$

Since $\sigma(x) = 0$ if $x < -1/2$ and $\sigma(x) = 1$ if $x > 1/2$, we have that $\sigma_\eta(x) = 0$ if $x < -\eta/2$ and $\sigma_\eta(x) = 1$ if $x > \eta/2$. Further recall that the u_i are ordered in increasing order. As a consequence,

$$\begin{aligned} a^\top H_\eta(u) a &\geqslant \sum_{i=0}^{m} \int_{u_i+\eta/2}^{u_{i+1}-\eta/2} \left(\sum_{k=0}^{i} a_k\right)^2 \mathrm{d}x \\ &= \sum_{i=0}^{m} (u_{i+1}-u_i-\eta) \left(\sum_{k=0}^{i} a_k\right)^2 \\ &\geqslant \frac{\Delta(u)}{2} \sum_{i=0}^{m} \left(\sum_{k=0}^{i} a_k\right)^2\,, \end{aligned} \tag{12}$$

where in the last step, we used that $\Delta(u) > 2\eta$ and thus $u_{i+1} - u_i - \eta \geqslant \Delta(u) - \eta \geqslant \Delta(u) - \Delta(u)/2 = \Delta(u)/2$. Now, denote $c_0 = 0$ and $c_i = \sum_{k=0}^{i-1} a_k$. Then $\|a\| = 1$ writes

$$\sum_{i=0}^{m} (c_{i+1}-c_i)^2 = 1.$$

Furthermore,

$$\sum_{i=0}^{m} (c_{i+1}-c_i)^2 = \sum_{i=0}^{m} c_{i+1}^2 + \sum_{i=0}^{m} c_i^2 - 2\sum_{i=0}^{m} c_{i+1}c_i \leqslant 4\sum_{i=0}^{m+1} c_i^2\,.$$

Hence

$$\sum_{i=0}^{m+1} c_i^2 \geqslant \frac{1}{4}\,,$$

which shows in conjunction with (12) that the smallest eigenvalue of $H_\eta(u)$ is lower-bounded by $\frac{\Delta u}{8}$.

B.2 Proof of Proposition 2

To show that $G(u) = (\nabla_u L_\eta)(a^*_\eta(u), u)$ is Lipschitz-continuous on $\mathcal{U}_\eta$, we show that it is differentiable on $\mathcal{U}_\eta$ and that its derivatives are uniformly bounded. The chain rule gives

$$\frac{\partial G_j}{\partial u_k} = \sum_{l=0}^{m} \frac{\partial a^*_{\eta,l}}{\partial u_k}(u) \frac{\partial^2 L_\eta}{\partial u_j \partial a_l}(a^*_\eta(u), u) + \frac{\partial^2 L_\eta}{\partial u_j \partial u_k}(a^*_\eta(u), u)\,.$$

From (10), using that σ is twice continuously differentiable, it can be checked that $\frac{\partial L_\eta}{\partial u_j}$ is differentiable in both its arguments and its derivatives are uniformly upper-bounded when a is bounded. Furthermore, for $u \in \mathcal{U}_\eta$,

$$\|a_\eta^*(u)\| \leqslant \|H_\eta(u)^{-1}\| \, \|b_\eta(u)\| \leqslant \frac{8M\sqrt{m+1}}{\Delta(u)} \,,$$

by Lemma 1 and Proposition 1. Finally, according to Lemma 3, a_η^* is differentiable with derivatives uniformly upper-bounded on $\mathcal{U}_\eta$. This concludes the proof.

B.3 Proof of Proposition 3

In this proof, we denote $u_i^{\mathrm{L}}(\tau)$ (resp. $u_i^{\mathrm{R}}(\tau)$) the position at time τ of the neuron that is *at initialization* closest to v_i to the left (resp. the right). Note that because of the movement of the neurons, it could be that u_i^{L} (resp. u_i^{R}) does not remain the neuron closest to the left (resp. the right) throughout the dynamics. Our proof discusses when this phenomenon occurs. Similarly, denote u_i^{LL} (resp. u_i^{RR}) the neuron second closest to the left (resp. the right) of v_i. Since the initialization is D-good, note that all these neurons are distinct.

Denote $\overline{\mathcal{T}}$ the minimal time $\tau \in [0, \mathcal{T}_{\max})$ such that $\Delta(u(\tau)) \leqslant D/2$ or there are less than two neurons in some piece $[v_i, v_{i+1}]$ of f^*. Note that by assumption, $\Delta(u(0)) \geqslant D > D/2$ and there are at least 6 neurons in each interval at initialization, thus $\overline{\mathcal{T}} > 0$. Furthermore, using the trivial inequalities $M \geqslant \Delta f/2$, $m+1 \geqslant 1$ and $\eta^{1/2} \geqslant \eta$, we have $\frac{D}{2} = \frac{2^{11/2} M\sqrt{m+1}\sqrt{\eta}}{\Delta f} \geqslant 8\eta > 2\eta$. Recall that 2η is the quantity defining the set $\mathcal{U}_\eta$ supporting the maximal solution of the equation (8). As a consequence, we do have $\overline{\mathcal{T}} < \mathcal{T}_{\max}$. At the end of the proof, we check that $\mathcal{T} < \overline{\mathcal{T}}$, by controlling carefully the movement of each neuron.

Let us first bound the difference between the dynamics of u and the dynamics that we would have if at each time τ, the weights a were given by $a_0^*(u(\tau))$, the best approximation of f^* by a piecewise constant function with subdivision $u(\tau)$. For any $\tau < \overline{\mathcal{T}}$ and $j \in \{1, \dots, m\}$, by Lemma 6, we have

$$\begin{aligned}
&\Big|\frac{\mathrm{d}u_j}{\mathrm{d}\tau}(\tau) + \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(\tau)), u(\tau))\Big| \\
&\quad = \Big|\frac{\partial L_\eta}{\partial u_j}(a_\eta^*(u(\tau)), u(\tau)) - \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(\tau)), u(\tau))\Big| \\
&\quad \leqslant 2M(\sqrt{m+1}+1)\|a_\eta^*(u(\tau)) - a_0^*(u(\tau))\| + \sqrt{m+1}\|a_\eta^*(u(\tau)) - a_0^*(u(\tau))\|^2 \,. \qquad (13)
\end{aligned}$$

We are therefore led to bounding $\|a_\eta^*(u(\tau)) - a_0^*(u(\tau))\|$, as follows:

$$\begin{aligned}
\|a_\eta^*(u(\tau)) - a_0^*(u(\tau))\| &\leqslant \frac{2^4 M\sqrt{m+1}\eta}{\Delta(u(\tau))} && \text{(by Lemma 7)} \\
&\leqslant \frac{2^5 M\sqrt{m+1}\eta}{D} && \text{(since } \Delta(u(\tau)) \geqslant D/2\text{)} \\
&= \frac{D(\Delta f)^2}{2^8 M\sqrt{m+1}} && \text{(by definition of } D\text{)}.
\end{aligned}$$

Then the first term in (13) is less than

$$\frac{(\sqrt{m+1}+1)D(\Delta f)^2}{2^7\sqrt{m+1}} \leqslant \frac{D(\Delta f)^2}{2^6} \,,$$

and the second term in (13) is less than

$$\frac{D^2(\Delta f)^4}{2^{16}M^2\sqrt{m+1}} \leqslant \frac{D(\Delta f)^2}{2^{14}} \,, \qquad \text{using } D \leqslant \Delta(u(0)) \leqslant 1,\ \Delta f \leqslant 2M \text{ and } m+1 \geqslant 1.$$

Hence we obtain, for any $\tau < \overline{\mathcal{T}}$ and $j \in \{1, \dots, m\}$,

$$\Big|\frac{\mathrm{d}u_j}{\mathrm{d}\tau}(\tau) + \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(\tau)), u(\tau))\Big| \leqslant \frac{D(\Delta f)^2}{60} =: \Delta g \qquad (14)$$

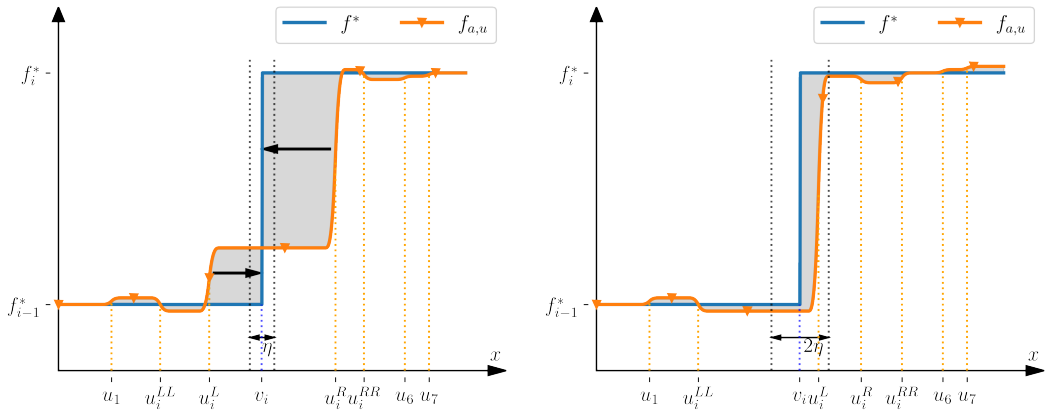
Now, let us examine how the neurons move, by leveraging Lemma 4 that gives exact formulae for $\frac{\partial L_\eta}{\partial u_j}(a_0^*(u(\tau)), u(\tau))$. First, if u_j is not next to a discontinuity, $\frac{\partial L_\eta}{\partial u_j}(a_0^*(u(\tau)), u(\tau)) = 0$, hence

$$|u_j(\tau) - u_j(0)| \leqslant (\Delta g)\tau\,.$$

Let us now study what happens next to a discontinuity v_i. Denote $(\delta f)_i = f_i^* - f_{i-1}^*$. W.l.o.g., assume that

$$u_i^{\mathrm{R}}(0) - v_i > v_i - u_i^{\mathrm{L}}(0)\,.$$

In the reverse case, the proof is the same by swapping the roles of u_i^{L} and u_i^{R}, and of u_i^{LL} and u_i^{RR}. We are going to show that the dynamics of u_i^{L} are divided into two phases. Define $\mathcal{T}_i$ as the minimal $\tau \in [0, \overline{\mathcal{T}}]$ such that $u_i^{\mathrm{L}}(\tau) = v_i - \eta/2$. In the first phase $[0, \mathcal{T}_i]$, we have $u_i^{\mathrm{L}}(\tau) < v_i - \frac{\eta}{2}$ and u_i^{L} moves towards v_i. In the second phase $[\mathcal{T}_i, \overline{\mathcal{T}}]$, we show below that $u_i^{\mathrm{L}}(\tau) \in [v_i - \eta, v_i + \eta]$. Note that we can have $\mathcal{T}_i = 0$ if $u_i^{\mathrm{L}}(0) \geqslant v_i - \frac{\eta}{2}$. It is also possible that $\mathcal{T}_i = \infty$ a priori; this means that the second phase does not exist. We show below that this case does not happen. Figure 6 depicts the two phases.



(a) Phase 1 (b) Phase 2

Figure 6: Dynamics of the neurons next to a discontinuity v_i. In the first phase, u_i^{L} and u_i^{R} move towards v_i, until the closest neuron (in this case u_i^{L}) reaches the interval of size η centered in v_i. In the second phase, u_i^{L} remains in an interval of size 2η around v_i, and none of the neurons move significantly.

Beginning by the first phase, we have, while $u_i^{\mathrm{L}}(\tau) < v_i - \frac{\eta}{2}$ and $u_i^{\mathrm{R}}(\tau) > v_i + \frac{\eta}{2}$, according to Lemma 4,

$$\frac{\partial L_\eta}{\partial u_i^{\mathrm{L}}}(a_0^*(u(\tau)), u(\tau)) = -\frac{1}{2}\frac{(u_i^{\mathrm{R}}(\tau) - v_i)^2(\delta f)_i^2}{(u_i^{\mathrm{R}}(\tau) - u_i^{\mathrm{L}}(\tau))^2}\,,$$
$$\frac{\partial L_\eta}{\partial u_i^{\mathrm{R}}}(a_0^*(u(\tau)), u(\tau)) = \frac{1}{2}\frac{(v_i - u_i^{\mathrm{L}}(\tau))^2(\delta f)_i^2}{(u_i^{\mathrm{R}}(\tau) - u_i^{\mathrm{L}}(\tau))^2}\,.$$

For ease of computation, let $d_i^{\mathrm{L}}(\tau) = v_i - u_i^{\mathrm{L}}(\tau)$ and $d_i^{\mathrm{R}}(\tau) = u_i^{\mathrm{R}}(\tau) - v_i$ be the distances between the neurons and v_i. Then, by (14),

$$\begin{aligned}\frac{\mathrm{d}d_i^{\mathrm{R}}}{\mathrm{d}\tau}(\tau) + \frac{\mathrm{d}d_i^{\mathrm{L}}}{\mathrm{d}\tau}(\tau) &\leqslant -\frac{1}{2}\frac{((d_i^{\mathrm{R}}(\tau))^2 + (d_i^{\mathrm{L}}(\tau))^2)(\delta f)_i^2}{(d_i^{\mathrm{L}}(\tau) + d_i^{\mathrm{R}}(\tau))^2} + 2\Delta g \\ &\leqslant -\frac{(\Delta f)^2}{4} + 2\frac{D(\Delta f)^2}{60} \leqslant -\frac{(\Delta f)^2}{5}\end{aligned}$$

since $D \leqslant \Delta(u(0)) \leqslant 1$. Thus, in some time less than $\mathcal{T} = \frac{6}{(\Delta f)^2}$, $d_i^{\mathrm{R}}(\tau) + d_i^{\mathrm{L}}(\tau) \leqslant \eta$, that is, either u_i^{L} reaches $v_i - \frac{\eta}{2}$ or u_i^{R} reaches $v_i + \frac{\eta}{2}$. Let us check that the second event cannot actually

happen: while $u_i^{\mathrm{L}}(\tau) < v_i - \frac{\eta}{2}$ and $u_i^{\mathrm{R}}(\tau) > v_i + \frac{\eta}{2}$, we also have

$$\begin{aligned}\frac{\mathrm{d}d_i^{\mathrm{R}}}{\mathrm{d}\tau}(\tau) - \frac{\mathrm{d}d_i^{\mathrm{L}}}{\mathrm{d}\tau}(\tau) &\geqslant \frac{((d_i^{\mathrm{R}}(\tau))^2 - (d_i^{\mathrm{L}}(\tau))^2)(\delta f)_i^2}{(d_i^{\mathrm{L}}(\tau) + d_i^{\mathrm{R}}(\tau))^2} - 2\Delta g \\ &= \frac{(d_i^{\mathrm{R}}(\tau) - d_i^{\mathrm{L}}(\tau))(\delta f)_i^2}{d_i^{\mathrm{L}}(\tau) + d_i^{\mathrm{R}}(\tau)} - 2\Delta g\,.\end{aligned}$$

By condition (c) of Definition 2 and by (14), we have $d_i^{\mathrm{R}}(0) - d_i^{\mathrm{L}}(0) \geqslant D = \frac{60\Delta g}{(\Delta f)^2} \geqslant \frac{60\Delta g}{(\delta f)_i^2}$, and furthermore $d_i^{\mathrm{L}}(\tau) + d_i^{\mathrm{R}}(\tau) \leqslant 1$. An easy reasoning then shows that $d_i^{\mathrm{R}} - d_i^{\mathrm{L}}$ is increasing. Therefore u_i^{R} must remain further away from v_i than u_i^{L}.

In summary, we showed that there exists some time $\mathcal{T}_i \leqslant \mathcal{T}$ when $u_i^{\mathrm{L}}(\mathcal{T}_i) = v_i - \frac{\eta}{2}$, which marks the end of the first phase, and we also have

$$d_i^{\mathrm{R}}(\mathcal{T}_i) - d_i^{\mathrm{L}}(\mathcal{T}_i) \geqslant d_i^{\mathrm{R}}(0) - d_i^{\mathrm{L}}(0) \geqslant D\,.$$

Moving on to the study of the second phase, let us show that $u_i^{\mathrm{L}}(\tau)$ stays in the interval $[v_i - \eta, v_i + \eta]$ for $\tau \in [\mathcal{T}_i, \overline{\mathcal{T}})$. Consider any $\tau \leqslant \overline{\mathcal{T}}$ such that $u_i^{\mathrm{L}}(\tau) = v_i - \eta$. Then we have by (14) and Lemma 4

$$\frac{\mathrm{d}u_i^{\mathrm{L}}}{\mathrm{d}\tau}(\tau) \geqslant \frac{(u_i^{\mathrm{R}}(\tau) - v_i)^2(\delta f)_i^2}{(u_i^{\mathrm{R}}(\tau) - v_i + \eta)^2} - \Delta g \geqslant \Delta g\,, \tag{15}$$

where the second upper bound comes from the fact that we have $u_i^{\mathrm{R}}(\tau) - v_i \geqslant \frac{D}{2} - \eta$ since $\Delta(u(\tau)) \geqslant D/2$, and furthermore, $x \mapsto \frac{x^2}{(x+\eta)^2}$ is increasing, hence

$$\frac{(u_i^{\mathrm{R}} - v_i)^2(\delta f)_i^2}{(u_i^{\mathrm{R}} - v_i + \eta)^2} \geqslant \Big(\frac{\frac{D}{2} - \eta}{\frac{D}{2}}\Big)^2 \Delta f^2 \underset{(D/2 \geqslant 2\eta)}{\geqslant} \frac{(\Delta f)^2}{4} \geqslant 2\Delta g\,.$$

Equation (15) implies that $u_i^{\mathrm{L}}(\tau) \geqslant v_i - \eta$ for all $\tau \in [\mathcal{T}_i, \overline{\mathcal{T}})$. Similarly, consider any $\tau \leqslant \overline{\mathcal{T}}$ such that $u_i^{\mathrm{L}}(\tau) = v_i + \eta$. Note that, for such a τ, $u_i^{\mathrm{L}}(\tau)$ is now on the right of v_i, and the neurons flanking v_i are u_i^{LL} and u_i^{L}. Thus we have by (14) and Lemma 4

$$\frac{\mathrm{d}u_i^{\mathrm{L}}}{\mathrm{d}\tau}(\tau) \leqslant -\frac{(v_i - u_i^{\mathrm{LL}}(\tau))^2(\delta f)_i^2}{(v_i + \eta - u_i^{\mathrm{LL}}(\tau))^2} + \Delta g \leqslant -\Delta g\,,$$

where the second lower bound unfolds similarly as previously. This shows that $u_i^{\mathrm{L}}(\tau) \leqslant v_i + \eta$ for all $\tau \in [\mathcal{T}_i, \overline{\mathcal{T}})$.

We now check that $\mathcal{T} < \overline{\mathcal{T}}$, that is, for all $\tau \leqslant \mathcal{T}$, $\Delta(u(\tau)) > D/2$ and there are at least two neurons in each interval $[v_i, v_{i+1}]$. Starting with the first condition, we say that neurons u_j and u_k collide if $|u_j(\tau) - u_k(\tau)| = D/2$ for some $\tau \leqslant \mathcal{T}$. Let us show that no pair of neurons collide.

We start by showing that there is no collision between u_i^{LL} and u_i^{L}. In the first phase $[0, \mathcal{T}_i]$, we have $\frac{\mathrm{d}u_i^{\mathrm{LL}}}{\mathrm{d}\tau}(\tau) \leqslant \Delta g$. Recall that we also have $\frac{\mathrm{d}u_i^{\mathrm{L}}}{\mathrm{d}\tau}(\tau) \geqslant -\Delta g$ and thus for $\tau \leqslant \mathcal{T}_i$,

$$u_i^{\mathrm{L}}(\tau) - u_i^{\mathrm{LL}}(\tau) \geqslant u_i^{\mathrm{L}}(0) - u_i^{\mathrm{LL}}(0) - 2\mathcal{T}\Delta g \geqslant \frac{4D}{5}$$

since $u_i^{\mathrm{L}}(0) - u_i^{\mathrm{LL}}(0) \geqslant D$ and $\mathcal{T}\Delta g = D/10$ by definition of $\mathcal{T}$ and Δg. As a consequence, u_i^{LL} and u_i^{L} do not collide during the first phase, and we have

$$u_i^{\mathrm{LL}}(\mathcal{T}_i) \leqslant u_i^{\mathrm{L}}(\mathcal{T}_i) - \frac{4D}{5} = v_i - \frac{\eta}{2} - \frac{4D}{5}\,. \tag{16}$$

In the second phase, we can have $u_i^{\mathrm{L}} \in [v_i, v_i + \eta]$ in which case u_i^{LL} becomes the neuron flanking v_i to the left and u_i^{L} the neuron flanking to the right. Then (14) and Lemma 4 give

$$\frac{\mathrm{d}u_i^{\mathrm{LL}}}{\mathrm{d}\tau} \leqslant \frac{(u_i^{\mathrm{L}}(\tau) - v_i)^2(\delta f)_i^2}{(u_i^{\mathrm{L}}(\tau) - u_i^{\mathrm{LL}}(\tau))^2} + \Delta g \leqslant \frac{16\eta^2 M^2}{D^2} + \Delta g\,.$$

Of course, this bound also holds when $u_i^{\mathrm{L}} \in [v_i - \eta, v_i]$, because then $\frac{\mathrm{d}u_i^{\mathrm{LL}}}{\mathrm{d}\tau} \leqslant \Delta g$. Thus, in the second phase $\tau \in [\mathcal{T}_i, \mathcal{T}]$, by the previous upperbound and the fact that $u_i^{\mathrm{L}}(\tau) \geqslant v_i - \frac{\eta}{2}$,

$$\begin{aligned} u_i^{\mathrm{L}}(\tau) - u_i^{\mathrm{LL}}(\tau) &\geqslant v_i - \frac{\eta}{2} - \Big(u_i^{\mathrm{LL}}(\mathcal{T}_i) + (\tau - \mathcal{T}_i)\Big(\frac{16\eta^2 M^2}{D^2} + \Delta g\Big)\Big) \\ &\geqslant \frac{4D}{5} - \mathcal{T}\Big(\frac{16\eta^2 M^2}{D^2} + \Delta g\Big)\,, \end{aligned}$$

by (16). Let us now upper-bound each of the last two terms by $D/10$ to conclude. By definition of D,

$$\eta = \frac{(\Delta f)^2 D^2}{2^{13}(m+1)M^2}\,.$$

Thus

$$\frac{16\eta^2 M^2 \mathcal{T}}{D^2} = \frac{3(\Delta f)^2 D^2}{2^{21}(m+1)^2 M^2} \leqslant \frac{D}{10}$$

using the definition of $\mathcal{T}$, $D \leqslant \Delta(u(0)) \leqslant 1$, $\Delta f \leqslant 2M$ and $m + 1 \geqslant 1$. Finally, $\mathcal{T}\Delta g = D/10$. Thus u_i^{LL} and u_i^{L} do not collide.

We now show that u_i^{L} and u_i^{R} do not collide. In the first phase $\tau \in [0, \mathcal{T}_i]$, we have

$$u_i^{\mathrm{R}}(\tau) - u_i^{\mathrm{L}}(\tau) \geqslant u_i^{\mathrm{R}}(\tau) - v_i = d_i^{\mathrm{R}}(\tau) \geqslant d_i^{\mathrm{R}}(\tau) - d_i^{\mathrm{L}}(\tau) \geqslant D\,.$$

As a consequence, u_i^{L} and u_i^{R} do not collide during the first phase, and we have

$$u_i^{\mathrm{R}}(\mathcal{T}_i) \geqslant D + u_i^{\mathrm{L}}(\mathcal{T}_i) = D + v_i - \frac{\eta}{2}\,. \tag{17}$$

In the second phase, u_i^{R} plays a role symmetric to u_i^{LL}: it can be, or not, the neuron closest to the right of v_i, depending on whether $u_i^{\mathrm{L}} \in [v_i - \eta, v_i]$ or $u_i^{\mathrm{L}} \in [v_i, v_i + \eta]$. As for u_i^{LL}, we can show that in any case, for $\tau \in [\mathcal{T}_i, \mathcal{T}]$,

$$\frac{\mathrm{d}u_i^{\mathrm{R}}}{\mathrm{d}\tau} \geqslant -\frac{16\eta^2 M^2}{D^2} - \Delta g\,.$$

Thus one concludes as before: for $\tau \in [\mathcal{T}_i, \mathcal{T}]$, by the previous lowerbound and the fact that $u_i^{\mathrm{L}}(\tau) \leqslant v_i + \frac{\eta}{2}$,

$$u_i^{\mathrm{R}}(\tau) - u_i^{\mathrm{L}}(\tau) \geqslant u_i^{\mathrm{R}}(\mathcal{T}_i) - (\tau - \mathcal{T}_i)\Big(\frac{16\eta^2 M^2}{D^2} + \Delta g\Big) - \Big(v_i + \frac{\eta}{2}\Big)\,.$$

Then, by (17),

$$u_i^{\mathrm{R}}(\tau) - u_i^{\mathrm{L}}(\tau) \geqslant D - \eta - \mathcal{T}\Big(\frac{16\eta^2 M^2}{D^2} + \Delta g\Big) > \frac{D}{2}\,,$$

where the last lower-bound unfolds similarly as for u_i^{LL} and u_i^{L}. Thus there is no collision between u_i^{L} and u_i^{R}.

The reader can check that all other pairs of neurons do not collide, including those involving $u_0 = -\eta/2$ and $u_{m+1} = 1 + \eta/2$. In fact, the proof is easier than for $u_i^{\mathrm{LL}}, u_i^{\mathrm{L}}$ and $u_i^{\mathrm{L}}, u_i^{\mathrm{R}}$ because the discontinuity at v_i attracts these neurons together.

Furthermore, we proved that before time $\mathcal{T}$ at most one neuron can escape on each side of a piece $[v_i, v_{i+1}]$ of f. Since we start with at least four (and even six) neurons per piece, there is always before $\mathcal{T}$ at least two neurons per piece.

This shows that $\mathcal{T} < \overline{\mathcal{T}}$, and we also proved that at time $\mathcal{T}$, all discontinuities have finished their first phase, hence there is a neuron at distance less than η from each discontinuity of the target function.

B.4 Proof of Theorem 2

Take $C = 2^{-19}$. Then by assumption of Theorem 2,

$$\eta \leqslant \frac{\delta^2 (\Delta f)^2}{2^{19} M^2 (m+1)^5}\,.$$

Moreover, by the definition of D from Proposition 3,

$$\eta = \frac{(\Delta f)^2 D^2}{2^{13} M^2 (m+1)} .$$

This implies that

$$D^2 \leqslant \frac{\delta^2}{2^6 (m+1)^4} ,$$

and in consequence

$$D \leqslant \frac{\delta}{6(m+1)^2} .$$

Then Lemma 10 shows that the initialization is D-good with probability at least $1-\delta$ (since the D-good property is monotonous in D).

Hence, with probability at least $1-\delta$, according to Proposition 3, the maximal solution to (8) is defined at least until $\mathcal{T}$ and at that time, there is a neuron at distance less than η from each discontinuity of the target function. Furthermore, $3\eta \leqslant \frac{1}{m+1} \leqslant \frac{1}{n} \leqslant \Delta v$, hence Lemma 9 applies. This implies that

$$\int_0^1 |f^*(x) - f(x; a^*(u(\mathcal{T})), u(\mathcal{T}))|^2 \mathrm{d}x \leqslant 6M^2 \eta n .$$

The assumption on η allow to conclude that the upper-bound is less than ξ.

Remark 2. *We did not try to optimize the value of C since our goal was to show convergence to a global optimum and the dependency of the dynamics on the parameters (for instance, it is remarkable that $\mathcal{T}$ does not depend on ξ).*

B.5 Proof of Proposition 4

For $s \leqslant t$, Proposition 1 holds since for $\Delta(u(s)) \geqslant 16\eta > 2\eta$. Thus $a_\eta^*(u(s))$ is well-defined and verifies

$$\nabla_a L_\eta(a_\eta^*(u(s)), u(s)) = 0 .$$

Let, for $s \leqslant t$, $V(s) = \|a(s) - a_\eta^*(u(s))\|$. Recall that, by (11),

$$\nabla_a L_\eta(a, u) = H_\eta(u) a - b_\eta(u) .$$

Hence, for $s \leqslant t$,

$$\begin{aligned}
\langle a(s) - a_\eta^*(u(s)), &\nabla_a L_\eta(a(s), u(s)) \rangle \\
&= \langle a(s) - a_\eta^*(u(s)), \nabla_a L_\eta(a(s), u(s)) - \nabla_a L_\eta(a_\eta^*(u(s)), u(s)) \rangle \\
&= \langle a(s) - a_\eta^*(u(s)), H_\eta(u(s))(a(s) - a_\eta^*(u(s))) \rangle \\
&\geqslant \frac{\Delta(u(s))}{8} V(s)^2 \\
&\geqslant \frac{D}{16} V(s)^2 ,
\end{aligned}$$

where the first lower bound is a consequence of Proposition 1. Then we have, for any $s \leqslant t$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\Big(\frac{1}{2} V(s)^2\Big) &= \Big\langle a(s) - a_\eta^*(u(s)), \frac{\mathrm{d}a}{\mathrm{d}s}(s) - \frac{\mathrm{d}}{\mathrm{d}s} a_\eta^*(u(s)) \Big\rangle \\
&= \Big\langle a(s) - a_\eta^*(u(s)), -\nabla_a L_\eta(a(s), u(s)) - \frac{\mathrm{d}}{\mathrm{d}s} a_\eta^*(u(s)) \Big\rangle \\
&\leqslant -\frac{D}{16} V(s)^2 + \Big\| \frac{\mathrm{d}}{\mathrm{d}s} a_\eta^*(u(s)) \Big\| V(s) .
\end{aligned}$$

Let us now upper bound the norm appearing in the second term. We first have by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}s} a_\eta^*(u(s)) = \frac{\partial a_\eta^*}{\partial u}(u(s)) \frac{\mathrm{d}u}{\mathrm{d}s}(s) .$$

By Lemma 3 (which holds since for $\Delta(u(s)) \geqslant 16\eta > 2\eta$),

$$\left\| \frac{\partial a_\eta^*}{\partial u}(u(s)) \right\| \leqslant \frac{8}{\Delta(u(s))} \big(2(m+1)\|a_\eta^*(u(s))\| + M\big) .$$

Besides,

$$\left\| \frac{\mathrm{d}u}{\mathrm{d}s}(s) \right\| \leqslant \varepsilon \|\nabla_u L_\eta(a(s), u(s))\| .$$

By Lemma 8,

$$\|\nabla_u L_\eta(a(s), u(s))\| \leqslant \sqrt{m+1}\|a(s)\|^2 + M\|a(s)\| . \tag{18}$$

Furthermore,

$$\|a(s)\| \leqslant \|a_\eta^*(u(s))\| + \|a(s) - a_\eta^*(u(s))\| = \|a_\eta^*(u(s))\| + V(s) .$$

By Lemmas 6 and 7, which apply since $\Delta(u(s)) > 2\eta$ and since there are at least two positions $u_j(s)$ in each interval $[v_i, v_{i+1}]$ for $s \leqslant t$,

$$\begin{aligned}
\|a_\eta^*(u(s))\| &\leqslant \|a_0^*(u(s))\| + \|a_0^*(u(s)) - a_\eta^*(u(s))\| \\
&\leqslant 2M\sqrt{m+1} + \frac{16M\sqrt{m+1}\eta}{\Delta(u(s))} \\
&\leqslant 2M\sqrt{m+1} + \frac{32M\sqrt{m+1}\eta}{D} \\
&\leqslant 3M\sqrt{m+1} ,
\end{aligned}$$

where the last upper bound is implied by the assumption $D \geqslant 32\eta$.

Now define $T_{\max} = \inf\big\{s \geqslant 0, V(s) > 3M\sqrt{m+1}\big\}$ and assume $s \leqslant \min(t, T_{\max})$ so that $V(s) \leqslant 3M\sqrt{m+1}$. Then we proved that $\|a(s)\| \leqslant 6M\sqrt{m+1}$. Going back to (18), we deduce that

$$\|\nabla_u L_\eta(a(s), u(s))\| \leqslant 36M^2(m+1)^{3/2} + 6M^2\sqrt{m+1} \leqslant 2^6 M^2 (m+1)^{3/2} . \tag{19}$$

Putting everything together, we obtain

$$\left\| \frac{\mathrm{d}}{\mathrm{d}s} a_\eta^*(u(s)) \right\| \leqslant \frac{2^9 M^2 (m+1)^{3/2}}{\Delta(u(s))} \big(6M(m+1)^{3/2} + M\big)\varepsilon \leqslant \frac{2^{13} M^3 (m+1)^3}{D}\varepsilon .$$

All in all,

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big(\frac{1}{2}V(s)^2\Big) \leqslant -\frac{D}{16}V(s)^2 + \frac{2^{13}M^3(m+1)^3}{D}\varepsilon V(s) .$$

Hence

$$\frac{\mathrm{d}}{\mathrm{d}s}(V(s)) = \frac{1}{V(s)}\frac{\mathrm{d}}{\mathrm{d}s}\big(\frac{1}{2}V(s)^2\big) \leqslant -\frac{D}{16}V(s) + \frac{2^{13}M^3(m+1)^3}{D}\varepsilon .$$

By Grönwall's inequality, for all $s \leqslant \min(t, T_{\max})$,

$$\begin{aligned}
V(s) &\leqslant \exp^{-\frac{D}{16}s} V(0) + \frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon(1 - \exp^{-\frac{D}{16}s}) && (20) \\
&\leqslant \exp^{-\frac{D}{16}s} V(0) + \frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon . && (21)
\end{aligned}$$

Finally note that $V(0) = \|a_\eta^*(0)\| \leqslant 2M\sqrt{m+1}$ and $\frac{2^{17}M^3(m+1)^3\varepsilon}{D^2} \leqslant 2M\sqrt{m+1}$ by the assumption of the Proposition on ε. Hence (20) implies that for all $s \leqslant \min(t, T_{\max})$, $V(s)$ is a (weighted) average of two terms less than $2M\sqrt{m+1}$ hence it is less than $2M\sqrt{m+1}$. This shows that $T_{\max} \geqslant t$, which concludes the proof since (21) is then valid for $s = t$.

B.6 Proof of Theorem 1

In the proof, we take $C_1 = 2^{-21}$ and $C_2 = 2^{-36}$. Denote

$$D = \frac{\delta}{6(m+1)^2}\,.$$

Lemma 10 shows that the initialization is D-good with probability at least $1-\delta$. In the following, we study the case where this event happens.

Denote $\overline{T}$ the minimal time $t > 0$ such that $\Delta(u(t)) \leqslant D/2$ or there are less than two neurons in some piece $[v_i, v_{i+1}]$ of f^* or $\|a(t)\| > 7M\sqrt{m+1}$. Note that $\overline{T} > 0$ since the initialization is D-good. By Lemma 8, $\nabla_u L_\eta$ and $\nabla_a L_\eta$ are Lipschitz-continuous on compacts, hence the solution of the gradient flow is well defined for $t < \overline{T}$ since $\overline{T}$ defines a compact set of parameters.

Then all the assumptions of Proposition 4 are satisfied on the time interval $[0, t]$ for any $t < \overline{T}$. More precisely, the assumptions that do not come directly from the definition of $\overline{T}$ are the lower bound for D and the upper bound for ε. The lower bound for D come from

$$D = \frac{\delta}{6(m+1)^2} \geqslant 32\eta \tag{22}$$

by (3) and the simple bounds $\delta \leqslant 1$, $\Delta f \leqslant 2M$, $m+1 \geqslant 1$. The upper bound for ε comes from (3) since

$$\varepsilon \leqslant \frac{\delta^3(\Delta f)^2}{2^{36}M^4(m+1)^{19/2}} \leqslant \frac{\delta^2}{36\cdot 2^{16}M^2(m+1)^{13/2}} = \frac{D^2}{2^{16}M^2(m+1)^{5/2}}\,,$$

where the second upper bound uses $m \geqslant 0$, $\delta \leqslant 1$ and $\Delta f \leqslant 2M$. Therefore, according to Proposition 4,

$$\|a(t) - a_\eta^*(u(t))\| \leqslant 3M\sqrt{m+1}\exp^{-\frac{D}{16}t} + \frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon\,, \tag{23}$$

Furthermore, the proof of Proposition 4 actually implies that

$$\|a_\eta^*(u(t))\| \leqslant 3M\sqrt{m+1} \quad \text{and} \quad \|a(t)\| \leqslant 6M\sqrt{m+1}. \tag{24}$$

The second bound implies that the condition $\|a(t)\| > 7M\sqrt{m+1}$ in the definition of $\overline{T}$ is actually never active. Let us distinguish between two phases: letting

$$T_0 = \frac{16}{D}\log\Big(\frac{2^{16}M^2(m+1)^3}{\delta(\Delta f)^2}\Big) = \frac{96(m+1)^2}{\delta}\log\Big(\frac{2^{16}M^2(m+1)^3}{\delta(\Delta f)^2}\Big),$$

then the first phase corresponds to $t \leqslant T_0$ and the second phase for $t \geqslant T_0$.

Analysis of the first phase. In the first phase, each neuron moves at most by

$$\varepsilon T_0 \max_j \Big|\frac{\partial L_\eta}{\partial u_j}(a(t), u(t))\Big| \leqslant \varepsilon T_0 \|\nabla_u L_\eta(a(s), u(s))\| \leqslant 2^6 \varepsilon T_0 M^2 (m+1)^{3/2}\,,$$

where the second upper bound comes from (19) in the proof of Proposition 4. This quantity is less than $\frac{D}{8}$ if

$$\frac{6144(m+1)^{7/2}M^2}{\delta}\log\Big(\frac{2^{16}M^2(m+1)^3}{\delta(\Delta f)^2}\Big)\varepsilon \leqslant \frac{\delta}{48(m+1)^2}\,.$$

Let us check this condition: we have

$$\begin{aligned}
&\frac{6144(m+1)^{7/2}M^2}{\delta}\log\Big(\frac{2^{16}M^2(m+1)^3}{\delta(\Delta f)^2}\Big)\varepsilon \\
&\qquad = \frac{16\cdot 6144(m+1)^{7/2}M^2}{\delta}\log\Big(\frac{2M^{1/8}(m+1)^{3/16}}{\delta^{1/16}(\Delta f)^{1/8}}\Big)\varepsilon \\
&\qquad \leqslant \frac{16\cdot 6144(m+1)^{7/2}M^2}{\delta}\log\Big(\frac{4M(m+1)}{\delta\Delta f}\Big)\varepsilon\,,
\end{aligned}$$

since $m+1 \geqslant 1$, $\delta \leqslant 1$, and $2M/\Delta f \geqslant 1$, hence $(2M/\Delta f)^{1/8} \leqslant 2M/\Delta f$. Next, upper-bounding $\log(x)$ by x, we have, by (3),

$$\begin{aligned}\frac{768(m+1)^{7/2}M^2}{\delta}\log\Big(\frac{2^{16}M^2(m+1)^3}{\delta(\Delta f)^2}\Big)\varepsilon &\leqslant \frac{64\cdot 6144(m+1)^{9/2}M^3}{\delta^2\Delta f}\varepsilon \\ &\leqslant \frac{6144\delta(\Delta f)}{2^{29}M(m+1)^5} \\ &\leqslant \frac{\delta}{48(m+1)^2}\end{aligned}$$

using $\Delta f \leqslant 2M$ and $m \geqslant 0$. Note that the upper bound $2^6\varepsilon T_0 M^2(m+1)^{3/2} \leqslant D/8$ also implies that

$$T_0 \leqslant \frac{D}{2^9\varepsilon M^2(m+1)^{3/2}} \leqslant \frac{1}{2\varepsilon(\Delta f)^2} = \frac{T}{12} \tag{25}$$

since $m \geqslant 0$, $D \leqslant 1$ and $\Delta f \leqslant 2M$. Since each neuron moves by at most $D/8$ in the time interval $[0, T_0]$ and since $\Delta(u(0)) \geqslant D$, we deduce that

$$\Delta(u(T_0)) \geqslant \frac{3}{4}D. \tag{26}$$

Similarly, by condition (c) of the definition of a D-good vector, for all discontinuities v_i,

$$|u_i^{\mathrm{R}}(0) + u_i^{\mathrm{L}}(0) - 2v_i| \geqslant D,$$

thus

$$|u_i^{\mathrm{R}}(T_0) + u_i^{\mathrm{L}}(T_0) - 2v_i| \geqslant \frac{3}{4}D\,. \tag{27}$$

Furthermore, there at least four neurons on each piece of f at T_0, because at most one neuron can move out of each piece by either side between 0 and T_0.

Analysis of the second phase. Let

$$\Delta a = \frac{D(\Delta f)^2}{2^9 M\sqrt{m+1}} = \frac{\delta(\Delta f)^2}{6\cdot 2^9 M(m+1)^{5/2}}\,.$$

In the second phase $t \geqslant T_0$, we are able to control by Δa the distance between $a(t)$ and the weights $a_0^*(u(t))$ that are the best approximation of f^* by a piecewise affine function with subdivision $u(t)$. To show this, first note that the first term in (23) is smaller than $\frac{\Delta a}{4}$ when

$$3M\sqrt{m+1}\exp^{-\frac{D}{16}t} \leqslant \frac{\Delta a}{4}\,.$$

which is equivalent to

$$t \geqslant \log\Big(\frac{12M\sqrt{m+1}}{\Delta a}\Big)\frac{16}{D}\,.$$

which is implied by $t \geqslant T_0$. Furthermore, the second term in (23) is smaller than $\frac{\Delta a}{4}$ because, by definition of D and by (3),

$$\frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon = \frac{36\cdot 2^{17}M^3(m+1)^7}{\delta^2}\varepsilon \leqslant \frac{6^2\delta(\Delta f)^2}{2^{19}M(m+1)^{5/2}} = \frac{6^3\Delta a}{2^{10}} \leqslant \frac{\Delta a}{4}\,.$$

Hence, for all $T_0 \leqslant t < \overline{T}$,

$$\|a(t) - a_\eta^*(u(t))\| \leqslant \frac{\Delta a}{2}\,.$$

Furthermore, note that the assumption of Lemma 7 applies for $t < \overline{T}$ since $\Delta(u(t)) \geqslant \frac{D}{2} > 2\eta$ by (22). Therefore, by Lemma 7 and by (3),

$$\begin{aligned}\|a_\eta^*(u(t)) - a_0^*(u(t))\| &\leqslant \frac{2^4 M\sqrt{m+1}}{\Delta(u(t))}\eta \\ &\leqslant \frac{2^5 M\sqrt{m+1}}{D}\eta \\ &= \frac{2^5 \cdot 6M(m+1)^{5/2}}{\delta}\eta \\ &\leqslant \frac{6\delta(\Delta f)^2}{2^{16}M(m+1)^{5/2}} \\ &= \frac{6^2\Delta a}{2^7} \leqslant \frac{\Delta a}{2}\,.\end{aligned}$$

By the triangular inequality, we deduce the upper bound that we were after, that is

$$\|a(t) - a_0^*(u(t))\| \leqslant \Delta a\,.$$

As in the proof of Proposition 3, we can now control the distance between the true dynamics and the one that we would have if the weights were equal to $a_0^*(u)$. Namely, for any $T_0 \leqslant t \leqslant \overline{T}$ and $j \in \{1, \dots, m\}$, by Lemma 6 (which applies since $\Delta(u(t)) > 2\eta$ by (22)), we have

$$\begin{aligned}&\Big|\frac{\mathrm{d}u_j}{\mathrm{d}t}(t) + \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(t)), u(t))\Big| \\ &\quad= \Big|\frac{\partial L_\eta}{\partial u_j}(a(t), u(t)) - \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(t)), u(t))\Big| \\ &\quad\leqslant 2M(\sqrt{m+1}+1)\|a(t) - a_0^*(u(t))\| + \sqrt{m+1}\|a(t) - a_0^*(u(t))\|^2\,.\end{aligned}$$

The first term is less than

$$2M(\sqrt{m+1}+1)\Delta a = \frac{(\sqrt{m+1}+1)D(\Delta f)^2}{2^8\sqrt{m+1}} \leqslant \frac{D(\Delta f)^2}{2^7}\,,$$

and the second term is less than

$$\sqrt{m+1}(\Delta a)^2 = \frac{D^2(\Delta f)^4}{2^{18}M^2\sqrt{m+1}} \leqslant \frac{D(\Delta f)^2}{2^{16}}\,,$$

using $D \leqslant \Delta(u(0)) \leqslant 1$, $\Delta f \leqslant 2M$ and $m+1 \geqslant 1$. Hence we obtain, for any $T_0 \leqslant t \leqslant \overline{T}$ and $j \in \{1, \dots, m\}$,

$$\Big|\frac{\mathrm{d}u_j}{\mathrm{d}t}(t) + \frac{\partial L_\eta}{\partial u_j}(a_0^*(u(t)), u(t))\Big| \leqslant \frac{D(\Delta f)^2}{120}\,.$$

We are therefore in a situation very similar to the proof of Proposition 3, starting from (14). One can check that all the arguments used in the proof also apply here. On top of the estimate above that resembles (14), the crucial facts that make the argument of Proposition 3 work here are the bounds (26) and (27) as well as the fact that there are at least four neurons on each piece f at T_0, which together are very similar to the conditions ensuring that $u(0)$ is D-good in the proof of Proposition 3. Another key point is (25), ensuring that a time at least equal to $11\overline{T}/12$ remains after the first phase of this proof, which is enough time for the dynamics described in the proof of Proposition 3 to unfold.

This yields that $T < \overline{T}$, and that at time T, there is a neuron at distance less than η from each discontinuity of f^*. Furthermore, $3\eta \leqslant \frac{1}{m+1} \leqslant \frac{1}{n} \leqslant \Delta v$, hence Lemma 9 applies. Thus

$$\int_0^1 (f_\eta(x; a_\eta^*(u(T)), u(T)) - f^*(x))^2 \mathrm{d}x \leqslant 6M^2\eta n \leqslant \frac{\xi}{2}\,,$$

where the second upper bound comes from $n \leqslant m+1$ and from (3). Furthermore, by (24) and by Lemma 8,

$$\begin{aligned}|L_\eta(a(T), u(T)) - L_\eta(a_\eta^*(u(T)), u(T))| &\leqslant \sqrt{m+1}(6M(m+1)+M)\|a(T) - a_\eta^*(u(T))\| \\ &\leqslant 16M(m+1)^{3/2}\|a(T) - a_\eta^*(u(T))\|\,.\end{aligned}$$

Let us show that this term is less than $\xi/4$. Recall that, by (23),

$$\|a(T) - a_\eta^*(u(T))\| \leqslant 3M\sqrt{m+1}\exp^{-\frac{D}{16}T} + \frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon\,.$$

By definition of D and T, by using $\exp(-x) \leqslant 1/x$ for $x \geqslant 1$ and by (3),

$$\begin{aligned}
16M(m+1)^{3/2} \cdot 3M\sqrt{m+1}\exp^{-\frac{D}{16}T} &= 48M^2(m+1)^2 \exp\Big(-\frac{\delta}{16(m+1)^2(\Delta f)^2\varepsilon}\Big) \\
&\leqslant \frac{48\cdot 16M^2(m+1)^4(\Delta f)^2}{\delta}\varepsilon \\
&\leqslant \frac{48(\Delta f)^2\delta}{2^{31}M^2(m+1)^{9/2}}\xi \\
&\leqslant \frac{\xi}{8}
\end{aligned}$$

using $\Delta f \leqslant 2M$, $\delta \leqslant 1$, and $m+1 \geqslant 1$. Furthermore, by (3), we get that

$$16M(m+1)^{3/2} \cdot \frac{2^{17}M^3(m+1)^3}{D^2}\varepsilon = \frac{36\cdot 2^{21}M^4(m+1)^{17/2}}{\delta^2}\varepsilon \leqslant \frac{\xi}{8}\,.$$

We therefore obtain the sought $\xi/4$ upper-bound and can conclude that

$$\begin{aligned}
\int_0^1 (f_\eta(x; a(T), u(T)) - f^*(x))^2 \mathrm{d}x &\leqslant \int_0^1 (f_\eta(x; a_\eta^*(u(T)), u(T)) - f^*(x))^2 \mathrm{d}x \\
&\quad + 2|L_\eta(a(T), u(T)) - L_\eta(a_\eta^*(u(T)), u(T))| \\
&\leqslant \xi\,.
\end{aligned}$$

C Experimental details

Setting Our code is available at **[XXX]**. To obtain Figures 3 and 4, we use the parameters of Table 1. For Figure 5, we use the parameters of Table 2.

Name	Value
m	20
ε	$2\cdot 10^{-5}$
η	$4\cdot 10^{-3}$
P	$1.8\cdot 10^{8}$
h	10^{-5}
Additive noise	Uniform on $[-1,1]$

Table 1: Parameters of Figures 3 and 4.

Name	Value
m	20
ε	1
η	$4\cdot 10^{-3}$
P	10^{6}
h	10^{-5}
Additive noise	Uniform on $[-1,1]$

Table 2: Parameters of Figure 5.

The number of iterations in Table 1 is much larger than the one in Table 2, due to the fact that the positions u evolve at a speed εh, which is much smaller in Table 1. However, note that it is possible

to increase h in Table 1 while keeping the same behavior (in our experiment, h is kept to the same value as in Table 2 in order to facilitate the comparison). More precisely, taking $h = 10^{-3}$ in Table 1 yields similar results while dividing the computational cost by 100.

Our target function is defined by $f^* = 1$ on $[0., 0.2], [0.35, 0.5], [0.65, 0.8]$, $f^* = 2$ on $[0.5, 0.65]$ and $f^* = 4$ elsewhere.

Additional plot We re-run the same SGD experiment as above twenty times, and plot the average $L2$ distance to the target as a function of ε, averaging over the initialization randomness and SGD randomness. This confirms that, in our setting, the SGD is able to recover the target function in the two-timescale regime ($\varepsilon \ll 1$), but fails outside of the two-timescale regime ($\varepsilon = 1$). The transition between the two regimes seems to occur for $\varepsilon \approx 0.1$.

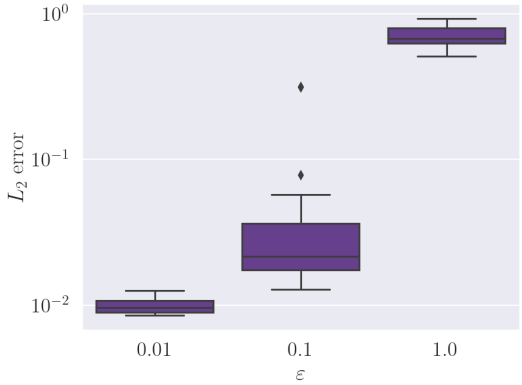


Figure 7: L_2 distance with the target as a function of ε, with 20 repeats