Supplement to the article
"Binary Classification as a Phase Separation Process"

Rafael Monteiro

Mathematics for Advanced Materials - Open Innovation Laboratory,
AIST, c/o Advanced Institute for Materials Research,
Tohoku University, Sendai, Japan
monteirodasilva-rafael@aist.go.jp, rafael.a.monteiro.math@gmail.com

September 17, 2020

Abstract

Supplementary material to the article "Binary Classification as a Phase Separation Process", by Rafael Monteiro. We present several tables with computational statistics data that can be visualized as figures in the paper. In addition to that, further details on the numerical implementation of the PSBC are given.

Contents

The tables in the first two sections are constructed using the files in the Statistics folder; see [1] for the code used to generate them, and README.pdf file for further information on how to access the data in the Statistics folder.

1 Non-diffusive PSBC in 1D (Sections 2 and 3)

The next table is related to tables 1a and 1b, with the main difference that it uses the parameters at the epoch of highest accuracy (on the training set), which is then applied to the test set.

γ^*	Average accuracy	
	Train	Test
0.6	0.957 ± 0.004	0.963 ± 0.004
0.7	0.917 ± 0.004	0.906 ± 0.003
0.8	0.859 ± 0.005	0.851 ± 0.007

(a) Weights-1-sharing (at best epoch).

γ^*	Average accuracy	
	Train	Test
0.6	1.0 ± 0.0	1.0 ± 0.0
0.7	1.0 ± 0.0	1.0 ± 0.0
0.8	1.0 ± 0.0	1.0 ± 0.0

(b) Weights-$\mathrm{N_t}$-sharing (at best epoch).

Table 1: Comparison between the accuracy in two versions of the PSBC in the form (2.1), with different types of weight sharing. The dataset obeys a train-test split of 80%-20%, and is made of 2000 points following an i.i.d. uniform distribution on $[0, 1]$ with labels (2.9). For each value of γ^* statistics were computed from a sample space of 100 simulations. Parameters are $\mathrm{N_t} = 20$, weights-$\mathrm{N_t}$-sharing, $\Delta_t^u = 0.1$ (initial), patience = $+\infty$, and learning rates $0.1 + 0.08 \cdot (0.93)^{\text{epoch}}$.

The next table is related to Tables 2a and 2b in the paper, with the main difference that it uses the parameters at the epoch of highest accuracy (on the training set), which is then applied to the test set.

γ^*	Average accuracy	
	Train	Test
0.6	0.577 ± 0.0	0.562 ± 0.0
0.7	0.668 ± 0.0	0.665 ± 0.0
0.8	0.75 ± 0.0	0.743 ± 0.0

(a) Without phase (at best epoch).

γ^*	Average accuracy	
	Train	Test
0.6	0.999 ± 0.003	0.999 ± 0.003
0.7	1.0 ± 0.0	1.0 ± 0.0
0.8	1.0 ± 0.0	1.0 ± 0.0

(b) With phase (at best epoch).

Table 2: Comparison between the accuracy in two versions of the PSBC, as evaluated at the best epoch: (a) in the form (2.1), and (b) in the form with phase, (3.4), that will be discussed in the next section. The dataset obeys a train-test split of 80%-20%, and is made of 2000 points following an i.i.d. uniform distribution on $[0, 1]$ with labels (2.12). For each value of γ^* statistics were computed from a sample space of 100 simulations. Parameters are $\mathrm{N_t} = 20$, weights-$\mathrm{N_t}$-sharing, $\Delta_t^u = 0.1$ (initial), patience = $+\infty$, and learning rates $0.1 + 0.08 \cdot (0.93)^{\text{epoch}}$.

2 Non-diffusive PSBC (Section 4)

Average of maximum throughout epochs (weights-1-sharing)				
N_{ptt}	Non-subordinate		Subordinate	
	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
78	1.7077 ± 0.058	17.7811 ± 0.0385	1.6628 ± 0.0365	17.9575 ± 0.081
87	1.6337 ± 0.0481	18.1841 ± 0.0731	1.6477 ± 0.0552	18.3613 ± 0.0617
98	1.6073 ± 0.0443	18.7144 ± 0.0769	1.6032 ± 0.0363	18.8444 ± 0.0741
112	1.5467 ± 0.0385	19.292 ± 0.0872	1.5494 ± 0.0523	19.5001 ± 0.0294
130	1.5516 ± 0.0728	20.1866 ± 0.0834	1.5533 ± 0.0552	20.4014 ± 0.0864
156	1.5178 ± 0.0505	21.3579 ± 0.0746	1.5416 ± 0.0657	21.5895 ± 0.0687
196	1.5125 ± 0.0519	23.1765 ± 0.1059	1.4703 ± 0.0343	23.351 ± 0.0977
261	1.4983 ± 0.0519	25.8914 ± 0.0901	1.4787 ± 0.0344	26.2361 ± 0.0744
392	1.4943 ± 0.0279	30.8686 ± 0.1385	1.4864 ± 0.0289	31.1317 ± 0.0902
784	1.0 ± 0.0	13.5509 ± 0.0098	1.0 ± 0.0	13.8641 ± 0.0103
Average of maximum throughout epochs (weights-N_t-sharing)				
N_{ptt}	Non-subordinate		Subordinate	
	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
78	1.4329 ± 0.031	15.3109 ± 0.1249	1.4274 ± 0.0663	15.7929 ± 0.1007
87	1.3975 ± 0.0571	15.7846 ± 0.1153	1.395 ± 0.0364	16.2141 ± 0.0628
98	1.3971 ± 0.0357	16.4919 ± 0.0772	1.4061 ± 0.0447	16.7979 ± 0.0868
112	1.3967 ± 0.0361	17.2627 ± 0.1326	1.4109 ± 0.0439	17.5361 ± 0.1452
130	1.3735 ± 0.0537	18.3488 ± 0.0719	1.3625 ± 0.0265	18.5868 ± 0.071
156	1.4004 ± 0.0324	19.8166 ± 0.0742	1.3659 ± 0.0369	20.0263 ± 0.1036
196	1.3896 ± 0.0342	21.8472 ± 0.1477	1.3924 ± 0.0371	22.1283 ± 0.1574
261	1.2867 ± 0.1487	21.5148 ± 5.2528	1.3815 ± 0.023	24.54 ± 0.0232
392	1.0 ± 0.0	10.9772 ± 0.0098	1.0 ± 0.0	11.7193 ± 0.0119
784	1.0 ± 0.0	10.9321 ± 0.011	1.0 ± 0.0	11.6689 ± 0.0077

Table 3: Average and standard deviation for the maximum value attained by the diameter of the set $\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor} := \operatorname{conv}\left(\{0,1\}\cup_{m=0}^{N_t-1}\{\alpha^{\lfloor m\rfloor}\}\right)$ over epochs; $\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}$ is defined similarly. This quantity immediately gives an estimate on the size of trainable weights (in ℓ^∞-norm) thanks to the relation $\max_{0\le n\le N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\} \le \operatorname{diam}(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}) \le 2\max_{0\le n\le N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\}$. The model in display is a non-diffusive PSBC with parameters $N_t = 2$, $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), and patience = 50. . Learning rates were chosen according to Appendix A, at $N_{ptt} = 196$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. Note that, even though each simulation has assigned to it a maximum value $Q = 600$ of epochs, each one of them may stop earlier due to *Early stopping*; see further information in Appendix A in the paper. For a visualization of the data in this table, see Figure 18 in the paper.

3 Diffusive PSBC, Neumann boundary conditions (Section 5)

We remark that values in the following tables agree when $N_t = 1$ because weights-1-sharing and weights-N_t-models coincide in that case.

Average accuracy (weights-1-sharing)				
	$N_t = 1$		$N_t = 2$	
ε	Train	Test	Train	Test
0	0.93824 ± 0.00016	0.93674 ± 0.00016	0.93821 ± 0.00015	0.93664 ± 0.00015
2^{-10}	0.93824 ± 0.00014	0.93674 ± 0.00014	0.93821 ± 0.00013	0.93674 ± 0.00013
2^{-9}	0.93823 ± 0.00011	0.93664 ± 0.00011	0.93823 ± 0.00015	0.93671 ± 0.00015
2^{-8}	0.93823 ± 9e-05	0.93667 ± 9e-05	0.9383 ± 0.00012	0.93684 ± 0.00012
2^{-7}	0.93827 ± 0.00012	0.93681 ± 0.00012	0.93808 ± 0.0001	0.93657 ± 0.0001
2^{-6}	0.93819 ± 9e-05	0.93664 ± 9e-05	0.93818 ± 0.00015	0.93684 ± 0.00015
2^{-5}	0.93819 ± 0.00012	0.93671 ± 0.00012	0.93823 ± 0.00011	0.93667 ± 0.00011
2^{-4}	0.93819 ± 0.0001	0.93667 ± 0.0001	0.93815 ± 7e-05	0.93657 ± 7e-05
2^{-3}	0.9382 ± 8e-05	0.93667 ± 8e-05	0.9382 ± 0.00011	0.93671 ± 0.00011
2^{-2}	0.93824 ± 0.0001	0.93667 ± 0.0001	0.93823 ± 9e-05	0.93664 ± 9e-05
2^{-1}	0.93817 ± 8e-05	0.93677 ± 8e-05	0.93827 ± 0.00014	0.93687 ± 0.00014
1	0.93822 ± 0.00012	0.93671 ± 0.00012	0.93815 ± 0.00014	0.93667 ± 0.00014
2	0.93819 ± 0.00011	0.93677 ± 0.00011	0.93791 ± 8e-05	0.93657 ± 8e-05
	$N_t = 4$		$N_t = 8$	
ε	Train	Test	Train	Test
0	0.93827 ± 0.00014	0.93667 ± 0.00014	0.93855 ± 0.00019	0.93677 ± 0.00019
2^{-10}	0.93838 ± 0.00014	0.93687 ± 0.00014	0.93846 ± 0.00017	0.93674 ± 0.00017
2^{-9}	0.93829 ± 0.00018	0.93681 ± 0.00018	0.93852 ± 9e-05	0.93681 ± 9e-05
2^{-8}	0.93828 ± 0.00018	0.93681 ± 0.00018	0.93861 ± 0.0001	0.93674 ± 0.0001
2^{-7}	0.93836 ± 0.00016	0.93691 ± 0.00016	0.93853 ± 0.00014	0.93674 ± 0.00014
2^{-6}	0.93842 ± 0.00011	0.93681 ± 0.00011	0.93855 ± 0.00014	0.93667 ± 0.00014
2^{-5}	0.93821 ± 0.00012	0.93671 ± 0.00012	0.93849 ± 0.00017	0.93671 ± 0.00017
2^{-4}	0.93828 ± 0.00025	0.93671 ± 0.00025	0.93862 ± 0.0002	0.93671 ± 0.0002
2^{-3}	0.93833 ± 0.00014	0.93671 ± 0.00014	0.9386 ± 0.00016	0.93674 ± 0.00016
2^{-2}	0.93834 ± 0.00016	0.93671 ± 0.00016	0.93864 ± 0.00013	0.93671 ± 0.00013
2^{-1}	0.93825 ± 0.00012	0.9366 ± 0.00012	0.93848 ± 0.00018	0.93671 ± 0.00018
1	0.93811 ± 0.00014	0.93657 ± 0.00014	0.93838 ± 0.00017	0.93674 ± 0.00017
2	0.93787 ± 0.00013	0.9365 ± 0.00013	0.93798 ± 0.00013	0.93657 ± 0.00013

Table 4: Average and standard deviation for the accuracy at the epoch with highest accuracy. The model in display is a diffusive PSBC with Neumann boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), patience = 50, $N_t \in \{1, 2, 4, 8\}$, and weights-1-sharing. Learning rates were chosen according to Appendix A in the paper, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. For a visualization of the data in this table, see Figure 15 in the paper.

Average accuracy (weights-N_t-sharing)				
ε	$N_t = 1$		$N_t = 2$	
	Train	Test	Train	Test
0	0.93824 ± 0.00016	0.93674 ± 0.00016	0.93819 ± 0.00015	0.93674 ± 0.00015
2^{-10}	0.93824 ± 0.00014	0.93674 ± 0.00014	0.93822 ± 0.00022	0.9366 ± 0.00022
2^{-9}	0.93823 ± 0.00011	0.93664 ± 0.00011	0.93817 ± 0.00012	0.93667 ± 0.00012
2^{-8}	$0.93823 \pm$ 9e-05	$0.93667 \pm$ 9e-05	0.93831 ± 0.00017	0.93677 ± 0.00017
2^{-7}	0.93827 ± 0.00012	0.93681 ± 0.00012	0.9383 ± 0.00023	0.93677 ± 0.00023
2^{-6}	$0.93819 \pm$ 9e-05	$0.93664 \pm$ 9e-05	0.93825 ± 0.00019	0.93681 ± 0.00019
2^{-5}	0.93819 ± 0.00012	0.93671 ± 0.00012	0.93821 ± 0.00021	0.93667 ± 0.00021
2^{-4}	0.93819 ± 0.0001	0.93667 ± 0.0001	0.93827 ± 0.00016	0.93674 ± 0.00016
2^{-3}	$0.9382 \pm$ 8e-05	$0.93667 \pm$ 8e-05	0.93827 ± 0.00015	0.93674 ± 0.00015
2^{-2}	0.93824 ± 0.0001	0.93667 ± 0.0001	0.93834 ± 0.00013	0.93677 ± 0.00013
2^{-1}	$0.93817 \pm$ 8e-05	$0.93677 \pm$ 8e-05	0.9383 ± 0.00017	0.93684 ± 0.00017
1	0.93822 ± 0.00012	0.93671 ± 0.00012	0.93816 ± 0.00013	0.93674 ± 0.00013
2	0.93819 ± 0.00011	0.93677 ± 0.00011	0.93801 ± 0.0001	0.9365 ± 0.0001
ε	$N_t = 4$		$N_t = 8$	
	Train	Test	Train	Test
0	0.93831 ± 0.00028	0.93687 ± 0.00028	0.93821 ± 0.00046	0.93684 ± 0.00046
2^{-10}	0.93826 ± 0.00031	0.93667 ± 0.00031	0.93801 ± 0.00037	0.93674 ± 0.00037
2^{-9}	0.93845 ± 0.00022	0.93677 ± 0.00022	0.93836 ± 0.00032	0.93704 ± 0.00032
2^{-8}	0.93813 ± 0.00028	0.9365 ± 0.00028	0.93797 ± 0.0004	0.93681 ± 0.0004
2^{-7}	0.93824 ± 0.00018	0.93664 ± 0.00018	0.93821 ± 0.00043	0.93725 ± 0.00043
2^{-6}	0.93833 ± 0.00015	0.9366 ± 0.00015	0.93822 ± 0.00029	0.93684 ± 0.00029
2^{-5}	0.93837 ± 0.00017	0.93681 ± 0.00017	0.93798 ± 0.00043	0.93674 ± 0.00043
2^{-4}	0.93831 ± 0.00031	0.93671 ± 0.00031	0.93821 ± 0.0004	0.93681 ± 0.0004
2^{-3}	0.93829 ± 0.00037	0.93684 ± 0.00037	0.93818 ± 0.00048	0.93691 ± 0.00048
2^{-2}	0.93829 ± 0.0002	0.9366 ± 0.0002	0.93819 ± 0.00035	0.93674 ± 0.00035
2^{-1}	0.93834 ± 0.00021	0.93698 ± 0.00021	0.93804 ± 0.00036	0.93687 ± 0.00036
1	0.93813 ± 0.00014	0.93674 ± 0.00014	0.93813 ± 0.00022	0.93677 ± 0.00022
2	0.93775 ± 0.00018	0.9364 ± 0.00018	0.93767 ± 0.0002	0.93623 ± 0.0002

Table 5: Average and standard deviation for the accuracy at the epoch with highest accuracy. The model in display is a diffusive PSBC with Neumann boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^P = 0.1$ (initial), patience = 50, $N_t \in \{1, 2, 4, 8\}$, and weights-N_t-sharing. Learning rates were chosen according to Appendix A in the paper, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. For a visualization of the data in this table, see Figure 15 in the paper.

Average of maximum throughout epochs (weights-1-sharing)				
ε	$N_t = 1$		$N_t = 2$	
	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$
0	2.2471 ± 0.0503	46.4559 ± 0.0705	1.4863 ± 0.0293	23.3581 ± 0.0574
2^{-10}	2.2336 ± 0.0394	46.3739 ± 0.1584	1.4682 ± 0.0198	23.4505 ± 0.0718
2^{-9}	2.2477 ± 0.059	46.4443 ± 0.1478	1.5303 ± 0.0467	23.3938 ± 0.0633
2^{-8}	2.2863 ± 0.0665	46.4056 ± 0.1125	1.492 ± 0.0387	23.361 ± 0.0881
2^{-7}	2.252 ± 0.0619	46.4102 ± 0.152	1.5083 ± 0.039	23.3573 ± 0.0887
2^{-6}	2.2687 ± 0.0682	46.4365 ± 0.1017	1.4867 ± 0.0439	23.3925 ± 0.0739
2^{-5}	2.2826 ± 0.0415	46.4389 ± 0.066	1.4811 ± 0.035	23.4263 ± 0.0802
2^{-4}	2.2309 ± 0.0777	46.3912 ± 0.0942	1.4972 ± 0.0503	23.3988 ± 0.0911
2^{-3}	2.2184 ± 0.0549	46.4794 ± 0.1239	1.5144 ± 0.0364	23.4082 ± 0.1008
2^{-2}	2.229 ± 0.0249	46.4126 ± 0.1217	1.5012 ± 0.0479	23.417 ± 0.0904
2^{-1}	2.2363 ± 0.0687	46.4252 ± 0.151	1.5014 ± 0.0599	23.4254 ± 0.0834
2^{0}	2.2371 ± 0.0222	46.3686 ± 0.0786	1.4962 ± 0.0399	23.469 ± 0.0702
2	2.2543 ± 0.0378	46.5026 ± 0.1338	1.461 ± 0.0301	23.5093 ± 0.0847

ε	$N_t = 4$		$N_t = 8$	
	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$
0	1.147 ± 0.0288	11.3702 ± 0.0643	1.0261 ± 0.0329	4.9041 ± 0.074
2^{-10}	1.1607 ± 0.024	11.3633 ± 0.0476	1.049 ± 0.0467	4.8718 ± 0.0766
2^{-9}	1.1714 ± 0.0367	11.3916 ± 0.1025	1.0321 ± 0.0404	4.8909 ± 0.07
2^{-8}	1.1447 ± 0.0369	11.3384 ± 0.0517	1.0284 ± 0.0269	4.8886 ± 0.0784
2^{-7}	1.1671 ± 0.0511	11.3607 ± 0.084	1.0287 ± 0.0263	4.8558 ± 0.0572
2^{-6}	1.1593 ± 0.0309	11.3685 ± 0.0673	1.0421 ± 0.0412	4.9097 ± 0.0949
2^{-5}	1.1507 ± 0.0321	11.3515 ± 0.0721	1.0215 ± 0.0219	4.9224 ± 0.0415
2^{-4}	1.1665 ± 0.0312	11.3383 ± 0.036	1.0398 ± 0.021	4.9444 ± 0.0852
2^{-3}	1.1644 ± 0.0335	11.3706 ± 0.0995	1.0265 ± 0.0242	4.9449 ± 0.0577
2^{-2}	1.1605 ± 0.0307	11.4076 ± 0.0683	1.0254 ± 0.0326	4.9526 ± 0.0573
2^{-1}	1.1602 ± 0.0282	11.4303 ± 0.0486	1.0177 ± 0.0262	5.1157 ± 0.0759
2^{0}	1.1424 ± 0.0249	11.5324 ± 0.0704	1.0451 ± 0.0287	5.3839 ± 0.0417
2	1.1565 ± 0.0243	11.6882 ± 0.0663	1.0291 ± 0.0295	5.6896 ± 0.0658

Table 6: Average and standard deviation for the maximum value attained by the diameter of the set $\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor} := \operatorname{conv}\left(\{0,1\} \cup_{m=0}^{N_t-1} \{\alpha^{\lfloor m\rfloor}\}\right)$ over epochs; $\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}$ is defined similarly. This quantity immediately gives an estimate on the size of trainable weights (in ℓ^{∞}-norm) thanks to the relation $\max\limits_{0\leq n\leq N_t-1} \left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^{\infty}}, 1\right\} \leq \operatorname{diam}(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}) \leq 2 \max\limits_{0\leq n\leq N_t-1} \left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^{\infty}}, 1\right\}$. The model in display is a diffusive PSBC with Neumann boundary conditions and parameters $\Delta_t^{u} = 0.1$ (initial), $\Delta_t^{p} = 0.1$ (initial), patience = 50, $N_t \in \{1,2,4,8\}$, and weights-1-sharing. Learning rates were chosen according to Appendix A, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. Note that, even though each simulation has assigned to it a maximum value $Q = 600$ of epochs, each one of them may stop earlier due to *Early stopping*; see further information in Appendix A in the paper.

Average of maximum throughout epochs (weights-N_t-sharing)				
ε	$N_t = 1$		$N_t = 2$	
	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
0	2.2471 ± 0.0503	46.4559 ± 0.0705	1.3728 ± 0.0378	22.0897 ± 0.1014
2^{-10}	2.2336 ± 0.0394	46.3739 ± 0.1584	1.3903 ± 0.0443	22.1332 ± 0.127
2^{-9}	2.2477 ± 0.059	46.4443 ± 0.1478	1.3891 ± 0.0448	22.085 ± 0.1051
2^{-8}	2.2863 ± 0.0665	46.4056 ± 0.1125	1.3834 ± 0.0368	22.0953 ± 0.0978
2^{-7}	2.252 ± 0.0619	46.4102 ± 0.152	1.3741 ± 0.0248	22.12 ± 0.1123
2^{-6}	2.2687 ± 0.0682	46.4365 ± 0.1017	1.37 ± 0.0394	22.1021 ± 0.1112
2^{-5}	2.2826 ± 0.0415	46.4389 ± 0.066	1.4061 ± 0.0392	22.0321 ± 0.1137
2^{-4}	2.2309 ± 0.0777	46.3912 ± 0.0942	1.415 ± 0.0691	22.0749 ± 0.1316
2^{-3}	2.2184 ± 0.0549	46.4794 ± 0.1239	1.3991 ± 0.05	22.1675 ± 0.147
2^{-2}	2.229 ± 0.0249	46.4126 ± 0.1217	1.3979 ± 0.0452	22.15 ± 0.1221
2^{-1}	2.2363 ± 0.0687	46.4252 ± 0.151	1.4135 ± 0.0792	22.1884 ± 0.1112
2^{0}	2.2371 ± 0.0222	46.3686 ± 0.0786	1.3845 ± 0.0405	22.1105 ± 0.1194
2	2.2543 ± 0.0378	46.5026 ± 0.1338	1.3795 ± 0.0329	22.2386 ± 0.1038
ε	$N_t = 4$		$N_t = 8$	
	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max\limits_{0\le q\le Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
0	1.076 ± 0.0302	10.9976 ± 0.1044	1.0 ± 0.0	4.5348 ± 0.0301
2^{-10}	1.0549 ± 0.044	11.0145 ± 0.0814	1.0029 ± 0.0077	4.5141 ± 0.0393
2^{-9}	1.0851 ± 0.0503	11.039 ± 0.1315	1.0 ± 0.0	4.5169 ± 0.0295
2^{-8}	1.0742 ± 0.0245	11.0067 ± 0.0886	1.0 ± 0.0	4.512 ± 0.0489
2^{-7}	1.0919 ± 0.0374	10.9999 ± 0.1257	1.0 ± 0.0	4.5006 ± 0.0273
2^{-6}	1.0622 ± 0.0228	10.998 ± 0.1314	1.0004 ± 0.0013	4.5194 ± 0.0404
2^{-5}	1.0662 ± 0.0329	11.0268 ± 0.1023	1.0 ± 0.0	4.5382 ± 0.0473
2^{-4}	1.0857 ± 0.0424	11.0645 ± 0.0598	1.0012 ± 0.0035	4.5243 ± 0.0334
2^{-3}	1.0725 ± 0.0339	11.0565 ± 0.131	1.002 ± 0.0059	4.5694 ± 0.0395
2^{-2}	1.0737 ± 0.0319	11.0437 ± 0.1098	1.0 ± 0.0	4.5834 ± 0.0302
2^{-1}	1.0563 ± 0.0282	11.0739 ± 0.1199	1.0 ± 0.0	4.7742 ± 0.0424
2^{0}	1.0884 ± 0.0405	11.1851 ± 0.1042	1.0 ± 0.0	5.0851 ± 0.0208
2	1.0649 ± 0.0349	11.3956 ± 0.0562	1.0 ± 0.0	5.4261 ± 0.019

Table 7: Average and standard deviation for the maximum value attained by the diameter of the set $\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor} := \operatorname{conv}\left(\{0,1\} \cup_{m=0}^{N_t-1} \{\alpha^{\lfloor m\rfloor}\}\right)$ over epochs; $\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}$ is defined similarly. This quantity immediately gives an estimate on the size of trainable weights (in ℓ^∞-norm) thanks to the relation $\max\limits_{0\le n\le N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\} \le \operatorname{diam}(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}) \le 2 \max\limits_{0\le n\le N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\}$. The model in display is a diffusive PSBC with Neumann boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), patience = 50, $N_t \in \{1, 2, 4, 8\}$, and weights-N_t-sharing. Learning rates were chosen according to Appendix A, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. Note that, even though each simulation has assigned to it a maximum value $Q = 600$ of epochs, each one of them may stop earlier due to *Early stopping*; see further information in Appendix A in the paper.

4 Diffusive PSBC, Periodic boundary conditions (Section 6.3)

We remark that values in the following tables agree when $N_t = 1$ because weights-1-sharing and weights-N_t-models coincide in that case.

Average accuracy (weights-1-sharing, Periodic)				
ε	$N_t = 1$		$N_t = 2$	
	Train	Test	Train	Test
0	0.93825 ± 0.0001	0.93674 ± 0.0001	0.9382 ± 0.00014	0.93671 ± 0.00014
2^{-10}	0.93817 ± 8e-05	0.9366 ± 8e-05	0.93824 ± 0.00013	0.93671 ± 0.00013
2^{-9}	0.93813 ± 9e-05	0.93671 ± 9e-05	0.93818 ± 0.00014	0.93667 ± 0.00014
2^{-8}	0.93819 ± 0.0001	0.9366 ± 0.0001	0.9382 ± 7e-05	0.93681 ± 7e-05
2^{-7}	0.93818 ± 0.00011	0.93667 ± 0.00011	0.93821 ± 9e-05	0.93667 ± 9e-05
2^{-6}	0.93816 ± 0.00012	0.93671 ± 0.00012	0.93827 ± 0.00014	0.93677 ± 0.00014
2^{-5}	0.93813 ± 0.0001	0.93677 ± 0.0001	0.93816 ± 0.00013	0.93667 ± 0.00013
2^{-4}	0.93823 ± 0.00014	0.93667 ± 0.00014	0.93825 ± 9e-05	0.93677 ± 9e-05
2^{-3}	0.93828 ± 0.00013	0.93667 ± 0.00013	0.93824 ± 0.00015	0.93671 ± 0.00015
2^{-2}	0.93818 ± 0.00011	0.93671 ± 0.00011	0.93819 ± 9e-05	0.9366 ± 9e-05
2^{-1}	0.9382 ± 7e-05	0.93667 ± 7e-05	0.9382 ± 8e-05	0.93677 ± 8e-05
1	0.93818 ± 0.00011	0.93671 ± 0.00011	0.93809 ± 5e-05	0.9366 ± 5e-05
2	0.93819 ± 0.00013	0.93667 ± 0.00013	0.93796 ± 0.00013	0.9364 ± 0.00013
ε	$N_t = 4$		$N_t = 8$	
	Train	Test	Train	Test
0	0.93835 ± 0.00021	0.93674 ± 0.00021	0.93862 ± 0.00014	0.93671 ± 0.00014
2^{-10}	0.9383 ± 0.00015	0.93671 ± 0.00015	0.93843 ± 0.00018	0.93681 ± 0.00018
2^{-9}	0.93829 ± 0.00012	0.93681 ± 0.00012	0.93868 ± 0.00012	0.93684 ± 0.00012
2^{-8}	0.93843 ± 0.00019	0.93698 ± 0.00019	0.93862 ± 0.00017	0.93667 ± 0.00017
2^{-7}	0.93822 ± 0.00013	0.93677 ± 0.00013	0.93862 ± 8e-05	0.93671 ± 8e-05
2^{-6}	0.93824 ± 0.00016	0.93674 ± 0.00016	0.93849 ± 0.00021	0.93674 ± 0.00021
2^{-5}	0.93833 ± 0.00019	0.93677 ± 0.00019	0.93866 ± 0.0002	0.93671 ± 0.0002
2^{-4}	0.93829 ± 0.00013	0.93677 ± 0.00013	0.9386 ± 0.00019	0.93674 ± 0.00019
2^{-3}	0.93835 ± 0.00019	0.93671 ± 0.00019	0.93861 ± 0.00013	0.93657 ± 0.00013
2^{-2}	0.93835 ± 0.00014	0.93677 ± 0.00014	0.93868 ± 0.00015	0.93681 ± 0.00015
2^{-1}	0.93824 ± 0.00013	0.93664 ± 0.00013	0.93844 ± 0.00021	0.93671 ± 0.00021
1	0.93815 ± 0.00014	0.9366 ± 0.00014	0.93842 ± 0.00014	0.93684 ± 0.00014
2	0.938 ± 8e-05	0.93647 ± 8e-05	0.93803 ± 8e-05	0.93657 ± 8e-05

Table 8: Average and standard deviation for the accuracy at the epoch with highest accuracy. The model in display is a diffusive PSBC with Periodic boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), patience = 50, $N_t \in \{1, 2, 4, 8\}$, and weights-1-sharing. Learning rates were chosen according to Appendix A in the paper, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. For a visualization of the data in this table, see Figure 16 in the paper.

Average accuracy (weights-N_t-sharing, Periodic)				
ε	$N_t = 1$		$N_t = 2$	
	Train	Test	Train	Test
0	0.9383 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0001	0.9367 ± 0.0001
2^{-10}	0.9382 ± 0.0001	0.9366 ± 0.0001	0.9383 ± 0.0001	0.9367 ± 0.0001
2^{-9}	0.9381 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0002	0.9367 ± 0.0002
2^{-8}	0.9382 ± 0.0001	0.9366 ± 0.0001	0.9383 ± 0.0001	0.9368 ± 0.0001
2^{-7}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0001	0.9367 ± 0.0001
2^{-6}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9382 ± 0.0002	0.9367 ± 0.0002
2^{-5}	0.9381 ± 0.0001	0.9368 ± 0.0001	0.9383 ± 0.0002	0.9367 ± 0.0002
2^{-4}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0002	0.9367 ± 0.0002
2^{-3}	0.9383 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0002	0.9368 ± 0.0002
2^{-2}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0002	0.9368 ± 0.0002
2^{-1}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0001	0.9368 ± 0.0001
2^{0}	0.9382 ± 0.0001	0.9367 ± 0.0001	0.9381 ± 0.0001	0.9365 ± 0.0001
2	0.9382 ± 0.0001	0.9367 ± 0.0001	0.938 ± 0.0001	0.9364 ± 0.0001
ε	$N_t = 4$		$N_t = 8$	
	Train	Test	Train	Test
0	0.9383 ± 0.0002	0.9368 ± 0.0002	0.9384 ± 0.0003	0.9369 ± 0.0003
2^{-10}	0.9382 ± 0.0003	0.9366 ± 0.0003	0.9382 ± 0.0003	0.937 ± 0.0003
2^{-9}	0.9384 ± 0.0003	0.9368 ± 0.0003	0.9379 ± 0.0005	0.9367 ± 0.0005
2^{-8}	0.9384 ± 0.0002	0.9368 ± 0.0002	0.938 ± 0.0004	0.9367 ± 0.0004
2^{-7}	0.9384 ± 0.0001	0.9367 ± 0.0001	0.9383 ± 0.0005	0.937 ± 0.0005
2^{-6}	0.9384 ± 0.0002	0.937 ± 0.0002	0.9383 ± 0.0004	0.9375 ± 0.0004
2^{-5}	0.9383 ± 0.0001	0.9368 ± 0.0001	0.938 ± 0.0005	0.9368 ± 0.0005
2^{-4}	0.9384 ± 0.0002	0.9367 ± 0.0002	0.9383 ± 0.0003	0.9367 ± 0.0003
2^{-3}	0.9383 ± 0.0002	0.9367 ± 0.0002	0.9382 ± 0.0003	0.9369 ± 0.0003
2^{-2}	0.9384 ± 0.0003	0.9366 ± 0.0003	0.9381 ± 0.0003	0.9369 ± 0.0003
2^{-1}	0.9382 ± 0.0002	0.9366 ± 0.0002	0.938 ± 0.0004	0.9369 ± 0.0004
2^{0}	0.9381 ± 0.0002	0.9366 ± 0.0002	0.9381 ± 0.0002	0.9369 ± 0.0002
2	0.9378 ± 0.0002	0.9364 ± 0.0002	0.9376 ± 0.0001	0.9364 ± 0.0001

Table 9: Average and standard deviation for the accuracy at the epoch with highest accuracy. The model in display is a diffusive PSBC with Periodic boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^P = 0.1$ (initial), patience = 50, $N_t \in \{1, 2, 4, 8\}$, and weights-N_t-sharing. Learning rates were chosen according to Appendix A in the paper, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. For a visualization of the data in this table, see Figure 16 in the paper.

Average of maximum throughout epochs (weights-1-sharing, Periodic)				
ε	$N_t = 1$		$N_t = 2$	
	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
0	2.2655 ± 0.0543	46.4456 ± 0.1654	1.4847 ± 0.0295	23.434 ± 0.0606
2^{-10}	2.2572 ± 0.0558	46.3963 ± 0.0979	1.4719 ± 0.0263	23.3742 ± 0.0577
2^{-9}	2.2475 ± 0.059	46.3758 ± 0.0852	1.495 ± 0.0555	23.4082 ± 0.077
2^{-8}	2.2623 ± 0.0354	46.351 ± 0.1384	1.5078 ± 0.0468	23.3969 ± 0.079
2^{-7}	2.2476 ± 0.0398	46.4304 ± 0.1513	1.5301 ± 0.0539	23.4165 ± 0.0698
2^{-6}	2.2443 ± 0.041	46.4176 ± 0.1396	1.5332 ± 0.0648	23.3601 ± 0.1011
2^{-5}	2.2449 ± 0.0336	46.4169 ± 0.1183	1.491 ± 0.0401	23.3763 ± 0.086
2^{-4}	2.2702 ± 0.0331	46.4255 ± 0.1553	1.4781 ± 0.0333	23.4263 ± 0.0861
2^{-3}	2.2205 ± 0.0658	46.4107 ± 0.0821	1.5138 ± 0.0386	23.3763 ± 0.0435
2^{-2}	2.2404 ± 0.0515	46.4571 ± 0.1305	1.4758 ± 0.0232	23.3989 ± 0.0711
2^{-1}	2.2493 ± 0.0569	46.4478 ± 0.1489	1.498 ± 0.0502	23.4418 ± 0.0768
2^{0}	2.2375 ± 0.0519	46.4828 ± 0.1084	1.4818 ± 0.04	23.4633 ± 0.1074
2	2.241 ± 0.0469	46.4303 ± 0.0941	1.488 ± 0.048	23.5118 ± 0.1166
ε	$N_t = 4$		$N_t = 8$	
	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}\right)_q$
0	1.151 ± 0.042	11.3637 ± 0.0607	1.0476 ± 0.0409	4.8581 ± 0.0349
2^{-10}	1.1722 ± 0.0513	11.3956 ± 0.0966	1.0281 ± 0.023	4.8932 ± 0.055
2^{-9}	1.1726 ± 0.0358	11.351 ± 0.0598	1.0502 ± 0.0505	4.9378 ± 0.1149
2^{-8}	1.1634 ± 0.0355	11.3766 ± 0.0498	1.0342 ± 0.0333	4.8624 ± 0.0964
2^{-7}	1.1446 ± 0.033	11.3496 ± 0.0637	1.0211 ± 0.0236	4.8737 ± 0.0559
2^{-6}	1.1553 ± 0.0221	11.3604 ± 0.0384	1.049 ± 0.0387	4.8903 ± 0.0932
2^{-5}	1.1529 ± 0.0321	11.3051 ± 0.0476	1.0376 ± 0.0325	4.8825 ± 0.0943
2^{-4}	1.1434 ± 0.034	11.3777 ± 0.0709	1.0497 ± 0.0291	4.883 ± 0.0642
2^{-3}	1.1647 ± 0.0345	11.3202 ± 0.0565	1.0289 ± 0.029	4.9716 ± 0.0856
2^{-2}	1.1856 ± 0.0539	11.3366 ± 0.0613	1.0304 ± 0.0316	4.9895 ± 0.0577
2^{-1}	1.1488 ± 0.0364	11.4419 ± 0.0459	1.0314 ± 0.0258	5.1011 ± 0.0436
2^{0}	1.164 ± 0.0303	11.5262 ± 0.0477	1.0277 ± 0.0241	5.3921 ± 0.0544
2	1.1572 ± 0.0405	11.6955 ± 0.0606	1.0285 ± 0.0242	5.6942 ± 0.0389

Table 10: Average and standard deviation for the maximum value attained by the diameter of the set $\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor} := \operatorname{conv}\left(\{0,1\} \cup_{m=0}^{N_t-1} \{\alpha^{\lfloor m\rfloor}\}\right)$ over epochs; $\mathscr{P}_\beta^{\lfloor N_t-1\rfloor}$ is defined similarly. This quantity immediately gives an estimate on the size of trainable weights (in ℓ^∞-norm) thanks to the relation $\max_{0\leq n\leq N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\} \leq \operatorname{diam}(\mathscr{P}_\alpha^{\lfloor N_t-1\rfloor}) \leq 2 \max_{0\leq n\leq N_t-1}\left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\}$. The model in display is a diffusive PSBC with Periodic boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), patience = 50, $N_t \in \{1,2,4,8\}$, and weights-1-sharing. Learning rates were chosen according to Appendix A, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. Note that, even though each simulation has assigned to it a maximum value $Q = 600$ of epochs, each one of them may stop earlier due to *Early stopping*; see further information in Appendix A in the paper.

Average of maximum throughout epochs (weights-N_t-sharing, Periodic)				
ε	$N_t = 1$		$N_t = 2$	
	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$
0	2.2655 ± 0.0543	46.4456 ± 0.1654	1.3775 ± 0.0429	22.1334 ± 0.1715
2^{-10}	2.2572 ± 0.0558	46.3963 ± 0.0979	1.3948 ± 0.0678	22.1239 ± 0.1769
2^{-9}	2.2475 ± 0.059	46.3758 ± 0.0852	1.3781 ± 0.0379	22.2032 ± 0.0829
2^{-8}	2.2623 ± 0.0354	46.351 ± 0.1384	1.4042 ± 0.0399	22.1296 ± 0.0824
2^{-7}	2.2476 ± 0.0398	46.4304 ± 0.1513	1.4067 ± 0.0383	22.1773 ± 0.19
2^{-6}	2.2443 ± 0.041	46.4176 ± 0.1396	1.3941 ± 0.0498	22.1501 ± 0.1372
2^{-5}	2.2449 ± 0.0336	46.4169 ± 0.1183	1.3984 ± 0.0387	22.0976 ± 0.1399
2^{-4}	2.2702 ± 0.0331	46.4255 ± 0.1553	1.4098 ± 0.0454	22.0997 ± 0.0977
2^{-3}	2.2205 ± 0.0658	46.4107 ± 0.0821	1.3639 ± 0.0273	22.1031 ± 0.1514
2^{-2}	2.2404 ± 0.0515	46.4571 ± 0.1305	1.3939 ± 0.0224	22.0829 ± 0.1908
2^{-1}	2.2493 ± 0.0569	46.4478 ± 0.1489	1.4122 ± 0.0354	22.0522 ± 0.1427
2^{0}	2.2375 ± 0.0519	46.4828 ± 0.1084	1.3639 ± 0.0352	22.1322 ± 0.1853
2	2.241 ± 0.0469	46.4303 ± 0.0941	1.4045 ± 0.0617	22.2923 ± 0.0991
ε	$N_t = 4$		$N_t = 8$	
	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}\right)_q$	$\max_{0\leq q\leq Q} \operatorname{diam}\left(\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}\right)_q$
0	1.0726 ± 0.0454	11.0186 ± 0.1212	1.0 ± 0.0	4.5323 ± 0.0401
2^{-10}	1.0796 ± 0.0476	11.0678 ± 0.1133	1.0 ± 0.0	4.531 ± 0.0408
2^{-9}	1.0746 ± 0.0434	11.0337 ± 0.0629	1.0 ± 0.0	4.5165 ± 0.0411
2^{-8}	1.0683 ± 0.0152	11.0202 ± 0.0873	1.0013 ± 0.0038	4.5277 ± 0.0207
2^{-7}	1.0583 ± 0.0314	10.9541 ± 0.1117	1.001 ± 0.0031	4.5236 ± 0.0433
2^{-6}	1.0803 ± 0.0261	10.9613 ± 0.1071	1.0 ± 0.0	4.5282 ± 0.0605
2^{-5}	1.0693 ± 0.0414	11.0465 ± 0.1129	1.0008 ± 0.0023	4.5216 ± 0.0391
2^{-4}	1.0717 ± 0.0351	11.0283 ± 0.1021	1.0 ± 0.0	4.5108 ± 0.0393
2^{-3}	1.0726 ± 0.0391	10.9933 ± 0.065	1.0 ± 0.0	4.5458 ± 0.0604
2^{-2}	1.0764 ± 0.0352	11.1219 ± 0.1005	1.0 ± 0.0	4.5904 ± 0.0373
2^{-1}	1.0912 ± 0.0558	11.1092 ± 0.1004	1.0 ± 0.0	4.7634 ± 0.0346
2^{0}	1.0488 ± 0.0189	11.2141 ± 0.1203	1.0 ± 0.0	5.1008 ± 0.0234
2	1.0735 ± 0.0303	11.3803 ± 0.1225	1.0 ± 0.0	5.4214 ± 0.0277

Table 11: Average and standard deviation for the maximum value attained by the diameter of the set $\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor} := \operatorname{conv}\left(\{0,1\} \cup_{m=0}^{N_t-1} \{\alpha^{\lfloor m\rfloor}\}\right)$ over epochs; $\mathscr{P}_{\beta}^{\lfloor N_t-1\rfloor}$ is defined similarly. This quantity immediately gives an estimate on the size of trainable weights (in ℓ^∞-norm) thanks to the relation $\max_{0\leq n\leq N_t-1} \left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\} \leq \operatorname{diam}(\mathscr{P}_{\alpha}^{\lfloor N_t-1\rfloor}) \leq 2 \max_{0\leq n\leq N_t-1} \left\{\|\alpha^{\lfloor n\rfloor}\|_{\ell^\infty}, 1\right\}$. The model in display is a diffusive PSBC with Periodic boundary conditions and parameters $\Delta_t^u = 0.1$ (initial), $\Delta_t^p = 0.1$ (initial), patience = 50, $N_t \in \{1,2,4,8\}$, and weights-N_t-sharing. Learning rates were chosen according to Appendix A, at $\varepsilon = 0$. Statistics were taken from a sample space of 10 simulations for each set of hyperparameters. Note that, even though each simulation has assigned to it a maximum value $Q = 600$ of epochs, each one of them may stop earlier due to *Early stopping*; see further information in Appendix A in the paper.

5 Auxiliary results for code vectorization

There are many challenges in the design and implementation of ML models. Our main goal in this section involves breaking the model into smaller subroutines, exploiting structure of computations that can be cast into a single framework.

Some additional tools to implement the PSBC as the algorithm 1 are given below. Our exposition relies on basic Calculus and Linear Algebra. We avoid technical tensorial algebra, that would not help in explaining how matrices should in fact be manipulated or computed. For that, we make use of the following definition.

Definition 5.1 (Flattening operator) *We define the Flattening operator* $\mathscr{F}_{flat}^{(a,b)} : \mathbb{R}^{a\times b} \to \mathbb{R}^{a\cdot b\times 1}$ *as an action on the space of matrices such that*

$$\mathscr{F}_{flat}^{(a,b)}\left(\begin{bmatrix} r_1 \\ \vdots \\ r_a \end{bmatrix}\right) = \left[\, r_1 \,|\, \ldots \,|\, r_a \,|\right].$$

One observes a clear redundancy in notation, because the space of matrices on which the flattening operator $\mathscr{F}_{\text{flat}}^{(a,b)}$ acts has well defined parameters a and b. Thus, for the sake of notation, in the remaining of this section we shall simply write $\mathscr{F}_{\text{flat}}(\cdot)$ to represent any flattening operator and $\mathscr{F}_{\text{flat}}^{-1}(\cdot) = (\mathscr{F}_{\text{flat}}^{(a,b)})^{-1}(\cdot)$ for its corresponding inverse.

The next result is crucial for an efficient vectorized implementation of the PSBC.

Lemma 5.2 *Let* $Z = \left(Z_{(1)}, \ldots, Z_{(\mathrm{N_d})}\right) \in \mathbb{R}^{\mathrm{N_u}} \oplus \ldots \oplus \mathbb{R}^{\mathrm{N_u}} \approx \mathbb{R}^{\mathrm{N_d}\cdot\mathrm{N_u}}$*, and a given function*

$$\mathscr{H}(Z) = \sum_{i=1}^{\mathrm{N_d}} \frac{1}{2\mathrm{N_d}} \|Z_{(i)}\|^2_{\ell^2(\mathbb{R}^{\mathrm{N_u}})},$$

where $Z_{(i)} = Z_{(i)}(U_{(i)}, P) \in \mathscr{C}^1(\mathbb{R}^{\mathrm{N_u}} \times \mathbb{R}^{\mathrm{N_{ptt}}}; \mathbb{R}^{\mathrm{N_u}})$*. For each* $1 \leq i \leq \mathrm{N_d}$*, assume that,*

$$A_{(i)} := \frac{\partial Z_{(i)}}{\partial U_{(i)}} = \mathscr{L} \cdot diag(v_{(i)}) \in \mathbb{R}^{\mathrm{N_u}\times\mathrm{N_u}}, \qquad B_{(i)} := \frac{\partial Z_{(i)}}{\partial P} = \mathscr{L} \cdot diag(v_{(i)}) \cdot \mathscr{K} \in \mathbb{R}^{\mathrm{N_u}\times\mathrm{N_{ptt}}}, \tag{1}$$

with $\mathscr{L} \in \mathbb{R}^{\mathrm{N_u}\times\mathrm{N_u}}$*,* $\mathscr{K} \in \mathbb{R}^{\mathrm{N_u}\times\mathrm{N_{ptt}}}$*, and* $v_{(i)} \in \mathbb{R}^{\mathrm{N_u}\times 1}$*. Clearly, the relations* $\frac{\partial \mathscr{H}}{\partial U_{(i)}} = A_{(i)}$*, and* $\frac{\partial \mathscr{H}}{\partial P} = \sum_{i=1}^{\mathrm{N_d}} B_{(i)}$ *hold. Last, define the matrix*

$$\mathscr{M}(Z) := \frac{1}{\mathrm{N_d}} \begin{bmatrix} Z_{(1)} \\ \vdots \\ Z_{(\mathrm{N_d})} \end{bmatrix} \cdot \mathscr{L} \circledast \begin{bmatrix} v_{(1)}^T \\ \vdots \\ v_{(\mathrm{N_d})}^T \end{bmatrix} \in \mathbb{R}^{\mathrm{N_d}\times\mathrm{N_u}}. \tag{2}$$

Then, we have

(i) $\nabla_Z \mathscr{H} = \mathscr{F}_{flat}\left(\begin{bmatrix} Z_{(1)} \\ \vdots \\ Z_{(\mathrm{N_d})} \end{bmatrix}\right).$

(ii) $\nabla_U \mathscr{H} = \nabla_Z \mathscr{H} \cdot \mathrm{diag}\left(A_{(i)}; 1 \leq i \leq \mathrm{N_d}\right) = \mathscr{F}_{flat}\left(\mathscr{M}(Z)\right).$

(iii) With $\mathbf{1} \in \mathbb{R}^{\mathrm{N_d}\times 1}$*, it holds that*[1] $\nabla_{\widetilde{P}} \mathscr{H} = \mathbf{1}^T \cdot \mathscr{M}(Z) \cdot \mathscr{K}.$

[1] When using Hadamard products and usual matrix products one should be careful, because operations are not associative. Therefore, $\mathscr{M}(Z)$ should be calculated first.

(iv) *(Hierarchical propagation) Consider*

$$U = \left(U_{(1)}, \ldots, U_{(\mathrm{N_d})}\right), V = \left(V_{(1)}, \ldots, V_{(\mathrm{N_d})}\right) \in \mathbb{R}^{\mathrm{N_u}} \oplus \ldots \oplus \mathbb{R}^{\mathrm{N_u}} \approx \mathbb{R}^{\mathrm{N_d} \cdot \mathrm{N_u}},$$

such that $U_{(i)} \in \mathscr{C}\left(\mathbb{R}^{\mathrm{N_u}}; \mathbb{R}^{\mathrm{N_u}}\right)$ *is a function of* $V_{(i)} \in \mathbb{R}^{\mathrm{N_u}}$ *only, i.e.,* $U_{(i)} = U_{(i)}\left(V_{(i)}\right)$. *Assume that*

$$\frac{\partial U_{(i)}}{\partial V_{(i)}} = \mathscr{R} \cdot \operatorname{diag}(d_{(i)}) \in \mathbb{R}^{\mathrm{N_u} \times \mathrm{N_u}}, \quad \text{with} \quad v_n \in \mathbb{R}^{\mathrm{N_u} \times 1}, \quad \mathscr{R} \in \mathbb{R}^{\mathrm{N_u} \times \mathrm{N_u}}. \tag{3}$$

(Recall from notation in Sec. 1.3 that vectors are stored as column matrices). Then, for any $U \mapsto \mathscr{H}(U) \in \mathscr{C}^1\left(\mathbb{R}^{\mathrm{N_d} \cdot \mathrm{N_u}}; \mathbb{R}\right)$, *it holds that*

$$\frac{\partial \mathscr{H}}{\partial V_{(i)}} = \frac{\partial \mathscr{H}}{\partial U_{(i)}} \cdot \mathscr{R} \cdot \operatorname{diag}(d_{(i)}) \in \mathbb{R}^{1 \times \mathrm{N_u}}, \quad d_{(i)} \in \mathbb{R}^{\mathrm{N_u} \times 1}. \tag{4}$$

In particular,

$$\begin{aligned} \nabla_V \mathscr{H} &= \mathscr{F}_{\mathit{flat}}\left(\begin{bmatrix} \frac{\partial \mathscr{H}}{\partial V_{(1)}} \\ \vdots \\ \frac{\partial \mathscr{H}}{\partial V_{(\mathrm{N_d})}} \end{bmatrix}\right) = \mathscr{F}_{\mathit{flat}}\left(\begin{bmatrix} \frac{\partial \mathscr{H}}{\partial U_{(1)}} \\ \vdots \\ \frac{\partial \mathscr{H}}{\partial U_{(\mathrm{N_d})}} \end{bmatrix} \cdot \mathscr{R} \circledast \begin{bmatrix} d_{(1)}^T \\ \vdots \\ d_{(\mathrm{N_d})}^T \end{bmatrix}\right). \\ &= \mathscr{F}_{\mathit{flat}}\left(\mathscr{F}_{\mathit{flat}}^{-1}\left(\nabla_U \mathscr{H}\right) \cdot \mathscr{R} \circledast \begin{bmatrix} d_{(1)}^T \\ \vdots \\ d_{(\mathrm{N_d})}^T \end{bmatrix}\right). \end{aligned} \tag{5}$$

(v) *With* $\mathbf{1} \in \mathbb{R}^{\mathrm{N_d} \times 1}$, *if* $U_{(i)} = U_{(i)}(\alpha) \in \mathscr{C}\left(\mathbb{R}^k; \mathbb{R}^{\mathrm{N_u}}\right)$ *and* $\dfrac{\partial U_{(i)}}{\partial \alpha} = \mathscr{R} \cdot \operatorname{diag}(v_{(i)}) \cdot \mathscr{K} \in \mathbb{R}^{\mathrm{N_u} \times k}$, *then*

$$\frac{\partial \mathscr{H}}{\partial \alpha} = \mathbf{1}^T \cdot \mathscr{M}(U) \cdot \mathscr{K}, \quad \text{where} \quad \mathscr{M}(U) = \begin{bmatrix} \frac{\partial \mathscr{H}}{\partial U_{(1)}} \\ \vdots \\ \frac{\partial \mathscr{H}}{\partial U_{(\mathrm{N_d})}} \end{bmatrix} \cdot \mathscr{R} \circledast \begin{bmatrix} d_{(1)}^T \\ \vdots \\ d_{(\mathrm{N_d})}^T \end{bmatrix}.$$

The previous lemmas are key ingredients for Backpropagation computations in all the models presented.

Corollary 5.3 *Given* $U = (U, P, Y) \in \mathbb{R}^{\mathrm{N_u}} \times \mathbb{R}^{\mathrm{N_p}} \times \{0, 1\}$, *let*

$$Z = Z(U, P, Y) = \mathcal{S}_{\mathrm{N_u}}^{(P)}(U) - Y\mathbf{1}, \quad \mathbf{1} \in \mathbb{R}^{\mathrm{N_u} \times 1},$$

where $(U, P) \mapsto \mathcal{S}_{\mathrm{N_u}}^{(P)}(U)$ *as in (4.10). For any pair* $(X_{(i)}, Y_{(i)})$ *in* $\mathcal{D}$ *with associated forward propagation* $\left(U^{\lfloor \cdot \rfloor}(X_{(i)}, \alpha^{\lfloor \cdot \rfloor}), P^{\lfloor \cdot \rfloor}\left(\frac{1}{2}\mathbf{1}; \beta^{\lfloor \cdot \rfloor}\right)\right)$ *and* $U^{\lfloor 0 \rfloor} = X_{(i)}$, *define*

$$Z_{(i)}(U, P) := Z\left(U^{\lfloor \mathrm{N_t} \rfloor}, \widetilde{P^{\lfloor \mathrm{N_t} \rfloor}}, Y_{(i)}\right), \tag{6}$$

where $\widetilde{P^{\lfloor \mathrm{N_t} \rfloor}} = \mathscr{K} \cdot P^{\lfloor \mathrm{N_t} \rfloor}$, *with* $\mathscr{K} = \mathbf{1} \in \mathbb{R}^{\mathrm{N_u} \times 1}$ *if the model is non-subordinated,* $\mathscr{K} = \mathscr{B}^*$ *if it is subordinated. Clearly,*

$$\mathrm{Cost}_{\mathscr{D}} = \sum_{i=1}^{\mathrm{N_d}} \frac{1}{2\mathrm{N_d}} \left\| Z_{(i)}\left(U^{\lfloor \mathrm{N_t} \rfloor}(X_{(i)}, \alpha^{\lfloor \mathrm{N_t}-1 \rfloor}), P^{\lfloor \mathrm{N_t} \rfloor}\left(\frac{1}{2}\mathbf{1}; \beta^{\lfloor \mathrm{N_t}-1 \rfloor}\right)\right) \right\|_{\ell^2(\mathbb{R}^{\mathrm{N_u}})}^2. \tag{7}$$

Thanks to Lemma B.1(v), it follows that all derivatives of $\mathrm{Cost}_{\mathscr{D}}$ *with respect to* $U^{\lfloor \mathrm{N_t} \rfloor}$ *can be computed taking* $A_{(i)}$ *in (1) of the form*

$$\mathscr{L} = Id_{\mathrm{N_u}}, \quad v_{(i)} = \mathbf{1} - 2\widetilde{P^{\lfloor \mathrm{N_t} \rfloor}}, \quad \text{and} \quad \mathscr{K} = \mathscr{B}^*.$$

Similarly, the derivatives with respect to $P^{\lfloor \mathrm{N_t} \rfloor}$ can be computed taking $B_{(i)}$ in (1) of the form

$$\mathscr{L} = Id_{\mathrm{N_u}}, \quad v_{(i)} = \mathbf{1} - 2U^{\lfloor \mathrm{N_t} \rfloor},$$

where $\mathscr{K} = \mathbf{1}$ (non-subordinated) or $\mathscr{K} = \mathscr{B}^*$ (subordinated). Still using Lemma B.1(v) we compute the derivatives of the cost function with respect to $\alpha^{\lfloor \mathrm{N_t}-1 \rfloor}$ and $\beta^{\lfloor \mathrm{N_t}-1 \rfloor}$. In the first case, recall that $\alpha^{\lfloor \mathrm{N_t}-1 \rfloor} = \mathscr{B}^* w^{\lfloor \mathrm{N_t}-1 \rfloor}$, therefore we aim to compute derivatives with respect to $w^{\lfloor \mathrm{N_t}-1 \rfloor}$, which we obtain by Lemma 5.2(v), with

$$\mathscr{R} = \mathscr{B}^* \quad \textit{and} \quad d_{(i)} = -\Delta_{\mathrm{t}} U^{\lfloor \mathrm{N_t}-1 \rfloor} \circledast (\mathbf{1} - U^{\lfloor \mathrm{N_t}-1 \rfloor}).$$

Similarly, derivatives of the cost function with respect to $\beta^{\lfloor \mathrm{N_t}-1 \rfloor}$ follow from Lemma B.1(v), applying the Chain Rule.

Finally, since forward propagation implies that $U^{\lfloor n \rfloor} = U^{\lfloor n \rfloor}(U^{\lfloor n-1 \rfloor})$ and $P^{\lfloor n \rfloor} = P^{\lfloor n \rfloor}(P^{\lfloor n-1 \rfloor})$ for all $n \in G_{\mathrm{N_t}}$, assuming that derivatives of the cost function with respect to $U^{\lfloor n \rfloor}$, $P^{\lfloor n \rfloor}$ have been computed, then:

(i) Derivatives with respect to $U^{\lfloor n-1 \rfloor}$ can be found using Lemma 5.2(iv) taking

$$\mathscr{R} = (\mathscr{L}_{\mathrm{N_u}})^{-1} \quad \textit{and} \quad d_{(i)} = \mathbf{1} + \Delta_{\mathrm{t}} \left[U^{\lfloor n-1 \rfloor} \circledast (\mathbf{1} - U^{\lfloor n-1 \rfloor}) + (U^{\lfloor n-1 \rfloor} - \alpha^{\lfloor n-1 \rfloor}) \circledast (\mathbf{1} - 2U^{\lfloor n-1 \rfloor}) \right],$$

which holds due to Lemma B.1(v).

(ii) Derivatives with respect to $\alpha^{\lfloor n-1 \rfloor}$ can be found similarly, using Lemma 5.2(iv) with

$$\mathscr{R} = \mathscr{B}^* \quad \textit{and} \quad d_{(i)} = -\Delta_{\mathrm{t}} U^{\lfloor n-1 \rfloor} \circledast (\mathbf{1} - U^{\lfloor n-1 \rfloor}).$$

(iii) Derivatives with respect to $P^{\lfloor n-1 \rfloor}$ and $\beta^{\lfloor n-1 \rfloor}$ are easily computed using the computations in Lemma B.1(v) and invoking the Chain Rule.

References

[1] R. Monteiro. Source code for the paper "Binary classification as a phase separation process". https://github.com/rafael-a-monteiro-math/Binary_classification_phase_separation, September 2020.