Journal of Statistical Mechanics:
Theory and Experiment

PAPER • OPEN ACCESS

An analytical theory of curriculum learning in teacher–student networks*

To cite this article: Luca Saglietti *et al J. Stat. Mech.* (2022) 114014

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: ML 2022

An analytical theory of curriculum learning in teacher–student networks*

Luca Saglietti[1,4,], Stefano Sarao Mannelli[2,4,**] and Andrew Saxe[1,3]**

[1] Institute for Data Science and Analytics, Bocconi University, Italy
[2] Gatsby Computational Neuroscience Unit and Sainsbury Wellcome Centre, University College, London, United Kingdom
[3] FAIR, Meta AI, United States of America
E-mail: luca.saglietti@unibocconi.it, s.saraomannelli@ucl.ac.uk and a.saxe@ucl.ac.uk

Received 4 October 2022
Accepted for publication 13 October 2022
Published 24 November 2022

Online at stacks.iop.org/JSTAT/2022/114014
https://doi.org/10.1088/1742-5468/ac9b3c

Abstract. In animals and humans, curriculum learning—presenting data in a curated order—is critical to rapid learning and effective pedagogy. A long history of experiments has demonstrated the impact of curricula in a variety of animals but, despite its ubiquitous presence, a theoretical understanding of the phenomenon is still lacking. Surprisingly, in contrast to animal learning, curricula strategies are not widely used in machine learning and recent simulation studies reach the conclusion that curricula are moderately effective or even ineffective in most cases. This stark difference in the importance of curriculum raises a fundamental theoretical question: when and why does curriculum learning help? In this work, we analyse a prototypical neural network model of curriculum learning in the high-dimensional limit, employing statistical physics methods. We study a task in which a sparse set of informative features are embedded amidst a large set of noisy features. We analytically derive average

*This article was also accepted in NeurIPS2022 Proceedings.
**Authors to whom any correspondence should be addressed.
[4]Equal contributions.

1742-5468/22/114014+32$33.00

learning trajectories for simple neural networks on this task, which establish a clear speed benefit for curriculum learning in the online setting. However, when training experiences can be stored and replayed (for instance, during sleep), the advantage of curriculum in standard neural networks disappears, in line with observations from the deep learning literature. Inspired by synaptic consolidation techniques developed to combat catastrophic forgetting, we propose curriculum-aware algorithms that consolidate synapses at curriculum change points and investigate whether this can boost the benefits of curricula. We derive generalisation performance as a function of consolidation strength (implemented as an L_2 regularisation/elastic coupling connecting learning phases), and show that curriculum-aware algorithms can yield a large improvement in test performance. Our reduced analytical descriptions help reconcile apparently conflicting empirical results, trace regimes where curriculum learning yields the largest gains, and provide experimentally-accessible predictions for the impact of task parameters on curriculum benefits. More broadly, our results suggest that fully exploiting a curriculum may require explicit adjustments in the loss.

Keywords: machine learning

Contents

1. Introduction

Presenting learning materials in a meaningful order according to a curriculum greatly helps learning in animals and humans [1–4], and is considered an essential aspect of good pedagogy [5]. For example, humans have been shown to learn visual discriminations faster when presented with examples that exaggerate the relevant difference between classes, a phenomenon known as 'fading' [6–8]. Beyond humans, curricula in the form of 'shaping' or 'staircase' procedures are a near-universal feature of task designs in animal studies, without which training often fails entirely. For instance, the International Brain Laboratory task, a standardised perceptual decision-making training paradigm in mice, involves six stages of increasing difficulty before reaching final performance [9].

Building from this intuition, a seminal series of papers proposed a similar curriculum learning approach for machine learning (ML) [10–12]. In striking contrast to the clear benefits of curriculum in biological systems, however, curriculum learning has generally yielded equivocal benefits in artificial systems. Experiments in a variety of domains [13, 14] have found usually modest speed and generalisation improvements from curricula. Recent extensive empirical analyses have found minimal benefits on standard datasets [15]. Indeed, a common intuition in deep learning practice holds that training distributions should ideally be as close as possible to testing distributions, a notion which runs counter to curriculum. Perhaps the only areas where curricula are actively used are in large language models [16] and certain reinforcement learning settings [17].

This gap between the effect of curriculum in biological and artificial learning systems poses a puzzle for theory. When and why is curriculum learning useful? What properties of a task determine the extent of possible benefits? What ordering of learning material is most beneficial? And can new learning algorithms better exploit curricula? Compared to the empirical investigations of curriculum learning, theoretical results on curriculum learning remain sparse. Most notably, [18, 19] show that curriculum can lead to faster learning in a simple setting, but the effects of curriculum on asymptotic generalisation and the dependence on task structure remain unclear. A hint that indeed curriculum learning might lead to statistically different minima comes from a connection between constraint-satisfaction problems and physics results on flow networks [20], but to our knowledge no direct result has been reported in the modern theoretical ML literature.

In this work we study the impact of curriculum using the analytically tractable teacher–student framework and the tools of statistical physics [21–24]. High-dimensional teacher–student models are a popular approach for systematically studying learning behaviour in neural networks [22, 25, 26], and have recently been leveraged to analyse a variety of phenomena [27–32]. Using a simple model to build structured data [12], we examine the impact of ordering examples by increasing difficulty (curriculum), decreasing difficulty (anti-curriculum), or standard shuffled training. We derive exact expressions for the online learning dynamics and the performance of batch learning. However, in the latter, curriculum confers no benefit under standard training. Motivated by theories of synaptic consolidation and elastic weight consolidation [33, 34], we

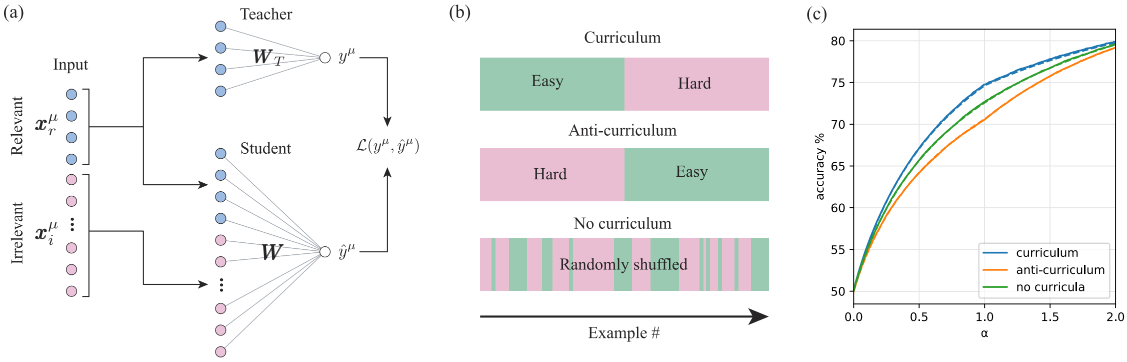


Figure 1. Teacher–student setting for curriculum learning. (a) Illustration of teacher–student setting in which a 'student' network is trained from i.i.d. inputs with labels from a 'teacher' network. Since the teacher network is sparse, its output depends only on a subset of *relevant* input features. (b) We consider curricula which order examples by difficulty, here taken to be the variance in the irrelevant feature dimensions. We refer to increasing, decreasing, and random difficulty order as curriculum, anti-curriculum, and no curriculum, respectively. (c) Example test error on hard examples for the student over training. The switch-point between easy and hard samples lies at $\alpha = 1/2$. Solid lines show numerical simulations, while dashed lines show theoretical predictions derived in section 3. For this particular parameter setting, curriculum speeds learning but only modestly improves final performance at $\alpha = 1$. Parameters: $\alpha_1 = 1$, $\alpha_2 = 1$, $\Delta_1 = 0$, $\Delta_2 = 1$, $\gamma = 10^{-5}$, $\eta = 3$.

introduce elastic penalties (Gaussian priors) that regularise training toward solutions obtained in prior curriculum phases. With these priors, curriculum yields benefits both in the online 3 and in the batch 4 settings.

2. Model definition and overview of approach

In the following, we revisit a prototypical model of curriculum learning from [12] that finds correspondence to the fading literature [6] as highlighted in section 5. Our setting is summarised in figure 1. The model entails a simple teacher–student setup, where teacher and student are each shallow one-layer neural networks of size N (also known as perceptrons). The learning task for the student is a binary classification problem, with dataset $\mathcal{D} = \{(y^\mu, \boldsymbol{x}^\mu)\}_{\mu=1}^M$, where the ground-truth labels are produced by the teacher network $y^\mu = \text{sign}\, \boldsymbol{W}_T \cdot \boldsymbol{x}^\mu$. A key feature of this model is that the teacher network is sparse, with only a fraction $\rho < 1$ of $\sim \mathcal{N}(0, 1)$ non-zero components. Therefore, in order to achieve a good test accuracy, the student has to guess which components should be set to zero and align the relevant weights in the correct direction. A large range of $0 < \rho < 1$ could give rise to the phenomenology we seek to analyse. In the remainder of the paper we will focus on the case $\rho = 0.5$.

We model the variable degree of difficulty in the samples by decomposing each input vector as $\boldsymbol{x}^\mu = [\boldsymbol{x}_r^\mu, \boldsymbol{x}_i^\mu] \in \mathbb{R}^N$, where $\boldsymbol{x}_r^\mu \in \mathbb{R}^{\rho N}$ denotes the relevant components of the

https://doi.org/10.1088/1742-5468/ac9b3c

input, and $\boldsymbol{x}_i^\mu \in \mathbb{R}^{(1-\rho)N}$ the irrelevant ones. Note that, crucially, the sparse teacher network is completely blind to the irrelevant part of the input: $y^\mu = \text{sign}\ \sum_{j=1}^{\rho N} W_{T,j} x_{r,j}^\mu$. While $x_{r,j}^\mu$ i.i.d. $\mathcal{N}(0,1), \forall\, \mu$,[5] we consider the variance for the irrelevant components to be sample-dependent $x_{i,j}^\mu \sim \mathcal{N}(0,\Delta^\mu)$. Note that a smaller variance in the irrelevant part induces a higher SNR in the student learning problem.

The dataset is partitioned according to difficulty levels given by the variances of the irrelevant inputs. For simplicity we consider only two partitions in most of our analysis, but generalisations to more difficulty levels follow straightforwardly. We have a dataset with $M = (\alpha_1 + \alpha_2)N = \alpha N$ total samples, in which the irrelevant inputs of the first $\alpha_1 N$ samples have variance Δ_1, and the remaining $\alpha_2 N$ samples have variance $\Delta_2 > \Delta_1$. In the curriculum learning condition we present the easy examples first, while in the anti-curriculum condition we present the hard examples first. Standard learning presents examples shuffled in random order.

3. Online dynamical solution in the large input limit

We start by focusing on the same online learning setting explored in [12]. We consider a one-layer student network with sigmoidal activation function, $\sigma(\cdot) = \text{erf}(\cdot/\sqrt{2})$, that learns to minimise a mean square error loss with L_2 regularisation of intensity γ, using gradient descent. This yields the updates

$$\boldsymbol{W}^{\mu+1} = \boldsymbol{W}^\mu - \frac{\eta}{\sqrt{N}}\sigma'\left(\frac{\boldsymbol{W}^\mu \cdot \boldsymbol{x}^\mu}{\sqrt{N}}\right)\left(\sigma\left(\frac{\boldsymbol{W}^\mu \cdot \boldsymbol{x}^\mu}{\sqrt{N}}\right) - y^\mu\right)\boldsymbol{x}^\mu - \gamma\,\boldsymbol{W}^\mu. \tag{1}$$

The dynamics of the model can be analysed in the high-dimensional limit $N, M \to \infty$ with $\alpha = M/N = \mathcal{O}(1)$. Generalising the results of [26, 35] on the online stochastic gradient descent dynamics in single-layer regression problems, we obtain a precise description of the performance at all times, as a function of several order parameters: the squared norm of the relevant and irrelevant part of the student weights $Q_r = \frac{1}{N}\boldsymbol{W}^r \cdot \boldsymbol{W}^r$ and $Q_i = \frac{1}{N}\boldsymbol{W}^i \cdot \boldsymbol{W}^i$, respectively; the overlap of the relevant weights of the student and teacher $R = \frac{1}{N}\boldsymbol{W}^r \cdot \boldsymbol{W}_T$; and the squared norm of the teacher vector $T = \frac{1}{N}\boldsymbol{W}_T \cdot \boldsymbol{W}_T$. In particular, given Q_r, Q_i, R and T, the test loss (i.e. average loss on a new example) on a dataset with variance Δ in the irrelevant inputs is given by

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\frac{Q_r + \Delta Q_i}{1 + Q_r + \Delta Q_i} - \frac{2}{\pi}\sin^{-1}\frac{R/\sqrt{T}}{\sqrt{Q_r + \Delta Q_i + 1}},$$

the accuracy by

$$\mathcal{A} = \mathbb{E}\left[\frac{1}{2}(y\ \ \text{sign}\,\hat{y} + 1)\right] = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left(\frac{R}{\sqrt{T(Q_r + \Delta Q_i)}}\right). \tag{2}$$

[5] In [12] the distribution of relevant and irrelevant inputs is uniform between 0 and 1, but this difference does not qualitatively change the results.

If the dataset contains a random mixture of different difficulty levels $\Delta_1, \Delta_2, \ldots$, the loss and accuracy can be obtained by taking a weighted average over the partitions.

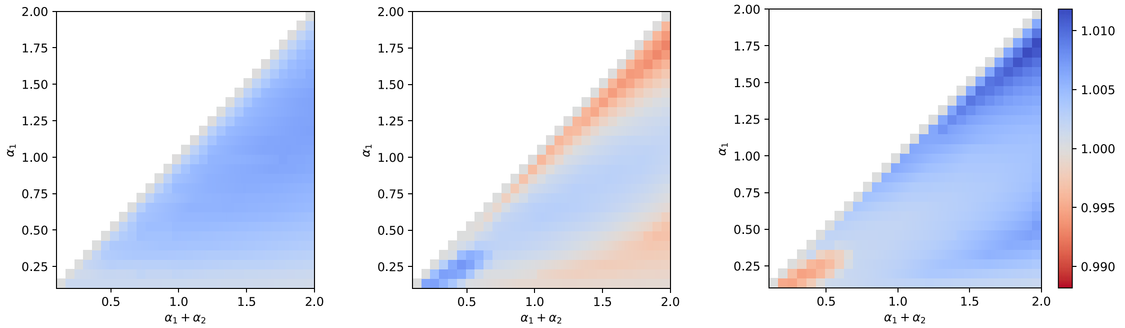
To understand how test performance changes through learning, we study the evolution of the order parameters. Combining their definition with the definition of the dynamics (1) and the fact that the random variables concentrate in the high-dimension as $N \to \infty$, we obtain an analytic form for the updates: $Q_r \leftarrow f_{Q_r}(Q_r, Q_i, R, T), Q_i \leftarrow f_{Q_i}(Q_r, Q_i, R, T), R \leftarrow f_R(Q_r, Q_i, R, T)$; where f_{Q_r}, f_{Q_i} and f_R are long but explicit expressions that are reported in the appendices.

Dynamical advantages of curriculum. With these theoretical results in hand, we now characterise the performance of curricula in the online setting. The dynamical equations have two key advantages relative to simulating models in this setting. First, they are free of finite size effects and stochastic fluctuations. And second, their evaluation is very fast, enabling systematic exploration of the parameter space of the problem, along with fine-grained optimisation over hyper-parameters such as learning rate, weight decay and scaling in the initialisation.

Optimising final test accuracy separately for each curriculum strategy, we find that curriculum learning is the optimal strategy, followed by baseline (no-curriculum) and lastly anti-curriculum. In figure 1(c) we show typical learning trajectories for a dataset with equal numbers of easy and hard samples. The results of the simulations (solid lines) are well-described by our theoretical equations (dashed lines), and show that the curriculum strategy leads to better performance throughout training. Figure 1(c) shows the evolution during training of the test accuracy computed on the whole dataset.

Next we systematically trace the effect of curriculum for a range of total dataset sizes $(\alpha_1 + \alpha_2)$ and number of easy examples α_1 in the phase diagram in figure 2. This diagram shows on the left (centre) the accuracy on hard instances reached at the end of training by curriculum learning (anti-curriculum learning respectively) normalised by the accuracy reached by the standard strategy. The two heatmaps show that curriculum learning always outperforms standard learning and that, on the other hand, anti-curriculum learning outperforms standard learning only in part of the diagram. Comparing the two strategies, figure 2 (right), we can observe that there is a region for small α and α_1 where anti-curriculum learning is the best strategy, while in the majority of the situations curriculum learning is the best strategy. Interestingly, there is a sizeable region of the diagram in which *both* curriculum and anti-curriculum help, possibly explaining why both have been recommended in prior work [12, 14, 36–38]. A possible intuition behind this counter-intuitive phonemenon highlighted by our analysis, is that, in some settings, the large amount of noise contained in the hard data will always be too disruptive for effective learning. Thus, leaving the easy (cleaner) data for last could allow the model to better exploit the easy data.

Further, we find that our setting, in which a small task-relevant signal is embedded in large task-irrelevant variation, is critical to the benefit of curriculum. Figure 4 shows performance as a function of sparsity ρ, additional details are deferred in appendix C. Non-sparse tasks do not benefit. Hence curriculum aids tasks with many irrelevant factors of variation. Interestingly, the literature from human psychology shows precisely



(a) Curriculum learning. (b) Anti-curriculum learning. (c) Curriculum vs anti-curriculum.

Figure 2. Phase diagram of online learning performance gap with optimal parameters. The colour scale shows the ratio of the accuracy on hard instances reached by curriculum over no-curriculum (a), anti-curriculum over no-curriculum (b), and curriculum over anti-curriculum (c), as a function of the total dataset size $(\alpha_1 + \alpha_2)$ and easy dataset size (α_1). Curriculum broadly benefits performance and anti-curriculum is effective in certain regions, but the size of the improvement is modest. Parameters: $\rho = 0.50, \Delta_1 = 0, \Delta_2 = 1$.

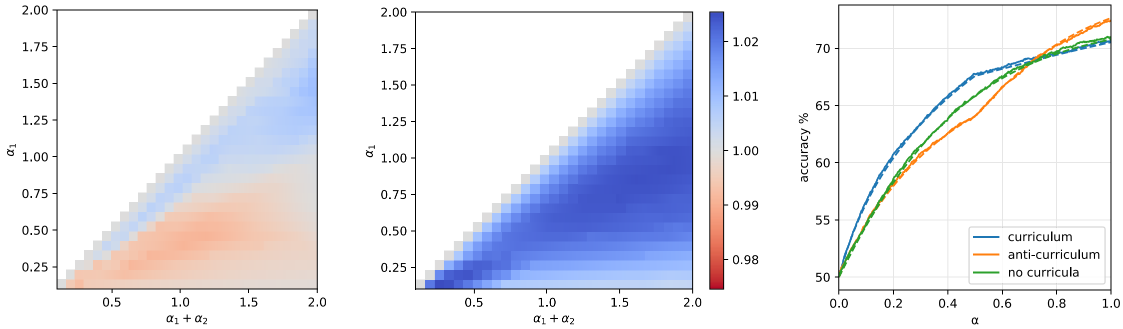
this: no curriculum benefits for low-dimensional tasks or tasks with no variation in irrelevant dimensions [6].

Our results also highlight the intricate dependence of curriculum on parameters of the learning setup. If not all parameters are correctly optimised, we can observe more complex scenarios. For instance, anti-curriculum learning is always the best strategy starting from a large variance in the weights' distribution, as figure 3 shows for weights of order 1. In this case, curriculum learning shows an advantage only in the first phase when easy examples are shown, which is consistent with the results of [19]. However, in the next phase when hard examples are shown, the curriculum strategy does not extract enough information and it is outperformed by the other two strategies. The fact that curriculum or anti-curriculum can look better depending on other learning parameters like initialisation might help explain the confusion in the literature over the best protocol [12, 14, 36–38]. At least in this model, better performance from anti-curriculum is a signature of a sub-optimal choice of the parameters. To summarise our findings in this online learning setting, curriculum mainly offers a *dynamical advantage*: it speeds learning, with only minimal impact on asymptotic performance.

4. Batch learning solution

The previous section discussed the online case where each example is used once and then discarded. However, in common ML practice, neural networks typically revisit each

https://doi.org/10.1088/1742-5468/ac9b3c



(a) Curriculum learning. (b) Anti-curriculum learning. (c) Accuracy hard samples.

Figure 3. Performance gap starting from high initialisation variance. The first two figures show the accuracy-gap on hard instances between curriculum learning and the baseline (a) and anti-curriculum learning and the baseline (b). Contrary to the phase diagram in figure 2, curriculum learning is not always the optimal and anti-curriculum is not always the worst strategy. The right panel shows the accuracy evaluated on the hard samples for $\alpha_1 = \alpha_2 = 0.5$.

sample repeatedly until convergence. Therefore an important question is: *can curricula lead to a generalisation improvement when trained on the same dataset until convergence?*

We investigate this question by considering a student that learns from slices of a dataset in distinct optimisation phases, where in each phase the student optimises a L_2-regularised logistic loss. Without further modification, curriculum can have no effect in this setting: due to the convex nature of the teacher–student setup [22], the network is bound to converge to a minimum uniquely determined by the final slice of data, with no memory of the progress made at intermediate steps. This simple observation may help explain empirical observations on real data, such as [15], which find no benefit of curriculum in standard settings. Despite curriculum could still influence non-convex problems [12], empirical results in the ML field are not showing clear signals of memory retention. A possible explanation is that relying on memory effects in the learning dynamics would require one to hit a sweet spot in the learning rate value and in the number of training epochs, and this seems hard to be achieved consistently. These observations raise the theoretical question of how curriculum learning could induce a non-vanishing effect in batch learning settings.

To instantiate a memory effect in our model, we propose biasing the optimisation landscape via a Gaussian prior, centered around the optimiser of the previous learning phase. The additional term in the loss acts as an elastic coupling between the successive phases, and the associated intensity γ_{12} is then an additional hyper-parameter of the model. This scheme is similar to regularisation methods proposed against catastrophic interference in continual learning, such as synaptic intelligence [39].

Tools from statistical physics can be used to analytically compute test performance under this scheme. In order to simplify the presentation, we first consider just two learning phases. It is natural to frame this setting as a two-level problem, involving two systems with independent copies of the network weights $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. In a typical statistical physics approach, we associate a Boltzmann–Gibbs measure to the systems, with an energy function determined by the regularised logistic loss $\mathcal{L}_\gamma$. While the statistical properties of the first system can be determined self-consistently, the added elastic interaction creates a dependence of the second measure on the configurations of the first system. In mathematical terms, the coupled system is represented by the following partition function:

$$\begin{aligned}\langle Z(\boldsymbol{W}_2, \boldsymbol{W}_1; \mathcal{D}_1, \mathcal{D}_2)\rangle_{\boldsymbol{W}_1} = &\int \mathrm{d}\,\boldsymbol{W}_1 \frac{\mathrm{e}^{-\beta_1 \mathcal{L}_{\gamma_1}(\boldsymbol{W}_1, \mathcal{D}_1)}}{Z_1(\boldsymbol{W}_1)} \log \\ &\times \int \mathrm{d}\,\boldsymbol{W}_2\, \mathrm{e}^{-\beta_2\left(\mathcal{L}_{\gamma_2}(\boldsymbol{W}_2, \mathcal{D}_2) + \frac{\gamma_{12}}{2}\|\boldsymbol{W}_2 - \boldsymbol{W}_1\|_2^2\right)} \end{aligned} \tag{3}$$

where $\mathcal{D}_1, \mathcal{D}_2$ denote the two dataset slices. This object represents the normalisation of the Boltzmann–Gibbs measure, and allows one to extract relevant information on the asymptotic behaviour of our model. The optimisations entailed in each learning phase can be described in the 'low noise' limit of $\beta_1, \beta_2 \to \infty$, where the measures focus on the minimisers of the respective losses. In order to study a self-averaging quantity that does not depend on a specific realisation of the dataset, we aim to compute the associated average free-energy:

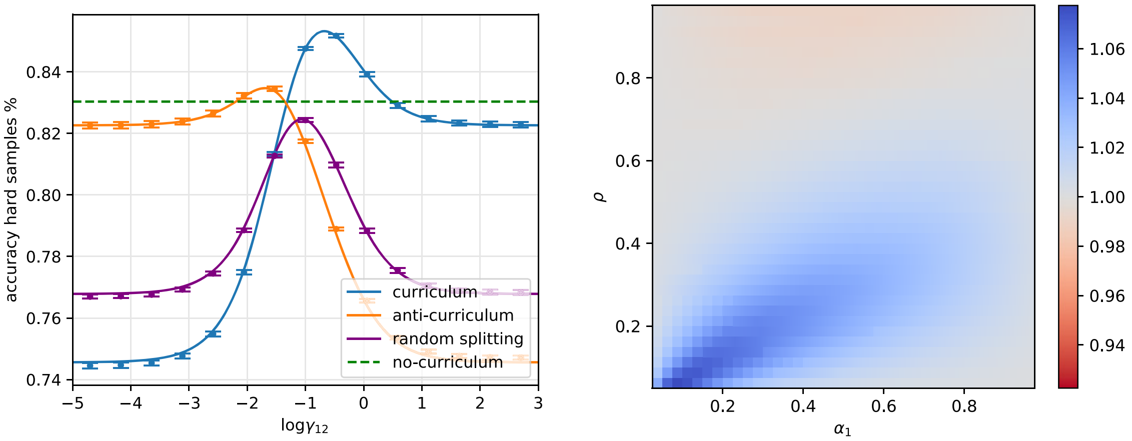
$$\Phi = \lim_{N\to\infty} \lim_{\beta_1,\beta_2\to\infty} \frac{1}{\beta_2 N} \left\langle \log \langle Z(\boldsymbol{W}_2, \boldsymbol{W}_1; \mathcal{D}_1, \mathcal{D}_2)\rangle_{\boldsymbol{W}_1} \right\rangle_{\mathcal{D}_1,\mathcal{D}_2}. \tag{4}$$

This quantity can be seen as a special case of the so-called Franz–Parisi potential computation [40, 41], and the entailed double average can be evaluated through the replica method. Refer to the appendices for details.

Similar to the online case, in high-dimensions the free-entropy concentrates on a deterministic function that depends on several order parameters that capture the geometrical distribution of teacher and student configurations. In addition to those already introduced in section 3, we also have δQ, which is linked to the variance of the student norm. Moreover, for each order parameter we also need to introduce a conjugate parameter, denoted in the following with the hat symbol. The final expression for the free-energy reads:

$$\begin{aligned}\Phi = \mathrm{extr}\Bigg[&-\left(\hat{R}R + \frac{1}{2}\left(\left(\hat{Q}\delta Q - \delta\hat{Q}Q\right)_{r+i}\right)\right) + g_S(\gamma_1, \gamma_2, \gamma_{12}) \\ &+ \alpha_1\, g_E(\Delta_1) + \alpha_2\, g_E(\Delta_2)\Bigg]\end{aligned} \tag{5}$$

where g_S and g_E are two scalar functions, often called entropic and energetic channels, that encode the dependence of the optimisation problem on the Gaussian prior and the logistic loss respectively. The extremum condition for the free-energy yields a system of fixed-point equations that converge to an asymptotic prediction for the order parameters, comparable with the results of numerical simulations on large instances, figure 4. At convergence, the order parameters can be inserted again in equation (2) to obtain an estimate of the test accuracy. Note that this formalism is not limited to two phases, but can be extended to the case of a discrete number of sequential stages.



(a) Learning with curricula

(b) Curriculum vs no-curriculum

Figure 4. Effect of elastic coupling (Gaussian prior) between curriculum phases. (a) Comparison between asymptotic performance of curricula (full lines) and single batch learning, at $\alpha_1 = 1, \alpha_2 = 1$, with a regularisation γ_1 that yields the best generalisation when learning the entire dataset (in principle not optimal for the other strategies). The points represent the results from 10 numerical simulations at size $N = 2000$. Parameters: $\rho = 0.50$, $\Delta_1 = 0$ and $\Delta_2 = 1$. (b) ratio between the accuracy reached by curriculum learning over anti-curriculum as a function of the number of easy samples in a dataset of dimension $\alpha_1 + \alpha_2 = 1$, and of the sparsity level of the teacher ρ. Note that ρ can also be seen as the fraction of relevant components in the inputs. $\Delta_1 = 0$ and $\Delta_1 = 1$. $\gamma_1 = \gamma_2$ and γ_{12} where set the values that optimise test performance.

The importance of sparsity. Sparsity is a key ingredient in determining the impact of curriculum strategies. It naturally introduces a notion of relevant and irrelevant input components, and defines a secondary learning goal, i.e. identifying what part of the presented data should be disregarded by the model. Curriculum learning can be extremely helpful in this identification process, since the easy samples are more transparent to this structure. This is also observed in human experiments [6]. However, the relative difficulty of the problem of inferring the support of the teacher and the problem of aligning with its non-zero components depends on the degree of sparsity ρ, so the effectiveness of curriculum can vary with it.

In the right panel of figure 4, we explore the interplay between the sparsity of the teacher ρ and the fraction of easy samples in the dataset α_1, comparing curriculum with the no-curriculum baseline. The phase diagram highlights the variability in the impact of the curriculum ordering:

- Curriculum is most effective at low values of ρ and close to the diagonal, where the fraction of easy examples in the dataset is comparable to the fraction of relevant dimensions.

https://doi.org/10.1088/1742-5468/ac9b3c

- When $\rho > 0.5$, the possible gain from ordering the samples according to difficulty is counterbalanced by the instrinsic cost of splitting the information content into two blocks, thus curriculum can become detrimental.
- When α_1 is too small compared to ρ (above diagonal), the first stage in the curriculum strategy can only help in the support identification problem, but will not allow a good estimation of the direction of the teacher. Because of the elastic prior, the second stage cannot improve too much over it and the effect of curriculum is small.
- When α is larger than the sparsity (below diagonal), the easy examples contain sufficient information for solving both the support and the teacher estimation problems, and this information is also exploited by the baseline. Thus the improvement of curriculum becomes negligible.

We refer to the appendices for an in-depth comparison with anti-curriculum.

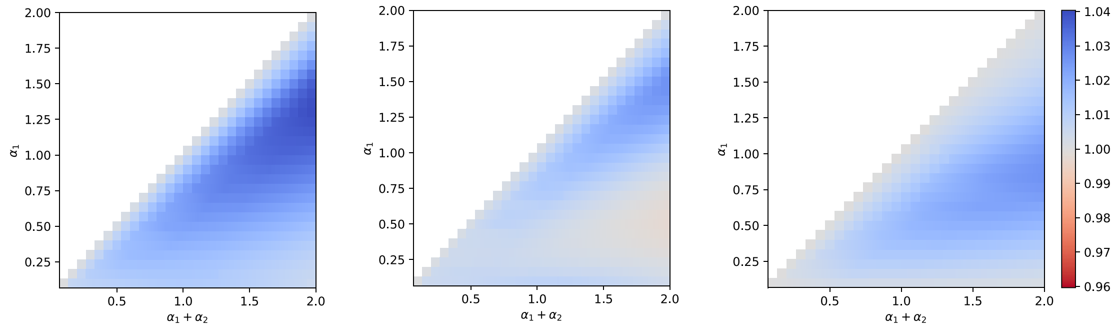
Asymptotic advantages of curriculum. Contrary to the online SGD case, if the fraction of relevant directions is small, batch learning with elastic coupling notably improves test accuracy of both curriculum and anti-curriculum above the baseline. This confirms the utility of curriculum strategies when the signal is partially 'hidden in clutter' [42]. Figure 5 shows similar phase diagrams to figure 2 but for the batch setting. At each point in the phase diagram the regularisation level $\gamma_1 = \gamma_2$ and the coupling γ_{12} are optimised to yield the best accuracy. In batch learning the performance order appears to be nearly always preserved: curriculum followed by anti-curriculum followed by baseline. We remark that this result is not trivial as splitting the learning process in two stages is not advantageous per se. Note that in the appendices we observe similar improvement applying the elastic coupling on the online setting and on real data.

5. Connection with experimental literature

Recent work has suggested that curriculum learning could provide an important window into the learning algorithms at work in biology [44]. Our analysis makes several predictions for curriculum effects. In this section we assess these predictions based on connections to extant experiments and propose future experimental tests.

First, we find that a curriculum strategy yields a speed up in learning in all the tested settings (see figure 1(c)). This acceleration is broadly consistent with the findings from cognitive science [1, 2, 6]. By contrast, our results show that the speed improvement does not necessarily translate into a sizeable generalisation error improvement, and the performance achieved at the end of training can even deteriorate when learning hyperparameters are not fully optimised (cf figure 3). Deterioration due to curricula has generally not been reported in the psychology literature, though it has been observed in ML [15]. This fact may suggest that animals naturally learn with near-optimal hyperparameters such that curricula generally confer benefits.

A more specific observation concerns the performance on different difficulties after learning. As reported in [43], human and rodent subjects trained in an auditory task using curricula showed the greatest improvement for intermediate level of difficulties as

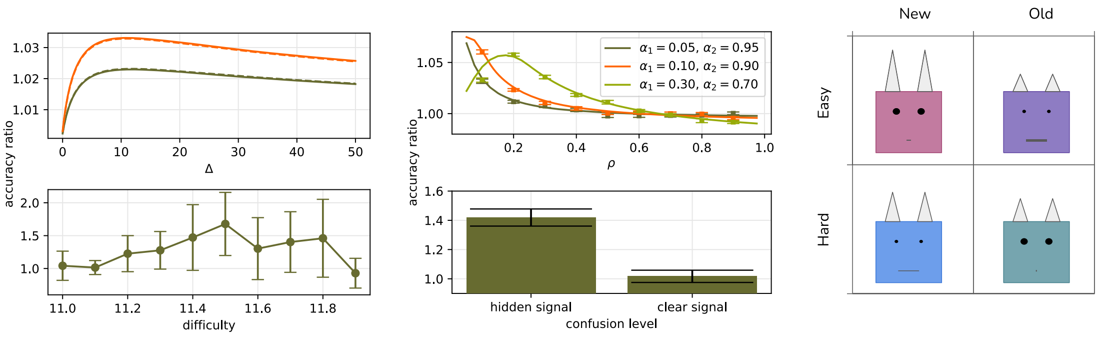


(a) Curriculum learning. (b) Anti-curriculum learning. (c) Curriculum vs anti-curriculum.

Figure 5. Phase diagram for the performance gap in the batch setting. The colour scale shows the ratio of the accuracy on hard instances for curriculum over no-curriculum (a), anti-curriculum over no-curriculum (b), and curriculum over anti-curriculum (c), as a function of the total dataset size $(\alpha_1 + \alpha_2)$ and easy dataset size (α_1). In contrast to the online case, performance benefits are greater and curriculum is strictly better than anti-curriculum. Both $\gamma_1 = \gamma_2$ and γ_{12} are optimised point-wise, in order to yield the best test accuracy. Parameters: $\rho = 0.50, \Delta_1 = 0$, $\Delta_2 = 1$.

depicted in figure 6(a) (bottom). The same conclusion can be drawn from the experiment of [7, 8], where, surprisingly, subjects trained with curricula to classify medical images showed poor performance in hard tasks compared to the control group. To address this phenomenon, we calculate accuracy as a function of difficulty in the model in figure 6(a) (top). Consistent with these experiments, we find regimes where the gap between curriculum learning and the baseline is non-monotonic, with the largest performance gain for intermediate difficulties. Contrary to [7, 8], however, we do not observe negative effects of curriculum for high difficulties. Further experiments that more systematically manipulate training and transfer difficulties could provide a stronger test of these predictions.

A key ingredient in our model is the role of sparsity, such that a small signal is embedded amidst many irrelevant features. Experimentally, the importance of having many factors of variation to obtaining a curriculum effect has been documented in the 'fading' experiments of [6]. Human subjects were trained on classification tasks involving stimuli with one task-relevant feature dimension and a variable number of task-irrelevant feature dimensions. Example cartoon 'daemon' stimuli are depicted in figure 6(c), where for instance horn height might be the distinguishing feature while colour, eye size, and mouth size might constitute task-irrelevant features. Without any irrelevant factors of variation ($\rho = 1$), they report no curriculum benefit. By contrast when 75% of features are irrelevant ($\rho = 0.25$), they record a strong curriculum effect, as shown in figure 6(b) bottom. This qualitative trend is also observed in our model (figure 6(b), top). While these experiments tested only two sparsity levels, further experiments could sample this dimension more extensively and test for interactions with the fraction of easy and hard



(a) Generalisation gap per difficulty.

(b) Role of sparsity.

(c) Fading experiment [6].

Figure 6. Connection with psychology experiments. (a) (Top) Accuracy ratio of different strategies in the model, with curriculum/no-curriculum in green and curriculum/anti-curriculum in orange. The ratio shows non-monotonic behaviour. (Bottom) The accuracy ratio obtained by [43]. Parameters $\rho = 0.5$, $\Delta_1 = 0.0$, $\Delta_2 = 1.0$, $\alpha_1 = 1$, $\alpha_2 = 1$ and optimal learning rate, variance at initialisation and weight decay. (b) (Top) Dependence on the sparsity of the generalisation gain of curriculum over no-curriculum, measured as ratio between final accuracy, for fixed total dataset size ($\alpha_1 + \alpha_2 = 1$). (Bottom) The ratio obtained from experiments 3 and 4 of [6]. (c) Example cartoon stimuli from the 'fading' paradigm used in [6], where participants distinguish daemons of the old world from daemons of the new world. The distinguishing feature (horn length) is diluted among many irrelevant features (colour, eye size, mouth size). Highlighting the relevant feature to participants leads to better and faster learning.

examples. We note that while the connectionist literature has addressed the effect of curriculum in several settings [10, 11, 45, 46], we found that easy-to-hard effects appear even in a simple setup without need for complex networks and/or dynamics.

Finally, our results may shed light on self-generated curricula during human development [47, 48]. Children undergo a spurt of vocabulary development that coincides with their ability to grasp and centre objects in the visual field [48]. Quantitative estimates of the amount of clutter (irrelevant objects) in self-generated views decrease due to this grasping ability, yielding a self-generated curriculum [42, 49]. Our model similarly predicts that reducing clutter should improve learning speed and performance. To verify this in a richer visual setting, we apply our curriculum-aware algorithm on real data where the loss is modified to keep track of the different phases of learning. We construct a simple cluttered object recognition task from the CIFAR10 dataset [50] by patching two images together into a 32×64 input image (figure 7(a)). The network has to learn that the classification depends only on the left image, while the right image is a distractor that is irrelevant to the classification. Easy and hard instances are obtained by reducing the brightness of the distractor. We train a single-layer network with the cross-entropy loss and the curriculum protocol with Gaussian prior between the two stages. Weights are optimised using SGD and momentum, with an annealed learning rate. All training

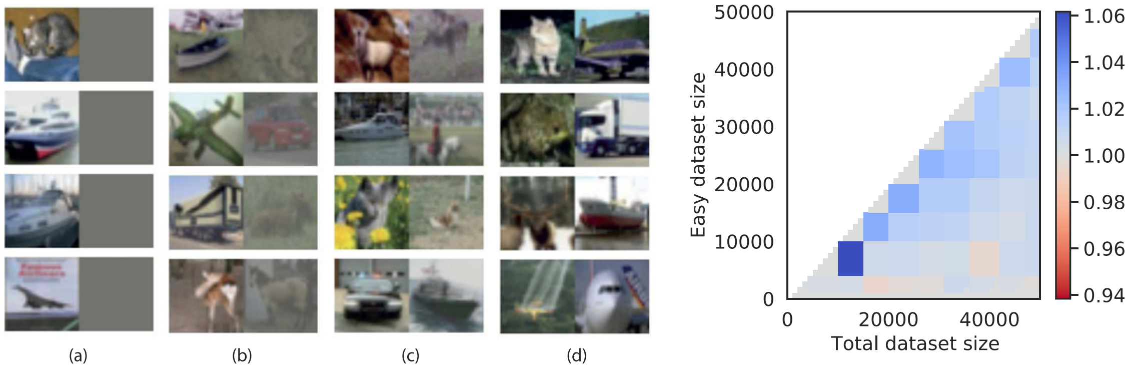


Figure 7. Experimental setting on CIFAR10-derived data. (a) Input samples combine a task-relevant image with a distractor image, and become progressively harder from left to right. (b) Ratio between final accuracy on hard instances for curriculum learning versus no curriculum. $\eta, \gamma, \gamma_{12}$, init, and stopping time are optimised.

parameters are optimised, and full details are presented in the appendices. Figure 7 shows a robust curriculum advantage in this setting, suggesting a possible functional benefit for children's self-generated curriculum.

6. Conclusions

We analysed a model of curriculum learning introduced by [12] and amenable of analytical treatment. This simple setting sheds light on results observed in the cognitive science and ML literature, and the theoretical tractability allows for exploration of a wide range of parameters that would be costly to obtain through experiments. Future work will need to move beyond models with simple loss landscapes to address the impact of curricula in complex tasks like reinforcement learning. Nevertheless, the model recapitulates a variety of observations in the literature [43, 51, 52], revealing that easy-to-hard effects can appear when a sparse signal is embedded in many irrelevant dimensions of variation. We find that making the algorithm curriculum-aware by modifying the loss can better exploit curricula, offering a potential route for improved practical algorithms. Other curriculum-aware approaches are possible such as adapting the learning algorithm [53] or the architecture [10]. On the psychology side, some of our predictions can help in designing new experiments. The benefit of anti-curriculum learning for intermediate sparsity, is a counter-intuitive result testable in animal experiments.

Acknowledgments

We thank Miguel Ruiz-Garcia and Ronald Dekker for precious discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and

Innovation Program Grant Agreement 714608-SMiLe, and a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (Grant No. 216386/Z/19/Z). A S is a CIFAR Azrieli Global Scholar in the Learning in Machines and Brains programme.

Appendix A. State evolution of the online dynamics

In this section we show how to derive the dynamical equations for the online dynamics. The equations given in an implicit form in the main text, f_{Q_r}, f_{Q_i}, f_R, are reported explicitly at the end of the next section, equations (A.22)–(A.24). Finally, in the subsequent section, we comment on how the state evolution is modified to deal with the Gaussian priors and we derive the new dynamical equations for that case.

Derivation. We follow the derivation proposed in [26, 35] to derive the averaged high-dimensional dynamical equations. The student is a one-layer network that minimises sample-wise the square error

$$\mathcal{L}^\mu = \frac{1}{2}(y^\mu - \hat{y}^\mu)^2 \doteq \frac{1}{2}(\delta^\mu)^2. \tag{A.1}$$

Given $\phi(\cdot) = \mathrm{sign}(\cdot), \sigma(\cdot) = \mathrm{erf}(\cdot/\sqrt{2})$, the online stochastic gradient descent updates are

$$\boldsymbol{W}^{\mu+1} = \boldsymbol{W}^\mu - \frac{\eta}{\sqrt{N}}\sigma'(\lambda_r^\mu + \lambda_i^\mu)\delta^\mu \boldsymbol{x}^\mu, \tag{A.2}$$

with

$$\lambda_r^\mu = \frac{1}{\sqrt{N}}\boldsymbol{W}_r \cdot \boldsymbol{x}_r^\mu, \tag{A.3}$$

$$\lambda_i^\mu = \frac{1}{\sqrt{N}}\boldsymbol{W}_i \cdot \boldsymbol{x}_i^\mu, \tag{A.4}$$

$$\rho^\mu = \frac{1}{\sqrt{N}}\boldsymbol{W}_T \cdot \boldsymbol{x}_r^\mu. \tag{A.5}$$

The evolution of the dynamics can be tracked using four order parameters:

$$Q_r = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_r, \tag{A.6}$$

$$Q_i = \frac{1}{N}\boldsymbol{W}_i \cdot \boldsymbol{W}_i, \tag{A.7}$$

$$R = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_T, \tag{A.8}$$

$$T = \frac{1}{N}\boldsymbol{W}_T \cdot \boldsymbol{W}_T\,; \tag{A.9}$$

representing the overlaps between the weights of student (relevant and irrelevant parts) and teacher.

The evolution of those follow from the definition of the dynamics equation (A.2). In the high-dimensional limit the random variables in the problem concentrates around

https://doi.org/10.1088/1742-5468/ac9b3c

the mean, therefor to the leading order we have the following equations

$$Q_r[k+1] = Q_r[k] + \frac{1}{N}\left[2\eta\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\lambda_r] + \rho\Delta\eta^2\mathbb{E}[\delta^2\ \sigma'(\lambda_r+\lambda_i)^2]\right]; \tag{A.10}$$

$$Q_i[k+1] = Q_r[k] + \frac{1}{N}\left[2\eta\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\lambda_i] + (1-\rho)\Delta\eta^2\mathbb{E}[\delta^2\ \sigma'(\lambda_r+\lambda_i)^2]\right]; \tag{A.11}$$

$$T[k+1] = Q_r[k] + \frac{1}{N}[\eta\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\rho]], \tag{A.12}$$

where the expectation acts with respect to all the stochastic variables. In order to obtain explicit formulae we need to evaluate those averages. The random variables in the equations—λ_r, λ_i and ρ—are Gaussian with zero mean, to characterise them we only need their covariance:

$$\Sigma_{\lambda_r,\lambda_i,\rho} = \begin{pmatrix} Q_r & 0 & R \\ 0 & Q_i & 0 \\ R & 0 & T \end{pmatrix}.$$

In order to derive analytical expression we must evaluate the expected values: $\mathbb{E}[\phi(\rho)\sigma'(\lambda)\rho]$, $\mathbb{E}[\phi(\rho)\sigma'(\lambda)\lambda]$, $\mathbb{E}[\sigma(\lambda)\sigma'(\lambda)\rho]$, $\mathbb{E}[\sigma(\lambda)\sigma'(\lambda)\lambda]$, $\mathbb{E}[\phi(\rho)^2\sigma'(\lambda)^2]$, $\mathbb{E}[\sigma(\lambda)^2\sigma'(\lambda)^2]$, and $\mathbb{E}[\phi(\rho)\sigma(\lambda)\sigma'(\lambda)^2]$, where σ is the activation function of the student and ϕ is the activation function of the teacher (in particular $\phi(\cdot) = \mathrm{sign}(\cdot)$ for classification).

$$\mathbb{E}[\phi(\rho)\sigma'(\lambda)\rho] = \frac{2}{\pi}\frac{\sqrt{T(Q_r+Q_i+1)-R^2}}{Q_r+Q_i+1} \tag{A.13}$$

$$\mathbb{E}[\phi(\rho)\sigma'(\lambda)\lambda_r] = \frac{2}{\pi}\frac{R(Q_i+1)}{Q_r+Q_i+1}\frac{1}{\sqrt{T(Q_r+Q_i+1)+R^2}}. \tag{A.14}$$

$$\mathbb{E}[\phi(\rho)\sigma'(\lambda)\lambda_i] = -\frac{2}{\pi}\frac{RQ_i}{Q_r+Q_i+1}\frac{1}{\sqrt{T(Q_r+Q_i+1)+R^2}}. \tag{A.15}$$

$$\mathbb{E}[\sigma(\lambda)\sigma'(\lambda)\rho] = \frac{2}{\pi}\frac{R}{Q_r+Q_i+1}\sqrt{\frac{Q_i+1}{2Q_i^2+2Q_rQ_i+3Q_i+2Q_r+1}}. \tag{A.16}$$

$$\mathbb{E}[\sigma(\lambda)\sigma'(\lambda)\lambda_r] = \frac{2}{\pi}\frac{Q_r}{Q_r+Q_i+1}\sqrt{\frac{Q_i+1}{2Q_i^2+2Q_rQ_i+3Q_i+2Q_r+1}}. \tag{A.17}$$

$$\mathbb{E}[\sigma(\lambda)\sigma'(\lambda)\lambda_i] = \frac{2}{\pi}\frac{Q_i}{Q_r+Q_i+1}\sqrt{\frac{Q_r+1}{2Q_r^2+2Q_rQ_i+3Q_r+2Q_i+1}}. \tag{A.18}$$

$$\mathbb{E}[\phi(\rho)^2\sigma'(\lambda)^2] = \frac{2}{\pi}\frac{1}{\sqrt{2Q_r+2Q_i+1}}. \tag{A.19}$$

https://doi.org/10.1088/1742-5468/ac9b3c

$$\mathbb{E}[\sigma(\lambda)^2\sigma'(\lambda)^2] = \frac{4}{\pi^2}\frac{1}{\sqrt{1+2(Q_r+Q_i)}}\sin^{-1}\left(\frac{Q_r+Q_i}{1+3(Q_r+Q_i)}\right). \tag{A.20}$$

$$\mathbb{E}[\phi(\rho)\sigma(\lambda)\sigma'(\lambda)^2] = \frac{4}{\pi^2}\frac{1}{\sqrt{2(Q_r+Q_i)+1}} \times \sin^{-1}\left(\frac{R\sqrt{Q_r+Q_i}}{\sqrt{3(Q_r+Q_i)+1}\sqrt{(2Q_r+2Q_i+1)[T(Q_r+Q_i)-R^2]+R^2}}\right). \tag{A.21}$$

Finally, we can substitute those equations into the equations (A.10)–(A.12) and obtained the state evolution equations used in the main section 3:

$$f_{Q_r}(Q_r[k],Q_i[k],R[k],T) = (1-\eta\gamma)^2 Q_r[k] + \frac{4\eta(1-\eta\gamma)}{N\pi(Q_r[k]+\Delta Q_i[k]+1)} \times\left[\frac{R[k](\Delta Q_i[k]+1)}{\sqrt{T(Q_r[k]+\Delta Q_i[k]+1)+R[k]^2}} - \frac{Q_r[k]}{\sqrt{2Q_r[k]+2\Delta Q_i[k]+1}}\right] + \frac{4}{\pi^2}\frac{\rho\eta^2}{N\sqrt{2(Q_r[k]+\Delta Q_i[k])+1}}\left[\frac{\pi}{2}+\sin^{-1}\left(\frac{Q_r[k]+\Delta Q_i[k]}{1+3(Q_r[k]+\Delta Q_i[k])}\right)+ - 2\sin^{-1}\left(\frac{R[k]}{\sqrt{3(Q_r[k]+\Delta Q_i[k])+1}\sqrt{T(2Q_r[k]+2\Delta Q_i[k]+1)-2R[k]^2}}\right)\right]; \tag{A.22}$$

$$f_{Q_i}(Q_r[k],Q_i[k],R[k],T) = (1-\eta\gamma)^2 Q_i[k] - \frac{4\eta(1-\eta\gamma)\Delta Q_i[k]}{N\pi(Q_r[k]+\Delta Q_i[k]+1)} \times\left[\frac{R[k]}{\sqrt{T(Q_r[k]+\Delta Q_i[k]+1)+R[k]^2}} + \frac{1}{\sqrt{2Q_r[k]+2\Delta Q_i[k]+1}}\right] + \frac{4}{\pi^2}\frac{(1-\rho)\Delta\eta^2}{N\sqrt{2(Q_r[k]+\Delta Q_i[k])+1}}\left[\frac{\pi}{2}+\sin^{-1}\left(\frac{Q_r[k]+\Delta Q_i[k]}{1+3(Q_r[k]+\Delta Q_i[k])}\right)+ - 2\sin^{-1}\left(\frac{R[k]}{\sqrt{3(Q_r[k]+\Delta Q_i[k])+1}\sqrt{T(2Q_r[k]+2\Delta Q_i[k]+1)-2R[k]^2}}\right)\right]; \tag{A.23}$$

$$f_R(Q_r[k], Q_i[k], R[k], T) = (1-\eta\gamma)R[k] + \frac{2\eta}{N\pi(Q_r[k]+\Delta Q_i[k]+1)} \times \left[\frac{T(Q_r[k]+\Delta Q_i[k]+1)-R[k]^2}{\sqrt{T(Q_r[k]+\Delta Q_i[k]+1)-R[k]^2}} - \frac{R[k]}{\sqrt{2Q_r[k]+2\Delta Q_i[k]+1}}\right]. \tag{A.24}$$

Elastic coupling. The introduction of the elastic coupling between stages of learning adds five new order parameters: three of them are just reminder of the previous stage and do not need to by updated $\tilde{Q}_r = \boldsymbol{W}_1^r \cdot \boldsymbol{W}_1^r/N$, $\tilde{Q}_i = \boldsymbol{W}_1^i \cdot \boldsymbol{W}_1^i/N$, and $\tilde{R} = \boldsymbol{W}_1^i \cdot \boldsymbol{W}^T/N$; two measure the correlation between the two stages $S_r = \boldsymbol{W}_1^r \cdot \boldsymbol{W}_2^r/N$ and $S_i = \boldsymbol{W}_1^i \cdot \boldsymbol{W}_2^i/N$ to the equations. These terms have associated their own state evolution equations slightly modified the updates of the other order parameters.

$$\begin{aligned} Q_r[k+1] &= (1-\eta\gamma+\eta\gamma_{12})^2 Q_r[k] + \frac{2\eta}{N}(1-\eta\gamma+\eta\gamma_{12})\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\lambda_r] \\ &\quad + \rho\Delta\frac{\eta^2}{N}\mathbb{E}[\delta^2\ \sigma'(\lambda_r+\lambda_i)^2] + 2\eta\gamma_{12}(1-\eta\gamma+\eta\gamma_{12})S_r[k] + \eta^2\gamma_{12}^2\tilde{Q}_r[k] \\ &\quad - \frac{2\eta^2\gamma_{12}}{N}\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\tilde{\lambda}_r]\,; \end{aligned} \tag{A.25}$$

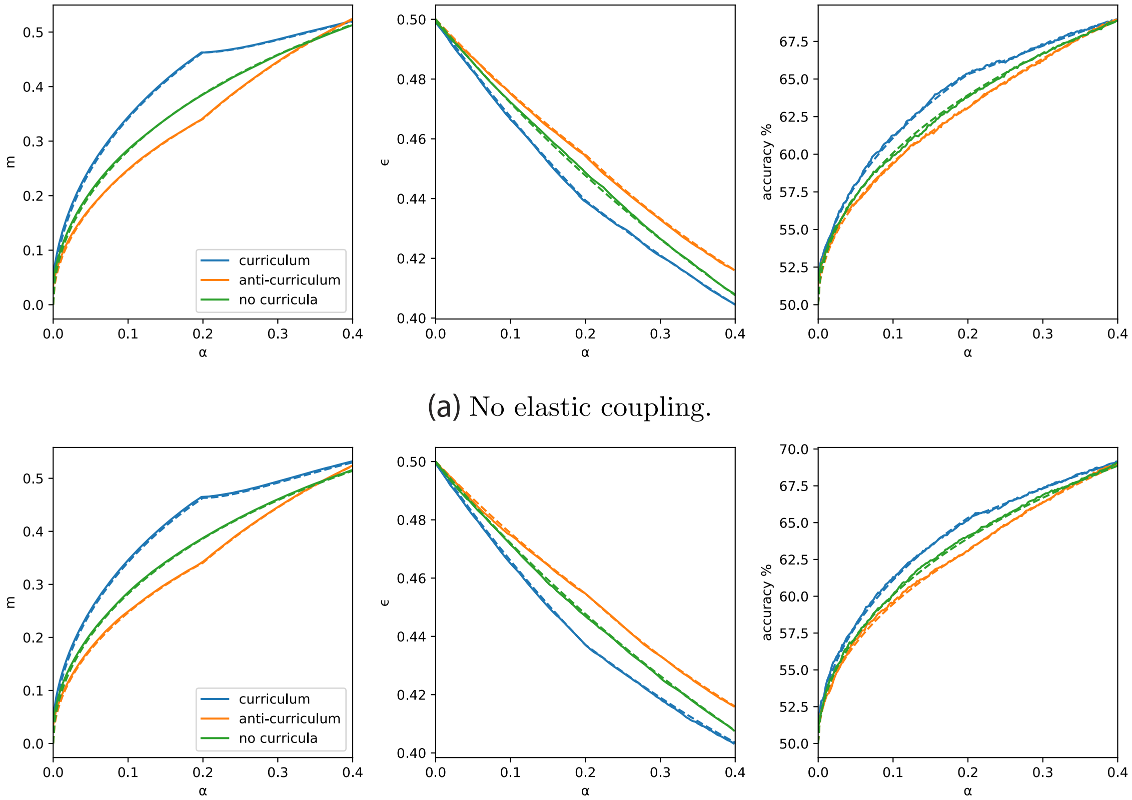
$$\begin{aligned} Q_i[k+1] &= (1-\eta\gamma+\eta\gamma_{12})^2 Q_i[k] + \frac{2\eta}{N}(1-\eta\gamma+\eta\gamma_{12})\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\lambda_i] \\ &\quad + (1-\rho)\Delta\frac{\eta^2}{N}\mathbb{E}[\delta^2\ \sigma'(\lambda_r+\lambda_i)^2] + 2\eta\gamma_{12}(1-\eta\gamma+\eta\gamma_{12})S_i[k] \\ &\quad + \eta^2\gamma_{12}^2\tilde{Q}_i[k] - \frac{2\eta^2\gamma_{12}}{N}\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\tilde{\lambda}_i]\,; \end{aligned} \tag{A.26}$$

$$R[k+1] = (1-\eta\gamma+\eta\gamma_{12})R[k] + \frac{\eta}{N}\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\rho] - \eta\gamma_{12}\tilde{R}[k]\,; \tag{A.27}$$

$$S_r[k+1] = (1-\eta\gamma+\eta\gamma_{12})S_r[k] + \frac{\eta}{N}\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\tilde{\lambda}_r] - \eta\gamma_{12}\tilde{Q}_r[k]\,; \tag{A.28}$$

$$S_i[k+1] = (1-\eta\gamma+\eta\gamma_{12})S_i[k] + \frac{\eta}{N}\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\tilde{\lambda}_i] - \eta\gamma_{12}\tilde{Q}_i[k]. \tag{A.29}$$

Introduced $\tilde{\lambda}_r = \frac{1}{\sqrt{N}}\boldsymbol{x}_r \cdot \tilde{\boldsymbol{W}}_r$ and $\tilde{\lambda}_i = \frac{1}{\sqrt{N}}\boldsymbol{x}_i \cdot \tilde{\boldsymbol{W}}_i$, this two additional random variables need to be averaged together with the others. The joint distribution of $\lambda_r, \lambda_i, \tilde{\lambda}_r, \tilde{\lambda}_i, \rho$ is still Gaussian with zero mean and covariance



(a) No elastic coupling.

(b) Optimal elastic coupling.

Figure A.1. Effect of elastic coupling in the curriculum. Figures showing the teacher–student cosine, the validation loss, and the accuracy of the three learning strategies. The two figures show the performance in presence (above) and absence (below) of elastic coupling. The dashed lines are obtained from the theoretical analysis, the full line come from the average of 500 simulations. The parameters η, γ, initialisation are set to the optimal values for each protocol. Parameters: $\rho = 0.5$, $\alpha_1 = 0.2$, $\alpha_2 = 0.2$, $\Delta_1 = 0$, $\Delta_2 = 1$.

$$\Sigma_{\lambda_r,\lambda_i,\tilde{\lambda}_r,\tilde{\lambda}_i,\rho} = \begin{pmatrix} Q_r & 0 & \tilde{S}_r & 0 & R \\ 0 & Q_i & 0 & \tilde{S}_i & 0 \\ \tilde{S}_r & 0 & \tilde{Q}_r & 0 & \tilde{R} \\ 0 & \tilde{S}_i & 0 & \tilde{Q}_i & 0 \\ R & 0 & \tilde{R} & 0 & T \end{pmatrix}.$$

Notice that, a part from a slight change of the existing equations, the coupling introduces only two additional integrals $\mathbb{E}[\delta\ \sigma'(\lambda_r + \lambda_i)\tilde{\lambda}_r]$ and $\mathbb{E}[\delta\ \sigma'(\lambda_r + \lambda_i)\tilde{\lambda}_i]$. After long, but straightforward, computations we obtain

$$\mathbb{E}[\delta\ \sigma'(\lambda_r + \lambda_i)\tilde{\lambda}_r] = \frac{2}{\pi}\frac{S_r}{Q_r + Q_i + 1}\frac{Q_i + 1}{2Q_i^2 + 2Q_rQ_i + 3Q_i + 2Q_r + 1} +$$

$$
\begin{aligned}
&-\frac{2}{\pi}\frac{TS_r - R\tilde{R}}{Q_rT - R^2}\frac{R(Q_i+1)}{Q_r+Q_i+1}\frac{1}{\sqrt{T(Q_r+Q_i+1)-R^2}}+\\
&-\frac{2}{\pi}\frac{T\tilde{R} - RS_r}{Q_rT - R^2}\frac{1}{\sqrt{T(Q_r+Q_i+1)-R^2}}\\
&\times\frac{1}{\frac{1}{T}+\frac{R^2}{Q_rT-R^2}\left(\frac{1}{T}-\frac{Q_i+1}{T(Q_r+Q_i+1)-R^2}\right)},
\end{aligned}
\tag{A.30}
$$

$$
\begin{aligned}
\mathbb{E}[\delta\ \sigma'(\lambda_r+\lambda_i)\tilde{\lambda}_i] = {}&\frac{2}{\pi}\frac{S_i}{Q_r+Q_i+1}\frac{Q_r+1}{2Q_r^2+2Q_rQ_i+3Q_r+2Q_i+1}+\\
&-\frac{2}{\pi}\frac{S_iR}{Q_r+Q_i+1}\frac{1}{\sqrt{T(Q_r+Q_i+1)-R^2}}.
\end{aligned}
\tag{A.31}
$$

Finally all the expected values are known and we can obtain the analytic updates equations (A.25)–(A.29) with the coupling. Figure A.1(a) shows an instance of the problem at $\alpha_1 = 0.2$ and $\alpha_2 = 0.2$, a situation that is particularly adversarial for curriculum according the phase diagram figure 2. This situation is treated by the introduction of Gaussian priors, figure A.1(b), consistently with the phase diagram in figure 7(c).

Appendix B. Replica computation for the batch case

We here the detailed replica computation employed to obtain the analytic description of curriculum learning in the batch case, in section 4. As mentioned in the main, we aim to study a coupled system, represented by the following partition function:

$$
\begin{aligned}
\langle Z(\boldsymbol{W}_2,\,\boldsymbol{W}_1\,;\mathcal{D}_1,\mathcal{D}_2)\rangle_{\boldsymbol{W}_1} = {}&\int \mathrm{d}\,\boldsymbol{W}_1\frac{\mathrm{e}^{-\beta_1\mathcal{L}_{\gamma_1}(\boldsymbol{W}_1,\mathcal{D}_1)}}{Z_1(\boldsymbol{W}_1)}\log\\
&\times\int \mathrm{d}\,\boldsymbol{W}_2\,\mathrm{e}^{-\beta_2\left(\mathcal{L}_{\gamma_2}(\boldsymbol{W}_2,\mathcal{D}_2)+\frac{\gamma_{12}}{2}\|\boldsymbol{W}_2-\boldsymbol{W}_1\|_2^2\right)},
\end{aligned}
\tag{B.1}
$$

where the examples is $\mathcal{D}_1, \mathcal{D}_2$ are characterised by a different variances in the irrelevant components.

This type of quantity is usually denoted as a ‘disordered’ partition function in statistical physics jargon, meaning that it is still dependent on a given realisation of the datasets, i.e. the source of disorder in this model. We want to characterise a typical realisation of this object, in the high-dimensional limit. However, because of its long-tailed statistics, the partition function turns out not to be a self-averaging quantity, i.e. its expectation over the dataset realisations will not correspond to the typical case scenario we are after. It is instead better to focus on the computation of the associated average free-entropy:

$$
\Phi = \lim_{N\to\infty}\lim_{\beta_1,\beta_2\to\infty}\frac{1}{\beta_2 N}\left\langle\log\langle Z(\boldsymbol{W}_2,\,\boldsymbol{W}_1\,;\mathcal{D}_1,\mathcal{D}_2)\rangle_{\boldsymbol{W}_1}\right\rangle_{\mathcal{D}_1,\mathcal{D}_2}.
\tag{B.2}
$$

What is immediately apparent is that we have to take the expectation of a logarithm, which is not tractable with rigorous mathematical methods. Moreover, we also have to average over the measure for $\boldsymbol{W}_1$, which is also a complicated operation.

Fortunately, replica theory offers a method for approaching this calculation [40, 41]. The idea is to exploit two separate replica tricks:

- In order to evaluate the disorder average, the logarithm can be removed by replicating the second weight configuration, i.e. introducing n identical replicas $\{\boldsymbol{W}_2^a\}_{a=1}^n$, and extrapolating the final result from the $n \to 0$ limit. This is based on the mathematical identity $\log x = \lim_{n\to 0} \partial_n x^n$.
- The average over the teacher can instead be computed by introducing $\tilde{n}-1$ non-interacting and a single interacting replica of the first weight configuration $\{\boldsymbol{w}_1^c\}_{c=1}^{\tilde{n}}$. Thus, only the $c=1$ replica will enter the Gaussian prior in the student measure. The sought statistical average is again recovered in the limit $\tilde{n} \to 0$.

Because of the high-dimensional limit we are considering, all typical realisations of the teacher vector with a given sparsity ρ will yield an identical free-entropy. Thus, we can avoid averaging and instead fix a gauge $\boldsymbol{W}_{T,i} = 1$ for $i = 1, \ldots, \rho N$ and $\boldsymbol{W}_{T,i} = 0$ elsewhere. In order to simplify the presentation, in the following we will assume that the datasets contain respectively α_1 and α_2 patterns, and that a curriculum ordering was employed, $\Delta_1 < \Delta_2$. Moreover, to avoid confusion with component and replica indices, we will denote with $\tilde{\boldsymbol{W}} = \boldsymbol{W}_1$ and $\boldsymbol{W} = \boldsymbol{W}_2$, so that all quantities with a tilde refer to the optimisation on the first dataset.

After the described replication procedures, we get the following expression for the average free-entropy:

$$
\begin{aligned}
\Phi = \frac{1}{N} \lim_{n,\tilde{n}\to 0} \partial_n \Bigg\langle \lim_{\tilde{\beta},\beta\to\infty} \frac{1}{\beta} \int \prod_{c=1}^{\tilde{n}} \mathrm{d}\, \tilde{\boldsymbol{W}}^c e^{-\frac{\tilde{\beta}\gamma_1}{2} \| \tilde{\boldsymbol{W}}^c \|_2^2} \prod_{\mu=1}^{\alpha_1 N} \\
\times \prod_{c=1}^{\tilde{n}} e^{-\frac{\beta}{2} \ell\left(\mathrm{sign}\left(\sum_{i=1}^{\rho N} \frac{x_i^\mu}{\sqrt{N}}\right), \sigma\left(\sum_{i=1}^{N} \frac{\tilde{W}_i^c x_i^\mu(\Delta_1)}{\sqrt{N}}\right)\right)} \\
\times \int \prod_{a=1}^{n} \mathrm{d}\, \boldsymbol{W}^a e^{-\frac{\beta\gamma_2}{2} \| \boldsymbol{W}^a \|_2^2} e^{-\frac{\beta\gamma_{12}}{2} \| \boldsymbol{W}^a - \tilde{\boldsymbol{W}}^1 \|_2^2} \prod_{\mu=1}^{\alpha_2} \\
\times \prod_{a} e^{-\frac{\beta}{2} \ell\left(\mathrm{sign}\left(\sum_{i=1}^{\rho N} \frac{x_i^\mu}{\sqrt{N}}\right), \sigma\left(\sum_{i=1}^{N} \frac{W_i^a x_i^\mu(\Delta_2)}{\sqrt{N}}\right)\right)} \Bigg\rangle_{\{\boldsymbol{x}^\mu\}},
\end{aligned}
\tag{B.3}
$$

where $\ell(y,\hat{y}) = \log(1+e^{-y\hat{y}})$ indicates the standard logistic loss. The next step is to explicitly compute the averages over the dataset realisations. Before doing that, we need to isolate the dependence of our expression on the patterns, and we achieve this by introducing Dirac's δ-functions for the pre-activations. We will use the integral representation of the δ, with integration variables u for the teacher preactivations λ for the student preactivations:

$$\begin{aligned}
&\frac{1}{N}\lim_{n,\tilde{n}\to 0}\partial_n \int \prod_{c=1}^{\tilde{n}} \mathrm{d}\,\tilde{\boldsymbol{W}}^c e^{-\frac{\beta\lambda}{2}\|\tilde{\boldsymbol{W}}^c\|_2^2} \int \prod_{a=1}^{\tilde{n}} \mathrm{d}\,\boldsymbol{W}^a e^{-\frac{\beta\lambda}{2}\|\boldsymbol{W}^a\|_2^2} e^{-\frac{\beta\lambda_{12}}{2}\|\boldsymbol{W}^a-\tilde{\boldsymbol{W}}^1\|_2^2}\\
&\quad\times\left\langle \int\prod_{\mu}\frac{\mathrm{d}\tilde{u}_{1\mu}\mathrm{d}\hat{\tilde{u}}_{1\mu}}{2\pi}\mathrm{e}^{\mathrm{i}\hat{\tilde{u}}_{1\mu}\left(\tilde{u}_{1\mu}-\sum_{i=1}^{\rho N}\frac{(\tilde{x}_1)_i^\mu}{\sqrt{N}}\right)} \int\prod_{\mu,c}\frac{\mathrm{d}\tilde{\lambda}_{1\mu}^c\mathrm{d}\hat{\tilde{\lambda}}_{1\mu}^c}{2\pi}\right.\\
&\quad\times\mathrm{e}^{\mathrm{i}\hat{\tilde{\lambda}}_{1\mu}^c\left(\lambda_{1\mu}^c-\sum_{i=1}^{N}\frac{\tilde{W}_i^c(\tilde{x}_1)_i^\mu}{\sqrt{N}}\right)} \int\prod_{\mu}\frac{\mathrm{d}u_{2\mu}\mathrm{d}\hat{u}_{2\mu}}{2\pi}\mathrm{e}^{\mathrm{i}\hat{u}_{2\mu}\left(u_{2\mu}-\sum_{i=1}^{\rho N}\frac{(x_2)_i^\mu}{\sqrt{N}}\right)}\\
&\quad\times\left.\int\prod_{\mu,a}\frac{d\lambda_{2\mu}^a\mathrm{d}\hat{\lambda}_{2\mu}^a}{2\pi}\mathrm{e}^{\mathrm{i}\hat{\lambda}_{2\mu}^a\left(\lambda_{2\mu}^a-\sum_{i=1}^{N}\frac{W_i^a(x_2)_i^\mu}{\sqrt{N}}\right)}\right\rangle_{\{\boldsymbol{x}^\mu\}}\\
&\quad\times\prod_{\mu,c}e^{-\frac{\beta}{2}\ell\left(\mathrm{sign}(\tilde{u}_{1\mu}),\sigma\left(\tilde{\lambda}_{1\mu}^c\right)\right)}\prod_{\mu,a}e^{-\frac{\beta}{2}\ell\left(\mathrm{sign}(u_{2\mu}),\sigma\left(\lambda_{2\mu}^a\right)\right)}. \qquad \text{(B.4)}
\end{aligned}$$

Thus, the disorder average is now factorised and only involves exponential terms. Since the two datasets are independent now that we made the teacher explicit, we can take the averages over each one separately. In both cases we get:

$$\begin{aligned}
\langle.\rangle &= \prod_{i=1}^{\rho N}\mathbb{E}_{(x_{\mathrm{rel}})_i^\mu}\mathrm{e}^{-\mathrm{i}\left(\frac{\hat{u}}{\sqrt{N}}+\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\right)(x_{\mathrm{rel}})_i^\mu}\prod_{i=\rho N+1}^{N}\mathbb{E}_{(x_{\mathrm{irr}})_i^\mu}\mathrm{e}^{-\mathrm{i}\left(\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\right)(x_{\mathrm{irr}})_i^\mu}\\
&= \prod_{i=1}^{\rho N}\left(1-i\left(\frac{\hat{u}}{\sqrt{N}}+\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\right)\overline{x_{\mathrm{rel}}}-\frac{1}{2}\left(\frac{\hat{u}}{\sqrt{N}}+\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\right)^2\mathrm{Var}(x_{\mathrm{rel}})\right)\\
&\quad\times\prod_{i=\rho N+1}^{N}\left(1-i\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\overline{x_{\mathrm{irr}}}-\frac{1}{2}\left(\sum_a\hat{\lambda}_a^\mu\frac{W_i^a}{\sqrt{N}}\right)^2\mathrm{Var}(x_{\mathrm{irr}})\right) \qquad \text{(B.5)}
\end{aligned}$$

$$\begin{aligned}
&= \prod_{i=1}^{\rho N}\left(1-\frac{1}{2N}(\hat{u}^\mu)^2-\frac{1}{N}\sum_a\hat{u}^\mu\hat{\lambda}_a^\mu W_i^a-\frac{1}{2N}\sum_{ab}\hat{\lambda}_a^\mu\hat{\lambda}_b^\mu W_i^a W_i^b\right)\\
&\quad\times\prod_{i=\rho N+1}^{N}\left(1-\frac{\Delta^\mu}{2N}\sum_{ab}\hat{\lambda}_a^\mu\hat{\lambda}_b^\mu W_i^a W_i^b\right)\\
&= e^{-\frac{1}{2}\sum_{ab}\hat{\lambda}_a^\mu\hat{\lambda}_b^\mu\left(\frac{\sum_{i=1}^{\rho N}W_i^aW_i^b}{N}+\Delta\frac{\sum_{i=\rho N+1}^{N}W_i^aW_i^b}{N}\right)-\frac{\rho}{2}(\hat{u}^\mu)^2-\hat{u}^\mu\sum_a\hat{\lambda}_a^\mu\frac{\sum_{i=1}^{\eta N}W_i^a}{N}}. \qquad \text{(B.6)}
\end{aligned}$$

This expression suggests what are the order parameters that capture the interactions of the model, namely:

- The teacher–student overlap at the end of the first learning phase: $\tilde{R}^c=\frac{\sum_{i=1}^{\rho N}\tilde{W}_i^c}{N}$.

- The teacher–student overlap at the end of the second learning phase: $R^a = \frac{\sum_{i=1}^{\rho N} W_i^a}{N}$.
- The norm of the student after the first stage, decomposed into relevant/irrelevant parts: $\tilde{Q}_r^{cd} = \frac{\sum_{i=1}^{\rho N} \tilde{W}_i^c \tilde{W}_i^d}{N}$, $\tilde{Q}_i^{cd} = \frac{\sum_{i=\rho N+1}^{N} \tilde{W}_i^c \tilde{W}_i^d}{N}$.
- The norm of the student after the second stage, decomposed into relevant/irrelevant parts: $Q_r^{ab} = \frac{\sum_{i=1}^{\rho N} W_i^a W_i^b}{N}$, $Q_i^{ab} = \frac{\sum_{i=\rho N+1}^{N} W_i^a W_i^b}{N}$.

Therefore, after introducing these definitions by means of Dirac's δ-functions, we can rewrite our replicated expression as:

$$\begin{aligned}\Omega^n = &\int \prod_c \frac{\mathrm{d}\tilde{R}^c \mathrm{d}\hat{\tilde{R}}^c}{2\pi/N} \int \prod_a \frac{\mathrm{d}R^a \mathrm{d}\hat{R}^a}{2\pi/N} \int \prod_{cd} \frac{\mathrm{d}\tilde{Q}_r^{cd} \mathrm{d}\hat{\tilde{Q}}_r^{cd}}{2\pi/N} \int \prod_{cd} \frac{\mathrm{d}\tilde{Q}_i^{cd} \mathrm{d}\hat{\tilde{Q}}_i^{cd}}{2\pi/N} \\ &\times \int \prod_{ab} \frac{\mathrm{d}Q_r^{ab} \mathrm{d}\hat{Q}_r^{ab}}{2\pi/N} \int \prod_{ab} \frac{\mathrm{d}Q_i^{ab} \mathrm{d}\hat{Q}_i^{ab}}{2\pi/N} G_i \, G_S\left(\hat{\tilde{R}}, \hat{R}, \hat{\tilde{Q}}_r, \hat{Q}_r\right)^{\rho N} \\ &\times G_S\left(0, 0, \tilde{Q}_i, Q_i\right)^{(1-\rho)N} G_E\left(\Delta_1, \tilde{Q}_r, \tilde{Q}_i, \tilde{R}, \tilde{n}\right)^{\alpha_1 N} G_E(\Delta_2, Q_r, Q_i, R, n)^{\alpha_2 N} \end{aligned} \tag{B.7}$$

where we introduced interaction, entropic and energetic potentials:

$$\begin{aligned} G_i = \exp\Bigg(-N\Bigg(&\sum_c \hat{\tilde{m}}^c \tilde{m}^c + \sum_a \hat{m}^a m^a + \sum_{cd} \hat{\tilde{Q}}_r^{cd} \tilde{Q}_r^{cd} + \sum_{cd} \hat{\tilde{Q}}_i^{cd} \tilde{Q}_i^{cd} \\ &+ \sum_{ab} \hat{Q}_r^{ab} Q_r^{ab} + \sum_{ab} \hat{Q}_i^{ab} Q_i^{ab}\Bigg)\Bigg) \end{aligned} \tag{B.8}$$

$$\begin{aligned} G_S\left(\tilde{R}, R, \tilde{Q}, Q\right) = &\int \prod_c \left[\mathrm{d}\tilde{W}^c e^{-\frac{\beta\gamma}{2}\left(\tilde{W}^c\right)^2}\right] e^{-\frac{n\beta\gamma_{12}}{2}(\tilde{W}^1)^2} \int \prod_a \left[\mathrm{d}W^a e^{-\frac{\beta(\gamma+\gamma_{12})}{2}(W^a)^2}\right] \\ &\times \exp\Bigg(\sum_c \hat{\tilde{R}}^c \tilde{W}^c + \sum_a \hat{R}^a W^a + \sum_{cd} \hat{\tilde{Q}}^{cd} \tilde{W}^c \tilde{W}^d \\ &+ \sum_{ab} \hat{Q}^{ab} W^a W^b + \beta\gamma_{12} W^a \tilde{W}^1\Bigg) \end{aligned} \tag{B.9}$$

$$\begin{aligned} G_E(\Delta, Q_r, Q_i, m, n) = &\int \frac{\mathrm{d}u\mathrm{d}\hat{u}}{2\pi} e^{iu\hat{u}} e^{-\frac{\varrho}{2}(\hat{u})^2} \int \prod_{a=1}^{n} \frac{d\lambda^a \mathrm{d}\hat{\lambda}^a}{2\pi} \mathrm{e}^{\mathrm{i}\lambda^a \hat{\lambda}^a} \\ &\times e^{-\frac{1}{2}\sum_{ab} \hat{\lambda}_a \hat{\lambda}_b \left(Q_r^{ab} + \Delta Q_i^{ab}\right) - \hat{u}\sum_a \hat{\lambda}_a R^a - \frac{\beta}{2}\ell(u, \lambda^a)}. \end{aligned} \tag{B.10}$$

Replica symmetric ansatz. The replica trick allowed us to express the average free-entropy as a function of the overlap order parameters. However, these objects are $n \times n$ matrices or n-dimensional vectors and in principle we have to average over all their possible realisations. Fortunately, the integrand function is exponential in N and in the thermodynamic limit $N \to \infty$ the integrals are dominated by the extremisers of the action, and thus can be approximated with the saddle-point method. Still, we need a guess for how to parameterise these order parameters. The simplest possible ansatz, which turns out to be the correct one in convex problems as the one at hand, is the so-called replica symmetric (RS) ansatz, given by:

- $\tilde{R}^c = \tilde{R}$.
- $R^a = R$.
- $\tilde{Q}^{cd}_{r/i} = \tilde{q}_{r/i}$, for $c \neq d$; $\tilde{Q}^{cd}_{r/i} = \tilde{Q}_{r/i}$ for $c = d$.
- $Q^{ab}_{r/n} = q_{r/n}$ for $a \neq b$; $Q^{ab}_{r/n} = Q_{r/n}$ for $a = b$.

We also perform a Wick rotation $-i\hat{Q}_{ac,bd} \to \hat{Q}_{ac,bd}$ in order to deal with real valued conjugate parameters and pose a similar ansatz for them. In the next paragraph we will compute the three terms separately, and finally put them together in the expression for the RS free-entropy.

Interaction term. We start by evaluating the interaction term, or better its normalised logarithm $g_i = \lim_{\tilde{n}\to 0} \log G_i/(nN)$:

$$g_i = -\lim_{\tilde{n}\to 0} \frac{1}{n}\left(\tilde{n}\hat{\tilde{R}}\tilde{R} + n\hat{R}R + \tilde{n}\left(\frac{\hat{\tilde{Q}}_r\tilde{Q}_r}{2} + \frac{\hat{\tilde{Q}}_i\tilde{Q}_i}{2}\right) + \frac{\tilde{n}(\tilde{n}-1)}{2}\left(\hat{\tilde{q}}_r\tilde{q}_r + \hat{\tilde{q}}_i\tilde{q}_i\right)\right.$$
$$\left. + n\left(\frac{\hat{Q}_rQ_r}{2} + \frac{\hat{Q}_iQ_i}{2}\right) + \frac{n(n-1)}{2}(\hat{q}_rq_r + \hat{q}_iq_i)\right) \tag{B.11}$$
$$= -\left(\hat{R}R + \frac{\left(\hat{Q}_rQ_r + \hat{Q}_iQ_i\right)}{2} - \frac{1}{2}(\hat{q}_rq_r + \hat{q}_iq_i)\right) \tag{B.12}$$

In order to recover the optimisation problems entailed in the curriculum procedure, we now have to consider the zero temperature limit of this expression. When $\beta \to \infty$, the order parameters follow non-trivial scaling laws:

- $\hat{Q} \to \beta^2\hat{Q} + \mathcal{O}(\beta)$, $\hat{q} \to \beta^2\hat{Q}$
- $(\hat{Q} - \hat{q}) \to -\beta\delta\hat{Q}$
- $\hat{R} \to \beta\hat{R}$
- $Q - q = \delta Q/\beta$

and similarly for the tilde parameters. Intuitively, looking at the last scaling law, we see that as the measure gets focused on the single minimiser of the loss, the overlap between different replicas q rapidly converges to the norm Q. Moreover, the scaling with the inverse temperature of the conjugate parameters prevents the interaction term from becoming sub-dominant in the saddle-point. If we substitute the rescaled parameters in the above expression we obtain:

$$g_i = -\beta\left(\hat{R}R + \frac{1}{2}\left(\hat{Q}_r\delta Q_r - \delta\hat{Q}_rQ_r\right) + \frac{1}{2}\left(\hat{Q}_i\delta Q_i - \delta\hat{Q}_iQ_i\right)\right). \tag{B.13}$$

Entropic term. We can now compute a similar quantity for the entropic potential, $g_i = \frac{\lim_{n\to 0}}{n} \log G_S\left(\tilde{R}, R, \tilde{Q}, Q\right)$. The general expression we will obtain can be specialised to the two cases $\left(\left\{\tilde{R}, R, \tilde{Q}_r, Q_r\right\}, \left\{0, 0, \tilde{Q}_i, Q_i\right\}\right)$ appearing in the free-entropy. After substituting the RS ansatz we find:

$$
\begin{aligned}
g_S = \lim_{\tilde{n}\to 0} \frac{1}{n} \log \int \prod_c \left[\mathrm{d}\tilde{W}^c e^{-\frac{\beta\gamma}{2}\left(\tilde{W}^c\right)^2}\right] e^{-\frac{n\beta\gamma_{12}}{2}(\tilde{W}^1)^2} \int \prod_a \left[\mathrm{d}W^a e^{-\frac{\beta(\gamma+\gamma_{12})}{2}(W^a)^2}\right] \\
\times \exp\left(\hat{\tilde{R}} \sum_c \tilde{W}^c + \hat{R} \sum_a W^a + \frac{1}{2}\left(\hat{\tilde{Q}} - \hat{\tilde{q}}\right) \sum_c \left(\tilde{W}^c\right)^2 + \frac{\hat{\tilde{q}}}{2}\left(\sum_c \tilde{W}^c\right)^2\right. \\
\left. + \frac{1}{2}\left(\hat{Q} - \hat{q}\right) \sum_a (W^a)^2 + \frac{\hat{q}}{2}\left(\sum_a W^a\right)^2 + \beta\gamma_{12} \sum_a \tilde{W}^1 W^a\right) \\
= \lim_{\tilde{n}\to 0} \frac{1}{n} \log \int \mathcal{D}z \int \mathcal{D}\tilde{z} \int \prod_c \left[\mathrm{d}\tilde{W}^c e^{-\frac{\beta\gamma}{2}\left(\tilde{W}^c\right)^2}\right] e^{-\frac{n\beta\gamma_{12}}{2}(\tilde{W}^1)^2} \int \prod_a \mathrm{d}W^a \\
\times e^{-\frac{\beta(\gamma+\gamma_{12})}{2}(W^a)^2} \exp\left(\frac{1}{2}\left(\hat{\tilde{Q}} - \hat{\tilde{q}}\right) \sum_c \left(\tilde{W}^c\right)^2 + \frac{1}{2}\left(\hat{Q} - \hat{q}\right) \sum_a (W^a)^2\right. \\
\left. + \left(\hat{\tilde{R}} + \sqrt{\hat{\tilde{q}}}\tilde{z}\right) \sum_c \tilde{W}^c + \left(\hat{R} + \beta\gamma_{12} W_1 + \sqrt{\hat{q}}z\right) \sum_a W^a\right)
\end{aligned}
\tag{B.14}
$$

$$
= \int \mathcal{D}z \int \mathcal{D}\tilde{z} \frac{\int \mathrm{d}\tilde{W} e^{-\frac{1}{2}(\beta\tilde{\gamma} - (\hat{\tilde{Q}} - \hat{\tilde{q}}))\tilde{W}^2 + (\hat{\tilde{R}} + \sqrt{\hat{\tilde{q}}}\tilde{z})\tilde{W}} \log\left(\int \mathrm{d}W\, e^{-\frac{1}{2}\left(\beta(\gamma+\gamma_{12}) - (\hat{Q} - \hat{q})\right)W^2 + \left(\hat{R} + \beta\gamma_{12}W_1 + \sqrt{\hat{q}}z\right)W}\right)}{\int \mathrm{d}\tilde{W} e^{-\frac{1}{2}(\beta\tilde{\gamma} - (\hat{\tilde{Q}} - \hat{\tilde{q}}))\tilde{W}^2 + (\hat{\tilde{R}} + \sqrt{\hat{\tilde{q}}}\tilde{z})\tilde{W}}}. \tag{B.15}
$$

In the zero-temperature limit, we consider the same rescaling of the order parameters we described above. The integrals over the weights become an extremum operation:

$$
g_s = \lim_{\beta\to\infty} \beta \int \mathcal{D}z \int \mathcal{D}\tilde{z} M_s^\star, \tag{B.16}
$$

where:

$$
M_s^\star = \max_W \left\{-\frac{1}{2}\left((\gamma + \gamma_{12}) + \delta\hat{Q}\right)W^2 + \left(\hat{R} + \gamma_{12}\tilde{W}^\star + \sqrt{\hat{Q}}z\right)W\right\} \tag{B.17}
$$

$$
= \frac{1}{2} \frac{\left(\hat{R} + \gamma_{12}\tilde{W}^\star + \sqrt{\hat{Q}}z\right)^2}{(\gamma + \gamma_{12}) + \delta\hat{Q}} \tag{B.18}
$$

and where: $\tilde{W}^\star = \mathrm{argmax}_{\tilde{W}} \left\{-\frac{1}{2}(\tilde{\gamma} + \delta\hat{\tilde{Q}})\tilde{W}^2 + (\hat{\tilde{R}} + \sqrt{\hat{\tilde{Q}}}\tilde{z})\tilde{W}\right\} = \frac{\hat{\tilde{R}} + \sqrt{\hat{\tilde{Q}}}\tilde{z}}{\tilde{\gamma} + \delta\hat{\tilde{Q}}}$.

Finally also the $\int \mathcal{D}z \int \mathcal{D}\tilde{z}$ integrations can be carried out, giving:

$$\beta \int \mathcal{D}z \int \mathcal{D}\tilde{z} M_s^{\star} = \beta \int \mathcal{D}z \int \mathcal{D}\tilde{z} \frac{1}{2} \frac{\left(\hat{R} + \gamma_{12} \frac{\hat{\tilde{R}} + \sqrt{\hat{\tilde{Q}}}\tilde{z}}{\tilde{\gamma} + \delta\hat{\tilde{Q}}} + \sqrt{\hat{Q}} z\right)^2}{(\gamma + \gamma_{12}) + \delta\hat{Q}} \tag{B.19}$$

$$= \frac{\beta}{2} \frac{\left(\hat{R} + \hat{\tilde{R}} \frac{\gamma_{12}}{\tilde{\gamma} + \delta\hat{\tilde{Q}}}\right)^2 + \left(\frac{\gamma_{12}\sqrt{\hat{\tilde{Q}}}}{\tilde{\gamma} + \delta\hat{\tilde{Q}}}\right)^2 + \hat{Q}}{(\gamma + \gamma_{12}) + \delta\hat{Q}}. \tag{B.20}$$

So, specialising to the two terms that appear in the free-entropy we get:

$$g_S(\gamma_1, \gamma_2, \gamma_{12}) = \rho\, g_s\left(\tilde{R}, R, \tilde{Q}_r, Q_r\right) + (1-\rho)\, g_s\left(0, 0, \tilde{Q}_i, Q_i\right)$$
$$= \frac{\beta}{2} \left(\rho \frac{\left(\hat{R} + \hat{\tilde{R}} \frac{\gamma_{12}}{\gamma_1 + \delta\hat{\tilde{Q}}_r}\right)^2 + \left(\frac{\gamma_{12}\sqrt{\hat{\tilde{Q}}_r}}{\gamma_1 + \delta\hat{\tilde{Q}}_r}\right)^2 + \hat{Q}_r}{(\gamma_2 + \gamma_{12}) + \delta\hat{Q}_r} + (1-\rho) \frac{\left(\frac{\gamma_{12}\sqrt{\hat{\tilde{Q}}_i}}{\gamma_1 + \delta\hat{\tilde{Q}}_i}\right)^2 + \hat{Q}_i}{(\gamma_2 + \gamma_{12}) + \delta\hat{Q}_i}\right). \tag{B.21}$$

Energetic term. Since one of the two energetic terms appearing in the replicated free-energy depends on the $\tilde{n}$ replicas of the first weight configuration, and there is no interaction, we can take the $\tilde{n} \to 0$ limit directly. Therefore we only have to evaluate the other contribution (dependent on the n replicas of the second weight configuration). Defining $Q = Q_r + \Delta Q_i$, $Q = Q_r + \Delta Q_i$, we evaluate $g_E = \lim_{n\to 0} \frac{1}{n} \log(G_E)$ in the RS ansatz:

$$g_E = \lim_{n\to 0} \frac{1}{n} \log \int \frac{\mathrm{d}u\mathrm{d}\hat{u}}{2\pi} e^{iu\hat{u}} e^{-\frac{\rho}{2}(\hat{u})^2} \int \prod_a \left(\frac{d\lambda^a \mathrm{d}\hat{\lambda}^a}{2\pi} \mathrm{e}^{\mathrm{i}\lambda^a\hat{\lambda}^a}\right)$$
$$\times\, e^{-\frac{1}{2}(Q-q)\sum_a \left(\hat{\lambda}_a\right)^2 - \frac{1}{2}q\left(\sum_a \hat{\lambda}_a\right)^2 - \hat{u}R\sum_a \hat{\lambda}_a - \beta\sum_a \ell(u,\lambda^a)} \tag{B.22}$$

$$= \lim_{n\to 0} \frac{1}{n} \log \int \frac{\mathrm{d}u}{\sqrt{2\pi\rho}} \int \prod_a \left(\frac{d\lambda^a \mathrm{d}\hat{\lambda}^a}{2\pi} \mathrm{e}^{\mathrm{i}\lambda^a\hat{\lambda}^a}\right)$$
$$\times\, e^{-\frac{1}{2}(Q-q)\sum_a \left(\hat{\lambda}_a\right)^2 - \frac{1}{2}q\left(\sum_a \hat{\lambda}_a\right)^2 - \beta\sum_a \ell(u,\lambda^a) - \frac{1}{2\rho}\left(u + i\, R\sum_a \hat{\lambda}_a\right)^2} \tag{B.23}$$

$$= \lim_{n\to 0} \frac{1}{n} \log \int \mathcal{D}z_0 \int \mathcal{D}u \left\{ \int \mathcal{D}\lambda \mathrm{e}^{-\beta\, \ell\left(\sqrt{\rho}\, u, \sigma\left(\sqrt{(Q-q)}\lambda + \sqrt{q - \frac{m^2}{\rho}} z_0 + \frac{m}{\sqrt{\rho}} u\right)\right)} \right\}^n$$
$$= \int \mathcal{D}z_0 \int \mathcal{D}u \log \int \mathcal{D}\lambda \mathrm{e}^{-\beta\, \ell\left(\sqrt{\rho}\, u, \sigma\left(\sqrt{Q-q}\lambda + \sqrt{q - \frac{m^2}{\rho}} z_0 + \frac{m}{\sqrt{\rho}} u\right)\right)}. \tag{B.24}$$

So in the $\beta \to \infty$ limit, with the proper rescalings, we get:

$$g_E = \beta \int \mathcal{D}z \int \mathcal{D}u \, M_E^\star, \tag{B.25}$$

where:

$$M_E^\star = \max_\lambda - \frac{\lambda^2}{2} - \ell\left(\mathrm{sign}(\sqrt{\rho}\, u), \sigma\left(\sqrt{\delta Q_r + \Delta \delta Q_i}\lambda + \sqrt{Q_r + \Delta Q_i - \frac{R^2}{\rho}} z + \frac{R}{\sqrt{\rho}} u \right)\right). \tag{B.26}$$

RS free-entropy. Finally, assuming the we can write down the RS free-entropy for the curriculum ordering as:

$$\begin{aligned}\Phi/\beta = -\mathrm{extr}\bigg(\hat{R}R + \frac{1}{2}\Big(\big(\hat{Q}\delta Q - \delta\hat{Q}Q\big)_r + \big(\hat{Q}\delta Q - \delta\hat{Q}Q\big)_i \Big)\bigg) \\ + \; g_S(\gamma_1, \gamma_2, \gamma_{12}) + \alpha_2 \, g_E(\Delta_2),\end{aligned} \tag{B.27}$$

where g_S is defined in equation (B.21) and g_E is defined in equation (B.25). The order parameters for the teacher system are obtained independently from identical equations, after substituting $\lambda_1 \to 0, \lambda_2 =\!\!\rightarrow \lambda_1$ and $\lambda_{12} \to 0$, $\alpha_2 \to \alpha_1$ and $\Delta_2 \to \Delta_1$, and after adding a tilde to the remaining parameters.

The saddle-point equations, yielding at convergence the asymptotic prediction for the order parameters, can be found by posing stationarity conditions for the free-entropy with respect to all overlaps.

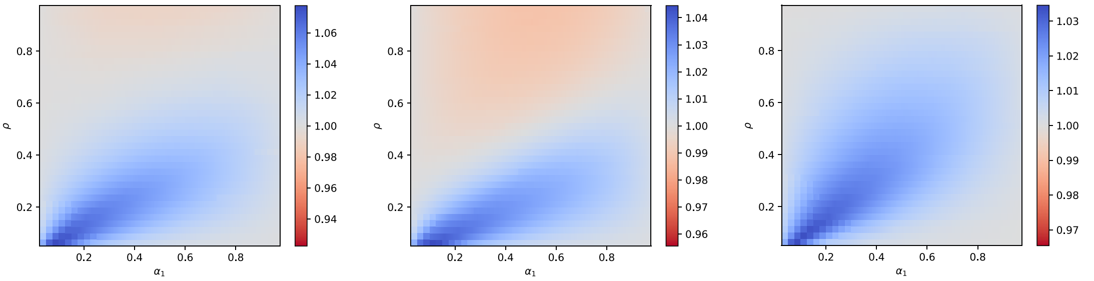
Note that, if instead of the simple setting just considered, where the data slice in the second stage has homogeneous variance for the irrelevant components, there are multiple subsets with different sizes and variances, the only variation in the free-entropy is in the energetic contribution. In general one will have a sum:

$$\sum_s \alpha_s g_E(\Delta_s) \tag{B.28}$$

over each of these subsets.

Moreover, if instead of two stages we consider multiple learning stages, the free-entropy for each successive step has an identical form, and one only has to substitute the tilde parameters with the order parameters obtained at the previous step. Note that the simplicity of nesting stages in this problem is connected to the convexity of this learning setting. Generally, adding more steps would increase the complexity of the calculation considerably.

Generalisation error. With the saddle-point values for the order parameters, one can easily evaluate the generalisation error on new datapoints, which is the measure of performance we are employing in the main. This performance can be obtained as:



(a) Curriculum learning. (b) Anti-curriculum learning. (c) Curriculum vs anti-curriculum.

Figure C.1. Effect of sparsity. Phase diagram on the effect of sparsity, figure 4(b), extended for all learning protocols.

$$1 - \epsilon_{\mathrm{g}} = \left\langle \Theta\left(\left(\frac{\boldsymbol{W}_T \cdot x}{\sqrt{N}}\right)\left(\frac{\boldsymbol{W}_2 \cdot x}{\sqrt{N}}\right)\right)\right\rangle_{x(\Delta)} \tag{B.29}$$

where Δ is the variance of the irrelevant components for the new pattern. A shortcut for evaluating this expression is to insert the order parameters in the expression through Dirac's δs. After a straightforward calculation, along the same lines of the one presented above, one obtains:

$$\epsilon_{\mathrm{g}} = \frac{1}{\pi} \arccos\left(\frac{R}{\sqrt{\rho(Q_r + \Delta Q_i)}}\right). \tag{B.30}$$

Of course, the generalisation accuracy is just the complementary quantity $1 - \epsilon_{\mathrm{g}}$.

Appendix C. Additional results on sparsity

We complement the discussion on the importance of sparsity, section 4, with the comparison with other learning protocols. Observe that anti-curriculum suffers the same issue of the curriculum method for sufficiently large fractions of relevant features ρ. In that regime, the splitting becomes sub-optimal because the solution found in the splitting does not provide enough information to help the other phase of learning. Consequently, the network is forced to set neglect the information in the batch in favour of exploring solutions further away from that one. This is outperform by standard learning, where all the bits of information are used (figure C.1).

Appendix D. Simulations on CIFAR10

Task design. Because a sparse set of relevant features is crucial to observing curriculum effects in our model, we created a task based on real data that has this property. In particular we create 32 × 64 pixel input examples by concatenating two images side-by-side from the CIFAR10 dataset (figure ??). The correct output label is given by the label of the image on the left, while the image on the right is an irrelevant distractor. To vary difficulty, we scale the contrast of the irrelevant image. This dataset is meant to instantiate a simple example of learning an object classification amidst clutter. We emphasise that, as in our synthetic data model, each training sample always contains the same relevant and distractor images (i.e., we are not considering a data augmentation setting where each relevant image appears with many non-relevant images). To ensure no cross-contamination of training and testing samples, the distractor images for the training and test sets are drawn only from the same set.

Model architecture and training regime. We train a single layer network with cross entropy loss (i.e. softmax regression), implemented in Pytorch Lightning by modifying the MIT-licensed `PyTorch_CIFAR10` repository (https://zenodo.org/record/4431043#.YLmz6zZKhsA) to ensure that training parameters accord with standard practice. Networks were trained with SGD and Nesterov momentum, under default parameters: a learning rate of $1e-2$, momentum parameter 0.9, batch size 256, and 100 epochs. The learning rate was annealed according to the 'WarmUpCosine' schedule used in `PyTorch_CIFAR10`, which linearly reduces the learning rate over the first 30% of training steps before switching to a cosine shaped schedule on the remainder.

Experiment details and hyperparameter optimisation. For the first phase of training, we used dataset sizes in 10 equal steps between 1000 and 50 000. For the second phase, we used nine dataset sizes in 9 equal steps between 5333 and 48 000. We optimised hyperparameters in each phase separately. In the first phase, we evaluated all combinations of initialisation scales of $\{0, 0.2, 0.5, 1.0\}$, weight decay parameters of $\{0, 0.2, 0.5, 1.0, 2.0\}$, and curriculum policy, for five random seeds. In the second phase, for each random seed and curriculum condition, we continued training from the best-performing model obtained in the first phase. We trained all combinations of five elastic penalties log spaced between $1e-3$ and $1e2$, and weight decay parameters $\{0, 0.2, 0.5\}$. We then compute the best performing model for each seed and take the mean over seeds. Finally, to evaluate the no-curriculum performance, we train shuffled dataset models with initialisation scales $\{0, 0.2, 0.5, 1.0\}$ and weight decay parameters $\{0, 0.2, 0.5\}$. For visualisation purposes, we used nearest-neighbors interpolation in the phase portrait to provide values for all points used in the synthetic experiments. Experiments were run on V100 GPUs and required approximately 10 000 GPU hours (including debugging and development), or ≈1110 kg CO_2 equation according to the MachineLearning Impact calculator of Lacoste *et al*, 2019.

Appendix E. Speed-up theory vs simulations

As remarked in the main text, one of the advantages of the theoretical analysis is a huge speed-up in the time to collect the results, without need of averaging to reduce the fluctuations. In this section, we briefly report a comparison between the time required for the lines from theory and simulations shown in the main text.

In order to obtain figure 1(c), a single run of the ODE equations takes 2 milliseconds and a run of the simulations takes 500 milliseconds. The figure is however obtained optimizing over all the hyperparameters (learning rate, initialization, weight decay) totalling 400 milliseconds for the analytical solution; while, due to noise, simulation results for a single set of hyperparameters requires averaging 5000 realizations totalling 41 min. We note that we did the hyperparameter optimization only once using the theoretical framework and then used the optima in the simulations in order to save compute time. The best comparison should therefore be done for a fixed set of hyperparameters and gives 2 milliseconds vs 41 min. Overall, the analytical solution is between 2 and 6 orders of magnitude faster.

References

[1] Lawrence D H 1952 The transfer of a discrimination along a continuum *J. Comparat. Physiol. Psychol.* **45** 511
[2] Baker R A and Osgood S W 1954 Discrimination transfer along a pitch continuum *J. Exp. Psychol.* **48** 241
[3] Elio R and Anderson J 1984 The effects of information order and learning mode on schema abstraction *Memory Cognition* **12** 20–30
[4] Wilson R C, Shenhav A, Straccia M and Cohen J D 2019 The eighty five percent rule for optimal learning *Nat. Commun.* **10** 4646
[5] Avrahami J, Kareev Y, Bogot Y, Caspi C, Dunaevsky S and Lerner S 1997 Teaching by examples: implications for the process of category acquisition *Q. J. Exp. Psychol.* A **50** 586–606
[6] Pashler H and Mozer M C 2013 When does fading enhance perceptual category learning? *J. Exp. Psychol.* **39** 1162–73
[7] Hornsby A N and Love B C 2014 Improved classification of mammograms following idealized training *J. Appl. Res. Memory Cognit.* **3** 72–6
[8] Roads B D, Xu B, Robinson J K and Tanaka J W 2018 The easy-to-hard training advantage with real-world medical images *Cogn. Res. Principles Implications* **3** 1–13
[9] The International Brain Laboratory *et al* 2021 Standardized and reproducible measurement of decision-making in mice *eLife* **10** e63711
[10] Elman J L 1993 Learning and development in neural networks: the importance of starting small *Cognition* **48** 71–99
[11] Krueger K A and Dayan P 2009 Flexible shaping: how learning in small steps helps *Cognition* **110** 380–94
[12] Bengio Y, Louradour J, Collobert R and Weston J 2009 Curriculum learning *Proc. 26th Annual Int. Conf. Machine Learning* pp 41–8
[13] Pentina A, Sharmanska V and Lampert C H 2015 *Curriculum Learning of Multiple Tasks* (IEEE Computer Society) pp 5492–500
[14] Hacohen G and Weinshall D 2019 On the power of curriculum learning in training deep networks *ICML* vol 97 (PMLR) pp 2535–44
[15] Wu X, Dyer E and Neyshabur B 2020 When do curricula work? *ICLR*
[16] Brown T *et al* 2020 Language models are few-shot learners *Adv. Neural Inf. Process. Syst.* **33** 1877–901
[17] Jiang M, Grefenstette E and Rocktäschel T 2021 Prioritized level replay (arXiv:2010.03934)
[18] Weinshall D, Cohen G and Amir D 2018 Curriculum learning by transfer learning: theory and experiments with deep networks *Int. Conf. Machine Learning* (PMLR) pp 5238–46
[19] Weinshall D and Amir D 2020 Theory of curriculum learning, with convex loss functions *J. Mach. Learn. Res.* **21** 1–19
[20] Ruiz-García M, Liu A J and Katifori E 2019 Tuning and jamming reduced to their minima *Phys. Rev.* E **100** 052608
[21] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* vol 9 (Singapore: World Scientific)
[22] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[23] Zdeborová L and Krzakala F 2016 Statistical physics of inference: thresholds and algorithms *Adv. Phys.* **65** 453–552

[24] Bahri Y, Kadmon J, Pennington J, Schoenholz S S, Sohl-Dickstein J and Ganguli S 2020 Statistical mechanics of deep learning *Ann. Rev. Condens. Matter Phys.* **11** 501–28

[25] Cugliandolo L F and Kurchan J 1993 Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model *Phys. Rev. Lett.* **71** 173

[26] Biehl M and Schwarze H 1995 Learning by on-line gradient descent *J. Phys. A: Math. Gen.* **28** 643

[27] Advani M S, Saxe A M and Sompolinsky H 2020 High-dimensional dynamics of generalization error in neural networks *Neural Netw.* **132** 428–46

[28] Goldt S, Advani M, Saxe A M, Krzakala F and Zdeborová L 2019 Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett (New York: Curran Associates)

[29] Mannelli S S, Krzakala F, Urbani P and Zdeborova L 2019 Passed and spurious: descent algorithms and local minima in spiked matrix-tensor models *ICML 2019* vol 97 pp 4333–42

[30] Mannelli S S, Biroli G, Cammarota C, Krzakala F and Zdeborová L 2019 Who is afraid of big bad minima? Analysis of gradient-flow in spiked matrix-tensor models *Advances in Neural Information Processing Systems* pp 8679–89

[31] Mannelli S S, Biroli G, Cammarota C, Krzakala F, Urbani P and Zdeborová L 2020 Complex dynamics in simple neural networks: understanding gradient flow in phase retrieval *Advances in Neural Information Processing Systems* vol 33 pp 3265–74

[32] Cui H, Saglietti L and Zdeborová L 2020 Large deviations for the perceptron model and consequences for active learning *Mathematical and Scientific Machine Learning* (PMLR) pp 390–430

[33] Zenke F, Poole B and Ganguli S 2017 Continual learning through synaptic intelligence *Int. Conf. Machine Learning* (PMLR) pp 3987–95

[34] Kirkpatrick J *et al* 2017 Overcoming catastrophic forgetting in neural networks *Proc. Natl Acad. Sci. USA* **114** 3521–6

[35] Saad D and Solla S A 1995 Exact solution for on-line learning in multilayer neural networks *Phys. Rev. Lett.* **74** 4337

[36] Kocmi T and Bojar O 2017 Curriculum learning and Minibatch bucketing in neural machine translation *Proc. Int. Conf. Recent Advances in Natural Language Processing, RANLP* (Varna, Bulgaria September 2017) (INCOMA) pp 379–86

[37] Schneider N, Hovy D, Johannsen A and Carpuat M 2016 Semeval-2016 task 10: detecting minimal semantic units and their meanings (dimsum) *Proc. 10th Int. Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016* ed S Bethard, D M Cer, M Carpuat, D Jurgens, P Nakov and (San Diego, CA, 16–17 June 2016) (The Association for Computer Linguistics) pp 546–59

[38] Zhang X, Shapiro P, Kumar G, Paul McN, Carpuat M and Duh K 2019 Curriculum learning for domain adaptation in neural machine translation *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers)* (Minneapolis, MN) (Association for Computational Linguistics) pp 1903–15

[39] Zenke F, Poole B and Ganguli S 2017 Continual learning through synaptic intelligence *Proc. 34th Int. Conf. Machine Learning, Vol 70 of Proc. Machine Learning Research* ed D Precup and Y Whye Teh (Sydney, Australia 6–11 Aug 2017) (International Convention Centre PMLR) pp 3987–95

[40] Franz S and Parisi G 1997 Phase diagram of coupled glassy systems: a mean-field study *Phys. Rev. Lett.* **79** 2486

[41] Saglietti L and Zdeborová L 2020 Solvable model for inheriting the regularization through knowledge distillation (arXiv:CoRR,abs/2012.00194)

[42] Clerkin E M, Hart E, Rehg J M, Chen Y and Smith L B 2017 Real-world visual statistics and infants first-learned object names *Phil. Trans. R. Soc. London* B **372**

[43] Liu E H, Mercado E III, Church B A and Orduña I 2008 The easy-to-hard effect in human (homo sapiens) and rat (rattus norvegicus) auditory identification *J. Compar. Psychol.* **122** 132

[44] Kepple D R, Engelken R and Kanaka R 2022 Curriculum learning as a tool to uncover learning principles in the brain *ICLR*

[45] Plunkett K, Marchman V and Knudsen S 1991 From rote learning to system building: acquiring verb morphology in children and connectionist nets *Connectionist Models* (Amsterdam: Elsevier) pp 201–19

[46] Plunkett K and Marchman V 1991 U-shaped learning and frequency effects in a multi-layered perception: implications for child language acquisition *Cognition* **38** 43–102

[47] Karmazyn Raz H, Abney D H, Crandall D, Chen Y and Smith L B 2019 How do infants start learning object names in a sea of clutter? *Annual Conf. Cognitive Science Society* pp 521–6

[48] Smith L B and Slone L K 2017 A developmental approach to machine learning? *Front. Psychol.* **8**

[49] Yu C and Smith L B 2012 Embodied attention and word learning by toddlers *Cognition* **125** 244–62

[50] Krizhevsky A 2009 *Learning Multiple Layers of Features from Tiny Images*

[51] Orduña I, Liu E H, Church B A, Eddins A C and Mercado E III 2012 Evoked-potential changes following discrimination learning involving complex sounds *Clin. Neurophysiol.* **123** 711–9

[52] Church B A, Mercado E III, Wisniewski M G and Liu E H 2013 Temporal dynamics in auditory perceptual learning: impact of sequencing and incidental learning *J. Exp. Psychol.* **39** 270

[53] Ruiz-Garcia M, Zhang G, Schoenholz S S and Liu A J 2021 Tilting the playing field: dynamical loss functions for machine learning *Int. Conf. Machine Learning* (PMLR) pp 9157–67