The Abdus Salam
International Centre
for Theoretical Physics

Youth in High-Dimensions | (SMR 3602)

15 Jun 2021 - 18 Jun 2021
Virtual, Virtual, Italy

P01 - ARORA Akhil

Low-rank Subspaces for Unsupervised Entity Linking

P02 - DE BRUYNE Benjamin

Generating constrained random walks

P03 - DOIMO Diego

Hierarchical nucleation in deep neural networks

P04 - DUTT Arkopal

Efficient Learning of Ising Models from Glauber Dynamics

P05 - ENGELKEN Rainer

Input correlations impede suppression of chaos and learning in balanced rate networks

P06 - IBRAHIM Maniru

Moving Least Squares Deformations and Applications in Multidimensional Projections

P07 - JAYAKUMAR Abhijith

Learning of Discrete Graphical Models with Neural Networks

P08 - MACOCCO Iuri

Intrinsic dimension estimator for discrete features datasets

P09 - PERDOMO ESCOBAR Alejandro Daniel

Tensor Networks for Text Classifications

P10 - PEREIRA Joao

Method of Moments using an efficient tensor decomposition algorithm

P11 - PETRINI Leonardo

Relative stability toward diffeomorphisms in deep nets indicates performance

P12 - RATHI Lata Dr. Manju

PCA and Regression Approach to Identify the Energy Efficiency Parameters of Cloud Environment

P13 - SAGLIETTI Luca

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

P14 - SARAO MANNELLI Stefano

Probing transfer learning with a model of synthetic correlated datasets

P15 - TOLEDO MARIN Quetzalcoatl Javier

Using Deep LSD to build operators in GANs latent space with meaning in real space

P16 - VILIMELIS ACEITUNO Pau

Minimizing costs in neuromorphic systems with m-of-n codes

P17 - WILINSKI Mateusz

Scalable Learning of Independent Cascade Dynamics from Partial Observations

P18 - ZAVATONE-VETH Jacob

Exact priors of finite neural networks

P01

Low-rank Subspaces for Unsupervised Entity Linking

Entity linking is an important problem with many applications. Most previous solutions were designed for settings where annotated training data is available, which is, however, not the case in numerous domains. We propose a light-weight and scalable entity linking method, Eigenthemes, that relies solely on the availability of entity names and a referent knowledge base. Eigenthemes exploits the fact that the entities that are truly mentioned in a document (the "gold entities") tend to form a semantically dense subset of the set of all candidate entities in the document. Geometrically speaking, when representing entities as vectors via some given embedding, the gold entities tend to lie in a low-rank subspace of the full embedding space. Eigenthemes identifies this subspace using the singular value decomposition and scores candidate entities according to their proximity to the subspace. On the empirical front, we introduce multiple strong baselines that compare favorably to the existing state of the art. Extensive experiments on benchmark datasets from a variety of real-world domains showcase the effectiveness of our approach.

P02

Generating constrained random walks

I will present a method to exactly generate bridge trajectories for discrete-time random walks, with arbitrary jump distributions, that are constrained to initially start at the origin and return to the origin after a fixed time. The method is based on an effective jump distribution that implicitly accounts for the bridge constraint. It is illustrated on various jump distributions and is shown to be very efficient in practice.

P03

Hierarchical nucleation in deep neural networks

Deep convolutional networks (DCNs) learn meaningful representations where data that share the same abstract characteristics are positioned closer and closer. Understanding these representations and how they are generated is of unquestioned practical and theoretical interest. In this work we study the evolution of the probability density of the ImageNet dataset across the hidden layers in some state-of-the-art DCNs. We find that the initial layers generate a unimodal probability density getting rid of any structure irrelevant for classification. In subsequent layers density peaks arise in a hierarchical fashion that mirrors the semantic hierarchy of the concepts. Density peaks corresponding to single categories appear only close to the output and via a very sharp transition which resembles the nucleation process of a heterogeneous liquid. This process leaves a footprint in the probability density of the output layer where the topography of the peaks allows reconstructing the semantic relationships of the categories.

P04

Efficient Learning of Ising Models from Glauber Dynamics

Input correlations impede suppression of chaos and learning in balanced rate networks

Cortical circuits exhibit complex activity patterns, both spontaneously and evoked by external stimuli. Information encoding and learning in neural circuits depend on how well time-varying input can control network activity. Previous work showed that input correlations can be detrimental to learning in balanced networks, but the reasons for this were not clear. We show that in firing-rate networks in the balanced state, external control of recurrent dynamics strongly depends on the correlations of the input: one might expect that driving all neurons with a common input helps to control network dynamics. Surprisingly, we find that the network is far easier to control with independent inputs into each neuron. We discover that this discrepancy is explained by the dynamic cancellation of common external input by recurrent feedback - a phenomenon previously described in binary networks in the balanced state. In contrast, no cancellation occurs for inputs independent across neurons, and a much weaker external input is sufficient for control. We present a time-dependent dynamic mean-field theory that explains for threshold-linear networks how these results depend on input frequency, recurrent coupling strength, and network size. In summary, we identified a novel link between recurrent network dynamics and chaos. This link can help to harness the computational capabilities of balanced recurrent circuits for plasticity and learning of stable trajectories. Specifically, we predict that uncorrelated inputs facilitate learning in balanced networks.

P06

Moving Least Squares Deformation and Applications in Multidimensional Projections

Maniru Ibrahim

Department of Mathematics, COMSATS University Islamabad, Pakistan

manirkhalil@gmail.com

Abstract

Multidimensional projections are among the most essential approaches in the visual analysis of multidimensional data. It transforms multidimensional data into two-dimensional representations that may be shown as scatter plots while preserving their similarity with the original data. In this work, we propose to study and improve on a well-known map called LAMP, which takes a multidimensional instance and embeds it in Cartesian space via moving least squares deformation. We propose solving an optimization problem to develop a new metric for evaluating the quality of multidimensional projections that combines three measures: silhouette coefficient, neighborhood preservation, and silhouette ratio, which were previously used to assess LAMP projections. Also, we will focus on trying to overcome a limitation of the method which requires a similar scale for control points and their counterparts in the Cartesian space.

Introduction

Given a set $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}$ with $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a dissimilarity distance function. Let the counterpart of X in the visual space be $Y = \{y_1, y_2, \dots, y_n\} \subset \mathbb{R}^2$ with the distance $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$. A multidimensional projection technique can be describe as a function which $f : X \to Y$ which minimizes $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$.

Multidimensional projection may also be defined using a subset of X called the control points $X_c = \{x_{c_1}, \dots, x_{c_k}\}$ and its counterpart in the visual space $Y_c = \{y_{c_1}, y_{c_2}, \dots, y_{c_k}\} \subset Y$ as a family of functions $f_{x_j} : X \to Y$ that minimizes $\sum_{i=1}^{k} \alpha_i \| f_{x_j}(x_{c_i}) - y_{c_i} \|^2$, where α_i is a scalar weight which depends on x_j.

Multidimensional projection methods using control points have setbacks, as they may produce high computational cost or do not produce effective techniques that allow fully interactive data manipulation. To overcome these problems [3] defined a multidimensional projection called LAMP using orthogonal mapping theory and moving least squares. LAMP was observed to have the ability to accurately map instances using few control points, it is very effective for interactive applications, and is robust and versatile. For LAMP to produce more pleasant layouts, the control points and their images have to be on the same scale.

We are going to focus on two main problems:

1 - a deeper comprehension of how the scale affects LAMP projections;

2 - and propose a new metric to evaluate projections, based on a machine learning approach, that could be employed in an algorithm to find the optimal scale hyperparameter of the LAMP projection.

In this work we tune the scale hyperparameter of the datasets so that a multidimensional algorithm such as LAMP provides high quality projections. To do that, we employ a learning approach, to learn the best weights to linearly combine known measures such as silhouette coefficient, neighborhood preservation, and silhouette ratio.

Quality Metrics

Silhoutte Coefficient

The silhouette coefficient [6] assesses both the cohesiveness and the separation of clustered occurrences. The average distance between instance $x \in \mathcal{D}$ and all other instances in the same group as x is used to compute x'^s cohesiveness a_x. The separation b_x is the shortest distance between x and all other instances of the same cluster. It is given by

$$Silh = \frac{1}{n} \sum_{x \in \mathcal{D}} \frac{(b_x - a_x)}{max(a_x, b_x)}.$$

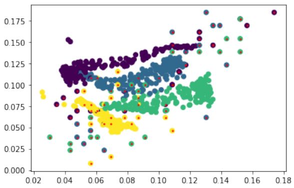


Figure 1: $SilhoutteCoefficient = 0.2505$.

Silhouette Ratio

Sometimes the multidimensional data may have poor clustering, this may affect the silhouette coefficient. Therefore, we define a new measure called the silhouette ratio which is the ratio of the silhouette coefficient compute on the original data to the silhouette coefficient on the projected data.

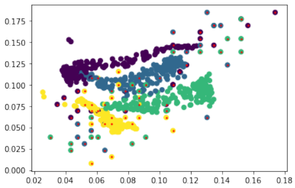


Figure 2: $SilhouetteRatio = 0.9036$.

Neighbourhood Preservation

The neighbourhood preservation metric [3] computes the fraction of an instance's k-nearest neighbours who are still neighbours in the visual space.

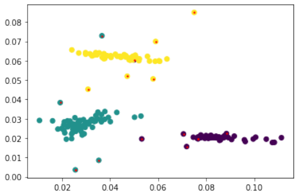


Figure 3: $NeighbourhoodPreservation = 0.1971$.

Proposed Method

Learning a New Measure to Evaluate Multidimensional Projections

To find the new metric, we first define the training dataset $X_{train} = \{x_i\}_{i=1}^{60}$ and its counterpart $Y_{train} = \{y_i\}_{i=1}^{60}$, where $x_i = (m_1^{(i)}, m_2^{(i)}, m_3^{(i)})$, m_1 is the silhouette coefficient, m_2 is the neighbourhood preservation, m_3 is the silhouette ratio, and y_i is the grade of x_i. We need to learn the weight of a regression function $f : \mathbb{R}^3 \to \mathbb{R}$, given by $f_{\bar{w}}(m_1, m_2, m_3) = w_1 m_1 + w_2 m_2 + w_3 m_3$, where $\bar{w} = \{w_1, w_2, w_3\}$. We the minimize the loss function

$$L(w_1, w_2, w_3) = \sum_{i=1}^{60} (f_{\bar{w}}(m_1^{(i)}, m_2^{(i)}, m_3^{(i)}) - y_i)^2 \quad (1)$$

$$= \sum_{i=1}^{60} (w_1 m_1^{(i)} + w_2 m_2^{(i)} + w_3 m_3^{(i)} - y_i)^2, \quad (2)$$

and then obtain the weights and the new metric as follows:

$$w_1 = 5.7097, \quad w_2 = 3.7741, \quad and \quad w_3 = -0.0106.$$

Therefore, our new metric is

$$M_{new} = 5.7097 m_1 + 3.77416 m_2 - 0.0106 m_3.$$

```
Algorithm 1: Hyperparameter tuning algorithm
1: Input: Dataset $D \subset \mathbb{R}^d$, Array $T$, New metric, Scale $[a, b]$ uniformly sample
2: for each scale $s$ in $[a, b]$ do
3:     Project the data set using Lamp
4:     Compute the new metric on the projection
5:     Store the value in the array $T$
6: end for
7: Find the highest value in the array $T$
8: Find the scale corresponding to the position of the highest value in array $T$
```

Results

Training a custom metric

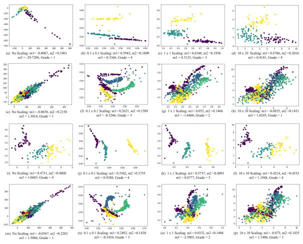


Figure 4: Figures showing projections used to train the new metric.

Visual Evaluation

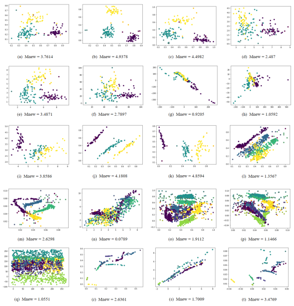


Figure 5: Figures showing projections with the result of the learnt metric.

Cross-validation

Table 1: The statistics of absolute errors of training and testing projections

Dataset	MAE	Median	Standard Deviation
Training	0.5414	0.4398	0.37
Testing	0.5630	0.4893	0.3878

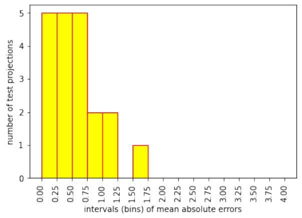


Figure 6: Histogram of Errors.

Conclusions

We developed a new metric for evaluating the impact of scale on the quality of a projection in this paper. In several scenarios, the proposed metric has been found to be very effective. We also show that the scales of the multidimensional dataset have an impact on the quality of the projection. As a result, we built an algorithm for determining the scales that produce the best projection for every given dataset. It was empirically observed that the optimal scale that gives the best projection lies in the interval $[0.1, 1]$.

Forthcoming Research

Another element that need to be look into more is determining the optimal number of neighbours to generate the desired layout. A radius of impact to each control point might be defined as an alternative to the $k-$nearest neighbours technique used in our approach.

References

[1] A. Frank and A. Asuncion. (2010). UCI machine learning repository.

[2] Gower, J., & Dijksterhuis, G. (2004). Procrustes Problems. Oxford University Press.

[3] Joia, P., Coimbra, D., Cuminato, J. Paulovich, F., & Nonato, L. (2011). Local affine multidimensional projection. *IEEE Trans. Vis. Comp. Graph.,* 17, 2563–2571.

[4] Pyae, A. (2019). Fish market: Database of common fish species for fish market.

[5] Schaefer, S., McPhail, T., & Warren, J. (2006). Image deformation using moving least squares. *ACM Trans. Graph.,* 25, 533-540.

[6] Tan, P., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining.

P07

Learning of Discrete Graphical Models with Neural Networks

The inverse problem of learning a discrete graphical model given i.i.d samples from its joint distribution can be solved with near-optimal sample complexity using a convex optimization method known as Generalized Regularized Interaction Screening Estimator (GRISE). But the computational cost of GRISE becomes prohibitive when the energy function of the true graphical model has higher order terms. We introduce NeurISE, a neural net based algorithm for graphical model learning, to tackle this limitation of GRISE. We use neural nets as function approximators in an Interaction Screening objective function. The optimization of this objective then produces a neural-net representation for the conditionals of the graphical model. NeurISE algorithm is seen to be a better alternative to GRISE when the energy function of the true model has a high order with a high degree of symmetry. In these cases NeurISE is able to find the correct parsimonious representation for the conditionals without being fed any prior information about the true model. NeurISE can also be used to learn the underlying structure of the true model with some simple modifications to its training procedure. We also demonstrate that NeurISE and its variants perform well when applied to certain learning problems that arise in quantum information.

P08

Intrinsic dimension estimator for discrete features datasets

Dealing with datasets in very high dimensional space is still a mess, relevant information is hard to extract without loosing important features of the data. Many approaches of dimensional reduction are capable of making this task a bit easier, however, before using them blindly, one should first have an idea of the intrinsic dimension (id) of the manifold the data are lying on. To this aim, lots of estimators have been proposed in the last few years. Such algorithms -mainly relying on the properties of the distances between the points- manage to perform pretty well on non-trivial dataset. However they have always been thought, built and applied on dataset with continuous features. Dealing with discrete features introduces further difficulties, like distance degeneracy and the unusual structure of the embedding manifold. In order to overcome these problems, I develop a new id estimator based on the relationship between the Poisson process statistics. I first demonstrate its behaviour and properties in the continuum: the potentially arbitrary precision together with the capability of investigating the I'd at different scales. I then extend its definitions to the discrete and test its performance on synthetic datasets, from points defined on a lattice to the simplest spin systems. The estimator was then used on metagenomic datasets and quantum computers states to find out non trivial (and otherwise invisible) properties of such complex systems.

ᴾ09

Tensor Networks for Text Classifications

The applicability of Tensor Networks, more specifically Matrix Product States (MPS), was studied in a supervised learning environment in the case of text classification. The classification model was built considering a mapping of non-linear characteristics of high dimension and a tensor of equal size that was represented as an MPS. As an interesting application we investigated the predictions made by the model on the collections of DBpedia and AG-News.

P10

Method of Moments using an efficient tensor decomposition algorithm

When analyzing a dataset, one of the first tools employed is PCA, which analyzes the first and second moment of the data (mean and covariance). However methods that use information from higher order moments are often disregarded, since these involve constructing tensors which occupy a lot of memory. Additionally, most tensor problems are NP-hard and algorithms for decomposing tensors are either lacking or not reliable. In this poster I will present a recently proposed algorithm for decomposing symmetric tensors, which not only is much faster than the state-of-the-art, it also enjoys a rich mathematical foundation. I also mention ongoing work on how to use this algorithm to decompose moment tensors implicitly, that is, extracting the relevant information directly from the samples without ever explicitly forming the moment tensors.

P11

Relative stability toward diffeomorphisms in deep nets indicates performance

Leonardo Petrini, Alessandro Favero, Mario Geiger, Matthieu Wyart

Institute of Physics
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland
`{name.surname}@epfl.ch`

Abstract

Understanding why deep nets can classify data in large dimensions remains a challenge. It has been proposed that they do so by becoming stable to diffeomorphisms, yet existing empirical measurements support that it is not the case. We revisit this question by defining a maximum-entropy distribution on diffeomorphisms, that allows to study typical diffeomorphisms of a given norm. We confirm that stability toward diffeomorphisms does not correlate to performance on four benchmark data sets of images. By contrast, we find that the *stability toward diffeomorphisms relative to that of generic transformations* R_f correlates remarkably with the test error ϵ_t. It is of order unity at initialization but decreases by several decades during training for state of the art architectures. For CIFAR10 and 15 known architectures we find $\epsilon_t \approx 0.2\sqrt{R_f}$, suggesting that obtaining a small R_f is necessary to achieve good performance. We study how R_f depends on the size of the training set and compare it to a simple model of invariant learning.

Preprint. Under review.

P12

PCA and Regression Approach to Identify the Energy Efficiency Parameters of Cloud Environment

Users can have access to applications, storages, communication, virtualization, collaboration, and infrastructure in an On-Demand basis in Cloud Computing Environment. With the continuous high growth and market share, Cloud computing is one of the hottest topics for the IT industry in general and specifically the energy usage scenarios. Cloud Computing users are increasing at a faster pace with a large number of organizations shifting to the cloud based services. However with the large number of deployments, the size of the cloud data centers is increasing across the globe. Though the increasing use of cloud computing and the resulting rise in the number of datacenters and hosting centers have brought forth many concerns regarding the energy consumption patterns of these data centers and the deployments. Present work has been dedicated to analyse the various scenarios and parameters referring to the energy efficiency of the Cloud Computing Environment. GreenCloud simulator has been chosen to simulate the cloud scenarios in the present work. This is an open source free software and can run through the web based interface. A set of 500 observations have been carried out by changing the values for 24 input parameters. Statistical analysis has been carried out to identify the significant parameters of energy efficiency. Multiple Regression and Principal Component Analysis models have been used for this purpose and the results have been reported.

P13

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

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

[1]SPOC laboratory, EPFL, Switzerland. [2]Gatsby Computational Neuroscience Unit, UCL, UK.

1. Abstract

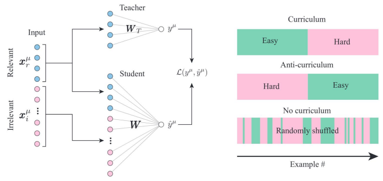
In humans and animals, curriculum learning - presenting data in a curated order - is critical to rapid learning and effective pedagogy. Yet in machine learning, curricula are not widely used and empirically often yield only moderate benefits. This stark difference in the importance of curriculum raises a fundamental theoretical question: when and why does curriculum learning help?

We analyse a prototypical neural network model of curriculum learning in the high-dimensional limit, employing statistical physics methods. Curricula could in principle change both the learning speed and asymptotic performance of a model.

- We provide an exact description of the online learning setting, confirming that curricula can modestly speed up learning.
- We derive the asymptotic performance in a batch learning setting, in which a network trains to convergence in successive phases. By connecting different learning phases through simple Gaussian priors, curriculum can yield a large improvement in test performance.

Our reduced analytical descriptions help reconcile apparently conflicting empirical results and trace regimes where curriculum learning yields the largest gains. More broadly, our results suggest that fully exploiting a curriculum may require explicit changes to the loss function at curriculum boundaries.

2. Model



We revisit a prototypical model of curriculum learning from [2] that entails a simple teacher-student setup, where teacher and student are shallow 1-layer neural networks (perceptrons). The learning task is a binary classification problem.

A key feature of this model is that the teacher network is sparse, with only a fraction **ρ**<1 of Gaussian non-zero components. Therefore, in order to achieve a good test accuracy, the student has to learn which components should be set to zero and align the relevant weights in the correct direction.

We model difficulty of a datum **x** defining *relevant* and *irrelevant* components, where the (ir)relevant components are those (not) used by the teacher for labelling. The entries of **x** are i.i.d. Gaussian with variance 1 and **Δ** for the relevant and irrelevant components respectively. **Δ** characterize the difficulty of each data point.

3. Online dynamics

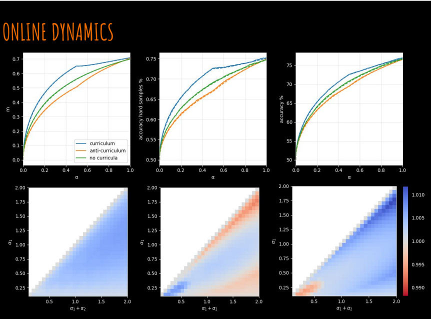
Using the tools from statistical physics [3,4] we study the online SGD dynamics on a square loss

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

The dynamics can be analysed tracking the evolution of few order parameters, in the high-dimensional limit the analysis leads to exact analytical equations for the updates at all times.

The upper panels show simulations for $\alpha_1=\alpha_2=0.5$, $\Delta_1=0$, $\Delta_2=1$ at the optimal learning rate, weight decay and initialization. In the order we have: the cosine between teacher and student, the accuracy on hard instances, the overall accuracy.

The phase diagrams below show the ratio between accuracy on hard instances of two protocols; in the order: curriculum vs no curricula, anti-curriculum vs baseline, curriculum vs anti-curriculum.
The curriculum strategy leads, not only to a dynamical advantage, but also to small final improvement in most of the phase diagram.

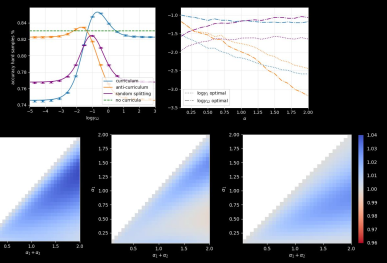


4. Batch learning

Employing again the tools of statistical physics [5,6] we consider the final generalization error when the networks can revisit all the samples. Due to the convex nature of the problem, the last stage of learning always dominate the landscape, therefore we propose a modification of the loss introducing elastic couplic to the previous stages. The coupling parameter is γ_{12}.

The results for $\alpha_1=\alpha_2=1$, $\Delta_1=0$, $\Delta_2=1$ shown in the upper-left figure show that curriculum learning, at the optimal coupling, achieve the best results followed by anti-curriculum and the baseline. The right panels shows the optimal γ_{12} and γ increasing the dataset size. The figure confirms the intuition that a positive coupling is always beneficial, since it is the only way of preserving the information gained in the former learning stage.

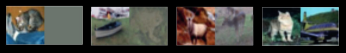
The phase diagrams below represent the same information as for the online dynamics and show that the strategy rank is preserved exploring the phase diagram of possible partitioning.

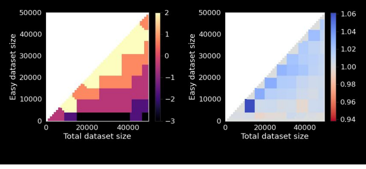


5. Rethinking curriculum learning

We tested curriculum learning with elastic coupling on a real dataset. We build easy and hard samples on the CIFAR10 dataset by attaching two images and using the labels of the left-ones. The right-images act as distractors and the difficulty is given by their brightness. Some examples are shown on the right.
A shallow network is trained on a logistic loss on the task, using both curriculum learning and standard learning. All training parameters are optimized.

The ratio curriculum/standard of the final accuracy on hard samples is shown on the right. The results is consistent with what observed in batch learning.
On the left we show the optimal log γ_{12} for the difference sizes, we see that (consistently with the expectations) a stronger coupling helps when a higher fraction of easy examples are present.





Contacts

luca.saglietti@gmail.com
stefano.sarao@gmail.com

References

[1] Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe. An Analytical Theory of Curriculum Learning in Teacher-Student Networks. Submitted arXiv:2106.?????.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

[3] Michael Biehl and Holm Schwarze. Learning by on-line gradient descent.Journal of Physics A: Mathematical and general, 28(3):643, 1995.

[4] David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. Physical Review Letters, 74(21):4337, 1995.

[5] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro.Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications, volume 9. World Scientific Publishing Company, 1987.

[6] Luca Saglietti and Lenka Zdeborová. Solvable model for inheriting the regularization through knowledge distillation.CoRR, abs/2012.00194, 2020.

P14

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

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

[1]SPOC laboratory, EPFL, Switzerland. [2]Gatsby Computational Neuroscience Unit, UCL, UK.

1. Abstract

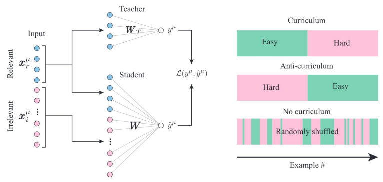
In humans and animals, curriculum learning - presenting data in a curated order - is critical to rapid learning and effective pedagogy. Yet in machine learning, curricula are not widely used and empirically often yield only moderate benefits. This stark difference in the importance of curriculum raises a fundamental theoretical question: when and why does curriculum learning help?

We analyse a prototypical neural network model of curriculum learning in the high-dimensional limit, employing statistical physics methods. Curricula could in principle change both the learning speed and asymptotic performance of a model.

- We provide an exact description of the online learning setting, confirming that curricula can modestly speed up learning.
- We derive the asymptotic performance in a batch learning setting, in which a network trains to convergence in successive phases. By connecting different learning phases through simple Gaussian priors, curriculum can yield a large improvement in test performance.

Our reduced analytical descriptions help reconcile apparently conflicting empirical results and trace regimes where curriculum learning yields the largest gains. More broadly, our results suggest that fully exploiting a curriculum may require explicit changes to the loss function at curriculum boundaries.

2. Model



We revisit a prototypical model of curriculum learning from [2] that entails a simple teacher-student setup, where teacher and student are shallow 1-layer neural networks (perceptrons). The learning task is a binary classification problem.

A key feature of this model is that the teacher network is sparse, with only a fraction **ρ**<1 of Gaussian non-zero components. Therefore, in order to achieve a good test accuracy, the student has to learn which components should be set to zero and align the relevant weights in the correct direction.

We model difficulty of a datum **x** defining *relevant* and *irrelevant* components, where the (ir)relevant components are those (not) used by the teacher for labelling. The entries of **x** are i.i.d. Gaussian with variance 1 and **Δ** for the relevant and irrelevant components respectively. **Δ** characterize the difficulty of each data point.

3. Online Dynamics

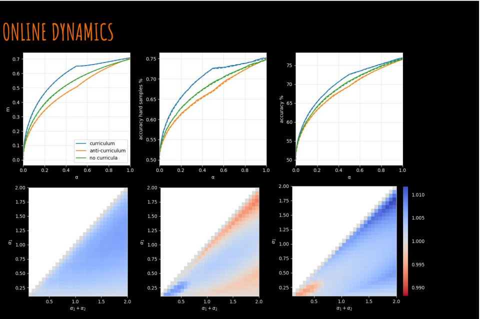
Using the tools from statistical physics [3,4] we study the online SGD dynamics on a square loss

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

The dynamics can be analysed tracking the evolution of few order parameters, in the high-dimensional limit the analysis leads to exact analytical equations for the updates at all times.

The upper panels show simulations for $\alpha_1=\alpha_2=0.5, \Delta_1=0, \Delta_2=1$ at the optimal learning rate, weight decay and initialization. In the order we have: the cosine between teacher and student, the accuracy on hard instances, the overall accuracy.

The phase diagrams below show the ratio between accuracy on hard instances of two protocols; in the order: curriculum vs no curricula, anti-curriculum vs baseline, curriculum vs anti-curriculum.
The curriculum strategy leads, not only to a dynamical advantage, but also to small final improvement in most of the phase diagram.

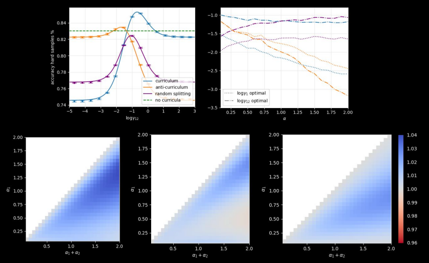


4. Batch Learning

Employing again the tools of statistical physics [5,6] we consider the final generalization error when the networks can revisit all the samples. Due to the convex nature of the problem, the last stage of learning always dominate the landscape, therefore we propose a modification of the loss introducing elastic couplic to the previous stages. The coupling parameter is γ_{12}.

The results for $\alpha_1=\alpha_2=1, \Delta_1=0, \Delta_2=1$ shown in the upper-left figure show that curriculum learning, at the optimal coupling, achieve the best results followed by anti-curriculum and the baseline. The right panels shows the optimal γ_{12} and γ increasing the dataset size. The figure confirms the intuition that a positive coupling is always beneficial, since it is the only way of preserving the information gained in the former learning stage.

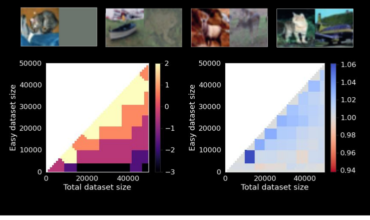
The phase diagrams below represent the same information as for the online dynamics and show that the strategy rank is preserved exploring the phase diagram of possible partitioning.



5. Rethinking Curriculum Learning

We tested curriculum learning with elastic coupling on a real dataset. We build easy and hard samples on the CIFAR10 dataset by attaching two images and using the labels of the left-ones. The right-images act as distractors and the difficulty is given by their brightness. Some examples are shown on the right.
A shallow network is trained on a logistic loss on the task, using both curriculum learning and standard learning. All training parameters are optimized.

The ratio curriculum/standard of the final accuracy on hard samples is shown on the right. The results is consistent with what observed in batch learning.
On the left we show the optimal log γ_{12} for the difference sizes, we see that (consistently with the expectations) a stronger coupling helps when a higher fraction of easy examples are present.



Contacts

luca.saglietti@gmail.com
stefano.sarao@gmail.com

References

[1] Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe. An Analytical Theory of Curriculum Learning in Teacher-Student Networks. Submitted arXiv:2106.?????.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

[3] Michael Biehl and Holm Schwarze. Learning by on-line gradient descent.Journal of Physics A: Mathematical and general, 28(3):643, 1995.

[4] David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. Physical Review Letters, 74(21):4337, 1995.

[5] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro.Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications, volume 9. World Scientific Publishing Company, 1987.

[6] Luca Saglietti and Lenka Zdeborová. Solvable model for inheriting the regularization through knowledge distillation.CoRR, abs/2012.00194, 2020.

ᑭ15

Using Deep LSD to build operators in GANs latent space with meaning in real space

Generative models rely on the key idea that data can be represented in terms of latent variables which are uncorrelated by definition. Lack of correlation is important because it suggests that the latent space manifold is simpler to understand and manipulate. Generative models are widely used in deep learning, e.g., variational autoencoders (VAEs) and generative adversarial networks (GANs). Here we propose a method to build a set of linearly independent vectors in the latent space of a GANs, which we call quasi-eigenvectors. These quasi-eigenvectors have two key properties: i) They span all the latent space, ii) A set of these quasi-eigenvectors map to each of the labeled features one-on-one. We show that in the case of the MNIST, while the number of dimensions in latent space is large by construction, 98% of the data in real space map to a sub-domain of latent space of dimensionality equal to the number of labels. We then show how the quasi-eigenvalues can be used for Latent Spectral Decomposition (LSD), which has applications in denoising images and for performing matrix operations in latent space that map to feature transformations in real space. We show how this method provides insight into the latent space topology. The key point is that the set of quasi-eigenvectors form a basis set in latent space and each direction corresponds to a feature in real space.

P16

Minimizing costs in neuromorphic systems with m-of-n codes

Neuromorphic engineering is one of the most promising technologies in the field of TinyML and Edge AI, leveraging the use of neuroscience-like sensing and computation to reduce energy cost in applications such as autonomous cars or space engineering. The potential of this technology has attracted investment and research from key industrial players such as Intel, IBM or Samsung as well as from many start-ups and academic institutions such as the HBP. However, despite this promise, there is little theoretical work done to explore the fundamental limitations of such approaches. In particular, we do not have a systematic way of finding the minimum energy costs of running those systems in the way that Shannon s Source and Channel theorems allow us to find the minimal code length in classical computing and sensing systems. In this work we propose a framework to decide how many neurons and how much energy should be used. Our approach is to find the minimum values of the total number of neurons and the number of active neurons that allow enough information to be transferred by leveraging fixed weight codes. The results on the noiseless case can be extended to codes with different number of active neurons, which save energy by assigning lower activation values to more frequent codes. On a second step, we add noise to the problem and derive the necessary conditions for achieving decodability almost certainly.

P17

Scalable Learning of Independent Cascade Dynamics from Partial Observations

Spreading processes play an increasingly important role in modeling for diffusion networks, information propagation, marketing and opinion setting. We address the problem of learning of a spreading model such that the predictions generated from this model are accurate and could be subsequently used for the optimization, and control of diffusion dynamics. Unfortunately, full observations of the dynamics are rarely available. As a result, standard approaches such as maximum likelihood quickly become intractable for large network instances. We introduce a computationally efficient algorithm, based on a scalable dynamic message-passing approach, which is able to learn parameters of the effective spreading model given only limited information on the activation times of nodes in the network. We show that tractable inference from the learned model generates a better prediction of marginal probabilities compared to the original model. We develop a systematic procedure for learning a mixture of models which further improves prediction quality of the model.

P18

Exact priors of finite neural networks

Bayesian neural networks are theoretically well-understood only in the infinite-width limit, where Gaussian priors over network weights yield Gaussian priors over network outputs. Recent work has suggested that finite Bayesian networks may outperform their infinite counterparts, but their non-Gaussian output priors have been characterized only though perturbative approaches. Here, we derive exact solutions for the output priors for individual input examples of a class of finite fully-connected feedforward Bayesian neural networks. For deep linear networks, the prior has a simple expression in terms of the Meijer G-function. The prior of a finite ReLU network is a mixture of the priors of linear networks of smaller widths, corresponding to different numbers of active units in each layer. Our results unify previous descriptions of finite network priors in terms of their tail decay and large-width behavior.