

Measure, Manifold, Learning, and Optimization: A Theory Of Neural Networks

Shuai Li *

lishuai918@gmail.com

November 30, 2018

Abstract

We present a formal measure-theoretical theory of neural networks (NN) built on *probability coupling theory*. Our main contributions are summarized as follows.

- Built on the formalism of probability coupling theory, we derive an algorithm framework, named Hierarchical Measure Group and Approximate System (HMGAS), nicknamed S-System, that is designed to learn the complex hierarchical, statistical dependency in the physical world, of which the hierarchical structure is formulated as measure-theoretical assumptions.
- We show that NNs are special cases of S-System when the probability kernels assume certain exponential family distributions. Activation Functions are derived formally. We further endow geometry on NNs through *information geometry*, show that intermediate feature spaces of NNs are stochastic manifolds, and prove that “distance” between samples is contracted as layers stack up.
- S-System shows NNs are inherently stochastic, and under a set of *realistic* boundedness and diversity conditions, it enables us to prove that for *large size nonlinear deep* NNs with a class of losses, including the hinge loss, all local minima are global minima with zero loss errors, and regions around the minima are flat basins where all eigenvalues of Hessians are concentrated around zero, using tools and ideas from *mean field theory*, *random matrix theory*, and *nonlinear operator equations*.
- S-System, the information-geometry structure and the optimization behaviors combined completes the analog between *Renormalization Group* (RG) and NNs. It shows that a NN is a complex adaptive system that estimates the statistic dependency of microscopic object, e.g., pixels, in multiple scales. Unlike clear-cut physical quantity produced by RG in physics, e.g., temperature, NNs renormalize/recompose manifolds emerging through learning/optimization that divide the sample space into highly semantically meaningful groups that are dictated by supervised labels (in supervised NNs).

However, the above contributions describe the theory present in a backward way, for that it is easy for readers to relate to the theory by debriefing what open problems are related. In a logical way, the paper describes *four* parts of the theory of NNs.

- **S-System**, a formal measure-theoretical framework that builds a hierarchical hypothesis space, of which NNs are special cases. It is motivated by the fact that nature is a complex system built hierarchically, and a mechanism is needed for any agents living in it to recognize and predict hierarchical events happening.
- The **geometry** of S-System. The objects in the hypothesis space are probability measures, thus have an information-geometry structure. It characterizes the phenomenon that NNs compose and recompose manifolds that have increasingly high level semantic meaning.
- The **learning framework** of S-System. It describes the objective functions to identify an element of the hypothesis space by learning the parameters. It gives a principled derivation of back propagation, and unifies supervised learning and unsupervised learning in NNs.
- The **optimization** landscape of S-System. It identifies principles and conditions that make the non-convex optimization of the learning problem of S-System benign; that is, all local minima are global minima.

Help is solicited in section 1.3.

*Research done during the stay at South China University of Technology, Chinese University of Hong Kong, Microsoft Research Asia, and University of Science and Technology of China, in reverse chronological order.

Contents

1	Introduction	3
1.1	A Theory of Neural Networks	3
1.1.1	Debrief on the four parts of the theory	3
1.1.2	Debrief on insights	4
1.1.3	Roadmap of future research	5
1.2	Why A Theory of Neural Networks?	6
1.2.1	The unsatisfactory progress of theory research and the problems it induces	6
1.2.2	Artificial intelligence and humanities	6
1.3	Call for Help	9
1.4	Acknowledgement	9
1.5	Notation	10
2	Neural Network, A Powerful Inference Machine of Plausibility of Events in the Physical World	10
3	S-System: Physics, Conditional Grouping Extension and S-System	12
3.1	Physical Probability Measure Space and Sensor	12
3.2	S-System: Hierarchical Measure Group and Approximate System	13
3.3	Related works	15
3.3.1	Hierarchy	15
3.3.2	Hierarchical Hypothesis Space of NNs	15
3.3.3	Machine Learning Algorithm Paradigm	16
4	Geometry of NNs (and S-System), and Neural Networks From the First Principle	17
4.1	Theoretical Derivation of Activation Functions and MLPs	17
4.2	NN Manifold and Contraction Effect of Conditional Grouping Extension	18
4.3	Related works	19
5	Learning Framework of S-System (and NNs)	20
5.1	Learning framework	20
5.2	Related works	22
6	Optimization Landscape of NNs (and S-System)	22
6.1	Hessian of NN Is Inherently A Huge Random Matrix	24
6.2	Eigenvalue Distribution of Symmetry Random Matrix with Slow Correlation Decay	25
6.3	Diversity Assumption is A Precondition to the Power of NNs and S-System	28
6.4	NN Loss Landscape: All Local Minima Are Global Minima with Zero Losses	29
6.5	Related works	32
	Appendices	40
A	Coupling Theory	40
B	Manifold and Diffeomorphism Group	41
C	Information Geometry	42
D	Statistical Learning Theory	43

1 Introduction

1.1 A Theory of Neural Networks

1.1.1 Debrief on the four parts of the theory

We present a measure-theoretical theory of NNs. We summarize the parts of the theory present in this paper in this section. To the best of our knowledge, we do not find works that are intimately close to ours, and each part of the theory has its own related works, which is present at the *end* of each part.

The theory is divided into the following *four* parts:

- **S-System**, a formal measure-theoretical framework that builds a hierarchical hypothesis space, of which NNs are special cases. It is motivated by the fact that nature is a complex system built hierarchically, and a mechanism is needed for any agents living in it to recognize and predict hierarchical events happening.
- The **geometry** of S-System. The objects in the hypothesis space are probability measures, thus have an information-geometry structure. It characterizes the phenomenon that NNs compose and recompose manifolds that have increasingly high level semantic meaning.
- The **learning framework** of S-System. It describes the objective functions to identify an element of the hypothesis space by learning the parameters. It gives a principled derivation of back propagation, and unifies supervised learning and unsupervised learning in NNs.
- The **optimization** landscape of S-System. It identifies principles and conditions that make the non-convex optimization of the learning problem of S-System benign; that is, all local minima are global minima.

We provide a short introduction to the ideas of the theory present, and defer a full informal introduction to section 2. The theory is a principle to assign probabilities to events, e.g., predicting the event that today would be a rainy or sunny day, given past experience. To recall what a principle to assign probabilities is, we recall the principle of symmetry, and the law of large number (LLN). The example of the principle of symmetry is ubiquitous in elementary probability theory, i.e., the equal probability assigned to the event that front, or back side is obtained when flipping a coin. So is the law of large number, which assigns normal probability distribution to that of the mean of a large number of random variables. However, the way that these principles assign probabilities grounds on a certain ensemble of repeated trials: the equal probability assigned to coin flipping is mostly grounded on the observation on thousands of repeated trials; and the LLN grounds on averaging repeated trials (with, or without the i.i.d. assumption). However, how can we assign probabilities to the events that are not as simple as coin flipping, and is not the average of an ensemble, like the one of weather prediction?



Figure 1: Cloud types: on the left is Cirrus, the type of cloud that normally wouldn't rain; on the right is the Cumulonimbus, the type of cloud that normally would lead to thunderstorms.

Given daily life experience, Cirrus (c.f. left picture of fig. 1) normally would not rain, but Cumulonimbus (c.f. right picture of fig. 1) would lead to thunderstorms. The prediction from daily experience comes from repeatedly observing the cloud shapes and the weather afterwards. The physics behind the cloud formation and raining is a complex interaction in the complex system of moisture, dust, gravity, temperature and wind, of which the outcome is highly chaotic thus uncertain. The past experience allows us to divide the cloud shapes into different groups by their salient features that would have a high, or low chance to induce different types of weather. **THUS**, in summary,

we have a mechanism to approximate the probability of outcomes of a complex physical system by observing certain features of the system. That is, to estimate the plausibility of the raining event based on the events that a certain group/type of cloud shape occurs.

The theory is a theory of the above mechanism. S-System is the mechanism to construct feature events, e.g., the shapes of the cloud, that are divided into groups/types and mirror the hierarchical relationship in the physical complex system in a self-organized way, e.g., the interaction between water molecules and dust particles. The geometry part characterizes the information geometry structure in the feature space. The relationship between it and S-System is like the one between Hilbert Space and functions: it provides a manifold structure. The learning framework characterizes how the past experience helps learn the grouping of events: the objective function that divides events into different cloud types/groups, and discovers feature events in a self-organized way. The optimization part investigates how such a grouping is implementable given objective functions in the learning framework through optimization. They overall make up a theory of NNs (and S-System): the mechanism to approximate/assign the probability of/to outcomes of events to estimate the plausibility of events, e.g., the occurrence of rain or not.

1.1.2 Debrief on insights

In this section, we summarize insights of our results that are related to existing research.

Inherent stochasticity of NNs. Supervised NNs, such as Multiple Layer Perceptron (MLP), has been normally taken as a deterministic model, a function approximator that approximates a probability distribution. However, underlying the superficial determinism, the deterministic forward propagation maximizes the expected data likelihood, while the backward propagation minimizes surrogates of KL divergence (details in section 5). Activation functions have principled derivation (details in section 4.1), and activations in the intermediate layers are a function of estimated realization of random variables, of which the true probability measure is transported from the input data. Activation functions used in practice actually assume the exponential family distributions, which explains why NNs learn templates, since the mean of an exponential family distribution uniquely determines its distribution. (details in section 4.2). S-System unifies and formalizes the interpretations proposed in Bottou et al. (1996) Ankit B. Patel et al. (2016), Lin and Tegmark (2017) and Trevisanutto (2018) — the connection with these works is explained in detail in section 3.3.2.

Stochastic manifolds composed and recomposed by NNs. It is observed that NNs gradually build representations that function similar with the “grandmother” cell found in a biological brain, and it has long been hypothesized that NNs build certain kinds of manifolds. The stochasticity discovered allows us to discover the information geometry structure in NNs, in which a hierarchy of manifolds are composed and recomposed to build increasingly high level semantic representations as layers stack up. More specifically, a stochastic manifold structure is endowed on the intermediate space of NNs, where the “distance” is defined to characterize the semantic difference between events/samples. *As the layers go deeper, the semantic difference potentially becomes gradually coarse-grained to reflect the higher level semantic difference, e.g., dogs vs cats, while ignoring lower level variations, e.g., textures (theorem 4.1).* This is true due to the information monotony phenomenon: by blocking half of the information from propagating (using ReLU as an example), information related to irrelevant variations could potentially be discarded. It also characterizes the phenomenon that samples/images can be compared in different criteria in different contexts (details in section 4.2), thus opening the new possibilities of metric learning.

Symmetry in NNs The event spaces of samples are the objects to study if one wants to study the symmetry in NNs. For example, robustness to deformation in images can be characterized as a close “distance” between two event collections where one is obtained by deforming all events in the other. An example is provided in example 4.1.

Optimization landscape of NNs. The stochasticity identified enables us to analyze NNs stochastically in its full complexity. It enables us to characterize the optimization behaviors of NNs in theorem 6.1. It explains the optimization myths of NNs that though being non-convex NNs can optimize the loss to zero, and why learning progresses slowly when approaching the minima. *Informally, a huge number of cooperative yet diverse neurons can divide samples into arbitrary groups corresponding to labels.* The assumptions made in the theorem are sufficient practice-guiding preconditions instead of unrealistic assumptions made to make the proof work. It explains why centering of neuron activation (Glorot and Bengio (2010)) is helpful, for it helps to let the eigenvalues of the Hessian of the risk function be symmetric w.r.t. y-axis (theorem 6.3), thus guaranteeing the existence of negative eigenvalues

to provide loss-minimizing directions; why normalization of neuron activation (Ioffe, Sergey and Szegedy (2015)) is helpful, for that the boundedness and diversity conditions 6.1 6.2 ask the correlation between neurons, formulated as cumulants, to be small, and normalization of standard deviation possibly maintains the conditions throughout the training (section 6.3); why the larger the network is, the easier for it to reach zero error, for our results on optimization is a probably-approximately-correct type result where the error is controlled by the size of the network — the underlying reason is complicated, and we refer readers to section 6.2.

Renormalization group implemented by NNs. NNs has been analogized with Renormalization Group (RG) (Mehta and Schwab, 2014). However, the analog is incomplete for that it does not identify a large scale property, quantity, or feature that is produced by such renormalization (Lin and Tegmark, 2017). RG is a Nobel-prize-winning tool that bridges phenomenon across scale in complex systems, e.g., spin glasses, in physics. In a similar vein, the theory completes the analog, and show that NNs is a complex adaptive system that estimates the statistic dependency of microscopic objects, e.g., pixels, in multiple scales. Unlike clear-cut physical quantity produced by RG in physics, e.g., temperature, NNs renormalize/recompose manifolds emerging through learning/optimization that divide the sample space into highly semantically meaningful groups that are dictated by supervised labels (in supervised NNs). It also formalizes the folk wisdom that NNs only care about a subset of the overall signal space, e.g., image space, in a measure theoretical way (details in section 3.1).

Discretization of continuous sample space done by NNs and formalization of semantics. This is the most important one, but also perhaps the hardest one to get. *S-System shows NNs work by grouping samples/events into groups — which emerges through optimizing an objective function — and estimate/approximate their true probability measure through empirical observations.* The group is a formalization of semantics, and explains how discrete labels emerge from continuous samples, i.e., a label identifies a group of events/samples. A proper implementation of the above process applied hierarchically creates an adaptive complex system that consists of a huge number of neurons. The neurons represent different groups of events of low mutual correlation, which preconditions the optimization results. It gives a formalization of the meaning of semantics in labels in supervised learning, which is a way to group samples/events in a way meaningful to humans (details can be found in section 5). This is the aim of the whole paper, and is described informally in section 2.

1.1.3 Roadmap of future research

It is straightforward that ongoing efforts are working on further perfecting of existing parts of the theory. Except for the first part, only the scaffold of the remaining three parts is worked out. To name an example, the optimization part has worked out the conditions for a NN with a binary loss to converge to global minima. The result analogizes with the situation where numbers have been defined, while the algebra on them has not. The understanding of the algebra would lead to results that cover all kinds of losses. Besides these short term goals, we would like to present a bigger picture that covers missing pieces, and future extension of the theory.

Short term goals: adversarial robustness, generalization behaviors, dynamical-sized self-generating networks, Lisp, C++, CUDA based NN library. 1) The adversarial samples phenomenon (Szegedy et al., 2013) is caused by an uncontrolled propagation of errors in a high dimensional space in an exponential way, a proper characterization of errors may solve the phenomenon. 2) Though the generalization myth discovered by Zhang et al. (2016) has been solved by Soudry et al. (2018) and Poggio et al. (2018), a neat generalization error (GE) bound like the one for Support Vector Machine has not been worked out. The insights obtained to control the error of adversarial samples will help us formulate a lean bound. 3) When the optimization and generalization behaviors are worked out, it would make it possible to formulate quantitative criteria concerning the training set performance and test set generalization solely based on network parameters. The criteria would make the training of dynamically-sized, self-generating networks possible. 4) Anticipating the coming of the self-generating networks, and more importantly, for the long term goal described below, we are designing a new NN library that can mix Lisp, C++ and CUDA arbitrarily. The idea is that a difference exists in the mathematical reasoning language (the front end), and the underlying hardware implementation language (the back end). The front end is universal, since it is based on mathematics. The back end is hardware-dependent, it could be GPU based for now, yet other types of hardware are possible, e.g., Tensor Processing Unit of Google. The back end idea has been around for a while, e.g., XLA in Tensorflow (Abadi et al., 2015), or even earlier, intermediate representation in LLVM, but our front end perhaps will be more than computational graph.

Long term goal: unification of logic and perception. For the above goals, we have a rough guess on how they could be achieved, thanks to the cracks opened by S-System. However, our ultimate goal is to combine logic with perception. What S-System achieves is an algorithm framework that builds a representation of the physical world. In other words, it is the mathematics of imagination. However, for now, it only characterizes rather basic hierarchical composition relationship between the representation of objects. The next big extension is to characterize how these representations are manipulated in a more universal way, which we believe is the key to rational reasoning found in human intelligence. That’s one key reason why we are working on a new library, since we believe the days of Lisp may come again when we reach this part of the road map. *This is the reason why we name the framework S-System: we envision one day S-Expression (McCarthy, 1960) and S-System would become dual views of a one.*

Temporary end goal: mathematics of languages. Languages are probabilistic logic rules to manipulate symbols that are representations of the physical world. The idea probably dates from Wittgenstein, who said in each sentence, there is a picture. The idea probably was taken very seriously in the Vienna Circle (Sigmund and Hofstadter, 2017) in logic empiricism/positivism. But they missed an important piece from Phenomenology (Dreyfus, 2015), which is what machine learning was set out to fill. The thread is rather simplified, and we will do a better survey when we finally reach the end of the road map. The unification of logic and perception would make possible the mathematics of languages, and consequently making true understanding of books possible. *It would lay the theoretical foundation to build a system that would organize all human knowledge that serves as an infrastructure may help humanities get out of the post-modern quandary.* More on this in section 1.2.2.

1.2 Why A Theory of Neural Networks?

1.2.1 The unsatisfactory progress of theory research and the problems it induces

The recent development in the algorithm family of neural networks (NN) (LeCun et al. (2015)) that aim to solve high dimensional perception problems, has led to results that sometimes outperform humans in particular datasets, e.g., vision (He et al. (2015)). It is a computational imitation of biological NNs (Rosenblatt (1958) Fukushima (1980) E. et al. (1986)). Researchers of various backgrounds have been intrigued by theoretical understanding of NNs, and made important contributions to it. From the perspective of physics, we have Mehta and Schwab (2014) Lin and Tegmark (2017) Choromanska et al. (2015a); from that of applied mathematics, we have Mallat (2016); from that of information theory, we have Amari (1995) Shwartz-Ziv and Tishby (2017); from that of theoretical computer science, we have Arora et al. (2014); and from the machine learning perspective, we have Anselmi et al. (2015) Anselmi et al. (2016) Jeffrey Pennington (2017) Ankit B. Patel et al. (2016) etc.

However, despite the above efforts, it has arguably progressed for six decades as a blackbox function approximator. It lacks a formal definition (Candes, 2002), and has opaque optimization behaviors, mysterious generalization errors (Zhang et al., 2016), and intriguing adversarial samples (Szegedy et al., 2013). The confusion of definition leads to the confusion of the capability, and more importantly the limitation of the algorithm family, which has led to criticism for not being able to undertake tasks that perhaps outside the domain of NNs (Pearl, 2018), the worry of technology wall in the industry and another AI winter, and ungrounded fear of singularity in the general public. The lack of understanding of the behaviors of NN leads to a proliferation of competing techniques and tricks that are hard to assess their relative merits (Lipton and Steinhardt, 2018). It is clear, for the field of deep learning to progress healthily, a theory underpinning is being called besides its intellectual reason.

1.2.2 Artificial intelligence and humanities

Zooming out from the academic world, intellectually, what is really significant about the success of artificial NNs, and the development of its science? It is perhaps the Homo Sapiens’ unusual opportunity to fix its own errors, drag itself out of the quandary of post-modernity, and build a future that has a better dual (Yin Yang) balance of liberty and equality.

The general public may be appalled, or perplexed by futurists’ attention-grabbing prediction on the technology singularity, in which the emerging of artificial superintelligence may disrupt human civilization, and make humans look like bugs. But intellectuals and scientists, we believe, are long at peace with the picture painted by science. The shock of the picture had been long ago manifested by the painting *Philosophy* (fig. 2) of Gustav Klimt, when he was commissioned to depict *The Triumph of Light over Darkness* to “deliver an optimistic glorification of progress”. Klimt instead paints “On the left a group of figures, the beginning of life, fruition, decay. On the right, the globe as mystery. Emerging below, a figure of light: knowledge.”, which presents an a dreamlike mass of humanity, referring neither to optimism nor rationalism, but to a “viscous void” (Wikipedia, 2018). The tension depicted by *philosophy*

dates back to the onset of the Romantic age (Barzun, 2001), for science's abstruseness and abstract, and perhaps more importantly, for that it shows man is not "the center of the universe" in primitive mythologies, not even "the measure of all things" in the Renaissance, but animals developed from stardust in a trial and error fashion in perhaps a corner of the vast, violent, lonely, impersonal universe. The tension gradually developed into a disillusion with science (Watson, 2002). The last straws are Godel's incomplete theorem and phenomenology. The two combined shows that logic reasoning cannot reach truth, all laws are human invention, and their correctness are entirely based on perceptual empirical observations. The uncertainty of truth, and the bust of the Utopian dream where people are allocated what they need gradually morphed the society into post-modernity, where "postmodernism resigns to the alienation, ephemerality, fragmentation and patent chaos of modern life and places individualized aesthetics over science, rationality, politics and morality" (Harvey, 1989).



Figure 2: The painting *Philosophy* by Klimt.

In the masterpiece of Barzun (2001), *From Dawn To Decadency*, in stating the reason why he judged the Western culture is in decadency, he wrote

But why should the story come to an end? It doesn't, of course, in the literal sense of stoppage or total ruin. All that is meant by Decadence is "falling off." It implies in those who live in such a time no loss of energy or talent or moral sense. On the contrary, it is a very active time, full of deep concerns, but peculiarly restless, for it sees no clear lines of advance. *The loss it faces is that of Possibility.* The forms of art as of life seem exhausted, the stages of development have been run through. Institutions function painfully. Repetition and frustration are the intolerable result. Boredom and fatigue are great historical forces.

It is a quite proper assessment, given our societies' situations. It even may give a sense of relief for expressing the subconscious public mood. Even Hacker Spirits (Himanen, 2001) has lost its moment (Levy, 1984), and was tamed by capital (Turner, 2010).

Yet, this is not the end of the story. The progress in modern biology, complex system, nonlinear science has drawn another humanitarian epic narrative (Wilson, 1999), out of the Darwin jungle one — indeed we are stardust, but we are also children of the sun: "Human social existence, unlike animal sociality, is based on the genetic propensity to form long-term contracts that evolve by culture into moral precepts and law. They evolved over tens or hundreds of millennia because they conferred upon the genes prescribing their survival and the opportunity to be

represented in future generations. We are not errant children who occasionally sin by disobeying instructions from outside our species. We are adults who have discovered which covenants are necessary for survival, and we have accepted the necessity of securing them by sacred oath.” The consilience of knowledge has the potential to build a shared culture for all (Wilson, 1999), beyond nationalism, tribalism, region fundamentalism, racism (Castells (2000) Castells (2011) Castells (2010)): “We are gaining in our ability to identify options in the political economy most likely to be ruinous. We have begun to probe the foundations of human nature, revealing what people intrinsically most need, and why. We are entering a new era of existentialism, not the old absurdist existentialism of Kierkegaard and Sartre, giving complete autonomy to the individual, but the concept that only unified learning, universally shared, makes accurate foresight and wise choice possible.(Wilson, 1999)”

However, the message is guarded in a high castle. To understand and act upon it requires tremendous efforts. “To which what one Crates’ said of the writings of Heraclitus falls pat enough, ‘that they required a reader who could swim well,’ so that the depth and weight of his learning might not overwhelm and stifle him.”(Montaigne and Screech, 2004) If the enlightenment project of consilience is to be successful, it has to be inclusive to all that are willing to take the leap: the ones sense that something deep in the contemporary culture is wrong, but do not have a thread or the knowledge to identify the exact source, then initialize the work to fix it. We *lack the infrastructure* to give the opportunities to those who want to be part of this Icarian flight: the bridge between the real problems and the skills that are conducive to a solution, instead of a cannily maintained tribalism elite culture that insecurely bars the door and concentrates all resources (Deresiewicz, 2014). Without it, the society in an optimistic guess may be gravitated towards self-perpetuating tech-meritocracy parishes, in which the masses are lived by a certain implementation of universal basic income without a meaningful goal for life as depicted in the drama *The Expanse* (Mark Fergus, 2018). Or even worse, the un-channeled negative emotion may build up, then explode to consume the earth through radical populism, or radical religion fundamentalism, as having happened repeatedly in the history. We believe the empowering is the Promethean fire that the enlightenment project was set out to bring in the beginning: the age that symbolically started by Francis Bacon.

There are reasons to believe that we are at the night before the dawn of a new age, as the Renaissance is built on the ashes of the corrupt monastery order (Mee, 2016), on the condition that we fix our errors — we already have a direction where our new philosophy may come from, thanks to forerunners like Wilson (1999). We have incrementally managed to build many new experimental infrastructures. To name an incomplete list: Linux, an operating system that is owned by all; Wikipedia, the open and free encyclopedia accessible to all; search engines to access to the world’s knowledge freely and instantly as long as it is online; 3D printing to bring the software power physical without heavy capitals to build a factory; RISC-V, an open source hardware counterpart to Linux; digital libraries that open research results to all, e.g., archive.org, and arxiv.org, where this paper is submitted; interplanetary file system (Benet, 2014), and the distributed Internet built on it; the possibilities of new content based business models enabled by blockchain based currency (Nakamoto, 2008) instead of the ad based one that feeds on traffic and appeases to baser instincts.

In this particular niche domain of artificial NNs, it is not about a gold mine for capitalistic opportunities, a boost in the efficiency of the current system, potential dangerous superintelligence or “smart” cities. It is a thread to decode how humans learn, so that if the science of learning is worked out, imagine what that would bring? The enlightenment was initiated by the printing press, which enables the massive dissemination of knowledge and information. In the contemporary society, the situation is reversed. It is not the access to information that becomes the bottleneck, but the capacity to process it: we are drowning in information, but starving for knowledge. The internet has pushed the situation to the extreme, and helped build polarized societies — Google is powerful on the condition that you ask the right question. If the science of learning is worked out, we may build an upgraded version of Google Books, or in other words, finish what Google Books set out to be: a system that organizes all human knowledge, and serves as a guide to everyone who is willing to learn and tackle challenges humanity facing, thus offers world-class education to everyone, and continues the enlightenment project initiated by the printing press, while fixing its errors.

Artificial Intelligence (AI) is part of the pieces mentioned above to build the infrastructures that help build solve the tension between liberty and equality, or laissez-faire and social welfare, by providing equal educational opportunities, at least offering a low cost thread to navigate the ocean of knowledge humanities has generated. Though how unconvincing the following statement may be, AI could not be a threat. All tools have two edges. If we are equipped the wisdom to wield it, it would not be a threat, and is fundamentally controllable, at least the part based on the mathematics in this paper, which covers deep NNs, the backbone of existing deep learning technology.

As a concluding remark, this section was initially planned as a short paragraph explaining the deeper motivation for this paper that has been there for years. It unexpectedly developed into a full section. The content is surely

underdeveloped, and may seem naive in the years to come. *This is the first time the idea is on paper, feedback and help are welcome, by sending them directly to the author, the email of which can be found at the author information in the first page. The help part is described in details in section 1.3 below, and the footnote of this page.*

1.3 Call for Help

We have some trouble identifying appropriate journals, or conference to publish this manuscript, in whole, or in parts. The theory has drawn techniques and ideas from a range of domains, and is rather interdisciplinary. The author is a self-taught outsider who navigates the domains involved all by himself. We are not deeply familiar with the established manner of communication in the domains involved, and trial-and-error submission would take too much time, which could be better used in other areas. **We very much appreciate experts in relevant domains to be our guide.**¹ In the following, we will debrief the domains involved, and where we want guide.

Expertise is needed in introducing *random matrix* techniques to the machine learning community. We may manage to get section 6 published on its own, since it solves a well-perceived open optimization problem of NNs — it does not mean that the rest does not solve open problems; it is just that the problems are less perceived. What's left to be done is to write an easy-to-follow introduction.

Expertise is needed in the boundary between *pure mathematics* and *applied mathematics*, between *probability* and *statistics*. The trickiest part is S-System in section 3. Coupling theory is traditionally a proof technique, here it is used as a computation technique. It is a new way to assign probabilities, with deep philosophical implications (which is not discussed in this paper), besides the symmetry principle and the law of large number. So this makes it pure mathematics. However, it is also an algorithm for concrete machine learning problems, which makes it applied mathematics. It is a marriage between probability measure and statistics in a way other than statistical learning theory (STL), or in other words, it is a missing piece of the theory of *algorithmic modeling in statistics* (Donoho, 2000) (Breiman, 2001), along with STL. We are not familiar with relevant venues, and are not sure where we should submit it: the feedback from the machine learning community says the idea is abstruse, while the math community seems not to have a clear community on the problem.

Expertise is needed in *information geometry* and *probabilistic graphical modeling*. Section 4 and section 5 cannot be understood without section 3, though they also involve different communities. The former on information geometry. The later perhaps is less clear — it belongs to a community that it sets out to prove ill-directed, i.e., probabilistic graphical modeling approach in machine learning; the task is obviously difficult. We need to get the ideas of section 3 through before expanding on these two parts. To publish them on their own, one approach is to demonstrate its usefulness in strong experiment results without requiring the reviewers to understand the nitty-gritty of the technical contents, but it would take some time to identify the proper practical problem to solve.

Perhaps the best way is to publish it in somewhere like the *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, where a paper (Mallat, 2016) similar in style published. But it would be even harder than the previous routes perhaps: not only expertise is needed, but also reputation; Mallat (2016) also published in piecemeal first, and took four years to reach the point to write a summary.

We are doing this in hackers' style: we work out an initial solution, and present it to the community, like the situation of GNU/Linux in the beginning. We hope the community would take notice, and help work on this in joint efforts.

1.4 Acknowledgement

The work would be indefinitely delayed, or even impossible without the support and tolerance of my past mentors. In reverse chronological order, I would like to thanks Kui Jia at the South China University of Technology, without whom to offer an extended stay in the academic world with a high degree of research freedom when I graduated my Master of Philosophy degree and decided to return to mainland China, I am not sure how long it would take, or even possible to reach this paper if I went off the industry; Xiaogang Wang, who was my master degree mentor at Chinese University of Hong Kong, and along with Kui to show me the works of Stephane Mallat (Mallat, 2012) to introduce me to the theory research in NNs, and helped me navigate the academic community; Jinhui Yuan,

¹The author is willing to take visiting scholarship, or research position to work on the publication of this manuscript. In addition, the author is looking for a strong community to get a PhD that is working on the fundamental theory of NNs, and is interested in the roadmap present in section 1.1.3. The author has been laboring in the darkness to get this preliminary theory out in the past six years, and is in great need of a community. If that is not an option, he is also seeking employment, for he is not just a researcher, but also an engineer in the first place — the theory partly started with the hacker idea to build a AI, one needs to understand it. For details, refer to <http://shawnleezx.github.io/employment>

my mentor at Microsoft Research Asia, to show me a broader world of machine learning, and share his enthusiasm to research; Linli Xu, who was along the first ones to introduce me to the field of machine learning, and be my undergraduate thesis advisor at University of Science and Technology of China — it was around that time that I started to be cognizant of and ponder on the theoretical problems of artificial intelligence, machine learning, and deep neural networks.

I am also equally in great debt to past great men. Without an imagined community built by them, and their stories to keep me company in the dark times, the mental wear and tear would have me give up on the project a long time ago. Without a particular order, they are Bertrand Russell, who spent ten years on *Principles of Mathematics* to search for truth in logic, and had to cover some of its publication expense on its own after those ten years' work; Nicola Tesla, who worked in a ditch to find the next meal when working on his three-phase electric power, and who used up all his life force on his worldwide wireless communication system; Albert Einstein, who spent seven years as a patent officer when working on his theory of relativity, light quanta, and Brownian motion; Voh Gogh, the father of modernism art who was so afflicted by modernity, and battled with it on the edge of insanity; Desiderius Erasmus who worked on the Latin and Greek translation of New Testament quite literally against the world (for religion reform) in chronic poverty; Montaigne who stayed a sane philosopher to write *Essays* in a world ripped apart by religion factions. Yangming Wang, a great sage who cannot really be characterized in one sentence without sufficient culture background. The list can go on, but I think I have sent the message. If not, to illustrate what I mean, Russell wrote in his autobiography:

At the time, I often wondered whether I should ever come out at the other end of the tunnel in which I seemed to be.... I used to stand on the footbridge at Kennington, near Oxford, watching the trains go by, and determining that tomorrow I would place myself under one of them. But when the morrow came I always found myself hoping that perhaps “Principia Mathematica” would be finished some day.

1.5 Notation

We note the notation used. All scalar functions are denoted as normal letters, e.g., f ; bold, lowercase letters denote vectors, e.g., \mathbf{x} ; bold, uppercase letters denote matrices, e.g., \mathbf{W} ; normal, uppercase letters denotes random elements/variables, all the remaining symbols are defined when needed, and should be self-clear in the context. r.e. and r.e.s are short for random element, random elements respectively, so are r.v. and r.v.s for random variable and random variables. To index entries of a matrix, following Erdős et al. (2017), we denote $\mathbb{J} = [N] := \{1, \dots, N\}$, the ordered pairs of indices by $\mathbb{I} = \mathbb{J} \times \mathbb{J}$. For $\alpha \in \mathbb{I}$, $A \subseteq \mathbb{I}$, $i \in \mathbb{J}$, given a matrix \mathbf{W} , w_α , \mathbf{W}_A , $\mathbf{W}_{i:}$, $\mathbf{W}_{:i}$ denotes the entry at α , the vector consists of entries at A , the i th row, i th column of \mathbf{W} respectively. Given two matrix \mathbf{A}, \mathbf{B} , the curly inequality \preceq between matrices, i.e., $\mathbf{A} \preceq \mathbf{B}$, means $\mathbf{B} - \mathbf{A}$ is a positive definite matrix. Similar statements apply between a matrix and a vector, and a matrix and a scalar. \succeq is defined similarly. κ denotes cumulant, whose definition and norms, e.g., $|||\kappa|||$, are reviewed at section 6.2. $:=$ is the “define” symbol, where $x := y$ defines a new symbol x by equating it with y . tr denotes matrix trace. $\text{dg}(\mathbf{h})$ denotes the diagonal matrix whose diagonal is the vector \mathbf{h} .

2 Neural Network, A Powerful Inference Machine of Plausibility of Events in the Physical World

This section aims to give an intuitive description of the formal definition of NN given and its behaviors without delving into mathematical details; and it also serves as the paper outline. *The rest of the paper characterizes the informal description in this section formally.*

To avoid philosophical debates, we start with a metaphysics assumption that an objective physical world exists independent of the perception/observation of any agents, yet we do not make any assumptions on the nature of the physical world, being it a simulation, or not. If an agent wants to interact with the world, for whatever reasons, it first needs to perceive it. A way to perceive the world is to measure certain physical objects of the world, which could be implemented as sensors of the agent, e.g., a camera measuring the spatial configuration of photon intensity. However, the measurement data (formalized in definition 1) record many events that happen in the same spatial and temporal span, and those events are entangled in the measurements (formalized in assumption 3.5). Let’s see an example. A core drive of a living creature, is to survive. When a lion is within 100m of a deer, the deer needs to run in order to avoid being eaten. But how a deer is supposed to know a lion is near by? A scheme could be through a photon sensor, i.e., eyes, of the deer to measure the photons, probably generated from the sun, that are reflected from the body surface of the lion. We call a lion exists nearby an *event*. However, a great number of other

events are also being measured by the sensor — the dark clouds that may rain, the delicious grass that makes food, and the hyenas that lurks around for a leftover meal. *To perceive events happening in the world, e.g., a lion is nearby, a mechanism is needed to recognize it from measurement data, e.g., sensed spatial photon patterns.*

This leads to the problem of the structure of the physical world, which the still unknown mechanism at least needs to relate to if it aims to recognize events in the world well. This leads us to Complex System (Bar-Yam (1997) Nicolis and Nicolis (2012) Newman (2009)). It is a vast and diverse field, but a summary would not be too misleading is that it is a science that studies the interaction of so many events that their collective behavior manifests at a scale beyond their characteristic scale. The research in the field allows us to safely say that one of the most important structures of the physical world is hierarchy (Amderson (1972)). Atoms form molecules; molecules form organism, and inorganic material; organisms form creatures, which form ecology system; inorganic material forms planets, then solar systems, then galaxies. *Of the scales across the hierarchy of the world, the events at a lower scale interacting with a particular way form events at a higher scale.* The hierarchy in nature is formulated as assumptions 3.1 3.2 3.3 3.4 3.5.

In section 3, S-System is introduced in definition 3 to recognize and represent the hierarchy of events from measurement data, formulated measure-theoretically and based on probability coupling theory (Thorisson (2000)). S-System formalizes the idea that a creature is not going to reproduce an impartial representation of the world, it only captures the events that cater to its need, e.g., the survival need to identify lion, and within its reach and capacity, i.e., the amount of measurements it can gather. *S-System recognizes a hierarchy of events from the measurements, not exactly in the sense of physical reality — if a creature never measured/saw a black swan, it does not mean there aren't any — but in the sense of a manually created hierarchical groups of events, and each group is perhaps given a tag/name.* For examples, some low level object groups are named as edges; groups in a relative higher level are named textures; groups in another higher level are named body parts; groups in an even higher one are named body, e.g., lion.

In section 4, a NN is shown to be an implementation of an S-system (shown in definition 4), when the measure is being approximated with compositional exponential family distributions (defined in definition 6). The unique properties of exponential family distributions enable us to define a geometry structure on the intermediate feature space of NNs based on information geometry. The most important discovery (for now) in such a structure is that it shows that the “distance” between events (intermediate features) is contracted as layers stack up, e.g., the intra-class difference between an object and a slight deformed object (c.f. theorem 4.1). In more details, Activation Function is derived formally in definition 5. The geometry structure of the representation built by NNs is defined in definition 7 as stochastic manifolds. The section shows that a NN is an inference system to infer how plausible groups of events forming a hierarchy have occurred, as S-System is defined.

In section 5, the learning framework of S-System (NNs) is introduced. It formalizes how past repeated observations are summarized to approximate the probability of groups of events given the current observation: through optimizing an objective function.

The practicality to optimize the objective function is addressed in section 6. The hierarchical organization of NNs to recognize and represent the hierarchical physical world has made itself a complex system. *Contrary to existing shallow models, or simple models, it is exactly the complexity built in hierarchy that makes NNs the most powerful inference model.* In a complex system, when the collective behavior of events is stable, in the sense that it emerges on the overall interaction of lower scale events, yet does not depends on any small subsets of them, it is termed an *emergent* behavior. For instance, the magnetic force of a magnet is a macroscopic phenomenon, yet it may emerge from the alignment of spin magnetic moment of many elementary particles. A misalignment/disorder in a fraction of the particles that does not exceed a critical points to cause phase transition won't destroy the magnetic field of the magnet, only weakening it. More examples and more thorough and in depth discussion about emergence could be found in Simon (1962) Nicolis and Nicolis (2012) Kadanoff (2000). Contrary to the high constrained interaction in the spinning particles, the neuron population in a NN is highly flexible and adaptable. A large collection of cooperative yet autonomous neurons, formalized as assumptions 6.1 6.2, gives NNs the ability to partition events into arbitrary groups and infer the plausibility of any groups of events (proved in theorem 6.1), which is the emergent behavior emerging from the disorder in the NN complex system. More technically, formulated as an optimization problem to minimize the error between the empirical probability distribution and its parametric representation, though being non-convex, NNs can reach zero error as long as NNs maintain its diversity while increasing its neuron population.

The four sections 3 4 5 6 respectively deal with the definition of a hypothesis space, the geometry of the space, learning objectives of the space, and the optimization landscape of the objective functions. They combined form a preliminary theory of NNs and S-System, a principle that assigns probability to physical events through learning on the past experience, to infer the plausibility of events might happen in the future.

Warning. We will give a theory drawing techniques and ideas from a range of fields — probability coupling theory Thorisson (2000), statistic physics Kadanoff (2000), information geometry Amari (2016), non-convex optimization Jain and Kar (2017), matrix calculus Magnus and Neudecker (2007), mean field theory Kadanoff (2000), random matrix theory Tao (2012), and nonlinear operator equations Helton et al. (2007). As a word of warning, based on the initial feedback obtained, it is very hard to understand respective parts of the paper without having solid background in the above fields. We suggest readers do prepare for a hard read.

3 S-System: Physics, Conditional Grouping Extension and S-System

In this section, a mechanism, Hierarchical Measure Group and Approximate System, nicknamed S-System, to recognize and represent events in the physical world from measurements is introduced with formalism from probability coupling theory (Thorisson (2000)) in a measure-theoretical way.

3.1 Physical Probability Measure Space and Sensor

To begin with, we formally define the assumptions made on the physical world.

Assumption 3.1 (Physics). *All events in the physical world makes a probability measure space $\mathcal{W} := (\Omega, \mathcal{F}^{\mathcal{W}}, \mu^{\mathcal{W}})$, where Ω denotes the event space, $\mathcal{F}^{\mathcal{W}}$ is the σ -algebra on Ω ; $\mu^{\mathcal{W}}$ is the probability measure on Ω . We call \mathcal{W} **Physical Probability Measure Space (PPMS)**. To avoid confusion, we note that Ω denotes the event space of PPMS throughout the paper.*

Remark. *For an example of the events, refer to section 2; given $\omega, \omega' \in \Omega$ the intersection $\omega \cap \omega'$, union $\omega \cup \omega'$ and complement operations $\neg \omega$ in \mathcal{F} respectively denotes both ω, ω'' occurs, either one of ω, ω'' occurs, and ω does not occur. We shy away from giving an assumption of the meaning μ , and merely rely on the axiomization by Kolmogorov, for it is philosophical and metaphysical for now, though a reasonable interpretation of the measure approximated by a NN is given later.*

Assumption 3.2 (Hierarchy). *Ω has a hierarchical structure, which means $\Omega = \cup_{s \in \mathcal{S}} \Omega_s$, where s is named **scaled parameter**, \mathcal{S} is a **poset**, i.e., a set with a partial order, Ω_s are event spaces, and for $s, s'''' \in \mathcal{S}, s < s'$, $\Omega_{s'} \subseteq \sigma(\Omega_s)$, where $\sigma(\Omega_s)$ is the σ -algebra generated by Ω_s .*

For $s, s' \in \mathcal{S}, s < s'$, we say $\Omega_{s'}$ is composed by Ω_s . Furthermore, when $\omega_{s''''} \in \sigma(\cup_{s \in I_s} \{\omega_s\})$, where I_s is an index set and for any $s \in I_s$, $\omega_s \in \Omega_s$, we say $\omega_{s''}$ is composed by ω_s . As motivated in section 2, to perceive the events happening in the world, measurements need to be collected, which is formalized as a r.e..

Definition 1 (Measurement Collection). *A measurement collection is a random function X that supported on PPMS \mathcal{W} with an induced probability measure space $\mathcal{X} := (\Omega^{\mathcal{X}}, \mathcal{E}^{\mathcal{X}}, \mu^{\mathcal{X}})$, where $\Omega^{\mathcal{X}} := \{x|x : \mathbb{U} \rightarrow \mathbb{V}\}$ and \mathbb{U}, \mathbb{V} are unspecified the domain and codomain.*

We make the following assumptions on X . It characterizes the capability and limitation of a sensor and the phenomenon that for an event $\omega_{s'}$ composed by a lower scale event ω_s , the time/place/support where $\omega_{s'}$ happens contains that of ω_s .

Assumption 3.3 (Resolution). *For any measurement collections, a lower bound of scale parameter s_0 exists, such that $\forall s \in \mathcal{S}$, s is comparable with s_0 , and $\forall \omega_s \in \Omega, s < s_0, \mu^{\mathcal{X}}(X(\omega_s)) = 0$. In other words, part of measure of Ω in singular w.r.t. to μ of \mathcal{X} . We call Ω_{s_0} **the events of the lowest measurable scale**.*

Assumption 3.4 (Measurability). *measurements are physical, i.e., $\mathcal{E}^{\mathcal{X}} \subseteq \sigma(\Omega_{s_0})$.*

Assumption 3.5 (Containment). *Given any two comparable scale parameter $s, s', s < s'$, $\omega_s \in \Omega_s, \omega_{s'} \in \Omega_{s'}$, where $\omega_{s'}$ is composed by ω_s , we have $\text{supp}(X(\omega_s)) \subseteq \text{supp}(X(\omega_{s'}))$, where supp denotes the support of $X(\omega) \in E$, i.e., the domain of $X(\omega)$ where $X(\omega) \neq 0$ (we assume the zero element is defined, and indicates nothing has been measured).*

Example 3.1. *Suppose the sensor is an image sensor. $E = \mathcal{L}^2(\mathbb{R}^2)$, the square integrable function space defined on \mathbb{R} , and the scale parameter could be interpreted as the distance $\|x - y\|_2$ between two coordinates $x, y \in \mathbb{R}^2$ for a image $f \in E$ obtained by the image sensor. In this case, a patch of the image is a lower scale event, which is spatially contained in the whole image.*

The containment phenomenon is troublesome, along with the phenomenon that it is possible for any events $\omega, \omega' \in \Omega$ to have overlapping support $\text{supp}(X(\omega)), \text{supp}(X(\omega'))$, even they do not have any composition relationship. This is an inherent problem of measuring: it has collapsed all the events across scales and within the same scale in the same measurement units of the sensor, e.g., pixels at the same location in the image sensor. To perceive certain event has occurred from X , a mechanism is needed to disentangle it from other events.

3.2 S-System: Hierarchical Measure Group and Approximate System

In this subsection, we introduce Hierarchical Measure Group and Approximate System, nicknamed as S-System. Following the motivation described of S-System in section 2, a.k.a. to perceive is to select and group events that serves certain needs of the creature but not to reconstruct faithfully, we create extensions of the probability measurable space \mathcal{X} that “reproduce” measure of higher scale events. The extensions are created hierarchically, by Conditional Grouping Extension (CGE). For a review of the coupling theory and probability measure space extension, please refer to appendix A.

Definition 2 (Event Representation; (Partial) Conditional Grouping Extension). *Let T be a r.e. in measurable space (E, \mathcal{E}) defined on a probability space (F, \mathcal{F}, μ) , a **Conditional Grouping Extension (CGE)** of T is created as the following by conditioning extension and splitting extension.*

First, a conditioning extension $(\Omega^{\hat{\mathcal{H}}^e}, \mathcal{H}^e, \mu^{\hat{\mathcal{H}}})$ of T is created with $((E, \mathcal{E}), (\Omega^{\hat{\mathcal{H}}}, \hat{\mathcal{H}}))$ probability kernel $\hat{Q}(\cdot, \cdot)$, of which an external r.e. \hat{H} in measurable space $(\Omega^{\hat{\mathcal{H}}}, \hat{\mathcal{H}})$ is created with law

$$\mu^{\hat{\mathcal{H}}}(\hat{H}|T) = \hat{Q}(T, \hat{H})$$

Then a splitting extension $(\Omega^{\mathcal{H}^e}, \mathcal{H}^e, \mu^{\mathcal{H}})$ of $(\Omega^{\hat{\mathcal{H}}^e}, \mathcal{H}^e, \mu^{\hat{\mathcal{H}}})$ is created with a $((E, \mathcal{E}) \otimes (\Omega^{\hat{\mathcal{H}}}, \hat{\mathcal{H}}), (\Omega^{\hat{\mathcal{H}}}, \hat{\mathcal{H}}))$ probability kernel $Q(\cdot, \cdot)$ to support an external random element H in measurable space $(\Omega^{\mathcal{H}}, \mathcal{H})$ with law ν , of which

$$\mu^{\mathcal{H}}(\hat{H}|H, T) = Q((T, H), \hat{H}), \text{ and } \mu^{\hat{\mathcal{H}}}(\hat{H}|T) = \int Q((T, H), \hat{H}) \nu(dH; T)$$

We assume that \hat{Q} is a kernel parameterized by $W(T; \theta)$, a transport map W applied on T parameterized by θ . The extension is well defined due to Thorisson (2000) Theorem 5.1.

*Let $\mathcal{M} := ((\Omega^{\mathcal{H}^e}, \mathcal{H}^e, \mu^{\mathcal{H}}), \{H, \hat{H}, T\})$, we call \mathcal{M} the event representation built on T through a CGE — we define formally an **event representation** is a pair, of which the first element is a probability measure space, and the second element is a set of r.e.s supported on the space, called **random element set** of \mathcal{M} . When absence of confusion, we just call \mathcal{M} the event representation built on T . T is called the **input random element** of \mathcal{M} ; H the **group indicator random element**; \hat{H} the **coupled random element**; (H, \hat{H}) the **output random elements** when we would like to refer to them in bunk; \hat{Q} the **coupling probability kernel**; Q the **group coupling probability kernel**; $(\Omega^{\mathcal{H}^e}, \mathcal{H}^e, \mu^{\mathcal{H}})$ the **coupled probability measure space**; $\mu^{\mathcal{H}}$ the **coupled probability measure**; ν the **conditional group indicator measure**; $W(T; \theta)$ the **transport map** of \mathcal{M} . Given an $\omega \in \Omega^{\mathcal{H}}$, we say $W^{-1}(\omega) \subset E$ is an **event represented/indexed/grouped by H** . Since CGE will be used recursively later, to emphasize, when \mathcal{M} only builds on a subset of output r.e.s of another event representation, to emphasize, \mathcal{M} is called an event representation built by a **Partial CGE**.*

We explain why they are named as Conditional Grouping Extension and Event Representation. By assumption, \hat{Q} is a probability kernel parameterized by $W(T)$, e.g., the exponential family probability kernel $e^{\mathbf{w}^T \mathbf{x} - \psi(\mathbf{w})}$, where $T := X, W(T) := \mathbf{w}^T X - \psi(\mathbf{w})$. Suppose X , a measurement collection r.e., is supported by PPMS \mathcal{W} , from definition 1. A transport map W applied on X is a deterministic coupling $(X, W(X))$ that transports the measure $\mu(A)$ of an event $A \in \mathcal{E}^{\mathcal{X}}$ to $W(A)$, of which $W(X)$ is a r.e. on a measurable space $(\Omega^{\hat{\mathcal{H}}}, \hat{\mathcal{H}})$ with law $\mu^{\hat{\mathcal{H}}}$ supported on \mathcal{W} where

$$\mu^{\hat{\mathcal{H}}}(W(A)) = \mu^{\mathcal{W}}(X^{-1}(A)), A \in \mathcal{E}^{\mathcal{X}}, X^{-1}(A) \in \mathcal{F}^{\mathcal{W}}$$

That is to say A is an event that are happening in the physical world, and is being measured by X . The goal of S-System is to estimate the plausibility of the event A . However, the problem is that we do not know $\mu^{\mathcal{W}}$ (that's not to say we do not have an estimation of $\mu^{\mathcal{W}}$ empirically). That's why CGE is needed. CGE hypothesizes a probability kernel Q that approximates the probability $\mu(A)$ of events being measured (conditioning extension) grouped by the r.e. H created by splitting extension. Notice two key constructions to deal with two key challenges here: for the enormity of the event space of PPMS, a.k.a. Ω , only events that happen along with current observation X is estimated through conditioning extension; for events happening along with X , probability is approximated in groups

indexed by H through splitting extension, which physically could be broken down into countless smaller scale events that compose A and some of the sub-events won't be estimated. The design could be understood as economic considerations, though probably it would be the only feasible solution to reasonably approximate $\mu^{\mathcal{W}}$. Then, the r.e. set of \mathcal{M} is the manipulable object that directly connects with events in the physical world, and is named event representation.

Yet, one more problem is looming around: how possibly Q approximates $\mu^{\mathcal{W}}(X^{-1}(A))$ reasonably? Suppose A is a top scale event, by assumption 3.2, $A \in \sigma(\Omega_{s_L}) \subset \sigma(\Omega_{s_{L-1}}) \subset \dots \subset \sigma(\Omega_{s_0})$, where $\mathcal{S}_l = \{s_l\}_{0 \leq l \leq L, l \in \mathbb{N}}$ is a finite set of scales. Thus, to approximate $\mu^{\mathcal{W}}(W^{-1}(A))$ is to approximate the joint distribution of events that compose A , which could be factorized into the probabilities of events that compose A and the probability of A conditioning on the sub-events. This asks to apply CGE recursively, through which we get an S-system.

Definition 3 (Hierarchical Measure Group and Approximate System). *A Hierarchical Measure Group and Approximate System (S-System) is a mechanism to extend the probability measure space of a measurement collection r.e. recursively according to a poset structure $\mathcal{S}^{\mathcal{X}}$ as described in algorithm 1. The poset is called the **scale poset** of the S-system. Ultimately, it creates an event representation $\mathcal{M}^{\bar{\mathcal{W}}} := (\bar{\mathcal{W}}, \mathcal{O}^{\bar{\mathcal{W}}})$, where $\bar{\mathcal{W}} := (\bar{\Omega}, \mathcal{F}^{\bar{\mathcal{W}}}, \mu^{\bar{\mathcal{W}}})$ is the extended probability measure space built and is called **Approximated Probability Measure Space (APMS)**, and $\mathcal{O}^{\bar{\mathcal{W}}}$ is a r.e. set indexed by elements of poset $\mathcal{S}^{\mathcal{X}}$. $\mathcal{M}^{\bar{\mathcal{W}}}$ is called the event representation built by S-System.*

Algorithm 1 S-System. In the algorithm below, the **predecessor(s)** returns a index set \mathbb{I}' that indexes elements in $\mathcal{S}^{\mathcal{X}}$ and $\forall i \in \mathbb{I}', s_i \leq s$ and **successor(s)** return a subset \mathcal{S}' of $\mathcal{S}^{\mathcal{X}}$, where $\forall s_i \in \mathcal{S}', s_i \geq s$. For examples of the functions, refer to algorithm 2.

Input: $\mathcal{S}^{\mathcal{X}} := \{s_i\}_{i \in \mathbb{I}}$ is a poset with a minimal element s_0 , whose elements are indexed by a set \mathbb{I} ; X is a measurement collection r.e. supported on $\mathcal{X} := (\Omega^{\mathcal{X}}, \mathcal{E}^{\mathcal{X}}, \mu^{\mathcal{X}})$, of which the events of the lowest measurable scale are Ω_{s_0}

Output: an event representation $\mathcal{M}^{\bar{\mathcal{W}}}$

$\mathcal{M}_{s_0} \leftarrow ((\Omega^{\mathcal{H}_{s_0}}, \mathcal{H}_{s_0}, \mu^{\mathcal{H}_{s_0}}) := \mathcal{X}, \mathcal{O}_{s_0} := (X)), \mathcal{T}_{\text{out}} \leftarrow \emptyset, \mathcal{S}^{\mathcal{X}_t} \leftarrow \text{SUCCESSOR}(s_0)$

while $\mathcal{S}^{\mathcal{X}_t}$ is not empty **do**

$\mathcal{S}^{\mathcal{X}_{t'}} \leftarrow \emptyset$

for $s \in \mathcal{S}^{\mathcal{X}_t}$ **do**

$\mathbb{I}' \leftarrow \text{PREDECESSOR}(s)$

$\mathcal{H} \leftarrow \bigotimes_{i \in \mathbb{I}'} (\Omega^{\mathcal{H}_{s_i}}, \mathcal{H}_{s_i}, \mu^{\mathcal{H}_{s_i}}), \mathcal{O} \leftarrow \bigcup_{i \in \mathbb{I}'} \mathcal{O}_i, \mathcal{M} \leftarrow (\mathcal{H}, \mathcal{O})$

 Build an event representation $\mathcal{M}_s := ((\Omega^{\mathcal{H}_s}, \mathcal{H}_s, \mu^{\mathcal{H}_s}), \mathcal{O} \cup \{H_s, \hat{H}_s\})$ on \mathcal{O} through conditional grouping extension that supports output r.e.s (H_s, \hat{H}_s)

if $\text{SUCCESSOR}(s)$ is empty **then**

$\mathcal{T}_{\text{out}} \leftarrow \mathcal{T}_{\text{out}} \cup \{\mathcal{M}_s\}$

else

$\mathcal{S}^{\mathcal{X}_{t'}} \leftarrow \mathcal{S}^{\mathcal{X}_{t'}} \cup \text{SUCCESSOR}(s)$

end if

end for

$\mathcal{S}^{\mathcal{X}_t} \leftarrow \mathcal{S}^{\mathcal{X}_{t'}}$

end while

$\mathcal{M}^{\bar{\mathcal{W}}} \leftarrow (\bigotimes_{i \in \mathbb{I}''} (\Omega^{\mathcal{H}_{s_i}}, \mathcal{H}_{s_i}, \mu^{\mathcal{H}_{s_i}}), \bigcup_{i \in \mathbb{I}''} \mathcal{O}_i)$, where \mathbb{I}'' indexes all event representations now in the set \mathcal{T}_{out}

Algorithm 2 Example implementations of S-System functions **successor** and **predecessor**.

function **SUCCESSOR**(s)

return the set of elements in $\mathcal{S}^{\mathcal{X}}$ that are the immediate successors of s (immediate successors of s is the set of smallest elements that are comparable with s and larger than s , though themselves are not comparable)

end function

function **PREDECESSOR**(s)

return the set of indices of elements in $\mathcal{S}^{\mathcal{X}}$ that are the immediate predecessors of s (immediate predecessors of s is the set of smallest elements that are comparable with s and smaller than s , though themselves are not comparable)

end function

3.3 Related works

3.3.1 Hierarchy

The idea that the data space that NNs process is hierarchically structured and NNs are only operating in a rather small subset of the space, has been more or less a folklore by the researchers in the neural network community. However, the wide recognition of hierarchy has come late, mostly because the seminal work by Krizhevsky et al. (2012) that proves the significance of hierarchy in NNs experimentally. The hierarchy is mostly motivated by the imitation of biological neural networks (Fukushima (1980) Riesenhuber and Poggio (1999) Riesenhuber and Poggio (2000)), where neuroscience shows that it has a hierarchical organization (Kruger et al. (2013)), and does not make the connection to the hierarchy in nature, which is reasonable since at the time NNs/Perceptron (Rosenblatt (1958)) was invented, the Complex System (Simon (1962) Amderson (1972)) that studies the hierarchy in nature did not exist yet. The connection between hierarchy in nature and NNs has been discussed qualitatively by physicists (Lin and Tegmark (2017) Mehta and Schwab (2014)), though to the best of our knowledge, a fully measure-theoretical characterization of the hierarchy in the data space, described in section 3.1 does not exist before. It gives a theoretical motivation of a hierarchically built hypothesis space, i.e., S-System, contrary to the motivation of artificial NNs, which is an imitation.

3.3.2 Hierarchical Hypothesis Space of NNs

Many works have been studying the hierarchical structure of the hypothesis space of NNs. Though perhaps surprisingly, an informal idea similar with S-System has been underlying the design of CNN (Lecun et al. (1998)) at the beginning, where in the unpublished report Bottou et al. (1996), they describe that it is better to defer normalization as much as possible since it “delimiting a priori the set of outcomes”, and pass scores as unnormalized log-probabilities. However, perhaps due to a lack of rigor, they removed the discussion in the formal publication. The passing of scores corresponds to the deterministic coupling that transports true measure in the PPMS, while normalization corresponds to assuming a probability kernel to approximate the true measure transported.

Further analysis on the hierarchical behavior of NNs waited for two decades. Early pioneers analyzes from the perspective of kernel space and harmonics. At the end of the dominant era of support vector machine (SVM), Smale et al. (2009) seeks to give NNs a theoretical foundation in Reproducible Kernel Hilbert Space (RKHS) (Vapnik (1999) Scholkopf and Smola (2001)), which is an analogy but may only give limited insights. We will discuss how RKHS relates to S-System later when we discuss the difference between S-System and RKHS based nonlinear algorithms. Many works in this direction have been done, either taking NNs as a recursively built RKHS (Daniely et al. (2016)), or applying the recursion idea to existing kernel methods (Mairal et al. (2014)). We do not aim to cover all kernel works. We envision it as a tool to aid analysis, and design probability kernels in S-System, yet not as the fundamental underpinning. A work (Anselmi et al. (2016)) in the line of RKHS has also sought foundation in probability measure theory, though its focus is the invariance and selectivity of the one layer representation built by NNs. It studies the measure transport due to compact group transformations, and points out that the output of the activation function of NNs could be the probability distribution of low dimensional projection of the measure of data and its transformations, which is similar to the case where S-System only couples group indicator r.e. — they both analyze the grouping of measure transported by transport maps — though when taking on the hierarchical behavior, it falls back to RKHS, and think recursion as “distributions on distributions” instead of coarse grained probability coupling. We believe the work could be inspirational to further refined analysis on r.e.s created by S-System. Under the umbrella of computational harmonics, Mallat (2012) Mallat (2016) understand NNs as a technique that learns a low dimensional function $\Phi(x), x \in \mathcal{X}$ that linearizes the function $f(x)$ to approximate on complex hierarchical symmetry groups from a high dimensional domain \mathcal{X} . It achieves this by progressively contracting space volume and linearizing transformation that consists of groups of local symmetries layer by layer. However, the group formalism used is an analogy that only rigorously characterizes Scattering Network (Mallat (2012)), a hierarchical hypothesis space simplified from NNs, and does not characterizes NNs. The group formalism is referred as the “mathematical ghost” in Mallat (2016). We believe these works are important to further incorporate symmetry structure in nature in S-System in future works.

More recently, Ankit B. Patel et al. (2016) interprets NNs in Probabilistic Graphical Model (PGM). It takes activation as log-probabilities that propagate in the net. As the description suggests, it confuses the transported measure to be approximated, and the approximated probability obtained by a probability kernel. Thus, it has to rely on the Gaussian assumption to justify the interpretation, of which the mean serves as templates, and the noise free assumption to justify ReLU activation function. Also, the assumption makes it a generative model that has to make assumptions on the data distribution, while an S-system is able to only make assumptions on how measure is supposed to group. From the spline theory perspective, Balestrierio and Baraniuk (2018) understands

NNs as a composition of max-affine spline operators, which implies NNs construct a set of signal-dependent, class-specific templates against which the signal is compared via an inner product. From S-System point of view, it is an analysis on the functional form of coupled r.e.s of an S-system that assumes compositional exponential probability kernels and does maximal estimation on group indicator r.e.s. It connects more with the function approximation results, that takes “signal-dependent” as a fact to see what that implies, than the goal of S-System, i.e., giving a theoretical formal definition and interpretation to NNs. We think it may contribute to the refined analysis of decision boundaries in S-System in the future. Analogizing with statistical mechanics, Trevisanutto (2018) takes the group indicator r.e.s. with binary values as gates, of which the expectation will multiply with the coupled r.e.s. to decide how much the “computation” done should be passed on to next layers. However, what is being computed is left unspecified. As in the definition of S-System, the computation is to extend the probability measure space of the measurement collection r.e. that aims to approximate probability measure of events in the event space of PPMS. The group indicator r.e.s. is not a gate, but serves to group measure. It behaves like a gate when its value is binary, yet underlying it serves to create further coupling of grouped measure. Thus, the analog does not unveil the deeper principles underlying, e.g., probability measure space extension and the probability estimation/learning happening in S-System (refer to section 3 section 5).

3.3.3 Machine Learning Algorithm Paradigm

We envision S-System as an attempt that tries to investigate a measure-theoretical foundation of algorithmic modeling methods (Breiman (2001)) for designing machine learning algorithms. Now we can see NNs as an implementation of S-System, which is a way to transport, group and approximate probability measure. From S-System, we can see that we do not need to make assumptions on the distribution of data to justify that our model is probabilistic — the randomness comes from the data source itself, and it is the probability measure space that a model is manipulating, not the probability values. Thus, we can break from statistics methods developed ever since Ronald Fisher that has to make assumptions on data, and proceed from there. This measure manipulation paradigm may be a promising candidate to the theoretical issue facing high dimensional data analysis (Donoho (2000)). Thus, we discuss current major algorithm paradigms in machine learning/high dimensional data analysis, i.e., Support Vector Machine with Kernels (SVMK) and Probabilistic Graphical Model (PGM).

It is well known that SVMK can be analogized to a NN with one hidden layer. The hypothesis f of SVM can be expressed as a linear combination of inner product between test samples \mathbf{x}_i and support vectors $f(\cdot) = \sum_i \alpha_i k(\cdot, \mathbf{x}_i)$, where k is the kernel function, and α_i scalars. Writing f in the form of $f(X) = \sum_i \alpha_i k(X, \mathbf{x}_i)$, it can be seen that the hypothesis is actually a deterministic coupling, where f is the transport map. As happening in section 5, the training of SVM is also minimizing a surrogate risk between the true data probability measure and the transported measure, though no probability distributions are ever introduced. The probability kernels in S-System is replaced by a positive semi-definite (PSD) kernel, whose output value is a real number indicating something similar with the coupled probability measure of S-System. This observation may seem surprising, however, it makes much sense when we notice the fact that probability is just a function. SVM is a function approximation techniques designed specifically for the case where the data are of high dimensional, yet the number of samples available is small. To combat the curse of dimensionality, it uses a PSD integral operator (Aronszajn (1950)) that maps the sample to a high dimensional space, which can be taken as templates, and only approximates measure that is in the vicinity of those templates and ignores the rest of the space. The kernel can also be built hierarchically, which is discussed in section 3.3.2. For the time being, S-System does not contain SVMK as a special case, while we envision by properly generalizing the probability kernels in CGE, a large class of algorithms may include SVM.

As for PGM, it is a special case of S-System. As mentioned repeatedly throughout the paper, S-System merely makes assumptions on how measure is supposed to group, without making assumptions on the actual distribution of the data. The learning framework of S-System described in section 5 is actually the same as PGM when only considering the unsupervised case, where assumptions on data distribution have to be made. Thus, S-System is a superset of algorithms including PGM. The graph in PGM is actually a poset. However, the insight comes from where they differ. Relying heavily on the assumptions on the distribution of data, which is in reality unknown, it introduces large model biases, which perhaps is the reason why it alone cannot compete with NNs on complex high dimensional data. Furthermore, S-System is naturally compatible with supervised labels, since hidden variables/group indicator r.e.s map one-to-one to labels, which dictates how measure should be grouped. This point is discussed more thoroughly in section 5, where supervised and unsupervised learning are taken as dual perspectives on the same object.

4 Geometry of NNs (and S-System), and Neural Networks From the First Principle

In the previous section, a mechanism S-System is introduced to transport, group and approximate probability measures that are of interest. It focuses on deriving a mechanism to recognize events through measurements from the first principle. In this section, we will show that Multiple Layer Perceptrons (MLP) (E. et al. (1986)) is an implementation of an S-system. *The derivation serves as a proof of concept, and as an example of S-System*, though we note that all existing NN architectures, e.g., Residual Network (He et al. (2016)), Convolutional Neural Network (Simard et al. (2003)), Recurrent Neural Network (Hochreiter et al. (1997)), Deep Belief Network (Hinton et al., 2006) (Hinton et al. (2006)) could be derived by using different measurable spaces, posets, probability kernels and successor, predecessor functions, along with manifold possibilities of new architectures. In the derivation, we will see classical activation functions emerging naturally. Then, we go further to endow geometry on event representations by defining the proper manifold structure on S-System using information geometry. It enables us to quantitatively prove the benefits of hierarchy that MLPs implement coarse graining that contracts the variations in the lower scale event spaces when creating higher scale event extensions, which plays the same role as RG in physics.

4.1 Theoretical Derivation of Activation Functions and MLPs

Let the CGE in definition 3 be MLPCGE (definition 4), the $\tilde{\mathbf{t}}$ in MLPCGE be obtained by transport map ReLU (definition 5), and the scale poset $\mathcal{S}^{\mathcal{X}}$ be a chain, i.e., a poset where all elements are comparable. By algorithm 1, we would obtain a MLP. The definitions are given in the following.

Definition 4 (MLP Conditional Grouping Extension). *An MLP Conditional Grouping Extension (MLPCGE) is a CGE with the following measurable space and parametric forms of probability kernels*

$$\begin{aligned} (E, \mathcal{E}) &= (\mathbb{D}^n, \mathcal{D}^n) \bigotimes (\mathbb{R}^n, \mathbb{B}(\mathbb{R}^n)), \nu(\mathbf{h}|\mathbf{t}) = e^{\mathbf{h}^T \mathbf{W}^T \tilde{\mathbf{t}}} / \sum_{\mathbf{h}} e^{\mathbf{h}^T \mathbf{W}^T \tilde{\mathbf{t}}} \\ \hat{Q}(T, \hat{H}) &= \hat{q}(\mathbf{t}, \hat{\mathbf{h}}) := e^{\mathbf{1}^T \hat{\mathbf{h}}(\mathbf{t})} / \left(\int e^{\mathbf{1}^T \hat{\mathbf{h}}(\mathbf{t})} d\mu(\mathbf{t}) \right) = e^{\mathbf{1}^T \mathbf{W}^T \tilde{\mathbf{t}}} / \left(\int e^{\mathbf{1}^T \mathbf{W}^T \tilde{\mathbf{t}}} d\tilde{\mu}(\tilde{\mathbf{t}}) \right) \\ Q((T, H), \hat{H}) &= q((\mathbf{t}, \mathbf{h}), \hat{\mathbf{h}}) := e^{\mathbf{h}^T \hat{\mathbf{h}}(\mathbf{t})} / \left(\int e^{\mathbf{h}^T \hat{\mathbf{h}}(\mathbf{t})} d\nu(\mathbf{h}|\mathbf{t}) d\mu(\mathbf{t}) \right) = e^{\mathbf{h}^T \mathbf{W}^T \tilde{\mathbf{t}}} / \left(\int e^{\mathbf{h}^T \mathbf{W}^T \tilde{\mathbf{t}}} d\nu(\mathbf{h}|\mathbf{t}) d\tilde{\mu}(\tilde{\mathbf{t}}) \right) \end{aligned}$$

where \mathbb{D}^n is a n -dimensional discrete-valued field, i.e., $\{0, 1\}^n$ or $\{-1, 1\}^n$, \mathcal{D}^n is the σ -algebra generated by \mathbb{D}^n , \mathbf{W} is a matrix (in this case, the transport map $W(T; \boldsymbol{\theta})$ is the matrix \mathbf{W} and parameters $\boldsymbol{\theta}$ are \mathbf{W}), $\mathbf{t}, \hat{\mathbf{h}}, \mathbf{h}$ are realizable values of r.e.s T, \hat{H}, H , and $\tilde{\mathbf{t}}$ is obtained by applying a yet unspecified transport map on \mathbf{t} — for now, it could be just taken as the output of an identity mapping and other possible forms are introduced when discussing activation functions — and $\tilde{\mu}(\tilde{\mathbf{t}})$ is the law on $\tilde{\mathbf{t}}$ induced by the law $\mu(\mathbf{t})$ on the input r.e. of MLPCGE. The meaning of the rest of the symbols is same with those in definition 2.

Note that it is not possible to compute $\hat{q}(\mathbf{t}, \hat{\mathbf{h}})$, for $\mu(\mathbf{t})$ is unknown. However, we can compute ν faithfully! This is because H is a manual creation/grouping instead of inherent events in PPMS. Here, with some further reasoning, we will have the marvelous trick done by NNs, i.e., the Activation Function (AF). The key is only to build a full, or partial CGE upon r.e.s created by a previous CGE, using an estimated value of H . The deeper principles of the estimation are described in section 5, which is the maximization of expected data log likelihood, and is part of the learning framework of S-System. When a full CGE is created upon output r.e.s. of a previous CGE, \mathbb{D} is $\{0, 1\}^n$, and the estimation is done through expectation or maximum, we recover the currently best performing activation function Swish (Ramachandran et al. (2017)) or ReLU (Glorot et al. (2011)) respectively; when a partial CGE is created on the group indicator r.e.s, the estimation is done through expectation, and \mathbb{D} is $\{0, 1\}$ or $\{1, -1\}$, we recover classical activation functions Sigmoid or Tanh respectively.

We derive ReLU as an example. The group indicator r.e.s H divides the measure transported from the event space of input r.e. T to the event space of \hat{H} into groups. Intuitively, if H divides the measure into two groups indexed by elements of \mathbb{D} , and we assume 1 collects the measure corresponds to an event collection while 0 collects the complement of the event collection (meaning the event collection does not occur), given a realization \mathbf{t} of T , to recognize higher scale events composed by lower scale events represented by H , we would like to *estimate* what events are present in \mathbf{t} , and create *further coupling* with another CGE on the events that are present. Formally,

Definition 5 (Rectified Linear Unit (ReLU)). Let $T := (H, \hat{H})$ be r.e.s created by a CGE, an estimation $\tilde{\mathbf{h}}$ of a realization of the r.e. H is obtained by

$$\tilde{h}_i = \arg \max_{h \in \mathbb{D}} \nu(h|\mathbf{t}) = e^{h\mathbf{W}_{:i}^T \tilde{\mathbf{t}}} / \sum_h e^{h\mathbf{W}_{:i}^T \tilde{\mathbf{t}}}$$

A further coupling is created by MLPCGE upon T , of which $\tilde{\mathbf{t}}$ is the estimated realization of the r.v. obtained by applying a transport map ReLU on T

$$\text{ReLU}(T) := H \odot \hat{H}$$

where \odot denotes Hadamard product. Operationally, ReLU is a binary mask \mathbf{h} applying on the outputs (preactivation) $\hat{\mathbf{h}}$ of the transport map \mathbf{W} . The i th dimension h_i of \mathbf{h} can be estimated separately because h_i, h_j are conditionally independent conditioning on \mathbf{t} (Hinton et al. (2006)).

As can be seen, AFs is not a well defined object, which is actually a combination of operations from two stages of computation.

4.2 NN Manifold and Contraction Effect of Conditional Grouping Extension

In Section 3, motivated by the hierarchy assumption 3.2, we designed S-System. Here, using MLP as an example, and also a proof of concept, we quantitatively show the benefits of hierarchical grouping done in S-System by showing that irrelevant/uninterested variations in the lower scale events could be gradually dropped by repeatedly applying CGE, characterized by “shrinking distance” between events.

To characterize the distance, we need a geometry structure on event representations. We give an initial construction built on information geometry (Amari (2016)). For a review of manifold and information geometry, please refer to appendix B C. To begin with, we define

Definition 6 (Compositional Exponential Family of Distributions). The form of **compositional probability distribution of exponential family** is given by the probability density function

$$p(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta}) d\mathbf{x} d\mathbf{h} = e^{(k(\mathbf{h}, \mathbf{x}; \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}))} d\mu(\mathbf{x}) d\nu(\mathbf{h}), k(\mathbf{h}, \mathbf{x}; \boldsymbol{\theta}) = \langle \mathbf{f}(\boldsymbol{\theta}; \mathbf{h}), \mathbf{g}(\mathbf{x}) \rangle$$

where \mathbf{x} is realizable values of a multivariate random variable, k is a function called **compositional kernel** that for a given \mathbf{h} , k is the inner product between certain vector function $\mathbf{g}(\mathbf{x})$, called **sufficient statistic**, (of which the component functions are linearly independent) and certain vector function $\mathbf{f}(\boldsymbol{\theta}; \mathbf{h})$, called **composition function**, $\psi(\boldsymbol{\theta})$ is the normalization factor, and μ, ν are the laws on r.v. \mathbf{x}, \mathbf{h} , respectively.

Conditioning on \mathbf{h} , $p(\mathbf{x}|\mathbf{h}; \boldsymbol{\theta}) = e^{k(\mathbf{h}, \mathbf{x}; \boldsymbol{\theta})} d\mu(\mathbf{x})$ is of the exponential family. Actually, it is of Curved Exponential Family (Amari (1995)). The parametric form of kernel Q of MLPCGE is of the compositional exponential family, where $k(\mathbf{h}, \mathbf{x}; \boldsymbol{\theta}) = \langle \mathbf{h}^T \mathbf{W}, \mathbf{x} \rangle$, $\mathbf{f}(\boldsymbol{\theta}; \mathbf{h}) = \mathbf{h}^T \mathbf{W}$, $\mathbf{g}(\mathbf{x}) = \mathbf{x}$.

Definition 7 (Neural Network Manifolds). Let \mathcal{M} be an event representation built on a measurement collection r.e. X through an S-system. If probability kernels Q, \hat{Q} of all CGE in the S-system are of the compositional exponential family, of which the composition kernel is parameterized by the CGE transport map, then the function space M_s of measure $\mu_s^{\mathcal{H}}(\hat{H}, T|H)$, $s \in \mathcal{S}^X$, where \mathcal{S}^X is the scale poset of the S-system, is a Riemannian manifold with the following properties:

- M_s has a coordinate system $\boldsymbol{\eta}_s|_{\mathbf{h}} = (\eta_1, \dots, \eta_n)$ that is the dual affine coordinate system of an exponential family distribution, where

$$\boldsymbol{\eta}_s|_{\mathbf{h}} := \nabla_{\mathbf{f}(\boldsymbol{\theta}; \mathbf{h})} \psi(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{t}|\mathbf{h}] = \int \mathbf{t} d\mu_s^{\mathcal{H}}(\hat{H}, T|H) = \int \mathbf{t} q((\mathbf{t}, \mathbf{h}), \hat{\mathbf{h}}) d\mu(\mathbf{t})$$

$\boldsymbol{\theta}$ is the parameters of the transport map $W(T; \boldsymbol{\theta})$, $\nabla_{\mathbf{f}(\boldsymbol{\theta}; \mathbf{h})}$ takes derivatives w.r.t. composition function of k and \mathbf{t} is realizations of T . We call the coordinates **neuron coordinates**.

- M_s has a Riemannian metric derived by the second order Taylor expansion of the dual Bregman divergence defined by

$$D_{\psi^*}[\boldsymbol{\eta}'_s|_{\mathbf{h}} : \boldsymbol{\eta}_s|_{\mathbf{h}}] := \psi^*(\boldsymbol{\eta}'_s|_{\mathbf{h}}) - \psi^*(\boldsymbol{\eta}_s|_{\mathbf{h}}) - \nabla \psi^*(\boldsymbol{\eta}_s|_{\mathbf{h}})^T (\boldsymbol{\eta}'_s|_{\mathbf{h}} - \boldsymbol{\eta}_s|_{\mathbf{h}})$$

where ψ^* is the Legendre dual of ψ . We call the divergence defined **neuron divergence**.

In the above definition, the stochastic manifold is defined by conditioning on group indicator r.e.s H . To appreciate the definition, let's return back to MLP. Let \mathcal{M} be the event representation built on a measurement collection r.e. X by a MLPCGE, i.e., the measure of output r.e.s being $\mu^{\mathcal{H}} = e^{\mathbf{h}^T \mathbf{W}^T \mathbf{x} - \psi(\mathbf{W})} \mu(\mathbf{x}) \nu(\mathbf{h}|\mathbf{x})$. When \mathbf{h} is fixed, letting $\mathbf{f}_0 = \mathbf{h}^T \mathbf{W}^T$ we have $\mu^{\mathcal{H}}(\mathbf{x}|\mathbf{h}) = e^{\mathbf{f}_0^T \mathbf{x} - \psi(\mathbf{W})} \mu(\mathbf{x})$. It is known (Nielsen and Garcia (2009)) that the expectation statistics, i.e., $\boldsymbol{\eta}|\mathbf{h} = \nabla_{\mathbf{f}_0} \psi(\mathbf{f}_0(\mathbf{W}))$, uniquely determines $\mu^{\mathcal{H}}(\mathbf{x}|\mathbf{h})$. It implies that given \mathbf{h} , $\mu^{\mathcal{H}}$ is a probability distribution, of which the most "salient" feature is the expectation. This explains why the visualization of NN representations tends to be templates (Mahendran and Vedaldi (2015) Zhang and Zhu (2018)), and the template based theories (Riesenhuber and Poggio (1999) Ankit B. Patel et al. (2016) Balestrieri and Baraniuk (2018)) are partially right. Thus, the group indicator \mathbf{h} represents the events of \mathcal{M} , of which the expectation is the representative. Let the transport map of \mathcal{M} be $W : \mathbf{x} \mapsto \mathbf{h}^T \mathbf{W}^T \mathbf{x}$, $\mu^{\mathcal{H}}(\mathbf{x}|\mathbf{h})$ approximates the measure $\mu^{\mathcal{W}}(X^{-1}W^{-1}(A))$, $A \subseteq \mathbb{R}^n$ in PPMS. That's why instead of using the canonical coordinate of exponential family distribution, we use its dual affine coordinate. Though essentially the two coordinate systems are dual views on the same object, we define the manifold this way to characterize the fact that for a given NN, D_{ψ^*} characterizes the degree of separation between two events $A, A' \in \Omega$ of which the probability measures $\mu^{\mathcal{W}}(A), \mu^{\mathcal{W}}(A')$ are transported by $W(X(A)), W(X(A'))$ and approximated by $\mu^{\mathcal{H}}$. Furthermore, the divergence is defined by conditioning reflects the fact events can be compared using multiple criteria, though to evaluate its implication more works are needed. For how the above definitions relate to classical definitions on NNs in information geometry, please refer to section 4.3.

As a last remark, in the above definition, we only define the function space of $\mu_s^{\mathcal{H}}(\hat{H}, T|H)$ is a manifold without touching the overall structure of their interaction, which we believe a much richer structure lies in, and is an important future direction.

By a directed application of theorem 14 of Liese and Vajda (2006), which is called information monotony in Amari (2016), we have

Theorem 4.1 (Contraction of divergence between events). *Let $\mathcal{A} := \mathcal{A}' \cup \mathcal{A}''$ be an event collection in event space Ω of PPMS consisting of two event collections, and two measurement collection r.e.s X', X'' are created for $\mathcal{A}', \mathcal{A}''$ respectively. Let \mathcal{S} be an S -system, $\mathcal{M}', \mathcal{M}''$ be event representations built on X', X'' by \mathcal{S} respectively, and $\mathcal{S}^{\mathcal{X}}$ be the scale poset of \mathcal{S} . Then $\forall s_1, s_2 \in \mathcal{S}^{\mathcal{X}}, s_1 < s_2$, we have*

$$D[\boldsymbol{\eta}'_{s_1}|\mathbf{h}_1 : \boldsymbol{\eta}''_{s_1}|\mathbf{h}_1] \geq D[\boldsymbol{\eta}'_{s_2}|\mathbf{h}_2 : \boldsymbol{\eta}''_{s_2}|\mathbf{h}_2]$$

where $\boldsymbol{\eta}'_{s_1}, \boldsymbol{\eta}'_{s_2}$ are the neuron coordinates at scale s_1, s_2 of \mathcal{M}' respectively; so are $\boldsymbol{\eta}''_{s_1}, \boldsymbol{\eta}''_{s_2}$ of those of \mathcal{M}'' ; $\mathbf{h}_1, \mathbf{h}_2$ are arbitrarily fixed realizations of group indicator r.e.s at scale s_1, s_2 respectively.

Example 4.1 (Contraction of divergence induced by deformation). *Let g be a diffeomorphism group, and $X' = g.X$, the deformed r.e. created by applying g on a r.e. X . By the above theorem, for event representations created by an S -system, coordinated as $\boldsymbol{\eta}|\mathbf{h}, \boldsymbol{\eta}'|\mathbf{h}$, their distance is gradually contracted in term of neuron divergence. For a review of diffeomorphism group, refer to appendix B.*

The theorem has twofold significance. First, it shows that a recursive application of CGE would shrink the discrepancy between events, thus possessing the capacity to contract irrelevant variations in the events, though further characterizations are needed to give operational guidance. It completes the incomplete analog between NNs and RG (Lin and Tegmark (2017)), which lacks a physical quantity to renormalize to produce large scale properties. The physical quantity is shown as the neuron divergence between event representations, or more informally, semantic difference between samples, *though we note essentially the large scale quantity is group indicator r.e.s of high scales that represent events that gradually have semantic meaning, of which the neuron divergence is a property*. Contrary to clear-cut physical quantities like temperature emerging in physics through RG, *a meaningful event group emerges through learning, which leads us to the section 5.6*. Second, along with definition 7, it identifies the proper object if the geometry of NNs is to be studied. For example, to study the symmetry in the geometry, the object to investigate is the symmetry in the event space, of which the diffeomorphism group is a type of symmetry, and invariance is the mapping of events to the same neuron coordinates. This is in contrast with existing works that study symmetry by studying the equivariance (Cohen and Welling (2016) Dieleman et al. (2016)), or invariance (Anselmi et al. (2016)), or linearization of diffeomorphism (Mallat (2016)) in NNs through studying the changes induced by group actions in feature maps in the intermediate layers of a sample, which is a rather ad hoc object. The event space perhaps is the "mathematical ghost" lurking in Mallat (2016).

4.3 Related works

In the related works on hierarchical hypothesis space discussed earlier, all of them have their own geometry, we only discuss related works in this subsection that are related to the geometry defined in section 4.2.

Most of the works we are aware of that try to endow a geometry on NNs through information geometry (IG) were done before the deep learning era, not surprisingly, by Amari, who developed IG. All the works study NNs with a single hidden layer. Amari and Nagaoka (2007) formulates the manifold parameterized by all parameters of a NN as neuromanifold, while in section 4.2, the manifold we defined focuses on the submanifolds indexed by a scale poset, which will be discussed more in the next paragraph. Actually, the neuromanifold is the stochastic manifold consisting of possible probability measure on the random element set of the event representation built by S-System. Two directions of analysis have been made. The first is to study the behavior of the curved exponential families obtained by conditioning, which is done in Amari (1995), and falls in the category of generative training. The other is to study supervised trained NNs, and study the neuromanifold, with a focus on the impact of singularities on training dynamics (Amari et al. (2006)). The later proposed the Natural Gradient Descent methods, and many works have been working on it thereafter, which we will not discuss. The study on the hierarchy has been limited on decomposition of high order interactions in a single hidden layer NN (Amari (2001)) without attacking the recursion in NNs, though we tend to think NNs with more layers unroll higher order interactions, but we do not find that they pursue this path. As mentioned, the study of hierarchical behaviors of NNs has been absent, which is the focus of this paper, and is emphasized in the paragraph below.

The geometry defined in section 4.2 is to investigate the hierarchical geometry of NNs. The compositional exponential family gives the definition of Neural Network Manifolds that properly identifies the curved exponential families, or in other words, submanifolds, in a probability family built by the overall parameter space of NNs, which is complicated, e.g., containing singularities (Amari et al. (2006)). Note that we do not differentiate submanifolds with manifolds in the main content to avoid clutters. As discussed, the submanifolds are well represented by their expectation statistics, and the definition identifies how coarse graining in divergence happens in theorem 4.1. Thus, definitions given are distinctive in characterizing the hierarchical geometry of NNs, which is absent in previous works, which either stay in the realm of single hidden layer (Amari (1995) Amari (2001)), or take the whole parameter space as the parameterization of a manifold that contains singularities (which rigorously is not a manifold) (Amari et al. (2006)), though we are well aware that the works present in this paper are merely scratching the surface. Our focus for now is merely to show the coarse graining contraction effect of CGE quantitatively, and much more works are to be done, e.g., the hierarchical and within-layer interactions between these submanifolds. As a concrete example, it is known that the EM algorithm has an IG interpretation (Amari (1995)). The expectation, KL divergence minimization interpretation of the back propagation algorithm in section 5 can be interpreted similarly from the IG perspective. Thus, Amari (1995) can be generalized to NNs with arbitrary number of layers, and in generative or supervised training settings. It implies the two directions mentioned in the previous paragraph can be unified, though further analysis on its impact, e.g., the analysis of singularities, needs more works.

Very recently, at the time of writing this paper, a few reports have been submitted on the archive that try to attack the supervised deep NNs (Amari et al. (2018a) Amari et al. (2018b)). But again, they follow their old idea that analyzes the whole neuromanifold. It assumes weights and biases of NNs to be Gaussian, and study how properties related to the distribution of activations of each layer change, e.g., fisher information matrix, without trying to formally define the geometry, or the submanifold structure in the intermediate layers of NNs.

Lastly, we note that Lin and Tegmark (2017) also tries to discuss the coarse graining effect in term of the information monotony phenomenon as “information distillation”, but it does that rather generally and qualitatively, does not put the phenomenon in an exact NN context, and not make the connection between it and the geometry in information.

5 Learning Framework of S-System (and NNs)

5.1 Learning framework

The previous sections defines S-System \mathbf{S} as a mechanism that given a measurement collection X , \mathbf{S} uses group coupling probability kernels to approximate the probability measure transported from X through the transport maps. Both the transport maps and the kernels are parameterized. To ensure that the approximation is relevant, a mechanism is needed to determine the parameters. In other word, how possibly can we ground the probability approximation created by coupling on “reality”? The learning framework that determines the parameters through optimizing an objective function, is present in this section. As a remark, supervised discriminative learning and unsupervised generative learning are two special cases of the learning objective present, and forward propagation is to maximize the expected data log likelihood, while back propagation is to minimize the KL divergence.

Let $\mathcal{M}^{\mathcal{W}}$ be an event representation built by an S-system with scale poset $\mathcal{S}^{\mathcal{Z}}$ built on a measurement collection r.e. $Z := (X, Y)$ with measure $\mu^{\mathcal{Z}}$ supported on the PPMS \mathcal{W} . The measurable space \mathcal{Z} is a product space $\mathcal{X} \times \mathcal{Y}$,

where \mathcal{X} denotes the data space, and \mathcal{Y} denotes label space. Let \mathcal{M}_s be the event representation built at scale s . Let $\mu_s^{\mathcal{H}}, s \in \mathcal{S}^Z$ be the probability measure on output r.e.s (H_s, \hat{H}_s) of \mathcal{M}_s , and ν_s the law on conditional group indicator r.e.. The learning of S-System is to minimize the discrepancy between measure $\mu^{\mathcal{W}}(X^{-1}(W_s^{-1}(A)))$ and $\mu_s^{\mathcal{H}}(A)$ assigned to a event $A \subset \Omega^{\mathcal{H}_s^e}$, where $\Omega^{\mathcal{H}_s^e}$ is the event space of the probability measure space of \mathcal{M}_s , and W_s is the transport map of the coupling probability kernel of \mathcal{M}_s . One way to characterize the discrepancy is Maximum Likelihood Estimation (MLE), where the parameters that most likely to generate the data consist the best estimator. The likelihood function of $\mathcal{M}^{\mathcal{W}}$ is

$$\begin{aligned} p(X; \theta) &= \sum_{\mathcal{O}^{\mathcal{W}} \setminus X} p(\mathcal{O}^{\mathcal{W}}; \theta) \\ &= \prod_{s_L \in \mathcal{S}_L} \sum_{\mathbf{h}_{s_L}} \int_{\hat{\mathbf{h}}_{s_L}} \sum_{\mathbf{h}_{s'}, s' \in p(s_L)} \int_{\hat{\mathbf{h}}_{s'}, s' \in p(s_L)} \sum_{\mathbf{h}_{s''}, s'' \in p(s')} \int_{\hat{\mathbf{h}}_{s''}, s'' \in p(s')} \dots \\ &\quad \mu_{s_L}^{\mathcal{H}}(H_{s_L}, \hat{H}_{s_L} | H_{p(s_L)}, \hat{H}_{p(s_L)}) \prod_{s' \in p(s_L)} \mu_{s'}^{\mathcal{H}}(H_{s'}, \hat{H}_{s'} | H_{p(s')}, \hat{H}_{p(s')}) \prod_{s'' \in p(s')} \dots \mu^{\mathcal{Z}}(X) \end{aligned}$$

where $\mathcal{O}^{\mathcal{W}}$ and θ are the r.e. set and the parameters of $\mathcal{M}^{\mathcal{W}}$ respectively, \mathcal{S}_L denotes the set of largest element in \mathcal{S}^Z , and $p(s)$ denotes the elements in \mathcal{S}^Z that are the predecessors of s . Depending on whether H_s, \hat{H}_s are discrete or not, the summation may be changed to integral, vice versus. It could be understood as getting the marginal probability distribution of X from a factorized probability of a direct acyclic graph in probabilistic graphical model (PGM) (Koller and Friedman (2009)).

Needless to say, the likelihood function is intractable when the r.e. set $\mathcal{O}^{\mathcal{W}}$ gets large. Perhaps more importantly, we do not know $\mu^{\mathcal{Z}}(X)$, so we do not know $\mu_s^{\mathcal{H}}$ since it is built on the transport map applied on X . Thus, to make the estimation tractable, and to faithfully estimate measure on events groups already seen without making assumptions on $\mu^{\mathcal{Z}}(X)$, we make the following decomposition of the log likelihood function to focus on estimating measures on group indicators r.e.s

$$\ln p(\mathcal{O}^{\mathcal{W}}; \theta) = \mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta) + \sum_{s \in \mathcal{S}^Z} D_{\text{KL}}^s(\nu_s || \mu^{\mathcal{X}}(H_s | W_s(X)))$$

where

$$\begin{aligned} \mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta) &= \sum_{\mathbf{h}_s, s \in \mathcal{O}^{\mathcal{W}} \setminus X} q(\mathcal{O}^{\mathcal{W}} \setminus X) \ln \frac{\mu^{\mathcal{Z}}(X)}{q(\mathcal{O}^{\mathcal{W}} \setminus X)} \\ q(\mathcal{O}^{\mathcal{W}} \setminus X) &= \prod_{s_L \in \mathcal{S}_L} \mu_{s_L}^{\mathcal{H}}(H_{s_L} | \hat{H}_{s_L}) \prod_{s' \in p(s_L)} \mu_{s'}^{\mathcal{H}}(H_{s'} | \hat{H}_{s'}) \prod_{s'' \in p(s')} \dots \\ &= \prod_{s_L \in \mathcal{S}_L} \nu_{s_L} \prod_{s' \in p(s_L)} \nu_{s'} \prod_{s'' \in p(s')} \dots \\ D_{\text{KL}}^s(\nu_s || \mu^{\mathcal{X}}(H_s | W_s(X))) &= \sum_{H_s} \nu_s \ln \frac{\mu^{\mathcal{X}}(H_s | W_s(X))}{\nu_s} \end{aligned} \tag{1}$$

Note that $\mu_s^{\mathcal{H}}(H_s | \hat{H}_s)$ is used instead of $\mu_s^{\mathcal{H}}(H_s | \hat{H}_s, \hat{H}_{p(s)})$ because in CGE, H_s is conditional independent with previous output r.e.s given \hat{H}_s . $\mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta)$ is called *expected data log likelihood*, $\ln p(\mathcal{O}^{\mathcal{W}}; \theta)$ the *complete data log likelihood* and $D_{\text{KL}}^s(\nu_s || \mu^{\mathcal{X}}(H_s | W_s(X)))$ is the *KL divergence at scale s* between estimated measure and true measure.

The decomposition has been used widely in PGM (Bishop (2006)). Successful techniques derived from it have been invented known as Variation Inference and Expectation Propagation etc. Yet, one remarkable difference in the above decomposition and existing decomposition is that here we decompose the probability measure on physical events in APMS $\tilde{\mathcal{W}}$, and estimate measure ν_s that aims to approximate the measure of groups of events in the event space of the PPMS \mathcal{W} , while in existing decomposition, their approaches are to hallucinate some parametric probabilistic models on $\mathcal{O}^{\mathcal{W}}$ (under the context of S-System), and because the “exact” inference is intractable, they use the decomposition to make the inference tractable. In essence, we are not making any assumptions on Z , but only on how they are supposed to group together, while existing approaches using the decomposition is solely about making assumptions on Z , and how to make the computation tractable, thus likely leading to significant model biases.

With the above decomposition, we can see what the training of a supervised NN is. Forward propagation (FP) is to estimate values of group indicator r.e.s by assigning H_s a value that maximizes the expected data log likelihood

$\mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta)$ w.r.t. $q(\mathcal{O}^{\mathcal{W}} \setminus X)$ through activation function (though depending on the activation function chosen, it may not always reach the maximum), while holding θ fixed. BP is to minimize the KL divergence D_{KL}^s at scale s w.r.t. θ whenever there is a supervisory information/label on H_s supervising how the events are supposed to group, while holding $q(\mathcal{O}^{\mathcal{W}} \setminus X)$ fixed.

The decomposition not only includes supervised NNs, but also includes variational autoencoder (VAE) (Welling (2014)), where further assumptions on probability measure of X are assumed. When absent of labels, a normal distribution on H_s is assumed, thus encouraging each group indicator r.e. to learn a grouping that is supposed to be disentangled with the rest. When some labels exist, we recover semi-supervised VAE.

Thus, supervised learning is never something that stands on its own, so is unsupervised learning. They are two perspectives to look at the same thing, or they are Yin and Yang of the Tao in Taoism (or The Book of Change), or the thesis and anti-thesis of dialectics. They are different ways with different assumptions to get information to approximate the measure of events groups, e.g., the group indicator r.e.s in S-System, which represents what has been recognized. Even pure supervised learning can do some unsupervised learning — by pushing KL divergence D_{KL}^s at some scales to zero, the expected data log likelihood will be closer to the complete data log likelihood. This partly explains the emergence of generic feature in NNs (though the maximization of $\mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta)$ perhaps is the main reason). So pure unsupervised learning can do some supervised learning — the maximization of $\mathcal{L}(\{\nu_s\}_{s \in \mathcal{S}^Z}, \theta)$ leads to a smaller KL divergence. We do not observe it in an obvious way in experiments because the grouping does not necessarily concur to the grouping we humans already have. By imposing some structure on the grouping scheme, e.g., imposing a normal distribution, we can discover manifolds that groups events that make sense, e.g., facial expression or digit variations (Welling (2014)).

Lastly, we note Bayesian aspects can be further included in the learning framework by endowing assumptions further on the parameter space.

5.2 Related works

The learning framework is an application of a general probability estimation framework on the particular case of S-System, thus, the reader may find the learning framework similar with variational inference widely used in existing probabilistic graphical models. However, the similarity lies in the fact that both S-System and PGM approximate probability, of which the decomposition of complete data likelihood is about probability to be estimated, not about specific hypotheses in use. The difference between S-System and PGM has been detailed in section 3.3.3. Previously, the BP algorithms have mostly been viewed as a heuristic tool, instead of having a theoretically rigorous derivation. The learning framework of S-System shows that the FP and BP are actually maximizing the complete data likelihood, and are not merely minimizing the discrepancy between the estimated conditional probability of labels given data with the true conditional probability through empirical risk minimization, but also maximizing the expected data likelihood through activation function.

6 Optimization Landscape of NNs (and S-System)

In the previous section, we formulate the learning objective of S-System to identify the parameters to hierarchically group and approximate measures faithfully. Yet, we still do not know whether optimization of the objective function is possible. This section addresses the issue by studying the optimization landscape of it.

We will show that despite possessing the normally undesirable complexity, non-identifiability and singularity (Amari (2016)) properties, S-System could be marvelously powerful. Stating in a more familiar language, the problem translates to how a many latent variable model is able to learn? This is the long standing optimization issue of NNs. We aim at investigating the principle underlying instead of proving the most general case. More specifically, when a set of boundedness and diversity conditions hold, we show that a NN can approximate the probability distribution of any binary group indicator r.e. given empirical measure of the r.e.; or in other words, for a class of losses, including the hinge loss, and a class of NNs, including MLP and CNN, we prove that all local minima of the empirical risk function are global minima with zero loss values.

The problem is formulated as the following. Let $\mathcal{M}^{\mathcal{W}}$ be an event representation built by an S-system \mathbf{S} on a measurement collection r.e. Z supported on the PPMS \mathcal{W} ; the measurable space \mathcal{Z} and measure on Z are $\mathcal{X} \times \mathcal{Y}$ and $\mu^{\mathcal{Z}}$ respectively, where $\mathcal{X} := \mathbb{R}^n, \mathcal{Y} := \{-1, 1\}$. Let the scale poset of \mathbf{S} be a chain, symbolically represented as an integer set $\mathcal{S}^{\mathcal{X}} = \{0, \dots, L\}, 0 < \dots < L$, and $(H_l, \hat{H}_l), l \in \mathcal{S}^{\mathcal{X}}$ the output r.e.s of $\mathcal{M}^{\mathcal{W}}$. A reward-penalty mechanism is introduced to give feedback on the “faithfulness” of approximated measure $\mu_L^{\mathcal{H}}$ as a discrepancy measure between $\nu(H_L|H_{\{1, \dots, L-1\}}, X)$ and $\mu^{\mathcal{Z}}(Y|X)$, where $(X, Y) \in \mathcal{Z}$. *We can see that supervised learning actually uses the group indicator r.e. H_L to approximate the grouping of samples arbitrated by labels. In a certain*

way, it formalizes semantics. The problem formulation is a part of the general learning framework of S-System described in section 5, a.k.a., the term $\nu_L \ln \frac{\mu^{\mathcal{X}(H_L|W_L(X))}}{\nu_L}$ in eq. (1). We only investigate well how the probability of a high level group indicator r.e. H_L can be approximated here.

The problem setting above is principally the same with the formulation of a binary supervised learning problem in statistic learning theory (SLT). For a classic-style formulation, the readers may refer to appendix D. The key insight of SLT is that instead of seeking a fully probabilistic formulation, the discrepancy can be formulated as an empirical risk that measures the discrepancy between H_L and Y , calculated on a set of training samples $\{X_i, Y_i\}_{i=1,\dots,m}$:

$$R(T) = \frac{1}{m} \sum_{i=1}^m l(T(X_i; \theta), Y_i) = \frac{1}{m} \sum_{i=1}^m l(\hat{H}_L(X_i), Y_i) \quad (2)$$

where T here denotes the hierarchical transport map built, l is a loss function, and $\hat{H}_L(X_i)$ is the coupled r.e. built on X at scale L .

Our goal is to investigate the fundamental principle that makes $R(T)$ tractable. To do this, we study the risk landscape by studying the Hessian of $R(T)$ of a particular class of loss functions, of which the eigenvalue spectrum dictates whether critical points of $R(T)$ are local minima, or saddle points. To motivate the class of losses, observing that

$$\frac{d^2}{d\theta^2} l(T(\mathbf{x}; \theta), y) = l''(T(\mathbf{x}; \theta), y) \frac{d}{d\mathbf{x}} T(\mathbf{x}; \theta) \frac{d}{d\mathbf{x}} T(\mathbf{x}; \theta)^T + l'(T(\mathbf{x}; \theta), y) \frac{d^2}{d\mathbf{x}} T(\mathbf{x}; \theta) \quad (3)$$

We study the class \mathcal{L}_0 of functions l and the class of NNs T , such that for $l \in \mathcal{L}_0$, it satisfies: 1) $l : \mathbb{R} \rightarrow \mathbb{R}^+$, when y is taken as a constant; 2) l is convex; 3) the second order derivatives $\frac{d^2}{d\mathbf{x}^2} l$ is zero; 4) $\min_{\mathbf{x}} l(\mathbf{x}, y) = 0$, while for T it satisfies: $\dim(T(\mathbf{x}; \theta)) = 1$. The restriction allows us to study the most critical aspect of the risk function of supervised NNs by making the first term above zero, while the second term a single matrix (instead of an addition of matrices). The class of l includes important loss functions like the hinge loss $\max(0, 1 - \hat{H}_L Y)$, and the absolute loss $|\hat{H}_L - Y|$, which were studied in Choromanska et al. (2015a) under unrealistic assumptions. The class of T is the NN with a single output neuron, which can be written as

$$T(\mathbf{x}; \theta) = \mathbf{x}^T \prod_{i=1}^{L-1} \mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i) \boldsymbol{\alpha} \quad (4)$$

where $\boldsymbol{\alpha}$ is a vector, $\text{dg}(\tilde{\mathbf{h}}_i)$ is the diagonal matrix whose diagonal is the estimated value of H_i , and the meaning of rest symbols is the same as those of MLPCGE in definition 4. Notice that both \mathbf{x} and \mathbf{h}_i are realizations or estimation of r.e.s, thus the Hessian of $T(\mathbf{x}; \theta)$ is a random matrix (Tao (2012)), which implies the Hessian \mathbf{H} of $R(T)$ is a random matrix created by summing random matrices, each of which is a gradient l' multiplies the Hessian of T .

Thus, the problem converts to study the eigen-spectrum of a random matrix \mathbf{H} . The conversion looks straightforward now, but is actually a major obstacle that stops Choromanska et al. (2015a) Jeffrey Pennington (2017), where a confusion about the source of randomness led them astray (the point is discussed in detail in section 6.5). With the following *realistic* assumptions on \mathbf{H} (for a discussion on the *practicality* of the assumptions, refer to section 6.3), we show that l has a surprising benign landscape. Suppose \mathbf{H} is a $N \times N$ matrix, and let

$$\mathbf{A} := \mathbb{E}[\mathbf{H}], \frac{1}{\sqrt{N}} \mathbf{W} := \mathbf{H} - \mathbf{A}, \mathcal{S}[\mathbf{R}] := \frac{1}{N} \mathbb{E}_{\mathbf{W}}[\mathbf{W} \mathbf{R} \mathbf{W}]$$

where the expectation in \mathcal{S} is taken w.r.t. \mathbf{W} while keeping \mathbf{R} fixed — it is a linear operator on the space of matrices.

Assumption 6.1 (Boundedness). 1) $\exists C \in \mathbb{R}, \forall N \in \mathbb{N}, \|\mathbf{A}\| \leq C$, where $\|\mathbf{A}\|$ denotes the operator norm. 2) $\exists \mu_q \in \mathbb{R}, \forall q \in \mathbb{N}, \forall \alpha \in \mathbb{J}, \mathbb{E}[\|\mathbf{W}_\alpha\|^q] \leq \mu_q$, where $\mathbb{J} = \mathbb{I} \times \mathbb{I}$, and $\mathbb{I} = \{1, \dots, N\}$. 3) $\exists C_1, C_2 \in \mathbb{R}, \forall \mathbf{R} \in \mathbb{N}, \epsilon > 0, \|\kappa\|_2^{iso} \leq C_1, \|\kappa\| \leq C_2 N^\epsilon$; 4) $\exists 0 < c < C, \forall \mathbf{T} \succ \mathbf{0}, c N^{-1} \text{tr } \mathbf{T} \preceq \mathcal{S}[\mathbf{T}] \preceq C N^{-1} \text{tr } \mathbf{T}$.

Assumption 6.2 (Diversity). There exists $\mu > 0$ such that the following holds: for every $\alpha \in \mathbb{I}$ and $q, R \in \mathbb{N}$, there exists a sequence of nested sets $\mathcal{N}_k = \mathcal{N}_k(\alpha)$ such that $\alpha \in \mathcal{N}_1 \subset \mathcal{N}_2 \subset \dots \subset \mathcal{N}_R = \mathcal{N} \subset \mathbb{I}, |\mathcal{N}| \leq N^{1/2-\mu}$ and $\kappa(f(\mathbf{W}_{\mathbb{I} \setminus \cup_j \mathcal{N}_{n_j+1}(\alpha_j)}), g_1(\mathbf{W}_{\mathcal{N}_{n_1}(\alpha_1) \setminus \cup_{j \neq 1} \mathcal{N}(\alpha_j)}), \dots, g_q(\mathbf{W}_{\mathcal{N}_{n_q}(\alpha_q) \setminus \cup_{j \neq q} \mathcal{N}(\alpha_j)})) \leq N^{-3q} \|f\|_{q+1} \prod_{j=1}^q \|g_j\|_{q+1}$, for any $n_1, \dots, n_q < R, \alpha_1, \dots, \alpha_q \in \mathbb{I}$ and real analytic functions f, g_1, \dots, g_q , where $\|\cdot\|_p$ is the L^p norm on function space. We call the set \mathcal{N} of α the **coupling set** of α .

Theorem 6.1. Let $R(T)$ be the risk function defined at eq. (2), where the loss function l is of class \mathcal{L}_0 , and the transport map T is a neural network defined at eq. (4). If the Hessian \mathbf{H} of $R(T)$ satisfies assumptions 6.1 6.2, $\mathbb{E}(\mathbf{H}) = \mathbf{0}$, and $N \rightarrow \infty$, then

1. all local minima are global minima with zero risk
2. A constant $\lambda_0 \in \mathbb{R}$ exists, such that the operator norm $\|\mathbf{H}\|$ of \mathbf{H} is upper bounded by $\mathbb{E}_m[l'(T(X), Y)]\lambda_0$, where $\mathbb{E}_m[l'(T(X), Y)]$ is the empirical expectation of l' . It implies the regions around the minima are flat basins, where the eigen-spectrum of \mathbf{H} is increasingly concentrated around zero.

For the impatient readers, refer to 6.4 for the proof directly and further discussion of the theorem. The error for finite N is discussed at the remarks of theorem 6.2. In the rest of this section, we elaborate and prove the theorem step by step.

6.1 Hessian of NN Is Inherently A Huge Random Matrix

As explained, to study the landscape of the loss function, we study the eigenvalue distribution of its Hessian \mathbf{H} at the critical points. First, we derive the Hessian \mathbf{H} of loss function of class \mathcal{L}_0 composed upon NNs with a single output. For a review of matrix calculus, the reader may refer to Magnus and Neudecker (2007).

The first partial differential of l w.r.t. \mathbf{W}_p is

$$\partial l(T\mathbf{x}, y) = l'(T\mathbf{x}, y)\boldsymbol{\alpha}^T \prod_{j=p+1}^{L-1} (\text{dg}(\tilde{\mathbf{h}}'_j)\mathbf{W}_j^T)\text{dg}(\tilde{\mathbf{h}}'_p) \otimes \mathbf{x}^T \prod_{i=1}^{p-1} (\mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i)) \partial \text{vec} \mathbf{W}_p$$

where \otimes denotes Kronecker product. Note that for clarity of presentation, we use the partial differential the same way as differential is defined and used in Magnus and Neudecker (2007), i.e., ∂l is a number instead of an infinitely small quantities, though in the book, partial differential is not defined explicitly. $\tilde{\mathbf{h}}'$ is an abuse of notation for clarity and needs some explanation. Recall that $\{\mathbf{h}_i\}_{i=1, \dots, L-1}$ are the group indicator r.e.s, and $\{\tilde{\mathbf{h}}_i\}_{i=1, \dots, L-1}$ are the estimation of them based on transport maps. $\tilde{\mathbf{h}}_i$ is a scalar function, and denote it as $\tilde{\mathbf{h}}_i(a)$, where a is the computed input. When computing the partial differential w.r.t. \mathbf{W}_p , by the chain rule, the differential of $\tilde{\mathbf{h}}_i(a)$ w.r.t. \mathbf{W}_p is $\partial \tilde{\mathbf{h}}_i(a)/\partial a$. To avoid introducing too much clutter, we denote $\tilde{\mathbf{h}}'_i$ as $\partial \tilde{\mathbf{h}}_i(a)/\partial a$.

Since \mathbf{H} is symmetric, we only need to compute the block matrices by taking partial differential w.r.t. \mathbf{W}_q , where $q > p$ — taking partial differential w.r.t. \mathbf{W}_p again gives zero matrix.

$$\begin{aligned} \partial^2 l(T\mathbf{x}, y) &= l'(T\mathbf{x}, y) \\ &[(\boldsymbol{\alpha}^T \prod_{k=q+1}^{L-1} (\text{dg}(\tilde{\mathbf{h}}''_k)\mathbf{W}_k^T)\text{dg}(\tilde{\mathbf{h}}'_q) \otimes \text{dg}(\tilde{\mathbf{h}}'_p) \prod_{j=p+1}^{q-1} (\mathbf{W}_j \text{dg}(\tilde{\mathbf{h}}'_j)) \partial \text{vec} \mathbf{W}_q)^T \\ &\otimes \mathbf{x}^T \prod_{i=1}^{p-1} (\mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i))] \partial \text{vec} \mathbf{W}_p \\ &= l'(T\mathbf{x}, y) \\ &(\partial \text{vec} \mathbf{W}_q)^T [\text{dg}(\tilde{\mathbf{h}}'_q) \prod_{k=q+1}^{L-1} (\mathbf{W}_k \text{dg}(\tilde{\mathbf{h}}''_k)) \boldsymbol{\alpha} \otimes \prod_{j=p+1}^{q-1} (\text{dg}(\tilde{\mathbf{h}}'_j)\mathbf{W}_j^T)\text{dg}(\tilde{\mathbf{h}}'_p) \\ &\otimes \mathbf{x}^T \prod_{i=1}^{p-1} (\mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i))] \partial \text{vec} \mathbf{W}_p \end{aligned}$$

where again $\tilde{\mathbf{h}}''$ is an abuse of notation, it is actually $\tilde{\mathbf{h}}' \odot \tilde{\mathbf{h}}'$, the hamadard product of partial differentials obtained by taking partial differential w.r.t. \mathbf{W}_p , and w.r.t. \mathbf{W}_q . The two partial differentials are the same because \mathbf{a} , the input the \mathbf{h} , is the same throughout. Notice that the estimation $\tilde{\mathbf{h}}$ of group indicator r.e.s \mathbf{H} is merely a function of \tilde{H} . It implies $\tilde{\mathbf{h}}$ is a realization of a r.e. created by a deterministic coupling by applying a transport map on \tilde{H} . The deeper principles of the estimation will be explained in section 5. For now, it suffices to stop with the fact that the entries of \mathbf{H} is a random variable. As an example, for estimation done by ReLU, $\tilde{\mathbf{h}}$ would be $\tilde{\mathbf{h}}_i = \max\{0, \tilde{\mathbf{h}}_i\}$. Thus, \mathbf{H} is an ensemble of real symmetric random matrix with correlated entries. Denote

$$\mathbf{H}_{pq} = l'(T\mathbf{x}, y) \text{dg}(\tilde{\mathbf{h}}'_q) \prod_{k=q+1}^{L-1} (\mathbf{W}_k \text{dg}(\tilde{\mathbf{h}}''_k)) \boldsymbol{\alpha} \otimes \prod_{j=p+1}^{q-1} (\text{dg}(\tilde{\mathbf{h}}'_j)\mathbf{W}_j^T)\text{dg}(\tilde{\mathbf{h}}'_p) \otimes \mathbf{x}^T \prod_{i=1}^{p-1} (\mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i)) \quad (5)$$

We have the Hessian of l as

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & \mathbf{H}_{12}^T & \cdots & \mathbf{H}_{1L}^T \\ \mathbf{H}_{12} & \mathbf{0} & \cdots & \mathbf{H}_{2L}^T \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{H}_{1L} & \mathbf{H}_{2L} & \cdots & \mathbf{0} \end{bmatrix} \quad (6)$$

6.2 Eigenvalue Distribution of Symmetry Random Matrix with Slow Correlation Decay

In this section, we show how to obtain the eigen-spectrum of \mathbf{H} through random matrix theory (RMT) (Tao (2012)). RMT has been born out of the study on the nuclei of heavy atoms, where the spacings between lines in the spectrum of a heavy atom nucleus is postulated the same with spacings between eigenvalues of a random matrix (Wigner (1957)). In a certain way, it seems to be the backbone math of complex systems, where the collective behaviors of sophisticated subunits can be analyzed stochastically when deterministic or analytic analysis is intractable.

The following definitions can be found in Tao (2012) unless otherwise noted.

The eigen-spectrum is studied as empirical spectral distribution (ESD) in RMT, define as

Definition 6.1 (Empirical Spectral Distribution). *Given a $N \times N$ random matrix \mathbf{H} , its **empirical spectral distribution** $\mu_{\mathbf{H}}$ is*

$$\mu_{\mathbf{H}} = \frac{1}{N} \sum_{i=1}^N \delta_{\lambda_i}$$

where $\{\lambda_i\}_{i=1,\dots,N}$ are all eigenvalues of \mathbf{H} and δ is the delta function.

Given a hermitian matrix \mathbf{H} , its ESD $\mu_{\mathbf{H}}(\lambda)$ can be studied via its resolvent \mathbf{G} .

Definition 6.2 (Resolvent). *Let \mathbf{H} be a normal matrix, and $z \in \mathbb{H}$ a spectral parameter. The **resolvent** \mathbf{G} of \mathbf{H} at z is defined as*

$$\mathbf{G} = \mathbf{G}(z) = \frac{1}{\mathbf{H} - z}$$

where

$$\mathbb{H} := \{z \in \mathbb{C} : \Im z > 0\}$$

\mathbb{C} denotes the complex field, and \Im is the function gets imaginary part of a complex number z .

\mathbf{G} compactly summarizes the spectral information of \mathbf{H} around z , which is normally analyzed by *functional calculus* on operators, and defined through Cauchy integral formula on operators

Definition 6.3 (Functions of Operators).

$$f(T) = \frac{1}{2\pi i} \int_C \frac{f(\lambda)}{\lambda - T} d\lambda \quad (7)$$

where f is an analytic scalar function and C is an appropriately chosen contour in \mathbb{C} .

The formula can be defined on a range of linear operators (Dunford and Schwartz (1957)) (Recall that a linear operator is a mapping whose domain and codomain are defined on the same field). Since the most complex case involved here will be a normal matrix, we stop at stating that the formula holds true when T is a normal matrix.

Resolvent \mathbf{G} is related to eigen-spectrum of \mathbf{H} through stieltjes transform of $\mu_{\mathbf{H}}(\lambda)$.

Definition 6.4 (Stieltjes Transform). *Let μ be a Borel probability measure on \mathbb{R} . Its **Stieltjes transform** at a spectral parameter $z \in \mathbb{H}$ is defined as*

$$m_{\mu}(z) = \int_{\mathbb{R}} \frac{d\mu(x)}{x - z}$$

With some efforts, it can be seen that the normalized trace of \mathbf{G} is stieltjes transform of eigen-spectrum of \mathbf{H}

$$m_{\mu_{\mathbf{H}}}(z) = \frac{1}{N} \text{tr } \mathbf{G}$$

For a proof, the reader may refer to proposition 2.1 in Alt et al. (2018). $\mu_{\mathbf{H}}$ can be recovered from $m_{\mu_{\mathbf{H}}}$ through the inverse formula of Stieltjes-Perron.

Lemma 6.1 (Inverse Stieltjes Transform). *Suppose that μ is a probability measure on \mathbb{R} and let m_μ be its Stieltjes transform. Then for any $a < b$, we have*

$$\mu((a, b)) + \frac{1}{2}[\mu(\{a\}) + \mu(\{b\})] = \lim_{\Im z \rightarrow 0} \frac{1}{\pi} \int_b^a \Im m_\mu(z) d\Re z$$

where \Re is the function that gets the real part of z . The proof can be found at Tao (2012) p. 144.

Consequently, the problem converts to obtain \mathbf{G} if we want to obtain $\mu_{\mathbf{H}}$. A recent advance in the RMT community has enabled the analysis of ESD of symmetric random matrix with correlation (Erdős et al. (2017)) from the perspective of mean field theory (Kadanoff (2000)), which we debrief in the following.

The resolvent \mathbf{G} holds an identity by definition

$$\mathbf{H}\mathbf{G} = \mathbf{I} + z\mathbf{G} \quad (8)$$

Note that in the above equation \mathbf{G} is a function $\mathbf{G}(\mathbf{H})$ of \mathbf{H} . When the average fluctuation of entries of \mathbf{G} w.r.t. to its mean is small as N grows large, eq. (8) can be turned into a solvable equation regarding \mathbf{G} instead of merely a definition. Formally, it is achieved by taking the expectation of eq. (8)

$$\mathbb{E}[\mathbf{H}\mathbf{G}] = \mathbf{I} + z\mathbb{E}[\mathbf{G}] \quad (9)$$

When fluctuation of moments beyond the second order are negligible, we can obtain a class of random matrices whose ESD can be obtained by solving a truncated cumulant expansion of eq. (9). With the above approach, using sophisticated multivariate cumulant expansion, Erdős et al. (2017) proves \mathbf{G} can be obtained as the unique solution to *Matrix Dyson Equation (MDE)* below

$$\mathbf{I} + (z - \mathbf{A} + \mathcal{S}[\mathbf{G}])\mathbf{G} = \mathbf{0}, \Im \mathbf{G} \succ \mathbf{0}, \Im z > 0 \quad (10)$$

where $\Im \mathbf{G} \succ \mathbf{0}$ means $\Im \mathbf{G}$ is positive definite,

$$\mathbf{A} := \mathbb{E}[\mathbf{H}], \frac{1}{\sqrt{N}}\mathbf{W} := \mathbf{H} - \mathbf{A}, \mathcal{S}[\mathbf{R}] := \frac{1}{N}\mathbb{E}_{\mathbf{W}}[\mathbf{W}\mathbf{R}\mathbf{W}] \quad (11)$$

\mathcal{S} is a linear map on the space of $N \times N$ matrices and \mathbf{W} is a random matrix with zero expectation. The expectation is taken w.r.t. \mathbf{W} , while taking \mathbf{R} as a deterministic matrix.

We describe their results formally in the following, which begins with some more definitions, adopted from Erdős et al. (2017).

Definition 6.5 (Cumulant). **Cumulants** of $\kappa_{\mathbf{m}}$ of a random vector $\mathbf{w} = (w_1, \dots, w_n)$ are defined as the coefficients of log-characteristic function

$$\log \mathbb{E} e^{it^T \mathbf{w}} = \sum_{\mathbf{m}} \kappa_{\mathbf{m}} \frac{it^{\mathbf{m}}}{\mathbf{m}!}$$

where $\sum_{\mathbf{m}}$ is the sum over all n -dimensional multi-indices $\mathbf{m} = (m_1, \dots, m_n)$.

To recall, a multi-indices is

Definition 6.6 (Multi-index). *a n -dimensional multi-index is an n -tuple*

$$\mathbf{m} = (m_1, \dots, m_n)$$

of non-negative integers. Note that $|\mathbf{m}| = \sum_{i=1}^n m_i$, and $\mathbf{m}! = \prod_{i=1}^n m_i!$.

Similar with the more familiar concept *moment*, cumulant is also a measure of statistic properties of r.e.s. Particularly, the k -order cumulant κ characterizes the k -way correlation of a set of r.v.s. The key of insight of the paper is to identify the condition where a matrix entry $w_\alpha, \alpha \in \mathbb{I}$ is only strongly correlated with a minority of $\mathbf{W}_{\mathbb{I} \setminus \{\alpha\}}$, and higher order cumulants tend to be weak and not influential in large N limit. Thus, a proper formulation of the correlation strength is needed, and is defined as the cumulant norms on entries of \mathbf{W} in the following. Given k entries \mathbf{W}_α at $\alpha = \{\alpha_i\}_{i=1, \dots, k}, \alpha_i \in \mathbb{I}$ of matrix \mathbf{W} , where duplication is allowed, denote $\kappa(\alpha_1, \dots, \alpha_k) = \kappa(w_{\alpha_1}, \dots, w_{\alpha_k})$.

Definition 6.7 (Cumulant Norms).

$$|||\kappa||| := |||\kappa|||_{\leq R} := \max_{2 \leq k \leq R} |||\kappa|||_k, |||\kappa|||_k := |||\kappa|||_k^{av} + |||\kappa|||_k^{iso}$$

$$|||\kappa|||_2^{av} := |||\kappa(*, *)|||, |||\kappa|||_k^{av} := N^{-2} \sum_{\alpha_1, \dots, \alpha_k} |\kappa(\alpha_1, \dots, \alpha_k)|, k \geq 4 \quad (12a)$$

$$|||\kappa|||_3^{av} := ||\sum_{\alpha_1} |\kappa(\alpha_1, *, *)||| + \inf_{\kappa = \kappa_{dd} + \kappa_{dc} + \kappa_{cd} + \kappa_{cc}} (|||\kappa_{dd}|||_{dd} + |||\kappa_{dc}|||_{dc} + |||\kappa_{cd}|||_{cd} + |||\kappa_{cc}|||_{cc}) \quad (12b)$$

$$|||\kappa|||_{cc} = |||\kappa|||_{dd} := N^{-1} \sqrt{\sum_{b_2, a_3} (\sum_{a_2, b_3} \sum_{\alpha_1} |\kappa(\alpha_1, a_2 b_2, a_3 b_3)|)^2}$$

$$|||\kappa|||_{cd} := N^{-1} \sqrt{\sum_{b_3, a_1} (\sum_{a_3, b_1} \sum_{\alpha_2} |\kappa(a_1 b_1, \alpha_2, a_3 b_3)|)^2}, |||\kappa|||_{dc} := N^{-1} \sqrt{\sum_{b_1, a_2} (\sum_{a_1, b_2} \sum_{\alpha_3} |\kappa(a_1 b_1, a_2 b_2, \alpha_3)|)^2}$$

$$|||\kappa|||_2^{iso} := \inf_{\kappa = \kappa_d + \kappa_c} (|||\kappa_d|||_d + |||\kappa_c|||_c), |||\kappa_d|||_d := \sup_{||\mathbf{x}|| \leq 1} |||\kappa(\mathbf{x}*, *)|||, |||\kappa_c|||_c := \sup_{||\mathbf{x}|| \leq 1} |||\kappa(\mathbf{x}*, \cdot)||| \quad (12c)$$

$$|||\kappa|||_k^{iso} := ||\sum_{\alpha_1, \dots, \alpha_{k-2}} |\kappa(\alpha_1, \dots, \alpha_{k-2}, *, *)|||, k \geq 3$$

where in eq. (12b), the infimum is taken over all decomposition of κ in four symmetric functions $\kappa_{dd}, \kappa_{dc}, \kappa_{cd}, \kappa_{cc}$; in eq. (12c) the infimum is taken over all decomposition of κ into the sum of symmetric κ_c and κ_d . The norms defined in eq. (12a) and eq. (12c) need some explanation on the notation. If, in place of an index $\alpha \in \mathbb{J}$, we write a dot (\cdot) in a scalar quantity then we consider the quantity as a vector indexed by the coordinate at the place of the dot. For example, $\kappa(a_1 \cdot, a_2 b_2)$ is a vector, the i -th entry of which is $\kappa(a_1 i, a_2 b_2)$ and therefore the inner norms in eq. (12a) indicate vector norms. In contrast, the outer norms indicate the operator norm of the matrix indexed by star (*). More specifically, $||A(*, *)||$ refers to the operator norm of the matrix with matrix elements A_{ij} . Thus $|||\kappa(\mathbf{x}*, \cdot)|||$ is the operator norm $||A||$ of the matrix A with matrix elements $A_{ij} = ||\kappa(\mathbf{x}i, j)||$. $\kappa(\mathbf{x}b_1, a_2 b_2)$ denotes $\sum_{a_1} \kappa(a_1 b_1, a_2 b_2) x_{a_1}$, where \mathbf{x} is a vector.

We do not want to explain the cumulant norms beyond what has been said, considering it is too technically involved and rather a distraction. For interested readers, we suggest reading the paper Erdős et al. (2017). Equipped with the cumulant norms, we would have the assumptions stated in assumption 6.1 6.2 that make MDE valid.

Remark. In Erdős et al. (2017), the functions f, g_1, \dots, g_q in assumption 6.2 are assumed to be functions without any qualifiers. We change it to analytic functions for further usage. In the proof of theorem 6.2, the functions are only required to be analytic, thus even if the assumptions are changed, the conclusion still holds.

The diversity assumption requires that a matrix entry w_α only couples with a minority of the overall entries, and for the rest of the entries, the coupling strength does not exceed a certain value $N^{-3q} ||f||_{q+1} \prod_{j=1}^q ||g_j||_{q+1}$ characterized by cumulants. For example, suppose $q = 1$, given a entry α_1 , the assumption essentially states that the entries in the coupling set $\mathcal{N}(\alpha_1)$ is not strongly coupled with the resting of the population $\mathbf{W}_{\mathbb{I} \setminus \mathcal{N}_{n_1+1}(\alpha_1)}$. The explanation goes similar as q grows, of which the coupling strength is characterized by higher order cumulants. While boundedness assumptions 1)2)3)4) states the expectation of \mathbf{H} is bounded, moments are finite, cumulants are bounded for the entries that do strongly couples, and $\mathcal{S}[\mathbf{W}]$ is bounded in the sense of eigenvalues.

When the assumptions satisfies, we have the resolvent \mathbf{G} of a random matrix \mathbf{H} close to the solution to the MDE probabilistically with some regular properties as the following, adopted in an informal style to ease reading from Erdős et al. (2017) theorem 2.2, Helton et al. (2007) theorem 2.1, and Alt et al. (2018) theorem 2.5.

Theorem 6.2. Let \mathbf{M} be the solution to the Matrix Dyson Equation eq. (10), and ρ the density function (measure) recovered from normalized trace $\frac{1}{N} \text{tr } \mathbf{M}$ through Stieljies inverse lemma 6.1. We have

1. The MDE has a unique solution $\mathbf{M} = \mathbf{M}(z)$ for all $z \in \mathbb{H}$.
2. suppp is a finite union of closed intervals with nonempty interior. Moreover, the nonempty interiors are called the **bulk** of the eigenvalue density function ρ .
3. The resolvent \mathbf{G} of \mathbf{H} converges to \mathbf{M} as $N \rightarrow \infty$.

Remark. theorem 6.2.3 is a probably-approximately-correct type result, where the error depends on N . We do not present the exact error bound here, for that it is rather complicated, and does not help understanding — since we are not working on finer behaviors of NNs with a particular size, and do not need such a fine granularity characterization yet. We refer interested readers to Erdős et al. (2017) theorem 2.1, 2.2, where the exact error bounds are present.

6.3 Diversity Assumption is A Precondition to the Power of NNs and S-System

Before we leverage MDE to obtain the eigen-spectrum of the Hessian \mathbf{H} of NNs derived at eq. (6), we explain the meaning of the assumptions 6.1, 6.2 in the NN and S-System context, so to point to the potential of the assumptions to give practical guidance on training NNs.

Recall that the objective function is the empirical risk function $R(T)$ at eq. (2). Given a set of i.i.d. training samples $(X_i, Y_i)_{i=1, \dots, m}$, $R(T)$ is a summation of the i.i.d. random matrices. Formally, reusing the notation to denote \mathbf{H} the Hessian of $R(T)$ and \mathbf{H}_i the Hessian of $l(T(X_i; \boldsymbol{\theta}), Y_i)$, we have

$$\mathbf{H} = \frac{1}{m} \sum_{i=1}^m \mathbf{H}_i \quad (13)$$

Acute reader may realize that by the multivariate central limiting theorem (Klenke (2012) theorem 15.57), \mathbf{H} will converge to a Gaussian ensemble, i.e., a random matrix of which the distribution of entries is a Gaussian process (GP), asymptotically as $m \rightarrow \infty$. For a GP, all higher order cumulants are zero, which greatly simplifies the picture, and gives much clearer meaning on the assumptions 6.1 6.2 made. In the following, we will explain the practicality, and also how they may serve as guidance to design and improve NNs, in the asymptotically large sample limit, which gives a picture that can be described using more widely used terms, i.e., mean and covariance.

The practicality of boundedness assumptions is obvious, since we do not want values to blow up. We only note for the two outliers. First, the lower bound in 4) in boundedness assumption, $cN^{-1} \text{tr } \mathbf{G} \preceq \mathcal{S}[\mathbf{G}]$, which is not about infinity. It asks the eigenvalues of $\mathcal{S}[\mathbf{G}]$ to stay close to its average value, so to let $\mathcal{S}[\mathbf{G}]$ stay in the cone of the positive definite matrices to ensure the stability of the MDE. It is essentially a constraint on the interaction of second order cumulants, and is realizable in a NN, though we are not clear on its physical meaning for the time being. Second, the boundedness assumption 3), as briefly discussed before, is a bound that bounds the strength of the entries of \mathbf{H} that do correlate, while the diversity assumption is about the weakness of the entries that are not correlated. The practicality of the former is straightforward. To see the practicality of diversity assumption in the NN context, first we come back to the concrete form of Hessian. We rewrite eq. (5) in the following form (the equation should be read vertically)

$$\mathbf{H}_{pq} = l'(T\mathbf{x}, y) \tilde{\mathbf{W}}_{q \sim (L-1)} \text{dg}(\tilde{\mathbf{h}}''_{L-1}) \boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}_q \quad (14a)$$

$$\otimes \tilde{\mathbf{W}}_{p \sim (q-1)} \quad \otimes \tilde{\mathbf{W}}_{p \sim (q-1)} \quad (14b)$$

$$\otimes \mathbf{x}^T \tilde{\mathbf{W}}_{1 \sim (p-1)} \quad \otimes \tilde{\mathbf{x}}_{p-1}^T \quad (14c)$$

where

$$\begin{aligned} \tilde{\mathbf{W}}_{q \sim (L-1)} &:= \text{dg}(\tilde{\mathbf{h}}''_q) \prod_{k=q+1}^{L-2} (\mathbf{W}_k \text{dg}(\tilde{\mathbf{h}}''_k)) \mathbf{W}_{L-1}, & \tilde{\boldsymbol{\alpha}}_q &:= \tilde{\mathbf{W}}_{q \sim (L-1)} \text{dg}(\tilde{\mathbf{h}}''_{L-1}) \boldsymbol{\alpha} l'(T\mathbf{x}, y) \\ \tilde{\mathbf{W}}_{p \sim (q-1)} &:= \prod_{j=p+1}^{q-1} (\text{dg}(\tilde{\mathbf{h}}'_j) \mathbf{W}_j^T) \text{dg}(\tilde{\mathbf{h}}'_p), \\ \tilde{\mathbf{W}}_{1 \sim (p-1)} &:= \prod_{i=1}^{p-1} (\mathbf{W}_i \text{dg}(\tilde{\mathbf{h}}_i)), & \tilde{\mathbf{x}}_{p-1}^T &:= \mathbf{x}^T \tilde{\mathbf{W}}_{1 \sim (p-1)} \end{aligned}$$

With some efforts, using NN terminologies, it can be viewed that eq. (14a) is a vector $\tilde{\boldsymbol{\alpha}}_q$ created by back propagating the vector $\text{dg}(\tilde{\mathbf{h}}''_{L-1}) \boldsymbol{\alpha} l'(T\mathbf{x}, y)$ to layer q by left multiplying $\tilde{\mathbf{W}}_{q \sim (L-1)}$ — note that if you replace \mathbf{h}''_k with \mathbf{h}'_k , you get the *back propagated gradient*; eq. (14b) is the *covariance matrix without removing the mean* between neurons at layer p and layer $q-1$, when taking expectation w.r.t. samples, i.e., $\mathbb{E}_{z \sim \mu} [\tilde{\mathbf{W}}_{p \sim (q-1)}]$; eq. (14c) is the *forward propagated activation* at layer $p-1$. Now it is quite clear what the correlation between entries of the Hessian is about. It is the correlation between *the product of forward propagated neuron activation at layer $p-1$, the back propagated "gradient" at layer q , and the strength of activation paths that connects the two sets of neurons*.

When \mathbf{H} is a Gaussian ensemble, all higher order cumulants vanishes, thus the diversity assumption is solely about the second order cumulants, and the case when $q=1$ and $q=2$. Since $f, \{g_i\}_{i=1, \dots, q}$ are analytic, $\kappa(f(\cdot), g_1(\cdot))$ is a generalized cumulant (McCullagh (1987) Chapter 3), which can be decomposed into a sum of cumulants of entries of the Hessian. So is $\kappa(f(\cdot), g_1(\cdot), g_2(\cdot))$. Considering that only first and second cumulants exist, which are means and covariance respectively, $\kappa(f(\cdot), g_1(\cdot))$ thus is a sum of means of entries of the Hessian and covariance between entries of the Hessian. So is $\kappa(f(\cdot), g_1(\cdot), g_2(\cdot))$. Using $\kappa(f(\cdot), g_1(\cdot))$ as an example, the diversity

assumption states that for any $\alpha \in \mathbb{I}$, nested sets $\mathcal{N}_1 \subset \mathcal{N}_2 = \mathcal{N}, |\mathcal{N}| \leq N^{1/2-\mu}$ exist (when only cumulants up to the second order exist, it suffices to let R be 2 instead of every $R \in \mathbb{N}$ (Erdős et al. (2017))), such that $\kappa(f(\mathbf{W}_{\mathbb{I} \setminus \mathcal{N}}), g_1(\mathbf{W}_{\mathcal{N}_1})) = \sum_{\beta \in \mathcal{N}'_1 \subset \mathcal{N}_1 \cup \mathbb{I} \setminus \mathcal{N}} \kappa(\beta) + \sum_{\beta, \gamma \in \mathcal{N}' \subset \mathcal{N}_1 \cup \mathbb{I} \setminus \mathcal{N}} \kappa(\beta, \gamma)$, where $\mathcal{N}'_1, \mathcal{N}'$ are subsets noted that depends on f, g_1 . The interpretation is qualitative. But even from the qualitative interpretation, it can see that *the diversity assumption is on the smallness of the mean and covariance of $\tilde{\alpha}_i \tilde{w}_{jk} \tilde{x}_l$, the product of “gradient”, activation path correlation strength, and forward activation*. Additionally, to prove theorem 6.1, we need a further assumption that $\mathbb{E}[\mathbf{H}] = \mathbf{0}$. It clearly connects to the experiment tricks used in the community, such as the early initialization schemes that tries to keep mean of gradient and activation zero, and standard deviation (std) small (Glorot and Bengio (2010) He et al. (2015)), the normalization schemes that keep the mean of activation zero and std small (Ioffe, Sergey and Szegedy (2015) Salimans and Kingma (2016)), though some more works are needed to reach there rigorously.

To recap, as stated similarly in section 2, section 6.2, the diversity assumption states that a diversity should exist in the neuron population, so that for any neurons, it does not strongly correlate with the majority of the neuron population. The diversity in r.e.s. of S-System is not a built-in feature, but a design choice in its implementations.

Qualitatively, we can see the design of NNs resonates with the diversity assumption: 1) different group indicator r.e.s. are assumed to be independent given input r.e.s., referring to definition 4 5, in which case, the coupling aims to group the measure that is distinctive w.r.t. other couplings created through grouping, thus, r.e.s that are not coupled together are likely to be uncorrelated; 2) activation function creates couplings that only couple higher scale events with the “active” lower scale events, thus implementing coarse graining that creates events that are composed by different lower scale events. The above design may not be the only choice, however, it helps create uncorrelated r.e.s within a scale and across scales, consequently making the product of the forward propagated activation, activation paths, and back propagated “gradient” tend to be uncorrelated.

Yet, this is a rather general explanation on why diversity occurs without taking into the finer statistics structure in the data. More improvements may still be made. For instance, the low correlation existed in CNN is the result of a coupling that considers the spatial symmetry, where output r.e.s in a large spatial distance simply does not couples, thus tending to be uncorrelated.

S-System is a fabulous mechanism that can indefinitely increase the number of parameters, thus its learning capacity, in a meaningful way, i.e., creating higher scale coupling yet maintaining the diversity of the r.e.s created. Such mechanism does not normally hold in other systems or algorithms. Taking linear NNs for example, though with the potential to infinitely increase its parameters, matrices that multiply together still have a highly correlation structure within, thus cannot create a population of diverse neurons that are of low correlation with a majority of the other neurons. Accompanying the result we will prove in the next section, which states $R(T)$ can be optimized to zero, assuming assumptions 6.1 6.2, we can see that *the diversity assumption actually characterizes a sufficient precondition to the optimization power of NNs*.

6.4 NN Loss Landscape: All Local Minima Are Global Minima with Zero Losses

We have obtained the operator equation to describe the eigen-spectrum of the Hessian of NNs and explained its assumptions. With one further assumption, we show in this section that for NNs with objective function belonging to the function class \mathcal{L}_0 , all local minima are global minima with zero loss values.

We outline the strategy first. Since MDE is a nonlinear operator equation, it is not possible to obtain a close form analytic solution. The only way to get its solution is an iterative algorithm (Helton et al. (2007)), which is not an easy task given the millions of parameters of a NN — remembering that we are dealing with large N limit — though it can serve as an exploratory tool. However, we do are able to get qualitative results by directly analyzing the equation. Our goal is to show all critical points are saddle points, except for the ones has zero loss values, which are global minima. To prove it, we prove that at the points where $R(T) \neq 0$, the eigen-spectrum $\mu_{\mathbf{H}}$ of the Hessian \mathbf{H} is symmetric w.r.t. the y -axis, which implies that as long as non-zero eigenvalues exist, half of them will be negative. To prove it, we prove the stieltjes transform $m_{\mu_{\mathbf{H}}}(z)$ of $\mu_{\mathbf{H}}$ satisfies $\Im m_{\mu_{\mathbf{H}}}(-z^*) = \Im m_{\mu_{\mathbf{H}}}(z)$, where z^* denote the complex conjugate of z . In the following, we present the proof formally.

Lemma 6.2. *Let $\mathbf{M}(z), \mathbf{M}'(-z^*)$ be the unique solution to the MDE at spectral parameter $z, -z^*$ defined at eq. (10) respectively, and $\mathbf{A} = \mathbf{0}$. We have*

$$\mathbf{M}' = -\mathbf{M}^*$$

where $*$ means taking conjugate transpose.

Proof. First, we rewrite the MDE. Note that $\mathcal{S}[\mathbf{G}]$ is positivity preserving, i.e., $\forall \mathbf{G} \succ \mathbf{0}, \mathcal{S}[\mathbf{G}] \succ \mathbf{0}$ by assumption 6.1 4). In addition, we have $\Im z > 0$, thus $\Im(z + \mathcal{S}[\mathbf{G}]) \succ \mathbf{0}$. Then, by Haagerup and Thorbjørnsen (2005) lemma

3.2, we have $z + \mathcal{S}[\mathbf{G}] \succ \mathbf{0}$, so it is invertable. Thus, we can rewrite the MDE into the following form

$$\mathbf{G} = -(z + \mathcal{S}[\mathbf{G}])^{-1} \quad (15)$$

Suppose \mathbf{M} is a solution to the MDE at spectral parameter z . The key to the proof is the fact that $\mathcal{S}[\mathbf{G}]$ is linear and commutes with taking conjugate, thus by replacing \mathbf{M} with $-\mathbf{M}^*$, and z with $-z^*$, we would get the same equation. We show it formally in the following.

First, note that $\mathcal{S}[\mathbf{M}]$ is a linear map of \mathbf{M} , so the we have

$$\mathcal{S}[-\mathbf{M}] = -\mathcal{S}[\mathbf{M}]$$

Also, $\mathcal{S}[\mathbf{M}]$ commutes with $*$, for the fact

$$\mathcal{S}[\mathbf{M}^*] = \mathbb{E}[\mathbf{W}\mathbf{M}^*\mathbf{W}] = \mathbb{E}[(\mathbf{W}\mathbf{M}\mathbf{W})^*] = \mathbb{E}[\mathbf{W}\mathbf{M}\mathbf{W}]^* = \mathcal{S}[\mathbf{M}]^*$$

Furthermore, we $*$ is commute with taking inverse, for the fact

$$\begin{aligned} & \mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \\ \implies & (\mathbf{A}\mathbf{A}^{-1})^* = \mathbf{I} \\ \implies & \mathbf{A}^{-1*}\mathbf{A}^* = \mathbf{I} \\ \implies & \mathbf{A}^{-1*} = \mathbf{A}^{*-1} \end{aligned}$$

With the commutativity results, we do the proof. The solution \mathbf{M} satisfies the equation

$$\mathbf{M} = -(z + \mathcal{S}[\mathbf{M}])^{-1}$$

Replacing \mathbf{M} with $-\mathbf{M}^*$, z with $-z^*$, we have

$$\begin{aligned} & -\mathbf{M}^* = -(-z^* + \mathcal{S}[-\mathbf{M}^*])^{-1} \\ \implies & \mathbf{M}^* = -(z^* + \mathcal{S}[\mathbf{M}^*])^{-1} \\ \implies & \mathbf{M}^* = -(z^* + \mathcal{S}[\mathbf{M}]^*)^{-1} \\ \implies & \mathbf{M}^* = -(z + \mathcal{S}[\mathbf{M}])^{-1*} \\ \implies & \mathbf{M} = -(z + \mathcal{S}[\mathbf{M}])^{-1} \end{aligned}$$

After the replacement, we actually get the same equation. Thus, $-\mathbf{M}^*, -z^*$ also satisfy eq. (15). Since the pair also satisfies the constrains $\Im \mathbf{M} \succ 0, \Im z > 0$, and by theorem 6.2, the solution is unique, we proved the solution \mathbf{M}' at the spectral parameter $-z^*$ is $-\mathbf{M}^*$. □

Theorem 6.3. *Let \mathbf{H} be a real symmetric random matrix satisfies assumptions 6.1 6.2, in addition to the assumption that $\mathbf{A} = \mathbf{0}$. Let the ESD of \mathbf{H} be $\mu_{\mathbf{H}}$. Then, $\mu_{\mathbf{H}}$ is symmetric w.r.t. to y -axis. In other words, half of the non-zero eigenvalues are negative. Furthermore, non-zero eigenvalues always exist, implying \mathbf{H} will always have negative eigenvalues.*

Proof. By theorem 6.2, the resolvent \mathbf{G} of \mathbf{H} is given by the the unique solution to eq. (10) at spectral parameter z . Let the solution to eq. (10) at spectral parameter $z, -z^*$ be \mathbf{M}, \mathbf{M}' , By lemma 6.2, we have the solutions satisfies

$$\mathbf{M}' = -\mathbf{M}^*$$

By 6.1, the ESD of \mathbf{H} at $\Re z$ is given at

$$\mu_{\mathbf{H}}(\Re z) = \lim_{\Im z \rightarrow 0} \frac{1}{\pi} \Im m_{\mu_{\mathbf{H}}}(z)$$

Since $m_{\mu_{\mathbf{H}}}(z) = \frac{1}{N} \text{tr } \mathbf{M}$, we have

$$\mu_{\mathbf{H}}(\Re z) = \lim_{\Im z \rightarrow 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } \mathbf{M}$$

Similarly,

$$\mu_{\mathbf{H}}(\Re(-z^*)) = \lim_{\Im(-z^*) \rightarrow 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } \mathbf{M}'$$

Note that

$$\begin{aligned}
\mu_{\mathbf{H}}(\Re(-z^*)) &= \lim_{\Im(-z^*) \rightarrow 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M' \\
\Rightarrow \mu_{\mathbf{H}}(\Re(-z^*)) &= \lim_{\Im(-z^*) \rightarrow 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } (-M^*) \\
\Rightarrow \mu_{\mathbf{H}}(-\Re z) &= \lim_{\Im z \rightarrow 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M
\end{aligned}$$

Thus, $\mu_{\mathbf{H}}(\lambda)$, $\lambda \in \mathbb{R}$ is symmetric w.r.t. y -axis. It follows that for all non-zero eigenvalues, half of them are negative.

By theorem 6.2 2, there are always bulks in $\text{supp} \mu_{\mathbf{H}}$, thus there are always non-zero eigenvalues. Since half of the non-zero eigenvalues are negative, it follows \mathbf{H} always has negative eigenvalues. \square

Proof of theorem 6.1. First, we prove part 1 of the theorem. The majority of the proof of part 1 have been dispersed earlier in the paper. What the proof here does mostly is to collect them into one piece.

The Hessian \mathbf{H} of the risk function eq. (2), can be decomposed into a summation of Hessians of loss functions of each training sample, which is described in eq. (13). For each Hessian in the decomposition, it is computed in eq. (6), and it has been shown that \mathbf{H} is a random matrix in section 6.1.

The analysis of the random matrix \mathbf{H} needs to break down into two cases: 1) for all training samples, at least one sample (x, y) has non-zero loss value; 2) and all training samples are classified properly with zero loss values.

We first analyze case 1), since the loss l belongs to function class \mathcal{L}_0 , l is convex and is valued zero at its minimum. When $l(x, y) \neq 0$, we have $l'(x, y) \neq 0$, thus \mathbf{H} is a random matrix — not a zero matrix. The analysis of this type of random matrix is undertaken in section 6.2. For a NN, the assumptions 6.1 6.2 can be satisfied, and the eigen-spectrum $\mu_{\mathbf{H}}$ of \mathbf{H} is given by the MDE defined at eq. (10). The practicality and its potential to guide real world NN optimization is discussed in section 6.3.

By theorem 6.3, $\mu_{\mathbf{H}}$ is symmetric w.r.t. y -axis, and half of its non-zero eigenvalues are negative. Thus, for all critical points of $R(T)$, its will have half of its non-zero eigenvalues negative. It implies all critical points are saddle points.

Now we turn to the case 2). In this case, all training samples are properly classified with zero loss value. Considering the lower bound of l is zero, we have reached the global minima. Also, since all critical points in case 1) are saddle points, local minima can only be reached in case 2), implying all local minima are global minima. Thus, the first part of the theorem is proved.

Now we prove part 2 of the theorem.

Note that the minima is reached for the fact that we have reached the situation where the Hessian \mathbf{H} has degenerated into a zero matrix. Thus, each local minimum is not a critical point, but an infimum, where in a local region around the infimum in the parameter space, all the eigenvalues are increasingly close to zero as the parameters of the NN approach the parameters at the infimum. We show it formally in the following.

Writing a block \mathbf{H}_{pq} (defined at eq. (5)) in the Hessian \mathbf{H}_i of one sample (defined at eq. (6)) in the form of

$$\mathbf{H}_{pq} = l'(T\mathbf{x}, y) \tilde{\mathbf{H}}_{pq}$$

where i is the index of the training samples, defined at eq. (2). Then, putting together $\tilde{\mathbf{H}}_{pq}$ together to form $\tilde{\mathbf{H}}_i$, \mathbf{H}_i is rewritten in the form of

$$\mathbf{H}_i = l'(T\mathbf{x}_i, y_i) \tilde{\mathbf{H}}_i$$

Then the Hessian \mathbf{H} (defined at eq. (13)) of the risk function defined eq. (2) can be rewritten in the form of

$$\mathbf{H} = \frac{1}{m} \sum_{i=1}^m l'(T\mathbf{x}_i, y_i) \tilde{\mathbf{H}}_i$$

Taking the operator norm on the both sides

$$\|\mathbf{H}\| = \left\| \frac{1}{m} \sum_{i=1}^m l'(T\mathbf{x}_i, y_i) \tilde{\mathbf{H}}_i \right\| \leq \frac{1}{m} \sum_{i=1}^m |l'(T\mathbf{x}_i, y_i)| \|\tilde{\mathbf{H}}_i\|$$

Denote $\max_i \{\|\tilde{\mathbf{H}}_i\|\}$ as λ_0 , we have

$$\begin{aligned}
\|\mathbf{H}\| &\leq \frac{1}{m} \sum_{i=1}^m |l'(T\mathbf{x}_i, y_i)| \lambda_0 \\
&= \mathbb{E}_m[l'(TX, Y)] \lambda_0
\end{aligned}$$

The above inequality shows that, as the risk decreases, more and more samples will have zero loss value, consequently $l' = 0$, thus $\mathbb{E}_m[l']$ will be increasingly small, thus the operator norm of \mathbf{H} . At the minima where all $l' = 0$, the Hessian degenerates to a zero matrix. \square

Remark. *It is not necessary for the assumption $\mathbf{A} = \mathbf{0}$ to be held for the theorem to hold, or more specifically, for \mathbf{H} to have negative eigenvalues at its critical points. By proposition 2.1 in Alt et al. (2018)*

$$\text{supp}\mu_{\mathbf{H}} \subset \text{Spec}\mathbf{A} + [-2\|\mathcal{S}\|^{1/2}, 2\|\mathcal{S}\|^{1/2}]$$

where $\text{Spec}\mathbf{A}$ denotes the support of the spectrum of the expectation matrix \mathbf{A} and $\|\mathcal{S}\|$ denotes the norm induced by the operator norm. Thus, it is possible for the spectrum $\mu_{\mathbf{H}}$ of \mathbf{H} to lie at the left side of the y-axis, as long as the spectrum of \mathbf{A} is not too way off from the origin. However, existing characterizations on $\text{supp}\mu_{\mathbf{H}}$ based on bound are too inexact to make sure the existence of support on the left of the y-axis. To get rid of the zero expectation assumption, more works are needed to obtain a better characterization, and could be a direction for future work.

The the phenomenon characterized by theorem 6.1.1 is rather remarkable, if not marvelous. It shows that instead of seeing non-convex optimization as something to avoid, a class of non-convex objective functions can be that powerful to the point of “solving” — minimizing the error to the point of vanishing — complex problems that nature is dealing with in a rather reliable fashion. We feel like this is how a brain is doing optimization. We envision that a much larger class of functions possess such benign loss landscapes than the one here we have studied. Actually, we have isolated a function class that represents some of the most essential characteristics of a more general class of function as shown in eq. (3), so that we can show the principle underlying. That is, diverse yet cooperative subunits aggregating together to form a system can optimize an objective consistently. This larger class of function could be as important as the concept of convexity, and would play an important role in optimization. The goal of the paper is to lay the backbone of the theory of the NNs that make the principles underlying clear, instead of presenting the theory in its complete form in one go. Thus, essential properties of the function class are yet to be identified, and will be part of our future work.

The theorem also contributes to explaining why depth is crucial. The large N limits of the Hessian can be achieved by adding more layers (in the terminology of S-System, using a scale poset having a longer chain as a subset), even though the number of neurons in each of the layers may be quite small compared with the overall number of neurons. The diversity of neurons is possible due to activation functions (in the terminology of S-System, conditional grouping extension on estimated realizations of previous created output r.e.s.).

The phenomenon characterized by theorem 6.1.2 explains why there are two phases in the training dynamics of NNs, i.e., the rapid error decreasing phase when loss value is high, and the slow error decaying phase when the loss value is close to minima. As the error decreases, the expectation of the derivative of loss values in $R(T)$ will increasingly approach zero, thus the $\text{supp}\mu_{\mathbf{H}}$ will concentrate around zero increasingly, making the landscape increasingly flat and the training process slowly. It probably also explains why we need to gradually decrease the step size in the gradient descent algorithms in practice. Very likely the flat regions are of a small volume compared with the overall parameter space. Thus, if the training goes conservatively, and inches towards the global minima, the risk will gradually decrease. But if we give a powerful kick to the training that induces a large shift in the parameter space, it may kick the current parameter out of the flat region that can inch toward the global minima, like kicking a ball from a valley to another mountain in the hyperspace, thus making the training starts all over again to find a valley to decrease the risk. A further characterization of the landscape goes beyond infinitesimal local regions may rigorously prove the conjecture. It even poses the possibility to move across the flat region rather swiftly, as long as we figure out how to stay in the valley as we stride big.

6.5 Related works

Similar to the study on the hierarchical hypothesis of NNs, the study on optimization gains its moment rather recently. We focus on the works that attack the full complexity of optimization problem of deep NNs, while for more related works, we refer the readers to related works discussed in Dauphin et al. (2014) Nguyen and Hein (2017) Liang et al. (2018) for works before the deep learning era, on shallow networks and NP-hardness of NN optimization.

Roughly, two approaches have been taken in analyzing the optimization of NNs, one from the linear algebra perspective, the other from mean field theory using random matrix theory. Our work falls in the latter approach. The linear algebra approach, as the name suggests, shies away from the nonlinear nature of the problem. Kawaguchi (2016) proves all local minima of a deep linear NN are global minima when some rank conditions of the weight matrices are held. Nguyen and Hein (2017) Nguyen and Hein (2018) prove that if in a certain layer of a NN, it

has more neurons than training samples, which makes it possible that the feature maps of all samples are linearly independent, then the network can reach zero training errors. A few works following in the linear-algebraic direction (Laurent and von Brecht (2018) Liang et al. (2018) Yun et al. (2018)) improve upon the two previous results, but using essentially the same approach. As the conditions in Kawaguchi (2016) indicate, the rank related linear algebraic condition does not transport to nonlinear NNs. While for Nguyen and Hein (2017), it characterizes a phenomenon that if in a layer of a NN, it can allocate a neuron to memorize each training sample, then based on the memorization, it can reach zero errors. In a certain way, we believe NNs are doing certain memorization, for the fact that the output elements in the intermediate event representations are learning template/mean of events, as discussed in section 4.2. However, it does it in a smart way, where the templates are decomposed hierarchically. Thus, it is likely we do not need so many linearly independent intermediate features, which would lead to poor generalization. Thus, to truly understand the optimization behavior of NNs, we need to step out of the comfort zone of linearity.

The mean field theory approach using the tools of random matrix theory can attack the optimization of NNs in its full complexity, though existing works tend to be confused on the source of randomness. Due to an inadequate understanding of the randomness induced by activation function, Choromanska et al. (2015a) tries to get rid of the group indicator r.e.s. by assuming that its value is independent of the input r.e.s. of CGE, which is unrealistic (Choromanska et al. (2015b)), nevertheless it is a brave attempt, and the first paper to attack a deep NN in its full complexity. After Choromanska et al. (2015a) which approaches by analogizing with spin glass systems — it is a complex system, as NNs are — some researchers start to study NNs from mean field theory from the first principle instead of by analog. Again, confused with the source of randomness in activation in the intermediate layers of NNs, Jeffrey Pennington (2017) just assumes data, weights and errors are of i.i.d. Gaussian distribution, which are mean field approach assumptions and unrealistic, and proceeds to analyze the Hessian of the loss function of NNs, though due to limitations of their assumptions, they can only analyze a NN with one hidden layer. By laying a theoretical foundation of NNs, S-System accurately points out where randomness arises in NNs, and what reminds to prove theorem 6.1 is to find the right random matrix tools.

References

- Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Jon Shlens, Benoit Steiner, Ilya Sutskever, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Oriol Vinyals, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems. Technical report, 2015.
- Johannes Alt, Laszlo Erdos, and Torben Krüger. The Dyson equation with linear self-energy: spectral bands, edges and cusps. Technical report, 2018. URL <http://arxiv.org/abs/1804.07752>.
- Amari. Information geometry of the $\{EM\}$ and $\{EM\}$ algorithm for neural networks. *Neural Networks*, 8(9): 1379–1408, 1995.
- S. Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016. ISBN 9784431559788.
- S. Amari and H. Nagaoka. *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society, 2007. ISBN 9780821843024.
- S. I. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001. ISSN 00189448. doi: 10.1109/18.930911.
- Shun Ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, 18(5):1007–1065, 2006. ISSN 1530888X. doi: 10.1162/neco.2006.18.5.1007.
- Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Statistical Neurodynamics of Deep Networks: Geometry of Signal Spaces. Technical report, 2018a. URL <http://arxiv.org/abs/1808.07169>.
- Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher Information and Natural Gradient Learning of Random Deep Networks. Technical report, 2018b. URL <http://arxiv.org/abs/1808.07172>.
- P W Amderson. More Is Different. *Science*, 177(4047):393–396, 1972. ISSN 0036-8075. doi: 10.1126/science.177.4047.393.
- Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A Probabilistic Framework for Deep Learning. In *NIPS*, 2016.
- Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations. *Theoretical Computer Science*, 633:112–121, jun 2015.
- Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning. *Information and Inference, Special Issue: Deep Learning*, may 2016.
- N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable Bounds for Learning Some Deep Representations. In *ICML*, 2014.
- R.B. Ash. *Real analysis and probability*. Probability and mathematical statistics. Academic Press, 1972.
- Randall Balestriero and Richard Baraniuk. A Spline Theory of Deep Networks (Extended Version). In *ICML*, 2018.
- A. Banyaga. *The structure of classical diffeomorphism groups*. Mathematics and its applications. Kluwer Academic, 1997. ISBN 9780792344759.
- Yaneer Bar-Yam. *Dynamics of Complex Systems*. Perseus Books, Cambridge, MA, USA, 1997. ISBN 0-201-55748-7.
- J. Barzun. *From Dawn to Decadence: 500 Years of Western Cultural Life 1500 to the Present*. HarperCollins, 2001. ISBN 9780060928834.
- Juan Benet. IPFS - Content Addressed , Versioned , P2P File System. Technical Report Draft 3, 2014.

- Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.
- Léon Bottou, Yoshua Bengio, and Yann LeCun. Document analysis with transducers. Technical Report Technical Memorandum HA615600-960701-01TM, 1996. URL <http://www.iro.umontreal.ca/~lisa/pointeurs/transducer-tm.ps.gz>.
- Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231, 2001. ISSN 08834237. doi: 10.2307/2676681.
- Ej Candes. *New ties between computational harmonic analysis and approximation theory*. 2002. ISBN 0826514162.
- M. Castells. *The Rise of The Network Society: The Information Age: Economy, Society and Culture*. Number v. 1 in Information Age Series. Wiley, 2000. ISBN 9780631221401.
- M. Castells. *End of Millennium*. Number v. 3. Wiley, 2010. ISBN 9781444323443.
- M. Castells. *The Power of Identity*. Information Age Series. Wiley, 2011. ISBN 9781444356298.
- Anna Choromanska, Mikael Henaff, and Michael Mathieu. The Loss Surfaces of Multilayer Networks. In *AISTATS*, 2015a.
- Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open Problem: The landscape of the loss surfaces of multilayer networks. In *COLT*, 2015b.
- Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks. In *ICML*, 2016.
- Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In *NIPS*, 2016.
- Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, 2014.
- W. Deresiewicz. *Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life*. Free Press, 2014. ISBN 9781476702735.
- Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting Cyclic Symmetry in Convolutional Neural Networks. In *ICML*, 2016.
- David L Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *American Math. Society Lecture-Math Challenges of the 21st Century*, pages 1–33, 2000. doi: 10.1.1.329.3392.
- H.L. Dreyfus. *What Computers Cant Do - Scholar’s Choice Edition*. Bibliolife DBA of Bibilio Bazaar II LLC, 2015. ISBN 9781298030009.
- N. Dunford and J.T. Schwartz. *Linear operators. Part 1: General theory*. Pure and Applied Mathematics. Interscience Publishers, 1957.
- Ruineihart David E., Hinton Geoffrey E., Williams J., and Ronald. Learning Internal Representations by Error Propagation. *Parallel distributed processing: explorations in the microstructure of cognition*, 1(V):318–362, 1986.
- László Erdős, Torben Krüger, and Dominik Schröder. Random Matrices with Slow Correlation Decay. Technical report, 2017. URL <http://arxiv.org/abs/1705.10661>.
- Kunihiko Fukushima. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36(4):193–202, 1980.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *AISTATS*, 2011. doi: 10.1.1.208.6449.
- Uffe Haagerup and Steen Thorbjørnsen. A new application of random matrices:. *Annals of Mathematics*, 162(2): 711–775, 2005. ISSN 0003-486X. doi: 10.4007/annals.2005.162.711.

- D. Harvey. *The Condition of Postmodernity*. Blackwell, 1989.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *ICCV*, 2015.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.
- J. William Helton, Reza Rashidi Far, and Roland Speicher. Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints. *International Mathematics Research Notices*, 2007:1–14, 2007. ISSN 10737928. doi: 10.1093/imrn/rnm086.
- Pekka Himanen. *The Hacker Ethic and the Spirit of the Information Age*. Random House Inc., New York, NY, USA, 2001. ISBN 0375505660.
- Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–54, 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527.
- Sepp Hochreiter, S Hochreiter, Jürgen Schmidhuber, and J Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–80, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.
- Christian Ioffe, Sergey and Szegedy. Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015.
- Prateek Jain and Purushottam Kar. Non-convex Optimization for Machine Learning. 2017. ISSN 1935-8237. doi: 10.1561/22000000058.
- Yasaman Bahri Jeffrey Pennington. Geometry of Neural Network Loss Surfaces via Random Matrix Theory. In *ICML*, 2017.
- L.P. Kadanoff. *Statistical Physics: Statics, Dynamics and Renormalization*. Statistical Physics: Statics, Dynamics and Renormalization. World Scientific, 2000. ISBN 9789810237646.
- Kenji Kawaguchi. Deep Learning without Poor Local Minima. In *NIPS*, 2016. ISBN 9781627480031. doi: 10.1177/0278364910371999.
- A. Klenke. *Probability Theory: A Comprehensive Course*. World Publishing Corporation, 2012. ISBN 9787510044113.
- Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193, 9780262013192.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.
- Norbert Kruger, Peter Janssen, Sinan Kalkan, Markus Lappe, Ales Leonardis, Justus Piater, Antonio J. Rodriguez-Sanchez, and Laurenz Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1847–1871, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.272.
- Thomas Laurent and James von Brecht. Deep linear neural networks with arbitrary loss: All local minima are global. In *ICML*, 2018.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. doi: 10.1038/nature14539.
- J. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2012. ISBN 9781441999825.
- S. Levy. *Hackers: Heroes of the Computer Revolution*. Anchor Press/Doubleday, 1984. ISBN 9780385191951.

- Shiyu Liang, Ruoyu Sun, Yixuan Li, and R Srikant. Understanding the Loss Surface of Neural Networks for Binary. In *ICML*, 2018.
- Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. ISSN 00189448. doi: 10.1109/TIT.2006.881731.
- Henry W. Lin and Max Tegmark. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, aug 2017.
- Zachary C Lipton and Jacob Steinhardt. Troubling Trends in Machine Learning Scholarship. In *ICML Debate Workshop*, 2018.
- J.R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics: 3rd. ed.* John Wiley & Sons, Limited, 2007.
- Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them. In *CVPR*, 2015.
- Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional Kernel Networks. In *NIPS*, 2014.
- Stéphane Mallat. Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, LXV:1331–1398, jan 2012.
- Stéphane Mallat. Understanding deep convolutional networks. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065), 2016. ISSN 1364-503X. doi: 10.1098/rsta.2015.0203.
- James S. A. Corey Mark Fergus, Hawk Ostby. The expanse, 2018. URL [https://en.wikipedia.org/wiki/The_Expanse_\(TV_series\)](https://en.wikipedia.org/wiki/The_Expanse_(TV_series)).
- John McCarthy. Recursive functions symbolic expressions and their computation by machine, Part I. *Communications of the ACM*, 3(4):184–195, 1960. ISSN 00010782. doi: 10.1145/367177.367199.
- P. McCullagh. *Tensor methods in statistics*. Monographs on statistics and applied probability. Chapman and Hall, 1987. ISBN 9780412274800.
- C.L. Mee. *Life in the Renaissance*. New Word City, 2016. ISBN 9781612309453.
- Pankaj Mehta and David J. Schwab. An exact mapping between the Variational Renormalization Group and Deep Learning. Technical report, oct 2014. URL <http://arxiv.org/abs/1410.3831>.
- M. Montaigne and M.A. Screech. *The Complete Essays*. Penguin classics. Penguin Books Limited, 2004. ISBN 9780141915937.
- Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash. Technical report, 2008.
- M E J Newman. Complex Systems: A Survey. *Physics Reports*, 79(I):10, 2009. ISSN 00029505. doi: 10.1119/1.3590372.
- Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *ICML*, 2017.
- Quynh Nguyen and Matthias Hein. Optimization Landscape and Expressivity of Deep CNNs. In *ICML*, 2018.
- Gregoire Nicolis and Catherine Nicolis. *Foundations of Complex Systems: Emergence, Information and Prediction*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2nd edition, 2012. ISBN 9789814366601, 9814366609.
- Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. Technical report, 2009. URL <http://arxiv.org/abs/0911.4863>.
- Judea Pearl. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. In *ICWSDM*, jan 2018.
- Tomaso Poggio, Qianli Liao, Brando Miranda, Andrzej Banburski, Xavier Boix, and Jack Hidary. Theory IIIb: Generalization in Deep Networks. Technical Report 90, 2018. URL <http://arxiv.org/abs/1806.11379>.

- Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for Activation Functions. Technical report, oct 2017. URL <http://arxiv.org/abs/1710.05941>.
- M Riesenhuber and T Poggio. Models of object recognition. *Nature Neuroscience. Special Issue: Computational approaches to brain function*, 3(Suppl):1199–1204, 2000. ISSN 1097-6256. doi: 10.1038/81479.
- Maximilian Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.
- F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519.
- Tim Salimans and Diederik P. Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *NIPS*, 2016.
- Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.
- Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. Technical report, mar 2017. URL <http://arxiv.org/abs/1703.00810>.
- K. Sigmund and D. Hofstadter. *Exact Thinking in Demented Times: The Vienna Circle and the Epic Quest for the Foundations of Science*. Basic Books, 2017. ISBN 9780465096961.
- P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *International Conference on Document Analysis and Recognition*. IEEE Comput. Soc, 2003. ISBN 0-7695-1960-1. doi: 10.1109/ICDAR.2003.1227801.
- Herbert A. Simon. The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106(6): 467–482, 1962. ISSN 1475-9551. doi: 10.1080/14759550302804.
- S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the Neural Response. *Foundations of Computational Mathematics*, 10(1):67–91, jun 2009. ISSN 1615-3375. doi: 10.1007/s10208-009-9049-1.
- Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. In *ICLR*, 2018.
- Christian Szegedy, W Zaremba, and I Sutskever. Intriguing properties of neural networks. *arXiv preprint arXiv: ...*, pages 1–10, 2013.
- T. Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc., 2012. ISBN 9780821885079.
- Hermann Thorisson. *Coupling, stationarity, and regeneration*. Springer-Verlag Inc, Berlin; New York, 2000.
- Paolo E Trevisanutto. Expectation propagation : a probabilistic view of Deep Feed Forward Networks. Technical report, 2018. URL <https://arxiv.org/abs/1805.08786>.
- F. Turner. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. University of Chicago Press, 2010. ISBN 9780226817439.
- V N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 10(5):988–99, 1999. ISSN 1045-9227. doi: 10.1109/72.788640.
- C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.
- P. Watson. *Modern Mind: An Intellectual History of the 20th Century*. HarperCollins, 2002. ISBN 9780060084387.
- Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.
- E. P. Wigner. Statistical properties of real symmetric matrices with many dimensions.pdf. In *Canadian Mathematical Congress Proceedings*, 1957.

- Wikipedia. Klimt university of vienna ceiling paintings, 2018. URL https://en.wikipedia.org/wiki/Klimt_University_of_Vienna_Ceiling_Paintings.
- Edward O. Wilson. *Consilience: The Unity of Knowledge*. ISSR library. Vintage Books, 1999. ISBN 9780679768678.
- Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. In *ICLR*, 2018.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2016. ISBN 1506.02142. doi: 10.1109/TKDE.2015.2507132.
- Quanshi Zhang and Song-Chun Zhu. Visual Interpretability for Deep Learning: a Survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, feb 2018.

Appendices

All definitions present in the appendices are adopted and reproduced from existing literature with sources cited, for the purpose of making exact the terminology used in the paper.

A Coupling Theory

The following definitions are adapted from Thorisson (2000) unless otherwise noted.

Definition A.1 (Random Element; Random Variable; Random Function). *A random element in a measurable space (E, \mathcal{E}) defined on a probability space (F, \mathcal{F}, μ) is a measurable mapping T from (F, \mathcal{F}, μ) to (E, \mathcal{E}) , where*

$$T^{-1}A \in \mathcal{F}, A \in \mathcal{E}; T^{-1}A := \{w \in F : T(w) \in A\}$$

*We say T is **supported by** probability measure space (F, \mathcal{F}, μ) , (F, \mathcal{F}, μ) is the support of T , T is an \mathcal{F}/\mathcal{E} **measurable mapping** from F to E , and the induced measure $\mu(T^{-1}(A))$ is the **law** of T . Some r.e.s have special names. When (E, \mathcal{E}) is the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where \mathbb{R} is the real number field and $\mathcal{B}(\mathbb{R})$ is the Borel set, X is also named as a **random variable**, whose abbreviation is r.v.; when (E, \mathcal{E}) is the multivariate real measurable space, T is named as a multivariate r.v.; when (E, \mathcal{E}) is a function space satisfying certain conditions (Ionescu-Tulcea theorem, or Kolmogorov's extension theorem (Klenke (2012))), T is named as a **random function**.*

Definition A.2 (Coupling). *A probability measure μ on $\bigotimes_{i \in \mathbb{I}} (E_i, \mathcal{E}_i)$ is a coupling of $\mu_i, i \in \mathbb{I}$, if μ_i is the i th marginal of μ , that is, if μ_i is induced by the i th projection mapping:*

$$\mu(\{x : x_i \in A\}) = \mu_i(A), A \in \mathcal{E}_i, i \in \mathbb{I}$$

where \mathbb{I} is an index set, μ_i is a probability measure on a measurable space (E_i, \mathcal{E}_i) , and $\bigotimes_{i \in \mathbb{I}} (E_i, \mathcal{E}_i) := (\prod_{i \in \mathbb{I}} E_i, \bigotimes_{i \in \mathbb{I}} \mathcal{E}_i)$, $\prod_{i \in \mathbb{I}} E_i$ is the Cartesian product of E_i and $\bigotimes_{i \in \mathbb{I}} \mathcal{E}_i$ is the product σ -algebra.

The general idea of coupling is to find a dependence structure (joint distribution) from fixed marginal distributions that fits one's purpose.

Definition A.3 (Coupling of Random Element; Deterministic Coupling; Transport Map). *Given two r.e.s X, Y , a coupling (X, Y) refers to the coupling of probability measure μ, ν of probability space $(E, \mathcal{E}, \mu), (F, \mathcal{F}, \nu)$, where μ, ν is the probability measure of r.e. X, Y respectively. The coupling (X, Y) is called **deterministic** if there exists a measurable function $T : \mathcal{E} \rightarrow \mathcal{F}$ such that $Y = T(X)$. T is normally referred as **transport map**. Informally, we say that T transports measure μ of X to measure ν of Y . The definitions are adapted from Villani (2008).*

To recognize an event ω_s that is composed of events of the lowest detectable scale Ω_{s_0} , the idea is to transport the probability measure of the event ω_s through a deterministic coupling, and construct a r.e. that represents possible states ω_s may take. We introduce concepts needed in the following.

Definition A.4 (Extension of Probability Space). *A probability space $(\bar{F}, \bar{\mathcal{F}}, \bar{\mu})$ is an extension of another probability space (F, \mathcal{F}, μ) , if $(\bar{F}, \bar{\mathcal{F}}, \bar{\mu})$ supports a r.e. ξ in (F, \mathcal{F}) having law μ . If T is a r.e. in (E, \mathcal{E}) defined on (F, \mathcal{F}, μ) , then it has a **copy** \bar{T} , i.e., the r.e. \bar{T} defined on $(\bar{F}, \bar{\mathcal{F}}, \bar{\mu})$ by $\bar{T}(\bar{\omega}) = T(\xi(\bar{\omega})), \bar{\omega} \in \bar{\Omega}$ and $\bar{\mu}(\bar{T}(\cdot)) = \mu(T(\cdot))$. \bar{T} is called **original** r.e.. New r.e.s may be created, which is called **external** r.e.s.*

The goal of S-System is to create extensions of a measurement collection r.e. \mathcal{X} to reconstruct Ω_s for some $s \in \mathcal{S}$ in PPMS, such that $\mu^{\mathcal{X}}(H_s = h_s) = \mu^{\mathcal{W}}(\omega_s)$, where H_s is a r.e. created through extension, and h_s is a realized value of it. However, we do not possess the p.d.f. of X , which we have to rely on realizations of X . What we can do is to leverage and only leverage available information through conditioning extension. Furthermore, it is likely for limitation of capacity and economy considerations, S-System has to group together events, thus discretizing them.

Definition A.5 (Probability Kernel). *Given two measurable space $(E_1, \mathcal{E}_1), (E_2, \mathcal{E}_2)$, a function $Q(\cdot, \cdot)$ from $E_1 \times \mathcal{E}_2$ to $[0, 1]$ is an $((E_1, \mathcal{E}_1), (E_2, \mathcal{E}_2))$ probability kernel if 1) $Q(\cdot, A)$ is $\mathcal{E}_1/\mathcal{B}([0, 1])$ measurable for each $A \in \mathcal{E}_2$ 2) and $Q(y, \cdot)$ is probability measure on (E_2, \mathcal{E}_2) for each $y \in E_1$. A probability kernel uniquely determines a probability measure on $(E_1, \mathcal{E}_1) \bigotimes (E_2, \mathcal{E}_2)$ (Ash (1972) Section 2.6.2).*

Definition A.6 (Conditioning Extension). Let T_1 be an r.e. in (E_1, \mathcal{E}_1) defined on a probability measure space (F, \mathcal{F}, μ) , and let $Q(\cdot, \cdot)$ be an $((E_1, \mathcal{E}_1), (E_2, \mathcal{E}_2))$ probability kernel. A conditioning extension $(\bar{F}, \bar{\mathcal{F}}, \bar{\mu})$ of (F, \mathcal{F}, μ) is created by letting

$$\begin{aligned} (\bar{F}, \bar{\mathcal{F}}) &:= (F, \mathcal{F}) \otimes (E_2, \mathcal{E}_2) & \bar{\mu}(A \times B) &:= \int_A Q(T_1(\omega), B) \mu(d\omega), A \in \mathcal{F}, B \in \mathcal{E}_2 \\ \xi(\omega, t) &:= \omega, \omega \in \Omega, t \in E_2 & \bar{T}_1(\omega, t) &:= T_1(\omega), \omega \in \Omega, t \in E_2 \quad \bar{T}_2(\omega, t) := t, t \in E_2 \end{aligned}$$

\bar{T}_1 is the original r.e.s in the new probability space, while \bar{T}_2 is a new external r.e. created. Conditioning extension can be repeated countably many times (Ash (1972) Section 2.7.2).

Definition A.7 (Splitting Extension). Let T_1, T_2 be r.e.s in $(E_1, \mathcal{E}_1), (E_2, \mathcal{E}_2)$ respectively, defined on a probability space (F, \mathcal{F}, μ) . Let ν be a probability measure on a Polish space (E_3, \mathcal{E}_3) , let $Q(\cdot, \cdot)$ be an $((E_3, \mathcal{E}_3), (E_2, \mathcal{E}_2))$ probability kernel, and suppose

$$\mu(T_2 \in A) = \int Q(t, A) \nu(dt), A \in \mathcal{E}_2$$

Then a splitting extension of (F, \mathcal{F}, μ) is to create an extension to support a r.e. T_3 in (E_3, \mathcal{E}_3) having distribution ν , and such that

$$\mu(T_2 \in \cdot | T_3 = t) = Q(t, \cdot), t \in E_3$$

Furthermore, T_1 is conditionally independent of T_3 given T_2 .

Remark. The extension is called a splitting extension for it splits the probability measure on (E_2, \mathcal{E}_2) into many groups of measure identified by t .

B Manifold and Diffeomorphism Group

The following definitions have been adapted from Lee (2012) unless otherwise noted.

Definition B.1 (Topological Manifold). Suppose M is a topological space. We say that M is a **topological manifold of dimension n** or a **topological n -manifold** or just a **n -manifold** if it has the following properties:

- M is a **Hausdorff space**: for every pair of distinct points $p, q \in M$, there are disjoint open subsets $U, V \subseteq M$ such that $p \in U$ and $q \in V$.
- M is **second-countable**: there exists a countable basis for the topology of M .
- M is **locally Euclidean of dimension n** : each point of M has a neighborhood that is homeomorphic to an open subset of \mathbb{R}^n .

Definition B.2 (Smooth Mapping; Diffeomorphism). If U, V are open subsets of Euclidean spaces \mathbb{R}^n and \mathbb{R}^m , respectively, a function $f : U \rightarrow V$ is said to be **smooth** (or C^∞ , or **infinitely differentiable**) if each of its component functions has continuous partial derivatives of all orders. If in addition f is bijective and has a smooth inverse map, it is called a **diffeomorphism**.

Definition B.3 (Chart; Coordinate Chart; Smooth Compatible). Let M be a topological n -manifold. A **coordinate chart** (or just a **chart**) on M is a pair (U, ψ) , where U is an open subset of M and $\psi : U \rightarrow \hat{U}$ is a homeomorphism from U to an open subset $\hat{U} = \psi(U) \subseteq \mathbb{R}^n$. U is called a **coordinate domain**, or just **domain**, ψ is called a **(local) coordinate map**, and the component functions (x^1, \dots, x^n) of ψ , defined by $p \in U, \psi(p) = (x^1(p), \dots, x^n(p))$, are called **local coordinates** on U . Two $(U, \phi), (V, \psi)$ are called **smoothly compatible** if either $U \cap V = \emptyset$, or $\psi \circ \phi^{-1}$ is a diffeomorphism.

Definition B.4 (Atlas; Smooth Atlas; Maximal Atlas). Let M be a topological manifold. An **atlas \mathcal{A}** for M is a collection of charts whose domain cover M . If any two charts in \mathcal{A} is smoothly compatible with each other, it is called a **smooth atlas**. A smooth atlas \mathcal{A} on M is **maximal** if it is not properly contained in any larger smooth atlas.

Definition B.5 (Smooth Manifold). A **smooth manifold** is a pair (M, \mathcal{A}) , where M is a topological manifold and \mathcal{A} is a maximal smooth atlas on M . When no confusion exists, we may just say “ M is a smooth manifold”.

Definition B.6 (Riemannian Metric). A **Riemannian metric** g of a smooth manifold M is a symmetric covariant 2-tensor field on M that is positive definite at each point. It defines an inner product on M , which informally, could be represented by a quadratic form $\boldsymbol{\eta}^T \mathbf{G} \boldsymbol{\eta}'$, where $\mathbf{G} = (g_{ij})$ is a matrix and $\boldsymbol{\eta}, \boldsymbol{\eta}'$ are the local coordinates of two points in M .

Definition B.7 (Riemannian Manifold). A **Riemannian manifold** is a pair (M, g) , where M is a smooth manifold and g is a Riemannian manifold. Or in short, we could say M is a Riemannian manifold if M is understood to be endowed with a specific Riemannian metric.

The following definition is adopted from Banyaga (1997).

Definition B.8 (C^∞ Diffeomorphism Group). Let $C^\infty(M, N)$ denote the space of all C^∞ mapping $f : M \rightarrow N$, where M, N are smooth manifolds. The diffeomorphism group, denoted by $\text{Diff}^\infty(M)$, is the set of all C^∞ diffeomorphisms of M , the group action of which is the mapping composition.

To make the definition more concrete to help understanding, we provide an example adopted from Mallat (2016).

Example B.1. The diffeomorphism group is the set of deformation that may be applied to objects, e.g., images, of which we can define a norm to characterize the deformation. A small diffeomorphism acting on a function $x(u)$ defined on \mathbb{R}^n can be written as a translation of u by a $g(u)$:

$$g.x(u) = x(u - g(u)), g \in \text{Diff}^\infty(\mathbb{R}^n)$$

Note that the smooth condition is not necessary, and is only used to avoid introducing further definitions. The diffeomorphism translates points by at most $\|g\|_\infty = \sup_{u \in \mathbb{R}^n} |g(u)|$. Small diffeomorphism corresponds to $\|\nabla g\|_\infty = \sup_u |\nabla g(u)| < 1$, where $|\nabla g|$ is the matrix norm of the Jacobian matrix of g at u . Thus, an element in a subset of $\text{Diff}^\infty(\mathbb{R}^n)$ can be understood as a small deformation of images where the deformation is bounded.

C Information Geometry

The following definitions are adapted from Amari (2016).

Definition C.1 (Divergence). Suppose that M is a n -manifold, of which the points have a local coordinates system $\boldsymbol{\eta}$. Given two points $\mathbf{p}, \mathbf{q} \in M$, the coordinates of which are $\boldsymbol{\eta}_p, \boldsymbol{\eta}_q$ respectively, a **divergence** is a function of $\boldsymbol{\eta}_p, \boldsymbol{\eta}_q$, written as $D[\mathbf{p} : \mathbf{q}]$ or $D[\boldsymbol{\eta}_p : \boldsymbol{\eta}_q]$, which satisfies the following criteria.

- $D[\mathbf{p} : \mathbf{q}] \geq 0$.
- $D[\mathbf{p} : \mathbf{q}] = 0$, if and only if $\mathbf{p} = \mathbf{q}$.

When \mathbf{p} and \mathbf{q} are sufficiently close, and D is differentiable, by denoting their coordinates by $\boldsymbol{\eta}_p$ and $\boldsymbol{\eta}_q = \boldsymbol{\eta}_p + d\boldsymbol{\eta}$, the Taylor expansion of D is written as

$$D[\boldsymbol{\eta}_p : \boldsymbol{\eta}_q + d\boldsymbol{\eta}] = \frac{1}{2} \sum g_{ij}(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_i \boldsymbol{\eta}_j + O(|d\boldsymbol{\eta}|^3)$$

, and matrix $\mathbf{G} = (g_{ij})$ is positive-definite, depending on $\boldsymbol{\eta}_p$.

Definition C.2 (Bregman Divergence). Given a convex function $\psi(\boldsymbol{\eta})$, a **Bregman divergence** derived from ψ is a divergence defined as

$$D_\psi[\boldsymbol{\eta} : \boldsymbol{\eta}'] = \psi(\boldsymbol{\eta}') - \psi(\boldsymbol{\eta}) - \nabla \psi(\boldsymbol{\eta})^T (\boldsymbol{\eta}' - \boldsymbol{\eta})$$

Definition C.3 (Legendre Dual; Legendre Transform). Given a convex function $\psi(\boldsymbol{\eta})$, the **Legendre dual** of ψ is the function ψ^*

$$\psi^*(\boldsymbol{\eta}^*) = \boldsymbol{\eta}^T \boldsymbol{\eta}^* - \psi(\boldsymbol{\eta})$$

where $\boldsymbol{\eta} = \mathbf{f}(\boldsymbol{\eta}^*)$ and \mathbf{f} is the inverse function of $\boldsymbol{\eta}^* = \nabla \psi(\boldsymbol{\eta})$. ψ^* is a convex function. $\nabla \psi(\boldsymbol{\eta})$ is called the **Legendre Transform** of $\boldsymbol{\eta}$.

Definition C.4 (Dual Bregman Divergence). Given a convex function $\psi(\boldsymbol{\eta})$, and D_ψ the Bregman divergence derived by ψ . Let ψ^* be the Legendre dual function of ψ , then the Bregman divergence D_{ψ^*} defined by ψ^* is called the **Dual Bregman Divergence** derived by ψ . We have

$$D_{\psi^*}[\boldsymbol{\eta}^* : \boldsymbol{\eta}'^*] = D_\psi[\boldsymbol{\eta}' : \boldsymbol{\eta}]$$

Definition C.5 (Exponential Family of Probability Distributions; Stochastic Manifold; Affine Coordinate System; Dual Affine Coordinate System). *The form of **probability distribution of exponential family** is given by the probability density function*

$$p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = e^{(\boldsymbol{\theta}^T \mathbf{h}(\mathbf{x}) - \psi(\boldsymbol{\theta}))} d\mu(\mathbf{x})$$

where \mathbf{x} is a realizable value of a multivariate random variable, $\mathbf{h}(\mathbf{x})$ is a vector function of \mathbf{x} which are linearly independent, $\psi(\boldsymbol{\theta})$ is the normalization factor, and μ is the law on r.v. \mathbf{x} .

Since ψ is a convex function w.r.t. $\boldsymbol{\theta}$, the exponential family distributions is a Riemannian manifold (M, g) with an **affine coordinate system** $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, and g is given by the second order Taylor expansion of the Bregman divergence derived from ψ . It is called the **stochastic manifold of exponential family distribution**. $\boldsymbol{\theta}$ is called natural or canonical parameters. An alternative coordinate system of M is given by the Legendre transform

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}) = \int \mathbf{h} e^{\boldsymbol{\theta}^T \mathbf{h}(\mathbf{x})} d\mu(\mathbf{x})$$

of $\boldsymbol{\theta}$, which is well known as the expectation parameter in statistics, and is called **dual affine coordinate system**. Correspondingly, an alternative Riemannian metric is derived from the Legendre dual of ψ . The Bregman divergence derived is the well known discrepancy measure on probability, the KL divergence.

D Statistical Learning Theory

Assume a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where \mathcal{X} is the space of input data, and \mathcal{Y} is the label space. We use $S_m = \{s_i = (\mathbf{x}_i, y_i)\}_{i=1}^m$ to denote the training set of size m whose samples are drawn independently and identically distributed (i.i.d.) according to an unknown distribution P . Given a loss function l , the goal of learning is to identify a function $f : \mathcal{X} \mapsto \mathcal{Y}$ in a hypothesis space (a class \mathcal{F} of functions) that minimizes the expected error

$$R(f) = \mathbb{E}_{z \sim P} [l(f(\mathbf{x}), y)],$$

where $z = (\mathbf{x}, y) \in \mathcal{Z}$ is sampled i.i.d. according to P . Since P is unknown, the observable quantity serving as the proxy to the true risk $R(f)$ is the empirical error

$$R_m(f) = \frac{1}{m} \sum_{i=1}^m l(f(\mathbf{x}_i), y_i).$$