Synthese (2024) 204:74
https://doi.org/10.1007/s11229-024-04702-z

ORIGINAL RESEARCH

Why and how to construct an epistemic justification of machine learning?

Petr Spelda[1] · **Vit Stritecky**[1]

Received: 17 March 2024 / Accepted: 5 July 2024 / Published online: 10 August 2024

Abstract

Consider a set of shuffled observations drawn from a fixed probability distribution over some instance domain. What enables learning of inductive generalizations which proceed from such a set of observations? The scenario is worthwhile because it epistemically characterizes most of machine learning. This kind of learning from observations is also inverse and ill-posed. What reduces the non-uniqueness of its result and, thus, its problematic epistemic justification, which stems from a one-to-many relation between the observations and many learnable generalizations? The paper argues that this role belongs to any complexity regularization which satisfies Norton's Material Theory of Induction (MTI) by localizing the inductive risk to facts in the given domain. A prime example of the localization is the Lottery Ticket Hypothesis (LTH) about overparameterized neural networks. The explanation of MTI's role in complexity regularization of neural networks is provided by analyzing the stability of Empirical Risk Minimization (ERM), an inductive rule that controls the learning process and leads to an inductive generalization on the given set of observations. In cases where ERM might become asymptotically unstable, making the justification of the generalization by uniform convergence unavailable, LTH and MTI can be used to define a local stability. A priori, overparameterized neural networks are such cases and the combination of LTH and MTI can block ERM's trivialization caused by equalizing the strengths of its inductive support for risk minimization. We bring closer the investigation of generalization in artificial neural networks and the study of inductive inference and show the division of labor between MTI and the optimality justifications (developed by Gerhard Schurz) in machine learning.

Keywords Lottery ticket hypothesis · Complexity regularization · Material theory of induction · Empirical risk minimization

✉ Petr Spelda
petr.spelda@fsv.cuni.cz

[1] Department of Security Studies, Institute of Political Studies, Faculty of Social Sciences, Charles University, U Kříže 8, 158 00 Praha 5, Czech Republic

1 Introduction

An epistemic justification of the inductive generalization in artificial neural networks can be achieved by connecting the state-of-the-art approaches (Bengio et al., 2021; LeCun et al., 2015; Schmidhuber, 2015) to recent theories of inductive inference (Norton, 2003, 2021; Schurz, 2019). If machine learning is considered as a kind of induction, then the epistemic justification is missing in machine learning as well as in epistemology debates.

In Sects. 2 and 3, the paper connects complexity regularization of (deep) artificial neural networks, the Lottery Ticket Hypothesis (Frankle & Carbin, 2019), and the Material Theory of Induction (Norton, 2003, 2014, 2021) to show that successful machine learning of inductive generalizations is epistemically justifiable by the localization of inductive risk. Sections 4 and 5 provide important qualifications to this epistemic justification, using Norton's work on the incompleteness of calculi of inductive inference (2019) to distinguish between asymptotic and local stability of inductive rules that facilitate the generalization learning in neural networks.

The Material Theory of Induction argues that retrodictive/predictive successes of induction stem from adapting general inductive schemas to material facts found in local domains, thus achieving the schemas' localization. The paper shows that neural network pruning described by the Lottery Ticket Hypothesis adapts a general architecture to a given local domain. By this, it transports the inductive risk from a schema (architecture) to local facts populating the evidence (training data), thus accomplishing the localization. If the requirements of statistical learning theory are met (i.i.d. [independent and identically distributed] samples from a fixed distribution over some instance domain, see Sect. 4), then any regularization method, satisfying the Material Theory of Induction by moving the inductive risk to local facts, can provide the inductive generalization in (deep) artificial neural networks with an epistemic justification. In case the requirements are not met, then local facts become unstable. The Material Theory of Induction can no longer provide the epistemic justification, which should be replaced with an optimality-based justification from the framework developed by Gerhard Schurz (see the conclusion of Sect. 5; also, Schurz, 2019, 2024; Spelda & Stritecky, 2021). This distinction shows a division of labor between John Norton's and Gerhard Schurz's theories of inductive inference (cf. Schurz & Thorn, 2020) and the limitation of the former theory in the machine learning context.

1.1 The motivation for an epistemic justification of inductive generalizations

Statistical learning theory treats training of a machine learning model as function estimation from a limited sample of training data (Vapnik, 1995). This means that instead of identifying the true model by estimating the function entirely, the true model is being 'imitated' (cf. Cherkassky & Dhar, 2015) by estimating the function at a given finite set of points of interest (Vapnik, 1995, pp. 167–170). Therefore, an epistemic justification for the ability of machine learning models to generalize is required.

To generalize is to perform correct inferences on new (yet unobserved) samples outside of training data by establishing a certain kind of connections among the observed

samples (training data). Formally (Definition 1), '$r\%$ of all so far observed Fs has been Gs, hence, with high (subjective) probability, approximately $r\%$ of all Fs are (will be) Gs' (Schurz, 2019, p. 2). Using this definition of inductive generalization, Gs are successful inferences by a machine learning model on new samples, provided all Fs are (will be) i.i.d. from a fixed distribution $\mathcal{D}$ on some domain $\mathcal{Z}$. Any function f from the class $\mathcal{F}$ expressible by a machine learning model is not estimated in full during training but rather it is estimated only at a given finite set of points of interest. A deductive inference from the entire function estimation to the points of interest thus cannot be performed (cf. Vapnik, 1995, p. 169, Fig. 6.1), motivating the desirability of an epistemic justification for the generalization capability of the trained model.

Determined by the class $\mathcal{F}$, different function estimations are possible on any finite training dataset, which leads to different generalization capabilities of the model. The problem of selecting the correct level of complexity of the model, determining its generalization performance via the fit to data (i.e., selecting $f \in \mathcal{F}$), is central. This is our core concern regarded as the one-to-many relation between the evidence (training data) and many learnable generalizations. The expressivity of artificial neural networks is not foundationally limited, since the universal approximation theorem established that multilayer feedforward neural networks using sigmoidal hidden layer activation functions are universal approximators (Cybenko, 1989; Hornik et al., 1989). Borel measurable functions from one finite dimensional space to another can be approximated to an arbitrary degree of accuracy if the network is large enough [see Barron (1993) for bounds on the sizes of single hidden layer networks for the approximation of various function classes, relatedly also Kůrková (1992); Schmidt-Hieber (2021) for more than single hidden layer ReLU networks). For various classes of functions, deep networks can improve the approximation efficiency, decreasing the number of hidden units while increasing the generalization performance (Goodfellow et al., 2016, pp. 192–95; Yarotsky, 2017; DeVore et al., 2021, pp. 398–423). Treating simplicity as a uniform guiding principle for selecting a function f from the class $\mathcal{F}$ is epistemically problematic, because simplicity depends on the local context, i.e., on the facts found in the training data (cf. Norton, 2021, Chap. 6; Roche, 2018).

The generalization capability measured as the error rate on new samples does not depend on a uniquely ordered sequence of the training data. For neural networks, learning starts from random initial conditions followed by a succession of epochs, each iteratively introducing the model to batches of randomly drawn samples from the training data (Goodfellow et al., 2016, pp. 270–73). This procedure is known as the minibatch method, during which the network's parameters (weights) are updated after evaluating its predictions on samples from a minibatch instead of on all samples to avoid costly updates based on processing the whole dataset (ibid.; an epoch is concluded once the network ran through the entire training dataset). Samples in minibatches and individual minibatches should be independent from each other to avoid updating the network's weights based on biased gradient estimates caused by potential dependencies between samples (ibid.). In practice, the training dataset is usually shuffled to simulate the effects of independence, and if several epochs are executed over the shuffled dataset, starting with the second epoch (assuming there no copies among the samples), estimates of the generalization error on minibatches will be biased (Goodfellow et al., 2016, pp. 273–75). The minibatch method is different

from the cross-validation or holdout method (Arlot & Celisse, 2010) because the latter require sample independence to estimate the generalization error in an unbiased way in order to serve as an epistemically justifiable model selection procedure. The minibatch methods have different objectives, e.g., balancing efficient training and overfitting.

This kind of learning does not distinguish among epistemic-temporal locations (indices) of the samples. It can be thus described as an ill-posed inverse problem of learning inductive generalizations from samples produced by not necessarily known empirical processes. De Vito et al. (2005) and Kůrková (2005) established an early connection between statistical learning theory and ill-posed inverse problems, building a formal link between learning from observations of some empirical process and complexity regularization. For perspectives on this kind of generalization learning from the causal inference point of view, one may refer to Kilbertus et al. (2018), Pearl (2019) or Schölkopf et al. (2021).

For ill-posed inverse problems, regularization methods seek to stabilize the learning algorithm by controlling the complexity of the machine learning model in terms of the expressible hypotheses/functions (cf. Shalev-Shwartz & Ben-David, 2014, Chap. 13). The stability is achieved if small variations in the inputs do not cause significant changes in the outputs the machine learning model (ibid.). Sections 4 and 5 analyze the difference between asymptotic and local stability to establish a connection between the latter and the Material Theory of Induction and thus unpack its impact on neural networks. Therefore, the epistemic problem lies in the one-to-many relation between a training dataset and many learnable generalizations, making the justification of selecting one of them by regularizing the network challenging. Next, we show that an epistemic justification of common regularization principles is not easy, which means that the same applies to the generalizations learned by overparameterized neural networks.

2 The epistemic justification of complexity regularization and its challenges

Most attempts to choose the representational capacity of a machine learning model in the hit-or-miss manner produce two outcomes. First, the selected version of the model underfits the evidence (training data) because its capacity (complexity) is too low, leaving the generalization underdetermined (cf. Goodfellow et al., 2016, pp. 107–113). Second, the model might overfit the evidence because its capacity is sufficient to capture insignificant patterns in data, leaving the generalization overdetermined (cf. ibid.). The least favorable outcome prevents the generalization capability because the model can capture negligible patterns, possibly culminating in the evidence memorization. In the case of overparameterization, which allows training 'modern' interpolating networks (Belkin, 2021), only overfitting is relevant, since the bias-variance trade-off transforms into a double descent generalization curve (ibid.). Overparameterization raises a contradiction in explaining the generalization capability (of networks perfectly fitting [noisy] training data) by uniform convergence. Sections 4 and 5 resolve the contradiction in a different way than the emerging theory of interpolation (cf. ibid.).

Regularization principles impose a priori constraints which limit the number of learnable generalizations by prohibiting certain kinds of generalization-establishing connections that can be formed among the pieces of evidence (cf. Wahba, 1995, p. 426). By impeding some connections to stimulate alternative ones, epitomized by the constraints, regularization seeks to prevent overdetermined generalizations. Such a priori constraints target the cases of overfitting emerging from the eliminated kinds of generalization-establishing connections (cf. ibid.). The constraints help to address the one-to-many relation, which makes an epistemic justification of the generalizations challenging. The epistemic justification of the constraints behind complexity regularization is, thus, important.

Regularization terms appended to loss functions used to train the models can also represent expert (domain) knowledge (Borghesi et al., 2020a, b; Lombardi et al., 2020; Silvestri et al., 2020). A regularized loss function then balances the accuracy of performed inferences and the level of satisfaction of the constraints, converted into a regularization term, that represent prior knowledge to ensure the resulting generalization possesses the desired properties (Borghesi et al., 2020a, pp. 5–6). Obtaining such a result directly from data might be difficult and the constraints help to achieve sample efficiency while ensuring that the generalization does not support improper inferences, considering the solutions allowed in the given domain (ibid.).[1]

The most common assumption about complexity regularization suggests that its effects come from smoothness and simplicity of function approximation (Chen & Haykin, 2002, p. 2792). Smoothness is accomplished by the generalization-establishing connections that create local stability (Goodfellow et al., 2016, pp. 152–153). The constraint seeks to encourage a stable decision boundary among individual pieces of the evidence to facilitate correct inferences on yet unobserved similar samples (ibid.). A good model for an evidence-task pair learns a function approximation that does not change rapidly in a small region (ibid.) to avoid increasing the estimation error by overfitting. The regularization effects of smoothness depend on the complexity of the selected model (i.e., a function f from the class $\mathcal{F}$), since any generalization is the result of a trade-off between the estimation error, $\mathbb{E}L(f_n) - \inf_{f\in\mathcal{F}} L(f)$, and approximation error, $\inf_{f\in\mathcal{F}} L(f) - L^*$, controlled by complexity regularization of the class $\mathcal{F}$ (Bartlett et al., 2002).[2]

Norton (2003, pp. 655–657) and others (e.g., Roche, 2018) showed the difficulties of maintaining a uniform (global) definition of simplicity. Thus, we need to ask how to epistemically justify complexity penalties when simplicity derives from local facts. Simplicity treated globally connected parsimony to the likelihood of achieving non-overdetermined generalizations for evidence-task pairs (cf. Sober, 2015, pp. 148–152, where the discussed fundamental epistemic goal is to learn a good generalization for

[1] It is important to note that in this case the regularization term does not score how simple is the hypothesis expressed by the network at the current step but rather whether the generated solution meets the application requirements expressed by the constraints; see, data-driven approaches for solving constrained problems with neural networks, for example, the Partial Latin Square completion problem (Silvestri et al., 2020).

[2] L, L^*—loss and loss of the optimal prediction rule respectively; $f_n \in \mathcal{F}$—a predictor with L as close as possible to L^* (Bartlett et al., 2002). The predictor is learned using an i.i.d. set $\{z_1, \ldots, z_n\}$ from an unknown distribution $\mathcal{D}$ over some domain $\mathcal{Z}$ (ibid.).

the given evidence-task pair). Since the epistemic goal is to learn generalizations that support correct inferences on yet unobserved samples, a goal which is distinct from training models as simple as possible, parsimony as a global principle should not be invoked to justify complexity regularization. This casts doubt on the epistemic indispensability of tools like Ockham's Razor and its variations, which are usually counted among the fundamental principles of regularization theory (cf. Chen & Haykin, 2002, p. 2832) and treat simplicity as a global rather than local matter.

Simplicity is often replaced with compression, which is understood in the identically global manner. Compression was used to connect regularization to complexity developed in information theory and its algorithmic variant (Chen & Haykin, 2002, pp. 2821–2823; pp. 2817–2818). By relying on the Kolmogorov complexity-based minimum description length, the latter theory expresses complexity as the length of the shortest program able to reconstruct the input object, with the intuition that increasing fidelity of the reconstruction accompanied by the decreasing program length creates regularization effects, i.e., reduces overfitting and thus the generalizations' overdetermination (Chen & Haykin, 2002, pp. 2817–2818). The former theory utilizes Shannon entropy and rate-distortion to show that entropy minimization has regularization effects which control the models' complexity (Chen & Haykin, 2002, pp. 2821–2823). It is expected that minimizing the conditional entropy between the evidence and the generalization creates sparse connections among the pieces of evidence (ibid.). Learning should spread out the generalization-establishing connections among a limited number of the network's hidden units, reducing its complexity (cf. ibid.).

Both theories imply that successful generalization learning minimizes the amount of information needed to produce a good model for the given evidence-task pair. This relationship between generalization and compression is described by the information bottleneck theory, positing a positive relation between maximizing the information about the task at hand and keeping the information about the evidence sparse (Tishby et al., 1999). Given an evidence-task pair, mutual information between the evidence and the generalization emerges from compression in an information bottleneck, representing a good model for the evidence-task pair. Considering the regularization's role in the model selection, compression can act as a drop-in replacement for simplicity treated globally, insomuch as it, too, offers a global remedy for overfitting and, thus, the generalizations' overdetermination.

Models which are best at compression, i.e., keep complexity at bay by minimizing the amount of information which needs to be retained to perform well on the given evidence-task pair, should be selected. Such a refocus from simplicity creates merely another general inductive inference schema (cf. Schurz, 2010, p. 269 [2]). Rather than seeking to perform correct inferences on yet unobserved samples, under this schema, *uniformly*, the best model outperforms all other models at compressing the evidence.

Empirically, simplicity and compression treated globally as uniform principles depend on non-local facts. To produce an epistemic justification for the complexity regularization based on simplicity or compression treated globally, every local and non-local evidence-task pair would have to confirm that simplicity or compression is the reliable guide for obtaining good models able to generalize in any environment. While promising a general inductive inference schema, the presupposition which underpins

it (simplicity or compression treated globally being the reliable guide) remains formal/abstract (cf. ibid.) by depending on unavailable non-local facts populating remote or unreachable epistemic locations. Hence, the epistemic justification for complexity regularization of neural networks cannot be obtained in this way because it remains incomplete or circular, i.e., completion by an epistemically unjustifiable inductive inference. It is also helpful to notice that if simplicity does not equal compression (both treated globally as uniform principles), then the epistemic justification of complexity regularization would face an additional puzzle of meta-selecting among these two and possibly other global principles. Since the selection process would be guided by predictive success of models developed according to the available global principles, it could be implemented as multiple-favorite meta-induction (Schurz, 2008, 2019) over the generalization success of candidate models. The selection based on past predictive successes provides optimality justification for the applied complexity regularization according to the foundation-theoretic epistemology by Gerhard Schurz (2022, 2024).

Here, we focus on situations in which the requirements of statistical learning theory are satisfied (i.i.d. samples from a fixed distribution over some instance domain), which justifies object-level induction but leaves open the epistemic justification of complexity regularization allowing overparameterized models to generalize. Norton's (2003) Material Theory of Induction is used to accomplish this. We also explain the distinction between local and optimal justifications in machine learning, depending on the satisfiability of the requirements of statistical learning theory.

3 The lottery ticket hypothesis

Deep artificial neural networks possess representational capacities which often suffice for memorization of the training data (Zhang et al., 2017). Yet when performing inferences on so far unobserved samples drawn from the same distribution on the instance domain that produced the training data, they generalize well. Hence, due to implicit and/or explicit regularization, the networks avoid overfitting even though their initial complexity invites it. Despite the lack of robust complexity measures (Dziugaite et al., 2020) that would provide reliable and accurate bounds on the generalization error,[3] experimentation uncovered the likely reason for why overparameterization does not hurt generalization and is, in fact, rather beneficial (Frankle & Carbin, 2019). During training, an overparameterized network can undergo principled or unstructured prune-expand cycles, producing a version of the network that generalizes well (Gordon et al., 2018; Frankle & Carbin, 2019 respectively, also Hoefler et al., 2021). The Lotter Ticket Hypothesis posits that a large network can morph into or contains a winning ticket whose structure fits the local facts found in the evidence and the altered network reinforces the inductive biases vital for the task at hand. The cycles that prune and (re)create parts of the network establish a local inference schema as close as possible to the optimal model. The regularized schema resulting from the 'prune-expand lottery' is the basis of the networks' generalization capability.

[3] The generalisation error equals the error rate of the inductive inference that underlies Definition 1 of inductive generalisation, i.e., $1 - r$.

Prune-expand cycles, morphing the initial networks into winning tickets, reflect Norton's notion of inductive risk localization (cf. Norton, 2003, pp. 664–665). First, a human expert forms a conjecture based on their experience and selects a neural network whose architecture represents a reasonable starting point for the given evidence-task pair. This architecture has been perhaps successfully applied to similar problems and it is regarded as generic enough to cover a broad range of learning scenarios. At this point, the inductive risk depends on the generic architecture and is not yet localized. Only guarantee regarding its generalization capability stems from non-local domains and from the intuition that networks with the right level of complexity generalize. Following Norton, in such a situation the inductive risk remains separated from the local domain (evidence) and 'resides within the schema' (Norton, 2003, p. 665).

Further, since it is difficult to come up with just the right complexity at the schema-level, reduction of the risk becomes difficult as well. Vacuous bounds on the generalization error of overparameterized neural networks support Norton's insight that it is hard to assess the involved inductive risk at the schema level (ibid.). Selecting a network from some family and predicting the correct level of the network's complexity that leads to a low generalization error before performing any localization of the inductive risk is hard. One way of achieving this is via scaling laws that can predict the generalization error from parameter counts, resulting from the density, depths, and width of neural networks from some family and for a dataset (Rosenfeld et al., 2021). Scaling laws and their parameters were derived using iterative magnitude pruning (ibid.), a kind of complexity regularization (see Sect. 5), which localizes the inductive risk. A scaling law fitted to a family and dataset can be used to find a network that minimizes the parameter count given a generalization error constraint without experimentation (ibid.). This predictive capability is epistemically justified by prior experiments that identified invariance, i.e., a local fact, among different networks in terms of their density, depth, and width, sharing the same generalization error on a dataset (ibid.). If this local fact holds, then re-localization of the inductive risk is unnecessary and inferences by the scaling law on the generalization capability of the candidate networks are epistemically justified. If the local domain changes, e.g., a different architecture-dataset pair, then re-localization of the inductive risk is necessary to find a new version of the invariance, a different local fact as per Norton, justifying the inductive inferences on the generalization error. It is important to note that not all the samples from the dataset impact the generalization error equally. The localization of inductive risk can be influenced not only by pruning the network but also by pruning the dataset (Paul et al., 2021).

The possible expand phase following the regularization (pruning) of the neural network challenges the role of sparsity as a uniform principle. The goal is not to produce the sparsest network but one which is regularized to satisfy additional requirements on the generalization. Often, an optimal model is required to generalize within a certain computational budget. Apart from measuring how well the network performs the task, there might be a limit on the number of computational operations per inference (Gordon et al., 2018). In case real-time reactions are required, every inference needs to fit a narrow time window. If, by using more operations, the inference misses the window, then the underlying generalization no longer serves its purpose. Relatedly, for embedded systems, lower energy consumption might be preferred to the network's accuracy

(Banbury et al., 2021). In this case, the process of localization of inductive risk can be cut off after reaching a certain number of operations per inference. This threshold then translates into a reduced generalization capability. However, given the task at hand, the generalization capability might be still sufficient. Since a rapid growth in the number of machine learning applications is anticipated, gratuitous generalization capabilities at the expense of increased energy consumption would make the localization of inductive risk dissipative. Principled prune-expand cycles that support multi-objective risk localization can lead to favorable trade-offs between the inductive risk and the energy consumption and/or latency of the inferences.

For example, pruning can be used to identify and eliminate parts of the neural network consuming the optimized resource unproductively (Gordon et al., 2018). If we aim at the number of operations per inference, then such an ablation induces rewiring of the network that reduces the computational demands while most likely hurting its generalization capability. The network might become too sparse and localize the inductive risk only imperfectly. In turn, during expand phases the network receives a targeted boost of representational capacity by re-growing some of the ablated parts (ibid.). The expansion aims at the effective parts of the network (ibid.) to ensure that the increase in its generalization capability does not come from a growth of the resource consumption that is being minimized. By repeating the prune-expand cycle several times, the network undergoes localization of inductive risk, and the resulting inference schema supports a balanced generalization at a reasonable cost.

The MorphNet algorithm (Gordon et al., 2018) is a good example of the cycle. The prune phase can, for instance, target inference costs (floating point operations per second) by using a regularizer that removes neurons (the neural network's nodes) or even whole layers according to their computational costs (ibid.). This will decrease the network's performance. To compensate for it, a width multiplier adds neurons uniformly to all layers (ibid., e.g., expands each layer by 40%). Heavily pruned layers will, thus, grow less than the important ones which were not severely impacted by pruning. This leads to a better distribution of resources in the network because its efficient parts will receive a boost at the expense of the rest of the network.

Localization of inductive risk can be also obtained by unstructured pruning of the neural network's weights. We focus on this type of pruning, please refer to Sect. 5, pages 24–25, and to Algorithms 1 and 2 for a detailed explanation. According to the Lottery Ticket Hypothesis (Frankle & Carbin, 2019), overparameterized networks contain 'winning tickets' (sparse subnetworks) responsible for the generalization capability, which should not be possible given the network's initial complexity. A winning ticket is produced by pruning negligible parts of an overparameterized network during complexity regularization that localizes the inductive risk and creates a local schema. To recover the generalization error of the overparameterized network, the winning ticket depends on the initialization lottery (ibid.). When trained in isolation, the weights of connections in the winning ticket subnetwork cannot be reinitialized but have to be reset to the values at or close to the initialization of the overparameterized network (ibid., more on the Reset and Rewind algorithms in Sect. 5). As a result, sparsity alone is insufficient to localize inductive risk and cannot epistemically justify complexity regularization.

Section 5 shows when iterative magnitude pruning of neural network connections satisfies the Material Theory of Induction and becomes an epistemically justified complexity regularization that does not depend on unjustifiable inductive inferences about the regularizing effects of simplicity. Since according to common complexity measures Empirical Risk Minimization is a priori asymptotically unstable for overparameterized networks, the epistemic justification for regularization is vital because uniform convergence is no longer certain. The uncertainty comes from uninformative bounds on the generalization error of overparameterized networks. In this situation, as will be explained, Empirical Risk Minimization (ERM) suffers from trivialization identical to one prescribed by Norton's No-Go theorem for inductive logic (2019), although in each case the cause for trivialization is different. For this reason, we speak about a No-Go-ERM result to distinguish it from the general No-Go result by Norton (2019). The two following sections show that as the general No-Go result can be blocked by the Material Theory of Induction, the same applies to No-Go-ERM, which opens a way for the desired epistemic justification of complexity regularization. Limits of the justification are discussed as well.

4 No-Go results for empirical risk minimization

The Material Theory of Induction (MTI) blocks No-Go results that follow from possible asymptotic instabilities of Empirical Risk Minimization (ERM; Vapnik, 1995, pp. 33–45 for the consistency conditions of learning processes) applied to overparameterized neural networks under increasing the size of the training dataset (increasing in the number of observations). Due to the possible asymptotic instabilities, two-sided uniform convergence of empirical risks to risks might fail to hold. In such a case, ERM, considered as an inductive rule, can become trivialized in the same way as inductive rules facing disjunctive refinements under Norton's No-Go theorem (2019). The trivialization of ERM, following from the absence of uniform convergence, can be blocked by MTI even though the trivialization is caused by evidence strengthening instead of disjunctive refinements discussed by Norton (2019). Instead of solving the asymptotic instability by 'flattening' (equalizing) all strengths of inductive support to allow trivial convergence (cf. Norton, 2019, pp. 1131–32), the No-Go-ERM result can be blocked by an external inductive supplement (Norton, 2019, pp. 1133–34). In the present case, it is the preference for the local context consisting of an overparameterized neural network which contains a winning ticket for the data distribution at hand. For such tickets, ERM establishes the strength of inductive support only locally, without requiring the asymptotic stability for learnability under the general setting (Vapnik, 1995, p. 18), which invites the No-Go-ERM result and trivialization of inductive rules.

To this end, we proceed as follows. First, Vladimir Vapnik's and Alexey Chervonenkis's work on the necessary and sufficient conditions for two-sided uniform convergence (1971) is linked with Norton's No-Go theorem for inductive logic (2019). It is shown that when two-sided uniform convergence fails because asymptotic instabilities cannot be ruled out, ERM becomes an instance of Norton's incomplete inductive rule that is unable to escape trivialization. Second, empirical consequences of the Lottery Ticket Hypothesis (LTH, and of other principles discussed in the previous

section) are identified as the external inductive supplement delivered by MTI to block the No-Go-ERM result caused by evidence strengthening. Therefore, global uniformities like simplicity or compression cannot block the No-Go-ERM result because they require a possibly trivial asymptotic stability (by equalizing the strengths of ERM's inductive support across different training dataset sizes) to secure two-sided uniform convergence on any data distribution $\mathcal{D}$. Any such requirement, trivial or otherwise, was shown to be violated for non-trivial learning problems that are learnable without uniform convergence (Shalev-Shwartz et al., 2010).

MTI overcomes the possible trivialization of ERM by replacing the asymptotic stability with a locally derived strength of inductive support for the strictly local convergence of the empirical risk to the risk given a fixed data distribution $\mathcal{D}$ and a sample $S \sim \mathcal{D}^m$ of the size m. Therefore, the locally winning lottery tickets block the No-Go-ERM (trivialization) result for ERM. Further, it is assumed that $S \sim \mathcal{D}^m$ consists of i.i.d. (independent and identically distributed) instances $z_1, \ldots, z_m$ drawn from $\mathcal{D}$. The i.i.d. requirement can be replaced with a less restrictive notion, i.e., exchangeability defined as invariance of the underlying ground-truths under changing conditions, allowing permutations of the instance indices (cf. Arjovsky et al., 2019), where the instances are drawn from a mixture of multiple data distributions and are no longer required to be i.i.d. MTI blocks the No-Go result in both situations. The following focuses on the i.i.d. presupposition due to its prevalence in the literature concerned with uniform convergence and two-sided uniform convergence bounds. The exposition assumes that $\mathcal{D}$ is fixed and unknown in line with the classical presuppositions of Statistical Learning Theory (Vapnik, 1999, p. 988). For an optimality-based justification of inductive rules under distribution shifts, one may refer to Spelda and Stritecky (2021), utilizing the work of Gerhard Schurz (2008, 2019) on the optimality of meta-induction which delivered the well-known result concerning Hume's Problem of Induction (1739/1978).

4.1 The absence of uniform convergence and the No-Go-ERM result

We begin with Definition 2 of two-sided uniform convergence by Shalev-Shwartz et al., (2010, pp. 2639–40), providing a common notation for the original result (cf. Vapnik, 1995, Chap. 2):

$$\sup_{\mathcal{D}} \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{\mathrm{h}\in\mathcal{H}}\left|R(\mathrm{h}) - \widehat{R}_S(\mathrm{h})\right|\right] \overset{m\to\infty}{\to} 0$$

where $\mathcal{H}$ represents a hypothesis class, h a particular hypothesis, $R(\mathrm{h}) = \mathbb{E}_{\mathrm{z}\sim\mathcal{D}}[\mathcal{L}(\mathrm{h};\mathrm{z})]$ the risk of the hypothesis h (i.e., the risk of the trained network estimated on a test set), and $\widehat{R}_S(\mathrm{h}) = \frac{1}{m}\sum_{\mathrm{z}\sim S}\mathcal{L}(\mathrm{h};\mathrm{z})$ the empirical risk of the hypothesis h (i.e., the empirical risk of the network on the training set). Additionally, $\mathcal{D}$ refers to a probability distribution over the input domain $\mathcal{Z}$ comprised of instances z. Further, S is a set of samples $S = \{\mathrm{z}_1, \ldots, \mathrm{z}_m\}$ resulting from m draws from the distribution $\mathcal{D}$, that is $S \sim \mathcal{D}^m$.

For a binary classification problem, involving a fixed unknown distribution $\mathcal{D}$ on an instance domain $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and the $0-1$ loss function ${}_{\{\mathrm{h}(\mathrm{x})\neq y\}}$, where $\mathrm{h} \in \mathcal{H}$ and $\mathrm{h} : \mathcal{X} \mapsto \{0,1\}$, uniform convergence and, thus, learnability follows from $\mathcal{H}$'s finite VC dimension (Vapnik & Chervonenkis, 1971; Shalev-Shwartz et al., 2010, p. 2640). VC dimension of the hypothesis class $\mathcal{H}$ is a combinatorial measure of $\mathcal{H}$'s capacity, which captures the number of possible separations of $S \sim \mathcal{D}^m$ between $\{0,1\}$ realizable by hypotheses from the class $\mathcal{H}$ (Vapnik & Chervonenkis, 1971). VC dimension can be used in $\mathcal{D}$-independent uniform convergence bounds which depend only on the hypothesis class $\mathcal{H}$ expressible by a given ML model (ibid.; Chervonenkis, 2015). However, the values of VC dimension for overparameterized (state-of-the-art) deep artificial neural networks are large while $R(\mathrm{h})$ remains stable or decreases (cf. Valle-Pérez & Louis, 2020, pp. 13–15; Zhang et al., 2017, 2021). As a result, the core component of $\mathcal{D}$-independent uniform convergence bounds, i.e., a ratio of the value of a $\mathcal{D}$-independent complexity measure to the sample size m, becomes vacuous and disconnected from the trend of $R(\mathrm{h})$. Therefore, ERM undergoes a special case of trivialization where the two-sided uniform convergence from Definition 2 becomes bounded by a trivial (large) limit which does not guarantee the conditions for convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$. By modifying Definition 2 accordingly, we obtain Definition 3 of a two-sided uniform convergence bound (cf. Nagarajan & Kolter, 2019, p. 5):

$$\forall \mathcal{D} \Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|R(\mathrm{h}) - \widehat{R}_S(\mathrm{h})\right| \leq \epsilon_{\mathrm{unif}}(m,\delta)\right] \geq 1-\delta$$

where $\epsilon_{\mathrm{unif}}(m,\delta)$ becomes trivial (large) considering the loss $\mathcal{L}$, with δ expressing the probability of drawing an abnormal $S \sim \mathcal{D}^m$ (ibid.). As an inductive rule, ERM is trivialized by a weak strength of inductive support for two-sided uniform convergence which might fail due to possible asymptotic instabilities of ERM within the loose upper bound $\epsilon_{\mathrm{unif}}(m,\delta)$ from Definition 3. This $\mathcal{D}$-independent No-Go result for ERM can be blocked by replacing the asymptotic stability with a local stability of winning tickets drawn from overparameterized networks to fit the data at hand (as prescribed by LTH). MTI blocks the No-Go result (trivialization) by justifying a strictly local convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$ in a winning ticket, which does not require the $\mathcal{D}$-independent asymptotic stability of ERM.

Considering $\mathcal{D}$-dependent uniform convergence bounds, the situation is equally problematic. Nagarajan and Kolter (2019) showed that uniform convergence bounds based on weight norms (e.g., the distance of the network's weights from their initialization) of fixed deep networks trained with stochastic gradient descent (SGD, batch size 1 or generally small) grow with the sample size m. Schematically, considering a particular $\mathcal{D}$, the harmful growth occurs on the right side of the following inequality—Definition 4 [cf. Nagarajan & Kolter, 2019, p. 4; for the full definition of the left- and right-side terms, cf. Nagarajan's and Kolter's equation (2019)]:

$$R(\mathrm{h}) - \widehat{R}_S(\mathrm{h}) \leq \mathcal{O}\left(\frac{some\ weight\ norms}{\sqrt{m}}\right)$$

To serve as a non-trivial uniform convergence bound on generalization error $R(\mathrm{h}) - \widehat{R}_S(\mathrm{h})$, the numerator of the right-hand side ratio must show an inverse trend to the value of the denominator (Nagarajan & Kolter, 2019, pp. 3–4). However, it has been observed that the opposite is the case, i.e., norm-based uniform convergence bounds increase with the sample size m because weight norms of a trained network (the numerator of the ratio on the right-hand side of Definition 4) increase with the sample size m (Nagarajan & Kolter, 2019). As a result, the uniform convergence bound on generalization error evolves in the opposite direction to the observed generalization error—the former increases as the latter decreases with the growing sample size m (ibid.).

This divergence causes the second type of No-Go results for ERM. Since the bound grows with the sample size m, the necessary and sufficient condition for ERM's consistency, guaranteeing the uniform convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$, is violated. If we repeatedly grow m to increase the size of the training dataset $S \sim \mathcal{D}^m$, the asymptotic stability does not hold because the strengths of ERM's inductive support do not converge to a single value,[4] guaranteeing two-sided uniform convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$. To reinstate the asymptotic stability in this situation requires to set a (trivial) limit equalizing all strengths of inductive support reflected by different values of the right-hand side of Definition 4 for each $S \sim \mathcal{D}^m$. Imposing such a limit would bring the asymptotic stability back, however, at the cost of trivializing ERM as an inductive rule, thus bringing about the second type of the No-Go-ERM result. For overparameterized networks under global uniformities such as simplicity or compression, the necessary and sufficient condition for two-sided uniform convergence can be brought back only if all strengths of inductive support become flattened to an arbitrary limit. This, however, defeats the purpose of speaking about ERM's consistency in the first place. Assigning to each value of the bound an equal prior probability of bringing about uniform convergence (the indifference principle) leads to equiprobability issues that will make ERM trivial or asymptotically unstable [for a foundational exposition as to why the indifference principle cannot be used to provide a non-circular justification for inductive rules see Gerhard Schurz (2019, Sect. 4.5)].

MTI can be once again used to block the No-Go result by turning ERM into a local inductive rule (cf. Schurz, 2010, pp. 268–69), where the strengths of its inductive support derive from a local stability. In terms of LTH, the local stability results from drawing a winning ticket from an overparameterized network such that the winning ticket fits a training dataset $S \sim \mathcal{D}^m$ at hand and generalizes well. Therefore, the epistemic justification of ERM is recovered by abandoning the two-sided uniform convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$ via the ERM's asymptotic stability which invites trivialization and the No-Go results.

To bring together both No-Go results for ERM identified in the literature, we can express the two discussed two-sided uniform convergence bounds on generalization

[4] In this case, strengthening of the evidence (iteratively growing m) has the same effect on ERM as disjunctive refinements on non-trivial inductive rules—trivialization of convergence. While the effect is identical, the No-Go-ERM result is caused by conjunctively increasing the sample size while Norton's No-Go theorem (2019) is caused by disjunctive refinements of a partition of the evidence. We are grateful to a referee for the journal for helping us to fully set these situations apart.

error as a function of the sample size m. For the first failure mode, concerning the $\mathcal{D}$-independent two-sided uniform convergence bound (VC dimension-based), the bound decreases with increasing m at the $\mathcal{O}\left(\frac{1}{m}\right)$ rate (cf. Bousquet et al., 2021). For the second failure mode, concerning the $\mathcal{D}$-dependent two-sided uniform convergence bound (weights norm-based), the bound increases with increasing m at the $\Omega\left(m^{0.68}\right)$ rate (Nagarajan & Kolter, 2019, p. 4–5). In the former case ($\mathcal{D}$-independent), where in practice we observe a decreasing generalization error $R(\mathrm{h}) - \widehat{R}_S(\mathrm{h})$ with an increasing m, the two-sided uniform convergence bound does not follow the trend due to the unfavorable upper bound on the rate of the learning curve convergence. In the latter case ($\mathcal{D}$-dependent), where in practice we observe a decreasing generalization error $R(\mathrm{h}) - \widehat{R}_S(\mathrm{h})$ with an increasing m, the two-sided uniform convergence bound increases due to the non-decaying lower bound on the rate of the learning curve convergence. In both cases, the consistency condition for ERM is not satisfied because the bounds on the learning curves convergence do not guarantee the asymptotic stability of ERM. Therefore, ERM can be de-trivialized by LTH and other principles with regularizing effects that satisfy MTI. Well-performing models are then produced by local empirical risk minimization, which does not require the asymptotic but merely a local stability. Such a stability becomes free of any dependence on global uniformities such as simplicity or compression.

5 The local stability under LTH and MTI

We now take a closer look at the explanations of LTH that guarantee a local stability of ERM for overparameterized networks and de-trivialize ERM under MTI. We rely on the fact that LTH and its core component, i.e., iterative magnitude pruning (IMP) of neural networks' weights (Frankle & Carbin, 2019) playing the regularization role, satisfies the MTI's requirements for inductive risk localization as explained in Sect. 3. The aim of this section is to provide a characterization of the strengths of ERM's inductive support in the local stability regime and its ability to deliver good models fitting training datasets at hand while generalizing well. The section is divided into two parts. First, the works connected to the instability analysis of lottery tickets (Frankle et al., 2020a), a major development following LTH (Frankle & Carbin, 2019), are reviewed to provide a measurable concept of local stability. Second, examples of LTH in various empirical contexts are given, including natural language processing (Chen et al., 2020; Yu et al., 2020), computer vision (Chen et al., 2021; Morcos et al., 2019), and reinforcement learning (Yu et al., 2020), to show that the local stability leads to successful local inductive schemas in different empirical contexts. Table 1 lists the results from both parts.

The original explanation behind the LTH's success builds on the assumption that with the increasing size of a neural network increases the likelihood that the network contains a winning ticket (cf. Frankle & Carbin, 2019). That is, as outlined in Sect. 3, a subnetwork trainable to the test accuracy of the original network if the parameters (weights of the connections between nodes) of the former are reset to their values at initialization of the latter (ibid.).

Table 1 An overview of recent LTH results

Paper	Result
Frankle and Carbin 2019	LTH & Reset Algorithm (Algorithm 1)
Frankle, Dziugaite, Roy, and Carbin 2020	Instability Analysis & Rewind Algorithm (Algorithm 2)
Frankle, Schwab, and Morcos 2020	Early Phase Training Dynamics and its relation to LTH, considering Algorithms 1 and 2
Renda, Frankle, and Carbin 2020	Comparison of Rewind Algorithm, including Learning Rate Schedules Rewinding, and Network Fine-Tuning
Frankle, Dziugaite, Roy, and Carbin 2021	Critical Investigation of Pruning at Initialization
Paul, Chen, Larsen, Frankle, Ganguli and Dziugaite 2023	What is Encoded in Pruning Masks?
Morcos, Yu, Paganini, and Tian 2019	Transferability of Winning Tickets
Yu, Edunov, Tian, and Morcos 2020	Winning Tickets for Reinforcement Learning and Natural Language Processing
Chen, Frankle, Chang, Liu, Zhang, Carbin, and Wang 2021	Matching Subnetworks in Pre-Trained Computer Vision Models, using Supervised and Self-Supervised Pre-Training
Chen, Frankle, Chang, Liu, Zhang, Wang, and Carbin 2020	Matching Subnetworks in Pre-Trained Language Models

Before we characterize the process of finding subnetworks in overparameterized networks, for which we are seeking ERM's local instead of asymptotic stability, we provide basics on the pruning method (IMP). IMP falls into the unstructured pruning category (Blalock et al., 2020). Compared to structured pruning strategies, which remove entire neurons (the network's nodes), unstructured pruning targets individual parameters, that is, the neural network's weights (ibid.). Considering Algorithms 1 and 2 below, each pruning iteration removes a fraction of the smallest magnitude, non-zero weights, which results in removing some connections between individual neurons located in different layers (ibid.). This creates the so-called pruning mask. At each iteration, the pruning mask delimits a sparse subnetwork by masking some connections between neurons by removing the fraction of the smallest magnitude, non-zero weights.

The aim is to find a subnetwork that will have a similar test accuracy as the original dense network. To achieve this, Algorithm 1 (Frankle & Carbin, 2019) resets the unpruned weights to their values at initialization of the dense network and trains the subnetwork to convergence. Pruning, resetting, and training is repeated several times to reach the final level of the subnetwork sparsity. Because Algorithm 1 was found to not work in every situation (see below), Algorithm 2 (Frankle et al., 2020a) was introduced. It replaces weights reset with the 'rewind' operation, which sets the value of unpruned weights in the mask to their value at the rewind point (ibid.). The rewind point is a state of the network after k training steps.

Sparse subnetworks produced by Algorithm 1 are considered winning tickets because they can be trained to the similar test accuracy as the dense network thanks to 'lucky' initialization of the weights that identify the subnetwork according to the final pruning mask (Frankle & Carbin, 2019). Sparse subnetworks produced by Algorithm 2 are considered 'matching' instead of winning because the weights in the pruning mask are not reset but changed back ('rewound') to their values at the training step $k > 0$ (Frankle et al., 2020a).

Following Frankle and Carbin (2019), and Blalock et al. (2020) for generics on neural network pruning, the algorithm searching for winning tickets using IMP is given as follows, starting with definitions:

$f(X; W_0)$ is the original neural network, where $W_0 \sim \mathcal{D}_W$ are its initial parameters; $m = 1^{|W|}$ is an initialized pruning mask; $f(X; m \odot W_0)$ is a sparse subnetwork created by applying a pruning mask $m \in \{0,1\}^{|W_0|}$ to the initial parameters W_0; $f(X; m \odot W_k)$ is a sparse subnetwork created by applying a pruning mask $m \in \{0,1\}^{|W_k|}$ after training for k iterations until obtaining the parameters W_k. Finally, X is the training dataset and $\odot$ element-wise product operator.

Algorithm 1—Reset (Frankle & Carbin, 2019)

1: create $f(X; W_0)$
2: create $m = 1^{|W|}$
3: train $f(X; W_0)$ to convergence; or for k iterations for Algorithm 2—Rewind (below)
4: **for** $n \in \{1, \ldots, N\}$ **do**
5: prune the $p^{\frac{1}{n}}\%$ smallest magnitude parameters, i.e., if $W_0[i]$ is pruned, then $m[i] = 0$ to get a revised m, and reset the rest of the weights to W_0
6: train $f(X; m \odot W_0)$ to convergence
7: **end for**
8: return $m \in \{0,1\}^{|W_0|}$, W_0

Lines 4–6 represent IMP searching for a winning ticket. That is, a non-trivially sparse subnetwork $f(X; m \odot W_0)$ capable of recovering the test accuracy of the original dense network. Such a winning ticket results from the local stability of ERM delivered by IMP which helps to localize the inductive risk on the dataset at hand as prescribed by MTI. However, with the increasing complexity of datasets and network architectures, the IMP search for winning tickets becomes challenging—the subnetworks can recover the test accuracy of the original dense network only at trivial levels[5] of sparsity (Frankle et al., 2020a), threatening to bring back the No-Go results for ERM. This led to the introduction of the rewind operation (Algorithm 2) together with the instability analysis of lottery tickets, which identified the cause behind LTH failures in complex settings (ibid.).

[5] The measure of triviality is established by drawing from the original dense network a random subnetwork matching the IMP subnetwork's accuracy (Frankle et al., 2020a).

Algorithm 2—Rewind (Frankle et al., 2020a)
Lines 1, 2, 4 are identical to Algorithm 1; Line 3—train for k iterations to get W_k, see Algorithm 1

5: train $f(X; m \odot W_k)$ to convergence (or for T steps)
6: prune the $p^{\frac{1}{n}}\%$ smallest magnitude parameters[a], i.e., if $W_k[i]$ is pruned, then $m[i] = 0$ to get a revised m, and rewind the rest of the weights to W_k
7: **end for**
8: return $m \in \{0,1\}^{|W_k|}$, W_k

[a]Alternatively, a fixed pruning ratio can be used, e.g., during each iteration, prune a fraction of the smallest magnitude, non-zero weights (Paul et al. 2023). In general, pruning can be based on scoring parameters (weights)—the absolute value approach is common, but there are alternatives, see Blalock et al. (2020)

The rewind on Line 6, which replaced the reset from Algorithm 1, allows the algorithm to find sparse subnetworks matching the test accuracies of the original dense networks in complex settings where Algorithm 1 fails. However, the sparsity level of the subnetworks will become non-trivial only if they remain robust to SGD noise (caused by augmentations and shuffling of the dataset between training runs, random seeds, ibid.). Frankle et al.'s (2020a) instability analysis is based on training two copies of a network with parameters W_k^1, W_k^2 and two different samples of SGD noise to W_T^1, W_T^2 and determining if their training errors remain non-increasing (ibid.). The robustness to SGD noise is indicated by linear mode connectivity which occurs if there is a non-increasing path which connects the two minima resulting from training the pair of networks (ibid.). The aim is to find an iteration $k \ll T$ at which the network becomes robust to SGD noise because then, if its parameters are rewound to k and an appropriate pruning mask m is applied, the resulting IMP network $f(X; m \odot W_k)$ can match the test accuracy of the dense network at a non-trivial level of sparsity in large scale settings (ibid.).

The possibility to distinguish between matching and non-matching IMP subnetworks via instability analysis provides the necessary condition for the local stability. That is, the No-Go results for ERM discussed in the previous section become blocked if the matching subnetworks, identified using MTI-satisfying IMP, are robust to SGD noise. The robustness guarantees a local risk minimization (i.e., the inductive risk localization in MTI terms) at non-trivial levels of networks' sparsity, which removes the requirement for ERM to be asymptotically stable for overparameterized networks. Hence, ERM is de-trivialized by removing its dependence on global uniformities that can no longer be used to guarantee convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$.

We now formally link ERM's local stability (and de-trivialization that blocks the No-Go-ERM result) with the condition for the IMP's success in finding matching subnetworks.

Definition 5 of the IMP success in finding matching subnetworks. Let an ϵ-linearly connected sublevel set (LCSS) of a network $f(X; W)$ be the set of all weights W' whose test error is ϵ-close to W, $\mathcal{E}(W') \leq \mathcal{E}(W) + \epsilon$, and where W and W' are connected in the weight space by a line without error barriers, that is, are linearly mode connected (Paul et al. 2023, Def. 2.4). Further, let $f(X; m \odot W)$ be a matching sparse subnetwork if $\mathcal{E}(m \odot W) \leq \mathcal{E}(W) + \epsilon$ (Paul et al. 2023, Def. 2.2). At a round n, IMP finds a matching sparse subnetwork if the axial subspace defined by the pruning mask m_{n+1} intersects the LCSS of W_n because, then, rewinding the remaining weights to the step k and training $f(X; m_{n+1} \odot W_k)$ produces W_{n+1} linearly connected to W_n (Paul et al. 2023, Sect. 3.1).

IMP does not find a matching subnetwork at a round n if the sparsity level of the pruning mask prevents the axial subspace from intersecting the LCSS (see, for example, Paul et al., 2023, Fig. 3). The robustness of SGD to perturbations is still required, as in Frankle et al. (2020a), but instead of testing linear mode connectivity between a pair of networks at single level of sparsity, Paul et al. (2023) is testing linear mode connectivity between subnetworks at successive levels of sparsity. ERM's local stability is a feature of the geometry of the IMP error landscape which enables local convergence of $\widehat{R}_S(\mathrm{h})$ to $R(\mathrm{h})$ on i.i.d. samples in sparse matching subnetworks identified by IMP inside overparameterized networks.

There is one additional point regarding justification of the local convergence. Let $\mathcal{H}$ be the hypothesis class expressible by an overparameterized neural network f. ERM can choose several $h \in \mathcal{H}$ hypotheses with the same test error, recall that each matching subnetwork achieves an ϵ-close test error at a different level of sparsity. Since axial subspaces corresponding to pruning masks of increasing sparsity are nested (Paul et al., 2023, Fig. 1), a pruning mask m_n produced by IMP already contains a sparser mask m_{n+1} which can be used to train a matching subnetwork if the mask's axial subspace intersects the LCSS. If we consider each matching subnetwork a hypothesis $h_i \in \mathcal{H}$ with the same empirical predictions (the subnetworks' test error is ϵ-close to each other), the epistemic justification of local convergence can be improved by Schurz's (2024, p. 262) 'Strengthened Optimality Principle' (SOP). Thanks to the i.i.d. assumption, ERM can find the optimal, that is, matching subnetworks, and since the axial subspaces defined by the pruning masks are nested, it is possible to use IMP to discard all but the last pruning mask whose axial subspace still intersects the LCSS and could be, thus, used to train the sparsest matching subnetwork. This characterization seems to fit SOP, which could give us a reason to believe that ERM de-trivialized by IMP produces strongly optimal models.[6]

[6] We are grateful to a referee for the journal for guiding us to focus on SOP.

Further results connected to the local stability include an exploration of the early phase training dynamics with IMP and the rewind algorithm (Frankle et al., 2020b), finding details about networks' non-robustness to weights manipulation, that reveals the early phase as crucial for winning/matching tickets performance (ibid.). Renda et al. (2020), by extending the rewind procedure to learning rates during training, established that not only weights rewinding outperforms network fine-tuning but also that rewinding of learning rate schedules combined with IMP can match or outperform Algorithm 2. This adds a new perspective on the factors contributing to the local stability. Frankle et al. (2021) assess methods attempting to prune networks at initialization, all of which currently underperform the lottery ticket rewinding, and investigate why the methods pruning at initialization fall short of IMP applied after training and what makes their purported justification suspect.

Under MTI, the strengths of ERM's inductive support derive from instability analysis within the LTH framework. It is natural to ask about empirical characteristics of winning/matching tickets in different contexts. Morcos et al. (2019) investigated the transferability of tickets found using a particular training setup, i.e., a dataset and an optimizer, to other settings. It was established that if the network topology, the empirical domain (natural images [standard benchmark datasets]), and the task to be performed on this domain (object classification) remain fixed, then the tickets identified using one setup can achieve similar performance in different training setups, i.e., datasets and optimizers, hence achieving transferability which improves with the dataset size (ibid.). A locally stable empirical risk minimization helps to learn transferable inductive biases that, as per MTI, remain conducive to good generalization performance as long as the facts in local domains remain stable and allow transferability.

Apart from supervised learning on the domain of natural images, Yu et al. (2020) established that it is possible to find winning tickets, performing similarly as their dense antecedents, also for neural network architectures used in natural language processing and reinforcement learning, providing evidence that LTH applies beyond the original empirical setting of Frankle and Carbin (2019) or Frankle et al. (2020a). Importantly, Chen et al. (2021) discovered that large computer vision networks, pretrained via supervised or self-supervised learning, contain subnetworks transferable to downstream tasks, such as classification or segmentation, that can match the accuracy of the networks using unpruned pre-trained weights. This second kind of transferability again shows that if the facts in local domains remain stable enough, MTI can be used to explain and justify the local stability of ERM. Finally, Chen et al. (2020) made a similar observation concerning matching subnetworks and pre-trained language models, also discerning the factors that limit the downstream transferability of the former.

In sum, instead of relying on global uniformities, such as simplicity or compression, the combination of MTI and LTH suggests that ERM can become locally stable if the following conditions are met. An overparameterized network, a training/testing dataset, and an algorithm, such as Algorithms 1 or 2, able to identify winning or matching subnetworks (respectively) that localize the inductive risk (refer to comments on Algorithms 1 and 2 and to Definition 5 of matching subnetworks provided above for precise information on the identification process). Additionally, for $\widehat{R}_S(\mathrm{h})$ to converge to $R(\mathrm{h})$, it is required that $S \sim \mathcal{D}^m$ consists of i.i.d. instances $z_1, \ldots, z_m \in \mathcal{Z}$ drawn from a fixed distribution $\mathcal{D}$ as prescribed by Statistical Learning Theory (Vapnik, 1999,

p. 988). The same requirement applies to any subsequent (testing or deployment-time) samples. Otherwise, it can no longer be epistemically justified that the inductive risk is localized. This represents a strong inductive inference on data uniformity past existing observations, which can be also identified at the core of MTI as the required stability of local facts. Local facts underpin instance domains as well as the probability distributions over them. What is to be done if the inference on uniformity cannot be reasonably justified and both the asymptotic and local stability of ERM as an inductive rule will become violated? The management of inductive risk has to be taken over by multiple-favorite meta-induction over candidate models inspired by online learning with expert advice (Schurz, 2008, 2019). Here, instead of expecting minimization of inductive risk based on predictions about data uniformity, candidate models are assigned weights according to their past predictive success and the optimality of multiple-favorite meta-induction maintains the justification for model selection at each time step without presupposing anything about the data distribution. The relation of Norton's MTI to higher-order accounts of induction, which should be applied if the inference on the uniformity failed [for the sake of sustaining optimal epistemic justifications (Schurz, 2022, 2024)], was established by Schurz and Thorn (2020) and its machine learning implications developed by Spelda and Stritecky (2021).

In sum, the inductive support for ERM inferred from uniform convergence, originally depending on the asymptotic stability of ERM, can be secured by a localization. The winning/matching subnetworks localize inductive risk, and even contain transferable inductive biases based on local facts, thus validating MTI under the LTH framework.

6 Conclusion

Any bound on the generalization error of a neural network indicates the strength of inductive support for ERM. The bound needs to stable and non-trivial to guarantee the ERM's asymptotic stability and uniform convergence that epistemically justifies the resulting generalization. If the inductive support becomes insufficient because the bound does not allow an inductive inference to uniform convergence, ERM cannot avoid the No-Go results. In that case, the asymptotic stability needs to be replaced with a local stability. It was shown that ERM will not become a stable local inductive rule by relying on the networks' simplicity alone. The networks' weights at or relatively close to initialization play a significant role in supporting the ERM's local stability as well. This reveals two things. First, iterative magnitude pruning is epistemically justified because it localizes the inductive risk in line with MTI. That is, without global uniformities for which it is impossible to find a complete inductive support free of circularity. Second, according to LTH experiments, the localization and the ERM's stability depends on 'lucky' weights of winning/matching subnetworks at or relatively close to initialization. The connection between sparsity and luck, together with a growing interest in robust complexity measures for modern neural networks, suggests that the investigation of generalization in artificial neural networks evolves in a direction which can benefit from the ongoing epistemological study of inductive inference.

Acknowledgements We would like to thank two reviewers for the journal for helpful suggestions that allowed us to improve the paper.

Funding Open access publishing supported by the National Technical Library in Prague. This output was supported by the NPO 'Systemic Risk Institute' No. [LX22NPO5101], funded by European Union—Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES).

Declarations

Conflict of interest Authors declare no conflict of interest.

References

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv:1907.02893v2 [stat.ML].

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

Banbury, C. R., Reddi, V. J., Lam, M., Fu, W., Fazel, A., Holleman, J., Huang, X., Hurtado, R., Kanter, D., Lokhmotov, A., Patterson, D., Pau, D., Seo, J. S., Sieracki, J., Thakker, U., Verhelst, M., & Yadav, P. (2021). Benchmarking TinyML systems: Challenges and direction. arXiv:2003.04821 [cs.PF].

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory, 39*(3), 930–945.

Bartlett, P. L., Boucheron, S., & Lugosi, G. (2002). Model selection and error estimation. *Machine Learning, 48*, 85–113.

Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica, 30*, 203–248.

Bengio, Y., LeCun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM, 64*(7), 58–65.

Blalock, D., Ortiz, J. J. G., Frankle, J., & Guttag, J. (2020). What is the state of neural network pruning? In I. Dhillon, D. Papailiopoulos, & V. Sze (Eds.), *Proceedings of the 2nd machine learning and systems conference*.

Borghesi, A., Baldo, F., Milano, M. (2020a) Improving deep learning models via constraint-based domain knowledge: A brief survey. arXiv:2005.10691 [cs.LG].

Borghesi, A., Baldo, F., Lombardi, M., & Milano, M. (2020b). Injective domain knowledge in neural networks for transprecision computing. In *6th International conference on machine learning, optimization, and data science.*

Bousquet, O., Hanneke, S., Moran, S., van Handel, R., & Yehudayoff, A. (2021). A theory of universal learning. In *Proceedings of the 53rd annual ACM symposium on theory of computing.*

Chen, Z. S., & Haykin, S. (2002). On different facets of regularization theory. *Neural Computation, 14*, 2791–2846.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Wang, Z., & Carbin, M. (2020). The lottery ticket hypothesis for pre-trained BERT networks. In H. Larochelle, M. Ranzato, M. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Proeedings of the 34th conference on neural information processing systems.*

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Carbin, M., & Wang, Z. (2021). The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Conference on computer vision and pattern recognition.*

Cherkassky, V., & Dhar, S. (2015). Interpretation of black-box predictive models. In V. Vovk, H. Papadopoulos, & A. Gammerman (Eds.), *Measures of complexity festschrift for Alexey Chervonenkis* (pp. 267–286). Springer.

Chervonenkis, A. Y. (2015). Measures of complexity in the theory of machine learning. In V. Vovk, H. Papadopoulos, & A. Gammerman (Eds.), *Measures of complexity festschrift for Alexey Chervonenkis* (pp. 171–184). Springer.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2*, 303–314.

De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., & Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research, 6*(30), 883–904.

DeVore, R., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica, 30*, 327–444.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., & Roy, D. M. (2020). In search of robust measures of generalization. In H. Larochelle, M. Ranzato, M. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Proeedings of the 34th conference on neural information processing systems*.

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the 7th international conference on learning representations.* arXiv:1803.03635v5 [cs.LG].

Frankle, J., Dziugaite, G. K., Roy, D. M., & Carbin, M. (2020a). Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 3259–3269). PMLR.

Frankle, J., Schwab, D. J., & Morcos, A. S. (2020b). The early phase of neural network training. In *8th international conference on learning representations*.

Frankle, J., Dziugaite, G. K., Roy, D. M., & Carbin, M. (2021). Pruning neural networks at initialization: Why are we missing the mark? In *9th international conference on learning representations*.

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gordon, A., Eban, E., Nachum, O., Chen, B., Wu, H., Yang, T.-J., & Choi, E. (2018). MorphNet: Fast & simple resource-constrained structure learning of deep networks. In *Proeedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 1586–1595).

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. arXiv:2102.00554 [cs.LG].

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Hume, D. (1739/1978). A treatise on human nature. In *On human understanding*. Oxford University Press.

Kilbertus, N., Parascandolo, G., Schölkopf, B. (2018). Generalization in anti-causal learning. In *Critiquing and correcting trends in machine learning workshop (NeurIPS'18).* arXiv:1812.00524 [cs.LG].

Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks, 5*, 501–506.

Kůrková, V. (2005). Neural network learning as an inverse problem. *Logic Journal of the IGPL, 13*(5), 551–559.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

Lombardi, M., Baldo, F., Borghesi, A., & Milano, M. (2020). An analysis of regularized approaches for constrained machine learning. In *1st international workshop on trustworthy AI—integrating learning, optimization and reasoning*.

Morcos, A. S., Yu, H., Paganini, M., & Tian, Y. (2019). One ticket to win them all: Generalizing lottery ticket initializations across datasets and optimizers. In H. Wallach, H. Larochelle, F. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Proeedings of the 32nd conference on neural information processing systems*.

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. In H. Wallach, H. Larochelle, F. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Proeedings of the 32nd conference on neural information processing systems.*

Norton, J. D. (2003). A material theory of induction. *Philosophy in Science, 70*(4), 647–670.

Norton, J. D. (2014). A material dissolution of the problem of induction. *Synthese, 191*(4), 671–690.

Norton, J. D. (2019). A Demonstration of the Incompleteness of Calculi of Inductive Inference. *The British Journal for the Philosophy of Science, 70*(4), 1119–1144.

Norton, J. D. (2021). *The material theory of induction*. University of Calgary Press.
Paul, M., Ganguli, S., Dziugaite, G. K. (2021). Deep learning on a data diet: Finding important examples early in training. arXiv:2107.07075 [cs.LG].
Paul, M., Chen, F., Larsen, B. W., Frankle, J., Ganguli, S., & Dziugaite, G. K. (2023). Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? In *The eleventh international conference on learning representations*.
Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54–60.
Renda, A., Frankle, J., & Carbin, M. (2020). Comparing rewinding and fine-tuning in neural network pruning. In *8th international conference on learning representations*.
Roche, W. (2018). The perils of parsimony. *The Journal of Philosophy, 115*(9), 485–505.
Rosenfeld, J., Frankle, J., Carbin, M., & Shavit. N. (2021). On the predictability of pruning across scales. In *Proceedings of the 38th international conference on machine learning* (p. 139). PMLR.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.
Schmidt-Hieber, J. (2021). The Kolmogorov-Arnold representation theorem revisited. *Neural Networks, 137*, 119–126.
Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE, 109*(5), 612–634.
Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy in Science, 75*(3), 278–305.
Schurz, G. (2010). Local, general and universal prediction methods: A game-theoretical approach to the problem of induction. In M. Suárez, M. Dorato, & M. Rédei (Eds.), *EPSA epistemology and methodology of science launch of the European philosophy of science association* (pp. 267–278). Springer.
Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. MIT Press.
Schurz, G. (2022). Optimality justifications and the optimality principle: New tools for foundation-theoretic epistemology. *Noûs, 56*(4), 972–999. https://doi.org/10.1111/nous.12390
Schurz, G. (2024). *Optimality justifications: New foundations for epistemology*. Oxford University Press.
Schurz, G., & Thorn, P. (2020). The material theory of object-induction and the universal optimality of meta-induction: Two complementary accounts. *Studies in History and Philosophy of Science Part A, 82*, 88–93.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research, 11*, 2635–2670.
Silvestri, M., Lombardi, M., & Milano, M. (2020). Injecting domain knowledge in neural networks: A controlled experiment on a constrained problem. In *Proceedings of the 18th international conference on integration of constraint programming, artificial intelligence, and operations research.*
Sober, E. (2015). *Ockham's razors: A user's manual*. Cambridge University Press.
Spelda, P., & Stritecky, V. (2021). Human induction in machine learning: A survey of the Nexus. *ACM Computing Surveys, 54*(3), 1–18.
Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th annual allerton conference on communication, control and computing* (pp. 368–377).
Valle-Pérez, G., & Louis, A. A. (2020). Generalization bounds for deep learning. arXiv:2012.04115v2.
Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab Appl, 16*(2), 264–280.
Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.
Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.
Wahba, G. (1995). Generalization and regularization in nonlinear learning systems. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 426–432). MIT Press.
Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks, 94*, 103–114.
Yu, H., Edunov, S., Tian, Y., & Morcos, A. S. (2020). Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP. In *8th international conference on learning representations*.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals. O. (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th international conference on learning representations.* arXiv:1611.03530v2 [cs.LG].

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115.