Philosophy of Science (2022), **89**, 20–41
doi:10.1017/psa.2021.12

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

Understanding Deep Learning with Statistical Relevance

Tim Räz

University of Bern, Institute of Philosophy, Bern, Switzerland
E-mail: tim.raez@posteo.de

(Received 08 February 2020; revised 15 June 2020; accepted 29 October 2020)

Abstract

This paper argues that a notion of statistical explanation, based on Salmon's statistical relevance model, can help us better understand deep neural networks. It is proved that homogeneous partitions, the core notion of Salmon's model, are equivalent to minimal sufficient statistics, an important notion from statistical inference. This establishes a link to deep neural networks via the so-called Information Bottleneck method, an information-theoretic framework, according to which deep neural networks implicitly solve an optimization problem that generalizes minimal sufficient statistics. The resulting notion of statistical explanation is general, mathematical, and subcausal.

1. Introduction

The present paper has two goals. The first goal is to formulate a notion of statistical explanation, based on the statistical relevance (SR) model proposed by Salmon (1971a). The centerpiece of Salmon's SR model is the concept of homogeneous partition. I will prove that homogeneous partitions are mathematically equivalent to minimal sufficient statistics, an important concept from statistics. This result is interesting because it establishes a conceptual bridge between philosophical discussions of explanation and understanding on the one hand, and a well-established method of statistical inference on the other. After establishing this conceptual bridge, I will explore its consequences.

The main upshot of the first part of the paper is an account of statistical explanation that is underpinned by both philosophy and statistics. It is a noncausal, general, and mathematical kind of explanation, according to which providing explanatory information amounts to providing a statistical analogue of necessary and sufficient conditions for prediction. From an information-theoretic perspective, the account can be interpreted as information compression without loss.

The second goal of the paper is to use the notion of statistical explanation proposed in the first part to explore the so-called Information Bottleneck (IB) method (Schwartz-Ziv and Tishby, 2017). The IB method is a recent proposal from deep learning theory that aims to explain some puzzling aspects of deep learning (DL) models,

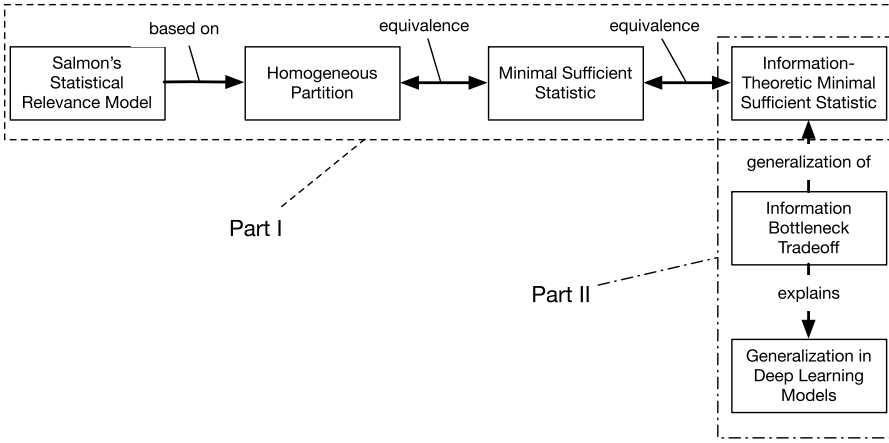

Figure 1. Concept map of Part I and Part II of the paper.

and in particular of deep neural networks (DNNs). Deep learning is a powerful technology with many successful applications, both in science and elsewhere. However, the reasons for this success are not very well understood. Why are DL models so successful, and how do they work? The IB method provides a partial answer to these questions. The IB method is a generalization of minimal sufficient statistics; this establishes a connection between the project of understanding DL models and the SR model of explanation.

The main upshot of the second part is that according to the IB method, DNNs implicitly solve an optimization problem that is a generalization of minimal sufficient statistics. The optimization is a tradeoff between the sufficiency and the necessity of minimal sufficient statistics. The kind of explanatory information we can gain on the basis of DNNs is weaker than the information we can gain according to the SR model.

The overall upshot of the present paper is that the core concept of Salmon's SR model, homogeneous partitions, capture what kind of understanding DNNs can provide on a very general level. The formal relation between statistical relevance and minimal sufficient statistics establishes that the kind of insight that DNNs provide is quite limited.

In order to facilitate orientation for the reader, the relations between the most important concepts of the paper are exhibited in figure 1. In the first part of the paper, the focus will be on the (horizontal) relations between homogeneous partitions and minimal sufficient statistics. In the second part, the focus will be on the (vertical) relations between minimal sufficient statistics and deep learning.

Part 1. Statistical relevance revisited

2. Homogeneous partitions are equivalent to minimal sufficient statistics

In this section, I provide an informal exposition of homogeneous partitions and minimal sufficient statistics and state the equivalence of these two concepts. I then discuss a third formulation of minimal sufficient statistics in information-theoretic terms. See the appendix for formal statements and proofs.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

Assume that we have a dataset A, and we want to predict a property B from A. For example, A could be a set of medical records, and B could be the presence or absence of some medical condition we want to predict. Assume that we know the conditional probability of B, given A, that is, the probability with which a patient in dataset A develops condition B. Now, two elements x and y of A, that is, different medical records, can have the same conditional probability. That is, even though x and y are different elements of A, x and y can be equivalent as far as the prediction of condition B is concerned. We can divide A into subsets such that two elements of A are in the same subset if and only if they have the same conditional probability. This means that we have grouped together all and only those elements of A that have the same "probabilistic behavior" with respect to B. Such a division of A into subsets is a *homogeneous partition* of A with respect to B.

The notion of homogeneous partition is the centerpiece of Salmon's statistical relevance model, first introduced in Salmon (1971a). Even without discussing Salmon's philosophical motivations, we can already guess why a homogeneous partition might be interesting from the perspective of statistical explanations: A homogeneous partition draws a distinction between elements of A exactly if this distinction makes a difference for the probabilistic prediction of B.

Now we turn from philosophy to statistics. A *statistic* of a random variable X is a function $T(X)$. A statistic of X induces a partition on X: members of X are in the same subset of the partition if they are sent to the same value by $T(X)$. In this sense, a statistic provides a summary of X. A *sufficient* statistic $T(X)$ of X for a variable Y is a summary of X such that, if you know the value of $T(X)$, this is all you need to know about X to predict Y–it is a "sufficient summary" of X for the purpose of predicting Y. Now, a sufficient statistic can give you too much information, in that the summary of X it provides is too fine-grained. A *minimal* sufficient statistic $T(X)$ is a summary of X for Y that gives you all the information in X for the prediction of Y, but not more. It is as coarse-grained as possible, while still being sufficient.[1]

Having read the above accounts of homogeneous partitions and minimal sufficient statistics, the reader may already have guessed that these two concepts are very similar. In fact, one can prove the following proposition:

Proposition 1. Let X, Y be random variables. A partition of X, represented by the statistic $T(X)$, is homogeneous with respect to Y if and only if $T(X)$ is a minimal sufficient statistic for Y.

What is interesting about this proposition is that Salmon's philosophically motivated notion of homogeneous partition has a counterpart in statistics, which opens up novel connections between philosophy and statistics. This formal relation between Salmon's work and statistics has gone unnoticed in the philosophical literature until

[1] The concept of sufficient statistic is due to Fisher; the theory of minimal sufficient statistic was initiated by Lehmann, Scheffé and Dynkin; see Lehmann and Casella (1998, p. 78). Traditionally, sufficient statistics serve as a principle of data reduction. Usually, X is interpreted as data, that is, $X = (X_1, \ldots, X_n)$, where the X_i are i.i.d., and Y is a vector of unknown parameters we wish to estimate. The concept of sufficiency is particularly useful if we make additional assumptions about the distribution of the data, e.g., that the X_i have a normal distribution; see Casella and Berger (2002, Ch. 6) for examples

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

now. In the present paper, I will explore the consequences of this formal relation. Note that, technically speaking, the formal characterization of minimal sufficient statistics in proposition 1 is not new; it is similar to a well-known characterization of minimal sufficient statistics due to Lehmann and Scheffé; see the appendix.

So far, we have seen two different characterizations of minimal sufficient statistics; now we turn to a third characterization of the same concept, the so-called information-theoretic formulation.[2] The information-theoretic formulation allows us to conceptualize minimal sufficient statistics in terms of statistical information: A sufficient statistic preserves all the information in variable X for the prediction of variable Y, and a minimal sufficient statistic preserves all and only the information in variable X for the prediction of variable Y. To make this idea precise, we need the notion of mutual information $I(X;Y)$ of random variables X, Y.[3] Mutual information quantifies how much we know about the behavior of one variable if we know about the behavior of the other. The notion of sufficient statistic can be formulated as follows: $S(X)$ is a sufficient statistic if and only if $I(Y;X) = I(Y;S(X))$, that is, a sufficient statistic of X for Y loses no mutual information about Y in comparison to X. Finally, a minimal sufficient statistic is a sufficient statistic that contains the least amount of mutual information about X among all sufficient statistics. We get the following, well-known[4] proposition:

Proposition 2. Let X, Y be random variables. A function $T(X)$ of X is a minimal sufficient statistic for Y if we have:

$$T(X) = \underset{S(X)}{arg\,min}\, I(S(X);X), \tag{1}$$

where $S(X)$ runs over all sufficient statistics for Y.

The information-theoretic formulation of minimal sufficient statistics is important for our purposes because, first, it opens up further interpretations of the concept of homogeneous partition, which will be useful in the philosophical discussion. Second, the information-theoretic formulation is important for the second part of the paper because it will make it easier to understand the relation between minimal sufficient statistics and the IB method.

3. Salmon's SR model revisited

In this section, I revisit Salmon's Statistical Relevance (SR) model of explanation and recapitulate the main objections against the SR model from the philosophical

[2] Information theory goes back to Shannon; the information-theoretic formulation of sufficient statistics is due to Kullback and Leibler; see Cover and Thomas (2006, pp. 54).

[3] Mutual information $I(X;Y)$ of a pair of discrete random variables (X, Y) is defined as $I(X;Y) = \sum_{x,y} P(X = x, Y = y) \log P(X = x, Y = y)/P(X = x)P(Y = y)$. Note that $I(X;Y)$ is nonnegative and symmetric.

[4] The proof of the proposition can be found in Shamir et al. (2011); see Cover and Thomas (2006) and Schwartz-Ziv and Tishby (2017) for background.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

discussion.[5] The revised account of statistical explanation formulated in the next section will be based on those aspects of the SR model that are not affected by these objections.

3.1 Salmon's SR model

Salmon's SR model is a theory of singular explanation; that is, Salmon analyzes *explananda* of the form: Why does x, which is a member of A, have attribute B? For example, the question could be why a patient with medical record $x \in A$ developed medical condition B. Salmon proposes the following definition:

Definition 3. A *Statistical Relevance (SR) Explanation* of why an instance x of A has attribute B consists of the following information:

1. A homogeneous partition $\{C_i\}_{i\in I}$ of A with respect to B,
2. the probabilities of the cells C_i of the partition with respect to B, $P(B|A \wedge C_i) =: p_i$ for $i \in I$,
3. the cell C_i to which the instance x belongs.

Note that the only substantive requirement of this definition is a homogeneous partition of A with respect to B. The probabilities in the second condition need to be known to make sure that a partition is homogeneous; there are no further constraints on these probabilities. The third requirement does not come with additional constraints either; simply knowing the cell C_i suffices. However, knowing the cell C_i has no particular significance. We might as well require that $x \in A$, which implies that x is in some C_i by definition. But the fact that x is in A, i.e., that x is part of the data, is a presupposition of the question we ask, not an additional constraint on the answers.

In what sense does the information required by the SR model constitute an explanation? Salmon writes:

"When an explanation [of this form] has been provided, we know exactly how to regard any A with respect to the property B. We know which ones to bet on, which to bet against, and at what odds. We know precisely what degree of expectation is rational. We know how to face uncertainty about an A's being a B in the most reasonable, practical, and efficient way. We know every factor that is relevant to an A having property B. We know exactly the weight that should have been attached to the prediction that this A will be a B. We know all of the regularities (universal or statistical) that are relevant to our original question. What more could one ask of an explanation?" (Salmon, 1971a, p. 78)

In these remarks on the explanatory nature of the SR model, Salmon only mentions relations between variables; instances of variables appear to play no role.

3.2 Objections against SR as singular statistical explanation

As we have seen, Salmon's SR model is an account of singular explanations, that is, why an x, which is an A, is also a B. We will now revisit objections against Salmon's model that are based on the fact that it is a *singular* statistical explanation. In

[5] See Salmon (1971a, Sec. 13) and the reconstruction and discussion in Woodward (2019, Sec. 3), on which the following account is based.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

particular, Salmon's model has consequences that disagree with intuitions about singular explanations; see Woodward (2019, Sec. 3.2. and 3.3.).

The first problem is that the SR model places no restrictions on probabilities other than that they form a homogeneous partition. In particular, if x belongs to C_i, there is no requirement that the probability $P(B|A \wedge C_i)$ is high, or higher than $P(B|A)$. This is counterintuitive if we subscribe to Hempel's idea that to explain x means to show that x was to be expected, that is, that x did occur with necessity or with high probability. Note that Salmon explicitly rejected Hempel's idea.

The second problem is that the SR model allows for the possibility that the same explanation applies to different instances x, y that are classified differently. For example, if we ask for an explanation of why patient p with medical record x is predicted to develop some medical condition B, whereas patient q with medical record y is predicted to *not* develop medical condition B, then, according to the SR model, it can be adequate to provide the same information to answer both of these questions, given that they belong to the same cell C_i in the homogeneous partition. This seems counterintuitive.

Thus, Salmon's model does not conform to some intuitions about singular explanation, and this speaks against using the SR model as an adequate account of *singular* statistical explanations. However, it does not speak against the idea that homogeneous partitions might provide a kind of *general* statistical explanation concerning probabilities, where general means that the *explananda* are not singular outcomes, but probabilistic relations between random variables. From here on, SR will be viewed as providing a kind of general statistical explanation in this sense.

3.3 Objection against SR as causal explanation

A further objection against the SR model is that it is not causal. There is a consensus in the philosophical literature that statistical relations between variables are not sufficient to infer unique causal relations between these variables; see Woodward (2019, Sec. 3.4.) and Spirtes et al. (2000). Whatever statistical relevance is, we cannot infer causal relations from it, and homogeneous partitions cannot be explanatory in virtue of identifying causal relations.

What does this mean for the explanatory significance of SR? It depends on how we see the relation between explanation and causality. Some philosophers believe that we can only explain by providing causal (or nomological) information. For example, Woodward (1987, Sec. 3) argues that SR is explanatorily irrelevant because it does not provide causal information. In the following quote, Woodward writes about the SR model: "[F]or the purpose of explanation, what matters is not just any information about (the frequency of occurrence of) the explanandum-phenomenon, but rather information that is causally or nomologically relevant." (p. 39) According to Woodward, the difference between statistical relevance and explanatory relevance is "between explaining and providing grounds for expecting or betting" (p. 39). SR does not capture causal relations and cannot provide causal explanations.

This means that if one wants to defend homogeneous partitions as providing a kind of explanation, it needs to be a noncausal kind. I will argue in the next section that we can view homogeneous partitions as providing a kind of mathematical explanation. This also means that I do not subscribe to the view, advocated by Woodward, that the only acceptable model of explanation is causal; see Lange (2016) for more on this. Note

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

that I do not wish to downplay the importance of causal explanations. Rather, I work under the assumption of explanatory pluralism, viz. that there is not just one model of explanation, but many; see, e.g., Reutlinger and Saatsi (2018, Ch. 3).

4. Explanation and understanding from statistical relevance

In this section, I articulate statistical relevance as an account of mathematical explanation. I tackle the following questions: Can we interpret homogeneous partitions as explanatory, or as providing understanding? What kind of explanatory information or understanding does it provide? And: What are the limitations of this account? Importantly, I will not defend the SR model in its original form, but only those aspects not threatened by the objections discussed in the last section. I will refer to the revised account as statistical relevance in order to emphasize the continuity with Salmon's work.

4.1 Characterizing statistical relevance

The model of statistical relevance I propose comprises homogeneous partitions. Why does a homogeneous partition provide explanatory insights about the variables involved? One way of seeing the explanatory nature of SR is by comparing it to a recent account of mathematical explanations articulated by Pincock (2015). The defining feature of Pincock's *abstract explanations* is that, in order to explain some property, we have to provide necessary and sufficient conditions for that property; see also Räz (2017, 2018). SR is analogous to abstract explanations in also requiring necessary and sufficient information with respect to probabilistic prediction. It is, however, disanalogous in not requiring *logically* necessary and sufficient conditions. Note that Salmon himself conceptualized SR in this way; in Salmon (1971a, p. 61), he describes the construction of a homogeneous partition as the "statistical analogue of the discovery of necessary and sufficient conditions". This characterization is also supported by statistics.[6] Thus, statistical relevance is explanatory in virtue of providing an analogue to necessary and sufficient conditions for probabilistic prediction.

What kind of explanation does statistical relevance provide? Building on the remarks above, statistical relevance is a variety of mathematical explanation (Mancosu, 2018), distinct from, but analogous to abstract explanations as proposed by Pincock. Furthermore, it is a kind of noncausal explanation because the information it provides is not sufficient for the identification of causal structures. It is, nevertheless, worthwhile to articulate such a subcausal notion of understanding, simply because in some cases, the kind of information captured by SR might be all that we can get.[7]

The information-theoretic formulation of minimal sufficient statistics provides us with further insights into the idea that statistical relevance provides all and only predictively relevant information; at the same time, this formulation provides a coherent

[6] A minimal sufficient statistic is equivalent to a statistic that is necessary and sufficient for prediction. Statisticians use these exact terms; see, e.g., Casella and Berger (2002, p. 308). In the context of statistics, a necessary statistic is defined as a statistic that is a function of every other statistic; thus a statistic that is sufficient and necessary is minimal sufficient.

[7] In the present paper, I restrict attention to explanatory understanding, i.e., understanding provided by explanations as opposed to, say, objectual understanding; cf. Baumberger et al. (2017).

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

information-theoretic model of explanation, a project that was started in the 1970s.[8] Information theory tells us that statistical relevance is data compression without loss: Finding a homogeneous partition amounts to finding a function $T(X)$ of the data X that retains all the information from X that is relevant for predicting Y, while discarding all information from X that is irrelevant for predicting Y. In this sense, a homogeneous partition is an optimal systematization of the data X for the purpose of predicting a variable Y.[9]

However, the information-theoretic characterization also suggests an objection to the idea that statistical relevance is explanatorily relevant. It could be argued that data compression is not what we are after when we are looking for an explanation. Compression only yields a summary of whatever information is given: If the dataset X is uninformative or not representative with respect to Y, then compressing X will not make it better. To put it bluntly, if X is trash, then $T(X)$ is a waste press. Now, I agree that the SR model will not tell us about causes, but only about correlations, which is the kind of understanding that statistical relevance can provide. Put differently, the SR model does not meet the standard of truth appropriate for causal explanations; it is not a veridical explanation in the causal sense. It is, however, veridical in always providing maximal compression of information without loss. By doing this, the SR model provides understanding in two ways. First, if the dataset X is, in some relevant sense, informative about Y, then we can expect the systematization $T(X)$ to provide us with with a more concise picture of this information, and it can play at least a heuristic role in constructing a causal explanation. Second, if X is not informative with respect to Y, constructing a summary in the form of a homogeneous partition might help us recognize a lack of informativeness in the first place, which is also valuable.

4.2 Limitations

A first limitation of statistical relevance has its origin in philosophical worries about homogeneous partitions. One problem, already mentioned above, is that a dataset X may not capture those phenomena that are relevant for the prediction of Y. This problem is related to the notion of *objective homogeneity*, the requirement that a homogeneous partition captures all factors that affect the prediction of the phenomenon

[8] It is instructive to compare the information-theoretic formulation of minimal sufficient statistics with an information-theoretic model of explanatory power proposed by Greeno (1970). Salmon (1971b, Preface) writes that his proposal and Greeno's presumably agree. However, a comparison of Greeno's notion with SR as codified in equation (1) shows that this is not so. In a nutshell, Greeno's proposal is that if S is the *explanans* variable and M the *explanandum* variable, then the mutual information $I(S;M)$ is a useful measure of the explanatory power of S with respect to M. Greeno thinks that, with certain caveats, higher values of mutual information are correlated with higher explanatory power. The main difference between the two proposals is that the idea of sufficiency is missing from Greeno's proposal. The reason for this omission might be that Greeno begins his investigation with an arbitrary partition of X, i.e., a statistic, and not with a variable X representing all the data. Consequently, he does not consider the question whether it is possible to construct a sufficient statistic that retains all relevant information from X.

[9] This characterization of SR suggests similarities with explanation as unification; see Kitcher (1989). Unification is also a best systematization account, but the optimization criteria of unification are different from those of SR. It is plausible that SR, as an account of explanation, shares some difficulties with Kitcher's model.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

described by Y; see Woodward (2019). This is a strong requirement because it is not relativized to a particular dataset, but requires that our variable X is maximally representative of the factors relevant for Y. Thus, we are typically dealing with *epistemically homogeneous* partitions, that is, we usually do not know whether there are further factors or variables that affect the prediction of Y.[10]

A second limitation of SR is statistical. Sufficient statistics are useful when the data have a fixed distribution of a certain kind; for example, $X = (X_1, \ldots, X_n)$, where the X_i are drawn from a fixed normal distribution with unknown mean and variance $Y = (\mu, \sigma^2)$. In this case, we can construct sufficient statistics of X for the unknown parameters Y of this normal distribution, the sample mean and variance. These statistics provide a useful summary of X because they are smaller than X. However, this is only the case if the X_i follow a particular kind of distribution. Casella and Berger (2002, p. 275) explain the problem as follows: "It turns out that outside of the exponential family of distributions, it is rare to have a sufficient statistic of smaller dimension than the size of the sample [...]". This means that outside of this family, a sufficient statistic will not yield an interesting amount of compression, i.e., $T(X)$ is not much smaller than X. This is the Pitman–Koopman–Darmois theorem.[11] Intuitively, it is plausible that if we do not understand an *explanandum* because of its sheer size, and the *explanans* $T(X)$ is not much smaller than the *explanandum*, then the explanation is not useful.

Unfortunately, many datasets we care about, in particular in the context of deep learning, do not have a distribution from the exponential family. Thus, statistical relevance *per se* is not of much help in understanding deep learning. However, there is hope–in the form of a generalization of minimal sufficient statistics. This brings us to the second part of the paper.

Part II. Understanding deep learning via information bottleneck

5. Deep learning

In this part of the paper, I explain how minimal sufficient statistics are related to deep learning via the Information Bottleneck (IB) method. The IB method is an information-theoretic framework that explains important features of deep learning models, in particular certain aspects of their learning and generalization behavior. In this section, I will recapitulate relevant aspects of deep learning, restricting attention to supervised learning in feedforward deep neural networks (DNNs).[12]

Abstractly speaking, a trained DNN for classification is a function $\hat{f} : X \to Y$, which takes inputs $x \in X$ and assigns one of finitely many values $y \in Y$ to that input. For

[10] Woodward considers only quantum mechanics to provide objectively homogeneous partitions: According to some interpretations of quantum mechanics, the probabilistic relations between the theory X and phenomena Y that can be deduced from it are complete such that no hidden variable T can screen off X and Y.

[11] The exponential family is a set of probability distributions that can be written in a particular parametric form. They have many nice mathematical properties, among them the fact that they allow for interesting minimal sufficient statistics; see Casella and Berger (2002); Lehmann and Casella (1998) for details.

[12] See Nielsen (2015) for an accessible introduction, LeCun et al. (2015) for an overview, Goodfellow et al. (2016) for a book-length discussion of deep learning. The account given here is based on these sources.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

example, X could be a set of medical records, and Y a set of medical conditions, in which case $\hat{f}$ provides medical diagnoses. The function $\hat{f}$ is computed by a DNN, a directed graph with several hidden layers between the input and the output; the network is called deep if there is a large number of hidden layers. If the network is fully connected, then the values are computed as follows: For each node, first compute a weighted sum of the values of all the nodes in the previous layer, plus a bias term. Then send the result through a (nonlinear) activation function. This is the value of the node, which is sent to the next layer.

A DNN will learn how to classify by adapting its parameters, the weights and biases. Assume we want an image classifier. For this, we need a training set of labeled pictures; abstractly speaking, this is a set $X = \{(x_i, y_i), i = 1 \dots n\}$, with n pairs of pictures x_i and labels y_i, indicating the correct classification of the image. The parameters of the DNN $\hat{f}$ are initiated randomly. Now, the network is given a small, random subset of the training set, called a batch. The DNN computes the values $\hat{f}(x_i) = \hat{y}_i$ for this batch. Now we use a cost function to measure the distance between the predicted values $\hat{f}(x_i) = \hat{y}_i$ (which may be inaccurate) and the label y_i (what is actually depicted). From the cost function, which is a measure of error, we can calculate how to modify the parameters such that the error becomes a little smaller, and we can propagate this error correction back through the network. This procedure is repeated until we have exhausted the training set. A training epoch can be repeated several times. The local optimization procedure is called stochastic gradient descent. It is stochastic because it depends on the random choice of batches, and it uses the gradient to make a descent in the error landscape, decreasing the error. In this way, the parameters of the DNN are gradually adapted, and the DNN "learns" to classify more accurately over time.

One of the most important metrics for evaluating a model is how well it generalizes. This is measured using a test set $X' = \{(x'_i, y'_i), i = 1 \dots m\}$, with data of the same kind as X, but disjoint from it. A model generalizes well if the classification error on the test set is small, that is, if the model is able to classify data that it has not seen before. DNNs are known to generalize very well for a variety of tasks, ranging from handwritten digits, over images of animals, to medical data.

However, there are a lot of open questions about DNNs. For one, we do not really understand why DNNs generalize so well; see, e.g., Zhang et al. (2017). The fact that DNNs perform well on test sets is just an empirical fact, and this fact is surprising, because DNNs have a lot of parameters, and prima facie, one would expect them to overfit to the training data and thus not perform particularly well on test sets. There are also many open questions about the learning process, that is, high-level properties of the local optimization procedure.[13]

6. The IB tradeoff as a generalization of MSS

In this section, I explain what the IB tradeoff is and how it generalizes minimal sufficient statistics; in the concept map of part II of the paper (figure 1), this corresponds

[13] Recently, a lot of work in computer science has engaged with the problem of formulating a theoretical framework for deep learning and the challenges raised in Zhang et al. (2017); the Information Bottleneck is just one of the proposed answers. See, e.g., Vidal et al. (2017) for an overview of open theoretical problems, and Achille and Soatto (2018) for a practical approach to the IB framework.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

to the upper arrow. As we have seen above, the concept of minimal sufficient statistic cannot be straightforwardly applied to the kinds of random variables relevant to DL models because these do not follow a nice distribution. Therefore, we cannot interpret DL models as constructing minimal sufficient statistics for these variables. This is where the IB method comes in.[14] The IB method partially answers the question why deep learning models generalize well by providing insights into the learning behavior of these models, and the explanation it provides is related to minimal sufficient statistics (MSS). Note that the generalization of MSS to the IB tradeoff can be formulated in information-theoretic terms and is, in principle, independent of the application of the IB tradeoff to DL models; the application of the IB tradeoff to DL models will be discussed in the next section.

The IB method generalizes MSS by making weaker assumptions: Instead of requiring that $T(X)$ is a function of X, the IB method only requires that there exists a conditional probability $P(T|X)$ of random variables X and T. Recall that X can be interpreted as the data, T as a representation of the data, and Y as a property we want to predict. The IB tradeoff, an optimization problem, can be formulated as follows:

$$\mathcal{L}_{IB}[P(T|X)] = \underset{P(T|X)}{argmin}\, I(X;T) - \beta I(T;Y), \tag{2}$$

where we require that Y and T are conditionally independent, given X. The idea behind this so-called IB Lagrangian is that we are looking for the distribution $P(T|X)$ that minimizes the expression on the right-hand side, which depends on $P(T|X)$ through the definition of mutual information. β is the Lagrange multiplier, a positive real number, which controls the tradeoff on the right-hand side between $I(X;T)$, the amount of information that the variable T contains about X (compression), and $I(T;Y)$, the amount of information in the variable T for predicting Y (retaining information). The conditional probability $P(T|X)$ provides a "soft partition" of X, analogous to the partition provided by an MSS $T(X)$.

The relation between the IB tradeoff and minimal sufficient statistics can be made formally precise. Compare equation (2) with the information-theoretic formulation of MSS in equation (1). The IB tradeoff combines two optimization problems, one for minimality, and one for sufficiency, both implicit in the MSS equation (1). The IB tradeoff lets us choose a relative weight for these two objectives through the value of β. If we increase β, we force T to retain more information about Y (sufficiency), and we relax the compression of X by T (minimality). If we let $\beta \to \infty$, T converges to a minimal sufficient statistic, formulated in equation (1). If we decrease β, we get coarser approximations of minimal sufficient statistics. In this sense, the IB tradeoff is a formal generalization of MSS, and the latter is a limit case of the former.

We have formulated the IB tradeoff in terms of probability distributions, independently of DL models. Eventually, we want to compare the behavior of DL models during and after training with the IB tradeoff. In order to do this, we have to calculate the tradeoff for different values of β. However, the conditional probabilities necessary to

[14] The IB method was introduced in Tishby et al. (1999); a detailed discussion of the relation between the IB method and minimal sufficient statistics can be found in Shamir et al. (2011); the relation between the IB method, minimal sufficient statistics and deep learning is discussed in Schwartz-Ziv and Tishby (2017). The account of the IB method given here is based on these papers.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

do so are generally not known, and estimating them is challenging. Describing how to estimate the IB tradeoff is beyond the scope of the present paper; see Schwartz-Ziv and Tishby (2017, Sec. 2.5) for references and Alemi et al. (2017) for an important recent method.

7. The IB method and deep learning

7.1 What the IB method tells us about deep learning

We are now in a position to understand how the IB method applies to DL models; in the concept map of part II of the paper (figure 1), this corresponds to the lower arrow. There are several aspects of DL models that can be elucidated with the IB method; see Schwartz-Ziv and Tishby (2017). Here we will focus on the (partial) explanation of the fact that DL models generalize well. This phenomenon requires an explanation because DNNs were not designed to have this property. It makes sense that DNNs minimize prediction error; this goal is given by the objective function. However, minimizing prediction error on the training set does not prevent overfitting.

To explain the generalization properties, we first have to establish a relation between the IB method and DL models. We interpret the random variables of the IB tradeoff as follows: X is the input of the DNN, the T_i are the hidden layers $i = 1 \ldots k$, and $\hat{Y}$ is the DNN's output. We can view the layers of a DNN abstractly as a Markov chain, $X \to T_1 \to \ldots \to T_k \to \hat{Y}$. This is a Markov chain because each layer only sends and receives information from its neighboring layers. The IB tradeoff can be calculated for each layer separately. Note that we have access to Y, the true label, for training and test data.

Now, to better understand why DNNs do not overfit, Schwartz-Ziv and Tishby (2017) examined how DNNs learn by tracking the quantities from the IB tradeoff during training. Specifically, they analyzed the evolution of $I(T_i, X)$, the information of a layer T_i about the input X, and $I(T_i, Y)$, the information of layers about the desired output. They found that the learning process of DNNs has two phases. In the first phase, called empirical error minimization, the network learns to classify as accurately as possible. This phase is characterized by a growth of $I(T_i; Y)$, that is, mutual information contained in the layers T_i about the label Y grows. In the second phase, called representation compression, the network compresses information in the layers T_i about the input variable X. This phase is characterized by a decrease in $I(X; T_i)$, the mutual information contained in the hidden layers about the input. Note that these relations hold for all layers T_i. At the end of training, the empirical values of each layer come close to the theoretical solutions of the IB tradeoff for appropriate values of β.

The first learning phase corresponds to error minimization; we expect this phase due to the objective function. It is the second learning phase that is surprising and that partially explains the fact that DNNs generalize well. In the second phase, the DNN "forgets" without losing information about the output–the available information in the layers is compressed. Schwartz-Ziv and Tishby (2017) write: "This compression occurs [..] without any other explicit regularization [..], and–we believe–is largely responsible for the absence of overfitting in DL" (Ibid., Introduction). The learning phase, which gets rid of irrelevant information, explains the absence of overfitting in DNNs.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

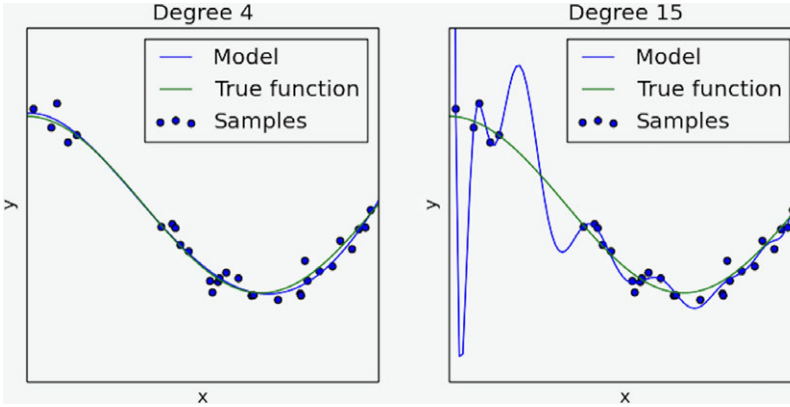

Figure 2. Image adapted from online version of Pedregosa et al. (2011).

Of course, we can ask further (explanatory) questions about the learning process and the two learning phases. In particular, it would be interesting to have a deeper, theoretical understanding of these phenomena. Such a theoretical understanding is still missing at this point – what we have so far is only a partial explanation. Schwartz-Ziv and Tishby (2017) provide more empirical findings, two of which should be mentioned here. First, the two distinct phases can be observed by plotting the evolution of the mean and the standard deviation of the weight gradients in the layers. In the first phase, the mean is larger, i.e., the DNN learns, while in the second phase, the standard deviation is larger, i.e., the DNN shows noisy behavior. Thus, the two objectives in the IB tradeoff each dominate one of the two learning phases. Second, most of the training time is spent in the second phase, i.e., the DNN spends most of the time compressing, or forgetting irrelevant information.

7.2 The IB method illustrated

In this section, I illustrate some aspects of the approximation of minimal sufficient statistics by the IB method, using an analogy with a simple regression problem; see figure 2. In a regression problem, the output variable is continuous, as opposed to discrete in a classification problem. Training a model essentially amounts to curve fitting. I use a very simple problem where both the input and the output are one-dimensional. The problem is as follows: Given a set of points (blue dots), drawn from a cosine function (green line) with a bit of noise, find a good approximation of the set of points with a polynomial.[15]

According to the IB method, deep learning models go through two learning phases. The first phase is empirical error minimization, i.e., the goal of this phase is to approximate the data as well as possible. In terms of curve fitting, the objective is

[15] It should be stressed that this is an illustration by analogy; only some properties of this example carry over to DNNs. Note the following differences between this toy problem and a genuine deep learning problem: In DL problems, the dimension of the input is much higher and more data are available; DL models have more free parameters and are not nearly as "nice" as polynomials; the notion of "simplicity" is not as clear cut in DL models; and finally, in DL we do not have access to the true data-generating function (green), we only have the dots.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

to find a curve (a model) that is as close to all the points as possible. A model that has completed the first phase looks like the blue curve in the figure on the right: the model comes reasonably close to the points, i.e., the data to which it is fit, without this fit being particularly nice where there are no points. This stage of learning corresponds to sufficiency: The model tries to capture as much information about the data as possible.

The second phase is compression, i.e., the model tries to get rid of irrelevant information. In terms of curve fitting, the objective is to make the curve (the model) smoother, or less wiggly, while retaining the information from the first phase. A model resulting from the compression phase looks like the blue curve in the figure on the left. The model still comes reasonably close to the dots, but it also does not wiggle unnecessarily in regions where there is no "reason" to wiggle. This stage of learning corresponds to minimal sufficiency: The model tries to forget as much irrelevant information (minimality) as it can while retaining the relevant information (sufficiency). So, retaining all and only relevant information has a natural counterpart in curve fitting. According to the IB method, the hard part of curve fitting is minimization, or compression, or getting rid of unnecessary wiggles in the curve.

The key difference between the left-hand side and the right-hand side of figure 2 is that the model on the left generalizes well, whereas the model on the right does not. The model on the right overfits the data–if we were to add new data points to the picture on the right by sampling from the green curve, the model's predictions would be off in regions where the model deviates from the green curve. The model on the left predicts points that it has not "seen" much better. Thus, it is the transition from the situation on the right to the situation on the left, which we can empirically observe during the second learning phase, that is responsible for good generalization behavior.

In practice, people often use regularization techniques to obtain models that do not overfit. Regularization enforces the choice of simple models. One way of implementing regularization is to add a term to the loss function which, intuitively speaking, punishes the choice of complex models. In the case of polynomials, this means that polynomials with a high degree, such as the one on the right, are punished more than polynomials with a lower degree. Now, the crux is that even if DL models are not regularized, we still observe that DL models compress in the second learning phase, which is surprising. DL models seem to perform regularization automatically–they are "implicit regularizers."

The discussion in this section can be rephrased as saying that DNNs perform well in terms of the bias–variance tradeoff. Intuitively, one would expect DNNs to be in the low-bias, high-variance regime due to the high number of parameters (Hastie et al., 2009, Sec. 2.9). The IB method provides a story as to why this is not the case. Note that in the present paper, we have only considered the IB method as a descriptive framework, i.e., it is used to understand properties of DNNs, not a prescriptive or heuristic framework to improve DNNs, as, say, in model selection. However, the IB framework has also been used in this prescriptive sense to improve DL models through explicit regularization; see, e.g., Achille and Soatto (2018).

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

7.3 Open questions

The IB method is a relatively recent proposal; as such, it faces difficulties and objections. One of the main difficulties of applying the IB method to deep learning is that the quantities in the IB tradeoff are theoretical, and we do not usually know them. Specifically, the two mutual information terms in the IB tradeoff depend on (conditional) probability distributions of the variables X, T and Y; however, we usually do not have direct access to these distributions. Thus, the true components of the IB tradeoff are unknown as well and have to be estimated; in general, this is a hard problem, as was noted above. In response to this problem, Schwartz-Ziv and Tishby (2017) write that a) they observe in simple cases that after training, the layers T_i of the network approximate their (estimated) optima according to the IB tradeoff and b) they expect this to be true for "real-life" cases as well.

The IB method is not universally accepted as an adequate answer to the questions it raises. For one, the method is not yet empirically supported by real-life examples. Additionally, the IB method's generality has been called into question. For example, Saxe et al. (2018) claim that the different learning phases and the connection between compression and generalization described in Schwartz-Ziv and Tishby (2017) cannot be reproduced for some relevant kinds of DNNs; in particular, they claim that the IB results only hold for activation functions that saturate, but not for other kinds of frequently used activation functions such as ReLUs. These claims have, in turn, been contested by Tishby et al. Thus, the jury is still out on the IB method.[16]

8. Explanation and understanding from deep learning

In this section, I put everything together and tackle the question of what kind of explanation the IB method provides from a philosophical perspective, how this translates to understanding aspects of DL models, and what this tells us about the philosophical issues of explanation and understanding.

What kind of explanatory insights does the IB method provide? To answer this question, let us compare the IB method with the SR model. Recall that the SR model provides insights about the prediction of the random variable Y from the random variable X by constructing a minimal sufficient statistic of X for Y, which is a partition of X. Sufficiency means retaining information that is relevant for prediction, whereas minimality means getting rid of information that is irrelevant for prediction. From the perspective of information theory, the construction of a minimal sufficient statistic amounts to achieving compression of X without loss of information about Y.

Analogously, the IB method provides insights about the (imperfect) prediction of a random variable Y from a random variable X by specifying conditional probabilities that constitute a tradeoff between minimality and sufficiency. In particular, the conditional probability $P(T|X)$ provides a soft partition of X. From the perspective of information theory, the IB method provides insights about the prediction of Y from X at a given level of loss of information about Y by specifying conditional probabilities that achieve, at that level, the best compression of X. The existence of conditions for

[16] It would be desirable to say more about the criticism of the IB method at a later point and also to contrast it with other proposed explanations of the generalization properties of DNNs, in order to better understand the mode of explanation at play here.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

minimality and sufficiency, albeit in the limited form of a tradeoff, is the main feature that the IB method inherits from statistical relevance.

Importantly, it is an empirical fact that the IB tradeoff is implicitly achieved by DNNs. If we examine conditional probabilities that DNNs approximate, we see that these probabilities correspond to theoretical optima given by the IB tradeoff. This suggests that the IB method provides insights about the implicit goal achieved by DNNs, namely, lossy compression. The existence of a compression phase in DNNs is a partial explanation of the generalization properties of DNNs. Compression, in turn, corresponds to minimality in minimal sufficient statistics.

The compromise between the two objectives of sufficiency and minimality is inevitable for the sorts of random variables we investigate with DNNs because actual minimal sufficient statistics cannot be found, but only approximated, for these kinds of variables. It is the tradeoff between sufficiency and minimality that connects the IB method to the debate on mathematical explanation because the tradeoff is a generalization of the well-known requirement of sufficiency and necessity for mathematical explanations; see section 4 above.

Taking a step back, what kind of explanatory insight does the IB method provide? If we look back at the philosophical discussion of the SR model in sections 2 and 3, we see that statistical relevance has traditionally been considered to be a relatively weak notion, which only provides limited insights about matters of explanation and understanding. Salmon himself suggested that statistical relevance is not sufficient for an account of (causal) explanation; see Salmon (1984). The IB method provides an even weaker kind of explanatory insight because it generalizes the SR model. In particular, just as the SR model, it does not provide a veridical causal explanation. Whatever modest understanding we may gain on the basis of minimal sufficient statistics cannot, in general, be had with the IB method. Finally, if the IB method correctly explains the implicit goal of DNNs, then the insights provided by DNNs are strictly weaker than those provided by the SR model, probabilistic necessity and sufficiency. Thus, one of the main lessons we can learn from the IB method is that we have to be modest about what DNNs can tell us if we approach them on the level of generality of the IB method.

Explanations provided by the IB method belong to the same general type as statistical relevance. Indeed, they generalize statistical relevance: they are noncausal, general, and mathematical explanations. In view of this classification, the IB method contributes to the debate on interpretability in machine learning. The focus of this debate has been to clarify what it means to understand machine learning models in general, and deep learning models in particular, and to find methods to facilitate understanding; see, e.g., Lipton (2016). In this debate, it has been pointed out that the concept of interpretability is heterogeneous and that it has to be spelled out what interpretability aims at. The IB method is an example of an answer to a clearly delineated problem of interpretability, namely, why DL models generalize well.[17]

What are the consequences of these insights? First, we should think about the ethical consequences of the IB tradeoff. The tradeoff tells us that if we want a substantive

[17] In a recent contribution to this debate, Krishnan (2020) has questioned whether the philosophical debate on scientific explanation has any bearing on interpretability. The connection between statistical relevance and the IB method shows that there is a close relation between the classical debate on scientific explanation and questions of interpretability, via minimal sufficient statistics.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

degree of compression, as provided by DNNs, the predictions of these models will necessarily discard some predictively relevant information. Maybe this is a consequence that is acceptable in some field of application, but not in others. Maybe a model that discards information should not be applied in cases where the stakes are very high.

Then, how can we make further progress toward a better understanding of what deep learning models can and cannot tell us? Progress can be made in two directions. First, we can modify DNNs such that they provide the kind of information that we require for stronger notions of explanation and understanding, be it information that is physically consistent, or nomological, or relevant to causal inference, or the like. Second, we can gain a better understanding of what DNNs in their current form can achieve. This would mean to further pursue the program started by Tishby et al. Clearly, there are still many open questions regarding the generalization and learning behavior of DNNs. The more theoretical insights into deep learning we have, the better our understanding of deep learning will be.

9. Conclusion

It is encouraging that philosophers and statisticians have come up with similar notions of statistical relevance. Salmon presumably proposed homogeneous partitions without knowing about minimal sufficient statistics.[18] This suggests that the underlying idea is robust and tracks a philosophically relevant feature of understanding how random variables are related. The equivalence also establishes a direct link between the debate on statistical explanation in philosophy and the project of understanding deep learning in computer science. Statistical relevance does provide a kind of explanatory insight that is strictly weaker than causal or nomological explanation. The link to statistics and deep learning theory provides an opportunity for philosophers to articulate a notion of statistical explanation that is descriptively adequate and philosophically deep. The present paper constitutes the first step of this project.

Acknowledgements. I thank the participants of the philosophy of science colloquium in the fall of 2019 in Bern for comments, Claus Beisbart and Michael Vock for valuable feedback on earlier versions of the paper, Samuel Portmann for discussions and help with the proof, and two anonymous reviewers for extensive and helpful comments. This work was funded by the cogito foundation.

References

Achille, Alessandro, and Stefano Soatto. 2018. "Information dropout: Learning optimal representations through noisy computation." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Alemi, Alexander A., Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. "Deep variational information bottleneck." arXiv:1612.00410v5.

Baumberger, Christoph, Claus Beisbart, and Georg Brun. 2017. "What is understanding? An overview of recent debates in epistemology and philosophy of science." In *Explaining Understanding: New Perspectives from Epistemolgy and Philosophy of Science*, edited by Stephen Grimm Christoph Baumberger and Sabine Ammon, 1–34. New York: Routledge.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Duxbury.

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory*. 2nd ed. Hoboken, NJ: Wiley.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

[18] Of course, there is no question of priority here: Fisher discovered the concept of sufficient statistics in 1922, whereas Salmon proposed his concept of homogeneous partition around 1970.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

Greeno, James G. 1970. "Evaluation of statistical hypotheses using information transmitted." *Philosophy of Science* 37 (2):279–94.
Hastie, Trevor, Roberto Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. Springer.
Kitcher, Philip. 1989. "Explanatory unification and the causal structure of the world." In *Scientific Explanation, Volume XIII of Minnesota Studies in the Philosophy of Science,* edited by Philip Kitcher and Wesley C. Salmon, 410–505. Minneapolis: University of Minnesota Press.
Kitcher, Kitcher, and Wesley C. Salmon, eds. 1989. *Scientific Explanation, Volume XIII of Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press.
Krishnan, Maya. 2016. "Against interpretability: a critical examination of the interpretability problem in machine learning." *Philosophy & Technology* 33:487–502.
Lange, Marc. 2016. *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford: Oxford University Press.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521:436–44.
Lehmann, E. L., and George Casella. 1998. *Theory of Point Estimation*. 2nd ed. *Springer Texts in Statistics*. New York, Berlin, Heidelberg: Springer.
Lipton, Zachary C. 2016. "The mythos of model interpretability." arXiv:1606.03490.
Mancosu, Paolo. 2018. "Explanation in mathematics." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Metaphysics Research Lab, Stanford University.
Nielsen, Michael A. 2015. *Neural Networks and Deep Learning*. Determination Press.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research* 12:2825–30.
Pincock, Christopher. 2015. "Abstract explanations in science." *British Journal for the Philosophy of Science* 66 (4):857–82.
Räz, Tim. 2017. "The Volterra principle generalized." *Philosophy of Science* 84 (4):737–60.
Räz, Tim. 2018. "Euler's Königsberg: the explanatory power of mathematics." *European Journal for Philosophy of Science* 8:331–46.
Reutlinger, Alexander, and Juha Saatsi, eds. 2018. *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford: Oxford University Press.
Salmon, W. C. 1971a. "Statistical Explanation." In *Statistical Explanation and Statistical Relevance,* edited by Wesley C. Salmon, 29–87. Pittsburgh: Pittsburgh University Press.
Salmon, Wesley C., ed. 1971b. *Statistical Explanation and Statistical Relevance*. Pittsburgh: Pittsburgh University Press.
Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
Saxe, Andrew M., Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. 2018. *On the information bottleneck theory of deep learning*. ICLR.
Schwartz-Ziv, Ravid, and Naftali Tishby. 2017. "Opening the black box of deep neural networks via information." arXiv:1703.00810.
Shamir, Ohad, Sivan Sabato, and Naftali Tishby. 2011. "Learning and generalization with the information bottleneck." *Theoretical Computer Science* 411:2696–2711.
Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction and Search*. Cambridge, MA: MIT Press.
Tishby, Naftali, Fernando C. Pereira, and William Bialek. 1999. "The information bottleneck method." In *Proc. of the 37th Allerton Conference on Communication, Control and Computing*, Allerton House, Monticello, Illinois, September 22-24, 1999.
Vidal, René, Joan Bruna, Raja Giryes, and Stefano Soatto. 2017. "Mathematics of deep learning." arXiv:1712.04741.
Woodward, James. 1987. "On an information-theoretic model of explanation." *Philosophy of Science* 54 (1):21–44.
Woodward, James. 2019. "Scientific explanation." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Metaphysics Research Lab, Stanford University.
Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. "Understanding deep learning requires rethinking generalization." arXiv:1611.03530.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

Proof that HP are equivalent to MSS

In the appendix, I show that the concepts of homogeneous partition and minimal sufficient statistic are equivalent. Note that only discrete random variables are considered.

A.1 Homogeneous partitions

The notion of homogeneous partition (HP) was introduced in Salmon (1971a); here I draw on the formulation in Woodward (2019). I first state the definitions using the same notation and terminology as Salmon and Woodward; a reformulation in more common statistical terms follows in the next section.

Definition 4. A *partition* $\{C_i\}_{i\in I}$ of a set A is a division of A into nonempty sets or cells C_i such that $C_i \cap C_j = \emptyset$ for $i \neq j$, i.e., different cells do not intersect, and such that $\bigcup_{i\in I} C_i = A$, i.e., the cells exhaust A.

Definition 5. *(HP)*: Let A be a population and B an attribute of members of that population. A partition $\{C_i\}_{i\in I}$ of A is *homogeneous* with respect to B if the following two conditions hold:

$$P(B|A \wedge C_i) \neq P(B|A \wedge C_j) \tag{3}$$

for all $i, j \in I$ such that $i \neq j$; and

$$P(B|A \wedge C_i \wedge D) = P(B|A \wedge C_i) \tag{4}$$

for all $i \in I$ and all D, where D is an attribute of the members of A.

The notion of homogeneous partition (HP) forms the core of Salmon's notion of statistical explanation. In Salmon's notation, both the random variable A and a subset C_i of values of A are part of the expression $P(B|A \wedge C_i)$. Below, the random variable A will be omitted because the C_i are represented by a function of A.

A.2 Homogeneous partitions and statistics

Here I explain how a function of a random variable encodes the notion of a partition–see Casella and Berger (2002, Ch. 6) for background–and I reformulate Salmon's definition in more common statistical terms.

Definition 6. Let X be a random variable. A *statistic* of X, denoted $T(X)$, is a function of X.

$T(X)$ is a function of X, which takes a value x of the random variable X and outputs the function value $T(x)$ in a deterministic manner. The range of T is not restricted, but it can be useful to take the real numbers as range.

Lemma 7. A partition $\{C_i\}_{i\in I}$ of X can be represented by a statistic $T(X)$ of X.

Proof: Given a partition $\{C_i\}_{i\in I}$ of X, choose one element x_i from each C_i. First we define $T(X)$ for each x_i such that $T(x_i) \neq T(x_j)$ for $i \neq j$, i.e., such that different elements are sent to different values. Now we define $T(X)$ as: $T(x) := T(x_i)$ for $x \in C_i$. $T(X)$ is defined on all of X because $\{C_i\}_{i\in I}$ is a partition. The statistic $T(X)$ represents the partition $\{C_i\}_{i\in I}$, in that elements of X are in the same cell C_i if and only if they are sent to the same value by $T(X)$. □

Note that the representation of a partition by a statistic is not unique. On the basis of this representation, we can express conditions (3) and (4) of (HP) using a statistic $T(X)$. A partition $\{C_i\}_{i\in I}$ of X,

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

represented by $T(X)$, is homogeneous with respect to a variable Y if two conditions hold. First, reformulating condition (3), we require that for all $i, j \in I$ such that $i \neq j$, there exists a $y \in Y$ such that:[19]

$$P(Y = y|T(X) = T(x_i)) \neq P(Y = y|T(X) = T(x_j)) \tag{5}$$

Second, reformulating condition (4), we require that for all $y \in Y$, it holds that:

$$P(Y = y|T(X) = T(x_i), U(X)) = P(Y = y|T(X) = T(x_i)) \tag{6}$$

for all partitions $U(X)$ of X and all $i \in I$.[20]

A.3 Sufficient statistics, minimal sufficient statistics

Now we define the notions of sufficient statistic and minimal sufficient statistic.[21]

Definition 8. Let X, Y be random variables and $T(X)$ a statistic of X. $T(X)$ is called *sufficient* for Y if X and Y are conditionally independent, given $T(X)$, i.e., if

$$P(Y|X, T(X)) = P(Y|T(X)). \tag{7}$$

Definition 9. A statistic $T(X)$ that is sufficient for Y is called *minimal sufficient* for Y if for any other sufficient statistic $U(X)$, there exists a function f such that:

$$T(X) = f(U(X)). \tag{8}$$

This means that $T(X)$ is a function of any other sufficient statistic $U(X)$. An equivalent formulation is that for any other statistic $U(X)$ that is sufficient for Y, we have that: if $U(x) = U(y)$, then $T(x) = T(y)$ for $x, y \in X$.

A.4 Homogeneous partitions are minimal sufficient statistics

I prove proposition 1, taking the two directions of the equivalence in turn.[22]

Proposition 10. If a statistic $T(X)$ represents a homogeneous partition $\{C_i\}_{i \in I}$ of X for Y, then $T(X)$ is a minimal sufficient statistic of X for Y.

[19] Note that the variable Y is existentially quantified: condition (5) is true if for all i, j with $i \neq j$, there is a value of Y such that the probabilities are different. This is one of two possible interpretations of Salmon's condition (3), the other being a universal quantification of y. I have chosen the existential quantification because it yields the desired equivalence in the end.

[20] Note that the partition of U may be finer than $\{C_i\}_{i \in I}$; this is the case if U takes different values in cells where T is constant–the extreme case being $U(X) = X$.

[21] See Casella and Berger (2002, Ch. 6) for an extensive treatment of sufficiency in a parametric setting, and Shamir et al. (2011) for a discussion of sufficiency in the context of the IB framework.

[22] In 1950, Lehmann and Scheff proved a characterization of minimal sufficient statistics that extends to continuous variables and that is very similar to the result proved here; see Casella and Berger (2002, Theorem 6.2.13). In fact, I believe that the result given here is a special case of the result by Lehmann and Scheffé. I will not prove this here.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

Proof: Given the homogeneous partition $\{C_i\}_{i\in I}$ of X for Y and the statistic $T(X)$ representing it, we have to show that $T(X)$ is minimal sufficient. First, we show that $T(X)$ is a sufficient statistic. $T(X)$ represents a homogeneous partition, it thus satisfies equation (6), that is, for all y, we have:

$$P(Y = y|T(X) = T(x_i), U(X)) = P(Y = y|T(X) = T(x_i))$$

for all partitions $U(X)$ of X and all $i \in I$. This means, in particular, that we can choose $U(X) = X$, i.e., the identity function, which represents the finest partition of X. We thus get:

$$P(Y|T(X) = T(x_i), X) = P(Y|T(X) = T(x_i))$$

for all $i \in I$. This condition, in turn, implies equation (7), the condition for $T(X)$ to be sufficient.

Second, we show that $T(X)$ is minimal sufficient, i.e., we show that if $U(X)$ is any other sufficient statistic, we have: if $U(x) = U(x')$, then $T(x) = T(x')$ for $x, x' \in X$. Assume, toward a contradiction, that for some sufficient statistic $U(X)$ and some $x, x' \in X$, we have $U(x) = U(x')$, but $T(x) \neq T(x')$. From $T(x) \neq T(x')$, and from the representation $T(X)$ of the partition $\{C_i\}_{i\in I}$ in Lemma 7, we can deduce that $x \in C_i$, $x' \in C_j$, for some $i, j \in I$, where $i \neq j$, which means that $T(x) = T(x_i)$ and $T(x') = T(x_j)$, where $i \neq j$. This, together with condition (5), which holds because $T(X)$ represents a homogeneous partition, implies that there is a y such that:

$$P(Y = y|T(X) = T(x)) \neq P(Y = y|T(X) = T(x')). \tag{9}$$

From the left-hand side, we can deduce:

$$\begin{aligned} P(Y = y|T(X) = T(x)) &= P(Y = y|X = x, T(X) = T(x)) \\ &= P(Y = y|X = x) \\ &= P(Y = y|X = x, U(X) = U(x)) \\ &= P(Y = y|U(X) = U(x)) \end{aligned} \tag{10}$$

where we use that $T(X)$, $U(X)$, and X are sufficient statistics. Similarly, we can deduce from the right-hand side of equation (9) that:

$$P(Y = y|T(X) = T(x')) = P(Y = y|U(X) = U(x')) \tag{11}$$

Putting these results from both sides of equation (9) together, we get:

$$P(Y = y|U(X) = U(x)) \neq P(Y = y|U(X) = U(x')). \tag{12}$$

This, however, is a contradiction with $U(x) = U(x')$. □

Proposition 11. Every minimal sufficient statistic $T(X)$ of X for Y represents a homogeneous partition $\{C_i\}_{i\in I}$ of X for Y.

Proof: We establish that a minimal sufficient statistic $T(X)$ representing a partition $\{C_i\}_{i\in I}$ satisfies conditions (5) and (6). First, condition (5). Assume, toward a contradiction, that the condition is violated, i.e., assume that for all y,

$$P(Y = y|T(X) = T(x_i)) = P(Y = y|T(X) = T(x_j)) \tag{13}$$

for some $i, j \in I$ with $i \neq j$. Define a function $U(X)$ such that $U(x) = T(x_i)$ for all $x \in C_i \cup C_j$ and such that $U(X)$ agrees with $T(X)$ everywhere else. $U(X)$ is a statistic that agrees with $T(X)$ everywhere, except in the cells C_i and C_j, where it takes one value, namely, $T(x_i)$. $U(X)$ is a sufficient statistic because it agrees with the sufficient statistic $T(X)$ outside $C_i \cup C_j$ and because within $C_i \cup C_j$, all values of X for the prediction of Y are the same according to equation (13), that is, the prediction of Y does not depend on X. At the same time, we have $U(x_i) = U(x_j)$. However, it also holds that $T(x_i) \neq T(x_j)$, in virtue of how we chose to represent partitions in Lemma 7, which means that $T(X)$ is not minimal sufficient, a contradiction.

https://doi.org/10.1017/psa.2021.12 Published online by Cambridge University Press

Second, condition (6). Starting from the left-hand side, we can deduce, for any i and any statistic $U(X)$:

$$P(Y|T(X) = T(x_i), U(X)) = P(Y|T(X) = T(x_i), U(X), X) \tag{14}$$

$$= P(Y|T(X) = T(x_i), X) \tag{15}$$

$$= P(Y|T(X) = T(x_i)), \tag{16}$$

where we used that T is sufficient and that X contains as much information as any statistic of X.

Cite this article: Räz, Tim. 2022. "Understanding Deep Learning with Statistical Relevance." *Philosophy of Science* 89 (1):20–41. https://doi.org/10.1017/psa.2021.12