Learning Optimal Representations with the Decodable Information Bottleneck

Yann Dubois
Facebook AI Research
yannd@fb.com

Douwe Kiela
Facebook AI Research
dkiela@fb.com

David J. Schwab
Facebook AI Research
CUNY Graduate Center
dschwab@fb.com

Ramakrishna Vedantam
Facebook AI Research
ramav@fb.com

Abstract

We address the question of characterizing and finding optimal representations for supervised learning. Traditionally, this question has been tackled using the Information Bottleneck, which compresses the inputs while retaining information about the targets, in a decoder-agnostic fashion. In machine learning, however, our goal is not compression but rather generalization, which is intimately linked to the predictive family or decoder of interest (e.g. linear classifier). We propose the Decodable Information Bottleneck (DIB) that considers information retention and compression from the perspective of the desired predictive family. As a result, DIB gives rise to representations that are optimal in terms of expected test performance and can be estimated with guarantees. Empirically, we show that the framework can be used to enforce a small generalization gap on downstream classifiers and to predict the generalization ability of neural networks.

1 Introduction

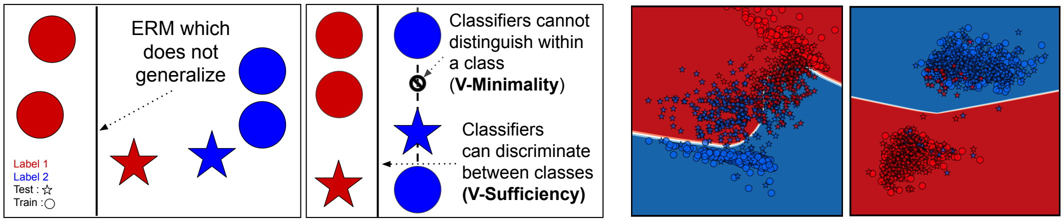
A fundamental choice in supervised machine learning (ML) centers around the data representation from which to perform predictions. While classical ML uses predefined encodings of the data [1–5] recent progress [6, 7] has been driven by learning such representations. A natural question, then, is what characterizes an "optimal" representation — in terms of generalization — and how to learn it.

The standard framework for studying generalization, statistical learning theory [8], usually assumes a fixed dataset/representation, and aims to restrict the predictive functional family $\mathcal{V}$ (e.g. linear classifiers) such that empirical risk minimizers (ERMs) generalize.[1] Here, we turn the problem on its head: we ask whether it is possible to enforce generalization by changing the representation of the inputs such that ERMs in $\mathcal{V}$ perform well, irrespective of the complexity of $\mathcal{V}$.

A common approach to representation learning consists of jointly training the classifier and representation by minimizing the empirical risk (which we call J-ERM). By only considering empirical risk, J-ERM is optimal in the infinite data limit (consistent; [10]), but the resulting representations do not favor classifiers that will generalize from finite samples. In contrast, the information bottleneck (IB) method [11] aims for representations that have *minimal* information about the inputs to avoid over-fitting, while having *sufficient* information about the labels [12]. While conceptually appealing and used in a range of applications [13–16], IB is based on Shannon's mutual information, which

[1] Rather than defining learning in terms of deterministic hypotheses $h \in \mathcal{H}$, we consider the more general [9] case of probabilistic predictors $\mathcal{V}$, in order to make a link with information theory.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) Schematic J-ERM (b) Schematic DIB (c) Empirical J-ERM (d) Empirical DIB

Figure 1: **Left two plots**: illustration of representations learned by joint empirical risk minimization (J-ERM) and our decodable information bottleneck (DIB), for classifiers $\mathcal{V}$ with linear vertical decision boundaries. (a) For representations learned by J-ERM, there may exist an ERM that does not generalize; (b) Representations learned by DIB ensure that any ERM will generalize to the test set ($\mathcal{V}$-minimality) . **Right two plots**: 2D representations encoded by an multi-layer perceptron (MLP) for odd-even classification of 200 MNIST [22] examples. The white decision boundary corresponds to a classifer which was trained to perform well on train but bad on test (see Sec. 4.2). (c) J-ERM allows such classifiers that cannot generalize; (d) DIB ensures that there are no such classifiers in $\mathcal{V}$.

was developed for communication theory [17] and does not take into account the predictive family $\mathcal{V}$ of interest. As a result, IB's sufficiency requirement does not ensure the existence of a predictor $f \in \mathcal{V}$ that can perform well using the learned representation; [2] while its minimality term is difficult to estimate, making IB impractical without resorting to approximations [18–21].

We resolve these issues by introducing the *decodable information bottleneck* (DIB) objective, which recovers minimal sufficient representations relative to a predictive family $\mathcal{V}$. Intuitively, it ensures that classifiers in $\mathcal{V}$ can predict labels ($\mathcal{V}$-sufficiency) but cannot distinguish examples with the same label ($\mathcal{V}$-minimality), as illustrated in Fig. 1. Our main contributions can be summarized as follows:

- We generalize notions of minimality and sufficiency to consider predictors $\mathcal{V}$ of interest.
- We prove that such representations are optimal — every downstream ERM in $\mathcal{V}$ reaches the best achievable test performance — and can be learned with guarantees using DIB.
- We experimentally demonstrate that using our representations can increase the performance and robustness of downstream classifiers in average and worst case scenarios.
- We show that the generalization ability of a neural network is highly correlated with the degree of $\mathcal{V}$-minimality of its hidden representations in a wide range of settings (562 models).

2 Problem Statement and Background

Throughout this paper, we provide a more informal presentation in the main body, and refer the reader to the appendices for more precise statements. For details about our notation, see Appx. A.

2.1 Problem Set-Up: Representation Learning as a Two-Player Game

Consider a game between Alice, who selects a classifier $f \in \mathcal{V}$, and Bob, who provides her with a representation to improve her performance. We are interested in Bob's optimal choice. Specifically:

1. Alice decides *a priori* on: a predictive family $\mathcal{V}$, a task of interest that consists of classifying labels Y given inputs X, and a score/loss function S measuring the quality of her predictions.
2. Given Alice's selections, Bob trains an encoder $P_{\mathrm{Z}\,|\,\mathrm{X}}$ to map inputs X to representations Z.
3. Using Bob's encoder and a dataset $\mathcal{D} \overset{\text{i.i.d.}}{\sim} P^M_{\mathrm{X,Y}}$ of M input-output pairs (x, y), Alice selects a classifier $\hat{f}$ from all ERMs $\hat{\mathcal{V}}(\mathcal{D}) := \arg\min_{f \in \mathcal{V}} \hat{\mathrm{R}}(f, \mathrm{Z}; \mathcal{D})$, where $\hat{\mathrm{R}}(f, \mathrm{Z}; \mathcal{D}) := \frac{1}{M} \sum_{y,x \in \mathcal{D}} \mathrm{E}_{z \sim P_{\mathrm{Z}\,|\,x}}[S(y, f[z])]$ is an estimate of the risk $\mathrm{R}(f, \mathrm{Z}) = \mathrm{E}_{\mathcal{D}}\left[\hat{\mathrm{R}}(f, \mathrm{Z}; \mathcal{D})\right]$.

[2] As an illustration, IB invariance to bijections suggests that a non-linearly entangled representation is as good as a linearly separable one if there is a bijection between them, even when classifying using a logistic regression.

Our goals are to: (i) characterize *optimal* representations Z^* that minimize Alice's expected loss $\mathrm{R}(\hat{f}, \mathrm{Z})$; (ii) derive an objective $\mathcal{L}$ that can be optimized to approximate the optimal encoder $P_{\mathrm{Z}^* \mid \mathrm{X}}$.

We assume that: (i) sample spaces $\mathcal{Y}, \mathcal{Z}, \mathcal{X}$ are finite; (ii) $S(y, f[z])$ is the log loss $-\log f[z](y)$, where $f[z](y)$ approximates $P_{\mathrm{Y}|\mathrm{Z}}(y|z)$ as in [23]; (iii) the family $\mathcal{V}$ satisfies mild constraints that hold for practical classifiers such as neural networks. See Appx. B for all assumptions.

2.2 Sufficiency, Minimality, and the Information Bottleneck (IB)

We review IB, an information theoretic method for supervised representation learning. IB is built upon the intuition that a representation Z should be maximally informative about Y (sufficient), but contain no additional information about X (minimal) to avoid possible over-fitting. Specifically, the set of **sufficient** representations $\mathcal{S}$ and **minimal sufficient** $\mathcal{M}$ representations are defined as:[3]

$$\mathcal{S} := \arg\max_{\mathrm{Z}'} \mathrm{I}[\mathrm{Y}; \mathrm{Z}'] \quad \text{and} \quad \mathcal{M} := \arg\min_{\mathrm{Z}' \in \mathcal{S}} \mathrm{I}[\mathrm{X}; \mathrm{Z}'] \tag{1}$$

The IB criterion (to minimize) can then be interpreted [12] as the Lagrangian relaxation of Eq. (1):

$$\mathcal{L}_{\mathrm{IB}} := -\,\mathrm{I}[\mathrm{Y}; \mathrm{Z}] + \beta * \mathrm{I}[\mathrm{X}; \mathrm{Z}] \tag{2}$$

Despite its intuitive appeal, IB suffers from the following theoretical and practical issues: (i) **Lack of optimality** guarantees for $\mathrm{Z} \in \mathcal{M}$. Generalization bounds based on $\mathrm{I}[\mathrm{Z}; \mathrm{X}]$ are a step towards such guarantees [12, 26] but current bounds are still vacuous [27]. The strong performance of invertible neural networks [28, 29] also shows that a small $\mathrm{I}[\mathrm{X}; \mathrm{Z}]$ is not required for generalization; (ii) $\mathcal{L}_{\mathrm{IB}}$ is **hard to estimate** with finite samples [12, 30]. One has to either restrict the considered setting [11, 31, 32], or optimize variational [18–20] or non-parametric [21] bounds; (iii) $\mathcal{L}_{\mathrm{IB}}$ is **invariant to bijections** and thus does not favor simple decision boundaries [33] that can be achieved by a $f \in \mathcal{V}$.

We stipulate that these known issues stem from a common cause: IB uses mutual information, which is agnostic to the predictive family $\mathcal{V}$ of interest. To remedy this, we leverage the recently proposed $\mathcal{V}$-information [23] to formalize the notion of $\mathcal{V}$-minimal $\mathcal{V}$-sufficient representations.

2.3 $\mathcal{V}$-information

From a predictive perspective, mutual information $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ corresponds to the difference in expected log loss when predicting Y with or without Z using the best possible probabilistic classifier.

$$\begin{aligned} \mathrm{I}[\mathrm{Y}; \mathrm{Z}] := \mathrm{H}[\mathrm{Y}] - \mathrm{H}[\mathrm{Y}|\mathrm{Z}] &= \mathrm{H}[\mathrm{Y}] - \mathrm{E}_{z,y \sim P_{\mathrm{Z},\mathrm{Y}}}\left[-\log P_{\mathrm{Y} \mid \mathrm{Z}}\right] && (3) \\ &= \mathrm{H}[\mathrm{Y}] - \inf_{f \in \mathcal{U}} \mathrm{E}_{z,y \sim P_{\mathrm{Z},\mathrm{Y}}}\left[-\log f[z](y)\right] \quad \text{Strict Properness [34]} && (4) \end{aligned}$$

where $\mathcal{U}$ is the collection of all predictors from $\mathcal{Z}$ to distributions over $\mathcal{Y}$, which we call *universal*. As the optimization in Eq. (4) is over $\mathcal{U}$, $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ measures information that might not be "usable" by $f \in \mathcal{V} \subset \mathcal{U}$. Xu et al.'s [23] resolve this issue by introducing $\mathcal{V}$-information $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ to only consider the information that can be decoded by a predictors of interest $f \in \mathcal{V}$ instead of $f \in \mathcal{U}$.[4]

$$\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}] := \mathrm{H}[\mathrm{Y}] - \mathrm{H}_{\mathcal{V}}[\mathrm{Y} \mid \mathrm{Z}] = \mathrm{H}[\mathrm{Y}] - \inf_{f \in \mathcal{V}} \mathrm{E}_{z,y \sim P_{\mathrm{Z},\mathrm{Y}}}\left[-\log f[z](y)\right] \tag{5}$$

$\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ has useful properties, it: recovers $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ for $\mathcal{V} = \mathcal{U}$, is non-negative, and is zero when Z is independent of Y. Importantly, $\mathcal{V}$-information is easier to estimate than Shannon's information; indeed, it corresponds to estimating the risk ($\mathrm{H}_{\mathcal{V}}[\mathrm{Y} \mid \mathrm{Z}]$) and thus inherits [23] probably approximately correct (PAC; [35]) estimation bounds that depend on the (Rademacher [36]) complexity of $\mathcal{V}$.

3 Methods

In this section, we define $\mathcal{V}$-minimal $\mathcal{V}$-sufficient representations, prove that they are optimal in the two-player representation learning game, and discuss how to approximately learn them in practice.

[3] As shown in Appx. C.1.1, this common [12, 24] definition is a generalization of minimal sufficient *statistics* [25] to stochastic statistics $\mathrm{Z} = T(\mathrm{X}, \epsilon)$ where ϵ is a source of noise independent of X.

[4] [23] also replace $\mathrm{H}[\mathrm{Y}]$ with $\mathrm{H}_{\mathcal{V}}[\mathrm{Y} \mid \varnothing]$. This requires that any $f \in \mathcal{V}$ can be conditioned on the empty set $\varnothing$. We keep $\mathrm{H}[\mathrm{Y}]$ for simplicity and show that both are equivalent in our setting (Appx. C.1.2).

3.1 $\mathcal{V}$-Sufficiency and Best Achievable Performance

Let us study Alice's best risk $\min_{f \in \mathcal{V}} \mathrm{R}(f, \mathrm{Z})$ using a representation Z. This tight lower bound on her performance looks strikingly similar to $\mathrm{H}_{\mathcal{V}}[\mathrm{Y} \mid \mathrm{Z}]$, which is controlled by $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ (see Eq. (5)). As a result, if Bob maximizes $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$, he will ensure that Alice can achieve the lowest loss.

Definition 1. A representation Z is said to be **$\mathcal{V}$-sufficient** if it maximizes $\mathcal{V}$-information with the labels. We denote all such representations as $\mathcal{S}_{\mathcal{V}} := \arg\max_{\mathrm{Z}'} \mathrm{I}_{\mathcal{V}}[\mathrm{Z}' \to \mathrm{Y}]$.

Proposition 1. Z is $\mathcal{V}$-sufficient $\iff$ there exists $f^* \in \mathcal{V}$ whose test loss when predicting from Z is the best achievable risk, i.e., $\mathrm{R}(f^*, \mathrm{Z}) = \min_{\mathrm{Z}} \min_{f \in \mathcal{V}} \mathrm{R}(f, \mathrm{Z})$.

Although the previous proposition may seem trivial, it bears important implications, namely that contrary to the sufficiency term of IB one should maximize $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ rather than $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ when predictors live in a constrained family $\mathcal{V}$. Indeed, ensuring sufficient $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ does not mean that there is a classifier $f \in \mathcal{V}$ that can "decode" that information. [5] We prove our claims in Appx. C.2.

3.2 $\mathcal{V}$-minimality and Generalization

We have seen that $\mathcal{V}$-sufficiency ensures that Alice *could* achieve the best loss. In this section, we study what representations Bob should chose to guarantee that Alice's ERMs *will* perform optimally by ensuring that any ERM generalizes beyond the training set.

IB suggests minimizing the information $\mathrm{I}[\mathrm{Z}; \mathrm{X}]$ between the representation Z and inputs X to avoid over-fitting. We, instead, argue that only the information that can be decoded by $\mathcal{V}$ matters, and would thus like to minimize $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{X}]$. However, the latter is not defined as X does not generally take value in (t.v.i.) $\mathcal{Y}$, the sample space of $\mathcal{V}$'s co-domain. For example, in a 32×32 image binary classification, $\mathcal{Y} = \{0, 1\}$ but $\mathcal{X} = [0, \ldots, 256]^{1024}$ so classifiers $f \in \mathcal{V}$ cannot predict $x \in \mathcal{X}$. To circumvent this, we decompose X into a collection of r.v.s N that t.v.i. $\mathcal{Y}$, so that $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{N}]$ is well defined. Specifically, let X_y,Z_y be "conditional r.v." s.t. $P_{\mathrm{X}_y} = P_{\mathrm{X}|y}$, $P_{\mathrm{Z}_y} = P_{\mathrm{Z}|y}$, $P_{\mathrm{X}_y,\mathrm{Z}_y} = P_{\mathrm{X},\mathrm{Z}|y}$. We define the y decomposition of X as r.v.s that arise by all possible labelings of X_y:[6]

$$\mathrm{Dec}(\mathrm{X}, y) := \{\mathrm{N} \mid \exists t' : \mathcal{X} \to \mathcal{Y} \; s.t. \; \mathrm{N} = t'(\mathrm{X}_y)\} \tag{6}$$

We can now define the average $\mathcal{V}$-information between Z and the y decompositions of X as:

$$\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})] := \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \frac{1}{|\mathrm{Dec}(\mathrm{X}, y)|} \sum_{\mathrm{N} \in \mathrm{Dec}(\mathrm{X}, y)} \mathrm{I}_{\mathcal{V}}[\mathrm{Z}_y \to \mathrm{N}] \tag{7}$$

$\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})]$ essentially measures how well predictors in $\mathcal{V}$ can predict arbitrary labeling $\mathrm{N} \in \mathrm{Dec}(\mathrm{X}, y)$ of examples with the same underlying label y. Replacing the minimality term $\mathrm{I}[\mathrm{X}; \mathrm{Z}]$ by $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})]$ we get our notion of $\mathcal{V}$-minimal $\mathcal{V}$-sufficient representations.

Definition 2. Z is **$\mathcal{V}$-minimal** $\mathcal{V}$-sufficient if it is $\mathcal{V}$-sufficient *and* has minimal average $\mathcal{V}$-information with y decompositions of X. We denote all such Z as $\mathcal{M}_{\mathcal{V}} := \arg\min_{\mathrm{Z} \in \mathcal{S}_{\mathcal{V}}} \mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})]$.

Intuitively, a representation is $\mathcal{V}$-minimal if no predictor in $\mathcal{V}$ can assign different predictions to examples with the same label. Consequently, predictors will not be able to distinguish train and test examples and must thus perfectly generalize. In this case, predictors will perform optimally as there is at least one which does ($\mathcal{V}$-sufficiency; Prop. 1). We formalize this intuition in Appx. C.3:

Theorem 1. (Informal) Let $\mathcal{V}$ be a predictive family, $\mathcal{D} \overset{\text{i.i.d.}}{\sim} P_{\mathrm{X,Y}}^{M}$ a dataset, and assume labels Y are a deterministic function of the inputs $t(\mathrm{X})$. If $\mathrm{Z} \in \mathcal{M}_{\mathcal{V}}$ be $\mathcal{V}$-minimal $\mathcal{V}$-sufficient, then the expected test loss of any ERM $\hat{f} \in \hat{\mathcal{V}}(\mathcal{D})$ is the best achievable risk, i.e., $\mathrm{R}(\hat{f}, \mathrm{Z}) = \min_{\mathrm{Z}} \min_{f \in \mathcal{V}} \mathrm{R}(f, \mathrm{Z})$.

As all ERMs reach the best risk, so does their expectation, i.e., any $\mathrm{Z} \in \mathcal{M}_{\mathcal{V}}$ is optimal. We also show in Appx. C.4 that $\mathcal{V}$-minimality and $\mathcal{V}$-sufficiency satisfy the following properties:

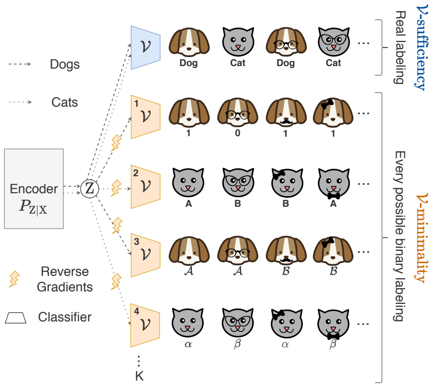
[5] Notice that $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ corresponds to the variational lower bound on $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$ used by Alemi et al. [19]. We view $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}]$ as the correct criterion rather than an estimate of $\mathrm{I}[\mathrm{Y}; \mathrm{Z}]$.

[6] Such (deterministic) labelings are also called "random labelings" [37] as they are semantically meaningless.

```
Function $\mathrm{H}(\mathcal{V}, \mathbf{x}, \mathbf{y}, P_{\mathrm{Z}|\mathrm{X}})$:
    $\mathbf{z} \leftarrow$ sample once from each $P_{\mathrm{Z}|\mathbf{x}}$
    return $\inf_{f \in \mathcal{V}} \sum_{z,y \in \mathbf{z},\mathbf{y}} \frac{-\log f[z](y)}{|\mathcal{D}|}$

Function $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{V}, \mathcal{D}, P_{\mathrm{Z}|\mathrm{X}}, K, \beta)$:
    $(\mathbf{x}, \mathbf{y}), \mathcal{L}_{\mathcal{V}\mathrm{min}}, \mathcal{Y} \leftarrow \mathcal{D}, 0, \mathrm{unique}(\mathbf{y})$
    $\mathcal{L}_{\mathcal{V}\mathrm{suff}} \leftarrow \mathrm{H}(\mathcal{V}, \mathbf{x}, \mathbf{y}, P_{\mathrm{Z}|\mathrm{X}})$
    for $y$ in $\mathcal{Y}$ do
        $\mathbf{x}_y \leftarrow \mathbf{x}[\mathbf{y} == y]$
        for $k \leftarrow 1$ to $K$ do
            $\mathbf{n} \leftarrow$ random_choice($\mathcal{Y}$, size=$|\mathbf{x}_y|$)
            $\mathcal{L}_{\mathcal{V}\mathrm{min}}$ += $\frac{\mathrm{H}(\mathcal{V}, \mathbf{x}_y, \mathbf{n}, P_{\mathrm{Z}|\mathrm{X}})}{|\mathcal{Y}| * K}$
    return (const) $+\mathcal{L}_{\mathcal{V}suff} - \beta * \mathcal{L}_{\mathcal{V}min}$
```

(a) Pseudo-code for $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{D})$



(b) DIB with neural networks

Figure 2: **Practical DIB** (a) Pseudo-code to compute the $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{D})$; (b) Illustration of DIB to train a neural encoder. $\mathcal{V}$-sufficiency corresponds to the standard log loss. $\mathcal{V}$-minimality heads are trained to classify K arbitrary labeling within each class but the gradients w.r.t. the encoder are reversed so that Z cannot be used for that task. Each head has different parameters but the same architecture $\mathcal{V}$.

Proposition 2. Let $\mathcal{V} \subseteq \mathcal{V}^+$ be two families and $\mathcal{U}$ the universal one. If labels are deterministic:

- **Recoverability** The set of $\mathcal{U}$-minimal $\mathcal{U}$-sufficient representations corresponds to the minimal sufficient representations that t.v.i. in the domain of $\mathcal{U}$, i.e., $\mathcal{M}_{\mathcal{U}} = \mathcal{M} \cap \mathcal{Z}$.
- **Monotonicity** $\mathcal{V}^+$-minimal $\mathcal{V}$-sufficient representations are $\mathcal{V}$-minimal $\mathcal{V}$-sufficient, i.e., $\arg\max_{\mathrm{Z} \in \mathcal{S}_{\mathcal{V}}} \mathrm{I}_{\mathcal{V}^+}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})] \subseteq \mathcal{M}_{\mathcal{V}}$.
- **Characterization** $\mathrm{Z} \in \mathcal{M}_{\mathcal{V}} \iff \mathrm{Z} \in \mathcal{S}_{\mathcal{V}}$ and $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})] = 0$.
- **Existence** At least one $\mathcal{U}$-minimal $\mathcal{U}$-sufficient representation always exists, i.e., $|\mathcal{M}_{\mathcal{V}}| > 0$.

The recoverability shows that our notion of $\mathcal{V}$-minimal $\mathcal{V}$-sufficiency is a generalization of minimal sufficiency. As a corollary, IB's representations are optimal when Alice is unconstrained in her choice of predictors $\mathcal{V} = \mathcal{U}$. The monotonicity implies that minimality with respect to (w.r.t.) a larger $\mathcal{V}^+$ is also optimal. Finally, the characterizations property gives a simple way of testing for $\mathcal{V}$-minimality.

3.3 Practical Optimization and Estimation

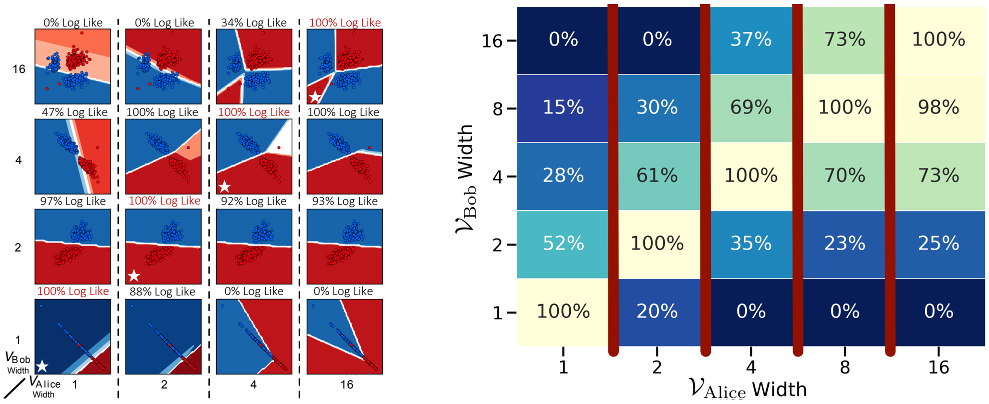
In the previous section we characterized optimal representations $\mathrm{Z}^* \in \mathcal{M}_{\mathcal{V}}$. Unfortunately, Bob cannot learn these Z^* as it requires the underlying distribution $P_{\mathrm{X},\mathrm{Y}}$. We will now show that he can nevertheless approximate Z^* in a sample- and computationally- efficient manner.

Optimization. Learning $\mathrm{Z} \in \mathcal{M}_{\mathcal{V}}$ requires solving a constrained optimization problem. Similarly to IB, we minimize the *decodable information bottleneck* (DIB), a Lagrangian relaxation of Def. 2:

$$\mathcal{L}_{\mathrm{DIB}} := -\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Y}] + \beta * \mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})] \tag{8}$$

Notice that each $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \cdot]$ has an internal optimization. In particular $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})]$ turns the problem into a min (over Z) - max (over $f \in \mathcal{V}$) optimization, which can be hard to optimize [38, 39]. We empirically compare methods for optimizing $\mathcal{L}_{\mathrm{DIB}}$ in Appx. E.2 and show that joint gradient descent ascent performs well if we ensure that the norm of the learned representation cannot diverge.

Estimation. A major benefit of $\mathcal{L}_{\mathrm{DIB}}$ over $\mathcal{L}_{\mathrm{IB}}$ is that it can be estimated with guarantees using finite samples. Namely, if Bob has access to a training set $\mathcal{D} \overset{\text{i.i.d.}}{\sim} P^M_{\mathrm{X},\mathrm{Y}}$, he can estimate $\mathcal{L}_{\mathrm{DIB}}$ reasonably well. In practice, we: (i) use $\mathcal{D}$ to estimate all expectations over $P_{\mathrm{X},\mathrm{Y}}$; (ii) use samples from Bob's encoder $z \sim P_{\mathrm{Z}\,|\,x}$; (iii) estimate the average over $\mathrm{N} \in \mathrm{Dec}(\mathrm{X}, y)$ in Eq. (7) using K samples. Figure 2a shows a (naive) algorithm to compute the resulting estimate $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{D})$. Despite these approximations, we show in Appx. C.5 that $\mathcal{L}_{\mathrm{DIB}}$ inherits $\mathcal{V}$-information's estimation bounds [23].



(a) 2D Visualization

(b) Scaling Up

Figure 3: **Optimality of $\mathcal{V}$-sufficiency.** Plots of Alice's best possible performance with different $\mathcal{V}_{Bob}$-sufficient representations. The log likelihood is column-wise scaled from 0 to 100, and vertical separators are present to discourage between-column comparison. The predictive families are MLPs with varying widths. (a) Samples of Bob's 2D representations along with Alice's decision boundaries for an odd-even CIFAR100 binary classification; (b) Same scaled log likelihood but using 8D representations, the standard CIFAR100 dataset, and averaging over 5 runs.

Proposition 3. (Informal) Let $\mathfrak{R}_{|\mathcal{D}|}$ denote the $|\mathcal{D}|$ samples Rademacher complexity. Assuming the loss is always bounded $|\log f[x](y)| < C$ then with probability at least $1-\delta$, $\hat{\mathcal{L}}_{\text{DIB}}(\mathcal{D})$ described in Fig. 2a approximates $\mathcal{L}_{\text{DIB}}$ with error less than $2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{V}) + \beta \log |\mathcal{Y}| + C\sqrt{\frac{2\log\frac{1}{\delta}}{M}}$.

The fact that the estimation error in Prop. 3 grows with the (Rademacher) complexity of $\mathcal{V}$, shows that the error is largest for $\mathcal{V} = \mathcal{U}$ corresponding to $\mathcal{L}_{\text{IB}}$. We also see a trade-off in Alice's choice of $\mathcal{V}$. A more complex $\mathcal{V}$ means the estimation of $\hat{\mathcal{L}}_{\text{DIB}}(\mathcal{D})$ is harder for Bob (Prop. 3), but Alice's prediction will improve (smaller $\min_{\text{Z}} \min_{f \in \mathcal{V}} \text{R}(f, \text{Z})$; Theorem 1).

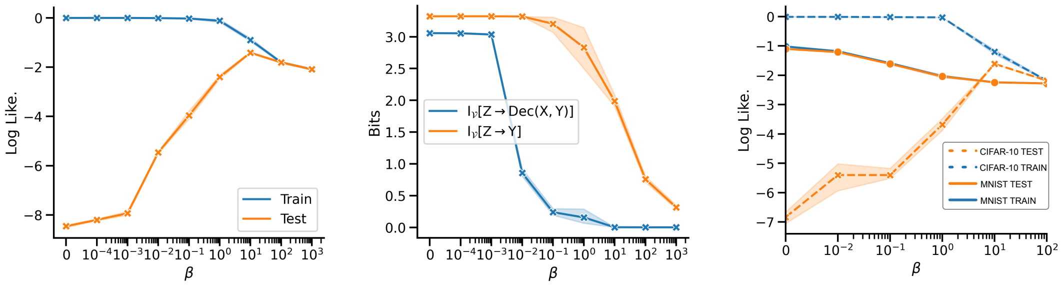
Case study: neural networks. Suppose that $\mathcal{V}$ is a specific neural architecture, the encoder $P_{\text{Z}\,|\,\text{X}}$ is parametrized by a neural network q_θ, and we are interested in cat-dog classification. As shown in Fig. 2b, training q_θ with DIB corresponds to fitting q_θ with multiple classification heads, each having exactly the same architecture $\mathcal{V}$ but different parameters. The $\mathcal{V}$-sufficiency head (in blue) tries to classify cats and dogs. Each of the K (typically 3-4, see Appx. E.4) $\mathcal{V}$-minimality heads (in orange) ensure that the representation cannot be used to classify an arbitrary (fixed) labeling of cats or dogs. In practice, the encoder and heads are trained jointly but gradients from $\mathcal{V}$-minimality heads are reversed. The $\mathcal{V}$-minimality losses are also multiplied by a hyper-parameter β.

4 Experiments

We evaluate our framework in practical settings, focusing on: (i) the relation between $\mathcal{V}$-sufficiency and Alice's best achievable performance; (ii) the relation between $\mathcal{V}$-minimality and generalization; (iii) the consequence of a mismatch between $\mathcal{V}_{Alice}$ and the functional family $\mathcal{V}_{Bob}$ w.r.t. which Z is sufficient or minimal — especially in IB's setting $\mathcal{V}_{Bob} = \mathcal{U}$; (iv) the use of our framework to predict generalization of trained networks. Many of our experiments involve sweeping over the complexity of families $\mathcal{V}^- \subseteq \mathcal{V} \subseteq \mathcal{V}^+$, we do this by varying widths of MLPs — with $\mathcal{V} \to \mathcal{U}$ in the infinite width limit [40, 41]. Alternative ways of sweeping over $\mathcal{V}$ are evaluated in Appx. E.1.

4.1 $\mathcal{V}$-sufficiency: Optimal Representations When the Data Distribution is Known

We study optimal representations when Alice has access to the data distribution $P_{\text{Z}\times\text{Y}}$. Alice's risk $\text{R}(f, \text{Z})$ in such setting is important as it is a tight lower bound on her performance in practical settings (see Sec. 3.1). We consider the following setting: Bob trains a ResNet18 encoder [42] by maximizing



(a) Generalization Gap (b) $\mathcal{V}$-Sufficiency,-Minimality (c) CIFAR10+MNIST

Figure 4: **Effect of DIB on generalization** Left two plots (CIFAR10): Impact of DIB's β on: (a) the train and test performance of Alice worst ERM; (b) the $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ and $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Y} \to \mathrm{Z}; \mathcal{D}]$ of Bob's representation Z. As β increases, Z becomes $\mathcal{V}$-minimal which increases the Alice's test performance until Z is far from $\mathcal{V}$-sufficient. Right plot (CIFAR10+MNIST): Same as (a) but images contain overlaid digits as distractors (see Appx. D.2). $\mathcal{V}$-minimality avoids over-fitting by removing spurious MNIST information. The shaded areas indicate 95% bootstrap confidence interval on 5 runs.

$\mathrm{I}_{\mathcal{V}_{\mathrm{Bob}}}[\mathrm{Z} \to \mathrm{Y}]$, Alice freezes it and trains her own classifier $f \in \mathcal{V}_{Alice}$ using the underlying $P_{\mathrm{Z} \times \mathrm{Y}}$, i.e., f is trained and evaluated on the *same* dataset. See Appx. D.1 for experimental details.

Which $\mathcal{V}$-sufficiency should Bob chose? Proposition 1 tells us that Bob's optimal choice is $\mathcal{V}_{Bob} = \mathcal{V}_{Alice}$. If he opts for a larger family $\mathcal{V}_{Alice} \subseteq \mathcal{V}_{Bob}$, the representation Z is unlikely to be decodable by Alice. If $\mathcal{V}_{Bob} \subseteq \mathcal{V}_{Alice}$, he will unnecessarily constrain Z. We first consider a setting that can be visualized: classifying the parity of CIFAR100 class index [43] using 2D representations. Figure 3a shows samples from Z and Alice's decision boundaries. To highlight the optimal $\mathcal{V}_{Bob}$ for a given $\mathcal{V}_{Alice}$, we scale the performance of each column from 0 to 100 in the figure. As predicted by Prop. 1, the best performance is achieved at $\mathcal{V}_{Alice} = \mathcal{V}_{Bob}$. The worst predictions arise when $\mathcal{V}_{Alice} \subseteq \mathcal{V}_{Bob}$, as the representations cannot be separated by Alice's classifier (e.g. $\mathcal{V}_{Bob}$ width 16 and $\mathcal{V}_{Alice}$ width 1). This suggests that IB's sufficiency (infinite width $\mathcal{V}_{Bob} = \mathcal{U}$) is undesirable when $\mathcal{V}_{Alice}$ is constrained. Figure 3b shows similar results in 8D across 5 runs. See Appx. E.8 for more settings.

4.2 $\mathcal{V}$-minimality: Optimal Representations for Generalization

Theorem 1 states that $\mathcal{V}$-minimality ensures all ERMs can generalize well. We investigate whether this is still approximately the case in practical settings, i.e., when Bob optimizes $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{D})$.

Experimental Details. Our claim concerns *all* ERMs $\hat{\mathcal{V}}^*(\mathcal{D})$, which cannot be supported by training a few $\hat{f} \in \hat{\mathcal{V}}^*(\mathcal{D})$. Instead, we evaluate the ERM that performs worst on the test set (Worst ERM), i.e., $\arg\max_{f \in \hat{\mathcal{V}}^*(\mathcal{D})} \mathrm{R}(f, \mathrm{Z})$. We do so by optimizing the following Lagrangian relaxation $\arg\min_{f \in \mathcal{V}} \hat{\mathrm{R}}(f, \mathrm{Z}; \mathcal{D}) - \gamma \mathrm{R}(f, \mathrm{Z})$ (see Appx. E.7). As our theory does not impose constraints on Z, we need an encoder close to a universal function approximator. We use a 3-MLP encoder with around 21M parameters and a 1024 dimensional Z. Since we want to investigate the generalization of ERMs resulting from Bob's criterion, we do not use (possibly implicit) regularizers such as large learning rate [44]. For more experimental details see Appx. D.1.

What is the impact of DIB's β ? We train representions on CIFAR10 with various β to investigate the effect of $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ and $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Y} \to \mathrm{Z}; \mathcal{D}]$ (Fig. 4b) on Alice's performance (Fig. 4a). Increasing β results in a decrease in $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ which monotonically shrinks the train-test gap. This suggests that, although our theory only applies for $\mathcal{V}$-minimality, $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y})]$ is tightly linked to generalization even when it is non-zero. After a certain threshold ($\beta = 10$) the generalization gains come at a large cost in $\mathrm{I}_{\mathcal{V}}[\mathrm{Y} \to \mathrm{Z}]$, which controls the best achievable loss. This shows that a trade-off (controlled by β) between $\mathcal{V}$-minimality (generalization) and $\mathcal{V}$-sufficient (lower bound) arises when Bob has to estimate $\mathcal{L}_{\mathrm{DIB}}$ using finite samples $\hat{\mathcal{L}}_{\mathrm{DIB}}(\mathcal{D})$.

$\mathcal{V}$-minimality and robustness to spurious correlations. We overlay MNIST digits as a distractor on CIFAR10 (see Appx. D.2). We run the same experiments as with CIFAR10, but we additionally train an ERM from $\mathcal{V}$ to predict MNIST labels, i.e., test whether Z contains decodable information about MNIST. Figure 4c shows that as β increases, predicting MNIST becomes harder. Indeed decreasing $\hat{\mathrm{I}}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ removes all $\mathcal{V}$-information in Z which is not useful for predicting Y. As a result, Alice's ERM must generalize better as it cannot over-fit spurious patterns.[7]

Table 1: Alice's worst and average case log loss given different representation schemes used by Bob (lower is better). Standard errors are across 5 runs.

	No Reg.	Stoch. Rep.	Dropout	Wt. Dec.	VIB	$\mathcal{V}^{-}$-DIB	$\mathcal{V}^{+}$-DIB	DIB
Worst	$10.23 \pm .13$	$8.61 \pm .05$	$1.90 \pm .00$	$10.25 \pm .03$	$1.82 \pm .02$	$1.54 \pm .03$	$1.94 \pm .33$	$\mathbf{1.41} \pm .01$
Avg.	$4.62 \pm .00$	$4.34 \pm .04$	$1.49 \pm .00$	$4.96 \pm .03$	$1.76 \pm .01$	$1.47 \pm .01$	$1.74 \pm .18$	$\mathbf{1.38} \pm .01$

Which $\mathcal{V}$-minimality should Bob chose? We study the effect of $\mathrm{Z} \in \mathcal{S}_{\mathcal{V}_{Alice}}$ being minimal w.r.t. families which are larger ($\mathcal{V}^{+}$-DIB), smaller ($\mathcal{V}^{-}$-DIB), and equal to $\mathcal{V}_{Alice}$. In theory, optimal representations would be $\mathcal{V}_{Alice}$-minimal (Theorem 1), which are achieved by DIB and $\mathcal{V}^{+}$-DIB (Monotonicity). $\mathcal{V}^{+}$-DIB should nevertheless be harder to estimate than DIB (Prop. 3). In the last 3 columns of Table 1 we indeed observe that DIB performs best. $\mathcal{V}^{+}$-DIB performs worse, suggesting that IB's minimality is undesirable in practice. We also minimize a known lower bound of $\mathrm{I}[\mathrm{Z}; \mathrm{X}]$ (VIB; [19]) and find that it performs worse than DIB.[8] We show results for different β in Appx. E.10.

Comparison to traditional regularizers. To ensure that the previous experimental gains support our theory and are not necessarily true for other regularizers, we test different regularizers on Bob and see whether they also learn representations that ensure Alice's ERM will generalize. In Table 1, we show the results of: (i) No regularization; (ii) Stochastic representations (DIB with $\beta = 0$); (iii) Dropout [45]; (iv) Weight decay. We find that DIB significantly outperforms other regularizers, which supports our claims that $\mathcal{V}$-minimality is well-suited for enforcing generalization. We emphasize that we evaluate the regularizers in a setting which is closer to our theory: two-stage game, no implicit regularizers, and evaluated on log likelihood. We show in Appx. E.11 that DIB is a descent regularizer in standard classification settings but performs a little worse than dropout.

4.3 Probing Generalization in Deep Learning

Methods that predict or correlate with the generalization of neural networks have been of recent theoretical and practical interest [46–51], as they can shed light on the inductive biases in deep learning [52–54] and prescribe better training procedures [55–59]. Having empirically shown a strong link between the degree of $\mathcal{V}$-minimality and generalization (Fig. 2b), it is natural to ask whether it can predict the generalization of a trained model. Specifically, consider the first L layers as an encoder from inputs X to representations Z_L, and subsequent layers as a classifier in $\mathcal{V}_L$. We hypothesize that $\hat{\mathrm{I}}_{\mathcal{V}_{\mathrm{L}}}[\mathrm{Z}_L \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ correlates well with the generalization of the network.

To test this, we follow Jiang et al. [50] and train convolutional networks (CNN) with varying hyperparameters (depth, width, dropout, batch size, weight decay, learning rate, dimensionality of Z) and retain those that reach 0.01 empirical risk. From this set of 562 models, we measure Kendall's rank correlation [60] between $\hat{\mathrm{I}}_{\mathcal{V}_{\mathrm{L}}}[\mathrm{Z}_L \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ and the generalization gap of each CNN, i.e., the difference between their train and test performance. For experimental details see Appx. D.3.

Table 2: Rank correlation τ between different measures and generalization gap (in terms of accuracy τ_{acc} and log loss $\tau_{logloss}$) of 562 CNNs. $\mathcal{V}_L$ denotes our $\hat{\mathrm{I}}_{\mathcal{V}_{\mathrm{L}}}[\mathrm{Z}_L \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$.

	$\mathcal{V}_L^{+}$	$\mathcal{V}_L^{-}$	$\mathcal{V}_L$	Entropy	Path Norm	Var. Grad.	Sharp. Mag.
$\tau_{acc.}$ [50]				0.148	0.373	0.311	**0.484**
τ_{acc} (ours)	**0.482**	0.391	0.471	0.234	0.347	0.332	0.385
$\tau_{logloss}$ (ours)	**0.505**	0.435	0.498	0.164	0.357	0.167	0.233

Does $\mathcal{V}$-minimality correlate with generalization? In the last five columns of Table 2, we compare our results ($\mathcal{V}_L$) to the best generalization measure from each categories investigated in [50]:

[7] Achille and Soatto [20] show that this happens for minimal Z. The novelty is that we obtain similar results when considering $\mathcal{V}$-minimality, which is less stringent (Prop. 2) and does not require the intractable $\mathrm{I}[\mathrm{X}; \mathrm{Z}]$.

[8] VIB is hard to compare to DIB as it is unclear w.r.t. which family, if any, VIB's solutions are minimal.

the entropy of the output [61], the path norm [62], the variance of the gradients after training (Var. Grad. ; [50]), and the "sharpness" of the minima (Sharp. Mag.; [57]).[9] As hypothesized, $\hat{\mathrm{I}}_{\mathcal{V}_{\mathrm{L}}}[\mathrm{Z}_L \to \mathrm{Dec}(\mathrm{X}, \mathcal{Y}); \mathcal{D}]$ correlates with generalization and even outperforms the baselines. Similarly to Table 1 we also evaluate minimality with respect to a family larger ($\mathcal{V}_L^+$) or smaller ($\mathcal{V}_L^-$) than $\mathcal{V}_L$. Surprisingly, $\mathcal{V}_L^+$ performs better than $\mathcal{V}_L$, which might be because larger networks can help optimization of sub-networks $\mathcal{V}_L \subseteq \mathcal{V}_L^+$ as suggested by the Lottery Ticket Hypothesis [63].

To the best of our knowledge $\mathcal{V}$-minimality is the first measure of generalization of a network that only considers a single internal representation Z_L. This could be of particular interest in transfer learning, as it can predict how well any model of a certain architecture will generalize when using a specific pretrained encoder. As $\mathcal{V}$-minimality is a property of a representation rather than the architecture, we show in Appx. E.12 that it can be meaningfully compared *across* different architectures and datasets.

5 Other Related Work

Generalized information, game theory and Bayes decision theory. If you need a distribution P_{X}^* to act as a representative $\Gamma \subseteq \mathcal{P}(\mathcal{X})$ you should follow the maximum entropy (MaxEnt) principle [64, 65] to minimize the worst-case log loss [66, 67]. Grünwald et al. [68] generalized MaxEnt to different losses by framing the problem as an adversarial game between nature and a decision maker. Robust supervised learning [69] can also be framed in a way that suggests to maximize *conditional* entropy [70, 71]. This line of work focuses on prediction rules (Alice). Our framing (Sec. 2.1) extends this literature by incorporating a co-operative agent (Bob), which learns representations to minimize the worst-case loss of the decision maker (Alice). Although [10, 72] also studied representations using generalized information, they focused on consistency rather than generalization.

Extended sufficiency and minimality. Linear sufficiency is well studied [73–75] but only considers linear encoders and predictors and is used for estimation rather than predictions. In ML, Cvitkovic and Koliander [76] incorporated the encoder's family (Bob) to characterize achievable Z. This is complementary to our incorporation of the decoder's family $\mathcal{V}$ (Alice) to characterize optimal Z.

Kernel Learning. There is a large literature in learning kernels [77–80] for support vector machines [2], which implicitly learns a data representation [81]. The learning is either done by minimizing estimates [82, 83] or bounds of the generalization error [84–88]. The major advantage of our work is that we are not restricted to predictors $\mathcal{V}$ that can be "kernelized" and provide an optimality proof.

6 Conclusion and Future Work

In this work, we propose a prescriptive theory for representation learning. We first characterize optimal representations Z^* for supervised learning, by defining minimal sufficient representation with respect to a family of classifiers $\mathcal{V}$. These representations Z^* guarantee that any downstream empirical risk minimizer $f \in \mathcal{V}$ will incur minimal expected test loss, by ensuring that f can correctly predict labels but cannot distinguish examples with the same label. We then provide the decodable information bottleneck objective to learn Z^* with PAC-style guarantees. We empirically show that using Z^* can improve the performance and robustness of image classifiers. We also demonstrate that our framework can be used to predict generalization in neural networks.

In addition to supporting our theory, our experiments raise interesting questions for future work. First, results in Sec. 4.2 suggest that performance is causally related with the degree of $\mathcal{V}$-minimality of a representation, even though we only prove it for "perfect" $\mathcal{V}$-minimality. A natural question, then, is whether generalization bounds can be derived for approximate $\mathcal{V}$-minimality. Second, the high correlation between generalization in neural networks and the degree $\mathcal{V}$-minimality (Table 2) suggest that it might be an important quantity to study for understanding generalization in deep learning.

More generally, our work shows that information theory in theoretical and applied ML can benefit from incorporating the predictive family $\mathcal{V}$ of interest. For example, we believe that many issues of mutual information [89] in self-supervised learning [90–92], and IB [33, 93, 94] in IB's theory of deep learning [14, 95] could be solved by taking into account $\mathcal{V}$. By extending $\mathcal{V}$-information to arbitrary r.v. (through decompositions) we hope to enable its use in those and many other domains.

[9] We report Jiang et al.'s [50] results since our experiments and Sharp. Mag. differs slightly from theirs.

Broader Impact

Our work takes the perspective that an "optimal" representation is one such that any classifier that fits the training data should generalize well to test. In terms of potential practical benefits, it is possible that using our optimal representations, one can alleviate the effort of hyperparameter search and selection currently required to tune deep learning models. This could be a step towards democratizing machine learning to sections of the society without large computational resources – since hyperparameter search is often computationally expensive. We do not anticipate that our work will advantage or disadvantage any particular group.

Acknowledgments and Disclosure of Funding

We would like to thank: Naman Goyal for early feedback and best engineering practices; Brandon Amos for support concerning min-max optimization; Stephane Deny for suggesting to look for the Worst ERM; Emile Mathieu, Chris Maddison, Sho Yaida, and Max Nickel for helpful discussions and feedback; Jakob Foerster for the name "decodable" information bottleneck; Dan Roy for suggesting to use the term "distinguishability" to understand $\mathcal{V}$-minimality; and Ari Morcos for tips to help Yann Dubois writing papers. DJS was partially supported by the NSF through the CPBF PHY-1734030, a Simons Foundation fellowship for the MMLS, and by the NIH under R01EB026943.

References

[1] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[3] T. K. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, 2001. [Online]. Available: https://doi.org/10.1023/A:1011126920638

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[5] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2007, pp. 1177–1184. [Online]. Available: http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines

[6] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2013.50

[7] G. Zhong, L. Wang, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *CoRR*, vol. abs/1611.08331, 2016. [Online]. Available: http://arxiv.org/abs/1611.08331

[8] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. [Online]. Available: http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms

[9] F. Gressmann, F. J. Király, B. A. Mateen, and H. Oberhauser, "Probabilistic supervised learning," *CoRR*, vol. abs/1801.00753, 2018. [Online]. Available: http://arxiv.org/abs/1801.00753

[10] J. Duchi, K. Khosravi, F. Ruan *et al.*, "Multiclass classification, information, divergence and surrogate risk," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3246–3275, 2018.

[11] N. Tishby, F. C. N. Pereira, and W. Bialek, "The information bottleneck method," *CoRR*, vol. physics/0004057, 2000. [Online]. Available: http://arxiv.org/abs/physics/0004057

[12] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," *Theor. Comput. Sci.*, vol. 411, no. 29-30, pp. 2696–2711, 2010. [Online]. Available: https://doi.org/10.1016/j.tcs.2010.04.006

[13] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, E. J. Yannakoudakis, N. J. Belkin, P. Ingwersen, and M. Leong, Eds. ACM, 2000, pp. 208–215. [Online]. Available: https://doi.org/10.1145/345508.345578

[14] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017. [Online]. Available: http://arxiv.org/abs/1703.00810

[15] P. Farajiparvar, A. Beirami, and M. S. Nokleby, "Information bottleneck methods for distributed learning," in *56th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2018, Monticello, IL, USA, October 2-5, 2018*. IEEE, 2018, pp. 24–31. [Online]. Available: https://doi.org/10.1109/ALLERTON.2018.8635884

[16] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, H. Larochelle, M. Botvinick, Y. Bengio, and S. Levine, "Infobot: Transfer and exploration via the information bottleneck," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=rJg8yhAqKm

[17] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, 1948. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[18] M. Chalk, O. Marre, and G. Tkacik, "Relevant sparse codes with variational information bottleneck," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1957–1965. [Online]. Available: http://papers.nips.cc/paper/6101-relevant-sparse-codes-with-variational-information-bottleneck

[19] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=HyxQzBceg

[20] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2897–2905, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2784440

[21] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019. [Online]. Available: https://doi.org/10.3390/e21121181

[22] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.4.541

[23] Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon, "A theory of usable information under computational constraints," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=r1eBeyHFDH

[24] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Mach. Learn. Res.*, vol. 19, pp. 50:1–50:34, 2018. [Online]. Available: http://jmlr.org/papers/v19/17-646.html

[25] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, no. 594-604, pp. 309–368, 1922.

[26] M. Vera, P. Piantanida, and L. R. Vega, “The role of information complexity and randomization in representation learning,” *CoRR*, vol. abs/1802.05355, 2018. [Online]. Available: http://arxiv.org/abs/1802.05355

[27] B. Rodriguez Galvez, “The information bottleneck: Connections to other problems, learning and exploration of the ib curve,” 2019.

[28] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham, “Reversible architectures for arbitrarily deep residual neural networks,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 2811–2818. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16517

[29] J. Jacobsen, A. W. M. Smeulders, and E. Oyallon, “i-revnet: Deep invertible networks,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=HJsjkMb0Z

[30] D. McAllester and K. Stratos, “Formal limitations on the measurement of mutual information,” in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 2020, pp. 875–884. [Online]. Available: http://proceedings.mlr.press/v108/mcallester20a.html

[31] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for gaussian variables,” *J. Mach. Learn. Res.*, vol. 6, pp. 165–188, 2005. [Online]. Available: http://jmlr.org/papers/v6/chechik05a.html

[32] M. Rey and V. Roth, “Meta-gaussian information bottleneck,” in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1925–1933. [Online]. Available: http://papers.nips.cc/paper/4517-meta-gaussian-information-bottleneck

[33] R. A. Amjad and B. C. Geiger, “Learning representations for neural network-based classification using the information bottleneck principle,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[34] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[35] L. G. Valiant, “A theory of the learnable,” *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984. [Online]. Available: https://doi.org/10.1145/1968.1972

[36] P. L. Bartlett and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results,” *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2002. [Online]. Available: http://jmlr.org/papers/v3/bartlett02a.html

[37] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Sy8gdB9xx

[38] I. J. Goodfellow, “On distinguishability criteria for estimating generative models,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6515

[39] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, “Unrolled generative adversarial networks,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=BydrOIcle

[40] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, Dec. 1989. [Online]. Available: https://doi.org/10.1007/BF02551274

[41] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991. [Online]. Available: https://doi.org/10.1016/0893-6080(91)90009-T

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[43] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[44] Y. Li, C. Wei, and T. Ma, “Towards explaining the regularization effect of initial large learning rate in training neural networks,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 11 669–11 680.

[45] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

[46] G. K. Dziugaite and D. M. Roy, “Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data,” in *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, G. Elidan, K. Kersting, and A. T. Ihler, Eds. AUAI Press, 2017. [Online]. Available: http://auai.org/uai2017/proceedings/papers/173.pdf

[47] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, “Stronger generalization bounds for deep nets via a compression approach,” in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 254–263. [Online]. Available: http://proceedings.mlr.press/v80/arora18b.html

[48] A. S. Morcos, D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick, “On the importance of single directions for generalization,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=r1iuQjxCZ

[49] Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio, “Predicting the generalization gap in deep networks with margin distributions,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=HJlQfnCqKX

[50] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, “Fantastic generalization measures and where to find them,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SJgIPJBFvH

[51] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, “Sensitivity and generalization in neural networks: an empirical study,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=HJC2SzZCW

[52] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1225–1234. [Online]. Available: http://proceedings.mlr.press/v48/hardt16.html

[53] S. L. Smith and Q. V. Le, “A bayesian perspective on generalization and stochastic gradient descent,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=BJij4yg0Z

[54] P. L. Bartlett, D. J. Foster, and M. Telgarsky, “Spectrally-normalized margin bounds for neural networks,” in *Advances in Neural Information Processing Systems 30: Annual*

Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6240–6249. [Online]. Available: http://papers.nips.cc/paper/7204-spectrally-normalized-margin-bounds-for-neural-networks

[55] S. Hochreiter and J. Schmidhuber, "Flat minima," *Neural Computation*, vol. 9, no. 1, pp. 1–42, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.1.1

[56] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. T. Chayes, L. Sagun, and R. Zecchina, "Entropy-sgd: Biasing gradient descent into wide valleys," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=B1YfAfcgl

[57] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=H1oyRlYgg

[58] C. Wei and T. Ma, "Data-dependent sample complexity of deep neural networks via lipschitz augmentation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 9722–9733. [Online]. Available: http://papers.nips.cc/paper/9166-data-dependent-sample-complexity-of-deep-neural-networks-via-lipschitz-augmentation

[59] B. Neyshabur, R. Salakhutdinov, and N. Srebro, "Path-sgd: Path-normalized optimization in deep neural networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2422–2430. [Online]. Available: http://papers.nips.cc/paper/5797-path-sgd-path-normalized-optimization-in-deep-neural-networks

[60] M. G. KENDALL, "A NEW MEASURE OF RANK CORRELATION," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 06 1938. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81

[61] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton, "Regularizing neural networks by penalizing confident output distributions," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=HyhbYrGYe

[62] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, ser. JMLR Workshop and Conference Proceedings, P. Grünwald, E. Hazan, and S. Kale, Eds., vol. 40. JMLR.org, 2015, pp. 1376–1401. [Online]. Available: http://proceedings.mlr.press/v40/Neyshabur15.html

[63] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=rJl-b3RcF7

[64] E. T. Jaynes and R. D. Rosenkrantz, *E.T. Jaynes : papers on probability, statistics, and statistical physics*. D. Reidel ; Sold and distributed in the U.S.A. and Canada by Kluwer Boston Dordrecht, Holland ; Boston : Hingham, MA, 1983.

[65] I. Csiszar *et al.*, "Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems," *The annals of statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

[66] F. Topsøe, "Information-theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. 8–27, 1979. [Online]. Available: http://www.kybernetika.cz/content/1979/1/8

[67] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

[68] P. D. Grünwald, A. P. Dawid *et al.*, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," *the Annals of Statistics*, vol. 32, no. 4, pp. 1367–1433, 2004.

[69] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, “A robust minimax approach to classification,” *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, 2002. [Online]. Available: http://jmlr.org/papers/v3/lanckriet02a.html

[70] A. Globerson and N. Tishby, “The minimum information principle for discriminative learning,” in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI ’04. Banff, Canada: AUAI Press, Jul. 2004, pp. 193–200.

[71] F. Farnia and D. Tse, “A minimax approach to supervised learning,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 4233–4241. [Online]. Available: http://papers.nips.cc/paper/6247-a-minimax-approach-to-supervised-learning

[72] X. Nguyen, M. J. Wainwright, and M. I. Jordan, “Estimating divergence functionals and the likelihood ratio by convex risk minimization,” *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010. [Online]. Available: https://doi.org/10.1109/TIT.2010.2068870

[73] H. Drygas, “Sufficiency and completeness in the general gauss-markov model,” *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 45, no. 1, pp. 88–98, 1983. [Online]. Available: http://www.jstor.org/stable/25050416

[74] J. K. Baksalary and R. Kala, “Linear transformations preserving best linear unbiased estimators in a general gauss-markoff model,” *The Annals of Statistics*, pp. 913–916, 1981.

[75] R. Kala, S. Puntanen, and Y. Tian, “Some notes on linear sufficiency,” *Statistical Papers*, vol. 58, no. 1, pp. 1–17, 2017.

[76] M. Cvitkovic and G. Koliander, “Minimal achievable sufficient statistic learning,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 1465–1474. [Online]. Available: http://proceedings.mlr.press/v97/cvitkovic19a.html

[77] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan, “Learning the kernel matrix with semidefinite programming,” *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004. [Online]. Available: http://jmlr.org/papers/v5/lanckriet04a.html

[78] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, “Multiple kernel learning, conic duality, and the SMO algorithm,” in *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, ser. ACM International Conference Proceeding Series, C. E. Brodley, Ed., vol. 69. ACM, 2004. [Online]. Available: https://doi.org/10.1145/1015330.1015424

[79] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, “Large scale multiple kernel learning,” *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006. [Online]. Available: http://jmlr.org/papers/v7/sonnenburg06a.html

[80] M. Gönen and E. Alpaydin, “Multiple kernel learning algorithms,” *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021071

[81] C. A. Micchelli and M. Pontil, “Feature space perspectives for learning the kernel,” *Mach. Learn.*, vol. 66, no. 2-3, pp. 297–319, 2007. [Online]. Available: https://doi.org/10.1007/s10994-006-0679-0

[82] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. A. Poggio, and V. Vapnik, “Feature selection for svms,” in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2000, pp. 668–674. [Online]. Available: http://papers.nips.cc/paper/1850-feature-selection-for-svms

[83] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, “Choosing multiple parameters for support vector machines,” *Mach. Learn.*, vol. 46, no. 1-3, pp. 131–159, 2002. [Online]. Available: https://doi.org/10.1023/A:1012450327387

[84] N. Srebro and S. Ben-David, “Learning bounds for support vector machines with learned kernels,” in *Learning Theory, 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006, Proceedings*, ser. Lecture Notes in Computer Science, G. Lugosi and H. U. Simon, Eds., vol. 4005. Springer, 2006, pp. 169–183. [Online]. Available: https://doi.org/10.1007/11776420_15

[85] C. Cortes, M. Mohri, and A. Rostamizadeh, “Generalization bounds for learning kernels,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, J. Fürnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 247–254. [Online]. Available: https://icml.cc/Conferences/2010/papers/179.pdf

[86] M. Kloft and G. Blanchard, “On the convergence rate of lp-norm multiple kernel learning,” *J. Mach. Learn. Res.*, vol. 13, pp. 2465–2502, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2503321

[87] C. Cortes, M. Kloft, and M. Mohri, “Learning kernels using local rademacher complexity,” in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 2760–2768. [Online]. Available: http://papers.nips.cc/paper/4896-learning-kernels-using-local-rademacher-complexity

[88] Y. Liu, S. Liao, H. Lin, Y. Yue, and W. Wang, “Infinite kernel learning: Generalization bounds and algorithms,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S. P. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 2280–2286. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14186

[89] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, “On mutual information maximization for representation learning,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rkxoh24FPH

[90] R. Linsker, “Self-organization in a perceptual network,” *IEEE Computer*, vol. 21, no. 3, pp. 105–117, 1988. [Online]. Available: https://doi.org/10.1109/2.36

[91] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[92] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[93] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, “On the information bottleneck theory of deep learning,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=ry_WPG-A-

[94] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk, “Caveats for information bottleneck in deterministic scenarios,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=rke4HiAcY7

[95] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*. IEEE, 2015, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ITW.2015.7133169

[96] G. Parmigiani, L. Inoue, and H. Lopes, *Decision Theory: Principles and Approaches*. Wiley Blackwell, Dec. 2010.

[97] A. P. Dawid, “Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design,” *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, vol. 139, 1998.

[98] J. M. Bernardo and A. F. Smith, *Bayesian theory*. John Wiley & Sons, 2009, vol. 405.

[99] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[100] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035. [Online]. Available: http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[101] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

[102] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1180–1189. [Online]. Available: http://proceedings.mlr.press/v37/ganin15.html

[103] B. A. Pearlmutter and J. M. Siskind, "Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator," *ACM Trans. Program. Lang. Syst.*, vol. 30, no. 2, pp. 7:1–7:36, 2008. [Online]. Available: https://doi.org/10.1145/1330017.1330018

[104] D. Maclaurin, D. Duvenaud, and R. P. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 2113–2122. [Online]. Available: http://proceedings.mlr.press/v37/maclaurin15.html

[105] E. Grefenstette, B. Amos, D. Yarats, P. M. Htut, A. Molchanov, F. Meier, D. Kiela, K. Cho, and S. Chintala, "Generalized inner loop meta-learning," *CoRR*, vol. abs/1910.01727, 2019. [Online]. Available: http://arxiv.org/abs/1910.01727

[106] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[107] W. R. Huang, Z. Emam, M. Goldblum, L. Fowl, J. K. Terry, F. Huang, and T. Goldstein, "Understanding generalization through visualizations," *CoRR*, vol. abs/1906.03291, 2019. [Online]. Available: http://arxiv.org/abs/1906.03291

[108] D. Kalimeris, G. Kaplun, P. Nakkiran, B. L. Edelman, T. Yang, B. Barak, and H. Zhang, "SGD on neural networks learns functions of increasing complexity," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3491–3501. [Online]. Available: http://papers.nips.cc/paper/8609-sgd-on-neural-networks-learns-functions-of-increasing-complexity