Learning Optimal Representations with the Decodable Information Bottleneck

Yann Dubois  
Facebook AI Research  
yann@fb.com

Douwe Kiela  
Facebook AI Research  
dkiela@fb.com

David J. Schwab  
Facebook AI Research  
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 datasetrepresentation, and aims to restrict the predictive functional family $\mathcal{V}$ (e.g. linear classifiers) such that empirical risk minimizers (ERMs) generalize. 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.

was developed for communication theory [17] and does not take into account the predictive family \( V \) of interest. As a result, IB’s sufficiency requirement does not ensure the existence of a predictor \( f \in V \) that can perform well using the learned representation, 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 \( V \). Intuitively, it ensures that classifiers in \( V \) can predict labels (\( V \)-sufficiency) but cannot distinguish examples with the same label (\( V \)-minimality), as illustrated in Fig. 1. Our main contributions can be summarized as follows:

- We generalize notions of minimality and sufficiency to consider predictors \( V \) of interest.
- We prove that such representations are optimal — every downstream ERM in \( V \) 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 \( V \)-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 \( f \in V \), and Bob, who provides her with a representation to improve her performance. We are interested in Bob’s optimal choice. Specifically:

1. Alice decides \textit{a priori} on: a predictive family \( V \), 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 \( P_{Z|X} \) to map inputs \( X \) to representations \( Z \).
3. Using Bob’s encoder and a dataset \( D \) \( \overset{\text{iid}}{\sim} P_{X,Y}^M \) of \( M \) input-output pairs \( (x, y) \), Alice selects a classifier \( \hat{f} \) from all ERMs \( \hat{V}(D) := \text{arg min}_{f \in V} \hat{R}(f, Z; D) \), where \( \hat{R}(f, Z; D) := \frac{1}{M} \sum_{y,x \sim D} E_z \sim P_Z(z) [S(y, f[z])] \) is an estimate of the risk \( R(f, Z) = E_D \hat{R}(f, Z; D) \).

\[ ^2 \text{As an illustration, IB invariance to bijections suggests that a non-linearly entangled representation is as good as a linearly separable one if there is a bijection between them, even when classifying using a logistic regression.} \]
We stipulate that these known issues stem from a common cause: IB uses mutual information, which

We review IB, an information theoretic method for supervised representation learning. IB is built

We assume that: (i) sample spaces \( \mathcal{Y}, \mathcal{Z}, \mathcal{X} \) are finite; (ii) the family \( \mathcal{Y} \) is independent of \( \mathcal{V} \) and thus does not favor simple decision boundaries [33]; (iii) \( \mathcal{L}_{\text{IB}} \) 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.

\[ \mathcal{V} \text{-information is easier to estimate than Shannon's information; (ii) } \mathcal{L}_{\text{IB}} \text{ is invariant to bijections and thus does not favor simple decision boundaries [33] that can be achieved by a } f \in \mathcal{V}. \]

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

\[ \begin{align*}
\text{3 Methods} & \quad \text{In this section, we define } \mathcal{V} \text{-minimal } \mathcal{V} \text{-sufficient representations, prove that they are optimal in the two-player representation learning game, and discuss how to approximately learn them in practice.} \\
\text{3 Methods} & \quad \text{In this section, we define } \mathcal{V} \text{-minimal } \mathcal{V} \text{-sufficient representations, prove that they are optimal in the two-player representation learning game, and discuss how to approximately learn them in practice.} \\
\end{align*} \]

\[ \text{As shown in Appx. C.1.1 this common [12] definition is a generalization of minimal sufficient statistics [23] to stochastic statistics } Z = T(X, \epsilon) \text{ where } \epsilon \text{ is a source of noise independent of } X. \]

\[ \text{We assume that: (i) sample spaces } \mathcal{Y}, \mathcal{Z}, \mathcal{X} \text{ are finite; (ii) the family } \mathcal{Y} \text{ is independent of } \mathcal{V} \text{ and thus does not favor simple decision boundaries [33]; (iii) } \mathcal{L}_{\text{IB}} \text{ is invariant to bijections and thus does not favor simple decision boundaries [33] that can be achieved by a } f \in \mathcal{V}. \]

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

Let us study Alice’s best risk minimization \(\min_{f \in \mathcal{V}} R(f, Z)\) using a representation \(Z\). This tight lower bound on her performance looks strikingly similar to \(H_\mathcal{V}[Y | Z]\), which is controlled by \(I_\mathcal{V}[Z \to Y]\) (see Eq. (5)). As a result, if Bob maximizes \(I_\mathcal{V}[Z \to Y]\), 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 \(S_\mathcal{V} := \arg \max_{Z'} I_\mathcal{V}[Z' \to Y]\).

**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., \(R(f^*, Z) = \min_{f \in \mathcal{V}} \min_{f \in \mathcal{V}} R(f, Z)\).

Although the previous proposition may seem trivial, it bears important implications, namely that contrary to the sufficiency term of IB one should maximize \(I_\mathcal{V}[Z \to Y]\) rather than \(I[Y; Z]\) when predictors live in a constrained family \(\mathcal{V}\). Indeed, ensuring sufficient \(I[Y; Z]\) does not mean that there is a classifier \(f \in \mathcal{V}\) that can “decode” that information. 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 choose to guarantee that Alice’s ERMs will perform optimally by ensuring that any ERM generalizes beyond the training set.

IB suggests minimizing the information \(I[Y; X]\) 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 \(I_\mathcal{V}[Z \to X]\). However, the latter is not defined as \(X\) does not generally take value in (t.v.i.) \(\mathcal{V}\), the sample space of \(Y\)’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 \(I_\mathcal{V}[Z \to N]\) is well defined. Specifically, let \(X_{y}, Z_{y}\) be “conditional r.v.” s.t. \(P_{X_{y}} = P_{X_{y} | y}, P_{Z_{y}} = P_{Z_{y} | y}, P_{X_{y} Z_{y}} = P_{X_{y} Z_{y}}\).

We define the \(y\) decomposition of \(X\) as r.v.s that arise by all possible labelings of \(X_{y}\):

\[
Dec(X, y) := \{N \mid \exists t' : \mathcal{X} \to \mathcal{Y} \ s.t. \ N = t'(X_{y})\}
\]

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

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

\(I_\mathcal{V}[Z \to Dec(X, Y)]\) essentially measures how well predictors in \(\mathcal{V}\) can predict arbitrary labeling \(N \in Dec(X, y)\) of examples with the same underlying label \(y\). Replacing the minimality term \(I[X; Z]\) by \(I_\mathcal{V}[Z \to Dec(X, Y)]\) 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 \(M_\mathcal{V} := \arg \min_{Z \in S_\mathcal{V}} I_\mathcal{V}[Z \to Dec(X, Y)]\).

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, \(D \overset{iid}{\sim} P_{X,Y}^M\) a dataset, and assume labels \(Y\) are deterministic functions of the inputs \(t(X)\). If \(Z \in M_\mathcal{V}\) be \(\mathcal{V}\)-minimal \(\mathcal{V}\)-sufficient, then the expected test loss of any ERM \(f \in \mathcal{V}(D)\) is the best achievable risk, i.e., \(R(f, Z) = \min_{f \in \mathcal{V}} \min_{f \in \mathcal{V}} R(f, Z)\).

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

---

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

2 Such (deterministic) labelings are also called “random labelings” [37] as they are semantically meaningless.
Function $\mathcal{H}(\mathcal{V}, x, y, P_{Z|X})$:
\[
\begin{align*}
    z &\leftarrow \text{sample once from each } P_{Z|X} \\
    \text{return } &\inf_{f \in \mathcal{V}} \sum_{x, y, z} \log f(z|y)
\end{align*}
\]

Function $\mathcal{L}_{\text{DIB}}(\mathcal{V}, D, P_{Z|X}, K, \beta)$:
\[
\begin{align*}
    (x, y), &\mathcal{L}_{\text{min}}, \mathcal{Y} \leftarrow D, 0, \text{unique}(y) \\
    \mathcal{L}_{\text{ suff}} &\leftarrow \mathcal{H}(\mathcal{V}, x, y, P_{Z|X}) \\
    \text{for } y \in \mathcal{Y} \text{ do} &\text{ } \\
    x_{y} &\leftarrow x | y = y \\
    \text{for } k \leftarrow 1 \text{ to } K \text{ do} &\text{ } \\
    n &\leftarrow \text{random_choice}(\mathcal{Y}, \text{size}=|x_{y}|) \\
    \mathcal{L}_{\text{min}} &\leftarrow \min \mathcal{H}(P_{Z|X}, n) \\
\end{align*}
\]
\[
\begin{align*}
    \text{return } &\text{(const)} + \mathcal{L}_{\text{ suff}} - \beta \star \mathcal{L}_{\text{min}}
\end{align*}
\]

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

Figure 2: Practical DIB (a) Pseudo-code to compute the $\hat{\mathcal{L}}_{\text{DIB}}(D)$; (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.s. 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.,
  \[
  \arg \max_{Z \in \mathcal{S}_{\mathcal{V}}} I_{\mathcal{V}}[Z \rightarrow \text{Dec}(X, \mathcal{Y})] \subseteq \mathcal{M}_{\mathcal{V}}.
  \]
- **Characterization** $Z \in \mathcal{M}_{\mathcal{V}} \iff Z \in \mathcal{S}_{\mathcal{V}}$ and $I_{\mathcal{V}}[Z \rightarrow \text{Dec}(X, \mathcal{Y})] = 0$.
- **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 $\mathcal{Y} = \mathcal{U}$. 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 $Z^{*} \in \mathcal{M}_{\mathcal{V}}$. Unfortunately, Bob cannot learn these $Z^{*}$ as it requires the underlying distribution $P_{X, Y}$. We will now show that he can nevertheless approximate $Z^{*}$ in a sample- and computationally-efficient manner.

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

\[
\mathcal{L}_{\text{DIB}} := - I_{\mathcal{V}}[Z \rightarrow Y] + \beta \star I_{\mathcal{V}}[Z \rightarrow \text{Dec}(X, \mathcal{Y})]
\]

Notice that each $I_{\mathcal{V}}[Z \rightarrow \cdot]$ has an internal optimization. In particular $I_{\mathcal{V}}[Z \rightarrow \text{Dec}(X, \mathcal{Y})]$ 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 $\mathcal{L}_{\text{DIB}}$ 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 $\mathcal{L}_{\text{DIB}}$ over $\mathcal{L}_{\text{IB}}$ is that it can be estimated with guarantees using finite samples. Namely, if Bob has access to a training set $D \overset{iid}{\leftarrow} P_{X,Y}$, he can estimate $\mathcal{L}_{\text{DIB}}$ reasonably well. In practice, we: (i) use $D$ to estimate all expectations over $P_{X,Y}$; (ii) use samples from Bob’s encoder $z \sim P_{Z|X}$; (iii) estimate the average over $N \in \text{Dec}(X, y)$ in Eq. (7) using $K$ samples. Figure 2a shows a (naive) algorithm to compute the resulting estimate $\mathcal{L}_{\text{DIB}}(D)$. Despite these approximations, we show in [Appx. C.3] that $\mathcal{L}_{\text{DIB}}$ inherits $\mathcal{V}$-information’s estimation bounds [23].
We study optimal representations when Alice has access to the data distribution $P_{Z|X}$. Alice’s risk $R(f, Z)$ 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 more complex $V$-sufficiency head (in blue) tries to classify cats and dogs. Each of the $K$ (typically 3-4, see Appx. E.4) $V$-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 $V$-minimality heads are reversed. The $V$-minimality losses are also multiplied by a hyper-parameter $\beta$.

**4 Experiments**

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

**4.1 $V$-sufficiency: Optimal Representations When the Data Distribution is Known**

We study optimal representations when Alice has access to the data distribution $P_{Z|X}$. Alice’s risk $R(f, Z)$ 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...
1. This suggests that IB’s sufficiency (infinite width \( \mathcal{V} \)) should Bob choose? 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 \( \mathcal{L}_{DIB}(D) \).

Experimental Details. Our claim concerns all ERMs \( \hat{\mathcal{V}}(D) \), which cannot be supported by training a few \( \hat{f} \in \hat{\mathcal{Y}}(D) \). Instead, we evaluate the ERM that performs worst on the test set (Worst ERM), i.e., \( \arg \max f \in \hat{\mathcal{V}}(D) R(f, Z) \). We do so by optimizing the following Lagrangian relaxation \( \arg \min f \in \hat{\mathcal{V}} \hat{R}(f, Z; D) = \gamma R(f, Z) \) (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 \( \beta \)? We train representations on CIFAR10 with various \( \beta \) to investigate the effect of \( I_{\mathcal{V}}(Z \rightarrow \hat{\text{Dec}}(X, Y); D) \) and \( I_{\mathcal{V}}(Y \rightarrow Z; D) \) (Fig. 4b) on Alice’s performance (Fig. 4a). Increasing \( \beta \) results in a decrease in \( I_{\mathcal{V}}(Z \rightarrow \hat{\text{Dec}}(X, Y); D) \) which monotonically shrinks the train-test gap. This suggests that, although our theory only applies for \( \mathcal{V} \)-minimality, \( I_{\mathcal{V}}(Z \rightarrow \hat{\text{Dec}}(X, Y)) \) 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 \( I_{\mathcal{V}}(Y \rightarrow Z) \), which controls the best achievable loss. This shows that a trade-off (controlled by \( \beta \)) between \( \mathcal{V} \)-minimality (generalization) and \( \mathcal{V} \)-sufficient (lower bound) arises when Bob has to estimate \( \mathcal{L}_{DIB}(D) \).

\( \mathcal{V} \)-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 $\beta$ increases, predicting MNIST becomes harder. Indeed decreasing $I_{V^+}[Z \rightarrow \text{Dec}(X, Y)]; D]$ removes all $V^+$-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.\footnote{Achille and Soatto \cite{achille2019information} show that this happens for minimal $Z$. The novelty is that we obtain similar results when considering $V^+$-minimality, which is less stringent \cite{achille2019information} and does not require the intractable $I[X; Z]$.}

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst</td>
<td>10.23 ± 1.3</td>
<td>8.61 ± 0.5</td>
<td>1.90 ± 0.0</td>
<td>10.25 ± 0.3</td>
<td>1.82 ± 0.2</td>
<td>1.54 ± 0.3</td>
<td>1.94 ± 0.3</td>
<td>1.41 ± 0.1</td>
</tr>
<tr>
<td>Avg.</td>
<td>4.62 ± 0.0</td>
<td>4.34 ± 0.4</td>
<td>1.49 ± 0.0</td>
<td>4.90 ± 0.3</td>
<td>1.76 ± 0.1</td>
<td>1.47 ± 0.1</td>
<td>1.74 ± 0.1</td>
<td>1.38 ± 0.1</td>
</tr>
</tbody>
</table>

Which $V^+$-minimality should Bob chose? We study the effect of $Z \in S_{V^+\text{Alice}}$ being minimal w.r.t. families which are larger ($V^+\text{-DIB}$), smaller ($V^+\text{+DIB}$), and equal to $V^+\text{Alice}$. In theory, optimal representations would be $V^+\text{Alice}$-minimal \cite{achille2019information}, which are achieved by DIB and $V^+\text{-DIB}$ (Monotonicity). $V^+\text{+DIB}$ should nevertheless be harder to estimate than DIB \cite{achille2019information}. In the last 3 columns of Table 1 we indeed observe that DIB performs best. $V^+\text{+DIB}$ performs worse, suggesting that IB’s minimality is undesirable in practice. We also minimize a known lower bound of $I[Z; X]$ (VIB; \cite{lee2017deep}) and find that it performs worse than DIB.\footnote{VIB is hard to compare to DIB as it is unclear w.r.t. which family, if any, VIB’s solutions are minimal.} We show results for different $\beta$ 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 \cite{srivastava2014dropout}; (iv) Weight decay. We find that DIB significantly outperforms other regularizers, which supports our claims that $V^+$-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 \cite{ben2019provably,shlezinger2019probing,cheng2019analyzing}, as they can shed light on the inductive biases in deep learning \cite{atzei2019information,lee2019bounds,deepmin,deepmin2} and prescribe better training procedures \cite{zhang2019lbc,li2019searching}. Having empirically shown a strong link between the degree of $V$-minimality and generalization (Fig. 2b), it is natural to ask whether we 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 $V_L$. We hypothesize that $I_{V_L}[Z_L \rightarrow \text{Dec}(X, Y)]; D]$ correlates well with the generalization of the network.

To test this, we follow Jiang et al. \cite{jiang2020empirical} 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 \cite{jiang2020empirical} between $I_{V_L}[Z_L \rightarrow \text{Dec}(X, Y)]; D]$ 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 $\tau$ between different measures and generalization gap (in terms of accuracy $\tau_{\text{acc}}$ and log loss $\tau_{\text{logloss}}$) of 562 CNNs. $V_L$ denotes our $I_{V_L}[Z_L \rightarrow \text{Dec}(X, Y)]; D]$

<table>
<thead>
<tr>
<th>$\tau_{\text{acc}}$</th>
<th>$V^+_L$</th>
<th>$V^-_L$</th>
<th>$V^-_L$</th>
<th>Entropy</th>
<th>Path Norm</th>
<th>Var. Grad.</th>
<th>Sharp Mag.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.482</td>
<td>0.391</td>
<td>0.471</td>
<td>0.234</td>
<td>0.347</td>
<td>0.332</td>
<td>0.385</td>
<td>\textbf{0.484}</td>
</tr>
<tr>
<td>\text{ours}</td>
<td>0.505</td>
<td>0.435</td>
<td>0.498</td>
<td>0.164</td>
<td>0.357</td>
<td>0.167</td>
<td>0.233</td>
</tr>
</tbody>
</table>

Does $V^+$-minimality correlate with generalization? In the last five columns of Table 2, we compare our results ($V_L$) to the best generalization measure from each categories investigated in \cite{jiang2020empirical}:
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.; \([57]\)). As hypothesized, \(I_{ZL}[Z_L \rightarrow \text{Dec}(X, Y); D]\) correlates with generalization and even outperforms the baselines. Similarly to Table 1 we also evaluate minimality with respect to a family larger (\(V_L^+\)) or smaller (\(V_L^-\)) than \(V_L\). Surprisingly, \(V_L^+\) performs better than \(V_L\), which might be because larger networks can help optimization of sub-networks \(V_L \subseteq V_L^+\) as suggested by the Lottery Ticket Hypothesis. \([63]\).

To the best of our knowledge \(V\)-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 \(V\)-minimality is a property of a representation rather than the architecture, we show in Appx. E.12 that it can be meaningfully compared \textit{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 \textit{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 \(V\) (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 \(V\) 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 \(V\). These representations \(Z^*\) guarantee that any downstream empirical risk minimizer \(f \in V\) 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 \(V\)-minimality of a representation, even though we only prove it for “perfect” \(V\)-minimality. A natural question, then, is whether generalization bounds can be derived for approximate \(V\)-minimality. Second, the high correlation between generalization in neural networks and the degree \(V\)-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 \(V\) of interest. For example, we believe that many issues of mutual information \([89]\) in self-supervised learning \([90-92]\), and IB \([93, 94]\) in IB’s theory of deep learning \([14, 25]\) could be solved by taking into account \(V\). By extending \(V\)-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 $V$-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


