Information Regularized Neural Networks

Tianchen Zhao[†] **Dejiao Zhang**[†] [*] **Zeyu Sun**[†] **Honglak Lee**[*,†]
[†]University of Michigan
[*]Google Brain
{ericolon,dejiao,zeyusun,honglak}@umich.edu

Abstract

We formulate an information-based optimization problem for supervised classification. For invertible neural networks, the control of these information terms is passed down to the latent feature and parameter matrix in the last fully connected layer, given that mutual information is invariant under invertible map. We propose an objective function and prove that it solves the optimization problem; our regularizer improves the classification performance by a noticeable margin.

1 Introduction

Information Bottleneck (IB) problem (Tishby et al. [1999]) is formulated as:

$$\textit{minimize} \;\; \mathrm{I}(X;T) - \lambda \mathrm{I}(Y;T) \; , \tag{1}$$

where the random variable T is interpreted as a minimal sufficient representation of signal X for label Y. The mutual information term $\mathrm{I}(X;T)$ has its origins in Lossy Compression and Rate-Distortion Theory (Cover and Thomas [2006]), conveying an simple idea of ”keep only what is relevant”.

However, Saxe et al. [2018] argued that the mutual information $\mathrm{I}(X;T)$ between signal X and feature T in intermediate layer is infinite, as the transformation from X to continuous random variable T is deterministic. In addition they showed experimentally that layers equipped with ReLU actually do not compress too much information, which is supported by many recent work on the invertibility of the neural network. This motivates us to consider a different problem with similar principle idea: we would like to establish a theoretically valid objective that allows the neural network to extract only the relevant information for classification from the data.

We focus on the discrete prediction random variable $\widehat{Y}$ inferred by the probabilistic model $\mathbb{P}(\widehat{Y}|X)$ and introduce the following information optimization problem for supervised classification:

$$\begin{aligned} &\textit{maximize} \;\; \mathrm{I}(Y;\widehat{Y}) \\ &\textit{subject to} \;\; \mathrm{I}(X;\widehat{Y}) - \mathrm{I}(Y;\widehat{Y}) < \tau \; , \text{ for some } \tau > 0 \; . \end{aligned} \tag{2}$$

The intuition behind this objective lies in two-fold:

Information perspective: A good classification model should be robust against irrelevant features of X, and prevent over-fitting in the learning process. In optimization problem (2) we maximize the relevant information $\mathrm{I}(Y;\widehat{Y})$, while constraining the irrelevant information $\mathrm{I}(X;\widehat{Y}) - \mathrm{I}(Y;\widehat{Y})$ that X has about $\widehat{Y}$. Although $\mathrm{I}(X;\widehat{Y}) - \mathrm{I}(Y;\widehat{Y})$ converges to zero as $\mathrm{I}(Y;\widehat{Y})$ approaching its maximum (see Figure 1), in practice it's never attained due to the limited capacity of the models or over-fitting. Our proposed constrain addresses the problem of over-fitting: if two models achieve

[*]Dejiao Zhang's participation was supported by DARPA grant 16-43-D3M-FP-037.

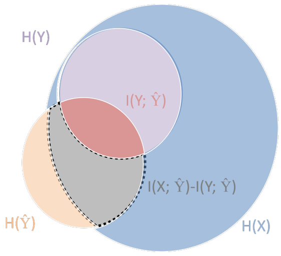

Figure 1: An information Venn diagram: three disks represent the entropy of $X, Y, \widehat{Y}$ respectively, the area in red is the relevant information $\mathrm{I}(Y;\widehat{Y})$, the area in grey is the irrelevant information $\mathrm{I}(X;\widehat{Y}) - \mathrm{I}(Y;\widehat{Y})$. The optimal solution is obtained when the smaller disks coincide, which is typically not achieved in practice. In particular, the trained model may be extremely confident in its prediction (when $H(\widehat{Y})$ lies inside of $H(X)$), but predicts the wrong label (having large grey area). Our optimization problem explicitly prohibits the growth of grey area throughout the training.

the similar classification accuracy, this constraint prefers the one that does not overfit to spurious factors of variation in X (e.g., pixel-level artifact/noise in the image that accidentally correlates to the labels in the training data).

Prediction confidence perspective: A good classification model should not be certain about its decision which is in fact wrong. However, modern neural networks are too confident in their predictions (Guo et al. [2017]). To be more precise, high capacity neural networks mostly assign labels of data with prediction confidence near 0 or 1. In particular, they assign 0 probability to correct labels for some data and therefore do not have enough flexibility to correct themselves from making the wrong prediction. We propose to compress the irrelevant information $\mathrm{I}(X;\widehat{Y}) - \mathrm{I}(Y;\widehat{Y})$, where minimizing $\mathrm{I}(X;\widehat{Y})$ decreases the confidence on all predictions but maximizing $\mathrm{I}(Y;\widehat{Y})$ increases the confidence on the correct predictions. Therefore the overall effect reduces the certainty on the false prediction of $\widehat{Y}$ (see Figure 1).

Our contribution: Our contribution lies in the following: (i) we formulate a novel information optimization problem for supervised classification and propose an objective function solving it; (ii) we formally justify the use of ℓ_1, ℓ_2 regularization from an information perspective. Different from the naive ℓ_2 regularization on parameters, regularization in our objective is novel and effective.

2 Related Work

Neyshabur et al. [2015] shows neural networks are equivalent up to some scaling factors passed among layers. As a consequence, there exist unbalanced neural networks with large ℓ_2 weights but are equivalent to those with small ℓ_2 weights. So the common belief that ℓ_2 regularization can ”simplify” a model does not necessarily make sense for deep model. Our idea of explicit regularization on w and $F(X)$ gives an interpretation to ℓ_1/ℓ_2 regularization methods in deep learning and is related to the margin based and stability based interpretations of generalization in deep learning respectively, studied by Arora et al. [2018], Bartlett and Mendelson [2003], Neyshabur et al. [2017], Sun et al. [2015].

Recent experimental work reports that neural networks with invertible structure have better performance. Dosovitskiy and Brox [2015] shows that images can be resconstructed from the latent features in AlexNet through an inverting process; this reconstruction is further improved by Zhang et al. [2016], where they also show reconstructive objective is beneficial to the performance of the neural network (e.g., VGGNet). Shang et al. [2016] proposes an invertible activation scheme named CReLU to preserve information; Gilbert et al. [2017] analyzes theoretically the invertibility of CNN; Jacobsen et al. [2018] build a theoretic invertible structure whose performance is comparable to ResNet He et al. [2015]. Invertibility seems to be an intriguing property or design principle that often emerges in the recent state-of-the-art deep architectures.

3 Main Results

Consider a classification problem where the training data $\mathcal{D} = \{(x_k, y_k)\}_{k=1:n}$ are sampled from random variable pair (X, Y) with unknown joint distribution. Each x_k is fed into a deep probabilistic model, which outputs probability densities and predicts $\widehat{y}_k$, a realization of the prediction random variable $\widehat{Y}$. Let $\mathcal{C}$ denote the label class and $\mathcal{X}$ denote the signal space.

Mutual information is bounded and its gradient with respect to logits is approximately zero over a large domain. In particular if the logits are initially small for true labels, gradient updates cannot effectively correct them. Therefore we choose to use a surrogate objective for information optimization problem (2).

We decompose the neural network into a feature map F and a fully connected matrix w in the last layer. Assuming F is invertible, neural network passes the control of $\mathrm{I}(X;\widehat{Y}) = \mathrm{I}(F(X);\widehat{Y})$ towards $F(X)$ and w in the last layer, where $F(X)$ can be interpreted as transformed signal that preserves information about the original signal X and the inference model becomes conceptually linear with classifier w. In Section 3.2 we derive a surrogate objective function (8) and prove that it solves the optimization problem (2).

The invertibility property has been empirically demonstrated for complex non-linear deep neural networks that are widely used in practice. We prove in Proposition A.1 that a lower bound of classification error is minimized if neural network is invertible and show in Section 4.2 how invertibility affects the performance of our proposed method.

3.1 Invertibility is Beneficial

We show in Proposition A.1 that the lower bound for the classification error is itself lower bounded by a constant, which is attained if F is invertible. Although the performance of the model also depends on the classifier w, our bound claims that an invertible feature map F could provide a better environment for the classifier w to perform well. The proof is provided in Appendix A.

Proposition 3.1. *(Fano's Inequality) The classification error is lower bounded as follows:*

$$\mathbb{P}(Y \neq \widehat{Y}) \geq \frac{H(Y|F(X)) - \log(2)}{\log(|\mathcal{C}| - 1)} . \tag{3}$$

The lower bound satisfies

$$\frac{H(Y|F(X)) - \log(2)}{\log(|\mathcal{C}| - 1)} \geq \frac{H(Y|X) - \log(2)}{\log(|\mathcal{C}| - 1)} \tag{4}$$

for all F, and the equality is attained if F is invertible.

3.2 Solving Optimization Problem with Feasible Terms

We first call out our assumptions used throughout our analysis. **(I)**: we assume the marginal densities of $Y, \widehat{Y}$ are uniform over $\mathcal{C}$; **(II)**: there exists a unique true label for every sample of X.

Without loss of generality, we consider the binary classification problem, i.e. the label class $\mathcal{C} = \{\pm 1\}$. To tract the population quantities $\mathrm{I}(F(X);Y)$ and $\mathrm{I}(Y;\widehat{Y})$, we decompose each of them into an empirical part and a probabilistic bound; the latter is negligible if sample size n is large. In Proposition 3.2, we show that in order to compress $\mathrm{I}(F(X);Y)$, we need to compress the norm of classifier w and feature $F(X)$. In particular, smaller $|w^T F(X)|$ represents lower confidence of the model on its predictions $\widehat{Y}$, indicating a smaller amount of mutual information $\mathrm{I}(F(X);\widehat{Y})$. The proof is provided in Appendix B.

Proposition 3.2. $\mathrm{I}(X;\widehat{Y}) = \mathrm{I}(F(X);Y)$ *is well estimated by its empirical version with high probability, which shares the same unique (global) minimum with $\sum_{k=1}^{n} |w^T F(x_k)|$ at $w^T F(x_k) = 0$, for all $k \in \{1, ..., n\}$.*

Proposition 3.3 establishes the relationship between maximization over mutual information $\mathrm{I}(Y;\widehat{Y})$ and minimization over cross entropy $-\sum_{k=1}^{n} \log \sigma(y_k w^T F(x_k))$ where σ is the sigmoid function; higher confidence of the model on its correct predictions indicates a larger value of $\mathrm{I}(Y;\widehat{Y})$. The proof is provided in Appendix C.

Proposition 3.3. $\mathrm{I}(Y;\widehat{Y})$ *is well estimated by its empirical version with high probability, which shares the same unique (global) maximum with $\sum_{k=1}^{n} \log \sigma(y_k w^T F(x_k))$ given that $y_k w^T F(x_k) > \frac{1}{2}$, for all $k \in \{1, ..., n\}$.*

3.3 Derivation of Objective Function

In Lagrangian form of optimization problem (2), the constant τ can be dropped and the objective becomes

$$\textit{maximize}\ \ (1+\lambda)\mathrm{I}(Y;\widehat{Y}) - \lambda\mathrm{I}(F(X);\widehat{Y}) \iff \textit{maximize}\ \ \mathrm{I}(Y;\widehat{Y}) - \frac{\lambda}{1+\lambda}\mathrm{I}(F(X);\widehat{Y})\,. \quad (5)$$

Consider a single signal x_k and its true label y_k, we propose the following objective function for binary supervised classification problem:

$$\mathcal{L}_k = \alpha\mathcal{R}\left(|w^T F(x_k)|\right) - \log\sigma(y_k w^T F(x_k))\,, \quad (6)$$

where $\mathcal{R}$ is some regularizer function. According to results in Section 3.2, minimizing (6) allows us to maximize $\mathrm{I}(Y;\widehat{Y})$ while constraining $\mathrm{I}(X;\widehat{Y})$.

Our regularizer should prefer a model that does not overfit among all the ones with high training accuracy. In this case neural networks assign only large logits $w_{y_k}^T F(x_k)$ to true label y_k for each signal x_k, and generalization of (6) to multi-class case for $\mathrm{I}(F(X);\widehat{Y})$ can be simplified to constraining $w_{y_k}^T F(x_k) - w_j^T F(x_k)$, where w_j is the jth column of w, assigning feature $F(x_k)$ a probability to label j. We propose to simply constrain $w_{y_k}^T F(x_k)$ and does not encourage the growth of $w_j^T F(x_k)$:

$$\mathcal{L}_k = \alpha\mathcal{R}\left(|w_{y_k}^T F(x_k)|\right) - \log\left(\frac{e^{w_{y_k}^T F(x_k)}}{e^{w_{y_k}^T F(x_k)} + \sum_{j\neq y_k} e^{w_j^T F(x_k)}}\right)\,. \quad (7)$$

In our experiment, we take the Elastic Net approach by Zou and Hastie [2005] using a combination of ℓ_2 and ℓ_1 regularizers: we use Holder's inequality to bound $|w_{y_k}^T F(x_k)|$ with both $\sup F(X)\cdot\|w_{y_k}\|_1$ and $\|w_{y_k}\|_2\cdot\|F(x_k)\|_2$. In practice we assume $\sup F(X)$ to be a (large) constant and is absorbed into the hyper-parameter. Our proposed objective function is of the form:

$$\mathcal{L} = \alpha_1||w||_1 + \frac{1}{n}\sum_{k=1}^{n}\left\{\alpha_2\|w_{y_k}\|_2\cdot\|F(x_k)\|_2 - \log\left(\frac{e^{w_{y_k}^T F(x_k)}}{e^{w_{y_k}^T F(x_k)} + \sum_{j\neq y_k} e^{w_j^T F(x_k)}}\right)\right\}\,. \quad (8)$$

4 Experiments

In our experiments we build the feature map F with ResNet. We argue that ResNet by He et al. [2015] is fairly invertible due to the intrinsic invertibility of the operator $I+\mathcal{L}$ in each block given $\|\mathcal{L}\| < 1$, which we experimentally verify to hold for all but the first layer in ResNet-32.

4.1 Performance

NaiveReg ResNet-32	$\alpha_2 = 0.0005$	$\alpha_2 = 0.001$	$\alpha_2 = 0.002$	$\alpha_2 = 0.004$
Best Accuracy %	70.06 ± 0.38	69.94 ± 0.33	69.86 ± 0.32	68.99 ± 0.52

ResNet-32	Original	$\alpha_2 = 0.01$	$\alpha_2 = 0.03$	$\alpha_2 = 0.05$
Best Accuracy %	70.15 ± 0.33	70.34 ± 0.27	**70.57** ± **0.20**	70.25 ± 0.23
Constrain	0.296 ± 0.044	0.293 ± 0.043	0.280 ± 0.040	**0.266** ± **0.038**

ResNetWide-28-10	Original	$\alpha_2 = 0.01$	$\alpha_2 = 0.05$	$\alpha_2 = 0.09$	$\alpha_2 = 0.15$
Best Accuracy %	78.51 ± 0.27	79.37 ± 0.18	79.62 ± 0.13	**79.64** ± **0.12**	79.45 ± 0.14
Constrain	0.254 ± 0.056	0.240 ± 0.051	0.213 ± 0.049	0.198 ± 0.044	**0.174** ± **0.038**

Table 1: Performance comparison on CIFAR100, Best Accuracy (%, test set) and the average values of the constrain $\mathrm{I}(X;Y) - \mathrm{I}(Y;\widehat{Y})$ throughout the training process are provided in the tables. α_1 is fixed to be 0 for ResNet-32 and $5e^{-6}$ for ResNet-Wide. All results are calculated from 10 samples.

We report the accuracy of ResNet on test data of CIFAR100 in Table 1.[2] Our regularizer outperforms the naive ℓ_2 regularization on w for ResNet-32. Note that the scales of hyper-parameters

[2] We use the open source code implemented by Xin Pan at https://github.com/tensorflow/models/tree/master/research/resnet and build our objective function based upon it. Our results for the baseline models match the reported ones on the website.

are different because naive ℓ_2 regularization acts on the full matrix w; intuitively naive ℓ_2 regularization should not be effective as the neural network can adapt by up-scaling the previous layers, which we observe in our experiments. Our regularization on $F(X), w$ improves the performance of ResNet-Wide by a noticeable margin. We conclude that our regularizer addresses the overfitting problem effectively.

4.2 The Role of Invertibility

Invertibility allows us to treat $F(X)$ as transformed data that preserves all the information from X, and therefore work on the information regularization problem under a linear scheme. In this section we build a PlainNet by using only $\mathcal{L}$ without residuals as the operator for each building block, so the theoretical guarantee for invertibility is not present for PlainNet. In Table 2, we see that PlainNet-32 can still benefit from our regularization, however, its performance is less stable compared to ResNet-32 if the hyper-parameters are too large. The reason is for PlainNet, the feature in the last layer $F(X)$ does not preserve information about X very well, so it has a higher demand on the capacity of the classifer w and is therefore more sensitive to our regularization.

Performance %	Original	$\alpha_2 = 0.01$	$\alpha_2 = 0.05$	$\alpha_2 = 0.09$	$\alpha_2 = 0.15$	$\alpha_2 = 0.3$
ResNet-32	92.49 ± 0.14	92.52 ± 0.30	92.76 ± 0.33	92.45 ± 0.37	88.36 ± 3.12	78.95 ± 4.33
PlainNet-32	90.06 ± 0.21	90.33 ± 0.24	90.25 ± 0.24	90.06 ± 0.31	85.97 ± 3.40	62.76 ± 16.22

Table 2: The performance statistics for ResNet-32 and PlainNet-32 without or under various regularizations; we hold $\alpha_1 = 1e^{-5}$ in the experiment. Results are calculated from 5 samples.

5 Conclusion

We give an interpretation of the deep learning dynamics by decomposing it into an signal transformation stage and feature classification stage, where we emphasis the importance of the classifier w in the last fully connected layer given that the feature map F is invertible. Then we take the advantage of the fact that mutual information quantities are invariant under invertible map to attack our proposed information optimization problem for supervised classification in deep learning. Our theory justifies the use of direct regularization terms on $w, F(X)$ for neural networks with decent invertibility property. Our regularization improves the performance of neural networks by a noticeable margin.

References

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. 1999.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. 2006.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *CoRR*, abs/1506.02617, 2015.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *CoRR*, abs/1802.05296, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3, March 2003.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *CoRR*, abs/1706.08947, 2017.

Shizhao Sun, Wei Chen, Liwei Wang, and Tie-Yan Liu. 2015.

Alexey Dosovitskiy and Thomas Brox. Inverting convolutional networks with convolutional networks. *CoRR*, abs/1506.02753, 2015.

Yuting Zhang, Kibok Lee, and Honglak Lee. Augmenting supervised neural networks with unsupervised objectives for large-scale image classification. *CoRR*, abs/1606.06582, 2016.

Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. *CoRR*, abs/1603.05201, 2016.

Anna C. Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards understanding the invertibility of convolutional neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 2017.

Jörn-Henrik Jacobsen, Arnold W. M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *CoRR*, abs/1802.07088, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 2005.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Marcus Hutter and Marco Zaffalon. Distribution of mutual information from complete and incomplete data. *Computational Statistics & Data Analysis*, 48(3):633–657, 2005.

A Invertibility is Beneficial

We show in Proposition A.1 that the lower bound for the classification error is itself lower bounded by a constant, which is attained if F is invertible. Although the performance of the model also depends on the classifier w, our bound claims that an invertible feature map F could provide a better environment for the classifier w to perform well. Intuitively, invertibility preserves the information of the signal X as it flows through the neural network and reaches the classifier w; on the other hand, w potentially performs better on the input that preserves all information of the data compared to the one that doesn't.

Proposition A.1. *(Fano's Inequality) The classification error is lower bounded as follows:*

$$\mathbb{P}(Y \neq \widehat{Y}) \geq \frac{H(Y|F(X)) - \log(2)}{\log(|\mathcal{C}| - 1)} \ . \tag{9}$$

The lower bound satisfies

$$\frac{H(Y|F(X)) - \log(2)}{\log(|\mathcal{C}| - 1)} \geq \frac{H(Y|X) - \log(2)}{\log(|\mathcal{C}| - 1)} \tag{10}$$

for all F, and the equality is attained if F is invertible.

Let $Z = F(X)$ and the machinery of deep learning can be decribed by the following Markov Chain:

$$Y \to X \to Z \to \widehat{Y} \ . \tag{11}$$

Lemma A.1 is a technical result that helps to prove Proposition A.1. The information $Z = F(X)$ has about the true labels Y is maximized when F is invertible, which is beneficial in the sense that the key information influential for classification can be well preserved.

Lemma A.1 (Chain Rule)**.** *Given the Markov Chain assumption equation 11, we have*

$$\mathrm{I}(Y; \widehat{Y}) \leq \mathrm{I}(Y; Z) \leq \mathrm{I}(Y; X) \ , \tag{12}$$

and the second equality is attained if F is invertible.

Proof. We will only prove the second inequality and the first inequality follows by a similar argument. Consider the decomposition

$$\begin{aligned}
\mathrm{I}(Y; X, Z) &= \int_{\mathcal{X}} \sum_{y} \int_{\mathcal{Z}} p(x,y,z) \log \frac{p(x,y,z)}{p(y)p(x,z)} dxdz \\
&= \int_{\mathcal{X}} \sum_{y} \int_{\mathcal{Z}} p(x|y,z)p(y,z) \log \frac{p(x|y,z)p(y,z)}{p(y)p(x|z)p(z)} dxdz \\
&= \mathrm{I}(Y; Z) + \int_{\mathcal{X}} \sum_{y} \int_{\mathcal{Z}} p(x|y,z)p(y,z) \log \frac{p(x|y,z)}{p(x|z)} dxdz \\
&= \mathrm{I}(Y; Z) + \int_{\mathcal{X}} \sum_{y} \int_{\mathcal{Z}} p(x,y|z)p(z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} dxdz \\
&= \mathrm{I}(Y; Z) + \mathrm{I}(X; Y|Z) \ .
\end{aligned} \tag{13}$$

Similarly we obtain

$$\mathrm{I}(Y; X, Z) = \mathrm{I}(X; Y) + \mathrm{I}(Y; Z|X) \ . \tag{14}$$

equation 13 together with equation 14 yields

$$\mathrm{I}(Y; Z) + \mathrm{I}(X; Y|Z) = \mathrm{I}(X; Y) + \mathrm{I}(Y; Z|X) \ . \tag{15}$$

According to the Markov Chain setting, Y and Z are conditionally independent given X, hence $\mathrm{I}(Y; Z|X) = 0$; in addition, the mutual information $\mathrm{I}(X; Y|Z)$ is nonnegative. It follows from (15) that

$$\mathrm{I}(Y; Z) \leq \mathrm{I}(Y; X) \ . \tag{16}$$

□

Next we present a lower bound for the classification error. This lower bound is negatively related to the mutual information $I(Y; F(X))$, and it attains its minimum if F is invertible. Although the performance also depends on the classifier w, Proposition A.1 implies that an invertible feature map F allows more chances for the classifier w to perform well.

Proof of Proposition A.1. Consider the random variable E defined as:

$$E = \begin{cases} 1, & \text{if } Y \neq \widehat{Y} \\ 0, & \text{otherwise .} \end{cases} \tag{17}$$

By the Chain Rule following from similar arguments presented in Lemma A.1, we have

$$\begin{aligned} H(E, Y|\widehat{Y}) &= H(Y|\widehat{Y}) + H(E|Y, \widehat{Y}) \\ H(E, Y|\widehat{Y}) &= H(E|\widehat{Y}) + H(Y|E, \widehat{Y}) \ . \end{aligned} \tag{18}$$

Note that $H(E|Y, \widehat{Y}) = 0$, since the value of E is determined given the knowledge of $Y, \widehat{Y}$. It then follows that

$$\begin{aligned} H(Y|\widehat{Y}) &= H(E|\widehat{Y}) + H(Y|E, \widehat{Y}) \\ &\leq \log(2) + H(Y|E = 0, \widehat{Y})\mathbb{P}(E = 0) + H(Y|E = 1, \widehat{Y})\mathbb{P}(E = 1) \\ &= \log(2) + H(Y|E = 1, \widehat{Y})\mathbb{P}(Y \neq \widehat{Y}) \\ &\leq \log(2) + \log(|\mathcal{C}| - 1)\mathbb{P}(Y \neq \widehat{Y}) \ . \end{aligned} \tag{19}$$

On the other hand, Lemma A.1 shows that

$$H(Y) - H(Y|\widehat{Y}) = \mathrm{I}(Y; \widehat{Y}) \leq \mathrm{I}(Z; Y) = H(Y) - H(Y|\widehat{Z}) \ , \tag{20}$$

which gives

$$H(Y|Z) \leq H(Y|\widehat{Y}) \ . \tag{21}$$

Substitute it into (19) yields the result

$$H(Y|Z) \leq \log(2) + \log(|\mathcal{C}| - 1)\mathbb{P}(Y \neq \widehat{Y}) \ . \tag{22}$$

As for the second statement, Lemma A.1 shows that

$$H(Y) - H(Y|Z) = \mathrm{I}(Y; Z) \leq \mathrm{I}(Y; X) = H(Y) - H(Y|X) \ . \tag{23}$$

It then follows that,

$$H(Y|Z) = H(Y|F(X)) \geq H(Y|X) = 0 \ , \tag{24}$$

where the equality is attained if F is invertible. □

B Proof of Proposition 3.2

Proposition 3.2 establishes a connection between $\mathrm{I}(X, \widehat{Y})$ and the absolute value of the logits $|w^T F(X)|$ for the binary case. Intuitively, decreasing the confidence of the model on its predictions will decrease the mutual information $\mathrm{I}(X; \widehat{Y})$.

Proposition 3.2. $\mathrm{I}(X; \widehat{Y}) = \mathrm{I}(F(X); Y)$ *is well estimated by its empirical version(Monte-carlo approximation) with high probability, which shares the same unique (global) minimum with* $\sum_{k=1}^{n} |w^T F(x_k)|$ *at* $w^T F(x_k) = 0$*, for all* $k \in \{1, ..., n\}$.

Proof. The mutual information $\mathrm{I}(X; \widehat{Y})$ is given as

$$\mathrm{I}(X; \widehat{Y}) = \int_{\mathcal{X}} \sum_{\widehat{y} \in \mathcal{C}} p(x, \widehat{y}) \log\left(\frac{p(x, \widehat{y})}{p(x)p(\widehat{y})}\right) dx \ . \tag{25}$$

Apply the assumption **(II)**, the marginal distribution of $\widehat{Y}$ is uniformly distributed:

$$\frac{p(x,\widehat{y})}{p(x)p(\widehat{y})} = \frac{p(\widehat{y}|x)}{p(\widehat{y})} = 2p(\widehat{y}|x)\ . \tag{26}$$

Substituting (26) into (25) yields

$$\begin{aligned}
\mathrm{I}(X;\widehat{Y}) &= \int_{\mathcal{X}} p(x) \sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x) \log\left(\frac{p(x,\widehat{y})}{p(x)p(\widehat{y})}\right) dx \\
&= \int_{\mathcal{X}} p(x) \sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x) \log(2p(\widehat{y}|x)) dx\ .
\end{aligned} \tag{27}$$

According to the Hoeffding's inequality for bounded random variables [Proposition 2.2.6, Vershynin [2018]], let M, m denote upper and lower bounds of the integrand of (27) correspondingly, we have

$$\mathbb{P}\left\{\left|\sum_{k=1}^{n}\left(\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x_k)\log(2p(\widehat{y}|x_k)) - \mathrm{I}(X;\widehat{Y})\right)\right| \geq t\right\} \leq 2e^{-\frac{2t^2}{n(M-m)^2}}\ . \tag{28}$$

Equivalently, with probability at least $1-\delta$,

$$\left|\frac{\sum_{k=1}^{n}\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x_k)\log(2p(\widehat{y}|x_k))}{n} - \mathrm{I}(X;\widehat{Y})\right| < \sqrt{-\frac{\log(\frac{\delta}{2})(M-m)^2}{2n}}\ . \tag{29}$$

Here $\sum_{k=1}^{n}\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x_k)\log(2p(\widehat{y}|x_k))/n$ is a Monte carlo estimation of RHS of $\mathrm{I}(X;\widehat{Y})$.

Recall that, for the binary case $p(\widehat{y}|x) = p(\widehat{y}|F(x))$ can expressed as

$$\begin{aligned}
p(\widehat{y}=1|F(x)) &= \sigma(w^T F(x)) \\
p(\widehat{y}=-1|F(x)) &= 1-\sigma(w^T F(x))\ .
\end{aligned} \tag{30}$$

Then we have

$$\begin{aligned}
\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|F(x_k))\log(2p(\widehat{y}|F(x_k))) &= \sigma(w^T F(x))\log(2\sigma(w^T F(x))) \\
&\quad + (1-\sigma(w^T F(x)))\log(2-2\sigma(w^T F(x))))\ ,
\end{aligned} \tag{31}$$

which is bounded by $[0, \log(2)]$.

Take $M = \log(2), m = 0$, we have

$$\mathrm{I}(X;\widehat{Y}) = \frac{\sum_{k=1}^{n}\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x_k)\log(2p(\widehat{y}|x_k))}{n} + \mathrm{O}\left(\sqrt{-\frac{\log(\frac{\delta}{2})\log(2)^2}{2n}}\right)\ . \tag{32}$$

hold with probability at least $1-\delta$.

The conclusion follows from the fact that$\sum_{k=1}^{n}\sum_{\widehat{y}\in\mathcal{C}} p(\widehat{y}|x_k)\log(2p(\widehat{y}|x_k))/n$ has a unique global minimum at $w^T F(x_k) = 0$ for each x_k. □

C Proof of Proposition 3.3

Consider the training samples $\{(x_k, y_k)\}_{k=1:n}$, each x_k is fed into a deep probabilistic model which outputs probability densities and predicts $\widehat{y}_k$, a realization of the prediction $\widehat{Y}$. Let $\mathcal{C} = \{\pm 1\}$ be the binary class and n_y be the counts of observed occurrences of k satisfying $y_k = y \in \mathcal{C}$, then $n = \sum_{y\in\mathcal{C}} n_y = \sum_y n_y$, where we omit the range over $\mathcal{C}$ for convenience.

We denote the true joint probability with $\pi_{y\widehat{y}} = p(y, \widehat{y})$, the marginal probabilities with $\pi_{y+} = \sum_{\widehat{y}} \pi_{y\widehat{y}}$ and $\pi_{+\widehat{y}} = \sum_{y} \pi_{y\widehat{y}}$. The mutual information $\mathrm{I}(Y; \widehat{Y})$ can be expressed as

$$\mathrm{I}(Y; \widehat{Y}) = \mathrm{I}(\pi) = \sum_{y\widehat{y}} \pi_{y\widehat{y}} \log \left(\frac{\pi_{y\widehat{y}}}{\pi_{y+}\pi_{+\widehat{y}}} \right) . \tag{33}$$

Our empirical mutual information $\mathrm{I}(\widehat{\pi})$ is defined as

$$\mathrm{I}(\widehat{\pi}) = \sum_{y\widehat{y}} \widehat{\pi}_{y\widehat{y}} \log \left(\frac{\widehat{\pi}_{y\widehat{y}}}{\widehat{\pi}_{y+}\widehat{\pi}_{+\widehat{y}}} \right) , \tag{34}$$

where $\widehat{\pi}_{y\widehat{y}} = \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(\widehat{y} w^T F(x_i))$.

Proposition 3.3 establishes a connection between $\mathrm{I}(Y; \widehat{Y})$ and the cross-entropy objective for the binary case. Intuitively, increasing the confidence of the model on its correct predictions will establish a more deterministic relationship between Y and $\widehat{Y}$ and thus increase the mutual information $\mathrm{I}(Y; \widehat{Y})$.

Proposition 3.3. *$\mathrm{I}(Y; \widehat{Y})$ is well estimated by $\mathrm{I}(\widehat{\pi})$ with high probability, which shares the same unique (global) maximum with $\sum_{k=1}^{n} \log \sigma(y_k w^T F(x_k))$ at $y_k w^T F(x_k) \to \infty$ given that $y_k w^T F(x_k) > \frac{1}{2}$, for all $k \in \{1, ..., n\}$.*

Proposition 3.3 follows from Proposition C.1 and Proposition C.2, where Proposition C.1 shows that $\mathrm{I}(Y; \widehat{Y})$ is well approximated by $\mathrm{I}(\widehat{\pi})$ with high probability and Proposition C.2 shows the remaining claims in Proposition 3.3.

As shown in Lemma C.1, $\widehat{\pi}_{y\widehat{y}} = \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(\widehat{y} w^T F(x_i))$ is an unbiased estimate of $\pi_{y\widehat{y}}$. Here σ to represent the sigmoid function defined by $\sigma(x) = e^x / (e^x + 1)$. By leveraging the concentration property of bounded variables, *i.e.,* $\sigma(\widehat{y} w^T F(x_i))$, the estimation error can be bounded with high probability (Lemma C.2).

Lemma C.1. *The empirical joint probability, defined as*

$$\widehat{\pi}_{y\widehat{y}} = \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(\widehat{y} w^T F(x_i)) , \tag{35}$$

is an unbiased estimate of the true joint distribution $\pi_{y\widehat{y}}$.

Lemma C.2. *With probability at least $1 - \delta$, we have*

$$|\Delta_{y\widehat{y}}| := |\widehat{\pi}_{y\widehat{y}} - \pi_{y\widehat{y}}| \leq \frac{1}{2} \sqrt{\frac{\log(\frac{2}{\delta}) \min\{\frac{1}{4} \sup_x (w^T F(x))^2, 1\}}{2n_y}} . \tag{36}$$

Proposition C.1. *With probability at least $1 - \delta$,*

$$\mathrm{I}(Y; \widehat{Y}) = \mathrm{I}(\pi) = \mathrm{I}(\widehat{\pi}) - \mathrm{O}\left(\sqrt{\frac{\log(\frac{8}{\delta}) \min\{\frac{1}{4} \sup_x (w^T F(x))^2, 1\}}{n}} \right) . \tag{37}$$

Proof. To estimate the empirical mutual information given fixed samples, we use the approach by Hutter and Zaffalon [2005]. In particular, taylor expansion gives

$$\mathrm{I}(\widehat{\pi}) = \mathrm{I}(\pi) + \sum_{y\widehat{y}} \log \left(\frac{\pi_{y\widehat{y}}}{\pi_{y+}\pi_{+\widehat{y}}} \right) \Delta_{y\widehat{y}} + \mathrm{O}(\Delta^2) , \tag{38}$$

where $\Delta_{y\widehat{y}} = \widehat{\pi}_{y\widehat{y}} - \pi_{y\widehat{y}}$. Hence, Eq (38) together with Lemma C.2 yield, with probability exceeding $1 - |\mathcal{C}|^2 \delta$,

$$\mathrm{I}(\pi) = \mathrm{I}(\widehat{\pi}) - \mathrm{O}\left(\sqrt{\frac{\log(\frac{2}{\delta}) \min\{\frac{1}{4} \sup_x (w^T F(x))^2, 1\}}{n}} \right) . \tag{39}$$

Notice that, in the binary case the cardinality $|\mathcal{C}| = 2$.

□

Next we prove the intermediate results, Lemmas C.1 and C.2.

Proof of Lemma C.1. Direct derivation on the true joint distribution $\pi_{y\widehat{y}}$ gives

$$\begin{aligned}\pi_{y\widehat{y}} = p(y,\widehat{y}) = \int_{\mathcal{X}} p(y,\widehat{y},x)dx &= \int_{\mathcal{X}} p(y|\widehat{y},x)p(\widehat{y}|x)p(x)dx \\ &= \int_{\mathcal{X}} p(y|x)p(\widehat{y}|x)p(x)dx \\ &= \int_{\mathcal{X}} p(x|y)\sigma(\widehat{y}w^T F(x))p(y)dx \ . \end{aligned} \tag{40}$$

Given assumption **(I)** which states that the marginal density of Y is uniform over $\mathcal{C}$, for every true label $y \in \mathcal{C}$ we have $p(y) = \frac{1}{2}$.

We can therefore rewrite (40) as

$$\pi_{y\widehat{y}} = \frac{1}{2}\int_{\mathcal{X}} \sigma(\widehat{y}w^T F(x))p(x|y)dx \ . \tag{41}$$

According to assumption **(II)**, $p(x|y)$ is a probability density over space of signal x with true label y.

The Monte Carlo estimation of (41) gives the empirical joint probability which is unbiased:

$$\widehat{\pi}_{y\widehat{y}} = \frac{1}{2n_y}\sum_{i=1}^{n_y} \sigma(\widehat{y}w^T F(x_i)) \ . \tag{42}$$

□

Proof of Lemma C.2. Again, by leveraging the Hoeffding's inequality for bounded random variables [Proposition 2.2.6 of Vershynin [2018]], we have

$$\mathbb{P}\left\{\left|\sum_{k_y=1}^{n_y}\left(\sigma(\widehat{y}w^T F(x_{k_y})) - \mathbb{E}_{X_y}\left[\sigma(\widehat{y}w^T F(X_y))\right]\right)\right| \geq t\right\} \leq 2e^{-\frac{2t^2}{n_y(M-m)^2}} \ , \tag{43}$$

where X_y is the data random variable whose true label is y.

Equivalently, with probability at least $1-\delta$,

$$\left|\frac{\sum_{k_y=1}^{n_y}\sigma(\widehat{y}w^T F(x_{k_y}))}{n_y} - \mathbb{E}_{X_y}\left[\sigma(\widehat{y}w^T F(X_y))\right]\right| < \sqrt{\frac{\log(\frac{2}{\delta})(M-m)^2}{2n_y}} \ , \tag{44}$$

where M, m are upper and lower bounds of random variable $\sigma(\widehat{y}w^T F(X_y))$, respectively.

Substitute (41) and (42) into (44), with probability at least $1-\delta$,

$$|\widehat{\pi}_{y\widehat{y}} - \pi_{y\widehat{y}}| \leq \frac{1}{2}\sqrt{\frac{\log(\frac{2}{\delta})(M-m)^2}{2n_y}} \ . \tag{45}$$

To estimate the upper and lower bounds M, m for $\sigma(\widehat{y}w^T F(X))$, we use the Taylor's theorem:

$$\begin{aligned}\sup_x \sigma(\widehat{y}w^T F(x)) &= \sigma(0) + \sup_x \sigma'(c)(\widehat{y}w^T F(x)) \\ &\leq \frac{1}{2} + \frac{1}{4}\sup_x |w^T F(x)| \end{aligned} \tag{46}$$

and

$$\begin{aligned}\inf_x \sigma(\widehat{y}w^T F(x)) &= \sigma(0) + \inf_x \sigma'(c)(\widehat{y}w^T F(x)) \\ &\geq \frac{1}{2} - \frac{1}{4}\sup_x |w^T F(x)| \ , \end{aligned} \tag{47}$$

given that the derivative of sigmoid function is bounded by $\frac{1}{4}$.

It follows that

$$\begin{aligned} |M - m| &= \sup_x \sigma(\widehat{y} w^T F(x)) - \inf_x \sigma(\widehat{y} w^T F(x)) \\ &\le \frac{1}{2} \sup_x |w^T F(x)| \; . \end{aligned} \tag{48}$$

Also notice that M, m are the bounds for sigmoid function, so their difference is at most 1.

From derivations above, we can rewrite (45) as

$$|\widehat{\pi}_{y\widehat{y}} - \pi_{y\widehat{y}}| \le \frac{1}{2}\sqrt{\frac{\log(\frac{2}{\delta}) \min\{\frac{1}{4}\sup_x (w^T F(x))^2, 1\}}{2n_y}} \; , \tag{49}$$

and the lemma follows. □

Proposition C.2. *The empirical mutual information* $\mathrm{I}(\widehat{\pi})$ *shares the same unique (global) maximum with* $\sum_{k=1}^{n} \log \sigma(y_k w^T F(x_k))$ *as* $y_k w^T F(x_k) \to \infty$ *given that* $y_k w^T F(x_k) > \frac{1}{2}$*, for all* $k \in \{1, ..., n\}$.

Proof. The empirical information $\mathrm{I}(\widehat{\pi})$ is defined by

$$\mathrm{I}(\widehat{\pi}) = \sum_{ij} \widehat{\pi}_{y\widehat{y}} \log\left(\frac{\widehat{\pi}_{y\widehat{y}}}{\widehat{\pi}_{y+}\widehat{\pi}_{+\widehat{y}}}\right) \; , \tag{50}$$

where the empirical joint probability is given by

$$\widehat{\pi}_{y\widehat{y}} = \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(\widehat{y} w^T F(x_i)) \; . \tag{51}$$

It then follows that for any $y \in \mathcal{C}$,

$$\begin{aligned} \widehat{\pi}_{y+} &= \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(w^T F(X_i)) + \frac{1}{2n_y} \sum_{i=1}^{n_y} \sigma(-w^T F(X_i)) \\ &= \frac{1}{2n_y} \sum_{i=1}^{n_y} (\sigma(w^T F(X_i)) + \sigma(-w^T F(X_i))) \\ &= \frac{1}{2n_y} \sum_{i=1}^{n_y} 1 = \frac{1}{2} \; . \end{aligned} \tag{52}$$

In binary case it means that

$$\widehat{\pi}_{1(-1)} = \frac{1}{2} - \widehat{\pi}_{11} \; , \quad \widehat{\pi}_{(-1)1} = \frac{1}{2} - \widehat{\pi}_{(-1)(-1)} \tag{53}$$

and the empirical mutual information can decomposed as

$$\begin{aligned} \mathrm{I}(\widehat{\pi}) = \widehat{\pi}_{11} \log\left(\frac{\widehat{\pi}_{11}}{\widehat{\pi}_{11} + \frac{1}{2} - \widehat{\pi}_{(-1)(-1)}}\right) + \left(\frac{1}{2} - \widehat{\pi}_{11}\right) \log\left(\frac{\frac{1}{2} - \widehat{\pi}_{11}}{\frac{1}{2} - \widehat{\pi}_{11} + \widehat{\pi}_{(-1)(-1)}}\right) + \\ \left(\frac{1}{2} - \widehat{\pi}_{(-1)(-1)}\right) \log\left(\frac{\frac{1}{2} - \widehat{\pi}_{(-1)(-1)}}{\widehat{\pi}_{11} + \frac{1}{2} - \widehat{\pi}_{(-1)(-1)}}\right) + \widehat{\pi}_{(-1)(-1)} \log\left(\frac{\widehat{\pi}_{(-1)(-1)}}{\frac{1}{2} - \widehat{\pi}_{11} + \widehat{\pi}_{(-1)(-1)}}\right) + \log(2) \; . \end{aligned} \tag{54}$$

We differentiate (54) with respect to $\widehat{\pi}_{11}$ and $\widehat{\pi}_{(-1)(-1)}$, and calculate the critical points over the domain $[0, \frac{1}{2}]$ for both variables, which gives

$$\widehat{\pi}_{11} = \widehat{\pi}_{(-1)(-1)} = \frac{1}{4} \; . \tag{55}$$

Observe that (55) is a global minimum over $[0, \frac{1}{2}] \times [0, \frac{1}{2}]$. Since this is the unique critical point where the derivative vanishes, the global maximums can only be obtained on the boundaries. In particular, if we restrict $(\widehat{\pi}_{11}, \widehat{\pi}_{(-1)(-1)})$ on $[\frac{1}{4}, \frac{1}{2}] \times [\frac{1}{4}, \frac{1}{2}]$, (54) is a strictly increasing function over $\widehat{\pi}_{11}, \widehat{\pi}_{(-1)(-1)}$ and the unique global maximum is obtained at

$$\widehat{\pi}_{11} = \widehat{\pi}_{(-1)(-1)} = \frac{1}{2} \, . \tag{56}$$

The proposition follows from the definition (51) of $\widehat{\pi}_{y\widehat{y}}$ that (56) is only approached when $y_k w^T F(x_k) \to \infty$, for all $k \in \{1, ..., n\}$.

□