Extracting robust and accurate features via a robust information bottleneck

Ankit Pensia* Varun Jog† Po-Ling Loh‡§

ankitp@cs.wisc.edu vjog@ece.wisc.edu ploh@stat.wisc.edu

Departments of Computer Science*, Electrical & Computer Engineering†, and Statistics‡

University of Wisconsin-Madison

Department of Statistics, Columbia University§

Abstract

We propose a novel strategy for extracting features in supervised learning that can be used to construct a classifier which is more robust to small perturbations in the input space. Our method builds upon the idea of the information bottleneck, by introducing an additional penalty term that encourages the Fisher information of the extracted features to be small when parametrized by the inputs. We present two formulations where the relevance of the features to output labels is measured using either mutual information or MMSE. By tuning the regularization parameter, we can explicitly trade off the opposing desiderata of robustness and accuracy when constructing a classifier. We derive optimal solutions to both robust information bottleneck formulations when the inputs and outputs are jointly Gaussian, proving that the optimally robust features are also jointly Gaussian in this setting. We also propose methods for optimizing variational bounds on the robust information bottleneck objectives in general settings using stochastic gradient descent, which may be implemented efficiently in neural networks. Our experimental results for synthetic and real data sets show that the proposed feature extraction methods indeed produce classifiers with increased robustness to perturbations.

1 Introduction

Over the past decade, deep learning algorithms have revolutionized modern machine learning, achieving superhuman performance in several diverse scenarios such as image classification [1], machine translation [2], and strategy games [3]. These algorithms are distinguished by their ability to solve complex problems by processing massive data sets efficiently with the help of large-scale computing power. On the other hand, as deep learning algorithms are gradually adopted in high-stakes applications such as autonomous driving, disease diagnosis, and legal analytics, it has become increasingly important to ensure their interpretability [4], fairness [5], and security [6]. In particular, the lack of "robustness" of neural networks (explained in more detail below) has become a significant concern.

It was observed in Szegedy et al. [7] that the high accuracy of trained neural networks may be compromised under small (nearly imperceptible) changes in the inputs [8]. Perhaps more alarmingly, empirical studies suggest the existence of certain "universal adversarial perturbations" that can thwart any neural network architecture [9]. Following these observations, the research area of *robust machine learning* has seen tremendous activity in recent years. Briefly stated, research in robust machine learning considers various threat models and proposes strategies to attack and defend neural networks. More recently, various researchers have proposed certifiable defenses; i.e., defenses that are provably robust against *all* possible adversaries. We briefly describe some relevant work below.

Data augmentation is a popular method for increasing the robustness of neural networks [10, 1, 11], wherein a training data set is enlarged using artificial training points constructed with small perturbations of the inputs. Some authors [12, 13] suggest augmenting the data set by carefully chosen perturbation directions that approximate worst-case perturbations, the latter of which are infeasible to compute exactly in high dimensions. Another approach is to smooth the decision boundaries of a trained neural network using a preprocessing step such as randomized smoothing [14, 15, 16], which may lead to computable certificates on the robustness of a neural network classifier [17, 18, 19, 15, 20]. On the other

hand, many of the methods for defending against adversarial attacks which initially showed promise have subsequently been broken [21].

Recent work suggests that high accuracy and high robustness may in fact be in conflict with each other [22, 23], which may even be a fundamental defect of any classifier [24, 25, 26, 27]. This has suggested certain tradeoffs between maximally robust and maximally accurate classification: If it is desirable to train a classifier which is robust to small perturbations in the inputs, it may be necessary to forego the level of accuracy obtained when such a restriction is not present.

Tsipras et al. [23] and Ilyas et al. [28] have suggested a dichotomy between "robust" and "non-robust" features. Although precise definitions of "robust features" are still elusive—with this paper providing a possible interpretation according to the magnitude of a conditional Fisher information term—intuitively, a robust feature is a function of the input that is robust to small perturbations of the input. Both robust and non-robust features might be useful for classification, but an adversary may perturb the input to render non-robust features irrelevant for classification. One approach for building a robust classifier would therefore be to train a classifier which only operates on the robust features.

Motivated by these lines of work, we propose a new method—the robust information bottleneck—for extracting features that are simultaneously robust and useful. We characterize robustness in terms of an appropriately defined notion of Fisher information, and quantify usefulness in terms of mutual information and estimation error. Our method is heavily inspired by the information bottleneck objective of Tishby et al. [29], which we will review in detail. A crucial difference between Tishby et al.'s objective and ours is the quantities being traded-off in the extracted features: accuracy and compression in Tishby et al. vs. accuracy and robustness in our work.

The explicit trade off between robustness and accuracy in the robust information bottleneck is reminiscent of the work of Zhang et al. [26], who propose a different regularizer to promote robustness at the cost of accuracy. Also worth mentioning is the work of Achille and

Soatto [30, 31], where both mutual information and Fisher information were used to measure the degree to which the parameters of a learning algorithm "memorize" the training data set. In this paper, we are concerned with the output of the algorithm; i.e., the features, rather than the parameters.

The remainder of the paper is organized as follows: In Section 2, we review the information bottleneck methodology and introduce the versions of the robust information bottleneck objective that will be studied in this paper. In Section 3, we derive properties of the proposed Fisher information regularizer, which encourages robustness of the extracted features. In Section 4, we rigorously derive solutions to the robust information bottleneck objective when the inputs and outputs are jointly Gaussian, and interpret the results. In Section 5, we present a variational optimization framework for obtaining approximate solutions in the case of general distributions. We provide simulation results on synthetic and real data sets in Section 6, and conclude with a discussion in Section 7.

Notation: Random variables will be denoted by capital letters (X, Y, Z), their support will be denoted by calligraphic letters $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$, and their densities will be denoted via subscripts (p_X, p_Y, p_Z). Random vectors will be written as column vectors, and when $X = (X_1, \ldots, X_n)^\top$ and $Y = (Y_1, \ldots, Y_m)^\top$, we will denote $(X, Y) = (X^\top, Y^\top)^\top$. For a vector $v \in \mathbb{R}^p$, we write $v^\downarrow$ to denote the vector with components rearranged in decreasing order. For two vectors $v, w \in \mathbb{R}^p$, we write $v^\downarrow \preceq w^\downarrow$ to indicate that $w^\downarrow$ majorizes $v^\downarrow$, meaning that for all $1 \leq k \leq p$, we have $\sum_{i=1}^k (v^\downarrow)_i \leq \sum_{i=1}^k (w^\downarrow)_i$, and $\sum_{i=1}^p (v^\downarrow)_i = \sum_{i=1}^p (w^\downarrow)_i$. We use $[n]$ to denote the set $\{1, \ldots, n\}$.

For a matrix $A \in \mathbb{R}^{p \times p}$, let $\lambda(A)$ denote the (multi)set of eigenvalues of A. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues, respectively. Let $\|A\|_F$ denote the Frobenius norm. Let $\mathrm{diag}(a_1, \ldots, a_p)$ denote the $p \times p$ diagonal matrix with $(a_1, \ldots, a_p)$ on the diagonal. We write I_d to denote the $d \times d$ identity matrix. In the linear algebraic statements throughout the paper, we will generally consider the singular value decomposition (SVD) to

be the "thin SVD." We write $\mathrm{Cov}(X)$ to denote the covariance matrix of a random vector X; and when Y is another random vector, we write $\mathrm{Cov}(X, Y)$ to denote the covariance matrix of the concatenated vector (X, Y). We write $\mathrm{Cov}(X|Y)$ to denote the average conditional covariance matrix of X, where the integral is taken with respect to the density of Y. We will denote the entropy of a discrete random variable X by $H(X)$, and the differential entropy of a continuous random variable X by $H(X)$, as well.

2 Problem formulation

Consider a data set $(X, Y) \sim p_{XY}$, where X is thought of as a sample corresponding to a label Y. The information bottleneck theory proposed in Tishby et al. [29] is a variational principle used for extracting as much relevant information about Y from X as possible, while achieving the largest possible compression of X. Using mutual information to measure "relevance" and "compression," Tishby et al. [29] proposed the optimization problem

$$\inf_{p_{T|X}(\cdot|\cdot)} \{I(T; X) - \gamma I(T; Y)\}. \tag{1}$$

The extracted feature, denoted by T, is a random function of X generated by the kernel $p_{T|X}$. Since it does not directly depend on Y, we have the Markov chain $Y \to X \to T$. The parameter $\gamma > 0$ trades off compression and relevance of the extracted features T. The information bottleneck principle has subsequently been applied to learning problems [32, 33, 34]. More recently, information bottleneck theory has also been used to gain insight into the training of deep neural networks. By measuring the information content of different layers in a network, it was observed that layers in a neural network undergo two separate phases, one consisting of a *memorization phase* where both $I(T; X)$ and $I(T; Y)$ increase, and a *compression phase* where $I(T; X)$ decreases while $I(T; Y)$ continues to increase [35, 36].

Broadly speaking, a "bottleneck" formulation trades off two quantities; in the information bottleneck, these quantities are relevance and compression, each measured using mutual

information. In this paper, we seek a formulation that trades off relevance and robustness. Depending on the specific learning problem under consideration, one may measure relevance and robustness using variety of metrics. We present two natural formulations below.

2.1 Measuring relevance

In the information bottleneck formulation, relevance is captured by the term $I(Y;T)$. Apart from mutual information being a natural quantity to consider, we may also justify $I(Y;T)$ via results such as Feder and Merhav [37, Theorem 1], which bounds the optimal classification error in terms of $I(Y;T)$. Additional discussion concerning the suitability of $I(Y;T)$ may be found in Shamir et al. [38].

As an alternative to mutual information, we will measure relevance via the minimum MSE for a predictor of Y constructed using T: $\mathsf{mmse}(Y|T) = \mathbb{E}[(Y - \mathbb{E}(Y|T))^2] = \operatorname{tr}(\operatorname{Cov}(Y|T))$. This notion is particularly useful when Y takes a continuum of values as opposed to a finite number of categories, and the goal is to estimate Y rather than pinpoint Y exactly.

2.2 Measuring robustness

Intuitively, a feature T is robust if small perturbations in X do not change the distribution of T significantly. We may think of the distribution of T as being parametrized by X. The sensitivity (being the opposite of robustness) of T to X may then be measured using the (statistical) Fisher information $\Phi(T|X)$, given below:

$$\Phi(T|X) = \int_{\mathcal{X}} \left(\int_{\mathcal{T}} \left\| \nabla_x \log p_{T|X}(t|x) \right\|_2^2 p_{T|X}(t|x) dt \right) p_X(x) dx := \int_{\mathcal{X}} \Phi(T|X=x) p_X(x) dx.$$

Under mild regularity conditions on the densities of X and T, we have $\Phi(T|X) = J(X|T) - J(X)$, where $J(X) = \mathbb{E}\left[\| \nabla_x \log p_X(x) \|_2^2 \right]$ and $J(X|T) = \int p_T(t) \left(\int \| \nabla_x \log p_{X|T}(x|t) \|_2^2 p_{x|t}(x|t) dx \right) dt$ (cf. Appendix E). The quantity $J(\cdot)$ is often called the information theorist's Fisher information, which is different from the statistical Fisher information $\Phi(\cdot|\cdot)$.

Naturally, Fisher information is not the only measure of robustness (or sensitivity) one may use. As we will show in Section 3, however, the Fisher information satisfies several properties which make it an attractive measure of sensitivity.

2.3 Robust information bottleneck objective

Since we want to extract features that are simultaneously relevant and robust, we define the features determined by the robust information bottleneck to be the optimum of

$$\inf_{p_{T|X}(\cdot|\cdot)} \{\mathsf{mmse}(Y|T) + \beta\Phi(T|X)\}, \quad \text{or} \tag{2}$$

$$\inf_{p_{T|X}(\cdot|\cdot)} \{-I(T;Y) + \beta\Phi(T|X)\}, \tag{3}$$

depending on what notion of "relevance" is being employed.

2.4 Examples

Before proceeding further, we describe two examples in the case when the input distribution is a Gaussian mixture. We will illustrate the instantiation of the Fisher information term as a regularizer, and return to these examples in the simulations to follow in Section 6.1.

Suppose Y takes values $+\mathbf{1} := (1,1)^\top$ and $-\mathbf{1} := (-1,-1)^\top$, with probability $1/2$ each. Conditioned on $Y = +\mathbf{1}$, the distribution of X is $\mathcal{N}(+\mathbf{1}, \mathrm{Diag}(\sigma_1^2, \sigma_2^2))$; and conditioned on $Y = -\mathbf{1}$, the distribution of X is $\mathcal{N}(-\mathbf{1}, \mathrm{Diag}(\sigma_1^2, \sigma_2^2))$. (This is identical to an example considered in Ilyas et al. [28].)

Example 1. *In the first setting of interest, we will consider random features T parametrized by $w := (w_1, w_2)^\top \in \mathbb{R}^2$ as $T = w^\top X + \xi$, where $\xi \sim \mathcal{N}(0,1)$.*

Example 2. *We will also consider a setting where T is a binary feature taking values ± 1, following a logistic distribution with parameter w: $\mathbb{P}(T = 1|X = x) = \frac{1}{1+\exp(-x^\top w)}$.*

The following two lemmas derive convenient closed-form expressions for the Fisher information, without making any assumptions on the distribution of X. However, we will use them to analyze the settings of Examples 1 and 2, respectively, when X follows a Gaussian mixture, in which case it will be simpler to assess the quality of the extracted features.

Lemma 1. *Suppose $T = AX + \epsilon$, where $\epsilon \sim N(0, I)$. Then $\Phi(T|X) = \|A\|_F^2$, so adding a Fisher information penalty is in this case equivalent to ℓ_2-regularization.*

As Lemma 1 shows, the Fisher information directly encodes the signal-to-noise ratio (SNR) of the channel from X to T. If the SNR is low, small changes in X have less of an effect on the distribution of T, meaning the features are more robust. In addition to quantifying this insight, Lemma 1 will be useful for our calculations later. The proof is contained in Appendix A.3. Now suppose we instead extract a binary feature. The proof of the following lemma is contained in Appendix A.4:

Lemma 2. *Suppose $T \in \{+1, -1\}$ is a binary feature such that $\mathbb{P}(T = 1 \mid X = x) = \frac{1}{1+\exp(-x^\top w)}$. Then $\Phi(T|X = x) = \|w\|_2^2 \cdot \mathbb{P}(T = 1 \mid X = x) \cdot \mathbb{P}(T = -1 \mid X = x)$.*

The empirical approximation to $\Phi(T|X) = \int \Phi(T|X = x)p_X(x)dx$ will be

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(T|X = x_i) = \frac{\|w\|_2^2}{n}\sum_{i=1}^{n} \mathbb{P}(T = 1|X = x_i) \cdot \mathbb{P}(T = -1|X = x_i). \tag{4}$$

We see from the formula in Lemma 2 that the Fisher penalty encourages more confident predictions. At the same time, the norm $\|w\|_2$ is encouraged to be small, relating to the discussion of the SNR following Lemma 1. In Section 6.1, we detail experiments that show how the Fisher penalty indeed encourages adversarial robustness.

Remark 1. *Note that the expression (4) has previously shown up in Wager et al. [39] as a "quadratic noising penalty," which is a first-order approximation of a regularizer obtained by adding noise to inputs when performing maximum likelihood estimation in logistic regression. This appears to be merely coincidental: A key difference in our setting is that the conditional*

probabilities appearing in the expression are for the feature T *conditioned on* $X = x_i$*, whereas the setting of Wager et al. [39] involves the probabilities of* Y *conditioned on* $X = x_i$*.*

3 Robustness properties of Fisher information

One of our motivations for using the Fisher information as a proxy for sensitivity is its amenability to analysis. Indeed, the Fisher information is a well-studied quantity in both information theory and estimation theory [40]. In this section, we collect several compelling reasons for using the Fisher information.

Relation to Cramér-Rao bound: The Cramér-Rao inequality [41] (or its generalization, the van Trees inequality) states that for a parameter $\Theta \sim p_\Theta$ and a family of distributions $p_{X|\Theta}$, we have $\mathsf{mmse}(\Theta|X) \geq \frac{1}{\Phi(X|\Theta)+J(\Theta)}$. In other words, high robustness (low $\Phi(X|\Theta)$) leads to lower accuracy (a larger lower bound on $\mathsf{mmse}(\Theta|X)$).

Scaling properties of Fisher information: Fisher information, mutual information, and MMSE are all invariant to changes in the scale (or indeed, any smooth bijective transformation) of T. The invariance under bijective transformations of T is critical—it would be unnatural to expect an extracted feature to become more (or less) robust by simply taking functions of that feature. The following standard lemma (see, for example Cover & Thomas [42]) makes this statement more precise. The proof is contained in Appendix A.1.

Lemma 3. *Let* $Y \to X \to T$ *be a Markov chain, such that* T *is an* $\mathbb{R}^d$*-valued random vector. Let* $f : \mathbb{R}^d \to \mathbb{R}^d$ *be a smooth bijection. Then the following equalities hold:*

1. $I(X;T) = I(X;f(T))$, $I(Y;T) = I(Y;f(T))$, $\mathsf{mmse}(Y|T) = \mathsf{mmse}(Y|f(T))$, *and* $\Phi(T|X) = \Phi(f(T)|X)$.

2. *If* $T = (T_1, T_2)$ *is such that* $T_2 \perp\!\!\!\perp (T_1, X, Y)$, *then* $I(X;T) = I(X;T_1)$, $I(Y;T) = I(Y;T_1)$, $\mathsf{mmse}(Y|T) = \mathsf{mmse}(Y|T_1)$, *and* $\Phi(T|X) = \Phi(T_1|X)$. *In other words, the*

independent component T_2 may be ignored when characterizing the optimal solution to the robust information bottleneck.

The standard information bottleneck formulation is invariant not only to transformations of T, but also to transformations of X and Y. This is not the case for the robust information bottleneck formulation, since $\Phi(T|X) \neq \Phi(T|f(X))$ in general. This is another attractive property of the robust information bottleneck formulation: If data are preprocessed so that the distribution $p_X(\cdot)$ is squeezed along a certain direction—for example, by multiplying X by a diagonal matrix $\mathrm{diag}(1, 1, \ldots, 1, \epsilon)$—the robustness with respect to perturbations along the final dimension should be reduced in comparison to the other directions. The standard information bottleneck formulation is blind to such transformations and extracts the same features regardless of transformations of X, whereas the robust information bottleneck adapts to the scaling of $p_X(\cdot)$.

Robustness implies compression: In the formulation (3), we do not have the $I(X;T)$ term that is present in the standard information bottleneck formulation. In the following lemma, we show that the $I(X;T)$ term is controlled by the $\Phi(T|X)$ term. Thus, the robust features learnt are also approximately compressed. A concern with this formulation could be that the value of $I(X;T)$ may be arbitrarily large at the optimum of formulation (3), leading to features that are not concise, although they may be robust. Our next lemma, proved in Appendix A.2, shows that this cannot happen and that robustness also implies compression:

Lemma 4. *Let $X \sim p_X$ be an $\mathbb{R}^p$-valued random variable, and let T be an extracted feature via the channel $p_{T|X}$. Then the following inequality holds:*

$$I(X;T) \leq H(X) - \frac{p}{2} \log \frac{2\pi e p}{\Phi(T|X) + J(X)}. \tag{5}$$

In particular, if $\Phi(T|X)$ is bounded from above, then so is $I(X;T)$.

Data processing inequality for Fisher information: Having extracted robust features T, we first note that *any* classifier that uses T to predict Y is guaranteed to be robust, as well. This supports the observation of Ilyas et al. [28], who show empirically that classifiers trained using "robust" features are also robust. Lemma 5 has previously appeared in Zamir [43], but we include a different proof in Appendix B.1.

Lemma 5. *Let $Y \to X \to T \to \widehat{Y}$ be a Markov chain. Here, we think of T as an extracted feature and $\widehat{Y}$ as a prediction of Y using T. The sensitivity of $\widehat{Y}$ to perturbations in X is measured by $\Phi(\widehat{Y}|X)$. Then $\Phi(\widehat{Y}|X) \leq \Phi(T|X)$. In other words, the output $\widehat{Y}$ is at least as robust as the extracted features T.*

Relation to mutual information: The following lemma, which follows from deBruijn's identity, is proved in Appendix B.2. It provides an interpretation of the term $\Phi(T|X)$ in terms of regularizing the effect of small perturbations to the mutual information:

Lemma 6. *Suppose $Z \sim N(0, I)$ is a standard normal random variable that is independent of (X, Y, T). Then $I(X;T) - I(X + \sqrt{\delta}Z;T) = \frac{\delta}{2}\Phi(T|X) + o(\delta)$.*

Lemma 6 shows that adding the Fisher information term $\Phi(T|X)$ encourages the mutual information between X and T to only change slightly under small Gaussian perturbations. Intuitively, this captures the idea that T cannot be too sensitive to X.

Relation to adversarial perturbations: Another way to interpret the Fisher information term is as follows: Let $\epsilon > 0$ and let u be a unit vector. An extracted feature T will be considered robust for a particular $X = x$ if the distributions $p_{T|X}(\cdot|X = x)$ and $p_{T|X}(\cdot|X = x + \epsilon u)$ are not too different for any choice of u and all small enough ϵ. The difference between these two distributions could be measured by a number of metrics, but we focus on the KL divergence here. Note that the KL divergence provides an upper bound on the total variation distance, and also bounds Wasserstein distances in certain special cases

[44]. (Wasserstein distance is the metric of study in recent work on distributional robustness [45, 46]; however, the goal of such studies is to directly learn neural network models that are distributionally robust to the inputs, rather than our intermediate step of extracting robust features.)

The proof of the following result is contained in Appendix B.3:

Lemma 7. *Let $\|u\|_2 = 1$. Let $x + \epsilon u$ be a small perturbation of x in the direction u. Then $D(p_{T|X=x+\epsilon u} \| p_{T|X=x}) = \frac{\epsilon^2}{2} \Phi(T|X = x) + o(\epsilon^2)$.*

Since the right-hand expression does not depend on the direction u, Lemma 7 shows that when x is perturbed arbitrarily in a ball of radius ϵ, the corresponding distribution of T lies in a KL-ball of radius $\frac{\epsilon^2}{2}\Phi(T|X = x)$ around the distribution $p_{T|X=x}(\cdot|X = x)$. Requiring $\Phi(T|X = x)$ to be small on average is equivalent to requiring $\Phi(T|X)$ to be small, so adding this term as a penalty encourages the algorithm to extract features that are robust to arbitrary ℓ_2-perturbations, on average. Note that this is identical to the objective of adversarial training in Madry et al. [13].

Finally, we show that the upper bound on the KL divergence in Lemma 7 can be translated into a direct guarantee on robustness. Consider a (deterministic) classifier $g : \mathcal{T} \to \mathcal{Y}$ that maps extracted features to a predicted label, and a classifier $f : \mathcal{X} \to \mathcal{Y}$ defined by $f(x) := \arg\max_y p_{T|X=x}(g(t) = y)$. In practice, we could approximate the value of f by generating random features according to the distribution $T|X = x$, applying the map g, and taking the majority vote over the result. The main idea, which is motivated by an argument found in Zhang and Liang [16], is to use the fact that an upper bound on the KL divergence between distributions implies an upper bound on total variation distance. Hence, if we have an input $x \in \mathcal{X}$ such that the *classification margin*, defined by $\text{margin}_f(x, y) := p_{T|X=x}(g(t) = y) - \max_{z \neq y} p_{T|X=x}(g(t) = z)$, is sufficiently large, then we should also have $f(x') = f(x)$ when x' is contained in a small ball around x.

For the result to follow, we assume that the $o(\epsilon^2)$ bound on the remainder in Lemma 7 is uniform over all choices of x, which holds if the third-degree differential of $p_{T|X=x}$ with

respect to x is uniformly bounded. The proof is provided in Appendix B.4.

Lemma 8. *For any $\epsilon, \eta > 0$, we have*

$$\mathbb{P}\Big(f(x') = f(x) \quad \forall x' \in B_\epsilon(x)\Big) \geq \mathbb{P}(x \in B^\eta) - \frac{\epsilon^2 \Phi(T|X) + o(\epsilon^2)}{\eta}, \tag{6}$$

where $B^\eta := \big\{x \in \mathcal{X} : \text{margin}_f(x, f(x)) > \sqrt{\eta}\big\}$.

The expression on the right side of inequality (6) provides a lower bound on the probability that a randomly chosen input is robust to perturbations of magnitude ϵ in any direction. Furthermore, the lower bound is higher when $\mathbb{P}(x \in B^\eta)$ is larger; i.e., the distribution on $\mathcal{X}$ is such that a larger fraction of points have high margin. To further interpret Lemma 8, suppose the distributions of X and $T|X$ are fixed, and consider the effect of adjusting the parameters ϵ or η. If we increase ϵ, the ball $B_\epsilon(x)$ in which the classifier is guaranteed to be robust becomes larger; however, the right side of inequality (6) decreases, leading to a weaker probabilistic guarantee. On the other hand, if we decrease η to increase the probability $\mathbb{P}(x \in B^\eta)$ appearing in the lower bound, the term $\frac{\epsilon^2 \Phi(T|X) + o(\epsilon^2)}{\eta}$ also increases. Thus, we see that tradeoffs exist in determining the optimal choices of both ϵ and η.

4 Jointly Gaussian variables

In general, it is impossible to obtain closed-form expressions for the solutions to the optimization problems (2) and (3). However, as in the case of the canonical information bottleneck, the optimization problems become more tractable when (X, Y) have a jointly Gaussian distribution [47]. In this section, we derive explicit formulas for the solutions to the optimization problems in order to develop some theoretical intuition for the similarities and differences between the robust information bottleneck formulations, and to verify that the extracted features are in fact meaningful in special cases. We will assume throughout this section that $\Sigma_x \succ 0$ and $\text{Cov}(Y|X) = \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} \succ 0$. We do not impose any restrictions on the

dimensionality of Y in relation to the dimensionality of X (which we will denote by p).

4.1 Information bottleneck formulation

We first study the information bottleneck formulation (3). The optimality of Gaussians in the standard information bottleneck formulation was proved in Globerson and Tishby [48]. The proof relies on the invariance of mutual information to linear bijective transformations and the optimality of Gaussians in the conditional entropy power inequality. The Fisher information term in our formulation precludes using such linear transformations or standard entropy inequalities. Instead, our proof uses a technique for establishing information inequalities pioneered by Geng and Nair [49] (see also Lieb [50] and Carlen [51]). Geng and Nair showed that it is enough to establish certain *subadditivity* relations for functionals in order to establish Gaussian optimality; this strategy has been used to prove a variety of entropy and information inequalities in the past few years [52, 53, 54, 55]. The proof of optimality is provided in detail in Appendix C, and we only provide a proof sketch here.

4.1.1 Optimality

Let (X_G, Y_G) be jointly Gaussian random variables. We express $Y_G = CX_G + \xi$, where ξ is independent of X, and rewrite the robust information bottleneck formulation as

$$\begin{aligned}&\sup_{p_{T|X_G}(\cdot|\cdot)} \{I(T;Y_G) - \beta\Phi(T|X_G)\} \\ &= \left[\sup_{p_{T|X_G}(\cdot|\cdot)} \Big\{ - H(Y_G|T) - \beta J(X_G|T)\Big\}\right] + H(Y_G) + \beta J(X_G).\end{aligned}$$

Since we are only concerned with the optimizing distribution $p_{T|X_G}$, we shall focus on the optimization problem in the square brackets. Consider the function f defined on the space of densities p_X over $\mathbb{R}^p$: $f(X) := -H(CX+\xi) - \beta J(X) := -H(Y) - \beta J(X)$, where we use $Y := CX + \xi$ to indicate the output channel that scales the input by C and adds Gaussian noise ξ to the scaled input. The upper-concave envelope of f, denoted by F is defined as

follows: For every distribution p, express p as a convex combination of distributions p_i such that $\sum_{i=1}^n p_i \lambda_i = p$, and define

$$F(p) = \sup_{n \geq 1} \sup_{\sum_{i=1}^n \lambda_i p_i = p} \sum_{i=1}^n \lambda_i f(p_i).$$

If T is a discrete random variable taking n values satisfying $p_{X|T=i} = p_i$ and $p_T(i) = \lambda_i$, then

$$\sum_{i=1}^n \lambda_i f(p_i) = f(X|T) := \sum_{i=1}^n p_T(i) f(X|T=i).$$

Thus, an equivalent way to think of the upper-concave envelope is through such auxiliary random variables T, as follows:

$$F(X) = \sup_{p_{T|X}} f(X|T) = \sup_{p_{T|X}(\cdot|\cdot)} \left\{ -H(Y|T) - \beta J(X|T) \right\},$$

where we allow $|\mathcal{T}|$ to be countably large for now. Note that the optimization problem in the square brackets above is equivalent to finding the optimizing T in the upper-concave envelope of f at the particular distribution X_G. Define a lifting of f to pairs of random variables (or equivalently, to probability distributions over $\mathbb{R}^p \times \mathbb{R}^p$): $f(X_1, X_2) := -H(Y_1, Y_2) - \beta J(X_1, X_2)$. As before, $Y_i = CX_i + \xi_i$ for $i \in \{1, 2\}$, where ξ_1 and ξ_2 are i.i.d. and independent of (X_1, X_2). Let $F(X_1, X_2)$ be the upper-concave envelope of $f(X_1, X_2)$. Our main result is the following subadditivity lemma (see Appendix C for a more accurate statement):

Lemma 9. *For any pair of random variables (X_1, X_2), we have $F(X_1, X_2) \leq F(X_1) + F(X_2)$.*

Next we prove the following lemma whose detailed proof is in Appendix C. (The proof of this lemma is lengthy, but the techniques employed are becoming relatively standard in the information theory literature.)

Lemma 10. *Consider the optimization problem $V(K) := \sup_{\mathrm{Cov}(X) \preceq K} f(X)$. The optimizer is a unique Gaussian random variable $X^* \sim \mathcal{N}(0, K^*)$, with $K^* \preceq K$. In particular, $f(X^*) = F(X^*) = V(K)$.*

Returning to the robust information bottleneck formulation for jointly Gaussian $(X_G, Y_G) \sim p_{X_G Y_G}$, let $\mathrm{Cov}(X_G) = K$. Let $X^* \sim \mathcal{N}(0, K^*)$ be the optimizer that achieves $V(K)$. Let $X' \perp\!\!\!\perp X$ be such that $X' \sim \mathcal{N}(0, K - K^*)$, so $X^* + X'$ has the same distribution as X_G. It is easy to check that $F(X_G) \geq f(X_G|X') = f(X^*) = V(K)$. However, we also have $F(X_G) \leq F(X^*) = V(K)$, where the first inequality comes from the fact that X^* maximizes both f and F. This shows that the optimal joint distribution (T, X_G) may be taken to be (X', X_G); i.e., $T = X'$. Since (X', X_G) are jointly Gaussian, this proves that it is enough to consider random variables T that are jointly Gaussian with X_G to solve the optimization problem (3). Note that the joint distribution of (X_G, T) has covariance $\begin{pmatrix} K & K - K^* \\ K - K^* & K - K^* \end{pmatrix}$. Thus, we may write $T = DX_G + N$, where $D = (K - K^*)K^{-1}$ and $N \sim \mathcal{N}(0, (K - K^*) - (K - K^*)K^{-1}(K - K^*))$. Since the scaling of T does not matter, we can also rewrite the optimizing T as $T = \tilde{D}X + \tilde{N}$, where

$$\tilde{D} = \left[(K - K^*) - (K - K^*)K^{-1}(K - K^*)\right]^{-1/2}(K - K^*)K^{-1}, \quad \text{and}$$
$$\tilde{N} \sim \mathcal{N}(0, I). \tag{7}$$

This completely identifies the optimal robust feature T in formulation (3).

4.1.2 Identity covariance

We now derive an explicit form of the optimal feature map in the case when Σ_x is a multiple of the identity. The proof of the following theorem is contained in Appendix D.1.

Theorem 1. *Suppose $\Sigma_x = \sigma_x^2 I$. Let $B = (\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1/2}\Sigma_{yx}\Sigma_x^{-1}$, and let $B = V\Lambda W^\top$ be the SVD. Let $\Lambda = \mathrm{diag}(\ell_1, \ldots, \ell_k)$, where the diagonal elements are sorted in decreasing order. For each $i \leq k$, define $d_i = \arg\min_{d \in [0,1]} \left\{ \frac{1}{2} \log\left(\frac{\sigma_x^2 d}{\ell_i} + 1 \right) + \frac{\beta}{\sigma_x^2 d} \right\}$, and let $D = \mathrm{diag}(d_1, \ldots, d_k)$. Let $\hat{U}$ be the permutation matrix which sorts the diagonal entries of D in increasing order, and let $U = W\hat{U}^\top$. An optimal feature map is then given by*

$T = \frac{1}{\sigma_x}(D^{-1} - I)^{1/2}U^\top X + \epsilon$, *where* $\epsilon \sim N(0, I_k)$.

To summarize, the optimal projection directions are given by a permutation/rearrangement of the right singular vectors W appearing in the SVD of B, together with appropriate rescalings obtained by optimizing univariate functions. As will be described in further detail in Section 4.3, this resembles the solution to the usual information bottleneck.

Remark 2. *As stated in Theorem 1, we can always find an optimal feature map into k dimensions, where $k = \mathrm{rank}(B)$. On the other hand, it is possible that the optimal feature map could be expressible in even fewer dimensions, e.g., if some of the d_i's are equal to 1.*

4.1.3 General covariance, small β

In the case when Σ_x is a general psd matrix, we can also derive a closed-form expression for the optimal feature map in settings where β is not too large. The following result is proved in Appendix D.2:

Theorem 2. *Suppose β is sufficiently small. Let $C = \Sigma_x^{-1/2}\Sigma_{xy}(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1/2}$ and suppose Σ_{xy} has full column rank. Consider the SVDs $C = W\Lambda V^\top$ and $(C^\top\Sigma_x^{-1}C)^{-1} = UDU^\top$. Define $\tilde{D} = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_k)$ to be a diagonal matrix with $\tilde{d}_i = \frac{1+\sqrt{1+4d_i/\beta}}{2d_i/\beta}$, where $D = \mathrm{diag}(d_1, \ldots, d_k)$. Let $S\Gamma S^\top$ be the SVD of $\Lambda V^\top U\tilde{D}^{-1}U^\top V\Lambda - I$. An optimal feature map is given by $T = \Gamma^{1/2}S^\top W^\top\Sigma_x^{-1/2}X + \epsilon$, where $\epsilon \sim N(0, I_k)$.*

As seen in the proof of the theorem, the required upper bound on β can be expressed in terms of the spectra of $(\Sigma_x, \Sigma_y, \Sigma_{xy})$.

Remark 3. *Note that the scenarios considered in Theorems 1 and 2 have a nonempty intersection—namely, when $\Sigma_x = \sigma_x^2 I$ and β is not too large. However, it is not entirely straightforward to compare the two expressions for T in the theorems, since the formulas for the two settings are derived using different proof strategies. Also note that our optimality proofs do not imply the uniqueness of an optimal feature map; indeed, as shown in Lemma 3 earlier, any bijective transformation of T leads to the same objective function value.*

4.2 MMSE formulation

We now consider the MMSE formulation (2). As in the previous section, we will provide a proof sketch for optimality that may be converted into a rigorous proof by following the steps in Appendix C. Moreover, we will derive the optimal form of T when (X_G, Y_G) are jointly Gaussian, expressed in terms of their associated covariance matrices.

4.2.1 Optimality

Using the chain rule for Fisher information in Appendix E, we have the following lemma:

Lemma 11. *The optimization objective in formulation (2) can be equivalently expressed as*

$$\min_{p_{T|X}(\cdot|\cdot)} \{\mathsf{mmse}(Y|T) + \beta\Phi(T|X)\} = \mathrm{tr}(\mathrm{Cov}(Y)) + \beta J(X) - \max_{p_{T|X}(\cdot|\cdot)} \{\mathrm{tr}(\mathrm{Cov}(Y|T)) - \beta J(X|T)\}.$$

Proof. Note that $\mathsf{mmse}(Y|T) = \mathrm{tr}\Big(\mathrm{Cov}(Y) - \mathrm{Cov}(Y|T)\Big)$ and $\Phi(T|X) = J(X|T) - J(X)$, where the second equation follows from Lemma 26. Since the joint distribution p_{XY} is fixed, we may remove the $\mathrm{tr}(\mathrm{Cov}(Y))$ and $J(X)$ terms from the optimization objective and arrive at the desired result. □

As in Section 4.1, we express Y_G as $Y_G = CX_G + \xi$. Define the function $f(X) := \mathrm{tr}(\mathrm{Cov}(CX + \xi)) - \beta J(X)$ on the space of densities p_X over $\mathcal{X}$, and let $F(\cdot)$ be the upper concave-envelope of f defined as $F(X) := \sup_{p_{T|X}(\cdot|\cdot)} \{\mathrm{tr}(\mathrm{Cov}(CX + \xi|T)) - \beta J(X|T)\}$. Define a lifting of F to pairs of random variables as

$$F(X_1, X_2) = \max_{p_{T|X_1,X_2}(\cdot|\cdot)} \{\mathrm{tr}(\mathrm{Cov}(CX_1 + \xi_1, CX_2 + \xi_2|T)) - \beta J(X_1, X_2|T)\}.$$

The main step is to establish a subadditivity lemma, analogous to Lemma 9:

Lemma 12. *The function F is subadditive, i.e., $F(X_1, X_2) \leq F(X_1) + F(X_2)$.*

The proof is essentially identical to that of Lemma 9, relying on the chain rule and data processing properties of Fisher information. We shall also omit the proof of the lemma below, since it follows the steps outlined in Geng and Nair [49], and also in our Appendix C:

Lemma 13. *Consider the optimization problem* $V(K) := \sup_{\mathrm{Cov}(X) \preceq K} f(X)$. *Then the optimizer of the above problem is a unique Gaussian random variable* $X^* \sim \mathcal{N}(0, K^*)$ *with* $K^* \preceq K$. *In particular,* $f(X^*) = F(X^*) = V(K)$.

Let $\mathrm{Cov}(X_G) = K$. Let X^* be the optimizer that achieves $V(K)$. Identifying the optimal T can now be done by following the exact same steps as in Section 4.1. In particular, we may take $T = DX_G + N$ (or $\tilde{D}X_G + \tilde{N}$), where D and N are as identified in Section 4.1.

4.2.2 Identity covariance

The following theorem derives a closed-form expression for the optimal feature map in the case when Σ_x is a multiple of the identity. The proof is contained in Appendix D.3.

Theorem 3. *Suppose* $\Sigma_x = \sigma_x^2 I$. *Let* $0 < \lambda_1 \leq \cdots \leq \lambda_k$ *denote the ordered nonzero eigenvalues of* $\Sigma_{xy}\Sigma_{yx}$. *Define* $U \in \mathbb{R}^{p \times k}$ *to be the matrix with columns equal to the ordered unit eigenvectors corresponding to* $(\lambda_1, \ldots, \lambda_k)$. *For* $1 \leq i \leq k$, *define* $d_i = \sqrt{\frac{\lambda_i}{\beta}} - 1$ *if* $\lambda_i \geq \beta$, *and* $d_i = 0$ *otherwise, and let* $D = \mathrm{diag}(d_1, \ldots, d_k)$. *Then an optimal choice of features is given by* $T = \frac{1}{\sigma_x} D^{1/2} U^\top X + \epsilon$, *where* $\epsilon \sim N(0, I_k)$.

4.3 Comparison between solutions

Now that we have derived explicit formulae for the optimal feature maps in several settings (Theorems 1, 2, and 3), it is instructive to compare the solutions. All of the feature maps may be expressed as $T = AX + \epsilon$, with $\epsilon \sim N(0, I_k)$, with $A = \tilde{D}\tilde{U}^\top \Sigma_x^{-1/2}$, where $\tilde{D} \in \mathbb{R}^{k \times k}$ is an appropriate diagonal matrix and $\tilde{U} \in \mathbb{R}^{p \times k}$ is a matrix with k orthonormal columns, taken from the spectral decomposition of some matrix function of $(\Sigma_x, \Sigma_y, \Sigma_{xy})$.

Digging a bit deeper, we see that the scaling matrix $\tilde{D}$ will generally depend critically on the value of the regularization parameter β. In particular, as $\beta \to \infty$, successive entries of $\tilde{D}$ will be truncated to 0 (e.g., $d_i \to 1$ in Theorem 1 and $d_i \to 0$ in Theorem 3). This same behavior is manifest in the canonical information bottleneck formulation for jointly Gaussian variables (cf. Theorem 3.1 of Chechik et al. [47]). The transition points are accordingly referred to as "critical points" for β. In our formulation, where the regularization parameter β trades off robustness and accuracy, it is natural that larger values of β will lead to zeroing out features (which are then very robust but completely useless in prediction); at the other extreme, small values of β lead to a full feature map which preserves all eigenvectors, regardless of the magnitude of the corresponding eigenvalues.

Turning to $\tilde{U}$, the matrix varies according to the robust bottleneck formulation. Comparing the two identity covariance cases, we see that for the information bottleneck formulation (Theorem 1), we are interested in the right singular vectors of $(\Sigma_y - \Sigma_{yx}\Sigma_{xy})^{-1/2}\Sigma_{yx}$. In the case of the MMSE formulation (Theorem 3), we are interested in the eigenvectors of $\Sigma_{xy}\Sigma_{yx}$.

For concreteness, we consider the following examples in which the optimal feature maps may be compared directly:

Example 3 (One-dimensional labels)**.** *First consider the case when Y is 1-dimensional. The formula given in Theorem 1 for the mutual information formulation when $\Sigma_x = \sigma_x^2 I$ results in the matrix B being a multiple of the vector Σ_{yx}, so that $U^T = W^T = \frac{\Sigma_{yx}}{\|\Sigma_{yx}\|_2}$. Thus, we have $T = \alpha_1 \Sigma_{yx} X + \epsilon$, where α_1 is a constant depending on β. Similarly, the formula in Theorem 3 for the MMSE formulation when Σ_x is a multiple of the identity implies that $U = \frac{\Sigma_{xy}}{\|\Sigma_{xy}\|_2}$, so that we also have $T = \alpha_2 \Sigma_{yx} X + \epsilon$ for a different constant α_2 depending on β. As remarked above, both α_1 and α_2 will become 0 when β exceeds an appropriate threshold, which differs depending on the formulation.*

In the case when Σ_x is arbitrary, Theorem 2 implies that the matrix C is a multiple of $\Sigma_x^{-1/2}\Sigma_{xy}$. Then $W = \frac{\Sigma_x^{-1/2}\Sigma_{xy}}{\|\Sigma_x^{-1/2}\Sigma_{xy}\|_2}$, and $T = \alpha_3 \Sigma_{yx}\Sigma_x^{-1} X + \epsilon$. Note in particular that the projection $\Sigma_{yx}\Sigma_x^{-1}X$ also arises as the solution to canonical correlation analysis (CCA).

Example 4 (Orthogonal covariance vectors)**.** *The formulas for optimal feature maps are somewhat more complicated when Y has more than one dimension. For illustration, we consider a somewhat contrived case where the columns of Σ_{xy} are orthogonal, so that the spectral decompositions are easier to analyze. (For example, this setup can be achieved by linearly transforming the data as in Chechik et al. [47].) First suppose $\Sigma_x = I$. Then the formula in Theorem 3 for the MMSE formulation implies that U is the matrix with columns equal to the renormalized columns of Σ_{xy}, and the feature map is given by $T = \widetilde{\Sigma}_{yx}X + \epsilon$, where we have used $\widetilde{\Sigma}_{yx}$ to denote a matrix with each row of Σ_{yx} scaled by a (possibly different) constant that depends on β.*

Turning to the formula in Theorem 1 for the mutual information formulation, we see that $B = (\Sigma_y - \Sigma_{yx}\Sigma_{xy})^{-1/2}\Sigma_{yx}$. In general, the columns of the matrix W may be somewhat different from the columns of Σ_{xy}; we will have $T = \widetilde{W}^T X + \epsilon$, where $\widetilde{W}$ again denotes a matrix with rescaled columns of W. However, note that the matrix $\Sigma_{yx}\Sigma_{xy}$ is diagonal, so if we further impose the constraint that Σ_y is a diagonal matrix, the columns of W are indeed rescaled versions of the columns of Σ_{xy}. (Similarly, note that if Σ_x is allowed to be an arbitrary matrix, the formula provided in Theorem 2 results in an optimal feature map which can look quite different from the projections involving rescaled columns of Σ_{xy}.)

5 Variational bounds

Although our work is motivated by robustness considerations in deep learning, the framework we have developed thus far does not involve any assumptions that the classifier we employ for predicting Y from T is a neural network. In this section, we see how properties of neural networks may be leveraged for the purpose of optimization.

The objectives (2) and (3) are intractable to minimize explicitly except in certain special cases, so we propose to minimize appropriate upper bounds. Inspired by a recent line of work on variational approximations to the information bottleneck objective [34], we describe

the upper bounds and a tractable optimization procedure that uses minibatch stochastic gradient descent. We shall restrict ourselves to kernels $p_{T|X}(\cdot|\cdot)$ that are parametrized by θ. Let $K \in \mathbb{N}$. Specifically, we consider $p_{T|X}(\cdot|x) = \mathcal{N}(\mu(x;\theta), \Sigma(x;\theta))$, where $\mu(\cdot;\theta)$ and $\Sigma(\cdot;\theta)$ are the mean and variance of a Gaussian density parametrized by θ. We shall also assume that $\Sigma(x;\theta)$ is a diagonal matrix with entries $\sigma_i^2(x;\theta)$, for $i \in [K]$. For neural networks, the parameters θ correspond to the weights of a network that takes inputs x and has $2K$ outputs corresponding to μ_i and σ_i^2.

5.1 Bound on $I(Y;T)$

We propose to use a variational bound for $I(Y;T)$ derived in Alemi et al. [34]. Let $\widehat{Y}$ be the estimate of Y based on T. This means that we have the Markov chain $Y \to X \to T \to \widehat{Y}$. A lower bound on $I(Y;T)$ is given by

$$I(Y;T) \geq \int p_{XY}(x,y) p_{T|X}(t|x) \log p_{\widehat{Y}|T}(y|t) dx dy dt.$$

Using the empirical distribution, this bound evaluates to $\frac{1}{N}\sum_{i=1}^{N} \int p_{T|X}(t|x_i) \log p_{\widehat{Y}|T}(y_i|t) dt$. In other words, the variational approximation to $I(Y;T)$ is essentially the cross-entropy loss.

5.2 Bound on $\mathsf{mmse}(Y|T)$

The MMSE has the variational characterization $\mathsf{mmse}(Y|T) = \inf_{f:\mathcal{T}\to\mathcal{Y}} \mathbb{E}(Y - f(T))^2$, where the infimum is achieved by the conditional expectation function $f^*(t) = \mathbb{E}(Y|T=t)$. Calculating f^* requires evaluating the posterior $p_{Y|T}(y|t)$, which we wish to avoid. Thus, we propose to use the upper bound $\mathsf{mmse}(Y|T) \leq \mathbb{E}(Y - \tilde{f}(T))^2$, for a suitable function $\tilde{f} : \mathcal{T} \to \mathcal{Y}$ that is easy to compute. This function $\tilde{f}$ may be parametrized by some parameters ϕ, which will be updated during iterations of stochastic gradient descent. (See Section 5.4 for details.)

5.3 Exact expression for $\Phi(T|X)$

The term $\Phi(T|X)$ may be efficiently optimized in its original form, and we do not need to derive variational bounds for it. To see this, note that $\Phi(T|X) = \int_{\mathcal{X}} \Phi(T|X = x)p_X(x)dx$. We have $p_{T|X}(t|x) = \frac{1}{\sqrt{(2\pi)^k \prod_{i=1}^k \sigma_i^2(x)}} \exp\left(-\sum_{j=1}^k \frac{(t_j-\mu_j(x))^2}{2\sigma_j(x)^2}\right)$. We may explicitly calculate $\Phi(T|X = x)$:

$$\begin{aligned}\Phi(T|X=x) &= \int_{\mathbb{R}^k} \|\nabla_x \log p_{T|X}(t|x)\|_2^2 p_{T|X}(t|x)dt \\ &= \int_{\mathbb{R}^k} \left\| -\sum_{j=1}^k \frac{\nabla_x \sigma_j(x)}{\sigma_j(x)} + \sum_{j=1}^k \frac{(t_j-\mu_j(x))}{\sigma_j(x)^2}\nabla_x \mu_j(x) + \sum_{j=1}^k \frac{(t_j-\mu_j(x))^2}{\sigma_j(x)^3}\nabla_x \sigma_j(x)\right\|_2^2 \\ &\qquad \frac{1}{\sqrt{(2\pi)^k \prod_{j=1}^k \sigma_j^2(x)}} \exp\left(-\sum_{j=1}^k \frac{(t_j-\mu_j(x))^2}{2\sigma_j(x)^2}\right) dt.\end{aligned}$$

What is essential is to compute the *derivative* of $\Phi(T|X)$ with respect to θ.

5.4 Evaluating stochastic gradients

Stochastic gradient for $\mathsf{mmse}(Y|T)$: Suppose we have a data set $\{(x_i, y_i) : i \in [n]\}$. The empirical distribution is $\mathbb{P}_n(x, y) := \frac{1}{n}\sum_{i=1}^n \delta(x_i, y_i)$, and the empirical version of the variational approximation to the MMSE term is $\mathbb{E}_{\mathbb{P}_n}\left[(Y - \tilde{f}(T))^2\right] = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^k}(y_i - \tilde{f}(t;\phi))^2 p_{T|X}(t|x_i;\theta)dt$, where we have have explicitly included the parameters ϕ and θ to indicate that they parametrize $\tilde{f}$ and $p_{T|X}$, respectively. Note that SGD involves calculating the gradient of this function in with respect to ϕ and θ, so it is critically important to evaluate these derivatives in a computationally feasible manner. We will employ the reparametrization trick of Kingma and Welling [56].

Note that we may write $T = \mu(x;\theta) + \Sigma(x;\theta)^{1/2}\epsilon =: \tau(x, \epsilon; \theta)$, where $\epsilon \sim \mathcal{N}(0, I)$. Rewriting the MMSE integral, we then have

$$\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^k}(y_i - \tilde{f}(t;\phi))^2 p_{T|X}(t|x_i;\theta)dt = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^k}(y_i - \tilde{f}(\tau(x_i,\epsilon;\theta);\phi))^2 p_\epsilon(\epsilon)d\epsilon.$$

Furthermore, we may approximate the integral over ϵ by resampling $\epsilon_{11}, \ldots, \epsilon_{mn}$ from the distribution p_ϵ, and computing $\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_i - \tilde{f}(\tau(x_i, \epsilon_{ij}; \theta); \phi))^2\epsilon_{ij}$.

Finally, note that the gradient of this function with respect to θ (or ϕ) may be calculated easily using backpropagation, since we may use the trained neural networks to evaluate the functions τ_θ and $\tilde{f}_\phi$, as well as the gradients of these functions with respect to either their parameters or their inputs. This shows how to take the gradient of the MMSE term.

Stochastic gradient for $I(Y;T)$ (as in Alemi et al. [34]): The reparametrization trick applied to the empirical version of the variational approximation to $I(Y;T)$ is given by

$$\frac{1}{N}\sum_{i=1}^{N}\int p_{T|X}(t|x_i;\theta)\log p_{\widehat{Y}|T}(y_i|t;\phi)dt = \frac{1}{N}\sum_{i=1}^{N}\int \log p_{\widehat{Y}|\tau(X,\epsilon;\theta)}(y_i|\tau(x_i,\epsilon;\theta);\phi)p(\epsilon)d\epsilon.$$

The gradient of the right hand side with respect to θ is given by

$$\nabla_\theta\left[\frac{1}{N}\sum_{i=1}^{N}\int \log p_{\widehat{Y}|\tau(X,\epsilon;\theta)}(y_i|\tau(x_i,\epsilon;\theta);\phi)p(\epsilon)d\epsilon\right] = \frac{1}{N}\sum_{i=1}^{N}\int \nabla_\theta\left[\log p_{\widehat{Y}|\tau(X,\epsilon;\theta)}(y_i|\tau(x_i,\epsilon;\theta);\phi)\right]p(\epsilon)d\epsilon.$$

For a given realization of ϵ, the gradient inside the integral is easily computed via backpropagation. An unbiased stochastic gradient is computed by sampling $\epsilon \sim \mathcal{N}(0, I)$ one or more times and averaging the calculated gradients.

Stochastic gradient for $\Phi(T|X)$: We now express the Fisher information term using the reparametrization above:

$$\begin{aligned}\Phi(T|X) &\approx \mathbb{E}_{\mathbb{P}_n}[\Phi(T|X=x)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^k}\left\| -\sum_{j=1}^{k}\frac{\nabla_x\sigma_j(x_i)}{\sigma_j(x_i)} + \sum_{j=1}^{K}\frac{(\tau(x_i,\epsilon;\theta)-\mu_j(x_i))}{\sigma_j(x_i)^2}\nabla_x\mu_j(x_i)\right.\end{aligned}$$

$$+\sum_{j=1}^{k}\frac{(\tau(x_i,\epsilon;\theta)-\mu_j(x_i))^2}{\sigma_j(x_i)^3}\nabla_x\sigma_j(x_i)\Bigg\|_2^2 p_\epsilon(\epsilon)d\epsilon. \tag{8}$$

Again, we may approximate the gradient by sampling from the distribution p_ϵ and then computing stochastic gradients with respect to θ using backpropagation. Note that this will require us to calculate expressions such as $\nabla_\theta\nabla_x\mu_j(x)$ and $\nabla_\theta\nabla_x\sigma_j(x)$, which may be computationally intensive depending on the dimension of x, but can still be obtained from the trained neural network classifier τ_θ.

Altogether, we conclude that the variational approximations to expressions (2) and (3) may be optimized using mini-batch SGD.

6 Experiments

We now provide simulation results showing the behavior of the feature extraction methods we have proposed. We begin with a variety of experiments involving synthetic data generated from a Gaussian mixture, and then provide experiments on MNIST data. In the case of the Gaussian mixture data, we optimized the MMSE formulation (2), the mutual information formulation (3), and the standard information bottleneck (1). For our MNIST data studies, we performed the variational optimization approach presented in the previous section applied to the mutual information formulation (3).

6.1 Gaussian mixture

We begin by conducting simulations for the setting described in Example 1. Figure 1 shows point clouds of 1000 points for the case of $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.2$. Note that in the absence of adversarial perturbations, the decision boundary of the optimal classifier should be close to the horizontal axis (Classifier 1 in Figure 1). Equivalently, the angle of w^* should be close to 90°. However, this classifier will not be optimal if we require robustness to adversarial ℓ_2-perturbations: Since the decision boundary is close to x_1-axis, it is easy for the adversary

to perturb the x_2-coordinate of a data point and cause the classifier to make an error, since a large number of points are *near* the boundary of Classifier 1. Thus, a robust classifier should tilt the boundary slightly to protect against ℓ_2-perturbations, leading to Classifier 2. This intuition is formalized in equation (10) and Figure 2(a) below, where we can see how the robustness of a classifier varies as the angle of the linear classifier tilts.

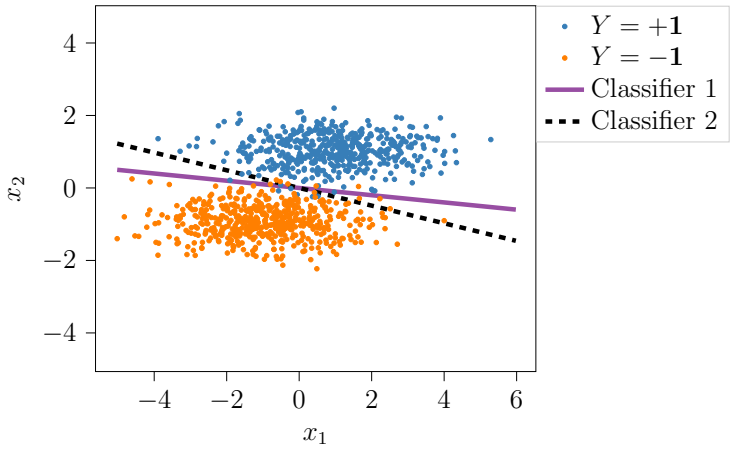


Figure 1: Plot showing point clouds of 1000 samples for Example 1, with $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.2$. The class $Y = +\mathbf{1}$ is centered at $(1, 1)$ and the class $Y = -\mathbf{1}$ is centered at $(-1, -1)$.

We will now show that imposing robustness via a Fisher information $\Phi(T|X)$ encourages a similar effect. By Lemma 1, we have $\Phi(T|X) = \|w\|_2^2$. Turning to the MMSE term, for a given w, we denote the features T by T_w. First, we note that conditioned on $Y = +\mathbf{1}$ and $-\mathbf{1}$, the distribution of T_w is $\mathcal{N}(w_1+w_2, w_1^2\sigma_1^2+w_2^2\sigma_2^2+1)$ and $\mathcal{N}(-w_1-w_2, w_1^2\sigma_1^2+w_2^2\sigma_2^2+1)$, respectively. Let $\mu_w := w_1 + w_2$ and $\sigma_w^2 := w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 1$. We see that

$$\mathbb{P}(Y = +\mathbf{1}|T_w = t) = \frac{\exp\left(-\frac{(t-\mu_w)^2}{2\sigma_w^2}\right)}{\exp\left(-\frac{(t-\mu_w)^2}{2\sigma_w^2}\right) + \exp\left(-\frac{(t+\mu_w)^2}{2\sigma_w^2}\right)} = \frac{1}{1 + \exp\left(-\frac{2\mu_w t}{\sigma_w^2}\right)} := \alpha_{t,w}.$$

Furthermore, we have $\mathbb{P}(Y = -\mathbf{1}|T_w = t) = 1 - \alpha_{t,w} := \bar{\alpha}_{t,w}$, so we can write

$$\begin{aligned}\mathsf{mmse}(Y|T_w) = \frac{1}{2}\int 2\left(1 - (\alpha_{t,w} - \bar{\alpha}_{t,w})\right)^2 p(t|Y = +\mathbf{1})dt \\ + \frac{1}{2}\int 2\left(-1 - (\alpha_{t,w} - \bar{\alpha}_{t,w})\right)^2 p(t|Y = -\mathbf{1})dt\end{aligned}$$

$$= \int_{\mathbb{R}} \frac{8\alpha_{t,w}^2}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(t+\mu_w)^2}{2\sigma_w^2}} dt = f\left(\frac{\mu_w}{\sigma_w}\right),$$

where $f(a) := \mathbb{E}_{Z\sim\mathcal{N}(a^2,a^2)} \frac{8}{(1+e^{2Z})^2}$. Thus, the MMSE of w depends only on the scalar $\frac{\mu_w^2}{\sigma_w^2}$. Monte Carlo approximation shows that $f(\cdot)$ is a decreasing function on positive reals.

Combining the two calculations, the optimal w_* solves the optimization problem

$$\begin{aligned} \arg\min_w \quad & f\left(\frac{\mu_w}{\sigma_w}\right) \quad \equiv \quad \arg\max_w \quad \frac{\mu_w^2}{\sigma_w^2} \\ s.t. \quad & \|w\|_2 \le R \qquad\qquad s.t. \quad \|w\|_2 \le R. \end{aligned} \tag{9}$$

Note that the optimal value in equation (9) is achieved at $\|w\|_2 = R$. A smaller value of R corresponds to increased robustness in features.

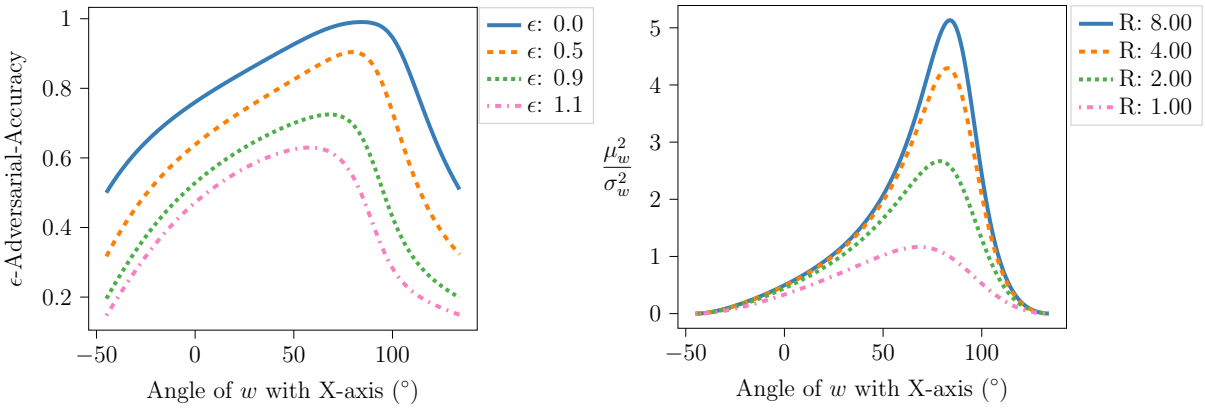
We briefly describe the performance of linear classifiers in the presence of an adversary. Consider linear classifiers of the form $\text{sign}(w^\top x)$. For such classifiers, adversarial accuracy in the presence of ϵ-corruption in the ℓ_2-metric is given by

$$\epsilon\text{-Adversarial-Accuracy} = \mathbb{P}\left\{Z \ge \frac{\epsilon\|w\|_1 - \mu_w}{\sqrt{w^\top \Sigma w}}\right\}, \tag{10}$$

where $\epsilon = 0$ corresponds to the accuracy of the classifier in the absence of any adversary, and Z is a standard normal random variable. It follows that the performance of such classifiers depends on w only through its direction. We parametrize the direction by θ, the angle between w and the horizontal axis, measured counter-clockwise. As the level of perturbation ϵ changes, the optimal θ^* changes considerably. Figure 2(a) shows the relationship between ϵ-Adversarial-Accuracy and the classifier angle for different values of ϵ. From the figure, we can see that the choice of classifier depends crucially on the desired level of robustness.

We now show that the same phenomenon occurs for the classifier obtained by solving equation (9) via a grid search. Comparing Figures 2(a) and 2(b), we observe that these are *qualitatively* the same: the angle of the optimal w^* (where the curves peak) reduces as

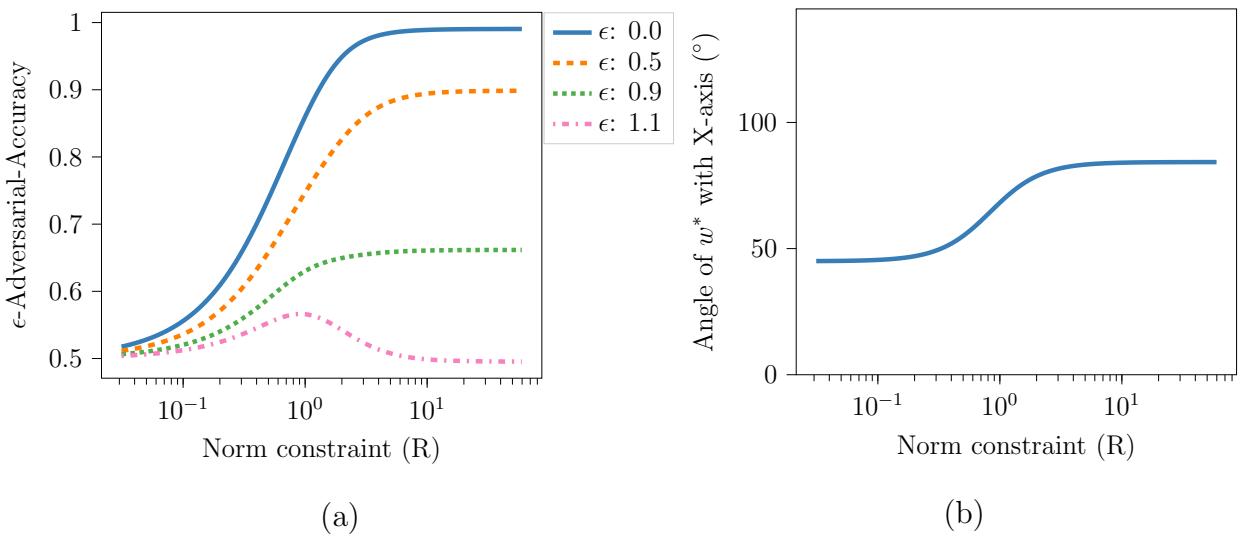
more robustness is desired. In Figure 2(a), additional robustness is imposed by increasing adversary's power ϵ; in Figure 2(b), it is imposed by reducing the norm constraint R.



(a) (b)

Figure 2: (a) Plot showing the adversarial accuracy for linear classifiers, $\text{sign}(w^\top x)$, as a function of the angle of w and ϵ, the maximum perturbation allowed in the ℓ_2-metric. The adversarial accuracy of such a classifier depends on w only through its direction, which we parametrize by the angle of w with the horizontal axis, measured counter-clockwise. Different curves correspond to different ϵ. Notice that the optimal classifier changes as ϵ increases. (b) Plot showing the value of the objective function in equation (9) as a function of the angle of w and R, where R corresponds to $\Phi(T|X)$ and $\frac{\mu_w^2}{\sigma_w^2}$ corresponds to $\text{MMSE}(Y|T)$. The plot shows that the behavior of the MMSE with a constraint on $\Phi(T|X)$ is similar to the ϵ-Adversarial-Accuracy as a function of ϵ. Primarily, the angle of the optimal w changes as the constraint ($\Phi(T|X)$) changes, mirroring the trend in (a) where the angle of w^* changes with the desired level of robustness (ϵ). We take $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.2$ for both plots.

Figure 3(a) shows the relation between the norm constraint R and the ϵ-Adversarial-Accuracy for several values of ϵ. As expected, as $R \to 0$, the extracted feature becomes independent of X and the accuracy tends to 50%. For $\epsilon = 0$, i.e., without any adversarial perturbation, the accuracy of the classifier degrades monotonically as we constrain the norm to be smaller. The curve corresponding to $\epsilon = 1.1$ is more insightful, showing that the performance of the classifier increases at first as we constrain $\|w\|_2$ to be smaller. If we further increase the constraint (making R smaller), the extracted feature tends toward Gaussian noise and the performance degrades. Figure 3(b) shows how the angle of the optimal w^* changes with the norm constraint on R. As $R \to \infty$, this angle is close to 90°, whereas as $R \to 0$, it tends to 45°.



(a) (b)

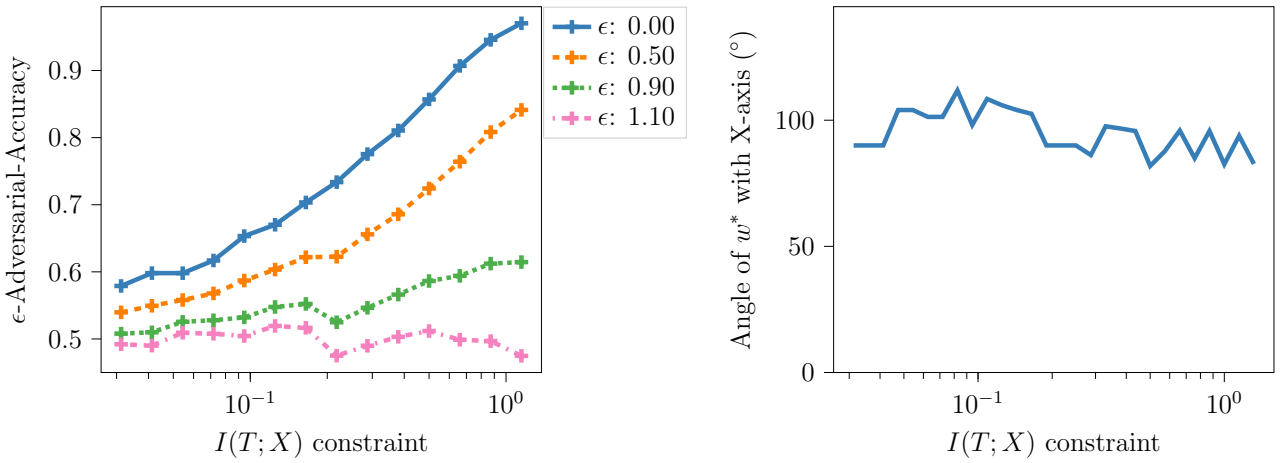
Figure 3: (a) Plot showing the adversarial accuracy for the solution of the robust information bottleneck (9) as a function of the norm constraint R and the maximum perturbation ϵ allowed in the ℓ_2-metric. For each R, we solve the optimization problem in equation (9) and then calculate its adversarial accuracy for different ϵ. (b) Plot showing the angle of the linear classifier as a function of the norm constraint R. For each R, we have $\|w^*\|_2 = R$, so we parametrize w^* by its angle with horizontal axis. The angle of the optimal w^* changes with $\Phi(T|X)$, the desired level of robustness.

We now compare the behavior of equation (9) with the usual information bottleneck of Tishby et al. [29]. We again consider the features T_w of the form $w^T X + Z$, where $Z \sim N(0, 1)$. We solve the optimization problem in equation (11) using a grid search:

$$\begin{aligned} \arg\min_{w} \quad & I(T_w; Y) \\ s.t. \quad & I(T_w; X) \leq Q. \end{aligned} \tag{11}$$

As T_w is a mixture of two univariate Gaussian distributions, we estimate the entropy of T using a Monte Carlo estimate. We report the results in Figure 4. Note that the angle of w^* *does not change* with the constraint on $I(T; X)$. As seen in Figure 2(a), the angle of the robust linear classifier changes with increasing ϵ—an intuitive trend that is successfully mimicked by the robust information bottleneck in Figure 2(b). Why does the information bottleneck behave so differently? Simply stated, unlike $\Phi(T|X)$, the term $I(X; T)$ is invariant

to linear bijective transformations of X. Thus, the information bottleneck formulation is blind to the skewed variances—which are crucial in our example—and returns the same linear classifier for different constraints on $I(X;T)$. This example also illustrates how compressed features may not always be robust.



(a) (b)

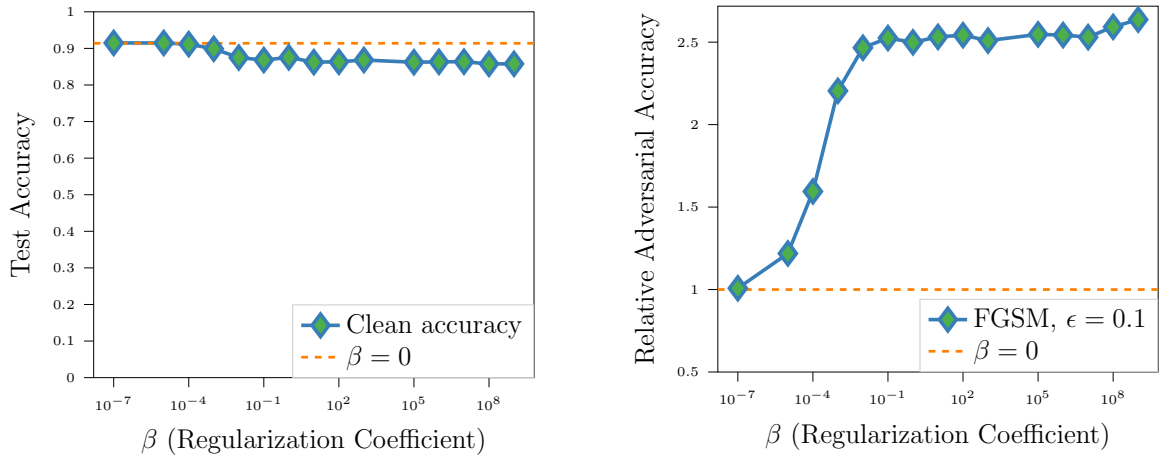
Figure 4: (a) Plot showing the adversarial accuracy of the standard information bottleneck as a function of the constraint on $I(T;X)$ and ϵ, the maximum perturbation allowed in the ℓ_2-metric. For each $I(T;X)$, we solve the optimization problem in equation (11) and then calculate its adversarial accuracy for different ϵ. (b) Plot showing the angle of the linear classifier as a function of the constraint on $I(T;X)$. The angle of the optimal w^* does not change with $I(T;X)$—a departure from the trends observed in Figure 2.

6.2 MNIST data

We now describe our experiments on the MNIST data set. We use the variational bounds described in Section 5 for the mutual information formulation (cf. Section 5.1) and implement the Fisher information term $\Phi(T|X)$ as a regularizer with coefficient β. Recall that we consider $p_{T|X}(\cdot|x) = \mathcal{N}(\mu(x;\theta), \Sigma(x;\theta))$, where $(\mu(x;\theta), \Sigma(x;\theta))$ are the mean and (diagonal) covariance matrix of a K-dimensional Gaussian distribution. We evaluate the adversarial robustness of the neural networks using the Fast Gradient Sign Method (FGSM) [12] with $\epsilon = 0.1$, with 10 random initializations for each example. As our model is inherently stochastic, we make the final prediction by taking an average over 12 samples from the posterior. This

allows the adversary to obtain a consistent estimate of the gradient. We approximate the stochastic integral in equation (8) with a single sample from the corresponding Gaussian distribution. As both the loss function and the regularizer have a sum structure, we use the Adam optimizer. We implemented our experiments using Tensorflow and the Adversarial robustness toolbox [57].

We first consider a simple one-layer architecture. The model architecture is $784 - 2K$ with $K = 10$, without any non-linearity. The model is thus a variant of multiclass logistic regression with stochastic logits. The first K values of the last layer encode the mean, and the remaining K values encode the variance of the features after a softplus transformation, similar to Alemi et al. [34]. For each value of β, we train the model for 150 epochs. Figure 5 reports the effect of the regularization coefficient β on clean accuracy and adversarial accuracy. As β grows, the adversarial accuracy improves, while the test accuracy decreases. We also ran experiments with the variational information bottleneck, i.e., $\gamma > 0$ and $\beta = 0$. The clean accuracy behaves (w.r.t. γ) similar to Figure 5(a), but we did not observe any trend similar to Figure 5(b). With the same setup as above, the best adversarial accuracy was 8%.

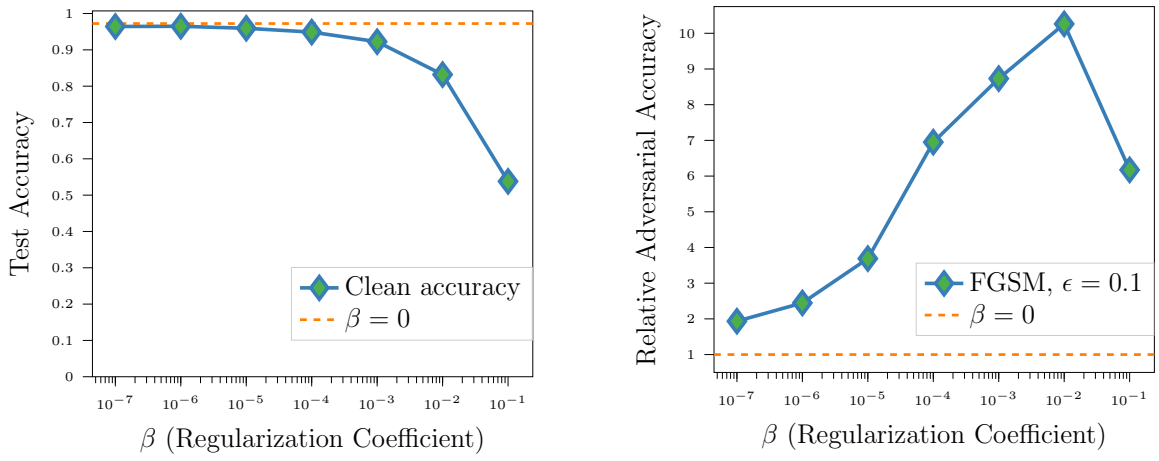


(a) (b)

Figure 5: Plots showing the clean accuracy and adversarial accuracy of multi-class logistic regression as a function of the regularization coefficient β. Panel (a) reports the test accuracy of the model, and panel (b) reports the relative adversarial accuracy as a function of β. We evaluate against the FGSM attack with $\epsilon = 0.1$ and 10 random initializations. The baseline accuracy corresponds to the case when $\beta = 0$, which is 14.31%.

To show that this phenomenon is also observed in more complicated networks, we consider a simple fully-connected multilayer architecture: $784 - 100 - 20 - 2K$. We use ReLU activations in all layers except the last layer, which is linear. Given the features $T = t$, the output of the classifier is a simple soft-max layer of the features (without any weights).

For each β, we train the network for 200 epochs. Figure 6 shows the effect of β on the adversarial accuracy. The adversarial accuracy increases at first as we increase the regularization coefficient, supporting the claim that Fisher regularization leads to increased adversarial robustness. If we further increase the regularization coefficient, increased robustness comes at the expense of accuracy and leads to degraded performance. This trend is similar to the case of $\epsilon = 1.1$ in Figure 3(a). We also tested this model against more powerful projected gradient descent (PGD) attacks. Although the absolute adversarial accuracy when using PGD attacks is lower compared to that obtained for an FGSM attack, the relative adversarial accuracy follows a trend identical to that in Figure 6.



(a) (b)

Figure 6: Plot showing the effect of β on clean accuracy and adversarial accuracy of multilayer neural networks. Panel (a) shows that clean accuracy decreases as we increase β. Panel (b) shows the relative adversarial accuracy of the neural networks as a function of the regularization coefficient β. We evaluate adversarial accuracy against the FGSM attack with $\epsilon = 0.1$ and 10 random initializations. The baseline accuracy corresponds to the case when $\beta = 0$, which is 2.6%.

7 Discussion

The research directions explored in this paper were inspired by recent work in adversarial machine learning. In particular, we were intrigued by the notion of a seemingly unavoidable tradeoff between robustness and accuracy, and the existence of a dichotomy between robust and non-robust features. A bottleneck formulation lends itself naturally to modeling a tradeoff between robustness and accuracy; quantifying these notions via information and estimation theory, we have proposed the robust information bottleneck as a new variational principle for extracting maximally useful robust features.

Like the standard information bottleneck, the robust information bottleneck formulation references only the data distribution, making it extremely general. Applying the principle for specific classes of features (e.g., linear or logistic) leads to feature-specific regularization terms. This means that one need not decide a priori to use an ℓ_1- or ℓ_2-regularizer, but may instead use a regularization penalty corresponding to the Fisher information term discussed in this paper. Furthermore, we showed that the Fisher information term satisfies a host of properties that make it ideally suited to characterize robustness. The robust information bottleneck is most clearly understood in the case of jointly Gaussian data: We showed that the optimally robust features in this setting are also jointly Gaussian with the data, and examined connections to the solution of the canonical information bottleneck.

Lastly, we showed that it is computationally easy to extract features via the robust information bottleneck optimization using a variational approximation, and that a classifier trained on robust features extracted via the robust information bottleneck principle is indeed robust to simple adversarial attacks. Although we were able to defeat the classifier using stronger classes of adversarial attacks, our work in this paper suggests that a deeper investigation of Fisher regularization in neural networks is likely to be fruitful.

Acknowledgments

We thank the AE and anonymous reviewers for their critical feedback that has led to a much improved manuscript. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing during Summer 2019. VJ thanks Chandra Nair for pointing him to [49, Corollary 1]. AP acknowledges support from the UW-Madison Institute for Foundations of Data Science (IFDS), NSF grant CCF-1740707. VJ acknowledges support from NSF grants CCF-1841190 and CCF-1907786, and the Nvidia GPU grant program. PL acknowledges support from NSF grant DMS-1749857.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in NIPS*, pages 1097–1105, 2012.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

[4] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

[5] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[6] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[8] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[9] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of CVPR*, pages 1765–1773, 2017.

[10] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *International Conference on Document Analysis*, volume 3, 2003.

[11] S. Rajput, Z. Feng, Z. Charles, P. Loh, and D. Papailiopoulos. Does data augmentation lead to positive margin? In *ICML*, pages 5321–5330, 2019.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

[14] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. *arXiv:1802.03471*, 2018.

[15] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *arXiv:1902.02918*, 2019.

[16] Y. Zhang and P. Liang. Defending against whitebox adversarial attacks via randomized discretization. *arXiv:1903.10586*, 2019.

[17] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. *arXiv:1801.09344*, 2018.

[18] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel. Towards fast computation of certified robustness for ReLU networks. *arXiv:1804.09699*, 2018.

[19] B. Li, C. Chen, W. Wang, and L. Carin. Second-order adversarial attack and certifiable robustness. *arXiv:1809.03113*, 2018.

[20] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv:1906.04584*, 2019.

[21] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[22] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. Adversarial spheres. *arXiv:1801.02774*, 2018.

[23] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv:1805.12152*, 2018.

[24] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In *Advances in NIPS*, pages 1178–1187, 2018.

[25] E. Dohmatob. Limitations of adversarial robustness: Strong no free lunch theorem. *arXiv:1810.04065*, 2018.

[26] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv:1901.08573*, 2019.

[27] V. Feldman. Does learning require memorization? A short tale about a long tail. *arXiv:1906.05271*, 2019.

[28] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *arXiv:1905.02175*, 2019.

[29] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[30] A. Achille and S. Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[31] A. Achille and S. Soatto. Where is the information in a deep neural network? *arXiv:1905.12213*, 2019.

[32] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 208–215. ACM, 2000.

[33] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *Advances in NIPS*, pages 881–888, 2003.

[34] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv:1612.00410*, 2016.

[35] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[36] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

[37] M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

[38] O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[39] S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. In *Advances in NIPS*, pages 351–359, 2013.

[40] D. Guo, S. Shamai, and S. Verdú. The interplay between information and estimation measures. *Foundations and Trends® in Signal Processing*, 6(4):243–429, 2013.

[41] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[42] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[43] R. Zamir. A proof of the Fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, 44(3):1246–1250, 1998.

[44] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–246, 2013.

[45] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv:1710.10571*, 2017.

[46] M. Staib and S. Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS Workshop on ML and Computer Security*, 2017.

[47] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6(Jan):165–188, 2005.

[48] A. Globerson and N. Tishby. On the optimality of the Gaussian information bottleneck curve. *The Hebrew University of Jerusalem, Tech. Rep*, 2004.

[49] Y. Geng and C. Nair. The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages. *IEEE Transactions on Information Theory*, 60(4):2087–2104, 2014.

[50] E. H. Lieb. Gaussian kernels have only Gaussian maximizers. In *Inequalities*, pages 595–624. Springer, 2002.

[51] E. A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 101(1):194–211, 1991.

[52] T. Courtade and J. Jiao. An extremal inequality for long Markov chains. In *52nd Annual Allerton Conference*, pages 763–770. IEEE, 2014.

[53] X. Zhang, V. Anantharam, and Y. Geng. Gaussian optimality for derivatives of differential entropy using linear matrix inequalities. *Entropy*, 20(3):182, 2018.

[54] T. A. Courtade. A strong entropy power inequality. *IEEE Transactions on Information Theory*, 64(4):2173–2192, 2018.

[55] V. Anantharam, V. Jog, and C. Nair. Unifying the Brascamp-Lieb inequality and the entropy power inequality. *arXiv:1901.06619*, 2019.

[56] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[57] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards. Adversarial robustness toolbox v1.0.0. *CoRR*, 1807.01069, 2018.

[58] E Aras, K-Y Lee, A Pananjady, and T. A. Courtade. A family of Bayesian Cramér-Rao bounds, and consequences for log-concave priors. *arXiv:1902.08582*, 2019.

[59] S. Y. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:24–39, 1980.

[60] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

[61] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[62] M. Costa. A new entropy power inequality. *Information Theory, IEEE Transactions on*, 31(6):751–760, 1985.

[63] C. Villani. A short proof of the "concavity of entropy power". *IEEE Transactions on Information Theory*, 46(4):1695–1696, 2000.

[64] N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271, 1965.

[65] S. N. Diggavi and T. M. Cover. The worst additive noise under a covariance constraint. *IEEE Transactions on Information Theory*, 47(7):3072–3081, 2001.

[66] K. K. Kim. A note on the convexity of $\log\det(i+kx^{-1})$ and its constrained optimization representation. *arXiv:1509.00777*, 2015.

[67] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2013.

[68] S. G. Bobkov, G. P. Chistyakov, and F. Götze. Fisher information and convergence to stable laws. *Bernoulli*, 20(3):1620–1646, 2014.

A Proofs for Section 2

In this appendix, we provide the proofs of several lemmas introduced in the setup of the robust information bottleneck formulation.

A.1 Proof of Lemma 3

1. Since $I(X; f(T)) = H(X) - H(X|f(T))$ and $I(X;T) = H(X) - H(X|T)$, it is enough to show that $H(X|T) = H(X|f(T))$. This is a standard fact that may be found in Cover and Thomas [42], but we include it here for completeness. If $S = f(T)$, then $p_S(s)ds = p_T(t)dt$ when $s = f(t)$. Hence,

$$\begin{aligned} H(X|S) &= \int_{\mathbb{R}^d} H(X|S=s)p_S(s)ds \\ &= \int_{\mathbb{R}^d} H(X|S=f^{-1}(t))p_T(t)dt \\ &= \int_{\mathbb{R}^d} H(X|T=t)p_T(t)dt \\ &= H(X|T). \end{aligned}$$

The proof for the equation involving Y is analogous.

Note that $\mathsf{mmse}(Y|T)$ has the equivalent formulation

$$\mathsf{mmse}(Y|T) = \inf_{g:\mathbb{R}^d \to \mathcal{Y}} \mathbb{E}[\|Y - g(T))\|^2],$$

where g is any measurable function from $\mathbb{R}^d$ to $\mathcal{Y}$. For any such function g, we have

$$\mathbb{E}[\|Y - g(T))\|^2] = \mathbb{E}[\|Y - g(f^{-1}(S))\|]^2.$$

Thus, taking an infimum over g, we obtain

$$\mathsf{mmse}(Y|T) = \inf_g \mathbb{E}[\|Y - g(f^{-1}(S))\|^2 \geq \inf_g \mathbb{E}[\|Y - g(S)\|^2] = \mathsf{mmse}(Y|S).$$

Analogously, we have $\mathsf{mmse}(Y|S) \geq \mathsf{mmse}(Y|T)$, proving the desired statement.

Observe that

$$\Phi(S|X) = \int \|\nabla_x \log p_{S|X}(s|x)\|_2^2 p_X(x) p_{S|X}(s|x) dx ds.$$

Using the change of variables $t = f^{-1}(s)$, we have

$$\begin{aligned}
\Phi(S|X) &= \int \left\| \nabla_x \log \frac{p_{T|X}(t|x)}{|\det \nabla_t f(t)|} \right\|_2^2 p_X(x) p_{T|X}(t|x) dx dt \\
&= \int \|\nabla_x \log p_{T|X}(t|x)\|^2 p_X(x) p_{T|X}(t|x) dx dt \\
&= \Phi(T|X),
\end{aligned}$$

as wanted.

2. Using the independence $T_2 \perp\!\!\!\perp (T_1, X, Y)$, we have

$$I(X;T) = I(X;T_1) + I(X;T_2|T_1) = I(X;T_1).$$

Similarly, we may obtain $I(Y;T) = I(Y;T_1)$. For the Fisher information term, observe that

$$\nabla_x \log p_{T|X=x}(t) = \nabla_x (\log p_{T_1|X=x}(t_1) + \log p_{T_2}(t_2)) = \nabla_x \log p_{T_1|X=x}(t_1).$$

Taking the expectation of the squared-norms, we conclude that $\Phi(T|X) = \Phi(T_1|X)$. Finally, we have $p_{Y|T=t}(y) = p_{Y|T_1=t_1}(y)$, so $\mathbb{E}[Y|T] = \mathbb{E}[Y|T_1]$. This shows that $\mathsf{mmse}(Y|T) = \mathsf{mmse}(Y|T_1)$.

A.2 Proof of Lemma 4

We use the entropic improvement of the Cramér-Rao bound (or Van Trees inequality) stated below:

$$\frac{1}{2\pi e}\exp\left(\frac{2}{p}H(X|T)\right) \geq \frac{p}{J(X)+\Phi(T|X)}. \tag{12}$$

The above inequality appeared recently in Aras et al. [58]. Aras et al. [58] noted that the entropic improvement does not appear to be well-known, although the univariate version was published in Efroimovich [59] in 1980. The claimed inequality immediately follows by noting that $I(X;T) = H(X) - H(X|T)$ and plugging in the lower bound for $H(X|T)$ from inequality (12).

A.3 Proof of Lemma 1

Note that

$$T \mid X = x \sim N(Ax, I),$$

so

$$p_{T|X}(t|x) = \frac{1}{(2\pi)^{k/2}}\exp\left(-\frac{1}{2}(t-Ax)^\top(t-Ax)\right).$$

In particular, we may calculate

$$\nabla_x p_{T|X}(t|x) = \nabla_x\left(-\frac{1}{2}(t-Ax)^\top(t-Ax)\right) = A^\top(t-Ax),$$

so

$$\begin{aligned}\Phi(T|X) &= \mathbb{E}_{T|X=x}\left[(t-Ax)^\top AA^\top(t-Ax)\right] = \mathbb{E}\left[\epsilon^\top AA^\top\epsilon\right] = \mathbb{E}_\epsilon\left[\mathrm{tr}(\epsilon^\top AA^\top\epsilon)\right] \\ &= \mathrm{tr}\left(AA^\top\mathbb{E}[\epsilon\epsilon^\top]\right) = \mathrm{tr}(AA^\top) = \|A\|_F^2,\end{aligned}$$

as wanted.

A.4 Proof of Lemma 2

We can compute

$$\nabla_x \mathbb{P}(T = 1 \mid X = x) = -\nabla_x \log\left(1 + \exp(-x^\top w)\right) = \frac{\exp(-x^\top w)w}{1 + \exp(-x^\top w)},$$
$$\nabla_x \mathbb{P}(T = -1 \mid X = x) = -\nabla_x \log\left(1 + \exp(x^\top w)\right) = \frac{-\exp(x^\top w)w}{1 + \exp(x^\top w)},$$

so that

$$\begin{aligned}
\Phi(T|X = x) &= \frac{1}{1 + \exp(-x^\top w)} \cdot \frac{\exp(-2x^\top w)\|w\|_2^2}{\left(1 + \exp(-x^\top w)\right)^2} + \frac{1}{1 + \exp(x^\top w)} \cdot \frac{\exp(2x^\top w)\|w\|_2^2}{\left(1 + \exp(x^\top w)\right)^2} \\
&= \|w\|_2^2 \left(\frac{\exp(-2x^\top w)}{\left(1 + \exp(-x^\top w)\right)^3} + \frac{\exp(2x^\top w)}{\left(1 + \exp(x^\top w)\right)^3}\right) \\
&= \|w\|_2^2 \cdot \frac{\exp(-2x^\top w) + \exp(-x^\top w)}{\left(1 + \exp(-x^\top w)\right)^3} \\
&= \|w\|_2^2 \cdot \frac{\exp(-x^\top w)}{\left(1 + \exp(-x^\top w)\right)^2} \\
&= \|w\|_2^2 \cdot \mathbb{P}(T = 1 \mid X = x) \cdot \mathbb{P}(T = -1 \mid X = x),
\end{aligned}$$

as claimed.

B Proofs for Section 3

In this appendix, we provide proofs of the lemmas concerning robustness properties of Fisher information.

B.1 Proof of Lemma 5

The proof is a direct application of the data processing inequality for Fisher information. We have the following sequence of inequalities:

$$\Phi(\widehat{Y}|X) \overset{(a)}{=} J(X|\widehat{Y}) - J(X)$$

$$
\begin{aligned}
&\overset{(b)}{\leq} J(X|\widehat{Y},T) - J(X) \\
&\overset{(c)}{=} J(X|T) - J(X) \\
&= \Phi(T|X).
\end{aligned}
$$

Here, (a) follows from Lemma 26, (b) follows from Lemma 27, and (c) follows because $X \to T \to \widehat{Y}$ is a Markov chain.

B.2 Proof of Lemma 6

Note that

$$
\begin{aligned}
I(X;T) - I(X+\sqrt{\delta}Z;T) &= H(X) - H(X|T) - H(X+\sqrt{\delta}Z) + H(X+\sqrt{\delta}Z|T) \\
&= [H(X) - H(X+\sqrt{\delta}Z)] + [-H(X|T) + [H(X+\sqrt{\delta}Z|T)] \\
&= -\frac{\delta}{2}J(X) + \frac{\delta}{2}J(X|T) + o(\delta) \\
&= \frac{\delta}{2}\Phi(T|X) + o(\delta),
\end{aligned}
$$

where the approximation follows by de Bruijn's identity [60].

B.3 Proof of Lemma 7

Up to a second-degree approximation, we have

$$
p_{T|X=x+\epsilon u}(t) = p_{T|X=x}(t) + \epsilon(\nabla_x p_{T|X=x}(t) \cdot u) + \epsilon^2(u^\top \nabla_x^2 p_{T|X=x}(t) u) + o(\epsilon^2).
$$

Thus, we may write

$$
\begin{aligned}
D(p_{T|X=x+\epsilon u} \| p_{T|X=x}) &= \int p_{T|X=x+\epsilon u}(t) \log \frac{p_{T|X=x+\epsilon u}(t)}{p_{T|X=x}} dt \\
&= \int \left[p_{T|X=x}(t) + \epsilon(\nabla_x p_{T|X=x}(t) \cdot u) + \epsilon^2(u^\top \nabla_x^2 p_{T|X=x}(t) u) + o(\epsilon^2)\right] \times
\end{aligned}
$$

$$
\begin{aligned}
&\log \frac{p_{T|X=x}(t) + \epsilon(\nabla_x p_{T|X=x}(t) \cdot u) + \epsilon^2 (u^\top \nabla_x^2 p_{T|X=x}(t) u) + o(\epsilon^2)}{p_{T|X=x}} dt \\
&= \int \left[p_{T|X=x}(t) + \epsilon(\nabla_x p_{T|X=x}(t) \cdot u) + \epsilon^2 (u^\top \nabla_x^2 p_{T|X=x}(t) u) + o(\epsilon^2) \right] \times \\
&\qquad \log \left(1 + \frac{\epsilon(\nabla_x p_{T|X=x}(t) \cdot u) + \epsilon^2 (u^\top \nabla_x^2 p_{T|X=x}(t) u) + o(\epsilon^2)}{p_{T|X=x}} \right) dt \\
&\overset{(a)}{=} \frac{\epsilon^2}{2} \left(\int p_{T|X=x}(t) \left(\frac{\nabla_x p_{T|X=x}(t) \cdot u}{p_{T|X=x}(t)} \right)^2 dt \right) + o(\epsilon^2) \\
&\overset{(b)}{=} \frac{\epsilon^2}{2} \Phi(T|X=x) + o(\epsilon^2). \qquad (13)
\end{aligned}
$$

Here, in (a) we have used the Taylor expansion of $\log(1+x) = x - \frac{x^2}{2} + o(x^2)$ and the equalities

$$
\int u \cdot \nabla_x p_{T|X=x}(t) dt = 0, \quad \text{and}
$$
$$
\int \left(u^\top \nabla_x^2 p_{T|X=x}(t) u \right) dt = 0,
$$

which hold under mild regularity conditions [41, Lemma 5.3, pg. 116]. In (b), we note that

$$
\begin{aligned}
\int p_{T|X=x}(t) \left(\frac{\nabla_x p_{T|X=x}(t) \cdot u}{p_{T|X=x}(t)} \right)^2 dt &\leq \int p_{T|X=x}(t) \left(\frac{\|\nabla_x p_{T|X=x}(t)\|}{p_{T|X=x}(t)} \right)^2 dt \\
&= \Phi(T|X=x).
\end{aligned}
$$

B.4 Proof of Lemma 8

Consider any $x, x' \in \mathcal{X}$ such that $\|x - x'\|_2 < \epsilon$. Note that for any y, we have

$$
|\operatorname{margin}_f(x,y) - \operatorname{margin}_f(x',y)| \leq 2 \left\| p_{T|X=x}(g(t)=\cdot) - p_{T|X=x'}(g(t)=\cdot) \right\|_{TV}.
$$

Indeed, we can write

$$
\operatorname{margin}_f(x,y) - \operatorname{margin}_f(x',y)
$$

$$
\begin{aligned}
&= \left(p_{T|X=x}(g(t)=y) - \max_{z\neq y} p_{T|X=x}(g(t)=z) \right) - \left(p_{T|X=x'}(g(t)=y) - \max_{z\neq y} p_{T|X=x'}(g(t)=z) \right) \\
&\leq \left(p_{T|X=x}(g(t)=y) - p_{T|X=x}(g(t)=y') \right) - \left(p_{T|X=x'}(g(t)=y) - p_{T|X=x'}(g(t)=y') \right) \\
&\leq \sum_y \left| p_{T|X=x}(g(t)=y) - p_{T|X=x'}(g(t)=y) \right| \\
&= 2 \left\| p_{Y|X=x}(g(t)=\cdot) - p_{Y|X=x'}(g(t)=\cdot) \right\|_{TV},
\end{aligned}
$$

where $y' = \arg\max_{z\neq y} p_{T|X=x'}(g(t)=z)$. A similar argument can be used to upper-bound the quantity $\text{margin}_f(x',y) - \text{margin}_f(x,y)$, from which the claim follows. In particular, Pinsker's inequality then implies that

$$
2 \left\| p_{T|X=x}(g(t)=\cdot) - p_{T|X=x'}(g(t)=\cdot) \right\|_{TV} \leq \sqrt{2D\left(p_{T|X=x'}(g(t)=\cdot) \| p_{T|X=x}(g(t)=\cdot) \right)}.
$$

Furthermore, by the data processing inequality, we have

$$
D\left(p_{T|X=x'}(g(t)=\cdot) \| p_{T|X=x}(g(t)=\cdot) \right) \leq D(p_{T|X=x'} \| p_{T|X=x}).
$$

Combining these inequalities with the bound (13), we then have

$$
\sup_{x'\in B_\epsilon(x)} |\,\text{margin}_f(x,y) - \text{margin}_f(x',y)| \leq \sqrt{\epsilon^2 \Phi(T|X=x) + o(\epsilon^2)}.
$$

We can then integrate over x to obtain

$$
\mathbb{E}\left[\sup_{x'\in B_\epsilon(x)} |\,\text{margin}_f(x,y) - \text{margin}_f(x',y)|^2 \right] \leq \epsilon^2 \Phi(T|X) + o(\epsilon^2).
$$

Define the event

$$
A^{\epsilon,\eta} := \left\{ x \in \mathcal{X} : \sup_{x'\in B_\epsilon(x)} |\,\text{margin}_f(x, f(x)) - \text{margin}_f(x', f(x))|^2 \leq \eta \right\}.
$$

By Markov's inequality, we can write

$$\mathbb{P}(x \notin A^{\epsilon,\eta}) = \mathbb{P}\left(\sup_{x' \in B_\epsilon(x)} |\operatorname{margin}_f(x,y) - \operatorname{margin}_f(x',y)|^2 > \eta\right) \leq \frac{\epsilon^2 \Phi(T|X) + o(\epsilon^2)}{\eta}.$$

Note that if $x \in A^{\epsilon,\eta} \cap B^\delta$, we must have $\operatorname{margin}_f(x', f(x)) > 0$ for any $x' \in B_\epsilon(x)$, implying that $f(x) = f(x')$. Hence,

$$\begin{aligned}\mathbb{P}\left(f(x') = f(x) \quad \forall x' \in B_\epsilon(x)\right) &\geq \mathbb{P}(x \in A^{\epsilon,\eta} \cap B^\eta) \geq \mathbb{P}(x \in B^\eta) - \mathbb{P}(x \notin A^{\epsilon,\eta}) \\ &\geq \mathbb{P}(x \in B^\eta) - \frac{\epsilon^2 \Phi(T|X) + o(\epsilon^2)}{\eta},\end{aligned}$$

which is inequality (6).

C Proof of Gaussian optimality

We now provide additional details on how to prove the optimality of Gaussian features when X and Y are jointly Gaussian, after having established a subadditivity result as in Lemmas 9 and 12. For brevity, we focus on the mutual information-based formulation of the robust information bottleneck; i.e., on Lemma 9. Throughout the proof, we shall use a common technique for smoothing densities to deal with technical issues. We do this by adding a small amount independent Gaussian noise to the random variables, and then taking the variance of the added Gaussian to 0 at the end [55]. Details of such proofs may be found in the references [49, 55].

Definition 1. Let X be an $\mathbb{R}^n$-valued random variable with density p_X. We say that X (or equivalently, p_X) belongs the the set $\mathcal{P}$ if:

1. $\int_{\mathbb{R}^n} p_X(x) \log(1 + p_X(x)) dx < \infty$,
2. $\mathbb{E}X = 0$, and
3. $\mathbb{E}\|X\|^2 < \infty$.

All random variables in $\mathcal{P}$ have well-defined entropies because of conditions (1) and (3). Condition (2) may be assumed without loss of generality, since the functions we consider are translation-invariant. Throughout the rest of the section, we shall assume that the random variables under consideration lie in $\mathcal{P}$.

Consider the function f_δ defined on the space of densities in $\mathcal{P}$ as

$$\begin{aligned} f_\delta(X) &= -H(C(X + \sqrt{\delta}N) + \xi) - \beta J(X + \sqrt{\delta}N) \\ &:= -H(Y_\delta) - \beta J(X_\delta), \end{aligned}$$

where $N \perp\!\!\!\perp X$ and $N \sim \mathcal{N}(0, I)$. The upper-concave envelope of f_δ is given by

$$F_\delta(X) = \sup_{p_{U|X}(\cdot|\cdot)} \left\{ -H(Y_\delta|U) - \beta J(X_\delta|U) \right\}.$$

We define a lifting of f to pairs of random variables, as follows:

$$f_\delta(X_1, X_2) = -H(Y_{1\delta}, Y_{2\delta}) - \beta J(X_{1\delta}, X_{2\delta}),$$

where we use $X_{i\delta} = X_i + \sqrt{\delta}N_i$, and $Y_{i\delta} = CX_{i\delta} + \xi_i$ for $i \in \{1, 2\}$. Note that (N_1, N_2, ξ_1, ξ_2) are independent of (X_1, X_2, U) and also of each other. Let $F_\delta(X_1, X_2)$ be the upper-concave envelope of $f_\delta(X_1, X_2)$. Just as in Lemma 9, we are able to show a subadditivity result for $F_\delta(\cdot, \cdot)$:

Lemma 14. *For any pair of random variables (X_1, X_2), we have that*

$$F_\delta(X_1, X_2) \leq F_\delta(X_1) + F_\delta(X_2).$$

Proof. Note that N_1 and N_2 are independent of (X_1, X_2, U) and also of each other. We have

$$f_\delta(X_1, X_2|U) := -H(Y_{1\delta}, Y_{2\delta}|U) - \beta J(X_{1\delta}, X_{2\delta}|U)$$

$$
\begin{aligned}
&\overset{(a)}{=} (-H(Y_{1\delta}|U) - \beta J(X_{1\delta}|X_{2\delta}, U)) \\
&\qquad + (-H(Y_{2\delta}|Y_{1\delta}, U) - \beta J(X_{2\delta}|X_{1\delta}, U)) \\
&\overset{(b)}{\leq} (-H(Y_{1\delta}|U) - \beta J(X_{1\delta}|U)) \\
&\qquad + (-H(Y_{2\delta}|X_{1\delta}, U) - \beta J(X_{2\delta}|X_{1\delta}, U)) \\
&= f_\delta(X_1|U) + f_\delta(X_2|X_{1\delta}, U) \\
&\overset{(c)}{\leq} F_\delta(X_1) + F_\delta(X_2).
\end{aligned}
$$

Here, (a) follows from the chain rule of Fisher information in Lemma 26, (b) follows from the data processing inequality for Fisher information in Lemma 27, and (c) follows from the definition of F. □

We also have a simple corollary of Lemma 14:

Corollary 1. *If X_1 and X_2 are independent random variables, then*

$$F_\delta(X_1, X_2) = F_\delta(X_1) + F_\delta(X_2).$$

Proof. Using Lemma 14, we have

$$F_\delta(X_1, X_2) \leq F_\delta(X_1) + F_\delta(X_2).$$

To prove the reverse direction, notice that we may always choose $U = (U_1, U_2)$ such that

$$(X_1, U_1) \perp\!\!\!\perp (X_2, U_2).$$

For such a U, we have

$$f_\delta(X_1, X_2|U) = f_\delta(X_1|U_1) + f_\delta(X_2|U_2).$$

Taking the supremum of the right hand side over all U_1 and U_2, we obtain

$$F_\delta(X_1, X_2) \geq \sup_{(U_1,U_2)} f_\delta(X_1, X_2|U) = F_\delta(X_1) + F_\delta(X_2),$$

which proves the claim. □

More importantly, we can also prove a "converse" to the above corollary:

Lemma 15. *If (X_1, X_2) and U are such that*

(a) $F_\delta(X_1, X_2) = F_\delta(X_1) + F_\delta(X_2)$, and

(b) $F_\delta(X_1, X_2) = f_\delta(X_1, X_2|U)$,

then the following results hold:

1. *For all $u \in \text{supp}(U)$, we have $X_1 \perp\!\!\!\perp X_2$ conditioned on $U = u$, and*
2. *$f_\delta(X_1|U) = F_\delta(X_1)$ and $f_\delta(X_2|U) = F_\delta(X_2)$.*

Proof. The key point is to identify the conditions under which equality holds in the proof of Lemma 9. We see that equality holds if and only if

$$J(X_{1\delta}|X_{2\delta}, U) = J(X_{1\delta}|U), \quad \text{and}$$
$$H(Y_{2\delta}|Y_{1\delta}, U) = H(Y_{2\delta}|X_{1\delta}, U).$$

The convexity of $J(\cdot)$ from Lemma 27 implies that $X_{1\delta} \perp\!\!\!\perp X_{2\delta}$ conditioned on all values of $U = u \in \text{supp}(U)$, and thus proves (1). It is not hard to show (using characteristic functions, for example) that $X_{1\delta} \perp\!\!\!\perp X_{2\delta}$ implies $X_1 \perp\!\!\!\perp X_2$. Now observe that

$$f_\delta(X_1, X_2|U) = f_\delta(X_1|U) + f_\delta(X_2|U).$$

The above equality, combined with the assumed equality $f_\delta(X_1, X_2|U) = F_\delta(X_1, X_2)$, immediately yields the equalities in (2). □

The rest of the proof closely follows the steps outlined in Geng and Nair [49, Appendix II] and Anantharam et al. [55, Section 4.3].

Definition 2. Define

$$v(\Sigma) := \sup_{\mathrm{Cov}(X)=\Sigma} f_\delta(X), \quad \text{and} \tag{14}$$

$$V(\Sigma) := \sup_{\mathrm{Cov}(X)\preceq\Sigma} F_\delta(X). \tag{15}$$

Lemma 16. *There exist random variables X^* and U^* satisfying*

1. $|\mathcal{U}^*| \leq \frac{n(n+1)}{2} + 1$, *and*
2. $\mathrm{Cov}(X^*) \preceq \Sigma$,

such that the following holds:

$$V(\Sigma) = f_\delta(X^*|U^*). \tag{16}$$

Proof sketch. Let $(X^{(t)}, t \geq 1)$ be a sequence of random variables such that $\mathrm{Cov}\, X^{(t)} = \widehat{\Sigma}$ and $f_\delta(X^{(t)}) \uparrow v(\widehat{\Sigma})$ as $t \to \infty$. Using tightness of this sequence [49, Proposition 17], we set $X^{\widehat{\Sigma}}$ to be the weak limit of $X^{(t)}$ as $t \to \infty$. Since $X_\delta^{(t)}$ satisfies the necessary regularity conditions as in Geng and Nair [49, Proposition 18], we have $H(X_\delta^{(t)}) \to H(X_\delta^{\widehat{\Sigma}})$, $H(Y_\delta^{(t)}) \to H(Y_\delta^{\widehat{\Sigma}})$. Moreover, we also have $J(X_\delta^{(t)}) \to J(X_\delta^{\widehat{\Sigma}})$ (see Lemma 20). Hence, we may conclude that $f_\delta(X^{\widehat{\Sigma}}) = v(\widehat{\Sigma})$.

Recall that $V(\Sigma)$ is defined as

$$\begin{aligned} V(\Sigma) &= \sup_{\mathrm{Cov}(X)\preceq\Sigma} F_\delta(X) \\ &= \sup_{(U,X),\mathrm{Cov}(X)\preceq\Sigma} f_\delta(X \mid U) \\ &\overset{(a)}{=} \sup_{\alpha_l\geq 0,\widehat{\Sigma}_l:\sum_{l=1}^M \alpha_l=1,\sum_{l=1}^M \alpha_l\widehat{\Sigma}_l\preceq\Sigma} \sum_{l=1}^M \alpha_l v(\widehat{\Sigma}_l), \end{aligned} \tag{17}$$

where, for the moment, M ranges over positive integers of arbitrary size. The equality in (a) is because we may restrict $p_{X|T}(\cdot|U)$ to the class of optimizers $X^{\widehat{\Sigma}}$ for $\widehat{\Sigma} \succeq 0$. Fenchel's extension of Carathéodory's Theorem [61, Theorem 1.3.7] implies that we can fix M to be $\binom{n+1}{2} + 1$ in equation (17) (see Lemma 4.3 in Anantharam et al. [55] for details.)

Consider any sequence of convex combinations $\left(\{\alpha_l^{(t)}\}_{l=1}^M, \{\widehat{\Sigma}_l^{(t)}\}_{l=1}^M\right)$ with $\sum_{l=1}^m \alpha_l^{(t)} \widehat{\Sigma}_l^{(t)} \preceq \Sigma$ for all $t \geq 1$, and such that $\sum_{l=1}^m \alpha_l^{(t)} v(\widehat{\Sigma}_l^{(t)})$ converges to $v(\Sigma)$ as $t \to \infty$. Since the M-dimensional simplex is compact, we may assume that $\alpha_l^{(t)} \to \alpha_l^*$ for all $l \in [M]$. Suppose that for some l, we have $\alpha_l^{(t)} \to 0$ as $t \to \infty$. Then noticing that $\alpha_l^{(t)} \widehat{\Sigma}_l^{(t)} \preceq \Sigma$, we have

$$\begin{aligned} v(\widehat{\Sigma}_l^{(t)}) &\overset{(a)}{\leq} -H(\xi) - \beta \operatorname{tr}((\widehat{\Sigma}_l^{(t)} + \delta I)^{-1}) \\ &\overset{(b)}{\leq} -H(\xi) - \beta \operatorname{tr}((\frac{\Sigma}{\alpha_l^{(t)}} + \delta I)^{-1}) \\ &\overset{(c)}{=} -H(\xi) - \beta \sum_{i=1}^n \left(\frac{\lambda_i}{\alpha_l^{(t)}} + \delta\right)^{-1}. \end{aligned}$$

In (a), we used $H(Y_\delta) \geq H(\xi)$, and the fact that Gaussians minimize Fisher information for a fixed variance. In (b), we used $\alpha_l^{(t)} \widehat{\Sigma}_l^{(t)} \preceq \Sigma$, and in (c) we let λ_i for $i \in [n]$ be the eigenvalues of Σ. It is now clear that if $\alpha_l^{(t)} \to 0$, then $\alpha_l^{(t)} v(\widehat{\Sigma}_l^{(t)}) \to 0$, as well. So we may assume without loss of generality that $\min_{l \in [M]} \alpha_l^* = \alpha_{\min} > 0$. This implies that $\widehat{\Sigma}_l^{(t)} \preceq \frac{2\Sigma}{\alpha_{\min}}$ for all large enough t, so we can find a convergent subsequence such that $\widehat{\Sigma}_l^{(t)} \to \Sigma_l^*$ for each $1 \leq l \leq M$ when $t \to \infty$. This leads us to the equation

$$V(\Sigma) = \sum_{l=1}^M \alpha_l^* v(\Sigma_l^*).$$

In other words, we can find a pair of random variables (X^*, U^*) with $|\mathcal{U}^*| \leq M$ such that $V(\Sigma) = f_\delta(X^*|U^*)$. This completes the proof. □

Having shown existence, we note a simple property stated below that is crucial for establishing Gaussian optimality:

Lemma 17. *Consider random variables (X_1, X_2, U), and define new random variables X_+ and X_- via*

$$X_+ := \frac{X_1 + X_2}{\sqrt{2}}, \quad \text{and} \quad X_- := \frac{X_1 - X_2}{\sqrt{2}}.$$

Then $f_\delta(X_1, X_2|U) = f_\delta(X_+, X_-|U)$.

Proof sketch. The proof follows by noticing that entropy and Fisher information functionals are invariant to unitary transformations, and we omit the details here. □

In the next lemma, we use all of the prior lemmas in a remarkable sequence of inequalities that yield Gaussian optimality.

Lemma 18. *Let the random variables X^* and U^* be as in Lemma 16; i.e., satisfying the equality $V(\Sigma) = f_\delta(X^*|U^*)$, and with $|\mathcal{U}^*| \leq M$. Consider two independent and identically distributed copies of (X^*, U^*), denoted by (X_1, U_1) and (X_2, U_2). Define new random variables X_+ and X_-, as follows:*

$$X_+ := \frac{X_1 + X_2}{\sqrt{2}}, \quad \text{and} \quad X_- := \frac{X_1 - X_2}{\sqrt{2}}.$$

Also define $U := (U_1, U_2)$. Then the following results hold:

(a) X_+ and X_- are conditionally independent given U.

(b) $V(\Sigma) = f_\delta(X_+|U)$ and $V(\Sigma) = f_\delta(X_-|U)$.

Proof. We have the following sequence of inequalities:

$$\begin{aligned} 2V(\Sigma) &\overset{(a)}{=} f_\delta(X_1|U_1) + f_\delta(X_2|U_2) \\ &\overset{(b)}{=} f_\delta(X_1, X_2|U_1, U_2) \\ &\overset{(c)}{=} f_\delta(X_+, X_-|U_1, U_2) \\ &\overset{(d)}{\leq} F_\delta(X_+, X_-) \end{aligned}$$

$$\overset{(e)}{\leq} F_\delta(X_+) + F_\delta(X_-)$$
$$\overset{(f)}{\leq} V(\Sigma) + V(\Sigma) = 2V(\Sigma).$$

Here, (a) follows from the assumption that $f_\delta(X^*|U^*) = V(\Sigma)$. Equality (b) follows from the independence $(X_1, U_1) \perp\!\!\!\perp (X_2, U_2)$. Equality (c) holds because of Lemma 17. Inequality (d) follows from the definition of $F_\delta(\cdot)$. Inequality (e) follows from the Lemma 14. Finally, inequality (f) follows from the definition in equation (15), and the fact that X_+ and X_- have the same covariance as X^*, which is bounded above by Σ in the positive semidefinite partial order.

Since the first and last expressions match, all the inequalities in the above sequence of inequalities must be equalities. In particular, equalities (d) and (e) combined with Lemma 15 imply that $X_+ \perp\!\!\!\perp X_-$ conditioned on (U_1, U_2), thus establishing part (a) of the lemma. Lemma 15 also gives $f_\delta(X_+|U_1, U_2) = F_\delta(X_+)$ and $f_\delta(X_-|U_1, U_2) = F_\delta(X_-)$. Finally, equality in (f) gives $F_\delta(X_+) = V(\Sigma)$ and $F_\delta(X_-) = V(\Sigma)$. This completes the proof of part (b). □

Finally, we show that the optimizer may be chosen to be a single and unique Gaussian distribution.

Lemma 19. *There exists $G^* \sim \mathcal{N}(0, \Sigma^*)$ such that $\Sigma^* \preceq \Sigma$ and $V(\Sigma) = f(G^*)$. Furthermore, the random variable G^* is the unique zero-mean element with covariance less than Σ satisfying $f_\delta(X) = V(\Sigma)$.*

Proof. The proof is identical in all respects to that of Geng and Nair [49, Theorem 1] and Anantharam et al. [55, Lemma 4.6], so we omit it here. □

Lemma 20. *Consider a sequence of random variables $\{X^{(t)}\}_{t\geq 0}$ such that $X^{(t)}$ converges weakly to X as $t \to \infty$. As before, denote $X^{(t)} + \sqrt{\delta}N := X_\delta^{(t)}$, and similarly for X_δ. Then we have*

$$\lim_{t\to\infty} J(X_\delta^{(t)}) = J(X_\delta).$$

Proof. Let G be a standard normal random variable that is independent of X. For $\tau \in (0, \infty)$, the function $H(X + \sqrt{\tau}G)$ is known to be a smooth, monotonically increasing, concave function of τ [62, 63]. Furthermore, de Bruijn's identity [64] states that

$$\frac{d}{d\tau} H(X + \tau G)\Big|_{\tau=\tau_0} = \frac{1}{2} J(X + \sqrt{\tau_0} G).$$

Using the concavity and monotonicity properties of this function, we observe that for any $t \geq 0$, any $\eta > 0$, and any $\delta' \in (0, \delta)$, we have

$$\begin{aligned} \frac{1}{2} J(X_\delta^{(t)}) &\geq \frac{H(X_\delta^{(t)} + \sqrt{\eta} G) - H(X_\delta^{(t)})}{\eta}, \quad \text{and} \\ \frac{1}{2} J(X_\delta^{(t)}) &\leq \frac{H(X_\delta^{(t)}) - H(X_{\delta'}^{(t)})}{\delta - \delta'}, \end{aligned}$$

where $\delta' \in (0, \delta)$. Taking the lim inf in the first inequality and lim sup in the second as $t \to \infty$, we obtain

$$\begin{aligned} \liminf_{t\to\infty} \frac{1}{2} J(X_\delta^{(t)}) &\geq \frac{H(X_\delta + \sqrt{\eta} G) - H(X_\delta)}{\eta}, \quad \text{and} \\ \limsup_{t\to\infty} \frac{1}{2} J(X_\delta^{(t)}) &\leq \frac{H(X_\delta) - H(X_{\delta'})}{\delta - \delta'}. \end{aligned}$$

Since the above statements are valid for any choice of $\eta > 0$ and $\delta' \in (0, \delta)$, we may take the limit as $\eta \to 0$ and $\delta' \to \delta$ to conclude that

$$\lim_{t\to\infty} \frac{1}{2} J(X_\delta^{(t)}) = \frac{1}{2} J(X_\delta).$$

□

Finally, we shall take $\delta \to 0$ and establish Gaussian optimality results for the function $f(\cdot)$.

Lemma 21. *Define the function* $f : \mathcal{P} \to \mathbb{R}$ *as*

$$f(X) = -H(CX + \xi) - \beta J(X) := -H(Y) - \beta J(X), \tag{18}$$

and consider the optimization problem

$$\sup_{\mathrm{Cov}(X) \preceq \Sigma} f(X).$$

The supremum of the above optimization problem is attained by a Gaussian random variable with covariance $\Sigma^* \preceq \Sigma$.

Proof. Let X be such that $\mathrm{Cov}(X) \preceq K$. We know that

$$\begin{aligned} f_\delta(X) &\leq \sup_{G \sim \mathcal{N}(0,\Sigma), \Sigma \preceq K} f_\delta(G) \\ &= \sup_{\Sigma \preceq K} -\frac{1}{2} \log(2\pi e)^k |C(\Sigma + \delta I)C^\top + \Sigma_\xi| - \frac{\beta}{2} \operatorname{tr}(\Sigma + \delta I)^{-1}. \end{aligned}$$

As $\delta \to 0$, we have

$$\begin{aligned} \lim_{\delta \to 0} H(CX_\delta + \xi) &= H(CX + \xi), \\ \limsup_{\delta} J(X_\delta) &\leq J(X). \end{aligned}$$

The first equality follows from the bounded variance assumption on X, and the last inequality follows from the convexity of $J(\cdot)$. We cannot assert the equality of the Fisher information limit without making additional smoothness assumptions on the density of X; however, inequality is sufficient for our purposes here, since we may conclude that

$$f(X) \leq \liminf_{\delta \to 0} f_\delta(X).$$

However, we also have

$$\liminf_{\delta\to 0} f_\delta(X) \;\leq\; \liminf_{\delta\to 0}\sup_{\Sigma\preceq K} -\frac{1}{2}\log(2\pi e)^k|C(\Sigma\,+\,\delta I)C^\top\,+\,\Sigma_\xi|\,-\,\frac{\beta}{2}\operatorname{tr}(\Sigma\,+\,\delta I)^{-1}.$$

We claim that as $\delta \to 0$, the following equality holds:

$$\begin{aligned}
&\lim_{\delta\to 0}\left[\sup_{\Sigma\preceq K} -\frac{1}{2}\log(2\pi e)^k|C(\Sigma+\delta I)C^\top+\Sigma_\xi| - \frac{\beta}{2}\operatorname{tr}(\Sigma+\delta I)^{-1}\right]\\
&\quad= \left[\sup_{\Sigma\preceq K} -\frac{1}{2}\log(2\pi e)^k|C\Sigma C^\top+\Sigma_\xi| - \frac{\beta}{2}\operatorname{tr}\Sigma^{-1}\right]\\
&\quad= \sup_{G\sim\mathcal{N}(0,\Sigma),\Sigma\preceq K} f(G).
\end{aligned}$$

To simplify notation, denote

$$\begin{aligned}
\theta_\delta(\Sigma) &= \left[-\frac{1}{2}\log|C(\Sigma+\delta I)C^\top+\Sigma_\xi| - \frac{\beta}{2}\operatorname{tr}(\Sigma+\delta I)^{-1}\right],\\
\theta(\Sigma) &= \left[-\frac{1}{2}\log|C\Sigma C^\top+\Sigma_\xi| - \frac{\beta}{2}\operatorname{tr}(\Sigma^{-1})\right].
\end{aligned}$$

Notice that for any fixed $\Sigma \succeq 0$, we have $\Theta_\delta(\Sigma) \to \Theta(\Sigma)$; i.e., we have pointwise convergence of these two functions on $\{\Sigma : \Sigma \preceq K\}$. Let $M = \Theta(K) - \epsilon_0$ for some small $\epsilon_0 > 0$ and pick δ_0 small enough so that $\Theta_\delta(K) \geq M$ for all $\delta < \delta_0$.

Notice that we are concerned with the maxima of $\Theta(\cdot)$ and $\Theta_\delta(\cdot)$ on the space $\Sigma \preceq K$. Without loss of generality, we may restrict to the space

$$\mathcal{R} := \{\Sigma \;:\; \Sigma \preceq K\} \cap \{\Sigma : \lambda_{\min}(\Sigma) > \lambda^*_{\min} - \delta_0\},$$

for a careful choice of $\lambda^*_{\min}$ and any small enough $\delta_0 < \lambda^*_{\min}$. This is because we have

$$\Theta(\Sigma) \leq -\frac{1}{2}\log|\Sigma_\xi| - \frac{\beta}{2}\operatorname{tr}(\Sigma^{-1})$$

$$= -\frac{1}{2}\log|\Sigma_\xi| - \beta\sum_{i=1}^{n}\frac{1}{\lambda_i}.$$

Note that $\lambda_{\max}(\Sigma) \leq \lambda_{\max}(K)$, so if $\lambda_i \to 0$ for any i, we will have $\Theta(\Sigma) \to -\infty$. So by imposing some threshold (M in this instance), we may rule out any matrix Σ whose smallest eigenvalue is so small that $\Theta(\Sigma) < M$, hence is not a contender for the supremum of $\Theta(\cdot)$. Suppose the lower bound obtained in this way is $\lambda^*_{\min}$. Observing that $\Theta_\delta(\Sigma) = \Theta(\Sigma + \delta I)$, we see that the eigenvalues of $\Sigma + \delta I$ should also be larger than $\lambda^*_{\min}$, so we may define the region $\mathcal{R}$ as shown above.

It is now easy to check that over the set $\mathcal{R}$, the function $\Theta(\cdot)$ and $\Theta_\delta(\cdot)$ converge uniformly, implying that

$$\sup_{\Sigma\preceq K}\Theta(\Sigma) = \sup_{\Sigma\in\mathcal{R}}\Theta(\Sigma) = \lim_{\delta\to 0}\sup_{\Sigma\in\mathcal{R}}\Theta_\delta(\Sigma) = \lim_{\delta\to 0}\sup_{\Sigma\preceq K}\Theta_\delta(\Sigma),$$

which proves the claim. □

D Derivations for Gaussians

In this appendix, we provide the details of the linear-algebraic derivations used to obtain the formulas for the optimal feature maps in the theorems of Section 4.

D.1 Proof of Theorem 1

Let (X_G, Y_G) be jointly Gaussian with $\mathrm{Cov}(X_G) = K$ and $Y_G = CX_G + \xi$, where $\mathrm{Cov}(\xi) = \Sigma_\xi = \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$. From Lemma 9 in the proof of Gaussian optimality, we see that it is enough to solve the optimization problem

$$\sup_{X\sim\mathcal{N}(0,\Sigma),\Sigma\preceq K}\left\{-H(CX+\xi) - \beta J(X)\right\},$$

since the optimum X^* of the above expression can be used to deduce the optimum T for (X_G, Y_G). For an $\mathbb{R}^d$-valued Gaussian random variable with covariance Σ, we have closed-form expressions for the entropy and Fisher information: $H(X) = \frac{d}{2}\log 2\pi e + \frac{1}{2}\log\det\Sigma$ and $J(X) = \operatorname{tr}(\Sigma^{-1})$. Using these formulas and ignoring constant terms, we can rewrite the optimization problem as

$$\inf_{\Sigma \preceq K} \left\{ \frac{1}{2}\log\det(C\Sigma C^\top + \Sigma_\xi) + \beta \operatorname{tr}(\Sigma^{-1}) \right\}.$$

Setting $B := \Sigma_\xi^{-1/2} C$ and ignoring terms that do not depend on Σ, we further obtain

$$\inf_{\Sigma \preceq K} \left\{ \frac{1}{2}\log\det(B\Sigma B^\top + I) + \beta \operatorname{tr}(\Sigma^{-1}) \right\}.$$

We now plug in $K = \sigma_x^2 I$. Writing the SVD $\frac{\Sigma}{\sigma_x^2} = UDU^\top$, we have the equivalent optimization problem

$$\min_{\substack{0 \preceq D \preceq I \\ U^\top U = I = UU^\top}} \left\{ \frac{1}{2}\log\det(\sigma_x^2 BUDU^\top B^\top + I) + \frac{\beta}{\sigma_x^2}\operatorname{tr}(D^{-1}) \right\}. \tag{19}$$

We now write the SVD $B = V\Lambda W^\top$, where $V \in \mathbb{R}^{k\times k}$ is an orthogonal matrix and $W \in \mathbb{R}^{p\times k}$ satisfies $W^\top W = I_k$. Then

$$\det(\sigma_x^2 BUDU^\top B^\top + I) = \det\left(V(\sigma_x^2 \Lambda W^\top UDU^\top W\Lambda + I)V^\top\right) = \det(\sigma_x^2 \Lambda W^\top UDU^\top W\Lambda + I).$$

Note in particular that this is the only term in the expression (19) which depends on U, so we fix D for the time being and focus on optimizing this term with respect to U. Let $\tilde{U} = U^\top W$. Note that $\tilde{U}^\top\tilde{U} = W^\top UU^\top W = W^\top W = I$, and furthermore,

$$\det(\sigma_x^2 \Lambda\tilde{U}^\top D\tilde{U}\Lambda + I) = (\det(\Lambda))^2 \det(\sigma_x^2\tilde{U}^\top D\tilde{U} + \Lambda^{-2}).$$

We will apply Lemma 29 with $A = \sigma_x^2 \tilde{U}^\top D \tilde{U}$ and $B = \Lambda^{-2}$. Let

$$(a_1, \ldots, a_k) := \lambda(A)^\downarrow,$$
$$(\ell_1, \ldots, \ell_k) := \lambda(B)^\downarrow.$$

We first have the following lemma:

Lemma 22. *Let $(d_1, \ldots, d_p) := \lambda(D)^\downarrow$. We have $\lambda(A)^\downarrow \preceq \sigma_x^2 \cdot (d_1, \ldots, d_k)$.*

Proof. For $1 \leq j \leq k$, we can write

$$\sum_{i=1}^{j} a_i = \operatorname{tr}(V^\top A V),$$

where the columns $V \in \mathbb{R}^{k \times j}$ are orthonormal eigenvectors of A with eigenvalues $\{a_1, \ldots, a_j\}$. Furthermore,

$$\operatorname{tr}(V^\top A V) = \operatorname{tr}(\sigma_x^2 V^\top \tilde{U}^\top D \tilde{U} V) \leq \max_{U \in \mathbb{R}^{k \times j} : U^\top U = I_j} \operatorname{tr}(\sigma_x^2 U^\top D U),$$

since $(\tilde{U}V)^\top (\tilde{U}V) = V^\top \tilde{U}^\top \tilde{U} V = V^\top V = I$. Finally, note that Lemma 31 implies that the last quantity is equal to $\sigma_x^2 \cdot \sum_{i=1}^{j} d_i$. Thus, the desired result follows. □

Thus, Lemmas 22 and 29 together imply that

$$\lambda\left(\sigma_x^2 \tilde{U}^\top D \tilde{U} + \Lambda^{-2}\right)^\downarrow \preceq \lambda(A)^\downarrow + \lambda(B)^\downarrow \preceq \sigma_x^2 \cdot (d_1, \ldots, d_k) + \lambda(B)^\downarrow. \tag{20}$$

Furthermore, recall that the function $f : \mathbb{R}_+^p \to \mathbb{R}$ defined by $f(x) = \prod_{i=1}^p x_i$ is Schur convex, so $x \preceq y$ implies that $f(x) \leq f(y)$. Consequently, inequality (20) implies that

$$\det(\sigma_x^2 \tilde{U}^\top D \tilde{U} + \Lambda^{-2}) \leq \prod_{i=1}^{k} (\sigma_x^2 d_i + \ell_i).$$

Further note that equality is clearly achieved when $\tilde{U}$ is a matrix with columns equal to the

canonical basis vectors of $\mathbb{R}^p$, which simply picks out the k maximal diagonal entries of D and sorts them according to the diagonal entries of Λ^{-2}. Finally, we argue that this matrix $\tilde{U}$ corresponds to a choice of the orthogonal matrix U. Indeed, define $\hat{W} \in \mathbb{R}^{p \times p}$ to be an orthogonal matrix obtained by taking the first k columns equal to the columns of W, and appending with $p-k$ column vectors to obtain an orthonormal basis of $\mathbb{R}^p$. Similarly, define $\hat{U}$ to denote an orthogonal matrix with first k columns equal to the columns of $\tilde{U}$, and the remaining columns chosen to form an orthonormal basis of $\mathbb{R}^p$. If we define $U = \hat{W}\hat{U}^\top$, we have

$$U^\top W = \hat{U}\hat{W}^\top W = \hat{U} \begin{bmatrix} I_k \\ 0_{(p-k)\times k} \end{bmatrix} = \tilde{U},$$

as wanted.

Returning to the objective (19) and substituting in the optimal choice of U, we obtain the following optimization problem in terms of D:

$$\min_{0 \preceq D \preceq I} \left\{ \frac{1}{2} \log \prod_{i=1}^{k} \left(\frac{\sigma_x^2 d_i}{\ell_i} + 1 \right) + \frac{\beta}{\sigma_x^2} \sum_{i=1}^{p} \frac{1}{d_i} \right\}.$$

Clearly, the constraint is equivalent to $0 \leq d_i \leq 1$ for all i. It is easy to see that the optimum is achieved when $d_i = 1$ for all $i > k$. Furthermore, for $i \leq k$, the optimal choice of d_i corresponds to minimizing the univariate function

$$f_i(d) = \frac{1}{2} \log \left(\frac{\sigma_x^2 d}{\ell_i} + 1 \right) + \frac{\beta}{\sigma_x^2 d}$$

over the interval $d \in [0, 1]$.

Finally, from the expression (7), we see that the optimal T is then given by

$$\begin{aligned} T &= \left((\sigma_x^2 I - \Sigma^*) - \frac{1}{\sigma_x^2}(\sigma_x^2 I - \Sigma^*)^2 \right)^{-1/2} \frac{\sigma_x^2 I - \Sigma^*}{\sigma_x^2} X + \epsilon \\ &= \left(\sigma_x^2 U(I-D)U^\top - \sigma_x^2 U(I-D)^2 U^\top \right)^{-1/2} U(I-D)U^\top X + \epsilon \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sigma_x}\left(U(D-D^2)U^\top\right)^{-1/2} U(I-D)U^\top X + \epsilon \\
&= \frac{1}{\sigma_x} U(D-D^2)^{-1/2}U^\top \cdot U(I-D)U^\top X + \epsilon \\
&= \frac{1}{\sigma_x} U(D^{-1}-I)^{1/2}U^\top X + \epsilon.
\end{aligned}
$$

We may apply the rotation $U^\top$ to obtain a slightly more attractive form:

$$
T = \frac{1}{\sigma_x}(D^{-1}-I)^{1/2}U^\top X + \epsilon.
$$

In fact, we can simplify this expression even further, since the latter $p-k$ coordinates of T are irrelevant. Indeed, since $d_i = 1$ for $i > k$, we see that the last $p-k$ diagonal entries of $(D^{-1}-I)^{1/2}$ are all 0. Furthermore, if we examine the factor $U^\top = \hat{U}\hat{W}^\top$, we see that this matrix has first k rows corresponding to a permuted version of the k rows of $W^\top$, and the remaining rows identical to the rows of $\hat{W}^\top$. Multiplication by $(D^{-1}-I)^{1/2}$ results in zeroing out the last $p-k$ rows. Thus, by Lemma 3(2), we conclude that we can also express the optimal features in terms of the truncated matrices $D = \mathrm{diag}(d_1, \ldots, d_k)$ and $U = W\hat{U}^\top$, as stated in the theorem.

D.2 Proof of Theorem 2

If (T, X, Y) are jointly Gaussian with $T = AX + \epsilon$, the problem (3) is equivalent to optimizing

$$
\min_A \left\{ \frac{1}{2}\log\det\left(\Sigma_y - \Sigma_{yx}A^\top(A\Sigma_x A^\top + I)^{-1}A\Sigma_{xy}\right) + \beta\,\mathrm{tr}(A^\top A)\right\},
$$

where we have used Lemma 1 to simplify the Fisher information term. We now make the substitution $B = A\Sigma_x^{1/2}$, so we may rewrite the optimization problem as

$$
\min_B \left\{ \frac{1}{2}\log\det\left(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1/2}B^\top(BB^\top + I)^{-1}B\Sigma_x^{-1/2}\Sigma_{xy}\right) + \mathrm{tr}\left(\Sigma_x^{-1/2}B^\top B\Sigma_x^{-1/2}\right)\right\}.
$$

Furthermore, using the fact that $B^\top(BB^\top + I)^{-1}B = I - (B^\top B + I)^{-1}$, we may transform the objective into

$$\min_B \left\{ \frac{1}{2} \log\det\left((\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}) + \Sigma_{yx}\Sigma_x^{-1/2}(B^\top B + I)^{-1}\Sigma_x^{-1/2}\Sigma_{xy} \right) + \beta \operatorname{tr}\left(\Sigma_x^{-1} B^\top B \right) \right\}.$$

Factoring out $(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})$ from the first term and then ignoring it (since the term does not depend on B), we then obtain

$$\min_B \left\{ \frac{1}{2} \log\det(I + C^\top (B^\top B + I)^{-1} C) + \beta \operatorname{tr}\left(\Sigma_x^{-1} B^\top B \right) \right\},$$

where $C := \Sigma_x^{-1/2}\Sigma_{xy}(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1/2}$.

Let $H = B^\top B + I$. Then the optimization problem further simplifies to

$$\min_H \left\{ \frac{1}{2} \log\det(I + C^\top H^{-1} C) + \beta \operatorname{tr}(\Sigma_x^{-1} H) \right\}, \tag{21}$$

where the optimization is over a restricted set of psd matrices H which may be represented as $B^\top B + I$.

The following lemma shows that the objective (21) is convex in H:

Lemma 23. *The function* $\log\det(I + C^\top H^{-1} C)$ *is convex in* H.

Proof. By Sylvester's determinant identity, we have $\det(I_m + A_{m\times n} B_{n\times m}) = \det(I_n + B_{n\times m} A_{m\times n})$. Thus, we have

$$\log\det(I + C^\top H^{-1} C) = \log\det(I + CC^\top H^{-1}).$$

Using the fact that $\log\det(I + KH^{-1})$ is convex in H for $K \succeq 0$ [65, 66], we conclude the proof. □

We now take a gradient with respect to H of the objective (21). We have the following lemma:

Lemma 24. *The gradient of* $\log\det(I + C^\top H^{-1} C)$ *in* H *is* $-H^{-1}C(I + C^\top H^{-1}C)^{-1}C^\top H^{-1}$.

Proof. Consider an ϵ-perturbation of H in a direction V given by $H + \epsilon V$. We shall expand out the function $\log\det(I + C^\top (H + \epsilon V)^{-1} C)$ and identify the coefficient of ϵ. Note that

$$\begin{aligned}(H + \epsilon V)^{-1} &= (H(I + \epsilon H^{-1} V))^{-1} \\ &= (I + \epsilon H^{-1} V)^{-1} H^{-1} \\ &= (I - \epsilon H^{-1} V) H^{-1} + o(\epsilon) \\ &= H^{-1} - \epsilon H^{-1} V H^{-1} + o(\epsilon).\end{aligned}$$

Hence, we have

$$\begin{aligned}\log\det(I + C^\top (H + \epsilon V)^{-1} C) &= \log\det(I + C^\top (H^{-1} - \epsilon H^{-1} V H^{-1} + o(\epsilon)) C) \\ &= \log\det(I + C^\top H^{-1} C - \epsilon C^\top H^{-1} V H^{-1} C + o(\epsilon)) \\ &\overset{(a)}{=} -\epsilon \operatorname{tr}((I + C^\top H^{-1} C)^{-1} C^\top H^{-1} V H^{-1} C) + o(\epsilon) \\ &= -\epsilon \operatorname{tr}(H^{-1} C (I + C^\top H^{-1} C)^{-1} C^\top H^{-1} V) + o(\epsilon).\end{aligned}$$

Here, the equality in (a) follows from the fact $\nabla_X \log\det X = X^{-1}$. Thus, the gradient must be $-H^{-1}C(I + C^\top H^{-1}C)^{-1}C^\top H^{-1}$. □

By Lemma 24, we have

$$\nabla_H \log\det(I + C^\top H^{-1} C) = -H^{-1} C (I + C^\top H^{-1} C)^{-1} C^\top H^{-1}.$$

Since $\nabla_H \operatorname{tr}(\Sigma_x^{-1} H) = \Sigma_x^{-1}$, this means the optimum occurs when

$$H^{-1} C (I + C^\top H^{-1} C)^{-1} C^\top H^{-1} = \beta \Sigma_x^{-1}.$$

If we left-multiply by $C^\top$ and right-multiply by C, we obtain

$$C^\top H^{-1} C (I + C^\top H^{-1} C)^{-1} C^\top H^{-1} C = \beta C^\top \Sigma_x^{-1} C.$$

We now derive the following useful lemma:

Lemma 25. *Suppose M is any positive definite matrix. Then $C^\top M C$ is also invertible.*

Proof. Recall that $\text{rank}(C^\top M C) = \text{rank}(M^{1/2} C) = \text{rank}(C)$, since M is invertible (cf. Chapter 0 of Horn and Johnson [67]). Since C has full column rank by assumption, we conclude that $C^\top M C$ is full-rank, hence invertible. □

In particular, Lemma 25 implies that both $C^\top H^{-1} C$ and $C^\top \Sigma_x^{-1} C$ are invertible. Hence, we can write

$$\frac{1}{\beta}(C^\top \Sigma_x^{-1} C)^{-1} = (C^\top H^{-1} C)^{-1}(I + C^\top H^{-1} C)(C^\top H^{-1} C)^{-1} = (C^\top H^{-1} C)^{-2} + (C^\top H^{-1} C)^{-1}. \tag{22}$$

Now let $(C^\top \Sigma_x^{-1} C)^{-1} = U D U^\top$ be the SVD. We know that U is a unitary matrix and D is positive definite. Left-multiplying equation (22) by $U^\top$ and right-multiplying by U then gives

$$\frac{D}{\beta} = U^\top (C^\top H^{-1} C)^{-2} U + U^\top (C^\top H^{-1} C)^{-1} U = (U^\top C^\top H^{-1} C U)^{-2} + (U^\top C^\top H^{-1} C U)^{-1}. \tag{23}$$

Let $\tilde{D}$ be the diagonal matrix which solves $\tilde{D}^{-2} + \tilde{D}^{-1} = \frac{D}{\beta}$. In other words, if $\{\tilde{d}_i\}$ and $\{d_i\}$ are the diagonal entries of $\tilde{D}$ and D, respectively, we have

$$\frac{1}{\tilde{d}_i^2} + \frac{1}{\tilde{d}_i} = \frac{d_i}{\beta},$$

implying that

$$1 + \tilde{d}_i = \frac{d_i}{\beta} \tilde{d}_i^{\,2},$$

so

$$\tilde{d}_i = \frac{1 + \sqrt{1 + 4d_i/\beta}}{2d_i/\beta}.$$

It remains to find a proper assignment of H, such that $U^\top C^\top H^{-1} C U = \tilde{D}$. Consider the SVD $C = W\Lambda V^\top$. We know that Λ is invertible and V is unitary, and that $W^\top W = I_k$. Define $\tilde{W}$ to be the matrix with first columns equal to the columns of W, and the remaining columns chosen arbitrarily to form a unitary matrix. Then define $H = \tilde{W} \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \tilde{W}^\top$, where

$$J := \Lambda V^\top U \tilde{D}^{-1} U^\top V \Lambda$$

is full-rank, since it is a product of invertible matrices. We can verify that

$$H^{-1} = \tilde{W} \begin{bmatrix} \Lambda^{-1} V^\top U \tilde{D} V \Lambda^{-1} & 0 \\ 0 & I \end{bmatrix} \tilde{W}^\top,$$

and

$$U^\top C^\top H^{-1} C U = U^\top V \Lambda W^\top \tilde{W} \begin{bmatrix} \Lambda^{-1} V^\top U \tilde{D} V \Lambda^{-1} & 0 \\ 0 & I \end{bmatrix} \tilde{W}^\top W \Lambda V^\top U = \tilde{D}.$$

Thus, equation (23) is satisfied.

It remains to choose B (and $A = B\Sigma_x^{-1/2}$) appropriately. Note that

$$\tilde{W}^\top B^\top B \tilde{W} = \tilde{W}^\top H \tilde{W} - I = \begin{bmatrix} J - I & 0 \\ 0 & 0 \end{bmatrix}.$$

If we write the SVD $J - I = S\Gamma S^\top$, assuming $J - I \succeq 0$, and let $\tilde{S} \in \mathbb{R}^{p \times k}$ be the matrix formed by taking $S \in \mathbb{R}^{k \times k}$ and zero-filling the remaining $p - k$ rows, we can take

$B = \Gamma^{1/2}\tilde{S}^\top \tilde{W}^\top = \Gamma^{1/2} S^\top W^\top$. Then

$$\tilde{W}^\top B^\top B \tilde{W} = \tilde{S}\Gamma\tilde{S}^\top = \begin{bmatrix} J - I & 0 \\ 0 & 0 \end{bmatrix},$$

which is what we want.

Lastly, note that $J - I \succeq 0$ holds when β is sufficiently small (as a function of $(\Sigma_x, \Sigma_y, \Sigma_{xy})$). Indeed, we have

$$\frac{1}{\tilde{d}_i} = \frac{2d_i/\beta}{1 + \sqrt{1 + 4d_i/\beta}} \geq \frac{2d_i/\beta}{3\sqrt{d_i/\beta}} = \frac{2}{3}\sqrt{\frac{d_i}{\beta}}$$

if $\beta < \frac{d_i}{2}$, so $\tilde{D}^{-1} \succeq \frac{2}{3}\sqrt{\frac{d_{\min}}{\beta}} I$, where $d_{\min} := \min_i d_i$, implying that

$$J \succeq \Lambda V^\top U \cdot \frac{2}{3}\sqrt{\frac{d_{\min}}{\beta}} I \cdot U^\top V \Lambda = \frac{2}{3}\sqrt{\frac{d_{\min}}{\beta}} \Lambda^2.$$

Hence,

$$\lambda_{\min}(J) \geq \frac{2}{3}\sqrt{\frac{d_{\min}}{\beta}} \lambda_{\min}^2(\Lambda),$$

and we can guarantee that the final expression is at least 1 by making β sufficiently small.

D.3 Proof of Theorem 3

Suppose $T = AX + \epsilon$ with $\epsilon \sim N(0, I)$ and $A \in \mathbb{R}^{k \times p}$, and denote the subblocks of the covariance matrix of (X, Y) by Σ_x, Σ_y, and Σ_{xy}.

Using standard formulas for jointly Gaussian variables, we have

$$\mathrm{Cov}(Y|T) = \Sigma_y - \Sigma_{yt}\Sigma_t^{-1}\Sigma_{ty} = \Sigma_y - \Sigma_{yx}(A\Sigma_x A^\top + I)^{-1}\Sigma_{xy}.$$

Furthermore, we have $\mathsf{mmse}(Y|T) = \mathrm{tr}(\mathrm{Cov}(Y|T))$. In addition, Lemma 1 implies that

$\Phi(X|T) = \operatorname{tr}(A^\top A)$. Hence, we can simplify the objective to

$$\min_A \left\{ -\operatorname{tr}\left(\Sigma_{yx} A^\top (A\Sigma_x A^\top + I)^{-1} A\Sigma_{xy}\right) + \beta \operatorname{tr}(A^\top A) \right\}.$$

Now let $B = A\Sigma_x^{1/2}$. The objective then becomes

$$\min_B \left\{ -\operatorname{tr}\left(\Sigma_{yx}\Sigma_x^{-1/2} B^\top (BB^\top + I)^{-1} B\Sigma_x^{-1/2}\Sigma_{xy}\right) + \beta \operatorname{tr}\left(\Sigma_x^{-1/2} B^\top B \Sigma_x^{-1/2}\right) \right\}. \tag{24}$$

Also note that by the Woodbury matrix formula, we can write

$$(B^\top B + I)^{-1} = I - B^\top (I + BB^\top)^{-1} B,$$

so that the objective (24) is equal to

$$-\operatorname{tr}\left(\Sigma_{yx}\Sigma_x^{-1/2}\left(I - (B^\top B + I)^{-1}\right)\Sigma_x^{-1/2}\Sigma_{xy}\right) + \beta \operatorname{tr}\left(\Sigma_x^{-1/2} B^\top B \Sigma_x^{-1/2}\right).$$

In particular, the optimization problem (24) may be viewed as an optimization problem over $B^\top B$. We write $B^\top B = UDU^\top$, where $D \in \mathbb{R}^{p\times p}$ is a diagonal psd matrix (possibly noninvertible) and $U \in \mathbb{R}^{p\times p}$ satisfies $U^\top U = I = UU^\top$. We may then equivalently optimize

$$\begin{aligned}
&\min_B \left\{ \operatorname{tr}\left(\Sigma_{yx}\Sigma_x^{-1/2}(B^\top B + I)^{-1}\Sigma_x^{-1/2}\Sigma_{xy}\right) + \beta \operatorname{tr}\left(\Sigma_x^{-1/2} B^\top B \Sigma_x^{-1/2}\right) \right\} \\
&= \min_{D,U} \left\{ -\operatorname{tr}\left(\Sigma_{yx}\Sigma_x^{-1/2} U(I + D)^{-1} U^\top \Sigma_x^{-1/2}\Sigma_{xy}\right) + \beta \operatorname{tr}\left(\Sigma_x^{-1/2} UDU^\top \Sigma_x^{-1/2}\right) \right\}.
\end{aligned}$$

A final transformation, using the cyclic property of trace, then gives

$$\min_{D,U} \left\{ \operatorname{tr}\left((D + I)^{-1} \cdot U^\top \Sigma_x^{-1/2}\Sigma_{xy}\Sigma_{yx}\Sigma_x^{-1/2} U\right) + \operatorname{tr}\left(D \cdot \beta U^\top \Sigma_x^{-1} U\right) \right\}. \tag{25}$$

Plugging in $\Sigma_x = \sigma_x^2 I$, we need to solve

$$\min_{D,U} \left\{ \operatorname{tr}\left((D+I)^{-1} \cdot U^\top \Sigma_{xy} \Sigma_{yx} U \right) + \operatorname{tr}(\beta D) \right\}. \tag{26}$$

We claim that the optimum U for a fixed D does not depend on the value of D. Indeed, suppose WLOG that $D = \operatorname{diag}(d_1, \ldots, d_p)$, where $d_1 \leq \cdots \leq d_p$. Applying the lower bound of Lemma 30 with $A = (D+I)^{-1}$ and $B = U^\top \Sigma_{xy} \Sigma_{yx} U$, we see that

$$\operatorname{tr}\left((D+I)^{-1} \cdot U^\top \Sigma_{xy} \Sigma_{yx} U \right) \geq \sum_{i=1}^{p} \frac{1}{d_i + 1} \cdot \lambda_i, \tag{27}$$

where $\lambda_1 \leq \cdots \leq \lambda_p$ are the eigenvalues of $U^\top \Sigma_{xy} \Sigma_{yx} U$. However, the eigenvalues $\{\lambda_i\}$ are also the eigenvalues of $\Sigma_{xy}\Sigma_{yx}$, which do not depend on U. Furthermore, the lower bound (27) is achieved when the columns of U are the eigenvectors of $\Sigma_{xy}\Sigma_{yx}$, in increasing order from left to right according to the values of the λ_i's.

We now turn to optimizing with respect to D. Returning to the objective (26), we need to minimize

$$\min_{d_1,\ldots,d_p} \sum_{i=1}^{p} \left(\frac{1}{d_i + 1} \cdot \lambda_i + \beta d_i \right).$$

Note that we may optimize each term separately; setting $f(d) = \frac{\lambda}{d+1} + \beta d$, we see that $f'(d) = \frac{-\lambda}{(d+1)^2} + \beta$, so taking $f'(d) = 0$ gives $d = \sqrt{\frac{\lambda}{\beta}} - 1$. Thus, we conclude that the optimum is achieved when

$$d_i = \begin{cases} \sqrt{\frac{\lambda_i}{\beta}} - 1 & \text{if } \lambda_i \geq \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

We now need to transform this joint solution (U, D) into a choice of the matrices B and A. If we construct B to be $D^{1/2}U^\top$, where D is the $p \times p$ matrix with values given by equation (28), then we will have $B^\top B = UDU^\top$. We may then take $A = B\Sigma_x^{-1/2} = \frac{1}{\sigma_x} D^{1/2} U^\top$. Lastly, note that by Lemma 3(2), we may safely ignore the entries of D and U corresponding to the zero eigenvalues, leading to the statement of the theorem.

E Properties of Fisher information

For the results in this appendix, we will assume all random variables have densities that are twice differentiable and have tails that decay sufficiently quickly. For instance, this happens when the random variables under consideration are preprocessed by convolving with a small amount of Gaussian noise. The precise technical conditions, which may be found in Lehmann and Casela [41, Lemma 5.3, pg. 116], essentially allow for interchanging of the integral and differential operators.

We begin with a few definitions:

Definition 3. The *Fisher information* of X is

$$J(X) = J(p_X) = \int_{\mathbb{R}^d} p_X(x) \|_2 \nabla \log p_X(x)\|^2 \, dx = \mathbb{E}\left[\|\nabla \log p_X(X)\|_2^2\right].$$

Definition 4. The *Fisher information matrix* of X is

$$\mathbf{J}(X) = \mathbf{J}(p_X) \;=\; \int_{\mathbb{R}^d} p_X(x)(\nabla \log p_X(x))(\nabla \log p_X(x))^\top dx \;=\; -\int_{\mathbb{R}^d} p_X(x)\nabla^2 \log p_X(x)\, dx,$$

so that $J(X) = \mathrm{tr}(\mathbf{J}(X))$.

Definition 5. The *mutual Fisher information* is the mutual version of the Fisher information:

$$J(X;Y) = J(Y \mid X) - J(Y),$$

where $J(Y|X) = \int J(Y|X = x)p_X(x)dx$.

Definition 6. The *mutual Fisher information matrix* is the mutual version of the Fisher information matrix:

$$\mathbf{J}(X;Y) = \mathbf{J}(Y \mid X) - \mathbf{J}(Y),$$

so that $J(X;Y) = \mathrm{tr}(\mathbf{J}(X;Y))$.

Definition 7. The *statistical Fisher information* of X with respect to Y is

$$\Phi(X \mid Y) = \int_{\mathbb{R}^n} p_Y(y) \int_{\mathbb{R}^n} p_{X|Y}(x \mid y) \|\nabla_y \log p_{X|Y}(x \mid y)\|_2^2 dx\, dy.$$

Definition 8. The *statistical Fisher information matrix* of X with respect to Y is

$$\begin{aligned}\mathbf{\Phi}(X \mid Y) &= \int_{\mathbb{R}^n} p_Y(y) \int_{\mathbb{R}^n} p_{X|Y}(x \mid y)(\nabla_y \log p_{X|Y}(x \mid y))(\nabla_y \log p_{X|Y}(x \mid y))^\top dx\, dy \\ &= -\int_{\mathbb{R}^n} p_Y(y) \int_{\mathbb{R}^n} p_{X|Y}(x \mid y) \nabla_y^2 \log p_{X|Y}(x \mid y) dx\, dy,\end{aligned}$$

so that $\Phi(X \mid Y) = \operatorname{tr}(\mathbf{\Phi}(X \mid Y))$.

Lemma 26. *Under mild regularity conditions, the random variables (X, Y) satisfy the identities*

$$\begin{aligned}J(X;Y) &= \Phi(X \mid Y), \\ \mathbf{J}(X;Y) &= \mathbf{\Phi}(X \mid Y).\end{aligned}$$

Proof. From the factorization

$$p_X(x) p_{Y|X}(y \mid x) = p_{XY}(x, y) = p_Y(y) p_{X|Y}(x \mid y),$$

we have

$$-\nabla_y^2 \log p_{Y|X}(y \mid x) = -\nabla_y^2 \log p_Y(y) - \nabla_y^2 \log p_{X|Y}(x \mid y),$$

where ∇_y^2 denotes the Hessian matrix of second derivatives with respect to y. We integrate both sides with respect to $p_{XY}(x, y)$. On the left-hand side, we obtain the expected Fisher information matrix $\mathbf{J}(Y \mid X)$. The first term on the right-hand side gives $\mathbf{J}(Y)$, while the

second term gives $\mathbf{\Phi}(X \,|\, Y)$. Therefore,

$$\mathbf{J}(Y \,|\, X) = \mathbf{J}(Y) + \mathbf{\Phi}(X \,|\, Y),$$

so that

$$\mathbf{J}(X; Y) = \mathbf{J}(Y \,|\, X) - \mathbf{J}(Y) = \mathbf{\Phi}(X \,|\, Y).$$

Taking the trace then gives

$$J(X; Y) = \mathrm{tr}(\mathbf{J}(X; Y)) = \mathrm{tr}(\mathbf{\Phi}(X \,|\, Y)) = \Phi(X; Y).$$

□

Lemma 27. *The Fisher information function J is convex; i.e., for any distributions p and q on $\mathbb{R}^n$ and any $\lambda \in (0, 1)$, we have*

$$J(\lambda p + (1 - \lambda)q) \leq \lambda J(p) + (1 - \lambda)J(q).$$

Furthermore, equality holds in the above inequality if and only if p and q are identical.

Proof. Our proof is adapted from Bobkov et al. [68]. Consider the function from $\mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$ given by $g(u, v) := g(u_1, u_2, \ldots, u_n, v) = \sum_{i=1}^n \frac{u_i^2}{v}$. It is easy to check that the Hessian of this function is

$$\frac{2}{v} \cdot \begin{pmatrix} I & a \\ a^\top & b \end{pmatrix},$$

where $b = \sum_{i=1}^n \frac{u_i^2}{v^2}$ and $a = \left(\frac{-u_1}{v}, \ldots, \frac{-u_n}{v}\right)^\top$. Using Schur complements, we see that the Hessian is positive semi-definite for all points in $\mathbb{R}^n \times \mathbb{R}_+$. Furthermore, the (unique) direction of zero curvature at (u, v) is seen to be (u, v). Using the convexity of g and considering the

points $(\nabla p, p)$ and $(\nabla q, q)$, we see that

$$\frac{\|\lambda\nabla p(x) + (1-\lambda)\nabla q(x)\|_2^2}{\lambda\nabla p(x) + (1-\lambda)\nabla q(x)} \leq \lambda\frac{\|\nabla p(x)\|_2^2}{p(x)} + (1-\lambda)\frac{\|\nabla q(x)\|_2^2}{q(x)}.$$

Integrating the above inequality over all x, we conclude that

$$J(\lambda p + (1-\lambda)q) \leq \lambda J(p) + (1-\lambda)J(q).$$

Equality holds if and only if for every $x \in \mathbb{R}^n$, we have $(\nabla p, p) \propto (\nabla q, q)$, which is only possible if p and q are identical. □

Corollary 2. *We have the inequality $J(X;Y) \geq 0$. Furthermore, $J(X;Y) = 0$ if and only if $X \perp\!\!\!\perp Y$.*

Proof. By convexity of J, we have $J(Y|X) \geq J(Y)$, or equivalently, $J(X;Y) \geq 0$. □

Corollary 3 (Data processing inequality)**.** *If $\Theta \to X \to Y$ is a Markov chain, then $J(\Theta|X) \geq J(\Theta|Y)$.*

Proof. By the above corollary, we have $J(\Theta|X,Y) \geq J(\Theta|Y)$. Using the Markov chain $\Theta \to X \to Y$ gives $J(\Theta|X,Y) = J(\Theta|X)$, completing the proof. □

Remark 4. *Convexity of $J(\cdot)$ is essentially identical to the data-processing properties of Fisher information. It is worth remarking that Zamir [43] established a chain rule for the statistical Fisher information, as follows:*

$$\mathbf{\Phi}(X,Y|\Theta) = \mathbf{\Phi}(X|\Theta) + \mathbf{\Phi}(Y|\Theta\|X),$$

where $\mathbf{\Phi}(Y|\Theta\|X)$ is interpreted as the average $\mathbf{\Phi}(p_{Y|X=x}|\Theta)$ over the marginal of X. This also implies that if $\Theta \to X \to Y$, then $\mathbf{\Phi}(X,Y|\Theta) = \mathbf{\Phi}(X|\Theta) = \mathbf{\Phi}(Y|\Theta) + \mathbf{\Phi}(X|\Theta\|Y)$, so that $\mathbf{\Phi}(X|\Theta) \geq \mathbf{\Phi}(Y|\Theta)$. Stated differently, $J(\Theta|X) \geq J(\Theta|Y)$.

Lemma 28. *Let X and Y be random variables taking values on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Then $J(X,Y) = J(X|Y) + J(Y|X)$.*

Proof. The Fisher information matrix $\tilde{J}(X,Y)$ is given by

$$\mathbf{J}(X,Y) \;=\; -\int_{\mathbb{R}^{m+n}} p_{XY}(x,y)\nabla^2 \log p_{XY}(x,y)\, dxdy.$$

Note that the trace of $\mathbf{J}(X,Y)$ is the same as the sum of the traces of

$$\mathbf{J}(X,Y)_{1:m,1:m} \qquad \text{and} \qquad \mathbf{J}(X,Y)_{m+1:m+n,m+1:m+n}.$$

To evaluate the first quantity, we factorize $p_{XY}(x,y) = p_Y(y)p_{X|Y}(x|y)$ and notice that the Hessian involves derivatives only in x. Thus, $\nabla_x^2 \log p_{XY}(x,y) = \nabla_x^2 \log p_{X|Y}(x|y)$. It is now easy to check that

$$\mathrm{tr}(\mathbf{J}(X,Y)_{1:m,1:m}) = J(X|Y).$$

A similar argument shows that

$$\mathbf{J}(X,Y)_{m+1:m+n,m+1:m+n} = J(Y|X).$$

□

F Linear algebraic lemmas

Finally, we collect several linear algebraic lemmas that are used in our derivations. Throughout this section, let M_n denote the space of $n \times n$ matrices with real-valued entries.

Lemma 29 (Fan inequality, Theorem 4.3.47 in Horn & Johnson [67])**.** *Let $A, B \in M_n$ be Hermitian. Let $\lambda(A), \lambda(B)$, and $\lambda(A+B)$ denote the real n-vectors of eigenvalues of A, B, and $A+B$, respectively. Then $\lambda(A)^{\downarrow} + \lambda(B)^{\downarrow}$ majorizes $\lambda(A+B)$.*

Lemma 30 (Theorem 4.3.53 in Horn & Johnson [67]). *Let $A, B \in M_n$ be Hermitian and have respective vectors of eigenvalues $\lambda(A) = \{\lambda_i(A)\}_{i=1}^n$ and $\lambda(B) = \{\lambda_i(B)\}_{i=1}^n$. Then*

$$\sum_{i=1}^{n} \lambda_i(A)^{\downarrow}\lambda_i(B)^{\uparrow} \leq \mathrm{tr}(AB) \leq \sum_{i=1}^{n} \lambda_i(A)^{\downarrow}\lambda_i(B)^{\downarrow}.$$

Lemma 31 (Corollary 4.3.39 in Horn & Johnson [67]). *Let $A \in M_n$ be Hermitian with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ and suppose $1 \leq m \leq n$. Then*

$$\lambda_1(A) + \cdots + \lambda_m(A) = \max_{V \in M_{n,m}: V^\top V = I_m} \mathrm{tr}(V^\top A V).$$