Reducing the Variance of Variational Estimates of Mutual Information by Limiting the Critic's Hypothesis Space to RKHS

by

P Aditya Sreekar Sreekar, Ujjwal Tiwari, Anoop Namboodiri

in

Machine Learning
: 1
-17

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2020

Reducing the Variance of Variational Estimates of Mutual Information by Limiting the Critic's Hypothesis Space to RKHS

P Aditya Sreekar, Ujjwal Tiwari, Anoop Namboodiri
Center for Visual Information Technology
International Institute of Information Technology, Hyderabad
Email: paditya.sreekar@research.iiit.ac.in, ujjwal.t@research.iiit.ac.in, anoop@iiit.ac.in

***Abstract*—Mutual information (MI) is an information-theoretic measure of dependency between two random variables. Several methods to estimate MI from samples of two random variables with unknown underlying probability distributions have been proposed in the literature. Recent methods realize parametric probability distributions or critic as a neural network to approximate unknown density ratios. These approximated density ratios are used to estimate different variational lower bounds of MI. While, these estimation methods are reliable when the true MI is low, they tend to produce high variance estimates when the true MI is high. We argue that the high variance characteristics is due to the uncontrolled complexity of the critic's hypothesis space. In support of this argument, we use the data-driven *Rademacher complexity* of the hypothesis space associated with the critic's architecture to analyse *generalization error bound* of variational lower bound estimates of MI. In the proposed work, we show that it is possible to negate the high variance characteristics of these estimators by constraining the critic's hypothesis space to *Reproducing Hilbert Kernel Space* (RKHS), which corresponds to a kernel learned using *Automated Spectral Kernel Learning* (ASKL). By analysing the generalization error bounds, we augment the overall optimisation objective with effective regularisation term. We empirically demonstrate the efficacy of this regularization in enforcing proper bias variance tradeoff on four different variational lower bounds of MI, namely NWJ, MINE, JS and SMILE.**

I. Introduction

Mutual information is a fundamental information theoretic measure that quantifies the dependency between two random variables (RVs). Given, two RVs, X and Y, mutual information (MI), denoted by $I(X;Y)$ is defined as:

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} log \frac{d\mathbb{P}_{XY}}{d\mathbb{P}_X \otimes \mathbb{P}_Y} d\mathbb{P}_{X,Y}$$

Where, $\mathbb{P}_{XY}$ is the joint probability distribution and, $\mathbb{P}_X$ and $\mathbb{P}_Y$ are the corresponding marginal distributions. Mutual information, $I(X;Y)$ between any two RVs ranges from 0 to $+\infty$. $I(X;Y)$ is high when X and Y share considerable information or in other words have a high degree of dependency and vice-versa. It is equal to zero *iff* X and Y are mutually independent. MI has found wide applications in representation learning [1]–[4], generative modeling [5], predictive modeling [6], and information bottleneck [7]–[9]. In the learning from data paradigm, data driven applications use sample based estimation of MI, where the key challenge is in estimating MI from the samples of two random variables with unknown joint and marginal distributions.

In the big data regime, with continuous increase in sample size and data dimentionality, reliable estimation of MI using mini-batch stochastic optimisation techniques is an area of active research [10]–[14]. Classical non-parametric MI estimators that used methods like binning [15], kernel density estimation [16] and K-Nearest Neighbour based entropy estimation [17] are computationally expensive, produce unreliable estimates, and do not conform to mini-batch based optimisation strategies. To overcome these difficulties, recent estimation methods [11]–[13] couple neural networks with variational lower bounds of MI [18], [19] for differential and tractable estimation of MI. In these methods, a critic parameterized as a neural network is trained to approximate unknown density ratios. The approximated density ratios are used to estimate different variational lower bounds of MI. Belghazi *et al*. [11], Poole *et al*. [12] and Song *et al*. [13] consider the universal approximation property of the critic neural network to estimate tighter variational lower bounds of MI. However, universal approximation ability of neural networks comes at the cost of neglecting the effect of critic's unbounded complexity on variational estimation of mutul information, which leads to unstable and highly fluctuating estimates. Similar observations have been reported in literature by Ghimire *et al*.in [14].

Nguyen *et al*. [18] by analysing the bias-variance tradeoff of variational lower bound estimates of MI showed the need to regularise the complexity of the critic's hypothesis space for stable and low variance estimation of MI. Motivated by their work, we argue that these variational lower bound estimators exhibit high sensitivity to the complexity of critic's (Neural Network) hypothesis space when optimised using mini-batch stochastic gradient strategy. To support this argument, we use a data-driven measure of hypothesis space complexity called *Rademacher complexity* to bound the generalization error for variational lower bounds of MI. Using these bounds, it is shown that higher complexity of critic's hypothesis space leads to higher generalization error and hence high variance estimates. In this proposal, our critic's hypothesis space is constructed in a smooth family of functions, the *Reproducing Kernel Hilbert Space* (RKHS). This corresponds to learning a kernel using *Automated Spectral Kernel Learning* (ASKL)

[20]. ASKL parameterized functions in the RKHS as a neural network with cosine activation in the hidden layer. By using the Rademacher complexity of ASKL-RKHS, an effective regularization to control the complexity of the critic's hypothesis space has also been proposed.

Rest of the paper is organised as follows. Related literature has been reviewed in section II. In section III, we explain some crucial concepts related to our work. The discussion on related work and preliminaries is followed by a detailed explanation of our approach in the section IV where we present a thorough theoretical analysis. Supporting experimental results are demonstrated in section V. For the sake of brevity, all proofs related to our proposal are included in the Appendix.

II. Related Work

A. Mutual Information Estimation

Mutual information can be characterized as the KL divergence between joint distribution $\mathbb{P}_{XY}$ and the product of marginal distributions $\mathbb{P}_X \otimes \mathbb{P}_Y$, $I(X;Y) = D_{KL}(\mathbb{P}_{XY} \| \mathbb{P}_X \otimes \mathbb{P}_Y)$. This is the central theme in the derivation of lower bounds of MI from variational lower bounds of KL divergence. KL divergence between two multivariate probability distributions, say $\mathbb{P}$ and $\mathbb{Q}$ belongs to broader class of divergences known as the f-divergences, which are characterized by convex function f of likelihood ratio ($d\mathbb{P}/d\mathbb{Q}$). Nguyen *et al*. [18] formulated variational lower bound of f-divergences by using the convex conjugate of f and leveraged convex empirical optimization to estimate f-divergences. Belghazi *et al*. [11] proposed a tighter variational lower bound of KL divergence which is derived from the Donsker-Varadhan [19] dual representation. In their work, two MI estimators are constructed by optimizing neural network critics to maximize (1) convex conjugate lower bound, and (2) Donsker-Varadhan lower bound. In the proposed work, convex conjugate based lower bound estimator is referred to as NWJ and Donsker-Varadhan based estimator as MINE. Poole *et al*. [12] developed a unified framework for different MI estimates and created an interpolated lower bound for better bias-variance tradeoff. They also proposed a lower bound which is optimized using GAN discriminator objective [21] to estimate density ratios. We refer to the estimator based on this lower bound as JS. Song *et al*. [13] showed that variance of both MINE and NWJ estimators increase exponentially with increase in the true magnitude of MI. The explained cause of this behaviour is the increase in variance of the partition function estimates [13]. They also proposed a lower bound estimator with improved bias-variance tradeoff by clipping the partition function estimate. In the proposed work, we refer to estimator based on this lower bound as SMILE.

In this approach, instead of designing a better lower bound estimate as proposed in [11]–[13], [18], we study the effect of restricting the hypothesis space of critics to RKHS for favourable bias-variance tradeoff. The comparative performance of the proposed work reflects the effectiveness of the proposed approach in learning low variance estimates of MI. Similar to this approach, Ghimire *et al*. [14] and Ahuja *et al*. [22] also restricted critic hypothesis space to RKHS. Their methods differ from ours in the choice of kernel functions under consideration. Convex combination of Gaussian kernels were considered in [22]. A stationary Gaussian kernel with inputs transformed by a neural network with randomly sampled output weights has been proposed in [14]. In contrast to the work, we learn a kernel belonging to a much broader class of non-stationary kernels rather than restricting the kernel to Gaussian kernels.

B. Kernel Learning

Kernel methods play an important role in machine learning [23], [24]. Initial attempts included learning convex [25], [26] or non linear combination [27] of multiple kernels. While the aforementioned kernel learning methods are an improvement over the isotropic kernels, they cannot be used to adapt any arbitrary stationary kernel. To alleviate this problem [28], [29] proposed approximating kernels by learning a spectral distribution. At the core of these methods is Bochner's theorem [30], which states that there exists a duality between stationary kernels and distributions in spectral domain (Fourier domain). Similarly, Yaglom's theorem [31] states that there is a duality between the class of kernels and positive semi-definite functions with finite variations in spectral domain. Kom Samo *et al*. [32] showed that kernels constructed using Yaglom's theorem are dense in the space of kernels. Ton *et al*. [33] used Monte- Carlo integration and Yaglom's theorem to construct non-stationary kernels for Gaussian Processes. Recent methods combine deep learning with kernel learning methods. Deep Kernel Learning [34] placed a plain deep neural network as the front-end of a spectral mixture kernel to extract features, which is further extended to a kernel interpolation framework [35] and stochastic variational inference [36]. Chun-Liang Li *et al*. [37] modeled the spectral distribution as an implicit generative model parameterized by a neural network and approximated a stationary kernel by performing Monte-Carlo integration using samples from the implicit model. Hui Xue *et al*. [38] and Jian Li *et al*. [20] (ASKL) represented a non-stationary kernel as Monte-Carlo integration of fixed samples which are optimized using gradient descent methods. In this work, ASKL is used to learn the kernel corresponding to the critic's hypothesis space in Reproducing Kernel Hilbert Space.

III. Preliminary

A. Variation Lower Bounds of Mutual Information

In this subsection, four different variational lower bounds namely I_{NWJ}, I_{MINE}, I_{JS} and I_{SMILE} based estimators of MI have been discussed. These estimators are used in throughout this work. In estimating variational lower bounds of MI, a parametric probability distribution or critic f_θ with trainable parameters θ is optimised to approximate the likelihood density ratio between the joint and product of marginal distributions ($d\mathbb{P}_{XY}/d\mathbb{P}_X \otimes \mathbb{P}_Y$). The approximated density ratio is used for sample based estimation of MI. The optimisation objective is to maximize the different variational lower bounds of MI with respect to the critic parameters θ to estimate

MI.

Donsker-Varadhan dual representation [19] based variational lower bound of MI, denoted as I_{DV} is given by:

$$I(X;Y) \geq I_{DV}(f_\theta) = \mathbb{E}_{\mathbb{P}_{XY}}[f_\theta(x,y)] - log\left(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}\left[e^{f_\theta(x,y)}\right]\right) \quad (1)$$

The optimal critic for which the equality $I_{DV} = I(X;Y)$ holds in (1) is given by $f^*_{DV} = log(d\mathbb{P}_{XY}/d\mathbb{P}_X \otimes \mathbb{P}_Y)$. I_{MINE} and I_{NWJ} lower bounds can be derived from Tractable Unnormalized Barber and Argakov (TUBA) lower bound, I_{TUBA}, considering only constant positive baseline in [12], that is $a > 0$ in the I_{TUBA} formulation defined as:

$$I(X;Y) \geq I_{TUBA}(f_\theta) = \mathbb{E}_{\mathbb{P}_{XY}}[f_\theta(x,y)] - \frac{\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}\left[e^{f_\theta(x,y)}\right]}{a} - log(a) + 1 \quad (2)$$

Optimal critic satisfying the equality $I_{TUBA} = I(X;Y)$ in equation 2 is given by, $f^*_{TUBA} = log(d\mathbb{P}_{XY}/d\mathbb{P}_X \otimes \mathbb{P}_Y) + log(a)$. In this work, I_{MINE} is formulated from I_{TUBA} by fixing the parameter a in (2) as exponential moving average of $e^{f_\theta(x,y)}$ across mini-batches. Similarly, I_{NWJ} is formulated from I_{TUBA} by substituting the parameter $a = e$.

Unlike the methods described above that maximize the variational lower bounds to learn likelihood density ratio, other methods [2], [12], [13] approximate the density ratio for sample based estimation of MI by optimizing GAN discriminator objective defined as:

$$\max_\theta \mathbb{E}_{\mathbb{P}_{XY}}[log(\sigma(f_\theta(x,y)))] + \mathbb{E}_{\mathbb{P}_X \times \mathbb{P}_Y}[log(1 - \sigma(f_\theta(x,y)))] \quad (3)$$

Where, $\sigma()$ is the sigmoid function. The optimal critic maximizing the GAN discriminator objective is given by, $f^*_{GAN} = log(d\mathbb{P}_{XY}/d\mathbb{P}_X \times \mathbb{P}_Y)$. Poole *et al.* [12] observed that $f^*_{NWJ} = f^*_{GAN} + 1$, where f^*_{NWJ} is the optimal critic for I_{NWJ} and constructed another variational lower bound I_{JS} by substituting $f_{GAN}(x,y) + 1$ as the critic function f_θ into (2). The f_{GAN} is optimized using the GAN discriminator objective. Similarly, Song *et al.* [13] constructed another lower bound of MI, denoted as I_{SMILE} by substituting f_{GAN} as critic f_θ in I_{DV} expressed in (1). In [13], the bias-variance tradeoff is controlled by clipping the critic output. It is essential to note that we do not clip the output of the ASKL critic to analyse the effectiveness of restricting the critic function f_θ hypothesis space to Reproducing Kernel Hilbert Space in controlling bias-variance tradeoff.

B. Automated Spectral Kernel Learning

In this subsection we discuss Reproducing Hilbert Kernel Spaces (RKHS) and Automated Spectral Kernel Learning (ASKL). *Hilbert space* $\mathcal{H}$, is an vector space of real valued functions $f : \mathcal{X} \to \mathbb{R}$ with defined inner product $\langle f, g\rangle_\mathcal{H}$ between two functions f and g. Function norm in the hilbert space is defined as $\|f\|_\mathcal{H} = \sqrt{\langle f, f\rangle_\mathcal{H}}$. *Reproducing kernel* of a hilbert space is a positive semi-definite function, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies the conditions: (1) $K(\cdot, x) \in \mathcal{H}\ \forall x \in \mathcal{X}$, and (2) $\langle f, K(\cdot, x)\rangle_\mathcal{H} = f(x)\ \ \forall f \in \mathcal{H}\ \&\ \forall x \in \mathcal{X}$. The latter of the two condition is known as the reproducing property of the kernel K [39]. A Hilbert space which posses a reproducing kernel is called a *Reproducing Kernel Hilbert Space*.

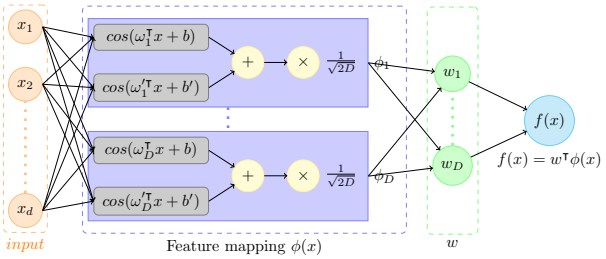


Fig. 1. Architecture of ASKL critic. The feature mapping ϕ is parameterized by the middle layer. Its weights are the frequency samples $\{\omega_i, \omega_i'\}_{i=1}^D$ sampled from spectral distribution $S(\omega, \omega')$. The output layer parameterizes the RKHS representation w of a function f such that $f(x) = w^\intercal \phi(x)$

There exist many feature mappings, $\varphi : \mathcal{X} \to \mathcal{F}$, where $\mathcal{F}$ is a Hilbert space, such that $K(x,y) = \langle \varphi(x), \varphi(y)\rangle_\mathcal{F}$ and $f(x) = \langle w, \varphi(x)\rangle_\mathcal{F}$, $w \in \mathcal{F}$, and $f \in \mathcal{H}$. A special case of such feature mappings known as implicit feature mapping is $\phi(x) = K(\cdot, x)$ and $K(x,y) = \langle \phi(x), \phi(y)\rangle_\mathcal{H}$.

Yaglom's theorem [31] as stated below shows that there exists a duality between a positive semidefinite kernel function and a non-negative Lebesgue-Stieltjes measure in spectral domain.

Theorem 1: (Yaglom's theorem) A kernel $K(x,y)$ is positive semi-definite *iff* it can be expressed as

$$K(x,y) = \int_{\mathcal{R}^d \times \mathcal{R}^d} e^{i(\omega^\intercal x - \omega'^\intercal y)} dS(\omega, \omega')$$

where, $S(\omega, \omega')$ is Lebesgue-Stieltjes measure associated to some positive semi-definite function $s(\omega, \omega')$ with bounded variations.

With appropriate scaling the Lebesgue-Stieltjes measure $S(\omega, \omega')$ can be treated as a probability distribution in spectral domain where ω and ω' are spectral variables. From here on, this probability distributions is referred to as spectral distribution. An implication of theorem 1 is that it is possible to learn an RKHS associated with a kernel by learning a spectral distribution.

Automated Spectral Kernel Learning (ASKL) [20] is a kernel learning method that used samples from the spectral distribution $S(\omega, \omega')$ to construct a feature mapping $\phi(x)$ defined as,

$$\phi(x) = \frac{1}{\sqrt{2D}}[cos(\Omega^\intercal x + b) + cos(\Omega'^\intercal x + b')] \quad (4)$$

Where, $\Omega = [\omega_1, \ldots, \omega_D]$ and $\Omega' = [\omega_1', \ldots, \omega_D']$ are $d \times D$ matrices of frequency samples $\{\omega_i, \omega_i'\}_{i=1}^D \overset{iid}{\sim} S(\omega, \omega')$ and b and b' are vectors of D uniform samples $\{b_i\}_{i=1}^D, \{b_i'\}_{i=1}^D \overset{iid}{\sim} \mathcal{U}[0, 2\pi]$. The kernel associated with the spectral distribution can be approximated using the feature mapping $\phi(x)$ defined above as $K(x,y) = \phi(x)^\intercal \phi(y)$. This feature mapping $\phi(x)$

produces a D-dimensional embedding in an RKHS for any input x. Any function in this RKHS is represented by a D-dimensional vector w, such that $f(x) = w^\intercal \phi(x)$.

ASKL represented the RKHS generated by the above feature mapping as a two layer neural network with cosine activations shown in Fig. 1. The hidden layer of this neural network represents the feature mapping $\phi(x)$, its trainable parameters are the frequency samples $\{\omega_i, \omega_i'\}$ from spectral distribution $S(\omega, \omega')$. The parameters w of the final output layer represent functions in the RKHS. The output of the final layer is the inner product $f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}}$. A RKHS can be learned by optimizing this neural network using a stochastic gradient descent method. During the optimization, a spectral distributions is learned implicitly through learning the parameters of the hidden layer $\{\omega_i, \omega_i'\}$. In this work, the critic's hypothesis space is restricted to an RKHS using the neural network architecture Fig. 1 and ASKL. For more information on ASKL refer to [20]. Any further reference to ASKL critic refers to the neural network architecture shown in Fig. 1.

IV. Theory & Our Approach

Our goal is to estimate the mutual information, $I(X;Y)$, between two RVs X and Y, from n i.i.d samples, $\{x_i, y_i\}_{i=0}^n$ from joint distribution $\mathbb{P}_{XY}$ and m i.i.d samples, $\{x_i', y_i'\}_{i=0}^m$ from the product of marginal distributions $\mathbb{P}_X \otimes \mathbb{P}_Y$. As, the true underlying probability distributions are unknown, we use empirical approximations of the variational lower bounds of MI defined as:

$$\hat{I}^{n,m}_{TUBA}(f_\theta, S) = \mathbb{E}_{\mathbb{P}^n_{XY}}[f_\theta(x,y)] - \frac{\mathbb{E}_{\mathbb{P}^m_X \otimes \mathbb{P}^m_Y}\left[e^{f_\theta(x,y)}\right]}{a} - log(a) + 1 \quad (5)$$

$$\hat{I}^{n,m}_{DV}(f_\theta, S) = \mathbb{E}_{\mathbb{P}^n_{XY}}[f_\theta(x,y)] - log\left(\mathbb{E}_{\mathbb{P}^m_X \otimes \mathbb{P}^m_Y}\left[e^{f_\theta(x,y)}\right]\right) \quad (6)$$

Where, S is the set of n,m i.i.d samples$\{x_i, y_i\}_{i=1}^n$. $\{x_i', y_i'\}_{i=1}^m$, $\mathbb{P}^n_{XY}$ and $\mathbb{P}^m_X \otimes \mathbb{P}^m_Y$ are empirical distributions corresponding to samples $\{x_i, y_i\}_{i=1}^n$ and $\{x_i', y_i'\}_{i=1}^m$, respectively, $\mathbb{E}_{\mathbb{P}^n_{XY}}[f(x,y)] = \frac{1}{n}\sum_{i=1}^n f(x_i, y_i)$ and $\mathbb{E}_{\mathbb{P}^m_X \otimes \mathbb{P}^m_Y}[f(x,y)] = \frac{1}{m}\sum_{i=1}^m f(x_i', y_i')$.

A. Theoretical Guarantees

In this subsection the generalization behaviour of the empirical estimates, $\hat{I}^{n,m}_{TUBA}$ and $\hat{I}^{n,m}_{DV}$ are discussed. We derive generalization error bound for the empirical estimates using data-driven Rademacher complexity of general critic's hypothesis space. We also bound the empirical Rademacher complexity of the ASKL critic's hypothesis space.

Generalization error quantifies the out of sample behaviour of an estimator. Formally, generalization error is defined as the maximum possible deviation of the empirical estimates from true values. If empirical estimate $\hat{I}$ is an unbiased estimate, then variance of this empirical estimate is upper bounded by the expectation of squared generalization error. Hence, generalization error is an indicator of the variance of the estimate. The following theorem bounds the generalization error of $\hat{I}^{n,m}_{TUBA}$ and $\hat{I}^{n,m}_{DV}$.

Theorem 2 (Generalization Error Bounds): Assume, that the hypothesis space $\mathcal{F}$ of the critic is uniformly bounded by M, that is $|f(x,y)| \leq M \; \forall f \in \mathcal{F} \; \& \; \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$, $M < \infty$. For a fixed $\delta > 0$ generalization errors of $\hat{I}^{n,m}_{TUBA}$ and $\hat{I}^{n,m}_{DV}$ can be bounded with probability of at least $1-\delta$, given by

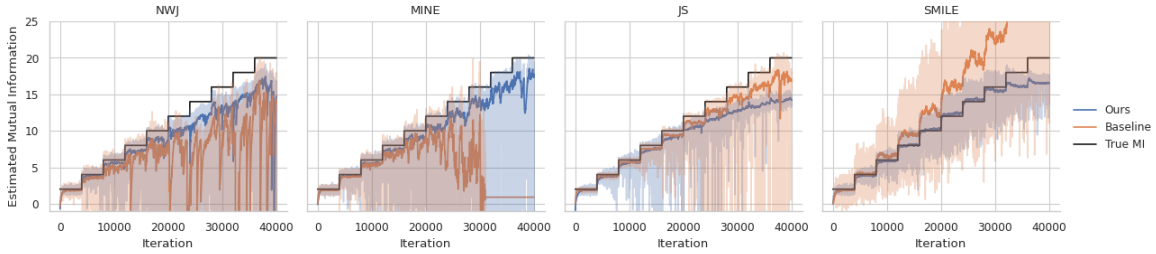
$$\sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f)\right) \leq 4\hat{\mathcal{R}}_n(\mathcal{F}) + \frac{8}{a}e^M \hat{\mathcal{R}}_m(\mathcal{F}) + \frac{4M}{n} log\left(\frac{4}{\delta}\right) + \frac{8Me^M}{am} log\left(\frac{4}{\delta}\right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^M - e^{-M})^2}{a^2 m}\right) log\left(\frac{2}{\delta}\right)}{2}} \quad (7)$$

$$\sup_{f \in \mathcal{F}} \left(I_{DV}(f) - \hat{I}^{n,m}_{DV}(f)\right) \leq 4\hat{\mathcal{R}}_n(\mathcal{F}) + 8e^{2M} \hat{\mathcal{R}}_m(\mathcal{F}) + \frac{4M}{n} log\left(\frac{4}{\delta}\right) + \frac{8Me^{2M}}{m} log\left(\frac{4}{\delta}\right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^{2M} - 1)^2}{m}\right) log\left(\frac{2}{\delta}\right)}{2}} \quad (8)$$

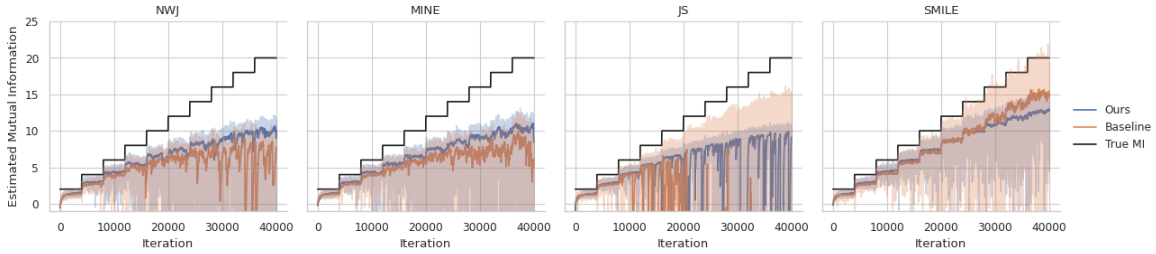
Where, sample set S for $\hat{I}^{n,m}_{TUBA}$ and $\hat{I}^{n,m}_{DV}$ is assumed to be known, and $\hat{\mathcal{R}}_n(\mathcal{F})$ and $\hat{\mathcal{R}}_m(\mathcal{F})$ are empirical Rademacher averages of the hypothesis space $\mathcal{F}$ for different sample sizes.

To formulate the generalization error bounds given in the above theorem, we used McDairmid's inequality to bound generalization error by expected generalization error over sample set S. Then we use lemma A5 given in [40] to bound the expected error by Rademacher complexity. Further, Rademecher concentration inequality, lemma A4 also given in [40] is used to arrive at the final theoretical guarantees. Refer to Appendix B for detailed proof. Error bounds for I_{NWJ} and I_{MINE} are derived by substituting the parameter a in bound 7 with e, and with exponential moving average of $e^f_\theta(x,y)$ across mini-batches, respectively. I_{JS} uses I_{NWJ} lower bound to estimate MI, hence generalization error of I_{JS} is bounded by generalization error bound of I_{NWJ}. Similarly, I_{SMILE} uses I_{DV} lower bound to estimate MI and its generalization error is bounded by error bound of I_{DV}

The generalization error bounds depend on the empirical Rademacher complexities and e^M. Our finding on the dependence of the generalization error on e^M is confirmed by similar observation made in [41] on the sample complexity of MINE estimator. From the error bounds, it can be inferred that high empirical Rademacher complexity of the critic's hypothesis space leads to high generalization error, hence high variance estimates. Therefore, variance of these estimates can be effectively reduced by choosing a hypothesis space for critic with low Rademacher complexity. However, it is also necessary to keep the hypothesis space rich enough to induce low bias. Though these bounds apply to all hypothesis spaces including the space of functions that are learned by a fully connected neural network, emperical estimation of



(a) Comparison on 20 dimensional correlated Gaussian dataset



(b) Comparision on cubed 20 dimensional correlated Gaussian dataset

Fig. 2. Qualitative comparison between ASKL and baseline critic on four diferent variational lower bounds of MI, I_{NWJ}, I_{MINE}, I_{JS}, and I_{SMILE}. MI estimates on Gaussian correlated and cubed Gaussian correlated datasets are plotted in (a) and (b), respectively. MI estimate by the proposed ASKL critic are in blue and the estimates of baseline critic are depicted in orange. The solid plotted lines are exponentially weighted moving average of these estimates. ASKL critic etimates are more stable in comparison to baseline estimates on all lower bounds of MI and both datasets. A specific case of estimation instability can be noticed in I_{MINE} (first row second plot) based estiamtion of MI using baseline critic architecture when the true MI is higher than 16, whereas, ASKL critic computes stable MI estimates even at higher values.

Rademacher complexity for a fully connected neural network is an open area of research. We restrict the critic neural networks hypothesis space to RKHS by using ASKL to gain insights into variational lower bound estimates of MI. The empirical Rademacher complexity of the ASKL critic's hypothesis space can be upper bounded as shown by the following theorem,

Theorem 3: The empirical Rademacher average of the RKHS $\mathcal{F}$ to which ASKL critic belongs can be bounded as following

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{B}{n}\sqrt{\sum_{i=1}^{n}\|\phi(x_i)\|_2^2} \leq \frac{B}{\sqrt{n}}$$

Where $B = \sup_{f\in\mathcal{F}}\|w\|_2$.

We used the Cauchy-Schwarz inequality to bound the complexity of the ASKL critic, for detailed proof refer to Appendix A. Note that, the second inequality in the above theorem is true only in the case of ASKL critic. Using the above theorem we can decrease the complexity by decreasing the largest possible norm of RKHS representation of functions w or decreasing the frobenius norm of the feature mapping matrix. In the next subsection, we present an optimization procedure to decrease the empirical Rademacher complexity by penalizing $\|w\|_2$ and $\|\phi(X)\|_F$ to control the bias-variance tradeoff. Using second inequality, and penalizing $\|w\|_2$ it is possible to carve out the regularisation used by Nguyen *et al*. [18] to control hypothesis space complexity.

TABLE I
REGULARIZATION WEIGHTS

Lower Bound	λ_1	λ_2
NWJ [42]	0.001	0.001
MINE [11]	0.001	0.001
JS [12]	1e-5	1e-5
SMILE [13]	1e-4	0.001

B. Training Methodology

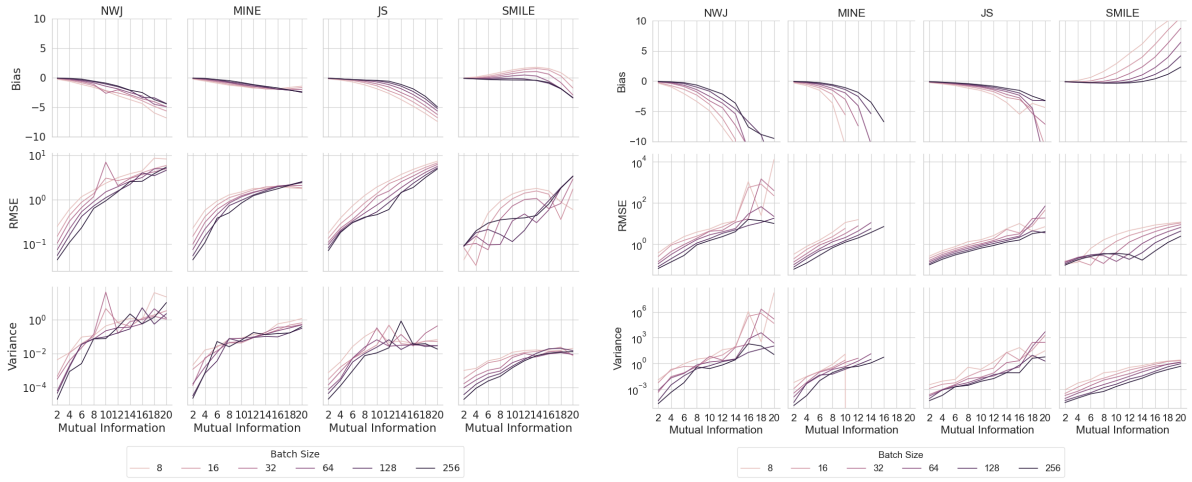
We train an ASKL critic neural network shown in Fig. 1 to simultaneously maximize empirical estimate of MI and minimize regularization terms defined below. The overall training objective is:

$$\underset{\theta}{\operatorname{argmin}} - \hat{I}(f_\theta, S) + \lambda_1\|w\|_2 + \lambda_2\|\phi(S;\theta)\|_F \qquad (9)$$

Where, $\hat{I}$ can be an empirical estimate of any variational lower bound of MI, $\hat{I}^{n,m}_{NWJ}$, $\hat{I}^{n,m}_{MINE}$, $\hat{I}^{n,m}_{JS}$ or $\hat{I}^{n,m}_{SMILE}$. And θ is the set of trainable parameters w, Ω, and Ω'. GAN discriminator objective is maximized in cases where $\hat{I}$ is $\hat{I}^{n,m}_{JS}$ or $\hat{I}^{n,m}_{SMILE}$. In this work, regularization terms $\|w\|_2$ and $\|\phi(S;\theta)\|_F$ appear in upper bound of empirical Rademacher complexity of ASKL critic's hypothesis space. Bias-variance tradeoff is controlled by tuning hyperparameters, λ_1 and λ_2. We use mini-batch stochastic gradient decent to train the estimator.

V. EXPERIMENTS

We empirically validate our claims on two different toy datasets which have been widely used by other MI estimation



(a) Bias, variance and RMSE of ASKL critic estimates for different batch sizes.

(b) Bias, variance and RMSE of baseline critic estimates for different batch sizes.

Fig. 3. Bias, variance, and RMSE values of ASKL critic and baseline critic estimates averaged over 50 experimental trials are shown in figures (a) and (b), respectively. In each figure first, second and third rows contain bias, RMSE and variance plots. Each column corresponds to different lower bound, and in each plot different plotted lines correspond to different batch sizes. ASKL critic etimates are less biased and exhibit lower variance compared to baseline critic estimates on all variational lower bounds.

methods [11]–[13], (1) correlated Gaussian dataset, where samples of two RVs (X, Y) are drawn from a 20 dimensional Gaussian distribution with correlation ρ between each dimension of X and Y. The correlation ρ is increased such that $I(X;Y)$ increases in steps of 2 every 4000 training steps, and (2) cubed Gaussian dataset, same as in (1) but we apply a cubic non-linearity to Y to get samples (x, y^3). As, mutual information remains unchanged by application of deterministic functions on random variables, $I\left(X;Y^3\right) = I(X;Y)$. Further, is it important to note that previous methods increased the correlation ρ till the true MI is increased to 10. In our experimental analysis, we increased the correlation ρ till the true MI is 20 to demonstrate that ASKL critic produces low variance estimates even at high values of MI.

For comparative analysis we train ASKL critic and a baseline critic on four different lower bounds, namely I_{NWJ}, I_{MINE}, I_{JS}, and I_{SMILE}. The baseline critic is a fully connected neural network with ReLU activations. This baseline has been used by previous estimation methods that consider the universal approximation property of neural networks [11]–[13]. ASKL critic with regularised space complexity computes low variance stable variational lower bound estimates of MI in comparison to baseline critic.

Code for this paper are available at https://cvit.iiit.ac.in/projects/mutualInfo/.

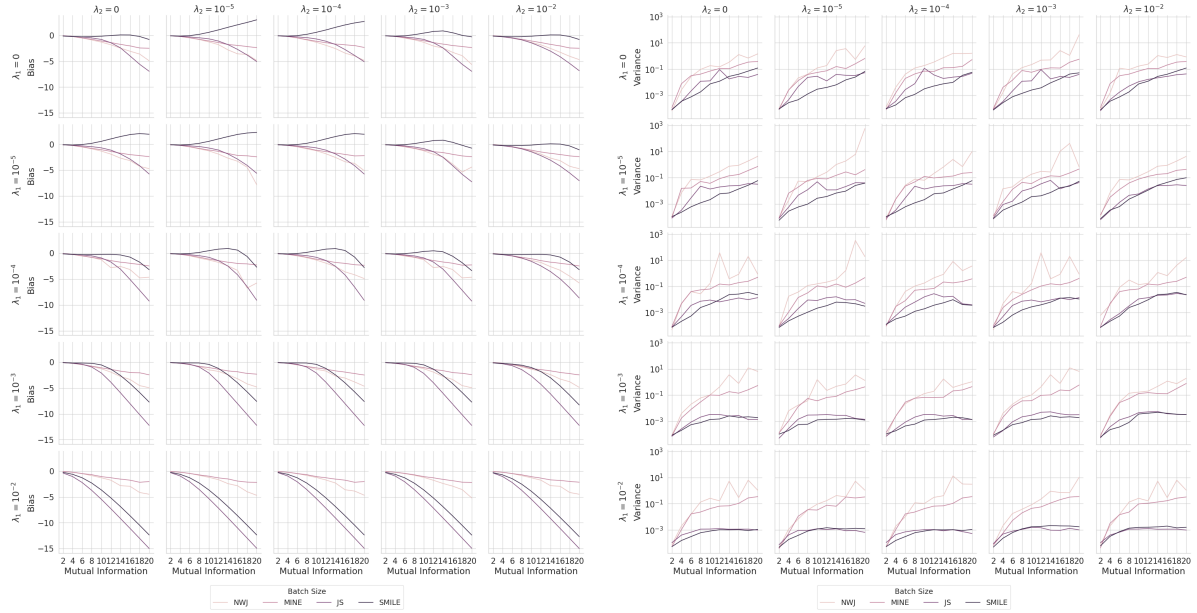
A. Training Details

For ASKL critic, D is set to 512, that is 512 spectral samples are used for estimation. The multiplicity factors for each of the regularization terms used for different estimators are given in Table I. For our baseline critic, we used a 3 layer neural network with 256 units in each hidden layer. Unless mentioned otherwise, batch size is set to 64. We use Adam optimizer [43] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Learning rates are set to 10^{-3} and 5×10^{-4} for ASKL and baseline critics, respectively.

We test the validity of our claim that constraining the critic to RKHS should lead to better bias-variance tradeoff in three different experimental setups, (1) qualitatively compare the variance of MI estimates between ASKL critic and baseline critic on four different variational lower bounds of MI. These experiments are performed on both toy datasets described above, batch size is fixed at 64 sample, (2) quantitatively compare the average bias, variance, and the root mean square error (RMSE) between the true and empirical estimates of MI over 50 experimental trials. These quantitative comparisons are made over a range of batch sizes to depict the robustness of our estimates with varying batch sizes, (3) quantitatively demonstrate the efficacy of the proposed regularisation terms in controlling bias-variance tradeoff of ASKL critic's space complexity by varying the regularisation hyperparameters λ_1 and λ_2 for $\|w\|_2$ and $\|\phi(X)\|_F$, respectively. In experiment (3), bias-variance values are estimated over 50 experiments. Both experients (1) and (2) are run on correlated Gaussian dataset. We further elaborate on each of these experimental results in the next subsection.

B. Results

Qualitative comparison between ASKL critic and baseline critic on four different variational lower bounds of MI has been shown in Fig. 2. Fig. 2(a) and Fig. 2(b) demonstrate the comparative results on the 20 dimensional correlated Gaussian dataset and the cubed correlated Gaussian dataset, respectively. In can be seen that maximisation using ASKL critic tends to produce stable estimates in comparison to their baseline counterpart. A particular instance of numerical



(a) Bias of ASKL critic based estimators for different configurations of regularization weights

(b) Variance of ASKL critic based estimators for different configurations of regularization weights

Fig. 4. Bias-variance tradeoff for different values of λ_1 and λ_2 for estimation using ASKL critic. Figures (a) and (b) show bias and variance plots, respectively. In both figures each row corresponds to a single λ_1 value and each column corresponds to a single λ_2 value. These figures quantitatively demonstrate the efficacy of the proposed regularisation terms (λ_1 and λ_2 for $\|w\|_2$ and $\|\phi(X)\|_F$, respectively) in controlling bias-variance tradeoff of ASKL critic's space complexity.

instability in baseline critic estimates can be observed in the plot corresponding to I_{MINE} when the true MI is higher than 16. Estimates by ASKL critic does not suffer from such instability and it is to be noted that the ASKL critic also produces comparatively low variance MI.

We compute bias, variance, and root mean square error of the estimated MI values to quantitatively evaluate the proposed ASKL critic's performance against the baseline. The bias, variance, and RMSE values have been averaged over 50 experimental trials. Fig. 3(a) and Fig. 3(b) show the computed values for the ASKL critic and the baseline, respectively. These plots conclusively demonstrate that the ASKL critic estimates have lower bias and variance characteristics in comparison to the baseline critic. Lower variance characteristics of the ASKL critic can be explained by observing that the empirical Rademacher complexity of ASKL critic's hypothesis space is bounded, theorem 3. Hence, generalization error is guaranteed to be upper bounded. Lower bias in estimates can be attributed to better control over bias-variance tradeoff.

Experimental results shown in Fig. 3, demonstrates the effect of change in batch size on the variance of ASKL and baseline critic estimates. It can be observed that with an increase in batch size the variance of both ASKL and baseline estimates decreases. This is due to the fact that the empirical Rademacher complexity is inversely proportional to the sample size (refer Appendix A for definition). Hence, an increase in batch size leads to a decrease in empirical Rademacher complexity and, corresponding decrease in variance of the MI estimates. Another key observation on the variance of MI estimates which holds for both critics is that with an increase in true MI the variance of the empirical estimates increases. This observations can be explained by noticing the effect of increase in the value of true MI on the log likelihood density ratio between the joint and product of marginal distributions, $log\left(d\mathbb{P}_{XY}/d\mathbb{P}_X \otimes \mathbb{P}_Y\right)$. The absolute value of the log density ratio evaluated at any given sample increases with increase in MI. The optimal critics for variational lower bound estimates of MI depend on the log density ratio. Hence, to match the increase in log density ratio the constant M which uniformly bounds the critic's hypothesis space also increases. As described in theorem 2, the generalization error bounds depend on both empirical Rademacher complexity and e^M, hence, an increase in M leads to an increase in variance of MI estimates.

Bias-variance tradeoff for different values of λ_1 and λ_2 in ASKL critic, figure 4. Figures 4(a) and 4(b) are the bias and variance plots, respectively. The left top most plots in both figures, 4(a) and 4(b) correspond to λ_1 and λ_2 set to 0, respectively. It can be seen in these plots that even without any explicit regularisation estimates using ASKL critic have lower bias and lower variance in comparison to the baseline critic. This verifies our claim that constraining the complexity of the hypothesis space leads to significant improvement in reliability of these estimates. It is evident from these plots that regularization weights are also effective in controlling the bias, as λ_1 and λ_2 increase the estimates get biased in negative direction. This demonstrates the efficacy of the proposed regularization

terms in inducing effective bias-variance tradeoff.

VI. Conclusion

In the proposed work, we successfully demonstrate the effect of controlling the complexity of critic's hypothesis space on the variance of sample based empirical estimates of mutual information. We negate the high variance characteristics of variational lower bound based estimates of MI by constructing the critic's hypothesis space in a Reproducing Kernel Hilbert Space, which corresponds to a critic learned using Automated Spectral Kernel Learning architecture. By analysing the generalisation bounds using Radmacher complexity of the constrained critic space, we demonstrate effective regularisation of bias-variance tradeoff on four different variational lower bounds of Mutual information. In larger scheme of Explainable-AI, this work theoretically motivates the implications of understanding the effect of regulating the complexity of deep neural network based critic hypothesis spaces on the bias-variance tradeoff of variational lower bound estimators of mutual information.

References

[1] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018, pp. 2610–2620.

[2] H. Kim and A. Mnih, "Disentangling by factorising," *arXiv preprint arXiv:1802.05983*, 2018.

[3] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo," *arXiv preprint arXiv:1711.00464*, 2017.

[4] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[6] Y. Li, "Which way are you going? imitative decision learning for path forecasting in dynamic scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[8] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

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

[10] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[11] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*, 2018, pp. 531–540.

[12] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On variational lower bounds of mutual information," in *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[13] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," *arXiv preprint arXiv:1910.06222*, 2019.

[14] S. Ghimire, P. K. Gyawali, and L. Wang, "Reliable estimation of kullback-leibler divergence by controlling discriminator complexity in the reproducing kernel hilbert space," *arXiv preprint arXiv:2002.11187*, 2020.

[15] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical review A*, vol. 33, no. 2, p. 1134, 1986.

[16] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, 1995.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[18] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[19] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[20] J. Li, Y. Liu, and W. Wang, "Automated spectral kernel learning," *arXiv preprint arXiv:1909.04894*, 2019.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[22] K. Ahuja, "Estimating kullback-leibler divergence using kernel machines," *arXiv preprint arXiv:1905.00586*, 2019.

[23] J. Shawe-Taylor, N. Cristianini *et al.*, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[24] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[25] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine learning research*, vol. 5, no. Jan, pp. 27–72, 2004.

[26] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," *arXiv preprint arXiv:1205.2653*, 2012.

[27] ——, "Learning non-linear combinations of kernels," in *Advances in neural information processing systems*, 2009, pp. 396–404.

[28] A. Wilson and R. Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *International conference on machine learning*, 2013, pp. 1067–1075.

[29] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum gaussian process regression," *The Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[30] W. Rudin, *Fourier analysis on groups*. Wiley Online Library, 1962, vol. 121967.

[31] A. M. Yaglom, "Correlation theory of stationary and related random functions." *Volume I: Basic Results.*, vol. 526, 1987.

[32] Y.-L. K. Samo and S. Roberts, "Generalized spectral kernels," *arXiv preprint arXiv:1506.02236*, 2015.

[33] J.-F. Ton, S. Flaxman, D. Sejdinovic, and S. Bhatt, "Spatial mapping with gaussian processes and nonstationary fourier features," *Spatial statistics*, vol. 28, pp. 59–78, 2018.

[34] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial Intelligence and Statistics*, 2016, pp. 370–378.

[35] A. Wilson and H. Nickisch, "Kernel interpolation for scalable structured gaussian processes (kiss-gp)," in *International Conference on Machine Learning*, 2015, pp. 1775–1784.

[36] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Stochastic variational deep kernel learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 2586–2594.

[37] C.-L. Li, W.-C. Chang, Y. Mroueh, Y. Yang, and B. Póczos, "Implicit kernel learning," *arXiv preprint arXiv:1902.10214*, 2019.

[38] H. Xue, Z.-F. Wu, and W.-X. Sun, "Deep spectral kernel learning," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 4019–4025.

[39] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[40] P. L. Bartlett, O. Bousquet, S. Mendelson *et al.*, "Local rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[41] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," *arXiv preprint arXiv:1811.04251*, 2018.

[42] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in neural information processing systems*, 2016, pp. 271–279.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Reducing the Variance of Variational Estimates of Mutual Information by Limiting the Critic's Hypothesis Space to RKHS Appendix

P Aditya Sreekar, Ujjwal Tiwari and Anoop Namboodri
Center for Visual Information Technology
International Institute of Information Technology, Hyderabad

A Rademacher Complexity

In problems pertinent to machine learning obtaining practical generalization error bound is crucial for proper model selection. Generalization error bounds are typically contained by a measure of the complexity of the learning model's hypothesis space, for example, the covering number of the hypothesis function space. The data-driven Rademacher's complexity used in this work is described as follows:

Let $(\mathcal{X}, \mathbb{P})$ be a probability space and $\mathcal{F}$ be the class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Consider $X_1, X_2, \ldots, X_n$ to be n i.i.d data samples from $\mathbb{P}$, with the corresponding empirical distribution denoted by $\mathbb{P}_n$. Now, let $\sigma_1, \sigma_2, \ldots, \sigma_n$ be n independent discrete random variables for which $Pr(\sigma = 1) = Pr(\sigma = -1) = \frac{1}{2}$ known as the Rademacher random variables. Then, for any $f \in \mathcal{F}$ we define

$$\begin{gathered} R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i),\ R_n\mathcal{F} = \sup_{f\in\mathcal{F}} R_n f \\ \hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma\left[R_n(\mathcal{F})\right],\ \mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[R_n(\mathcal{F})\right] \end{gathered} \tag{1}$$

Where, $\mathbb{E}_\sigma$ denotes expectation with respect to the Rademacher random variables, $\{\sigma_i\}_{i=1}^n$. And $\mathbb{E}$ is the expectation with respect to Rademacher random variables and data samples, $\{X_i\}_{i=1}^n$. $\mathcal{R}_n(\mathcal{F})$ and $\hat{\mathcal{R}}_n(\mathcal{F})$ are the Rademacher average and empirical (conditional) Rademacher average of $\mathcal{F}$, respectively. Intuitive reason for $\mathcal{R}_n(\mathcal{F})$ as a measure of complexity is that it quantifies the extent to which a function from the class $\mathcal{F}$ can correlate to random noise, a function belonging to a complex set can correlate to any random sequence. For a comprehensive overview of Rademacher averages and it's properties refer to [1–3]. Results from the aforementioned research work that have been used in the proofs related to our work are mentioned below.

The following is the concentration inequality that depicts the relation between Rademacher averages and empirical Rademacher averages. The deriva-

tion utilizes Talagrand's inequality, kindly refer to Lemma A.4 in [3] for full derivation.

Lemma A.1. *Let $\mathcal{F}$ be a class functions with range $[a, b]$. For fixed $\delta > 0$, with probability of atleast $1 - \delta$,*

$$\mathcal{R}_n(\mathcal{F}) \leq \inf_{\alpha \in (0,1)} \left(\frac{1}{1-\alpha} \hat{\mathcal{R}}_n(\mathcal{F}) + \frac{(b-a) log(\frac{1}{\delta})}{4n\alpha(1-\alpha)} \right)$$

The expected maximum deviation of empirical means from actual can be bounded by Rademacher averages as shown in the following bound. Check Lemman A.5 in [3] for derivation.

Lemma A.2. *For any class of function $\mathcal{F}$ we have,*

$$\max \left(\mathbb{E} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbb{P}} [f] - \mathbb{E}_{\mathbb{P}_n} [f] \right), \mathbb{E} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbb{P}_n} [f] - \mathbb{E}_{\mathbb{P}} [f] \right) \right) \leq 2\mathcal{R}_n(\mathcal{F})$$

Where, $\mathbb{E}_{\mathbb{P}_n} [f]$ is the empirical mean given n samples from $\mathbb{P}$ given by $\frac{1}{n} \sum_{i=1}^{n} f(X_i)$. Using Lemma A.1 and Lemma A.2, one can relate expected maximum deviation of empirical estimate from actual value to the empirical Rademacher averages.

We would like to point a minor error in the derivation of the generalization error bound in *et al.* [4] where Lemma A.2 has been used. In their work left hand side of the bound has been misinterpreted as maximum deviation instead of expected maximum deviation. To relate maximum deviation to Rademacher average we need another bound before Lemma A.2 which relates maximum deviation to expected maximum deviation. We will look at this corrected approach in the next section where we derive the generalization results for our work.

The following simple structural result can be used to express Rademacher averages for a complex class of functions in terms of Rademacher averages of simple class of functions.

Lemma A.3. *If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L_ϕ and satisfies $\phi(0) = 0$, then $\mathcal{R}_n(\phi \circ \mathcal{F}) \leq 2L_\phi \mathcal{R}_n(\mathcal{F})$*

Next, we look at the empirical Rademacher average for the class of functions represented by our ASKL critic.

Theorem A.4. *The empirical Rademacher average of the RKHS $\mathcal{F}$ learned by the ASKL critic can be bounded and is described as follows,*

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{B}{n} \sqrt{\sum_{i=1}^{n} \|\phi(x_i)\|_2^2} \leq \frac{B}{\sqrt{n}}$$

Where, $B = \sup_{f \in \mathcal{F}} \|w\|_2$.

Proof.

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right)\right] \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(w^\intercal\Phi_\sigma\right)\right]\end{aligned}$$

Here $\Phi_\sigma = \sum_{i=1}^{n}\sigma_i\phi(x_i)$ is a D dimensional vector

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(w^\intercal\Phi_\sigma\right)\right] \\ &\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\|w\|_2\|\Phi_\sigma\|_2\right)\right] \qquad (2)\\ &\leq \frac{B}{n}\mathbb{E}_\sigma\left[\|\Phi_\sigma\|_2\right] \\ &\leq \frac{B}{n}\sqrt{\mathbb{E}_\sigma\left[\|\Phi_\sigma\|_2^2\right]} \qquad (3)\end{aligned}$$

Where step 2 is a direct implication of the Cauchy-Schwarz inequality.

$$\begin{aligned}\mathbb{E}_\sigma\left[\|\Phi_\sigma\|_2^2\right] &= \mathbb{E}_\sigma\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j\phi\left(x_i\right)^\intercal\phi\left(x_j\right)\right] = \sum_{i=1}^{n}\|\phi\left(x_i\right)\|_2^2 \\ &= \frac{1}{2D}\sum_{i=1}^{n}\sum_{j=1}^{D}cos\left(\left(\omega_j-\omega_j'\right)^\intercal x_i\right)+1 \leq n\end{aligned} \qquad (4)$$

From 3 and 4 we have the final result. □

B Generalization Error Bounds

In this section we derive the generalization error bounds contributed in the scope of paper. We represent joint distribution, $\mathbb{P}_{XY}$ as $\mathbb{P}$ and the product of marginal distributions, $\mathbb{P}_X \otimes \mathbb{P}_Y$, as $\mathbb{Q}$. Both distribution are define on measurable space $(\mathcal{X}\times\mathcal{Y}, \Sigma_{XY})$. $\mathbb{P}_n$ and $\mathbb{Q}_m$ represents the corresponding empirical distributions and the pair (x, y) is referred as z. The proofs use McDiarmid's inequality which is described as follows:

Lemma B.1 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathcal{X}$, and assume that $\phi : \mathcal{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i'\in\mathcal{X}}|\phi(x_1,\ldots,x_n)-\phi(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i$$

for every $1 \leq i \leq n$.

Then, for every $t > 0$*,*

$$Pr\left\{\phi\left(X_1,\ldots,X_n\right) - \mathbb{E}\left[\phi\left(X_1,\ldots,X_n\right)\right] \geq t\right\} \leq e^{-2t^2/\sum_{i=1}^{n} c_i^2}$$

Stated in another way, for some fixed $\delta > 0$ *and with probability of at least* $1-\delta$*:*

$$\phi\left(X_1,\ldots,X_n\right) \leq \mathbb{E}\left[\phi\left(X_1,\ldots,X_n\right)\right] + \sqrt{\frac{\sum_{i=1}^{n} c_i^2\ log\left(\frac{1}{\delta}\right)}{2}}$$

In this section generalization error bounds for two lower bounds of mutual information I_{TUBA} and I_{DV} have been derived.

$$I_{TUBA}\left(f\right) = \mathbb{E}_{\mathbb{P}}\left[f(z)\right] - \frac{\mathbb{E}_{\mathbb{Q}}\left[e^{f(z)}\right]}{a} - log\left(a\right) + 1 \tag{5}$$

$$I_{DV}\left(f\right) = \mathbb{E}_{\mathbb{P}}\left[f(z)\right] - log\left(\mathbb{E}_{\mathbb{Q}}\left[e^{f(z)}\right]\right) \tag{6}$$

Where in Eq.5 the baseline $a(y)$ is restricted to a constant a, this is because both I_{MINE} and I_{NWJ} lower bounds considered in this work correspond to constant baseline case. As the true distributions $\mathbb{P}$ and $\mathbb{Q}$ are unknown, we approximate the true expectation, with expectation with respect to empirical distributions $\mathbb{P}_n$ and $\mathbb{Q}_m$ corresponding to n independent samples, $\{z_i\}_{i=1}^{n}$, from $\mathbb{P}$ and m independent samples, $\{z_i'\}_{i=1}^{m}$, from $\mathbb{Q}$ respectively.

$$\hat{I}_{TUBA}^{n,m}\left(f,S\right) = \mathbb{E}_{\mathbb{P}_n}\left[f(z)\right] - \frac{\mathbb{E}_{\mathbb{Q}_m}\left[e^{f(z)}\right]}{a} - log\left(a\right) + 1 \tag{7}$$

$$\hat{I}_{DV}^{n,m}\left(f,S\right) = \mathbb{E}_{\mathbb{P}_n}\left[f(z)\right] - log\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^{f(z)}\right]\right) \tag{8}$$

Where S is the set of samples $\{z_i\}_{i=1}^{n}$ and $\{z_i'\}_{i=1}^{m}$. The goal of generalization error bound is to bound the maximum deviation between I_{TUBA} and $\hat{I}_{TUBA}$ or between I_{DV} and $\hat{I}_{DV}$.

Theorem B.2 (Generalization bound for I_{TUBA})**.** *Assume that the function space* $\mathcal{F}$ *learnt by the critic is uniformly bounded by* M*, that is* $|f(z)| \leq M\ \forall f \in \mathcal{F}\ \&\ \forall z \in \mathcal{X}\times\mathcal{Y}$*, where* $M < \infty$*. For any fixed* $\delta > 0$ *generalization error of TUBA estimate can be bounded with probability of at least* $1-\delta$

$$\sup_{f\in\mathcal{F}}\left(I_{TUBA}(f) - \hat{I}_{TUBA}^{n,m}(f)\right) \leq 4\hat{\mathcal{R}}_n\left(\mathcal{F}\right) + \frac{8}{a}e^{M}\hat{\mathcal{R}}_m\left(\mathcal{F}\right) + \frac{4M}{n}log\left(\frac{4}{\delta}\right) + \frac{8Me^{M}}{am}log\left(\frac{4}{\delta}\right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^{M}-e^{-M})^2}{a^2 m}\right)log\left(\frac{2}{\delta}\right)}{2}}$$

Where, sample set S *for* $\hat{I}_{TUBA}^{n,m}$ *has been implicitly assumed to be given.*

Proof. Let,

$$\phi(S) = \sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f, S) \right)$$

Letting $\tilde{S}_i$ represent another set of samples which differ from S at only one sample z_i when $i \in [1, n]$ or at sample z'_i when $i \in [n+1, n+m]$, where the first case is when differing sample is sampled from $\mathbb{P}$ and the second case is when the differing sample is sampled from $\mathbb{Q}$. Now, when $i \in [1, n]$

$$\begin{aligned}
\left| \phi(S) - \phi(\tilde{S}_i) \right| &= | \sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f, S) \right) \\
&\quad - \sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f, \tilde{S}_i) \right) | \\
&\leq \sup \left| \hat{I}^{n,m}_{TUBA}(f, \tilde{S}_i) - \hat{I}^{n,m}_{TUBA}(f, S) \right| \\
&= \frac{1}{n} sup \left| f(\tilde{z}_i) - f(z_i) \right| \\
\left| \phi(S) - \phi(\tilde{S}_i) \right| &\leq \frac{2M}{n}
\end{aligned} \tag{9}$$

Where, step 9 is because the maximum difference between values of a function bounded between $[-M, M]$ is $2M$, when $i \in [n+1, n+m]$.

$$\begin{aligned}
\left| \phi(S) - \phi(\tilde{S}_i) \right| &\leq \sup_{f \in \mathcal{F}} \left| \hat{I}^{n,m}_{TUBA}(f, \tilde{S}_i) - \hat{I}^{n,m}_{TUBA}(f, S) \right| \\
&= \frac{1}{am} \sup \left| e^{f\left(\tilde{z}'_i\right)} - e^{f\left(z'_i\right)} \right| \\
\left| \phi(S) - \phi(\tilde{S}_i) \right| &\leq \frac{e^M - e^{-M}}{am}
\end{aligned} \tag{10}$$

and, step 10 is due to the fact the maximum difference between values of exponential of a function bounded in $[-M, M]$ is $e^M - e^{-M}$. Because, there exists a c_i such that $\left| \phi(S) - \phi\left(\tilde{S}_i\right) \right| < c_i \ \forall \ i \in [1, m+n]$ we can apply McDiarmid's inequality, lemma B.1, and write for a fixed $\delta > 0$ with probability of at least $1 - \delta/2$,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f) \right) \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f) \right) \right] + \\
\sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^M - e^{-M})^2}{a^2 m} \right) log\left(\frac{2}{\delta}\right)}{2}}
\end{aligned} \tag{11}$$

By using lemma A.2 we get,

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{P}_n}[f] \right) \right] \leq 2\mathcal{R}_n(\mathcal{F}) \tag{12}$$

Similarly, if we consider a family of functions $\psi \circ \mathcal{F} = \{\psi(f(z)) : \forall f \in \mathcal{F}\}$ where $\psi(x) = e^x - 1$.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right] - \mathbb{E}_{\mathbb{Q}}\left[e^f\right]\right)\right] = \mathbb{E}\left[\sup_{g\in\psi\circ\mathcal{F}}\left(\mathbb{E}_{\mathbb{Q}_m}\left[g\right] - \mathbb{E}_{\mathbb{Q}}\left[g\right]\right)\right]$$
$$\leq 2\mathcal{R}_m\left(\psi \circ \mathcal{F}\right) \tag{13}$$
$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right] - \mathbb{E}_{\mathbb{Q}}\left[e^f\right]\right)\right] \leq 4e^M\mathcal{R}_m\left(\mathcal{F}\right) \tag{14}$$

Step 13 is from lemma A.2 and step 14 is in implication of lemma A.3 and the fact that $\phi(x) = e^x - 1$ is Lipschitz with constant e^M when $x \in [-M, M]$. Now, we are in a position to relate expectation of maximum deviation to Rademacher average.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f)\right)\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{P}}\left[f\right] - \mathbb{E}_{\mathbb{P}_n}\left[f\right]\right)\right] +$$
$$\frac{1}{a}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right] - \mathbb{E}_{\mathbb{Q}}\left[e^f\right]\right)\right]$$
$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(I_{TUBA}(f) - \hat{I}^{n,m}_{TUBA}(f)\right)\right] \leq 2\mathcal{R}_n\left(\mathcal{F}\right) + \frac{4e^M}{a}\mathcal{R}_m\left(\mathcal{F}\right) \tag{15}$$

The last step 15 is in tandem with steps 14 and 12. Using lemma A.1 and setting $\alpha = 1/2$ we get with probability of at least $1 - \delta/4$:

$$\mathcal{R}_n\left(\mathcal{F}\right) \leq 2\hat{\mathcal{R}}_n\left(\mathcal{F}\right) + \frac{2Mlog\left(\frac{4}{\delta}\right)}{n} \tag{16}$$

Similar, relationship can also be stated between $\mathcal{R}_m\left(\mathcal{F}\right)$ and $\hat{\mathcal{R}}_m\left(\mathcal{F}\right)$
Combining 11, 15 and 16 we get with probability of at least $1 - \delta$

$$\sup_{f\in\mathcal{F}}\left(I_{TUBA}\left(f\right) - \hat{I}^{n,m}_{TUBA}\right) \leq 4\hat{\mathcal{R}}_n\left(\mathcal{F}\right) + \frac{8e^M}{a}\hat{\mathcal{R}}_m\left(\mathcal{F}\right) + \frac{4Mlog\left(\frac{4}{\delta}\right)}{n} +$$
$$\frac{8Me^Mlog\left(\frac{4}{\delta}\right)}{am} + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^M - e^{-M})^2}{a^2m}\right)log\left(\frac{2}{\delta}\right)}{2}} \tag{17}$$

□

Li *et al.* [4] in derivation of generalization error bounds incorrectly replaced expectation of maximum deviation with maximum deviation in lemma A.2. To

rectify that error, we used McDiarmid's Inequality to bound maximum deviation with expected maximum deviation, this adds and additional term inside square root in the bound in theorem B.2.

Next, we are going to look at the generalization error bounds of Donsker Varadhan estimates, it is used to estimate mutual information in I_{SMILE} estimate. We follow the same procedure used for deriving generalization error bounds of I_{TUBA} to keep the proof brief.

Theorem B.3 (Generalization bound for I_{DV})**.** *Assume that the function space $\mathcal{F}$ learnt by the critic is uniformly bounded by M, that is $|f(z)| \leq M \ \forall f \in \mathcal{F}$ & $\forall z \in \mathcal{X} \times \mathcal{Y}$, where $M < \infty$. For a fixed $\delta > 0$ generalization error of Donsker Varadhan estimate can be bounded with probability of at least $1 - \delta$*

$$\sup_{f \in \mathcal{F}} \left(I_{DV}(f) - \hat{I}^{n,m}_{DV}(f) \right) \leq 4\hat{\mathcal{R}}_n(\mathcal{F}) + 8e^{2M}\hat{\mathcal{R}}_m(\mathcal{F}) + \frac{4M}{n} log\left(\frac{4}{\delta}\right) \frac{8Me^{2M}}{m} log\left(\frac{4}{\delta}\right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^{2M}-1)^2}{m}\right) log\left(\frac{2}{\delta}\right)}{2}}$$

Where, sample set S for $\hat{I}^{n,m}_{DV}$ is implicitly assumed.

Proof. Let

$$\phi(S) = \sup_{f \in \mathcal{F}} \left(I_{DV}(f) - \hat{I}^{n,m}_{DV}(f, S) \right)$$

Letting $\tilde{S}_i$ represent another set of samples which differ from S at only one sample. When $i \in [1, n]$ $\left|\phi(S) - \phi(\tilde{S}_i)\right| \leq \frac{2M}{n}$ from equation 9, when $i \in [n+1, n+m]$

$$\begin{aligned}
\left|\phi(S) - \phi(\tilde{S}_i)\right| &\leq \sup_{f \in \mathcal{F}} \left| \hat{I}_{DV}\left(\tilde{S}_i\right) - \hat{I}_{DV}\left(S_i\right) \right| \\
&= \sup \left| log\left(\mathbb{E}_{\mathbb{Q}_{m,i}}\left[e^{f(z)}\right] \right) - log\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^{f(z)}\right] \right) \right| \qquad (18) \\
&\leq e^M \sup \left| \mathbb{E}_{\mathbb{Q}_{m,i}}\left[e^{f(z)}\right] - \mathbb{E}_{\mathbb{Q}_m}\left[e^{f(z)}\right] \right| \qquad (19) \\
&= \frac{e^M}{m} \sup \left| e^{f\left(\tilde{z}'_i\right)} - e^{f\left(z'_i\right)} \right| \\
&\leq \frac{e^{2M} - 1}{m} \qquad (20)
\end{aligned}$$

In step 18 $\mathbb{Q}_{m,i}$ refers to the empirical distribution corresponding to the sample set $\tilde{S}'_i$. Where, inequality 19 is due to the fact that $log(x)$ is Lipschitz with constant e^M when $x \in \left[e^{-M}, e^M\right]$. Inequality 20 is due to the fact that the maximum difference between values of exponential of a function bounded in $[-M, M]$ is $e^M - e^{-M}$. Now, apply McDairmid's inequality to $\phi(S)$, for a fixed $\delta > 0$ with probability of at least $1 - \delta/2$:

$$\sup_{f\in\mathcal{F}}\left(I_{DV}(f)-\hat{I}^{n,m}_{DV}(f)\right)\leq\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(I_{DV}(f)-\hat{I}^{n,m}_{DV}(f)\right)\right]+\sqrt{\frac{\left(\frac{4M^2}{n}+\frac{(e^{2M}-1)^2}{m}\right)log\left(\frac{2}{\delta}\right)}{2}} \tag{21}$$

We use 12 and 14 to bound expected maximum deviation with Rademacher averages

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(I_{DV}(f)-\hat{I}^{n,m}_{DV}(f)\right)\right]&\leq\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f]\right)\right]+\\&\quad\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(log\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right]\right)-log\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right]\right)\right)\right]\\&\leq\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f]\right)\right]+\\&\quad e^M\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right]-\mathbb{E}_{\mathbb{Q}_m}\left[e^f\right]\right)\right]\\\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(I_{DV}(f)-\hat{I}^{n,m}_{DV}(f)\right)\right]&\leq 2\mathcal{R}_n(\mathcal{F})+4e^{2M}\mathcal{R}_m(\mathcal{F})\end{aligned} \tag{22}$$

Where last step is in tandem with 12 and 14. Combining 21, 22 with 16 we get the final result with probability of at least $1-\delta$ described below as,

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left(I_{DV}(f)-\hat{I}^{n,m}_{DV}(f)\right)&\leq 4\hat{\mathcal{R}}_n(\mathcal{F})+8e^{2M}\hat{\mathcal{R}}_m(\mathcal{F})+\frac{4M}{n}log\left(\frac{4}{\delta}\right)\\&\quad\frac{8Me^{2M}}{m}log\left(\frac{4}{\delta}\right)+\sqrt{\frac{\left(\frac{4M^2}{n}+\frac{(e^{2M}-1)^2}{m}\right)log\left(\frac{2}{\delta}\right)}{2}}\end{aligned} \tag{23}$$

□

References

[1] S. Mendelson, "A few notes on statistical learning theory," in *Advanced lectures on machine learning*. Springer, 2003, pp. 1–40.

[2] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[3] P. L. Bartlett, O. Bousquet, S. Mendelson *et al.*, "Local rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[4] J. Li, Y. Liu, and W. Wang, "Automated spectral kernel learning." in *AAAI*, 2020, pp. 4618–4625.