Regularized Training of Convolutional Autoencoders using the Rényi-Stratonovich Value of Information

Isaac J. Sledge,[1,2] *Member, IEEE* José C. Príncipe,[2,3] *Life Fellow, IEEE*

[1] Advanced Signal Processing and Automated Target Recognition Branch, Naval Surface Warfare Center
[2] Department of Electrical and Computer Engineering, University of Florida
[3] Department of Biomedical Engineering, University of Florida

***Abstract*—We propose an information-theoretic cost function for the regularized training of convolutional autoencoders that imposes an organization on the bottleneck-layer-projected samples so as to facilitate discrimination. This function is based on a continuous-space, Rényi-mutual-information version of Stratonovich's value of information. It quantifies the maximum benefit that can be obtained for a given bottleneck-layer representation compression amount. The compression amount is controlled by a single hyperparameter that trades off between the autoencoder reconstruction quality and the hidden-layer representation uncertainty.**

***Index Terms*—Value of information, deep neural networks, deep learning, information-theoretic learning, information theory**

I. Introduction

Unsupervised and semi-supervised deep learning is a widely popular approach for uncovering features from data. It does, however, have issues that can inhibit performance. Deep autoencoder networks may not uncover parsimonious, intermediate feature representations that lend themselves well to classification. The bottleneck-layer-projected samples may hence be organized arbitrarily. As well, the autoencoder representations are usually not minimally sufficient statistics in an information-theoretic sense. They might not be maximally compressive intermediate mappings that preserve the most information about the desired response and hence have the chance of being overfit.

For unsupervised and semi-supervised deep learning to be effective, the aforementioned learning concerns should be addressed. We believe that this can be partly resolved by utilizing details about the training samples to inform parameter selection. This objective, which may be viewed as a type of lossy source compression, can be handled by applying information-theoretic concepts that trade off between the network prediction quality and the representation uncertainty.

Here, we consider an information-theoretic [1] cost function for the regularized training of semi-supervised, convolutional autoencoder networks for object recognition. This cost is based on a novel Rényi-information version of Stratonovich's value of information [2, 3]. The value of information quantifies the maximum benefit that can be obtained from a given quantity of information to improve the average penalties. This has the effect of guiding the choice of network parameters so that, in the bottleneck-layer of an autoencoder network, samples from the same class are typically grouped while samples from different classes are well separated for appropriate choices of penalty functions.

More specifically, the value of information is a two-term optimization problem that significantly pre-dates and generalizes the information-bottleneck method [4, 5] to arbitrary cost functions. When applied to training autoencoder networks, we show that the value of information facilitates an optimal trade-off between the conciseness of the bottleneck-layer representation and that network's expected predictive capability through the chosen network parameters. This compromise is dictated by a Rényi-mutual-information bound that specifies the mutual dependence between the intermediate representation and the model response. The higher the bound, the greater the representation uncertainty. This leads to a bottleneck-layer representation organization for each object class that can be increasingly arbitrary but facilitates high model accuracy. That is, the autoencoder network learns to mimic the training samples well for appropriate penalty functions. The intermediate representation may have highly overlapping class regions, though, which complicates learning a good classifier from the bottleneck features. Lowering the bound imposes constraints on the representation organization, which often speeds up training due to an aggregation of Markov chains [6] underlying the network layers. The representation for each object class tends to become increasingly compact and well separated, despite not using class labels during training. This sample-separation property aids in subsequent discrimination between the different classes; labels can be utilized to often enhance discrimination.

For the value of information to be effective for network training, an efficient optimization strategy is required. We have already begun investigating this topic in the context of reinforcement learning [7–10]. Any of these schemes could be directly applied to train deep architectures. They would be computationally prohibitive, though, given the great number of parameters that need to be tuned, let alone the continuous nature of the network random variables. Here, we estimate Rényi mutual information using positive-definite kernels with specific properties. This allows us to obtain information-like quantities without assuming that the underlying conditional probabilities associated with the network layers and marginal probabilities related to the data are either known or estimated. We then show that this kernel-based representation admits gradients that can be used to efficiently train the regularized networks via back-propagation gradient descent with mini-batches.

II. Methodology

In what follows, we build up to and describe the notion of the Rényi-Stratonovich value of information. We then show how to optimize a principled approximation of the criterion.

A. Criterion Definition

Consider a system defined by an input, intermediate encoding, and output space, all of which are measurable; this will be an abstraction of an autoencoder network. We let x_t and y_t be continuous random variables for the input and output, at iteration t; we use the same variables to denote the random-variable values, where appropriate. We assume that these random variables are distributed according to some joint probability distribution $p_{x_t,y_t;\theta}$ with marginal probabilities indicated by $p_{x_t} = \int_{\mathcal{Y}} p_{x_t,dy_t;\theta}$ and $p_{y_t} = \int_{\mathcal{X}} p_{dx_t,y_t;\theta}$. We also assume that the internal representation z_t of a hidden layer is a stochastic encoding of the input, which is specified by the conditional probability $p_{z_t|x_t;\theta} = p_{x_t,z_t;\theta} / \int_{\mathcal{Y}} p_{x_t,dy_t;\theta}$, with parameters $\theta \in \mathbb{R}^p$.

After observing the input x_t, output y_t, and encoding z_t random variables, an estimate of the network parameters could be obtained by minimizing the conditional expected penalty,

$$\inf_\theta \mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t}] = \inf_\theta \int_{\mathcal{X}} g_{x_t,y_t,z_t;\theta} p_{dx_t},$$

where $g_{x_t,y_t,z_t;\theta}$ is a real-valued penalty function. The obtained estimate would, however, be independent of the stochastic encoding $p_{z_t|x_t;\theta}$. A

The work of the first and second authors was funded by grants N00014-15-1-2013 (Jason Stack), N00014-14-1-0542 (Marc Steinberg), and N00014-19-WX-00636 (Marc Steinberg) from the US Office of Naval Research. The first author was additionally supported by in-house laboratory independent research (ILIR) grant N00014-19-WX-00687 (Frank Crosby) from the US Office of Naval Research and a Naval Innovation in Science and Engineering (NISE) grant from NAVSEA.

more appropriate expression to minimize is $\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t,z_t;\theta}] = \mathbb{E}[\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{z_t|x_t;\theta}]; p_{x_t}]$, the total expected penalty,

$$\inf_\theta \mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t,z_t;\theta}] = \inf_\theta \int_{\mathcal{X}}\int_{\mathcal{Z}} g_{x_t,y_t,z_t;\theta} p_{dz_t|x_t;\theta} p_{dx_t},$$

as it accounts for the influence of the stochastic encoding. For it to prove useful for regularized training, though, additional constraints are required. If they are not imposed, then there is the chance, for arbitrary penalty functions, that the network will simply implement an identity mapping without regards to hidden-layer organization.

We adopt the view that information-theoretic constraints can regularize network training. One way of doing this is to ensure that any mappings which minimize the total expected penalty also obey a mutual information bound between the intermediate random variable and either the input or output random variables. Such a bound limits the amount of information that is propagated through the network and hence helps choose parameter values that cause the network to remove superfluous details.

When using a mutual information constraint, there are two extreme cases to consider. The first is when the intermediate layer random variables carry no information about either the autoencoder input or output random variables. In this case, there is only one way to choose parameters, which is by using the marginal. The total expected penalty hence becomes equivalent to the conditional expected penalty, $\inf_\theta \mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t,z_t;\theta}] = \inf_\theta \mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t}]$. There is hence complete uncertainty, and the network will not optimize the stochastic encoding against the penalty function. If the intermediate layer possess complete information about either the input or the output random variables, then the total expected penalty becomes $\inf_\theta \mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t,z_t;\theta}] = \mathbb{E}[\inf_\theta g_{x_t,y_t,z_t;\theta}; p_{x_t}]$, where

$$\mathbb{E}[\inf_\theta g_{x_t,y_t,z_t;\theta}; p_{x_t}] = \int_{\mathcal{X}} \inf_\theta g_{x_t,y_t,z_t;\theta}.$$

The network will optimize the penalty function and the best performance will be achieved, as there is no uncertainty. This does not imply that the network will maximize class discrimination, though; in the context of autoencoder networks, it will merely imply that the input is reconstructed well at the output.

The transition between no information to complete information, and hence a reduction of penalties, is not immediate. There is a smooth, non-linear transition [11] between these two extremes for varying levels of information. Stratonovich [2, 3] proposed an expression for these intermediate cases: the value of information. The value of information is independent of the chosen information measure; here we consider a Rényi mutual information term to ease optimizing the criterion in the continuous random variable case.

For autoencoder networks, the value of information quantifies the expected decrease in penalties from the baseline case, as quantified by the conditional expected penalty, where the mutual dependence of the random variables for the stochastic encoding is constrained.

Definition 2.1. Let x_t, y_t, z_t be random variables with marginals $p_{x_t}, p_{y_t}, p_{z_t}$ and conditional distributions $p_{z_t|x_t;\theta}, p_{z_t|y_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. Let $g_{x_t,y_t,z_t;\theta}$ be a real-valued penalty function. The Rényi-Stratonovich value of information is the difference between the conditional expected, $\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t}]$, and total expected, $\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t,z_t;\theta}]$, penalties,

$$\inf_\theta \int_{\mathcal{X}} g_{x_t,y_t,z_t;\theta} p_{dx_t} - \inf_\theta \int_{\mathcal{X}}\int_{\mathcal{Z}} g_{x_t,y_t,z_t;\theta} p_{dz_t|x_t;\theta} p_{dx_t},$$

where the minimization is subject to a Rényi mutual-information bound $I_\alpha(y_t; z_t) = \beta$, $\beta \geq 0$; this bound is referred to as the β-information amount.

Definition 2.2. Let y_t, z_t be random variables with a marginal and parameterized conditional distribution. For $\alpha \in (0,1) \cup (1,\infty)$, a version of Rényi's mutual information is given by the difference between the Rényi entropy, $\log(\mathbb{E}[p_{z_t}^{\alpha-1}; p_{z_t}])/(1-\alpha)$, and the conditional Rényi entropy, $\log(\mathbb{E}[\mathbb{E}[p_{z_t|y_t;\theta}^{\alpha}/p_{z_t}; p_{z_t}]^{1/\alpha}; p_{y_t}])^{\alpha/(\alpha-1)}$,

$$I_\alpha(y_t; z_t) = \log(\textstyle\int_{\mathcal{Z}} p_{dz_t}^{\alpha})/(1-\alpha) + \log(\textstyle\int_{\mathcal{Y}}(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha})^{1/\alpha} p_{dy_t})^{\alpha/(\alpha-1)}.$$

Here, we have considered a version of Rényi's mutual information defined by a novel conditional Rényi's entropy. This version has several advantageous properties over existing definitions, and we theoretically motivate it in the appendix.

The value of information pre-dates and generalizes several regularizations schemes. One is the information-bottleneck method [4, 5]. The latter uses on a non-fixed penalty function that relies on a so-called optimal mapping $g_{x_t,y_t,z_t;\theta} = D_{\mathrm{KL}}[p_{y_t|x_t;\theta} \| p_{y_t|z_t;\theta}]$; it measures the dissimilarity between the conditionals $p_{y_t|x_t;\theta}$ and $p_{y_t|z_t;\theta} = \int_{\mathcal{X}} p_{y_t|x_t;\theta} p_{z_t|x_t;\theta} p_{dx_t}/p_{z_t}$ so that maximally relevant information about the output random variables can be retained. Such a penalty may not be appropriate in all situations, though, as we show in our simulations. The value of information, in contrast, permits the use of arbitrary penalties that can be tailored to specific applications, such as semi-supervised learning.

Criterion Interpretation. The value of information specifies a constrained difference in penalties. The first term, $\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{x_t}]$, quantifies the expected penalty when the divergence between the intermediate-layer and input random variables are zero. It establishes the worst autoencoder performance. This is because the network parameters are only be guided by the marginal probability, which is not a sufficient regularizer. This second term, $\mathbb{E}[\mathbb{E}[g_{x_t,y_t,z_t;\theta}; p_{z_t|x_t;\theta}]; p_{x_t}]$, specifies the total expected penalty for case where the random variables have partial to no uncertainty. Both terms highlight the possible improvement in the autoencoder's ability to mimic the identity mapping.

The constraint term, an expected Rényi's α-information, stipulates the amount of informational overlap between conditional stochastic encoding $p_{z_t|y_t;\theta}$ and the marginal probability p_{z_t}. This term is bounded by the β-information amount, which dictates the amount that the marginal probability p_{z_t} can change to become a conditional probability $p_{z_t|y_t;\theta}$. If it is zero, then there is no overlap between the two random variables and the transformation costs are infinite, so $p_{z_t|y_t;\theta}$ does not change from p_{z_t}. This corresponds to the no-information case, which implies that the penalty function is not minimized due to the lack of information flowing through the network. The hidden-layer features tend to be distributed in a compact, well-separated manner, though, when this occurs. If the β-information is larger than the output random-variable entropy, $-\mathbb{E}[\log(p_{y_t}); p_{y_t}]$, then the highest amount of overlap is achieved and the transformation costs are ignored. The marginal probability p_{z_t} is therefore free to change to a stochastic encoding $p_{z_t|y_t;\theta}$ that leads to the greatest reduction of penalties. This yields an autoencoder that mimics well the identity mapping for the chosen penalty function. The hidden-layer features often belong to overlapping distributions, though, due to the lack of random-variable quantization.

Taken together, the value of information highlights the decrease in the penalty function for a given compression of the hidden-layer random variables. The use of a joint-random-variable uncertainty measure facilitates quantization, as described in [12, 13].

B. Criterion Optimization

Our previous approaches to optimizing the value of information have been through expectation-maximization-like updates where the mutual-information term does not need to be explicitly estimated [7–10]. While these updates are suitable for discrete spaces, they are intractable for continuous ones. Another issue is that they require knowledge of the underlying probability densities; these can be difficult, let alone computationally expensive, to estimate.

Here, we sidestep these issues by approximating the Rényi mutual-information constraint with a matrix-based version in a reproducing kernel Hilbert space. As a consequence of this approximation, we can derive efficient, mini-batch-based gradient updates.

Matrix-Based Rényi Information. We begin by establishing the notion of a matrix-based Rényi entropy for positive-definite kernels; we refer readers to [14] for more details.

Definition 2.3. Let $\kappa \succeq 0$, where $\kappa \in \mathbb{R}^{n\times n}$ is a Gram matrix computed from mini-batch of n samples. The parameterized matrix functional for Rényi's entropy is given by $s_\alpha(\kappa) = \log(\mathrm{tr}(\kappa^\alpha))/(1-\alpha)$, where $\alpha \in (0,1) \cup (1,\infty)$.

We can view this version of entropy as measuring the lack of statistical regularities in a transformed version of a mini-batch, as represented by

the Gram matrix. It hence quantifies uncertainty.

This notion can be extended to the joint and conditional cases; here, we consider the case of two random variables, but the definitions can be posed for any finite number of them [15].

Definition 2.4. Let $\kappa_k \succeq 0$, $k = 1, 2$, where $\kappa_k \in \mathbb{R}^{n\times n}$. The matrix functional for the joint Rényi entropy, for $\alpha \in (0, 1) \cup (1, \infty)$, is

$$s_\alpha(\kappa_1, \kappa_2) = s_\alpha(\kappa_1 \circ \kappa_2 / \text{tr}(\kappa_1 \circ \kappa_2)).$$

The matrix functional for the conditional Rényi entropy is

$$s_\alpha(\kappa_1|\kappa_2) = s_\alpha(\kappa_1 \circ \kappa_2 / \text{tr}(\kappa_1 \circ \kappa_2)) - s_\alpha(\kappa_1).$$

An advantage of this formulation is that the random-variable probabilities do not need to be known, let alone estimated estimated. This property is immensely useful when dealing with massive datasets, let alone large networks with many parameters.

The following proposition establishes that these matrix-derived quantities behave similarly to entropy.

Proposition 2.1. Let $\kappa_k \succeq 0$, $[\kappa_k]_{i.j} \geq 0$, $\forall i, j$ and $k = 1, 2$, where $\text{tr}(\kappa_k) = 1$ and $[\kappa_k]_{i,i} = 1/n$, $\forall i$. We have the following generalizations for:

(i) Monotonicity: $s_\alpha(\kappa_1 \circ \kappa_2 / \text{tr}(\kappa_1 \circ \kappa_2)) \geq s_\alpha(\kappa_2)$

(ii) Chain rule: $s_\alpha(\kappa_1 \circ \kappa_2 / \text{tr}(\kappa_1 \circ \kappa_2)) \leq s_\alpha(\kappa_1) + s_\alpha(\kappa_2)$.

Due to the generalization of the chain rule, we have a partial guarantee of non-negativeness for the following Rényi mutual-information approximation.

Definition 2.5. Let $\kappa_k \succeq 0$, $[\kappa_k]_{i.j} \geq 0$, $\forall i, j$, where $\text{tr}(\kappa_k) = 1$ and $[\kappa_k]_{i,i} = 1/n$, $\forall i$ and $k = 1, 2$. The matrix-based Rényi mutual-information approximation is $s_\alpha(\kappa_1) + s_\alpha(\kappa_2) - s_\alpha(\kappa_1, \kappa_2)$.

To guarantee non-negativity, we assume that the kernels specifying the Gram matrices are infinity divisible. This constraint can be satisfied by choosing Gaussian kernel functions, for instance.

Proposition 2.2. Let $\kappa_k \succeq 0$, $[\kappa_k]_{i.j} \geq 0$, $\forall i, j$ and $k = 1, 2$, where $\text{tr}(\kappa_k) = 1$ and $[\kappa_k]_{i,i} = 1/n$, $\forall i$. Let $\kappa_k^{\circ r}$ denote the entry-wise rth power: $[\kappa_k^{\circ r}]_{i,j} = ([\kappa_k]_{i,j})^r$. If $\kappa_k^{\circ r} \succeq 0$, for $\forall r \geq 0$, then:

(i) Non-negativity: $s_\alpha(\kappa_1) + s_\alpha(\kappa_2) - s_\alpha(\kappa_1, \kappa_2) \geq 0$

(ii) Monotonicity: $s_\alpha(\kappa_1) \geq s_\alpha(\kappa_1) + s_\alpha(\kappa_2) - s_\alpha(\kappa_1, \kappa_2)$.

This formulation of Rényi's mutual information is well defined since the set of positive semi-definite matrices is closed under the Hadamard product. It also stems from the well-definedness of $s_\alpha(\kappa_k^{\circ r} / \text{tr}(\kappa_k^{\circ r}))$ due to the use of infinitely divisible kernels.

Value-of-Information Gradients. We now re-state the value of information in a way that facilitates deriving meaningful gradient-based updates for parameterized encoding and decoding mappings with a fully connected bottleneck layer of arbitrary dimensionality. These expressions are general and encompass convolutional maps for certain mapping functions.

Definition 2.6. Let x_t, z_t, y_t be random variables. A parameterized autoencoder network is given by the encoder mapping $z_t = \psi(\theta_{x_t} x_t + \rho_{x_t})$ and decoder mapping $y_t = \psi^{-1}(\theta_{y_t} z_t + \rho_{y_t})$ for some real-valued, differentiable function ψ; here, $\theta_{x_t}, \theta_{y_t}$ are weight matrices and ρ_{x_t}, ρ_{y_t} are biases.

Definition 2.7. Let x_t, y_t, z_t be random variables with marginals $p_{x_t}, p_{y_t}, p_{z_t}$ and conditional distributions $p_{z_t|x_t;\theta}, p_{z_t|y_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. Let $g_{x_t,y_t,z_t;\theta}$ be a real-valued penalty function. The unconstrained Rényi-Stratonovich value of information is

$$\inf_\theta \int_{\mathcal{X}} g_{x_t,y_t,z_t;\theta} p_{dx_t} - \int_{\mathcal{X}} \int_{\mathcal{Z}} g_{x_t,y_t,z_t;\theta} p_{dz_t|x_t;\theta} p_{dx_t} - \gamma(s_\alpha(\kappa_{z_t}) + s_\alpha(n\kappa_{z_t} \circ \kappa_{y_t}) - s_\alpha(\kappa_{y_t})),$$

where γ is a Lagrange multiplier and $\kappa_{x_t}, \kappa_{y_t}$ are normalized Gram matrices for the respective random variables.

Proposition 2.3. Assume the above definition for the form of the parameterized encoder and decoder mappings. The gradients of $s_\alpha(\kappa_{y_t})$ and $s_\alpha(n\kappa_{x_t} \circ \kappa_{y_t})$ with respect to κ_{y_t} are

$$\partial s_\alpha(\kappa_{y_t}) / \partial \kappa_{y_t} = \alpha U \Lambda^{\alpha-1} U^\top / (1-\alpha) \text{tr}(\kappa_{y_t}^\alpha)$$

$$\partial s_\alpha(n\kappa_{x_t} \circ \kappa_{y_t}) / \partial \kappa_{y_t} = \alpha (n\kappa_{x_t} \circ (V \Gamma^{\alpha-1} V^\top)) / (1-\alpha) \text{tr}((n\kappa_{x_t} \circ \kappa_{y_t})^\alpha)$$

where $U \Lambda U^\top$ and $V \Gamma V^\top$ are the eigenvalue decompositions of κ_{y_t} and $n\kappa_{x_t} \circ \kappa_{y_t}$, respectively. The partial derivatives of the conditional entropy with respect to the autoencoder parameters are

$$\partial s_\alpha(\kappa_{y_t}) / \partial \theta_{y_t} = -4(\theta_{y_t} Z_t^\top)(D_{y_t} - P_{y_t}) Z$$

$$\partial s_\alpha(\kappa_{y_t}) / \partial \theta_{x_t} = -4(\theta_{x_t}^\top \theta_{x_t} Z_t^\top (D_{y_t} - P_{y_t}) \circ \Psi_{z_t}^\top) X$$

$$\partial s_\alpha(\kappa_{y_t}) / \partial \rho_{y_t} = -4(\theta_{x_t}^\top \theta_{x_t} Z_t^\top (D_{y_t} - P_{y_t}) \circ \Psi_{z_t}^\top) 1_n$$

where $D_{y_t} = \text{diag}(P_{y_t} 1_n) - P_{y_t}$ and $P_{y_t} = \partial / \partial \kappa_{y_t} s_\alpha(\kappa_{y_t}) \circ \kappa_{y_t} / 2\sigma^2$, assuming a Gaussian kernel with variance σ^2. $\Psi_{z_t} = \psi'(Z_t)$ are the derivatives of the encoder non-linearity at $Z_t = X_t \Theta_{x_t}^\top + P_{x_t}$.

For $\alpha \in (0, 1]$, Rényi's mutual information is provably convex; for $\alpha \in (1, \infty)$, it is quasi-convex. The value of information is hence either convex or quasi-convex and a global minimum can be found.

III. Simulations

In this section, we assess the capability of convolutional autoencoder networks to recognize various objects when value-of-information regularization is used. Learning discriminative, reduced-dimensionality feature representations for these images is challenging due to the varieties in the objects' appearances, among other issues.

We have multiple simulation aims. Our main objective is to demonstrate that the value of information yields bottleneck-layer organizations which innately facilitate discrimination. Tied with this objective is that of understanding how the hyperparameter influences the organization. We additionally highlight that value-of-information regularized networks outperform existing regularization schemes for both unsupervised and supervised learning models. Many of these regularizers can be seen as a special case of the value of information, which permits direct comparisons and analyses of their behavior. Fewer epochs are also needed to achieve good classification performance when using the value of information. Our results discussions in this section focus on an understanding of why these various methodologies behave as they do.

A. Hyperparameter Effects Results and Discussions

In what follows, we assess the effects of the hyperparameter β on the discrimination performance for value-of-information-based networks applied to the CIFAR-10 dataset. We show that large values of β lead to better-performing features for classification, as they promote a disentanglement of the latent generating factors for the imagery.

Simulation Setup. Our convolutional autoencoder networks were implemented using the TensorFlow framework. For optimization purposes, we used ADAM-based gradient descent with mini batches [16]. ADAM maintains an average of the first two gradient moments and performs a bias correction to adjust the per-parameter learning rates, thereby facilitating quick convergence to good parameter values.

Except where otherwise noted, we considered the following parameter values and update schemes in our simulations. These parameters were informed by prior experiments on vision-based problems. An initial learning rate of 10^{-3} was chosen for ADAM. The learning rate was decreased by half every twenty epochs. We used exponential decay rates of 9.0×10^{-1} and 9.9×10^{-1} for the first- and second-order moments, respectively. An epsilon additive factor of 10^{-8} was employed to preempt division by zero. Lastly, a mini-batch size of 32 samples was used to ensure that the gradient estimation was sufficiently noisy to bias against terminating in poor local minima [17].

The results presented in this section were averaged across 100 Monte Carlo simulations where the network's initial parameters are randomly chosen. For each simulation, we randomly split the CIFAR-10 dataset into training, testing, and validation sets, where ratios of 70%, 15%, and 15% were used, respectively. Each simulation was terminated once the error on the validation set monotonically increased for 10 epochs. We report classification results for both an unsupervised cross-entropy-error and semi-supervised learning-by-association [18] cost function.

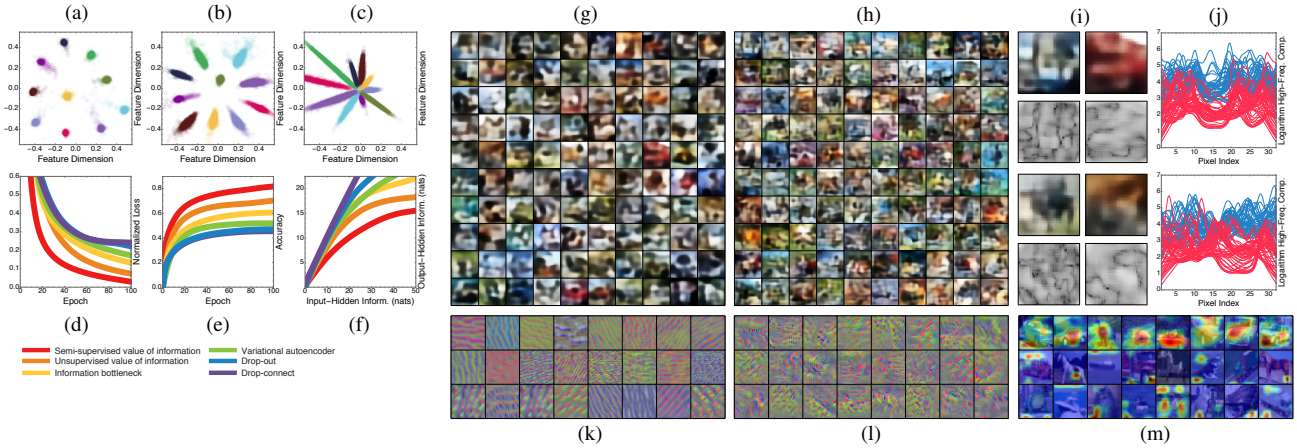

Figure 1: Regularized training results for the CIFAR-10 dataset. (a)–(c) Class-colored scatterplots highlighting the bottleneck-layer sample organization when using the Rényi-based value of information for training, where $\alpha = 0.99$ and $\gamma = 0.1$, 0.075, and 0.001, respectively. Higher values of γ, up to some data-dependent threshold, lead to better discriminability, due to the autoencoder mainly encoding features that are predictive of the class label. Lower values of γ favor autoencoders that implement an increasingly accurate identity map. (d)–(f) Plots of the smoothed, averaged normalized loss, validation set accuracy, and layer-wise mutual information for the various regularization schemes; averages were obtained over 100 random runs when using good, empirically-derived parameters. These statistics highlight that the value of information converges more quickly, leads to more predictive features, and filters out unnecessary information better than the alternatives. (g)–(h) Decoder reconstructions of random samples from the bottleneck-layer feature space for $\gamma = 0.1$ and 0.001, respectively. High values of γ only preserve low-frequency characteristics of the images, like rough shape and color information. Low values incorporate more details about textures and sharp edges. This is quantitatively illustrated in (i) and (j). The images in (i) include two images from (h) (left) and two from (i) (right) along with a visualization of their high-frequency components (below); it is clear from the zero-crossings (dark contours) in the high-frequency-component images that the images from (h) have more texture content. The plots in (j) quantitatively illustrate this behavior: the red and blue curves correspond, respectively, to the average logarithm of the high-frequency components for the images from (g) and (h). (k)–(l) Activation maps for the networks which generated (g) and (h). High values of γ lead to networks which capture only rough directional and color information from the images. Low values of γ permit reproducing complex patterns to produce better autoencoders. (m) Class activation heatmaps for the value of information (top row), variational autoencoders (middle row), and information bottleneck (bottom row). The networks trained using our criterion consistently focus their attention on regions where there is a target of interest. The other unsupervised regularization schemes often either are not stimulated by the target or focus on unimportant details.

We relied on an eleven-layer convolutional autoencoder topology for our simulations. The encoder stage network consisted of a $24\times24\times26$ convolution layer followed by a $12\times12\times26$ max-pooling layer. Following this was an $8\times8\times36$ convolution layer, a $4\times4\times36$ max-pooling layer, a 250-element fully-connected layer with leaky-rectified-linear-unit activation functions, and a 2-element bottleneck layer for visualization purposes. In some of our tests, we considered a higher-dimensional bottleneck layer to allow for more degrees of freedom for an appended classification layer consisting of linear activation functions. The decoder portion of the network was a reverse of this. It was composed of a 250-element fully connected layer, a $4\times4\times36$ de-convolution layer, an $8\times8\times36$ un-pooling layer, a $12\times12\times26$ de-convolution layer, and a $24\times24\times26$ un-pooling layer. Both the encoder and decoder parts of the network have a total of about 3.38×10^5 parameters. We found that alternate topologies extracted either too few or too many unnecessary image-based features for adequately representing the objects.

Results. The plots in figures 1(a)–(c) showcase the representation discriminability property of our criterion. High values of γ promote distributional compactness and separability, leading to trivially linearly separable bottleneck codes regardless of which of the two cost functions were used. The autoencoder performance, however, is poor. As γ decreases, the distributions begin to overlap, reducing the recognition rate but improving the autoencoder performance.

Figures 1(h)–(l) summarize the effects of the hyperparameter on the network response. For high values of γ, the decoder responses only preserve the most salient object details, such as shape and color. Few high-frequency components, such as texture and sharp edges, remain, as captured in figures 1(i)–1(j). The network filters are mainly sensitive to only directional and color details. For low values of γ, more high-frequency components remain, as the goal becomes to mimic the identity function; such components are needed to ensure an accurate reconstruction of the original image. The network filters converge to representations that retain details about complex patterns.

Discussions. Through these simulation results, we have demonstrated the effects of the hyperparameter γ on the regularization process. This parameter also naturally influenced the discrimination capability of the bottleneck-layer features.

Our results in figure 1(c) indicate that near-zero values of γ lead to convolutional autoencoders with little to no class-distribution organization. Significant class overlap can be encountered, which complicates the ensuing classification for both linear and non-linear appended layers. This was expected. As γ tends to zero, the importance of the total expected penalty term, $\mathbb{E}[\mathbb{E}[g_{x_t,y_t,z_t;\theta}|p_{z_t|x_t;\theta}]|p_{x_t}]$, diminishes. The network attempts to solely maximize the mutual dependence, $\mathbb{E}[D_{\mathrm{KL}}(p_{z_t|y_t;\theta}\|p_{z_t})|p_{y_t}]$, between the bottleneck-layer and output random variables. The bottleneck-layer representation hence carries combinations of details about the various objects beyond the type. These include, but are not limited to, the objects' orientations, outline and shape, which is corroborated by both the distributional overlap and the covariances. All of these characteristics aid in optimally reconstructing the images at the network output layer, but do not necessarily lend themselves to discrimination.

High values of γ, which tend toward the random-variable entropy, fair better for class discrimination, as we showed in figure 1(a)–(c). Such values assign a high weight to the total expected penalty, $\mathbb{E}[\mathbb{E}[g_{x_t,y_t,z_t;\theta}|p_{z_t|x_t;\theta}]|p_{x_t}]$. For our chosen penalty measure, the value of information is encouraging the bottleneck layer to maximally quantize the pseudo-information associated with the input while minimizing the uncertainty between the bottleneck-layer representation and the output. Minimally sufficient statistics, in a Neyman-Fisher sense, about the output from the input are therefore recovered. The resulting bottleneck-layer representation primarily encodes the objects type from a non-linear combination of various objects' shape and textural features. Samples from each class are mapped to compact, well-separated distributions in the reduced-order space, which are trivially linearly separable. The per-class distribution variance decreases as γ rises due to the additional imposed quantization. If γ is increased high enough, then each sample would be mapped to a distribution with zero variance. The resulting reconstruction is then just an approximate average of the class samples.

Values of γ between these two extremes lead to a compromise between the objective of being maximally predictive and maximally compressive. Particular ranges of γ where one objective dominates the other are data dependent. To help choose good values, we augmented the value-of-information criterion so that the rate-distortion-like curve was provably convex with respect to γ. Values of γ for which this augmented criterion achieved a maximal value correspond to the 'knee' region of the original criterion, which is where the two competing objectives become balanced. The resulting autoencoders are capable of faithfully reproducing the

images while still possessing a bottleneck-layer representation that is near-completely linearly separable, which we demonstrated in figure 1(a)–(b). Such optimal values of γ could be deduced algebraically, assuming that all of the data are available a priori. For streaming-data problems where the statistics of the images change dramatically with each batch, γ would need to be re-assessed periodically to ensure that a near-optimal trade-off between the quantization amount and prediction quality is maintained.

At a higher level, value-of-information-regularized convolutional autoencoders can be seen as learning a disentangled representation. That is, individual latent processing elements are sensitive only to changes in single generative factors while being mostly invariant to changes in other factors [19, 20]. In the context of object recognition, this implies that the value-of-information-trained networks learned independent latent units that are sensitive to generative factors like the object identity, its position and orientation, shape and scale, and so on, thus acting as an inverse graphics model [21]. In a disentangled representation, knowledge about one factor can generalize to novel configurations of other factors [22], which explains the good generalization performance that we obtained. As we elaborate on below, the hidden-unit disentanglement behavior also supports why the networks organized the bottleneck-layer features in the manner presented in figure 1(a).

We can prove that the value of information actively promotes disentanglement. To do this, we consider the total-correlation measure $\mathbb{E}[\log(q_{z_t})|q_{z_t}] - \sum_j \mathbb{E}[\log(q_{z_t^j})|q_{z_t^j}]$. This measure is zero if the components j of the intermediate-layer random variables are mutually independent and hence disentangled. Here, q_{z_t} is a factorized product distribution related to the marginal p_{z_t}, the latter of which is intractable to exactly compute. The value of information is implicitly forcing the total correlation to be zero in certain cases, which follows from re-writing the Shannon information $\mathbb{E}[D_{\mathrm{KL}}(p_{z_t|y_t;\theta}\|q_{z_t})|p_{y_t}]$, for the factorized distribution q_{z_t}, as $\mathbb{E}[D_{\mathrm{KL}}(q_{z_t|y_t;\theta}\|p_{z_t})|p_{y_t}]$, which can be further simplified to $D_{\mathrm{KL}}(p_{z_t|x_t;\theta}\|\prod_j q_{z_t^j}^j)$ and hence $D_{\mathrm{KL}}(q_{z_t}\|\prod_j q_{z_t^j}) + \mathbb{E}[D_{\mathrm{KL}}(q_{z_t|y_t;\theta}\|q_{z_t})|p_{y_t}]$, assuming mutual independence is adopted. Total correlation thus emerges as a constraint in the original objective function. The influence of total correlation on learning is dictated by the Lagrange-multiplier value. As the corresponding Lagrange multiplier rises, the network is increasingly forcing the total correlation toward zero, implying that the latent units are primarily sensitive to just the object identity and shape. This naturally yielded compact representations with a high between-class distance and minimal overlap, since the generating factors behind the objects were separated. As the multiplier decreased, it became difficult to disassociate the underlying factors, which led to class distributions that overlapped and had large covariances.

In summary, the class-sample quantization behaviors emerged due to the use of an information measure that restricts the amount of information passing through the network. The penalty function also played a role, and alternate choices may organize the features in ways that can possibly impede discrimination. If, for instance, a mean-squared-error-like penalty term was used within the value of information, then the network would just attempt to mimic the identity mapping for both small and large values of γ. No constraints would hence be imposed on the bottleneck feature distribution. This corresponds to the conventional learning case for autoencoder networks where class discrimination can hence be poor. Additionally, other information measures may yield worse behaviors, especially if they do not promote disentanglement. We show this in the next section, where we compare against conventional network regularization techniques that turn out to be special instances of the value of information with degenerate information measures that impede disentanglement.

B. Comparative Results and Discussions

We now demonstrate that value-of-information-trained networks can better discriminate between objects types than other popular regularization schemes. We also highlight that they generalize better to untrained samples, even ones with markedly different statistics. Both behaviors are due to a Markov-chain-aggregation property of the criterion.

Simulation Setup. For the ensuing simulations, we relied on the same training protocols and network configuration as in the previous section. We considered a range of alternate regularization approaches that used correlation-based error measures. These included network drop-out, network drop-connect, which we incorporated into ADAM-based gradient descent. We also considered the information-bottleneck method where mutual information is estimated via a variational approximation.

Both network drop-out [23] and drop-connect [24] operate by stochastically altering the topology of the network during training to prevent the processing elements from co-adapting too greatly. Drop-out does this by temporarily removing, at random, both processing elements and their connections. Drop-connect instead randomly selects a subset of connections between processing elements and sets their corresponding weights to zero. For our simulations, the chance of removing a connection was 5.0×10^{-1} and 8.0×10^{-1} for hidden-layer and input-layer processing elements, respectively, when using drop-out. For drop-connect, the probability of temporarily severing a connection was 4.0×10^{-1}. Such values were guided by previous empirical studies.

Results. Quantitative comparisons of the value of information with other conventional regularization schemes, for the same convolutional architecture, are provided in figures 1(d)–1(f). Figure 1(d) shows that the value of information cost function converges more quickly than the alternate regularization approaches, such as drop-out and drop-connect. Figure 1(e) highlights that our criterion generalizes better, regardless of the chosen cost function, in fewer epochs. They also encode more relevant information, which is captured in figure 1(e).

We also compared against both variational and information-bottleneck-regularized convolutional autoencoders. Neither approach fared as well as our criterion; see figures 1(e) and 1(m).

Discussions. The above results indicate that the value of information often outperforms common network regularizers.

In the context of parameter selection, optimizing the value of information attempts to minimize the 'distance' between a joint and marginal distribution such that the representations are increasingly sufficient, in a Neyman-Fisher sense. Both drop-out and drop-connect propagate a possibly non-optimal amount of information through the network and hence do not necessarily yield representations that are minimally sufficient statistics.

The fact that the value of information will outperform these regularizers can be definitively proved; here we provide a sketch of this fact. Both drop-out and drop-connect are special cases of the value of information where the hidden-layer random variables are, respectively, corrupted by multiplicative Bernoulli noise with unit mean and multiplicative log-normal noise with zero mean. In this case, the Rényi information term is marginalized, leading to a Rényi α-divergence, $D_{\mathrm{R}}^{\alpha}(p_{z_t|x_t;\theta}|p_{z_t})$. As well, the penalty function is an augmented cross-entropy term $\mathbb{E}[-\log(p_{y_t|z_t;\theta})|p_{z_t|x_t;\theta}]$. For the same penalty function, the value of information yields equal or better parameter values, which follows since $\mathbb{E}[D_{\mathrm{R}}^{\alpha}(p_{z_t|y_t;\theta}\|p_{z_t})|p_{y_t}] \leq D_{\mathrm{KL}}^{\alpha}(p_{z_t|x_t;\theta}|p_{z_t})$ due to Jensen's inequality and some additional arguments. Hence, global solutions to drop-out and drop-connect lie on or above the value-of-information's rate-distortion-like curve.

Variational autoencoders [25] can be considered as instances of the value of information. Similar arguments to those we used above can be used to prove that the value of information will either perform similarly or better than it. This explains our empirical findings.

The value of information converges to good network parameter values more quickly than other regularizers. This is due to its Markov-chain aggregation properties [6]. As we noted in [26, 27], autoencoder networks can be viewed as analogues of Markov chains. In [6], we proved that the value of information quantizes the state space of Markov chains. The criterion maps state groups from the original chain to states of a reduced-order chain and provides probabilistic one-to-many correspondences between those two group sets. The number of state groups is dictated entirely by the Lagrange multiplier. For low multiplier values, many state groups exist in the reduced-order chain. This complicates the training process, since a wide gamut of parameter values are considered during training. The parameter values are sought that are specialized to groups containing a few input samples in an attempt to model them well and recover the identity function. This case leads to a reduced-order Markov chain with similar long-run dynamics as the original chain;

there is little to no regularization since a large parameter space must be explored. For high parameter values, the chains are compressed greatly and parameter regularization occurs. This implicitly leads to training over a topologically simpler network.

V. Conclusions

The value of information is an information-theoretic criterion that describes the maximum benefit that can be obtained from a piece of information for either increasing expected rewards or reducing average costs. We have previously shown that this property facilitates optimal decision-making under uncertainty. Here, we have exploited it to regularize parameter selection when training convolutional autoencoder networks for object classification.

More specifically, the value of information trades off between the competing objectives of mimicking the identity function and compressing the autoencoder representation, as dictated by the class organization. The emphasis placed on either objective is dictated by a single user-selectable hyperparameter. Near-zero values of this hyperparameter do not regularize parameter selection. There are hence no constraints on the bottleneck-layer class organization, which corresponds to conventional training for autoencoder networks. As the hyperparameter values is increased, the samples from each class start to become grouped in the bottleneck layer. High values ensure that the each class' samples are compactly distributed and well separated. That is, features are extracted from the input samples that facilitate discrimination, which should have been an explicit behavior implemented by autoencoder networks since their inception.

Efficiently optimizing the value of information is imperative for choosing thousands to millions of network parameters. Our previous work on solving the value of information relied on either expectation-maximization updates or numerical continuation updates, neither of which would be computationally tractable for the complicated network topologies. They also would not be able to easily handle the continuous random variable case. We thus developed an alternate, more general optimization scheme that is based on approximating value-of-information solutions. That is, we showed how to nearly tightly bound the criterion while still preserving its differentiability. This allowed us to use mini-batch-based gradient descent for tuning the network parameters in the continuous random-variable case. This approach is also applicable to discrete random variables.

In our simulations, we highlighted the appropriateness of this training regularization for both supervised and unsupervised learning. We demonstrated that, regardless of the bottleneck-layer dimensionality, the value of information promoted the formation of compact, well-separated class distributions for appropriate hyperparameter values. That is, the features extracted by the networks led to classes that were trivially linearly separable. This permitted recognition for a variety of diverse object types. The features also were relatively insensitive to nuisances, such as background variations. As a consequence, they tended to perform similarly to features extracted by unsupervised convolutional autoencoder networks. In the latter case, this was because conventional autoencoder networks can produce features that lead to highly overlapping class distributions. This occurs even if the class labels are used to help choose the weights in the encoder and decoder portions of the networks.

In the future, we will explore alternate applications of the value of information for network training. We will show that the value of information can be used to perform layer-wise parameter selection in conventional neural networks. In this context, the value of information will be limiting the amount of information that flows through each layer of the network. This should have the effect of speeding up network training time compared to conventional deep architectures. It should also guarantee that the network achieves a certain level of performance.

A. Appendix

We first demonstrate that our Rényi conditional α-entropy can be related to Shannon conditional entropy.

Proposition A.1. Let y_t, z_t be random variables with marginals p_{y_t}, p_{z_t} along with a conditional distribution $p_{z_t|y_t;\theta}$ and joint $p_{y_t,z_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. We have that our Rényi conditional α-entropy is equivalent to Shannon conditional entropy in the limit of $\alpha \to 1$,

$$\lim_{\alpha\to 1}\log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)} = -\int_{\mathcal{Y}}\int_{\mathcal{Z}} p_{dy_t,dz_t;\theta}\log(p_{y_t,z_t;\theta}/p_{dz_t}).$$

As a consequence of this equality, our Rényi α-mutual-information is equivalent to Shannon mutual information in the same limit. This property follows from the equivalency of Rényi's α-entropy and Shannon entropy in this case.

We can also show consistency of our Rényi's conditional α-entropy with conditional min- and max-entropy.

Proposition A.2. Let x_t, y_t be random variables with marginals p_{y_t}, p_{z_t} and a conditional distribution $p_{z_t|y_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. We have that our Rényi conditional α-entropy is equivalent to,

(i) Conditional min-entropy: When $\alpha \to \infty$,

$$\lim_{\alpha\to\infty}\log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)} = -\log\Big(\int_{\mathcal{Y}} p_{dy_t}\inf_{\mathcal{Z}} p_{z_t|y_t;\theta}\Big).$$

(ii) Conditional max-entropy: When $\alpha \to 0$,

$$\lim_{\alpha\to 0}\log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)} = \log\Big(\sup_{y_t\in\mathcal{Y}}\Big|\sup_{z_t\in\mathcal{Z}} p_{z_t|y_t;\theta}\Big|\Big).$$

We now show that our Rényi's conditional α-entropy is monotonically decreasing with respect to α; this is a similar behavior to the non-conditional α-entropy case.

Proposition A.3. Let x_t, y_t be random variables with marginals p_{y_t}, p_{z_t} along with a conditional distribution $p_{z_t|y_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. For $\alpha \leq \beta$, $\alpha, \beta \in [0, \infty]$, we have that

$$\log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)} \leq \log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\beta}\Big)^{1/\beta} p_{dy_t}\Big)^{\beta/(\beta-1)},$$

which is naturally less than or equal to $\log(|\mathcal{Z}|)$.

It is apparent that our Rényi's conditional α-entropy $H^{\alpha}_{z_t|y_t;\theta}$ is non-negative, so these quantities are bounded below by zero.

Two important properties that we verify are that conditioning on one variable has the potential to reduce the uncertainty in another, on average, and that the weak chain rule of conditional entropy is satisfied. It therefore behaves similarly to Shannon entropy.

Proposition A.4. Let x_t, y_t be random variables with marginals p_{y_t}, p_{z_t} along with a conditional distribution $p_{z_t|y_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. For $\alpha \in [0, \infty]$, we have that

(i) Conditioning reduces average uncertainty: $H^{\alpha}_{z_t} \geq H^{\alpha}_{z_t|y_t;\theta}$,

$$\log\Big(\int_{\mathcal{Z}} p^{\alpha}_{dz_t}/(1-\alpha)\Big) \geq \log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)},$$

(ii) Weak chain rule: $H^{\alpha}_{z_t|y_t;\theta} \geq H^{\alpha}_{z_t,y_t;\theta} - H^{0}_{z_t}$,

$$\log\Big(\int_{\mathcal{Y}}\Big(\int_{\mathcal{Z}} p_{dz_t|y_t;\theta}^{\alpha}\Big)^{1/\alpha} p_{dy_t}\Big)^{\alpha/(\alpha-1)} \geq \int_{\mathcal{Y}}\int_{\mathcal{Z}} p^{\alpha}_{dy_t,dz_t;\theta}/(1-\alpha) - H^{0}_{z_t}.$$

Demonstrating that the chain rule is satisfied is important. For two random variables y_t, z_t, the joint α-entropy given by the expression $H^{\alpha}_{y_t,z_t} = \log(\mathbb{E}[p^{\alpha}_{y_t,z_t;\theta}/p_{y_t}\, p_{z_t}|p_{y_t}\, p_{z_t}])/(1-\alpha)$ describes the number of bits, on average, needed to describe the exact system state. If we first learn the value of z_t, then we have gained $H^{0}_{z_t}$ bits of information and at least $H^{\alpha}_{y_t,z_t} - H^{0}_{z_t}$ remaining bits are needed to describe the entire system state. If an information-theoretic measure does not satisfy the chain rule, then a greater number of bits will be needed. We therefore would not be able to directly express the joint α-entropy in terms of the difference between the conditional and marginal, which would complicate our learning strategy, since we rely on this property to re-write our cost function in a more easy-to-optimize form.

It can be shown that existing definitions of Rényi's conditional α-entropy only satisfy either the weak chain rule property or the uncertainty reduction property, not both. They would therefore be poorly suited as regularizers for restricting the flow of information through deep networks.

Proposition A.5. Let x_t, y_t be random variables with marginals p_{y_t}, p_{z_t}

along with a conditional distribution $p_{z_t|y_t;\theta}$ and joint $p_{y_t,z_t;\theta}$ parameterized by $\theta \in \mathbb{R}^p$. For $\alpha \in [0, \infty]$, we have that

(i) The following conditional α-entropies do not reduce average uncertainty by conditioning

$$\int_{\mathcal{Y}} \log(\int_{\mathcal{Z}} p_{dy_t,dz_t;\theta}^{\alpha} / p_{dy_t}^{\alpha}) p_{dy_t} / (1-\alpha),$$
$$\log(\int_{\mathcal{Y}} \int_{\mathcal{Z}} p_{dy_t,dz_t;\theta}^{\alpha} / \int_{\mathcal{Y}} p_{dy_t}^{\alpha}) / (1-\alpha).$$

(ii) The following conditional α-entropies do not satisfy either a weak or strong chain rule,

$$\log(\int_{\mathcal{Y}} (\int_{\mathcal{Z}} p_{dy_t,dz_t;\theta}^{\alpha} / p_{dy_t}^{\alpha}) p_{dy_t}) / (1-\alpha),$$
$$\log(\int_{\mathcal{Y}} (\int_{\mathcal{Z}} p_{dy_t,dz_t;\theta}^{\alpha} / p_{dy_t}^{\alpha})^{1/(\alpha-1)} p_{dy_t})^{\alpha-1} / (1-\alpha).$$

REFERENCES

[1] J. C. Príncipe, *Information Theoretic Learning*. New York City, NY, USA: Springer-Verlag, 2010.

[2] R. L. Stratonovich, "On value of information," *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, vol. 5, no. 1, pp. 3–12, 1965.

[3] R. L. Stratonovich and B. A. Grishanin, "Value of information when an estimated random variable is hidden," *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, vol. 6, no. 1, pp. 3–15, 1966.

[4] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 22-24 1999, pp. 368–377.

[5] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 165–188, 2005.

[6] I. J. Sledge and J. C. Príncipe, "Reduction of Markov chains using a value-of-information-based approach," *Entropy*, vol. 21, no. 4, pp. 349(1–34), 2019. Available: http://dx.doi.org/10.3390/e21040349

[7] ——, "An analysis of the value of information when exploring stochastic, discrete multi-armed bandits," *Entropy*, vol. 20, no. 3, pp. 155(1–34), 2018. Available: http://dx.doi.org/10.3390/e20030155

[8] ——, "Analysis of agent expertise in Ms. Pac-Man using value-of-information-based policies," *IEEE Transactions on Computational Intelligence and Artificial Intelligence in Games*, vol. 11, no. 2, pp. 142–158, 2019. Available: http://dx.doi.org/10.1109/TG.2018.2808201

[9] I. J. Sledge, M. S. Emigh, and J. C. Príncipe, "Guided policy exploration for Markov decision processes using an uncertainty-based value-of-information criterion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2080–2098, 2018. Available: http://dx.doi.org/10.1109/TNNLS.2018.2812709

[10] I. J. Sledge and J. C. Príncipe, "Adapting the exploration rate for value-of-information-based reinforcement learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, (under review). Available: http://arxiv.org/abs/1702.08628

[11] R. V. Belavkin, "Asymmetry of risk and value of information," in *Dynamics of Information Systems*, C. Vogiatzis, J. Walteros, and P. Pardalos, Eds. New York, NY, USA: Springer-Verlag, 2014, pp. 1–20.

[12] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1249–1257, 1992. Available: http://dx.doi.org/10.1109/18.144705

[13] ——, "Constrained clustering as an optimization method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 8, pp. 785–794, 1993. Available: http://dx.doi.org/10.1109/34.236251

[14] L. G. Sánchez-Giraldo, M. Rao, and J. C. Príncipe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 535–548, 2015. Available: http://dx.doi.org10.1109/TIT.2014.2370058

[15] S. Yu, L. G. Sánchez-Giraldo, R. Jenssen, and J. C. Príncipe, "Multivariate extension of matrix-based Renyí's α-order entropy functional," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, (under review). Available: https://arxiv.org/abs/1808.07912

[16] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 7-9 2015, pp. 1–15. Available: https://arxiv.org/abs/1412.6980

[17] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Proceedings of the International Conference on Machine Learning (ICML)*, New York, NY, USA, June 19-24 2016, pp. 1225–1234.

[18] P. Haeusser, A. Mordvintsev, and D. Cremers, "Learning by association: A versatile semi-supervised training method for neural networks," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 21-26 2017, pp. 89–98. Available: http://dx.doi.org/10.1109/CVPR.2017.74

[19] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, April 24-26 2017, pp. 1–22.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. Available: http://dx.doi.org/10.1109/TPAMI.2013.50

[21] T. D. Kulkami, W. F. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems (NIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 2539–2547.

[22] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that think like people," *Behavioral and Brain Sciences*, vol. 40, no. 1, pp. e253(1–58), 2017. Available: http://dx.doi.org/10.1017/S0140525X16001837

[23] S. Wang and C. Manning, "Fast dropout training," in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, GA, USA, June 16-21 2013, pp. 118–126.

[24] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of neural networks using DropConnect," in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, GA, USA, June 16-21 2013, pp. 1058–1066.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA, May 2-4 2013, pp. 1–14. Available: https://arxiv.org/abs/1312.6114

[26] S. Yu and J. C. Príncipe, "Understanding autoencoders with information-theoretic concepts," *Neural Networks*, vol. 117, no. 1, pp. 104–123, 2019. Available: http://dx.doi.org/10.1016/j.neunet.2019.05.003

[27] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe, "Understanding convolutional neural network training with information theory," *IEEE Transactions on Information Theory*, 2019, (under review). Available: https://arxiv.org/abs/1804.06537