Gradient Information for Representation and Modeling

Jie Ding, Robert Calderbank, Vahid Tarokh

***Abstract*—Motivated by Fisher divergence, in this paper we present a new set of information quantities which we refer to as gradient information. These measures serve as surrogates for classical information measures such as those based on logarithmic loss, Kullback-Leibler divergence, directed Shannon information, etc. in many data-processing scenarios of interest, and often provide significant computational advantage, improved stability and robustness. As an example, we apply these measures to the Chow-Liu tree algorithm, and demonstrate remarkable performance and significant computational reduction using both synthetic and real data.**

***Index Terms*—Capacity; Fisher divergence; Hyvarinen loss; Information; Stability; Tree Approximation.**

I. Introduction

A standard step in data fitting and statistical model selection is the application of a loss function (sometimes referred to as scoring function) of the form $s : (y, p) \mapsto s(y, p)$, where y is the observed data and $p(\cdot)$ is a density function. In this context, it is assumed that the smaller $s(y, p)$, the better y fits p. A class of such functions that exhibit desirable statistical properties has been studied in the context of proper scoring functions [1]. As a special case, the logarithmic loss $s_L(y, p) = -\log p(y)$ has served as the cornerstone of classical statistical analysis because of the intimate relation between logarithmic loss and the Kullback-Leibler (KL) divergence. In fact, the KL divergence from a density function p to the true data-generating density function p_* can be written as $\mathbb{E}\{s_L(y, p)\} + c$, where the expectation of $s_L(y, p)$ is taken under p_*), and c is a constant that only depends on p_*. Therefore, minimizing the sample average $n^{-1}\sum_{i=1}^{n} s_L(y_i, p)$ over a set of candidates density functions p asymptotically amounts to finding the closest candidate $\hat{p}$ to p_* in KL divergence.

This research is funded by the Defense Advanced Research Projects Agency (DARPA) under grant numbers W911NF1810134 and HR00111890040.

J. Ding is with the School of Statistics, University of Minnesota, Minneapolis, Minnesota 55414, United States. V. Tarokh and R. Calderbank are with the Information Initiative at Duke University, Durham, North Carolina 27708, United States.

A notable use of the logarithmic loss function is in maximum likelihood estimation for parametric models. By minimizing $n^{-1}\sum_{i=1}^{n} s_L(y_i, p_\theta)$ over $\theta \in \Theta$ for some parameter space Θ, an estimate $\hat{\theta}$ is obtained to represent the data generating model. Some commonly used objective loss function such as cross-entropy loss for classification and squared loss for regression can be regarded as special cases of the logarithmic loss function. Another important use of the logarithmic loss is in model comparison and model selection. In the presence of multiple candidate models, data analysts have to follow a model selection principle to select the most appropriate model for interpretation or prediction. The log-likelihood function, which can be regarded as logarithmic loss evaluated at observed data, play a crucial role in most state-of-the-art principles, including information criteria, Bayesian likelihood, Bayes factors (see, e.g., [2], [3] and the references therein). The logarithmic loss and KL divergence are also foundational in inference and information processing, exemplified by their use in variational inference [4], contrastive divergence learning [5], learning with information gains [6]–[8], etc.

Is the logarithmic loss always the best choice? In a series of recent works, a new loss function (also referred to as the Hyvarinen scoring function) [9] has been proposed as a surrogate for logarithmic loss function for statistical inference and machine learning. It is defined by

$$s_H(y, p) = \frac{1}{2}\|\nabla_y \log p(y)\|^2 + \Delta_y \log p(y), \quad (1)$$

where ∇ denotes the gradient and Δ denotes the Laplacian, and p is defined over an unbounded domain. It was first proposed in the context of parameter inference, which can produce (in a way similar to the logarithmic loss) a consistent estimation of θ of a probability density function $p_\theta(y)$ [9]. It was shown that $s_H(y, p)$ is computationally simpler to calculate in most cases, particularly in the case of intractable normalizing constants. This enables a richer class of models for prediction given the same amount of computational resources. It was discovered that the Hyvarinen loss also enjoys desirable properties in Bayesian model comparison and provides

better interpretability for Bayesian model selection in the presence of vague or improper priors, compared with classical Bayesian principles based on marginal likelihoods or Bayes factors [10], [11].

On the other hand, from an information theoretic view, the differential entropy (for a random variable) is the expectation of its logarithmic loss, the mutual information (between two random variables) is the KL divergence from the product density function to the joint density function, and the mutual information is linked to differential entropy through the chain rule. This view is crucial to motivate our new information measures. Motivated by the definition of Shannon's differential entropy, it is natural to define the "entropy" for the Hyvarinen loss. This turns out to be intimately related to Fisher divergence. In fact the classical mutual information may be re-defined based on Fisher divergence instead of KL divergence. It turns out that these quantities still can be interpreted in information theoretic manner.

The main contributions of our work are described next. First, motivated by some recent advances in scoring functions, we propose a set of information quantities to measure the uncertainty, dependence, and stability of random variables. We study their theoretical properties, resemblance and difference to the existing counterpart measures widely used in machine learning. Second, we provide interpretations and applications of the proposed gradient information, e.g. to fast tree approximation and community discovery. Third, we point out some interesting directions enabled by gradient information, including a new form of causality for predictive modeling, channel coding where the stability of channel capacity is of a major concern, and general inequalities that could have been highly nontrivial from an algebraic point of view.

The paper is outlined as follows. In Section II, we introduce the Hyvarinen loss function and extend its scope from unbounded to bounded continuous random variables. We introduce a set of quantities in Section II-C referred to as gradient information measures, and study their properties, interpretations, and implications for machine learning. In Section III, we provide some applications of the proposed concepts, including a new algorithm for graphical modeling that parallels the classical Chow-Liu tree algorithm (but from an alternative perspective that can enjoy computational benefit), a new supervised learning algorithm, and a community discovery algorithm. We conclude the paper and share our thoughts on some future research in Section IV.

II. Gradient Information and Its Properties

A. Fisher divergence and Hyvarinen loss

We focus on multidimensional continuous random variables (often denoted by $Y \in \mathbb{R}^d$) in this paper unless otherwise stated. We use p and $\mathbb{E}$ to denote the density function and expectation with respect to the distribution of Y, respectively. The jth entry of Y is denoted by Y_j ($j = 1, \ldots, d$). Let $[Y, Z]$ denote the joint vector that consists of Y and Z. We often use upper and a lower case letters to respectively denote a random variable and its realizations. We consider a class of distributions $\mathcal{P}$ over $\mathbb{R}^d$ that consists of distributions whose Lebesgue density $p(\cdot)$ is a twice continuously differentiable function. We use $\mathcal{N}(\mu, V)$ to denote a Gaussian distribution of mean μ and covariance V. We use $\|\cdot\|$ to denote the Euclidean norm. For a joint density function $p(\cdot)$ of (Y, Z), let $p_{Z|Y}$ denote $(y, z) \mapsto p(z \mid y)$, a function of both y and z.

Suppose p denotes the true-data generating distribution that is usually unknown. Given observed data $y_1, \ldots, y_n$, a typical machine learning problem is to search for a density function q (usually parameterized) over some space that is the most representative of the data. For that purpose, a measure of difference between probability distributions are needed. The existing literature largely replies on the KL divergence, which is defined by

$$D_{\mathrm{KL}}(p, q) = \mathbb{E}\{\log p(Y)/\log q(Y)\}$$

(to log base e), where $p(\cdot), q(\cdot)$ are two probability density functions and $\mathbb{E}$ is with respect to p. Note that

$$D_{\mathrm{KL}}(p, q) = -\mathbb{E}\{\log q(Y)\} + \mathbb{E}\{\log p(Y)\}$$

and it is only minimized at $q = p$ (almost everywhere). This implies that $-\mathbb{E}\{\log q(Y)\}$ is minimized at $q = p$. A direct consequence of the above observation is the use of maximum likelihood estimation that minimizes $-\sum_{i=1}^n \log q(y_i)$. By the law of large numbers, the estimator $\hat{q}$ can be proved to be close to p for large n. A possible alternative to KL divergence is the following. The Fisher divergence from a probability density function $q(\cdot)$ to another $p(\cdot)$ is

$$D_{\mathrm{F}}(p, q) = \frac{1}{2} \int_{\mathbb{R}^d} \|\nabla_y \log q(y) - \nabla_y \log p(y)\|^2 p(y) dy,$$

where ∇ is the gradient. It is also referred to as generalized Fisher information distance in physics, and has found application in statistical inference (see, e.g., [12]), with the difference that the ∇ operator is with respect to the parameter instead of data. By similar argument as in the KL divergence, the following result was proved in [9].

Proposition 1. *Suppose that the following regularity conditions hold:*

1) $p(\cdot)$ *and* $\nabla_y \log q(\cdot)$ *on* $(-\infty, \infty)$ *are continuously differentiable,*

2) $\mathbb{E}\|\log \nabla_y p(y)\|^2$ *and* $\mathbb{E}\|\log \nabla_y q(y)\|^2$ *are finite,*

3) $\lim_{|y_j|\to\infty} p(y) \cdot \partial_j \log q(y) = 0$*,* $(j = 1, \ldots, d)$

then we have

$$D_{\mathrm{F}}(p, q) = \mathbb{E}\{s_H(y, q)\} + \frac{1}{2}\mathbb{E}\|\nabla_y \log q(y)\|^2. \quad (2)$$

where $s_H(y, q)$*, referred to as the Hyvarinen loss function, is defined in (1).*

Suppose that a set of observations $y_1, \ldots, y_n$ are drawn from some unknown p, a sample analog of $\mathbb{E}\{s_H(y, q)\}$ can be used to search for the q from a space of density functions to approximate p (just like the maximum likelihood estimation). Hyvarinen loss function is particularly powerful when the probability density function is only known up to a multiplicative normalization constant. This is because (1) is invariant under a constant multiplication of $p(y)$. There has been an extension to discrete random variables (see, e.g., [13], [14]). We refer to [9], [11] for more details of Hyvarinen loss in the context of i.i.d. and time series settings.

B. Extension of Hyvarinen loss to partially bounded random variables

The definition in (1) only applies to unbounded random variables. Following an extension to nonnegative variables [15], we further extend the Hyvarinen loss to general continuous variables including unbounded, partially-bounded, and fully bounded cases. Suppose that $y_j \in [a_j, b_j]$, where a_j, b_j $(j = 1, \ldots, d)$ may be at infinity (unbounded case). Suppose there exist nonnegative integers α_j, β_j such that

$$p(y) \cdot \partial_{y_j} \log q(y) \cdot (y_j - a_j)^{2\alpha_j}(y_j - b_j)^{2\beta_j} \to 0 \quad (3)$$

as $y \to a_j^+$ or $y \to b_j^-$ for all densities q within the specified model class and true data-generating density p. In the above condition, when $a_j = -\infty$ (resp. $b_j = \infty$), it is understood that $\alpha_j = 0$, $(y_j - a_j)^{2\alpha_j} = 1$ (resp. $\beta_j = 0$, $(y_j - b_j)^{2\beta_j} = 1$). Note that the assumption made in [9] corresponds to the unbounded case with $\alpha_j = \beta_j = 0$. For any two vectors $u, v \in \mathbb{R}^d$, and a vector of nonnegative integers $w \in \mathbb{N}^d$, let $u \circ v$ and u^α denote vectors in $\mathbb{R}^d$ whose jth entry is $u_j v_j$, $u_j^{\alpha_j}$, respectively. As a special case, $v^0 = 1$ for any scalar v. Clearly, the operation $\circ$ is associative. Let $a, b, \alpha, \beta \in \mathbb{R}^d$, and consider

$$D_\nabla(p, q) = \frac{1}{2}\int_{\mathcal{D}_y} \|\{\nabla \log q(y) - \nabla \log p(y)\} \circ \{(y - a)^\alpha\} \circ \{(y - b)^\beta\}\|^2 p(y) dy. \quad (4)$$

Theorem 1. $D_\nabla(p, q)$ *defined as above equals zero if and only if* p *equals* q *almost everywhere. Moreover, assume condition (3) holds,* $p(\cdot)$ *and* $\nabla_y \log q(\cdot)$ *are continuously differentiable, and*

$$\max_{1\le j\le d} \mathbb{E}|y^{\alpha_j+\beta_j}\partial_j \log h(y)|^2 < \infty, \quad \text{for } h = p, q.$$

Then $D_\nabla(p, q)$ *can be written as* $s_\nabla(y, q) + c_p$*, where* c_p *is a constant that only depends on* p*, and* $s_\nabla(y, q)$ *is defined as*

$$\frac{1}{2}\|\{\nabla \log q(y)\} \circ \{(y - a)^\alpha\} \circ \{(y - b)^\beta\}\|^2 + \sum_{j=1}^d \partial_{y_j}\left\{(\partial_{y_j} \log q(y))(y_j - a_j)^{2\alpha_j}(y_j - b_j)^{2\beta_j}\right\}. \quad (5)$$

The regularity conditions made in Theorem 1 are mild and hold for many commonly used distributions such as sub-Gaussian and sub-exponential families. By Theorem 1, the extended $s_\nabla(y, q)$ in (5) inherits the properties of (1) and is applicable to a wide range of continuous random variables. To summarize its desirable properties, $s_\nabla(y, q)$ is

(P1) only a function of $y, \nabla q(y), \nabla^2 q(y)$, which usually has an analytic form to evaluate (for parametric q);

(P2) invariant under scaling of q, which can be quite favorable when the parameterized density q is known up to a normalizing constant;

(P3) statistically proper [1] in the sense that its expectation is only minimized at $q = p$ (almost everywhere) where q is the true data generating density;

(P4) applicable to a wide range of continuous random variables whose entries may be bounded or unbounded or a mixture of them.

C. Information quantities and properties

The classical information quantities largely rely on KL divergence. For instance, the Shannon entropy is the expectation of the log loss function, and the mutual information is the KL divergence between product of marginal densities and joint density. Likewise, starting from the Hyvarinen loss, we define the following entropy, conditional entropy and mutual information that are generally referred to as gradient information.

Definition 1 (Gradient information)**.** *For a continuous random variable* $Y = [Y_1, Y_2]$ *we define the following information quantities:*

- *Entropy:* $H_\nabla(Y) = \mathbb{E}\{s_\nabla(Y, p_Y)\}$
- *Conditional entropy:* $H_\nabla(Z \mid Y) = E_p\{s_\nabla([Y, Z], p_{Z|Y})\}$
- *Mutual information:* $I_\nabla(Y, Z) = D_\nabla(p_{YZ}, p_Y p_Z)$.

where p_Y, p_Z denotes the marginal densities and p_{YZ} denotes the joint density of random variables Y, Z.

For partially bounded random variables in Subsection II-B, we let α, β be the smallest nonnegative integers such that (3) holds. The above definition can be extended to discrete random variables but we leave this extension to a future work. The gradient entropy (denoted by 'G-entropy') along with the Shannon entropy ('S-entropy') for some common distribution families are tabulated in Table I.

Let $q = p$ in Theorem 1, it is not difficult to observe the following identity.

$$H_\nabla(Y) = -\frac{1}{2} J(Y) \tag{6}$$

where

$$J(Y) = \|\{\nabla \log p(y)\} \circ \{(y-a)^\alpha\} \circ \{(y-b)^\beta\}\|^2.$$

For unbounded Y and zero vectors α, β, $J(Y)$ reduces to $\int_{\mathbb{R}} p(y) \|\nabla_y \log p(y)\|^2 dy$ which is sometimes called the Fisher information of Y that has many implications in physics (see, e.g. [16]). A related but different definition of Fisher information is the variance of the partial derivative with respect to the parameter of a log-likelihood function. As a consequence of (6), we have $H_\nabla(Y) \leq 0$. Its interpretation is elaborated in Subsection II-D. Also, $I_\nabla(Y, Z)$ equals to zero if and only if Y and Z are independent.

Next we show that the above information quantities enjoy desirable quantities such as chain rule and conditioning reduces entropy that are reminiscent of the properties of Shannon Information. However, they can be more suitable for machine learning due to computational and interpretation advantages we shall point out.

Theorem 2. *Suppose the assumptions in Theorem 1 hold. We have the chain rules*

$$I_\nabla(Y; Z) = H_\nabla(Y) + H_\nabla(Z) - H_\nabla(Y, Z) \tag{7}$$

$$H_\nabla(Y, Z) = H_\nabla(Y) + H_\nabla(Z \mid Y) \tag{8}$$

As a by product of Theorem 2, $I_\nabla(Y, Z) = H_\nabla(Z) - H_\nabla(Z \mid Y) \geq 0$ (i.e. conditioning reduces entropy), with equality if and only if Y, Z are independent.

We may also define the following "generalized association":

$$I_c(Y; Z) = -\frac{I_\nabla(Y; Z)}{H_\nabla(Y, Z)} \tag{9}$$

between two random variables Y, Z. It can be proved that $I_c \in [0, 1)$, and $I_c = 0$ if and only if Y, Z are independent. In the bivariate Gaussian case, $I_c = \rho^2$ where ρ is the usual correlation.

It is worth mentioning that not all properties of gradient information are counterparts of those in classical Shannon information. Examples are given in the following proposition that are used in proving our results in the supplementary material.

Proposition 2. *For any two unbounded random variables Y, Z whose joint distribution exists and satisfies conditions of Proposition 1, we have*

$$H_\nabla(Z \mid Y) \leq \mathbb{E}\{H_\nabla(Z \mid Y = y)\} \tag{10}$$

where the expectation on the right hand side is with respect to Y.

Suppose further that Y and Z are independent, then

$$H_\nabla(Y + Z \mid Z) = 2H_\nabla(Y). \tag{11}$$

Compared with (Shannon) differential entropy of a random variable that measures its descriptive complexity, the entropy and conditional entropy in Definition 1 measure the uncertainty in prediction. The following Proposition 3 serves as a continuous analog of Fano's inequality that bounds the mean-squared prediction error, and is also intimately related to the Cramér-Rao bound. It can be extended to multidimensional case but we do not pursue the details here.

Proposition 3. *Suppose that Y is an unbounded scalar random variable to predict, and X is a variable (providing side information about Y) such that the joint distribution of (X, Y) exists. Suppose that $\hat{Y}(X)$ is any estimate of Y which is only a function of X. Then the expected $\mathbb{L}^2$ prediction error satisfies*

$$\mathbb{E}(Y - \hat{Y}(X))^2 \geq \frac{1}{-2H_\nabla(Y \mid X)} \tag{12}$$

with equality if and only if Y is Gaussian and independent with X, and $\hat{Y}(X) = \mathbb{E}Y$.

D. Stability interpretation

In modern machine learning systems with uncertainty in data generating processes, stability is a key issue of concern. We introduce a relationship between the gradient information and KL divergence based information that is widely used in machine learning. For brevity, we narrow our scope to unbounded continuous random variables and introduce the following definition and results. Its proof follows from the de Bruijn's identity, Theorem 1 of [17], and Theorem 2 in Section II.

Definition 2 (Perturbed random variable)**.** *Let Y be any random variable with a finite variance with density $p(\cdot)$. Let e be an standard Gaussian random variable independent of Y. Let $Y_v = Y + \sqrt{v}e$ with density $p_v(\cdot)$.*

TABLE I: Examples of gradient entropy versus Shannon entropy for common distributions

DISTRIBUTION	PARAMETER	DENSITY	SUPPORT	G-ENTROPY	S-ENTROPY
GAUSSIAN	MEAN μ, VARIANCE σ^2	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(y-\mu)^2/2\sigma^2}$	$(-\infty,\infty)$	$-\frac{1}{2\sigma^2}$	$\frac{1}{2}\log(2\pi e\sigma^2)$
GAMMA	SHAPE α AND RATE β	$\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}$	$[0,\infty)$	$-\frac{1}{2}(\alpha+1)$	$\alpha+\log\frac{\Gamma(\alpha)}{\beta}+(1-\alpha)\psi(\alpha)$
EXPONENTIAL	RATE λ	$\lambda e^{-\lambda y}$	$[0,\infty)$	-1	$1-\log(\lambda)$
UNIFORM	RANGE a,b	$\frac{1}{b-a}$	$[a,b]$	0	$\log(b-a)$
PARETO	SCALE a, SHAPE γ	$\frac{\gamma a^\gamma}{y^{\gamma+1}}$	$[a,\infty)$	$-\frac{1+\gamma}{2+\gamma}$	$\log\frac{a}{\gamma}+1+\frac{1}{\gamma}$

Proposition 4. *We have*

$$H_\nabla(Y) = -\frac{d}{dv}H(Y_v)\mid_{v=0},$$
$$D_\nabla(p,q) = -\frac{d}{dv}D_{\mathrm{KL}}(p_v,q_v)\mid_{v=0}$$
$$H_\nabla(Z\mid Y) = -\frac{d}{dv}H(Z_v\mid Y_v)\mid_{v=0}$$
$$I_\nabla(Y;Z) = -\frac{d}{dv}I(Y_v;Z_v)\mid_{v=0}$$

The identities in Proposition 4 indicates that gradient information describes the non-equilibrium dynamics of the its counterpart under KL divergence. It further indicates that minimizing gradient information is intimately related to enhancing stability. We refer to [17] for more discussions on this. Moreover, the non-negativeness of $I_\nabla(Y;Z)$ means that perturbing Y and Z with independent noises will decrease their mutual information. Similar interpretations apply to other three identities.

E. Implication on causality

The identification of causality usually serves as a key step towards simplified modeling and learning. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be two sequences of data. In general, we say a series X_t causes another series Y_t if knowing the past $\{X_1,\ldots,X_{t-1}\}$ can provide information on the future of Y_t given the past of $\{Y_1,\ldots,Y_{t-1}\}$. This school of thoughts is exemplified by the seminal work of Granger [18] in identifying causal relations between multivariate times series. The Granger causality is typically tested in linear models between Y_t and X_t (with lags) and the two processes are assumed to be stationary [19]. In general, this type of of causality can be unified by Kolmogorov complexity $\mathbb{K}(\cdot)$ which, not only extends Granger causality to nonstationary and nonlinear processes, but also includes various other approximations of Kolmogorov information in the literature, such as Shannon's mutual information, Renyi's information, directed Shannon information, directed Renyi information, combinatorial measures of information (e.g. Lempel-Ziv information).

Intuitively, the quantity measures how much complexity of the series $\{Y_t\}$ is reduced by knowing $\{X_t\}$. The past $\{X_1,\ldots,X_{t-1}\}$ provide information about Y_t if $\mathbb{K}(Y_t|X_{t-1},\ldots,X_1,Y_{t-1},\ldots,Y_1) < \mathbb{K}(Y_t|Y_{t-1},\ldots,Y_1)$ for Kolmogorov complexity measure $\mathbb{K}(\cdot)$. In this case, additional predictive information is provided by the series $\{X_t\}$ if

$$\mathbb{K}(Y_t|Y_{t-1},\ldots,Y_1) - \mathbb{K}(Y_t|Y_{t-1},\ldots,Y_1,X_{t-1},\ldots,X_1)$$

is greater than zero. We define the left hand side of the above term to be the Kolmogorov causal information provided by series $\{X_t\}$ for predicting $\{Y_t\}$.

However, Kolmogorov information is in general not computable and its surrogates may be used instead. For example, consider replacing the complexity measure $\mathbb{K}$ by Shannon entropy $H(Y) = -\int p(y)\log p(y)dy$ for a continuous random variable Y. In this case, Kolmogorov causal information reduces to the directed Shannon information [20].

Using gradient entropy as the surrogate for the Kolmogorov information, we can define gradient-directed information (as a surrogate for Kolmogorov causal information) as:

$$H_\nabla(Y_t|Y_{t-1},\ldots,Y_1) - H_\nabla(Y_t|Y_{t-1},\ldots,Y_1,X_{t-1},\ldots,X_1).$$

The above measure provides an alternative method of measuring the causality from the sensitivity point of view. It also provides significant computational advantages particularly when density normalizing constants are unknown.

III. APPLICATIONS OF GRADIENT INFORMATION

A. Tree approximation of joint distributions

Many machine learning tasks involve high dimensional data where the number of observations is large compared to the number of variables. One approach that

we often undertake is divide and conquer (e.g. grouping different dimensions in smaller subsets, removing irreverent connections). We then bootstrap the models we learn for the subsets or sparse graphs to the entire data dimensions. An effective approach is to assume a tree dependence structure among the variables to simplify the learning, compress data, or to find good initial states for more complex graphical models. In light of the properties (P1)-(P4) of $s_P(y,p)$ (in Subsection II-B), we expect to develop faster and more stable algorithms for approximating high-dimensional data distributions based on gradient information.

In particular, given an nth-order probability distribution $p(X_1,\ldots,X_n)$ with X_i being continuous random variables, we wish to find the following optimal first-order dependence tree p_τ.

Definition 3. *A distribution $p_{\tau_0}(X_1,\ldots,X_n)$ ($\tau_0 \in T_n$) follows the optimal first-order dependence tree, if*

$$D_\nabla(p,p_{\tau_0}) \le D_\nabla(p,p_\tau)$$

for all $\tau \in T_n$, where T_n is the set of all possible first-order dependence trees (see Fig. 1 for an illustration).

An analogous definition is given in the pioneering work of Chow [21], but we use D_∇ instead of D_{KL} here to measure the discrepancy of approximation. Exhaustive search is a not computationally feasible for any moderately large n, since there are n^{n-2} dependence trees in T_n (by Cayley's formula). Motivated by the Chow-Liu algorithm [21] (based on the KL divergence), we provide the following theorem that enables tree approximation of joint distributions using a fast greedy algorithm with $O(n)$-complexity. The problem is formulated as the search for a maximum spanning tree. The procedure is outlined in Algo. 1.

Theorem 3. *Under the assumptions made in Theorem 2, the best tree approximation of a joint distribution $p(y_1,\ldots,y_n)$ under Fisher divergence $D_\nabla(\cdot,\cdot)$ is the maximum spanning tree with weight $H_\nabla(Y_i,Y_{j(i)})$, where $(i,j(i))$ is an edge that denotes $p(Y_i \mid Y_{j(i)})$.*

We apply our algorithm to a protein signaling flow cytometry dataset. The dataset encodes the presence of $p=11$ proteins in $n=7466$ cells. It was first analyzed using Bayesian networks in [23] who fit a directed acyclic graph to the data, later studied in [24] using different methods. The tree can be used as a graphical visualization tool to highlight the most highly correlated genes in the correlation network.

We suppose that any pair of random variables Y_1, Y_2 follow the following exponential family distribution

$$p(y_1,y_2) \propto \exp\{\theta_1 y_1^2 y_2^2 + \theta_2 y_1^2 + \theta_3 y_2^2 + \theta_4 y_1 y_2 + \theta_5 y_1 + \theta_6 y_2\} \quad (13)$$

Note that the constant is a function of θ and it does not have a closed form. The above distribution is a special case of a class of exponential family distributions with normal conditionals. This family is intriguing from the perspective of graphical modeling as, in contrast to the Gaussian case, conditional dependence may also express itself in the variances [24]. To estimate the density, we minimize the sample average of (1), and obtain a closed form solution $\hat{\theta}$ (to be elaborated in the supplement). Based on the estimates, we can obtain a consistent estimator of the entropy $H_\nabla(Y_1,Y_2)$ by a sample analog of (6), using Monte Carlo samples generated from the estimated density. To calculate $H_\nabla(Y_1)$ and $H_\nabla(Y_2)$, we calculate the marginal distributions in closed form, and obtain a consistent estimation of entropy and mutual information by sample analogs. The details are elaborated in the "derivations for exponential family example" section in the supplement. Figure 2 shows the network structure after applying our method to the data using the proposed approach. Our result is consistent with the estimated graph structure in [23].

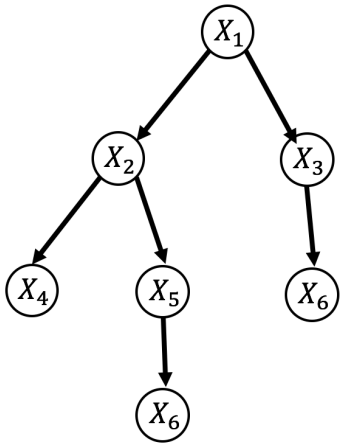

Fig. 1: Illustration of joint distribution with tree structure: $p(x_1,x_2,x_3,x_4,x_5,x_6) = p(x_1)\ p(x_2 \mid x_1)\ p(x_4 \mid x_2)\ p(x_5 \mid x_2)\ p(x_6 \mid x_5)\ p(x_3 \mid x_1)\ p(x_6 \mid x_3)$.

We record the computational time by running different number of variables (p from 2 to 11) in Fig 3, which shows the gradient information based algorithm is more than 100 times faster than Shannon information based algorithm.

It is worth noting that the method of Algo. 1 can also be used to perform supervised classification. Given features $X_1,\ldots,X_p$ and their corresponding label Y, we

Algorithm 1 Generic tree approximation based on gradient information

input Observations of $Y_1, \ldots, Y_p$

output A first-order dependence trees τ, and (optionally) a joint density $p(y_1, \ldots, y_p)$ built on τ

1: Estimate $I_\nabla(Y_i, Y_j)$ for each $i \neq j, i, j = 1, \ldots, p$.
2: Build an undirected weighted graph with p vertices representing $Y_1, \ldots, Y_p$, where the weight between vertices i, j is $I_\nabla(Y_i, Y_j)$.
3: Apply Kruskal's algorithm [22] (or alternative algorithms) to obtain a maximum spanning tree τ.
4: Derive or approximate the conditional distribution $p(y_i \mid y_{j(i)})$ that corresponds to each edge $(i, j(i))$, and thus obtain $p(y_1, \ldots, y_p)$.

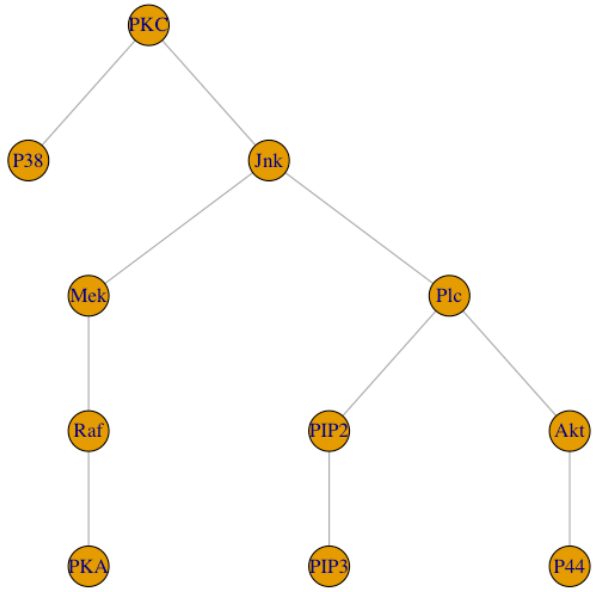

Fig. 2: The tree discovered from the protein data.

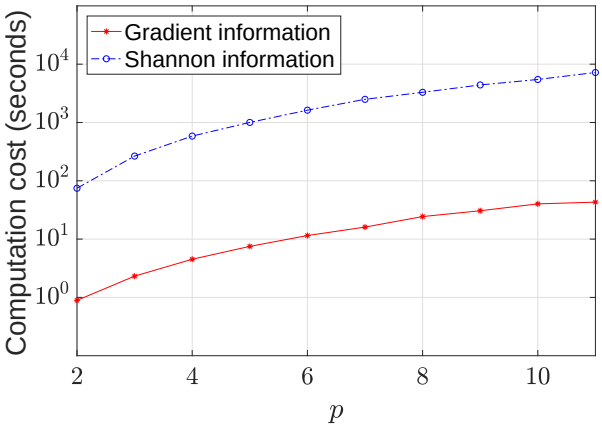

Fig. 3: Comparison of computational costs of Chow-Liu tree approximation using classical Shannon information [21] and gradient information (proposed here), depicted in logarithmic scale.

TABLE II: Classification accuracy of three methods for data with different levels of correlation

	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
CHOW-LIU TREE	0.610	0.829	0.965	0.994
RANDOM FOREST	0.607	0.759	0.886	0.953
ELASTIC NET	0.535	0.537	0.541	0.526

calculate the tree distribution that approximates the joint distribution of $p(x_1, \ldots, x_p)$ for each class of Y. We then perform a likelihood ratio test to decide which class a given feature vector $x_1, \ldots, x_p$ is associated with. In calculating the joint density, we let the spanning tree be rooted at a node with the largest geodesic distance and the rest of the nodes are ranked according to the height directed preorder (HDP) traversal of the tree [25]. In a synthetic data experiment, we generate two classes of data from an independent Gaussian vector $[X_1, \ldots, X_p]$. The covariance $\mathrm{Cov}(X_i, X_j)$ of the first class of data is $\rho^{|i-j|}$, and the covariance of the second class is $(-\rho)^{|i-j|}$ (for $i, j = 1, \ldots, p$). We generated 100 data in each of the 1000 replications and record the cross validation accuracy (with 30% test data) in Table II. We also compared our method (denoted by "Chow-Liu tree") with two popular classification methods, random forest [26] and elastic net [27]. The results indicate the superior performance of our method. Elastic net does not work well for this example mainly because the two classes of data are not linearly separable.

B. Community discovery

Many real-world networks of data exhibit a community structure: the vertices of the network are partitioned into groups such that the statistical dependence is high among vertices in the same group and low otherwise. Most of the community detection methods (e.g. stochastic block models [28]) focus on the concentration of linkages in random graphs. Here we provide an alternative perspective using gradient mutual information.

Algorithm 2 Community discovery based on mutual information

input Observations of $Y_1, \ldots, Y_p$, number of communities k ($1 \leq k \leq p$)
output A partition of $\{1, \ldots, p\}$ into k subsets
1: Apply Algo. 1 to obtain the spanning tree τ (with weights being the mutual information).
2: Remove the k edges that have the smallest weights from τ.
3: The output partition is represented by the connected components of the current τ.

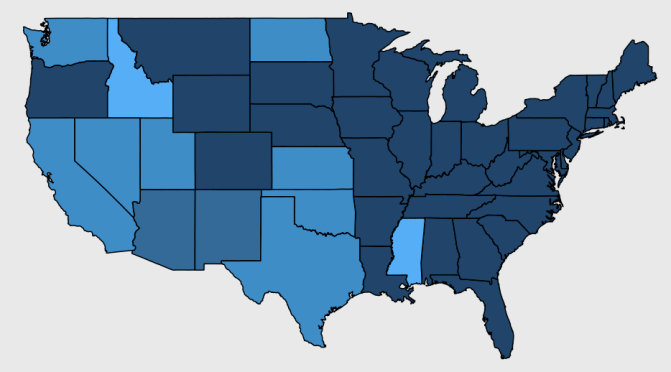

Fig. 4: The communities of states detected from quarterly growth rates of payroll employment for the U.S. states (the number of communities is set to be four).

More precisely, we do not perform communities discovery from edge connections, but from dependence among variables/vertices (assuming multiple observations at each vertex). Such dependence is quantified by mutual information. And the obtained communities can be understood as disjoint subsets of variables that exhibit large within-community dependence and small inter-community dependence. We propose a fast community discovery approach based on Algo. 1. The main idea, summarized in Algo. 2, is to first obtain a spanning tree that best represents the joint distribution, and then construct communities by removing weak-dependence connections.

In a data study, we considered a dataset constructed in [29]. The data was also studied in [30] using an algorithm that recovers the communities using the eigenvectors of the sample covariance matrix. The data consists of quarterly growth rates of payroll employment for the U.S. states (excluding Alaska and Hawaii) from the second quarter of 1956 to the fourth of 2007, which results in a panel of $n = 48$ time series over $T = 207$ periods. The data are seasonally adjusted and annualized. We show the results of applying our clustering algorithm to the sample in Figure 4. The communities roughly match the clusters of Fig. 3 in [30] using a partial correlation network model.

IV. Conclusions

Representation and modeling of data from limited observations are key issues in machine learning. Existing methods for determining randomness, conditional randomness, stability, causality, discrepancy, information gains, etc. in data representation and modeling largely depend on KL divergence and logarithmic loss. For enhanced stability and reduced computationally cost, at least in some occasions, we introduced gradient information as a possible surrogate to classical information measures that have been widely used in machine learning practice (e.g. mutual information in feature selection, KL divergence in variational inference). Gradient information is motivated from a perspective different from that of the KL divergence, and provides new theoretical tools for representation of information and data modeling.

Some future directions for research are outlined below. First, the propose gradient information and its properties can be extended to discrete random variables. Second, it would be interesting to apply the tree approximation algorithm based on gradient information to more complicated graphical models (e.g. better initialization or dimension reduction). Third, gradient information may be utilized to a broad range of machine learning frameworks such as variational inference, information gains, neural networks, etc. to ensure stability in estimation, reduced amount of training data, and computational complexity.

Acknowledgement

We would like to thank Prof. Yu Xiang for inspiring discussions.

Supplementary Material

In the supplement, we include proofs of theoretical results, and provide two additional applications of gradient information. One application is about channel stability. The second application is in derivation of some general inequalities using gradient information (which are otherwise highly nontrivial to establish).

References

[1] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Ann. Stat.*, pp. 561–592, 2012.

[2] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: a new criterion for autoregression," *IEEE Trans. Inf. Theory*, 2017.

[3] ——, "Model selection techniques–an overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, 2018.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Am. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[5] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning." in *Aistats*, vol. 10. Citeseer, 2005, pp. 33–40.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[7] A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. E. Socolar, "Mutual information in random boolean models of regulatory networks," *Physical Review E*, vol. 77, no. 1, p. 011901, 2008.

[8] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[9] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. Apr, pp. 695–709, 2005.

[10] A. P. Dawid, M. Musio *et al.*, "Bayesian model selection based on proper scoring rules," *Bayesian Anal.*, vol. 10, no. 2, pp. 479–499, 2015.

[11] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, "Bayesian model comparison with the Hyvarinen score: computation and consistency," *J. Am. Stat. Assoc.*, 2018.

[12] C. Holmes and S. Walker, "Assigning a value to a power likelihood in a general bayesian model," *Biometrika*, vol. 104, no. 2, pp. 497–503, 2017.

[13] A. D. Hendrickson and R. J. Buehler, "Proper scores for probability forecasters," *Ann. Math. Stat.*, pp. 1916–1921, 1971.

[14] A. P. Dawid, S. Lauritzen, M. Parry *et al.*, "Proper local scoring rules on discrete sample spaces," *Ann. Stat.*, vol. 40, no. 1, pp. 593–608, 2012.

[15] A. Hyvärinen, "Some extensions of score matching," *Computational statistics & data analysis*, vol. 51, no. 5, pp. 2499–2512, 2007.

[16] B. R. Frieden, "Fisher information, disorder, and the equilibrium distributions of physics," *Physical Review A*, vol. 41, no. 8, p. 4265, 1990.

[17] S. Lyu, "Interpretation and generalization of score matching," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 359–366.

[18] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, pp. 424–438, 1969.

[19] ——, "Some recent development in a concept of causality," *Journal of Econometrics*, vol. 39, pp. 199–211, 1988.

[20] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*. Citeseer, 1990, pp. 303–305.

[21] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[22] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Am. Math. Soc.*, vol. 7, no. 1, pp. 48–50, 1956.

[23] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[24] M. Yu, M. Kolar, and V. Gupta, "Statistical inference for pairwise graphical models using score matching," in *Advances in Neural Information Processing Systems*, 2016, pp. 2829–2837.

[25] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *Ann. Stat.*, pp. 697–717, 1979.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, vol. 67, no. 2, pp. 301–320, 2005.

[28] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, vol. 84, no. 6, p. 066106, 2011.

[29] J. D. Hamilton and M. T. Owyang, "The propagation of regional recessions," *Rev. Econ. Stat.*, vol. 94, no. 4, pp. 935–947, 2012.

[30] C. T. Brownlees, G. Gudmundsson, and G. Lugosi, "Community detection in partial correlation network models," 2017.

[31] J.-F. Bercher and C. Vignat, "On minimum fisher information distributions with restricted support and fixed variance," *Information Sciences*, vol. 179, no. 22, pp. 3832–3842, 2009.

[32] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. 11, no. 2, pp. 267–271, 1965.

[33] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

Supplementary Material for "Gradient Information for Representation and Modeling"

Proof of Theorem 1

It can be seen that $D_\nabla(p,q)=0$ if and only if $\nabla p(y)=\nabla q(y)$ almost everywhere. Since p and q are densities and integrate to one, $\nabla p(y)=\nabla q(y)$ is equivalent to $p(y)=q(y)$ almost everywhere.

Using direct calculation and integration by parts, we have

$$\begin{aligned}D_\nabla(p,q)=&\frac{1}{2}\int_{[a,b]}p(y)|(\nabla\log q(y))\circ(y-a)^{\alpha_j}\circ(y-b)^{\beta_j}|^2dx\\&-\sum_{j=1}^{d}\int_{[a_j,b_j]}p(y)\ (\partial_{y_j}\log p(y))(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_j)^{2\beta_j}dy_j+C\end{aligned}$$

and the j-th term in the above summation is

$$\begin{aligned}&-\int_{[a_j,b_j]}p(y)\ (\partial_{y_j}\log p(y))(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_j)^{2\beta_j}dy_j\\&=-\int_{[a_j,b_j]}(\partial_{y_j}p(y))\,(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_j)^{2\beta_j}dy_j\\&=-p(y)(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_j)^{2\beta_j}\ |_{a_j}^{b_j}+\int_{[a_j,b_j]}p(y)\,\partial_{y_j}\left\{(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_j)^{2\beta_j}\right\}dy_j\\&=\int_{[a_j,b_j]}p(y)\,\partial_{y_j}\left\{(\partial_{y_j}\log q(y))(y_j-a_j)^{2\alpha_j}(y_j-b_d)^{2\beta_j}\right\}dy_j\end{aligned}$$

where the last identity holds under (3).

Proof of Theorem 2

We use $\nabla_{y,z}$ and ∇_y to highlight that the derivative is taken with regard to $[y,z]$ and y, respectively. We only prove for unbounded Y,Z. The proof of the extended case is similar, as discussed in Subsection II-B.

We first prove Identity (7). Applying Proposition 1 and Identity (6), we obtain the following identities (where expectations are with respect to p_{YZ}):

$$\begin{aligned}D_\nabla(p_{YZ},p_Yp_Z)&=\mathbb{E}\{s_\nabla([Y,Z],p_Yp_Z)\}+\frac{1}{2}\mathbb{E}\|\nabla_{y,z}\log p_{Y,Z}(Y,Z)\|^2\\&=\frac{1}{2}\mathbb{E}\bigg(\|\nabla_y\log\{p_Y(Y)p_Z(Z)\}\|^2+\|\nabla_z\log\{p_Y(Y)p_Z(Z)\}\|^2\bigg)\\&\quad+\Delta_y\log\{p_Y(Y)p_Z(Z)\}+\Delta_z\log\{p_Y(Y)p_Z(Z)\}-\mathbb{E}\{s_\nabla([Y,Z],p_{YZ})\}\\&=\mathbb{E}\{s_\nabla(Y,p_Y)\}+\mathbb{E}\{s_\nabla(Z,p_Z)\}-\mathbb{E}\{s_\nabla([Y,Z],p_{YZ})\}.\end{aligned}$$

We then prove Identity (8). Direct calculations give

$$\begin{aligned}\mathbb{E}\{s_\nabla([Y,Z],p_{Y,Z})\}&=\frac{1}{2}\mathbb{E}\|\nabla_{y,z}\{\log p_Y(Y)+\log p_{Z|Y}(Z\mid Y)\}\|^2+\mathbb{E}\big(\Delta_{y,z}\{\log p_Y(Y)+\log p_{Z|Y}(Z\mid Y)\}\big)\\&=\frac{1}{2}\mathbb{E}\big(\|\nabla_y\log p_Y(Y)\|^2\\&\quad+\|\nabla_{y,z}\log p_{Z|Y}(Z\mid Y)\|^2\big)+\mathbb{E}\{\Delta_y\log p_Y(Y)\}+\mathbb{E}\{\Delta_{y,z}\log p_{Z|Y}(Z\mid Y)\}+c\\&=H_\nabla(Y)+H_\nabla(Z\mid Y)+c\end{aligned}$$

where c denotes

$$c=\mathbb{E}\{\nabla_y\log p_Y(Y)^{\mathrm{T}}\cdot\nabla_y\log p_{Z|Y}(Z\mid Y)\}.$$

It remains to show that $c = 0$. We use d_y, $\mathcal{D}_y$, $\mathcal{D}_{y_{(-j)}}$ to denote the dimension of Y, domain of Y, domain of the subvector of Y excluding dimension j, respectively. $\mathcal{D}_z$ and $\mathcal{D}_{y,z}$ are similarly defined. We have

$$\begin{aligned}
c &= \mathbb{E}\{\nabla_y \log p_Y(Y)^{\mathrm{T}} \cdot \nabla_y \log p_{Z|Y}(Z \mid Y)\} \\
&= \sum_{j=1}^{d_y} \int_{\mathcal{D}_{y,z}} p_{Y,Z}(y,z) \frac{\partial_{y_j} p_Y(y)}{p_Y(y)} \cdot \frac{\partial_{y_j} p_{Z|Y}(z \mid y)}{p(z \mid y)} dy dz \\
&= \sum_{j=1}^{d_y} \int_{\mathcal{D}_{y,z}} \partial_{y_j} p_Y(y) \cdot \partial_{y_j} p(z \mid y) dy dz \\
&= \sum_{j=1}^{d_y} \int_{\mathcal{D}_{y_{(-j)},z}} \left(p_Y(y) \cdot \partial_{y_j} p(z \mid y) \mid_{-\infty}^{\infty} - \int_{\mathcal{D}_j} p_Y(y) \partial_{y_j}^2 p(z \mid y) dy_j \right) \prod_{k \neq j} dy_k dz \\
&= -\int_{\mathcal{D}_y} p_Y(y) \left(\int_{\mathcal{D}_z} \Delta_y p(z \mid y) dz \right) dy = -\int_{\mathcal{D}_y} p_Y(y) \left(\Delta_y \int_{\mathcal{D}_z} p(z \mid y) dz \right) dy = 0.
\end{aligned}$$

Proof of Proposition 2

We first prove (10). By a derivation similar to the proof of Theorem 2, we obtain

$$H_\nabla(Z \mid Y) = -\frac{1}{2} \mathbb{E} \| \nabla_{[z,y]} \log p(z \mid y) \|^2 \tag{14}$$

for any two random variables Z, Y whose joint distribution exists. Therefore,

$$H_\nabla(Z \mid Y) = -\frac{1}{2} \mathbb{E} \| \nabla_{Y,Z} \log p_{Z|Y}(Z \mid Y) \|^2 \leq -\frac{1}{2} \mathbb{E} \| \nabla_Z \log p_{Z|Y}(Z \mid Y) \|^2 = \mathbb{E}\{H_\nabla(Z \mid Y = y)\}.$$

We then prove (11). Suppose $W = Y + Z$. It follows from (14) that

$$\begin{aligned}
H_\nabla(W \mid Z) &= -2 \times \frac{1}{2} \mathbb{E} \| \nabla_w \log p(w - z) \|^2 \\
&= -\mathbb{E} \| \nabla_y \log p(y) \|^2 = 2 H_\nabla(Y).
\end{aligned}$$

Proof of Proposition 3

Lemma 1. *Given a fixed covariance matrix V of a random variable Y supported on $\mathbb{R}^d$, the distribution that maximizes $H_\nabla(Y)$ is Gaussian (with an arbitrary mean), and the maximum is $-Tr(V^{-1})/2$.*

We now prove that the maximum entropy distribution on $\mathbb{R}^d$ is Gaussian given second moment constraints. The results are readily observable from the known results that the distribution with a fixed variance that minimizes the Fisher information is the Gaussian distribution, typically proved using calculus of variations and differential equations (see, e.g. [31]). Here we provide a much simpler proof.

Proof: Suppose that Y_1, Y_2 are two i.i.d. random variables following the maximum entropy distribution. Then $2Y_1$ follows the maximum entropy distribution with variance $2V$, and by definition, $J(2Y_1) \leq J(Y_1 + Y_2)$. Direction calculations show that $J(2Y_1) = J(Y_1)/2$, therefore, it follows from the convolution inequality [32] that

$$\frac{J(Y_1)}{2} = J(2Y_1) \leq J(Y_1 + Y_2) \leq \frac{1}{J(Y_1)^{-1} + J(Y_2)^{-1}} = \frac{J(Y_1)}{2}$$

with equality only if the equality for convolution inequality for Fisher information holds, which implies that Y_1, Y_2 must be Gaussian. ∎

Proof of Proposition 3:

By Lemma 1, we have

$$\mathbb{E}(Y - \hat{Y}(X))^2 = \mathbb{E}_X \mathbb{E}_{Y|X}(Y - \hat{Y}(x) \mid X = x)^2 \geq \mathbb{E}_X Var(Y \mid X = x) \geq \mathbb{E}_X \frac{1}{-2H_\nabla(Y \mid X = x)}.$$

Moreover, applying Cauchy's inequality and Identity 10, we obtain

$$\mathbb{E}_X \frac{1}{-2H_\nabla(Y \mid X = x)} \geq \frac{1}{\mathbb{E}_X\{-2H_\nabla(Y \mid X = x)\}} \geq \frac{1}{-2H_\nabla(Y \mid X)}.$$

This concludes the proof.

PROOF OF THEOREM 3

Let p^a denote a distribution with first-order dependence tree structure. Using Proposition 1 and Identity 6, we have

$$\mathbb{E}\{D_\nabla(p,p^a)\} = \mathbb{E}\{s_\nabla(Y,p^a)\} + \frac{1}{2}\mathbb{E}\|\nabla_y \log p(y)\|^2 = \mathbb{E}\{s_\nabla(Y,p^a)\} - H_\nabla(Y) \tag{15}$$

In order to minimize $E\{D_\nabla(p,p^a)\}$, we only need to minimize $\mathbb{E}\{s_\nabla(Y,p^a)\}$. Let $j(i)$ index the parent node of i on the tree that represents p^a. Without loss of generality, let Y_1 denote the root of the tree, and E_a denote the set of edges. By Identity (8) in Theorem 2, the Markovity of p^a, and the fact that $p^a(\cdot)$ agrees with $p(\cdot)$ on all the first and second order marginal distributions, we can rewrite $\mathbb{E}\{s_\nabla(Y,p^a)\}$ as

$$\begin{aligned}
\mathbb{E}\{s_\nabla(Y,p^a)\} &= \mathbb{E}\{s_\nabla(Y_1,p^a_1)\} + \sum_{\{i,j(i)\}\in E_a} \mathbb{E}\{s_\nabla(Y_i,p^a_{i|j(i)})\} \\
&= H_\nabla(Y_1) + \sum_{\{i,j(i)\}\in E_a} \{H_\nabla(Y_i) - I_\nabla(Y_i;Y_{j(i)})\} \\
&= \sum_{i=1}^n H_\nabla(Y_i) - \sum_{\{i,j(i)\}\in E_a} I_\nabla(Y_i;Y_{j(i)}).
\end{aligned}$$

This concludes the proof.

DERIVATIONS FOR EXPONENTIAL FAMILY EXAMPLE

The sample average of (1) is a quadratic function of $\theta = [\theta_1,\ldots,\theta_5]$. The closed form solution $\hat\theta$ is derived as

$$\hat\theta = -\left\{\sum_{j=1}^n (a_{1,j}a_{1,j}^{\mathrm{T}} + a_{2,j}a_{2,j}^{\mathrm{T}})\right\}^{-1} \sum_{j=1}^n (a_{3,j} + a_{4,j}) \tag{16}$$

where

$$\begin{aligned}
a_{1,j} &= [2y_{1,i}y_{2,i}^2, 2y_{1,i}, 0, y_{2,i}, 1, 0]^{\mathrm{T}}, \quad a_{2,j} = [2y_{1,i}^2 y_{2,i}, 0, 2y_{2,i}, y_{1,i}, 0, 1]^{\mathrm{T}}, \\
a_{3,j} &= [2y_{2,i}^2, 2, 0, 0, 0, 0]^{\mathrm{T}}, \quad a_{4,j} = [2y_{1,i}^2, 0, 2, 0, 0, 0]^{\mathrm{T}},
\end{aligned}$$

The distribution density of Y_2 is

$$p(y_2) \propto (-\theta_1 y_2^2 - \theta_2)^{-1/2} \exp\left\{-\frac{(\theta_4 y_2 + \theta_5)^2}{4(\theta_1 y_2^2 + \theta_2)} + \theta_3 y_2^2 + \theta_6 y_2\right\} \tag{17}$$

So its entropy can be estimated by

$$-\frac{1}{2n}\sum_{j=1}^n \|\nabla \log p_{\hat\theta}(y_2)\|^2 = -\frac{1}{2n}\sum_{j=1}^n \left\{-\frac{1}{2}\frac{2\hat\theta_1 y_{2,j}}{\hat\theta_1 y_{2,j}^2 + \hat\theta_2} + \frac{1}{4}\frac{2\hat\theta_1 y_{2,i}(\hat\theta_4 y_{2,i} + \hat\theta_5)^2}{(\theta_1 y_{2,i}^2 + \theta_2)^2} - \frac{1}{4}\frac{2\hat\theta_4(\hat\theta_4 y_{2,i} + \hat\theta_5)}{\theta_1 y_{2,i}^2 + \theta_2} + 2\hat\theta_3 y_{2,i} + \hat\theta_6\right\}^2.$$

The value of $H_\nabla(Y_1)$ can be similarly estimated. We therefore get an consistent (under some moment conditions) estimate $I_\nabla(Y_1,Y_2)$ from Proposition 7.

The conditional distribution density $p(y_1 \mid y_2)$ can be calculated from (13) and (17):

$$p(y_1 \mid y_2) \propto (-\theta_1 y_2^2 - \theta_2)^{1/2} \exp\left\{\theta_1 y_1^2 y_2^2 + \theta_2 y_1^2 + \theta_4 y_1 y_2 + \theta_5 y_1 + \frac{(\theta_4 y_2 + \theta_5)^2}{4(\theta_1 y_2^2 + \theta_2)}\right\} \tag{18}$$

APPLICATION: STABILITY OF CHANNEL CAPACITY

Consider input X, output Y, and a time invariant channel described by conditional distribution $p_{Y|X}$. The channel capacity in many cases is achieved by maximizing $I(X;Y)$ over the marginal distribution of X, p_X (see for example the channel coding theorem [33]). In practical applications, we may also be interested in the stability of channel capacity, in the sense that the capacity is not very sensitive to perturbations at both ends of the channel. One possible way to define the channel stability is through the definition of $I_\nabla(X;Y)$ which can be interpreted as the sensitivity of $I(X;Y)$: the smaller the better (see Subsection II-D).

It is thus reasonable to assume that solutions of the following types of optimization problem would lead to channel coding that is both efficient in transmitting signals and robust against noise.

$$\max_{p_X} I(X;Y)/I_\nabla(X;Y). \tag{19}$$

It has been well known that Gaussian input maximizes the Gaussian channel capacity under power constraints. In the following theorem, we show that Gaussian input also minimizes $I_\nabla(X;Y)$ for Gaussian channels under power constraints. The result further indicates that Gaussian input achieves the optimum of (19), i.e. it enjoys *not only the largest information capacity but also the smallest instability*.

Theorem 4. *For Gaussian channel $Y = X + N$ where $N \sim \mathcal{N}(0, v_N)$ and $var(X) = v_X$, the p_X that achieves the minimum of $I_\nabla(X;Y)$ is Gaussian.*

$$\min_{p_X} I_\nabla(X;Y). \tag{20}$$

Moreover, the smallest mutual information is

$$I_\nabla(X;Y) = H_\nabla(Y) - H_\nabla(Y \mid X) = -\frac{1}{2(v_X+v_N)} + \frac{1}{v_N} = \frac{2v_X+v_N}{2(v_X+v_N)v_X},$$

which is increasing in v_X with range $[v_N^{-1}, 2v_N^{-1})$.

Proof: By Identity (8) in Theorem 2, we have $I_\nabla(X;Y) = H_\nabla(Y) - H_\nabla(Y \mid X)$. Using Identity 11, we have

$$H_\nabla(Y \mid X) = H_\nabla(X+N \mid X) = 2H_\nabla(N) = -\frac{1}{v_N}$$

which is a constant that does not depend on X. Thus, minimizing (20) is equivalent to minimizing $H_\nabla(Y)$ under the constraint that *Var*$(Y) \le v_N + v_X$. Using Lemma 1 concludes the proof. ■

APPLICATION: SOME ELEMENTARY INEQUALITIES

Proposition 5. *Under the same assumptions of Theorem 2, we have*

$$H_\nabla(Y_1,\ldots,Y_n) = \sum_{i=1}^n H_\nabla(Y_i \mid Y_1,\ldots,Y_{i-1}) \le \sum_{i=1}^n H_\nabla(Y_i). \tag{21}$$

Proof: The result can be obtained by recursively applying Identity (8) in Theorem 2 for n random variables $Y_1,\ldots,Y_n$. ■

We can generalize Proposition 5 to show how the monotonicity of the average entropy rates of subsets as the size of the subsets increases.

Proposition 6. *Suppose that $Y_1,\ldots,Y_n$ have a joint distribution. For every $S \subseteq \{1,\ldots,n\}$, let $Y_S = \{Y_i : i \in S\}$, and $Y_{S^c} = \{Y_i : i \notin S\}$. Let*

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:card(S)=k} \frac{H_\nabla(Y_S)}{k}, \tag{22}$$

$$t_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:card(S)=k} \frac{1}{-H_\nabla(Y_S)/k}, \tag{23}$$

Then

$$h_1^{(n)} \geq \cdots \geq h_n^{(n)}$$
$$t_1^{(n)} \geq \cdots \geq t_n^{(n)}$$

Proof: Inequalities (22) can be proved using the same arguments as in the proof of Theorem 16.8.1 in [33] (which only uses Proposition 5). We only prove the Inequality (23) here. The proof of Theorem 16.8.1 in [33] implies that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{H_\nabla(Y_{-i})}{n-1} \geq \frac{1}{n}H_\nabla(Y_1,\ldots,Y_n) \tag{24}$$

where Y_{-i} denotes $[Y_1,\ldots,Y_{i-1},Y_{i+1},Y_n]$. Thus, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\frac{-H_\nabla(Y_{-i})}{n-1}} \geq \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{-H_\nabla(Y_{-i})}{n-1}} \geq \frac{1}{-\frac{1}{n}H_\nabla(Y_1,\ldots,Y_n)} \tag{25}$$

where the first inequality of (25) follows from the fact that $H_\nabla(\cdot)$ is a negative function and the harmonic mean is no larger than the arithmetic mean, and the second inequality follows from (24). Inequality (25) is equivalent to $t_{n-1}^{(n)} \geq t_n^{(n)}$. To prove (23), we first condition on a k-element subset, and apply the existing result to obtain $t_{k-1}^{(k)} \geq t_n^{(k)}$. We then take a uniform choice over the k-element subset and its $k-1$-element subsets. ■

Consider the specific case where Y_i's are jointly distributed according to Gaussian distribution with covariance V. We can have inequalities for traces, as Gaussian gradient entropy is

$$-\frac{1}{2}\text{Tr}(V^{-1}).$$

Throughout the remainder of this chapter, we will assume that V is a positive definite symmetric $n \times n$ matrix.

Proposition 7. *For any positive definite matrix V, we have*

$$Tr(V^{-1}) \geq \sum_{i=1}^{n} Tr(V_i^{-1}). \tag{26}$$

for any set of block matrices $\{V_1,\ldots,V_n\}$ of V whose rows (resp. columns) form a partition of the rows (resp. columns) of V. Moreover, the equality holds if and only if V are block-diagonal with blocks $\{V_1,\ldots,V_n\}$.

Inequality (26) immediately follows from (21). We note that the inequality in (26) can also be proved by using block matrix inversion and Woodbury matrix identity, but it is much more involved compared with the simple proof here using entropy inequality.

Proposition 8. *If h_k denotes the product of the determinants of all the principal k-rowed minors of a positive definite $n \times n$ matrix V, i.e.,*

$$h_k = \sum_{1\leq i_1<\cdots<i_k\leq n} Tr(V(i_1,\ldots,i_k)^{-1})$$

then

$$h_1 \leq \cdots \leq \frac{h_k}{\binom{n-1}{k-1}} \leq \cdots \leq h_n$$

with equality if and only if V is a diagonal matrix.

Proof: Let $X \sim \mathcal{N}(0,V)$. Then the inequality follows directly from Proposition 6 and $k\binom{n}{k} = n\binom{n-1}{k-1}$. The proof of Inequality (22) implies that if X is Gaussian, the equality holds if and only if the entries of X are independent, i.e., V is diagonal. ■