Renormalized Mutual Information for Artificial Scientific Discovery

Leopoldo Sarra,[1,*] Andrea Aiello,[1] and Florian Marquardt[1,2]
[1]Max Planck Institute for the Science of Light, 91058 Erlangen, Germany
[2]Department of Physics, Friedrich-Alexander Universität Erlangen-Nürnberg, 91058 Erlangen, Germany

(Received 21 May 2020; revised 23 February 2021; accepted 2 April 2021; published 17 May 2021)

We derive a well-defined renormalized version of mutual information that allows us to estimate the dependence between continuous random variables in the important case when one is deterministically dependent on the other. This is the situation relevant for feature extraction, where the goal is to produce a low-dimensional effective description of a high-dimensional system. Our approach enables the discovery of collective variables in physical systems, thus adding to the toolbox of artificial scientific discovery, while also aiding the analysis of information flow in artificial neural networks.

DOI: 10.1103/PhysRevLett.126.200601

Introduction.—One of the most useful general concepts in the analysis of physical systems is the notion of collective coordinates. In many cases, ranging from statistical physics to hydrodynamics, the description of a complex many-particle system can be dramatically simplified by considering only a few collective variables like the center of mass, an order parameter, a flow field, or vortex positions. However, in new situations, it is not clear *a priori* which low-dimensional "feature" $y = f(x)$ is best suited as a compact description of the high-dimensional data x. This is the domain of unsupervised feature extraction in computer science, where large datasets like images or time series are to be analyzed [1]. Future frameworks of artificial scientific discovery [2–5] will have to rely on general approaches like this, adding to the rapidly developing toolbox of machine learning for physics [6–8].

The simplest and most known algorithm to obtain such features is the principal component analysis (PCA) [9]. The idea is to project the input into the directions of largest variance. However, its power is limited, since it can only extract linear features. A general approach to estimate the quality of a proposed feature is given by mutual information [10,11]. In general, the mutual information $I(x, y)$ answers the following question: if two random variables y and x are dependent on one another, and we are provided with the value of y, how much do we learn about x? Technically, it is defined via $I(x, y) = I(y, x) = H(y) - H(y|x)$, where $H(y|x)$ is the conditional entropy of y given x [11]. Here and in what follows, we use $H(y)$ to indicate the entropy associated to the probability density of the random variable y. Maximization of mutual information can be used to extract "optimal" features [12], as sketched in Fig. 1.

There exists, however, a well-known important problem in evaluating the mutual information for *continuous* variables with a *deterministic* dependence [13,14], which is exactly the case relevant for feature extraction. In this case, $I(x, y)$ diverges, and it is not clear how to properly cure this divergence without losing important properties of I. Specifically, reparametrization invariance turns out to be crucial: applying a bijective function to obtain $y' = g(y)$ does not change the information content, and thus $I(x, y') = I(x, y)$.

In this work, we introduce a properly *renormalized* version of mutual information for the important case of feature extraction with continuous variables

$$\tilde{I}(x, y) = H(y) - \int dx P_x(x) \ln \sqrt{\det \nabla f(x) \cdot \nabla f(x)}, \quad (1)$$

where $x \in \mathbb{R}^N$, $y = f(x) \in \mathbb{R}^K$; we use $\nabla f(x) \cdot \nabla f(x)$ as a short-hand notation for $\sum_i \partial_i f_\mu \partial_i f_\nu$, with $1 \le i \le N$ and $1 \le \mu, \nu \le K$, i.e., the $K \times K$ matrix resulting from the product of the $(K \times N)$ Jacobian matrix $\nabla f(x)$ and its transpose. The quantity $\tilde{I}$ is well defined and finite.

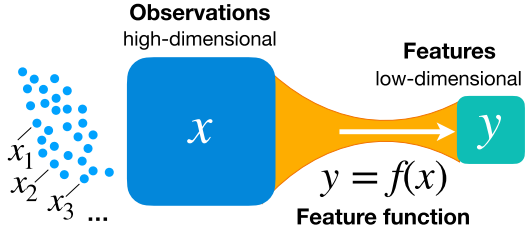

FIG. 1. Feature extraction, where a high-dimensional "microscopic" description x (such as the configuration of a many-particle system) is mapped to a low-dimensional feature $y = f(x)$. This is the case where the renormalized mutual information presented in this Letter is needed for feature optimization.

In addition, it preserves fundamental properties of mutual information—among which the invariance under reparametrization of the features:

$$\tilde{I}(x, g(y)) = \tilde{I}(x, y), \tag{2}$$

for a bijective function $g: \mathbb{R}^K \to \mathbb{R}^K$. We will derive and discuss below the meaning and usefulness of the renormalized quantity $\tilde{I}$.

Mutual information is used in many cutting edge machine learning applications, helping to improve the intermediate layers of a neural network [15,16], to increase the interpretability of generative adversarial networks [17], to analyze the behavior of neural networks during training [18,19] through the information bottleneck method [20,21], and for feature extraction via mutual information optimization [22]. It can be also used to characterize the variables in a renormalization group procedure [23]. Its practical estimation is not trivial [24], but recently derived bounds [25] permit its evaluation even in high-dimensional spaces, with the help of neural networks [26].

However, there is a problem with deterministically dependent continuous features: the conditional entropy $H(y|x)$ formally diverges as $-\log \delta(0)$ whenever y is a deterministic function of x. To understand why, it is enough to take its definition, $H(y|x) = -\int dxdy P_x(x)P(y|x)\ln P(y|x)$, and plug in $P(y|x) = \delta(y - f(x))$. This is specific to continuous variables: with discrete variables, conditional entropy would be zero and mutual information would coincide with the entropy of one of the variables. It is clear that, to deal with a deterministic continuous dependence, it is necessary to somehow redefine mutual information. Past remedies involved adding noise to the feature y or (equivalently) to simply consider the nondiverging term $H(y)$ [22,27], as briefly suggested in the InfoMax seminal paper [12]. However, they all lead to a very undesirable property: they break the fundamental reparametrization invariance of mutual information. In this scheme, any two features can be made to have the same entropy $H(y)$ simply by rescaling. Thus, in the context of feature optimization, they would be considered equally favorable, even if they represent very different information about x. The reason is that such a scheme completely ignores the diverging quantity $H(y|x)$. In contrast, we show that $H(y|x)$ contains a nontrivial finite dependence on the feature $f(x)$, which must be taken into account to obtain consistent results.

Renormalized mutual information.—In any physical system, there are small preexisting measurement uncertainties associated with extracting the microscopic observables x. Thus, loosely speaking, when trying to deduce information about x given the value of y, we have to be content with resolving x up to some spread ε. Motivated by this, we first consider a finite regularized quantity $I_\varepsilon(x, y)$. It is defined as the mutual information between the observable x and the feature function applied to a noisy version of the observable: $y = f(x + \varepsilon\lambda)$, where $\varepsilon \in \mathbb{R}$ is the noise strength and $\lambda \in \mathbb{R}^N$ is a random multidimensional Gaussian of zero mean and unit covariance matrix. In the limit $\varepsilon \to 0$ we recover the original definition of mutual information, which diverges logarithmically. Even in that limit, the nature of the adopted noise distribution (e.g., isotropy, independence of x) still matters, and corresponds to imposing some hypotheses about the observed quantities x (e.g., same measurement uncertainty in all variables). We discuss these generalizations at the end of this work.

Consider

$$P(y|x) = \int d\lambda P_\lambda(\lambda)\delta(y - f(x + \varepsilon\lambda)). \tag{3}$$

When $\varepsilon \ll 1$, we can expand $f(x+\varepsilon\lambda) \simeq f(x) + \varepsilon\lambda \cdot \nabla f(x)$. By explicit calculation, it can be easily found that $P(y|x)$ is a Gaussian distribution of zero mean and covariance matrix $\varepsilon^2 \nabla f(x) \cdot \nabla f(x) = \varepsilon^2 \sum_i \partial_i f_\mu \partial_i f_\nu$. We can calculate the conditional entropy and get

$$H(y|x) = \int dx P_x(x) \ln \sqrt{\det \nabla f(x) \cdot \nabla f(x)} + K H_\varepsilon, \tag{4}$$

where H_ε is the entropy of a one-dimensional Gaussian with variance ε^2. The first term only depends on the features, and the second only on the noise. Only this term diverges when $\varepsilon \to 0$. Therefore,

$$\tilde{I}_\varepsilon(x, y) = I_\varepsilon(x, y) + K H_\varepsilon \tag{5}$$

has a well-defined limit $\varepsilon \to 0$ and still contains all the dependence on $f(x)$. By performing the limit we obtain our main result, Eq. (1).

We can easily show that Eq. (1) is invariant under feature reparametrization. Consider an invertible function $z = g(y): \mathbb{R}^K \to \mathbb{R}^K$. We can rewrite the entropy of z as the entropy of y plus an extra term, which cancels with that obtained by differentiating $\ln \det \nabla g(f(x))$, leading to Eq. (2). We emphasize the importance of this property: after an invertible transformation on the variable y, no information should be lost, and the new variable should have the same mutual information with x as the old one. In contrast, by adding Gaussian noise η to the feature y instead of to x, i.e., $y = f(x) + \varepsilon\eta$, the final result would depend on the feature only via $H(y)$. Reparametrization invariance would not hold anymore under this alternative regularization: we have $I_\varepsilon(x, g(f(x) + \varepsilon\eta)) = I_\varepsilon(x, f(x) + \varepsilon\eta)$ but not $I_\varepsilon(x, g(f(x)) + \varepsilon\eta) = I_\varepsilon(x, f(x) + \varepsilon\eta)$ as Eq. (2) would require.

The price for a finite mutual information between two deterministically dependent variables is that when there is no dependence, e.g., $y = \text{const}$, we get $-\infty$ instead of 0. In addition, given the different roles that x and y play, renormalized mutual information is no longer symmetric in

its arguments. From a different perspective [28], Eq. (1) can be expressed as a particular kind of *information loss* [32,33].

Mutual information obeys inequalities like $I(x,(y_1,y_2)) \geq I(x,y_1)$, which translate to the regularized version I_ε. However, naively taking $\varepsilon \to 0$ results in an empty inequality $\tilde{I}(x,(y_1,y_2)) + \infty \geq \tilde{I}(x,y_1)$. By contrast, starting from $I(x,(y_1,y_2)) \geq I(x,y_1) + I(x,y_2) - I(y_1,y_2)$, we can take the same limit and obtain a useful finite result:

$$\tilde{I}(x,(y_1,y_2)) \geq \tilde{I}(x,y_1) + \tilde{I}(x,y_2) - I(y_1,y_2). \quad (6)$$

In the special case where the dimensions of y_1 and y_2 add up to the dimension of x, and the mapping $x \mapsto (y_1,y_2)$ is bijective, reparametrization invariance produces $\tilde{I}(x,(y_1,y_2)) = \tilde{I}(x,x) = H(x)$, and so

$$H(x) \geq \tilde{I}(x,y_1) + \tilde{I}(x,y_2) - I(y_1,y_2). \quad (7)$$

If one constructs y_2 to be independent of y_1, the third term on the right-hand side vanishes. However, it would be impermissible to drop $\tilde{I}(x,y_2)$, since it can have any sign.

Feature comparison.—The renormalized mutual information can be used to find out how useful any given "macroscopic" quantity [i.e., a feature $y = f(x)$] would be in characterizing the system. The result depends on the statistical distribution of x. It might be the Boltzmann distribution in equilibrium or a distribution of "snapshots" of the system configuration during some arbitrary time evolution. When control parameters such as temperature or external fields change the distribution of x, the optimal feature can change. Intuitively, observing a feature with higher $\tilde{I}$ is more effective in narrowing down the set of underlying configurations x compatible with the observed value, thus yielding more information about the system.

We show proof-of-concept examples in the most common domains of physics that deal with many degrees of freedom: fluctuating fields and many-particle systems. One important goal is to discover, without prior knowledge, that a given fluctuating field is dominated by certain localized excitations (like solitons and vortices) and to robustly estimate their properties (position, shape, velocity, etc.). The simplest example is a 1D field on a lattice with a wave packet of fixed shape at a random position [Figs. 2(a,b)] [28]. For now, we evaluate $\tilde{I}$ for a variety of handcrafted features, turning to feature optimization further below. Because of reparametrization invariance [Eq. (2)], the scaling of any of them is irrelevant, as is any bijective nonlinear transformation. For comparison, we also consider PCA [9], which in our context corresponds to a feature $f(x) = \sum_j x_j u_j$, where u is the eigenvector associated to the largest eigenvalue of the covariance matrix $\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$, and the bottleneck of a contractive autoencoder [34].

In a many-particle system (molecule, star cluster, plasma, etc.), the goal is to discover the most meaningful collective coordinates. A simple prototypical example is a liquid drop of fluctuating shape and orientation, made of atoms with known force fields [Figs. 2(c,d)].

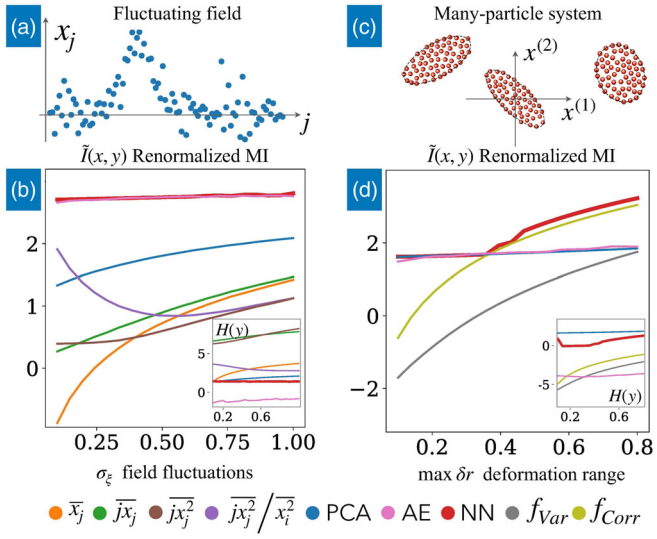

FIG. 2. Comparing renormalized mutual information $\tilde{I}$ for several features in two representative physical scenarios. (a) Fluctuating 1D field on a lattice, with a randomly placed "wave packet" (we depict one single sample). (b) $\tilde{I}$ as a function of the size of the field fluctuations σ_ξ for several features. Let $\overline{A_j} = 1/N \sum_{j=1}^{N} A_j$. We consider the average field $f(x) = \overline{x_j}$, the position j weighted by the field amplitude, $\overline{jx_j}$, or weighted by the field intensity, $\overline{jx_j^2}$, as well as the "normalized" feature $\overline{jx_j^2}/\overline{x_i^2}$ (similar to an expectation value in quantum mechanics) and the first PCA component. (c) Two-dimensional "drops" with elliptical shapes of fixed area but with fluctuating deformation amplitude δr and orientation θ (we depict three samples). (d) $\tilde{I}$ vs max. deformation spread for the 2D feature given by PCA and for two nonlinear features sensitive to shape deformations, $f_{\mathrm{Var}} = [\overline{(x_j^{(1)})^2}, \overline{(x_j^{(2)})^2}]$ and $f_{\mathrm{Corr}} = [\overline{(x_j^{(1)})^2}, \overline{x_j^{(1)} x_j^{(2)}}]$, where $x_j^{(1)}, x_j^{(2)}$ are the coordinates of particle j. In both (b) and (d) AE represents the bottleneck of a contractive autoencoder trained to reconstruct the input and NN corresponds to the feature given by a neural network optimized to maximize $\tilde{I}$. In the insets, we show the entropy $H(f(x))$. This quantity is not reparametrization invariant.

Feature optimization.—Instead of comparing different plausible features, we can consider a class of parametrized features and optimize $\tilde{I}$ over the parameters. We opted for a multilayer neural network [35], where $f(x) = f_\theta(x)$ with θ representing the parameters of the network. Intuitively, meaningful features are those that provide the largest information without overengineering. While handcrafted features, like in the previous section, are unarguably simple, the optimization of an excessively powerful feature function could lead to encode additional (nonrelevant) information by means of very nonlinear transformations. The tradeoff between the simplicity of the feature and the

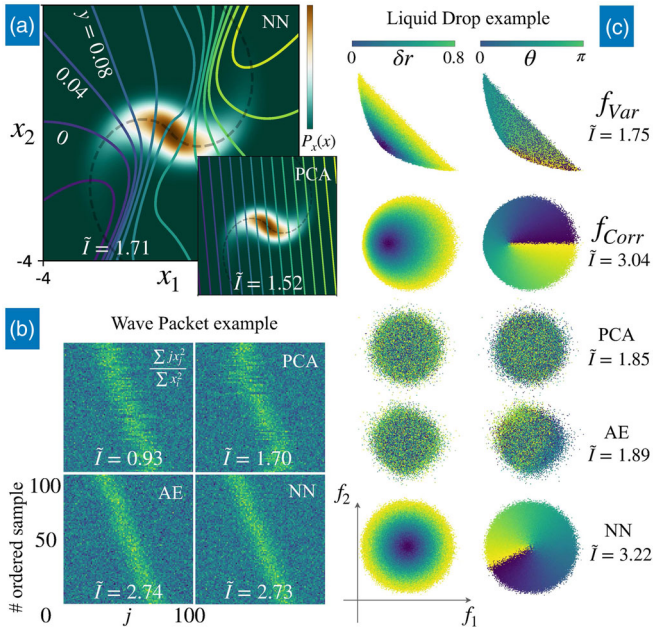

FIG. 3. Feature optimization and visual assessment of quality. (a) 2D non-Gaussian distribution. The obtained 1D feature $y = f(x_1, x_2)$, shown as contour lines atop the distribution $P_x(x)$, is parametrized with a neural network. Inset: PCA feature. (b) Wave packets as in Fig. 2(a), one by row, ordered by increasing value of the feature. The NN feature is clearly very powerful to sort the samples. (c) Liquid drops as in Fig. 2(c). We show how different 2D features map the deformation and the orientation of the drop. The NN builds up a representation very similar to our best handcrafted feature f_{Corr}.

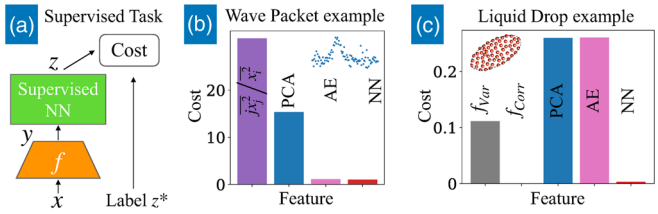

FIG. 4. Comparing the performance of a supervised regression task for different features as input. (a) For each batch of samples x we calculate the feature $y = f(x)$ and train a supervised neural network to predict the provided label z^*. (b) Predicting the center of the wave packet [example from Fig. 2(a)]. (c) Predicting the orientation and deformation of the drop [example from Fig. 2(c)]. The optimized NN feature achieves the best performance in (b) and a performance very close to that of our best handcrafted feature (c).

amount of preserved information can be adjusted both by the choice of network architecture and by adding a small additional regularization penalty (in practice, this can be achieved by punishing features with large gradients). The optimization of $\tilde{I}(x, f_\theta(x))$ can be implemented easily with gradient ascent algorithms [35]. The first term in Eq. (1) can be estimated with a histogram; for the second term, one can immediately obtain the required ∇f, since neural networks are differentiable functions, and rely on statistical sampling of x. Note that also the extra degree of freedom of feature space due to reparametrization invariance [Eq. (2)] can be exploited to enforce additional constraints [28].

In Fig. 3(a) we show the optimization of a nonlinear 1D feature for a 2D non-Gaussian distribution. Such a low-dimensional setting allows to visualize the shape of the feature and to compare it with PCA. We apply the same technique also to the physical examples [see "NN" in Figs. 2(b,d)].

One way to assess the quality of features is by suitable visualization [see Figs. 3(b,c)]. The optimized NN feature is clearly able, better than (or at least as good as) other features, to identify the relevant properties of the system. A more quantitative, well-known approach is to perform supervised training for a regression task with the feature as input and analyze the resulting performance [28]. In the physics examples shown here, one is naturally interested in predicting underlying parameters, like the wave packet location. Figures 4(b,c) illustrate superior or very good performance of the network.

In our illustrative examples we only considered 1D or to 2D features. For higher-dimensional features, the numerical estimation of Eq. (1) is more challenging, but in principle still feasible [28]—for example, through adversarial techniques [36].

Also, all the components x_j had the same physical meaning (e.g., particle coordinates). For components with different dimensions (e.g., positions and momenta), one needs to decide how to compare fluctuations along different components. A slight change in the regularization procedure is required. Most generally, we can consider the noise distribution $P(\lambda|x)$ to have an arbitrary covariance matrix $\Sigma(x)$, even allowing for a location-dependent "resolution." We find that it is necessary to replace the matrix $\nabla f(x) \cdot \nabla f(x)$ in Eq. (1) with $\nabla f(x) \Sigma(x) \nabla f(x)$, thus effectively introducing a metric on x space [28]. This changes the inequality mentioned above [Eq. (7)].

Outlook.—Renormalized mutual information can be useful in many areas of statistical analysis, machine learning, and physics.

It can be directly applied in diverse physical scenarios, with many interesting variations and extensions. In statistical physics, one expects that different phases of matter yield different optimal features. Moreover, one could optimize for feature *fields* (order parameter fields) by using convolutional layers in the neural network. The locations of defects like domain walls and vortices could be discovered as relevant features. In general, an optimized low-dimensional description of a high-dimensional system can be used to make partial predictions for the time evolution. In dynamical systems the renormalized mutual information could help to discover the underlying regularities of the system. Even in the presence of chaos, the evolution of collective variables can be predictable (and still

nontrivial) [37]. Quantum-mechanical systems could be analyzed as well, e.g., by sampling configurations x according to a many-body state, or sampling parameters in the Hamiltonian and looking at the expectation values x of a set of commuting observables in the corresponding ground state.

Renormalized mutual information can be used to analyze deterministic representations of a dataset. Here we illustrated the approach only in settings with at most two-dimensional features, but it should be feasible to efficiently evaluate $\tilde{I}$ also with high-dimensional feature spaces. This approach could be used to study the behavior of a neural network from an information-theoretic perspective, for example, by analyzing the renormalized mutual information between the input and an intermediate layer of a neural network. This could be helpful for concepts like the "information bottleneck" [20,38], which is known to be affected by the problems we discussed. Moreover, the important challenge of representation learning for high-dimensional datasets (like images) can benefit: our optimized features are purely defined by their information content and not by the capability to accomplish selected tasks. Thus, they could be useful in transfer learning scenarios, in which many classifiers are built from the same representation. We emphasize that the method advocated here should be especially useful when the dimensionality is so drastically reduced that autoencoders [34,39,40] would not plausibly work very well, since it would be impossible for a decoder to produce an approximation of the input from so few latent variables [see Fig. 3(c)]. This is precisely the situation important for collective variables and similar strongly reduced descriptions.

The code of this paper is publicly available [41].

We thank Andreas Maier for discussions.

*leopoldo.sarra@mpl.mpg.de

[1] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).
[2] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, The automation of science, Science **324**, 85 (2009).
[3] M. Schmidt and H. Lipson, Distilling free-Form natural laws from experimental data, Science **324**, 81 (2009).
[4] T. Wu and M. Tegmark, Toward an artificial intelligence physicist for unsupervised learning, Phys. Rev. E **100**, 033311 (2019).
[5] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner, Discovering Physical Concepts with Neural Networks, Phys. Rev. Lett. **124**, 010508 (2020).
[6] V. Dunjko and H. J. Briegel, Machine learning and artificial intelligence in the quantum domain, arXiv:1709.02779.
[7] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, Phys. Rep. **810**, 1 (2019).
[8] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).
[9] I. T. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. R. Soc. A **374**, 20150202 (2016).
[10] T. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, Hoboken, NJ, 2006).
[11] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. (McGraw-Hill, Boston, Massachusetts, 2009).
[12] A. J. Bell and T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. **7**, 1129 (1995).
[13] R. A. Amjad and B. C. Geiger, Learning representations for neural network-based classification using the information bottleneck principle, IEEE Trans. Pattern Anal. Mach. Intell., **42** 2225 (2019).
[14] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk, Caveats for information bottleneck in deterministic scenarios, in *International Conference on Learning Representations* (OpenReview.net, New Orleans, 2019).
[15] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, Learning deep representations by mutual information estimation and maximization, arXiv:1808.06670.
[16] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, On Mutual information maximization for representation learning, in *International Conference on Learning Representations* (OpenReview.net, Addis Ababa, 2020).
[17] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS16 (Curran Associates Inc., Red Hook, NY, USA, 2016), pp. 2180–2188.
[18] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv:1703.00810.
[19] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, On the information bottleneck theory of deep learning, J. Stat. Mech. (2019) 124020.
[20] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing* (Univ. of Illinois, Urbana-Champaign, 1999), pp. 368–377.
[21] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, Entropy and mutual information in models of deep neural networks, J. Stat. Mech. (2019) 124014.
[22] S. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ, 1999).

[23] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, Nat. Phys. **14**, 578 (2018).

[24] A. Kraskov, H. Stögbauer, and P. Grassberger, Estimating mutual information, Phys. Rev. E **69**, 066138 (2004).

[25] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, On variational bounds of mutual information, in *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, Long Beach, California, USA, 2019), pp. 5171–5180.

[26] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, Mutual information neural estimation, in *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, edited by J. Dy and A. Krause (PMLR, Stockholmsmässan, Stockholm Sweden, 2018), pp. 531–540.

[27] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*, edited by J. Taylor and C. Mannion, Perspectives in Neural Computing (Springer, New York, New York, NY, 1996).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.126.200601 for the equation of renormalized mutual information in terms of information loss, the derivation of the general case with position-dependent noise and for technical details on the examples, on the implementation of neural-network-based feature optimization and on the comparison with other techniques, which includes Refs. [29–31].

[29] M. Abadi, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard *et al.*, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015).

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*, Springer series in statistics (Springer, New York, 2001).

[31] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, edited by Y. Bengio and Y. LeCun (Yoshua Bengio, Yann LeCun, San Diego, 2015).

[32] B. C. Geiger and G. Kubin, On the information loss in memoryless systems: The multivariate case, in *Proc. Int. Zurich Seminar on Communications* (ETH Zurich, Zurich, 2011), pp. 32–35.

[33] B. C. Geiger, C. Feldbauer, and G. Kubin, Information loss in static nonlinearities, in *2011 8th International Symposium on Wireless Communication Systems* (IEEE, Aachen, Germany, 2011), pp. 799–803.

[34] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11 (OmniPress, Madison, WI, USA, 2011), pp. 833–840.

[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Curran Associates, Inc., Montreal, 2014), pp. 2672–2680.

[37] D. Forster, *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*, 1st ed. (CRC Press, Boca Raton, 2018).

[38] D. Strouse and D. J. Schwab, The deterministic information bottleneck, Neural Comput. **29**, 1611 (2017).

[39] G. Hinton, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[40] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (Association for Computing Machinery, New York, NY, USA, 2008), pp. 1096–1103.

[41] https://github.com/lsarra/rmi.