Osvaldo Simeone

Introducing Information Measures via Inference

Information measures, such as the entropy and the Kullback–Leibler (KL) divergence, are typically introduced using an abstract viewpoint based on a notion of "surprise." Accordingly, the entropy of a given random variable (rv) is larger if its realization, when revealed, is on average more "surprising" (see, e.g., [1]–[3]). The goal of this lecture note is to describe a principled and intuitive introduction to information measures that builds on inference, i.e., estimation and hypothesis testing. Specifically, entropy and conditional entropy measures are defined using variational characterizations that can be interpreted in terms of the minimum Bayes risk in an estimation problem. Divergence metrics are similarly described using variational expressions derived via mismatched estimation or binary hypothesis testing principles. The classical Shannon entropy and the KL divergence are recovered as special cases of more general families of information measures.

Relevance

Information measures are among the criteria most commonly used to derive pattern recognition and machine-learning methods, including blind source separation and variational inference. An understanding of information measures in terms of inference principles can clarify their significance and illuminate the implications of their adoption for signal processing and learning problems.

Digital Object Identifier 10.1109/MSP.2017.2766239
Date of publication: 9 January 2018

Prerequisites

This lecture note requires basic knowledge in probability and statistics.

Problem statement

We consider the following three questions.

1) Given an rv X distributed according to a known probabilistic model $p_X(x)$, i.e., $X\sim p_X$, how can we measure the information associated with its observation? Addressing this question leads to the definition of generalized entropy as the minimum average loss, or Bayes risk, attainable on the estimate of X based only on the knowledge of the probabilistic model p_X [4].
2) Given two rvs X and Y jointly distributed according to a known probabilistic model $p_{XY}(x, y)$, i.e., $(X, Y)\sim p_{XY}$, how can we measure the information associated with the observation of X when Y is already known? This leads to the definition of the generalized conditional entropy as the minimum average loss, or Bayes risk, attainable on the estimate of X given the knowledge of Y and of the probabilistic model p_{XY} [4].
3) Given two probabilistic models p_X and q_X defined over the same alphabet $\mathcal{X}$, how can we quantify how "different" they are? Tackling this question leads to the definition of divergence measures, such as the KL divergence, based on the inference problems of mismatched estimation [4] and binary hypothesis testing [5], [6].

Throughout this lecture note, we focus on the case of discrete rvs taking values in finite alphabets indicated by calligraphic letters, as in $X \in \mathcal{X}$ for an rv X. For extensions to more general alphabets, we refer to the bibliography. We will denote the probability mass function (pmf) of a discrete rv X as p_X. The conditional pmf of X given the observation $Y = y$ of a jointly distributed rv Y is indicated as $p_{X|Y=y}$, so that $p_{X|Y}$ is a random pmf indexed by Y. The notation $\mathrm{E}_{X\sim p_X}[\cdot]$ indicates the expectation of the argument with respect to the rv $X\sim p_X$, and the conditional expectation is defined in a similar way. $\mathrm{var}(\cdot)$ represents the variance of the argument pmf. The notation log represents the logarithm in base two.

Solution

Generalized entropy

As proposed by Claude Shannon, the amount of information received from the observation of a discrete rv $X\sim p_X$ defined over a finite alphabet $\mathcal{X}$ should be measured by the amount of uncertainty about its value prior to its measurement [7]. This is typically done

by introducing the "surprise" associated with the occurrence of an outcome x as $-\log p_X(x)$. According to intuition, this is an increasing function of $p_X(x)^{-1}$: the more unlikely x is, the larger is its associated surprise. The average surprise is the Shannon entropy

$$H(X) = \mathrm{E}_{X\sim p_X}[-\log p_X(X)]. \quad (1)$$

The logarithmic surprise measure $-\log p_X(x)$ can be justified based on engineering arguments as well as by using an axiomatic approach (see [3] for a review).

Taking a step back, we would like to outline a more direct approach for quantifying the information associated with the observation of an rv X. To this end, we consider the problem of estimating the value of X when one only knows the probabilistic model p_X. The key idea is that the observation of an rv X is more informative if its value is more difficult to predict a priori, that is, based only on the knowledge of p_X.

To formalize this notion, we need to specify 1) the type of estimate that one is allowed to make on the value of X; and 2) the loss function ℓ that is used to measure the accuracy of the estimate. We will proceed by considering two types of estimates: point estimates, whereby one needs to commit to a specific value $\hat{x} \in \mathcal{X}$ as the estimate of X; and distributional estimates, in which instead we are allowed to produce a pmf $\hat{p}_X$ over alphabet $\mathcal{X}$, hence defining a profile of "beliefs" over the possible values of X.

Point estimate

Given a point estimate $\hat{x} \in \mathcal{X}$ and an observed value $x \in \mathcal{X}$, the estimation error can be measured by a nonnegative loss function $\ell(x,\hat{x})$. Examples include the quadratic loss function $\ell_2(x,\hat{x}) = (x-\hat{x})^2$, and the 0–1 loss function, or detection error, $\ell_0(x,\hat{x}) = |x-\hat{x}|_0$, where $|a|_0 = 0$ if $a = 0$ and $|a|_0 = 1$ otherwise. For any given loss function ℓ, based on the aforementioned discussion, we can measure the information accrued by the observation of rv $X\sim p_X$ by evaluating the average loss that is incurred by the best possible a priori estimate of X. This leads to the definition of generalized entropy [4]

$$H_\ell(X) = H_\ell(p_X) = \min_{\hat{x}} \mathrm{E}_{X\sim p_X}[\ell(X,\hat{x})], \quad (2)$$

where the estimate $\hat{x}$ is generally not constrained to lie in the alphabet $\mathcal{X}$. As highlighted by the notation $H_\ell(p_X)$, the generalized entropy depends on the pmf p_X and on the loss function ℓ. The notion of generalized entropy (2) coincides with that of minimum Bayes risk for the given loss function ℓ.

Let us consider the examples of the quadratic and 0–1 loss functions. For the former, the generalized entropy can be computed as

$$H_{\ell_2}(p_X) = \mathrm{var}(p_X), \quad (3)$$

where we have imposed the optimality condition $\mathrm{dE}[(X-\hat{x})^2]/\mathrm{d}\hat{x} = 0$ to conclude that the optimal point estimate is the mean $\hat{x} = \mathrm{E}_{X\sim p_X}[X]$. Under the quadratic loss function, the generalized entropy is hence simply the variance of the distribution. As for the 0–1 loss, we can write

$$\begin{aligned} H_{\ell_0}(p_X) &= \min_{\hat{x}} \sum_{x\neq\hat{x}} p_X(x) \\ &= 1 - \max_{\hat{x}} p_X(\hat{x}), \end{aligned} \quad (4)$$

since the optimal estimate is the mode, i.e., the value $\hat{x}$ with the largest probability $p_X(\hat{x})$. The generalized entropy (4) equals the minimum probability of error for the detection of X.

Distributional estimate

We now consider a different type of estimation problem in which we are permitted to choose a pmf $\hat{p}_X$ on the alphabet $\mathcal{X}$ as the estimate for the outcome of variable X. To ease intuition, we can imagine $\hat{p}_X(x)$ to represent the fraction of one's wager that is invested on the outcome of X being a specific value x. Note that it may not be necessarily optimal to put all of one's money on one value x! In fact, this depends on how we measure the reward, or conversely the cost obtained when a value x is realized.

To this end, we define a nonnegative loss function $\ell(x,\hat{p}_X)$ representing the loss, or the "negative gain," suffered when the value x is observed. This loss should sensibly be a decreasing function of $\hat{p}_X(x)$—we register a smaller loss or, conversely, a larger gain, when we have wagered more on the actual outcome x. As a fairly general class of loss functions, we can hence define

$$\ell(x,\hat{p}_X) = f(\hat{p}_X(x)), \quad (5)$$

where f is a decreasing function. Note that a more general class of loss functions can be defined based on the notion of scoring rule [3].

Denote as $\Delta(\mathcal{X})$ the simplex of pmfs defined over alphabet $\mathcal{X}$. The generalized entropy can now be defined in a way that is formally equivalent to (2), with the only difference being the optimization over pmf $\hat{p}_X$ rather than over the point estimate $\hat{x}$:

$$\begin{aligned} H_\ell(X) &= H_\ell(p_X) \\ &= \min_{\hat{p}_X\in\Delta(\mathcal{X})} \mathrm{E}_{X\sim p_X}[\ell(X,\hat{p}_X)]. \end{aligned} \quad (6)$$

A key example of loss function $\ell(x, \hat{p}_X)$ in class (5) is the log-loss $\ell(x,\hat{p}_X) = -\log\hat{p}_X(x)$. The log-loss has a strong motivation in terms of lossless compression. In fact, by Kraft's inequality [1], it is possible to design a prefix-free—and hence decodable without delay—lossless compression scheme that uses $\lceil -\log\hat{p}_X(x)\rceil$ bits to represent value x. As a result, the choice of a pmf $\hat{p}_X$ is akin to the selection of a prefix-free lossless compression scheme that requires a description of around $-\log\hat{p}_X(x)$ bits to represent value x. The expectation in (6) measures the corresponding average number of bits required for lossless compression by the given scheme.

Using the log-loss in (2), we obtain

$$H(p_X) = \min_{\hat{p}_X\in\Delta(\mathcal{X})} \mathrm{E}_{X\sim p_X}[-\log\hat{p}_X(x)], \quad (7)$$

where $H(p_X)$ is the Shannon entropy (1). In fact, imposing the optimality condition on the right-hand side of (7) yields the optimal pmf $\hat{p}_X(x)$ as $\hat{p}_X(x) = p_X(x)$. Equation (7) reveals that the entropy (1) is the minimum average log-loss when optimizing over all possible pmfs $\hat{p}_X$. As a note, when the alphabet $\mathcal{X}$ has more than two elements, it can be proved that the log-loss is the only loss function of the form (5) for which $\hat{p}_X(x) = p_X(x)$ is optimal, up to multiplicative and additive constants [8, Th. 1].

Remark

When p_X is the empirical distribution of the data and the optimization over the pmf $\hat{p}_X$ is constrained to lie in a given set of parameterized pmfs, the cost function in (7) is typically referred to as the *cross-entropy* loss and the resulting problem coincides with the maximum likelihood (ML) estimation of the parameterized model $\hat{p}_X$ [2].

Remark

The generalized entropy $H_\ell(p_X)$ can be proved to be a concave function of p_X. This implies that a variable $X \sim \lambda p_X + (1-\lambda) q_X$ distributed according to the mixture of two distributions is more "random," i.e., is more difficult to estimate, than both constituent variables $X \sim p_X$ and $Y \sim q_X$.

Generalized conditional entropy and mutual information

Given two rvs X and Y jointly distributed according to a known probabilistic model $p_{XY}(x,y)$, i.e., $(X,Y) \sim p_{XY}$, we now discuss how to quantify the information that the observation of one variable, say Y, brings about the other, i.e., X. Following the same approach adopted previously, we can distinguish two inferential scenarios for this purpose: in the first, a point estimate $\hat{x}(y)$ of X needs to be produced based on the observation of a value $Y = y$ and the knowledge of the joint pmf p_{XY}; while, in the second, we are allowed to choose a pmf $\hat{p}_{X|Y=y}$ as the estimate of X given the observation $Y = y$.

Point estimate

Assuming point estimates and given a loss function $\ell(x,\hat{x})$, the generalized conditional entropy for an observation $Y = y$ is defined as the minimum average loss as shown in (8) in the box at the bottom of the page.

Note that this definition is consistent with (2) as applied to the conditional pmf $p_{X|Y=y}$. Averaging over the distribution of the observation Y yields the generalized conditional entropy

$$H_\ell(X|Y) = \mathrm{E}_{Y\sim p_Y}[H_\ell(p_{X|Y})]. \quad (9)$$

It is emphasized that the generalized conditional entropy depends on the joint distribution p_{XY}, while (8) depends only on the conditional pmf $p_{X|Y=y}$.

For the squared error, the generalized conditional entropy can be easily seen to be the average conditional variance $H_{\ell_2}(X|Y) = \mathrm{E}_{Y\sim p_Y}[\mathrm{var}(p_{X|Y})]$, since the a posteriori mean $\hat{x}(y) = \mathrm{E}_{X\sim p_{X|Y=y}}[X|Y=y]$ is the optimal estimate. For the 0–1 loss, the generalized conditional entropy $H_{\ell_0}(X|Y)$ is instead equal to the minimum probability of error for the detection of X given Y, and the maximum a posteriori estimate $\hat{x}(y) = \mathrm{argmax}_{\hat{x}\in\mathcal{X}} p_{X|Y}(\hat{x}|y)$ is optimal.

Distributional estimate

Assume now that we are allowed to choose a pmf $\hat{p}_{X|Y=y}$ as the estimate of X given the observation $Y = y$, and that we measure the estimation loss via a function $\ell(x,\hat{p}_X)$ as in (5). The definition of generalized conditional entropy for a given value of $Y = y$ follows directly from the aforementioned arguments and is given as $H_\ell(p_{X|Y=y})$, while the generalized conditional entropy is (9). With the log-loss function, generalized conditional entropy can be again seen to coincide with Shannon conditional entropy $H(X|Y) = \mathrm{E}_{X,Y\sim p_{X,Y}}[-\log p_{X|Y}(X|Y)]$.

Remark

If X and Y are independent, we have the equality $H_\ell(X|Y) = H_\ell(X)$. Furthermore, since in (8) we can always choose estimates that are independent of Y, we generally have the inequality $H_\ell(X|Y) \le H_\ell(X)$: observing Y, on average, can only decrease the entropy. Note, however, that it is not true that $H_\ell(p_{X|Y=y})$ is necessarily smaller than $H_\ell(X)$ [1, Ch. 2].

Remark

Assume that p_{XY} is the empirical distribution of the data, typically partitioned into as domain variables X and labels Y, and that the optimization over the conditional pmf $\hat{p}_{X|Y}$ is constrained to lie in a given set of parameterized pmfs. In this case, the cost function $\mathrm{E}_{X,Y\sim p_{XY}}[-\log \hat{p}_{X|Y}(X|Y)]$ is again defined as the *cross-entropy* loss, and the resulting problem coincides with the ML supervised learning of the parameterized model $\hat{p}_{X|Y}$, as in, e.g., logistic regression [2].

Mutual information

The inequality $H_\ell(X|Y) \le H_\ell(X)$ justifies the definition of generalized mutual information with respect to the given loss function ℓ as

$$I_\ell(X;Y) = H_\ell(X) - H_\ell(X|Y). \quad (10)$$

The mutual information measures the decrease in average loss that is obtained by observing Y as compared to having only prior information about p_X. This notion of mutual information is in line with the concept of statistical information proposed by DeGroot [10]. With the log-loss, the generalized mutual information (10) reduces to Shannon's mutual information.

Divergence measures

Here we discuss how to quantify the "difference" between two given probabilistic models p_X and q_X defined over the same alphabet $\mathcal{X}$. We will take two different inferential viewpoints that will lead to different definitions of divergence between two distributions. The first is based on mismatched inference and naturally follows the approach used previously to define generalized entropy, conditional entropy, and mutual information. In contrast, the second is based on the conceptually distinct inferential scenario of binary hypothesis testing.

Mismatched inference

Assume that the correct probabilistic model p_X, from which the observation $X \sim p_X$ is drawn, is not known, but only an approximation q_X is available. The point estimate $\hat{x}$ can hence depend only on q_X, and is selected by minimizing the mismatched average loss as

$$\hat{x}^{(q_X)} = \mathrm{argmin}_{\hat{x}} \mathrm{E}_{X\sim q_X}[\ell(X,\hat{x})]. \quad (11)$$

In a similar manner, for the distributional estimate $\hat{p}_X$, we have the mismatched

$$H_\ell(p_{X|Y=y}) = \min_{\hat{x}(y)} \mathrm{E}_{X\sim p_{X|Y=y}}[\ell(X,\hat{x}(y))|Y=y]. \quad (8)$$

estimate $\hat{p}_X^{(q_X)} = \text{argmin}_{\hat{p}_X \in \Delta(\mathcal{X})} \mathrm{E}_{X\sim q_X}[\ell(X,\hat{p}_X)]$. The difference between the average loss obtained with the mismatched estimate and the minimum loss $H_\ell(X)$ can be adopted as a measure of the divergence between the two distributions.

For a given loss function ℓ, this approach yields the following definition of divergence between two distributions

$$D_\ell(p_X \| q_X) = \mathrm{E}_{X\sim p_X}[\ell(X,\hat{x}^{(q_X)})] - H_\ell(p_X) \quad (12)$$

in the case of point estimates, and

$$D_\ell(p_X \| q_X) = \mathrm{E}_{X\sim p_X}[\ell(X,\hat{p}_X^{(q_X)})] - H_\ell(p_X) \quad (13)$$

for distributional inference. It is noted that the divergence $D_\ell(p_X \| q_X)$ equals zero if and only if the mismatched estimate performs as well as the optimal estimate in terms of average loss.

For the quadratic loss, the divergence is given as $D_{\ell_2}(p_X \| q_X) = (\mathrm{E}_{X\sim p_X}[X] - \mathrm{E}_{X\sim q_X}[X])^2$, which measures the difference in the means of the two pmfs. In the special case of log-loss, the definition (12) coincides with the conventional KL divergence

$$D(p_X \| q_X) = \mathrm{E}_{X\sim p_X}\left[\log \frac{p_X(X)}{q_X(X)}\right]. \quad (14)$$

By comparing (12) and (13) with the definition of mutual information (10), it can be seen that the following general relationship holds between the generalized mutual information and the divergence (12), (13)

$$I_\ell(X;Y) = \mathrm{E}_{Y\sim p_Y}[D_\ell(p_{X|Y} \| p_X)]. \quad (15)$$

Hence, the generalized mutual information measures the average divergence between the conditional pmf $p_{X|Y=y}$ and the marginal pmf p_X.

Binary hypothesis testing

We now consider the different inferential set-up of binary hypothesis testing: given an observation X, decide whether X was generated from pmf p_X or from pmf q_X. To proceed, we define a decision rule $T(x)$, which should increase with the confidence that a value x is generated from p_X rather than q_X. In this way, in practice, one may impose a threshold on the rule $T(x)$ so that, for $T(x)$ larger than the threshold, a decision is made that X was generated from p_X.

To design the decision rule $T(x)$, we again minimize a loss function or, equivalently, maximize a merit function. For convenience, here we take the latter approach, and define the problem of maximizing the merit function

$$\mathrm{E}_{X\sim p_X}[T(X)] - \mathrm{E}_{X\sim q_X}[g(T(X))] \quad (16)$$

over the rule $T(x)$, where g is a convex increasing function. This criterion can be motivated as follows: 1) the expression (16) increases if $T(x)$ is large, on average, for values of X generated from p_X; and 2) it decreases if, upon expectation, $T(x)$ is large for values of X generated from q_X. The function g can be used to define the relative importance of errors made in favor of one distribution or the other. We note that the merit function (16) can also be formally related to the error probability of binary hypothesis testing [11].

From this discussion, the optimal value of (16) can be taken to be a measure of the distance between the two pmfs. This yields the following definition of divergence between two pmfs:

$$D_f(p_X \| q_X) = \max_{T(x)} \mathrm{E}_{X\sim p_X}[T(X)] - \mathrm{E}_{X\sim q_X}[g(T(X))], \quad (17)$$

where the subscript f will be justified next.

Under suitable differentiability assumptions on function g (see [6] for generalizations), taking the derivative with respect to $T(x)$ for all $x \in \mathcal{X}$ yields the optimality condition $g'(T(x)) = p_X(x)/q_X(x)$. This relationship reveals the connection between the optimal detector $T(x)$ and the likelihood ratio $p_X(x)/q_X(x)$. Plugging this result into (17), it can be directly checked that the following equality holds [5]:

$$D_f(p_X \| q_X) = \mathrm{E}_{X\sim q_X}\left[f\left(\frac{p_X(X)}{q_X(X)}\right)\right], \quad (18)$$

where the function $f(x) = g^*(x)$ is the convex conjugate of $g(t)$, which is defined as $g^*(x) = \sup_t(xt - g(t))$. Note that the convex conjugate is a convex function.

Under the additional constraint $f(1) = 0$, definition (18) describes a large class of divergence measures parameterized by the convex function f, which are known as *f-divergences* or *Ali–Silvey distance measures* [9]. The constraint $f(1) = 0$ ensures that the divergence is zero when the pmfs p_X and q_X are identical. Among their key properties, f-divergences satisfy the data processing inequality [1], [9].

As a specific example, the choice $g(t) = \exp(t-1)$, which gives the convex conjugate $f(x) = x\log x$, yields the optimal detector $T(x) = 1 + \log(p_X(x)/q_X(x))$ and the corresponding divergence measure (18) is the standard KL divergence $\mathrm{KL}(p_X \| q_X)$ in (14). Another instance of f-divergence, obtained with $g(t) = -\log(2 - \exp(t))$ and the optimal detector $T(x) = \log(2p_X(x)/p_X(x) + q_X(x))$, is the Jensen–Shannon divergence. Further examples include the class of α-divergences [6], [9].

We finally mention the related divergence class of integral probability metrics, which measure the difference $\mathrm{E}_{X\sim p_X}[f(X)] - \mathrm{E}_{X\sim q_X}[f(X)]$ upon maximization over all functions f within a given class. This leads, among other metrics, to the maximum mean discrepancy measure, and to the Wasserstein (or Earth mover) divergence based on optimal transport theory [12].

Remark

When p_X is the empirical distribution of the data, q_X is the (explicit or implicit) distribution of a model to be learned and $T(x)$ is a parametric detector, problem (17) is a key step of generative adversarial networks [6].

Conclusions

In this lecture note, we have presented an introduction of information measures in terms of inferential problems—estimation for entropy and conditional entropy, and mismatched estimation and binary hypothesis testing for divergence metrics. This approach allows the definition of general classes of information measures including, as special cases, Shannon's entropy and KL divergence, in an intuitive way that reveals their operational significance. The variational formulations that define the

information measures as optimal inference problems can be used to derive learning algorithms, such as in [6], as well as estimates of information measures [5], [11].

Author

Osvaldo Simeone (osvaldo.simeone@kcl.ac.uk) received his M.Sc. degree (with honors) and his Ph.D. degree, both in information engineering, from Politecnico di Milano, Italy, in 2001 and 2005, respectively. He is a professor of information engineering with the Centre for Telecommunications Research in the Department of Informatics of King's College London. From 2006 to 2017, he was a faculty member with the Electrical and Computer Engineering Department at the New Jersey Institute of Technology. His research interests include wireless communications, information theory, optimization, and machine learning. He is a corecipient of the 2017 JCN Best Paper Award, the 2015 IEEE Communication Society Best Tutorial Paper Award, and the Best Paper Awards of IEEE SPAWC 2007 and IEEE WRECOM 2007. He was awarded a European Research Council Consolidator Grant in 2016. He is a Fellow of the IEEE.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 2012.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[3] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, Sept. 2008.

[4] P. Grünwald and P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Ann. Statist.*, vol. 32, no. 4, pp. 1367–1433, 2004.

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

[6] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Neural Information Processing Systems Conf.*, Barcelona, Spain, 2016, pp. 4240–4248.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, July 1948.

[8] J. Jiao, T. A. Courtade, A. No, K. Venkat, and T. Weissman, "Information measures: The curious case of the binary alphabet," *IEEE Trans. Inform. Theory*, vol. 60, no. 12, pp. 7616–7626, Dec. 2014.

[9] J. C. Duchi, *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford, CA.

[10] M. H. DeGroot, "Changes in utility as information," *Theory Decis.*, vol. 17, no. 3, pp. 287–303, Nov. 1994.

[11] V. Berisha, A. Wisler, A. Hero, and A. Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Trans. Signal Proc.*, vol. 64, no. 3, pp. 580–591, Feb. 2016.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv Preprint, arXiv:1701.07875, Jan. 2017.

SP

FROM THE EDITOR

(continued from page 4)

objective of creating content that can be used to promote the wealth of signal processing applications in lower-level undergraduate courses. I believe such articles would provide additional motivation for students and professors in signal processing. Content from industry will enhance the offerings of *SPM*, making it more valuable to the readership and enhancing the underlying industrial support of contributors.

Standards and commercialization are important in many areas of signal processing, which are not covered extensively in *SPM*. For example, communication standards are driving the development of fifth-generation signal processing algorithms while the commercialization of virtual reality by several firms is encouraging further development in low-power multimedia signal processing for consumers. I envision more columns or articles where the authors provide a top-level view of the standards development including what is being standardized, the time line, how the meetings are conducted, and where information is shared. This will be of great value for anyone not attending the meetings. In terms of commercialization, I foresee contributions related to emerging consumer products (explaining the signal processing connection, not just blanket marketing) or highlights of different technology start-ups. Content in these new areas will enhance the appeal of the magazine.

Signal processing can be funny. Many graduate students remember Ph.D. comics or the cartoons in signal processing books about the terrors on convolution. I will work to develop some lighthearted content in *SPM*, composed of contributions from the readership. These could take the form of short articles or cartoons, including plays on words, the trials of graduate school, or even fun puzzles. Such contributions will provide a hook that will keep readers looking through every issue and will help create a fun culture that brings people together.

SPM is a magazine for all of us. I look forward to your feedback and ideas.

SP