DOI:
10.1109/MSP.2017.2766239

Document Version
Peer reviewed version

Introducing Information Measures via Inference [Lecture Notes]

Osvaldo Simeone

Information measures, such as the entropy and the Kullback-Leibler (KL) divergence, are typically introduced in Information Theory, Pattern Recognition and Machine Learning books using an abstract viewpoint based on a notion of "surprise": the entropy of a given random variable is larger if its realization, when revealed, is on average more "surprising" (see, e.g., [1], [2], [3]). The goal of these lecture notes is to describe a principled and intuitive introduction to information measures that builds on inference, namely 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.

Prerequisites

These notes require basic knowledge in probability and statistics.

PROBLEM STATEMENT

We consider the following three questions.

1. Given a random variable (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 random variables 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 two different inference problems, namely mismatched estimation [4] and binary hypothesis testing [5], [6].

Throughout these notes, we focus on the case of discrete rvs taking values in finite alphabets indicated by calligraphic letters, as in $X \in \mathcal{X}$ for a rv X. For extensions to more general alphabets, we refer to the bibliography. We will denote to 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

1. 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 a 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 $-\log p_X(x)$, which is indeed an increasing function of $p_X(x)^{-1}$: the more unlikely x is, the larger is its induced surprise. The average surprise is the Shannon entropy

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

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 a random variable 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 a 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: (*i*) the type of estimates that one is allowed to make on the value of X; and (*ii*) the loss function ℓ that is used to measure the accuracy of the estimate. We will proceed by considering two types of estimates, namely *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 Estimates: 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 non-negative 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 discussion above, we can measure the information accrued by the observation of $X \sim p_X$ 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]

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

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

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

where we have imposed the optimality condition $\mathrm{d}\mathrm{E}[(X - \hat{x})^2]/\mathrm{d}\hat{x} = 0$ 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

$$H_{\ell_0}(p_X) = \min_{\hat{x}} \sum_{x \neq \hat{x}} p_X(x) = 1 - \max_{\hat{x}} p_X(\hat{x}), \tag{4}$$

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 non-negative 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

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

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 $\hat{x}$:

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

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

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

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, Theorem 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 parametrized 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 parametrized 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., it is more difficult to estimate, than both variables $X \sim p_X$ and $Y \sim q_X$.

2. 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, namely X. Following the same approach adopted above, 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

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

Note that this definition is consistent with (8) as applied to the conditional pmf $p_{X|Y=y}$. Averaging over the distribution of the observation Y yields the generalized conditional entropy

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

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 (MAP) 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 arguments above and is given as $H_\ell(p_{X|Y=y})$, while the generalized conditional entropy is (9). With the log-loss function, the definition above can be again seen to coincide with Shannon conditional entropy $H(X|Y) = \mathrm{E}_{X,Y\sim p_{X,Y}}[-\log p_{X|Y}(X)]$.

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 $H_\ell(X|Y) \leq H_\ell(X)$: 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, Chapter 2].

Remark: Assume that $p_{X,Y}$ 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 parametrized pmfs. In this case, the cost function $\mathrm{E}_{X,Y\sim p_{X,Y}}[-\log \hat{p}_{X|Y}(X)]$ is again defined as the *cross-entropy* loss, and the resulting problem coincides with the ML supervised learning of the parametrized model $\hat{p}_{X|Y}$, as in, e.g., logistic regression [2].

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

$$I_\ell(X;Y) = H_\ell(X) - H_\ell(X|Y). \tag{10}$$

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.

3. 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 follows naturally the approach used above to define generalized entropy, conditional entropy and mutual information; while 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

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

In a similar manner, for the distributional estimate $\hat{p}_X$, we have the mismatched estimate $\hat{p}_X^{(q_X)} = \operatorname{argmin}_{\hat{p}_X\in\Delta(\mathcal{X})} \mathrm{E}_{X\sim q_X}[\ell(X,\hat{p}_X)]$. 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

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

in the case of point estimates, and

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

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

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

By comparing (12)-(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)

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

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.

In order 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

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

over the rule $T(x)$, where g is a convex increasing function. This criterion can be motivated as follows: (*i*) It increases if $T(x)$ is large, on average, for values of X generated from p_X; and (*ii*) 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

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

where the subscript f will be justified below.

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]

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

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 convex conjugate is convex.

Under the additional constraint $f(1) = 0$, definition (18) describes a large class of divergence measures parametrized 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 (MMD) measure and 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 empirical distribution obtained from a model to be learned and $T(x)$ is a parametric detector, problem (17) is a key step of Generative Adversarial Networks (GANs) [6].

Conclusions

In these lecture notes, we have presented an introduction of information measures in terms of inferential problems, namely estimation for entropy and conditional entropy, as well as 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 [11], [5].

REFERENCES

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

[2] C. M. Bishop, *Pattern recognition and machine learning,* Springer, 2006.

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

[4] P. Grünwald and P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *The Annals of Statistics*, 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 Conference*, pp. 4240-4248, Barcelona, Spain, Dec. 2016.

[7] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, Jul. 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 University.

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

[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.

AUTHOR

Osvaldo Simeone is a Professor of Information Engineering with the Centre for Telecommunications Research at the Department of Informatics of King's College London. He received an M.Sc. degree (with honors) and a Ph.D. degree in information engineering from Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. He was previously a Professor at the New Jersey Institute of Technology (NJIT). Dr Simeone is a co-recipient of the 2017 JCN Best Paper Award, the 2015 IEEE Communication Society Best Tutorial Paper Award and of the Best Paper Awards of IEEE SPAWC 2007 and IEEE WRECOM 2007. He was awarded a Consolidator grant by the European Research Council (ERC) in 2016. He is a Fellow of the IEEE.