Free Energy Minimization: A Unified Framework for Modelling, Inference, Learning, and Optimization [Lecture Notes]

Sharu Theresa Jose and Osvaldo Simeone

The goal of these lecture notes is to review the problem of free energy minimization as a unified framework underlying the definition of maximum entropy modelling, generalized Bayesian inference, learning with latent variables, statistical learning analysis of generalization, and local optimization. Free energy minimization is first introduced, here and historically, as a thermodynamic principle. Then, it is described mathematically in the context of Fenchel duality. Finally, the mentioned applications to modelling, inference, learning, and optimization are covered starting from basic principles.

Relevance

Free energy minimization is often invoked, implicitly or explicitly, in different domains by taking an ad hoc approach that hides the generality of the formulation and the common structure of the problem. Mathematical details, e.g., in terms of measure theory, also often make some of the material not easily accessible. These notes are intended to provide an accessible reference for researchers interested in connecting the dots among various standard problems in modelling, inference, learning, and optimization within a common mathematical framework.

Prerequisites

These notes require basic knowledge in probability and statistics.

PROBLEM STATEMENT

The maximum entropy modelling principle, generalized Bayesian inference, maximum likelihood learning with latent variables, Probably Approximately Correct (PAC) Bayes theory, mirror descent optimization, as well as recent theories of human behavior, have all in common their underlying reliance on an optimization principle first enunciated at the end of the 19th century –

the minimization of the free energy. Denoting as $E_{x\sim p(x)}[\cdot]$ the expectation operator for a random variable $x \sim p(x)$, the problem can be mathematically formulated as

$$\underset{q}{\text{minimize }} J(q) = \underbrace{E_{x\sim q(x)}[L(x)]}_{\text{average loss ("energy")}} + T \times \underbrace{D(q)}_{\text{information "complexity"}}, \tag{1}$$

where:

• $q(\cdot)$, also denoted as q, is the probability distribution under optimization;

• $L(\cdot)$ is a loss function, which is also known as "energy" (or Hamiltonian in statistical physics);

• $T \geq 0$ is a "temperature" parameter; and

• $D(q)$ is a convex penalty term, measuring the information-theoretic "complexity" of distribution $q(\cdot)$.

The free energy $J(q)$ reflects a trade-off between performance, as measured by loss function $L(\cdot)$, and complexity, with the balance between the two terms being dictated by the temperature level T. The complexity penalty $D(q)$ may take different forms, depending on the application. Notable examples are the Kullback-Leibler divergence $D(q) = \text{KL}(q||p) = E_{x\sim q(x)}[\log(p(x)/q(x))]$ with respect to a reference distribution p, which represents a state of minimal complexity; and the negative entropy $D(q) = -\text{H}(q) = E_{x\sim q(x)}[\log q(x)]$. Informally, the information complexity $D(q)$ can be taken to measure the amount of knowledge, or information processing, needed to generate a sample $x \sim q(x)$. For instance, with $D(q) = -\text{H}(q)$, a small $D(q)$, and hence a large entropy, amounts to a "random" pick, while a concentrated distribution with a large $D(q)$ describes a more deliberate, and hence complex, choice.

The rest of these notes will tackle the following questions:

• *What is the historical origin of problem (1)?* (The answer will briefly take us back to the 19th century.)

• *What is the general form of the solution q_{opt} to problem (1) and its corresponding optimal value $J_{\text{opt}} = J(q_{\text{opt}})$?* (The answer will involve Fenchel duality.)

• *What are its applications to modelling, inference, learning, and optimization?* (The answer will cover the maximum entropy principle, generalized Bayesian inference, variational learning, PAC Bayes theory, and mirror descent.)

As some remark on notation, the ith element of a vector a is denoted as a_i or $[a]_i$. Given finite dimensional vectors $a = [a_i]$ and $b = [b_i]$ with the same number of elements, we use $a^T b = \sum_i a_i b_i$ to denote the dot product, with superscript T indicating the transpose operator. For a vector a, $||a||_2 = \sqrt{a^T a} = \sqrt{\sum_i a_i^2}$ is the l_2 norm, and $||a||_1 = \sum_i |a_i|$ is the l_1 norm.

We write $c^+ = \max\{0, c\}$ for a scalar c. $1(\mathcal{E})$ denotes the indicator function that equals 1 when event $\mathcal{E}$ is true and zero otherwise. Finally, we denote random variables with Roman fonts and corresponding realizations with standard fonts.

I. ORIGIN

The concept of free energy goes back to the roots of thermodynamics, where it was introduced by von Helmholtz and Gibbs in the 1870s and 1880s to measure the maximum amount of work that can be extracted from a closed thermodynamic system at a constant temperature. Originally called "available energy", it captures the fact that only part of the energy of a system can be used for work. If, for example, all the molecules of a gas in a box move to the left, the corresponding kinetic energy can be used to drive a turbine. If, instead, the same kinetic energy is distributed as random molecular motion, it cannot be fully transformed into work. The maximum usable part of the energy is the (Helmholtz) free energy $F = E - TS$, where E is the overall (kinetic) energy of the system; T represents the system's temperature; and S is the thermodynamic entropy. The entropy measures the level of "complexity" of the system in terms of the number of micro-states, e.g., the number of molecules' positions and momenta, that produce the given observed macro-state.

II. MINIMIZING THE FREE ENERGY

In this section, we will see that problem (1) amounts to the computation of the Fenchel dual of the convex function $D(q)$, and that, as a consequence, solutions can be in principle easily defined – if not computed – for a variety of information complexity measures $D(q)$.

To start, we recall that, given a function $f(u)$, with u being a finite-dimensional real vector, its convex, or Fenchel, dual function $f^*(v)$ is given as the solution of the problem

$$f^*(v) = \max_u u^T v - f(u), \tag{2}$$

where vector v has the same dimension as u. For a convex function $f(u)$, this can be interpreted as evaluating the intercept $-f^*(v)$ for the tangent of slope v of function $f(u)$ [1]. From now on, we will assume that function $f(u)$ is convex, closed[1], and differentiable. From standard results

---

[1] $f(u)$ is closed if its epigraph, i.e., the set $\{(u, y) : f(u) \le y\}$, is a closed set.

Fenchel duality	free energy minimization
u	$q \doteq [q(1), ..., q(N_x)]^T$
v	$-L/T \doteq -[L(1), ..., L(N_x)]^T/T$
$f(u)$	$D(q) + I_S(q)$
$f^*(v)$	$-J_{\text{opt}}/T$

TABLE I: Mapping between free energy minimization (1) and Fenchel duality (2).

$D(q)$	minimum free energy J_{opt}	optimal solution $q_{\text{opt}}(x)$
$-\text{H}(q)$	$-\log \sum_x \exp(-L(x)/T)$	$\exp(-L(x)/T)/\sum_x \exp(-L(x)/T)$
$\text{KL}(q\|\|p)$	$-\log \text{E}_{\text{x}\sim p(x)}[\exp(-L(\text{x})/T)]$	$p(x)\exp(-L(x)/T)/\text{E}_{\text{x}\sim p(x)}[\exp(-L(\text{x})/T)]$
$\frac{1}{2}\|\|q-p\|\|^2$	$\min_q\|\|q-(p-L/T)\|\|^2$	$(p(x)-L(x)/T-\tau)^+$

TABLE II: Optimal solution of the free energy minimization problem (1) for different choices of the information complexity $D(q)$. Parameter τ satisfies $\sum_x(p(x) - L(x)/T - \tau)^+ = 1$.

in Fenchel duality, the optimal solution $u_{\text{opt}}(v)$ to problem (2) for any fixed vector v is related to the dual function $f^*(v)$ by the relationship

$$u_{\text{opt}}(v) = \nabla f^*(v). \tag{3}$$

Furthermore, definition (2) implies the Fenchel-Young inequality

$$f^*(v) \geq u^T v - f(u), \tag{4}$$

which holds for any vectors u and v, with equality attained at $u = u_{\text{opt}}(v)$.

How is the free energy minimization problem (1) related to the computation (2) of the Fenchel dual? To elaborate, assume that the alphabet $\mathcal{X}$ of random variable x in (1) is discrete and finite, i.e., $\mathcal{X} = \{1, ..., N_x\}$. We can then write the negative free energy scaled by the temperature T as

$$-\frac{J(q)}{T} = -\text{E}_{\text{x}\sim q(x)}\left[\frac{L(\text{x})}{T}\right] - D(q) = q^T\left(-\frac{L}{T}\right) - D(q), \tag{5}$$

where vectors q and L are defined in Table I. Therefore, with the correspondence detailed in Table I, the maximization of (5), and hence the minimization (1), is an instance of the Fenchel duality problem (2). Note that in Table I the notation $I_S(q) = 1(q \geq 0 : \|\|q\|\|_1 = 1)$ represents the support function for the set of all possible distributions q, i.e., the probability simplex. From (3) and known Fenchel duality relationships, we obtain the solution of problem (1) listed in Table II for the most common information complexity measures $D(q)$.
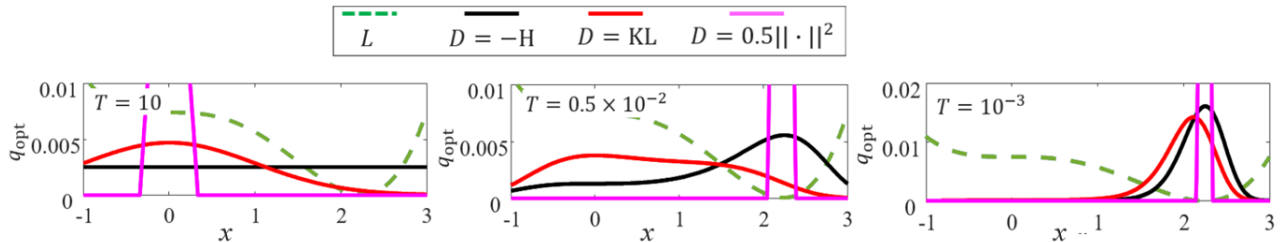
Fig. 1: Loss function $L(x)$ and optimal solution q_{opt} of problem (1) under the three complexity measures $D(q) = -\text{H}(q)$, $\text{KL}(q||p)$ and $0.5||q - p||^2$ with prior distribution $p(x) = \mathcal{N}(x; 0, 1)$ as a function of x, for various values of the temperature T.

As an example, Figure 1 illustrates the trade-off between accuracy and complexity that results from free energy minimization when varying the temperature T under the three complexity measures indicated in Table II with prior $p(x) = \mathcal{N}(x; 0, 1)$ being the zero-mean unit-variance Gaussian distribution. For $T = 10$, the optimizing distribution q_{opt} puts more weight on minimizing the complexity term $D(q)$. As a result, for $D(q) = -\text{H}(q)$, the optimal distribution q_{opt} is close to a uniform distribution, whereas it approximates the prior p for $D(q) = \text{KL}(q||p)$ and $0.5||q - p||^2$. For smaller temperature values T, the emphasis shifts on minimizing $L(x)$ and the optimal distribution q_{opt} increasingly concentrates on the minimizer of the loss function.

III. APPLICATIONS

1. *Maximum Entropy Principle*

The maximum entropy principle provides a standard approach for the selection of a probabilistic model when the only information available on the quantity of interest consists of the averages of given statistics. Formally, suppose that we are given the expected values α_k of functions $f_k(x)$ for $k = 1, \ldots, m$ of a random observation x with unknown distribution. The maximum entropy principle selects the probability distribution q that solves the following optimization problem [2]

$$\max_{q} \quad \text{H}(q)$$

$$\text{s.t.} \quad \text{E}_{x \sim q(x)}[f_k(x)] = \alpha_k \quad \text{for } k = 1, \ldots, m. \tag{6}$$

Introducing a Lagrange multiplier λ_k for each of the average constraints in (6), the optimization of the Lagrangian for problem (6) can be written as the free energy minimization (with $T = 1$)

$$\min_q J(q) = \mathrm{E}_{\mathrm{x} \sim q(x)}[\underbrace{-\sum_k \lambda_k(f_k(\mathrm{x}))}_{L(\mathrm{x})}] + \underbrace{(-\mathrm{H}(q))}_{D(q)}. \tag{7}$$

The solution of problem (6) then directly follows from Table II as

$$q_{\mathrm{opt}}(x) \propto \exp\left(\sum_k \lambda_k f_k(x)\right). \tag{8}$$

The optimizing distribution in (8) belongs to the exponential family with natural parameter vector $\lambda = (\lambda_1, \dots, \lambda_m)$ and sufficient statistics vector $f = (f_1, \dots, f_m)$ [3].

2. *Generalized Bayesian Inference*

Consider a probabilistic model given by a joint distribution $p(x, y)$. Bayesian inference of x given an observation y $= y$ amounts to the computation of the posterior distribution $p(x|y) = p(x, y)/p(y)$. This requires the evaluation of the partition function, or marginal likelihood, $p(y) = \sum_x p(x, y)$ of the observation. More generally, given an unnormalized distribution $\tilde{p}(x)$ – e.g., $p(x, y)$ with a fixed value of y – Bayesian inference requires the computation of the normalizing partition function $Z = \sum_x \tilde{p}(x)$.

Define as the loss function $L(x) = -\log \tilde{p}(x)$ and as the information complexity penalty $D(q) = -\mathrm{H}(q)$. With this choice, by Table II, the minimum free energy with $T = 1$ equals the negative log-partition function, i.e., $J_{\mathrm{opt}} = -\log Z = -\log \sum_x \tilde{p}(x)$ and the optimal distribution $q_{\mathrm{opt}}(x)$ equals the desired normalized distribution $p(x) = \tilde{p}(x)/Z$. Furthermore, by Table I and the Fenchel-Young inequality (4), we obtain the following inequality, also known as *Evidence Lower BOund (ELBO)* [3]

$$\log Z \geq -\mathrm{E}_{\mathrm{x} \sim q(x)}[\underbrace{-\log \tilde{p}(\mathrm{x})}_{L(\mathrm{x})}] + \underbrace{\mathrm{E}_{\mathrm{x} \sim q(x)}[-\log q(\mathrm{x})]}_{\mathrm{H}(q)} = -J(q). \tag{9}$$

To sum up, we have shown that: (*i*) the log-partition function is lower bounded by the negative free energy with the log-loss function $L(x) = -\log \tilde{p}(\mathrm{x})$ and an entropy penalty; and (*ii*) the maximum of the lower bound is achieved for distribution $q(x)$ equal to the desired normalized distribution $p(x)$.

Returning to the problem of Bayesian inference, the derivation above shows that the problem of computing the posterior $p(x|y)$ can be equivalently framed as the minimization of the free

energy $J(q) = \mathrm{E}_{\mathrm{x} \sim q(x)}[-\log p(\mathrm{x}, y)] - \mathrm{H}(q)$. When this optimization is not tractable, restricting the optimization space to a parametric, or non-parametric, family of distributions yields an approximate Bayesian inference approach known as *variational inference*, which is central to state-of-the-art scalable solutions to Bayesian inference [3]–[5]. Furthermore, allowing for a larger class of divergence penalty terms $D(q)$, for any loss function $L(x)$ – not necessarily the log-loss – and for any value of the temperature $T > 0$ yields the framework of *generalized Bayesian inference* [6]. We finally note that the ELBO (9) is also central to Bayesian and Minimum Description Length (MDL) model selection strategies [4], [7].

3. *Parametric Learning with Latent Variables*

Consider now the problem of learning a probabilistic model in the presence of latent, or unobserved, variables. The probabilistic model is defined by a joint distribution $p_\theta(x, y)$ over observations y and latent variables x that is parameterized by a vector θ. Maximum likelihood learning is based on the maximization of the marginal log-likelihood $\log p_\theta(y) = \log \sum_x p_\theta(x, y)$ for an observation y $= y$. Denoting the ELBO (9) as $J_\theta(q) = \mathrm{E}_{\mathrm{x} \sim q(x)}[-\log p_\theta(\mathrm{x}, y)] - \mathrm{H}(q)$ in order to highlight its dependence on the model parameter vector θ, this problem can be reformulated as the minimax optimization

$$\min_\theta \max_q \quad J_\theta(q). \tag{10}$$

This is because, as we have discussed in the previous section, the inner maximization yields the marginal log-likelihood $\log p_\theta(y)$ with the optimum achieved for $q(x) = p_\theta(x|y)$.

The Expectation Maximization (EM) algorithm tackles the ML problem (10) by iteratively optimizing over distribution q for fixed θ – which amounts to the Bayesian inference problem covered in the previous section – and optimizing over the parameter vector θ for a fixed q. In modern applications, neither of the two steps is typically tractable, and approximate solutions are used in which the inner maximization is replaced by variational inference and stochastic gradient-based descent-ascent methods are implemented [5]. Variants that consider alternative loss measures and the divergence metrics include the variational information bottleneck approach [8].

4. *Statistical Learning Theory*

Assume that we have a training data set s $= (\mathrm{z}_1, \ldots, \mathrm{z}_\mathrm{m})$ of m examples generated independently as $\mathrm{z}_i \sim p(z)$, with $i = 1, \ldots, m,$ from some unknown distribution $p(z)$. A (probabilistic)

learning algorithm is described by a stochastic mapping $q(x|s)$ between the training data set and a model parameter x. The algorithm is said to generalize if it "performs well" on a new, independently generated test example $z' \sim p(z)$. Denoting as $\ell(x, z)$ the loss accrued by model x when applied to data z, the performance criterion of interest for a model parameter x is hence the test loss $\mathcal{L}(x) = \mathrm{E}_{z' \sim p(z)}[\ell(x, z')]$ [3]–[5]. For a given training set s $= s$, the learner, however, has only access to the training loss $\mathcal{L}_s(x) = \frac{1}{m} \sum_{i=1}^{m} \ell(x, z_i)$. The difference between the test loss and the training loss $\Delta \mathcal{L}_s(x) = \mathcal{L}(x) - \mathcal{L}_s(x)$ is known as the *generalization gap* or generalization error. In the case of a probabilistic algorithm, one is typically interested in its average $\Delta \mathcal{L}_s = \mathrm{E}_{x \sim q(x|s)}[\Delta \mathcal{L}_s(x)]$. This is an important metric to gauge the generalization capacity of a training procedure $q(x|s)$. In fact, if the generalization gap is small, a low training loss, which can be guaranteed through the design of the training procedure $q(x|s)$, implies a low test loss.

While the average generalization gap $\Delta \mathcal{L}_s$ cannot be evaluated due to its dependence on the unknown test loss, one of the main concerns of statistical learning theory is obtaining computable upper bounds on it. As we will discuss, these bounds can be used both to define learning criteria and to obtain generalization measures.

To see how the framework of free energy minimization can be useful for this purpose, define as the loss function $L(x) = -\beta \Delta \mathcal{L}_s(x)$ and as the information complexity penalty $D(q) = \mathrm{KL}(q(x|s)||p(x))$ for some constant $\beta > 0$ and prior distribution $p(x)$. With this choice, by Table II, the minimum free energy with $T = 1$ is given as $J_{\mathrm{opt}} = -\log \mathrm{E}_{x \sim p(x)}[\exp(\beta \Delta \mathcal{L}_s(x))]$. This is the negative *log-moment generating function*, also known as negative cumulant generating function, of random variable $\Delta \mathcal{L}_s(x)$ under the prior distribution x $\sim p(x)$. Furthermore, by Table I and the Fenchel-Young inequality (4), we obtain the following upper bound – an instance of the *Donsker-Varadhan inequality* – on the average generalization gap

$$\mathrm{E}_{x \sim q(x|s)}[\beta \Delta \mathcal{L}_s(x)] \leq \mathrm{KL}(q(x|s)||p(x)) + \log \mathrm{E}_{x \sim p(x)}[\exp(\beta \Delta \mathcal{L}_s(x))]. \tag{11}$$

Let us now use the bound (11) to define a learning criterion. To this end, we note that this inequality directly implies that the average test loss $\mathrm{E}_{x \sim q(x|s)}[\mathcal{L}(x)]$ – the true objective of a learning procedure – can be upper bounded by the regularized (average) training loss $\mathrm{E}_{x \sim q(x|s)}[\mathcal{L}_s(x)] + \beta^{-1}\mathrm{KL}(q(x|s)||p(x))$ plus an additional term – the log-moment generating function – that does not depend on the learning algorithm $q(x|s)$ under optimization. Hence, for

a given prior distribution $p(x)$ and $\beta > 0$, a learning algorithm $q(x|s)$ that minimizes an upper bound on the average test loss is obtained by addressing the free energy minimization problem

$$\min_{q(x|s)} \mathrm{E}_{\mathrm{x}\sim q(x|s)}[\mathcal{L}_s(\mathrm{x})] + \frac{1}{\beta}\mathrm{KL}(q(x|s)||p(x)). \tag{12}$$

This approach is known as *information risk minimization* [9], and it can be seen to be a special case of generalized Bayesian inference (see Sec. II.2). The optimal choice of the learning algorithm follows from Table II as the Gibbs distribution $q_{\mathrm{opt}}(x|s) \propto p(x)\exp(-\beta\mathcal{L}_s(\mathrm{x}))$.

The inequality (11) can also be used to obtain an explicit upper bound on the generalization gap. Specifically, PAC Bayes theory seeks bounds that hold with high probability with respect to $\mathrm{s} \sim p(s)$ for any data distribution $p(s)$. This derivation is made difficult by the presence of the log-moment generating function (second term on the right-hand side in (11)), which depends on the unknown test loss. To proceed, one needs to make additional assumptions on the distribution of the loss function. The most basic derivation of PAC Bayes bounds assumes that the loss function is bounded, i.e., $a \leq l(\cdot, \cdot) \leq b$ for some parameters $0 \leq a < b$. In this case, an application of the Markov and Hoeffding inequalities followed by an optimization over β yields the following result [10]: With probability at least $1 - \delta$ over the training data $\mathrm{s} \sim p(s)$ for any distribution $p(s)$, the following upper bound on the average generalization gap holds

$$\mathrm{E}_{\mathrm{x}\sim q(x|s)}[\Delta\mathcal{L}_{\mathrm{s}}(\mathrm{x})] \leq \sqrt{\frac{(b-a)^2}{2m}\left(\mathrm{KL}(q(x|s)||p(x)) + \log\frac{1}{\delta}\right)}. \tag{13}$$

5. *Local Optimization*

The standard gradient descent algorithm for a differentiable function $g(x)$ produces a sequence of iterates $x^{(i)}$, for $i = 1, 2, ...$, with the $(i+1)$th iterate obtained as

$$x^{(i+1)} = \arg\min_{x} \ \nabla g(x^{(i)})^T x + \frac{1}{2\alpha^{(i)}}||x - x^{(i)}||^2, \tag{14}$$

where $\alpha^{(i)} > 0$ is a step size. The $(i+1)$th iterate in (14) is hence the minimum of a strongly convex approximation of function $g(x)$ that has the same gradient $\nabla g(x^{(i)})$ at the current iterate $x^{(i)}$. The quadratic term $||x - x^{(i)}||^2$ penalizes deviations from the current iterate by an amount that depends on the inverse of the step size. This ensures that the next iterate is localized within a "trust region" defined by the current iterate. It can be easily seen that the solution of problem (14) yields the familiar update $x^{(i+1)} = x^{(i)} - \alpha^{(i)}\nabla g(x^{(i)})$.

By (14), gradient descent penalizes deviations from the current iterate by assuming that the relevant geometry is Euclidean so that the distance is measured by the term $||x - x^{(i)}||^2$. In

problems involving the optimization over a probability distribution q, the Euclidean space, and associated distance metric $||x-x^{(i)}||^2$, are no longer natural choices, and other "distance" metrics, such as the KL divergence, have a more direct relevance.

Based on the above, considering a discrete distribution $q = [q(1), \ldots q(N_x)]^T$, one can define a variant of gradient descent that addresses, at each iteration i, the problem

$$q^{(i+1)} = \arg\min_q \ \nabla g(q^{(i)})^T q + \frac{1}{\alpha^{(i)}}\left(\text{KL}(q||q^{(i)})1_S(q)\right). \tag{15}$$

This is an instance of the minimization (1) of the free energy with $L(x) = [\nabla g(q^{(i)})]_x$, $T = 1/\alpha^{(i)}$, and $D(q) = \text{KL}(q||q^{(i)})1_S(q)$. It is also an instance of *mirror descent*, which, more generally, tackles a problem in the form (15) with a Bregman divergence as penalty and minimization constrained within a convex set [11]. By Table II, the solution to problem (15) is given as

$$q^{(i+1)}(x) \propto q^{(i)}(x) \exp(-\alpha^{(i)}[\nabla g(q^{(i)})]_x). \tag{16}$$

yielding the Normalized Exponentiated Gradient algorithm [1].

CONCLUSIONS

The minimization of the free energy is a general principle that underlies many of the technique that are commonly used in signal processing and machine learning for the key tasks of model selection, inference, learning from data, and local optimization. These notes have aimed at elucidating this unifying thread in simple terms by starting from first principles via Fenchel duality. More connections not covered here include the use of the Donsker-Varadhan inequality (11) for the estimation of divergences and mutual information from data (see [12] and references therein).

AUTHORS

Sharu Theresa Jose (sharu.jose@kcl.ac.uk) is a Postdoctoral Research Associate with the Department of Engineering at King's College London. Her research interests include information theory, statistical learning theory, stochastic control and optimization.

Osvaldo Simeone (osvaldo.simeone@kcl.ac.uk) is a Professor of Information Engineering with the Centre for Telecommunications Research at the Department of Informatics of King's College London, where he directs King's Communications, Learning, and Information Processing (KCLIP) lab. He is a Fellow of the IEEE.

REFERENCES

[1] S. Shalev-Shwartz *et al.*, "Online Learning and Online Convex Optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition,*. Wiley-Interscience, July 2006.

[3] O. Simeone, "A Brief Introduction to Machine Learning for Engineers," *Foundations and Trends® in Signal Processing*, vol. 12, no. 3-4, pp. 200–431, 2018.

[4] D. J. MacKay, *Information theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[5] E. Angelino, M. J. Johnson, and R. P. Adams, "Patterns of Scalable Bayesian Inference," *Foundations and Trends® in Signal Processing*, vol. 9, pp. 119–247, Nov 2016.

[6] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized Variational Inference," *arXiv preprint arXiv:1904.02063*, 2019.

[7] P. D. Grünwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[8] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep Variational Information Bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[9] T. Zhang, "Information-Theoretic Upper and Lower Bounds for Statistical Estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[10] B. Guedj and L. Pujol, "Still No Free Lunches: The Price to Pay for Tighter PAC-Bayes Bounds," *arXiv preprint arXiv:1910.04460*, 2019.

[11] A. Beck, *First-Order Methods in Optimization*, vol. 25. SIAM, 2017.

[12] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. Katsoulakis, "Cumulant GAN," *arXiv preprint arXiv:2006.06625*, 2020.