# A Brief Introduction to Machine Learning for Engineers

## Other titles in Foundations and Trends® in Signal Processing

*Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*
Emil Bjornson, Jakob Hoydis and Luca Sanguinetti
ISBN: 978-1-68083-985-2

*Using Inertial Sensors for Position and Orientation Estimation*
Manon Kok, Jeroen D. Hol and Thomas B. Schon
ISBN: 978-1-68083-356-0

*Computational Visual Attention Models*
Milind S. Gide and Lina J. Karam
ISBN: 978-1-68083-280-8

*Video Coding: Part II of Fundamentals of Source and Video Coding*
Thomas Wiegand and Heiko Schwarz
ISBN: 978-1-68083-178-8

Full text available at: http://dx.doi.org/10.1561/2000000102

# A Brief Introduction to Machine Learning for Engineers

**Osvaldo Simeone**
Department of Informatics
King's College London
osvaldo.simeone@kcl.ac.uk

the essence of knowledge
Boston — Delft

# Foundations and Trends® in Signal Processing

*Published, sold and distributed by:*
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

*Outside North America:*
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

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.

ISBN: 978-1-68083-473-4

Full text available at: http://dx.doi.org/10.1561/2000000102

**Foundations and Trends® in Signal Processing**

Volume 12, Issue 3-4, 2018

# Editorial Scope

## Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
  - Classification and detection
  - Estimation and regression
  - Tree-structured methods

## Information for Librarians

Foundations and Trends® in Signal Processing, 2018, Volume 12, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

# Contents

# A Brief Introduction to Machine Learning for Engineers

Osvaldo Simeone[1]

[1]*Department of Informatics, King's College London;*
*osvaldo.simeone@kcl.ac.uk*

ABSTRACT

This monograph aims at providing an introduction to key concepts, algorithms, and theoretical results in machine learning. The treatment concentrates on probabilistic models for supervised and unsupervised learning problems. It introduces fundamental concepts and algorithms by building on first principles, while also exposing the reader to more advanced topics with extensive pointers to the literature, within a unified notation and mathematical framework. The material is organized according to clearly defined categories, such as discriminative and generative models, frequentist and Bayesian approaches, exact and approximate inference, as well as directed and undirected models. This monograph is meant as an entry point for researchers with an engineering background in probability and linear algebra.

Osvaldo Simeone (2018), "A Brief Introduction to Machine Learning for Engineers", Foundations and Trends® in Signal Processing: Vol. 12, No. 3-4, pp 200–431. DOI: 10.1561/2000000102.

# Notation

- Random variables or random vectors – both abbreviated as rvs – are represented using roman typeface, while their values and realizations are indicated by the corresponding standard font. For instance, the equality $\mathrm{x} = x$ indicates that rv $\mathrm{x}$ takes value $x$.
- Matrices are indicated using uppercase fonts, with roman typeface used for random matrices.
- Vectors will be taken to be in column form.
- $X^T$ and $X^\dagger$ are the transpose and the pseudoinverse of matrix $X$, respectively.
- The distribution of a rv $\mathrm{x}$, either probability mass function (pmf) for a discrete rv or probability density function (pdf) for continuous rvs, is denoted as $p_{\mathrm{x}}$, $p_{\mathrm{x}}(x)$, or $p(x)$.
- The notation $\mathrm{x} \sim p_{\mathrm{x}}$ indicates that rv $\mathrm{x}$ is distributed according to $p_{\mathrm{x}}$.
- For jointly distributed rvs $(\mathrm{x}, \mathrm{y}) \sim p_{\mathrm{xy}}$, the conditional distribution of $\mathrm{x}$ given the observation $\mathrm{y} = y$ is indicated as $p_{\mathrm{x}|\mathrm{y}=y}$, $p_{\mathrm{x}|\mathrm{y}}(x|y)$ or $p(x|y)$.
- The notation $\mathrm{x}|\mathrm{y} = y \sim p_{\mathrm{x}|\mathrm{y}=y}$ indicates that rv $\mathrm{x}$ is drawn according to the conditional distribution $p_{\mathrm{x}|\mathrm{y}=y}$.
- The notation $\mathrm{E}_{\mathrm{x}\sim p_{\mathrm{x}}}[\cdot]$ indicates the expectation of the argument with respect to the distribution of the rv $\mathrm{x} \sim p_{\mathrm{x}}$. Accordingly, we will also write $\mathrm{E}_{\mathrm{x}\sim p_{\mathrm{x}|\mathrm{y}}}[\cdot|y]$ for the conditional expectation with respect to

the distribution $p_{\mathrm{x}|\mathrm{y}=y}$. When clear from the context, the distribution over which the expectation is computed may be omitted.

- The notation $\Pr_{\mathrm{x}\sim p_{\mathrm{x}}}[\cdot]$ indicates the probability of the argument event with respect to the distribution of the rv $\mathrm{x} \sim p_{\mathrm{x}}$. When clear from the context, the subscript is dropped.

- The notation log represents the logarithm in base two, while ln represents the natural logarithm.

- $\mathrm{x} \sim \mathcal{N}(\mu, \Sigma)$ indicates that random vector x is distributed according to a multivariate Gaussian pdf with mean vector $\mu$ and covariance matrix $\Sigma$. The multivariate Gaussian pdf is denoted as $\mathcal{N}(x|\mu, \Sigma)$ as a function of $x$.

- $\mathrm{x} \sim \mathcal{U}(a, b)$ indicates that rv x is distributed according to a uniform distribution in the interval $[a, b]$. The corresponding uniform pdf is denoted as $\mathcal{U}(x|a, b)$.

- $\delta(x)$ denotes the Dirac delta function or the Kronecker delta function, as clear from the context.

- $||a||^2 = \sum_{i=1}^{N} a_i^2$ is the quadratic, or $l_2$, norm of a vector $a = [a_1, \ldots, a_N]^T$. We similarly define the $l_1$ norm as $||a||_1 = \sum_{i=1}^{N} |a_i|$, and the $l_0$ pseudo-norm $||a||_0$ as the number of non-zero entries of vector $a$.

- $I$ denotes the identity matrix, whose dimensions will be clear from the context. Similarly, 1 represents a vector of all ones.

- $\mathbb{R}$ is the set of real numbers; $\mathbb{R}^+$ the set of non-negative real numbers; $\mathbb{R}^-$ the set of non-positive real numbers; and $\mathbb{R}^N$ is the set of all vectors of $N$ real numbers.

- $1(\cdot)$ is the indicator function: $1(x) = 1$ if $x$ is true, and $1(x) = 0$ otherwise.

- $|\mathcal{S}|$ represents the cardinality of a set $\mathcal{S}$.

- $x_{\mathcal{S}}$ represents a set of rvs $x_k$ indexed by the integers $k \in \mathcal{S}$.

# Acronyms

AI: Artificial Intelligence
AMP: Approximate Message Passing
BN: Bayesian Network
DAG: Directed Acyclic Graph
ELBO: Evidence Lower BOund
EM: Expectation Maximization
ERM: Empirical Risk Minimization
GAN: Generative Adversarial Network
GLM: Generalized Linear Model
HMM: Hidden Markov Model
i.i.d.: independent identically distributed
KL: Kullback-Leibler
LASSO: Least Absolute Shrinkage and Selection Operator
LBP: Loopy Belief Propagation
LL: Log-Likelihood
LLR: Log-Likelihood Ratio
LS: Least Squares
MC: Monte Carlo
MCMC: Markov Chain Monte Carlo
MDL: Minimum Description Length
MFVI: Mean Field Variational Inference
ML: Maximum Likelihood

MRF: Markov Random Field
NLL: Negative Log-Likelihood
PAC: Probably Approximately Correct
pdf: probability density function
pmf: probability mass function
PCA: Principal Component Analysis
PPCA: Probabilistic Principal Component Analysis
QDA: Quadratic Discriminant Analysis
RBM: Restricted Boltzmann Machine
SGD: Stochastic Gradient Descent
SVM: Support Vector Machine
rv: random variable or random vector (depending on the context)
s.t.: subject to
VAE: Variational AutoEncoder
VC: Vapnik–Chervonenkis
VI: Variational Inference

# Part I

# Basics

# 1

# Introduction

Having taught courses on machine learning, I am often asked by colleagues and students with a background in engineering to suggest "the best place to start" to get into this subject. I typically respond with a list of books – for a general, but slightly outdated introduction, read this book; for a detailed survey of methods based on probabilistic models, check this other reference; to learn about statistical learning, I found this text useful; and so on. This answer strikes me, and most likely also my interlocutors, as quite unsatisfactory. This is especially so since the size of many of these books may be discouraging for busy professionals and students working on other projects. This monograph is an attempt to offer a basic and compact reference that describes key ideas and principles in simple terms and within a unified treatment, encompassing also more recent developments and pointers to the literature for further study.

## 1.1 What is Machine Learning?

A useful way to introduce the machine learning methodology is by means of a comparison with the conventional engineering design flow. This starts with a in-depth analysis of the problem domain, which culminates with the definition of a mathematical model. The mathematical model is meant to capture the key features of the problem under study, and is typically the result of the work of a number of experts. The mathematical model is finally leveraged to derive hand-crafted solutions to the problem that offer given optimality guarantees.

For instance, consider the problem of defining a chemical process to produce a given molecule. The conventional flow requires chemists to leverage their knowledge of models that predict the outcome of individual chemical reactions, in order to craft a sequence of suitable steps that synthesize the desired molecule. Another example is the design of speech translation or image/video compression algorithms. Both of these tasks involve the definition of models and algorithms by teams of experts, such as linguists, psychologists, and signal processing practitioners, not infrequently during the course of long standardization meetings.

The engineering design flow outlined above may be too costly and inefficient for problems in which faster or less expensive solutions are desirable. The machine learning alternative is to collect large data sets, e.g., of labelled speech, images or videos, and to use this information to train general-purpose learning machines to carry out the desired task. While the standard engineering flow relies on domain knowledge and on design optimized for the problem at hand, machine learning lets large amounts of data dictate algorithms and solutions. To this end, rather than requiring a precise model of the set-up under study, machine learning requires the specification of an objective, of a generic model to be trained, and of an optimization technique.

Returning to the first example above, a machine learning approach would proceed by training a general-purpose machine to predict the outcome of known chemical reactions based on a large data set, and by then using the trained algorithm to explore ways to produce more complex molecules. In a similar manner, large data sets of images or videos would be used to train a general-purpose algorithm with the aim of obtaining compressed representations from which the original input can be recovered with some distortion.

## 1.2 When to Use Machine Learning?

Based on the discussion above, machine learning can offer an efficient alternative to the conventional engineering flow when development cost and time are the main concerns, or when the problem appears to be too complex to afford the development of solutions with optimality guarantees. On the flip side, the approach has the key disadvantages of providing generally suboptimal performance, of producing black-box, and hence non-interpretable, solutions, and of applying only to a limited set of problems.

In order to identify tasks for which machine learning methods may be useful, reference [31] suggests the following criteria:

1. the task involves a function that maps well-defined inputs to well-defined outputs;
2. large data sets exist or can be created containing input-output pairs;
3. the task provides clear feedback with clearly definable goals and metrics;
4. the task does not involve long chains of logic or reasoning that depend on diverse background knowledge or common sense;
5. the task does not require detailed explanations for how the decision was made;
6. the task has a tolerance for error and no need for provably correct or optimal solutions;
7. the phenomenon or function being learned should not change rapidly over time; and
8. no specialized dexterity, physical skills, or mobility is required.

These criteria are useful guidelines for the decision of whether machine learning methods are suitable for a given task of interest. They also offer a convenient demarcation line between machine learning as is intended today, with its focus on training and computational statistics tools, and more general notions of Artificial Intelligence (AI) based on knowledge and common sense [86] (see [126] for an overview on AI research).

### 1.2.1 Learning Tasks

We can distinguish among three different main types of machine learning problems, which are briefly introduced below. The discussion reflects the focus of this monograph on parametric probabilistic models, as further elaborated on in the next section.

**1. Supervised learning:** We have $N$ labelled training examples $\mathcal{D}=\{(x_n, t_n)\}_{n=1}^N$, where $x_n$ represents a covariate, or explanatory variable, while $t_n$ is the corresponding target label, or response. For instance, variable $x_n$ may represent the text of an email, while the label $t_n$ may be a binary variable indicating whether the email is spam or not. The goal of supervised learning is to predict the value of the label $t$ for an input $x$ that is not in the training set. In other words, supervised learning aims at generalizing the observations in the data set $\mathcal{D}$ to new inputs. For example, an algorithm trained on a set of emails should be able to classify a new email not present in the data set $\mathcal{D}$.

We can generally distinguish between *classification* problems, in which the label $t$ is discrete, as in the example above, and *regression* problems, in which variable $t$ is continuous. An example of a regression task is the prediction of tomorrow's temperature $t$ based on today's meteorological observations $x$.

An effective way to learn a predictor is to identify from the data set $\mathcal{D}$ a predictive distribution $p(t|x)$ from a set of parametrized distributions. The conditional distribution $p(t|x)$ defines a profile of beliefs over all possible of the label $t$ given the input $x$. For instance, for temperature prediction, one could learn mean and variance of a Gaussian distribution $p(t|x)$ as a function of the input $x$. As a special case, the output of a supervised learning algorithm may be in the form of a deterministic predictive function $t = \hat{t}(x)$.

**2. Unsupervised learning:** Suppose now that we have an unlabelled set of training examples $\mathcal{D}=\{x_n\}_{n=1}^N$. Less well defined than supervised learning, unsupervised learning generally refers to the task of learning properties of the mechanism that generates this data set. Specific tasks and applications include clustering, which is the problem of grouping similar examples $x_n$; dimensionality reduction, feature extraction, and representation learning, all related to the problem of representing the data in a smaller or more convenient space; and generative modelling, which is the problem of learning a generating mechanism to produce artificial examples that are similar to available data in the data set $\mathcal{D}$.

As a generalization of both supervised and unsupervised learning, *semi-supervised learning* refers to scenarios in which not all examples are labelled, with the unlabelled examples providing information about the distribution of the covariates $x$.

**3. Reinforcement learning:** Reinforcement learning refers to the problem of inferring optimal sequential decisions based on rewards or punishments received as a result of previous actions. Under supervised learning, the "label" $t$ refers to an action to be taken when the learner is in an informational state about the environment given by a variable $x$. Upon taking an action $t$ in a state $x$, the learner is provided with feedback on the immediate reward accrued via this decision, and the environment moves on to a different state. As an example, an agent can be trained to navigate a given environment in the presence of obstacles by penalizing decisions that result in collisions.

Reinforcement learning is hence neither supervised, since the learner is not provided with the optimal actions $t$ to select in a given state $x$; nor is it fully unsupervised, given the availability of feedback on the quality of the chosen action. Reinforcement learning is also distinguished from supervised and unsupervised learning due to the influence of previous actions on future states and rewards.

This monograph focuses on supervised and unsupervised learning. These general tasks can be further classified along the following dimensions.

- *Passive vs. active learning*: A passive learner is given the training examples, while an active learner can affect the choice of training examples on the basis of prior observations.

- *Offline vs. online learning*: Offline learning operates over a batch of training samples, while online learning processes samples in a streaming fashion. Note that reinforcement learning operates inherently in an online manner, while supervised and unsupervised learning can be carried out by following either offline or online formulations.

This monograph considers only passive and offline learning.

## 1.3 Goals and Outline

This monograph aims at providing an introduction to key concepts, algorithms, and theoretical results in machine learning. The treatment concentrates on probabilistic models for supervised and unsupervised learning problems. It introduces fundamental concepts and algorithms by building on first principles, while also exposing the reader to more advanced topics with extensive pointers to the literature, within a unified notation and mathematical framework. Unlike other texts that are focused on one particular aspect of the field, an effort has been made here to provide a broad but concise overview in which the main ideas and techniques are systematically presented. Specifically, the material is organized according to clearly defined categories, such as discriminative and generative models, frequentist and Bayesian approaches, exact and approximate inference, as well as directed and undirected models. This monograph is meant as an entry point for researchers with a background in probability and linear algebra. A prior exposure to information theory is useful but not required.

Detailed discussions are provided on basic concepts and ideas, including overfitting and generalization, Maximum Likelihood and regularization, and Bayesian inference. The text also endeavors to provide intuitive explanations and pointers to advanced topics and research directions. Sections and subsections containing more advanced material that may be skipped at a first reading are marked with a star $(*)$.

The reader will find here neither discussions on computing platform or programming frameworks, such as map-reduce, nor details on specific applications involving large data sets. These can be easily found in a vast and growing body of work. Furthermore, rather than providing exhaustive details on the existing myriad solutions in each specific category, techniques have been selected that are useful to illustrate the most salient aspects. Historical notes have also been provided only for a few selected milestone events.

Finally, the monograph attempts to strike a balance between the algorithmic and theoretical viewpoints. In particular, all learning algorithms are introduced on the basis of theoretical arguments, often based on information-theoretic measures. Moreover, a chapter is devoted to statistical learning theory, demonstrating how to set the field of supervised learning on solid theoretical foundations. This chapter is more theoretically involved than the others, and proofs of some key results are included in order to illustrate the theoretical underpinnings of learning. This contrasts with other chapters, in which proofs of the few theoretical results are kept at a minimum in order to focus on the main ideas.

The rest of the monograph is organized into five parts. The first part covers introductory material. Specifically, Chapter 2 introduces the frequentist, Bayesian and Minimum Description Length (MDL) learning frameworks; the discriminative and generative categories of probabilistic models; as well as key concepts such as training loss, generalization, and overfitting – all in the context of a simple linear regression problem. Information-theoretic metrics are also briefly introduced, as well as the advanced topics of interpretation and causality. Chapter 3 then provides an introduction to the exponential family of probabilistic models, to Generalized Linear Models (GLMs), and to energy-based models, emphasizing main properties that will be invoked in later chapters.

The second part concerns supervised learning. Chapter 4 covers linear and non-linear classification methods via discriminative and generative models, including Support Vector Machines (SVMs), kernel methods, logistic regression, multi-layer neural networks and boosting. Chapter 5 is a brief introduction to the statistical learning framework of the Probably Approximately Correct (PAC) theory, covering the Vapnik–Chervonenkis (VC) dimension and the fundamental theorem of PAC learning.

The third part, consisting of a single chapter, introduced unsupervised learning. In particular, in Chapter 6, unsupervised learning models are described by distinguishing among directed models, for which Expectation Maximization (EM) is derived as the iterative maximization of the Evidence Lower BOund (ELBO); undirected models, for which Restricted Boltzmann Machines (RBMs) are discussed as a representative example; discriminative models trained using the InfoMax principle; and autoencoders. Generative Adversarial Networks (GANs) are also introduced.

The fourth part covers more advanced modelling and inference approaches. Chapter 7 provides an introduction to probabilistic graphical models, namely Bayesian Networks (BNs) and Markov Random Fields (MRFs), as means to encode more complex probabilistic dependencies than the models studied in previous chapters. Approximate inference and learning methods are introduced in Chapter 8 by focusing on Monte Carlo (MC) and Variational Inference (VI) techniques. The chapter briefly introduces in a unified way techniques such as variational EM, Variational AutoEncoders (VAE), and black-box inference. Some concluding remarks are provided in the last part, consisting of Chapter 9.

We conclude this chapter by emphasizing the importance of probability as a common language for the definition of learning algorithms [34]. The centrality of the probabilistic viewpoint was not always recognized, but has deep historical roots. This is demonstrated by the following two quotes, the first from the first AI textbook published by P. H. Winston in 1977, and the second from an unfinished manuscript by J. von Neumann (see [126, 63] for more information):

> “Many ancient Greeks supported Socrates opinion that deep, inexplicable thoughts came from the gods. Today’s equivalent to those gods is the erratic, even probabilistic neuron. It is more likely that increased randomness of neural behavior is the problem of the epileptic and the drunk, not the advantage of the brilliant.”

from *Artificial Intelligence*, 1977;

> "All of this will lead to theories of computation which are much less rigidly of an all-or-none nature than past and present formal logic... There are numerous indications to make us believe that this new system of formal logic will move closer to another discipline which has been little linked in the past with logic. This is thermodynamics primarily in the form it was received from Boltzmann."

from *The Computer and the Brain*, 1958.

Full text available at: http://dx.doi.org/10.1561/2000000102

# References

[1] Abadi, M., Ú. Erlingsson, I. Goodfellow, H. Brendan McMahan, I. Mironov, N. Papernot, K. Talwar, and L. Zhang. 2017. "On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches". *ArXiv e-prints*. Aug. arXiv: 1708.08022 [stat.ML].

[2] Abu-Mostafa, Y. S., M. Magdon-Ismail, and H.-T. Lin. 2012. *Learning from data*. Vol. 4. AMLBook New York, NY, USA.

[3] Agakov, F. 2005. *Variational Information Maximization in Stochastic Environments (PhD thesis)*. University of Edinburgh.

[4] Alemi, A. A., B. Poole, and E. a. Fischer. 2017. "An Information-Theoretic Analysis of Deep Latent-Variable Models". *ArXiv e-prints*. Nov. arXiv: 1711.00464v1.

[5] Amari, S.-I. 1998. "Natural gradient works efficiently in learning". *Neural computation*. 10(2): 251–276.

[6] Amari, S.-I. 2016. *Information geometry and its applications*. Springer.

[7] Angelino, E., M. J. Johnson, and R. P. Adams. 2016. "Patterns of scalable Bayesian inference". *Foundations and Trends® in Machine Learning*. 9(2-3): 119–247.

[8] Arjovsky, M., S. Chintala, and L. Bottou. 2017. "Wasserstein GAN". *arXiv preprint arXiv:1701.07875*.

[9] Arulkumaran, K., M. P. Deisenroth, M. Brundage, and A. A. Bharath. 2017. "Deep Reinforcement Learning: A Brief Survey". *IEEE Signal Processing Magazine*. 34(6): 26–38. ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2743240.

[10] Azoury, K. S. and M. K. Warmuth. 2001. "Relative loss bounds for on-line density estimation with the exponential family of distributions". *Machine Learning*. 43(3): 211–246.

[11] Bagheri, A., O. Simeone, and B. Rajendran. 2017. "Training Probabilistic Spiking Neural Networks with First-to-spike Decoding". *ArXiv e-prints*. Oct. arXiv: 1710.10704 [stat.ML].

[12] Baldi, P., P. Sadowski, and Z. Lu. 2016. "Learning in the machine: Random backpropagation and the learning channel". *arXiv preprint arXiv:1612.02734*.

[13] Bamler, R., C. Zhang, M. Opper, and S. Mandt. 2017. "Perturbative Black Box Variational Inference". *ArXiv e-prints*. Sept. arXiv: 1709.07433 [stat.ML].

[14] Baraniuk, R. G. 2007. "Compressive sensing [lecture notes]". *IEEE signal processing magazine*. 24(4): 118–121.

[15] Barber, D. 2012. *Bayesian reasoning and machine learning*. Cambridge University Press.

[16] Beal, M. J. 2003. *Variational algorithms for approximate Bayesian inference*. University of London, London.

[17] Bekkerman, R., M. Bilenko, and J. Langford. 2011. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press.

[18] Belghazi, I., S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville. 2018. "MINE: Mutual Information Neural Estimation". *arXiv preprint arXiv:1801.04062*.

[19] Bengio, Y. 2012. "Deep learning of representations for unsupervised and transfer learning". In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 17–36.

[20] Bengio, Y., A. Courville, and P. Vincent. 2013. "Representation learning: A review and new perspectives". *IEEE transactions on pattern analysis and machine intelligence*. 35(8): 1798–1828.

[21] Berisha, V., A. Wisler, A. O. Hero, and A. Spanias. 2016. “Empirically estimable classification bounds based on a nonparametric divergence measure”. *IEEE Transactions on Signal Processing*. 64(3): 580–591.

[22] Bertsekas, D. P. 2011. “Incremental gradient, subgradient, and proximal methods for convex optimization: A survey”. *Optimization for Machine Learning*. 2010(1-38): 3.

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

[24] Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2017. “Variational inference: A review for statisticians”. *Journal of the American Statistical Association*. (just-accepted).

[25] Blei, D., R. Ranganath, and S. Mohamed. “Variational Inference: Foundations and Modern Methods”.

[26] Blitzer, J., K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. 2008. “Learning bounds for domain adaptation”. In: *Advances in neural information processing systems*. 129–136.

[27] Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra. 2015. “Weight uncertainty in neural networks”. *arXiv preprint arXiv:1505.05424*.

[28] Boyd, S. and L. Vandenberghe. 2004. *Convex optimization*. Cambridge University Press.

[29] Brakel, P. and Y. Bengio. 2017. “Learning Independent Features with Adversarial Nets for Non-linear ICA”. *ArXiv e-prints*. Oct. arXiv: 1710.05050 [stat.ML].

[30] Bronstein, M. M., J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. 2017. “Geometric deep learning: going beyond euclidean data”. *IEEE Signal Processing Magazine*. 34(4): 18–42.

[31] Brynjolfsson, E. and T. Mitchell. 2017. “What can machine learning do? Workforce implications”. *Science*. 358(6370): 1530–1534.

[32] Burda, Y., R. Grosse, and R. Salakhutdinov. 2015. “Importance weighted autoencoders”. *arXiv preprint arXiv:1509.00519*.

[33] Cevher, V., S. Becker, and M. Schmidt. 2014. "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics". *IEEE Signal Processing Magazine*. 31(5): 32–43.

[34] Cheeseman, P. C. 1985. "In Defense of Probability." In: *IJCAI*. Vol. 85. 1002–1009.

[35] Cichocki, A., D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. 2015. "Tensor decompositions for signal processing applications: From two-way to multiway component analysis". *IEEE Signal Processing Magazine*. 32(2): 145–163.

[36] Collins, M., S. Dasgupta, and R. E. Schapire. 2002. "A generalization of principal components analysis to the exponential family". In: *Advances in neural information processing systems*. 617–624.

[37] Cortes, C. and V. Vapnik. 1995. "Support-vector networks". *Machine learning*. 20(3): 273–297.

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

[39] Cristianini, N. and J. Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

[40] Csiszár, I. and P. C. Shields. 2004. "Information theory and statistics: A tutorial". *Foundations and Trends® in Communications and Information Theory*. 1(4): 417–528.

[41] Davidson-Pilon, C. 2015. "Probabilistic Programming & Bayesian Methods for Hackers".

[42] Dayan, P., G. E. Hinton, R. M. Neal, and R. S. Zemel. 1995. "The helmholtz machine". *Neural computation*. 7(5): 889–904.

[43] De, S., G. Taylor, and T. Goldstein. 2015. "Variance Reduction for Distributed Stochastic Gradient Descent". *arXiv preprint arXiv:1512.01708*.

[44] Di Lorenzo, P. and G. Scutari. 2016. "Next: In-network nonconvex optimization". *IEEE Transactions on Signal and Information Processing over Networks*. 2(2): 120–136.

[45] Duchi, J. 2016. “Lecture Notes for Statistics 311/Electrical Engineering 377”.

[46] Duchi, J. C., K. Khosravi, and F. Ruan. 2016. “Information Measures, Experiments, Multi-category Hypothesis Tests, and Surrogate Losses”. *arXiv preprint arXiv:1603.00126*.

[47] Dumoulin, V., I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. 2016. “Adversarially learned inference”. *arXiv preprint arXiv:1606.00704*.

[48] Efron, B. and T. Hastie. 2016. *Computer Age Statistical Inference*. Vol. 5. Cambridge University Press.

[49] Fedus, W., M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow. 2017. “Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step”. *ArXiv e-prints*. Oct. arXiv: 1710.08446 [stat.ML].

[50] Feutry, C., P. Piantanida, Y. Bengio, and P. Duhamel. 2018. “Learning Anonymized Representations with Adversarial Neural Networks”. *ArXiv e-prints*. Feb. arXiv: 1802.09386 [stat.ML].

[51] Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

[52] Fu, Y., T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong. 2017. “Recent advances in zero-shot recognition”. *arXiv preprint arXiv:1710.04837*.

[53] Gal, Y. 2016. “Uncertainty in Deep Learning”. *PhD thesis*. University of Cambridge.

[54] Gersho, A. and R. M. Gray. 2012. *Vector quantization and signal compression*. Vol. 159. Springer Science & Business Media.

[55] Goodfellow, I. J., J. Shlens, and C. Szegedy. 2014a. “Explaining and harnessing adversarial examples”. *ArXiv e-prints*. arXiv: 1412.6572.

[56] Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. MIT Press.

[57] Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014b. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2672–2680.

[58] Grant, M., S. Boyd, and Y. Ye. 2009. “cvx users’ guide”.

[59] Grathwohl, W., D. Choi, Y. Wu, G. Roeder, and D. Duvenaud. 2017. “Backpropagation through the Void: Optimizing control variates for black-box gradient estimation”. *ArXiv e-prints*. Oct. arXiv: 1711.00123 [cs.LG].

[60] Grunwald, P. D. 2007. *The minimum description length principle*. MIT Press.

[61] Grünwald, P. D. and A. P. Dawid. 2004. “Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory”. *The Annals of Statistics*. 32(4): 1367–1433.

[62] Haveliwala, T. H. 2003. “Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search”. *IEEE transactions on knowledge and data engineering*. 15(4): 784–796.

[63] Hinton, G. 2016. “Neural Networks for Machine Learning (online course)”.

[64] Hinton, G. E., P. Dayan, B. J. Frey, and R. M. Neal. 1995. “The "wake-sleep" algorithm for unsupervised neural networks”. *Science*. 268(5214): 1158.

[65] Hochreiter, S. and J. Schmidhuber. 1997. “Flat minima”. *Neural Computation*. 9(1): 1–42.

[66] Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew. 2006. “Extreme learning machine: theory and applications”. *Neurocomputing*. 70(1): 489–501.

[67] Huszár, F. “Everything that Works Works Because it’s Bayesian: Why Deep Nets Generalize?” URL: http://www.inference.vc/.

[68] Huszár, F. 2017a. “Choice of Recognition Models in VAEs: a regularisation view”. URL: http://www.inference.vc/.

[69] Huszár, F. 2017b. “Is Maximum Likelihood Useful for Representation Learning?” URL: http://www.inference.vc/.

[70] Huszár, F. 2017c. “Variational Inference using Implicit Distributions”. *arXiv preprint arXiv:1702.08235*.

[71] Jain, P. and P. Kar. 2017. "Non-convex Optimization for Machine Learning". *Foundations and Trends® in Machine Learning*. 10(3-4): 142–336. ISSN: 1935-8237. URL: http://dx.doi.org/10.1561/2200000058.

[72] Jang, E., S. Gu, and B. Poole. 2016. "Categorical reparameterization with gumbel-softmax". *arXiv preprint arXiv:1611.01144*.

[73] Jiao, J., T. A. Courtade, A. No, K. Venkat, and T. Weissman. 2014. "Information measures: the curious case of the binary alphabet". *IEEE Transactions on Information Theory*. 60(12): 7616–7626.

[74] Jiao, J., T. A. Courtade, K. Venkat, and T. Weissman. 2015. "Justification of logarithmic loss via the benefit of side information". *IEEE Transactions on Information Theory*. 61(10): 5357–5365.

[75] Johnson, R. and T. Zhang. 2013. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in neural information processing systems*. 315–323.

[76] Karpathy, A. "Deep Reinforcement Learning: Pong from Pixels". URL: http://karpathy.github.io/2016/05/31/rl/.

[77] Kawaguchi, K., L. Pack Kaelbling, and Y. Bengio. 2017. "Generalization in Deep Learning". *ArXiv e-prints*. Oct. arXiv: 1710.05468 [stat.ML].

[78] Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. 2016. "On large-batch training for deep learning: Generalization gap and sharp minima". *arXiv preprint arXiv:1609.04836*.

[79] Kingma, D. P. and M. Welling. 2013. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114*.

[80] Koller, D. and N. Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT Press.

[81] Korenkevych, D., Y. Xue, Z. Bian, F. Chudak, W. G. Macready, J. Rolfe, and E. Andriyash. 2016. "Benchmarking quantum hardware for training of fully visible boltzmann machines". *arXiv preprint arXiv:1611.04528*.

[82] LeCun, Y., S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. 2006. “A tutorial on energy-based learning”. *Predicting structured data*. 1.

[83] Lee, J. H., T. Delbruck, and M. Pfeiffer. 2016. “Training deep spiking neural networks using backpropagation”. *Frontiers in neuroscience*. 10.

[84] Lee, T.-W., M. Girolami, and T. J. Sejnowski. 1999. “Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources”. *Neural computation*. 11(2): 417–441.

[85] Lehmann, E. L. and G. Casella. 2006. *Theory of point estimation*. Springer Science & Business Media.

[86] Levesque, H. J. 2017. *Common Sense, the Turing Test, and the Quest for Real AI*. MIT University Press.

[87] Levin, S. 2016. *A beauty contest was judged by AI and the robots didn’t like dark skin*. URL: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[88] Levine, S. 2017. *Deep Reinforcement Learning*. URL: http://rll.berkeley.edu/deeprlcourse/#lecture-videos.

[89] Li, Y. “Topics in Approximate Inference”. URL: http://yingzhenli.net/home/pdf/topics_approx_infer.pdf.

[90] Li, Y. and R. E. Turner. 2016. “Rényi divergence variational inference”. In: *Advances in Neural Information Processing Systems*. 1073–1081.

[91] Loh, P.-L. 2017. “On Lower Bounds for Statistical Learning Theory”. *Entropy*. 19(11): 617.

[92] Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2012. *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press.

[93] Maaten, L. v. d. and G. Hinton. 2008. “Visualizing data using t-SNE”. *Journal of Machine Learning Research*. 9(Nov): 2579–2605.

[94] MacKay, D. J. 2003. *Information theory, inference and learning algorithms*. Cambridge University Press.

[95] Maddison, C. J., A. Mnih, and Y. W. Teh. 2016. “The concrete distribution: A continuous relaxation of discrete random variables”. *arXiv preprint arXiv:1611.00712*.

[96] Minka, T. 2005. “Divergence measures and message passing”. *Tech. rep.* Technical report, Microsoft Research.

[97] Minsky, M. and S. Papert. 1969. “Perceptrons.”

[98] Mnih, A. and K. Gregor. 2014. “Neural variational inference and learning in belief networks”. *arXiv preprint arXiv:1402.0030*.

[99] Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013. “Playing atari with deep reinforcement learning”. *arXiv preprint arXiv:1312.5602*.

[100] Mohamed, S. and B. Lakshminarayanan. 2016. “Learning in implicit generative models”. *arXiv preprint arXiv:1610.03483*.

[101] Mokhtari, A. and A. Ribeiro. 2017. “First-Order Adaptive Sample Size Methods to Reduce Complexity of Empirical Risk Minimization”. *ArXiv e-prints*. Sept. arXiv: 1709.00599 [cs.LG].

[102] Montavon, G., W. Samek, and K.-R. Müller. 2017. “Methods for interpreting and understanding deep neural networks”. *arXiv preprint arXiv:1706.07979*.

[103] Mott, A., J. Job, J.-R. Vlimant, D. Lidar, and M. Spiropulu. 2017. “Solving a Higgs optimization problem with quantum annealing for machine learning”. *Nature*. 550(7676): 375.

[104] Murphy, K. P. 2012. *Machine learning: a probabilistic perspective.* MIT Press.

[105] Nguyen, X., M. J. Wainwright, and M. I. Jordan. 2010. “Estimating divergence functionals and the likelihood ratio by convex risk minimization”. *IEEE Transactions on Information Theory*. 56(11): 5847–5861.

[106] Nielsen, F. 2011. “Chernoff information of exponential families”. *arXiv preprint arXiv:1102.2684*.

[107] Nowozin, S., B. Cseke, and R. Tomioka. 2016. “f-GAN: Training generative neural samplers using variational divergence minimization”. In: *Advances in Neural Information Processing Systems*. 271–279.

[108] Odena, A., C. Olah, and J. Shlens. 2016. "Conditional image synthesis with auxiliary classifier gans". *arXiv preprint arXiv:1610.09585*.

[109] O'Neil, K. 2016. *Weapons of Math Destruction*. Penguin Books.

[110] Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. "The PageRank citation ranking: Bringing order to the web." *Tech. rep.* Stanford InfoLab.

[111] Papernot, N., P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami. 2016. "Practical Black-Box Attacks against Machine Learning". *ArXiv e-prints*. Feb. arXiv: 1602.02697 [cs.CR].

[112] Pearl, J. 2018. "Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution". *ArXiv e-prints*. Jan. arXiv: 1801.04016 [cs.LG].

[113] Pearl, J., M. Glymour, and N. P. Jewell. 2016. *Causal inference in statistics: a primer*. John Wiley & Sons.

[114] Pereyra, M., P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin. 2016. "A survey of stochastic simulation and optimization methods in signal processing". *IEEE Journal of Selected Topics in Signal Processing*. 10(2): 224–241.

[115] Peters, J., D. Janzing, and B. Scholkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press (available on-line).

[116] Pinker, S. 1997. *How the Mind Works*. Penguin Press Science.

[117] Rabiner, L. and B. Juang. 1986. "An introduction to hidden Markov models". *IEEE ASSP magazine*. 3(1): 4–16.

[118] Raginsky, M. 2011. "Directed information and Pearl's causal calculus". In: *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE. 958–965.

[119] Raginsky, M., A. Rakhlin, M. Tsao, Y. Wu, and A. Xu. 2016. "Information-theoretic analysis of stability and bias of learning algorithms". In: *Information Theory Workshop (ITW), 2016 IEEE*. IEEE. 26–30.

[120] Ranganath, R., S. Gerrish, and D. Blei. 2014. "Black box variational inference". In: *Artificial Intelligence and Statistics*. 814–822.

[121] Ranganath, R., L. Tang, L. Charlin, and D. Blei. 2015. "Deep exponential families". In: *Artificial Intelligence and Statistics*. 762–771.

[122] Rezende, D. J., S. Mohamed, and D. Wierstra. 2014. "Stochastic backpropagation and approximate inference in deep generative models". *arXiv preprint arXiv:1401.4082*.

[123] Roth, K., A. Lucchi, S. Nowozin, and T. Hofmann. 2017. "Stabilizing Training of Generative Adversarial Networks through Regularization". *arXiv preprint arXiv:1705.09367*.

[124] Rudolph, M., F. Ruiz, S. Athey, and D. Blei. 2017. "Structured Embedding Models for Grouped Data". *ArXiv e-prints*. Sept. arXiv: 1709.10367 [stat.ML].

[125] Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1988. "Learning representations by back-propagating errors". *Cognitive modeling*. 5(3): 1.

[126] Russel, S. and P. Norvig. 2009. *Artificial Intelligence: A Modern Approach*. Pearson.

[127] Salakhutdinov, R., A. Mnih, and G. Hinton. 2007. "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th international conference on Machine learning*. ACM. 791–798.

[128] Salimans, T., J. Ho, X. Chen, and I. Sutskever. 2017. "Evolution strategies as a scalable alternative to reinforcement learning". *arXiv preprint arXiv:1703.03864*.

[129] Samadi, A., T. P. Lillicrap, and D. B. Tweed. 2017. "Deep Learning with Dynamic Spiking Neurons and Fixed Feedback Weights". *Neural Computation*. 29(3): 578–602.

[130] Scutari, G., F. Facchinei, L. Lampariello, and P. Song. 2014. "Distributed methods for constrained nonconvex multi-agent optimization-part I: theory". *arXiv preprint arXiv:1410.4754*.

[131] Scutari, M. 2017. “Bayesian Dirichlet Bayesian Network Scores and the Maximum Entropy Principle”. *ArXiv e-prints*. arXiv: 1708.00689.

[132] Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. 2016. “Taking the human out of the loop: A review of bayesian optimization”. *Proceedings of the IEEE*. 104(1): 148–175.

[133] Shalev-Shwartz, S. and S. Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

[134] Shannon, C. E. 1948. “A mathematical theory of communication”. *The Bell System Technical Journal*. 27(3): 379–423.

[135] Silver, D. 2015. *Course on reinforcement learning*. URL: http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html.

[136] Smith, S. L., P.-J. Kindermans, and Q. V. Le. 2017. “Don’t Decay the Learning Rate, Increase the Batch Size”. *ArXiv e-prints*. Nov. arXiv: 1711.00489 [cs.LG].

[137] Spectrum, I. *Will the Future of AI Learning Depend More on Nature or Nurture?* URL: https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-and-psychology-researchers-debate-the-future-of-deep-learning.

[138] Stigler, S. M. 2016. *The seven pillars of statistical wisdom*. Harvard University Press.

[139] Subramaniam, S., T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. 2006. “Online outlier detection in sensor data using non-parametric models”. In: *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment. 187–198.

[140] Sugiyama, M., T. Suzuki, and T. Kanamori. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.

[141] Sun, Y., P. Babu, and D. P. Palomar. 2017. “Majorization-minimization algorithms in signal processing, communications, and machine learning”. *IEEE Transactions on Signal Processing*. 65(3): 794–816.

[142] Tegmark, M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Allen Lane.

[143] Thrun, S. 1996. "Is learning the n-th thing any easier than learning the first?" In: *Advances in neural information processing systems*. 640–646.

[144] Times, T. N. Y. 1958. *NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser*. URL: http://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html.

[145] Tishby, N., F. C. Pereira, and W. Bialek. 2000. "The information bottleneck method". *arXiv preprint physics/0004057*.

[146] Tsybakov, A. B. 2009. "Introduction to nonparametric estimation".

[147] Turner, R. E. and M. Sahani. 2011. "Two problems with variational expectation maximisation for time-series models". *Bayesian Time series models*: 115–138.

[148] Uber. *Pyro: Deep universal probabilistic programming*. URL: http://pyro.ai/.

[149] Venkateswara, H., S. Chakraborty, and S. Panchanathan. 2017. "Deep-Learning Systems for Domain Adaptation in Computer Vision: Learning Transferable Feature Representations". *IEEE Signal Processing Magazine*. 34(6): 117–129. ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2740460.

[150] Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. 2010. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion". *Journal of Machine Learning Research*. 11(Dec): 3371–3408.

[151] Wainwright, M. J. and M. I. Jordan. 2008. "Graphical models, exponential families, and variational inference". *Foundations and Trends® in Machine Learning*. 1(1–2): 1–305.

[152] Watt, J., R. Borhani, and A. Katsaggelos. 2016. *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press.

[153] Welling, M., M. Rosen-Zvi, and G. E. Hinton. 2005. “Exponential family harmoniums with an application to information retrieval”. In: *Advances in neural information processing systems*. 1481–1488.

[154] Wikipedia. *AI Winter*. URL: https://en.wikipedia.org/wiki/AI_winter.

[155] Wikipedia. *Conjugate priors*. URL: https://en.wikipedia.org/wiki/Conjugate_prior.

[156] Wikipedia. *Exponential family*. URL: https://en.wikipedia.org/wiki/Exponential_family.

[157] Wilson, A. C., R. Roelofs, M. Stern, N. Srebro, and B. Recht. 2017. “The Marginal Value of Adaptive Gradient Methods in Machine Learning”. *arXiv preprint arXiv:1705.08292*.

[158] Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[159] Zhang, C., J. Butepage, H. Kjellstrom, and S. Mandt. 2017. “Advances in Variational Inference”. *ArXiv e-prints*. Nov. arXiv: 1711.05597 [cs.LG].

[160] Zhang, Y., J. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. “Information-theoretic lower bounds for distributed statistical estimation with communication constraints”. In: *Advances in Neural Information Processing Systems*. 2328–2336.

[161] Zhao, X. and A. H. Sayed. 2015. “Asynchronous adaptation and learning over networks—Part I: Modeling and stability analysis”. *IEEE Transactions on Signal Processing*. 63(4): 811–826.