Learning with Limited Samples: Meta-Learning and Applications to Communication Systems

Other titles in Foundations and Trends® in Signal Processing

An Introduction to Quantum Machine Learning for Engineers
Osvaldo Simeone
ISBN: 978-1-63828-058-3

Bilevel Methods for Image Reconstruction
Caroline Crockett and Jeffrey A. Fessler
ISBN: 978-1-63828-002-6

Operating Characteristics for Classical and Quantum Binary Hypothesis Testing
Catherine A. Medlock and Alan V. Oppenheim
ISBN: 978-1-68083-882-4

Foundations of User-Centric Cell-Free Massive MIMO
Özlem Tugfe Demir, Emil Björnson and Luca Sanguinetti
ISBN: 978-1-68083-790-2

Data-Driven Multi-Microphone Speaker Localization on Manifolds
Bracha Laufer-Goldshtein, Ronen Talmon and Sharon Gannot
ISBN: 978-1-68083-736-0

Learning with Limited Samples: Meta-Learning and Applications to Communication Systems

Lisha Chen
Rensselaer Polytechnic Institute

Sharu Theresa Jose
King's College London

Ivana Nikoloska
King's College London

Sangwoo Park
King's College London

Tianyi Chen
Rensselaer Polytechnic Institute

Osvaldo Simeone
King's College London
osvaldo.simeone@kcl.ac.uk

now
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

L. Chen *et al.*. *Learning with Limited Samples: Meta-Learning and Applications to Communication Systems*. Foundations and Trends® in Signal Processing, vol. 17, no. 2, pp. 79–208, 2023.

ISBN: 978-1-63828-137-5

Foundations and Trends® in Signal Processing

Volume 17, Issue 2, 2023

Editorial Board

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, 2023, Volume 17, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Contents

Learning with Limited Samples: Meta-Learning and Applications to Communication Systems

Lisha Chen[1], Sharu Theresa Jose[2], Ivana Nikoloska[2], Sangwoo Park[2], Tianyi Chen[1] and Osvaldo Simeone[2]

[1]*Rensselaer Polytechnic Institute, USA*
[2]*King's College London, UK*

ABSTRACT

Deep learning has achieved remarkable success in many machine learning tasks such as image classification, speech recognition, and game playing. However, these breakthroughs are often difficult to translate into real-world engineering systems because deep learning models require a massive number of training samples, which are costly to obtain in practice. To address labeled data scarcity, few-shot meta-learning optimizes learning algorithms that can efficiently adapt to new tasks quickly. While meta-learning is gaining significant interest in the machine learning literature, its working principles and theoretic fundamentals are not as well understood in the engineering community.

This review monograph provides an introduction to meta-learning by covering principles, algorithms, theory, and engineering applications. After introducing meta-learning in comparison with conventional and joint learning, we describe the main meta-learning algorithms, as well as a general bilevel optimization framework for the definition of

Lisha Chen, Sharu Theresa Jose, Ivana Nikoloska, Sangwoo Park, Tianyi Chen and Osvaldo Simeone (2023), "Learning with Limited Samples: Meta-Learning and Applications to Communication Systems", Foundations and Trends® in Signal Processing: Vol. 17, No. 2, pp 79–208. DOI: 10.1561/2000000115.

meta-learning techniques. Then, we summarize known results on the generalization capabilities of meta-learning from a statistical learning viewpoint. Applications to communication systems, including decoding and power allocation, are discussed next, followed by an introduction to aspects related to the integration of meta-learning with emerging computing technologies, namely neuromorphic and quantum computing. The monograph is concluded with an overview of open research challenges.

1

Introduction and Background

1.1 Introduction

One of the main principles underlying the design of data-efficient machine learning is **knowledge sharing** across learning tasks. As an example, consider the problem of **few-shot classification**. In it, one is interested in designing a classifier based on few examples for each class. The limited availability of data is typically an insurmountable problem for conventional machine learning solutions, unless one has detailed information about the structure of the problem that can be used to handcraft a well-performing classifier. When such domain knowledge is not available, it may be, however, possible to collect data sets from distinct classification tasks that are deemed to be related to the task of interest. Transferring knowledge from such auxiliary tasks to the target task may compensate for the lack of sufficient data or domain knowledge.

The specific way in which knowledge sharing can be realized depends on the setting of interest and on the availability of data. Central to these distinctions is the notion of a **learning task**. A learning task generally refers to a specific supervised, unsupervised, or reinforcement learning instance characterized by an underlying data-generation distribution and loss or reward function. For instance, a learning task may amount to the problem of classifying images in a number of categories based on labelled examples. With this definition, at a high level, we can distinguish the following methodologies (see, e.g., [160]).

- **Transfer learning**: In transfer learning, one is concerned with two learning tasks – a source task and a target task. Data are typically available for both tasks, although data for the target task may be limited. The goal is to address the target task by utilizing also data from the source task with the aim of reducing data requirements for the target task. In the image classification example, transfer learning would facilitate the optimization of a classifier for a target classification task, e.g., distinguishing images of cats and dogs, using data for another classification task, e.g., distinguishing images of teapots and mugs.

- **Multi-task learning and joint learning**: In multi-task learning, there are $K > 1$ learning tasks, and one is interested in learning a machine learning model that is able to address *all* the tasks based on data pooled from all the tasks. Generally, the machine learning model has some shared components, e.g., layers of a neural network, and also separate parts pertaining each task, e.g., "heads" of a classifier. When the model is fully shared across tasks, multi-task learning is also known as **joint learning**. In the image classification example, multi-task learning would optimize a classifier producing decisions for a set of classification tasks.

- **Meta-learning**: In meta-learning, we have access to data for a number of tasks, but we are not interested in training a machine learning model for them as in multi-task learning. Rather, we would like to use data from multiple tasks in order to design a **training procedure**, and not to produce a single machine learning model. Specifically, the goal is to ensure that the *meta-learned* training procedure can efficiently optimize a machine learning model for *any*, a priori unknown, learning task. Accordingly, in a meta-learning setting, one does not know a priori what the target task will be, although one expects it to be similar to those for which data are available. By optimizing the learning process, meta-learning implements a form of **learning to learn**. In the image classification example, meta-learning would produce a procedure able to optimize a classifier for *any new classification task* by using data from a pool of other similar classification tasks.

This review monograph provides an introduction to meta-learning by covering principles, algorithms, theory, and engineering applications. In this section, we start by providing a first exposition to meta-learning by contrasting it with conventional machine learning and multi-task learning. The section concludes with a description of the organization of the rest of the monograph.

1.2 Meta-Learning

In meta-learning, we target an entire **class of tasks**, also known as the **task environment**, and we wish to "prepare" for any new task that may be encountered from this class. As we will review in this subsection, conventional learning aims at optimizing model parameters, such as the weights of a neural network, by applying a given training algorithm, which is defined by a set of **hyperparameters**. Training algorithms typically involve local search procedures, e.g., based on gradient information, and hyperparameters include the learning rate – i.e., the size of the updates at each iteration – and the initialization. In contrast, the goal of meta-learning is to optimize **hyperparameters** with the goal of identifying a training algorithm that may perform well on new tasks.

1.2.1 Meta-Training and Meta-Testing

The working assumption underlying meta-learning is that, prior to observing the – typically small – training data set for a new task, one has access to a larger data set of examples from related tasks. This is known as the **meta-training data set**. Meta-learning consists of two distinct phases:

- **Meta-training**: Given the meta-training data set, a set of hyperparameters is optimized;

- **Meta-testing**: After the meta-learning phase is completed, data for a target task, known as **meta-test task**, is revealed, and model parameters are optimized using the meta-trained hyperparameters.

As such, the meta-training phase aims at optimizing hyperparameters that enable efficient training on a new, a priori unknown, target task in the meta-testing phase.

1.2.2 Reviewing Conventional Learning

In order to introduce the notation necessary to describe meta-learning, let us briefly review the operation of conventional machine learning.

Training and testing. In conventional machine learning, the starting point is the selection of a model class $\mathcal{H}$ and of a training algorithm. The choice of model class and training algorithm determines the **inductive bias** applied by the learning procedure to generalize from training to test data. The model class $\mathcal{H}$ contains models parameterized by a vector ϕ, such as neural networks. Model class and training algorithm are ideally tailored to information available about the problem of interest.

Furthermore, both model class and training algorithm generally depend on a *fixed* vector of hyperparameters, denoted as θ. Thereafter, hyperparameters may specify, for instance, a mapping defining the vector of features to be used in a linear model, or the initialization and learning rate of an iterative optimizer.

The training algorithm is applied to a training set $\mathcal{D}^{\text{tr}}$, which may include also a separate validation set. The training algorithm produces a model parameter vector ϕ by minimizing the **training loss**

$$L_{\mathcal{D}^{\text{tr}}}(\phi), \tag{1.1}$$

which is obtained by evaluating an empirical average of the loss accrued over the data points in the training set $\mathcal{D}^{\text{tr}}$. Note that regularized versions of the training loss can also be used. Finally, the trained model is tested on a separate test data set $\mathcal{D}^{\text{va}}$ by evaluating the **validation loss** $L_{\mathcal{D}^{\text{va}}}(\phi)$, in which the loss is averaged over the test data in data set $\mathcal{D}^{\text{va}}$. The overall process is summarized in Figure 1.1.

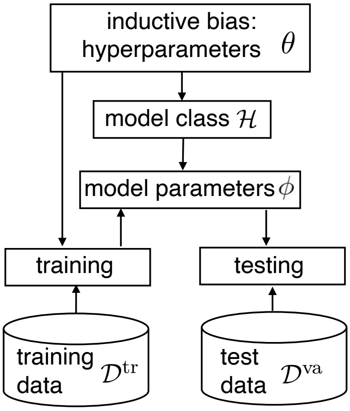
inductive bias:
hyperparameters θ
model class $\mathcal{H}$
model parameters ϕ
training
testing
training data $\mathcal{D}^{\mathrm{tr}}$
test data $\mathcal{D}^{\mathrm{va}}$

Figure 1.1: Illustration of conventional machine learning.

Drawbacks of conventional learning. As anticipated, conventional machine learning suffers from two main potential shortcomings that meta-learning can help address, namely:

- Large **sample complexity**: By training a model "from scratch", conventional learning generally requires a large number of training samples, N, to obtain a suitable test performance. The number of samples needed to obtain some level of accuracy is known as sample complexity.

- Large **iteration complexity**: By relying on a generic optimization procedure, conventional learning may require a large number of iterations to converge to a well-performing model.

Both issues can be potentially mitigated if the inductive bias – i.e., the selection of model class and training algorithm – is tailored to the problem under study based on domain knowledge. For instance, as part of the inductive bias, we may choose an architecture for a neural network model that satisfies known symmetries in the data; or select an initialization point for the model parameters that ϕ is suitably adapted to the learning task at hand. With such informed inductive biases, one we can generally reduce both sample and iteration complexities.

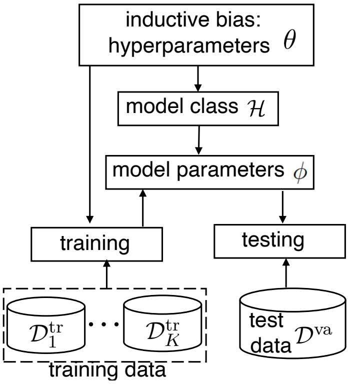
inductive bias:
hyperparameters θ
model class $\mathcal{H}$
model parameters ϕ
training
testing
$\mathcal{D}_1^{\mathrm{tr}}$
$\mathcal{D}_K^{\mathrm{tr}}$
training data
test data $\mathcal{D}^{\mathrm{va}}$

Figure 1.2: Illustration of joint learning.

When one does not have access to sufficient information about the problem to identify a tailored inductive bias, it may become useful to transfer knowledge from data pertaining related tasks.

1.2.3 Joint Learning

Suppose that we have access to training data sets $\mathcal{D}_k^{\mathrm{tr}}$ for a number of distinct learning tasks in the same task environment that are indexed by the integer $k = 1, ..., K$. Each data set $\mathcal{D}_k^{\mathrm{tr}}$ contains N training examples. We now review the idea of joint learning, which is a special case of multi-task learning in which a common model is trained for all K learning tasks.

Training and testing. Joint learning pools together all the training sets $\{\mathcal{D}_k^{\mathrm{tr}}\}_{k=1}^K$, and uses the resulting aggregate training loss

$$L_{\{\mathcal{D}_k^{\mathrm{tr}}\}_{k=1}^K}(\phi) = \frac{1}{K}\sum_{k=1}^{K} L_{\mathcal{D}_k^{\mathrm{tr}}}(\phi) \tag{1.2}$$

as the learning criterion to train a shared model parameter ϕ.

As illustrated in Figure 1.2, joint learning inherently caters only to the K tasks in the original pool, and is hence generally unable to provide desirable performance for new, as of yet unknown, tasks.

Joint learning is a natural first attempt to transfer knowledge across tasks with the aim of improving sample and iteration complexities. First, by pooling together data from K tasks, the overall size of the training set is $K \cdot N$, which may be large even when the available data per task is limited, i.e., when N is small. Second, training only once for K tasks amortizes the iteration complexity across the tasks, yielding a potential reduction of the number of iterations by a factor equal to K.

Drawbacks of joint learning. Joint learning has two potentially critical shortcomings.

- **Bias**: The jointly trained model may improve the performance of conventional learning only if there is a single model parameter ϕ that "works well" for all tasks. This may not be the case if the tasks are sufficiently distinct.
- **Lack of adaptation**: Even if there is a single model parameter ϕ that yields desirable test results on all K tasks, this does not guarantee that the same is true for a new task. In fact, by focusing on training a common model for all tasks, joint learning is not designed to enable adaptation to a new task.

As a remedy for the second shortcoming just highlighted, one could use the jointly trained model parameter ϕ to initialize the training process on a new task – a process known as **fine-tuning**. However, there is generally no guarantee that this would yield a desirable outcome, since the training process used by joint learning does not account for the subsequent step of adaptation on a new task. This is a key distinction between joint learning and meta-learning, which will be introduced next.

1.2.4 Introducing Meta-Learning

As for joint learning, in meta-learning one assumes the availability of data from K related tasks from the same task environment, which are referred to as **meta-training tasks**. However, unlike joint learning, data from these tasks are kept separate, and a distinct model parameter ϕ_k is trained for each k task. As illustrated in Figure 1.3, meta-learning tasks only share a **common hyperparameter vector** θ that is optimized based on meta-training data. As a result, meta-training data is not used to optimize a common model, but only a **shared inductive bias**. In

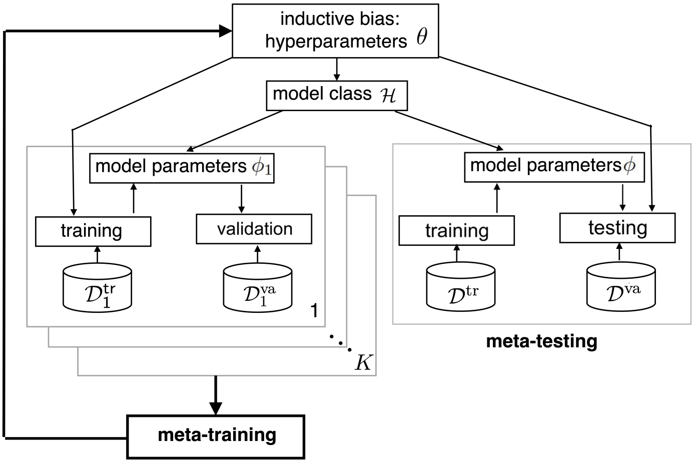
inductive bias:
hyperparameters θ
model class $\mathcal{H}$
model parameters ϕ_1
training
validation
$\mathcal{D}_1^{\text{tr}}$
$\mathcal{D}_1^{\text{va}}$
1
K
model parameters ϕ
training
testing
$\mathcal{D}^{\text{tr}}$
$\mathcal{D}^{\text{va}}$
meta-testing
meta-training

Figure 1.3: Illustration of meta-learning.

other words, the optimization carried out by meta-learning operates at a higher level of abstraction, leaving the model parameters free to adapt to each individual task.

We now introduce meta-learning by emphasizing the differences with respect to joint learning and by detailing the meta-training and meta-testing phases.

Inductive bias and hyperparameters. As discussed, the goal of meta-learning is optimizing the hyperparameter vector θ and, through it, the inductive bias that is applied for the training of each task. To simplify the discussion and focus on the most common setting, let us assume that the model class $\mathcal{H}$ is fixed, while the training algorithm is a mapping $\phi^{\text{tr}}(\mathcal{D}|\theta)$ between a training set $\mathcal{D}$ and a model parameter vector ϕ that depends on the hyperparameter vector θ, i.e.,

$$\phi = \phi^{\text{tr}}(\mathcal{D}|\theta). \tag{1.3}$$

As an example, the training algorithm $\phi^{\text{tr}}(\mathcal{D}|\theta)$ could output the last iterate of an optimizer.

The hyperparameter θ can affect the output $\phi^{\text{tr}}(\mathcal{D}|\theta)$ of the training procedure in different ways. For instance, it can determine the regularization constant; the learning rate and/or the initialization of an iterative training procedure; the mini-batch size; a subset of the parameters in vector ϕ, e.g., used to define a shared feature extractor; the parameters of a prior distribution; and so on.

The output $\phi^{\mathrm{tr}}(\mathcal{D}|\theta)$ of a training algorithm is generally random. This is the case, for instance, if the algorithm relies on stochastic gradient descent (SGD). In the following discussion, we will assume for simplicity a deterministic training algorithm, but the approach carries over directly to the more general case of a random training procedure by adding an average over the randomized of the trained model $\phi^{\mathrm{tr}}(\mathcal{D}|\theta)$.

Meta-training. To formulate meta-training, a natural idea is to use as the optimization criterion the aggregate training loss

$$\mathcal{L}_{\{\mathcal{D}_k^{\mathrm{tr}}\}_{k=1}^K}(\theta) = \frac{1}{K}\sum_{k=1}^{K} L_{\mathcal{D}_k^{\mathrm{tr}}}(\phi^{\mathrm{tr}}(\mathcal{D}_k^{\mathrm{tr}}|\theta)), \tag{1.4}$$

which is a function of the hyperparameter θ. This quantity is known as the **meta-training loss**. The resulting problem

$$\min_{\theta} \mathcal{L}_{\{\mathcal{D}_k^{\mathrm{tr}}\}_{k=1}^K}(\theta) \tag{1.5}$$

of minimizing the meta-training loss over the hyperparameter θ is different from the ERM problem $\min_{\phi} L_{\{\mathcal{D}_k^{\mathrm{tr}}\}_{k=1}^K}(\phi)$ tackled in joint learning for the following reasons:

- First, optimization is over the **hyperparameter** vector θ and not over a shared model parameter ϕ.

- Second, the model parameter ϕ is trained **separately** for each task k through the parallel applications of the training function $\phi^{\mathrm{tr}}(\cdot|\theta)$ to the training set $\mathcal{D}_k^{\mathrm{tr}}$ of each task $k = 1, ..., K$.

As a result of these two key differences with respect to joint training, the minimization of the meta-training loss (1.4) inherently caters for **adaptation**: The hyperparameter vector θ is optimized in such a way that the trained model parameter vectors $\phi_k = \phi^{\mathrm{tr}}(\mathcal{D}_k^{\mathrm{tr}}|\theta)$, adapted separately to the data of each task k, minimize the aggregate loss across all meta-training tasks $k = 1, ..., K$.

Advantages of meta-training over joint training. While retaining the advantages of joint learning in terms of sample and iteration complexity, meta-learning addresses the two shortcomings of joint learning:

- **Knowledge sharing via hyperparameters**: Meta-learning does not assume that there is a single model parameter ϕ that "works well" for all tasks. It only assumes that there exists a common model class and a common training algorithm, as specified by **hyperparameters** θ, that can be effectively applied across the class of tasks of interest.

- **Optimization for adaptation**: Meta-learning prepares the training algorithm $\phi^{\text{tr}}(\mathcal{D}|\theta)$ to **adapt** to potentially new tasks through the selection of the hyperparameters θ. This is because the model parameter vector ϕ is left free by design to be adapted to the training data $\mathcal{D}_k^{\text{tr}}$ of each task k.

Meta-testing. As mentioned, the goal of meta-learning is ensuring generalization to any new task that is drawn at random from the same task environment. For any new task, during the meta-testing phase, we have access to training set $\mathcal{D}^{\text{tr}}$ and validation set $\mathcal{D}^{\text{va}}$. The new task is referred to as the **meta-test task**, and is illustrated in Figure 1.3 along with the meta-training tasks.

The training data $\mathcal{D}^{\text{tr}}$ of the meta-test task is used to adapt the model parameter vector to the meta-test task, obtaining $\phi^{\text{tr}}(\mathcal{D}^{\text{tr}}|\theta)$. Importantly, the training algorithm depends on the hyperparameter θ. The performance metric of interest for a given hyperparameter θ is the test loss for the meta-test task, or **meta-test loss**, given by

$$L_{\mathcal{D}^{\text{va}}}(\phi^{\text{tr}}(\mathcal{D}^{\text{tr}}|\theta)). \tag{1.6}$$

In (1.6), the population loss of the trained model is estimated via the test loss evaluated with the test set $\mathcal{D}^{\text{va}}$.

We have just seen that meta-testing requires a split of the data for the new task into a training part, used for adaptation, and a validation part, used to estimate the population loss (1.6). We now discuss how the idea of splitting per-task data sets into training and validation parts can be useful also during the meta-training phase.

As explained in Section 1.2.4, the training algorithm $\phi(\mathcal{D}^{\text{tr}}|\theta)$ is defined by an optimization procedure for the problem of minimizing the training loss on the training set $\mathcal{D}^{\text{tr}}$. We can write the learning procedure informally as

$$\phi^{\text{tr}}(\mathcal{D}^{\text{tr}}|\theta) \underset{\theta}{\leftarrow} \min_{\phi} L_{\mathcal{D}^{\text{tr}}}(\phi), \tag{1.7}$$

highlighting the dependence of the training algorithm on the training loss $L_{\mathcal{D}^{\text{tr}}}(\phi)$ and on the hyperparameter θ.

Because of (1.7), in problem (1.5) one is effectively optimizing the training losses $L_{\mathcal{D}_k^{\text{tr}}}(\phi)$ for the meta-training tasks $k = 1, ..., K$ twice, first over the model parameters in the inner optimization (1.7) and then over the hyperparameters θ in the outer optimization (1.5). This reuse of the meta-training data for both adaptation and meta-learning may cause overfitting to the meta-training data, and thus result in a training algorithm $\phi^{\text{tr}}(\cdot|\theta)$ that fails to generalize to new tasks.

The problem highlighted above is caused by the fact that the meta-training loss (1.4) does not provide an unbiased estimate of the sum of the population losses across the meta-training tasks. The bias is a consequence of the reuse of the same data for both adaptation and hyperparameter optimization. To address this problem, for each meta-training task k, we can partition the available data into two data sets, a training data set $\mathcal{D}_k^{\text{tr}}$ and a validation data set $\mathcal{D}_k^{\text{va}}$. Therefore, the overall meta-training data set is given as $\mathcal{D}^{\text{mtr}} = \{(\mathcal{D}_k^{\text{tr}}, \mathcal{D}_k^{\text{va}})_{k=1}^{K}\}$.

The key idea is that the training data set $\mathcal{D}_k^{\text{tr}}$ is used for adaptation using the training algorithm (1.7), while the test data set $\mathcal{D}_k^{\text{va}}$ is kept aside to estimate the population distribution of task k for the trained model. The hyperparameter θ is not optimized to minimize the sum of the training losses as in (1.5). Rather, they target the sum of the test losses, which provides an unbiased estimate of the corresponding sum of population losses.

Meta-learning as nested optimization. To summarize, the general procedure followed by many meta-learning algorithms consists of a nested optimization of the following form:

- **Inner loop**: For a fixed hyperparameter vector θ, training on each task k is done separately, producing per-task model parameters

$$\phi_k = \phi^{\text{tr}}(\mathcal{D}_k^{\text{tr}}|\theta) \underset{\theta}{\leftarrow} \min_{\phi} L_{\mathcal{D}_k^{\text{tr}}}(\phi) \tag{1.8}$$

  for $k = 1, ..., K$;

- **Outer loop**: The hyperparameter vector θ is optimized as

$$\theta_{\mathcal{D}^{\text{mtr}}} = \arg\min_{\theta} \mathcal{L}_{\mathcal{D}^{\text{mtr}}}(\theta), \tag{1.9}$$

where the **meta-training loss** is (re-)defined as

$$\mathcal{L}_{\mathcal{D}^{\text{mtr}}}(\theta) = \frac{1}{K}\sum_{k=1}^{K} L_{\mathcal{D}_k^{\text{va}}}(\phi^{\text{tr}}(\mathcal{D}_k^{\text{tr}}|\theta)). \tag{1.10}$$

As we will detail in Section 2, the specific implementation of a meta-learning algorithm depends on the selection of the training algorithm $\phi^{\text{tr}}(\mathcal{D}|\theta)$ and on the method used to solve the outer optimization.

1.2.5 Meta-Inductive Bias

While the inductive bias underlying the training algorithm used in the inner loop is optimized by means of meta-learning, the meta-learning process itself assumes a **meta-inductive bias**. The meta-inductive bias encompasses the choices of the hyperparameters to optimize in the outer loop – e.g., the initialization of an SGD training algorithm – as well as the optimization algorithm used in the outer loop. There is of course no end to this nesting of inductive biases: any new learning level brings its own assumptions and biases. Meta-learning moves the potential cause of bias at the outer level of the meta-learning loop, which may improve the efficiency of training.

It is important, however, to note that the selection of a meta-inductive bias may cause **meta-overfitting** in a similar way as the choice of an inductive bias can cause overfitting in conventional learning. In a nutshell, if the meta-inductive bias is too broad and the number of tasks insufficient, the meta-trained inductive bias may overfit the meta-training data and fail to prepare for adaptation to new tasks.

1.3 Organization of the Monograph

The rest of the monograph is organized as follows.

Section 2. Meta-Learning Algorithms: This section provides a taxonomy and an introduction to the most common meta-learning algorithms, including model agnostic meta-learning (MAML).

Section 3. Bilevel Optimization for Meta-Learning: This section presents a general optimization-based perspective on meta-learning, which views meta-learning as a form of stochastic bilevel optimization.

Section 4. Statistical Learning Theory for Meta-learning: This section revisits meta-learning through the different perspective of generalization. Specifically, it investigates from a theoretical viewpoint the performance of meta-learning algorithms in terms of their capacity to generalize outside the meta-training data set to new tasks.

Section 5. Meta-Learning Applications to Communications: This section turns to several examples of applications of meta-learning to the engineering problem of designing communication systems. Examples of reviewed applications include demodulation and power control.

Section 6. Integration with Emerging Computing Technologies: This section highlights the potential synergies between meta-learning and two emerging computing technologies, namely neuromorphic and quantum computing.

Section 7. Outlook: The last section presents an outlook on the area of meta-learning by offering a brief review of open problems and further directions for reading and research.

References

[1] 3GPP, "Enhancement for Unmanned Aerial Vehicles," *TS 22.289 V17.1.0*, 2019.

[2] 3GPP, "Study on channel model for frequencies from 0.5 to 100 ghz (3gpp tr 38.901 version 16.1.0 release 16)," *TR 38.901*, 2020.

[3] M. Abbas, Q. Xiao, L. Chen, P.-Y. Chen, and T. Chen, "Sharp-MAML: Sharpness-aware model-agnostic meta learning," in *Proc. Intl. Conf. on Machine Learning*, 2022.

[4] A. Abdi and M. Kaveh, "A space-time correlation model for multielement antenna systems in mobile fading channels," *IEEE Journal on Selected Areas in communications*, vol. 20, no. 3, 2002, pp. 550–560.

[5] A. Agrawal, J. G. Andrews, J. M. Cioffi, and T. Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Transactions on wireless communications*, vol. 4, no. 3, 2005, pp. 878–884.

[6] F. Alet, T. Lozano-Pérez, and L. P. Kaelbling, "Modular meta-learning," in *Proc. Conference on Robot Learning*, pp. 856–868, 2018.

[7] F. Alet, E. Weng, T. Lozano-Pérez, and L. P. Kaelbling, "Neural relational inference with fast modular meta-learning," in *Proc. Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] P. Alquier, "User-friendly introduction to PAC-Bayes bounds," *arXiv preprint arXiv: 2110.11216*, 2021.

[9] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine learning*, vol. 50, no. 1, 2003, pp. 5–43.

[10] F. A. Aoudia and J. Hoydis, "Model-free training of end-to-end communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, 2019, pp. 2503–2516.

[11] F. A. Aoudia and J. Hoydis, "End-to-end learning for ofdm: From neural receivers to pilotless communication," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, 2021.

[12] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," in *Proc. Advances in Neural Information Processing Systems*, 2019.

[13] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, 2019, pp. 505–510.

[14] Y. Bai, M. Chen, P. Zhou, T. Zhao, J. Lee, S. Kakade, H. Wang, and C. Xiong, "How important is the train-validation split in meta-learning?" In *Proc. Intl. Conf. on Machine Learning*, pp. 543–553, 2021.

[15] J. F. Bard, *Practical bilevel optimization: algorithms and applications*, vol. 30. Springer Science & Business Media, 2013.

[16] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, 2020, pp. 30 063–30 070.

[17] J. Baxter, "Theoretical models of learning to learn," in *Learning to learn*, Springer, 1998, pp. 71–94.

[18] P. Benioff, "Quantum mechanical hamiltonian models of turing machines," *Journal of Statistical Physics*, vol. 29, no. 3, 1982, pp. 515–546.

[19] G. Berseth, Z. Zhang, G. Zhang, C. Finn, and S. Levine, "Comps: Continual meta policy search," in *Proc. Intl. Conf. on Learning Representations*, 2021.

[20] H. Beyer and H. Schwefel, "Evolution strategies - a comprehensive introduction," *Natural Computing*, vol. 1, no. 1, 2002, pp. 3–52.

[21] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, 4. Springer, 2006.

[22] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM*, vol. 36, no. 4, 1989, pp. 929–965.

[23] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in o-ran for data-driven nextg cellular networks," *IEEE Communications Magazine*, vol. 59, no. 10, 2021, pp. 21–27.

[24] Z. Borsos, M. Mutny, and A. Krause, "Coresets via bilevel optimization for continual learning and streaming," in *Proc. Advances in Neural Information Processing Systems*, 2020.

[25] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, 2019, pp. 567–579.

[26] O. Bousquet, "New approaches to statistical learning theory," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 2, 2003, pp. 371–389.

[27] J. Bracken and J. T. McGill, "Mathematical programs with optimization problems in the constraints," *Operations Research*, vol. 21, no. 1, 1973, pp. 37–44.

[28] S. Cammerer, F. A. Aoudia, S. Dörner, M. Stark, J. Hoydis, and S. Ten Brink, "Trainable communication systems: Concepts and prototype," *IEEE Transactions on Communications*, vol. 68, no. 9, 2020, pp. 5489–5503.

[29] L. Chen and T. Chen, "Is Bayesian model-agnostic meta learning better than model-agnostic meta learning, provably?" In *Proc. Intl. Conf. on Artificial Intelligence and Statistics*, pp. 1733–1774, 2022.

[30] L. Chen, S. Lu, and T. Chen, "Understanding benign overfitting in gradient-based meta learning," in *Proc. Advances in Neural Information Processing Systems*, 2022.

[31] T. Chen, Y. Sun, and W. Yin, "Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems," in *Proc. Advances in Neural Information Processing Systems*, vol. 34, 2021.

[32] T. Chen, Y. Sun, Q. Xiao, and W. Yin, "A single-timescale method for stochastic bilevel optimization," in *Proc. Intl. Conf. on Artificial Intelligence and Statistics*, vol. 151, pp. 2466–2488, 2022.

[33] T. Chen, Y. Sun, and W. Yin, "Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization," *IEEE Transactions on Signal Processing*, vol. 69, 2021, pp. 4937–4948.

[34] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Intl. Conf. on Learning Representations*, 2018.

[35] K. Chua, Q. Lei, and J. D. Lee, "How fine-tuning allows for effective meta-learning," in *Proc. Advances in Neural Information Processing Systems*, vol. 34, 2021.

[36] M. Cicerone, O. Simeone, and U. Spagnolini, "Channel estimation for mimo-ofdm systems by modal analysis/filtering," *IEEE Transactions on Communications*, vol. 54, no. 11, 2006, pp. 2062–2074.

[37] K. M. Cohen, S. Park, O. Simeone, and S. Shamai, "Learning to learn to demodulate with uncertainty quantification via bayesian meta-learning," in *International ITG Workshop on Smart Antennas*, pp. 1–6, 2021.

[38] K. M. Cohen, S. Park, O. Simeone, and S. Shamai, "Towards reliable and efficient ai for 6g: Bayesian active meta-learning for few pilot demodulation and equalization," *arXiv preprint arXiv: 2108.00785*, 2021.

[39] L. Collins, A. Mokhtari, and S. Shakkottai, "Why does MAML outperform ERM? An optimization perspective," *arXiv preprint: 2010.14672*, 2020.

[40] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, 2007, pp. 235–256.

[41] T. M. Cover, *Elements of information theory.* John Wiley & Sons, 1999.

[42] M. Davies, A. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risbud, "Advancing neuromorphic computing with loihi: A survey of results and outlook," *Proceedings of the IEEE*, vol. 109, no. 5, 2021, pp. 911–934.

[43] C. L. Degen, F. Reinhard, and P. Cappellaro, "Quantum sensing," *Reviews of modern physics*, vol. 89, no. 3, 2017, p. 035 002.

[44] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," in *Proc. Intl. Conf. on Machine Learning*, pp. 179–186, 2012.

[45] S. Dempe, J. Dutta, and B. S. Mordukhovich, "New necessary optimality conditions in optimistic bilevel programming," *Optimization*, vol. 56, no. 5-6, 2007, pp. 577–604.

[46] S. Dempe, V. Kalashnikov, G. A. Perez-Valdes, and N. Kalashnykova, *Bilevel Programming Problems: Theory, Algorithms and Applications to Energy Networks*, vol. 10. Springer, 2015.

[47] S. Dempe, B. S. Mordukhovich, and A. B. Zemkoho, "Necessary optimality conditions in pessimistic bilevel programming," *Optimization*, vol. 63, no. 4, 2014, pp. 505–533.

[48] S. Dempe and A. Zemkoho, *Bilevel Optimization.* Springer, 2020.

[49] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil, "Learning to learn around a common mean," in *Proc. Advances in Neural Information Processing Systems*, vol. 31, 2018.

[50] L. Deng, G. Wu, J. Fu, Y. Zhang, and Y. Yang, "Joint resource allocation and trajectory control for uav-enabled vehicular communications," *IEEE Access*, vol. 7, 2019, pp. 132 806–132 815.

[51] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *Intl. Conf. on Learning Representations*, 2020.

[52] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "Rl^2: Fast reinforcement learning via slow reinforcement learning," *arXiv preprint arXiv: 1611.02779*, 2016.

[53] M. Eisen and A. Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks," *IEEE Transactions on Signal Processing*, vol. 68, 2020, pp. 2977–2991.

[54] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proc. Intl. Conf. on Machine Learning*, 2017.

[55] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online meta-learning,” in *Proc. Intl. Conf. on Machine Learning*, pp. 1920–1930, 2019.

[56] C. Finn, K. Xu, and S. Levine, “Probabilistic model-agnostic meta-learning,” in *Proc. Advances in Neural Information Processing Systems*, 2018.

[57] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-aware minimization for efficiently improving generalization,” in *Proc. Intl. Conf. on Learning Representations*, 2020.

[58] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, “Forward and reverse gradient-based hyperparameter optimization,” in *Proc. Intl. Conf. on Machine Learning*, pp. 1165–1173, 2017.

[59] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, “Bilevel programming for hyperparameter optimization and meta-learning,” in *Proc. Intl. Conf. on Machine Learning*, pp. 1568–1577, 2018.

[60] S. Frei, N. S. Chatterji, and P. L. Bartlett, “Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data,” *arXiv preprint arXiv: 2202.05928*, 2022.

[61] F. Futami, T. Iwata, N. Ueda, I. Sato, and M. Sugiyama, “Excess risk analysis for epistemic uncertainty with application to variational inference,” *arXiv preprint arXiv: 2206.01606*, 2022.

[62] K. Gao and O. Sener, “Modeling and optimization trade-off in meta-learning,” in *Proc. Advances in Neural Information Processing Systems*, vol. 33, 2020.

[63] S. Ghadimi and G. Lan, “Stochastic first-and zeroth-order methods for nonconvex stochastic programming,” *SIAM Journal on Optimization*, vol. 23, no. 4, 2013, pp. 2341–2368.

[64] S. Ghadimi and M. Wang, “Approximation methods for bilevel programming,” *arXiv preprint arXiv: 1802.02246*, 2018.

[65] S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[66] M. Goutay, F. A. Aoudia, and J. Hoydis, “Deep hypernetwork-based mimo detection,” in *Proc. International Workshop on Signal Processing Advances in Wireless Communications*, 2020.

[67] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, “Recasting gradient-based meta-learning as hierarchical bayes,” in *Proc. Intl. Conf. on Learning Representations*, 2018.

[68] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo, “On the iteration complexity of hypergradient computation,” in *Proc. Intl. Conf. on Machine Learning*, pp. 3748–3758, 2020.

[69] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. Intl. Conf. on Machine Learning*, 2017.

[70] Z. Guo and T. Yang, “Randomized stochastic variance-reduced methods for stochastic bilevel optimization,” *arXiv preprint: 2105.02266*, 2021.

[71] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[72] S. Hochreiter, A. S. Younger, and P. R. Conwell, “Learning to learn using gradient descent,” in *Proc. Intl. Conf. on Artificial Neural Networks*, pp. 87–94, 2001.

[73] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, “A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic,” *arXiv preprint:2007.05170*, 2020.

[74] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, “Meta-learning in neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[75] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, “Distributed multi-agent meta learning for trajectory design in wireless drone networks,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, 2021, pp. 3177–3192.

[76] Y. Hu, S. Zhang, X. Chen, and N. He, "Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning," in *Proc. Advances in Neural Information Processing Systems*, pp. 2759–2770, 2020.

[77] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck, "Dvs benchmark datasets for object tracking, action recognition, and object recognition," *Frontiers in neuroscience*, vol. 10, 2016, p. 405.

[78] F. Huang and H. Huang, "Biadam: Fast adaptive bilevel optimization methods," *arXiv preprint:2106.11396*, 2021.

[79] Y. Huang, Y. Liang, and L. Huang, "Provable generalization of overparameterized meta-learning trained with sgd," in *Proc. Advances in Neural Information Processing Systems*, 2022.

[80] E. Hüllermeier, "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?" *arXiv preprint:2209.03302*, 2022.

[81] M. Humphries, *The Spike: An Epic Journey Through the Brain in 2.1 Seconds*. Princeton University Press, 2021.

[82] H. Jang, O. Simeone, B. Gardner, and A. Gruning, "An introduction to probabilistic spiking neural networks: Probabilistic models, learning rules, and applications," *IEEE Signal Processing Magazine*, vol. 36, no. 6, 2019, pp. 64–77.

[83] K. Ji, J. Yang, and Y. Liang, "Multi-step model-agnostic meta-learning: Convergence and improved algorithms," *arXiv preprint arXiv: 2002.07836*, 2020.

[84] K. Ji, J. Yang, and Y. Liang, "Provably faster algorithms for bilevel optimization and applications to meta-learning," in *Proc. Intl. Conf. on Machine Learning*, 2021.

[85] K. Ji, J. Yang, and Y. Liang, "Theoretical convergence of multi-step model-agnostic meta-learning.," *Journal of Machine Learning Research*, vol. 23, 2022, pp. 29–1.

[86] W. Jiang and H. D. Schotten, "A comparison of wireless channel predictors: Artificial intelligence versus kalman filter," in *Proc. Intl. Conf. on Communications*, pp. 1–6, 2019.

[87] W. Jiang, M. Strufe, and H. D. Schotten, "Long-range mimo channel prediction using recurrent neural networks," in *Proc. IEEE Annual Consumer Communications & Networking Conference*, pp. 1–6, 2020.

[88] Y. Jiang, H. Kim, H. Asnani, and S. Kannan, "Mind: Model independent neural decoder," in *Proc. International Workshop on Signal Processing Advances in Wireless Communications*, pp. 1–5, 2019.

[89] S. T. Jose, S. Park, and O. Simeone, "Information-theoretic analysis of epistemic uncertainty in bayesian meta-learning," in *Proc. Intl. Conf. on Artificial Intelligence and Statistics*, pp. 9758–9775, 2022.

[90] S. T. Jose and O. Simeone, "An information-theoretic analysis of the impact of task similarity on meta-learning," in *Proc. IEEE International Symposium on Information Theory*, pp. 1534–1539, 2021.

[91] S. T. Jose and O. Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization," *IEEE Signal Processing Magazine*, vol. 38, no. 2, 2021, pp. 120–125.

[92] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, 2021.

[93] S. T. Jose, O. Simeone, and G. Durisi, "Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization," *IEEE Transactions on Information Theory*, vol. 68, no. 1, 2021, pp. 474–501.

[94] J. Kaddour, S. Sæmundsson, *et al.*, "Probabilistic active meta-learning," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 20 813–20 822, 2020.

[95] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, *et al.*, "Model based reinforcement learning for atari," in *Proc. Intl. Conf. on Learning Representations*, 2019.

[96] A. Karni, G. Meyer, C. Rey-Hipolito, P. Jezzard, M. M. Adams, R. Turner, and L. G. Ungerleider, "The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex," *Proceedings of the National Academy of Sciences*, vol. 95, no. 3, 1998, pp. 861–868.

[97] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, "A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization," in *Proc. Advances in Neural Information Processing Systems*, 2021.

[98] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive mimo channel prediction: Kalman filtering vs. machine learning," *IEEE Transactions on Communications*, vol. 69, no. 1, 2020, pp. 518–528.

[99] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized variational inference: Three arguments for deriving new posteriors," *arXiv preprint arXiv: 1904.02063*, 2019.

[100] V. Konda and V. Borkar, "Actor-critic-type learning algorithms for markov decision processes," *SIAM Journal on Control and Optimization*, vol. 38, no. 1, 1999, pp. 94–123.

[101] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh, "Meta-learning for mixed linear regression," in *Proc. Intl. Conf. on Machine Learning*, pp. 5394–5404, 2020.

[102] D. Kudithipudi, M. Aguilar-Simon, J. Babb, M. Bazhenov, D. Blackiston, J. Bongard, A. P. Brna, S. Chakravarthi Raja, N. Cheney, J. Clune, *et al.*, "Biological underpinnings for lifelong learning machines," *Nature Machine Intelligence*, vol. 4, 2022.

[103] G. Kunapuli, K. P. Bennett, J. Hu, and J.-S. Pang, "Classification model selection via bilevel programming," *Optimization Methods & Software*, vol. 23, no. 4, 2008, pp. 475–489.

[104] K. Kunisch and T. Pock, "A bilevel optimization approach for parameter learning in variational models," *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, 2013, pp. 938–983.

[105] W. W. Lee, Y. J. Tan, H. Yao, S. Li, H. H. See, M. Hon, K. A. Ng, B. Xiong, J. S. Ho, and B. C. Tee, "A neuro-inspired artificial peripheral nervous system for scalable electronic skins," *Science Robotics*, vol. 4, no. 32, 2019.

[106] J. Li, B. Gu, and H. Huang, "A fully single loop algorithm for bilevel optimization without hessian inverse," in *Proc. Association for the Advancement of Artificial Intelligence*, pp. 7426–7434, 2022.

[107] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change," in *Proc. IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pp. 2060–2069, 2006.

[108] G. Lindsay, *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain.* Bloomsbury Publishing, 2021.

[109] J. Liu, Y. Fan, Z. Chen, and Y. Zheng, "Pessimistic bilevel optimization: A survey," *International Journal of Computational Intelligence Systems*, vol. 11, no. 1, 2018, pp. 725–736.

[110] J. Liu, Y. Fan, Z. Chen, and Y. Zheng, "Methods for pessimistic bilevel optimization," in *Bilevel Optimization*, Springer, 2020, pp. 403–420.

[111] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," in *Proc. Advances in Neural Information Processing Systems*, 2016.

[112] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[113] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang, "A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton," in *Proc. Intl. Conf. on Machine Learning*, pp. 6305–6315, 2020.

[114] W. Liu, L.-L. Yang, and L. Hanzo, "Recurrent neural network based narrowband channel prediction," in *Proc. IEEE 63rd Vehicular Technology Conference*, vol. 5, pp. 2173–2177, 2006.

[115] Y. Liu and O. Simeone, "Learning how to transfer from uplink to downlink via hyper-recurrent neural network for fdd massive mimo," *IEEE Transactions on Wireless Communications*, 2022.

[116] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints.* Cambridge University Press, 1996.

[117] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proc. Intl. Conf. on Machine Learning*, F. Bach and D. Blei, Eds., vol. 37, pp. 2113–2122, 2015.

[118] A. Mancoo, S. Keemink, and C. K. Machens, "Understanding spiking networks through convex optimization," in *Proc. Advances in Neural Information Processing Systems*, 2020.

[119] R. Marini, S. Park, O. Simeone, and C. Buratti, "Continual meta-reinforcement learning for uav-aided vehicular wireless networks," *arXiv preprint arXiv: 2207.06131*, 2022.

[120] S. J. Martin, P. D. Grimwood, and R. G. Morris, "Synaptic plasticity and memory: An evaluation of the hypothesis," *Annual review of neuroscience*, vol. 23, no. 1, 2000, pp. 649–711.

[121] A. Masegosa, "Learning under model misspecification: Applications to variational and ensemble methods," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 5479–5491, 2020.

[122] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Intl. Conf. on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[123] A. Mehonic and A. J. Kenyon, "Brain-inspired computing needs a master plan," *Nature*, vol. 604, no. 7905, 2022, pp. 255–260.

[124] R. Mendonca, A. Gupta, R. Kralev, P. Abbeel, S. Levine, and C. Finn, "Guided meta-policy search," in *Proc. Advances in Neural Information Processing Systems*, 2019.

[125] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. Intl. Conf. on Learning Representations*, 2018.

[126] A. Nagabandi, I. Clavera, S. Liu, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," *arXiv preprint arXiv: 1803.11347*, 2018.

[127] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, 2019, pp. 51–63.

[128] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," *arXiv preprint arXiv: 1412.6614*, 2014.

[129] C. Nguyen, T.-T. Do, and G. Carneiro, "Uncertainty in model-agnostic meta-learning using variational inference," in *Proc. Winter Conference on Applications of Computer Vision*, pp. 3090–3100, 2020.

[130] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv: 1803.02999*, 2018.

[131] A. Nichol and J. Schulman, "Reptile: A scalable meta learning algorithm," *arXiv preprint arXiv: 1803.02999*, 2018.

[132] I. Nikoloska and O. Simeone, "Bayesian active meta-learning for black-box optimization," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications*, 2022.

[133] I. Nikoloska and O. Simeone, "Modular meta-learning for power control via random edge graph neural networks," *IEEE Transactions on Wireless Communications*, 2022.

[134] I. Nikoloska and O. Simeone, "Quantum-aided meta-learning for bayesian binary neural networks via Born machines," *arXiv preprint arXiv: 2203.17089*, 2022.

[135] T. O'shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, 2017, pp. 563–575.

[136] S. Park, H. Jang, O. Simeone, and J. Kang, "Learning to demodulate from few pilots via offline and online meta-learning," *IEEE Transactions on Signal Processing*, vol. 69, 2020, pp. 226–239.

[137] S. Park and O. Simeone, "Predicting flat-fading channels via meta-learned closed-form linear filters and equilibrium propagation," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 8817–8821, 2022.

[138] S. Park, O. Simeone, and J. Kang, "End-to-end fast training of communication links without a channel model via online meta-learning," in *Proc. International Workshop on Signal Processing Advances in Wireless Communications*, 2020.

[139] S. Park, O. Simeone, and J. Kang, "Meta-learning to communicate: Fast end-to-end training for fading channels," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 5075–5079, 2020.

[140] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *Proc. Intl. Conf. on Machine Learning*, pp. 737–746, 2016.

[141] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. Intl. Conf. on Machine Learning*, pp. 5171–5180, 2019.

[142] K. Pratik, R. A. Amjad, A. Behboodi, J. B. Soriaga, and M. Welling, "Neural augmentation of kalman filter with hypernetwork for channel tracking," in *Proc. IEEE Global Communications Conference*, pp. 1–6, 2021.

[143] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 7229–7238, 2018.

[144] M. Rabinovich, E. Angelino, and M. I. Jordan, "Variational consensus monte carlo," in *Proc. Advances in Neural Information Processing Systems*, vol. 28, 2015.

[145] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, "Meta-learning with implicit gradients," in *Proc. Advances in Neural Information Processing Systems*, 2019.

[146] K. Rakelly, A. Zhou, C. Finn, S. Levine, and D. Quillen, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," in *Proc. Intl. Conf. on Machine Learning*, pp. 5331–5340, 2019.

[147] S. Ravi and A. Beatson, "Amortized bayesian meta-learning," in *Proc. Intl. Conf. on Learning Representations*, 2019.

[148] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, "Online meta-learning for hybrid model-based deep receivers," *arXiv preprint arXiv: 2203.14359*, 2022.

[149] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, 1951, pp. 400–407.

[150] B. Rosenfeld, B. Rajendran, and O. Simeone, "Fast on-device adaptation for spiking neural networks via online-within-online meta-learning," in *Proc. IEEE Data Science and Learning Workshop*, pp. 1–6, 2021.

[151] S. Sabach and S. Shtern, "A first order method for solving convex bilevel optimization problems," *SIAM Journal on Optimization*, vol. 27, no. 2, 2017, pp. 640–660.

[152] J. Schmidhuber, "A neural network that embeds its own meta-levels," in *Proc. IEEE Intl. Conf. on Neural Networks*, 407–412 vol.1, 1993.

[153] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots, "Truncated back-propagation for bilevel optimization," in *Proc. Intl. Conf. on Artificial Intelligence and Statistics*, pp. 1723–1732, 2019.

[154] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[155] H. Shen and T. Chen, "A single-timescale analysis for stochastic approximation with multiple coupled sequences," in *Proc. Advances in Neural Information Processing Systems*, 2022.

[156] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, "Model-based deep learning: On the intersection of deep learning and optimization," *Proceedings of the National Academy of Sciences of the United States of America*, 2022.

[157] N. Shlezinger, N. Farsad, Y. C. Eldar, and A. J. Goldsmith, "Viterbinet: A deep learning based viterbi algorithm for symbol detection," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, 2020, pp. 3319–3331.

[158] N. Shlezinger, R. Fu, and Y. C. Eldar, "Deepsic: Deep soft interference cancellation for multiuser mimo detection," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, 2020, pp. 1349–1362.

[159] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, 2018, pp. 648–664.

[160] O. Simeone, *Machine Learning for Engineers*. Cambridge University Press, 2022.

[161] O. Simeone, S. Park, and J. Kang, "From learning to meta-learning: Reduced training overhead and complexity for communication systems," in *6G Wireless Summit*, pp. 1–5, 2020.

[162] O. Simeone and U. Spagnolini, "Lower bound on training-based channel estimation error for frequency-selective block-fading rayleigh mimo channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 11, 2004, pp. 3265–3277.

[163] A. Sinha, P. Malo, and K. Deb, "A review on bilevel optimization: From classical to evolutionary approaches and applications," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, 2017, pp. 276–295.

[164] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Advances in Neural Information Processing Systems*, pp. 4080–4090, 2017.

[165] X. Song, W. Gao, Y. Yang, K. Choromanski, A. Pacchiano, and Y. Tang, "Es-maml: Simple hessian-free meta learning," in *Proc. Intl. Conf. on Learning Representations*, 2019.

[166] N. Soures, P. Helfer, A. Daram, T. Pandit, and D. Kudithipudi, "Tacos: Task agnostic continual learning in spiking neural networks," in *Proc. Intl. Conf. on Machine Learning*, 2021.

[167] D. Sow, K. Ji, Z. Guan, and Y. Liang, "A constrained optimization approach to bilevel optimization with multiple inner minima," *arXiv preprint arXiv: 2203.01123*, 2022.

[168] H. V. Stackelberg, *The Theory of Market Economy*. Oxford University Press, 1952.

[169] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[170] R. S. Sutton, M. H. Bowling, and P. M. Pilarski, "The Alberta plan for AI research," *arXiv preprint arXiv: 2208.11173*, 2022.

[171] D. Tandur and M. Moonen, "Joint adaptive compensation of transmitter and receiver iq imbalance under carrier frequency offset in ofdm-based systems," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, 2007, pp. 5246–5252.

[172] D. A. Tarzanagh and L. Balzano, "Online bilevel optimization: Regret analysis of online alternating gradient methods," *arXiv preprint:2207.02829*, 2022.

[173] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *arXiv preprint arXiv: 2009.14286*, 2020.

[174] G. Verdon, M. Broughton, J. R. McClean, K. J. Sung, R. Babbush, Z. Jiang, H. Neven, and M. Mohseni, "Learning to learn with quantum neural networks via classical neural networks," *arXiv preprint arXiv: 1907.05415*, 2019.

[175] L. N. Vicente and P. H. Calamai, "Bilevel and multilevel programming: A bibliography review," *Journal of Global optimization*, vol. 5, no. 3, 1994, pp. 291–306.

[176] P. Vicol, J. P. Lorraine, F. Pedregosa, D. Duvenaud, and R. B. Grosse, "On implicit bias in overparameterized bilevel optimization," in *Proc. Intl. Conf. on Machine Learning*, pp. 22 234–22 259, 2022.

[177] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," in *Proc. Advances in Neural Information Processing Systems*, vol. 29, pp. 3630–3638, 2016.

[178] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, 1967, pp. 260–269.

[179] K. Wang, V. Muthukumar, and C. Thrampoulidis, "Benign overfitting in multiclass classification: All roads lead to interpolation," in *Proc. Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[180] Z. Wang, Y. Zhao, P. Yu, R. Zhang, and C. Chen, "Bayesian meta sampling for fast uncertainty adaptation," in *Proc. Intl. Conf. on Learning Representations*, 2020.

[181] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, 1985, pp. 387–392.

[182] Wikipedia, *Heinrich freiherr von stackelberg*, 2013. URL: https://en.wikipedia.org/wiki/Heinrich_Freiherr_von_Stackelberg.

[183] M. Wilson, R. Stromswold, F. Wudarski, S. Hadfield, N. M. Tubman, and E. G. Rieffel, "Optimizing quantum heuristics with meta-learning," *Quantum Machine Intelligence*, vol. 3, no. 1, 2021, pp. 1–14.

[184] J. Xia, D. Deng, and D. Fan, "A note on implementation methodologies of deep learning-based signal detection for conventional mimo transmitters," *IEEE Transactions on Broadcasting*, vol. 66, no. 3, 2020, pp. 744–745.

[185] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[186] J. Yang, K. Ji, and Y. Liang, "Provably faster algorithms for bilevel optimization," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 13 670–13 682.

[187] J. Ye and D. Zhu, "Optimality conditions for bilevel programming problems," *Optimization*, vol. 33, no. 1, 1995, pp. 9–27.

[188] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, "Meta-learning without memorization," in *Proc. Intl. Conf. on Learning Representations*, 2020.

[189] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, "Bayesian model-agnostic meta-learning," in *Proc. Advances in Neural Information Processing Systems*, 2018.

[190] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma, "Mopo: Model-based offline policy optimization," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 14 129–14 142, 2020.

[191] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive mimo with channel aging," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, 2020, pp. 2960–2973.

[192] M. Zecchin, S. Park, O. Simeone, M. Kountouris, and D. Gesbert, “Robust bayesian learning for reliable wireless ai: Framework and applications,” *arXiv preprint arXiv: 2207.00300*, 2022.

[193] J. Zhang, Y. Yuan, G. Zheng, I. Krikidis, and K.-K. Wong, “Embedding model based fast meta learning for downlink beamforming adaptation,” *IEEE Transactions on Wireless Communications*, 2021.

[194] T. Zhang, “Information-theoretic upper and lower bounds for statistical estimation,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, 2006, pp. 1307–1321.

[195] P. Zhou, X. Yuan, H. Xu, S. Yan, and J. Feng, “Efficient meta learning via minibatch proximal update,” in *Proc. Advances in Neural Information Processing Systems*, 2019.

[196] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, “Fast context adaptation via meta-learning,” in *Proc. Intl. Conf. on Machine Learning*, pp. 7693–7702, 2019.