Bilevel Programming for Hyperparameter Optimization and Meta-Learning

Luca Franceschi [1 2]   **Paolo Frasconi** [3]   **Saverio Salzo** [1]   **Riccardo Grazzi** [1]   **Massimiliano Pontil** [1 2]

Abstract

We introduce a framework based on bilevel programming that unifies gradient-based hyperparameter optimization and meta-learning. We show that an approximate version of the bilevel problem can be solved by taking into explicit account the optimization dynamics for the inner objective. Depending on the specific setting, the outer variables take either the meaning of hyperparameters in a supervised learning problem or parameters of a meta-learner. We provide sufficient conditions under which solutions of the approximate problem converge to those of the exact problem. We instantiate our approach for meta-learning in the case of deep learning where representation layers are treated as hyperparameters shared across a set of training episodes. In experiments, we confirm our theoretical findings, present encouraging results for few-shot learning and contrast the bilevel approach against classical approaches for learning-to-learn.

1. Introduction

While in standard supervised learning problems we seek the best hypothesis in a given space and with a given learning algorithm, in hyperparameter optimization (HO) and meta-learning (ML) we seek a configuration so that the optimized learning algorithm will produce a model that generalizes well to new data. The search space in ML often incorporates choices associated with the hypothesis space and the features of the learning algorithm itself (e.g., how optimization of the training loss is performed). Under this common perspective, both HO and ML essentially boil down to *nesting two search problems*: at the inner level we seek a good

hypothesis (as in standard supervised learning) while at the outer level we seek a good configuration (including a good hypothesis space) where the inner search takes place. Surprisingly, the literature on ML has little overlap with the literature on HO and in this paper we present a unified framework encompassing both of them.

Classic approaches to HO (see e.g. Hutter et al., 2015, for a survey) have been only able to manage a relatively small number of hyperparameters, from a few dozens using random search (Bergstra and Bengio, 2012) to a few hundreds using Bayesian or model-based approaches (Bergstra et al., 2013; Snoek et al., 2012). Recent gradient-based techniques for HO, however, have significantly increased the number of hyperparameters that can be optimized (Domke, 2012; Maclaurin et al., 2015; Pedregosa, 2016; Franceschi et al., 2017) and it is now possible to tune as hyperparameters entire weight vectors associated with a neural network layer. In this way, it becomes feasible to design models that possibly have more hyperparameters than parameters. Such an approach is well suited for ML, since parameters are learned from a small dataset, whereas hyperparameters leverage multiple available datasets.

HO and ML only differ substantially in terms of the experimental settings in which they are evaluated. While in HO the available data is associated with a single task and split into a training set (used to tune the parameters) and a validation set (used to tune the hyperparameters), in ML we are often interested in the so-called *few-shot* learning setting where data comes in the form of short episodes (small datasets with few examples per class) sampled from a common probability distribution over supervised tasks.

Early work on ML dates back at least to the 1990's (Schmidhuber, 1992; Baxter, 1995; Thrun and Pratt, 1998) but this research area has received considerable attention in the last few years, mainly driven by the need in real-life and industrial scenarios for learning quickly a vast multitude of tasks. These tasks, or *episodes*, may appear and evolve continuously over time and may only contain few examples (Lake et al., 2017). Different strategies have emerged to tackle ML. Although they do overlap in some aspects, it is possible to identify at least four of them. The *metric strategy* attempts to use training episodes to construct embeddings such that examples of the same class are mapped into similar repre-

[1] Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa, Italy [2] Department of Computer Science, University College London, London, UK [3] Department of Information Engineering, Università degli Studi di Firenze, Florence, Italy. Correspondence to: Luca Franceschi <luca.franceschi@iit.it>.

Table 1. Links and naming conventions among different fields.

Bilevel programming	Hyperparameter optimization	Meta-learning
Inner variables	Parameters	Parameters of Ground models
Outer variables	Hyperparameters	Parameters of Meta-learner
Inner objective	Training error	Training errors on tasks (Eq. 3)
Outer objective	Validation error	Meta-training error (Eq. 4)

sentations. It has been instantiated in several variants that involve non-parametric (or instance-based) predictors (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017). In the related *memorization strategy*, the meta-learner learns to store and retrieve data points representations in memory. It can be implemented either using recurrent networks (Santoro et al., 2016) or temporal convolutions (Mishra et al., 2018). The use of an attention mechanism (Vaswani et al., 2017) is crucial both in (Vinyals et al., 2016) and in (Mishra et al., 2018). The *initialization strategy* (Ravi and Larochelle, 2017; Finn et al., 2017) uses training episodes to infer a good initial value for the model's parameters so that new tasks can be learned quickly by fine tuning. The *optimization strategy* (Andrychowicz et al., 2016; Ravi and Larochelle, 2017; Wichrowska et al., 2017) forges an optimization algorithm that will find it easier to learn on novel related tasks.

A main contribution of this paper is a unified view of HO and ML within the natural mathematical framework of bilevel programming, where an outer optimization problem is solved subject to the optimality of an inner optimization problem. In HO the outer problem involves hyperparameters while the inner problem is usually the minimization of an empirical loss. In ML the outer problem could involve a shared representation among tasks while the inner problem could concern classifiers for individual tasks. Bilevel programming (Bard, 2013) has been suggested before in machine learning in the context of kernel methods and support vector machines (Keerthi et al., 2007; Kunapuli et al., 2008), multitask learning (Flamary et al., 2014), and more recently HO (Pedregosa, 2016), but never in the context of ML. The resulting framework outlined in Sec. 2 encompasses some existing approaches to ML, in particular those based on the initialization and the optimization strategies.

A technical difficulty arises when the solution to the inner problem cannot be written analytically (for example this happens when using the log-loss for training neural networks) and one needs to resort to iterative optimization approaches. As a second contribution, we provide in Sec. 3 sufficient conditions that guarantee good approximation properties. We observe that these conditions are reasonable and apply

to concrete problems relevant to applications.

In Sec. 4, by taking inspiration on early work on representation learning in the context of multi-task and meta-learning (Baxter, 1995; Caruana, 1998), we instantiate the framework for ML in a simple way treating the weights of the last layer of a neural network as the inner variables and the remaining weights, which parametrize the representation mapping, as the outer variables. As shown in Sec. 5, the resulting ML algorithm performs well in practice, outperforming most of the existing strategies on MiniImagenet.

2. A bilevel optimization framework

In this paper, we consider bilevel optimization problems (see e.g. Colson et al., 2007) of the form

$$\min\{f(\lambda) : \lambda \in \Lambda\}, \qquad (1)$$

where function $f : \Lambda \to \mathbb{R}$ is defined at $\lambda \in \Lambda$ as

$$f(\lambda) = \inf\{E(w_\lambda, \lambda) : w_\lambda \in \arg\min_{u \in \mathbb{R}^d} L_\lambda(u)\}. \qquad (2)$$

We call $E : \mathbb{R}^d \times \Lambda \to \mathbb{R}$ the *outer objective* and, for every $\lambda \in \Lambda$, we call $L_\lambda : \mathbb{R}^d \to \mathbb{R}$ the *inner objective*. Note that $\{L_\lambda : \lambda \in \Lambda\}$ is a class of objective functions parameterized by λ. Specific instances of this problem include HO and ML, which we discuss next. Table 1 outlines the links among bilevel programming, HO and ML.

2.1. Hyperparameter Optimization

In the context of hyperparameter optimization, we are interested in minimizing the validation error of a model $g_w : \mathcal{X} \to \mathcal{Y}$ parameterized by a vector w, with respect to a vector of hyperparameters λ. For example, we may consider representation or regularization hyperparameters that control the hypothesis space or penalties, respectively. In this setting, a prototypical choice for the inner objective is the regularized empirical error

$$L_\lambda(w) = \sum_{(x,y) \in D_{\mathrm{tr}}} \ell(g_w(x), y) + \Omega_\lambda(w),$$

where $D_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^n$ is a set of input/output points, ℓ is a prescribed loss function, and Ω_λ a regularizer parameterized by λ. The outer objective represents a proxy for the generalization error of g_w, and it may be given by the average loss on a validation set D_{val}

$$E(w, \lambda) = \sum_{(x,y) \in D_{\mathrm{val}}} \ell(g_w(x), y).$$

or, in more generality, by a cross-validation error, as detailed in Appendix B. Note that in this setting, the outer objective E does not depend explicitly on the hyperparameters λ,
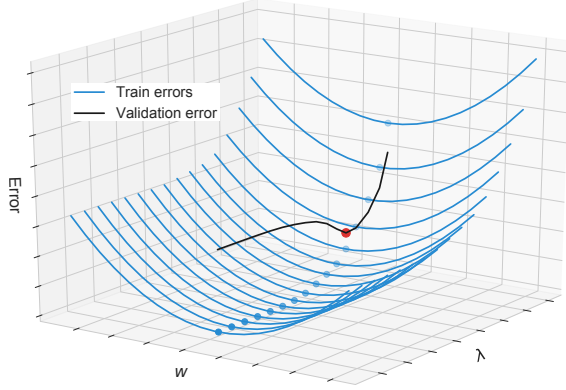
Figure 1. Each blue line represents the average training error when varying w in $g_{w,\lambda}$ and the corresponding inner minimizer is shown as a blue dot. The validation error evaluated at each minimizer yields the black curve representing the outer objective $f(\lambda)$, whose minimizer is shown as a red dot.

since in HO λ is instrumental in finding a good model g_w, which is our final goal. As a more specific example, consider linear models, $g_w(x) = \langle w, x \rangle$, let ℓ be the square loss and let $\Omega_\lambda(w) = \lambda \|w\|^2$, in which case the inner objective is ridge regression (Tikhonov regularization) and the bilevel problem optimizes over the regularization parameter the validation error of ridge regression.

2.2. Meta-Learning

In meta-learning (ML) the inner and outer objectives are computed by averaging a training and a validation error over multiple tasks, respectively. The goal is to produce a learning algorithm that will work well on novel tasks[1]. For this purpose, we have available a meta-training set $\mathcal{D} = \{D^j = D^j_{\text{tr}} \cup D^j_{\text{val}}\}_{j=1}^N$, which is a collection of datasets, sampled from a meta-distribution $\mathcal{P}$. Each dataset $D^j = \{(x_i^j, y_i^j)\}_{i=1}^{n_j}$ with $(x_i^j, y_i^j) \in \mathcal{X} \times \mathcal{Y}^j$ is linked to a specific task. Note that the output space is task dependent (e.g. a multi-class classification problem with variable number of classes). The model for each task is a function $g_{w^j,\lambda} : \mathcal{X} \to \mathcal{Y}^j$, identified by a parameter vectors w^j and hyperparameters λ. A key point here is that λ is shared between the tasks. With this notation the inner and outer objectives are

$$L_\lambda(w) = \sum_{j=1}^N L^j(w^j, \lambda, D^j_{\text{tr}}), \qquad (3)$$

$$E(w, \lambda) = \sum_{j=1}^N L^j(w^j, \lambda, D^j_{\text{val}}) \qquad (4)$$

respectively. The loss $L_j(w_j, \lambda, S)$ represents the empirical error of the pair (w_j, λ) on a set of examples S. Note that the

---

[1]The ML problem is also related to multitask learning, however in ML the goal is to extrapolate from the given tasks.

inner and outer losses for task j use different train/validation splits of the corresponding dataset D^j. Furthermore, unlike in HO, in ML the final goal is to find a good λ and the w^j are now instrumental.

The cartoon in Figure 1 illustrates ML as a bilevel problem. The parameter λ indexes an hypothesis space within which the inner objective is minimized. A particular example, detailed in Sec. 4, is to choose the model $g_{w,\lambda} = \langle w, h_\lambda(x) \rangle$, in which case λ parameterizes a feature mapping. Yet another choice would be to consider $g_{w^j,\lambda}(x) = \langle w + \lambda, x \rangle$, in which case λ represents a common model around which task specific models are to be found (see e.g. Evgeniou et al., 2005; Finn et al., 2017; Khosla et al., 2012; Kuzborskij et al., 2013, and reference therein).

2.3. Gradient-Based Approach

We now discuss a general approach to solve Problem (1)-(2) when the hyperparameter vector λ is real-valued. To simplify our discussion let us assume that the inner objective has a unique minimizer w_λ. Even in this simplified scenario, Problem (1)-(2) remains challenging to solve. Indeed, in general there is no closed form expression w_λ, so it is not possible to directly optimize the outer objective function. While a possible strategy (implicit differentiation) is to apply the implicit function theorem to $\nabla L_\lambda = 0$ (Pedregosa, 2016; Koh and Liang, 2017; Beirami et al., 2017), another compelling approach is to replace the inner problem with a dynamical system. This point, discussed in (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2017), is developed further in this paper.

Specifically, we let $[T] = \{1, \ldots, T\}$ where T is a prescribed positive integer and consider the following approximation of Problem (1)-(2)

$$\min_\lambda f_T(\lambda) = E(w_{T,\lambda}, \lambda), \qquad (5)$$

where E is a smooth scalar function, and[2]

$$w_{0,\lambda} = \Phi_0(\lambda), \ w_{t,\lambda} = \Phi_t(w_{t-1,\lambda}, \lambda), \ t \in [T], \qquad (6)$$

with $\Phi_0 : \mathbb{R}^m \to \mathbb{R}^d$ a smooth initialization mapping and, for every $t \in [T]$, $\Phi_t : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ a smooth mapping that represents the operation performed by the t-th step of an optimization algorithm. For example, the optimization dynamics could be gradient descent: $\Phi_t(w_t, \lambda) = w_t - \eta_t \nabla L_\lambda(\cdot)$ where $(\eta_t)_{t \in [T]}$ is a sequence of steps sizes.

The approximation of the bilevel problem (1)-(2) by the procedure (5)-(6) raises the issue of the quality of this approximation and we return to this issue in the next section.

---

[2]In general, the algorithm used to minimize the inner objective may involve auxiliary variables, e.g., velocities when using gradient descent with momentum, so w should be intended as a larger vector containing both model parameters and auxiliary variables.

However, it also suggests to consider the inner dynamics as a form of approximate empirical error minimization (e.g. early stopping) which is valid in its own right. From this perspective – conversely to the implicit differentiation strategy – it is possible to include among the components of λ variables which are associated with the optimization algorithm itself. For example, λ may include the step sizes or momentum factors if the dynamics Φ_t in Eq. (6) is gradient descent with momentum; in (Andrychowicz et al., 2016; Wichrowska et al., 2017) the mapping Φ_t is implemented as a recurrent neural network, while (Finn et al., 2017) focus on the initialization mapping by letting $\Phi_0(\lambda) = \lambda$.

A major advantage of this reformulation is that it makes it possible to compute efficiently the gradient of f_T, which we call *hypergradient*, either in time or in memory (Maclaurin et al., 2015; Franceschi et al., 2017), by making use of reverse or forward mode algorithmic differentiation (Griewank and Walther, 2008; Baydin et al., 2017). This allows us to optimize a number of hyperparameters of the same order of that of parameters, a situation which arise in ML.

3. Exact and Approximate Bilevel Programming

In this section, we provide results about the existence of solutions of Problem (1)-(2) and the approximation properties of Procedure (5)-(6) with respect to the original bilevel problem. Proofs of these results are provided in the supplementary material.

Procedure (5)-(6), though related to the bilevel problem (1)-(2), may not be, in general, a good approximation of it. Indeed, making the assumptions (which sound perfectly reasonable) that, for every $\lambda \in \Lambda$, $w_{T,\lambda} \to w_\lambda$ for some $w_\lambda \in \arg\max L_\lambda$, and that $E(\cdot, \lambda)$ is continuous, one can only assert that $\lim_{T\to\infty} f_T(\lambda) = E(w_\lambda, \lambda) \geq f(\lambda)$. This is because the optimization dynamics converge to some minimizer of the inner objective L_λ, but not necessarily to the one that also minimizes the function E. This is illustrated in Figure 2. The situation is, however, different if the inner problem admits a unique minimizer for every $\lambda \in \Lambda$. Indeed in this case, it is possible to show that the set of minimizers of the approximate problems converge, as $T \to +\infty$ and in an appropriate sense, to the set of minimizers of the bilevel problem. More precisely, we make the following assumptions:

(i) Λ is a compact subset of $\mathbb{R}^m$;

(ii) $E : \mathbb{R}^d \times \Lambda \to \mathbb{R}$ is jointly continuous;

(iii) the map $(w, \lambda) \mapsto L_\lambda(w)$ is jointly continuous and such that $\arg\min L_\lambda$ is a singleton for every $\lambda \in \Lambda$;

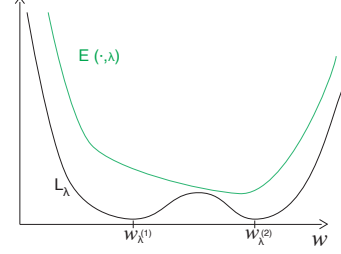(iv) $w_\lambda = \arg\min L_\lambda$ remains bounded as λ varies in Λ.

Figure 2. In this cartoon, for a fixed λ, $\arg\min L_\lambda = \{w_\lambda^{(1)}, w_\lambda^{(2)}\}$; the iterates of an optimization mapping Φ could converge to $w_\lambda^{(1)}$ with $E(w_\lambda^{(1)}, \lambda) > E(w_\lambda^{(2)}, \lambda)$.

Then, problem (1)-(2) becomes

$$\min_{\lambda \in \Lambda} f(\lambda) = E(w_\lambda, \lambda), \qquad w_\lambda = \arg\min_u L_\lambda(u). \quad (7)$$

Under the above assumptions, in the following we give results about the existence of solutions of problem (7) and the (variational) convergence of the approximate problems (5)-(6) towards problem (7) — relating the minima as well as the set of minimizers. In this respect we note that, since both f and f_T are nonconvex, $\arg\min f_T$ and $\arg\min f$ are in general nonsingleton, so an appropriate definition of set convergence is required.

Theorem 3.1 (Existence). *Under Assumptions* (i)-(iv) *problem* (7) *admits solutions.*
Proof See Appendix A.

The result below follows from general facts on the stability of minimizers in optimization problems (Dontchev and Zolezzi, 1993).

Theorem 3.2 (Convergence). *In addition to Assumptions* (i)-(iv)*, suppose that:*

(v) $E(\cdot, \lambda)$ *is uniformly Lipschitz continuous;*

(vi) *The iterates* $(w_{T,\lambda})_{T\in\mathbb{N}}$ *converge uniformly to* w_λ *on* Λ *as* $T \to +\infty$.

Then

(a) $\inf f_T \to \inf f$,

(b) $\arg\min f_T \to \arg\min f$, *meaning that, for every* $(\lambda_T)_{T\in\mathbb{N}}$ *such that* $\lambda_T \in \arg\min f_T$, *we have that:*

    - $(\lambda_T)_{T\in\mathbb{N}}$ *admits a convergent subsequence;*
    - *for every subsequence* $(\lambda_{K_T})_{T\in\mathbb{N}}$ *such that* $\lambda_{K_T} \to \bar{\lambda}$, *we have* $\bar{\lambda} \in \arg\min f$.

Proof See Appendix A.

We stress that assumptions (i)-(vi) are very natural and satisfied by many problems of practical interests. Thus, the above results provide full theoretical justification to the proposed approximate procedure (5)-(6). The following remark discusses assumption (vi), while the subsequent example will be relevant to the experiments in Sec. 5.

Algorithm 1. Reverse-HG for Hyper-representation

Input: λ, current values of the hyperparameter, T number of iteration of GD, η ground learning rate, $\mathcal{B}$ mini-batch of episodes from $\mathcal{D}$

Output: Gradient of meta-training error w.r.t. λ on $\mathcal{B}$

for $j = 1$ **to** $|\mathcal{B}|$ **do**
$\quad w_0^j = 0$
$\quad$ **for** $t = 1$ **to** T **do**
$\qquad w_t^j \leftarrow w_{t-1} - \eta \nabla_w L^j(w_{t-1}^j, \lambda, D_{\text{tr}}^j)$
$\quad \alpha_T^j \leftarrow \nabla_w L^j(w_T^j, \lambda, D_{\text{val}})$
$\quad p^j \leftarrow \nabla_\lambda L^j(w_T^j, \lambda, D_{\text{val}})$
$\quad$ **for** $t = T - 1$ **downto** 0 **do**
$\qquad p^j \leftarrow p^j - \alpha_{t+1}^j \eta \nabla_\lambda \nabla_w L^j(w_t^j, \lambda, D_{\text{tr}}^j)$
$\qquad \alpha_t^j \leftarrow \alpha_{t+1}^j \left[I - \eta \nabla_w \nabla_w L^j(w_t^j, \lambda, D_{\text{tr}}^j) \right]$
return $\sum_j p^j$

Remark 3.3. If L_λ is strongly convex, then many gradient-based algorithms (e.g., standard and accelerated gradient descent) yield linear convergence of the iterates $w_{T,\lambda}$'s. Moreover, in such cases, the rate of linear convergence is of type $(\nu_\lambda - \mu_\lambda)/(\nu_\lambda + \mu_\lambda)$, where ν_λ and μ_λ are the Lipschitz constant of the gradient and the modulus of strong convexity of L_λ respectively. So, this rate can be uniformly bounded from above by $\rho \in]0, 1[$, provided that $\sup_{\lambda \in \Lambda} \nu_\lambda < +\infty$ and $\inf_{\lambda \in \Lambda} \mu_\lambda > 0$. Thus, in these cases $w_{T,\lambda}$ converges uniformly to w_λ on Λ (at a linear rate).

Example 3.4. Let us consider the following form of the inner objective:

$$L_H(w) = \|y - XHw\|^2 + \rho\|w\|^2, \qquad (8)$$

where $\rho > 0$ is a fixed regularization parameter and $H \in \mathbb{R}^{d \times d}$ is the hyperparameter, representing a linear feature map. L_H is strongly convex, with modulus $\mu = \rho > 0$ (independent on the hyperparameter H), and Lipschitz smooth with constant $\nu_H = 2\|(XH)^\top XH + \rho I\|$, which is bounded from above, if H ranges in a bounded set of square matrices. In this case assumptions (i)-(vi) are satisfied.

4. Learning Hyper-Representations

In this section, we instantiate the bilevel programming approach for ML outlined in Sec. 2.2 in the case of deep learning where representation layers are shared across episodes. Finding good data representations is a centerpiece in machine learning. Classical approaches (Baxter, 1995; Caruana, 1998) learn both the weights of the representation mapping and those of the ground classifiers jointly on the same data. Here we follow the bilevel approach and split each dataset/episode in training and validation sets.

Our method involves the learning of a cross-task intermedi-

ate representation $h_\lambda : \mathcal{X} \to \mathbb{R}^k$ (parametrized by a vector λ) on top of which task specific models $g^j : \mathbb{R}^k \to \mathcal{Y}^j$ (parametrized by vectors w^j) are trained. The final ground model for task j is thus given by $g^j \circ h$. To find λ, we solve Problem (1)-(2) with inner and outer objectives as in Eqs. (3) and (4), respectively. Since, in general, this problem cannot be solved exactly, we instantiate the approximation scheme in Eqs. (5)-(6) as follows:

$$\min_\lambda f_T(\lambda) = \sum_{j=1}^N L^j(w_T^j, \lambda, D_{\text{val}}^j) \qquad (9)$$

$$w_t^j = w_{t-1}^j - \eta \nabla_w L^j(w_{t-1}^j, \lambda, D_{\text{tr}}^j), \ t, j \in [T], \ [N]. \quad (10)$$

Starting from an initial value, the weights of the task-specific models are learned by T iterations of gradient descent. The gradient of f_T can be computed efficiently in time by making use of an extended reverse-hypergradient procedure (Franceschi et al., 2017) which we present in Algorithm 1. Since, in general, the number of episodes in a meta-training set is large, we compute a stochastic approximation of the gradient of f_T by sampling a mini-batch of episodes. At test time, given a new episode $\bar{D}$, the representation h is kept fixed, and all the examples in $\bar{D}$ are used to tune the weights $\bar{w}$ of the episode-specific model $\bar{g}$.

Like other initialization and optimization strategies for ML, our method does not require lookups in a support set as the memorization and metric strategies do (Santoro et al., 2016; Vinyals et al., 2016; Mishra et al., 2018). Unlike (Andrychowicz et al., 2016; Ravi and Larochelle, 2017) we do not tune the optimization algorithm, which in our case is plain empirical loss minimization by gradient descent, and rather focus on the hypothesis space. Unlike (Finn et al., 2017), that aims at maximizing sensitivity of new task losses to the model parameters, we aim at maximizing the generalization to novel examples during training episodes, with respect to λ. Our assumptions about the structure of the model are slightly stronger than in (Finn et al., 2017) but still mild, namely that some (hyper)parameters define the representation and the remaining parameters define the classification function. In (Munkhdalai and Yu, 2017) the meta-knowledge is distributed among fast and slow weights and an external memory; our approach is more direct, since the meta-knowledge is solely distilled by λ. A further advantage of our method is that, if the episode-specific models are linear (e.g. logistic regressors) and each loss L^j is strongly convex in w, the theoretical guarantees of Theorem 3.2 apply (see Remark 3.3). These assumptions are satisfied in the experiments reported in the next section.

5. Experiments

The aim of the following experiments is threefold. First, we investigate the impact of the number of iterations of the optimization dynamics on the quality of the solution on a
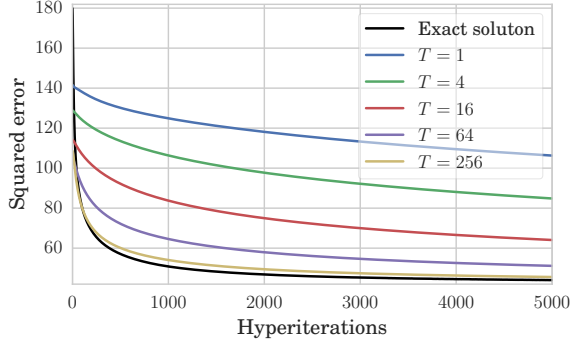
Figure 3. Optimization of the outer objectives f and f_T for exact and approximate problems. The optimization of H is performed with gradient descent with momentum, with same initialization, step size and momentum factor for each run.
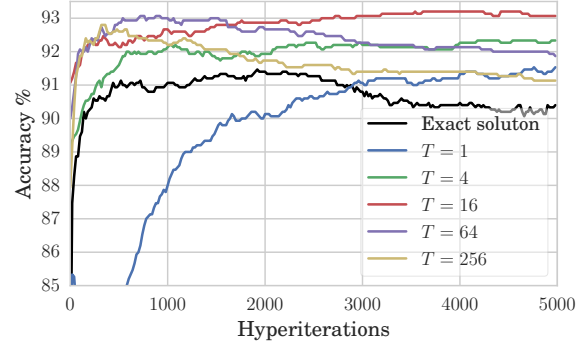

Figure 4. Accuracy on D_{test} of exact and approximated solutions during optimization of H. Training and validation accuracies reach almost 100% already for $T = 4$ and after few hundred hyperiterations, and therefore are not reported.

simple multiclass classification problem. Second, we test our hyper-representation method in the context of few-shot learning on two benchmark datasets. Finally, we constrast the bilevel ML approach against classical approaches to learn shared representations [3].

5.1. The Effect of T

Motivated by the theoretical findings of Sec. 3, we empirically investigate how solving the inner problem *approximately* (i.e. using small T) affects convergence, generalization performances, and running time. We focus in particular on the linear feature map described in Example 3.4, which allows us to compare the approximated solution against the closed-form analytical solution given by

$$w_H = [(XH)^T XH + \rho I]^{-1} (XH)^T Y.$$

In this setting, the bilevel problem reduces to a (non-convex) optimization problem in H.

We use a subset of 100 classes extracted from Omniglot dataset (Lake et al., 2017) to construct a HO problem aimed at tuning H. A training set D_{tr} and a validation set D_{val}, each consisting of three randomly drawn examples per class, were sampled to form the HO problem. A third set D_{test}, consisting of fifteen examples per class, was used for testing. Instead of using raw images as input, we employ feature vectors $x \in \mathbb{R}^{256}$ computed by the convolutional network trained on one-shot five-ways ML setting as described in Sec. 5.2.

For the approximate problems we compute the hypergradient using Algorithm 1, where it is intended that $\mathcal{B} = \{(D_{\text{tr}}, D_{\text{val}})\}$. Figure 3 shows the values of functions f and f_T (see Eqs. (1) and (5), respectively) during the optimization of H. As T increases, the solution of the approximate

---

[3] The code for reproducing the experiments, based on the package FAR-HO (https://bit.ly/far-ho), is available at https://bit.ly/hyper-repr

Table 2. Execution times on a NVidia Tesla M40 GPU.

T	1	4	16	64	256	Exact
Time (sec)	60	119	356	1344	5532	320

problem approaches the true bilevel solution. However, performing a small number of gradient descent steps for solving the inner problem acts as implicit regularizer. As it is evident from Figure 4, the generalization error is better when T is smaller than the value yielding the best approximation of the inner solution. This is to be expected since, in this setting, the dimensions of parameters and hyperparameters are of the same order, leading to a concrete possibility of overfitting the outer objective (validation error). An appropriate, problem dependent, choice of T may help avoiding this issue (see also Appendix C). As T increases, the number of hyperiterations required to reach the maximum test accuracy decreases, further suggesting that there is an interplay between the number of iterations used to solve the inner and the outer objective. Finally, the running time of Algorithm 1, is linear in T and the size of w and independent of the size of H (see also Table 2), making it even more appealing to reduce the number of iterations.

5.2. Few-shot Learning

We new turn our attention to learning-to-learn, precisely to few-shot supervised learning, implementing the ML strategy outlined in Sec. 4 on two different benchmark datasets:

● OMNIGLOT (Lake et al., 2015), a dataset that contains examples of 1623 different handwritten characters from 50 alphabets. We downsample the images to 28×28.

● MINIIMAGENET (Vinyals et al., 2016), a subset of ImageNet (Deng et al., 2009), that contains 60000 downsampled images from 100 different classes.

Following the experimental protocol used in a number of recent works, we build a meta-training set $\mathcal{D}$, from which we sample datasets to solve Problem (9)-(10), a meta-validation

set $\mathcal{V}$ for tuning ML hyperparameters, and finally a meta-test set $\mathcal{T}$ which is used to estimate accuracy. Operationally, each meta-dataset consists of a pool of samples belonging to different (non-overlapping between separate meta-dataset) classes, which can be combined to form ground classification datasets $D^j = D^j_{\text{tr}} \cup D^j_{\text{val}}$ with 5 or 20 classes (for Omniglot). The D^j_{tr}'s contain 1 or 5 examples per class which are used to fit w^j (see Eq. 10). The D^j_{val}'s, containing 15 examples per class, is used either to compute $f_T(\lambda)$ (see Eq. (9)) and its (stochastic) gradient if $D^j \in \mathcal{D}$ or to provide a generalization score if D^j comes from either $\mathcal{V}$ or $\mathcal{T}$. For MiniImagenet we use the same split and images proposed in (Ravi and Larochelle, 2017), while for Omniglot we use the protocol defined by (Santoro et al., 2016).

As ground classifiers we use multinomial logistic regressors and as task losses ℓ^j we employ cross-entropy. The inner problems, being strongly convex, admit unique minimizers, yet require numerical computation of the solutions. We initialize ground models parameters w^j to 0 and, according to the observation in Sec. 5.1, we perform T gradient descent steps, where T is treated as a ML hyperparameter that has to be validated. Figure 6 shows an example of meta-validation of T for one-shot learning on MiniImagenet. We compute a stochastic approximation of $\nabla f_T(\lambda)$ with Algorithm 1 and use Adam with decaying learning rate to optimize λ.

Regarding the specific implementation of the representation mapping h, we employ for Omniglot a four-layers convolutional neural network with strided convolutions and 64 filters per layer as in (Vinyals et al., 2016) and other successive works. For MiniImagenet we tried two different architectures:

• *C4L*, a four-layers convolutional neural network with max-pooling and 32 filters per layer;

• *RN*: a residual network (He et al., 2016) built of four residual blocks followed by two convolutional layers.

The first network architecture has been proposed in (Ravi and Larochelle, 2017) and then used in (Finn et al., 2017), while a similar residual network architecture has been employed in a more recent work (Mishra et al., 2018). Further details on the architectures of h, as well as other ML hyperparameters, are specified in the supplementary material. We report our results, using *RN* for MiniImagenet, in Table 3, alongside scores from various recently proposed methods for comparison.

The proposed method achieves competitive results highlighting the relative importance of learning a task independent representation, on the top of which logistic classifiers trained with very few samples generalize well. Moreover, utilizing more expressive models such as residual network as representation mappings, is beneficial for our proposed strategy and, unlike other methods, does not result in overfitting

of the outer objective, as reported in (Mishra et al., 2018). Indeed, compared to *C4L*, *RN* achieves a relative improvement of 6.5% on one-shot and 4.2% on five-shot. Figure 5 provides a visual example of the goodness of the learned representation, showing that MiniImagenet examples (the first from meta-training, the second from the meta-testing sets) from similar classes (different dog breeds) are mapped near each other by h and, conversely, samples from dissimilar classes are mapped afar.
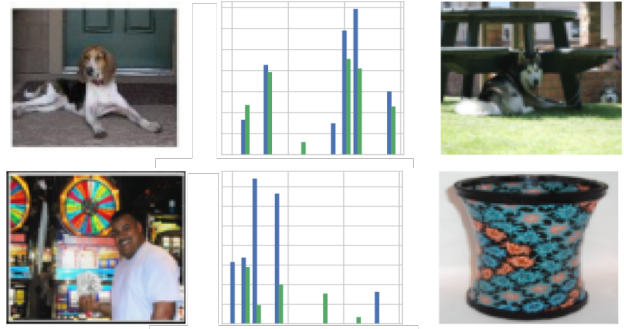

Figure 5. After sampling two datasets $D \in \mathcal{D}$ and $D' \in \mathcal{T}$, we show on the top the two images $x \in D$, $x' \in D'$ that minimize $||h_\lambda(x) - h_\lambda(x')||$ and on the bottom those that maximize it. In between each of the two couples we compare a random subset of components of $h_{\lambda(x)}$ (blue) and $h_\lambda(x')$ (green).
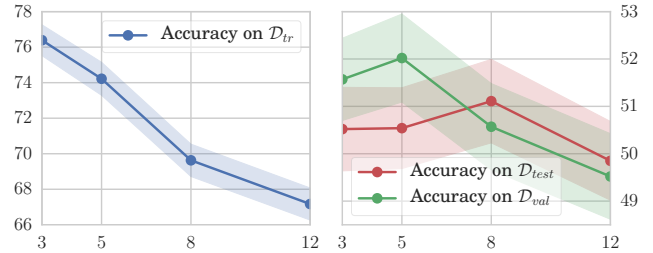

Figure 6. Meta-validation of the number of gradient descent steps (T) of the ground models for MiniImagenet using the *RN* representation. Early stopping on the accuracy on meta-validation set during meta-training resulted in halting the optimization of λ after 42k, 40k, 22k, and 15k hyperiterations for T equal to 3, 5, 8 and 12 respectively; in line with our observation in Sec. 5.1.

5.3. On Variants of Representation Learning Methods

In this section, we show the benefits of learning a representation within the proposed bilevel framework compared to other possible approaches that involve an explicit factorization of a classifier as $g^j \circ h$. The representation mapping h is either pretrained or learned with different meta-learning algorithms. We focus on the problem of one-shot learning on MiniImagenet and we use *C4L* as architecture for the representation mapping. In all the experiments the ground models g^j are multinomial logistic regressor as in Sec. 5.2, tuned with 5 steps of gradient descent. We ran the following experiments:

• *Multiclass*: the mapping $h : \mathcal{X} \to \mathbb{R}^{64}$ is given by the

Table 3. Accuracy scores, computed on episodes from $\mathcal{T}$, of various methods on 1-shot and 5-shot classification problems on Omniglot and MiniImagenet. For MiniImagenet 95% confidence intervals are reported. For Hyper-representation the scores are computed over 600 randomly drawn episodes. For other methods we show results as reported by their respective authors.

Method	OMNIGLOT 5 classes		OMNIGLOT 20 classes		MINIIMAGENET 5 classes	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Siamese nets (Koch et al., 2015)	97.3	98.4	88.2	97.0	—	—
Matching nets (Vinyals et al., 2016)	98.1	98.9	93.8	98.5	43.44 ± 0.77	55.31 ± 0.73
Neural stat. (Edwards and Storkey, 2016)	98.1	99.5	93.2	98.1	—	—
Memory mod. (Kaiser et al., 2017)	98.4	99.6	95.0	98.6	—	—
Meta-LSTM (Ravi and Larochelle, 2017)	—	—	—	—	43.56 ± 0.84	60.60 ± 0.71
MAML (Finn et al., 2017)	98.7	99.9	95.8	98.9	48.70 ± 1.75	63.11 ± 0.92
Meta-networks (Munkhdalai and Yu, 2017)	98.9	—	97.0	—	49.21 ± 0.96	—
Prototypical Net. (Snell et al., 2017)	98.8	99.7	96.0	98.9	49.42 ± 0.78	68.20 ± 0.66
SNAIL (Mishra et al., 2018)	99.1	99.8	97.6	99.4	55.71 ± 0.99	68.88 ± 0.92
Hyper-representation	98.6	99.5	95.5	98.4	50.54 ± 0.85	64.53 ± 0.68

linear outputs before the softmax operation of a network[4] pretrained on the totality of examples contained in the training meta-dataset (600 examples for each of the 64 classes). In this setting, we found that using the second last layer or the output after the softmax yields worst results;

• *Bilevel-train*: we use a bilevel approach but, unlike in Sec. 4, we optimize the parameter vector λ of the representation mapping by minimizing the loss on the training sets of each episode. The hypergradient is still computed with Algorithm 1, albeit we set $D_{\text{val}}^j = D_{\text{tr}}^j$ for each training episodes;

• *Approx* and *Approx-train*: we consider an approximation of the hypergradient $\nabla f_{\mathcal{T}}(\lambda)$ by disregarding the optimization dynamics of the inner objectives (i.e. we set $\nabla_\lambda w_T^j = 0$). In *Approx-train* we just use the training sets;

• *Classic*: as in (Baxter, 1995), we learn h by jointly optimize $\hat{f}(\lambda, w^1, \ldots, w^N) = \sum_{j=1}^{N} L^j(w^j, \lambda, D_{\text{tr}}^j)$ and treat the problem as standard multitask learning, with the exception that we evaluate $\hat{f}$ on mini-batches of 4 episodes, randomly sampled every 5 gradient descent iterations.

In settings where we do not use the validation sets, we let the training sets of each episode contain 16 examples per class. Using training episodes with just one example per class resulted in performances just above random chance. While the first experiment constitutes a standard baseline, the others have the specific aim of assessing (i) the importance of splitting episodes of meta-training set into training and validation and (ii) the importance of computing the hypergradient of the approximate bilevel problem with Algorithm 1. The results reported in Table 4 suggest that both the training/validation splitting and the full computation of the hypergradient constitute key factors for learning a good representation in a meta-learning context. On the other side, using pretrained representations, especially in a low-dimensional space, turns out to be a rather effective baseline. One possible explanation is that, in this context,

---

[4]The network is similar to C4L but has 64 filters per layer.

Table 4. Performance of various methods where the representation is either transfered or learned with variants of hyper-representation methods. The last raw reports, for comparison, the score obtained with hyper-representation.

Method	# filters	Accuracy 1-shot
Multiclass	64	43.02
Bilevel-train	32	29.63
Approx	32	41.12
Approx-train	32	38.80
Classic-train	32	40.46
Hyper-representation-C4L	32	47.51

some classes in the training and testing meta-datasets are rather similar (e.g. various dog breeds) and thus ground classifiers can leverage on very specific representations.

6. Conclusions

We have shown that both HO and ML can be formulated in terms of bilevel programming and solved with an iterative approach. When the inner problem has a unique solution (e.g. is strongly convex), our theoretical results show that the iterative approach has convergence guarantees, a result that is interesting in its own right. In the case of ML, by adapting classical strategies (Baxter, 1995) to the bilevel framework with training/validation splitting, we present a method for learning hyper-representations which is experimentally effective and supported by our theoretical guarantees.

Our framework encompasses recently proposed methods for meta-learning, such as learning to optimize, but also suggests different design patterns for the inner learning algorithm which could be interesting to explore in future work. The resulting inner problems may not satisfy the assumptions of our convergence analysis, raising the need for further theoretical investigations. An additional future direction of research is the study of the statistical properties of bilevel strategies where outer objectives are based on the generalization ability of the inner model to new (validation) data. Ideas from (Maurer et al., 2016; Denevi et al., 2018) may be useful in this direction.

References

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., and de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3981–3989.

Bard, J. F. (2013). *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media. 01251.

Baxter, J. (1995). Learning internal representations. In *Proceedings of the 8th Annual Conference on Computational Learning Theory (COLT)*, pages 311–320. ACM.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:153:1–153:43.

Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. (2017). On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3458–3468.

Bergstra, J. and Bengio, Y. (2012). Random search for hyperparameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning, (ICML)*, pages 115–123.

Caruana, R. (1998). Multitask learning. In *Learning to learn*, pages 95–133. Springer. 02683.

Colson, B., Marcotte, P., and Savard, G. (2007). An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*. To appear in UAI 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Domke, J. (2012). Generic Methods for Optimization-Based Modeling. In *AISTATS*, volume 22, pages 318–326.

Dontchev, A. L. and Zolezzi, T. (1993). *Well-posed optimization problems*, volume 1543 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Edwards, H. and Storkey, A. (2016). Towards a Neural Statistician. *arXiv:1606.02185 [cs, stat]*. 00027 arXiv: 1606.02185.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, pages 1126–1135.

Flamary, R., Rakotomamonjy, A., and Gasso, G. (2014). Learning constrained task similarities in graph-regularized multi-task learning. *Regularization, Optimization, Kernels, and Support Vector Machines*, page 103.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. (2017). Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, pages 1165–1173.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AIStat)*, pages 249–256.

Griewank, A. and Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hutter, F., Lcke, J., and Schmidt-Thieme, L. (2015). Beyond Manual Tuning of Hyperparameters. *KI - Künstliche Intelligenz*, 29(4):329–337.

Kaiser, L., Nachum, O., Roy, A., and Bengio, S. (2017). Learning to remember rare events.

Keerthi, S. S., Sindhwani, V., and Chapelle, O. (2007). An efficient method for gradient-based adaptation of hyperparameters in svm models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 673–680.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. 00127.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1885–1894.

Kunapuli, G., Bennett, K., Hu, J., and Pang, J.-S. (2008). Classification model selection via bilevel programming. *Optimization Methods and Software*, 23(4):475–489.

Kuzborskij, I., Orabona, F., and Caputo, B. (2013). From n to n+ 1: Multiclass transfer incremental learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3358–3365.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40. 00152.

Maclaurin, D., Duvenaud, D. K., and Adams, R. P. (2015). Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning, (ICML*, pages 2113–2122.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884.

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. (2018). A simple neural attentive meta-learner.

Munkhdalai, T. and Yu, H. (2017). Meta networks. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, pages 2554–2563.

Pedregosa, F. (2016). Hyperparameter optimization with approximate gradient. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pages 737–746.

Ravi, S. and Larochelle, H. (2017). Optimization as a model for few-shot learning. In *In International Conference on Learning Representations (ICLR)*.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *Proceedings ofthe 33rd International Conference on Machine Learning*, pages 1842–1850.

Schmidhuber, J. (1992). Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. *Neural Computation*, 4(1):131–139. 00082.

Snell, J., Swersky, K., and Zemel, R. S. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4080–4090.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2951–2959.

Thrun, S. and Pratt, L. (1998). *Learning to learn*. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6000–6010.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3630–3638.

Wichrowska, O., Maheswaranathan, N., Hoffman, M. W., Colmenarejo, S. G., Denil, M., Freitas, N., and Sohl-Dickstein, J. (2017). Learned optimizers that scale and generalize. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3751–3760.

A. Proofs of the Results in Sec. 3

Proof of Theorem 3.1. Since Λ is compact, it follows from Weierstrass theorem that a sufficient condition for the existence of minimizers is that f is continuous. Thus, let $\bar{\lambda} \in \Lambda$ and let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in Λ such that $\lambda_n \to \bar{\lambda}$. We prove that $f(\lambda_n) = E(w_{\lambda_n}, \lambda_n) \to E(w_{\bar{\lambda}}, \bar{\lambda}) = f(\bar{\lambda})$. Since $(w_{\lambda_n})_{n\in\mathbb{N}}$ is bounded, there exists a subsequence $(w_{k_n})_{n\in\mathbb{N}}$ such that $w_{k_n} \to \bar{w}$ for some $\bar{w} \in \mathbb{R}^d$. Now, since $\lambda_{k_n} \to \bar{\lambda}$ and the map $(w, \lambda) \mapsto L_\lambda(w)$ is jointly continuous, we have

$$\forall w \in \mathbb{R}^d, \quad L_{\bar{\lambda}}(\bar{w}) = \lim_n L_{\lambda_{k_n}}(w_{k_n})$$
$$\leq \lim_n L_{\lambda_{k_n}}(w) = L_{\bar{\lambda}}(w).$$

Therefore, $\bar{w}$ is a minimizer of $L_{\bar{\lambda}}$ and hence $\bar{w} = w_{\bar{\lambda}}$. This prove that $(w_{\lambda n})_{n\in\mathbb{N}}$ is a bounded sequence having a unique cluster point. Hence $(w_{\lambda n})_{n\in\mathbb{N}}$ is convergence to its unique cluster point, which is $w_{\bar{\lambda}}$. Finally, since $(w_{\lambda_n}, \lambda_n) \to (w_{\bar{\lambda}}, \bar{\lambda})$ and E is jointly continuous, we have $E(w_{\lambda_n}, \lambda_n) \to E(w_{\bar{\lambda}}, \bar{\lambda})$ and the statement follows. $\square$

We recal a fundamental fact concerning the stability of minima and minimizers in optimization problems (Dontchev and Zolezzi, 1993). We provide the proof for completeness.

Theorem A.1 (Convergence). *Let φ_T and φ be lower semicontinuous functions defined on a compact set Λ. Suppose that φ_T converges uniformly to φ on Λ as $T \to +\infty$. Then*

(a) $\inf \varphi_T \to \inf \varphi$,

(b) $\operatorname{argmin} \varphi_T \to \operatorname{argmin} \varphi$, *meaning that, for every $(\lambda_T)_{T\in\mathbb{N}}$ such that $\lambda_T \in \operatorname{argmin} \varphi_T$, we have that:*

- *$(\lambda_T)_{T\in\mathbb{N}}$ admits a convergent subsequence;*
- *for every subsequence $(\lambda_{K_T})_{T\in\mathbb{N}}$ such that $\lambda_{K_T} \to \bar{\lambda}$, we have $\bar{\lambda} \in \operatorname{argmin} \varphi$.*

Proof. Let $(\lambda_T)_{T\in\mathbb{N}}$ be a sequence in Λ such that, for every $T \in \mathbb{N}$, $\lambda_T \in \operatorname{argmin} \varphi_T$. We prove that

1) $(\lambda_T)_{T\in\mathbb{N}}$ admits a convergent subsequence.

2) for every subsequence $(\lambda_{K_T})_{T\in\mathbb{N}}$ such that $\lambda_{K_T} \to \bar{\lambda}$, we have $\bar{\lambda} \in \operatorname{argmin} \varphi$ and $\varphi_{K_T}(\lambda_{K_T}) \to \inf \varphi$.

3) $\inf \varphi_T \to \inf \varphi$.

The first point follows from the fact that Λ is compact. Concerning the second point, let $(\lambda_{K_T})_{T\in\mathbb{N}}$ be a subsequence such that $\lambda_{K_T} \to \bar{\lambda}$. Since φ_{K_T} converge uniformly to φ, we have

$$|\varphi_{K_T}(\lambda_{K_T}) - \varphi(\lambda_{K_T})| \leq \sup_{\lambda\in\Lambda}|\varphi_{K_T}(\lambda) - \varphi(\lambda)| \to 0.$$

Therefore, using also the continuity of φ, we have

$$\forall \lambda \in \Lambda, \quad \varphi(\bar{\lambda}) = \lim_T \varphi(\lambda_{K_T}) = \lim_T \varphi_{K_T}(\lambda_{K_T})$$
$$\leq \lim_T \varphi_{K_T}(\lambda) = \varphi(\lambda).$$

So, $\bar{\lambda} \in \operatorname{argmin} \varphi$ and $\varphi(\bar{\lambda}) = \lim_T \varphi_{K_T}(\lambda_{K_T}) \leq \inf \varphi = \varphi(\bar{\lambda})$, that is, $\lim_T \varphi_{K_T}(\lambda_{K_T}) = \inf \varphi$.

Finally, as regards the last point, we proceed by contradiction. If $(\varphi_T(\lambda_T))_{T\in\mathbb{N}}$ does not convergce to $\inf f$, then there exists an $\varepsilon > 0$ and a subsequence $(\varphi_{K_T}(\lambda_{K_T}))_{T\in\mathbb{N}}$ such that

$$|\varphi_{K_T}(\lambda_{K_T}) - \inf \varphi| \geq \varepsilon, \quad \forall T \in \mathbb{N} \qquad (11)$$

Now, let $(\lambda_{K_T^{(1)}})$ be a convergent subsequence of $(\lambda_{K_T})_{T\in\mathbb{N}}$. Suppose that $\lambda_{K_T^{(1)}} \to \bar{\lambda}$. Clearly $(\lambda_{K_T^{(1)}})$ is also a subsequence of $(\lambda_T)_{T\in\mathbb{N}}$. Then, it follows from point 2) above that $\varphi_{K_T^{(1)}}(\lambda_{K_T^{(1)}}) \to \inf \varphi$. This latter finding together with equation (11) gives a contradiction. $\square$

Proof of Theorem 3.2. Since $E(\cdot, \lambda)$ is uniformly Lipschitz continuous, there exists $\nu > 0$ such that for every $T \in \mathbb{N}$ and every $\lambda \in \Lambda$

$$|f_T(\lambda) - f(\lambda)| = |E(w_{T,\lambda}, \lambda) - E(w_\lambda, \lambda)|$$
$$\leq \nu \|w_{T,\lambda} - w_\lambda\|.$$

It follows from assumption (vi) that $f_T(\lambda)$ converges to $f(\lambda)$ uniformly on Λ as $T \to +\infty$. Then the statement follows from Theorem A.1 $\square$

B. Cross-validation and Bilevel Programming

We note that the (approximate) bilevel programming framework easily accommodates also estimations of the generalization error generated by a cross-validation procedures. We describe here the case of K-fold cross-validation, which includes also leave-one-out cross validation.

Let $D = \{(x_i, y_i)\}_{i=1}^n$ be the set of available data; K-fold cross validation, with $K \in \{1, 2, \ldots, N\}$ consists in partitioning D in K subsets $\{D^j\}_{j=1}^K$ and fit as many models g_{w^j} on training data $D_{\text{tr}}^j = \bigcup_{i\neq j} D^i$. The models are then evaluated on $D_{\text{val}}^j = D^j$. Denoting by $w = (w^j)_{j=1}^K$ the vector of stacked weights, the K-fold cross validation error is given by

$$E(w, \lambda) = \frac{1}{K} \sum_{j=1}^K E^j(w^j, \lambda)$$

where $E^j(w^j, \lambda) = \sum_{(x,y)\in D^j} \ell(g_{w^j}(x), y)$. E can be treated as the outer objective in the bilevel framework, while the inner objective L_λ may be given by the sum of regularized empirical errors over each D_{tr}^j for the K models.

Under this perspective, a K-fold cross-validation procedure closely resemble the bilevel problem for ML formulated in Sec. 2.2, where, in this case, the meta-distribution collapses on the data (ground) distribution and the episodes are sampled from the same dataset of points.

By following the procedure outlined in Sec. 2.3 we can approximate the minimization of L_λ with T steps of an optimization dynamics and compute the hypergradient of $f_T(\lambda) = \frac{1}{K}\sum_j E^j(w_T^j, \lambda)$ by training the K models and proceed with either forward or reverse differentiation. The models may be fitted sequentially, in parallel or stochastically. Specifically, in this last case, one can repeatedly sample one fold at a time (or possibly a mini-batch of folds) and compute a stochastic hypergradient that can be used in a SGD procedure in order to minimize f_T. At last, we note that similar ideas for leave-one out cross-validation error are developed in (Beirami et al., 2017), where the hypergradient of an approximation of the outer objective is computed by means of the implicit function theorem.

C. The Effect of T: Ridge Regression

In Sec. 5.1 we showed that increasing the number of iterations T leads to a better optimization of the outer objective f through the approximations f_T, which converge uniformly to f by Proposition 3.2. This, however, does not necessary results in better test scores due to possible overfitting of the training and validation errors, as we noted for the linear hyper-representation multiclass classification experiment in Sec. 5.1. The optimal choice of T appears to be, indeed, problem dependent: if, in the aforementioned experiment, a quite small T led to the best test accuracy (after hyperparameter optimization), in this section we present a small scale linear regression experiment that benefits from an increasing number of inner iterations.

We generated 90 noisy synthetic data points with 30 features, of which only 5 were informative, and divided them equally to form training, validation and test sets. As outer objective we set the mean squared error $E(w) = ||X_{\text{val}}w - y_{\text{val}}||^2$, where X_{val} and y_{val} are the validation design matrix and targets respectively and $w \in \mathbb{R}^{30}$ is the weight vector of the model. We set as inner inner objective

$$L_\lambda(w) = ||X_{\text{tr}}w - y_{\text{tr}}||^2 + \sum_{i=1}^{30} e^{\lambda_i} w_i^2$$

and optimized the L^2 vector of regularization coefficients λ (equivalent to a diagonal Tikhonov regularization matrix). The results, reported in Table 5, show that in this scenario overfitting is not an issue and 250 inner iterations yield the best test result. Note that, as in Sec. 5.1, also this problem admits an analytical solution, from which it is possible to compute the hypergradient analytically and perform exact

hyperparameter optimization, as reported in the last row of the Table 5.

Table 5. Validation and test mean absolute percentage error (MAPE) for various values of T.

T	Validation MAPE	Test MAPE
10	11.35	43.49
50	1.28	5.22
100	0.55	1.26
250	0.47	0.50
Exact	0.37	0.57

D. Further Details on Few-shot Experiments

This appendix contains implementation details of the representation mapping h and the meta-learning hyperparameters used in the few-shot learning experiments of Sec. 5.2.

To optimize the representation mapping, in all the experiments, we use Adam with learning rate set to 10^{-3} and a decay-rate of 10^{-5}. We used the initialization strategy proposed in (Glorot and Bengio, 2010) for all the weights λ of the representation.

For Omniglot experiments we used a meta-batch size of 32 episodes for five-way and of 16 episodes for 20-way. To train the episode-specific classifiers we set the learning rate to 0.1.

For one set of experiments with Mini-imagenet we used an hyper-representation (*C4L*) consisting of 4 convolutional layers where each layer is composed by a convolution with 32 filters, a batch normalization followed by a ReLU activation and a 2x2 max-pooling. The classifiers were trained using mini-batches of 4 episodes for one-shot and 2 episodes for five-shot with learning rate set to 0.01.

The other set of experiments with Mini-imagenet employed a Residual Network (*RN*) as the representation mapping, built of 4 residual blocks (with 64, 96, 128, 256 filters) and then the block that follows $\{1 \times 1$ conv (2048 filters), avg pooling, 1×1 conv (512 filters) $\}$. Each residual block repeats the following block 3 times $\{1 \times 1$ conv, batch normalization, leaky ReLU (leak 0.1)$\}$ before the residual connection. In this case the classifiers were optimized using mini-batches of 2 episodes for both one and five-shot with learning rate set to 0.04.