Mediterr. J. Math. (2023) 20:300
https://doi.org/10.1007/s00009-023-02497-w
1660-5446/23/060001-13
published online September 9, 2023

Mediterranean Journal of Mathematics

Theoretical Aspects in Penalty Hyperparameters Optimization

Flavia Esposito, Laura Selicato and Caterina Sportelli

Abstract. Learning processes play an important role in enhancing understanding and analyzing real phenomena. Most of these methodologies revolve around solving penalized optimization problems. A significant challenge arises in the choice of the penalty hyperparameter, which is typically user-specified or determined through Grid search approaches. There is a lack of automated tuning procedures for the estimation of these hyperparameters, particularly in unsupervised learning scenarios. In this paper, we focus on the unsupervised context and propose a bi-level strategy to address the issue of tuning the penalty hyperparameter. We establish suitable conditions for the existence of a minimizer in an infinite-dimensional Hilbert space, along with presenting some theoretical considerations. These results can be applied in situations where obtaining an exact minimizer is unfeasible. Working on the estimation of the hyperparameter with the gradient-based method, we also introduce a modified version of Ekeland's principle as a stopping criterion for these methods. Our approach distinguishes from conventional techniques by reducing reliance on random or black-box strategies, resulting in stronger mathematical generalization.

Mathematics Subject Classification. 68Q32, 46N10, 90C46, 49J27, 90C48.

Keywords. Hyperparameters optimization, learning approaches, existence results.

1. Introduction

Learning models are important approaches successfully applied in real-life applications. These processes often require the specification of several variables by the users, namely hyperparameters, which must be set before the learning procedure starts. Hyperparameters govern the whole learning process and play a crucial role in guaranteeing good model performances. They are often manually specified, and the lack of an automatic tuning procedure makes

This work was completed with the support of our TEX-pert.

the field of hyperparameter optimization (HPO) an ever-evolving topic. The literature offers various solutions for hyperparameters tuning, from gradient-based to black-box or Bayesian approaches, besides some naive but daily used methods such as Grid and Random search. A brief overview of existing methods can be found in Ref. [6]. Hyperparameters can be of different types (discrete, continuous, categorical), and in most cases, the number of their configurations to explore is infinite. This paves the way for a mathematical formalization of the HPO in learning contexts with abstract spaces, such as Hilbert spaces.

A learning algorithm may be represented as a map $\mathscr{A}$ that takes a configuration of hyperparameters, $\lambda \in \Lambda$, and a dataset D, and returns a hypothesis $h \in \mathscr{H}$:

$$\mathscr{A} : \Lambda \times D \to \mathscr{H}; \qquad \mathscr{A}(\lambda, D) = h, \tag{1.1}$$

where Λ is a hyperparameter space, and $\mathscr{H}$ is a hypothesis space [11]. A quite standard claim for the hypotheses set is to be a linear function space, endowed with a suitable norm (more binding arising from an inner product): two requirements satisfied when $\mathscr{H}$ is a Hilbert space of functions over the input space $\mathscr{X}$.[1] Assuming a Hilbert space structure on the hypothesis space has some advantages: (i) practical computations reduced to ordinary linear algebra operations and (ii) self-duality; that is for any $x \in \mathscr{X}$ a representative of x can be found, i.e., $k_x \in \mathscr{H}$ exists such that

$$h(x) = \langle k_x, h \rangle \quad \text{for all } h \in \mathscr{H}, \tag{1.2}$$

where k_x is a suitable positive definite "kernel". This construction gives the chance to connect the abstract structure of $\mathscr{H}$ and what its elements actually are, flipping the construction of the hypotheses set from the kernel. Providing a suitable positive function k on $\mathscr{X}$, $\mathscr{H}$ can be set as the minimal complete space of functions involving all $\{k_x\}_{x \in \mathscr{X}}$ equipped with the scalar product in (1.2). Thus, $\mathscr{H}$ is outlined in a unique way, and it is named the Representing Kernel Hilbert Space mapped to the kernel k.

Starting from this abstract scenario, HPO can be formulated as the problem of minimizing the goodness of the solution given by the algorithm $\mathscr{A}$ that, implicitly or explicitly, depends on the hyperparameter λ. In particular, for supervised learning contexts, the optimal hyperparameter λ^* can be found in the literature as the solution for the following optimization task:

$$\lambda^* = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \, \mathscr{V}(\mathscr{A}(\lambda, D_{tr}), D_{val}), \tag{1.3}$$

where $\mathscr{V} : \mathscr{H} \times \mathscr{X} \to \mathbb{R}$ evaluates the goodness of $\mathscr{A}$ measuring discrepancy between $\mathscr{A}$ on a given training dataset D_{tr}, and a validation dataset D_{val} [10].

In this study, we will focus on HPO in unsupervised context, by using bi-level programming formalization. Bi-level approaches solve an outer optimization problem subject to the optimality of an inner optimization problem [1,3,5,11]. In particular, we will consider as a hyperparameter the penalty

[1] If X is an infinite-dimensional space the boundedness is needed, too.

coefficient in penalized optimization problems. It is important to note that penalization functions are essential tools in optimization and learning problems. They are used to introduce a bias towards simpler or more general solutions. In particular, they can help to prevent overfitting or enforce feature selection operations while controlling the sparsity, to stabilize the solution and prevent noise amplification in inverse problems with regularization conditions, deal with multicollinearity in regression models, or to improve visualization tasks with orthogonality constraints. We already treat this aspect and solve the problem in the specific case of the Nonnegative Matrix Factorization task [7]. By the way, some generalizations are needed to overcome theoretical restrictions and made the strategy broadly and cross-sectional applicable to other learning approaches. In particular, this work extends the existence and uniqueness theorems for the solution of the hyperparameters bi-level problem to the more general framework of infinite-dimensional Hilbert space. This latter also allows the application of Ekeland's variational principle to state that whenever a functional is not guaranteed to have a minimum, under suitable assumptions, a "good" substitute can be found, namely the best one can get as an approximate minimum. One of the purposes of this paper is to use this theoretical tool as a stopping criterion for the update of the hyperparameters as we will see later.

The outline of the paper is as follows. Section 2 introduces the classical bi-level formalization of HPO and some preliminary notions in a supervised context. Section 3 illustrates our proposal, an extension of the unsupervised context. A general framework addressing HPO in Hilbert space is also set, and some general abstract tools are stated in Sect. 4. Finally, Sect. 5 summarizes the obtained results and draws some conclusions and future works.

2. Previous Works and Preliminaries

As briefly mentioned in the introduction, in a supervised learning scenario, HPO can be addressed through a bi-level formulation. This approach looks for the hyperparameters λ such that the minimization of the regularized training leads to the best performance of the trained data-driven model on a validation set. Accordingly, to the ideas introduced in [12,20], the best hyperparameters for a data learning task can be selected as the solution to the following bi-level problem:

$$\min_{\lambda\in\Lambda} J(\lambda) \qquad J(\lambda) = \inf\{\mathscr{E}(w_\lambda,\lambda) : w_\lambda \in \operatorname*{argmin}_{u\in\mathbb{R}^r} \mathscr{L}_\lambda(u)\}, \tag{2.1}$$

where $w \in \mathbb{R}^r$ are r parameters, $J : \Lambda \to \mathbb{R}$ is the so-called *Response Function* of the outer problem with objective function $\mathscr{E} : \mathbb{R}^r \times \Lambda \to \mathbb{R}$, and for every $\lambda \in \Lambda \subset \mathbb{R}^p$, $\mathscr{L}_\lambda : \mathbb{R}^r \to \mathbb{R}$ is the inner objective function.
One way to solve the reformulation of HPO as a bi-level optimization problem is the adoption of gradient-based (GB) methods. In particular, in GB methods, HPO is addressed with classical procedure for continuous optimization, in which the sequence is generated by the following rule:

$$\lambda_{t+1} = \lambda_t - \alpha \mathbf{h}_t(\lambda), \tag{2.2}$$

where $\mathbf{h}_t$ is an approximation of the gradient of the function J and α is a step size, that converges to the optimal hyperparameter. In this context, it is known that the main challenge is the computation of $\mathbf{h}_t$, called hypergradient. In several cases, this numerical approximation can be calculated for real-valued hyperparameters with iterative algorithms. There are two main strategies for computing the hypergradient: iterative differentiation [12,13,17] and implicit differentiation [16,18]. The former requires calculating the exact gradient of an approximate objective. This is defined through the recursive application of optimization dynamics that aims to replace and approximate the learning algorithm $\mathscr{A}$. The latter involves the numerical application of the implicit function theorem to the solution mapping $\mathscr{A}$ when it is expressible through an appropriate equation [11].

In this study, we follow the iterative strategy, so that problem in (2.1) can be addressed through a dynamical system type approach.

If the following hypothesis hold:

Hypothesis 1. *1. the set Λ is a compact subset of $\mathbb{R}$;*
2. the Error Function $\mathscr{E} : \mathbb{R}^r \times \Lambda \to \mathbb{R}$ is jointly continuous;
3. the map $(w, \lambda) \to \mathscr{L}_\lambda(w)$ is jointly continuous, and then the problem $\operatorname{argmin} \mathscr{L}_\lambda$ *is a singleton for every $\lambda \in \Lambda$;*
4. $w_\lambda = \operatorname*{argmin}_{u \in \mathbb{R}^r} \mathscr{L}_\lambda(u)$ remains bounded as λ varies in Λ;

the problem in (2.1) becomes:

$$\min_{\lambda \in \Lambda} J(\lambda) = \mathscr{E}(w_{\lambda^*}, \lambda^*), \quad w_\lambda = \operatorname*{argmin}_{u \in \mathbb{R}^r} \mathscr{L}_\lambda(u). \tag{2.3}$$

It can be proved that the optimal solution $(w_{\lambda^*}, \lambda^*)$ of problem (2.3) exists [13].

Considering the optimization problem in which hyperparameter is the penalty coefficient $\lambda \in \mathbb{R}_+$, the *Inner Problem* is associated with the penalized empirical error represented by $\mathscr{L}$, defined as

$$\mathscr{L}_\lambda(w) = \sum_{(x,y) \in D_{tr}} \ell(g_w(x), y) + \lambda \mathscr{r}(w), \tag{2.4}$$

where ℓ is a loss function, $g_w : \mathscr{X} \to \mathscr{Y}$, is a parameterized model from the input to the output space, $D_{tr} \subset \mathscr{X} \times \mathscr{Y}$ the training set, and $\mathscr{r} : \mathbb{R}^r \to \mathbb{R}$ is a penalty function. While the *Outer Problem* is related to the generalized error of g_w represented by $\mathscr{E}$:

$$\mathscr{E}(w, \lambda) = \sum_{(x,y) \in D_{val}} \ell(g_w(x), y), \tag{2.5}$$

where $D_{val} \subset \mathscr{X} \times \mathscr{Y}$ is the validation set. Note that $\mathscr{E}$ does not explicitly depend on λ.

This work will allow us to extend these issues to the unsupervised context overcoming some assumptions of Hypothesis 1 (such as compactness) that are difficult to satisfy in real data learning contexts, and also to use some theoretical results as Ekeland's variational principle, stated in the following section.

3. Our Proposal

The bi-level HPO framework can be modified to include unsupervised learning paradigms, generally designed to detect some useful latent structure embedded in data. Tuning hyperparameters for unsupervised learning models is more complex than the supervised case due to the lack of output space, which defines the ground truth collected in the validation set.

This section describes a general framework to address HPO in Hilbert spaces for the unsupervised case and a corollary of Ekeland's variational principle used to derive a useful stopping criterion for iterative algorithms solving the HPO problem.

Let $X \in \mathbb{R}^{n\times m}$ be a data matrix, with reference to the problem (2.1), where now $J : \Lambda \to \mathbb{R}$ is a suitable functional and Λ a Hilbert space equipped with the scalar product $(\cdot,\cdot)$. With these presumptions, the outer problem is defined by the following function:

$$\mathscr{E} : \mathbb{R}^r \times \Lambda \to \mathbb{R} \qquad \mathscr{E}(w,\lambda) = \sum_{x\in X} \ell(g_w(x)), \tag{3.1}$$

and for every $\lambda \in \Lambda$, the inner relate problem is

$$\mathscr{L}_\lambda : \mathbb{R}^r \to \mathbb{R} \qquad \mathscr{L}_\lambda(w) = \sum_{x\in X} \ell(g_w(x)) + \mathscr{R}(\lambda, w), \tag{3.2}$$

where $\mathscr{R} : \Lambda \times \mathbb{R}^r \to \mathbb{R}$ is a penalty function. We want to emphasize that in this new formulation, all optimization is performed on the data matrix X, and the penalty hyperparameter has been included as a variable in the penalty function $\mathscr{R}$. With this new formulation, the process of optimizing the hyperparameters is integrated directly into the broader optimization problem. This integration may streamline the optimization process and improve the overall efficiency of finding the best hyperparameters for the given problem. Furthermore, by partitioning the reference matrix X, it becomes possible to penalize each partition of the matrix with different magnitudes, potentially leading to better model performance or more refined results.

The bi-level problem associated with (3.1)–(3.2) can be solved with a dynamical system approach in which a numerical approximation of the hypergradient is computed. Once the hypergradient is achieved a GB approach can be used to find the optimum λ^*.

Ekeland's variational principle can be used to construct an appropriate stopping criterion for iterative algorithms, with the aim of justifying and setting the hyperparameters related to the stopping criterion more appropriately.

Theorem 3.1. (Ekeland's variational principle) [9] *Let (Λ, d) be a complete metric space and $J : \Lambda \to \bar{\mathbb{R}}$ be a lower semi-continuous function which is bounded from below. Suppose that $\varepsilon > 0$ and $\tilde{\lambda} \in \Lambda$ exist such that*

$$J(\tilde{\lambda}) \le \inf_{\Lambda} J + \varepsilon.$$

Then, given any $\rho > 0$, $\lambda_\rho \in \Lambda$ exists such that

$$J(\lambda_\rho) \le J(\tilde{\lambda}), \qquad d(\lambda_\rho, \tilde{\lambda}) \le \frac{\varepsilon}{\rho},$$

and

$$J(\lambda_\rho) < J(\lambda) + \rho\, d(\lambda_\rho, \lambda) \qquad \forall\, \lambda \neq \lambda_\rho.$$

Roughly speaking, this variational principle asserts that, under assumptions of lower semi-continuity and boundedness from below, if a point $\tilde{\lambda}$ is an "almost minimum point" for a function J, hence a small perturbation of J exists which attains its minimum at a point "near" to $\tilde{\lambda}$. It is important to note that a variation of the Theorem 3.1 can be used to reduce the number of user-dependent factors for the stopping criterion. In particular, a fruitful selection of ρ (for $\rho = \sqrt{\varepsilon}$) restricts the number of hyperparameters to the precision error only, allowing us to use the following corollary.

Corollary 3.2. *Let (Λ, d) be a complete metric space and $J : \Lambda \to \bar{\mathbb{R}}$ be a lower semi-continuous function which is bounded from below. Suppose that $\varepsilon > 0$ and $\tilde{\lambda} \in \Lambda$ exist such that*

$$J(\tilde{\lambda}) \leq \inf_{\Lambda} J + \varepsilon.$$

Then, $\tilde{z} \in \Lambda$ exists such that

$$J(\tilde{z}) \leq J(\tilde{\lambda}), \qquad d(\tilde{z}, \tilde{\lambda}) \leq \sqrt{\varepsilon}$$

and

$$J(\tilde{z}) < J(\lambda) + \sqrt{\varepsilon}\, d(\tilde{z}, \lambda) \quad \forall\, \lambda \neq \tilde{z}.$$

4. Main Abstract Results

In this section, we are ready t o weaken the assumptions we discussed earlier and provide results related to the use of Ekeland's principle as a stopping criterion. We mention an abstract result of the existence of a minimizer in Hilbert spaces which has great importance and a wide range of applications in several fields. Just one example is represented by Riesz's Representation Theorem, that, even if implicitly, makes use of the existence of a minimizer [4]. This is a widely relevant issue about Hilbert spaces, which makes them nicer than Banach spaces or other topological vector spaces.

4.1. Abstract Existence Theorem

It is well known that each bounded sequence in a normed space Λ has a norm convergent subsequence if and only if it is a finite-dimensional normed space.

Thus, given a normed space Λ, as the strong topology (i.e., the one induced by the norm) is too strong to provide any widely appropriate subsequential extraction procedure, one can consider other weak topologies joined with the linear structure of the space and look for subsequential extraction processes therein.

In Banach spaces, as well as in Hilbert spaces, the two most relevant weaker-than-norm topologies are the weak-star topology and the weak topology. As the former is established in dual spaces, the latter is set up in every normed space. The notions of these topologies are not self-contained but fulfill a leading role in many features of the Banach space theory. In this regard,

here we state some results we will use shortly. The next one is straightforward (see, e.g., [4, Chapter 3]).

Proposition 4.1. *If Λ is a finite-dimensional space, the strong and weak topologies coincide. In particular, it follows that the weak topology is normable, and then clearly metrizable, too.*

If Λ is an infinite-dimensional space, the weak topology is strictly contained in the strong topology, namely open sets for the strong topology exist which are not open for the weak topology. Furthermore, the weak topology turns out to be not metrizable in this case.

Definition 4.2. A functional $J : \Lambda \to \bar{\mathbb{R}}$ with Λ topological space, is said to be lower semi-continuous on Λ if for each $a \in \mathbb{R}$, the sublevel sets

$$J^{-1}(]-\infty, a]) = \{\lambda \in \Lambda : J(\lambda) \leq a\}$$

are closed subsets of Λ.

In the following, we introduce a "generalized Weierstrass Theorem" which gives a criterion for the existence of a minimum for a functional defined on a Hilbert space. For this reason, the incoming results will be provided for the abstract framework of a Hilbert space although, in some cases, they apply in the more general context of Banach spaces. Thus, throughout the remaining part of this section, we denote by Λ any real infinite-dimensional Hilbert space.

In an infinite-dimensional setting, the following definitions are strictly related to the different notions of weak and strong topology.

Definition 4.3. A functional $J : \Lambda \to \bar{\mathbb{R}}$ is said to be strongly (weakly, respectively) lower semi-continuous if J is lower semi-continuous when Λ is equipped with the strong (weak, respectively) topology.

Definition 4.4. A functional $J : \Lambda \to \bar{\mathbb{R}}$ is said to be strongly (weakly, respectively) sequentially lower semi-continuous if

$$\liminf_{n \to +\infty} J(\lambda_n) \geq J(\lambda)$$

for any sequence $(\lambda_n)_n \subset \Lambda$ such that $\lambda_n \to \lambda$ ($\lambda_n \rightharpoonup \lambda$, respectively).

We proceed by providing some useful results.

Proposition 4.5. *The following statements are equivalent:*

(i) $J : \Lambda \to \mathbb{R}$ is sequentially weakly lower semi-continuous functional;

(ii) the epigraph of J is weakly sequentially closed, where, by definition, it is

$$\mathrm{epi}(J) = \{(\lambda, t) \in \mathrm{dom}(J) \times \mathbb{R} : J(\lambda) \leq t\}.$$

Remark 4.6. As a further consequence of the preliminary Proposition 4.1, we have that sequential weak lower semi-continuity and weak lower semi-continuity do not match if Λ is infinite-dimensional since weak topology is not metrizable. However, the weaker concept of sequential weak lower semi-continuity meets our needs. For the proof of the next result, we refer the interested reader to [2, Theorem 3.32].

Proposition 4.7. *Let $\mathscr{C} \subseteq \Lambda$ be a closed and convex subset. Then, $\mathscr{C}$ is weakly sequentially closed, too.*

Since a sequentially weakly closed set is also strongly closed, it follows that a sequentially weakly lower semi-continuous functional is also (strongly) lower semi-continuous. Instead, the converse holds under an additional assumption. In particular, Proposition 4.7 allows us to infer the following results.

Proposition 4.8. *If $J : \Lambda \to \mathbb{R}$ is a strongly lower semi-continuous convex functional; thus J is weakly sequentially lower semi-continuous, too.*

Proof. Since J is lower semi-continuous, thus $\mathrm{epi}(J)$ is closed. On the other hand, since J is convex, so it is $\mathrm{epi}(J)$, hence Proposition 4.7 ensures that $\mathrm{epi}(J)$ is weakly sequentially closed, i.e., J is weakly sequentially lower semi-continuous. □

Thus, we are able to state the main result of this section.

Theorem 4.9. *Let $\mathscr{C} \subset \Lambda$ be a non-empty, closed, bounded, and convex subset. Let $J : \Lambda \to \mathbb{R}$ be a lower semi-continuous and convex functional. Thus J achieves its minimum in $\mathscr{C}$, i.e., $\bar{\lambda} \in \mathscr{C}$ exists such that $J(\bar{\lambda}) = \inf_{\lambda \in \mathscr{C}} J(\lambda)$.*

Proof. Let $m := \inf_{\lambda \in \mathscr{C}} J(\lambda)$; hence, $(\lambda_n)_n \subset \mathscr{C}$ exists such that

$$J(\lambda_n) \to m \quad \text{as } n \to +\infty. \tag{4.1}$$

Now, our boundness assumption on $\mathscr{C}$ implies that, up to subsequences, $\lambda \in \mathscr{C}$ exists such that $\lambda_n \rightharpoonup \lambda$ as $n \to +\infty$. Actually, since $\mathscr{C}$ is a closed and convex subset of Λ, thus Proposition 4.7 applies, which guarantees that $\lambda \in \mathscr{C}$.

Finally, from (4.1), Proposition 4.8 and Definition 4.4 we infer that $J(\bar{\lambda}) \leq m$, which gives the desired result. □

Remark 4.10. We observe that Theorem 4.9 still holds if the subset $\mathscr{C}$ is not bounded as long as we ask for an additional assumption on the functional J. In fact, requiring J to be coercive[2] (and if at least $\bar{\lambda} \in \mathscr{C}$ exists such that $J(\bar{\lambda}) < +\infty$), then any minimizer of J on $\mathscr{C}$ necessarily lies in some closed ball of radius $r > 0$. In fact, since $J(\bar{\lambda}) < +\infty$, any minimizer λ of J must have $J(\lambda) \leq J(\bar{\lambda})$; furthermore, since J is coercive, a sufficient large radius $r > 0$ exists such that $J(\lambda) > J(\bar{\lambda})$ for all $\lambda \in \mathscr{C}$ with $\|\lambda\| > r$. Thus, any minimizer, if exists, lies in the ball $\{\lambda \in \mathscr{C} : \|\lambda\| \leq r\}$.
In particular, Theorem 4.9 applies to the intersection between $\mathscr{C}$ and a closed ball of suitable radius, since it turns to be convex if we formally require $\mathscr{C}$ to be closed and convex.

Namely, the following result holds.

Corollary 4.11. *Let $\mathscr{C} \subset \Lambda$ be a non-empty, closed, and convex subset. Let $J : \Lambda \to \mathbb{R}$ be a lower semi-continuous, convex, and coercive functional. Thus J achieves its minimum, i.e., $\bar{\lambda} \in \mathscr{C}$ exists such that $J(\bar{\lambda}) = \inf_{\lambda \in \mathscr{C}} J(\lambda)$.*

[2] We say that a functional $J : H \to \mathbb{R}$ is coercive if $J(u) \to \infty$ as $\|u\| \to \infty, u \in H$.

Now we introduce a couple of results that are a direct consequence of Ekeland's variational principle. For the sake of completeness, here we provide them with all the details (see [8] for the original statements).

Let Λ be a complete metric space and $J : \Lambda \to \mathbb{R}$ be the lower semi-continuous response function on Λ. Suppose that a point $\lambda \in \Lambda$ exists such that $J(\lambda) < +\infty$. Thus, the following results hold.

Theorem 4.12. (Perturbation Result) *Let $J_\lambda : \Lambda \to \bar{\mathbb{R}}$ be a lower semi-continuous differentiable function such that the inequality*

$$|J_\lambda(\gamma) - J(\gamma)| \leq \zeta(d(\gamma, \lambda)) \quad \textit{holds} \quad \forall \gamma \in \Lambda, \tag{4.2}$$

where $J_\lambda(\cdot)$ denote model function,[3] *ζ is some growth function,*[4] *and let λ^+ be a minimizers of J_λ. If λ^+ coincides with λ, then $|\nabla J(\lambda)| = 0$. On the other hand, if λ and λ^+ are distinct, then a point $\hat{\lambda} \in X$ exists which satisfies*

1. $d(\lambda^+, \hat{\lambda}) \leq 2 \cdot \frac{\zeta(d(\lambda^+, \lambda))}{\zeta'(d(\lambda^+, \lambda))}$ *(point proximity)*
2. $J(\hat{\lambda}) \leq J(\lambda^+) + \zeta(d(\lambda^+, \lambda))$ *(value proximity).*

Proof. By Taylor's theorem, it is simple to verify that $|\nabla J_\lambda|(\lambda) = |\nabla J|(\lambda)$. Now, since λ is a minimizer, we have $|\nabla J(\lambda)| = 0$ if $\lambda^+ = \lambda$. On the other hand, if $\lambda^+ \neq \lambda$, from inequality (4.2) and the definition of λ^+, it follows that

$$J(\gamma) \geq J_\lambda(\lambda^+) - \zeta(d(\gamma, \lambda)).$$

Let us define the new function

$$G(\gamma) := J(\gamma) + \zeta(d(\gamma, \lambda)).$$

Thus, from assumption (4.2) and inequality $\inf G \geq J_\lambda(\lambda^+)$, we infer that

$$G(\lambda^+) - \inf G \leq J(\lambda^+) - J_\lambda(\lambda^+) + \zeta(d(\lambda^+, \lambda)) \leq 2\zeta(d(\lambda^+, \lambda)).$$

Hence, Theorem 3.1 applies and, having $\varepsilon := 2\zeta(d(\lambda^+, \lambda))$, for all $\rho > 0$ λ_ρ exists such that

$$G(\lambda_\rho) \leq G(\lambda^+) \quad \text{and} \quad d(\lambda^+, \lambda_\rho) \leq \frac{\varepsilon}{\rho}.$$

The desired result follows simply by placing $\rho = \zeta'(d(\lambda^+, \lambda))$ with $\hat{\lambda} = \lambda_\rho$. □

An immediate consequence of Theorem 4.12 is the following subsequence convergence result.

Corollary 4.13. (Subsequence convergence to stationary points) *Consider a sequence of points λ_k and closed functions $J_{\lambda_k} : \Lambda \to \bar{\mathbb{R}}$ satisfying $\lambda_{k+1} = \operatorname{argmin}_\gamma J_{\lambda_k}(\gamma)$ and $d(\lambda_{k+1}, \lambda_k) \to 0$. Moreover, suppose that the inequality*

$$|J_{\lambda_k}(\gamma) - J(\gamma)| \leq \zeta(d(\lambda_k, \gamma)) \quad \textit{holds} \quad \forall k \in \mathbb{N} \quad \textit{and} \quad \gamma \in \Lambda, \tag{4.3}$$

[3] As *model function* we mean the Taylor's expansion of J in λ, stopped to the first order.
[4] A differentiable univariate function $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ is called a growth function if it satisfies $\zeta(0) = \zeta'(0) = 0$ and $\zeta' > 0$ on $(0, +\infty)$. If in addition, equalities $\lim_{t\to 0} \zeta'(t) = \lim_{t\to 0} \zeta(t)/\zeta'(t) = 0$ hold, we say that ζ is a proper growth function.

where ζ is a proper growth function. If $(\lambda^*, J(\lambda^*))$ is a limit point of the sequence $(\lambda_k, J(\lambda_k))$, then λ^* is stationary for J.

Two interesting consequences for convergence analysis flow from there. Suppose that the models are chosen in such a way that the step-sizes $\|\lambda_{k+1} - \lambda_k\|$ tend to zero. This assumption is often enforced by ensuring that $J(\lambda_{k+1}) < J(\lambda_k)$ by at least a multiple of $\|\lambda_{k+1} - \lambda_k\|^2$ (sufficient decrease condition). Then, assuming for simplicity that J is continuous on its domain, any limit point λ^* of the iterate sequence λ_k will be stationary for the problem (Corollary 4.13).

Thus, by choosing an error ε, we can stop update (2.2) for GB algorithms in the context of bi-level HPO for penalty hyperparameter, according to the pseudo-code described in Algorithm 1.

Algorithm 1 Pseudo-code

Require: Tolerance ε. Some starting points λ_0, λ_1.
Ensure: Optimum λ^*
1: **while** $\|\lambda_t - \lambda_{t-1}\| > \varepsilon$ **do**
2: Compute the hypergradient with iterative differentiation $\mathbf{h}(\lambda)$;
3: update λ_{t+1} according to GB approach (2.2);
4: Increment the iteration t.
5: **end while**

5. Conclusions

In this paper, we studied the task of penalty HPO and we provided a mathematical formulation, based on Hilbert spaces, to address this issue in an unsupervised context. We want to emphasize that moving to infinite-dimensional Hilbert spaces is not a mere abstract pretense, but it is also widely used in supervised contexts. For example, when Support Vector Machine (SVM) is taken into consideration, a well-known "kernel trick" permits the interpretation of a Gaussian kernel as an inner product in a feature space. This is potentially infinite-dimensional, allowing us to read the SVM classifier function as a linear function in the feature space [19]. Another example is provided by the quantum system possible states problem, in which the state of a free particle can be described as vectors residing in a complex separable Hilbert space [21].

In this work, we considered as hyperparameter the penalty coefficient of the constrained objective function and set up a bi-level strategy for its automatic tuning. Indeed, the strength of this article lies in theory. We showed some relaxed theoretical results both to weaken the hypotheses necessary for the existence of the solution and also proposed a variant of Ekeland's principle as a stopping criterion of GB methods. Our approach differs from the more standard techniques in reducing the random or black-box strategies giving stronger mathematical generalization suitable also when it is not possible to

obtain an exact minimizer. Both the existence theorem and the stopping criterion allow us to build an approach based on solid mathematical foundations useful for future extensions and generalizations to other problems, too. For example, infinite-dimensional Covariance Descriptors (CovDs) for classification are a fertile application arena for the extensions developed here. This finds motivation in the fact that CovDs could be mapped to Reproducing Kernel Hilbert Space (RKHS) via the use of SPD-specific kernels [14]. Also, the generalization of this approach related to a particular constrained matrix factorization problem, defined with Bregman divergence on Hilbert spaces, are subject of future works with experiments evaluating the goodness of the novel stopping criterion [15].

Author Contributions All the authors contributed equally to this work.

Funding Open access funding provided by Universitá degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement. The author F. E. was funded by REFIN Project, grant number 363BB1F4, Reference project idea UNIBA027 "Un modello numerico-matematico basato su metodologie di algebra lineare e multilineare per l'analisi di dati genomici". The author C. S. was partially supported by PRIN project "Qualitative and quantitative aspects of nonlinear PDEs" (2017JPCAPN_005) funded by Ministero dell'Istruzione, dell'Università e della Ricerca.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose.

References

[1] Bard, J.F.: Practical Bilevel Optimization: Algorithms and Applications, vol. 30. Springer, Boston (2013)

[2] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, New York (2011)

[3] Bertrand, Q., Klopfenstein, Q., Blondel, M., Vaiter, S., Gramfort, A., Salmon, J.: Implicit differentiation of lasso-type models for hyperparameter optimization. In: International Conference on Machine Learning, PMLR, pp. 810–821 (2020)

[4] Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2010)

[5] Colson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Ann. Oper. Res. **153**(1), 235–256 (2007)

[6] Del Buono, N., Esposito, F., Selicato, L.: Methods for hyperparameters optimization in learning approaches: An overview. In: International Conference on Machine Learning, Optimization, and Data Science, Springer, pp. 100–112 (2020)

[7] Del Buono, N., Esposito, F., Selicato, L., Zdunek, R.: Bi-level algorithm for optimizing hyperparameters in penalized nonnegative matrix factorization. Appl. Math. Comput. **457**, 128184 (2023)

[8] Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. Math. Progr. (2019). https://doi.org/10.1007/s10107-019-01432-w

[9] Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**(2), 324–353 (1974)

[10] Feurer, M., Hutter, F.: Hyperparameter optimization. In: Automated Machine Learning. Springer, Cham (2019)

[11] Franceschi, L.: A unified framework for gradient-based hyperparameter optimization and meta-learning. Ph.D. thesis, University College London (2021)

[12] Franceschi, L., Donini, M., Frasconi, P., Pontil, M.: Forward and reverse gradient-based hyperparameter optimization. In: International Conference on Machine Learning, PMLR, pp. 1165–1173 (2017)

[13] Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., Pontil, M.: Bilevel programming for hyperparameter optimization and meta-learning. In: International Conference on Machine Learning, PMLR, pp. 1568–1577 (2018)

[14] Harandi, M., Salzmann, M., Porikli, F.: Bregman divergences for infinite dimensional covariance matrices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1003–1010 (2014)

[15] Kim, Kyungsup: Understanding non-negative matrix factorization in the framework of Bregman divergence. J. Korean Soc. Ind. Appl. Math. **25**(3), 107–116 (2021)

[16] Lorraine, J., Vicol, P., Duvenaud, D.: Optimizing millions of hyperparameters by implicit differentiation. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 1540–1552 (2020)

[17] Maclaurin, D., Duvenaud, D., Adams, R.: Gradient-based hyperparameter optimization through reversible learning. In: Proc. of ICML, pp. 2113–2122 (2015)

[18] Pedregosa, F.: Hyperparameter optimization with approximate gradient, pp. 737–746. PMLR, ICML (2016)

[19] Rossi, F., Villa, N.: Classification in Hilbert spaces with support vector machines. In: Proceedings of ASMDA, pp. 635–642 (2005)

[20] Vincent, D., Gelly, S., Nicolas Le, R., Bousquet, O.: Online hyper-parameter optimization (2018)

[21] Ying, M.: Foundations of Quantum Programming. Morgan Kaufmann, San Francisco (2016)

Flavia Esposito and Laura Selicato
Department of Mathematics
Università degli Studi di Bari Aldo Moro
via Orabona, 4
70125 Bari
Italy
e-mail: flavia.esposito@uniba.it

Laura Selicato
e-mail: laura.selicato@uniba.it

Caterina Sportelli
Department of Mathematics and Statistics
University of Western Australia
35 Stirling Highway
Crawley
WA 6009
Australia
e-mail: caterina.sportelli@uwa.edu.au

Received: May 23, 2023.
Revised: August 2, 2023.
Accepted: August 18, 2023.