Relax and penalize: a new bilevel approach to mixed-binary hyperparameter optimization

Marianna de Santis, Jordan Frecon, Francesco Rinaldi, Saverio Salzo, Martin Schmidt

Relax and penalize: a new bilevel approach to mixed-binary hyperparameter optimization

A Preprint

Marianna De Santis* **Jordan Frecon**† **Francesco Rinaldi**‡ **Saverio Salzo*** **Martin Schmidt**§

Abstract

In recent years, bilevel approaches have become very popular to efficiently estimate high-dimensional hyperparameters of machine learning models. However, to date, binary parameters are handled by *continuous relaxation and rounding* strategies, which could lead to inconsistent solutions. In this context, we tackle the challenging optimization of mixed-binary hyperparameters by resorting to an equivalent continuous bilevel reformulation based on an appropriate penalty term. We propose an algorithmic framework that, under suitable assumptions, is guaranteed to provide mixed-binary solutions. Moreover, the generality of the method allows to safely use existing continuous bilevel solvers within the proposed framework. We evaluate the performance of our approach for a specific machine learning problem, i.e., the estimation of the group-sparsity structure in regression problems. Reported results clearly show that our method outperforms state-of-the-art approaches based on *relaxation and rounding*.

1 Introduction

Nowadays, machine learning systems tend to incorporate an increasing number of hyperparameters with the purpose of improving the overall performance of learning tasks and achieving a higher flexibility. Then, optimizing such high-dimensional hyperparameters becomes a crucial step for devising efficient and fully parameter-free machine learning systems. In recent years, bilevel approaches to hyperparameter optimization have become very popular as an effective way to estimate high-dimensional hyperparameters [1, 2, 3, 6, 10, 16, 18]. On the other hand, in many circumstances binary hyperparameters are included in the model to allow the pruning of the irrelevant variables or the discovery of sparsity structures. Interesting examples are given by the pruning of large-scale deep learning models [22], the identification of the group-sparsity structures in regression problems [8, 20], and learning the discrete structure of a graph neural networks [7]. For these cases the usual optimization approach is that of *relaxing* the respective parameter over the unit interval $[0, 1]$, solve the continuous optimization problem, and then *rounding* the solution so to get a binary output. This is essentially a heuristic, which overcomes the challenge of dealing with integer variables, but does not offer any theoretical guarantee.

The aim of the present work is that of providing a more principled way of approaching mixed-binary hyperparameter optimization.

Related works. In the context of machine learning, bilevel optimization problems with binary variables arise in a number of situations. In [8] the estimation of group-sparsity structure in multi-task regression is addressed by a mixed-binary bilevel optimization model, which is handled by a continuous relaxation and approximation of the problem. The output of the optimization procedure is a vector of continuous variables, which are then rounded to the closest binary values, so to provide the final grouping of the features. In Section 5, we tackle this same problem and show the advantage of our approach. In the work [22], a new model pruning, based on bilevel optimization, is proposed, where the upper level variable is a binary mask. The related iterative algorithm performs a gradient descent

*DIAG, Sapienza University of Rome, 00185 Rome, Italy

†Université Jean Monnet Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

‡Department of Mathematics, University of Padova, 35121 Padova, Italy

§Department of Mathematics, Trier University, 54296 Trier, Germany

step on the continuous relaxation of the problem followed by a projection step onto a discrete set, which is indeed a hard-thresholding operation. No convergence guarantees are provided.

Beyond the machine learning literature, there is a number of works related to mixed-integer bilevel programming (see, e.g., Section 5.3 of the recent survey [14]). An important drawback one needs to take into account when applying those methods to hyperparameter optimization problems is however their limited scalability. Common approaches like, e.g., the outer-approximation based method in [13], or the algorithm proposed in [17], which requires the global solution of a significant number of mixed integer nonlinear programs, have indeed a prohibitive cost when the dimensionality grows. Furthermore, it should be noted that they aim to achieve global optimality—an overly ambitious goal in the context of machine learning applications.

Contributions and Outline. In this paper, we analyze more carefully the mixed-binary setting and propose a *relax and penalize* method, which produces a mixed-binary output and relies on improved mathematical grounds. More precisely, we consider in Section 3 a general mixed continuous-binary bilevel problem and show that it is equivalent, in terms of global minima and minimizers, to a fully continuous and penalized optimization problem. Next, in Section 4, we propose an algorithmic framework, which consists of iteratively solving a sequence of continuous and penalized problems which, under suitable assumptions, is guaranteed to provide mixed-binary local solutions. The performance of the proposed approach are quantitatively assessed on the estimation of the group structure in the group lasso problem. Numerical experiments are reported in Section 5 and show how the *relax and penalize* method outperforms state-of-the-art approaches based on relaxation and rounding. Finally, conclusions and perspectives are drawn in Section 6.

Notation. For every integer $n \geq 1$, $[n]$ denotes the set $\{1, \ldots, n\}$. We denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^n$ and by $\|\cdot\|_\infty$ the infinity norm, meaning $\|x\|_\infty = \max_{1 \leq u \leq n} |x_i|$. If $x \in \mathbb{R}^n$ and $\rho > 0$ we denote by $B_\rho(x)$ the closed ball in $\mathbb{R}^n$ with center x and radius ρ, i.e., $B_\rho(x) = \{x' \in \mathbb{R}^n \colon \|x' - x\| \leq \rho\}$. The standard $(n-1)$-simplex is denoted by $\Delta^{n-1} = \{x \in \mathbb{R}^n_+ \colon \sum_{i=1}^n x_i = 1\}$. Also, $\odot$ is the Hadamard product, meaning the component-wise multiplication of vectors in $\mathbb{R}^d$. If $\Psi \colon \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $\Omega \subseteq \mathbb{R}^n$, we denote by $\operatorname{argmin}_\Omega \Psi$ the set of minimizers of Ψ over Ω and with a slight abuse of notation also the minimizer itself when it is unique.

2 Problem statement

We consider mixed-binary bilevel problems of the form

$$\min_{\lambda,\theta} \quad F(\lambda, \theta, w(\lambda, \theta)) \tag{1a}$$

$$\text{s.t.} \quad \lambda \in \Lambda \subseteq \mathbb{R}^m, \ \theta \in \Theta_{\text{bin}} \subseteq \{0,1\}^p, \tag{1b}$$

$$w(\lambda, \theta) = \operatorname*{argmin}_{w \in W(\lambda,\theta)} f(\lambda, \theta, w), \tag{1c}$$

where $F, f \colon \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ and $W(\lambda, \theta) \subseteq \mathbb{R}^d$. We note that the lower-level problem is supposed to admit a unique solution and that the hyperparameters λ and θ are continuous and binary variables, respectively. In the context of machine learning problems, the functions F and f often are the loss over a validation set and training set, respectively. We will provide a major application of this situation in Section 5.

In the remainder of this paper we assume that the binary set Θ_{bin} is embedded in a larger continuous set Θ and that the lower-level problem admits a unique solution also when $\theta \in \Theta$. Then, we can consider the following more compact formulation

$$\min_{(\lambda,\theta) \in \Lambda \times \Theta_{\text{bin}}} G(\lambda, \theta) \tag{2}$$

of the above problem, where we set $G(\lambda, \theta) := F(\lambda, \theta, w(\lambda, \theta))$ and require that the following assumption holds.

Assumption 1.

(i) $\Lambda \subseteq \mathbb{R}^m$ *is nonempty and compact.*

(ii) $\Theta_{\text{bin}} := \Theta \cap \{0,1\}^p \neq \varnothing$, *where* $\Theta \subseteq [0,1]^p$ *is convex and compact and* $\Theta \setminus \Theta_{\text{bin}} \neq \varnothing$.

(iii) $G \colon \Lambda \times \Theta \to \mathbb{R}$ *is continuous.*

(iv) *For all* $\lambda \in \Lambda$*, the map* $G(\lambda, \cdot)$ *is Lipschitz continuous with constant* $L > 0$ *on* Θ.

Assumption 1 can be met with appropriate hypothesis on the functions F and f. For instance, a sufficient condition for (iii) is that the functions F and f are jointly continuous, the function $f(\lambda, \theta, \cdot)$ is strongly convex with a modulus of convexity which is uniform for every (λ, θ), and the set-valued mapping $(\lambda, \theta) \mapsto W(\lambda, \theta)$ is closed and such that, for every $(\bar{\lambda}, \bar{\theta}) \in \Lambda \times \Theta$, $\mathrm{dist}(w(\bar{\lambda}, \bar{\theta}), W(w, \theta)) \to 0$ as $(\lambda, y) \to (\bar{\lambda}, \bar{\theta})$ [4, Proposition 4.4]. Additional conditions can ensure the validity of (iv) too (see, [4, Section 4.4].)

3 Restating the problem via a smooth penalty function

In order to deal with the binary variables in problem (2), we relax the integrality constraints on θ via a classic penalty term. This leads to the continuous optimization problem

$$\min_{(\lambda,\theta)\in\Lambda\times\Theta} G(\lambda, \theta) + \frac{1}{\varepsilon}\varphi(\theta) \tag{3}$$

in which we use the penalty function

$$\varphi(\theta) = \sum_{i=1}^{p} \theta_i(1 - \theta_i). \tag{4}$$

Note that the function in (4) is a smooth, concave, and quadratic function with the following properties:

$$\forall\, \theta \in [0, 1]^p \colon \varphi(\theta) \geq 0 \quad \text{and} \quad \forall\, \theta \in \{0, 1\}^p \colon \varphi(\theta) = 0.$$

This penalty has been introduced in [19] to define equivalent continuous reformulations of mixed-integer linear programming problems. In Section A we give the main properties of this penalty function that we use to prove the main results of this and the next section.

We start with a result establishing the equivalence of Problems (2) and (3) above in terms of global minimizers. It is in line with a stream of works analyzing the use of concave penalty functions in the framework of nonlinear optimization problems with binary/integer variables (see, e.g., [9, 11, 15] and references therein). For the reader's convenience we provide the proof of this result in the appendix.

Theorem 1. *Suppose that Assumption 1 is satisfied. Then, there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in \,]0, \bar{\varepsilon}]$, problems* (2) *and* (3) *have the same global minimizers, i.e.,*

$$\operatorname*{argmin}_{(\lambda,\theta)\in\Lambda\times\Theta_{\mathrm{bin}}} G(\lambda, \theta) = \operatorname*{argmin}_{(\lambda,\theta)\in\Lambda\times\Theta} G(\lambda, \theta) + \frac{1}{\varepsilon}\varphi(\theta).$$

The conclusion of Theorem 1 is remarkable since it guarantees that despite the fact that (3) is a purely continuous optimization problem, for ε sufficiently small, all of its global minimizers are mixed-binary feasible and are exactly the global minimizers of the original problem (2).

Remark 1.

(i) *Set $G_\varepsilon \colon \Lambda \times \Theta \to \mathbb{R}$ such that $G_\varepsilon(\lambda, \theta) = G(\lambda, \theta) + \varepsilon^{-1}\varphi(\theta)$ and let $\delta_{\Lambda\times\Theta_{\mathrm{bin}}} \colon \Lambda \times \Theta \to \mathbb{R}$ be the indicator function of the set $\Lambda \times \Theta_{\mathrm{bin}}$, i.e., the function that is zero on $\Lambda \times \Theta_{\mathrm{bin}}$ and $+\infty$ otherwise. Then, it is easy to see that G_ε Γ-converges[5] to $G + \delta_{\Lambda\times\Theta_{\mathrm{bin}}}$ as $\varepsilon \to 0$. Moreover, the family of functions $(G_\varepsilon)_{\varepsilon>0}$ is clearly equicoercive since they are all defined on the compact set $\Lambda \times \Theta$. Therefore, it holds*

$$\operatorname*{argmin}_{\Lambda\times\Theta} G_\varepsilon \to \operatorname*{argmin}_{\Lambda\times\Theta} G + \delta_{\Lambda\times\Theta_{\mathrm{bin}}} = \operatorname*{argmin}_{\Lambda\times\Theta_{\mathrm{bin}}} G \quad \text{as } \varepsilon \to 0$$

in the sense of set convergence. This is a standard result from variational analysis [5] and it is always true provided that G and φ are continuous functions as well as that $\varphi \geq 0$ and $\varphi(\theta) = 0 \iff \theta \in \Theta_{\mathrm{bin}}$ holds.

(ii) *In view of* (i)*, which gives an asymptotic result, the statement of Theorem 1 is stronger in the sense that, for the special function* (4) *and for ε small enough,* $\operatorname{argmin}_{\Lambda\times\Theta} G_\varepsilon = \operatorname{argmin}_{\Lambda\times\Theta_{\mathrm{bin}}} G$ *holds.*

The previous theorem provides a justification to address problem (3) instead of (2). However, because the objective function in (3) is nonconvex, only local minimizers are computationally approachable. Thus, the idea is that of looking for local minimizers of (3) which are also mixed-binary —since the global minimizers of (2) lie among them.

The next result is entirely new and addresses the issue of identifying mixed-binary local minimizers of the objective in (3), providing a sufficient condition for that purpose.

[5]This type of convergence of functions is also known as epiconvergence.

Theorem 2. *Suppose that Assumption 1 holds. Let* $c \in]0, 1/2[$ *and* $0 < \varepsilon < (1-2c)/L$. *Moreover, let* $(\bar{\lambda}, \bar{\theta})$ *be a local minimizer of*

$$G(\lambda, \theta) + \frac{1}{\varepsilon}\varphi(\theta) \ \text{ on } \ \Lambda \times \Theta.$$

If $\operatorname{dist}_\infty(\bar{\theta}, \Theta_{\text{bin}}) := \inf_{\theta \in \Theta_{\text{bin}}} \|\bar{\theta} - \theta\|_\infty < c$, *then* $\bar{\theta} \in \Theta_{\text{bin}}$.

Proof. Since $\operatorname{dist}_\infty(\bar{\theta}, \Theta_{\text{bin}}) < c$, there exists $\theta \in \Theta_{\text{bin}}$ such that $\|\bar{\theta} - \theta\|_\infty < c$. Let

$$\theta_t := (1-t)\bar{\theta} + t\theta = \bar{\theta} + t(\theta - \bar{\theta}) \quad \text{with } t \in [0, 1].$$

In particular, $\|\theta_t - \bar{\theta}\|_\infty = t\|\theta - \bar{\theta}\|_\infty < tc \leq c$ and $\theta_t \in \Theta$, since Θ is convex. By Lemma 3,

$$\varphi(\bar{\theta}) - \varphi(\theta_t) \geq (1-2c)\|\theta_t - \bar{\theta}\| \tag{5}$$

holds. Moreover, there exists $\rho > 0$ such that for all $(\lambda', \theta') \in B_\rho(\bar{\lambda}, \bar{\theta}) \cap (\Lambda \times \Theta)$, it holds

$$G(\bar{\lambda}, \bar{\theta}) + \frac{1}{\varepsilon}\varphi(\bar{\theta}) \leq G(\lambda', \theta') + \frac{1}{\varepsilon}\varphi(\theta').$$

Now, take $t \in]0, 1[$ such that $t < \rho/(c\sqrt{p})$. Then, $\theta_t \in \Theta$ and $\|\theta_t - \bar{\theta}\| \leq \sqrt{p}\|\theta_t - \bar{\theta}\|_\infty < \sqrt{p}\,tc < \rho$. Therefore, $(\bar{\lambda}, \theta_t) \in B_\rho(\bar{\lambda}, \bar{\theta}) \cap (\Lambda \times \Theta)$ and we obtain

$$\begin{aligned} G(\bar{\lambda}, \theta_t) - G(\bar{\lambda}, \bar{\theta}) + \frac{1}{\varepsilon}\varphi(\theta_t) - \frac{1}{\varepsilon}\varphi(\bar{\theta}) &\leq L\|\theta_t - \bar{\theta}\| + \frac{1}{\varepsilon}(\varphi(\theta_t) - \varphi(\bar{\theta})) \\ &\overset{(5)}{\leq} L\|\theta_t - \bar{\theta}\| - \frac{1-2c}{\varepsilon}\|\theta_t - \bar{\theta}\| \\ &= \underbrace{\left(L - \frac{1-2c}{\varepsilon}\right)}_{<0}\|\theta_t - \bar{\theta}\|. \end{aligned} \tag{6}$$

Moreover, if $\bar{\theta} \notin \{0, 1\}^p$, since $\theta \in \{0, 1\}^p$, we have $\|\theta - \bar{\theta}\| > 0$ and hence $\|\theta_t - \bar{\theta}\| > 0$ for $t > 0$. Thus, (6) is strictly negative and it holds

$$G(\bar{\lambda}, \theta_t) + \frac{1}{\varepsilon}\varphi(\theta_t) < G(\bar{\lambda}, \bar{\theta}) + \frac{1}{\varepsilon}\varphi(\bar{\theta}),$$

which gives a contradiction. Thus, necessarily $\bar{\theta} \in \{0, 1\}^p$. □

Remark 2.

(i) *Theorem 2 essentially says that, if* ε *is small enough, within the distance of* $1/2$ *measured with the infinity norm, there are no other local minimizers of* (3) *than the ones that are mixed-binary feasible.*

(ii) *Note that it does not make much sense to consider local minimizers of the function* G *over* $\Lambda \times \Theta_{\text{bin}}$, *since any point in* Θ_{bin} *is an isolated point and thus one can find a corresponding local minimizer for each one of them.*

4 An iterative penalty method

We now present an iterative method addressing problem (2). The idea is that of solving a sequence of problems of the form (3), indexed with k, with decreasing parameters ε_k. Hence, the problem to be solved in each iteration reads

$$\min_{(\lambda, \theta) \in \Lambda \times \Theta} G(\lambda, \theta) + \frac{1}{\varepsilon^k}\varphi(\theta). \tag{P^k}$$

Then, thanks to Theorem 1, it is clear that after a finite number of iterations the original mixed-binary optimization problem and the relaxed and penalized one (P^k) become equivalent in terms of global minimizers. Moreover, as we have already discussed in the previous section, in practice we can only target the computation of local minimizers, but we can restrict the search to the mixed-binary ones. In the following, we make this strategy more precise.

Theorem 3. *Suppose that Assumption 1 holds. Let* $(\varepsilon_k)_{k \in \mathbb{N}}$ *be a vanishing sequence of positive numbers and, for every* $k \in \mathbb{N}$, *let* (λ^k, θ^k) *be a local minimizer of* (P^k). *Then,*

$$\liminf_{k \to +\infty} \operatorname{dist}_\infty(\theta^k, \Theta_{\text{bin}}) < 1/2 \ \Rightarrow \ \exists\, k \in \mathbb{N} \text{ s.t. } \theta^k \in \Theta_{\text{bin}}.$$

Moreover, if $\theta^k \in \Theta_{\text{bin}}$, *then we have that* λ^k *is a local minimizer of*

$$\min_{\lambda \in \Lambda} G(\lambda, \theta^k).$$

Algorithm 1: Penalty method

Input: Problem (2), $\varepsilon^0 > 0$, $\beta \in]0,1[$.
for $k = 0, 1, 2, \dots$ **do**
  Let (λ^k, θ^k) be a solution (either local or global) of problem (P^k).
  if $\theta^k \notin \{0,1\}^p$ **then**
    update $\varepsilon^{k+1} = \beta \varepsilon^k$
  else
    return (λ^k, θ^k).

Proof. Suppose that $\liminf_{k\to+\infty} \mathrm{dist}_\infty(\theta^k, \Theta_{\mathrm{bin}}) < 1/2$ and let $c > 0$ such that $\liminf_{k\to+\infty} \mathrm{dist}_\infty(\theta^k, \Theta_{\mathrm{bin}}) < c < 1/2$. Then, there exists a subsequence $(\theta^{n_k})_{k\in\mathbb{N}}$ such that

$$\forall\, k \in \mathbb{N}\colon\ \ \mathrm{dist}_\infty(\theta^{n_k}, \Theta_{\mathrm{bin}}) < c \quad \text{and} \quad \varepsilon^{n_k} \to 0.$$

Thus, there exists $k \in \mathbb{N}$ such that

$$\mathrm{dist}_\infty(\theta^{n_k}, \Theta_{\mathrm{bin}}) < c \quad \text{and} \quad \varepsilon_{n_k} < \frac{1-2c}{L}$$

and this, in view of Theorem 2, gives that $\theta^{n_k} \in \Theta_{\mathrm{bin}}$. Concerning the second part of the statement, suppose that $\theta^k \in \Theta_{\mathrm{bin}}$, where (λ^k, θ^k) is a local minimizer of (P^k). Then, $\theta^k \in \{0,1\}^p$ and there exists $\rho_k > 0$ such that

$$\forall\,(\lambda,\theta) \in B_{\rho_k}(\lambda^k,\theta^k) \cap (\Lambda \times \Theta)\colon\ G(\lambda^k,\theta^k) + \frac{1}{\varepsilon_k}\varphi(\theta^k) \le G(\lambda,\theta) + \frac{1}{\varepsilon_k}\varphi(\theta).$$

Therefore, taking $\theta = \theta^k$ in the above inequality and noting that $\varphi(\theta^k) = 0$, we have

$$\forall\,\lambda \in B_{\rho_k}(\lambda^k) \cap \Lambda\colon\ G(\lambda^k,\theta^k) + \frac{1}{\varepsilon_k}\underbrace{\varphi(\theta^k)}_{=0} \le G(\lambda,\theta^k) + \frac{1}{\varepsilon_k}\underbrace{\varphi(\theta^k)}_{=0},$$

which shows that λ^k is a local minimizer of $G(\cdot,\theta^k)$ over Λ. □

Remark 3.

(i) *In the experiment given in Section 5, we checked that the condition considered in Theorem 3 always occurs, meaning that the distance* $\mathrm{dist}_\infty(\theta^k, \Theta_{\mathrm{bin}})$*, where* θ^k *was obtained by solving problem* (P^k) *via a gradient-based subroutine, remains well-below the threshold* $1/2$*.*

(ii) *We note that there might indeed exist points* $\theta \in \Theta$ *such that* $\mathrm{dist}_\infty(\theta, \Theta_{\mathrm{bin}}) > 1/2$*. For instance, if we take the standard* $(p-1)$*-simplex*

$$\Theta = \Delta^{p-1} = \left\{\theta \in \mathbb{R}^p_+ :\ \sum_{i=1}^{p} \theta_i = 1\right\},$$

we have that

$$\big\{\theta \in \Theta :\ \mathrm{dist}_\infty(\theta, \Theta_{\mathrm{bin}}) \ge 1/2\big\} = \Delta^{p-1} \cap [0,1/2]^p, \tag{7}$$

which for $p = 3$ *is the full equilateral triangle with vertices* $(e_1+e_2)/2, (e_1+e_3)/2$ *and* $(e_2+e_3)/2$*, where the* e_i*'s are the vectors of the canonical basis of* $\mathbb{R}^p$*. In general, the set in* (7) *is a polytope of dimension* $p-1$ *with* $2p$ *facets and* $(p(p-1)/2)$ *vertices.*

The method is formally given in Algorithm 1.

5 Application to the estimation of the group-structure in regression problems

We address the problem of estimating the group structure in group-sparse regression problems. First, we detail in Section 5.1 how this problem fits our mathematical formulation. Then, in Section 5.2 we provide numerical experiments on both synthetic data datasets.

5.1 Problem setting and formulation

Given an output vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n\times d}$, the group lasso problem can be formulated as follows

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|Xw - y\|^2 + \lambda \sum_{l=1}^{L} \|\theta_l \odot w\|_2,$$

where $\lambda > 0$ is a regularization parameter and θ_l is a binary vector (with entries in $\{0,1\}$) indicating the features (components) of w belonging to the lth group, meaning $\mathcal{G}_l = \{i \in [d] \colon \theta_{i,l} = 1\}$[6]. In case of nonoverlapping groups it is assumed that $\sum_{l=1}^{L} \theta_{j,l} = 1$ for every $j \in [d]$. In the classic literature on the topic the groups are assumed to be known a priori [21, 23], but often in practice there is no clue about the structure of the groups and actually the problem is to infer this group structure from the data. However, this amounts to estimating the binary variables θ_l's, which in general poses a challenge.

Continuous relaxation. In view of the discussion above, in the related literature a common approach is to relax the problem allowing the θ_l's to vary in the continuum $[0,1]^d$. This approach was followed in [8], in which the bilevel optimization problem

$$\begin{aligned} &\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} C_t(w_t(\lambda,\theta)), \quad \text{with} \quad \Theta = \Big\{\theta \in [0,1]^{d\times L} \colon \sum_{l=1}^{L} \theta_l = \mathbf{1}_d\Big\} = (\Delta^{L-1})^d, \\ &\text{and} \quad w(\lambda,\theta) = \operatorname*{argmin}_{(w_1,\dots,w_T)\in\mathbb{R}^{d\times T}} \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\|X_t w_t - y_t\|^2 + \lambda \sum_{l=1}^{L}\|\theta_l \odot w_t\|_2 + \frac{\eta}{2}\|w_t\|^2\right) \end{aligned} \tag{8}$$

is proposed, where $(X_t, y_t)_{1\le t\le T}$ defines T regression problems in which the regressors share the same group-sparsity structure and C_t is a smooth cost function acting as a validation error for the t-th task. Note that in this formulation, λ is supposed to be fixed (possibly determined by a cross-validation procedure). The regularization terms $\eta\|w_t\|^2$, with $\eta \ll 1$, are added to ensure uniqueness of the solution for the lower level problem and to devise a dual algorithmic procedure generating a sequence $(w^{(q)}(\lambda,\theta))_{q\in\mathbb{N}}$ with smooth updates (w.r.t. θ) such that $w^{(q)}(\lambda,\theta) \to w(\lambda,\theta)$ uniformly on Θ as $q \to +\infty$ (see [8, Section 3.2]). Ultimately, the groups are estimated by solving the problem

$$\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} C_t(w^{(q)}(\lambda,\theta)),$$

with q large enough, and by appropriately thresholding (a posteriori) the solution θ in order to recover binary variables θ_l's.

Proposed method. Our general approach, as described in Section 4, allows to bypass the last thresholding step and to directly address the more challenging problem

$$\min_{(\lambda,\theta)\in\Lambda\times\Theta_{\text{bin}}} \frac{1}{T}\sum_{t=1}^{T} C_t(w^{(q)}(\lambda,\theta)), \text{ with } \begin{cases} \Lambda = [\lambda_{\min}, \lambda_{\max}] \quad \text{with } 0 < \lambda_{\min} < \lambda_{\max}, \\ \Theta_{\text{bin}} = \Big\{\theta \in \{0,1\}^{d\times L} \colon \forall\, j \in [d] \ \sum_{l=1}^{L} \theta_{j,l} = 1\Big\}, \end{cases} \tag{9}$$

which also involve the estimation of the regularization parameter λ. Now, since $w^{(q)}(\lambda,\theta)$ is smooth w.r.t. θ, the objective in (9) satisfies Assumption 1 and hence, in view of Theorem 1 and Theorem 3, we can consider the *relaxed and penalized* version of problem (9), that is

$$\min_{(\lambda,\theta)\in\Lambda\times\Theta} \frac{1}{T}\sum_{t=1}^{T}\left(C_t(w^{(q)}(\lambda,\theta)) + \frac{1}{\varepsilon}\varphi(\theta)\right), \tag{10}$$

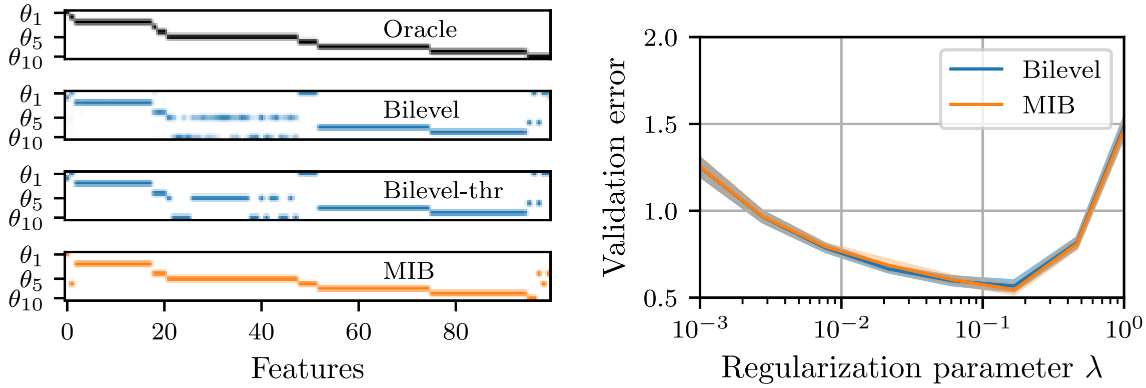
and state that, if ε is small enough, the two problems (9) and (10) share the same global minimizers. By leveraging this equivalence, we study in the next section the added benefits of the proposed Algorithm 1.

Remark 4. *According to [8, Theorem 3.1] we have that $w^{(q)}(\lambda,\theta) \to w(\lambda,\theta)$ as $q \to +\infty$ uniformly on $\Lambda \times \Theta_{\text{bin}}$, so that, similarly to [8, Theorem 2.1], one can prove that problem* (9) *converges as $q \to +\infty$ to problem*

$$\min_{(\lambda,\theta)\in\Lambda\times\Theta_{\text{bin}}} \frac{1}{T}\sum_{t=1}^{T} C_t(w(\lambda,\theta))$$

in terms of optimal values and sets of global minimizers. This provides a justification for addressing problem (9).

[6] The vectors θ_l are thought as columns of a matrix $d \times L$



(a) Learned groups $\theta^\top = [\theta_{l,j}]_{\substack{l\in\{1,\dots,10\}\\ j\in\{1,\dots,100\}}}$ (b) Validation error

Figure 1: Comparison of the proposed approach (MIB) with the one given in [8] (Bilevel).

5.2 Numerical experiments

In this section, we provide some preliminary numerical experiments on the estimation of group-sparsity structure detailed in Section 5.1. We compare our approach with the one proposed in [8], addressing the same bilevel problem that is, we optimize only over the hyperparameter θ, while leaving the scalar hyperparameter λ to be determined by cross-validation. Details are given below.

The experimental setting is similar to that of [8] but differs by two aspects to spice up the problem. First, we consider random groups of possibly very different sizes. Second, we stand in a setting made of few tasks T which was shown in [8] to be critical. In particular, we restrict to $T = 50$ which is 10 times less than the setting of [8]. We follow the next procedure to generate the training, validation and test datasets.

Synthetic dataset generation. We consider the problem of Section 5.1 where each task amounts to predict an oracle regressor $w^\star \in \mathbb{R}^d$, made of $d = 100$ features, from $n = 50$ observations. Each regressor yields a group structure dictated by $L = 10$ oracle groups $\{\mathcal{G}_l\}_{l=1}^L$ randomly generated with largely different sizes. An example is provided in Figure 1a (top panel) where the L groups form a partitioning of the d features. For instance, the first and second feature belong to $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, whereas the next 17 features belong to $\mathcal{G}_3$. For every task $t \in \{1,\dots,T\}$, we generate the oracle regressor $w_t^\star$ and the data (X_t, y_t) as follows. The regressor $w_t^\star$ is generated such that its values are non-zero in at most 2 groups chosen at random, the design matrix $X_t \in \mathbb{R}^{n\times d}$ is randomly drawn according to a standard normal distribution and then normalized column-wise, and the output y_t is such that $y_t \sim \mathcal{N}(X_t w_t^\star, 0.1\mathrm{I}_d)$. Validation and test sets are generated similarly.

Algorithmic setting. We implement the framework of Algorithm 1 by solving a sequence of optimization problems of type (10) w.r.t. θ only and with $\varepsilon^0 = \infty$, $\varepsilon^1 = 10^3$ and $\varepsilon^{k+1} = 0.5\varepsilon^k$ for every $k \in \{1,\dots,K = 10\}$. For each problem of type (10), we use the same differentiable algorithm of [8] for solving the lower-level problem in (8) with $\eta = 10^{-3}$ and $q = 500$ inner iterations. Whereas, for the upper optimization in (10), since the proposed framework is general enough to handle any algorithm, here we resort to Adam [12]. We embrace the finite-sum nature of the objective in (10) by setting the batch size to 10 and the maximum number of epochs equal to 500. The step-size is set to 10^{-2}, the weight decay to 0 and the running average coefficients are kept by default, i.e., $(0.9, 0.999)$. As for the model proposed in [8], for a fair comparison, we let algorithm Adam run for a maximum number of $500 \times (K+1)$ epochs.

Results. Both methods are run on 5 different realizations multiple times for 10 different values of λ equally spaces, in logarithmic scale, in the interval $[10^{-3}, 10]$. The averaged validation errors, as functions of λ, are reported in Figure 1b and are used to perform cross-validation of λ. Once the best λ is selected, we illustrate the recovered groups in Figure 1a. In the second and third rows are reported the relaxed and *relaxed and rounded* solution of [8], respectively. In the last row, we report the proposed estimate, coined MIB.

In addition, the two methods are compared according to two performance measures: the test error and the reconstruction error, that is $(1/d)\|w - w^\star\|_F^2$. The averaged results are reported in Table 1. These results clearly show that our method outperforms the *relaxed and rounding* strategy.

Table 1: Test performance

Method	Test error	Reconstruction error
MIB	0.5487 ± 0.0522	0.2955 ± 0.0459
Bilevel	0.5668 ± 0.0626	0.3162 ± 0.0535

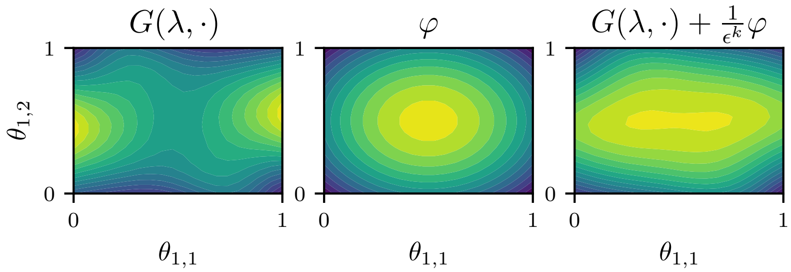


Figure 2: Illustration of the loss landscape. Adding the penalty φ to G allows to match the global minima with the oracle solutions ($\theta_{1,2} = 0, \theta_{1,1} = 1$) and ($\theta_{1,2} = 1, \theta_{1,1} = 0$).

Further insights about the added penalty. We perform some small scale experiment to visualize the effect of the penalization term φ. We consider a two-dimensional problem (i.e., $d = 2$) with only $L = 2$ groups. In Figure 2 (left panel), we show the contour lines of the original objective loss landscape, which has two global minimizers at the top left and bottom right corners. In Figure 2 (right panel), we illustrate how the loss landscape changes with the added penalty while preserving the global minimizers.

6 Conclusion

In this paper, we studied the idea of relaxing the integrality constraints and using a penalty term to handle mixed-binary bilevel optimization problems arising in hyperparameter tuning of machine learning systems. Besides a result concerning the equivalence in terms of global minimizers, sufficient conditions for identifying mixed-binary local minimizers are stated. These theoretical results naturally lead to devise a penalty method that is, under suitable assumptions, guaranteed to provide mixed-binary solutions. The novel approach is shown to outperform classic approaches based on relaxation and rounding using the example of the group lasso problem.

References

[1] Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *International Conference on Learning Representations (ICLR)*, 2022.

[2] Juhan Bae and Roger B Grosse. Delta-stn: Efficient bilevel optimization for neural networks using structured response jacobians. *Advances in Neural Information Processing Systems*, 33:21725–21737, 2020.

[3] Kristin P Bennett, Jing Hu, Xiaoyun Ji, Gautam Kunapuli, and Jong-Shi Pang. Model selection via bilevel optimization. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1922–1929. IEEE, 2006.

[4] J.Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer-Verlag, New York, 2000.

[5] Asen L. Dontchev and Tullio Zolezzi. *Well-posed optimization problems*. Springer-Verlag, Berlin., 1993.

[6] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

[7] Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1972–1982. PMLR, 09–15 Jun 2019.

[8] Jordan Frecon, Saverio Salzo, and Massimiliano Pontil. Bilevel learning of the group lasso structure. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8301–8311. Curran Associates, Inc., 2018.

[9] Franco Giannessi and F Niccolucci. Connections between nonlinear and integer programming problems. In *Symposia Mathematica*, volume 19, pages 161–176. Academic Press New York, 1976.

[10] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3748–3758. PMLR, 13–18 Jul 2020.

[11] Bahman Kalantari and J Ben Rosen. Penalty formulation for zero-one nonlinear programming. *Discrete Applied Mathematics*, 16(2):179–182, 1987.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR (Poster), ArXiv:1412.6980*, 2015.

[13] Thomas Kleinert, Veronika Grimm, and Martin Schmidt. Outer approximation for global optimization of mixed-integer quadratic bilevel problems. *Mathematical Programming (Series B)*, 2021.

[14] Thomas Kleinert, Martine Labbé, Ivana Ljubić, and Martin Schmidt. A survey on mixed-integer programming techniques in bilevel optimization. *EURO Journal on Computational Optimization*, 9:100007, 2021.

[15] Stefano Lucidi and Francesco Rinaldi. Exact penalty functions for nonlinear integer programming problems. *Journal of optimization theory and applications*, 145(3):479–488, 2010.

[16] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2113–2122, Lille, France, 07–09 Jul 2015. PMLR.

[17] Alexander Mitsos. Global solution of nonlinear mixed-integer bilevel programs. *Journal of Global Optimization*, 47(4):557–582, 2010.

[18] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 737–746, New York, New York, USA, 20–22 Jun 2016. PMLR.

[19] Madabhushi Raghavachari. On connections between zero-one integer programming and concave programming under linear constraints. *Operations Research*, 17(4):680–684, 1969.

[20] Yingjie Wang, Hong Chen, Feng Zheng, Chen Xu, Tieliang Gong, and Yanhong Chen. Multi-task additive models for robust estimation and automatic structure discovery. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11744–11755. Curran Associates, Inc., 2020.

[21] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.

[22] Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu. Advancing model pruning via bi-level optimization. In *Advances in Neural Information Processing Systems*, 2022.

[23] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468 – 3497, 2009.

Supplementary material

In this part we provide technical results we used in the paper and the proof of Theorem 1.

A Auxiliary results

In this section, we give a number of technical results used throughout the paper related to the penalty function φ. We recall that $\varphi\colon [0,1]^p \to \mathbb{R}$ is defined as

$$\varphi(\theta) = \sum_{i=1}^{p} \theta_i(1-\theta_i).$$

Moreover, we denote by $[p]$ the set $\{1,\dots,p\}$ and by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^p$. Occasionally, we will also used the following norms

$$\|\theta\|_1 = \sum_{i=1}^{p}|\theta_i| \quad \text{and} \quad \|\theta\|_\infty = \max_{1\le i\le p}|\theta_i|.$$

Lemma 1. *Let $\psi\colon [0,1] \to \mathbb{R}$ be such that*

$$\forall\, t \in [0,1]\colon \quad \psi(t) = t(1-t).$$

Let $\sigma \in \,]0,1/2]$. Then, for every $t_1, t_2 \in [0,1]$ we have

$$\left|\frac{t_1+t_2}{2} - \frac{1}{2}\right| \ge \sigma \implies |\psi(t_2) - \psi(t_1)| \ge 2\sigma|t_2 - t_1|.$$

Proof. Let $t_1, t_2 \in [0,1]$. One can easily check that

$$\psi(t_2) - \psi(t_1) = (1 - (t_1+t_2))(t_2 - t_1).$$

Therefore,

$$|\psi(t_2) - \psi(t_1)| = |t_1 + t_2 - 1||t_2 - t_1| = 2\left|\frac{t_1+t_2}{2} - \frac{1}{2}\right||t_2 - t_1| \ge 2\sigma|t_2 - t_1|.$$

□

Lemma 2. *Let $\sigma \in \,]0,1/2]$ and $\theta, \theta' \in [0,1]^p$ be such that the following holds:*

$$\forall\, i \in [p]\colon\ \theta'_i \neq \theta_i \implies \left|\frac{\theta_i+\theta'_i}{2} - \frac{1}{2}\right| \ge \sigma \quad \textit{and} \quad \left|\theta'_i - \frac{1}{2}\right| \le \left|\theta_i - \frac{1}{2}\right|.$$

Then, $\varphi(\theta') - \varphi(\theta) \ge 2\sigma\|\theta' - \theta\|$.

Proof. Let $\theta, \theta' \in [0,1]^p$ be as in the statement and let $I = \{i \in [p] \mid \theta_i \neq \theta'_i\}$. Then,

$$\forall\, i \in I\colon \quad \left|\frac{\theta_i+\theta'_i}{2} - \frac{1}{2}\right| \ge \sigma \quad \text{and} \quad \psi(\theta_i) \le \psi(\theta'_i)$$

and hence, by Lemma 1, we have

$$\begin{aligned}
\varphi(\theta') - \varphi(\theta) &= \sum_{i\in I} \psi(\theta'_i) - \psi(\theta_i) \\
&= \sum_{i\in I}|\psi(\theta'_i) - \psi(\theta_i)| \ge 2\sigma \sum_{i\in I}|\theta'_i - \theta_i| \\
&= 2\sigma\|\theta' - \theta\|_1 \ge 2\sigma\|\theta' - \theta\|,
\end{aligned}$$

where we used $\|\cdot\| \le \|\cdot\|_1$ for the last inequality. □

Remark 5. *Lemma 2 says that if the components of the mid point $(\theta+\theta')/2$ are bounded away from $1/2$ and the componentwise distance from θ to $1/2$ is larger than that of θ' to $1/2$, then $\varphi(\theta') - \varphi(\theta)$ can be bounded from below by $\|\theta' - \theta\|$; up to a multiplicative constant.*

Remark 6. *The conditions on θ and θ' required by Lemma 2 are satisfied if*

$$\forall\, i \in [p]\colon\ \theta_i \neq \theta'_i \implies \begin{cases} (\theta_i - 1/2)(\theta'_i - 1/2) \geq 0, \\ |\theta'_i - 1/2| \leq |\theta_i - 1/2|, \\ 2\sigma \leq |\theta_i - 1/2|. \end{cases}$$

Indeed, in such case we have

$$\begin{aligned} \left|\frac{\theta_i + \theta'_i}{2} - \frac{1}{2}\right| &= \frac{1}{2}|\theta_i + \theta'_i - 1| \\ &= \frac{1}{2}\left|\theta_i - \frac{1}{2} + \theta'_i - \frac{1}{2}\right| \overset{(*)}{=} \frac{1}{2}\left(\left|\theta_i - \frac{1}{2}\right| + \left|\theta'_i - \frac{1}{2}\right|\right) \geq \frac{1}{2}\left|\theta_i - \frac{1}{2}\right| \geq \sigma, \end{aligned}$$

where the equality in $(*)$ is due to the fact that

$$\begin{cases} (\theta_i - 1/2)(\theta'_i - 1/2) \geq 0 \\ |\theta'_i - 1/2| \leq |\theta_i - 1/2| \end{cases} \implies \big(\theta_i \leq \theta'_i \leq 1/2 \ \text{ or }\ \theta_i \geq \theta'_i \geq 1/2\big).$$

Corollary 1. *Let $\sigma \in\,]0, 1/2]$ and $\theta \in \{0,1\}^p$. Then*

$$\forall\, \theta' \in [0,1]^p\colon\ \ \|\theta' - \theta\| \leq 1 - 2\sigma \implies \varphi(\theta') \geq 2\sigma\|\theta' - \theta\|.$$

Proof. Let $\theta \in \{0,1\}^p$ and $\theta' \in [0,1]^p$. We will check the conditions in Lemma 2. Let $i \in [p]$. Since $|\theta_i - 1/2| = 1/2$, the condition $|\theta' - 1/2| \leq |\theta_i - 1/2|$ is automatically satisfied. Now, we note that

$$\begin{aligned} \theta_i = 0 &\implies \left|\frac{\theta_i + \theta'_i}{2} - \frac{1}{2}\right| = \left|\frac{\theta'_i}{2} - \frac{1}{2}\right| = \frac{1}{2}|\theta'_i - 1| = \frac{1}{2}(1 - |\theta'_i - \theta_i|) \\ \theta_i = 1 &\implies \left|\frac{\theta_i + \theta'_i}{2} - \frac{1}{2}\right| = \left|\frac{\theta'_i}{2}\right| = \frac{1}{2}|\theta'_i| = \frac{1}{2}(1 - |\theta'_i - \theta_i|). \end{aligned}$$

Therefore,

$$\left|\frac{\theta_i + \theta'_i}{2} - \frac{1}{2}\right| \geq \sigma \iff 1 - |\theta'_i - \theta_i| \geq 2\sigma \iff |\theta'_i - \theta_i| \leq 1 - 2\sigma$$

and the first condition in Lemma 2 is then equivalent to the condition $\|\theta' - \theta\|_\infty \leq 1 - 2\sigma$. The statement follows by recalling that $\|\cdot\|_\infty \leq \|\cdot\|$. ☐

Remark 7. *It is clear from the proof of Corollary 1 that in fact it holds*

$$\forall\, \theta' \in [0,1]^p\colon\ \ \|\theta' - \theta\|_\infty \leq 1 - 2\sigma \implies \varphi(\theta') \geq 2\sigma\|\theta' - \theta\|.$$

Lemma 3. *Let $\theta \in \{0,1\}^p$, $\bar\theta \in [0,1]^p$, and $c \in \mathbb{R}$ be such that $\|\theta - \bar\theta\|_\infty < c < \frac{1}{2}$. Let*

$$\theta^t = (1-t)\bar\theta + t\theta \quad \text{with} \quad t \in [0,1].$$

Then

$$\varphi(\bar\theta) - \varphi(\theta^t) \geq (1 - 2c)\|\theta^t - \bar\theta\|.$$

Proof. Since $|\theta_i - \bar\theta_i| < c < \frac{1}{2}$ holds for all $i = 1, \ldots, n$, we have

$$\frac{1}{2} = \left|\theta_i - \frac{1}{2}\right| \leq |\theta_i - \bar\theta_i| + \left|\bar\theta_i - \frac{1}{2}\right| < c + \left|\bar\theta_i - \frac{1}{2}\right|,$$

which implies

$$\left|\bar\theta_i - \frac{1}{2}\right| > \frac{1}{2} - c.$$

Moreover, since $\theta_i \in \{0,1\}$ and $|\bar\theta_i - \theta_i| < c < \frac{1}{2}$, we have

$$\begin{cases} \theta_i = 0 & \implies \theta_i = 0 \leq \bar\theta_i < c < \frac{1}{2} \\ \theta_i = 1 & \implies \frac{1}{2} < 1 - c < \bar\theta_i \leq 1 = \theta_i \end{cases}$$

and, hence, since θ_i^t is between θ_i and $\bar{\theta}_i$, it holds

$$\left|\theta_i^t - \frac{1}{2}\right| \geq \left|\bar{\theta}_i - \frac{1}{2}\right| > \frac{1}{2} - c$$

so that

$$\begin{aligned}\left|\frac{\theta_i^t + \bar{\theta}_i}{2} - \frac{1}{2}\right| &= \left|\frac{1}{2}(\theta_i^t + \bar{\theta}_i - 1)\right| = \frac{1}{2}\left|\left(\theta_i^t - \frac{1}{2}\right) + \left(\bar{\theta}_i - \frac{1}{2}\right)\right| \\ &= \frac{1}{2}\left(\left|\theta_i^t - \frac{1}{2}\right| + \left|\bar{\theta}_i - \frac{1}{2}\right|\right) \geq \left|\bar{\theta}_i - \frac{1}{2}\right| > \frac{1}{2} - c.\end{aligned}$$

Therefore, by Lemma 2, $\varphi(\bar{\theta}) - \varphi(\theta^t) \geq (1 - 2c)\|\theta^t - \bar{\theta}\|$ holds. □

B Proof of Theorem 1

For the sake of brevity we set

$$S = \underset{(\lambda,\theta)\in\Lambda\times\Theta_{\text{bin}}}{\operatorname{argmin}} G(\lambda,\theta) \quad \text{and} \quad S(\varepsilon) = \underset{(\lambda,\theta)\in\Lambda\times\Theta}{\operatorname{argmin}} G(\lambda,\theta) + \frac{1}{\varepsilon}\varphi(\theta).$$

Recall that $\Theta_{\text{bin}} = \Theta \cap \{0,1\}^p$. Let $\rho \in]0,1[$ and let $\hat{\varepsilon} \in]0,(1-\rho)/L[$. We define the open set

$$U = \bigcup_{\theta\in\Theta_{\text{bin}}} B_\rho(\theta),$$

where ρ is chosen small enough so to ensure that $\Theta \setminus U \neq \emptyset$[7]. Let $\bar{\theta}$ be a minimizer of φ over the compact set $\Theta \setminus U$. Then, clearly

$$\forall\, \theta' \in \Theta \setminus U\colon \quad \varphi(\theta') \geq \varphi(\bar{\theta}) > 0. \tag{11}$$

(Note that, since $\bar{\theta} \notin U$, then $\bar{\theta} \notin \{0,1\}^p$, and hence there exists $i \in [p]$ such that $\bar{\theta}_i \in]0,1[$, which implies that $\varphi(\bar{\theta}) \geq \bar{\theta}_i(1-\bar{\theta}_i) > 0$.) Thus, since

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\varphi(\bar{\theta}) = +\infty$$

there exists $\tilde{\varepsilon} \in]0,\hat{\varepsilon}]$ such that

$$\forall\, \varepsilon \in]0,\tilde{\varepsilon}] : \quad \frac{1}{\varepsilon}\varphi(\bar{\theta}) > \sup_{\Lambda\times\Theta_{\text{bin}}} G - \inf_{\Lambda\times(\Theta\setminus U)} G. \tag{12}$$

Now, we let $\varepsilon \in]0,\tilde{\varepsilon}]$ and show that $S(\varepsilon) \subset \Theta_{\text{bin}}$. Let $(\lambda^*,\theta^*) \in S(\varepsilon)$ and suppose, by contradiction, that $(\lambda^*,\theta^*) \notin \Theta_{\text{bin}}$. We can have the following two cases:

(a) Let $\theta^* \in U$. Then, there exists $\theta \in \Theta_{\text{bin}}$ such that $\theta^* \in \Theta \cap B_\rho(\theta)$. Thus, in view of Assumption 1 and Corollary 1, we have

$$G(\lambda^*,\theta) - G(\lambda^*,\theta^*) \leq L\|\theta^* - \theta\| < \frac{1-\rho}{\varepsilon}\|\theta^* - \theta\| \leq \frac{1}{\varepsilon}\varphi(\theta^*) = \frac{1}{\varepsilon}\varphi(\theta^*) - \frac{1}{\varepsilon}\varphi(\theta)$$

and hence

$$G(\lambda^*,\theta) + \frac{1}{\varepsilon}\varphi(\theta) < G(\lambda^*,\theta^*) + \frac{1}{\varepsilon}\varphi(\theta^*),$$

which yields a contradiction since (λ^*,θ^*) is a global minimizer of Problem (3).

(b) Let $\theta^* \notin U$. Then $(\lambda^*,\theta^*) \in \Lambda \times (\Theta \setminus U)$ and hence, recalling (11) and (12), we have

$$\begin{aligned} G(\lambda^*,\theta^*) + \frac{1}{\varepsilon}\varphi(\theta^*) &\geq \inf_{\Lambda\times(\Theta\setminus U)} G + \frac{1}{\varepsilon}\varphi(\theta^*) \\ &\geq \inf_{\Lambda\times(\Theta\setminus U)} G + \frac{1}{\varepsilon}\varphi(\bar{\theta}) \\ &> \sup_{\Lambda\times\Theta_{\text{bin}}} G \\ &\geq G(\lambda,\theta) + \frac{1}{\varepsilon}\underbrace{\varphi(\theta)}_{=0} \end{aligned}$$

[7] This means that $\exists\, \theta^* \in \Theta$ such that for every $\theta \in \Theta_{\text{bin}}$ it holds $\|\theta - \theta^*\| > \rho$. This condition is met if we pick $\theta^* \in \Theta \setminus \{0,1\}^p$ (see Assumption 1(ii)) and (taking into account that $\Theta_{\text{bin}} = \Theta \cap \{0,1\}^p$ is a finite set) choose ρ such that $0 < \rho < \inf_{\theta\in\Theta_{\text{bin}}}\|\theta - \theta^*\|$.

for any $(\lambda, \theta) \in \Lambda \times \Theta_{\text{bin}} \subseteq \Lambda \times \Theta$, which gives again a contradiction.

Thus, in both cases we get a contradiction and therefore necessarily $(\lambda^*, \theta^*) \in \Lambda \times \Theta_{\text{bin}}$. Now, if we take $(\lambda^*, \theta^*) \in S(\varepsilon)$, since $(\lambda^*, \theta^*) \in \Lambda \times \Theta_{\text{bin}}$, we have

$$G(\lambda^*, \theta^*) + \frac{1}{\varepsilon} \underbrace{\varphi(\theta^*)}_{=0} \leq G(\lambda, \theta) + \frac{1}{\varepsilon} \underbrace{\varphi(\theta)}_{=0} \quad \forall\, (\lambda, \theta) \in \Lambda \times \Theta_{\text{bin}} \subseteq \Lambda \times \Theta.$$

Therefore, for all $(\lambda, \theta) \in \Lambda \times \Theta_{\text{bin}}$, $G(\lambda^*, \theta^*) \leq G(\lambda, \theta)$, meaning $(\lambda^*, \theta^*) \in S$. Vice versa, let $(\lambda^*, \theta^*) \in S$. Choosing $(\tilde{\lambda}^*, \tilde{\theta}^*) \in S(\varepsilon)$, since $(\tilde{\lambda}^*, \tilde{\theta}^*) \in \Lambda \times \Theta_{\text{bin}}$, we have

$$\begin{aligned} G(\lambda^*, \theta^*) + \frac{1}{\varepsilon} \underbrace{\varphi(\theta^*)}_{=0} &= G(\lambda^*, \theta^*) \leq G(\tilde{\lambda}^*, \tilde{\theta}^*) \\ &\leq G(\tilde{\lambda}^*, \tilde{\theta}^*) + \frac{1}{\varepsilon} \varphi(\tilde{\theta}^*) = \min_{(\lambda, \theta)) \in \Lambda \times \Theta} G(\lambda, \theta) + \frac{1}{\varepsilon} \varphi(\theta). \end{aligned}$$

Thus, $(\lambda^*, \theta^*) \in S(\varepsilon)$.