CONSTRAINED VARIABLE PROJECTION FOR STRUCTURED PROBLEMS

Emanuele Zangrando
Gran Sasso Science Institute, L'Aquila, Italy.

Sara Venturini
MOBS Lab, Northeastern University, Boston, USA.

Francesco Rinaldi
University of Padova, Padova, Italy.

Francesco Tudisco
Gran Sasso Science Institute, L'Aquila, Italy.
University of Edinburgh, Edinburgh, UK.

ABSTRACT

Variable projection is a classical technique for separable nonlinear least-squares problems, in which variables that enter linearly are eliminated exactly, yielding a reduced nonlinear problem. By expressing this framework as a particular instance of a broader class of bilevel optimization problems, we develop a constrained variable-projection framework for data-science models, where the remaining variables are subject to convex constraints and the eliminated variables arise from a lower-level least-squares problem. In particular, by interpreting variable projection as a collapsed bilevel optimization problem, we derive exact reduced-gradient formulas compatible with automatic differentiation and propose a conditional-gradient algorithm for the resulting constrained reduced problem. We establish convergence guarantees under standard smoothness and compactness assumptions, and discuss extensions to structured lower-level variables. Numerical experiments on sparse autoencoding, dictionary learning, blind deconvolution, and few-shot learning suggest that the method can improve wall-clock efficiency and data efficiency relative to natural joint-optimization baselines.

1 Introduction

Many problems in modern data science involve fitting models with two distinct types of variables: a variable block that enters a loss function linearly or quadratically, and a second block of nonlinear variables that controls features, representations, constraints, or physical parameters. This structure appears in applications such as dictionary learning [49], inverse problems [15], signal recovery, representation learning [11], and neural network training [46]. In these settings, the linear variable can often be optimized exactly once the nonlinear variables are fixed, yielding a reduced optimization problem of smaller dimension.

A classical framework for exploiting this structure is the variable projection method, or VarPro, initially introduced for separable nonlinear least-squares problems [31, 30, 29]. In its standard form, VarPro applies to problems of the type

$$\min_{w\in\mathbb{R}^N,\,\theta\in\mathbb{R}^p} \frac{1}{2}\|M(\theta)w - y(\theta)\|^2 + \frac{\lambda}{2}\|\Omega w\|^2 + \frac{\mu}{2}R(\theta), \tag{1}$$

where w is a linear variable and θ is a nonlinear variable. For fixed θ, the minimizer with respect to w can be computed by solving a least-squares problem. Substituting this minimizer into the objective produces a reduced problem in θ alone. This reduction can substantially decrease the dimension of the optimization problem and can improve numerical conditioning and computational efficiency.

The classical VarPro theory, however, is primarily designed for *unconstrained* separable nonlinear least-squares problems. This leaves open an important question: how should variable projection be used when the remaining nonlinear variables are constrained? Such constraints are common and often essential, as structured feasible sets arise naturally in signal processing, inverse problems, and machine learning. For instance, sparsity, rank, and norm constraints promote efficiency, interpretability, and regularization [13, 22, 33, 48, 51, 57], while simplex, box, and non-negativity constraints encode feasibility or statistical structure [14, 23, 28, 38, 43].

The goal of this paper is to develop a constrained variable-projection framework for such problems. We reinterpret and generalize VarPro through a collapsed bilevel optimization problem: the lower-level problem eliminates the variable w through a structured least-squares solve, while the upper-level problem optimizes the remaining variable θ over a convex feasible set. This viewpoint allows us to combine the dimension-reduction benefits of VarPro with projection-free constrained optimization methods. In particular, we derive reduced-gradient formulas that can be evaluated efficiently using a combination of closed-form lower-level solvers and automatic differentiation through vector-Jacobian products. We then use these gradients inside a conditional gradient, or Frank-Wolfe, method for the reduced constrained problem.

The resulting framework is motivated by applications in which the eliminated variable has a clear statistical or computational meaning. In sparse autoencoding, the eliminated variable may correspond to a linear decoder or readout map. In dictionary learning, the eliminated variables encode representation coefficients or dictionary-dependent least-squares updates. In blind deconvolution and inverse problems, they correspond to structured linear operators or signal components. In few-shot learning, a pretrained nonlinear representation can be combined with an optimized linear head, leading naturally to a variable-projection formulation.

1.1 Contributions

The main contributions of this paper are as follows.

- We observe that the problem treated in the original Variable-Projection works is the collapsed version of a specific class of bilevel optimization problems. Following this insight, we formulate constrained variable projection as a collapsed bilevel optimization problem, in which the lower-level problem is a structured least-squares problem and the upper-level problem imposes convex constraints on the remaining variables.
- We derive explicit reduced-gradient formulas for the constrained reduced objective. The formulas combine closed-form lower-level solutions with automatic differentiation through vector-Jacobian products, avoiding, when possible, the need to differentiate naively through ill-conditioned normal equations.
- We propose a tailored conditional-gradient method for the constrained reduced problem. The method is projection-free and is therefore well-suited to feasible sets for which linear minimization oracles are cheaper than Euclidean projections. We further establish convergence guarantees for the proposed method under standard assumptions.
- We demonstrate the flexibility of the framework on representative data-science and machine-learning problems, including sparse autoencoding, dictionary learning, blind deconvolution, and few-shot learning with a pretrained neural representation.

1.2 Organization

The rest of the paper is organized as follows. In Section 2, we discuss related work on variable projection, bilevel optimization, implicit differentiation, and conditional-gradient methods. In Section 3, we introduce the general constrained variable-projection problem formulation and its bilevel interpretation. We then derive the reduced-gradient expressions used by our method and discuss extensions to structured lower-level variables. The convergence analysis of the proposed conditional-gradient scheme is presented next. Finally, we report numerical experiments on several representative data-science problems.

2 Related Work

The variable projection method was developed for separable nonlinear least-squares problems in which some variables enter the model linearly and can therefore be eliminated exactly. The observation that such models can be reduced to a nonlinear problem in the remaining variables appears in [41], which credits an unpublished 1965 work by N. E. Dahl for the original idea. Golub and Pereyra developed the method systematically in [31], deriving formulas for the Jacobian of the reduced residual and proposing practical Gauss-Newton-type algorithms. The method and its developments are surveyed in [29].

A key computational issue in VarPro is the cost of forming the reduced Jacobian. Kaufman introduced a cheaper approximation for the Jacobian under a small-residual assumption [39]. More recent work has revisited this approximation and its effect on convergence. In particular, [17] studies the large-residual regime and proposes improved variants with stronger convergence properties, [25] analyzes the effect of approximate Jacobians and proposes practical stopping criteria that preserve convergence, and [55] extends the analysis for nonsmooth objectives. Global optimality results for a Riemannian relaxed version of the problem have been proposed in [24], while an algorithm for sparse regularization has been proposed in [56].

VarPro has been used successfully in a range of applications, including atmospheric remote sensing [12], neural network training [21, 46, 45], semi-blind image deblurring [50], low-rank matrix approximation problems [53, 44], and polynomial nearness problems [54]. These works show that eliminating linear variables can improve the numerical behavior of large-scale learning and inverse problems.

Our work follows this general direction, with the emphasis on constrained data-science models in which the remaining variables typically satisfy convex structural constraints. Sparse autoencoding highlights the interaction between representation learning and regularized linear reconstruction. Dictionary learning illustrates the role of constraints in structured matrix factorization. Blind deconvolution connects the framework to inverse problems and signal processing. Few-shot learning shows how a nonlinear representation map can be combined with an optimized linear head.

Bilevel optimization and implicit differentiation The formulation considered in this paper can be interpreted as a bilevel optimization problem in which the lower-level problem is solved exactly and then substituted into the upper-level objective (see, e.g., [4, 18, 20]). This places constrained VarPro within the broader class of collapsed bilevel methods, argmin differentiation, and differentiable optimization layers (see, e.g., [1, 3, 7, 26, 32] for further details on these topics). In contrast to generic bilevel optimization, the lower-level problems considered here have a least-squares structure, which allows explicit characterization of the lower-level solution and efficient computation of reduced gradients. This structure is central to the algorithmic efficiency and theoretical developments of the paper.

Conditional-gradient methods for constrained data science The upper-level feasible sets that arise in data science are often convex but not necessarily easy to project onto. Conditional-gradient methods are attractive in this setting because they replace Euclidean projections with linear minimization oracles [9, 10, 27]. Such methods are particularly useful for constraints involving simplices, norm balls, linear structures, and other sets for which linear optimization is cheaper than projection (see, e.g., [19, 37]). Our algorithm applies a tailored conditional-gradient method to the reduced VarPro objective, thereby combining projection-free constrained optimization with exact elimination of the least-squares variable.

3 Problem setting

3.1 Notation

Throughout the paper, $\|\cdot\|$ denotes the ℓ^2 componentwise norm, unless stated otherwise. For a full column-rank matrix Z, we write Z^+ for its Moore-Penrose pseudoinverse and $\mathcal{P}(Z) := ZZ^+$ for the orthogonal projector onto the range of Z.

3.2 General bilevel problem formulation

We consider constrained separable nonlinear least-squares problems of the form

$$\min_{w\in\mathbb{R}^N,\,\theta\in\mathcal{C}} \frac{1}{2}\|M(\theta)w - y(\theta)\|^2 + \frac{\lambda}{2}\|\Omega w\|^2 + \frac{\mu}{2}R(\theta), \tag{2}$$

where $\mathcal{C} \subseteq \mathbb{R}^p$ is a convex feasible set, $M : \mathbb{R}^p \to \mathbb{R}^{d\times N}$, $y : \mathbb{R}^p \to \mathbb{R}^d$, $\Omega \in \mathrm{GL}(\mathbb{R}^N)$, and $\lambda, \mu \geq 0$ are regularization parameters. The variable w enters the data-fitting term linearly, whereas θ enters nonlinearly through $M(\theta)$ and $y(\theta)$.

The Tikhonov term in w can be absorbed into the least-squares residual. Indeed, (2) is equivalent to

$$\min_{w\in\mathbb{R}^N,\,\theta\in\mathcal{C}} \frac{1}{2}\|M_\lambda(\theta)w - y_\lambda(\theta)\|^2 + \frac{\mu}{2}R(\theta), \tag{3}$$

where $M_\lambda : \mathbb{R}^p \to \mathbb{R}^{(N+d)\times N}$, $M_\lambda(\theta) = [M(\theta)^\top, \sqrt{\lambda}\Omega^\top]^\top$, $y_\lambda(\theta) = [y(\theta)^\top, 0_N^\top]^\top$. When $\lambda > 0$, the matrix $M_\lambda(\theta)$ has full column rank for every θ, as $M_\lambda(\theta)^\top M_\lambda(\theta) \succeq \lambda\,\Omega^\top\Omega$, and therefore its smallest eigenvalue is bounded below by λ times the smallest eigenvalue of $\Omega^\top\Omega$, which must be positive.

For this reason, and to simplify notation, we henceforth work directly with the augmented formulation. That is, unless stated otherwise, M and y denote the augmented quantities M_λ and y_λ, and we assume that $M : \mathbb{R}^p \to \mathbb{R}^{m\times N}$ and $y : \mathbb{R}^p \to \mathbb{R}^m$, with $m = d + N$ and $M(\theta)$ full column rank for every $\theta \in \mathcal{C}$. With this convention, the problem becomes

$$\min_{w\in\mathbb{R}^N,\,\theta\in\mathcal{C}} \frac{1}{2}\|M(\theta)w - y(\theta)\|^2 + \frac{\mu}{2}R(\theta). \tag{4}$$

Problem (4) admits an equivalent bilevel interpretation. For a fixed value of θ, the optimal linear variable is obtained by solving the lower-level least-squares problem

$$\widehat{w}(\theta) \in \operatorname*{arg\,min}_{w\in\mathbb{R}^N} \frac{1}{2}\|M(\theta)w - y(\theta)\|^2.$$

Thus (4) can be written as

$$\begin{cases} \min\limits_{\theta\in\mathcal{C}} \dfrac{1}{2}\|M(\theta)\widehat{w} - y(\theta)\|^2 + \dfrac{\mu}{2}R(\theta), \\ \widehat{w} \in \operatorname*{arg\,min}\limits_{w\in\mathbb{R}^N} \dfrac{1}{2}\|M(\theta)w - y(\theta)\|^2. \end{cases} \tag{5}$$

Since $M(\theta)$ has full column rank, the lower-level solution is unique and is given by

$$\widehat{w}(\theta) = \big(M(\theta)^\top M(\theta)\big)^{-1} M(\theta)^\top y(\theta) = M(\theta)^+ y(\theta). \tag{6}$$

Substituting (6) into the upper-level objective collapses the bilevel problem to the reduced single-level problem

$$\min_{\theta\in\mathcal{C}} f(\theta) := \frac{1}{2}\big\|\big(\mathcal{P}(M(\theta)) - I\big)y(\theta)\big\|^2 + \frac{\mu}{2}R(\theta), \tag{7}$$

where $\mathcal{P}(M(\theta)) = M(\theta)M(\theta)^+$ is the orthogonal projector onto $\mathrm{range}(M(\theta))$. This is the constrained analogue of the classical variable-projection formulation of [31].

The bilevel viewpoint also suggests a broader class of problems. In particular, we will consider a generalized formulation in which the lower-level problem is used to define $\widehat{w}(\theta)$ and the upper-level objective need not involve the same least-squares operator:

$$\begin{cases} \min\limits_{\theta\in\mathcal{C}} f_U(\widehat{w},\theta) := \dfrac{1}{2}\|M_U(\theta)\widehat{w} - y_U(\theta)\|^2 + \dfrac{\mu}{2}R(\theta), \\ \widehat{w} \in \operatorname*{arg\,min}\limits_{w\in\mathbb{R}^N} f_L(w,\theta) := \dfrac{1}{2}\|M_L(\theta)w - y_L(\theta)\|^2. \end{cases} \tag{8}$$

Here $M_L, M_U : \mathbb{R}^p \to \mathbb{R}^{m\times N}$ are assumed to have full column rank for all relevant θ, and $y_L, y_U : \mathbb{R}^p \to \mathbb{R}^m$. The classical constrained VarPro formulation corresponds to the special case $M_L = M_U = M$ and $y_L = y_U = y$.

In the remainder of the paper, we focus on the case in which the feasible set $\mathcal{C} \subseteq \mathbb{R}^p$ factorizes as

$$\mathcal{C} = \mathcal{C}_A \times \mathcal{C}_B \subseteq \mathbb{R}^{p_A} \times \mathbb{R}^{p_B}, \qquad p_A + p_B = p,$$

where $\mathcal{C}_A \subset \mathbb{R}^{p_A}$ is compact and convex, and $\mathcal{C}_B = \mathbb{R}^{p_B}$. This structure separates the variables subject to explicit constraints from those that remain unconstrained, and it will be used in the design and analysis of the conditional-gradient method below.

3.3 Constrained VarPro and conditional gradient methods

Under the assumption that the feasible set $\mathcal{C}$ is convex, the reduced problem can be naturally approached using conditional-gradient methods. In particular, Frank-Wolfe-type methods [9, 27] require only the solution of a linear minimization oracle (LMO) at each iteration; that is, they minimize a first-order approximation of the objective over the original feasible set. This makes them especially attractive when projections onto $\mathcal{C}$ are expensive, but linear minimization over $\mathcal{C}$ is efficient.

The main algorithmic requirement is therefore the ability to evaluate the gradient of the reduced upper-level objective, $\nabla \widehat{f}_U(\theta) = \nabla f_U(\widehat{w}(\theta), \theta)$. Assuming that ∇R can be computed efficiently, the central difficulty is the evaluation of the contribution coming from the dependence of $\widehat{w}(\theta)$ on θ. In principle, this derivative could be computed by automatic differentiation through the closed-form expression for $\widehat{w}(\theta)$. However, this approach may be numerically unstable when $M^\top M$ has small singular values.

To retain flexibility across applications while avoiding this instability, we use partial automatic differentiation. The starting point is the compositional structure of the reduced objective[1]:

$$\begin{aligned}\widehat{f}_U(\theta) =& \|M_U(\theta)M_L(\theta)^+ y_L(\theta) - y_U(\theta)\|^2 = \|(M_U M_L^+ - \mathcal{P}(M_U))y_L\|^2 + \|\mathcal{P}(M_U)(y_L - y_U)\|^2 + \\ &+ \|(\mathcal{P}(M_U) - I)y_U\|^2 + 2\langle (M_U M_L^+ - \mathcal{P}(M_U))y_L, \mathcal{P}(M_U)(y_L - y_U)\rangle,\end{aligned}$$

whose gradient can be written by the chain rule as

$$\nabla \widehat{f}_U(\theta) = \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)} + \partial_\theta \widehat{w}(\theta)^\top \nabla_w f_U|_{(\widehat{w}(\theta),\theta)}. \tag{9}$$

The derivative $\partial_\theta \widehat{w}(\theta)$ can be obtained by differentiating the first-order stationarity conditions for the lower-level problem. Namely,

$$\partial_\theta \nabla_w f_L(\widehat{w}(\theta), \theta) = \nabla^2_{w\theta} f_L|_{(\widehat{w}(\theta),\theta)} + \nabla^2_{ww} f_L|_{(\widehat{w}(\theta),\theta)} \partial_\theta \widehat{w}(\theta),$$

which implies $\partial_\theta \widehat{w}(\theta) = -(\nabla^2_{ww} f_L|_{(\widehat{w}(\theta),\theta)})^{-1} \nabla^2_{w\theta} f_L|_{(\widehat{w}(\theta),\theta)}$. Substituting this expression into the chain rule (9), and setting $\Gamma := M_U(M_L^\top M_L)^{-1}$, gives

$$\begin{aligned}\nabla \widehat{f}_U(\theta) =& \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)} - \nabla^2_{w\theta} f_L|^\top_{(\widehat{w}(\theta),\theta)} (\nabla^2_{ww} f_L|_{(\widehat{w}(\theta),\theta)})^{-\top} \nabla_w f_U|_{(\widehat{w}(\theta),\theta)} \\ =& \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)} - \nabla^2_{w\theta} f_L|^\top_{(\widehat{w}(\theta),\theta)} (M_L^\top M_L)^{-\top} \left[M_U^\top M_U (M_L^\top M_L)^{-1} M_L^\top y_L - M_U^\top y_U\right] \\ =& \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)} - \nabla^2_{\theta w} f_L|_{(\widehat{w}(\theta),\theta)} \left[\Gamma^\top \Gamma M_L^\top y_L - \Gamma^\top y_U\right] \\ =& \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)} - \nabla^2_{\theta w} f_L|_{(\widehat{w}(\theta),\theta)} \Gamma^\top \left[M_U M_L^+ y_L - y_U\right].\end{aligned} \tag{10}$$

In the special case where $M_L = M_U$ and $y_L = y_U$, one has $\Gamma^\top \left[M_U M_L^+ y_L - y_U\right] = 0$, and the corresponding coupling term in Equation (10) vanishes, leading to $\nabla \widehat{f}_U(\theta) = \nabla_\theta f_U|_{(\widehat{w}(\theta),\theta)}$.

The form of Equation (10) is useful computationally. In any application where $\widehat{w}(\theta)$ can be evaluated efficiently and where one can compute the action of $\nabla^2_{\theta w} f_L|_{(\widehat{w}(\theta),\theta)}$ on a tangent vector, the reduced gradient $\nabla \widehat{f}_U(\theta)$ can be evaluated exactly using automatic differentiation. As we will discuss in Section 6, this covers several applications of interest. In particular, to compute the required action of the Hessian on a vector, it is sufficient to notice that

$$\nabla^2_{w\theta} f_L^\top[v] = \partial_\theta \langle \nabla_w f_L, v\rangle = \partial_\theta \langle M_L(\theta)^\top (M_L(\theta)w - y_L(\theta)), v\rangle.$$

Therefore, if we have access to an automatic differentiation system, the vector-Jacobian products can be queried easily through the previous formula.

[1] For notational simplicity, in the calculations we only display the least-squares part of the function.

3.4 Extension for structured matrix problems

In several applications, the eliminated variable w is not free in the ambient space, but is constrained to belong to a lower-dimensional set, which we denote by $\mathscr{W}$. This set may have the structure of a differentiable manifold, an affine space, or a linear subspace. The preceding framework continues to apply whenever the lower-level problem has a unique solution. In particular, the same analysis can be used for any geometric class $\mathscr{W}$ such that

$$\min_{w \in \mathscr{W}} \|Mw - y\|_2^2 + \lambda \|\Omega w\|_2^2$$

admits a unique minimizer. Up to an additive constant, this lower-level problem can be equivalently written as

$$\min_{w \in \mathscr{W}} \|w - \widehat{w}\|_Q^2,$$

where $Q = (M^\top M + \lambda \Omega^\top \Omega), \|x\|_Q^2 = \|Q^{1/2}x\|_2^2, \widehat{w} = Q^{-1}M^\top y$. Thus, the lower-level solution is the projection of $\widehat{w}$ onto $\mathscr{W}$ with respect to the Q-inner product. When $\lambda > 0$, the matrix Q is positive definite and $\|\cdot\|_Q$ is a norm; otherwise, it may only define a seminorm. In particular, if $\mathscr{W}$ is an affine subspace, the corresponding projection depends smoothly on M.

The situation is more delicate when $\mathscr{W}$ is only convex. In finite dimension, closed convex sets are Chebyshev, so the projection is single-valued and continuous. However, differentiability of the projection may fail. In particular, the next result characterizes when the differentiability of the lower-level solution holds globally.

Proposition 3.1. *(Differentiability of optimal solution for constrained problems) Let $\mathscr{W}$ be a Chebyshev subset of $E := \mathbb{R}^N$ and consider the problem*

$$\mathcal{P}_{\mathscr{W}}^Q(\widehat{w}) := \arg\min_{w \in \mathscr{W}} \|w - \widehat{w}\|_Q^2.$$

Then, for $Q = (M^\top M + \lambda \Omega^\top \Omega), \|x\|_Q^2 = \|Q^{1/2}x\|_2^2, \widehat{w} = Q^{-1}M^\top y$, and assuming that $\lambda > 0, m \geq N$ and Ω full rank, the map

$$\Pi : \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^N, \quad \Pi(M, y) = \mathcal{P}^{Q(M)}(\widehat{w}(M, y))$$

is differentiable for all $(M, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ if and only if $\mathscr{W}$ is affine.

Proof. Let $E = \mathbb{R}^N$, endowed with the ℓ^2 inner product. Since $\mathscr{W}$ is a Chebyshev subset of a finite-dimensional Euclidean space, $\mathscr{W}$ is closed and convex. Hence, for every positive definite matrix Q, the Q-metric projection onto $\mathscr{W}$ is well-defined and single-valued.

Because $\Omega^\top \Omega \succ 0$ and $\lambda > 0$, we have

$$Q(M) = M^\top M + \lambda \Omega^\top \Omega \succ 0$$

for every M. Thus, $Q(M)^{-1}$ depends smoothly on M, and so does $\widehat{w}(M, y) = Q(M)^{-1}M^\top y$.

We first prove the easy direction. Suppose that $\mathscr{W}$ is affine, i.e., $\mathscr{W} = w_* + \mathscr{L}$, where $\mathscr{L} \subset E$ is a linear subspace. Fix a basis $v_1, \ldots, v_r$ of $\mathscr{L}$. For fixed $Q \succ 0$, the projection of $\widehat{w}$ onto $\mathscr{W}$ has the form

$$\mathcal{P}_{\mathscr{W}}^Q(\widehat{w}) = w_* + \sum_{i=1}^r \alpha_i v_i.$$

The coefficients are determined by the normal equations

$$\langle \Big(\widehat{w} - w_* - \sum_{j=1}^m \alpha_j v_j \Big), Q v_i \rangle = 0, \qquad i = 1, \ldots, r.$$

Equivalently, $G(Q)\alpha = b(Q, \widehat{w})$, where

$$G(Q)_{ij} = \langle v_i, v_j \rangle_Q, \qquad b(Q, \widehat{w})_i = \langle \widehat{w} - w_*, Q v_i \rangle.$$

Since $Q \succ 0$, the Gram matrix $G(Q)$ is positive definite, hence invertible. Therefore

$$\alpha = G(Q)^{-1} b(Q, \widehat{W})$$

depends smoothly on $(Q, \widehat{w})$. Since $Q(M)$ and $\widehat{w}(M, y)$ depend smoothly on (M, y), the map

$$\Pi(M, y) = \mathcal{P}_{\mathscr{W}}^{Q(M)}(\widehat{w}(M, y))$$

is differentiable, indeed smooth.

Conversely, assume that Π is differentiable for every (M, y). Choose $M_0 \in \mathbb{R}^{m \times N}$ with $\operatorname{rank}(M_0) = N$, which is possible because $N \leq m$. Set $Q_0 := Q(M_0) \succ 0$ and consider the linear map

$$L : \mathbb{R}^m \to \mathbb{R}^N, \qquad L(y) = Q_0^{-1} M_0^\top y.$$

Since $\operatorname{rank}(M_0) = N$, the matrix $Q_0^{-1} M_0^\top$ has rank N, hence L is surjective. Thus L admits a linear right inverse $R : \mathbb{R}^N \to \mathbb{R}^m$. For fixed M_0, the map $y \mapsto \Pi(M_0, y) = \mathcal{P}_{\mathscr{W}}^{Q_0}(L(y))$ is differentiable by assumption. Since $L \circ R = \mathrm{Id}_E$, we get $\mathcal{P}_{\mathscr{W}}^{Q_0}(z) = \Pi(M_0, Rz)$. Therefore, the full Q_0-metric projection $z \mapsto \mathcal{P}_{\mathscr{W}}^{Q_0}(z)$ is differentiable everywhere on E.

Now reduce to the ordinary Frobenius projection. Define the invertible linear map

$$T : E \to E, \qquad T(w) = Q_0^{1/2} w,$$

and set $C := T(\mathscr{W}) = Q_0^{1/2} \mathscr{W}$. Since $\|w - \widehat{w}\|_{Q_0} = \|T(w) - T(\widehat{w})\|_2$, we have $T\big(\mathcal{P}_{\mathscr{W}}^{Q_0}(\widehat{W})\big) = \mathcal{P}_C(T\widehat{W})$, or equivalently

$$\mathcal{P}_C = T \circ \mathcal{P}_{\mathscr{W}}^{Q_0} \circ T^{-1}.$$

Hence the ordinary Frobenius projection $\mathcal{P}_C$ is differentiable everywhere.

We now show that this forces C to be affine. Let $c \in C$, and let $\mathcal{K}_c C := \overline{\operatorname{cone}(C - c)}$ be the tangent cone of C at c. For closed convex sets in a finite dimensional Hilbert space, the directional derivative of the metric projection at a point $c \in C$ is given by $d\mathcal{P}_C(c; h) = \mathcal{P}_{\mathcal{K}_c C}(h)$, where $\mathcal{P}_{\mathcal{K}_c C}$ denotes the orthogonal projection onto the closed convex cone $\mathcal{K}_c C$.

But $\mathcal{P}_C$ is Frechet differentiable at c, so $h \mapsto d\mathcal{P}_C(c; h)$ is linear, i.e., $h \mapsto \mathcal{P}_{\mathcal{K}_c C}(h)$ is linear. The range of this linear map is exactly $\mathcal{K}_c C$, which is therefore a linear subspace. Let now $A := \operatorname{aff}(C)$ be the affine hull of C, and let $V := A - c = \operatorname{span}(C - c)$ be its associated direction space. Since $\mathcal{K}_c C \subset V$ and $C - c \subset \mathcal{K}_c C$, we get $V = \operatorname{span}(C - c) \subset \mathcal{K}_c C \subset V$. Therefore, $\mathcal{K}_c C = V$ for every $c \in C$.

We claim that every $c \in C$ is in the relative interior of C inside A. Indeed, if some $c \in C$ were not in $\operatorname{int}_A(C)$, the supporting hyperplane theorem would give a nonzero $u \in V$ such that $\langle u, z - c \rangle_2 \leq 0$ for every $z \in C$. Passing to the tangent cone gives $\langle u, h \rangle_2 \leq 0$ for every $h \in \mathcal{K}_c C$. But $\mathcal{K}_c C = V$, and $u \in V$, so choosing $h = u$ gives $\|u\|_2^2 \leq 0$, a contradiction. Hence $C = \operatorname{int}_A(C)$. Thus C is both relatively open and relatively closed in its affine hull A. Since A is connected and $C \neq \emptyset$, it follows that $C = A$. Therefore C is affine. Finally, since T^{-1} is an invertible linear map, $\mathscr{W} = T^{-1}(C)$ is affine as well. This proves the converse direction, and hence the proposition. □

The result in Proposition 3.1 is particularly relevant for structured versions of Equation (5) in which the eliminated variable is constrained to a structured matrix set $\mathscr{W}$ and first-order methods are used. The proposition shows that global differentiability of the lower-level solution map is restrictive: for arbitrary right-hand sides, it essentially forces the constraint set to be affine. In Section 6.3 we present a numerical example in which the constraint set $\mathscr{W}$ is given by an affine subspace, representing convolutional linear transforms.

4 The algorithm

In this section, we present the proposed Constrained Regularized Variable Projection method. The method is applied to the reduced problem

$$\min_{\theta \in \mathcal{C}_A \times \mathbb{R}^{p_B}} \widehat{f}_U(\theta), \qquad \theta = (\theta_A, \theta_B), \tag{11}$$

Algorithm 1 Constrained Regularized Variable Projection (CR-VarPro)

Input: $\theta^{(0)} \in \mathcal{C}$, T_{max}
Set $t = 0$
while $t \leq T_{max}$ **do**
  Assemble hypergradient $\nabla \widehat{f}_U(\theta^{(t)})$
  Set $\bar{\theta}^{(t)} = \text{BLMO}(\theta^{(t)}, \nabla \widehat{f}_U(\theta^{(t)}))$
  if $\theta^{(t)}$ stationary **then**
    STOP
  end if
  Set $\theta^{(t+1)} = \theta^{(t)} + \alpha_t(\bar{\theta}^{(t)} - \theta^{(t)})$, with $\alpha_t \in (0, 1]$ stepsize chosen via a line search
  Set $t = t + 1$
end while
return $\theta^{(t)}$

Algorithm 2 Block Linear Minimization Oracle (BLMO)

Input: $\tilde{\theta} = (\tilde{\theta}_A, \tilde{\theta}_B) \in \mathcal{C}_A \times \mathbb{R}^{p_B}$, $\nabla \widehat{f}_U(\tilde{\theta})$
Get $\bar{\theta}_A := \underset{\theta_A \in \mathcal{C}_A}{\arg\min} \langle \theta_A, \nabla_{\theta_A} \widehat{f}_U(\tilde{\theta}) \rangle$
Get $\bar{\theta}_B := \underset{\|\theta_B - \tilde{\theta}_B\| \leq 1}{\arg\min} \langle \theta_B, \nabla_{\theta_B} \widehat{f}_U(\tilde{\theta}) \rangle$
return $(\bar{\theta}_A, \bar{\theta}_B)$

where $\mathcal{C}_A$ is compact and convex, while the second block is unconstrained. The reduced objective is defined as

$$\widehat{f}_U(\theta) := f_U(\widehat{w}(\theta), \theta), \qquad \widehat{w}(\theta) \in \arg\min_w f_L(w, \theta).$$

Therefore, the lower-level variable w is eliminated and the algorithm only updates the outer variable θ. The gradient used in the algorithm is the hypergradient $\nabla \widehat{f}_U(\theta)$, which includes the dependence of the lower-level solution $\widehat{w}(\theta)$ on θ. The detailed scheme is reported in Algorithm 1.

Starting from an initial feasible point $\theta^{(0)} \in \mathcal{C} := \mathcal{C}_A \times \mathbb{R}^{p_B}$, Algorithm 1 sets $t = 0$ and repeats the following operations until either stationarity is reached or the maximum number of iterations is exceeded. At iteration t, the first step consists in assembling the hypergradient $\nabla \widehat{f}_U(\theta^{(t)})$.

Once the hypergradient has been computed, Algorithm 1 calls the Block Linear Minimization Oracle described in Algorithm 2 and sets

$$\bar{\theta}^{(t)} = \text{BLMO}\left(\theta^{(t)}, \nabla \widehat{f}_U(\theta^{(t)})\right).$$

The algorithm then checks whether $\theta^{(t)}$ is stationary. Equivalently, one may use the Frank-Wolfe gap

$$g_t := -\left\langle \nabla \widehat{f}_U(\theta^{(t)}), d^{(t)} \right\rangle$$

as a stationarity measure, with $d^{(t)} := \bar{\theta}^{(t)} - \theta^{(t)}$. If g_t is lower than a given threshold, the algorithm stops. Otherwise, a stepsize $\alpha_t \in (0, 1]$ is chosen by line search and the new iterate is computed as $\theta^{(t+1)} = \theta^{(t)} + \alpha_t \left(\bar{\theta}^{(t)} - \theta^{(t)}\right)$.

We now clarify the role of the BLMO reported in Algorithm 2. At iteration t, Algorithm 1 computes the point $\bar{\theta}^{(t)}$ and defines the search direction

$$d^{(t)} := \bar{\theta}^{(t)} - \theta^{(t)} = \left(d_A^{(t)}, d_B^{(t)}\right),$$

where

$$d_A^{(t)} := \bar{\theta}_A^{(t)} - \theta_A^{(t)}, \qquad d_B^{(t)} := \bar{\theta}_B^{(t)} - \theta_B^{(t)}.$$

The first block of the BLMO coincides with the usual Frank-Wolfe LMO over the compact set $\mathcal{C}_A$. The second block is different: instead of minimizing the linear model over all of $\mathbb{R}^{p_B}$, it minimizes over the unit ball centered at the current iterate. Indeed, a standard LMO over the full feasible set would require solving

$$\min_{\theta_A \in \mathcal{C}_A,\, \theta_B \in \mathbb{R}^{p_B}} \left\langle \theta_A, \nabla_{\theta_A} \widehat{f}_U(\theta^{(t)}) \right\rangle + \left\langle \theta_B, \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) \right\rangle.$$

The first term is well-defined because $\mathcal{C}_A$ is compact, but the second term is unbounded from below whenever $\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) \neq 0$. Thus, the standard Frank-Wolfe LMO is not well-defined on $\mathcal{C}_A \times \mathbb{R}^{p_B}$. The BLMO avoids this issue by keeping the Frank-Wolfe oracle on the compact set $\mathcal{C}_A$ and replacing the unbounded linear minimization problem by a local one on the unconstrained variables. In particular, if $\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) \neq 0$, then

$$\bar{\theta}_B^{(t)} = \theta_B^{(t)} - \frac{\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)})}{\left\| \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) \right\|},$$

while if $\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) = 0$, one can choose $\bar{\theta}_B^{(t)} = \theta_B^{(t)}$. Consequently, the BLMO combines a Frank-Wolfe step on $\mathcal{C}_A$ with a normalized gradient descent step on $\mathbb{R}^{p_B}$. The latter can be interpreted as a norm-constrained LMO step, since it is equivalent to

$$d_B^{(t)} \in \arg\min_{\|d_B\| \leq 1} \left\langle d_B, \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) \right\rangle.$$

This is the update principle used in SCION-type methods, where descent directions are generated through LMOs over norm balls rather than by using the raw gradient directly [47].

5 Theoretical analysis

We now analyze Algorithm 1. Throughout this section, the objective is the reduced function $\widehat{f}_U$. We assume that $\nabla \widehat{f}_U$ is Lipschitz continuous with constant $L > 0$ on the level set generated by the algorithm, and that $\widehat{f}_U$ is bounded from below on $\mathcal{C}_A \times \mathbb{R}^{p_B}$. We denote

$$\widehat{f}_U^* := \inf_{\theta \in \mathcal{C}_A \times \mathbb{R}^{p_B}} \widehat{f}_U(\theta).$$

Let $\Delta_A := \max_{\theta_A, \tilde{\theta}_A \in \mathcal{C}_A} \|\theta_A - \tilde{\theta}_A\|$ be the diameter of the compact block. Since the BLMO satisfies $\|d_B^{(t)}\| \leq 1$, we define $\bar{\Delta} := \sqrt{\Delta_A^2 + 1}$. Then, for every iteration t,

$$\|d^{(t)}\|^2 = \|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \leq \Delta_A^2 + 1 = \bar{\Delta}^2.$$

Note that g_t represents a valid stationarity measure for the product set $\mathcal{C}_A \times \mathbb{R}^{p_B}$. Indeed, by the construction of the BLMO,

$$-\left\langle \nabla_{\theta_A} \widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle \geq 0, \qquad -\left\langle \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle \geq 0.$$

Moreover, $g_t = 0$ if and only if $\left\langle \nabla_{\theta_A} \widehat{f}_U(\theta^{(t)}), \theta_A - \theta_A^{(t)} \right\rangle \geq 0$, $\forall \theta_A \in \mathcal{C}_A$, and $\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) = 0$. Thus, $g_t = 0$ is equivalent to the first-order stationarity condition for the constrained block $\mathcal{C}_A$ together with the unconstrained stationarity condition for the block $\mathbb{R}^{p_B}$.

Finally, we assume that the stepsize rule satisfies the conditions:

$$\alpha_t \geq \bar{\alpha}_t := \min\left\{1, \frac{g_t}{L\bar{\Delta}^2}\right\}, \tag{12}$$

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) \geq \rho \bar{\alpha}_t g_t \tag{13}$$

for some fixed $\rho > 0$.

Theorem 5.1. *Let $\{\theta^{(t)}\}$ be the sequence generated by Algorithm 1. Assume that $\nabla \widehat{f}_U$ is Lipschitz continuous with constant $L > 0$, that $\widehat{f}_U$ is bounded below by $\widehat{f}_U^*$, and that the stepsize satisfies* (12) *and* (13)*. Define $g_T^* := \min_{0\le t\le T-1} g_t$. Then, for every $T \in \mathbb{N}$,*

$$g_T^* \le \max\left\{ \sqrt{\frac{L\bar{\Delta}^2\left(\widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^*\right)}{\rho T}}, \frac{2\left(\widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^*\right)}{T} \right\}. \tag{14}$$

Proof. In order to prove the result, we distinguish two different cases.

Case 1. $\bar{\alpha}_t < 1$.

Then, by definition of $\bar{\alpha}_t$, we have $\bar{\alpha}_t = \frac{g_t}{L\bar{\Delta}^2}$. Using the sufficient decrease condition (13), we get

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) = \widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \alpha_t d^{(t)}) \ge \frac{\rho g_t^2}{L\bar{\Delta}^2}. \tag{15}$$

Case 2. $\bar{\alpha}_t = 1$.

Since $\alpha_t \in (0, 1]$ and $\alpha_t \ge \bar{\alpha}_t$, the condition $\bar{\alpha}_t = 1$ implies $\alpha_t = 1$. By the standard descent lemma [6, Proposition 6.1.2] applied to $\widehat{f}_U$ with center $\theta^{(t)}$ and direction $d^{(t)}$, we have

$$\begin{aligned}
\widehat{f}_U(\theta^{(t+1)}) &= \widehat{f}_U(\theta^{(t)} + d^{(t)}) \\
&\le \widehat{f}_U(\theta^{(t)}) + \left\langle \nabla \widehat{f}_U(\theta^{(t)}), d^{(t)} \right\rangle + \frac{L}{2}\|d^{(t)}\|^2 \\
&= \widehat{f}_U(\theta^{(t)}) + \left\langle \nabla_{\theta_A} \widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle + \left\langle \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle + \frac{L}{2}\left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2\right) \\
&\le \widehat{f}_U(\theta^{(t)}) - g_t + \frac{L}{2}\bar{\Delta}^2.
\end{aligned}$$

In the last inequality, we used the definition $g_t := -\left\langle \nabla \widehat{f}_U(\theta^{(t)}), d^{(t)} \right\rangle$ and the bound

$$\|d^{(t)}\|^2 = \|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \le \Delta_A^2 + 1 = \bar{\Delta}^2.$$

Moreover, since $\bar{\alpha}_t = \min\left\{1, \frac{g_t}{L\bar{\Delta}^2}\right\} = 1$, we have $\frac{g_t}{L\bar{\Delta}^2} \ge 1$, and hence $g_t \ge L\bar{\Delta}^2$.

Therefore,

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) \ge g_t - \frac{L}{2}\bar{\Delta}^2 \ge \frac{g_t}{2}. \tag{16}$$

Now, based on the two cases above, we partition the iterations $\{0, 1, \ldots, T-1\}$ into

$$N_1 := \{t < T : \bar{\alpha}_t < 1\}, \qquad N_2 := \{t < T : \bar{\alpha}_t = 1\}.$$

Using (15) and (16), we obtain

$$\begin{aligned}
\widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^* &\ge \sum_{t=0}^{T-1}\left(\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)})\right) \\
&= \sum_{t\in N_1}\left(\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)})\right) + \sum_{t\in N_2}\left(\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)})\right) \\
&\ge \sum_{t\in N_1}\frac{\rho g_t^2}{L\bar{\Delta}^2} + \sum_{t\in N_2}\frac{g_t}{2} \\
&\ge |N_1| \min_{t\in N_1}\frac{\rho g_t^2}{L\bar{\Delta}^2} + |N_2| \min_{t\in N_2}\frac{g_t}{2} \\
&\ge (|N_1| + |N_2|)\min\left\{\frac{\rho (g_T^*)^2}{L\bar{\Delta}^2}, \frac{g_T^*}{2}\right\} \\
&= T\min\left\{\frac{\rho (g_T^*)^2}{L\bar{\Delta}^2}, \frac{g_T^*}{2}\right\},
\end{aligned}$$

where in the last inequality we used the definition of g_T^*. Hence,

$$T \min\left\{ \frac{\rho(g_T^*)^2}{L\bar{\Delta}^2}, \frac{g_T^*}{2} \right\} \leq \widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^*.$$

To finish, if $T \min\left\{ \frac{\rho(g_T^*)^2}{L\bar{\Delta}^2}, \frac{g_T^*}{2} \right\} = T\frac{g_T^*}{2}$, then

$$g_T^* \leq \frac{2\left(\widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^*\right)}{T}. \tag{17}$$

Otherwise,

$$g_T^* \leq \sqrt{\frac{L\bar{\Delta}^2\left(\widehat{f}_U(\theta^{(0)}) - \widehat{f}_U^*\right)}{\rho T}}. \tag{18}$$

The claim follows by taking the maximum in the system formed by (17) and (18). □

The optimality condition follows from the definition of g_t. Indeed, by construction of the BLMO, both quantities

$$-\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle, \qquad -\left\langle \nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle$$

are nonnegative. Therefore, if $g_t = 0$, then both of them must be equal to zero.

For the constrained block, we have

$$-\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle = 0.$$

Since $d_A^{(t)} = \bar{\theta}_A^{(t)} - \theta_A^{(t)}$ is generated by the Frank-Wolfe linear oracle on $\mathcal{C}_A$, it satisfies

$$\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), \bar{\theta}_A^{(t)} \right\rangle \leq \left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), \theta_A \right\rangle \qquad \forall \theta_A \in \mathcal{C}_A.$$

Equivalently,

$$\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle \leq \left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), \theta_A - \theta_A^{(t)} \right\rangle \qquad \forall \theta_A \in \mathcal{C}_A.$$

Since the left-hand side is zero, we obtain

$$\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), \theta_A - \theta_A^{(t)} \right\rangle \geq 0 \qquad \forall \theta_A \in \mathcal{C}_A.$$

Thus, the first-order optimality condition holds with respect to the constrained block θ_A.

For the unconstrained block, we have

$$-\left\langle \nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle = 0.$$

The BLMO defines $d_B^{(t)}$ as the solution of

$$d_B^{(t)} \in \arg\min_{\|d_B\| \leq 1} \left\langle d_B, \nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}) \right\rangle.$$

Hence, if $\nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}) \neq 0$, then $d_B^{(t)} = -\frac{\nabla_{\theta_B}\widehat{f}_U(\theta^{(t)})}{\left\|\nabla_{\theta_B}\widehat{f}_U(\theta^{(t)})\right\|}$, and therefore

$$-\left\langle \nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle = \left\|\nabla_{\theta_B}\widehat{f}_U(\theta^{(t)})\right\| > 0,$$

which contradicts the equality above. Consequently, $\nabla_{\theta_B}\widehat{f}_U(\theta^{(t)}) = 0$. Therefore, $g_t = 0$ is equivalent to the stationarity conditions

$$\left\langle \nabla_{\theta_A}\widehat{f}_U(\theta^{(t)}), \theta_A - \theta_A^{(t)} \right\rangle \geq 0 \qquad \forall \theta_A \in \mathcal{C}_A,$$

and

$$\nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}) = 0.$$

In the following lemma, we show that the conditions (12) and (13) can be satisfied by standard stepsize rules. In particular, this is true when $\alpha_t = \bar{\alpha}_t$, and also when α_t is determined by an Armijo line search (see [8, 9] for further details). The Armijo rule defines

$$\alpha_t = \delta^j, \tag{19}$$

where j is the smallest nonnegative integer such that

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \alpha_t d^{(t)}) \geq \gamma \alpha_t g_t. \tag{20}$$

Here, $\gamma \in (0, \frac{1}{2})$ and $\delta \in (0, 1)$ are fixed constants.

Lemma 5.2. *The bound condition on the stepsize*

$$\alpha_t \geq \bar{\alpha}_t := \min\left\{1, \frac{g_t}{L\bar{\Delta}^2}\right\}, \tag{21}$$

and the sufficient decrease condition

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) \geq \rho \bar{\alpha}_t g_t \tag{22}$$

hold under the following stepsize rules:

- *If $\alpha_t = \bar{\alpha}_t$, then the condition holds with $\rho = \frac{1}{2}$.*
- *If α_t is determined by the Armijo line search rule* (20)*, then the condition holds with*

$$\rho = \gamma \min\{1, 2\delta(1-\gamma)\}.$$

Proof. By the standard descent lemma [6, Proposition 6.1.2], for every $\alpha \in [0, 1]$, we have

$$\begin{aligned}
\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \alpha d^{(t)}) &\geq -\alpha \left\langle \nabla \widehat{f}_U(\theta^{(t)}), d^{(t)} \right\rangle - \alpha^2 \frac{L}{2} \|d^{(t)}\|^2 \\
&= -\alpha \left\langle \nabla_{\theta_A} \widehat{f}_U(\theta^{(t)}), d_A^{(t)} \right\rangle - \alpha \left\langle \nabla_{\theta_B} \widehat{f}_U(\theta^{(t)}), d_B^{(t)} \right\rangle \\
&\quad - \alpha^2 \frac{L}{2} \left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \right) \\
&= \alpha g_t - \alpha^2 \frac{L}{2} \left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \right).
\end{aligned} \tag{23}$$

First, assume that $\alpha_t = \bar{\alpha}_t$. Then the stepsize lower bound (21) is trivially satisfied. Moreover, from (23), it is immediate that

$$\alpha g_t - \alpha^2 \frac{L}{2} \left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \right) \geq \alpha \frac{g_t}{2} \tag{24}$$

for every

$$0 \leq \alpha \leq \frac{g_t}{L \left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \right)}.$$

We can apply (24) to $\bar{\alpha}_t$, since

$$0 \leq \bar{\alpha}_t \leq \frac{g_t}{L\bar{\Delta}^2} \leq \frac{g_t}{L \left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \right)},$$

where the last inequality follows from $\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \leq \bar{\Delta}^2$. Therefore,

$$\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) = \widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \bar{\alpha}_t d^{(t)}) \geq \bar{\alpha}_t \frac{g_t}{2}.$$

Thus, the sufficient decrease (22) condition holds with $\rho = \frac{1}{2}$.

Now assume that α_t is determined by the Armijo line search rule. From (23), the Armijo condition $\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \alpha d^{(t)}) \geq \gamma \alpha g_t$, is satisfied whenever

$$0 \leq \alpha \leq 2(1-\gamma)\frac{g_t}{L\left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2\right)}.$$

By the standard backtracking argument, the accepted stepsize satisfies

$$\alpha_t \geq \min\left\{1, 2\delta(1-\gamma)\frac{g_t}{L\left(\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2\right)}\right\}.$$

Using again $\|d_A^{(t)}\|^2 + \|d_B^{(t)}\|^2 \leq \bar{\Delta}^2$, we get

$$\alpha_t \geq \min\left\{1, 2\delta(1-\gamma)\frac{g_t}{L\bar{\Delta}^2}\right\} \geq \min\{1, 2\delta(1-\gamma)\}\bar{\alpha}_t, \tag{25}$$

we thus have $\alpha_t \geq \min\left\{1, c\frac{g_t}{L\Delta^2}\right\}$,

for some $c > 0$. We have two cases: if $c \geq 1$ the lower bound (21) is trivially satisfied. If $c < 1$ we can still satisfy equation (21) by considering $\tilde{L} = L/c$ instead of L as Lipschitz constant. Finally, using the Armijo condition (20) and then (25), we obtain

$$\begin{aligned}\widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t+1)}) = \widehat{f}_U(\theta^{(t)}) - \widehat{f}_U(\theta^{(t)} + \alpha_t d^{(t)}) \geq \gamma\alpha_t g_t \geq \\ \geq \gamma \min\{1, 2\delta(1-\gamma)\}\bar{\alpha}_t g_t.\end{aligned}$$

Hence the sufficient decrease condition (22) holds with $\rho = \gamma \min\{1, 2\delta(1-\gamma)\}$. □

6 Experimental results

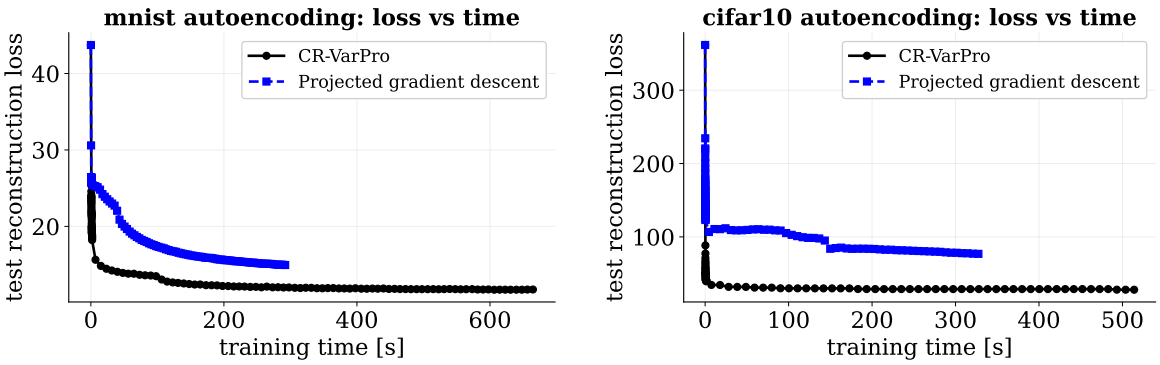
6.1 Example 1: Sparse autoencoder and latent dimension selection

Autoencoding has a variety of different applications, from representation learning [52], dimensionality reduction [35], image denoising and generation [5], to neural network interpretability [36, 2]. The problem of autoencoding has a number of variants, but in its basic version, it consists of reconstructing the identity map on a dataset through the composition of learnable encoding and decoding functions as a solution of the optimization problem $\min_{\theta \in \mathcal{C}} \frac{1}{2}\|D \circ E(X) - Y\|_F^2$, where E and D are the encoder and decoder neural networks, respectively. In this experiment, we consider the asymmetric setting with $D_\phi(x) = Wx, W \in \mathbb{R}^{D\times d}$, and $E_\theta(x)$ is a feedforward neural network with a hyperbolic tangent activation function. In this setting, the autoencoding problem can be reformulated in the setting of Equation (8) as

$$\begin{cases} \min\limits_{e^\top s \leq \rho, s \geq 0, \phi \in \mathbb{R}^p} \|\widehat{W}\,\mathrm{diag}(s) E_\phi(X_1) - X_1\|_F^2 + \lambda\|s\|^2 \\ s.t.\ \widehat{W} \in \operatorname{arg\,min}\limits_{W \in \mathbb{R}^{D\times d}} \|W\,\mathrm{diag}(s) E_\phi(X_2) - X_2\|_F^2, \end{cases}$$

which fits exactly the formulation in Equation (8) with $\theta = (s, \phi) \in \mathbb{R}^{d+p}$,$w = \mathrm{vec}(W)$, $M_U(\theta) = E_\phi(X_1)^\top \mathrm{diag}(s) \otimes I$,$M_L(\theta) = E_\phi(X_2)^\top \mathrm{diag}(s) \otimes I$, $y_U \equiv X_1, y_L \equiv X_2$ and $\mathcal{C}_A = B_{\|\cdot\|_1}(0, \rho] \cap \mathbb{R}^d_{\geq 0}, \mathcal{C}_B = \mathbb{R}^p$.

In Figure 1, we report the results of the method presented in Algorithm 1 in the setting in which E is a four-layer neural network with intermediate dimensions $[256, 128, 64, 32]$, $\rho = 1, \lambda = 10^{-5}$, and the datasets X_1, X_2 are two splits of the MNIST training dataset. We compare the performance of Algorithm 1 with a standard projected gradient descent on the joint problem 1 with a budget of 20K optimization steps. As we can observe from the results in Figure 1, despite the higher per-iteration computational cost, Bilevel VarPro is able to achieve a lower test reconstruction loss given a fixed time budget. We repeat the same experiment with the CIFAR10 dataset for 12K optimization steps, with the only difference that E is a four-layer neural network with doubled intermediate, i.e., $[512, 256, 128, 64]$.



(a) MNIST [42] (b) CIFAR10 [40]

Figure 1: Convergence of CR-VarPro against Projected gradient descent for the Autoencoding problem of MNIST and CIFAR10 on a fully connected neural network.

6.2 Example 2: Dictionary learning

We consider here the dictionary-learning problem [49], where the goal is to find a factorization of a given matrix $X \in \mathbb{R}^{D\times N}$ into a dictionary $D \in \mathbb{R}^{D\times d}$ and a norm-constrained representation matrix $R \in \mathbb{R}^{d\times N}$. This problem has many applications in data science and machine learning, including recommender systems [16], data compression [49], and interpretability of large language models [36]. We consider here the specialized setting in which we are interested in the reconstruction error in terms of the Frobenius norm

$$\min_{D\in\mathbb{R}^{D\times d}, R\in\mathbb{R}^{d\times N}} \|DR - X\|^2 + \lambda\|\Omega R\|^2, \quad \text{s.t. } \|D\|_1 \le \rho, \tag{26}$$

where $\|\cdot\|_1$ indicates the entrywise L^1 norm. Notice that this problem is exactly a special case of the formulation Equation (2) with $\theta = \mathrm{vec}(D), M(\theta) = I \otimes D, w = \mathrm{vec}(R), y = \mathrm{vec}(X)$. In particular, the minimizers of Equation (26) are minimizers of the bilevel-reformulated version

$$\begin{cases} \min\limits_{D\in\mathbb{R}^{D\times d}} \|D\widehat{R} - X\|^2 + \lambda\|\Omega\widehat{R}\|^2, \\ \text{s.t.} \quad \widehat{R} \in \arg\min\limits_{R\in\mathbb{R}^{d\times N}} \|DR - X\|^2 + \lambda\|\Omega R\|^2, \quad \|D\|_1 \le \rho. \end{cases} \tag{27}$$

The lower-level optimization problem can be solved in closed form, leading to $\widehat{R} = (D^\top D + \lambda\Omega^\top\Omega)^{-1}D^\top X$. In particular, Equation (27) can be reformulated as the following optimization problem

$$\begin{cases} \min\limits_{D\in\mathbb{R}^{D\times d}} \widehat{f}(D) := \|D(D^\top D + \lambda\Omega^\top\Omega)^{-1}D^\top X - X\|^2 + \lambda\|\Omega(D^\top D + \lambda\Omega^\top\Omega)^{-1}D^\top X\|^2, \\ \text{s.t.} \quad \|D\|_1 \le \rho. \end{cases} \tag{28}$$

For this numerical experiment, we set $N = 3000, d = 128, D = 784$, $\Omega = I, \rho = 1, \lambda = 10^{-5}$ and $X \in \mathbb{R}^{D\times N}$ is a randomly sampled subset of the MNIST dataset [42]. In Figure 2 we present the numerical results of Equation (28), comparing Algorithm 1 with Projected gradient descent on the joint problem Equation (26) for a total of 10K optimization steps (with objective value reported every 25 steps).

6.3 Example 3: Blind deconvolution

Blind deconvolution is a common problem in image processing, which requires reconstructing a corrupted signal without prior knowledge on the smoothing process. More precisely, assume we are given a batch of N measured signals $Y \in \mathbb{R}^{D\times N}$, and we know that they have been corrupted by a smoothing process and additional noise $y_i = w * x_i + \varepsilon_i$, and we assume to have no knowledge on the kernel w. The problem of blind deconvolution is trying to recover the batch of original signals $X = [x_1, \dots, x_N]$ given the corrupted ones $Y = [y_1, \dots, y_N]$. We remark that, even without any additional noise (i.e., $\varepsilon_i = 0$)

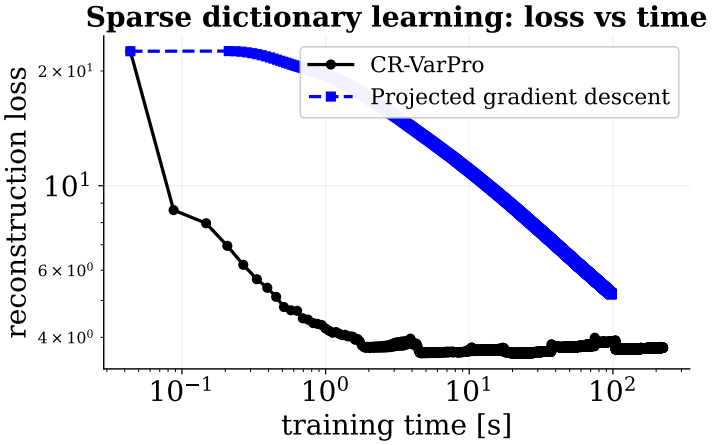


Figure 2: Convergence of CR-VarPro against Projected gradient descent for the Dictionary Learning problem. The number of iterations is the same for the two methods, a budget of 10000 optimization steps.

the problem is still ill posed. In particular, assuming $\varepsilon_i = 0$, the problem becomes to find x_i, w such that $w * x_i = y_i$ for all $i = 1, \dots, N$. However, note that given a pair $(x_i^\star, w^\star)$ that solves the problem, then $(\alpha x_i^\star, \alpha^{-1} w^\star)$ is still a solution for all $\alpha \neq 0$. More than this scalar invariance (which already produces a continuum of solutions), there is a full diagonal invariance, as $\mathcal{F}(w * x_i) = \mathcal{F}(w) \odot \mathcal{F}(x_i) = \mathcal{F}(y_i)$ and any couple $(\mathcal{F}(w) \odot d, \mathcal{F}(x_i) \odot d^{\odot -1})$ is still a solution. In order to make the recovery problem well-posed, some additional constraints are typically needed. Assuming that the noise $\varepsilon_i \underset{i.i.d}{\sim} p$, and assuming that we have a prior knowledge on $w \sim q$ and we know that the original signal is constrained $x_i \in \mathcal{C}$ (e.g., box constraint or norm-based), reconstruction can be posed as a constrained maximum a posteriori estimation

$$\min_{w, x_i \in \mathcal{C}} \sum_{i=1}^{N} p(y_i - w * x_i) + q(w).$$

In the specific case in which we assume a Gaussian prior $q(w) \propto e^{-\lambda \|w\|_2^2}$ and we assume that the noise is Gaussian as well, we get the objective function $\min_{W \in \mathscr{W}, X \in \mathcal{C}} \|WX - Y\|_F^2 + \lambda \|W\|_F^2$, leading back to the problem formulation we presented in Section 3.4. In particular, if we assume to use circular convolution and $w \in \mathbb{R}^D, X \in \mathbb{R}^{D \times N}$, the problem can be solved in closed form in W in the Fourier domain, and it reduces to a non-linear constrained minimization problem in X,

$$\min_{X \in \mathcal{C}} \|w^*(X) * X - Y\|_F^2 + \lambda \|w^*(X)\|_F^2, \quad w^*(X) := \mathcal{F}^{-1} \left\{ \frac{\operatorname{diag}(\mathcal{F}\{Y\}\mathcal{F}\{X\}^H)}{\operatorname{diag}(\mathcal{F}\{X\}\mathcal{F}\{X\}^H + \lambda I)} \right\}, \quad (29)$$

which can be solved using the Frank-Wolfe algorithm. In Figure 3, we present convergence results (wall-time against objective function) for Algorithm 1 and Projected gradient descent. To do this, we considered the specific instance of the problem presented in Equation (29) in the case in which $\mathcal{C}_A := B_{\|\cdot\|_F}(0, \rho), \mathcal{C}_B = \{0\}$. In Figure 3 we report the results of wall time against objective function value for the blind deconvolution problem, for a total of 5K iterations $\rho = 1, N = 50, D = 128, \lambda = 10^{-3}$.

6.4 Example 4: Few-shot learning

As a final numerical experiment, we test the performance of Algorithm 1 in few-shot learning on CIFAR10 [40] starting from a pretrained ResNet-18 [34], and compare it with other fine-tuning approaches. In particular, the few-shot learning problem fits the formulation Equation (8) with $M_U(\theta) = \phi_\theta(X_1)^\top \otimes I, M_L(\theta) = \phi_\theta(X_2)^\top \otimes I, w = \operatorname{vec}(W)$ where ϕ is the backbone of the Resnet-18 architecture (all layers but the final linear classifier), and $X_1, X_2 \in \mathbb{R}^{32 \times 32 \times N}$ are two different splits of the CIFAR10 dataset. In this setting, θ is not composed of all trainable parameters of the backbone, but just the parameters of the fourth layer. For this particular experiment, we use a 10-way 1-shot setup, i.e.,

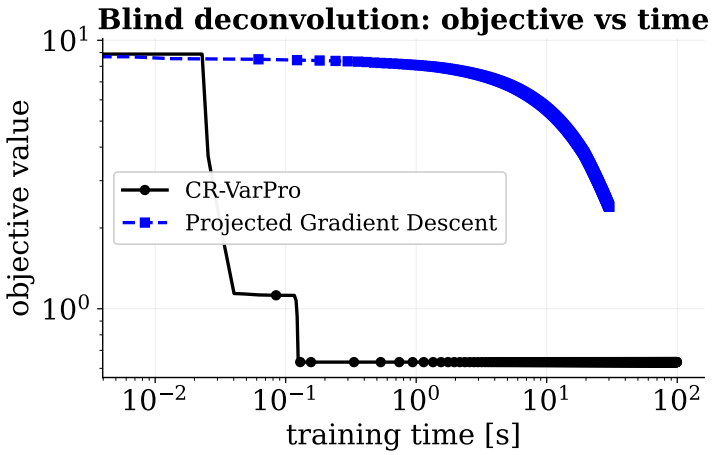


Figure 3: Convergence of CR-VarPro against Projected gradient descent for the blind deconvolution problem.

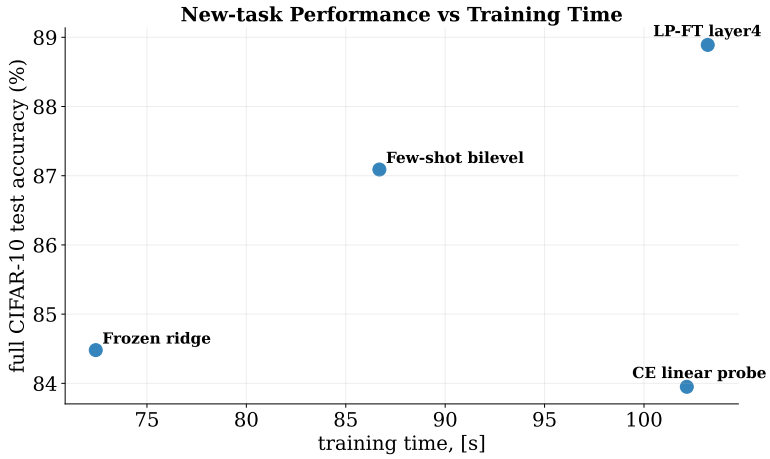


Figure 4: Fine tuning results ResNet-18 on CIFAR10. Frozen Ridge fits only the linear head as a solution of an L^2 regression problem, CE linear probe fits only the linear head through cross-entropy minimization, LP-FT layer4 fine-tunes the linear probe and the fourth intermediate layer. Few-shot bilevel performs 10-way 1-shot 1-query learning using the bilevel formulation in Equation (8).

$N = 10$, and we use one example per class on each split. We compare the performance of Bilevel VarPro with fine-tuning only the last linear layer using an L^2 loss (Frozen ridge in Figure 4), with cross-entropy loss (CE linear probe in Figure 4), and by fine-tuning the last linear layer and the fourth layer of ϕ. In Figure 4, we compare the final test accuracy of each method against the effective training time (not accounting for full accuracy calculation) in seconds, from which we can observe the effectiveness of the bilevel formulation in terms of data efficiency, training time, and overall performance.

7 Conclusions

We developed a constrained variable-projection framework for structured data-science models in which a least-squares block is eliminated exactly and the remaining variables are optimized over a convex feasible set. By interpreting variable projection as a collapsed bilevel problem, we derived reduced-gradient formulas that combine closed-form lower-level solves with automatic differentiation through vector-Jacobian products. This yields exact hypergradients without differentiating naively through normal equations, and naturally accommodates extensions in which the eliminated variable has additional affine structure.

We proposed a projection-free conditional-gradient method for the resulting reduced problem, combining a Frank–Wolfe oracle on the constrained block with a normalized descent step on unconstrained variables. Under standard smoothness and boundedness assumptions, we established convergence to first-order stationary points in terms of a Frank–Wolfe-type gap. The numerical experiments on sparse autoencoding, dictionary learning, blind deconvolution, and few-shot learning indicate that constrained variable projection can improve computational efficiency and data efficiency relative to natural joint-optimization baselines. Future work includes inexact lower-level solves, stochastic variants, and broader classes of structured lower-level constraints.

Acknowledgements The work of FT is partially funded by the PRIN-MUR project MOLE code 2022ZK5ME7 and by the PRIN-PNRR project FIN4GEO within the European Union's Next Generation EU framework, Mission 4, Component 2, CUP P2022BNB97. The work of E. Zangrando was funded by the MUR-PNRR project "Low-parametric machine learning".

References

[1] A. AGRAWAL, B. AMOS, S. BARRATT, S. BOYD, S. DIAMOND, AND J. Z. KOLTER, *Differentiable convex optimization layers*, in Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 9558–9570.

[2] G. ALAIN AND Y. BENGIO, *Understanding intermediate layers using linear classifier probes*, 2017.

[3] B. AMOS AND J. Z. KOLTER, *OptNet: Differentiable optimization as a layer in neural networks*, in Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 136–145.

[4] J. F. BARD, *Practical Bilevel Optimization: Algorithms and Applications*, vol. 30 of Nonconvex Optimization and Its Applications, Springer, Boston, MA, 1998, https://doi.org/10.1007/978-1-4757-2836-1.

[5] Y. BENGIO, L. YAO, G. ALAIN, AND P. VINCENT, *Generalized denoising auto-encoders as generative models*, 2013, https://arxiv.org/abs/1305.6663.

[6] D. P. BERTSEKAS, *Convex optimization algorithms*, Athena Scientific, Belmont, 2015.

[7] M. BLONDEL, Q. BERTHET, M. CUTURI, R. FROSTIG, S. HOYER, F. LLINARES-LÓPEZ, F. PEDREGOSA, AND J.-P. VERT, *Efficient and modular implicit differentiation*, in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 5230–5242.

[8] I. M. BOMZE, F. RINALDI, AND D. ZEFFIRO, *Active set complexity of the away-step Frank–Wolfe algorithm*, SIAM Journal on Optimization, 30 (2020), pp. 2470–2500, https://doi.org/10.1137/19M1309419.

[9] I. M. BOMZE, F. RINALDI, AND D. ZEFFIRO, *Frank–Wolfe and friends: a journey into projection-free first-order optimization methods*, 4OR, 19 (2021), pp. 313–345, https://doi.org/10.1007/s10288-021-00493-y.

[10] G. BRAUN, A. CARDERERA, C. W. COMBETTES, H. HASSANI, A. KARBASI, A. MOKHTARI, AND S. POKUTTA, *Conditional Gradient Methods: From Core Principles to AI Applications*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2025, https://doi.org/10.1137/1.9781611978568.

[11] J. J. BRUST, *Nonlinear least-squares for large-scale machine learning using stochastic jacobian estimates*, arXiv preprint arXiv:1412.6980, (2021).

[12] A. BÄRLIGEA, P. HOCHSTAFFL, AND F. SCHREIER, *A generalized variable projection algorithm for least squares problems in atmospheric remote sensing*, Mathematics, 11 (2023), https://doi.org/10.3390/math11132839.

[13] E. J. CANDÈS, X. LI, Y. MA, AND J. WRIGHT, *Robust principal component analysis?*, Journal of the ACM, 58 (2011), pp. 11:1–11:37, https://doi.org/10.1145/1970392.1970395.

[14] N. CARLINI AND D. WAGNER, *Towards evaluating the robustness of neural networks*, in 2017 IEEE Symposium on Security and Privacy, IEEE, 2017, pp. 39–57, https://doi.org/10.1109/SP.2017.49.

[15] G. CHAVENT, *Nonlinear least squares for inverse problems*, Scientific Computation, Springer, Dordrecht, Netherlands, 2010 ed., Oct. 2009, https://doi.org/10.1007/978-90-481-2785-6.

[16] C. CHEN, D. LI, J. YAN, AND X. YANG, *Modeling dynamic user preference via dictionary learning for sequential recommendation*, IEEE Transactions on Knowledge and Data Engineering, 34 (2022), pp. 5446–5458, https://doi.org/10.1109/TKDE.2021.3050407.

[17] G. CHEN, P. XUE, M. GAN, J. CHEN, W. GUO, AND C. P. CHEN, *Variable projection algorithms: Theoretical insights and a novel approach for problems with large residual*, Automatica, 177 (2025), p. 112300, https://doi.org/10.1016/j.automatica.2025.112300.

[18] B. COLSON, P. MARCOTTE, AND G. SAVARD, *An overview of bilevel optimization*, Annals of Operations Research, 153 (2007), pp. 235–256, https://doi.org/10.1007/s10479-007-0176-2.

[19] C. W. COMBETTES AND S. POKUTTA, *Complexity of linear minimization and projection on some sets*, Operations Research Letters, 49 (2021), pp. 565–571, https://doi.org/10.1016/j.orl.2021.06.019.

[20] S. DEMPE AND A. B. ZEMKOHO, eds., *Bilevel Optimization: Advances and Next Challenges*, vol. 161 of Springer Optimization and Its Applications, Springer, Cham, 2020, https://doi.org/10.1007/978-3-030-52119-6.

[21] S. DONG AND J. YANG, *Numerical approximation of partial differential equations by a variable projection method with artificial neural networks*, Computer Methods in Applied Mechanics and Engineering, 398 (2022), p. 115284, https://doi.org/10.1016/j.cma.2022.115284.

[22] D. L. DONOHO, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306, https://doi.org/10.1109/TIT.2006.871582.

[23] J. DUCHI, S. SHALEV-SHWARTZ, Y. SINGER, AND T. CHANDRA, *Efficient projections onto the ℓ_1-ball for learning in high dimensions*, in Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 272–279, https://doi.org/10.1145/1390156.1390191.

[24] M. DUS, *Grassmannian geometry and global convergence of variable projection for neural networks*, 2026, https://arxiv.org/abs/2601.22897.

[25] M. I. ESPANOL AND G. JERONIMO, *Local convergence analysis of a variable projection method for regularized separable nonlinear inverse problems*, SIAM Journal on Matrix Analysis and Applications, 46 (2025), pp. 858–878, https://doi.org/10.1137/24M1639087.

[26] L. FRANCESCHI, P. FRASCONI, S. SALZO, R. GRAZZI, AND M. PONTIL, *Bilevel programming for hyperparameter optimization and meta-learning*, in Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 1568–1577.

[27] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Nav. Res. Logist. Q., 3 (1956), pp. 95–110, https://doi.org/10.1002/nav.3800030109.

[28] N. GILLIS, *Nonnegative Matrix Factorization*, vol. 2 of Data Science, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2020, https://doi.org/10.1137/1.9781611976410.

[29] G. GOLUB AND V. PEREYRA, *Separable nonlinear least squares: the variable projection method and its applications*, Inverse problems, 19 (2003), p. R1, https://doi.org/10.1088/0266-5611/19/2/201.

[30] G. H. GOLUB AND R. LEVEQUE, *Extensions and uses of the variable projection algorithm for solving nonlinear least squares problems.*, in Proceedings of the 1979 Army Numerical Analysis and Computers Conference, 1979.

[31] G. H. GOLUB AND V. PEREYRA, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM Journal on Numerical Analysis, 10 (1973), pp. 413–432, https://doi.org/10.1137/0710036.

[32] S. GOULD, B. FERNANDO, A. CHERIAN, P. ANDERSON, R. SANTA CRUZ, AND E. GUO, *On differentiating parameterized argmin and argmax problems with application to bi-level optimization*, arXiv preprint arXiv:1607.05447, (2016), https://arxiv.org/abs/1607.05447.

[33] T. HASTIE, R. TIBSHIRANI, AND M. WAINWRIGHT, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC, Boca Raton, FL, 2015, https://doi.org/10.1201/b18401.

[34] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[35] G. E. HINTON AND R. R. SALAKHUTDINOV, *Reducing the dimensionality of data with neural networks*, Science, 313 (2006), pp. 504–507, https://doi.org/10.1126/science.1127647.

[36] R. HUBEN, H. CUNNINGHAM, L. R. SMITH, A. EWART, AND L. SHARKEY, *Sparse autoencoders find highly interpretable features in language models*, in The Twelfth International Conference on Learning Representations, 2024.

[37] M. JAGGI, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, in Proceedings of the 30th International Conference on Machine Learning, S. Dasgupta and D. McAllester, eds., vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA, 17–19 Jun 2013, PMLR, pp. 427–435.

[38] M. JAGGI, *An equivalence between the lasso and support vector machines*, in Regularization, Optimization, Kernels, and Support Vector Machines, J. A. K. Suykens, ed., Taylor & Francis, 2014, pp. 1–26, https://doi.org/10.1201/b17558-4.

[39] L. KAUFMAN, *A variable projection method for solving separable nonlinear least squares problems*, BIT Numerical Mathematics, 15 (1975), pp. 49–57, https://doi.org/10.1007/BF01932995.

[40] A. KRIZHEVSKY AND G. HINTON, *Learning multiple layers of features from tiny images*, Tech. Report 0, University of Toronto, Toronto, Ontario, 2009.

[41] W. H. LAWTON AND E. A. SYLVESTRE, *Elimination of linear parameters in nonlinear regression*, Technometrics, 13 (1971), pp. 461–467, https://doi.org/10.2307/1267160.

[42] Y. LECUN, C. CORTES, AND C. J. BURGES, *The mnist database of handwritten digits*, (1998).

[43] C. H. LIM AND S. J. WRIGHT, *A box-constrained approach for hard permutation problems*, in Proceedings of the 33rd International Conference on Machine Learning, vol. 48 of Proceedings of Machine Learning Research, PMLR, 2016, pp. 2454–2463.

[44] I. MARKOVSKY, *Recent progress on variable projection methods for structured low-rank approximation*, Signal Processing, 96 (2014), p. 406–419, https://doi.org/10.1016/j.sigpro.2013.09.021.

[45] E. NEWMAN, J. CHUNG, M. CHUNG, AND L. RUTHOTTO, *slimtrain—a stochastic approximation method for training separable deep neural networks*, SIAM Journal on Scientific Computing, 44 (2022), pp. A2322–A2348, https://doi.org/10.1137/21M1452512.

[46] E. NEWMAN, L. RUTHOTTO, J. HART, AND B. VAN BLOEMEN WAANDERS, *Train like a (var)pro: Efficient training of neural networks with variable projection*, SIAM Journal on Mathematics of Data Science, 3 (2021), pp. 1041–1066, https://doi.org/10.1137/20M1359511.

[47] T. PETHICK, W. XIE, K. ANTONAKOPOULOS, Z. ZHU, A. SILVETI-FALLS, AND V. CEVHER, *Training deep learning models with norm-constrained LMOs*, in Proceedings of the 42nd International Conference on Machine Learning, vol. 267 of Proceedings of Machine Learning Research, PMLR, 2025, pp. 49069–49104.

[48] B. RECHT, M. FAZEL, AND P. A. PARRILO, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review, 52 (2010), pp. 471–501, https://doi.org/10.1137/070697835.

[49] R. RUBINSTEIN, A. M. BRUCKSTEIN, AND M. ELAD, *Dictionaries for sparse representation modeling*, Proceedings of the IEEE, 98 (2010), pp. 1045–1057, https://doi.org/10.1109/JPROC.2010.2040551.

[50] D. B. C. SALZER, M. I. ESPAÑOL, AND G. JERONIMO, *Variable projection methods for solving regularized separable inverse problems with applications to semi-blind image deblurring*, 2026, https://arxiv.org/abs/2601.05224.

[51] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), 58 (1996), pp. 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[52] M. TSCHANNEN, O. BACHEM, AND M. LUCIC, *Recent advances in autoencoder-based representation learning*, 2018, https://arxiv.org/abs/1812.05069.

[53] K. USEVICH AND I. MARKOVSKY, *Variable projection for affinely structured low-rank approximation in weighted 2-norms*, Journal of Computational and Applied Mathematics, 272 (2014), pp. 430–448, https://doi.org/10.1016/j.cam.2013.04.034.

[54] K. USEVICH AND I. MARKOVSKY, *Variable projection methods for approximate (greatest) common divisor computations*, Theoretical Computer Science, 681 (2017), pp. 176–198, https://doi.org/10.1016/j.tcs.2017.03.028. Symbolic Numeric Computation.

[55] T. VAN LEEUWEN AND A. Y. ARAVKIN, *Variable projection for nonsmooth problems*, SIAM Journal on Scientific Computing, 43 (2021), pp. S249–S268, https://doi.org/10.1137/20M1348650.

[56] H.-L. XU, G.-Y. CHEN, S.-Q. CHENG, M. GAN, AND J. CHEN, *Variable projection algorithms with sparse constraint for separable nonlinear models*, Control Theory and Technology, 22 (2024), pp. 135–146, https://doi.org/10.1007/s11768-023-00194-3.

[57] E. ZANGRANDO, S. VENTURINI, F. RINALDI, AND F. TUDISCO, *dEBORA: Efficient bilevel optimization-based low-rank adaptation*, in International Conference on Learning Representations, 2025.