Machine Learning (2026) 115:87
https://doi.org/10.1007/s10994-026-06997-0

Smoothing the Edges: Smooth Optimization for Sparse Regularization Using Hadamard Overparametrization

Chris Kolb[1,2] · Christian L. Müller[1,2,3,4] · Bernd Bischl[1,2] · David Rügamer[1,2]

Received: 5 November 2024 / Revised: 7 January 2026 / Accepted: 15 January 2026
© The Author(s) 2026

Abstract
In recent years, overparametrization has received considerable attention in various fields, shown to accelerate training and promote simplicity. However, few works study the induced sparse regularization of the original parameters that is caused by combining overparametrization with explicit smooth regularization. Here, we present a unifying framework for smooth optimization of explicitly regularized objectives for (structured) sparsity. These non-smooth and possibly non-convex problems typically rely on solvers tailored to specific models or regularizers and have not been widely adopted in deep learning. In contrast, our method promises fully differentiable and approximation-free optimization for sparse regularizers and is thus compatible with the ubiquitous gradient descent paradigm. The proposed optimization transfer comprises overparameterization of selected parameters and a change of penalties. We prove that the surrogate objective is equivalent in the sense of identical global and local minima, thereby avoiding the introduction of spurious solutions. We comprehensively review sparsity-inducing parametrizations across different fields and combine them in our explicit surrogate regularization framework. We further extend their scope, point out improvements and present novel parametrizations. Numerical experiments further demonstrate the correctness and effectiveness of our approach on several sparse learning problems from high-dimensional regression to sparse neural network training.

Keywords Overparametrization · Sparse regularization · Smooth optimization · Hadamard product parametrization · Gradient descent · Neural networks

1 Introduction and Background

As a result of the recent proliferation of high-dimensional and unstructured data, methods for sparse or low-rank representations have become increasingly important in fields such as machine learning, statistics, and signal processing. Parsimonious models are commonly

Editor: Gang Niu.

Extended author information available on the last page of the article

Published online: 02 April 2026

used to incorporate prior knowledge about the complexity of the underlying phenomenon, to obtain interpretable sparse approximations of non-sparse ground truths (Bach et al., 2012), or to regularize otherwise intractable inverse problems (Benning & Burger, 2018). In deep learning (DL), the reduction in computational burden for large-scale optimization or inference is an important motivation for model sparsification, from both the perspective of efficiency and sustainability (Blalock et al., 2020; Hoefler et al., 2021). Structured or group sparsity naturally generalizes the notion of unstructured sparsity to enable *structural* prior information about parameter group complexity into the optimization problem (Huang et al., 2009; Jenatton et al., 2011). .

1.1 Convex and Non-convex Sparse Regularization

$\boldsymbol{\ell_0}$ and $\boldsymbol{\ell_1}$ regularization In sparse estimation problems of a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ given any objective function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}_0^+$, the classical optimization problem using explicit regularization is

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\beta}) + \lambda \mathcal{R}(\boldsymbol{\beta}), \tag{1}$$

with regularization or penalty function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}_0^+$ to $\mathcal{L}(\boldsymbol{\beta})$, whose strength is controlled by $\lambda \geqslant 0$. A natural choice for the regularizer is $\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$, i.e., the cardinality of the support of $\boldsymbol{\beta}$ counting its non-zero entries. However, this best-subset approach is infeasible due to its non-convex, and non-continuous NP-hard nature (Chen et al., 2017; Natarajan, 1995). To overcome these difficulties, convex relaxations of ℓ_0 regularization have been proposed that enable optimization via, e.g., coordinate descent or projected gradient methods (Schmidt et al., 2007; Tropp, 2006). The tightest convex relaxation of $\|\boldsymbol{\beta}\|_0$ is given by its convex envelope $\|\boldsymbol{\beta}\|_1$, resulting in

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1. \tag{2}$$

This formulation is known as ℓ_1 regularization today. In the context of linear models, it has been introduced as the lasso to the statistics community (Tibshirani, 1996) and as Basis Pursuit Denoising in signal processing (Chen & Donoho, 1994; Chen et al., 2001). For convex $\mathcal{L}$, such as in linear regression, the well-developed machinery of convex optimization can be utilized to solve (2). ℓ_1 regularization has also been shown to have some favorable theoretical properties, such as consistent recovery of the true support of $\boldsymbol{\beta}$ under restricted conditions (Donoho & Elad, 2003; Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006). However, using ℓ_1 regularization to achieve sparsity also comes with a disadvantage: whereas $\|\boldsymbol{\beta}\|_0$ is constant on the support of $\boldsymbol{\beta}$, the ℓ_1 penalty increases linearly in the magnitude of its components. This leads to estimation bias for large parameters (Zhang & Huang, 2008) and inconsistent support recovery (Chartrand, 2007; Xu et al., 2012). To mitigate the challenges posed by ℓ_1 and ℓ_0 regularization, the seminal work of Fan and Li (2001) proposed smoothly clipped absolute deviations (SCAD), one of the earliest examples of non-convex regularizers. Another popular non-convex penalty that enables feature selec-

tion and nearly unbiased estimation is the minimax concave penalty (MCP) introduced by Zhang (2010).

$\ell_{p,q}$regularization In this work, however, we focus on a generalization of the ℓ_1 penalty based on the ℓ_q quasi-norm, $\|\cdot\|_q$, for $0 < q \leqslant 1$. This approach was initially described by Frank and Friedman (1993) and subsequently popularized as the bridge penalty by Fu (1998). Non-convex bridge regularization is defined by a regularization term of the form $\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q$ for $0 < q < 1$. For the case of structured sparsity, ℓ_q regularization can be straightforwardly extended to mixed-norm $\ell_{p,q}$ regularization for $0 < q < p \leqslant 2$, studied, e.g., in Hu et al. (2017). A number of important desirable theoretical results have been established for non-convex ℓ_q and $\ell_{p,q}$ regularization, such as requiring fewer linear measurements for support recovery and permitting sparser solutions compared to convex ℓ_1 and $\ell_{2,1}$ (group-wise) regularization (Chartrand, 2007; Fu, 1998; Fu & Knight, 2000; Xu et al., 2012). Moreover, the regularity conditions required for consistent recovery are weaker than typically required for ℓ_1 (Chartrand & Staneva, 2008; Loh & Wainwright, 2017) or $\ell_{2,1}$ penalties (Hu et al., 2017).

Optimization using ℓ_q regularization, however, poses a non-smooth and non-convex problem for $0 < q < 1$ and is thus difficult to solve efficiently. Ge et al. (2011) show that identification of the global minimum is strongly NP-hard. Still, computing local minima of the non-convex regularization problem usually performs better compared to convex regularization approaches (Lyu et al., 2013; Wen et al., 2018; Xu et al., 2012, 2010). A variety of optimization techniques such as the local quadratic approximation or majorization-minimization algorithms (Hunter & Li, 2005; Lange et al., 2000), and various flavors of coordinate or subgradient descent methods have been discussed in the literature (see Wen et al., 2018, for a survey of optimization with non-convex regularization).

The need for specialized optimization routines for non-smooth and non-convex regularized optimization problems has arguably hindered the widespread use of ℓ_q and $\ell_{p,q}$ regularization, despite their favorable theoretical properties and the limitations of convex regularizers (see, e.g., Freijeiro-González et al., 2022). In contrast, smooth first-order methods have become the go-to optimization tool for many researchers and practitioners, not limited to the field of DL anymore. This can be attributed to their applicability to a vast class of problems using automatic differentiation, their scalability to large data sets, and their surprising effectiveness despite using only cheaply computed gradient information. While in practice, popular DL platforms offer implementations of ℓ_1 regularization, this essentially reduces to applying stochastic gradient descent (SGD) to a non-differentiable problem. Unsurprisingly, this mismatch typically results in oscillating parameter updates, slow convergence, and a failure of parameter iterates to approach zero values **(see Section 9)**.

1.2 Our Contributions

To overcome the obstacles and complexities of using optimization routines tailored for specific non-smooth and potentially non-convex regularized problems, we apply a smooth variational form (SVF) that allows expressing the non-smooth regularizer as the constrained minimum of a smooth surrogate regularizer, where the constraint involves an overparametrization of model parameters (see Fig. 1). In our framework, we construct a general template for exact smooth surrogate optimization of non-smooth and potentially non-convex regularized problems. This optimization transfer is based on finding SVFs of the respective

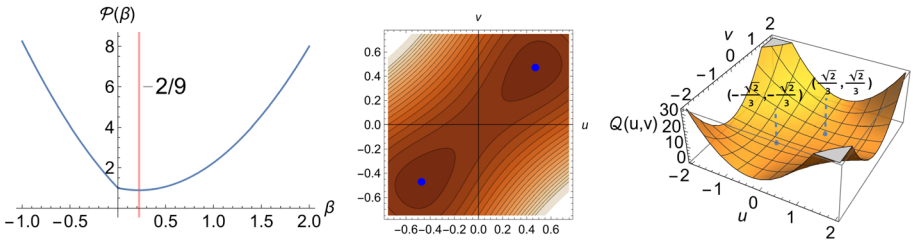


Fig. 1 Illustration of smooth optimization transfer. Left: univariate lasso problem $\mathcal{P}(\beta) = (1 - \frac{3}{2}\beta)^2 + 2|\beta|$ (red line indicates the global minimizer $\hat{\beta}$). Middle: contours of the equivalent smooth surrogate $\mathcal{Q}(u, v) = (1 - \frac{3}{2}uv)^2 + u^2 + v^2$ using a Hadamard product parametrization (10) with $\mathcal{K}(u, v) = uv = \beta$. Both global minimizers (dots) map to $\mathcal{K}(\hat{u}, \hat{v}) = \hat{\beta}$. Right: non-convex surface of higher-dimensional $\mathcal{Q}(u, v)$

regularizers, which entail a smooth parametrization map together with a smooth surrogate regularizer. Combined, an equivalent smooth surrogate objective can be constructed. Specifically, we

- provide a comprehensive review of the loosely connected works on Hadamard parametrizations, relating literature across DL, statistics, and optimization.
- introduce a smooth surrogate optimization framework for non-smooth and non-convex regularization of arbitrary parameters, including a matching local minima property. While previous works often exploit properties particular to their setting, our main results (Thm. 1 and Lemma 1-4) are stated broadly and hold for arbitrary losses, learning models, and regularizers given our assumptions.
- apply our template method to a wide array of (group-)sparse ℓ_q and $\ell_{p,q}$ regularized problems, expanding the collection of sparsity-inducing parametrizations to Hadamard powers and shared parameters.
- present different SVFs with variable amounts of overparametrization for the same induced regularizer, highlighting that it is not overparametrization *per se* inducing sparsity, but rather its effect on the curvature of the loss landscape.
- identify parametrizations with specific neural network structures, generalizing previous findings on linear models to modular components within arbitrary networks. This enables the integration of sparse regularization into the prevalent SGD-based optimization paradigm in DL using sparse "drop-in" replacements.
- evaluate our smooth optimization transfer approach on various sparse learning applications and demonstrate its correctness and practical feasibility.

Outline Section 2 establishes a set of theoretical results that prove the validity of our general framework and provide a construction template for various regularizers. Sections 3 and 4 apply our optimization transfer to construct equivalent smooth surrogates for convex ℓ_1 and structured $\ell_{2,1}$ sparse regularization. Section 5 discusses deeper factorizations involving more than two Hadamard factors, enabling smooth optimization of a restricted class of non-convex ℓ_q and $\ell_{p,q}$ regularized problems. Additionally, mitigation strategies to reduce the computational complexity of overparametrization, such as parameter sharing, are discussed. Section 6 leverages the concept of Hadamard powers to broaden the expressivity of previous parametrizations, thereby lifting the aforementioned restrictions

on the class of induced regularizers. Section 7 discusses specifics regarding the practical optimization of the constructed smooth surrogates. Related work is discussed and compared with our approach in Sect. 8. In Sect. 9, we showcase numerical experiments demonstrating the practical feasibility and competitiveness of our approach on a variety of model classes, ranging from sparse linear regression to (convolutional) neural network architectures. Section 10 concludes by assessing the merits and limitations of our framework and identifying promising directions for future research.

2 Set-Up for Transfer and Theoretical Results

Notation We represent vectors using bold lowercase letters and bold capital letters for matrices. We use $\boldsymbol{\beta} \in \mathbb{R}^d$ to denote the parameter vector which is subject to regularization, and $\boldsymbol{\psi} \in \mathbb{R}^{d_\psi}$ for the remaining parameters, so that all model parameters are collected in $(\boldsymbol{\psi}, \boldsymbol{\beta})$. We make this notational distinction to emphasize that our approach can be applied to arbitrary subsets of parameters of an optimization problem, irrespective of the presence of other parameters or the structure of the main objective $\mathcal{L}$. Thus, sparse regularization using our optimization transfer framework can be applied to, e.g., specific layers of a neural network. Further, let $\|\boldsymbol{\beta}\|_q \triangleq (\sum_{j=1}^d |\beta_j|^q)^{1/q}$ denote the ℓ_q norm $\forall \boldsymbol{\beta} \in \mathbb{R}^d, q \in (0, \infty)$. Note that for $0 < q < 1$, only a quasi-norm is defined as the subadditivity does not hold. For $q = 0$, the ℓ_0 "norm" penalty $\|\boldsymbol{\beta}\|_0$ counts the number of non-zero elements in $\boldsymbol{\beta}$. Given a partition $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$ of $[d] \triangleq \{1, \ldots, d\}$, the $\ell_{p,q}$ group (quasi-)norm is defined as $\|\boldsymbol{\beta}\|_{p,q} \triangleq (\sum_{j=1}^L (\sum_{i \in \mathcal{G}_j} |\beta_i|^p)^{q/p})^{1/q} = (\sum_{j=1}^L \|\boldsymbol{\beta}_j\|_p^q)^{1/q} \, \forall \, \boldsymbol{\beta} \in \mathbb{R}^d$, $p, q > 0$, where $\boldsymbol{\beta}_j$ contains the components corresponding to $\mathcal{G}_j$. The regularization term for an $\ell_{p,q}$ penalty is given by the q-th power of the $\ell_{p,q}$ mixed-norm, $\|\boldsymbol{\beta}\|_{p,q}^q = \sum_{j=1}^L \|\boldsymbol{\beta}_j\|_p^q$. Further, we use various notations to define Hadamard product-like operations, introduced in the following. Let $\odot : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ denote the classical Hadamard product, defined as $(\boldsymbol{u}, \boldsymbol{v}) \mapsto (u_1 v_1, \ldots, u_d v_d)^\top$, and $\bigodot_{l=1}^k \boldsymbol{u}_l$ the Hadamard product of k vectors, for which we also use the shorthand notation $\boldsymbol{u}_l^{\odot k}$. For parameter vectors with more than one index, e.g., $\boldsymbol{u}_{jl}^{\odot k}$, the Hadamard product is always taken over the second index. The self-Hadamard product $\boldsymbol{u} \odot \boldsymbol{u}$ is simply written as $\boldsymbol{u}^2$. A generalization of the self-Hadamard product to non-integer exponents $k > 0$, i.e., element-wise raising the entries of $\boldsymbol{u}$ to the k-th power, is denoted as $\boldsymbol{u}^{\circ k}$. Given a partition $\mathcal{G}$ of $[d]$ into $L \leqslant d$ subsets, we define the group Hadamard product $\odot_{\mathcal{G}}$ of two vectors $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{\nu} \in \mathbb{R}^L$ as $\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} \triangleq (\boldsymbol{u}_j \nu_j)_{j \in \mathcal{G}}$, or more explicitly as

$$\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} \triangleq \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_L \end{pmatrix} \odot \begin{pmatrix} \nu_1 \mathbf{1}_{|\mathcal{G}_1|} \\ \vdots \\ \nu_L \mathbf{1}_{|\mathcal{G}_L|} \end{pmatrix},$$

where $\mathbf{1}_{|\mathcal{G}_j|}$ denotes the 1-vector of size $|\mathcal{G}_j|$. To make the distinction between vectors of size d and L more clear where necessary, we denote vectors in $\mathbb{R}^d$ as $\boldsymbol{v}$, and alternatively, use $\boldsymbol{\nu}$ for vectors in $\mathbb{R}^L$. In case $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L)^\top$ is constant within groups $\mathcal{G}_j$, both are related as $\boldsymbol{v}_j = \nu_j \mathbf{1}_{|\mathcal{G}_j|}$ for $j = 1, \ldots, L$, and $\boldsymbol{u} \odot \boldsymbol{v}$ equals the group Hadamard product $\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$. Further, $\mathcal{B}(\boldsymbol{\beta}, \varepsilon) \subseteq \mathbb{R}^d$ is used to denote an open ball with radius ε centered at

$\boldsymbol{\beta} \in \mathbb{R}^d$, for a Euclidean space endowed with the standard topology induced by the Euclidean metric. The non-negative reals are abbreviated as $\mathbb{R}_0^+$. Given a differentiable function $f : \mathbb{R}^m \to \mathbb{R}$, $\boldsymbol{a} \mapsto f(\boldsymbol{a})$, the gradient $\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) \in \mathbb{R}^m$ of f at $\boldsymbol{a}$ contains partial derivatives $\partial f(\boldsymbol{a})/\partial a_j$ for $j \in [m]$. The Hessian $\mathcal{H}_f(\boldsymbol{a})$ of f at $\boldsymbol{a}$ is the $m \times m$ matrix containing second partial derivatives $(\mathcal{H}_f(\boldsymbol{a}))_{ij} \triangleq \partial^2 f/\partial a_i \partial a_j$. For vector-valued differentiable maps $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n, \boldsymbol{a} \mapsto \boldsymbol{f}(\boldsymbol{a})$, the Jacobian $\mathcal{J}_{\boldsymbol{f}}(\boldsymbol{a})$ of $\boldsymbol{f}$ at $\boldsymbol{a}$ is the $n \times m$ matrix containing partial derivatives $(\mathcal{J}_{\boldsymbol{f}}(\boldsymbol{a}))_{ij} \triangleq \partial \boldsymbol{f}_i/\partial a_j$. If we say a function is **smooth**, we require it merely to be $\mathcal{C}^r$-smooth, $r \geqslant 1$, i.e., at least continuously differentiable. Local solutions to an optimization problem over $(\boldsymbol{\psi}, \boldsymbol{\beta})$ are denoted by $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. Finally, the complete proofs are deferred to the appendix.

Set-up Before discussing the applications of our proposed framework to specific sparse regularizers, we first provide a number of general results on the equivalence of (regularized) optimization problems under reparametrization and a change of penalties, which will be applied throughout the paper. Let

$$\mathcal{P} : \mathbb{R}^{d_\psi} \times \mathbb{R}^d \to \mathbb{R}_0^+,\ (\boldsymbol{\psi}, \boldsymbol{\beta}) \mapsto \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \cdot \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}), \tag{3}$$

denote the regularized objective function in its base parametrization $(\boldsymbol{\psi}, \boldsymbol{\beta}) \in \mathbb{R}^{d_\psi} \times \mathbb{R}^d$, where $\boldsymbol{\beta}$ is an arbitrary subset of all model parameters, and $\boldsymbol{\psi}$ comprises the complementary components. In a typical empirical risk minimization setting, $\mathcal{L}$ can be written more explicitly as $\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \sum_{i=1}^n \mathscr{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i|\boldsymbol{\psi}, \boldsymbol{\beta}))$, with independently sampled data $\mathcal{D} \triangleq \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote generic feature and label spaces, $\mathscr{L} : \mathcal{Y} \times \mathbb{R}^{d_y} \to \mathbb{R}_0^+$ is an arbitrary loss contribution, and the (parametric) model $f : \mathcal{X} \to \mathbb{R}^{d_y}$ is parametrized by $(\boldsymbol{\psi}, \boldsymbol{\beta})$.

The non-smooth and potentially non-convex regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ is defined as $\mathcal{R}_{\boldsymbol{\beta}} : \mathbb{R}^d \to \mathbb{R}_0^+$, $\boldsymbol{\beta} \mapsto \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, with $\lambda \geqslant 0$ controlling the amount of regularization. In this work, we consider classical norm-based sparsity-inducing regularizers.

Optimization transfer To transfer the optimization of $\mathcal{P}$ to a surrogate $\mathcal{Q}$, first consider a continuous and surjective parametrization of $\boldsymbol{\beta}$ defined by $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d, \boldsymbol{\xi} \mapsto \mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}$. Moreover, we define a surrogate regularization function $\mathcal{R}_{\boldsymbol{\xi}} : \mathbb{R}^{d_\xi} \to \mathbb{R}_0^+, \boldsymbol{\xi} \mapsto \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$. Together, $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$ define our proposed two-step optimization transfer approach to construct an equivalent surrogate $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ from the original objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$:

Definition 1 *(Construction of surrogate $\mathcal{Q}$)* Let the objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ as in (3), $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ the non-smooth regularizer, $\mathcal{K}(\boldsymbol{\xi})$ a parametrization of $\boldsymbol{\beta}$, and $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ a surrogate regularizer for $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$. The variational surrogate $\mathcal{Q}$ can then be constructed from the tuple $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$ as follows. First, **i)** parametrize $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}$ to get a "lifted" $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$, and **ii)** substitute $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ for $\mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\boldsymbol{\xi}))$:

$$\mathcal{Q} : \mathbb{R}^{d_\psi} \times \mathbb{R}^{d_\xi} \to \mathbb{R}_0^+,\ (\boldsymbol{\psi}, \boldsymbol{\xi}) \mapsto \mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}). \tag{4}$$

Further, if $\mathcal{L}$, $\mathcal{K}$, and $\mathcal{R}_{\boldsymbol{\xi}}$ are $\mathcal{C}^1$ functions, we call $\mathcal{Q}$ a smooth surrogate for $\mathcal{P}$.

The next definition explicitly states our notion of equivalence between $\mathcal{P}$ and $\mathcal{Q}$:

Definition 2 *(Equivalence of optimization problems)* We say the two optimization problems

$$\underset{\boldsymbol{\psi},\boldsymbol{\beta}}{\text{minimize}}\ \mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta}) \quad \text{and} \quad \underset{\boldsymbol{\psi},\boldsymbol{\xi}}{\text{minimize}}\ \mathcal{Q}(\boldsymbol{\psi},\boldsymbol{\xi}),$$

are equivalent if the following conditions hold:

a) $\inf_{\boldsymbol{\psi},\boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta}) = \inf_{\boldsymbol{\psi},\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{\psi},\boldsymbol{\xi})$, i.e., their globally optimal values coincide.

b) If $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$, then there is a local minimizer $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$ of $\mathcal{Q}(\boldsymbol{\psi},\boldsymbol{\xi})$ with $\hat{\boldsymbol{\xi}} \in \mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$ and $\mathcal{Q}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$.

c) If $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi},\boldsymbol{\xi})$, then $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\beta}} = \mathcal{K}(\hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$ and $\mathcal{Q}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$.

Equivalence of local minima is particularly important for non-convex regularization, encompassed by the second and third conditions, as finding global minima in non-convex optimization is challenging and, for the most part, intractable. Moreover, in non-convex regularization, local minima have been observed to generalize similarly or even better than global minima on test data (Chartrand & Yin, 2008; Chartrand, 2007; Olsson et al., 2017).

This so-called matching of local minima (Levin, 2020) ensures that the transfer from the original objective $\mathcal{P}$ to a surrogate objective $\mathcal{Q}$ preserves all the properties of the local minima structure of $\mathcal{P}$. By focusing only on global minima, important information about the structure of the problem is neglected. Importantly, we do not introduce spurious local minima in $\mathcal{Q}$, which would artificially increase the difficulty of the optimization problem. Through the matching property, we can further use the surjection $\mathcal{K}(\boldsymbol{\xi})$ to reconstruct all (local) minimizers of $\mathcal{P}$ from local minimizers of $\mathcal{Q}$ as $\mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$. To guarantee this matching of local minima property for $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$ under the parametrization $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{\xi}))$, local openness of the parametrization mapping $\mathcal{K}(\boldsymbol{\xi})$ at all local minimizers $\hat{\boldsymbol{\xi}}$ of $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{\xi}))$ is a crucial property (Nouiehed & Razaviyayn, 2022). Levin et al. (2024) show that it is both a necessary and sufficient condition for the preservation of local minima under the parametrization $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}$,[1]

Definition 3 *(Local openness)* A mapping $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d$, $\boldsymbol{\xi} \mapsto \mathcal{K}(\boldsymbol{\xi})$ is locally open at $\boldsymbol{\xi}$ if for every $\varepsilon > 0$ we can find $\delta > 0$ such that $\mathcal{B}(\mathcal{K}(\boldsymbol{\xi}),\delta) \subseteq \mathcal{K}(\mathcal{B}(\boldsymbol{\xi},\varepsilon))$. Further, the map $\mathcal{K}$ is called globally open if it is locally open at all $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$.

General theoretical results Using the function characterizations encapsulated in Definition 1, we can now prove the following results. Note that these hold for any potentially unregularized objective $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$ under reparametrization:

Lemma 1 *If $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$, and $\mathcal{K}(\boldsymbol{\xi})$ is a continuous surjection, then all $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$ such that $\hat{\boldsymbol{\xi}} \in \mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$ are local minimizers of $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{\xi}))$ with $\mathcal{P}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}}) = \mathcal{P}(\hat{\boldsymbol{\psi}},\mathcal{K}(\hat{\boldsymbol{\xi}}))$.*

Lemma 2 *If $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{\xi}))$, and the continuous surjection $\mathcal{K}(\boldsymbol{\xi})$ is locally open at $\hat{\boldsymbol{\xi}}$, then $(\hat{\boldsymbol{\psi}},\mathcal{K}(\hat{\boldsymbol{\xi}})) = (\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$ with $\mathcal{P}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}}) = \mathcal{P}(\hat{\boldsymbol{\psi}},\mathcal{K}(\hat{\boldsymbol{\xi}}))$.*

[1] Local openness is closely related to, but distinct from the notion of continuity, which is defined as $\forall\,\varepsilon > 0\,\exists\,\delta > 0 :\ \mathcal{K}(\mathcal{B}(\boldsymbol{\xi})\delta) \subseteq \mathcal{B}(\mathcal{K}(\boldsymbol{\xi}),\varepsilon)$ using the same notation.

Their proofs are given in Appendices A.1 and A.2. Together, both results show that the set of local minima of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ and $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$ are equal if $\mathcal{K}(\boldsymbol{\xi})$ is locally open at all local minimizers of $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$, and the local minimizers are related via $(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. Still, for non-smooth regularizers, smoothly parametrizing $\boldsymbol{\beta}$ will not result in a smooth optimization problem. Nevertheless, the results using local openness of the map $\mathcal{K}(\boldsymbol{\xi})$ can be applied to guarantee matching local minima under smooth and surjective parametrizations in general problems, e.g., for parametrizations used in the implicit regularization literature. In our optimization transfer approach, however, we further replace the parametrized regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\boldsymbol{\xi}))$ by $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ to obtain the surrogate objective $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$. The surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$ and the non-smooth regularizer $\mathcal{R}_{\boldsymbol{\beta}}$ are related as follows:

Definition 4 *(Smooth variational form)* A (smooth) variational form is an expression of a function $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ as the minimum of a (smooth) surrogate $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ over a feasible set given by the fiber $\mathcal{K}^{-1}(\boldsymbol{\beta})$ of a surjective parametrization $\mathcal{K}(\boldsymbol{\xi})$ at $\boldsymbol{\beta}$, i.e.,

$$\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d. \tag{5}$$

By definition, $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ majorizes $\mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\boldsymbol{\xi}))$, i.e., $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \geqslant \mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\boldsymbol{\xi})) \forall \boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$. Importantly, finding the appropriate SVF is non-trivial and will be derived for each considered regularizer $\mathcal{R}_{\boldsymbol{\beta}}$ in the respective section. To give a canonical example, the ℓ_1 penalty $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\|\boldsymbol{\beta}\|_1$ can be expressed as the minimum of $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$, where $\boldsymbol{\xi} = (\boldsymbol{u}, \boldsymbol{v})^\top \in \mathbb{R}^{2d}$, subject to the parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^d$ (cf. Sect. 3.1). It should be noted that the SVFs for an arbitrary regularizer $\mathcal{R}_{\boldsymbol{\beta}}$ are not necessarily unique, as evidenced by several distinct SVFs, defined by pairs $(\mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$, yielding the same induced regularizer $\mathcal{R}_{\boldsymbol{\beta}}$ in Table 1. Further, determining conditions for the existence of an SVF, defined by a pair $(\mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$, for *arbitrary* regularizers $\mathcal{R}_{\boldsymbol{\beta}}$, is challenging, primarily as the induced regularizer is determined jointly by the parametrization $\mathcal{K}$ and the surrogate regularizer $\mathcal{R}_{\boldsymbol{\xi}}$. Hence, we leave this relevant question to future research.
To preserve local minima in our approach with replaced surrogate regularization $\mathcal{R}_{\boldsymbol{\xi}}$, we further require stability of the solutions $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ to the SVF with respect to the regularized parameter $\boldsymbol{\beta}$, i.e., continuous dependence of the minimizers $\hat{\boldsymbol{\xi}} \in \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ on $\boldsymbol{\beta}$, formalized through lower hemicontinuity of the set-valued solution mapping:

Definition 5 *(Lower hemicontinuity)* A set-valued map $\hat{\boldsymbol{\xi}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^{d_\xi},\ \boldsymbol{\beta} \mapsto \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$, is said to be lower hemicontinuous (l.h.c.) at $\boldsymbol{\beta} \in \mathbb{R}^d$ if

$$\forall \boldsymbol{\xi} \in \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})\, \forall \varepsilon > 0\, \exists \delta > 0 : \forall \tilde{\boldsymbol{\beta}} \in \mathcal{B}(\boldsymbol{\beta}, \delta)\, \exists \tilde{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\tilde{\boldsymbol{\beta}}) \cap \mathcal{B}(\boldsymbol{\xi}, \varepsilon).$$

Hence, for every point $\boldsymbol{\xi} \in \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\boldsymbol{\beta}, \delta)$, the set $\hat{\boldsymbol{\xi}}(\tilde{\boldsymbol{\beta}})$ contains at least one point in the ball $\mathcal{B}(\boldsymbol{\xi}, \varepsilon)$.
Instead of requiring local openness, as previously for parametrizations without a change of regularizers, we use the lower hemicontinuity of the solution map $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$, a property that is easily obtained as a by-product in the construction of our smooth variational forms. For details on set-valued analysis, we refer to Aubin and Frankowska (2009). In the following,

Table 1 Overview of induced regularizers $\mathcal{R}_{\boldsymbol{\beta}}$ obtained by parametrizing $\boldsymbol{\beta}$ through $\mathcal{K}(\boldsymbol{\xi})$ and adding a smooth surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$.

Abbreviation	$\mathcal{K}(\cdot) = \boldsymbol{\beta}$	$\mathcal{R}_{\boldsymbol{\xi}}$	$\mathcal{R}_{\boldsymbol{\beta}}$	Type	References
HPP	$\boldsymbol{u} \odot \boldsymbol{v}$	$\lVert\boldsymbol{u}\rVert_2^2 + \lVert\boldsymbol{v}\rVert_2^2$	$2\lVert\boldsymbol{\beta}\rVert_1$	ℓ_1	Grandvalet (1998); Hoff (2017)
HDP	$\boldsymbol{\gamma}^2 - \boldsymbol{\delta}^2$	$\lVert\boldsymbol{\gamma}\rVert_2^2 + \lVert\boldsymbol{\delta}\rVert_2^2$	$\lVert\boldsymbol{\beta}\rVert_1$	ℓ_1	Vaskevicius et al. (2019)
GHPP	$\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$	$\sum_{j=1}^{L}(\lVert\boldsymbol{u}_j\rVert_2^2 + \nu_j^2)$	$2\lVert\boldsymbol{\beta}\rVert_{2,1}$	$\ell_{2,1}$	Tibshirani (2021)
Adj. GHPP	$\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$	$\sum_{j=1}^{L}(\lVert\boldsymbol{u}_j\rVert_2^2 + \lvert\mathcal{G}_j\rvert\nu_j^2)$	$2\sum_{j=1}^{L}\sqrt{\lvert\mathcal{G}_j\rvert}\lVert\boldsymbol{\beta}_j\rVert_2$	$\ell_{2,1}$	–
$k \in \mathbb{N},\ k_1 \in \mathbb{N},\ k_2 \triangleq k - k_1 \in \mathbb{N}$:					
HPP_k	$\bigodot_{l=1}^{k} \boldsymbol{u}_l$	$\sum_{l=1}^{k} \lVert\boldsymbol{u}_l\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	Hoff (2017)
GHPP_k	$\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}$	$\sum_{j=1}^{L} \lVert\boldsymbol{u}_j\rVert_2^2 + \sum_{r=1}^{k-1} \nu_{jr}^2$	$k\lVert\boldsymbol{\beta}\rVert_{2,2/k}^{2/k}$	$\ell_{2,2/k}$	Tibshirani (2021)
$\mathrm{GHPP}_{k_1,k}$	$\boldsymbol{\mu}_t^{\odot k_1} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot k_2}$	$\sum_{t=1}^{k_1} \lVert\boldsymbol{\mu}_t\rVert_2^2 + \sum_{r=1}^{k_2} \lVert\boldsymbol{\nu}_r\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2/k_1,2/k}^{2/k}$	$\ell_{2/k_1,2/k}$	Dai et al. (2021)
HDP_k	$\boldsymbol{u}_l^{\odot k} - \boldsymbol{v}_l^{\odot k}$	$\sum_{l=1}^{k} \lVert\boldsymbol{u}_l\rVert_2^2 + \lVert\boldsymbol{v}_l\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	Chou et al. (2023)
HPP_k^{shared}	$\boldsymbol{u} \odot \boldsymbol{v}^{k-1}$	$\lVert\boldsymbol{u}\rVert_2^2 + (k-1)\lVert\boldsymbol{v}\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	–
HDP_k^{shared}	$\boldsymbol{u}^k - \boldsymbol{v}^k$	$\lVert\boldsymbol{u}\rVert_2^2 + \lVert\boldsymbol{v}\rVert_2^2$	$\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	Chou et al. (2023)
$k \in \mathbb{R}_{>2},\ k_1 \in \mathbb{R}_{>1},\ k_2 \triangleq k - k_1 \in \mathbb{R}_{>1},\ (k \in \mathbb{R}_{>1}$ for Powerprop.) :					
HPowP_k	$\boldsymbol{u} \odot \lvert\boldsymbol{v}\rvert^{\circ(k-1)}$	$\lVert\boldsymbol{u}\rVert_2^2 + (k-1)\lVert\boldsymbol{v}\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	–
Powerprop	$\boldsymbol{v} \odot \lvert\boldsymbol{v}\rvert^{\circ(k-1)}$	$\lVert\boldsymbol{v}\rVert_2^2$	$\lVert\boldsymbol{\beta}\rVert_{2/k}^{2/k}$	$\ell_{2/k}$	Schwarz et al. (2021)
GHPowP_k	$\boldsymbol{u} \odot_{\mathcal{G}} \lvert\boldsymbol{\nu}\rvert^{\circ(k-1)}$	$\lVert\boldsymbol{u}\rVert_2^2 + (k-1)\lVert\boldsymbol{\nu}\rVert_2^2$	$k\lVert\boldsymbol{\beta}\rVert_{2,2/k}^{2/k}$	$\ell_{2,2/k}$	–
$\mathrm{GHPowP}_{k_1,k}$	$(\boldsymbol{\mu} \odot \lvert\boldsymbol{\mu}\rvert^{\circ(k_1-1)}) \odot_{\mathcal{G}} \lvert\boldsymbol{\nu}\rvert^{\circ k_2}$	$k_1\lVert\boldsymbol{\mu}\rVert_2^2 + k_2\lVert\boldsymbol{\nu}\rVert_2^2$	$\lVert\boldsymbol{\beta}\rVert_{2/k_1,2/k}^{2/k}$	$\ell_{2/k_1,2/k}$	–

The letter "H" stands for "Hadamard", the letter "G" for "Group", "PP" for "Product Parametrization", "DP" for "Difference Parametrization", and "PowP" abbreviates "Power Parametrization". Novel results are underlined.

we state a minimal but sufficient set of assumptions to establish matching of local minima, aiming to remain as general as possible.

Assumption 1 *(Minimal assumptions for optimization transfer)* Let $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ be the original objective as in (3) and $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ the surrogate as in (4). Let $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}$ be a continuous surjection, $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ a surrogate regularizer such that $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \, \forall \boldsymbol{\beta}$, and let the set-valued solution map $\boldsymbol{\beta} \mapsto \arg\min_{\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ be l.h.c.

Lemma 3 *If $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then all $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ such that $\hat{\boldsymbol{\xi}} \in \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ are local minimizers of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ with $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ under Assumption 1.*

Lemma 4 *If $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$, then $(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ under Assumption 1.*

In the proof of Lemma 4, it is also established that for all local minimizers $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$, $\hat{\boldsymbol{\xi}}$ must also minimize the SVF over the fiber $\mathcal{K}^{-1}(\mathcal{K}(\hat{\boldsymbol{\xi}}))$. Figure 2a provides some intuition behind the preceding results, assuming no additional parameters $\boldsymbol{\psi}$. It illustrates the relationship between a local minimizer $\hat{\boldsymbol{\xi}}$ of $\mathcal{Q}(\boldsymbol{\xi})$ and the corresponding local minimizer $\mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$ of $\mathcal{P}(\boldsymbol{\beta})$. Note that at points $\tilde{\boldsymbol{\xi}}$ around $\hat{\boldsymbol{\xi}}$ that are in the image of $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ (dashed green), we have equality of $\mathcal{Q}(\tilde{\boldsymbol{\xi}})$ and $\mathcal{P}(\tilde{\boldsymbol{\beta}})$. By Lemma 1, if $\hat{\boldsymbol{\beta}}$ is a local minimizer of $\mathcal{P}(\boldsymbol{\beta})$, then any $\hat{\boldsymbol{\xi}} \in \mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$ (red curve) is a local minimizer of the non-smooth overparametrized $\mathcal{P}(\mathcal{K}(\boldsymbol{\xi}))$. But only those $\hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}}) \in \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \subset \mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$ (red dot at vertex) are also local minimizers of $\mathcal{Q}(\boldsymbol{\xi})$ due to the majorization property $\mathcal{P}(\mathcal{K}(\boldsymbol{\xi})) \leqslant \mathcal{Q}(\boldsymbol{\xi}) \forall \boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$, combined with $\mathcal{Q}(\hat{\boldsymbol{\xi}}) = \mathcal{P}(\mathcal{K}(\hat{\boldsymbol{\xi}}))$. Conversely, if $\hat{\boldsymbol{\xi}}$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\xi})$, then by continuity of the solution map $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ at $\mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$, if there existed $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\hat{\boldsymbol{\beta}}, \delta)$ such that $\mathcal{P}(\tilde{\boldsymbol{\beta}}) < \mathcal{P}(\hat{\boldsymbol{\beta}})$, this would imply existence of $\tilde{\boldsymbol{\xi}} \in \mathcal{B}(\hat{\boldsymbol{\xi}}, \varepsilon)$ with $\mathcal{Q}(\tilde{\boldsymbol{\xi}}) < \mathcal{Q}(\hat{\boldsymbol{\xi}})$, contradicting that $\hat{\boldsymbol{\xi}}$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\xi})$. Figure 2b shows a specific choice of functions for $\mathcal{K}$ and $\mathcal{R}_{\boldsymbol{\xi}}$.

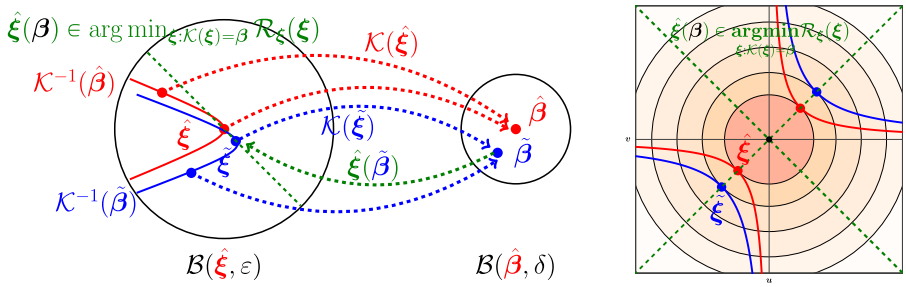
Assumption 1 requires only a continuous surjection $\mathcal{K}$ and, in principle, poses no restrictions on the functions $\mathcal{L}$ and $\mathcal{R}_{\boldsymbol{\xi}}$, as long as the variational expression $\mathcal{R}_{\boldsymbol{\beta}}$ holds for any $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is l.h.c. However, without further smoothness assumptions, the surrogate $\mathcal{Q}$ might have the same differentiability issues as the base problem. So from now on, we only consider cases where $\mathcal{L}, \mathcal{K}$, and $\mathcal{R}_{\boldsymbol{\xi}}$ are smooth, so that $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ is non-smooth, but $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ is smooth, and we can tackle the problem with (S)GD. The previous results let us now state our main result:

Theorem 1 *(Smooth optimization transfer for sparse regularization) Let the non-smooth objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ and its smooth surrogate $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ be defined as in Eqs. (3) and (4). Under Assumption 1 the optimization problems*

$$\underset{\boldsymbol{\psi}, \boldsymbol{\beta}}{minimize} \; \mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) \triangleq \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \,, \tag{6}$$

$$\underset{\boldsymbol{\psi}, \boldsymbol{\xi}}{minimize} \; \mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi}) \triangleq \mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \,, \tag{7}$$

are equivalent by Definition 2.



(a) Schematic illustration of smooth optimization transfer.

(b) Contours of HPP in (10) and surrogate ℓ_2 penalty.

Fig. 2 Relationship between local minimizer $\hat{\boldsymbol{\xi}}$ of $\mathcal{Q}$, the induced minimizer $\mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$ of $\mathcal{P}$, and the cont. solution mapping $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ of the SVF. Left: solid curves show two fibers $\mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$ (red) and $\mathcal{K}^{-1}(\tilde{\boldsymbol{\beta}})$ (blue). The solution map $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ (dashed green) maps to minimizers of the SVF for varying $\boldsymbol{\beta}$. Right: concrete example showing scalar parametrization $\beta_j = \mathcal{K}(u_j, v_j)$ with surrogate ℓ_2 penalty

Proof For the first point of Definition 2, we show that the infima of both $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ and $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ coincide. Because $\mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$ is constant on the fiber of $\mathcal{K}$ at $\boldsymbol{\beta} = \mathcal{K}(\boldsymbol{\xi})$, we can pull in the infimum and re-state it in terms of $\boldsymbol{\beta}$:

$$\inf_{\boldsymbol{\psi},\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \inf_{\boldsymbol{\psi},\boldsymbol{\xi}} \left\{\mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi})) + \lambda \cdot \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})\right\} = \inf_{\boldsymbol{\psi},\boldsymbol{\beta}} \left\{\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \inf_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \{\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})\}\right\}$$

By Assumption 1, we have $\inf_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, and thus

$$\inf_{\boldsymbol{\psi},\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \inf_{\boldsymbol{\psi},\boldsymbol{\beta}} \{\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})\} = \inf_{\boldsymbol{\psi},\boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}).$$

This shows the first point. For the second and third points of Definition 2, we can apply Lemma 3 together with Lemma 4 under Assumption 1 to obtain the required matching of local minima with corresponding minimizers. □

From Theorem 1 it follows that there is a surjective mapping from the set of local minimizers of $\mathcal{Q}$ to the set of local minimizers of $\mathcal{P}$, obtained by restricting the domain of the parametrization $\mathcal{K}$ to the set of local minimizers of $\mathcal{Q}$. Table 1 shows an (incomplete) summary of the different parametrizations $\mathcal{K}(\boldsymbol{\xi})$ of $\boldsymbol{\beta}$ that can be represented in our framework, together with the sparse regularization terms $\mathcal{R}_{\boldsymbol{\beta}}$ that are induced by applying the smooth regularizers $\mathcal{R}_{\boldsymbol{\xi}}$ to the surrogate parameters $\boldsymbol{\xi}$.

It should be noted that although this work only considers convex and non-convex regularizers based on (quasi-)norm- and mixed-norm penalties $\mathcal{R}_{\boldsymbol{\beta}}$, Theorem 1 provides a more general result that holds for any smooth optimization transfer fulfilling the assumptions. To concretize the setting for the remainder of our work, given a partition $\mathcal{G}$ of parameter indices $[d]$ into $L \leqslant d$ groups, we consider sparsity-inducing regularizers of the form $\{\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \in \{\mathcal{R} : \mathbb{R}^d \to \mathbb{R}_0^+, \boldsymbol{\beta} \mapsto \sum_{j=1}^{L} \omega_j \|\boldsymbol{\beta}_j\|_p^q \mid 0 < q \leqslant p \leqslant 2,\ \omega_j > 0\, \forall\, j\}$. Note that setting $p = q$, $L = d$ and $\omega_j = 1\, \forall\, j$ reduces the expression to the familiar ℓ_q regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q$. Merely setting $\omega_j = 1$ results in the $\ell_{p,q}$ regularizer, whereas, e.g., $\omega_j = \sqrt{|\mathcal{G}_j|} \in \mathbb{N}, p = 2, q = 1$ yields the $\ell_{2,1}$

group lasso (Yuan & Lin, 2006). Similarly, we take the smooth surrogate regularizers $\mathcal{R}_{\boldsymbol{\xi}}$ to be of the form $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \in \left\{\mathcal{R} : \mathbb{R}^{d_{\boldsymbol{\xi}}} \to \mathbb{R}_0^+,\ \boldsymbol{\xi} \mapsto \sum_{j=1}^{d_\xi} \tilde{\omega}_j\, \xi_j^2 \;\middle|\; \tilde{\omega}_j > 0\ \forall\, j\right\}$.
As the parametrizations considered in this work are based on Hadamard products and variations thereof, the following smoothness and separability assumptions on $\mathcal{K}$, as well as a specific monomial-like structure, describe the multiplicative parametrizations in the overview of parametrizations shown in Table 1:

Assumption 2 *(Power-Product Parametrizations $\mathcal{K}$)* The parametrization map $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d, \boldsymbol{\xi} \mapsto \boldsymbol{\beta}$, is a $\mathcal{C}^r$-smooth surjection, $r \geqslant 1$, with the following properties:

a) $\mathcal{K}$ is block-separable, i.e., for a partition of $[d]$ into $L \leqslant d$ groups of size $|\mathcal{G}_j|, j \in [L]$, the corresponding $\boldsymbol{\beta}_j$ are parametrized by disjoint subsets $\boldsymbol{\xi}_j \in \mathbb{R}^{d_{\boldsymbol{\xi}_j}}$ of the entries of $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$, where $d_{\boldsymbol{\xi}} = d_{\boldsymbol{\xi}_1} + \ldots + d_{\boldsymbol{\xi}_L}$. That is, $\mathcal{K}$ is the Cartesian function product $\mathcal{K}(\boldsymbol{\xi}) = (\mathcal{K}_1(\boldsymbol{\xi}_1), \ldots, \mathcal{K}_L(\boldsymbol{\xi}_L))$ of block-wise parametrizations $\mathcal{K}_j(\boldsymbol{\xi}_j) = \boldsymbol{\beta}_j$.
b) Each $\boldsymbol{\xi}_j$ can further be grouped into k factors $\boldsymbol{\xi}_{jl} \in \mathbb{R}^{d_{jl}}, l \in [k]$, so that $\sum_{l=1}^k d_{jl} = d_{\boldsymbol{\xi}_j}$ and $d_{jl} \in \{1, |\mathcal{G}_j|\}$, i.e., each factor is either a scalar or a vector of the same dimension as $\boldsymbol{\beta}_j$.
c) $\mathcal{K}_j(\boldsymbol{\xi}_{j1}, \ldots, \boldsymbol{\xi}_{jk})$ has a power-product structure such that each coordinate $\mathcal{K}_{ji}(\boldsymbol{\xi}_j) = \beta_{ji}$, $i \in \mathcal{G}_j$, can be written as $\mathcal{K}_{ji}(\boldsymbol{\xi}_j) = \prod_{l=1}^k \operatorname{sign}(\xi_{jl}^{(i)}) \cdot |\xi_{jl}^{(i)}|^{\alpha_l}$, where $\xi_{jl}^{(i)} \in \boldsymbol{\xi}_{jl}$, and $\alpha_l \geqslant 1$ are the (entry-wise) exponents for factor $l \in [k]$. Some parametrizations omit the signs or absolute values, e.g., pure monomials.

3 Smooth ℓ_1 Regularization Using Hadamard Products

In this section, we introduce two smooth surrogate approaches for sparsity-inducing ℓ_1 regularization and provide some intuition on the underlying geometry.

3.1 Hadamard Product Parametrization

We first present a canonical example of our optimization transfer framework based on the so-called Hadamard product parametrization (Hoff, 2017). This approach enables smooth optimization of ℓ_1 regularized objectives by applying an overparametrization $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$ and imposing ℓ_2 regularization on the surrogate parameters. As the prototype case of our framework, this connection between ℓ_1 and ℓ_2 regularization under reparametrization will be re-derived in the following for illustrative purposes. Assume a non-smooth ℓ_1 regularized objective $\mathcal{P}$ with $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\,\|\boldsymbol{\beta}\|_1$ and consider the following overparametrized smooth surrogate $\mathcal{Q}$:

$$\mathcal{P} : \mathbb{R}^{d_\psi} \times \mathbb{R}^d \to \mathbb{R}_0^+,\ (\boldsymbol{\psi}, \boldsymbol{\beta}) \mapsto \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + 2\lambda \|\boldsymbol{\beta}\|_1 = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + 2\lambda \textstyle\sum_{j=1}^d |\beta_j|\,, \tag{8}$$

$$\mathcal{Q} : \mathbb{R}^{d_\psi} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+,\ (\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}) \mapsto \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot \boldsymbol{v}) + \lambda \textstyle\sum_{j=1}^d \left(u_j^2 + v_j^2\right). \tag{9}$$

In (9), the HPP map is defined as

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}, \tag{10}$$

while the surrogate penalty is the plain ℓ_2 regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$ with $\boldsymbol{\xi} = (\boldsymbol{u}, \boldsymbol{v})^\top$. Our goal is to show that the minimization of (8) and (9) is equivalent according to Definition 2. In our smooth optimization transfer framework, the main assumption of Theorem 1 requires that the HPP $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$ and the surrogate regularization $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$ together define an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\,\|\boldsymbol{\beta}\|_1$ (cf. Definition 4). The inequality of arithmetic and geometric means (AM-GM) provides a simple but powerful tool for the construction of SVFs using ℓ_2 regularization as the surrogate penalty and is repeatedly applied throughout the paper. It states that, given a list of $n \in \mathbb{N}$ non-negative numbers x_i, $i = 1, \ldots, n$, it holds that $\frac{x_1 + \ldots + x_n}{n} \geqslant \sqrt[n]{x_1 \cdots x_n}$ with equality if and only if $x_1 = \ldots = x_n$.

In the case of the HPP, it allows us to determine the minimum of the surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$ under the constraint $\boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^d$.

Lemma 5 *Given the parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}$, the minimum of surrogate ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$ subject to $\boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\,\|\boldsymbol{\beta}\|_1$ in (8) and is given by* $\min_{\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}} \quad \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 = 2\,\|\boldsymbol{\beta}\|_1 \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d$.

Proof Because the HPP defines element-wise multiplication, we can minimize $u_j^2 + v_j^2$ such that $u_j v_j = \beta_j$ for some $\beta_j \in \mathbb{R}$ and $j = 1, \ldots, d$. Using the AM-GM inequality for $n = 2$ and the non-negative numbers u_j^2 and v_j^2, we obtain

$$\frac{u_j^2 + v_j^2}{2} \geqslant \sqrt{u_j^2 v_j^2} = \sqrt{(u_j v_j)^2} = \sqrt{\beta_j^2} = |\beta_j|\,,$$

which reduces to equality if and only if $u_j^2 = v_j^2$, yielding a minimum value of $u_j^2 + v_j^2 = 2|\beta_j|$. Repeating this procedure for all $j = 1, \ldots, d$ shows that the constrained minimum of the surrogate penalty is indeed equal to $2\|\boldsymbol{\beta}\|_1$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$. □

The optimality conditions $u_j^2 = v_j^2 = |\beta_j|$ further ensure that we can derive continuous solutions $(\hat{u}_j, \hat{v}_j)$ as functions of $\beta_j = u_j v_j$. Analytically, $(\hat{u}_j, \hat{v}_j)$ are of the form

$$\underset{(u_j, v_j) : u_j v_j = \beta_j}{\arg\min} \quad u_j^2 + v_j^2 = \begin{cases} \left(\sqrt{|\beta_j|}, \sqrt{|\beta_j|}\right) \text{ and } \left(-\sqrt{|\beta_j|}, -\sqrt{|\beta_j|}\right) & \text{for } \beta_j > 0 \\ (0, 0) & \text{for } \beta_j = 0 \\ \left(\sqrt{|\beta_j|}, -\sqrt{|\beta_j|}\right) \text{ and } \left(-\sqrt{|\beta_j|}, \sqrt{|\beta_j|}\right) & \text{for } \beta_j < 0\,. \end{cases}$$

Further, we can determine the number of equivalent solutions in the surrogate problem, corresponding to a specific solution in the original problem, using the AM-GM inequality. Due to this duplicity for each $j = 1, \ldots, d$, there are a total of 2^s equivalent local minimizers $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ of $\mathcal{Q}$ for each local minimizer $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ of $\mathcal{P}$, where $s = ||\hat{\boldsymbol{\beta}}||_0$. Moreover, we can establish the stability of a solution mapping in a more general setting for solutions that are characterized by necessary optimality conditions similar to the above, obtained from applying the AM-GM inequality to the squared surrogate parameters u_j^2 and v_j^2 for $j \in [d]$.

Lemma 6 *(Lower hemicontinuity of $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta}_j)$ under Ass. 2) Let $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta}) : \mathbb{R}^d \rightrightarrows \mathbb{R}^{d_\xi}$ denote the solution mapping $\boldsymbol{\beta} \mapsto \arg\min_{\boldsymbol{\xi} \in \mathcal{K}^{-1}(\boldsymbol{\beta})} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, where the parametrization $\mathcal{K}$ satisfies*

Assumption 2, so that each coordinate β_{ji} depends on the $l \in [k]$ factors only via the ith entries of the vectors $\boldsymbol{\xi}_{jl} \in \mathbb{R}^{|\mathcal{G}_j|}$ and the scalars ξ_{jl} in a power-product structure with exponents $\alpha_l \geqslant 1$ for $l \in [k], j \in [L]$. Let the surrogate penalty be $\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j) = \sum_{l=1}^{k} \alpha_l \|\boldsymbol{\xi}_{jl}\|_2^2$. Then $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is lower hemicontinuous on $\mathbb{R}^d$.

Note that for the simple HPP, we have $L = d$ and $k = 2$ scalars with exponents $\alpha_l = 1$ for each $j \in [d]$. This ensures Assumption 1 holds in this case, and we can combine Lemma 5 with Theorem 1 to obtain the final equivalence.

Corollary 1 *The optimization of $\mathcal{P}$ in (8) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}$ in (9) by Definition 2, and solutions to the base problem can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}} \odot \hat{\boldsymbol{v}})$.*

Inspired by the implicitly regularized elastic net (Zhao et al., 2022), we propose the following explicitly regularized differentiable variant:

Remark 1 *(Smooth elastic net formulation via HPP)* We can readily extend the HPP optimization transfer for ℓ_1 regularization to the Elastic Net penalty $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2$, $\alpha \in (0,1)$, as introduced in Zou and Hastie (2005). To do this, we merely redefine $\tilde{\mathcal{L}}(\boldsymbol{\psi}, \boldsymbol{\beta}) \triangleq \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda\alpha\|\boldsymbol{\beta}\|_2^2$ and $\lambda\tilde{\mathcal{R}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq \lambda(1-\alpha)\|\boldsymbol{\beta}\|_1$. Applying the HPP, we minimize

$$\tilde{\mathcal{L}}(\boldsymbol{\psi}, \boldsymbol{u} \odot \boldsymbol{v}) + \frac{\lambda}{2}\tilde{\mathcal{R}}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot \boldsymbol{v}) + \lambda\alpha\|\boldsymbol{u} \odot \boldsymbol{v}\|_2^2 + \frac{\lambda(1-\alpha)}{2}(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2)$$

over $(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v})$ instead of $(\boldsymbol{\psi}, \boldsymbol{\beta})$. Solutions to the Elastic Net-regularized problem can be reconstructed after optimization of the smooth surrogate as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}} \odot \hat{\boldsymbol{v}})$.

Remark 2 *(Smooth formulations of SCAD/MCP/TL1 via HPP)* As in the previous remark, we can construct smooth surrogates for objectives with non-convex SCAD (Fan & Li, 2001), MCP (Zhang, 2010), or transformed ℓ_1 (Zhang & Xin, 2018) regularization via the HPP and a corresponding surrogate regularizer. The surrogate regularizer is constructed by replacing all terms involving $|\beta_j|$ with the left-hand side of the inequality $(u_j^2 + v_j^2)/2 \geqslant |\beta_j|$. A detailed derivation can be found in Appendix A.21.

Further, for general smoothly parametrized objectives $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}))$ using the HPP, one requirement for equivalence to $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ by Lemma 2 is that the parametrization $\mathcal{K}$ is locally open at the local minimizers $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) \in \mathbb{R}^d \times \mathbb{R}^d$ of $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}))$. In fact, the Hadamard product of two d-dimensional real-valued vectors is a (uniformly) open map everywhere (Balcerzak et al., 2016), meaning that under $\mathcal{K}$, the image of any open ball around $(\boldsymbol{u}, \boldsymbol{v})$ in $\mathbb{R}^d \times \mathbb{R}^d$ contains an open ball around $\boldsymbol{u} \odot \boldsymbol{v}$ in $\mathbb{R}^d$ for all $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^d \times \mathbb{R}^d$ (cf. Def. 3).

3.2 The Hadamard Difference Parametrization

An alternative smooth optimization transfer approach for ℓ_1 regularization is based on the Hadamard difference parametrization (HDP), which is defined as

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d,\ (\boldsymbol{\gamma}, \boldsymbol{\delta}) \mapsto \boldsymbol{\gamma} \odot \boldsymbol{\gamma} - \boldsymbol{\delta} \odot \boldsymbol{\delta} = \boldsymbol{\beta}. \tag{11}$$

This variant of the HPP is often employed in studying the implicit regularization effects of GD in linear neural networks, facilitating an easier theoretical analysis (Vaskevicius et al., 2019; Vivien et al., 2022; Woodworth et al., 2020). In our framework, applying the HDP and imposing explicit surrogate ℓ_2 regularization on $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ corresponds to ℓ_1 regularization of $\boldsymbol{\beta}$, albeit without the scaling factor of 2 present in the HPP. To establish a connection between the HDP and the HPP, let $\boldsymbol{\gamma} = \frac{\boldsymbol{u}+\boldsymbol{v}}{2}$ and $\boldsymbol{\delta} = \frac{\boldsymbol{v}-\boldsymbol{u}}{2}$, or equivalently, $\boldsymbol{u} = \boldsymbol{\gamma} - \boldsymbol{\delta}$ and $\boldsymbol{v} = \boldsymbol{\gamma} + \boldsymbol{\delta}$. Then it is easy to confirm that $\boldsymbol{\gamma} \odot \boldsymbol{\gamma} - \boldsymbol{\delta} \odot \boldsymbol{\delta} = \boldsymbol{u} \odot \boldsymbol{v}$.
For an objective $\mathcal{P}$ with ℓ_1 regularization of $\boldsymbol{\beta}$, we can construct a smooth surrogate $\mathcal{Q}$ applying the HDP and surrogate ℓ_2 regularization. Both objectives can be written as

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \textstyle\sum_{j=1}^d |\beta_j|\,, \tag{12}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\gamma}^2 - \boldsymbol{\delta}^2) + \lambda(\|\boldsymbol{\gamma}\|_2^2 + \|\boldsymbol{\delta}\|_2^2) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\gamma}^2 - \boldsymbol{\delta}^2) + \lambda \textstyle\sum_{j=1}^d \left(\gamma_j^2 + \delta_j^2\right). \tag{13}$$

To show equivalence of the smooth surrogate, we first establish that $\mathcal{K}$ and $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \|\boldsymbol{\gamma}\|_2^2 + \|\boldsymbol{\delta}\|_2^2$ together define an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$.

Lemma 7 *Given the parametrization map $\mathcal{K}(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \boldsymbol{\gamma} \odot \boldsymbol{\gamma} - \boldsymbol{\delta} \odot \boldsymbol{\delta}$ as defined in Eq. (11), the minimum of the surrogate ℓ_2 regularization $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\gamma}, \boldsymbol{\delta})) = \|\boldsymbol{\gamma}\|_2^2 + \|\boldsymbol{\delta}\|_2^2$ subject to $\boldsymbol{\gamma} \odot \boldsymbol{\gamma} - \boldsymbol{\delta} \odot \boldsymbol{\delta} = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ in (12):*

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\delta} : \boldsymbol{\gamma}^2 - \boldsymbol{\delta}^2 = \boldsymbol{\beta}} \|\boldsymbol{\gamma}\|_2^2 + \|\boldsymbol{\delta}\|_2^2 \ = \|\boldsymbol{\beta}\|_1 \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d. \tag{14}$$

For each β_j in $\boldsymbol{\beta}$, either γ_j^2 or δ_j^2 must equal zero at the minimum, depending on the sign of β_j, with the square of the non-zero parameter being equal to $|\beta_j|$. The minimizers $(\hat{\gamma}_j, \hat{\delta}_j)$ hence form a continuous set-valued function of β_j.

Corollary 2 *Optimization of $\mathcal{P}$ (12) is equivalent to optimization of the smooth surrogate $\mathcal{Q}$ (13), and solutions to the $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}^2 - \hat{\boldsymbol{\delta}}^2)$.*

For the preservation of local minima in general, potentially unregularized objectives $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ parametrized using the HDP, Lemma 2 again requires local openness of the HDP at local minimizers $(\hat{\gamma}, \hat{\delta})$ of $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\gamma}, \boldsymbol{\delta}))$. Recall that rotating a point $(u, v) \in \mathbb{R}^2$ by $45°$ clockwise about the origin defines the transformation $(\gamma, \delta) \triangleq (\frac{u+v}{\sqrt{2}}, \frac{v-u}{\sqrt{2}})$. Evaluating the HDP at the rotated point yields $\gamma^2 - \delta^2 = 2uv$, showing that the HDP constitutes a rotation of the HPP scaled by a factor of 2, with both actions preserving the openness. Details on the difference between HPP and HDP can be found in Appendix C.1.

3.3 Geometric Intuition and Induced Network Architectures

Correspondence to diagonal linear networks The HPP $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$ and HDP $\boldsymbol{\beta} = \boldsymbol{\gamma} \odot \boldsymbol{\gamma} - \boldsymbol{\delta} \odot \boldsymbol{\delta}$ parametrizations reveal close connections to diagonal linear networks and linear regression (Tibshirani, 2021; Woodworth et al., 2020). Assume a simple linear

model $f(\boldsymbol{x}_i|\boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ with no additional parameters $\boldsymbol{\psi}$. This can be represented as a simplistic neural network with d input neurons, a single output neuron, and d edges for the weights β_j with linear activations and no additional bias terms. Applying the respective parametrization, e.g., using the HPP, $f(\boldsymbol{x}_i|\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{x}_i^\top(\boldsymbol{u} \odot \boldsymbol{v})$, which represents a diagonal linear network with linear activations, a single hidden layer, and no bias terms. By Corollary 1, this is equivalent to a linear regression with ℓ_1 regularization of $\boldsymbol{\beta}$ under ℓ_2 regularization of the weights. Figure 3 shows two such linear networks, with the diagonal network corresponding to the HPP on the left, and the diagonal network corresponding to the HDP in the middle. This correspondence, however, is not limited to overparametrized linear models. For example, we can "stretch out" a network architecture by inserting additional diagonal layers at certain locations, promoting localized sparse representations. More generally, we can overparametrize any layer of a DNN by replacing its weights $\boldsymbol{\beta}$ by $\mathcal{K}(\boldsymbol{\xi})$. Imposing suitable surrogate regularization on the weights $\boldsymbol{\xi}$ then induces sparsity in the original layer in its parametrization.

Geometric intuition A graphical analysis of our optimization transfer approach for ℓ_1 regularization using the HPP provides additional insights into the underlying geometry. Figure 4a illustrates why the minimum of the surrogate ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$ over $\{(\boldsymbol{u}, \boldsymbol{v}) : \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}\}$ equals $2\|\boldsymbol{\beta}\|_1$. The setting in Fig. 4a shows the HPP $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}$ (blue/green), the majorizing surrogate ℓ_2 penalty (orange), as well as the feasible set defined by the fiber $\mathcal{K}^{-1}(\boldsymbol{\beta})$ (black hyperbola). In this example, we set $d = 1$ and fix $\beta = 10$. Alternatively, we can interpret the plot as an illustration for only a single entry $\beta_j = 10$, $j \in [d]$. The shape of $\mathcal{K}(u_j, v_j) = u_j v_j$ is a hyperbolic paraboloid, and the fiber $\mathcal{K}^{-1}(\beta_j) \subset \mathbb{R} \times \mathbb{R}$ for $\beta_j = 10$ is obtained by intersecting $\mathcal{K}$ with the horizontal plane $\beta_j = 10$. The geometric shape of the resulting set is a rectangular hyperbola, forming an unbounded feasible set in the constrained minimization problem stated in Lemma 5. Since the surrogate regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v})$ defines an elliptic paraboloid for each $j \in [d]$, the constrained minimization problem is solved by searching the entry-wise feasible sets with respect to the smallest surrogate penalty. For a hyperbola defined by $u_j(v_j) = \frac{\beta_j}{v_j}, \beta_j > 0$, this is achieved at the vertices $(\sqrt{\beta_j}, \sqrt{\beta_j})$ and $(-\sqrt{\beta_j}, -\sqrt{\beta_j})$, with a minimal distance of $\sqrt{2\beta_j}$. Similarly, for $\beta_j < 0$, minimal distance of $\sqrt{2|\beta_j|}$ is attained at $(-\sqrt{|\beta_j|}, \sqrt{|\beta_j|})$ and $(\sqrt{|\beta_j|}, -\sqrt{|\beta_j|})$. For $\beta_j = 0$, the fiber $\mathcal{K}^{-1}(0)$ contains all points on the coordinate axes, with 0 minimal distance at (0, 0). The majorization property is visualized in Fig. 4b.

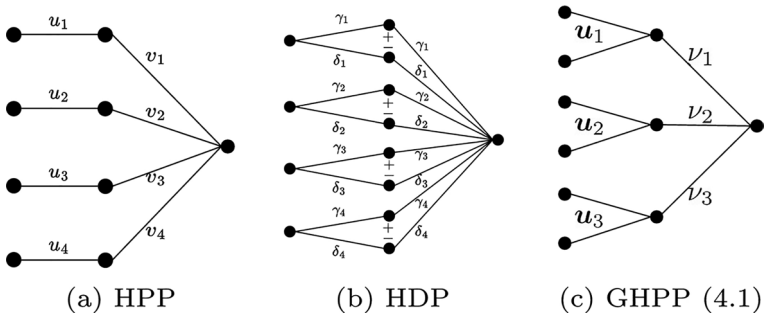
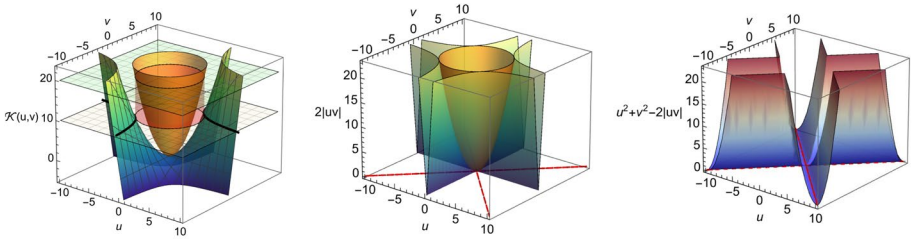


Fig. 3 Diagonal linear networks corresponding to different parametrizations of a linear predictor: **a** HPP (ℓ_1), **b** HDP (ℓ_1), **c** Structure-inducing parametrization (GHPP for $\ell_{2,1}$, cf. 4.1) with grouping layer. Left nodes are inputs, and the right-most node the output



(a) Optim. Transfer (HPP) (b) Majorization of ℓ_1 via surrogate ℓ_2 penalty (HPP)

Fig. 4 Illustration of ℓ_1 optimization transfer using HPP and surrogate ℓ_2 regularization on a scalar $\beta_j = 10$ (lower plane). The hyperbolic paraboloid (blue/green) shows the parametrization $\mathcal{K}(u_j, v_j) = u_j v_j$ and the elliptic paraboloid (orange) the ℓ_2 surrogate. The fiber $\mathcal{K}^{-1}(10)$ defines a hyperbola (black), whose two vertices achieve minimal a min. ℓ_2 penalty of $2|10| = 20$ (upper plane) over $\mathcal{K}^{-1}(10)$. **b)** Majorization of overparametrized ℓ_1 term $2|u_j v_j|$ (blue/green) through ℓ_2 penalty. The ℓ_2 (orange) is tightly "hugged" by the ℓ_1 term. The difference of both regularizers attains zero at the red perpendicular continuous lines intersecting at the origin, illustrating the l.h.c. of the SVF solution map. The lines are defined by $|u_j| = |v_j|$. It can be seen that for any (u_j, v_j) with $|u_j| = |v_j|$, and any $\beta'_j = u'_j v'_j$ close to $\beta_j = u_j v_j$, there is (u'_j, v'_j) near (u_j, v_j) with $|u'_j| = |v'_j|$. This is because the solutions continuously increase in $|\beta|$ as we move away from 0

We further demonstrate how the proposed optimization transfer to an equivalent smooth surrogate transforms the loss landscape using a simple toy objective in Fig. 1 and a more detailed visualization in Fig. 19.

4 Hadamard Group Lasso for Structured Sparsity

In many applications, we have additional *a priori* structural information on the parameters, e.g., that certain gene pathways can only be jointly relevant or that a set of dummy-coded features representing a categorical variable should either be included in the model or fully selected out. To obtain structured sparsity, we make use of parametrization maps "tying together" groups of parameters through shared factors, with the property that adding smooth ℓ_2 regularization on the surrogate parameters induces an $\ell_{2,1}$ (group lasso) penalty $2\sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2$ in the base parametrization $\boldsymbol{\beta}$. This is fundamentally different from the structured HPP approach discussed in Hoff (2017), where structure is induced through dependent Gaussian priors on $\boldsymbol{u}$ and $\boldsymbol{v}$ instead of mixed-norm regularization.

Set-up for structured sparsity regularization Let $[d]$ denote the index set corresponding to the entries of $\boldsymbol{\beta} \in \mathbb{R}^d$, and define $\mathcal{G}_j \subseteq [d]$ to be the subsets of indices corresponding to groups $j = 1, \ldots, L$. Let $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$ form a partition of $[d]$, i.e. $\cup_{j=1}^{L} \mathcal{G}_j = \{1, \ldots, d\}$ and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ for $i \neq j$, so that $|\mathcal{G}_1| + \ldots + |\mathcal{G}_L| = d$. The parameter vector $\boldsymbol{\beta}$ contains the group-wise vectors $\boldsymbol{\beta}_j$, i.e., $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L)^\top$, where $\boldsymbol{\beta}_j = (\beta_{ji})_{i \in \mathcal{G}_j} \in \mathbb{R}^{|\mathcal{G}_j|}$ for $j \in [L]$.

4.1 Group Hadamard Product Parametrization

For the group Hadamard product parametrization (GHPP), we again use the parametrization structure $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$, but now with the elements of $\boldsymbol{v}$ (and thus also $\boldsymbol{\beta}$) constrained to reflect the group membership. Noting that $\mathbb{R}^d = \mathbb{R}^{|\mathcal{G}_1|+\ldots+|\mathcal{G}_L|}$, the Hadamard factors are

$$\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top \in \mathbb{R}^d,\ \boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L)^\top = \begin{pmatrix} \nu_1 \mathbf{1}_{|\mathcal{G}_1|} \\ \vdots \\ \nu_L \mathbf{1}_{|\mathcal{G}_L|} \end{pmatrix} \in \mathbb{R}^d\ .$$

Then we have $\boldsymbol{\beta}_j = \boldsymbol{u}_j \odot \boldsymbol{v}_j = \nu_j \cdot (u_{j1}, \ldots, u_{j|\mathcal{G}_j|})^\top \in \mathbb{R}^{|\mathcal{G}_j|}$ for $j \in [L]$.

Note that in this parametrization, the second Hadamard factor $\boldsymbol{v}$ is a d-dimensional vector containing values $\nu_1, \ldots, \nu_L$, where each ν_j is repeated $|\mathcal{G}_j|$ times in $\boldsymbol{v}$. Comparing this to the Hadamard factor $\boldsymbol{v} = (v_1, \ldots, v_d)^\top$ in the HPP, the d distinct entries of $\boldsymbol{v}$ are replaced by entries that are constant within groups $\mathcal{G}_1, \ldots, \mathcal{G}_L$, thereby "tying" together the parameters in each $\mathcal{G}_j$. The first Hadamard factor $\boldsymbol{u}$ remains unconstrained as in the HPP, i.e., $\boldsymbol{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$. Letting $\boldsymbol{\nu} \in \mathbb{R}^L$ denote $(\nu_1, \ldots, \nu_L)^\top$, the GHPP map is defined as:

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{\nu}) \mapsto \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_L \end{pmatrix} \odot \begin{pmatrix} \nu_1 \mathbf{1}_{|\mathcal{G}_1|} \\ \vdots \\ \nu_L \mathbf{1}_{|\mathcal{G}_L|} \end{pmatrix} = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix} = \boldsymbol{\beta}, \quad (15)$$

where we use the notation $\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} \triangleq (\boldsymbol{u}_j \nu_j)_{j \in \mathcal{G}}$. Given an objective $\mathcal{P}$ with non-smooth regularization $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\,\|\boldsymbol{\beta}\|_{2,1}$, defining the surrogate regularization as $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}) \triangleq \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{\nu}\|_2^2$ provides a smooth optimization transfer $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$, from which we construct the smooth surrogate $\mathcal{Q}$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + 2\lambda \|\boldsymbol{\beta}\|_{2,1} = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + 2\lambda \textstyle\sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2\,, \quad (16)$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{v}) + \lambda(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{\nu}\|_2^2) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{v}) + \lambda \textstyle\sum_{j=1}^{L} \left(\|\boldsymbol{u}_j\|_2^2 + \nu_j^2\right). \quad (17)$$

The functions $\mathcal{K}$ and $\mathcal{R}_{\boldsymbol{\xi}}$ are chosen so that we obtain an SVF for $\mathcal{R}_{\boldsymbol{\beta}}$:

Lemma 8 *Given the parametrization map $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, the minimum of the surrogate ℓ_2 regularization $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}) = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{\nu}\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ in (16) and is*

$$\min_{\boldsymbol{u}_j, \nu_j : \boldsymbol{\beta}_j = \nu_j \boldsymbol{u}_j} \textstyle\sum_{j=1}^{L} \|\boldsymbol{u}_j\|_2^2 + \nu_j^2 = 2 \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2 \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d, \quad (18)$$

According to Lemma 6, the optimality conditions $\|\boldsymbol{u}_j\|_2^2 = \nu_j^2 = \|\boldsymbol{\beta}_j\|_2$ of the AM-GM inequality in the proof allow us to derive the minimizers $(\hat{\boldsymbol{u}}_j, \hat{\nu}_j)$ as a lower hemicontinuous function of $\boldsymbol{\beta}_j$ for all $\boldsymbol{\beta}_j \in \mathbb{R}^{|\mathcal{G}_j|}$:

$$\arg\min_{\substack{(\boldsymbol{u}_j,\nu_j):\\ \boldsymbol{\beta}j=\nu j\boldsymbol{u}j}} \|\boldsymbol{u}j\|_2^2+\nu_j^2=\begin{cases}\pm\left(\boldsymbol{\beta}_j/\sqrt{\|\boldsymbol{\beta}j\|2},\sqrt{\|\boldsymbol{\beta}j\|_2}\right) & \|\boldsymbol{\beta}_j\|_2,0\\ (\mathbf{0},0) & \|\boldsymbol{\beta}_j\|_2=0\end{cases} \tag{19}$$

The tuple $(\mathcal{R}_{\boldsymbol{\beta}},\mathcal{K},\mathcal{R}_{\boldsymbol{\xi}})$ is thus a valid optimization transfer for $\ell_{2,1}$ group sparsity:

Corollary 3 *The optimization of $\mathcal{P}$ in (16) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}$ in (17) by Definition 2, and solutions to the base problem can be obtained as $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})=(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{u}}\odot_{\mathcal{G}}\hat{\boldsymbol{\nu}})$.*

Note that there are only two equivalent minimizers for each $\boldsymbol{\beta}_j$ given $\|\boldsymbol{\beta}_j\|_2>0$, as the sign of $\hat{\nu}_j$ uniquely determines the sign of all $\hat{u}_{ji}$ in $\hat{\boldsymbol{u}}_j$ for $i\in\mathcal{G}_j$. Thus, for each minimizer $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})$ of $\mathcal{P}$, there will be 2^s equivalent corresponding solutions $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{u}},\hat{\boldsymbol{\nu}})$ to $\mathcal{Q}$, where s is the number of groups $\mathcal{G}_j$ with $\|\hat{\boldsymbol{\beta}}_j\|_2>0$.

For linear predictors, structure-inducing overparametrization was also studied in Tibshirani (2021) and Dai et al. (2021), however, without proving the matching local minima property or going beyond linearity. Similar to the HPP approach to smooth ℓ_1 regularization, the GHPP corresponds to a particular network structure with linear activations and a grouping layer when applied to a linear model, as shown in Fig. 3c. The ℓ_2 regularized network then corresponds to a linear model with an $\ell_{2,1}$ penalty.

Considering the preservation of local minima in a general objective $\mathcal{P}(\boldsymbol{\psi},\boldsymbol{\beta})$ under smooth parametrization of $\boldsymbol{\beta}$ using the GHPP, local openness of $\mathcal{K}$ at the local solutions to $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{\xi}))$ is a crucial requirement for Lemma 2. This assumption, however, is not straightforward for the GHPP. While the Banach open mapping theorem states that every continuous linear surjection between Banach spaces is globally open, it is known that this openness principle can not be extended to *bilinear* continuous surjections (Balcerzak et al., 2013; Horowitz, 1975). A widely used counterexample of a bilinear continuous surjection that is not open everywhere is given, e.g., in (Rudin, 1991, Chapter 2, Exercise 11), corresponding to the GHPP for $L=1$ and $d=2$. Therefore, as opposed to the HPP, whose global openness is discussed at the end of Sect. 3.1, the GHPP is not globally open in general. Whereas local minimality would be preserved under the HPP for any function due to its global openness, this does not hold for the GHPP, depending on which points qualify as local minimizers. The points of openness for the GHPP are as follows:

Lemma 9 *(Local openness of the GHPP) The parametrization map defined by $\mathcal{K}:\mathbb{R}^d\times\mathbb{R}^L$, $(\boldsymbol{u},\boldsymbol{\nu})\mapsto\boldsymbol{u}\odot_{\mathcal{G}}\boldsymbol{\nu}$ is locally open at $(\boldsymbol{u},\boldsymbol{\nu})$, with $\boldsymbol{u}=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L)^\top$ and $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_L)^\top$, if the $(\boldsymbol{u}_j,\nu_j)$ are such that $\nu_j=0$ implies $\|\boldsymbol{u}_j\|_2=0$ for all $j\in[L]$.*

To establish matching local minima, we thus need to ensure that local solutions $(\hat{\boldsymbol{u}},\hat{\boldsymbol{\nu}})$ to $\mathcal{P}(\boldsymbol{\psi},\mathcal{K}(\boldsymbol{u},\boldsymbol{\nu}))$ are indeed points of openness. Note that all minimizers of $\mathcal{Q}(\boldsymbol{u},\boldsymbol{\nu})$ are of the form stated in (19), i.e., either $(\hat{\boldsymbol{u}}_j,\hat{\nu}_j)=(\mathbf{0},0)$, or the $(\hat{\boldsymbol{u}}_j,\hat{\nu}_j)$ are such that $\|\hat{\boldsymbol{u}}_j\|_2>0$ and $|\hat{\nu}_j|>0$ for all $j\in[L]$. Then, by Lemma 9, $\mathcal{K}(\boldsymbol{u},\boldsymbol{\nu})$ is locally open at all potential local minimizers of $\mathcal{Q}$.

4.2 Adjusting the GHPP for Variable Group Sizes

The well-known group lasso, initially proposed by Yuan and Lin (2006), does not employ plain $\ell_{2,1}$ regularization, but includes additional weights accounting for the variable group sizes $|\mathcal{G}_j|, j \in [L]$. With this modification, we can define the non-smooth penalty as $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{L} \sqrt{|\mathcal{G}_j|}\, \|\boldsymbol{\beta}_j\|_2$. Interestingly, this regularizer can be obtained as a simple extension to the previous approach by introducing a scaling factor in the surrogate penalty. The derivation is deferred to Appendix A.9. This results in the following smooth objective $\mathcal{Q}$ and corresponding equivalent group lasso regularized objective $\mathcal{P}$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + 2\lambda \textstyle\sum_{j=1}^{L} \sqrt{|\mathcal{G}_j|} \|\boldsymbol{\beta}_j\|_2\,, \tag{20}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}) + \lambda \textstyle\sum_{j=1}^{L} \left(\|\boldsymbol{u}_j\|_2^2 + |\mathcal{G}_j| \nu_j^2\right). \tag{21}$$

5 Going Deeper: Non-convex Regularization with Hadamard Product Parametrizations of Depth *k*

The Hadamard product parametrizations factorizing $\boldsymbol{\beta}$ using two factors $\boldsymbol{u}, \boldsymbol{v}$ can be naturally extended to deeper factorizations of depth $k > 2,\ k \in \mathbb{N}$. For a suitable surrogate penalty, these parametrizations induce (a restricted class) of non-convex ℓ_q and $\ell_{p,q}$ regularizers for $0 < q < 1$ and $0 < q < p \leqslant 2$ in the base parametrization $\boldsymbol{\beta}$.

5.1 Hadamard Product Parametrization of Depth *k*

First, consider a multilinear extension of the bilinear HPP termed the HPP_k,

$$\mathcal{K} : \textstyle\prod_{l=1}^{k} \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \mapsto \bigodot_{l=1}^{k} \boldsymbol{u}_l = \boldsymbol{\beta}, \tag{22}$$

where $\prod_{l=1}^{k} \mathbb{R}^d$ denotes the kth Cartesian power of $\mathbb{R}^d$ and $k > 2$. The depth two case recovers the simple HPP (10). Each $\beta_j,\ j \in [d]$, is parametrized as the product $\prod_{l=1}^{k} u_{jl}$, where each factor u_{jl} is taken from a different $\boldsymbol{u}_l$. Further, we define $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \triangleq \sum_{l=1}^{k} \|\boldsymbol{u}_l\|_2^2$. Then, minimizing $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$ subject to the constraint imposed by the parametrization map $\mathcal{K}$ yields an SVF for non-convex ℓ_q regularization with $q = 2/k$:

Lemma 10 *Given the parametrization map $\mathcal{K}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \boldsymbol{u}_l^{\odot k}$, the minimum surrogate ℓ_2 regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \sum_{l=1}^{k} \|\boldsymbol{u}_l\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\, \|\boldsymbol{\beta}\|_{2/k}^{2/k}$ and is given by $\min_{\boldsymbol{u}_l : \boldsymbol{\beta} = \boldsymbol{u}_l^{\odot k}} \sum_{l=1}^{k} \|\boldsymbol{u}_l\|_2^2 = k \|\boldsymbol{\beta}\|_{2/k}^{2/k}\ \forall \boldsymbol{\beta} \in \mathbb{R}^d$.*

A visualization of the HPP_k for $k = 3$ can be found in Appendix C.2, illustrating the shape of the fibers of $\mathcal{K}$ and the majorization of the non-smooth $\ell_{2/3}$ penalty by the smooth surrogate ℓ_2 penalty. Given an objective $\mathcal{P}$ with smooth $\mathcal{L}$ and non-convex $\ell_{2/k}$ regularization $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, applying the optimization transfer defined by $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$ yields the corresponding $\mathcal{Q}$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \|\boldsymbol{\beta}\|_{2/k}^{2/k} = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \sum_{j=1}^{d} |\beta_j|^{2/k}, \tag{23}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{u}_l^{\odot k}\big) + \lambda \sum_{l=1}^{k} \|\boldsymbol{u}_l\|_2^2 = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{u}_l^{\odot k}\big) + \lambda \sum_{j=1}^{d} \sum_{l=1}^{k} u_{jl}^2 \,. \tag{24}$$

The optimality conditions of the AM-GM inequality ensure lower hemicontinuity of the solution map in Lemma 10 by Lemma 6, implying equivalence of $\mathcal{P}$ and $\mathcal{Q}$:

Corollary 4 *The optimization of $\mathcal{P}$ (23) is equivalent to optimization of the smooth surrogate $\mathcal{Q}$ (24) by Def. 2, and solutions to $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}}_l^{\odot k})$.*

This result is also shown in Hoff (2017), but its application there is limited to simple linear models that can be optimized with alternating ridge regression. Extending the HPP, the HPP_k also corresponds to a horizontally "stretched" diagonal network structure with increased depth, as shown in Fig. 5a. The relation of parametrization and corresponding network structure for linear models was also studied in simpler settings and without proof of our general result (Dai et al., 2021; Tibshirani, 2021). Besides these works in explicit regularization, a strand of literature in DL uses diagonal linear networks to study the implicit regularization of GD (Gissin et al., 2019; Gunasekar et al., 2018; Li et al., 2021; Moroshko et al., 2020; Woodworth et al., 2020).
Regarding applications of the HPP_k to general objectives $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ without surrogate regularization, we can establish the global openness of the k-linear surjection $\mathcal{K}$:

Lemma 11 *The map $\mathcal{K} : \prod_{l=1}^{k} \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \mapsto \bigodot_{l=1}^{k} \boldsymbol{u}_l$ is globally open.*

Consequently, applying Lemma 2, smoothly parametrizing any continuous objective using the HPP_k preserves the local minima of $\mathcal{P}$.

5.2 Group Hadamard Product Parametrizations of Depth *k*

The smooth optimization transfer for $\ell_{2,1}$ regularized problems can be naturally extended to structured sparsity with non-convex $\ell_{2,2/k}$ regularization. We start with the same set-up as in Sect. 4, but now consider deeper factorizations of $\boldsymbol{\beta}$. Recall that the GHPP is defined

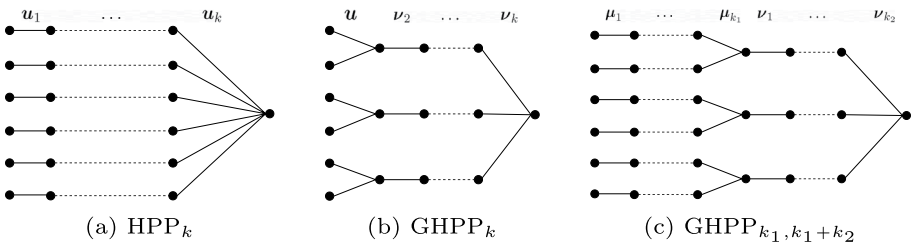


Fig. 5 Deep diagonal linear networks corresponding to parametrizations of a linear predictor. **a)** HPP (for $\ell_{2/k}$), **b)** GHPP_k (for $\ell_{2,2/k}$), **c)** $\mathrm{GHPP}_{k_1,k_1+k_2}$ (for $\ell_{2/k_1,2/(k_1+k_2)}$). The depth up to and including the grouping layer is k_1, followed by $k_2 = k - k_1$ more diagonal layers. Nodes on the left represent input features and the single node on the right the output

as $\mathcal{K}(\boldsymbol{u},\boldsymbol{\nu}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} = \boldsymbol{\beta}$. Further factorizing the grouping parameter $\boldsymbol{\nu}$ into $k-1$ Hadamard factors, i.e., $\boldsymbol{\nu} = \bigodot_{r=1}^{k-1} \boldsymbol{\nu}_r$, defines the GHPP_k map:

$$\mathcal{K} : \mathbb{R}^d \times \prod_{r=1}^{k-1} \mathbb{R}^L \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) \mapsto \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_L \end{pmatrix} \odot \begin{pmatrix} \mathbf{1}_{|\mathcal{G}_1|} \prod_{r=1}^{k-1} \nu_{1r} \\ \vdots \\ \mathbf{1}_{|\mathcal{G}_L|} \prod_{r=1}^{k-1} \nu_{Lr} \end{pmatrix}$$

Equivalently, the parametrization on the group level reads $\boldsymbol{\beta}_j = \boldsymbol{u}_j \prod_{r=1}^{k-1} \nu_{jr}$, where $\boldsymbol{\beta}_j, \boldsymbol{u}_j \in \mathbb{R}^{|\mathcal{G}_j|}$ and $\nu_{jr} \in \mathbb{R}$, for $j = 1, \ldots, L$ and $r = 1, \ldots, k-1$. Applying plain ℓ_2 regularization under this parametrization, i.e., $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) \triangleq \|\boldsymbol{u}\|_2^2 + \sum_{r=1}^{k-1} \|\boldsymbol{\nu}_r\|_2^2$, induces the non-smooth and non-convex regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = k \|\boldsymbol{\beta}\|_{2,2/k}^{2/k}$ for structured sparsity in the base parametrization. To show this, we first prove that the minimum ℓ_2 penalty under the parametrization map constraint equals $\mathcal{R}_{\boldsymbol{\beta}}$:

Lemma 12 *Given the parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}$, the minimum of the surrogate ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) \triangleq \|\boldsymbol{u}\|_2^2 + \sum_{r=1}^{k-1} \|\boldsymbol{\nu}_r\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}_1 \ldots, \boldsymbol{\nu}_{k-1}) = \boldsymbol{\beta}$ constitutes the following SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k \|\boldsymbol{\beta}\|_{2,2/k}^{2/k}$:*

$$\min_{\substack{\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1} : \\ \boldsymbol{\beta} = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}}} \sum_{j=1}^{L} \left(\|\boldsymbol{u}_j\|_2^2 + \sum_{r=1}^{k-1} \nu_{jr}^2 \right) = k \|\boldsymbol{\beta}\|_{2,2/k}^{2/k} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d. \tag{25}$$

For an objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with non-convex $\ell_{2,2/k}$ regularization,

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \|\boldsymbol{\beta}\|_{2,2/k}^{2/k} = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2^{2/k}, \tag{26}$$

the smooth surrogate $\mathcal{Q}$ obtained from the tuple $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$ is given by

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}\big) + \lambda \sum_{j=1}^{L} \left(\|\boldsymbol{u}_j\|_2^2 + \sum_{r=1}^{k-1} \nu_{jr}^2 \right). \tag{27}$$

By Lemma 6, the optimality conditions obtained in the proof of Lemma 12 imply a lower hemicontinuous solution map as a function of $\boldsymbol{\beta}$, so that we can state:

Corollary 5 *The optimization of $\mathcal{P}$ in (26) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}$ in (27) by Definition 2, and solutions to $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}} \odot_{\mathcal{G}} \hat{\boldsymbol{\nu}_r}^{\odot(k-1)})$.*

We can think of the parametrization $\mathcal{K}$ as a composition involving the GHPP and the HPP_{k-1} for $\boldsymbol{\nu}$ to gain insights into the network architecture corresponding to a linear model overparametrized by $\mathcal{K}$. Compared to the depth-two network matching the GHPP in Fig. 3c, the network for the GHPP_k shown in Fig. 5b adds $k-1$ diagonal layers after the initial layer, corresponding to the additional deeper factorization of $\boldsymbol{\nu}$ in the GHPP_k. In the previously mentioned less general setting, Tibshirani (2021) first discovered that optimizing a

network as in Fig. 3c with weight decay induces an objective with the same global minimum as an $\ell_{2,2/k}$ regularized linear model.

Regarding the preservation of local minima when applying the GHPP_k to a general objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ without surrogate regularization, we can use the compositional nature of the GHPP_k to obtain points of local openness, as required by Lemma 2:

Corollary 6 *(Points of local openness of the $GHPP_k$) The parametrization mapping $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}$ is locally open at $(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1})$ whenever the GHPP (15) is locally open at $(\boldsymbol{u}, \boldsymbol{\nu}_r^{\odot(k-1)})$.*

Besides the proof in Appendix A.13, conditions for the local openness of the GHPP are stated in Lemma 9. Note that the optimality conditions in the proof of Lemma 12 thus also imply the local openness of $\mathcal{K}$ at all local minimizers of $\mathcal{Q}$.

5.3 Generalizing the GHPP to Mixed $\ell_{p,q}$ Quasi-Norms

We can extend the principle behind the construction of the GHPP_k, i.e., starting with the GHPP and factorizing the $\boldsymbol{\nu}$ parameter, to deeper parametrizations factorizing both $\boldsymbol{u}$ and $\boldsymbol{\nu}$ simultaneously into k_1 and k_2 Hadamard factors. In the following, we establish that smooth ℓ_2 regularization of the resulting surrogate parameters induces non-convex $\ell_{p,q}$ mixed-norm regularization in the base parametrization, with $(p, q) \in \{(2/k_1, 2/(k_1 + k_2)) : k_1, k_2 \in \mathbb{N}\}$. We start with the same structured parameter set-up as in Sect. 4, partitioning the components of $\boldsymbol{\beta}$ into L groups. Consider the GHPP map given by $\boldsymbol{\beta} = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, with $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$, together comprising L pairs of group-wise parameters $(\boldsymbol{u}_j, \nu_j)$. Factorizing each $\boldsymbol{u}_j$ into a product of k_1 Hadamard factors $\boldsymbol{\mu}_{jt}$, $t = 1, \ldots, k_1$, and each ν_j into a product of k_2 scalar factors ν_{jr}, $r = 1, \ldots, k_2$, we can define the following surjective parametrization mapping $\mathcal{K}$ termed the $\mathrm{GHPP}_{k_1, k_1+k_2}$:

$$\begin{aligned}\mathcal{K} : \textstyle\prod_{t=1}^{k_1} \mathbb{R}^d \times \prod_{r=1}^{k_2} \mathbb{R}^L \to & \mathbb{R}^d,\ (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}) \mapsto \boldsymbol{\mu}_t^{\odot k_1} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot k_2} \\ &= \begin{pmatrix} \boldsymbol{\mu}_{1t}^{\odot k_1} \\ \vdots \\ \boldsymbol{\mu}_{Lt}^{\odot k_1} \end{pmatrix} \odot \begin{pmatrix} \mathbf{1}_{|\mathcal{G}_1|} \prod_{r=1}^{k_2} \nu_{1r} \\ \vdots \\ \mathbf{1}_{|\mathcal{G}_L|} \prod_{r=1}^{k_2} \nu_{Lr} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix} = \boldsymbol{\beta}\,, \end{aligned} \tag{28}$$

where $\boldsymbol{\mu}_t \triangleq (\boldsymbol{\mu}_{1t}, \ldots, \boldsymbol{\mu}_{Lt})^\top \in \mathbb{R}^d$ and $\boldsymbol{\nu}_r \triangleq (\nu_{1r}, \ldots, \nu_{Lr})^\top \in \mathbb{R}^L$. Note that each $\boldsymbol{\mu}_{jt}$ is the $t-$th factor of the j-th parameter group with entries $(\mu_{j1t}, \ldots, \mu_{j|\mathcal{G}_j|t})^\top \in \mathbb{R}^{|\mathcal{G}_j|}$. On the group level, the parametrization reads $\boldsymbol{\beta}_j = \boldsymbol{u}_j \nu_j = (\bigodot_{t=1}^{k_1} \boldsymbol{\mu}_{jt}) \prod_{r=1}^{k_2} \nu_{jr} = \boldsymbol{\mu}_{jt}^{\odot k_1} \prod_{r=1}^{k_2} \nu_{jr}$, for $j \in [L]$. Further, let $k \triangleq k_1 + k_2$ denote the total factorization depth. To derive the non-convex group-sparse regularizer for $\boldsymbol{\beta}$ induced through ℓ_2 regularization of $\boldsymbol{\mu}_{jt}, \nu_{jr}$ for $j \in [L], t \in [k_1], r \in [k_2]$, a simple generalization of the AM-GM inequality is required. Defining the surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$ as plain ℓ_2 regularization, we can show that $\mathcal{R}_{\boldsymbol{\xi}}$ and $\mathcal{K}$ induce an SVF for mixed-norm $\ell_{p,q}$ regularization.

Lemma 13 *Given a parametrization $\mathcal{K}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}) = \boldsymbol{\mu}_t^{\odot k_1} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot k_2}$, the minimum surrogate ℓ_2 regularization $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1} \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}) \triangleq \sum_{t=1}^{k_1} \|\boldsymbol{\mu}_t\|_2^2 + \sum_{r=1}^{k_2} \|\boldsymbol{\nu}_r\|_2^2$ subject*

to $\mathcal{K}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}) = \boldsymbol{\beta}$ *constitutes an SVF for* $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k \|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k}$ *and is given by*

$$\min_{\substack{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2} : \\ \boldsymbol{\beta} = \boldsymbol{\mu}_t^{\odot k_1} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot k_2}}} \sum_{j=1}^{L} \left(\sum_{t=1}^{k_1} \|\boldsymbol{\mu}_{jt}\|_2^2 + \sum_{r=1}^{k_2} \nu_{jr}^2 \right) = k \|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d. \tag{29}$$

Note that by Lemma 6, the optimality conditions in the proof above ensure lower hemicontinuity of the solution map to the SVF. Assuming an objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with non-convex $\ell_{2/k_1,2/k}$ regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, the optimization transfer $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$ defines the following equivalent smooth surrogate $\mathcal{Q}$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k} = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/k}, \tag{30}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{\mu}_t^{\odot k_1} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot k_2}\big) + \lambda \sum_{j=1}^{L} \left(\sum_{t=1}^{k_1} \|\boldsymbol{\mu}_{jt}\|_2^2 + \sum_{r=1}^{k_2} \nu_{jr}^2 \right). \tag{31}$$

Corollary 7 *The objective* $\mathcal{P}$ *in (30) is equivalent to the smooth surrogate* $\mathcal{Q}$ *in (31) by Definition 2, and solutions to* $\mathcal{P}$ *can be constructed as* $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\mu}}_t^{\odot k_1} \odot_{\mathcal{G}} \hat{\boldsymbol{\nu}}_r^{\odot k_2}) = \big(\hat{\boldsymbol{\psi}}, \big(\bigodot_{t=1}^{k_1} \hat{\boldsymbol{\mu}}_t\big) \odot_{\mathcal{G}} \big(\bigodot_{r=1}^{k_2} \hat{\boldsymbol{\nu}}_r\big)\big)$.

Figure 5c shows an exemplary network architecture corresponding to the $\mathrm{GHPP}_{k_1,k_1+k_2}$ applied to an LM (Dai et al., 2021). The architecture also provides an intuitive visualization of mixed-norm regularization for structured sparsity as a whole. While the depth of the first block of diagonal layers, factorizing $\boldsymbol{u}$ into k_1 Hadamard factors $\boldsymbol{\mu}_t$, determines the induced *within-group* norm, the depth of the group-wise constant parameters in $\boldsymbol{\nu}$ into k_2 Hadamard factors determines the induced *between-group* norm.

5.4 Parametrizations with Parameter Sharing

Parameter or weight sharing enables interesting modifications of the previously presented parametrizations, as the parameter redundancy caused by overparametrization can be greatly reduced by allowing for shared parameters between the Hadamard factors. Parameter sharing can be defined as identifying two or more parameters of an objective function as a single parameter, i.e., interpreting them as identical. For example, the group structure-inducing GHPP, $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, is essentially the HPP $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}$, but with shared parameters $\boldsymbol{v}_j = \nu_j \mathbf{1}_{|\mathcal{G}_j|}$ within groups $j \in [L]$, collapsed into the scalar ν_j. Despite requiring many fewer additional parameters, these parametrizations still define a valid SVF $\mathcal{R}_{\boldsymbol{\beta}}$ like their fully overparametrized counterparts.

Deep HPP with shared parameters Consider the parametrization map for the HPP_k, defined as $\mathcal{K}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \bigodot_{l=1}^{k} \boldsymbol{u}_l$. By introducing parameter sharing between $(k-1)$ Hadamard factors, i.e., replacing the Hadamard product of $k-1$ separate factors with a self-Hadamard product, we retain enough freedom to ensure surjectivity of the parametrization.

We use $\boldsymbol{u} \in \mathbb{R}^d$ to denote the first Hadamard factor, and $\boldsymbol{v}^{k-1} \in \mathbb{R}^d$ for the other factors that are collapsed into a single shared vector $\boldsymbol{v} \in \mathbb{R}^d$. The following defines the HPP_k^{shared}

$$\complement : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u} \odot (\boldsymbol{v} \odot \cdots \odot \boldsymbol{v}) = \boldsymbol{u} \odot (\bigodot_{l=1}^{k-1} \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}^{k-1} = \boldsymbol{\beta}. \quad (32)$$

The suitable surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$ to obtain an SVF is a re-weighted ℓ_2 penalty accounting for the increased contribution of the shared parameter to the parametrization. More precisely, the shared parameter $\boldsymbol{v}$ is counted $(k-1)$ times, thereby ensuring the appropriate re-weighting for $\mathcal{R}_{\boldsymbol{\xi}}$ to define an SVF for non-convex $\ell_{2/k}$ regularization:

Lemma 14 *Given the parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}^{k-1}$, the minimum surrogate ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2/k}^{2/k}$, i.e., $\min_{\boldsymbol{u},\boldsymbol{v}:\boldsymbol{u}\odot\boldsymbol{v}^{k-1}=\boldsymbol{\beta}} \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2 = k\|\boldsymbol{\beta}\|_{2/k}^{2/k}\ \forall \boldsymbol{\beta} \in \mathbb{R}^d$.*

However, despite constituting a valid SVF with less overparametrization, parameter sharing breaks the balance and symmetry in the parametrization, with unclear consequences for the optimization. Yet, we can relate the GD optimization dynamics for the HPP_k^{shared} to its fully overparametrized counterpart HPP_k under identical initialization of the to-be-shared parameters. Using a rescaled learning rate for the shared factors, we derive identical updates for both variants, as detailed in Appendix A.16.

Moreover, initializing *all* k Hadamard factors of the HPP_k identically prohibits them from changing their sign over the iterations for sufficiently small step sizes, since the gradient updates vanish as the reconstructed coefficients β_j approach zero. This is because under identical initialization, the gradient of the entry-wise parametrization $\mathcal{K}_j((u_{jl})_{l=1}^k) = \prod_{l=1}^k u_{jl}$ is given by a vector of identical entries $\prod_{l' \neq l} u_{jl'}$ for $l \in [k]$. Hence, since all u_{jl}^0 are identical, they also receive identical updates and thus stay identical, $u_{jl}^{t+1} = u_{jl'}^{t+1}\ \forall l \in [k]$. In this case, the only way for the product to change signs is by passing through the origin, which is prohibited by the vanishing products in the parametrization gradient. This can be exploited to solve non-negative least squares (Chou et al., 2022; Gissin et al., 2019).

HDP of depth *k* without and with shared weights Similar to how the HPP can be generalized to the deeper parametrization HPP_k, the HDP from 3.2 can be generalized to deeper variants inducing $\ell_{2/k}$ regularization in the base parametrization under ℓ_2 regularization of the surrogate parameters. Chou et al. (2023) mention this fully-overparametrized generalization of the HDP, here named HDP_k: $\mathcal{K} : \prod_{l=1}^{2k} \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \mapsto \boldsymbol{u}_l^{\odot k} - \boldsymbol{v}_l^{\odot k} = \boldsymbol{\beta}$. In their analysis of gradient dynamics they restrict themselves to the case of identical initialization, effectively giving rise to the following parametrization termed the HDP_k^{shared}, incorporating parameter sharing between the $\boldsymbol{u}_l$ and the $\boldsymbol{v}_l$ for $l \in [k]$, respectively: $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u}^k - \boldsymbol{v}^k = \boldsymbol{\beta}$. In DL, these parametrizations are widely applied in the implicit regularization literature to obtain simple-to-analyze depth-*k* networks that exhibit rich optimization and implicit regularization dynamics see, e.g.,, (Woodworth et al., 2020; Li et al., 2021).

6 Hadamard Powers: Non-integer Factorization Depths for Unrestricted ℓ_q and $\ell_{p,q}$ Regularization

The parametrizations based on (group) Hadamard products can induce ℓ_q and $\ell_{p,q}$ regularization under surrogate ℓ_2 regularization for the restricted class $q \in \{2/k | k \in \mathbb{N}\}$ and $(p,q) \in \{(2/k_1, 2/(k_1+k_2) | k_1, k_2 \in \mathbb{N}\}$. Extending Hadamard product-based parametrizations to Hadamard powers permits a more flexible choice of the induced regularizer, allowing selection of the previously restricted p and q arbitrarily from $q \in (0,1]$ and $0 < q < p \leqslant 2$. Thus, smooth optimization for non-convex sparse regularization can be achieved using our framework for any feasible real-valued choices of q and p, extending previous results to non-integer factorization depths.

6.1 Hadamard Power Parametrization

To construct a parametrization that induces ℓ_q regularization of $\boldsymbol{\beta}$ under (slightly modified) ℓ_2 regularization of the surrogate parameters for any $q \in (0,1]$, we extend the notion of self-Hadamard products to Hadamard powers. For powers v_j^k with positive, real-valued exponents k to be well-defined, we require positivity of the base v_j, e.g., by designing parametrizations of the form $\beta_j = u_j |v_j|^{k-1}$. The resulting HPowP_k map is

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u} \odot |\boldsymbol{v}|^{\circ(k-1)} = \boldsymbol{\beta}, \tag{33}$$

where $|\boldsymbol{v}|^{\circ(k-1)}$ denotes element-wise raising the $|v_j|$ to the $(k-1)$-th power, with $k > 2$ to ensure $\mathcal{K} \in \mathcal{C}^1$. This generalizes the self-Hadamard product $\bigodot_{l=1}^{k-1} \boldsymbol{v} = \boldsymbol{v}^{k-1}$, defined for $k \in \mathbb{N}$, to real-valued positive exponents, with $\circ(k-1)$ denoting non-integer exponents.

Lemma 15 *Given the parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot |\boldsymbol{v}|^{\circ(k-1)}$, the minimum surrogate ℓ_2 regularization $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2/k}^{2/k}$, i.e., $\min_{\boldsymbol{u},\boldsymbol{v}:\boldsymbol{u}\odot|\boldsymbol{v}|^{\circ(k-1)}=\boldsymbol{\beta}} \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2 = k\|\boldsymbol{\beta}\|_{2/k}^{2/k} \forall \boldsymbol{\beta} \in \mathbb{R}^d$.*

Note that the sign of the constrained minimizer $\hat{u}_j$ is uniquely determined by the sign of β_j due to the non-negativity of $|\hat{v}_j|^{k-1}$. By the optimality conditions, the squared coefficients u_j^2 and $|v_j|^2$ must equal $|\beta_j|^{2/k}$ at the minimum, so that by Lemma 6, the set-valued solution map is lower hemicontinuous and Assumption 1 is satisfied. Thus, for any $k > 2$, given an $\ell_{2/k}$ regularized base objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$,

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k \|\boldsymbol{\beta}\|_{2/k}^{2/k}, \tag{34}$$

we can construct an equivalent differentiable $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v})$ from the tuple $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$:

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{u} \odot |\boldsymbol{v}|^{\circ(k-1)}\big) + \lambda \left(\|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2\right) . \tag{35}$$

Corollary 8 *The optimization of $\mathcal{P}$ in (34) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}$ in (35) by Definition 2, and solutions to $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}} \odot |\hat{\boldsymbol{v}}|^{\circ(k-1)})$.*

Note that similar to Lemma 14, we modify the usual ℓ_2 regularization by multiplying each of the $|v_j|^2$ by $(k-1)$ to reflect the imbalance of $\boldsymbol{u}$ and $\boldsymbol{v}$ in the parametrization $\boldsymbol{\beta} = \boldsymbol{u} \odot |\boldsymbol{v}|^{\circ(k-1)}$.

6.2 Invertible Reparametrization with Hadamard Powers

In Schwarz et al. (2021), a differentiable sparsity-promoting parametrization termed Powerpropagation was introduced, aligning with discussions of related approaches in mathematical optimization (Ramlau & Zarzer, 2012). The underlying motivation is to artificially increase the curvature of the loss landscape, which induces optimization- and initialization-dependent "rich get richer" dynamics for sparse training of DNNs: the key idea is that applying a power parametrization makes the gradient with respect to the surrogate parameters critically depend on their current values (cf. Fig. 19a).

Intuitively, this promotes the accumulation of weights either close to or far away from zero, however, Schwarz et al. (2021) did not realize the induced sparse regularization in the base parametrization under explicit ℓ_2 regularization. Being bijective, Powerpropagation is not an over- but rather a reparametrization given by

$$\mathcal{K} : \mathbb{R}^d \to \mathbb{R}^d, \boldsymbol{v} \mapsto \boldsymbol{v} \odot |\boldsymbol{v}|^{\circ(k-1)} = \boldsymbol{\beta} \quad , k > 1. \tag{36}$$

Note that with a single parameter $\boldsymbol{v}$, it suffices to require $k > 1$ to ensure $\mathcal{K} \in \mathcal{C}^1$ contrasting the previous HPowP_k (33). To see the "rich get richer" effect, consider a generic objective $\mathcal{P}(\boldsymbol{\beta})$, whose gradient under Powerporpagation is given by $\nabla_{\boldsymbol{v}} \mathcal{P}(\mathcal{K}(\boldsymbol{v})) = \nabla_{\boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\beta}) \cdot \mathrm{diag}(k|\boldsymbol{v}|^{\circ(k-1)})$. This additional factor causes amplification of gradients for v_j with large magnitudes and attenuation for small magnitudes. Considering ℓ_2 regularization for this parametrization, the feasible set of the problem $\min_{\boldsymbol{v}:\boldsymbol{v}\odot|\boldsymbol{v}|^{\circ(k-1)}=\boldsymbol{\beta}} \|\boldsymbol{v}\|_2^2$ is a singleton containing $\hat{\boldsymbol{v}}$ such that $\hat{v}_j = \sqrt[k]{|\beta_j|}$ for $\beta_j \geqslant 0$ and $\hat{v}_j = -\sqrt[k]{|\beta_j|}$ for $\beta_j < 0$, $j \in [d]$. Hence, $\|\hat{\boldsymbol{v}}\|_2^2$ contains d summands $\hat{v}_j^2 = |\beta_j|^{2/k}$, and we conclude $\min_{\boldsymbol{v}:\boldsymbol{v}\odot|\boldsymbol{v}|^{\circ(k-1)}=\boldsymbol{\beta}} \|\boldsymbol{v}\|_2^2 = \|\boldsymbol{\beta}\|_{2/k}^{2/k}$. Since the solution map is continuous in $\boldsymbol{\beta}$, Assumption 1 holds. Thus, for an $\ell_{2/k}$ regularized objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with real-valued $k > 1$, we can construct an equivalent smooth $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{v})$ as follows:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_{2/k}^{2/k}, \tag{37}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{v}) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{v} \odot |\boldsymbol{v}|^{\circ(k-1)}\big) + \lambda \|\boldsymbol{v}\|_2^2 \,. \tag{38}$$

Corollary 9 *The optimization of $\mathcal{P}$ in (37) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}$ in (38) by Definition 2, and solutions to $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{v}} \odot |\hat{\boldsymbol{v}}|^{\circ(k-1)})$.*

This result shows that it is the functional shape of the parametrization and its warping effect on the loss surface that induces sparsity, not overparametrization *per se*.

6.3 Hadamard Group Powers (GHPowP)

We can naturally extend the Hadamard power parametrization presented in 6.1 to structured sparsity, thereby obtaining a more flexible choice of the hyperparameters p and q in $\ell_{p,q}$ regularization. The following two subsections are structured analogously to their Hadamard product-based counterparts discussed in Sect. 4. As before, we consider the parameter vector with group structure $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L)^\top$. Consider the following parametrization mapping, named the GHPowP_k,

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^d, (\boldsymbol{u}, \boldsymbol{\nu}) \mapsto \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_L \end{pmatrix} \odot \begin{pmatrix} |\nu_1|^{k-1} \mathbf{1}_{|\mathcal{G}_1|} \\ \vdots \\ |\nu_L|^{k-1} \mathbf{1}_{|\mathcal{G}_L|} \end{pmatrix} = \boldsymbol{\beta}. \quad (39)$$

On the group level, we have $\boldsymbol{\beta}_j = |\nu_j|^{k-1}(u_{j1}, \ldots, u_{j|\mathcal{G}_j|})^\top \, \forall j \in [L]$, where $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$, and $k > 2$. Now, define $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2,2/k}^{2/k}$ and $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}) \triangleq \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{\nu}\|_2^2$.

Lemma 16 *Given the parametrization map $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)}$, the minimum of the surrogate ℓ_2 regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}) \triangleq \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{\nu}\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2,2/k}^{2/k}$ and is given by*

$$\min_{\boldsymbol{u}, \boldsymbol{\nu} : \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)} = \boldsymbol{\beta}} \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{\nu}\|_2^2 = k\|\boldsymbol{\beta}\|_{2,2/k}^{2/k} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d. \quad (40)$$

Using Lemma 6, the optimality conditions provided in the proof ensure lower hemicontinuity of the solution map. Therefore, Assumption 1 holds and we can construct an equivalent smooth surrogate $\mathcal{Q}$ to an $\ell_{2,2/k}$ regularized base objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ for any $k > 2$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k\|\boldsymbol{\beta}\|_{2,2/k}^{2/k}, \quad (41)$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu}) = \mathcal{L}\big(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)}\big) + \lambda\big(\|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{\nu}\|_2^2\big). \quad (42)$$

Corollary 10 *The optimization of $\mathcal{P}$ in (41) is equivalent to the optimization of $\mathcal{Q}$ in (42) by Def. 2, and solutions to $\mathcal{P}$ can be constructed as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{u}} \odot_{\mathcal{G}} |\hat{\boldsymbol{\nu}}|^{\circ(k-1)})$.*

6.4 Mixed Norm Regularization with Hadamard Group Powers

Analogous to the previous subsection, we can further apply Hadamard powers to induce $\ell_{p,q}$ mixed-norm regularization for arbitrary feasible values $0 < q < p \leqslant 2$. As a starting point, we again revisit the structured group set-up $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L)^\top$. However, to allow for non-integer factorization depths, a more complex nested power parametrization is required and constructed in the following. Consider the parametrization $\boldsymbol{\beta} = \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ k_2}$, $k_2 > 1$, with $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top \in \mathbb{R}^d$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top \in \mathbb{R}^L$, corresponding to parametrization (39). Additionally, the auxiliary parameter $\boldsymbol{u}$ is parametrized using an invert-

ible pre-composition, i.e., $\boldsymbol{u} = \boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)}$ with $k_1 > 1$ and surrogate parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L)^\top \in \mathbb{R}^{|\mathcal{G}_1|+\ldots+|\mathcal{G}_L|} = \mathbb{R}^d$. We can then define the $\mathrm{GHPowP}_{k_1,k_1+k_2}$ as

$$\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^d, (\boldsymbol{\mu}, \boldsymbol{\nu}) \mapsto \boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ k_2} = \begin{pmatrix} \boldsymbol{\mu}_1 \odot |\boldsymbol{\mu}_1|^{\circ(k_1-1)} \\ \vdots \\ \boldsymbol{\mu}_L \odot |\boldsymbol{\mu}_L|^{\circ(k_1-1)} \end{pmatrix} \odot \begin{pmatrix} |\nu_1|^{k_2} \mathbf{1}_{|\mathcal{G}_1|} \\ \vdots \\ |\nu_L|^{k_2} \mathbf{1}_{|\mathcal{G}_L|} \end{pmatrix},$$

or equivalently on the group level, $\boldsymbol{\beta}_j = \boldsymbol{u}_j |\nu_j|^{k_2} = \boldsymbol{\mu}_j \odot |\boldsymbol{\mu}_j|^{\circ(k_1-1)} \cdot |\nu_j|^{k_2}$ for groups $j \in [L]$. The parametrization of $\boldsymbol{u}_j$ via $\boldsymbol{\mu}_j$ is bijective, so that for each $u_{ji}, i \in \mathcal{G}_j$ in $\boldsymbol{u}_j$, it holds $\mu_{ji} = \mathrm{sign}(u_{ji}) \cdot |u_{ji}|^{1/k_1}$. Thus, we can express the squared Euclidean norm of $\boldsymbol{\mu}_j$ as

$$\|\boldsymbol{\mu}_j\|_2^2 = \sum_{i \in \mathcal{G}_j} \mu_{ji}^2 = \sum_{i \in \mathcal{G}_j} |u_{ji}|^{2/k_1} = \|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1}.$$

Letting $k \triangleq k_1 + k_2 > 2$, we define the non-convex base regularizer as $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k}$ and the surrogate as $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \triangleq k_1\|\boldsymbol{\mu}\|_2^2 + k_2\|\boldsymbol{\nu}\|_2^2$. Together, $\mathcal{K}$ and $\mathcal{R}_{\boldsymbol{\xi}}$ form an SVF for $\mathcal{R}_{\boldsymbol{\beta}}$:

Lemma 17 *For a parametrization $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ k_2}$, the minimum of the surrogate ℓ_2 regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\nu}) \triangleq k_1\|\boldsymbol{\mu}\|_2^2 + k_2\|\boldsymbol{\nu}\|_2^2$ subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) = \boldsymbol{\beta}$ constitutes an SVF for $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq k\|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k}$ and is given by*

$$\min_{\boldsymbol{\mu}, \boldsymbol{\nu} : \boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ k_2} = \boldsymbol{\beta}} k_1\|\boldsymbol{\mu}\|_2^2 + k_2\|\boldsymbol{\nu}\|_2^2 = k\|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d, \tag{43}$$

The optimality conditions obtained in the proof of this result further ensure Assumption 1 holds by establishing lower hemicontinuity of the set-valued solution map of the SVF according to Lemma 6. Given an $\ell_{2/k_1,2/k}$ regularized objective $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, we can construct a surrogate $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\mu}, \boldsymbol{\nu})$ from the tuple $(\mathcal{R}_{\boldsymbol{\beta}}, \mathcal{K}, \mathcal{R}_{\boldsymbol{\xi}})$:

$$\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda k\|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k}, \tag{44}$$

$$\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{L}\big(\boldsymbol{\psi}, (\boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)}) \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ k_2}\big) + \lambda \left(k_1\|\boldsymbol{\mu}\|_2^2 + k_2\|\boldsymbol{\nu}\|_2^2\right). \tag{45}$$

Corollary 11 *The optimization of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ in (44) is equivalent to the optimization of the smooth surrogate $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu})$ in (45) for any $k_1, k_2 > 1$ according to Definition 2, and solutions to $\mathcal{P}$ can be constructed from solutions to $\mathcal{Q}$ as $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \boldsymbol{\mu} \odot |\boldsymbol{\mu}|^{\circ(k_1-1)} \odot_{\mathcal{G}} |\hat{\boldsymbol{\nu}}|^{\circ k_2})$.*

7 Optimization Details

In this section, we discuss optimization details of our smooth optimization transfer approach and provide some guidance regarding practical implementations.

Iterative optimization using (S)GD A considerable body of literature has established desirable convergence properties of (S)GD that hold in overparametrized non-convex settings,

such as provably almost always escaping (strict) saddle points under random initialization and mild regularity conditions (Lee et al., 2016, 2019). For full-batch GD, however, Du et al. (2017) show that it might take exponentially long to escape saddle points. This can be reduced to polynomial time in the presence of sufficient perturbation in the gradient updates (Ge et al., 2015; Jin et al., 2017), emphasizing the benefit of SGD in efficiently optimizing non-convex problems.

The effect of applying a power-product parametrization on the optimization landscape is to transfer the problem to a more curved space, which impacts the optimization geometry of (S)GD in a way that has been termed the "rich get richer" effect in the Powerpropagation literature Schwarz et al. (2021), but can also take other forms depending on the parametrization. The effect hinges on the multiplicative structure of the parametrizations $\mathcal{K}$, leading to additional multiplicative dependence of the gradient updates of one factor on the current values of a subset of all factors. For the bijective Powerpropagation (36), there is only a single surrogate parameter. Hence, the multiplicative dependence on current values becomes self-reinforcing: As discussed in Sect. 6.2, loss gradients for large-magnitude parameters are magnified, while the gradients for small-magnitude parameters are shrunken, promoting a heavy-tailed weight distribution. The same holds for the HDP (11), but separately for each surrogate parameter: its (entry-wise) gradient is $\nabla\mathcal{K}_j(\gamma_j, \delta_j) = (2\gamma_j, 2\delta_j)$, resulting in self-reinforcing "rich get richer" dynamics for γ_j and δ_j, respectively. Interestingly, for the HPP (10), the multiplicative dependence implies dynamics that are more akin to a "Robin Hood" effect that balances the effective learning rates for both parameters, because its entry-wise gradient is (v_j, u_j), i.e., the partial derivatives contain the *other* factor. For an unbalanced factorization $|u_j| \gg |v_j|$, u_j moves slowly because the respective gradient is attenuated by v_j, and v_j moves fast because the gradient is magnified by u_j. If $|u_j| = |v_j|$, then both parameters have the same effective learning rate. These adaptive dynamics also show in the gradient of the smooth surrogate (46) where the product-structured Jacobian $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi})$ essentially acts as a parameter-dependent preconditioner leading to adaptive step sizes and momentum (Arora et al., 2019).

Besides overparametrization, our approach also imposes differentiable surrogate regularization, which fundamentally differentiates our approach from the mere (unpenalized) overparametrization in implicit regularization. In these methods, the optimization dynamics of the overparametrized problem are tweaked, e.g., via impractically small initialization scales (Vaskevicius et al., 2019; Woodworth et al., 2020; Zhao et al., 2022), so that the optimizer converges to a specific solution on the global minima manifold, such as the minimum ℓ_1-norm solution. In contrast, our optimization transfer approach transforms the sparsity-regularized problem while exactly preserving the solution structure, and hence, in theory, is agnostic to how this solution is reached. This means our smooth surrogate can, in principle, be solved with any optimizer and does not rely on changing the gradient flow dynamics in a specific way to induce implicit regularization behavior.

Critical points Due to the results obtained in Lemma 3 and Lemma 4, any local minimum of the surrogate optimization problem corresponds to a local minimum in the base parametrization. As a result, if the base optimization problem $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ is convex, e.g., for a convex $\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with ℓ_1 or $\ell_{2,1}$ regularization, every local minimum of the surrogate problem $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ is necessarily global. For non-convex base problems, our approach ensures no spurious minima are created in the optimization transfer.

However, such a matching property does not necessarily hold for critical points of the surrogate $\mathcal{Q}$, owed to the zero-product property of the parametrizations $\mathcal{K}$. Without loss of generality, consider a non-smooth regularized objective $\mathcal{P}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ with smooth loss $\mathcal{L}(\boldsymbol{\beta})$ and no additional unregularized parameters $\boldsymbol{\psi}$. Applying the proposed smooth optimization transfer, we construct the surrogate $\mathcal{Q}(\boldsymbol{\xi}) = \mathcal{L}(\mathcal{K}(\boldsymbol{\xi})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ using a smooth parametrization $\mathcal{K}(\boldsymbol{\xi})$ and further imposing surrogate ℓ_2 regularization on $\boldsymbol{\xi}$. The gradient of $\mathcal{Q}$ with respect to $\boldsymbol{\xi}$ is then given by

$$\nabla_{\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{\xi}) = \mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}^{\top}(\boldsymbol{\xi}) \nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi})) + \lambda \nabla_{\boldsymbol{\xi}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}), \tag{46}$$

where $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi})$ is the $d \times d_{\xi}$-dimensional Jacobian of $\mathcal{K}$ at $\boldsymbol{\xi}$, and the gradients $\nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi}))$ and $\nabla_{\boldsymbol{\xi}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ are d- and d_{ξ}-dimensional vectors, respectively. For the parametrizations we consider, the Jacobian $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\mathbf{0})$ at $\boldsymbol{\xi} = \mathbf{0}$ is the null matrix. As $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ is a type of ℓ_2 penalty, we have $\nabla_{\boldsymbol{\xi}} \mathcal{R}_{\boldsymbol{\xi}}(\mathbf{0}) = \mathbf{0}$, and it follows $\nabla_{\boldsymbol{\xi}} \mathcal{Q}(\mathbf{0}) = \mathbf{0}$. Therefore, $\boldsymbol{\xi} = \mathbf{0}$ is a critical point of $\mathcal{Q}$, irrespective of the gradient $\nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi}))$ of $\mathcal{L}$ in the base objective. A derivation of the Hessian of $\mathcal{Q}$ is given in Appendix D.

Regarding the nature of potentially spurious critical points, it is known that parametrizations of depth $k = 2$, such as the HPP or HDP, only induce strict saddle points at $\boldsymbol{\xi} = \mathbf{0}$, since their Hessian evaluated at the origin $\mathcal{H}_{\mathcal{K}}(\mathbf{0})$ contains parameter-independent non-zero constants that ensure a strictly negative eigenvalue (Zhao et al., 2022).[2] Through construction of a counterexample, (Kawaguchi, 2016, Corollary 2.4) shows that the strict saddle property does not necessarily hold for deep factorizations with depth $k > 2$. In our framework, this corresponds to those parametrizations $\mathcal{K}$ that induce non-convex ℓ_q or $\ell_{p,q}$ regularization in the base objective under surrogate ℓ_2 regularization. For this class of non-convex regularizers with unbounded derivatives approaching the origin, $\hat{\boldsymbol{\beta}} = \mathbf{0}$ is always a local minimizer in the base problem $\mathcal{P}(\boldsymbol{\beta})$, regardless of $\mathcal{L}(\boldsymbol{\beta})$ (Loh & Wainwright, 2015). In the constructed smooth surrogate $\mathcal{Q}(\boldsymbol{\xi})$, this is reflected in the Hessian $\mathcal{H}_{\mathcal{Q}(\boldsymbol{\xi})}(\boldsymbol{\xi})$. For $k > 2$ and $\lambda = 0$, the Hessian at $\boldsymbol{\xi} = \mathbf{0}$ degenerates to a null matrix, $\mathcal{H}_{\mathcal{Q}(\boldsymbol{\xi})}(\mathbf{0}) = \mathbf{0}$, inducing a higher-order saddle point. For the regularized problems we are interested in, the strong convexity of $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ guarantees that $\mathcal{H}_{\mathcal{Q}(\boldsymbol{\xi})}(\mathbf{0})$ has only positive eigenvalues. Thus, $\hat{\boldsymbol{\xi}} = \mathbf{0}$ is a local minimizer of $\mathcal{Q}$, corresponding to the local minimizer $\hat{\boldsymbol{\beta}} = \mathbf{0}$ in $\mathcal{P}(\boldsymbol{\beta})$ that is induced by the non-convex regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. Hence, the additional ℓ_2 regularization in our smooth surrogate avoids the problematic spurious non-strict saddle point at $\boldsymbol{\xi} = \mathbf{0}$ induced by $\mathcal{K}$, even for non-convex regularization with $k > 2$. Importantly, $\hat{\boldsymbol{\xi}} = \mathbf{0}$ being a local minimizer of the surrogate $\mathcal{Q}(\boldsymbol{\xi})$ for non-convex ℓ_q and $\ell_{p,q}$ regularization in the base problem $\mathcal{P}(\boldsymbol{\beta})$ is not a property of our proposed method, but of the non-convex regularizer $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$.

Initialization Another relevant question concerns finding effective and well-founded initializations for the surrogate parameters, and how they relate to an appropriate initialization of the base parameter $\boldsymbol{\beta}^t$ at $t = 0$. A natural approach would be to initialize the surrogate parameters functionally equivalent to a standard initialization scheme for the base parameter $\boldsymbol{\beta}^0$. However, in the case of overparametrization, there are many such options, and it is *a priori* unclear how to optimally select among feasible initializations of $\boldsymbol{\beta}^0$. It seems natural to initialize the surrogate parameters $\boldsymbol{\xi}$ according to the optimality conditions provided by the implemented SVF, i.e., $\boldsymbol{\xi}^0 = \hat{\boldsymbol{\xi}}(\boldsymbol{\beta}^0)$, where $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is the set-valued solution mapping of

[2] A strict or ridable saddle point is a saddle at which the Hessian has at least one strictly negative eigenvalue, i.e., there is a direction of descent.

the SVF, and $\boldsymbol{\beta}^0$ is obtained from a standard initialization scheme for the base parameters. This ensures that the optimization is initialized at a minimizer of the surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ over $\{\boldsymbol{\xi} : \mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}^0\}$.

To provide two examples, consider a parametrization of $\boldsymbol{\beta}$ using HPP_k, i.e., $\boldsymbol{\beta} = \boldsymbol{u}_l^{\odot k}$ with surrogate ℓ_2 regularization. One approach then entails initializing the surrogate factors $\boldsymbol{u}_l$ identically as $\boldsymbol{u}_l^0 = \sqrt[k]{|\boldsymbol{\beta}^0|}$, and subsequently multiplying one (arbitrary) factor $\boldsymbol{u}_l^0$ by the respective signs of $\boldsymbol{\beta}^0$.[3] For structured sparsity using the GHPP, $\boldsymbol{\beta} = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, the surrogate parameters are initialized as $\nu_j^0 = \sqrt{\|\boldsymbol{\beta}_j^0\|_2}$ and $\boldsymbol{u}_j^0 = \boldsymbol{\beta}_j^0 / \sqrt{\|\boldsymbol{\beta}_j^0\|_2}$, again equivalently for sign patterns $\pm(\boldsymbol{u}_j^0, \nu_j^0)$. Another option would be to randomly initialize all factors, but increase the initialization scale so that the product is initialized at a desired scale, a variant of which is proposed in Kolb et al. (2025).

Effects on optimization landscape The parametrizations $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{\beta}$ considered in this work (cf. Assumption 2) are based on Hadamard products and powers. This has a notable effect on the loss landscape, primarily due to a modification of curvature induced by the multiplicative nature of the parametrizations. For the bijective Powerpropagation (36), the warping effect of the reparametrization takes place in the same space as the base parameters and can thus be disentangled from overparametrization. Appendix C.3 contains more details.

Practical implementation It is important to consider the case when the surrogate parameters $\boldsymbol{\xi}_j$ corresponding to some base parameter β_j are initialized in an orthant that maps to an incorrect sign under the multiplicative parametrization $\mathcal{K}$ compared to the solution $\hat{\beta}_j$. In these cases, it is crucial to use large learning rates during early iterations, as previously suggested by, e.g., Li et al. (2023). Otherwise, the respective parameter iterates will gradually approach zero from the side of the initial orthant. This occurs due to the "rich get richer" effect, resulting in diminishing gradient magnitudes as the parameter approaches zero, making it difficult to "step over" the zero boundary. Although for most DNNs the sign pattern is not identified, large step sizes in DL have been found to drive SGD toward simpler structures (Andriushchenko et al., 2023; Chen et al., 2024) and balanced Hadamard factors (Ziyin, 2024), thus facilitating sparse optimization. Besides initially large learning rates, we further emphasize the importance of the commonplace recommendation of using either small batch sizes in SGD or perturbing the gradient updates via additional noise injection for faster convergence and the improved ability to escape saddles and local minima (Jin et al., 2017). Further, note that using (S)GD to optimize the differentiable surrogate does not have an inherent proximal step. Consequently, the iterates can only asymptotically approach theoretically zero values in a finite number of steps. However, this issue is benign, as with standard training hyperparameters, numerically zero or negligibly small floating point representations can be attained. For additionally accelerated and more adaptive optimization, a dynamic thresholding schedule can be implemented, in which during training the surrogate parameters $\boldsymbol{\xi}_j = \mathbf{0}$ parametrizing a scalar parameter are thresholded if the reconstructed parameter $|\beta_j| < \varepsilon_{tiny}$. In this way, the weights are automatically removed from future updates early on, allowing the optimization more time exploring sparser subnetworks without having to implement a hard mask or change the architecture. This is not required due to the multiplicative structure of the gradients of $\mathcal{K}$, which necessarily all become zero for all future iterations once $\boldsymbol{\xi}_j = \mathbf{0}$.

[3] Applying any sign pattern to the $\boldsymbol{u}_l$ that respects the signs of $\boldsymbol{\beta}^0$ under the parametrization $\mathcal{K}$ is valid.

8 Related Work

In this section, we relate our optimization transfer framework and the presented parametrizations to prior art. In recent years, parametrizations based on Hadamard products have attracted considerable interest in several fields, including DL, statistics, signal processing, and optimization. The context in which they are applied varies significantly, so we focus on works that use these parametrizations for sparsity-inducing regularization with explicit surrogate regularization, i.e., not relying on manipulating the optimization dynamics as in implicit regularization approaches.

Comparison to prior work using explicit regularization In our work, we extend the literature on approximation-free, differentiable optimization for sparse regularization using a combination of overparametrization and surrogate regularization. An early connection between ℓ_1 and an adaptive variant of ℓ_2 regularization using the HPP was first observed by Grandvalet (1998). In statistics, the basic idea was re-discovered for a restricted problem class by Hoff (2017) using the HPP_k, however, without noting its compatibility with SGD or its applicability to non-linear models. Unlike our work, their suggested method is limited to linear models, for which the overparametrized objective can be optimized using alternating ridge regressions due to the multi-convex nature of the optimization problem. This multi-convexity is lost in more complex models like neural networks, requiring less restrictive optimizers like SGD. Besides these, Tibshirani (2021) additionally studies the GHPP_k in linear models and finds that they have identical global minima to certain weight-decayed network architectures (cf. Fig. 3). Notably, a simple weight-decayed diagonal linear network with one hidden layer has the same global minimum as the lasso, which, in turn, is equivalent to applying the Hadamard product parametrization and ℓ_2 regularization. This observation can also be implicitly inferred from the representation cost analysis of the same architecture presented in Dai et al. (2021). Building on the results of Hoff (2017), the work of Ziyin and Wang (2023) constitutes a first endeavor to adapt differentiable sparse regularization to the dominant SGD optimization paradigm for DNNs, however, using only a two-parameter factorization.

Previous works, however, exhibit several limitations. First, their scope is limited to single or a small subset of known sparsity-inducing parametrizations (cf. Table 1), while we are the first to provide a comprehensive account. Among the works incorporating ℓ_2 regularization to induce sparsity, the proposed methods are either confined to linear models by scope (Dai et al., 2021; Tibshirani, 2021) or restrictions of their optimization procedure (Hoff, 2017). The only work applying overparametrization with ℓ_2 regularization to broader objectives is Ziyin and Wang (2023), whose approach is limited to a simple overparametrization for induced convex ℓ_1-type regularization. Further, we place particular emphasis on ensuring matching local minima as a crucial property to preserve structure in the overparametrized problem, which aligns with the work of Levin (2020) and Nouiehed and Razaviyayn (2022) and was previously only discussed by Hoff (2017); Ziyin and Wang (2023).

Moreover, although the implicit regularization literature also studies several overparametrizations Woodworth et al. (2020); Gunasekar et al. (2018); Nacson et al. (2022); Vaskevicius et al. (2019); Li et al. (2021); Gissin et al. (2019); Moroshko et al. (2020); Chou et al. (2023); Vivien et al. (2022), these works do not consider the induced regularizer under explicit surrogate regularization and base their implicit regularization on the manipulation of optimization dynamics, mainly through vanishing initialization scales. This

renders implicit regularization approaches impractical and fundamentally different from our approach, besides their restriction to convex ℓ_1 regularization in important settings (Nacson et al., 2022). Due to these limitations, the goal of implicit regularization works is mainly to understand optimization dynamics, and few works are specifically geared toward practical applications Zhao et al. (2022); Chou et al. (2022). In contrast, our method does not rely on manipulation of the optimization dynamics to reach a specific minimizer, but has the same solution structure as the sparsity-regularized problem, a property that is independent of optimization dynamics.

Table 2 compares the most closely related prior works. In the following, we further describe the related literature on Hadamard parametrizations in different subfields.

DL literature In the theoretical DL community, the surge in activity can be ascribed to the correspondence of Hadamard product-based parametrizations of linear models and simple, easy-to-analyze network architectures with linear activations (Dai et al., 2021; Tibshirani, 2021), predominantly studied under the name of diagonal linear networks (Even et al., 2023; Gissin et al., 2019; Gunasekar et al., 2018; Li et al., 2021; Pesme et al., 2021; Wang & Jacot, 2024), as well as similar stylized architecture for structured sparsity (Li et al., 2023). These networks are primarily analyzed in the context of implicit regularization effects and the representation cost of neural networks. The first phenomenon studies initialization and trajectory-based regularization effects of (S)GD without any explicit regularization term (Vaskevicius et al., 2019; Woodworth et al., 2020), whereas the latter is concerned with measuring the cost that is required for a DNN to represent particular functions in terms of norms of network weights (Dai et al., 2021; Jacot et al., 2022). Implicit regularization through Hadamard product-based overparametrization of linear models was further extended to robust and sparse linear regression in Ma and Fattahi (2022) using subgradient descent. In the absence of explicit regularization, Chou et al. (2023) study the implicit regularization of two variants of Hadamard parametrizations on gradient flow under vanishing initialization, obtaining improved sample complexity for compressed sensing problems. The "rich" gradient dynamics (Woodworth et al., 2020) caused by identical small initialization are further developed for overparametrized non-negative least squares problems in Chou et al. (2022). Contrasting our explicit surrogate regularization, an important shortcoming of implicit regularization approaches is that it is limited to convex ℓ_q norms for $q \geqslant 1$ for common

Table 2 Overview of related works using Hadamard parametrizations for explicit and implicit sparse regularization.

Reference	Regularization	Induced sparse regularizers	Matching Local Min.	Corre-sponding NN struct.	Application to arbitrary model sub-components
Grandvalet (1998)	Explicit ℓ_2 (adaptive)	ℓ_1	✗	✗	✗ (LM)
Hoff (2017)	Explicit ℓ_2	ℓ_q (restricted)	✓	✗	✗ (LM)
Ziyin and Wang (2023)	Explicit ℓ_2	$\ell_1, \ell_{2,1}$	✓ (ℓ_1)	✗	✓
Tibshirani (2021)	Explicit ℓ_2	$\ell_q, \ell_{2,q}$ (restricted)	✗	✓	✗ (LM)
Zhao et al. (2022)	Implicit (GD)	min-ℓ_1-solution	✓	✗	✗ (LM)
Our framework	Explicit ℓ_2 (weighted)	$\ell_q, \ell_{p,q}$	✓	✓	✓

GD stands for gradient descent, and LM for linear model. In the third column, the addition (restricted) refers to a choice of $q = 2/k,\ k \in \mathbb{N}$.

losses such as the square loss, ruling out non-convex ℓ_q regularization for $q < 1$ (Nacson et al., 2022; Woodworth et al., 2020). Li and Lin (2024) improve the adaptivity of sequence models using multiplicative overparametrization, whereas Kolb et al. (2025a, 2025b) apply structured overparametrization to induce, e.g., attention sparsity in transformer models.

Statistics literature The implicit ℓ_1 regularization effect of applying a simple Hadamard product parametrization to the parameters of a linear model under vanishing initialization and GD was studied by Zhao et al. (2022). Under the name "neuronized priors", Shin and Liu (2022) studies similar parameter factorizations in a Bayesian modeling framework, whereas Cheltsov et al. (2024) proposes Hadamard Langevin dynamics for sampling ℓ_1 priors. (Kaushik et al., 2024) derive asymptotics for a family of reweighted least-squares approaches for linear models with Hadamard product parametrization.

Signal processing literature Li et al. (2023) recently applied the Hadamard Product Parametrization (HPP) to solve the tail-ℓ_1 problem in compressed sensing. In a more general setting, Yang et al. (2022), and subsequently Parhi and Nowak (2023), analyze the sparse functional representations learned by ℓ_2 regularized neural networks with homogeneous activation functions from a signal processing perspective, employing a similar line of reasoning to our work to show the equivalence of group sparse $\ell_{2,1}$ and surrogate ℓ_2 regularization using their Neural Balance Theorem. Similarly, starting with Neyshabur et al. (2015a, 2015b), several works advanced the understanding of ℓ_2 regularized networks and the inductive biases in the learned representations (see, e.g. Pilanci and Ergen, 2020; Ergen and Pilanci, 2021, 2023; Jagadeesan et al., 2022).

Optimization literature Micchelli et al. (2013) study the optimization of convex regularizers such as the ℓ_1 penalty by smoothly approximating the absolute value using a quadratic variational formulation of the regularizer involving an additional surrogate parameter $\boldsymbol{\eta}$. This concept is similarly discussed in Bach et al. (2012) under the umbrella term sub-quadratic norms. Formal connections between the so-called $\boldsymbol{\eta}$-trick and the Hadamard product (over) parametrizations studied in Hoff (2017) are established for convex and lower semicontinuous proper loss functions in Poon and Peyré (2021), who subsequently leverage Hadamard parametrizations to smooth bilevel programming (Poon & Peyré, 2023). Recently, Ouyang et al. (2025) studied smooth ℓ_1 regularization using Hadamard parametrizations and derive the surrogate Kurdyka-Lojasiewicz exponent at second-order stationary points from that of the original objective. An analysis of the optimization dynamics of the HPP_k applied to a linear model under gradient flow is presented in Labarrière et al. (2024) and connections to a corresponding mirror flow are made by, e.g., Poon and Peyré (2023); Labarrière et al. (2024); Jacobs and Burkholz (2025); Jacobs et al. (2025). Another branch of literature in optimization that is related to our approach is the perspective functions framework, a versatile tool for constructing proximal methods (Combettes & Müller, 2018, 2020).

9 Numerical Experiments

In this section, we present experimental findings supporting our theoretical results and demonstrating the generality of our method by applying it to various learning problems ranging from non-convex regularized linear regression to enhanced DNN pruning and filter-sparse convolutional neural networks (CNNs). The main goal of these experiments is not to establish the superiority of our method over other approaches but rather to demonstrate the prac-

tical feasibility and competitiveness of using SGD to solve non-smooth regularization.[4] Details on optimization settings and architectures can be found in Appendix B.

9.1 Failure of (Sub)GD to Solve Sparse Regularization

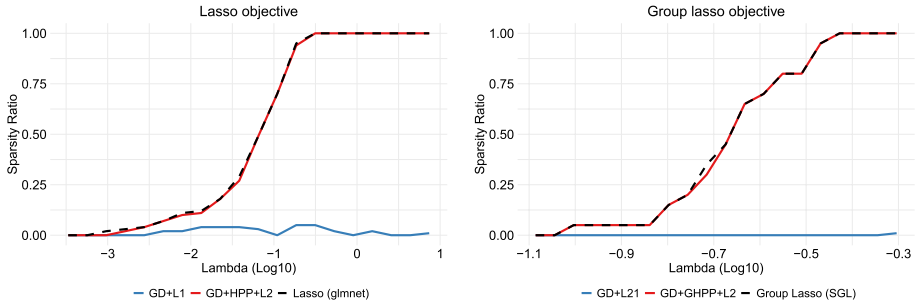
First, we illustrate the failure of directly applying GD to solve both unstructured and structured sparsity regularization, even in the case of a convex (group) lasso objective with linear predictor and independent features. In DL libraries, the gradient at non-differentiable points is typically assigned zero in the GD update, effectively constituting subgradient descent. To this end, we draw $\boldsymbol{X} \in \mathbb{R}^{1000\times 100}$, $\boldsymbol{\beta} \in \mathbb{R}^{100}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{1000}$ from independent Gaussians and compose the noisy outcome as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For the group lasso, the parameters are partitioned into $L = 20$ groups. The objectives in the base parametrization for both regularizers are $\mathcal{P}_{\ell_1}(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ and $\mathcal{P}_{\ell_{2,1}}(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{L}\|\boldsymbol{\beta}_j\|_2$, and we compare three optimization approaches: directly applying GD to the non-smooth objective, GD under smooth optimization transfer using the (G)HPP, and a highly efficient specialized combination of non-smooth methods, implemented in `glmnet` (Friedman et al., 2010) and `SGL` (Simon et al., 2013). The equivalent differentiable objectives of the second approach are defined as $\mathcal{Q}_{\ell_1}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{u} \odot \boldsymbol{v})\|_2^2 + \frac{\lambda}{2}(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2)$ and $\mathcal{Q}_{\ell_{2,1}}(\boldsymbol{u}, \boldsymbol{\nu}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu})\|_2^2 + \frac{\lambda}{2}(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{\nu}\|_2^2)$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{100}$ and $\boldsymbol{\nu} \in \mathbb{R}^{20}$.

Figure 6 shows the failure of direct GD to achieve parameter (group) sparsity. In contrast, applying GD to the equivalent smooth objective $\mathcal{Q}$ matches the regularization paths of the specialized optimizers, providing numerical evidence that by optimizing the equivalent surrogate, the non-smooth base problem can be solved exactly using fully differentiable standard GD. Figure 11 further plots the parameter norms as a function of λ, complementing previous findings. For direct GD, the weight norm even starts to increase for large values of λ, raising serious concerns about the actual effect achieved by direct GD optimization for ℓ_1 regularized DNNs e.g.,, (Han et al., 2015; Wen et al., 2016; Liu et al., 2017).

9.2 Comparison with Convex and Non-convex Regularizers

Next, we investigate the behavior of our smooth optimization method for ℓ_q regularization under SGD in a high-dimensional ($d > n$) sparse linear regression simulation setting, comparing against widely-used convex and non-convex regularizers. The ℓ_q regularized sparse linear regression problem we consider is defined as $\mathcal{P}(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{2/k}^{2/k}$. Smooth optimization of this objective is achieved by overparametrization of $\boldsymbol{\beta}$ using the HPP_k for factorization depths $k \in \{2, 3, 4, 6\}$. Combined with ℓ_2 regularization of the surrogate parameters, equivalent smooth surrogates for SGD optimization are given by $\mathcal{Q}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{u}_l^{\odot k}\|_2^2 + \frac{\lambda}{k}\sum_{l=1}^{k}\|\boldsymbol{u}_l\|_2^2$. We compare our models against widely used implementations of convex ℓ_1 and non-convex SCAD and MCP regularizers, as well as an oracle model that is obtained as the least squares estimator using only the true informative features. All models are evaluated with respect to their standardized estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/\|\boldsymbol{\beta}^*\|_2^2$, as well as their test root mean squared error $\sqrt{n^{-1}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_2^2}$ (RMSE).

[4] We stress that the proposed method offers a differentiable formulation of sparse regularizers, thus inherently tying its performance to that of the induced regularizer.



(a) HPP vs direct GD for lasso objective (b) GHPP vs direct GD for group lasso

Fig. 6 Comparison of regularization paths of (G)HPP-based GD and direct (Sub)GD optimization of the non-smooth ℓ_1 regularized lasso **a** and $\ell_{2,1}$ regularized group lasso **b** objectives. Dashed lines indicate (optimal) solutions of the non-smooth optimizer. Parameters (groups) with magnitude (ℓ_2 norm) below 1×10^{-6} are considered 0.

Figure 7 shows the distribution of estimation and test prediction errors over 30 simulation runs. The results indicate that the performance of our differentiable method for ℓ_q regularization improves monotonically with the factorization depth k, outperforming ℓ_1 regularization for $k > 2$, and surpassing or matching both SCAD and MCP. These results are noteworthy considering the use of vanilla SGD without tuning. Comparing the performance of the Hadamard parametrized model of depth $k = 2$ and the standard implementation of the lasso, we find virtually identical results, empirically validating our theoretical results.

Besides estimation and prediction error, the support recovery of our approach is also of interest for variable selection. In line with previous findings, we demonstrate empirically that deeper factorizations improve support recovery. Appendix B contains the corresponding results, as well as additional experiments for both a low-dimensional ($d < n$) setting and varying sparsity of the ground-truth parameter, whose findings are consistent with previous results.

9.3 Unstructured Sparsity: Enhanced DNN Pruning

In this application, we demonstrate how one-shot pruning of DNNs can be enhanced with differentiable (non-convex) sparse regularization using the HPP_k. Pruning (Lecun et al., 1989) is the dominant sparsification technique for DNNs (Hoefler et al., 2021) and selectively removes components according to some saliency criterion, typically chosen to be the weight magnitude. Our method, as any sparse regularizer, can be easily combined with other sparsification schemes, e.g., by additionally applying global magnitude pruning (Blalock et al., 2020) after training the overparametrized sparse network.

To evaluate this approach, we train a LeNet-300-100 on the MNIST image classification task (Deng, 2012) using Adam. The fully connected network has two hidden layers with 300 and 100 units and ReLU activation. We apply the HPP_k to all 266, 610 weights and biases for depths $k \in \{2, 3, 4\}$. After training, the Hadamard factors are collapsed and the reconstructed model is further pruned to desired sparsity levels without finetuning. Figure 8 (left) shows the pruning curves for $\lambda = 0$ and different depths k. The plot reveals that factorizing the parameters without surrogate ℓ_2 regularization already improves the pruning performance, in line with the arguments provided for the mechanism of Power-

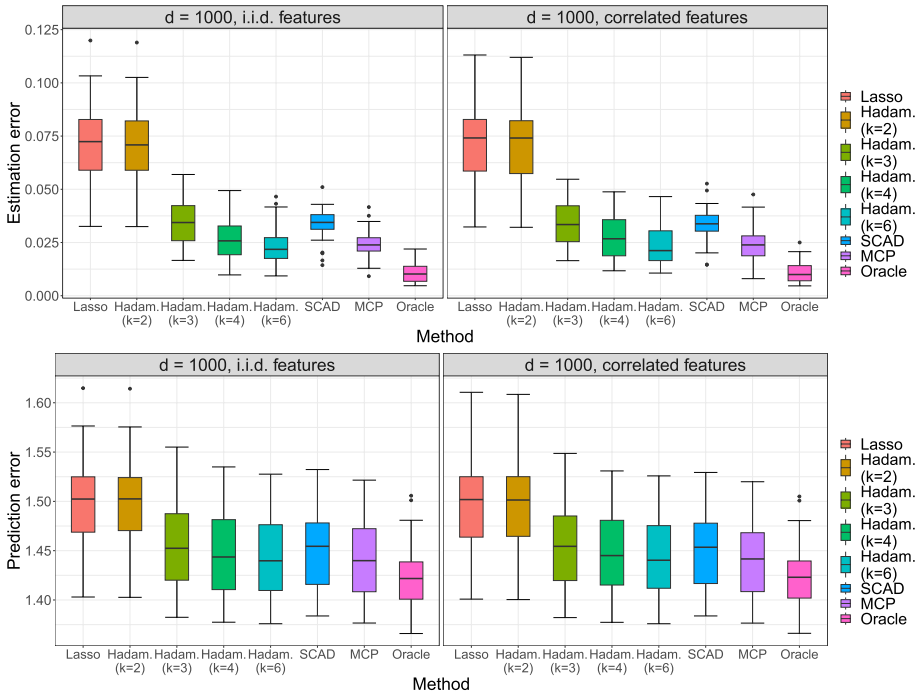


Fig. 7 Stand. estimation error (top row) and test prediction error (bottom row) for two $\mathbf{\Sigma}$ settings (columns) of our approach for depths $k \in \{2, 3, 4, 6\}$, compared with specialized optimizers for ℓ_1 and non-convex SCAD and MCP penalties

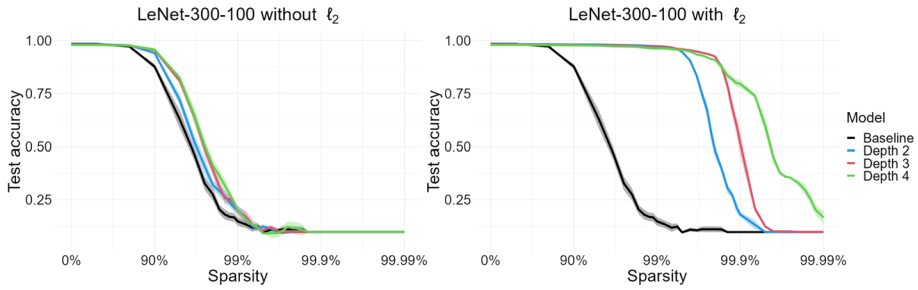


Fig. 8 One-shot pruning curves obtained by overparametrizing the weights and biases of a LeNet-300-100 trained on MNIST using the HPP_k. Left: results for unregularized models. Right: adding smooth ℓ_2 regularization perhaps counterintuitively produces profound sparsity-inducing effects. Magnitude-based pruning constitutes the baseline and the error bars show standard errors over five random initializations

propagation (Schwarz et al., 2021). This is surprising since the model expressivity has not changed, highlighting important trajectory-dependent effects. The right plot is with active ℓ_2 regularization, inducing sparse $\ell_{2/k}$ regularization according to our theory. The Pareto curves are taken as the best performance over a grid of λ values for each sparsity level. The results show drastic improvements over both the baseline (magnitude pruning) and the unregularized overparametrization, with induced non-convex regularization ($k > 2$) further

outperforming induced ℓ_1 sparsity. At a fixed accuracy of 75%, magnitude pruning still uses $\approx 24,000$ param., while the models for $k = 2, 4$ require only $\approx 1,800$ and 230 parameters, respectively. Similarly, at a fixed sparsity of 99.9%, the model performance for $k = 2$ almost degrades to random guessing, while the depth 4 model retains $> 80\%$ test accuracy.

9.4 Structured Sparsity: Filter-Sparse CNNs

The next experiment applies the structured Hadamard power parametrization from Sect. 6.3 to a small VGG-style CNN to obtain filter sparsity. The network has a total of 99, 178 parameters of which 64, 800 are filter weights. Although structured sparsity in DL generally leads to poorer performance-sparsity trade-offs than unstructured sparsity, its capacity to jointly remove whole model components permits a much greater reduction in computational footprint and is thus of particular interest for practical applications. Writing the regularized CNN training objective for filter sparsity as $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_{2,2/k}^{2/k}$, all biases and the weights of fully-connected layers are contained in $\boldsymbol{\psi}$ while $\boldsymbol{\beta}$ comprises the grouped filter weights of the convolutional layers. Applying the GHPowP_k as defined in (39) to $\boldsymbol{\beta}$, the equivalent differentiable objective reads $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)}) + \frac{\lambda}{k} \sum_{j=1}^{L} (\|\boldsymbol{u}_j\|_2^2 + (k-1)\nu_j^2)$, where L is the total number of filters. Effectively, the weights of each filter are multiplied by a shared scalar $|\nu_j|^{k-1}$, inducing the group structure. Note that by using a structured Hadamard power parametrization, only one additional parameter per filter is introduced for any factorization depth k, resulting in minimal overparametrization (99, 370 parameters). Figure 9 shows the regularization path for the overparametrized CNNs trained on MNIST using real-valued depths $k \in \{2, 2.5, 3\}$. The models are trained using SGD without any post-hoc pruning and compared to (structured) magnitude pruning of the original CNN based on the ℓ_2 norm of the filter weights. The results show a $> 90\%$ filter reduction at a negligible drop in accuracy, with deeper factorizations allowing for slightly higher sparsity. In comparison, structured magnitude pruning already starts degrading sharply at 50% sparsity.

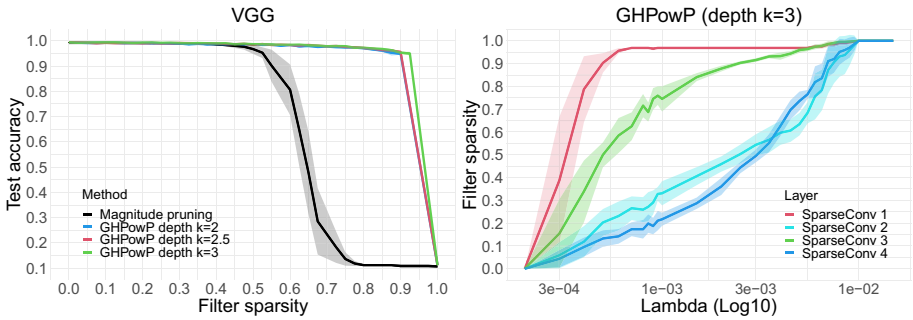


Fig. 9 Structured Sparsity using GHPowP. Left: regularization paths for (structured) filter sparsity using the GHPowP_k for $k \in \{2, 2.5, 3\}$ to overparametrize the filter weights of a small VGG architecture trained on MNIST. Structured magnitude pruning based on filter norms constitutes the baseline. Right: layer-wise sparsity patterns for the GHPowP_3. Error bars show standard errors over ten random initializations

9.5 Computational Complexity

An important question is how the overparametrization in our method affects the runtime complexity of DNN training using SGD. Since the networks are reduced to their base parametrization after training and sparse components are removed, the inference time complexity is reduced by the extent of the achieved sparsity. During training, the overparametrization increases both model size and computational complexity, which is heavily dependent on the architecture, hardware, and specific choice of $\mathcal{K}$.

To evaluate the impact of our approach, we train a fully-connected ReLU network with four hidden layers on the Fashion MNIST data set (Xiao et al., 2017). Figure 10 reports the mean wall-clock training time per sample for different batch sizes and factorization depths k of both the HPP_k (22) and its parameter-sharing counterpart the HPP_k^{shared} (32). The results show that the computational overhead increases sublinearly in k, but with diminishing effects for larger batches. For the HPP_k, training time is at worst roughly tripled for $k = 8$, whereas parameter sharing affords significant improvements over the full HPP_k: for batch sizes $\geqslant 64$, there is no discernible increase in training time for the tested depth levels. Details on architecture, hardware, and additional results for a ResNet are provided in Appendix B.4.

10 Summary and Discussion

In this work, we propose a general framework for smooth optimization of objectives that involve non-smooth and potentially non-convex sparse regularization of parameter subsets. Being model- and loss-agnostic, our approach is applicable to a wide range of scenarios. The key idea underlying our method is to find a smooth variational form of the non-smooth sparse regularizer. Applying a smooth parametrization map and a change of regularizers enables the construction of an equivalent smooth surrogate objective, eliminating the need for specialized optimization routines for non-smooth and non-convex problems. Moreover, our framework can be easily integrated into existing differentiable structures such as DNNs. Our general template is applied to the smooth optimization of a broad range of non-smooth ℓ_q and $\ell_{p,q}$ regularized optimization problems for (structured) sparsity. Numerical experi-

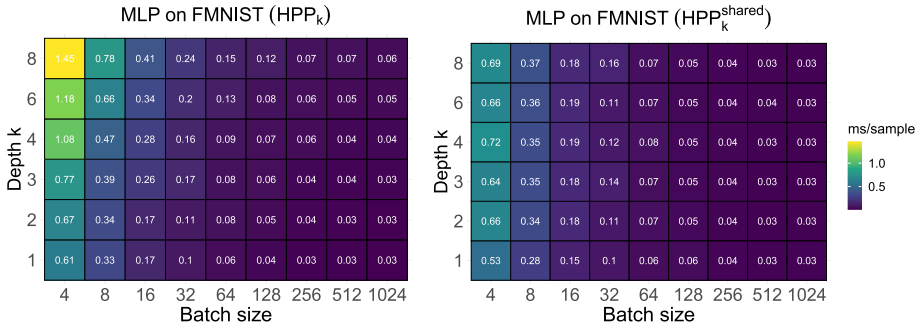


Fig. 10 Time per sample (training) for different factorization depths and batch sizes. Left: full overparametrization using the HPP_k. Right: parameter sharing significantly reduces computational overhead. Averages over four epochs are displayed

ments demonstrate the practical feasibility and effectiveness of our method in various sparse learning problems and in comparison with other methods.

Our approach also presents certain limitations that merit discussion. One limitation pertains to the initialization of the surrogate parameters, where an optimal choice is not straightforward. In addition, while our approach enables efficient optimization using SGD, obtaining (numerically) exact zeros is not guaranteed for small λ. This is a characteristic of SGD and not a limitation of the optimization transfer *per se*. For variable selection, we recommend a post-thresholding step. It is worth emphasizing that these challenges do not inherently limit the potential of our approach; instead, they underline key areas where additional research is needed.

There are several promising avenues for future research. Notably, our approach offers the flexibility to construct reparametrized, sparse "drop-in" replacements for network components, allowing for modular sparse regularization in differentiable network structures. This makes our method especially well-suited for exploring applications across various domains, such as input-sparse DNNs. Although a heuristic initialization performed well in our experiments, there is further great interest in understanding how to construct initialization schemes tailored to the surrogate parameters. There is also an opportunity to investigate the relationship between our smooth optimization transfer approach and implicit regularization methods in the DL literature. Our approach enforces a balanced parameter norm condition through surrogate ℓ_2 regularization, which bears similarities to the balanced weight conditions employed in implicit regularization techniques. Investigating this relationship could reveal valuable insights and potential synergies between the two approaches. Lastly, establishing theoretical conditions for the existence of differentiable parameterizations $\mathcal{K}$ and surrogate regularizers $\mathcal{R}_{\boldsymbol{\xi}}$ inducing a given $\mathcal{R}_{\boldsymbol{\beta}}$ constitutes an interesting direction for future research.

Appendix A Proofs of Lemmas and Theorems

A.1 Proof of Lemma 1

Proof Assume $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then $\exists\, \varepsilon > 0 : \forall (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon) : P(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}')$. Since $\mathcal{K}(\boldsymbol{\xi})$ is a continuous surjection, so is $\tilde{\mathcal{K}}(\boldsymbol{\psi}, \boldsymbol{\xi}) \triangleq (\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$. Pick any $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) \in \tilde{\mathcal{K}}^{-1}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \{\hat{\boldsymbol{\psi}}\} \times \{\hat{\boldsymbol{\xi}} : \mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}\}$. By continuity of $\tilde{\mathcal{K}}$, there $\exists \delta > 0 : \tilde{\mathcal{K}}(\mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)) \subseteq \mathcal{B}(\tilde{\mathcal{K}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon) = \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. This means $\forall\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta) : (\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}')) = (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. Since by assumption, $P(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}')$ for all $(\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$, and by continuity all $(\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)$ map to some $(\boldsymbol{\psi}', \boldsymbol{\beta}')$ in $\mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$ under $\tilde{\mathcal{K}}$, we conclude that

$$\forall\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta) : \mathcal{P}(\tilde{\mathcal{K}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}') = \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}'))\,.$$

Therefore, if $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then all $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ in the fiber $\tilde{\mathcal{K}}^{-1}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ are local minimizers of $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$ with equivalent local minima $\mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}}))$. □

A.2 Proof of Lemma 2

Proof Assume $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$, then $\exists\, \varepsilon > 0 : \forall (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon) : \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) \leqslant \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}'))$. Since $\mathcal{K}(\boldsymbol{\xi})$ is locally open at $\hat{\boldsymbol{\xi}}$, so is $\tilde{\mathcal{K}}(\boldsymbol{\psi}, \boldsymbol{\xi}) \triangleq (\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$ at $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$. By local openness, we can find $\delta > 0$ such that $\mathcal{B}(\tilde{\mathcal{K}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta) \subseteq \tilde{\mathcal{K}}(\mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon))$. Thus, $\forall\, (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}(\tilde{\mathcal{K}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)\ \exists\ (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon)$ such that $(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}')) = (\boldsymbol{\psi}', \boldsymbol{\beta}')$. But since we have by assumption that $\forall\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon) : \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}'))$, and we established $\forall (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta)\, \exists\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon) : (\boldsymbol{\psi}', \boldsymbol{\beta}') = (\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}'))$, it follows

$$\forall\, (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta) : \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) \leqslant \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}')) = \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}')\,.$$

Thus, $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}}))$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with corresponding local minimum $\mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}}))$. □

A.3 Proof of Lemma 3

Proof Assume $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then $\exists\, \varepsilon > 0 : \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}')$ $\forall\, (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. Since $\mathcal{K}(\boldsymbol{\xi})$ is a continuous surjection, so is $\tilde{\mathcal{K}}(\boldsymbol{\psi}, \boldsymbol{\xi}) \triangleq (\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$. By assumption of the variational form in Assumption 1, $\exists \hat{\boldsymbol{\xi}} \in \arg\min \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \subseteq \{\boldsymbol{\xi} : \mathcal{K}(\boldsymbol{\xi}) = \hat{\boldsymbol{\beta}}\}$ so that $\mathcal{R}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}) = \mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\hat{\boldsymbol{\xi}})) = \mathcal{R}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})$, and therefore also $\mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \mathcal{L}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = \mathcal{L}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) + \lambda \mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\hat{\boldsymbol{\xi}})) = \mathcal{L}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}) = \mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$.

By continuity of $\tilde{\mathcal{K}}$, there $\exists \delta > 0 : \tilde{\mathcal{K}}(\mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)) \subseteq \mathcal{B}(\tilde{\mathcal{K}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon) = \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. This means $\forall\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta) : (\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}')) = (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. Because $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, $\mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \leqslant \mathcal{P}(\boldsymbol{\psi}', \boldsymbol{\beta}')$ for all $(\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$, and by continuity of $\tilde{\mathcal{K}}$, all $(\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)$ map to some $(\boldsymbol{\psi}', \boldsymbol{\beta}')$ in $\mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \varepsilon)$. Then we can conclude $\mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) \leqslant \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}'))$ for all $(\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta)$. Lastly, using the majorization property of the surrogate penalty, $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \geqslant \mathcal{R}_{\boldsymbol{\beta}}(\mathcal{K}(\boldsymbol{\xi}))\, \forall\, \boldsymbol{\xi}$, we obtain the following chain of inequalities:

$$\forall\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \delta) : \ \mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}})) \leqslant \mathcal{P}(\boldsymbol{\psi}', \mathcal{K}(\boldsymbol{\xi}')) \leqslant \mathcal{Q}(\boldsymbol{\psi}', \boldsymbol{\xi}').$$

Thus, $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$. Therefore, if $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then all $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ such that $\hat{\boldsymbol{\xi}} \in \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ are local minimizers of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$ with $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. □

A.4 Proof of Lemma 4

Proof Assume $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$, then $\exists\, \varepsilon_0 > 0$ such that $\forall (\boldsymbol{\psi}', \boldsymbol{\xi}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon_0) : \mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) \leqslant \mathcal{Q}(\boldsymbol{\psi}', \boldsymbol{\xi}')$. First, we show that for each local minimizer $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ of $\mathcal{Q}$, letting $\mathcal{K}(\hat{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$, it must also hold that $\hat{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, i.e., $\hat{\boldsymbol{\xi}}$ is a minimizer of the SVF given $\hat{\boldsymbol{\beta}}$. Suppose for contradiction that $\hat{\boldsymbol{\xi}}$ is not a minimizer of $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ over the fiber $\mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$. Then $\forall\, \varepsilon > 0\, \exists\, \tilde{\boldsymbol{\xi}} \in \mathcal{B}(\hat{\boldsymbol{\xi}}, \varepsilon)$ such that $\mathcal{K}(\tilde{\boldsymbol{\xi}}) = \hat{\boldsymbol{\beta}}$ and $\mathcal{R}_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\xi}}) < \mathcal{R}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}})$. Let $\varepsilon = \varepsilon_0$. Because $\mathcal{K}(\tilde{\boldsymbol{\xi}}) = \mathcal{K}(\hat{\boldsymbol{\xi}})$, and the loss $\mathcal{L}(\boldsymbol{\psi}, \mathcal{K}(\boldsymbol{\xi}))$ is constant over all $\boldsymbol{\xi} \in \mathcal{K}^{-1}(\hat{\boldsymbol{\beta}})$, we have $\mathcal{L}(\hat{\boldsymbol{\psi}}, \mathcal{K}(\tilde{\boldsymbol{\xi}})) = \mathcal{L}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. But then $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) = \mathcal{L}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\xi}}) < \mathcal{L}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}) = \mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$, with $(\hat{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon_0)$, contra-

dicting local minimality of $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$. Thus if $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$, then $\hat{\boldsymbol{\xi}} \in \arg\min_{\boldsymbol{\xi}:\mathcal{K}(\boldsymbol{\xi})=\hat{\boldsymbol{\beta}}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, with $\mathcal{R}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}) = \mathcal{R}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})$, and so $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$.
Using this result, we now proceed to prove that $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ by contradiction. Suppose $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ is not a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$, then $\forall \delta > 0\ \exists (\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta) :\ \mathcal{P}(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) < \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. By Assumption 1, the set-valued solution map $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is lower hemicontinuous at $\hat{\boldsymbol{\beta}}$, and Lemma 18 extends this property to the identity-augmented product map $g : (\boldsymbol{\psi}, \boldsymbol{\beta}) \mapsto \{(\boldsymbol{\psi}, \boldsymbol{\xi}) : \boldsymbol{\xi} \in \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})\}$. Since we know $\hat{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}})$, it follows from lower hemicontinuity of $g(\boldsymbol{\psi}, \boldsymbol{\beta})$ at $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ that for all $\varepsilon > 0$ there is $\delta > 0$ so that:

$$\forall\ (\boldsymbol{\psi}', \boldsymbol{\beta}') \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta)\, \exists\, (\boldsymbol{\psi}', \boldsymbol{\xi}') \in g(\boldsymbol{\psi}', \boldsymbol{\beta}') \cap \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon).$$

Letting $\varepsilon = \varepsilon_0$ and $(\boldsymbol{\psi}', \boldsymbol{\beta}') = (\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}})$, this implies there is also $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon_0)$ with $\tilde{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\tilde{\boldsymbol{\beta}})$. As $\tilde{\boldsymbol{\xi}}$ is a minimizer of $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ over the fiber $\mathcal{K}^{-1}(\tilde{\boldsymbol{\beta}})$, we have $\mathcal{R}_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\xi}}) = \mathcal{R}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}})$ and thus $\mathcal{P}(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) = Q(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}})$. But then we have found $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon_0)$ such that $\mathcal{Q}(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) = \mathcal{P}(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) < \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = \mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$, contradicting that $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}$. This shows that if $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\xi})$, then $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\psi}}, \mathcal{K}(\hat{\boldsymbol{\xi}}))$ is a local minimizer of $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ with $\mathcal{Q}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \mathcal{P}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. □

Lemma 18 *(Lower hemicontinuity of product-augmented maps) Let $\hat{\boldsymbol{\xi}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^{d_\xi}$ be lower hemicontinuous at $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$. Augmenting $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ by the identity function $\mathrm{id}_{\boldsymbol{\psi}} : \mathbb{R}^{d_\psi} \to \mathbb{R}^{d_\psi}, \boldsymbol{\psi} \mapsto \boldsymbol{\psi}$, we can define the Cartesian product map $g : \mathbb{R}^{d_\psi} \times \mathbb{R}^d \rightrightarrows \mathbb{R}^{d_\psi} \times \mathbb{R}^{d_\xi}$, $(\boldsymbol{\psi}, \boldsymbol{\beta}) \mapsto g(\boldsymbol{\psi}, \boldsymbol{\beta}) \triangleq \{(\boldsymbol{\psi}, \boldsymbol{\xi}) : \boldsymbol{\xi} \in \hat{\boldsymbol{\xi}}(\boldsymbol{\beta})\}$. Then g is lower hemicontinuous at $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}) \in \mathbb{R}^{d_\psi} \times \mathbb{R}^d$ for any $\hat{\boldsymbol{\psi}} \in \mathbb{R}^{d_\psi}$.*

Proof Let $\varepsilon > 0$ and let $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) \in g(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$, i.e., $\hat{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}})$, be arbitrary. Since $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is lower hemicontinuous at $\hat{\boldsymbol{\beta}}$, there is $\delta_1 > 0$ such that for all $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\hat{\boldsymbol{\beta}}, \delta_1)$ and all $\boldsymbol{\xi}' \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}})$, there exist $\tilde{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\xi}} \in \mathcal{B}(\boldsymbol{\xi}', \varepsilon/\sqrt{2})$, in particular for $\hat{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}})$. Set $\delta \triangleq \min(\delta_1, \varepsilon/\sqrt{2})$. Let $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta)$. Then $\|\tilde{\boldsymbol{\psi}} - \hat{\boldsymbol{\psi}}\|_2^2 + \|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2^2 < \delta^2 \leqslant \frac{\varepsilon^2}{2}$, so in particular, $\|\tilde{\boldsymbol{\psi}} - \hat{\boldsymbol{\psi}}\|_2 < \delta \leqslant \varepsilon/\sqrt{2}$, and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\hat{\boldsymbol{\beta}}, \delta_1)$. Hence there exists $\tilde{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\tilde{\boldsymbol{\beta}})$ with $\|\tilde{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}\|_2 < \varepsilon/\sqrt{2}$. Combining these, we obtain

$$\|(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) - (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})\|_2^2 = \|\tilde{\boldsymbol{\psi}} - \hat{\boldsymbol{\psi}}\|_2^2 + \|\tilde{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}\|_2^2 < \frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2} = \varepsilon^2,$$

so that $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon)$. Since $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in g(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}})$, this shows that for all $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) \in \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}), \delta)$, there exists $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\xi}}) \in g(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\beta}}) \cap \mathcal{B}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}), \varepsilon)$. Since $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\xi}} \in \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}})$ were arbitrary, this shows the lower hemicontinuity of g at $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$. □

A.5 Proof of Lemma 6

Proof We prove lower hemicontinuity for each $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$, $j \in [L]$, separately, and then extend to the entire solution map $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$. Our argument proceeds by (i) deriving the minimal $\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j)$ over the fiber $\mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$ given our structural assumptions, (ii) analytically constructing solution magnitudes attaining this minimum, and (iii) verifying the solution map is lower hemicontinuous.

Fix $j \in [L]$ and consider $\boldsymbol{\beta}_j \in \mathbb{R}^{|\mathcal{G}_j|}$. As each of the k factors $\boldsymbol{\xi}_{jl}$ is either a scalar or a vector in $\mathbb{R}^{|\mathcal{G}_j|}$, let $S_j \subseteq [k]$ denote the subset of scalar indices, $V_j = [k] \setminus S_j$ the vector indices, and let the respective sum of exponents be $k_1 = \sum_{l \in V_j} \alpha_l$ and $k_2 = \sum_{l \in S_j} \alpha_l$. By the power-product assumption on $\mathcal{K}_j$, each entry $i \in \mathcal{G}_j$ can be written as

$$\mathcal{K}_{ji}(\boldsymbol{\xi}_j) = \underbrace{\prod_{l=1}^{k} \operatorname{sign}(\xi_{jl}^{(i)})\, |\xi_{jl}^{(i)}|^{\alpha_l}}_{\text{Assumption 2}} = \underbrace{\prod_{l \in V_j} \operatorname{sign}(\xi_{jli})\, |\xi_{jli}|^{\alpha_l}}_{\triangleq \beta_{ji}^V = \beta_{ji}/\beta_j^S} \cdot \underbrace{\prod_{l \in S_j} \operatorname{sign}(\xi_{jl})\, |\xi_{jl}|^{\alpha_l}}_{\triangleq \beta_j^S} = \beta_{ji},$$

where $\xi_{jl}^{(i)}$ equals the ith entry ξ_{jli} of $\boldsymbol{\xi}_{jl}$ for $l \in V_j$ and $\xi_{jl}^{(i)} = \xi_{jl}$ for scalars $l \in S_j$. The partial product β_{ji}^V written as the fraction β_{ji}/β_j^S is well-defined for $\boldsymbol{\beta}_j \neq \mathbf{0}$, as the product-structure implies $\xi_{jl} \neq 0 \,\forall\, l \in S_j$ and hence $\beta_j^S \neq 0$.

(i) Minimum of $\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j)$ over $\mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$. For $\boldsymbol{\beta}_j = \mathbf{0}$, the unique norm-minimizing solution $\hat{\boldsymbol{\xi}}_j(\mathbf{0})$ is the singleton $\{\mathbf{0}\}$, yielding $\mathcal{R}_{\boldsymbol{\xi}_j}(\mathbf{0}) = \mathbf{0}$. Next, consider $\boldsymbol{\beta}_j \neq \mathbf{0}$. We reorder the terms in $\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j)$ as follows:

$$\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j) = \sum_{l=1}^{k} \alpha_l \|\boldsymbol{\xi}_{jl}\|_2^2 = \sum_{l \in V_j} \alpha_l \|\boldsymbol{\xi}_{jl}\|_2^2 + \sum_{l \in S_j} \alpha_l \xi_{jl}^2 = \sum_{i \in \mathcal{G}_j} \sum_{l \in V_j} \alpha_l\, \xi_{jli}^2 + \sum_{l \in S_j} \alpha_l\, \xi_{jl}^2.$$

Using the weighted AM-GM inequality (Prop. 4) on the first term and inserting the parametrization constraints $\mathcal{K}_{ij}(\boldsymbol{\xi}_j) = \beta_{ji} = \beta_{ji}^V \cdot \beta_j^S$, the constrained minimum is

$$\sum_{i \in \mathcal{G}_j} \sum_{l \in V_j} \alpha_l\, \xi_{jli}^2 \geqslant \sum_{i \in \mathcal{G}_j} k_1 \Big(\prod_{l \in V_j} \big(\xi_{jli}^2\big)^{\alpha_l} \Big)^{1/k_1} = \sum_{i \in \mathcal{G}_j} k_1 \Big(\Big| \prod_{l \in V_j} \operatorname{sign}(\xi_{jli}) |\xi_{jli}|^{\alpha_l} \Big| \Big)^{2/k_1} = \sum_{i \in \mathcal{G}_j} k_1 |\beta_{ji}/\beta_j^S|^{2/k_1} = k_1 \|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{2/k_1},$$

Applying the same inequality to the partially minimized sum and inserting $\mathcal{K}_j$, it holds

$$\begin{aligned} k_1 \|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{2/k_1} + \textstyle\sum_{l \in S_j} \alpha_l\, \xi_{jl}^2 &\geq (k_1 + k_2) \big(\|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^2 \cdot \textstyle\prod_{l \in S_j} (\xi_{jl}^2)^{\alpha_l} \big)^{1/(k_1+k_2)} \\ &= (k_1 + k_2) \big(\|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1} |\beta_j^S| \big)^{2/(k_1+k_2)} = (k_1 + k_2) \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/(k_1+k_2)}. \end{aligned}$$

Combined, this yields the minimum value of $\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j)$ over $\mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$:

$$\mathcal{R}_{\boldsymbol{\xi}_j}(\boldsymbol{\xi}_j) = \sum_{l=1}^{k} \alpha_l\, \|\boldsymbol{\xi}_{jl}\|_2^2 \geqslant k_1 \|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{2/k_1} + \sum_{l \in S_j} \alpha_l\, \xi_{jl}^2 \geqslant (k_1 + k_2) \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/(k_1+k_2)},$$

with equality holding if and only if both AM-GM optimality conditions are met, i.e., $|\xi_{jli}| = |\xi_{jl'i}| = |\beta_{ji}/\beta_j^S|^{1/k_1}\ \forall l, l' \in V_j$ and all $i \in \mathcal{G}_j$, as well as $|\xi_{jl}| = |\xi_{jl'}| = \|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{1/k_1} = \|\boldsymbol{\beta}_j\|_{2/k_1}^{1/(k_1+k_2)}\ \forall\, l, l' \in S_j$, while $\mathcal{K}_j(\boldsymbol{\xi}_j) = \boldsymbol{\beta}_j$.

(ii) Construction of solutions. The AM–GM optimality conditions and the multiplicative structure of $\mathcal{K}_j$ determine the absolute values of all solution parameters. Applying correct sign configurations to $\boldsymbol{\xi}_j$ respecting $\operatorname{sign}(\boldsymbol{\beta}_j)$ under $\mathcal{K}_j$ then produces full solutions. For a direct construction for any $\boldsymbol{\beta}_j \neq \mathbf{0}$, define for $k_1, k_2 \geqslant 1$ the scale at which the scalar factors must balance at optimality, $T(\boldsymbol{\beta}_j) \triangleq \|\boldsymbol{\beta}_j\|_{2/k_1}^{1/(k_1+k_2)}$. Then the absolute values of the solution parameters are given by:

$$\hat{\boldsymbol{\xi}}_j^{abs} : \mathbb{R}^{|\mathcal{G}_j|} \to \mathbb{R}^{d_{\boldsymbol{\xi}_j}}, \ \boldsymbol{\beta}_j \mapsto |\boldsymbol{\xi}_j| = \begin{cases} |\xi_{jl}| = T(\boldsymbol{\beta}_j) & \text{for } |\xi_{jl}| \in \mathbb{R}, \, l \in S_j \\ |\boldsymbol{\xi}_{jl}| = T(\boldsymbol{\beta}_j)^{-k_2/k_1} |\boldsymbol{\beta}_j|^{\circ(1/k_1)} & \text{for } |\boldsymbol{\xi}_{jl}| \in \mathbb{R}^{|\mathcal{G}_j|}, \, l \in V_j \end{cases}$$

Since $\hat{\boldsymbol{\xi}}_j^{abs}$ is a composition of continuous functions on $\mathbb{R}^{|\mathcal{G}_j|}$ (norms, powers, multiplication), its continuity follows immediately. Note for vector-valued $|\boldsymbol{\xi}_{jl}|$, each ith entry is $T(\boldsymbol{\beta}_j)^{-k_2/k_1}|\beta_{ji}|^{1/k_1}$, $i \in \mathcal{G}_j$. To verify the balancedness optimality conditions, we have by construction $|\xi_{jli}| = |\xi_{jl'i}| = T(\boldsymbol{\beta}_j)^{-k_2/k_1}|\beta_{ji}|^{1/k_1} \, \forall \, l, l' \in V_j, i \in \mathcal{G}_j$, as well as $|\xi_{jl}| = T(\boldsymbol{\beta}_j) \, \forall \, l \in S_j$. To show that the ξ_{jli} balance at the optimal scale in the first AM-GM application, see

$$|\beta_{ji}/\beta_j^S|^{1/k_1} = |\beta_{ji}/T(\boldsymbol{\beta}_j)^{k_2}|^{1/k_1} = T(\boldsymbol{\beta}_j)^{-k_2/k_1}|\beta_{ji}|^{1/k_1} = |\xi_{jli}| \quad \forall \, l \in V_j \ i \in \mathcal{G}_j.$$

Finally, to confirm that the lower bound of the penalty for the factors in V_j, $k_1\|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{2/k_1}$, is balanced in the second AM-GM application, we use both the definition of $T(\boldsymbol{\beta}_j)$ and $|\beta_j^S| = T(\boldsymbol{\beta}_j)^{k_2}$ to obtain

$$\begin{aligned} \|\boldsymbol{\beta}_j/\beta_j^S\|_{2/k_1}^{1/k_1} &= T(\boldsymbol{\beta}_j)^{-k_2/k_1}\|\boldsymbol{\beta}_j\|_{2/k_1}^{1/k_1} = \|\boldsymbol{\beta}_j\|_{2/k_1}^{-k_2/(k_1(k_1+k_2))}\|\boldsymbol{\beta}_j\|_{2/k_1}^{1/k_1} = \|\boldsymbol{\beta}_j\|_{2/k_1}^{k_1/(k_1(k_1+k_2))} \\ &= \|\boldsymbol{\beta}_j\|_{2/k_1}^{1/(k_1+k_2)} = T(\boldsymbol{\beta}_j). \end{aligned}$$

To show feasibility of the constructed solutions, select any $\beta_{ji} = \mathcal{K}_{ji}(\boldsymbol{\xi}_j), i \in \mathcal{G}_j$. Then:

$$\begin{aligned} \mathcal{K}_{ji}(\hat{\xi}_{ji}^{abs}(\boldsymbol{\beta}_j)) &= \prod_{l \in S_j} T(\boldsymbol{\beta}_j)^{\alpha_l} \prod_{l \in V_j} (T(\boldsymbol{\beta}_j)^{-k_2/k_1}|\beta_{ji}|^{1/k_1})^{\alpha_l} = T(\boldsymbol{\beta}_j)^{k_2}\left(|\beta_{ji}|^{1/k_1}T(\boldsymbol{\beta}_j)^{-k_2/k_1}\right)^{k_1} \\ &= T(\boldsymbol{\beta}_j)^{k_2}(|\beta_{ji}|^{1/k_1})^{k_1}T(\boldsymbol{\beta}_j)^{-k_2} = |\beta_{ji}|. \end{aligned}$$

The non-uniqueness of the solution signs and the multiplicative parametrization result in sign-flip symmetries. This also implies that, in addition to balanced magnitudes, solutions also require suitable sign configurations to ensure $\mathcal{K}_j(\boldsymbol{\xi}_j) = \boldsymbol{\beta}_j$ and not a sign-permuted version.

(iii) Lower hemicontinuity of solution mapping. To establish lower hemicontinuity of $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$, our goal is to show: for all $\bar{\boldsymbol{\xi}}_j \in \hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$ and any $\varepsilon > 0$, there is $\delta > 0$ so that for all $\boldsymbol{\beta}_j' \in \mathcal{B}(\boldsymbol{\beta}_j, \delta)$ there exists $\boldsymbol{\xi}_j' \in \hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j') \cap \mathcal{B}(\bar{\boldsymbol{\xi}}_j, \varepsilon)$. The continuity of the $\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j)$ significantly simplifies the argument.

Fix an arbitrary $\boldsymbol{\beta}_j \neq \mathbf{0}$ and any solution $\bar{\boldsymbol{\xi}}_j \in \hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$. By our construction of the absolute values of the solutions, it follows that $|\bar{\xi}_{jl}| = T(\boldsymbol{\beta}_j)$ and $|\bar{\xi}_{jli}| = T(\boldsymbol{\beta}_j)^{-k_2/k_1}|\beta_{ji}|^{1/k_1}$ for all $i \in \mathcal{G}_j, l \in [k]$. For any $\boldsymbol{\beta}_j'$ near $\boldsymbol{\beta}_j$, we can use the construction $\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j')$ to obtain $|\boldsymbol{\xi}_j'|$. For a fixed $\bar{\boldsymbol{\xi}}_j \in \hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$, define the signed function $\hat{\boldsymbol{\xi}}_j^{sign}(\boldsymbol{\beta}_j') \triangleq \operatorname{sign}(\bar{\boldsymbol{\xi}}_j) \odot \hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j')$ by applying the signs of $\bar{\boldsymbol{\xi}}_j$ to the absolute values. Trivially, $\hat{\boldsymbol{\xi}}_j^{sign}(\boldsymbol{\beta}_j) = \bar{\boldsymbol{\xi}}_j$. Note that this is a slight abuse of notation: in the case of a zero entry $\beta_{ji} = 0$, we have $\bar{\xi}_{jli} = 0 \, \forall \, l \in V_j$, and the respective entries in $\operatorname{sign}(\bar{\boldsymbol{\xi}}_j)$ are undefined. In this case, we set those signs to any pattern that respects $\operatorname{sign}(\beta_{ji}')$ to ensure $\mathcal{K}_{ji}(\hat{\xi}_{ji}^{sign}(\boldsymbol{\beta}_j')) = \beta_{ji}'$ for all $\boldsymbol{\beta}_j'$ sufficiently close to $\boldsymbol{\beta}_j$. This is permissible for our argument because the sign of ξ_{jli}' becomes irrelevant

when measuring distance to $\bar{\xi}_{jli} = 0$. For simplicity, we can hence assume $\operatorname{sign}(\bar{\boldsymbol{\xi}}_j)$ to be well-defined. Because $\hat{\boldsymbol{\xi}}_j^{abs}$ is continuous at $\boldsymbol{\beta}_j$, $\forall \varepsilon > 0 \exists \delta > 0 : \forall \boldsymbol{\beta}' \in \mathcal{B}(\boldsymbol{\beta}_j, \delta)$: $\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}'_j) \in \mathcal{B}(\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j), \varepsilon)$. But then $\boldsymbol{\xi}'_j \triangleq \hat{\boldsymbol{\xi}}_j^{sign}(\boldsymbol{\beta}'_j) \in \mathcal{B}(\bar{\boldsymbol{\xi}}_j, \varepsilon)$ must hold, because

$$\begin{aligned}\|\bar{\boldsymbol{\xi}}_j - \boldsymbol{\xi}'_j\|_2 = \|\operatorname{sign}(\bar{\boldsymbol{\xi}}_j) \odot |\bar{\boldsymbol{\xi}}_j| - \operatorname{sign}(\bar{\boldsymbol{\xi}}_j) \odot |\boldsymbol{\xi}'_j|\|_2 &= \|\operatorname{sign}(\bar{\boldsymbol{\xi}}_j) \odot (\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j) - \hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}'_j))\|_2 \\ &= \|(\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j) - \hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}'_j)\|_2 < \varepsilon\end{aligned}$$

Therefore, $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$ is lower hemicontinuous at $\boldsymbol{\beta}_j \neq 0$. Now, for $\boldsymbol{\beta}_j = \mathbf{0}$ we have $\hat{\boldsymbol{\xi}}_j(\mathbf{0}) = \{\mathbf{0}\}$. Because we know $\hat{\boldsymbol{\xi}}_j^{abs}(\boldsymbol{\beta}_j)$ continuously approaches $\mathbf{0}$ as $\boldsymbol{\beta}_j \to \mathbf{0}$ we can infer that for all $\varepsilon > 0$ there is $\delta > 0$ such that for any $\boldsymbol{\beta}'_j \in \mathcal{B}(\mathbf{0}, \delta)$, *all* solutions in $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}'_j)$ are in $\mathcal{B}(\mathbf{0}, \varepsilon)$, and thus $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$ is also lower hemicontinuous at $\mathbf{0}$.

This shows $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$ is lower hemicontinuous on $\mathbb{R}^{|\mathcal{G}_j|}$. Since each $\hat{\boldsymbol{\xi}}_j(\boldsymbol{\beta}_j)$ is lower hemicontinuous on $\mathbb{R}^{|\mathcal{G}_j|}$ and the solution map is the Cartesian function product $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta}) = (\hat{\boldsymbol{\xi}}_1(\boldsymbol{\beta}_1), \ldots, \hat{\boldsymbol{\xi}}_L(\boldsymbol{\beta}_L))$, it follows that $\hat{\boldsymbol{\xi}}(\boldsymbol{\beta})$ is lower hemicontinuous at every $\boldsymbol{\beta} \in \mathbb{R}^d$. □

A.6 Proof of Lemma 7

Proof As in Lemma 5, we can proceed by finding the minimum element-wise. Since the constraint implies that the difference of two non-negative numbers equals β_j for $j = 1, \ldots, d$, we further differentiate by the sign of β_j. For $\beta_j = 0$, the constraint reduces to $\gamma_j^2 = \delta_j^2$, which provides a unique minimizer $(\hat{\gamma}_j, \hat{\delta}_j) = (0, 0)$, resulting in a minimum ℓ_2 regularization term of $0 = |\beta_j|$. For $\beta_j > 0$, the constraint gives us $\gamma_j^2 = \beta_j + \delta_j^2 \geqslant \beta_j \implies |\gamma_j| \geqslant \sqrt{|\beta_j|} = \sqrt{\beta_j}$. Thus, we consider $\gamma_j = \pm\sqrt{|\beta_j|}$ and $\delta_j = 0$. This choice trivially satisfies the constraint, and it is easy to see that any other pair (γ_j, δ_j) satisfying $\beta_j = \gamma_j^2 - \delta_j^2$ needs to have a strictly larger magnitude in both γ_j and δ_j, resulting in a larger sum of the squared 2-norms. Thus the minimizers for $\beta_j > 0$ are given by $(\hat{\gamma}_j, \hat{\delta}_j) = (\pm\sqrt{|\beta_j|}, 0)$, resulting in a minimum regularization term of $|\beta_j|$. For $\beta_j < 0$, an analogous argument holds: By the constraint $\gamma_j^2 - \delta_j^2 = \beta_j$ we have $\delta_j^2 = \gamma_j^2 - \beta_j \geqslant -\beta_j = |\beta_j| \implies |\delta_j| \geqslant \sqrt{|\beta_j|}$. Considering $\delta_j = \pm\sqrt{|\beta_j|}$ and $\gamma_j = 0$, we again observe that any other pair (γ_j, δ_j) satisfying the constraint has strictly larger magnitude in γ_j and δ_j, resulting in a larger ℓ_2 regularization term. Thus, the minimizers for $\beta_j < 0$ are given by $(\hat{\gamma}_j, \hat{\delta}_j) = (0, \pm\sqrt{|\beta_j|})$, yielding a minimum ℓ_2 penalty of $|\beta_j|$.

In all three cases, the minimum of $\gamma_j^2 + \delta_j^2$ subject to $\gamma_j^2 - \delta_j^2 = \beta_j$ is given by $|\beta_j|$. The proof is completed by iterating over $j = 1, \ldots, d$. □

A.7 Proof of Lemma 8

Proof Due to the separable structure of the parametrization, we can proceed by finding the minimizer for each summand $j \in [L]$. Using the AM-GM on $\|\boldsymbol{u}_j\|_2^2$ and ν_j^2,

$$\frac{\|\boldsymbol{u}_j\|_2^2 + \nu_j^2}{2} \geqslant \sqrt{\nu_j^2 \cdot \|\boldsymbol{u}_j\|_2^2} = \sqrt{(\nu_j \cdot \|\boldsymbol{u}_j\|_2)^2} = |\nu_j| \cdot \|\boldsymbol{u}_j\|_2 = \|\nu_j \boldsymbol{u}_j\|_2 = \|\boldsymbol{\beta}_j\|_2 \,,$$

where we used the absolute homogeneity of norms. The expression reduces to equality if and only if $\|\boldsymbol{u}_j\|_2^2 = \nu_j^2 = \|\boldsymbol{\beta}_j\|_2$. Iterating over all groups $j = 1, \ldots, L$ shows that the constrained minimum in (18) is indeed $2\|\boldsymbol{\beta}\|_{2,1}$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$. □

A.8 Proof of Lemma 9

Proof We show that $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^d$, $(\boldsymbol{u}, \boldsymbol{\nu}) \mapsto \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, is locally open at $(\boldsymbol{u}, \boldsymbol{\nu})$, with $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$, if the $(\boldsymbol{u}_j, \nu_j)$ are such that $\nu_j = 0$ implies $\|\boldsymbol{u}_j\|_2 = 0$ for all $j \in [L]$. Recall that $d = |\mathcal{G}_1| + \ldots + |\mathcal{G}_L|$. We proceed in two steps. First, we find the points of openness for the group-wise parametrizations $\mathcal{K}_j : \mathbb{R}^{|\mathcal{G}_j|} \times \mathbb{R} \to \mathbb{R}^{|\mathcal{G}_j|}$, $(\boldsymbol{u}_j, \nu_j) \mapsto \boldsymbol{u}_j \nu_j$. In a second step, we then show that local openness of $\mathcal{K}_j$ at $(\boldsymbol{u}_j, \nu_j)$ for $j \in [L]$ implies local openness of the GHPP

$$\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}) \triangleq \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu} = (\boldsymbol{u}_1 \nu_1, \ldots, \boldsymbol{u}_L \nu_L)^\top = (\mathcal{K}_1(\boldsymbol{u}_1, \nu_1), \ldots, \mathcal{K}_L(\boldsymbol{u}_L, \nu_L))^\top$$

at $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$. For the first step, we show that the $\mathcal{K}_j$ are open at all points $(\boldsymbol{u}_j, \nu_j) \in \mathbb{R}^{|\mathcal{G}_j|} \times \mathbb{R}$ except $(\boldsymbol{u}_j, \nu_j) \in (\mathbb{R}^{|\mathcal{G}_j|} \times \{0\}) \setminus \{(\boldsymbol{0}, 0)\}$. To do this, we use the following result on the local openness of matrix multiplication:

Proposition 2 (Prop. 1 in Nouiehed and Razaviyayn (2022), rephrased) Let $\mathcal{M} : \mathbb{R}^{m \times z} \times \mathbb{R}^{z \times n} \to \mathbb{R}^{m \times n}, (\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \boldsymbol{M}_1 \boldsymbol{M}_2$, denote the bilinear matrix multiplication mapping such that $z \geqslant \min\{m, n\}$. Then $\mathcal{M}$ is locally open at $(\boldsymbol{M}_1, \boldsymbol{M}_2)$ if and only if

$$\begin{aligned} &\exists \tilde{\boldsymbol{M}}_1 \in \mathbb{R}^{m \times z} : \tilde{\boldsymbol{M}}_1 \boldsymbol{M}_2 = \boldsymbol{0}_{m \times n} \wedge \tilde{\boldsymbol{M}}_1 + \boldsymbol{M}_1 \text{ is full row-rank or} \\ &\exists \tilde{\boldsymbol{M}}_2 \in \mathbb{R}^{z \times n} : \boldsymbol{M}_1 \tilde{\boldsymbol{M}}_2 = \boldsymbol{0}_{m \times n} \wedge \tilde{\boldsymbol{M}}_2 + \boldsymbol{M}_2 \text{ is full column-rank}\,. \end{aligned}$$

Letting $m = |\mathcal{G}_j| > 1$, $z = 1$ and $n = 1$, we can apply this result to the group-wise functions $\mathcal{K}_j$: $\mathcal{K}_j$ is open at $(\boldsymbol{0}, 0) \in \mathbb{R}^{|\mathcal{G}_j|} \times \mathbb{R}$ if $\exists \tilde{\nu}_j : \boldsymbol{0}\tilde{\nu}_j = \boldsymbol{0}$ and $0 + \tilde{\nu}_j$ has full column-rank, i.e., $\tilde{\nu}_j \neq 0$. This holds for all $\tilde{\nu}_j \neq 0$. Further, $\mathcal{K}_j$ is open at $(\boldsymbol{u}_j, \nu_j)$, with $\|\boldsymbol{u}_j\|_2 \geqslant 0$, $\nu_j \neq 0$, if $\exists \tilde{\nu}_j : \boldsymbol{u}_j \tilde{\nu}_j = \boldsymbol{0}$, and $\nu_j + \tilde{\nu}_j \neq 0$. This holds for $\tilde{\nu}_j = 0$. Finally, $\mathcal{K}_j$ were to be open at $(\boldsymbol{u}_j, 0)$ with $\|\boldsymbol{u}_j\|_2 > 0$, if either $\exists \tilde{\nu}_j : \boldsymbol{u}_j \tilde{\nu}_j = \boldsymbol{0}$ and $\nu_j + \tilde{\nu}_j \neq 0$, or $\exists \tilde{\boldsymbol{u}}_j : \tilde{\boldsymbol{u}}_j \nu_j = \boldsymbol{0}$ and $\boldsymbol{u}_j + \tilde{\boldsymbol{u}}_j$ has full row-rank. The first condition implies $\tilde{\nu}_j = 0$, but then $0 + \tilde{\nu}_j = 0$, contradicting $\nu_j + \tilde{\nu}_j \neq 0$. Also, there is no such $\tilde{\boldsymbol{u}}_j$ as in the second condition, since $\boldsymbol{u}_j + \tilde{\boldsymbol{u}}_j \in \mathbb{R}^{|\mathcal{G}_j| \times 1}$ can not be full row-rank for $|\mathcal{G}_j| > 1$. Therefore, we have shown that the $\mathcal{K}_j$ are locally open at all points in $\mathbb{R}^{|\mathcal{G}_j|} \times \mathbb{R}$ except $(\boldsymbol{u}_j, \nu_j) \in (\mathbb{R}^{|\mathcal{G}_j|} \times \{0\}) \setminus \{(\boldsymbol{0}, 0)\}$.
For the second step, let the Cartesian product of two Euclidean spaces be endowed with the norm $\|\|\cdot\|_2, \|\cdot\|_2\|_2$. We now show that if $\mathcal{K}_j$ is open at $(\boldsymbol{u}_j, \nu_j)$ for $j \in [L]$, then $\mathcal{K}$ is open at $(\boldsymbol{u}, \boldsymbol{\nu})$, i.e.,

$$\forall \varepsilon > 0 \, \exists \, \tilde{\delta} > 0 : \, \mathcal{B}(\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}), \tilde{\delta}) \subseteq \mathcal{K}(\mathcal{B}((\boldsymbol{u}, \boldsymbol{\nu}), \varepsilon))\,.$$

Let $\varepsilon > 0$ be arbitrary. Define $\varepsilon_j \triangleq \varepsilon/\sqrt{L}$. By the local openness of the $\mathcal{K}_j$ at $(\boldsymbol{u}_j, \nu_j)$, there are δ_j such that $\mathcal{B}(\mathcal{K}_j(\boldsymbol{u}_j, \nu_j), \delta_j) \subseteq \mathcal{K}_j(\mathcal{B}((\boldsymbol{u}_j, \nu_j), \varepsilon_j))$ for all $j \in [L]$. Let $\tilde{\delta} \triangleq \min_j\{\delta_j\}$ and let $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}), \tilde{\delta})$ be arbitrary. Writing $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_L)^\top$, we then have

$$\|\tilde{\boldsymbol{\beta}} - \mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu})\|_2^2 = \textstyle\sum_{j=1}^L \|\tilde{\boldsymbol{\beta}}_j - \mathcal{K}_j(\boldsymbol{u}_j, \nu_j)\|_2^2 < \tilde{\delta}^2 \,,$$

which implies $\|\tilde{\boldsymbol{\beta}}_j - \mathcal{K}_j(\boldsymbol{u}_j, \nu_j)\|_2 < \tilde{\delta} \leqslant \delta_j$. By local openness of the $\mathcal{K}_j$, there then exist $(\tilde{\boldsymbol{u}}_j, \tilde{\nu}_j)$ such that $\mathcal{K}_j(\tilde{\boldsymbol{u}}_j, \tilde{\nu}_j) = \tilde{\boldsymbol{\beta}}_j$, with $\|(\boldsymbol{u}_j, \nu_j) - (\tilde{\boldsymbol{u}}_j, \tilde{\nu}_j)\| = \|\|\boldsymbol{u}_j - \tilde{\boldsymbol{u}}_j\|_2, |\nu_j - \tilde{\nu}_j|\,\|_2 < \varepsilon_j = \varepsilon/\sqrt{L}$. Defining $\tilde{\boldsymbol{u}} = (\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_L)^\top$ and $\tilde{\boldsymbol{\nu}} = (\tilde{\nu}_1, \ldots, \tilde{\nu}_L)^\top$, we find

$$\begin{aligned}
\|(\boldsymbol{u}, \boldsymbol{\nu}) - (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\nu}})\|^2 = \|\,\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_2,\, \|\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}}\|_2\,\|_2^2 &= \textstyle\sum_{j=1}^L \|\boldsymbol{u}_j - \tilde{\boldsymbol{u}}_j\|_2^2 + \sum_{j=1}^L |\nu_j - \tilde{\nu}_j|^2 \\
&= \textstyle\sum_{j=1}^L \|\,\|\boldsymbol{u}_j - \tilde{\boldsymbol{u}}_j\|_2,\, |\nu_j - \tilde{\nu}_j|\,\|_2^2 \\
&< \textstyle\sum_{j=1}^L \left(\frac{\varepsilon}{\sqrt{L}}\right)^2 = \varepsilon^2 \,,
\end{aligned}$$

and thus $\|(\boldsymbol{u}, \boldsymbol{\nu}) - (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\nu}})\| = \|\,\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_2,\, \|\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}}\|_2\,\|_2 < \varepsilon$. By definition of $\mathcal{K}$, we have

$$\mathcal{K}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\nu}}) = (\mathcal{K}_1(\tilde{\boldsymbol{u}}_1, \tilde{\nu}_1), \ldots, \mathcal{K}_L(\tilde{\boldsymbol{u}}_L, \tilde{\nu}_L))^\top = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_L)^\top = \tilde{\boldsymbol{\beta}} \in \mathbb{R}^d \,.$$

Taking both results together, we obtain $\tilde{\boldsymbol{\beta}} \in \mathcal{K}(\mathcal{B}((\boldsymbol{u}, \boldsymbol{\nu}), \varepsilon))$. Because $\tilde{\boldsymbol{\beta}}$ was chosen without loss of generality, it follows that $\mathcal{B}(\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}), \tilde{\delta}) \subseteq \mathcal{K}(\mathcal{B}((\boldsymbol{u}, \boldsymbol{\nu}), \varepsilon))$. As $\varepsilon > 0$ was arbitrary, we have shown the second step, i.e., that local openness of $\mathcal{K}_j$ at $(\boldsymbol{u}_j, \nu_j)$ for all $j \in [L]$ implies local openness of $\mathcal{K}$ at $(\boldsymbol{u}, \boldsymbol{\nu})$, with $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$.

Combining both steps completes the proof, and it is shown that $\mathcal{K}$ is locally open at $(\boldsymbol{u}, \boldsymbol{\nu})$, if for all $(\boldsymbol{u}_j, \nu_j), j \in [L]$, it holds that ν_j is zero only if $\|\boldsymbol{u}_j\|_2 = 0$ as well. □

A.9 Derivation of Group Size-Adjusted GHPP

We can induce the group size-adjusted group lasso penalty $\mathcal{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \triangleq \sum_{j=1}^L \sqrt{|\mathcal{G}_j|}\,\|\boldsymbol{\beta}_j\|_2$ as a simple extension to the previous GHPP approach, by counting each entry in $\boldsymbol{v}_j$ as its own parameter for the surrogate regularization, instead of subsuming all entries of the Hadamard factor under the scalar parameter ν_j as in 4.1. In this setting, the surrogate ℓ_2 regularization term counts ν_j not once, but $|\mathcal{G}_j| \triangleq p_j$ times, and is written as follows: $\widetilde{\mathcal{R}_{\boldsymbol{\xi}}}(\boldsymbol{u}, \boldsymbol{\nu}) = \sum_{j=1}^L \left(\|\boldsymbol{u}_j\|_2^2 + p_j \nu_j^2\right)$. Applying the AM-GM inequality to $\|\boldsymbol{u}_j\|_2^2$ and $(\sqrt{p_j}\nu_j)^2$ for $j \in [L]$, it holds

$$\begin{aligned}
\textstyle\sum_{j=1}^L \left(\|\boldsymbol{u}_j\|_2^2 + (\sqrt{p_j}\nu_j)^2\right) \geqslant 2\textstyle\sum_{j=1}^L \sqrt{\|\boldsymbol{u}_j\|_2^2 (\sqrt{p_j}\nu_j)^2} &= 2\textstyle\sum_{j=1}^L \sqrt{\left(\|\boldsymbol{u}_j\|_2 (\sqrt{p_j}\nu_j)\right)^2} \\
&= 2\textstyle\sum_{j=1}^L \left|\|\boldsymbol{u}_j\|_2 \cdot (\sqrt{p_j}\nu_j)\right| = 2\textstyle\sum_{j=1}^L \sqrt{p_j} \cdot |\nu_j| \cdot \|\boldsymbol{u}_j\|_2 \\
&= 2\textstyle\sum_{j=1}^L \sqrt{p_j}\|\nu_j \boldsymbol{u}_j\|_2 = 2\textstyle\sum_{j=1}^L \sqrt{p_j}\|\boldsymbol{\beta}_j\|_2 \,,
\end{aligned}$$

with equality if and only if $\|\boldsymbol{u}_j\|_2^2 = (\sqrt{p_j}\nu_j)^2 = \sqrt{p_j}\|\boldsymbol{\beta}_j\|_2$. The constrained minimizers $\hat{\boldsymbol{u}}_j$ and $\hat{\nu_j}$ corresponding to some $\boldsymbol{\beta}_j$ are obtained as

$$\underset{\substack{(\boldsymbol{u}_j,\nu_j):\\ \boldsymbol{\beta}_j=\nu_j\boldsymbol{u}_j}}{\arg\,min}\ \|\boldsymbol{u}_j\|_2^2+(\sqrt{p_j}\nu_j)^2=\begin{cases}\pm\left(\frac{\boldsymbol{\beta}_j}{\sqrt{\|\boldsymbol{\beta}_j\|_2/\sqrt{p_j}}},\sqrt{\|\boldsymbol{\beta}_j\|_2/\sqrt{p_j}}\right)\|\boldsymbol{\beta}_j\|_2 & 0\\ (\mathbf{0},0)\|\boldsymbol{\beta}_j\|_2=0\end{cases}$$

for each $j\in[L]$. Using identical arguments as for the unadjusted GHPP in 4.1, we can construct the equivalent smooth surrogate $\mathcal{Q}$ in Eq. (21) for the non-smooth objective $\mathcal{P}$ regularized with the adjusted $\ell_{2,1}$ penalty in Eq. (20). Minimizing $\mathcal{Q}$ over $(\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{\nu})$ yields (local) solutions to $\mathcal{P}$ in (20), which can be reconstructed using $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\beta}})=(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{u}}\odot_{\mathcal{G}}\hat{\boldsymbol{\nu}})$ as defined above.

A.10 Proof of Lemma 10

Proof Applying the AM-GM inequality for each $j=1,\ldots,d$ to the squared parameters u_{jl}^2, $l=1,\ldots,k$, we obtain

$$\frac{u_{j1}^2+\ldots+u_{jk}^2}{k}\geqslant\sqrt[k]{(u_{j1}^2)\cdot\ldots\cdot(u_{jk}^2)}=\sqrt[k]{(u_{j1}\cdot\ldots\cdot u_{jk})^2}=\sqrt[k]{|\beta_j|^2}=|\beta_j|^{2/k}\ ,$$

with equality holding if and only if $u_{j1}^2=\ldots=u_{jk}^2=|\beta_j|^{2/k}$. Summing over all $j\in[d]$ then shows the result. □

A.11 Proof of Lemma 11

Proof To prove the global openness of the *k*-linear function $\mathcal{K}:\prod_{l=1}^k\mathbb{R}^d\to\mathbb{R}^d,(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k)\mapsto\bigodot_{l=1}^k\boldsymbol{u}_l=\boldsymbol{\beta}$, defining the HPP_k, we make use of an existing result for scalar-valued multilinear maps and then generalize it to the *d*-dimensional real-valued case.

Proposition 3 (Theorem 1.2 in Balcerzak et al. (2016), rephrased) Let $X_1,\ldots,X_k$ be normed spaces over the scalar field $\mathbb{K}\in\{\mathbb{R},\mathbb{C}\}$, and let *T* from $X_1\times\cdots\times X_k$ to $\mathbb{K}$ be a nontrivial *k*-linear functional. Then *T* is globally open.

Using this result, the global openness of the HPP_k for $d=1$ follows directly, or equivalently, for a single entry of the general *d*-dimensional HPP_k. We define the entry-wise parametrizations as $\mathcal{K}_j:\prod_{l=1}^k\mathbb{R}\to\mathbb{R},(u_{j1},\ldots,u_{jk})\mapsto\prod_{l=1}^k u_{jl}=\beta_j$ for $j\in[d]$, such that

$$\begin{aligned}&\mathcal{K}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k)=(\mathcal{K}_1(u_{11},\ldots,u_{1k}),\ldots,\mathcal{K}_d(u_{d1},\ldots,u_{dk}))^\top\\ \Longrightarrow\ &\bigodot_{l=1}^k\boldsymbol{u}_l=(\textstyle\prod_{l=1}^k u_{1l},\ldots,\prod_{l=1}^k u_{dl})^\top\ ,\end{aligned}$$

where $\boldsymbol{u}_l=(u_{1l},\ldots,u_{dl})^\top\in\mathbb{R}^d$ contains the parameters in each Hadamard factor $l\in[k]$. Let $\boldsymbol{u}_j=(u_{j1},\ldots,u_{jk})^\top\triangleq(u_{jl})_{l=1}^k\in\mathbb{R}^k$ collect the parameters of the entry-wise parametrizations $\mathcal{K}_j$ for $j\in[d]$. For clarity, we also use $(\boldsymbol{u}_l)_{l=1}^k$ to abbreviate $(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k)$, and

further endow the k-times Cartesian product of Euclidean spaces with the norm $\|(\boldsymbol{u}_l)_{l=1}^k\| \triangleq \| \|\boldsymbol{u}_1\|_2, \ldots, \|\boldsymbol{u}_k\|_2 \|_2$.
We now proceed to show that local openness of $\mathcal{K}_j$ at $\boldsymbol{u}_j = (u_{jl})_{l=1}^k$, i.e.,

$$\forall \varepsilon_j > 0 \exists \delta_j > 0 : \mathcal{B}(\mathcal{K}_j((u_{jl})_{l=1}^k), \delta_j) \subseteq \mathcal{K}_j(\mathcal{B}((u_{jl})_{l=1}^k, \varepsilon_j)) \quad \forall\, j \in [d],$$

implies local openness of $\mathcal{K}$ at $(\boldsymbol{u}_l)_{l=1}^k$, i.e.,

$$\forall \varepsilon > 0 \exists \tilde{\delta} > 0 : \ \mathcal{B}(\mathcal{K}((\boldsymbol{u}_l)_{l=1}^k), \tilde{\delta}) \subseteq \mathcal{K}(\mathcal{B}((\boldsymbol{u}_l)_{l=1}^k, \varepsilon)),$$

where each $\boldsymbol{u}_l$ is constructed as $\boldsymbol{u}_l = (u_{1l}, \ldots, u_{dl})^\top$ from the points of openness $\boldsymbol{u}_j = (u_{jl})_{l=1}^k$ of the $\mathcal{K}_j$. Let $\varepsilon > 0$ be arbitrary and define $\varepsilon_j \triangleq \varepsilon/\sqrt{d}$. By our assumption, there are δ_j such that Preposition (3) holds for each $j \in [d]$ with ε_j. Let $\tilde{\delta} \triangleq \min_j\{\delta_j\}$ and pick any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}(\mathcal{K}((\boldsymbol{u}_l)_{l=1}^k), \tilde{\delta})$. It then holds by definition

$$\|\tilde{\boldsymbol{\beta}} - \mathcal{K}((\boldsymbol{u}_l)_{l=1}^k)\|_2^2 = \textstyle\sum_{j=1}^d |\tilde{\beta}_j - \mathcal{K}_j((u_{jl})_{l=1}^k)|^2 < \tilde{\delta}^2 \leqslant \delta_j^2,$$

implying $\tilde{\beta}_j \in \mathcal{B}(\mathcal{K}_j((u_{jl})_{l=1}^k), \tilde{\delta}) \subseteq \mathcal{B}(\mathcal{K}_j((u_{jl})_{l=1}^k), \delta_j)$. By local openness of the $\mathcal{K}_j$, it follows that $\tilde{\beta}_j \in \mathcal{K}_j(\mathcal{B}((u_{jl})_{l=1}^k, \varepsilon_j))$. This means that $\forall\, j \in [d]$ we have

$$\exists (\tilde{u}_{jl})_{l=1}^k : \mathcal{K}_j((\tilde{u}_{jl})_{l=1}^k) = \tilde{\beta}_j \text{ and } \|(u_{jl})_{l=1}^k - (\tilde{u}_{jl})_{l=1}^k\|^2 = \| \, |u_{j1} - \tilde{u}_{j1}|, \ldots, |u_{jk} - \tilde{u}_{jk}| \, \|_2^2 < \varepsilon_j^2.$$

Collecting the $\tilde{u}_{jl}$ as $\tilde{\boldsymbol{u}}_l = (\tilde{u}_{1\,l}, \ldots, \tilde{u}_{dl})^\top$ for $l \in [k]$, and evaluating $\mathcal{K}$ at these arguments, we obtain

$$\mathcal{K}((\tilde{\boldsymbol{u}}_l)_{l=1}^k) = (\mathcal{K}_1((\tilde{u}_{1\,l})_{l=1}^k), \ldots, \mathcal{K}_d((\tilde{u}_{dl})_{l=1}^k))^\top = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)^\top = \tilde{\boldsymbol{\beta}} \in \mathbb{R}^d,$$

as well as

$$\begin{aligned}
\|(\boldsymbol{u}_l)_{l=1}^k - (\tilde{\boldsymbol{u}}_l)_{l=1}^k\|^2 &= \| \, \|\boldsymbol{u}_1 - \tilde{\boldsymbol{u}}_1\|_2, \ldots, \|\boldsymbol{u}_k - \tilde{\boldsymbol{u}}_k\|_2 \, \|_2^2 = \textstyle\sum_{l=1}^k \|\boldsymbol{u}_l - \tilde{\boldsymbol{u}}_l\|_2^2 \\
&= \textstyle\sum_{l=1}^k \sum_{j=1}^d |u_{jl} - \tilde{u}_{jl}|^2 = \sum_{j=1}^d \|(u_{jl})_{l=1}^k - (\tilde{u}_{jl})_{l=1}^k\|_2^2 \\
&< \textstyle\sum_{j=1}^d \varepsilon_j^2 = d\left(\frac{\varepsilon}{\sqrt{d}}\right)^2 = \varepsilon^2,
\end{aligned}$$

i.e., $\|(\boldsymbol{u}_l)_{l=1}^k - (\tilde{\boldsymbol{u}}_l)_{l=1}^k\| < \varepsilon$. Taking both findings together, it follows $\tilde{\boldsymbol{\beta}} \in \mathcal{K}(\mathcal{B}((\boldsymbol{u}_l)_{l=1}^k, \varepsilon))$. Because $\tilde{\boldsymbol{\beta}}$ was arbitrary, we have $\mathcal{B}(\mathcal{K}((\boldsymbol{u}_l)_{l=1}^k), \tilde{\delta}) \subseteq \mathcal{K}(\mathcal{B}((\boldsymbol{u}_l)_{l=1}^k, \varepsilon))$. Finally, because $\varepsilon > 0$ was arbitrary, local openness of $\mathcal{K}_j$ at $(u_{jl})_{l=1}^k$ for all $j = 1, \ldots, L$ implies local openness of $\mathcal{K}$ at $(\boldsymbol{u}_l)_{l=1}^k$. Since the $\mathcal{K}_j$ are globally open, it follows that $\mathcal{K}$ is also globally open, completing the proof. □

A.12 Proof of Lemma 12

Proof Using the AM-GM on the group-wise parameters $j \in [L]$, it holds

$$\begin{aligned}\sum_{j=1}^{L} \frac{\|\boldsymbol{u}_j\|_2^2 + \sum_{r=1}^{k-1} \nu_{jr}^2}{k} &\geqslant \sum_{j=1}^{L} (\|\boldsymbol{u}_j\|_2^2 \cdot \nu_{j2}^2 \cdot \ldots \cdot \nu_{jk}^2)^{1/k} = \sum_{j=1}^{L} \left(\sqrt{(\|\boldsymbol{u}_j\|_2 \cdot \nu_{j2} \cdot \ldots \cdot \nu_{jk})^2}\right)^{2/k} \\ &= \sum_{j=1}^{L} |\|\boldsymbol{u}_j\|_2 \cdot \nu_{j2} \cdot \ldots \cdot \nu_{jk}|^{2/k} = \sum_{j=1}^{L} (|\nu_{j2} \cdot \ldots \cdot \nu_{jk}| \cdot \|\boldsymbol{u}_j\|_2)^{2/k} \\ &= \sum_{j=1}^{L} \|\boldsymbol{u}_j \cdot \nu_{j2} \cdot \ldots \cdot \nu_{jk}\|_2^{2/k} = \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2^{2/k} = \|\boldsymbol{\beta}\|_{2,2/k}^{2/k}\end{aligned}$$

with equality if and only if $\|\boldsymbol{u}_j\|_2^2 = \nu_{j2}^2 = \ldots = \nu_{jk}^2 = \|\boldsymbol{\beta}_j\|_2^{2/k}$. The result follows. □

A.13 Proof of Corollary 6

Proof First, we note that local openness is preserved under composition. Given two maps $\mathcal{K}_1 : \mathcal{M} \to \mathcal{N}$ and $\mathcal{K}_2 : \mathcal{N} \to \mathcal{O}$ between (Cartesian products of) Euclidean spaces, if $\mathcal{K}_1$ is open at $m \in \mathcal{M}$, and $\mathcal{K}_2$ is open at $n = \mathcal{K}_1(m) \in \mathcal{N}$, then $\mathcal{K}_2 \circ \mathcal{K}_1$ is open at *m*.

To obtain points of local openness of $\mathcal{K}(\boldsymbol{u}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}_r^{\odot(k-1)}$, we utilize the preservation of local openness under composition by reducing $\mathcal{K}$ to a composition involving two parametrizations of which we already know the points of openness, the HPP_k, and the GHPP. Specifically, we express $\mathcal{K}$ as the composition $\mathcal{K} = \mathrm{GHPP} \circ \mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}}$, where $\mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}}$ is an auxiliary *globally* open map constructed in the following.

We have that $\mathcal{K}_{\boldsymbol{\nu}} : \prod_{r=1}^{k-1} \mathbb{R}^L \to \mathbb{R}^L$, $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}) \mapsto \boldsymbol{\nu}_r^{\odot(k-1)}$, is globally open due to Lemma 11, as it can be recognized to be the HPP_{k-1} for vectors in $\mathbb{R}^L$. Then, we can define the identity-augmented map

$$\mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}} : \mathbb{R}^d \times \prod_{r=1}^{k-1} \mathbb{R}^L \to \mathbb{R}^d \times \mathbb{R}^L, (\boldsymbol{u}, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_k) \mapsto (\boldsymbol{u}, \mathcal{K}_{\boldsymbol{\nu}}(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}))$$

that is simply the Cartesian product function of $\mathcal{K}_{\boldsymbol{\nu}}$ and the identity function $\mathrm{id}_{\boldsymbol{u}} : \mathbb{R}^d \to \mathbb{R}^d, \boldsymbol{u} \mapsto \boldsymbol{u}$, and maps inputs for the GHPP_k to the input domain of the GHPP by adding an independent extra entry $\boldsymbol{u}$ and multiplying the remaining $k-1$ inputs $\boldsymbol{\nu}_r \in \mathbb{R}^L$ element-wise to obtain a single $\boldsymbol{\nu} \in \mathbb{R}^L$, so the image of $(\boldsymbol{u}, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_k)$ under $\mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}}$ is $(\boldsymbol{u}, \boldsymbol{\nu}) \in \mathbb{R}^d \times \mathbb{R}^L$. Since local openness is trivially preserved for an identity-augmented Cartesian product map, and $\mathcal{K}_{\boldsymbol{\nu}}$ is globally open, $\mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}}$ is also globally open. Due to the composability property, $\mathcal{K}$ is thus locally open at $(\boldsymbol{u}, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_k)$ if the GHPP is locally open at $(\boldsymbol{u}, \mathcal{K}_{\boldsymbol{\nu}}(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}))$ by Lemma 9. □

Remark 3 (Points of Openness for the $\mathrm{GHPP}_{k_1,k_1+k_2}$) To establish preservation of local minima under smooth parametrization of $\boldsymbol{\beta}$ using $\mathrm{GHPP}_{k_1,k_1+k_2}$ (28) in general objectives $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta})$ using Lemma 2, we can make essentially the same line of arguments as in the previous result regarding the points of openness for the GHPP_k. First, we define nested parametrizations $\mathcal{K}_{\boldsymbol{u}}$ and $\mathcal{K}_{\boldsymbol{\nu}}$, both of which are globally open maps since they correspond to a HPP_k mapping with depths k_1 and k_2, respectively. These are combined in the globally open pre-composition $\mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}} \triangleq (\mathcal{K}_{\boldsymbol{u}}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_1}), \mathcal{K}_{\boldsymbol{\nu}}(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k_2}))$. This allows us to express $\mathcal{K}$ as the composition $\mathcal{K} = \mathrm{GHPP} \circ \mathcal{K}_{\boldsymbol{u},\boldsymbol{\nu}}$. By the preservation of local openness under composition, if points in the domain of $\mathcal{K}$ are such that the conditions for local openness of the GHPP in Lemma 9 apply for $\mathcal{K}_{\boldsymbol{u}}$ and $\mathcal{K}_{\boldsymbol{\nu}}$, then $\mathcal{K}$ is also locally open at that point.

A.14 Proof of Lemma 13

Proof This proof requires a simple weighted generalization of the AM-GM inequality:

Proposition 4 (Weighted AM-GM inequality) Let $n \in \mathbb{N}$, $x_1, \ldots, x_n$ non-negative real values, $w_1, \ldots, w_n$ non-negative real weights, and $w \triangleq \sum_{i=1}^{n} w_i$. Then

$$\frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w} \geqslant \sqrt[w]{x_1^{w_1} x_2^{w_2} \cdots x_n^{w_n}},$$

with equality holding if and only if $x_1 = \ldots = x_n$.
We proceed in two steps. First, the AM-GM inequality is applied to the k_1 squared parameters μ_{jti} present in the parametrization of a single scalar entry β_{ji} of $\boldsymbol{\beta}$, $i \in G_j$, for each $j = 1, \ldots, L$. From this, the minimum of $\sum_{t=1}^{k_1} \|\boldsymbol{\mu}_{jt}\|_2^2$ as a function of the auxiliary parameter $\boldsymbol{u}_j$ can be inferred. In the second step, the weighted AM-GM inequality is used to obtain the minimum of the overall regularization term:

$$\begin{aligned}
\sum_{j=1}^{L} \Big(\sum_{t=1}^{k_1} \|\boldsymbol{\mu}_{jt}\|_2^2 + \sum_{r=1}^{k_2} \nu_{jr}^2 \Big) &= \sum_{j=1}^{L} \big(\sum_{i \in \mathcal{G}_j} \sum_{t=1}^{k_1} \mu_{jti}^2 + \sum_{r=1}^{k_2} \nu_{jr}^2 \big) \overset{(i)}{\geqslant} \sum_{j=1}^{L} \Big(\sum_{i \in \mathcal{G}_j} k_1 \big(\prod_{t=1}^{k_1} \mu_{jti}^2 \big)^{1/k_1} + \sum_{r=1}^{k_2} \nu_{jr}^2 \Big) \\
&= \sum_{j=1}^{L} \left(k_1 \sum_{i \in \mathcal{G}_j} \big| \underbrace{\prod_{t=1}^{k_1} \mu_{jti}}_{u_{ji}} \big|^{2/k_1} + \sum_{r=1}^{k_2} \nu_{jr}^2 \right) = \sum_{j=1}^{L} \left(k_1 \|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1} + \sum_{r=1}^{k_2} \nu_{jr}^2 \right) \\
&\overset{(ii)}{\geqslant} \sum_{j=1}^{L} (k_1 + k_2) \left[\Big(\|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1} \Big)^{k_1} \cdot \prod_{r=1}^{k_2} \nu_{jr}^2 \right]^{1/(k_1+k_2)} = \sum_{j=1}^{L} k \big| \|\boldsymbol{u}_j\|_{2/k_1} \cdot \prod_{r=1}^{k_2} \nu_{jr} \big|^{2/k} \\
&= \sum_{j=1}^{L} k \big\| \boldsymbol{u}_j \cdot \prod_{r=1}^{k_2} \nu_{jr} \big\|_{2/k_1}^{2/k} = k \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/k} = k \|\boldsymbol{\beta}\|_{2/k_1, 2/k}^{2/k}
\end{aligned}$$

The first inequality (*i*) using the AM-GM inequality holds with equality if and only if $\mu_{jti}^2 = |u_{ji}|^{2/k_1} \; \forall \, t = 1, \ldots, k_1, \; i \in \mathcal{G}_j$ and $j = 1, \ldots, L$. The second inequality (*ii*) applies Proposition 4 and reduces to equality if and only if $\|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1} = \nu_{j1}^2 = \ldots = \nu_{jk_2}^2 = \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/k} \; \forall \, j = 1, \ldots, L$. □

A.15 Proof of Lemma 14

Proof We apply the AM-GM inequality to each summand $j = 1, \ldots, d$ of the surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$:

$$\frac{\|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2}{k} = \sum_{j=1}^{d} \frac{u_j^2 + (k-1)v_j^2}{k} \geqslant \sum_{j=1}^{d} \sqrt[k]{u_j^2 \prod_{l=1}^{k-1} v_j^2} = \sum_{j=1}^{d} \sqrt[k]{\left(u_j \cdot v_j^{k-1}\right)^2}$$
$$= \sum_{j=1}^{d} |\beta_j|^{2/k} = \|\boldsymbol{\beta}\|_{2/k}^{2/k} ,$$

with equality holding if and only if $u_j^2 = v_j^2 = |\beta_j|^{2/k}$ for all $j = 1, \ldots, d$. □

A.16 Parameter Sharing and Identical Initialization

In (S)GD, dynamics with shared Hadamard factors can be related to their fully overparametrized counterparts through identical initialization of the to-be-shared parameters. We define a differentiable surrogate $\mathcal{Q}$ based on the HPP_k, i.e., $\boldsymbol{\beta} = \boldsymbol{u}_l^{\odot k}$, with surrogate ℓ_2 regularization for $\mathcal{R}_{\boldsymbol{\xi}}$ and no additional unregularized parameters $\boldsymbol{\psi}$:

$$\mathcal{Q}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \mathcal{L}(\mathcal{K}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) .$$

Consider an updating scheme for the $\boldsymbol{u}_l$, given by $\boldsymbol{u}_l^{t+1} = \boldsymbol{u}_l^t - \alpha \nabla_{\boldsymbol{u}_l} \mathcal{Q}(\boldsymbol{u}_1^t, \ldots, \boldsymbol{u}_k^t)$, where α denotes the learning rate. Assume identical initialization for $k-1$ factors, i.e., $\boldsymbol{u}_1^0 = \tilde{\boldsymbol{u}}$ and $\boldsymbol{u}_2^0 = \ldots = \boldsymbol{u}_k^0 = \tilde{\boldsymbol{v}}$. Then we have for $l = 1, \ldots, k$,

$$\nabla_{\boldsymbol{u}_l} \mathcal{Q}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \left(\partial \boldsymbol{\beta} / \partial \boldsymbol{u}_l\right)^\top \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) + \lambda \nabla_{\boldsymbol{u}_l} \mathcal{R}_{\boldsymbol{\xi}}(\cdot) = \operatorname{diag}\left(\bigodot_{l' \in [k] \setminus \{l\}} \boldsymbol{u}_{l'}\right) \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) + 2\lambda \boldsymbol{u}_l,$$

where $\left(\partial \boldsymbol{\beta} / \partial \boldsymbol{u}_l\right)$ is a $d \times d$ matrix containing partial derivatives $(\partial \beta_i / \partial u_{jl})_{ij}$, $i, j \in [d]$. At initialization, the gradients of $\mathcal{Q}$ with respect to the $\boldsymbol{u}_l$ are given by $\nabla_{\boldsymbol{u}_1} \mathcal{Q}(\boldsymbol{u}_1^0, \ldots, \boldsymbol{u}_k^0) = \operatorname{diag}\left(\tilde{\boldsymbol{v}}^{k-1}\right) \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) + 2\lambda \tilde{\boldsymbol{u}}$ and $\nabla_{\boldsymbol{u}_l} \mathcal{Q}(\boldsymbol{u}_1^0, \ldots, \boldsymbol{u}_k^0) = \operatorname{diag}\left(\tilde{\boldsymbol{u}} \odot \tilde{\boldsymbol{v}}^{k-2}\right) \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) + 2\lambda \tilde{\boldsymbol{v}}$, where $l = 2, \ldots, k$. Note that the gradient is constant over the identically initialized factors. It thus follows from the updating rule that $\boldsymbol{u}_2^t = \ldots = \boldsymbol{u}_k^t \; \forall t \in \mathbb{N}$.

Compare this to the gradient of an alternative surrogate $\tilde{\mathcal{Q}}$ based on the shared parametrization $\tilde{\mathcal{K}}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v}^{k-1}$, with initialization $(\boldsymbol{u}^0, \boldsymbol{v}^0) = (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ and penalty $\tilde{\mathcal{R}_{\boldsymbol{\xi}}} = \|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{v}\|_2^2$. It is easy to see that $\nabla_{\boldsymbol{u}} \tilde{\mathcal{Q}}(\boldsymbol{u}, \boldsymbol{v}) = \nabla_{\boldsymbol{u}_1} \mathcal{Q}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$, and under identical initialization for $l = 2, \ldots, k$, we have $\nabla_{\boldsymbol{v}} \tilde{\mathcal{Q}}(\boldsymbol{u}, \boldsymbol{v}) = (k-1) \nabla_{\boldsymbol{u}_l} \mathcal{Q}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$. Therefore, updating $\boldsymbol{u}^{t+1} = \boldsymbol{u}^t - \alpha \nabla_{\boldsymbol{u}} \tilde{\mathcal{Q}}(\boldsymbol{u}^t, \boldsymbol{v}^t)$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t - \frac{\alpha}{k-1} \nabla_{\boldsymbol{v}} \tilde{\mathcal{Q}}(\boldsymbol{u}^t, \boldsymbol{v}^t)$, using a scaled learning rate $\frac{\alpha}{k-1}$ for $\boldsymbol{v}$, results in identical updates compared to running gradient descent on $\mathcal{Q}$ with identical initialization for the $k-1$ (shared) factors.

A.17 Proof of Lemma 15

Proof We apply the weighted AM-GM inequality to each summand $j \in [d]$ of $\mathcal{R}_{\boldsymbol{\xi}}$:

$$\frac{\|\boldsymbol{u}\|_2^2 + (k-1)\,\||\boldsymbol{v}|\|_2^2}{k} = \sum_{j=1}^{d} \frac{u_j^2 + (k-1)|v_j|^2}{k} \geqslant \sum_{j=1}^{d} \sqrt[k]{u_j^2\,(|v_j|^2)^{k-1}} = \sum_{j=1}^{d} \sqrt[k]{(u_j \cdot |v_j|^{k-1})^2}$$
$$= \sum_{j=1}^{d} \big(\underbrace{|u_j \cdot |v_j|^{k-1}|}_{=\beta_j}\big)^{2/k} = \sum_{j=1}^{d} |\beta_j|^{2/k} = \|\boldsymbol{\beta}\|_{2/k}^{2/k}\,,$$

with equality holding if and only if $u_j^2 = |v_j|^2 = |\beta_j|^{2/k}$ for all $j = 1, \ldots, d$. □

A.18 Proof of Lemma 16

Proof We again apply the weighted AM-GM inequality on the group level for each $j \in [L]$ of the surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$ and find

$$\frac{\|\boldsymbol{u}\|_2^2 + (k-1)\|\boldsymbol{\nu}\|_2^2}{k} = \textstyle\sum_{j=1}^{L} \frac{\|\boldsymbol{u}_j\|_2^2 + (k-1)|\nu_j|^2}{k} \geqslant \sum_{j=1}^{L} \Big(\|\boldsymbol{u}_j\|_2^2 \cdot \big(|\nu_j|^2\big)^{k-1}\Big)^{1/k}$$
$$= \textstyle\sum_{j=1}^{L} \big(\|\boldsymbol{u}_j\|_2 \cdot |\nu_j|^{k-1}\big)^{2/k} = \sum_{j=1}^{L} \||\nu_j|^{k-1} \cdot \boldsymbol{u}_j\|_2^{2/k} = \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2^{2/k}\,,$$

with equality holding if and only if $\|\boldsymbol{u}_j\|_2^2 = |\nu_j|^2 = \|\boldsymbol{\beta}_j\|_2^{2/k}\ \forall j = 1, \ldots, L$. □

A.19 Proof of Lemma 17

Proof Applying the weighted AM-GM inequality to the surrogate regularizer $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ on the group-level for each $j \in [L]$, we find

$$\frac{k_1\|\boldsymbol{\mu}\|_2^2 + k_2\|\boldsymbol{\nu}\|_2^2}{k} = \sum_{j=1}^{L} \frac{k_1\|\boldsymbol{\mu}_j\|_2^2 + k_2|\nu_j|^2}{k} = \sum_{j=1}^{L} \frac{k_1\|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1} + k_2|\nu_j|^2}{k}$$
$$\geqslant \sum_{j=1}^{L} \left(\left(\|\boldsymbol{u}_j\|_{2/k_1}^{2/k_1}\right)^{k_1} \cdot \big(|\nu_j|^2\big)^{k_2}\right)^{1/(k_1+k_2)} = \sum_{j=1}^{L} \left(\Big|\|\boldsymbol{u}_j\|_{2/k_1} \cdot |\nu_j|^{k_2}\Big|\right)^{2/k}$$
$$= \sum_{j=1}^{L} \|\boldsymbol{u}_j \cdot |\nu_j|^{k_2}\|_{2/k_1}^{2/k} = \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/k} = \|\boldsymbol{\beta}\|_{2/k_1,2/k}^{2/k}\,,$$

with equality holding if and only if $\|\boldsymbol{\mu}_j\|_2^2 = |\nu_j|^2 = \|\boldsymbol{\beta}_j\|_{2/k_1}^{2/k}\ \forall j = 1, \ldots, L$. □

A.20 Fibers and Structure of Product Parametrizations

The fibers or level sets $\mathcal{K}^{-1}(\boldsymbol{\beta})$ of the parametrizations considered in our work (Assumption 2) are well-behaved and exhibit regularity properties worth discussing. Let $\boldsymbol{\beta} \in \mathbb{R}^d$ be a parameter that can be partitioned into $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L)$ for $j \in [L]$, $L \leqslant d$, so that $\boldsymbol{\beta}_j \in \mathbb{R}^{|\mathcal{G}_j|}$ and $|\mathcal{G}_1| + \ldots + |\mathcal{G}_L| = d$. Further, let $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d$ be a $\mathcal{C}^1$-smooth surjective parametrization of $\boldsymbol{\beta}$. For product and power-type structures such as the HPP_k, GHPP, or the GHPowP_k, the following holds: All $\boldsymbol{\beta}_j \in \mathbb{R}^{|\mathcal{G}_j|} \setminus \{\mathbf{0}\}$ are *regular* values of $\mathcal{K}_j(\boldsymbol{\xi}_j)$ for each $j \in [L]$, i.e., the Jacobian $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j)$ has full row rank $|\mathcal{G}_j|$ for all $\boldsymbol{\xi}_j \in \mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$. In the following, we derive this for three exemplary parametrizations:

Example 1 *(Regularity of* HPP_k) As the canonical multiplicative parametrization, consider the HPP_k. The Jacobian $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j)$ has full row rank $|\mathcal{G}_j| = 1$ at all $\boldsymbol{\xi}_j \in \mathcal{K}_j^{-1}(\beta_j)$ with $\beta_j \neq 0$. This follows from $\beta_j = \prod_{l=1}^{k} \xi_{jl}$: each entry β_j depends only on the k factors $(\xi_{jl})_{l\in[k]}$, and its gradient is nonzero whenever $\beta_j \neq 0$, implying $\xi_{jl} \neq 0 \,\forall\, l \in [k]$. The partial derivatives $\partial\beta_j/\partial\xi_{jl}$ are themselves non-zero products of the remaining factors, $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j) = [\prod_{l\neq 1} \xi_{jl}, \ldots, \prod_{l\neq k} \xi_{jl}] \in \mathbb{R}^{1\times k}$. Hence, the Jacobian of each $\mathcal{K}_j$ for the HPP_k has full row-rank for all $\boldsymbol{\xi}_j \notin \mathcal{K}_j^{-1}(0)$.

Example 2 (*Regularity of* GHPP) The GHPP parametrizes $\boldsymbol{\beta}$ as $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{u} \odot_{\mathcal{G}} \boldsymbol{\nu}$, or groupwise $\mathcal{K}_j(\boldsymbol{\xi}_j) = \boldsymbol{u}_j \cdot \nu_j$ with $\boldsymbol{u}_j \in \mathbb{R}^{|\mathcal{G}_j|}$ and $\nu_j \in \mathbb{R}$. The Jacobian $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j) \in \mathbb{R}^{|\mathcal{G}_j|\times(|\mathcal{G}_j|+1)}$ with respect to $\boldsymbol{\xi}_j = (\boldsymbol{u}_j, \nu_j)$ is given by

$$\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j) = \begin{bmatrix} \nu_j & 0 & \cdots & 0 & u_{j1} \\ 0 & \nu_j & \cdots & 0 & u_{j2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \nu_j & u_{j|\mathcal{G}_j|} \end{bmatrix}.$$

$\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j)$ has full row rank $|\mathcal{G}_j|$ whenever $\nu_j \neq 0$, since the diagonal entries ν_j are nonzero and span the row space. Thus, all $\boldsymbol{\beta}_j \neq \mathbf{0}$ are regular values of $\mathcal{K}_j$, as $\boldsymbol{\beta}_j = \boldsymbol{u}_j \cdot \nu_j$ implies $\nu_j \neq 0$ whenever $\boldsymbol{\beta}_j \neq \mathbf{0}$. Note that the diagonal part is always of full rank as long as $\nu_j \neq 0$, even if some entries $u_{ji} = 0$.

Example 3 *(Regularity of* GHPP_k) The GHPowP_k parametrizes $\boldsymbol{\beta}$ as $\mathcal{K}(\boldsymbol{\xi}) = \boldsymbol{u} \odot_{\mathcal{G}} |\boldsymbol{\nu}|^{\circ(k-1)}$, or groupwise $\mathcal{K}_j(\boldsymbol{\xi}_j) = \boldsymbol{u}_j \cdot |\nu_j|^{k-1}$, $k > 2, \boldsymbol{u}_j \in \mathbb{R}^{|\mathcal{G}_j|}, \nu_j \in \mathbb{R}$. The Jacobian $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j) \in \mathbb{R}^{|\mathcal{G}_j|\times(|\mathcal{G}_j|+1)}$ with respect to $\boldsymbol{\xi}_j = (\boldsymbol{u}_j, \nu_j)$ is given by

$$\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j) = \begin{bmatrix} |\nu_j|^{k-1} & & & u_{j1}(k-1)\,\mathrm{sign}(\nu_j)|\nu_j|^{k-2} \\ & |\nu_j|^{k-1} & & u_{j2}(k-1)\,\mathrm{sign}(\nu_j)|\nu_j|^{k-2} \\ & & \ddots & \vdots \\ & & |\nu_j|^{k-1} & u_{j|\mathcal{G}_j|}(k-1)\,\mathrm{sign}(\nu_j)|\nu_j|^{k-2} \end{bmatrix}.$$

Hence $\mathcal{J}_{\mathcal{K}_j}(\boldsymbol{\xi}_j)$ has row rank $|\mathcal{G}_j|$ as long as $\nu_j \neq 0$, similar to the GHPP, and all $\boldsymbol{\beta}_j \neq \mathbf{0}$ are regular values of $\mathcal{K}_j$.

Further, the $\mathcal{K}_j(\boldsymbol{\xi}_j)$ considered in our work are positively homogeneous of degree $k \geqslant 2$, i.e., $\mathcal{K}_j(c\,\boldsymbol{\xi}_j) = c^k\,\mathcal{K}_j(\boldsymbol{\xi}_j)$ for all $\boldsymbol{\xi}_j$ and $c > 0$. In the context of the fibers of $\mathcal{K}_j$, the positive homogeneity implies scale-invariance of the shape of each individual fiber at the regular values $\boldsymbol{\beta}_j$. That is, a fiber associated with a regular value $\boldsymbol{\beta}_j$ has the same "shape" as the fiber of any scaled version of $\boldsymbol{\beta}_j$. Moreover, continuity of $\mathcal{K}_j$ ensures that the fibers $\mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$ are closed sets. However, they are not compact in the case of overparametrization, since the product structure of the $\mathcal{K}_j$ implies unbounded fibers that extend to infinity. The following well-known result in differential geometry can be used to characterize the fibers of $\mathcal{K}_j$ at regular values, which is a consequence of the inverse function theorem and is also known as the preimage theorem e.g., Theorem 3.2 in, (Hirsch, 2012):

Proposition 5 (Regular value Theorem) Let $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d$, $\boldsymbol{\xi} \mapsto \boldsymbol{\beta}$, be a $\mathcal{C}^r, r \geqslant 1$, function and let $\boldsymbol{\beta} \in \mathbb{R}^d$ be a regular value of $\mathcal{K}$. Then the fiber $\mathcal{K}^{-1}(\boldsymbol{\beta})$ is a $(d_\xi - d)$-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^{d_\xi}$.

Thus, the fiber of $\mathcal{K}_j$ at a regular value $\boldsymbol{\beta}_j$ forms a $(d_{\xi_j} - |\mathcal{G}_j|)$-dimensional smooth manifold. For product-like parametrizations $\mathcal{K}$ with sign-flip symmetries, this manifold consists of disjoint connected components, each embedded within an orthant of suitable sign configuration. For any non-zero $\boldsymbol{\beta}_j$, each connected component of the fiber $\mathcal{K}_j^{-1}(\boldsymbol{\beta}_j)$ contains a unique minimal-norm point. By the AM-GM inequality, this point is attained when the Euclidean norms of the factors are balanced. Figure 2b visualizes the disjoint components (branches of the hyperbola) and their minimal-norm points for the scalar HPP.
As we traverse the fiber away from the minimal-norm points, i.e., the more unbalanced the factorization of $\boldsymbol{\beta}_j$ becomes, the fiber exhibits increasing curvature. Figure 18c illustrates such a fiber for a scalar-valued HPP_k parametrization of depth $k = 3$ (22).
In addition, we can prove the local connectedness of the fibers of $\mathcal{K}$ even at the non-regular value $\mathbf{0}$, where the Jacobian degenerates to the null matrix for product-type parametrizations $\mathcal{K}$:

Lemma 19 *(Local Connectedness of fibers of $\mathcal{K}$ for product-type parametrizations) Under Assumption 2 and assuming all $\boldsymbol{\beta}_j \neq \mathbf{0}$ are regular values of the $\mathcal{K}_j, j \in [L]$, the fibers of $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d$ are (locally) connected sets at every $\boldsymbol{\beta} \in \mathbb{R}^d$.*

Proof By Assumption 2, $\mathcal{K}(\boldsymbol{\xi})$ is block-separable into parametrizations $\mathcal{K}_j(\boldsymbol{\xi}_j)$ and further assume the $|\mathcal{G}_j| \times d_{\xi_j}$-dimensional Jacobian of each $\mathcal{K}_j$ has full row rank for all $\boldsymbol{\xi}_j \notin \mathcal{K}_j^{-1}(\mathbf{0})$. Therefore, the fibers of $\mathcal{K}_j$ at non-zero $\boldsymbol{\beta}_j \in \mathbb{R}^{|\mathcal{G}_j|}$ are $\mathcal{C}^r$ manifolds by Proposition 5, and thus locally connected sets whenever $\mathcal{K}_j$ is regular. Further, by the product-power structure assumed for $\mathcal{K}_j$, the parametrization $\mathcal{K}_j(\boldsymbol{\xi}_j)$ maps to the non-regular value $\boldsymbol{\beta}_j = \mathbf{0}$ if at least one of its factors is zero. Thus, the fiber $\mathcal{K}_j^{-1}(\mathbf{0})$ contains all $\boldsymbol{\xi}_{j1}, \ldots, \boldsymbol{\xi}_{jk}$ such that at least one $\boldsymbol{\xi}_{jl} = \mathbf{0}$. This fiber, while not a manifold, is thus a connected set, since all of the contained hyperplanes intersect at $\boldsymbol{\xi}_{j1} = \ldots = \boldsymbol{\xi}_{jk} = \mathbf{0}$.

Considering the separable Cartesian product structure of $\mathcal{K}$, the fiber of $\mathcal{K}$ at $\boldsymbol{\beta}$ is a Cartesian product of the fibers of $\mathcal{K}_j$ at their respective values $\boldsymbol{\beta}_j$. Each of these fibers is either locally connected (for regular values of $\boldsymbol{\beta}_j$) or connected (for $\boldsymbol{\beta}_j = \mathbf{0}$), and thus the fiber of $\mathcal{K}$ at $\boldsymbol{\beta}$ is also locally connected. □

A.21 Differentiable SCAD, MCP, and TL1 via the HPP

For the popular non-convex regularizers SCAD, MCP, and the transformed ℓ_1 (TL1) penalty, we can derive smooth surrogates based on the HPP and a differentiable surrogate penalty $\mathcal{R}_{\boldsymbol{\xi}}$, replacing absolute value terms in $\mathcal{R}_{\boldsymbol{\beta}}$ by their variational quadratic formulation. All three regularizers are defined as separable functions of the entry-wise absolute values, to each of which we can apply the surrogate $(u_j^2 + v_j^2)/2 \geqslant |\beta_j|$. In the following, we derive only smooth variational forms of those regularizers; the equivalence of the resulting overparametrized optimization problems follows from the fact that the solution

map $\hat{\boldsymbol{\xi}}$ coincides with that for the HPP with induced ℓ_1 regularization. Thus, Theorem 1 can be applied directly.

TL1 penalty The transformed ℓ_1 penalty is a non-convex regularizer defined as

$$\mathcal{R}_{\boldsymbol{\beta}}^{TL1}(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{d} \frac{(a+1)|\beta_j|}{a+|\beta_j|}, \quad a > 0,$$

where the hyperparameter a steers the degree of non-convexity and interpolates between the ℓ_0 penalty (for $a \to 0$) and the ℓ_1 penalty (for $a \to \infty$). The following result provides an SVF of $\mathcal{R}_{\boldsymbol{\beta}}^{TL1}$ using the HPP, thereby enabling the construction of differentiable equivalent surrogate objectives.

Lemma 20 *(SVF for the TL1 penalty) Given the HPP $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}$, the minimum of the surrogate penalty*

$$\mathcal{R}_{\boldsymbol{\xi}}^{TL1}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{d} \frac{(a+1)(u_j^2+v_j^2)}{2a+(u_j^2+v_j^2)}, \quad a > 0,$$

subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\beta}$ constitutes the following SVF for the TL1 penalty $\mathcal{R}_{\boldsymbol{\beta}}^{TL1}(\boldsymbol{\beta})$:

$$\min_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d:\boldsymbol{u}\odot\boldsymbol{v}=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}^{TL1}(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{R}_{\boldsymbol{\beta}}^{TL1}(\boldsymbol{\beta})$$

Proof Since both $\mathcal{R}_{\boldsymbol{\beta}}^{TL1}$ and $\mathcal{R}_{\boldsymbol{\xi}}^{TL1}$ are coordinate-wise separable, it suffices to show the equality for a generic summand $j \in [d]$, and the result immediately follows for all summands. First, note that the scalar function $h(t) = \frac{(a+1)t}{2a+t}$ is strictly increasing on $[0, \infty)$ for all $a > 0$. Therefore, minimizing the corresponding summand of $\mathcal{R}_{\boldsymbol{\xi}}^{TL1}$ subject to $u_j v_j = \beta_j$ is equivalent to minimizing $(u_j^2 + v_j^2)$ subject to the same constraint. Using the AM-GM inequality, we obtain a constrained minimum of $2|\beta_j|$ attained if and only if $|u_j| = |v_j| = \sqrt{|\beta_j|}$ and $\operatorname{sign}(u_j v_j) = \operatorname{sign}(\beta_j)$. Thus,

$$\min_{u_j,v_j:u_jv_j=\beta_j} \frac{(a+1)(u_j^2+v_j^2)}{2a+(u_j^2+v_j^2)} = \frac{(a+1)2|\beta_j|}{2a+2|\beta_j|} = \frac{(a+1)|\beta_j|}{a+|\beta_j|}.$$

Summing over $j \in [d]$, we obtain the SVF

$$\min_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d:\boldsymbol{u}\odot\boldsymbol{v}=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}^{TL1}(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{R}_{\boldsymbol{\beta}}^{TL1}(\boldsymbol{\beta}).$$

□

It follows that TL1 regularized objectives of the form $\mathcal{P}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \lambda \mathcal{R}_{\boldsymbol{\beta}}^{TL1}(\boldsymbol{\beta})$ are equivalent to the smooth surrogate $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{u} \odot \boldsymbol{v}) + \lambda \mathcal{R}_{\boldsymbol{\xi}}^{TL1}(\boldsymbol{u}, \boldsymbol{v})$.

Minimax Concave Penalty The MCP is a non-convex separable regularizer with a piecewise definition, relaxing the penalization rate to 0 away from the origin. In its non-smooth formulation, it is given by

$$\mathcal{R}_{\boldsymbol{\beta}}^{MCP}(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{d} \rho_{\lambda,\gamma}^{MCP}(\beta_j), \quad \rho_{\lambda,\gamma}^{MCP}(\beta_j) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}, & |\beta_j| \leqslant \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & |\beta_j| > \gamma\lambda, \end{cases}$$

where $\gamma > 1$ controls the degree of non-convexity. A fully differentiable surrogate of $\mathcal{R}_{\boldsymbol{\beta}}^{MCP}$ can be obtained by replacing $|\beta_j|$ by $(u_j^2 + v_j^2)/2$ and β_j^2 by $(u_j^2 + v_j^2)^2/4$. This yields the following result:

Lemma 21 *(SVF for the MCP) Given $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}$ (HPP), the minimum of*

$$\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \sum_{j=1}^{d} \tilde{\rho}_{\lambda,\gamma}^{MCP}(u_j, v_j), \quad \tilde{\rho}_{\lambda,\gamma}^{MCP}(u_j, v_j) \triangleq \begin{cases} \lambda(u_j^2 + v_j^2)/2 - \dfrac{((u_j^2 + v_j^2)/2)^2}{2\gamma}, & (u_j^2 + v_j^2)/2 \leqslant \gamma\lambda, \\ \dfrac{1}{2}\gamma\lambda^2, & (u_j^2 + v_j^2)/2 > \gamma\lambda. \end{cases}$$

subject to $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\beta}$ forms an SVF for the MCP regularizer:

$$\min_{\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d : \boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{R}_{\boldsymbol{\beta}}^{MCP}(\boldsymbol{\beta})$$

Proof Fix $j \in [d]$ and set $s_j \triangleq (u_j^2 + v_j^2)/2$. On the first, non-constant, branch of $\tilde{\rho}_{\lambda,\gamma}^{MCP}$,

$$\tilde{\rho}_{\lambda,\gamma}^{MCP}(u_j, v_j) = \lambda s_j - \frac{s_j^2}{2\gamma} = h(s_j), \qquad h(t) \triangleq \lambda t - \frac{t^2}{2\gamma},$$

and $h'(t) = \lambda - t/\gamma > 0$ for $t \in [0, \gamma\lambda]$, so h is strictly increasing there. Thus, whenever the first branch is feasible under $u_j v_j = \beta_j$, minimizing $\tilde{\rho}_{\lambda,\gamma}^{MCP}$ reduces to minimizing s_j on the fiber.

By AM-GM, for all (u_j, v_j) with $u_j v_j = \beta_j$, $s_j = (u_j^2 + v_j^2)/2 \geq |u_j v_j| = |\beta_j|$, and equality is attained for $|u_j| = |v_j|$ and correct signs for β_j.
If $|\beta_j| > \gamma\lambda$, then $s_j \geq |\beta_j| > \gamma\lambda$ for all feasible (u_j, v_j), so only the second branch of $\tilde{\rho}_{\lambda,\gamma}^{MCP}$ applies and $\min_{u_j v_j = \beta_j} \tilde{\rho}_{\lambda,\gamma}^{MCP}(u_j, v_j) = \frac{1}{2}\gamma\lambda^2$.

If $|\beta_j| \leq \gamma\lambda$, the constrained minimizer of $s_j = (u_j^2 + v_j^2)/2$ over $u_j v_j = \beta_j$ satisfies $s_j = |\beta_j| \leqslant \gamma\lambda$, so the first branch is feasible and, by monotonicity of h on $[0, \gamma\lambda]$,

$$\min_{u_j v_j = \beta_j} \tilde{\rho}_{\lambda,\gamma}^{MCP}(u_j, v_j) = h\big(\min_{u_j v_j = \beta_j} s_j\big) = \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}.$$

Moreover, $\frac{1}{2}\gamma\lambda^2 - \big(\lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}\big) \geq 0$, so the constant branch cannot improve upon this value. Therefore, the entry-wise minima coincide with $\rho_{\lambda,\gamma}^{MCP}(\beta_j)$, and summing over $j \in [d]$ gives the desired SVF

$$\min_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d:\,\boldsymbol{u}\odot\boldsymbol{v}=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{u},\boldsymbol{v}) = \mathcal{R}_{\boldsymbol{\beta}}^{MCP}(\boldsymbol{\beta}).$$

□

Continuity at the boundary $s_j = \gamma\lambda$ follows from $\lambda(\gamma\lambda) - \frac{(\gamma\lambda)^2}{2\gamma} = \frac{1}{2}\gamma\lambda^2$, which coincides with the second branch. Differentiability of $\tilde{\rho}_{\lambda,\gamma}^{MCP}$ with respect to (u_j, v_j) at the boundary $s_j = \gamma\lambda$ follows by observing that the gradient of the first branch is $\nabla_{(u_j,v_j)}\big(\lambda s_j - \frac{s_j^2}{2\gamma}\big) = \big(\lambda - \frac{s_j}{\gamma}\big)\nabla_{(u_j,v_j)} s_j = \big(\lambda - \frac{s_j}{\gamma}\big)(u_j, v_j)$, which vanishes on $s_j = \gamma\lambda$, matching the zero gradient of the constant second branch.

Smoothly Clipped Absolute Deviations The SCAD penalty is a non-convex, separable regularizer that transitions from ℓ_1 penalization near the origin to an unpenalized regime away from 0. In its original non-smooth form, it is defined as

$$\mathcal{R}_{\boldsymbol{\beta}}^{SCAD}(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{d} \rho_{\lambda,a}^{SCAD}(\beta_j),$$

where $a > 2$ is a hyperparameter and

$$\rho_{\lambda,a}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j|, & |\beta_j| \le \lambda, \\ \dfrac{-\beta_j^2 + 2a\lambda|\beta_j| - \lambda^2}{2(a-1)}, & \lambda < |\beta_j| \le a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & |\beta_j| > a\lambda. \end{cases}$$

As the construction of a smooth surrogate is completely analogous to the previous MCP derivation, only a brief description is provided here. The variational equivalence and boundary reduction follow the same arguments as in Lemma 21. The fully differentiable surrogate for $\mathcal{R}_{\boldsymbol{\beta}}^{SCAD}$ is constructed via replacing $|\beta_j|$ by $(u_j^2 + v_j^2)/2$ and β_j^2 by $(u_j^2 + v_j^2)^2/4$, yielding a smooth, piecewise surrogate penalty. As in the MCP construction, we abbreviate $s_j \triangleq (u_j^2 + v_j^2)/2$ and obtain the following smooth surrogate for each β_j:

$$\mathcal{R}_{\boldsymbol{\xi}}^{SCAD}(\boldsymbol{u},\boldsymbol{v}) \triangleq \sum_{j=1}^{d} \tilde{\rho}_{\lambda,a}^{SCAD}(u_j, v_j), \quad \tilde{\rho}_{\lambda,a}^{SCAD}(u_j, v_j) \triangleq \begin{cases} \lambda s_j, & s_j \le \lambda, \\ \dfrac{-s_j^2 + 2a\lambda s_j - \lambda^2}{2(a-1)}, & \lambda < s_j \le a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & s_j > a\lambda. \end{cases}$$

This yields an SVF of the SCAD penalty under the HPP:

$$\min_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d:\boldsymbol{u}\odot\boldsymbol{v}=\boldsymbol{\beta}} \mathcal{R}_{\boldsymbol{\xi}}^{SCAD}(\boldsymbol{u},\boldsymbol{v}) = \mathcal{R}_{\boldsymbol{\beta}}^{SCAD}(\boldsymbol{\beta}) \;\; \forall\, \boldsymbol{\beta} \in \mathbb{R}^d,$$

which is continuously differentiable in $(\boldsymbol{u}, \boldsymbol{v})$ and separable across coordinates. Using the same AM-GM arguments as for the MCP, the feasible values of s_j under the con-

straint $u_j v_j = \beta_j$ satisfy $s_j \in [|\beta_j|, \infty)$, with the lower bound attained for balanced factorizations. Consequently, the branch conditions of $\tilde{\rho}^{SCAD}_{\lambda,a}(u_j, v_j)$ in terms of s_j reduce to thresholds in $|\beta_j|$: if $|\beta_j| > a\lambda$, all feasible points s_j lie on the constant branch. If $\lambda < |\beta_j| \leqslant a\lambda$, the minimal feasible s_j lies in the middle branch. Finally, if $|\beta_j| \leqslant \lambda$, the minimal feasible s_j lies in the first branch. Since $\tilde{\rho}^{SCAD}_{\lambda,a}$ is non-decreasing in s_j on each piecewise branch, the constrained minimum over $u_j v_j = \beta_j$ is attained at the smallest feasible value $s_j = |\beta_j|$, recovering exactly the scalar SCAD penalty $\rho^{SCAD}_{\lambda,a}(\beta_j)$.

Appendix B Details on Numerical Experiments

Table 3 provides a detailed overview of the optimization hyperparameters, simulation settings, and data/task-specific information for the numerical experiments in Sect. 9.

B.1 Comparison of (G)HPP vs SubGD Optimization

Figure 11 shows the norm-based regularization paths for the first experiment. The plot confirms that the (group) lasso objective can be effectively optimized using our smooth surrogate method, matching the optimal trajectory and inducing numerically exact zeros, while direct GD struggles to even shrink parameters near zero.

Table 3 Hyperparameters and details for experiments in Section. 9

Config./experiment	**Optim. ℓ_1**	**Optim. $\ell_{2,1}$**	**Sparse Lin. Reg.**	**LeNet-300-100 Pruning**	**Filter-sparse CNN**
Optimizer	SGD	SGD	SGD	Adam	SGD
Learning rate	0.18	0.1	0.005	0.001	0.01
LR scheduler	cosine	cosine	decay (10^{-6})	cosine	decay (10^{-5})
Momentum	0.9	0.9	0	✗	0.9
Epochs	3000	2000	2000	75	100
Early stopping	✗	✗	✓ (200)	✓ (10)	✓ (6)
Batch size	full batch	full batch	32	128	32
Loss	MSE	MSE	MSE	cross-entropy	cross-entropy
Init. 1st factor	He Normal	He Normal	He Normal	He Normal (adj.)	Glorot Unif
Init. rem. factors	**1** (ones)	**1** (ones)	**1** (ones)	He Normal (adj.)	**1** (ones)
Threshold (0)	10^{-6}	10^{-6}	val. optimal	`float32.eps`	`float32.eps`
Repetitions	✗	✗	30	5	10
Data and tasks					
Task type	regression	regression	regression	classif	classif
Sparsity type	unstruct. (ℓ_1)	struct. ($\ell_{2,1}$)	unstruct. ($\ell_{2/k}$)	unstruct. ($\ell_{2/k}$)	struct. ($\ell_{2,2/k}$)
Train samples	1000	1000	500	60, 000	60, 000
Test samples	✗	✗	500	10, 000	10, 000
Input dim	100	100	$\{100, 1000\}$	784	784
Output dim	1	1	1	10	10

The parentheses for early stopping indicate the patience in epochs.float32.eps $\approx 1.19 \times 10^{-7}$.

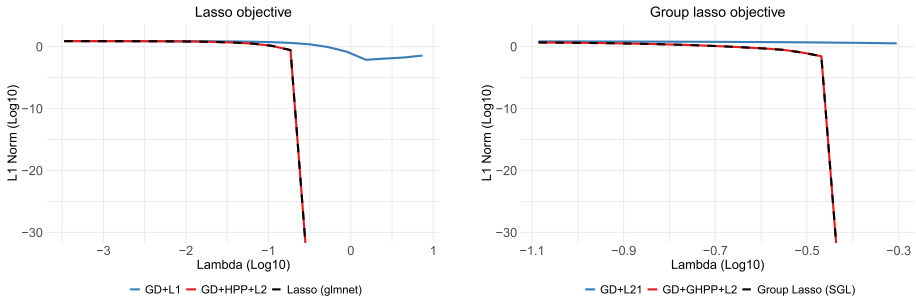
B.2 Sparse Linear Regression

Detailed Set-Up In our experiments, we simulate 30 data sets $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \mathbb{R}$ the scalar outcome. Each dataset contains $n = 500$ samples each for training, validation, and testing. We focus on the high-dimensional $d > n$ setting, in which the number features, $d = 1000$, exceeds the number of samples. We set the number of informative features in the true parameter vector, $\boldsymbol{\beta}^* \in \mathbb{R}^d$, to $s = \|\boldsymbol{\beta}^*\|_0 = 10$. The magnitudes of the true signals range from the smallest signal value $|\beta^{\min}| \triangleq \frac{1}{2}\sigma\sqrt{(2/n)\log(d)}$ to a "large" signal $|\beta^{large}| \triangleq 2\log(d)\sigma\sqrt{(2/n)\log(d)}$, where $\sigma > 0$ is the standard deviation of the additive noise in (B1). This range is based on the information-theoretic lower bound for recoverable signals (Wainwright, 2009; Zhang, 2010). Our data are simulated according to the following data-generating process:

$$(\boldsymbol{X}, \boldsymbol{\varepsilon}) \sim P_{\boldsymbol{X}} \times P_{\boldsymbol{\varepsilon}}, \quad \boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\varepsilon} \in \mathbb{R}^n, \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{Y} \in \mathbb{R}^d, \tag{B1}$$

where $P_{\boldsymbol{\varepsilon}}$ is a spherical Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, i.e., $\sigma = 1$, and $P_{\boldsymbol{X}}$ corresponds to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. We consider two settings for the design matrix $\boldsymbol{X}$: in the first setting, independent features with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ are used, whereas the second setting investigates correlated features drawn from a multivariate Gaussian with Toeplitz power covariance structure, i.e., $\boldsymbol{\Sigma}_{i,j} \triangleq \rho^{|i-j|}$ with correlation $\rho = 0.5$. The s informative signals' indices are randomly assigned in each simulation. To optimize the ℓ_1, SCAD, and MCP regularized problems, we choose a specialized routine based on the Convex-Concave Procedure and the Modified Local Quadratic Approximation implemented in the `ncpen` R package (Kim et al., 2021).[5]

Variable selection SGD does not have an in-built mechanism to produce (theoretically) exact zeros, although given a sufficiently large number of iterations, a zero floating-point representation can be obtained. To circumvent inefficient training times to obtain numerically zero parameters in our SGD-optimized models, we use early stopping combined



(a) HPP vs direct GD for lasso objective (b) GHPP vs direct GD for group lasso

Fig. 11 Comparison of parameter norms of (G)HPP-based GD and direct (Sub)GD optimization of the non-smooth ℓ_1 regularized lasso (**a**) and $\ell_{2,1}$ regularized group lasso (**b**) objectives. Dashed lines indicate optimal solutions

[5] We also performed experiments with more widespread coordinate descent algorithms implemented in the glmnet (Friedman et al., 2010) and ncvreg (Breheny & Huang, 2011) packages with qualitatively identical results.

with a post-thresholding step as suggested in, (Zhao et al., 2022), whose optimal cut-off for the reconstructed parameters is determined on the validation loss.

Figure 12 shows the support recovery, defined as the classification accuracy w.r.t. informative signals, as well as the number of false positives (FP) for the models in Sect. 9, after applying thresholding to our overparametrized models. To disentangle the effects of the optimization transfer and the thresholding step, we also apply the same operation to the conventional lasso. We can make two basic observations: first, support recovery, and FP both improve monotonically with increasing factorization depth k. Second, we observe that the thresholding step also significantly improves variable selection performance.

Low-dimensional simulation setting Besides the $d > n$ setting we previously analyzed, we repeat the experiment for a low-dimensional $d < n$ setting while keeping the number of true signals constant. The magnitudes of the non-zero parameters are adjusted to the new setting using the provided definitions. Figure 13 shows the results for an identical simulation set-up as before, but in an alternative lower-dimensional setting with $d = 100$ features, $s = 10$ non-zero parameters, and $n = 500$ samples in the training data. Qualitatively, the results are consistent with those for the high-dimensional setting, with minor instabilities at large factorization depths k, suggesting a trade-off between depth and stability.

Further simulations with varying ground-truth parameter Complementary to the previously analyzed low- and high-dimensional settings, we further repeat all simulations by varying the structure and sparsity of the ground-truth vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$.

In our first additional setting, we keep the number of true signals at $s = \|\boldsymbol{\beta}^*\|_0 = 10$, but modify their structure and scale by fixing their values to $(-0.75, -0.25, -2, -2, -2, 2, 2, 2, 0.25, 0.75)^\top$. Figure 14 contains the full results for

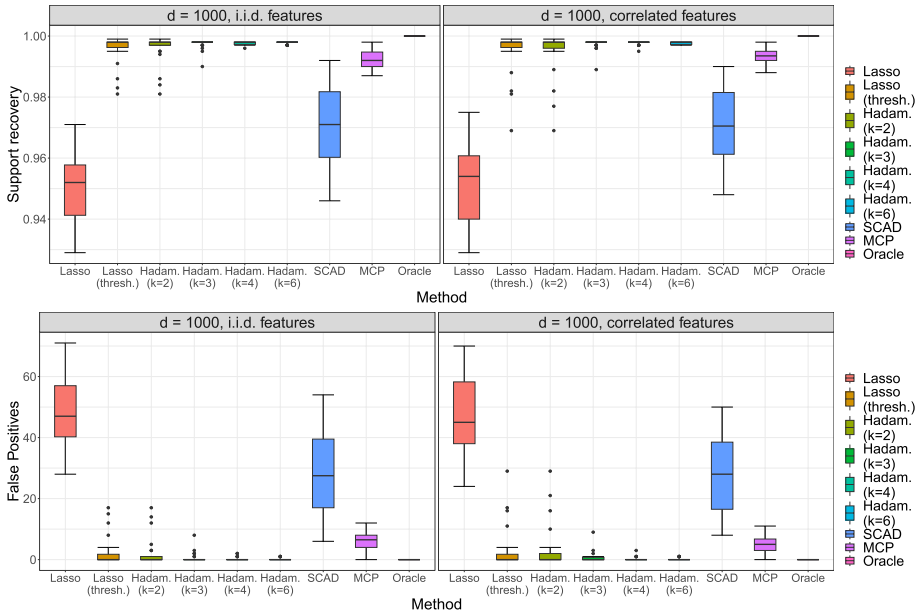


Fig. 12 Support recovery accuracy (top row) and false positives (bottom row) for different $\boldsymbol{\Sigma}$ settings (columns) and increasing factorization depths, compared against standard implementations of convex ℓ_1 and non-convex SCAD and MCP penalties

both low- and high-dimensional settings and independent vs correlated features, with 30 simulation repetitions for each model, feature dimension, and correlation structure. Qualitatively, the results are consistent with previous findings in Figs. 7, 12, 13.

In our second additional setting, we vary the sparsity of the ground-truth vector and increase the number of non-zero signals to $s = \|\boldsymbol{\beta}^*\|_0 = 40$. We specify half the coefficients to be "small" and logarithmically spaced between $\beta^{\min}/2$ and $\beta^{large}/2$ as defined in Appendix B.2. Note that this includes true signals below the recoverable threshold. The remaining 20 "large" signals are selected to have magnitudes between 2 and 5. Half of the signals have positive and half have negative signs. Figure 15 contains the full results, again with 30 repetitions for each model, feature dimension, and correlation structure. Qualitatively, the results are largely consistent with previous findings and further highlight the trade-off between depth and stability, as the performance for deep factorizations with $k = 6$ becomes brittle and degrades on average.

B.3 Details on CNN Architecture

Our small VGG-style CNN implementation consists of two blocks of two convolutional layers after each of which max pooling is applied. The convolutional layers have a kernel size of 3 and stride 1. ReLU activation is used for all hidden layers. The classification head consists of two hidden layers followed by the softmax output. Dropout (Srivastava et al., 2014) is applied after each convolutional block and dense layer. The full architecture is:

$$[[\mathrm{Input}((28,28)), \mathrm{Conv2D}(32), \mathrm{Conv2D}(32), \mathrm{MaxPool2D}(2), \mathrm{Dropout}(0.25)],$$
$$[\mathrm{Conv2D}(64), \mathrm{Conv2D}(64), \mathrm{MaxPool2D}(2), \mathrm{Dropout}(0.25)],$$
$$[\mathrm{Dense}(32), \mathrm{Dropout}(0.25), \mathrm{Dense}(32), \mathrm{Dropout}(0.25), \mathrm{Dense}(10)]]$$

B.4 Additional Results on Computational Complexity

The experiments on computational overhead are performed on a single 16GB RTX A4000 GPU using `TensorFlow 2.9`. The time per sample is the average wall-clock time for a single epoch normalized by sample size. The fully-connected network (MLP) has four hidden ReLU layers with 128 units each, containing ≈ 0.15m parameters. The input

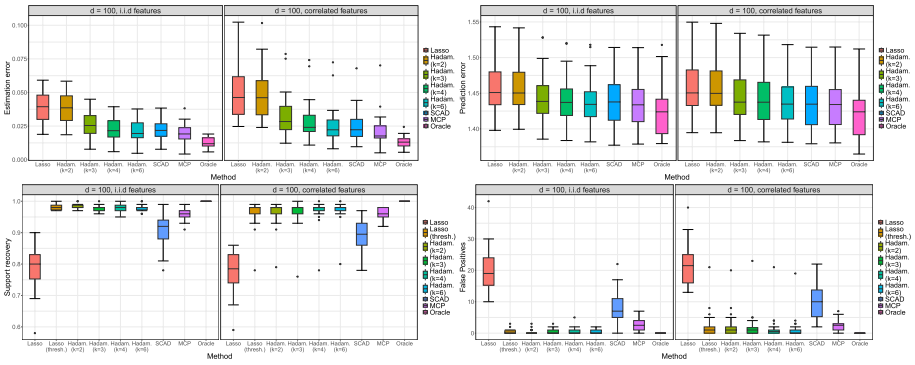
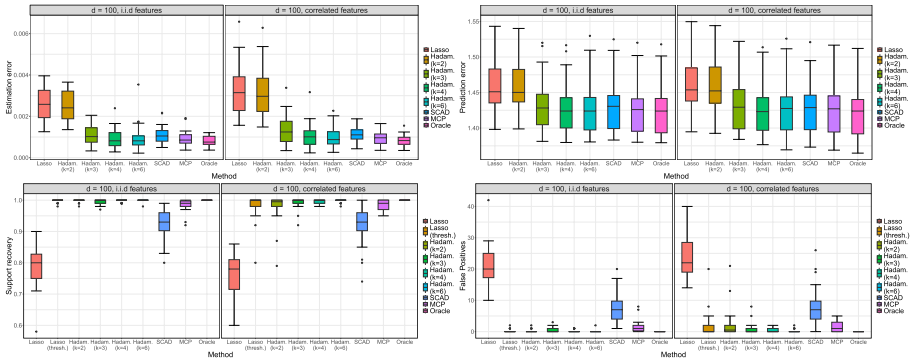
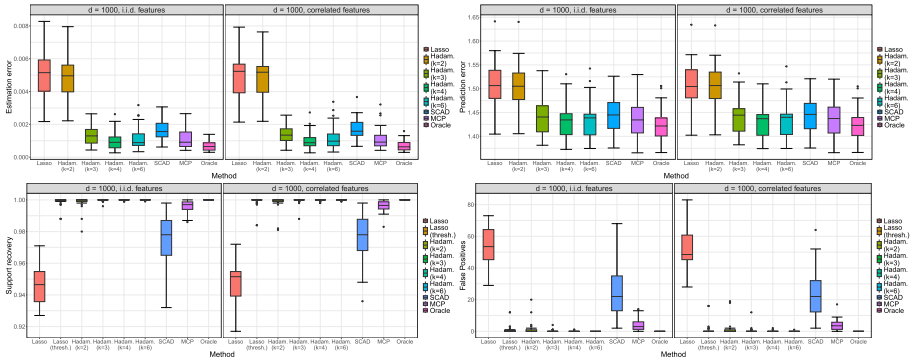


Fig. 13 Left: standardized estimation error (top row) and support recovery accuracy (bottom row) for different $\boldsymbol{\Sigma}$ settings (columns) of our approach for increasing factorization depths, compared against standard implementations of convex ℓ_1 and non-convex SCAD and MCP penalties. Right: test prediction error (top row) and false positives (bottom row) for different settings of $\boldsymbol{\Sigma}$ (columns)



(a) Low-dimensional setting ($d = 100$)



(b) High-dimensional setting ($d = 1000$)

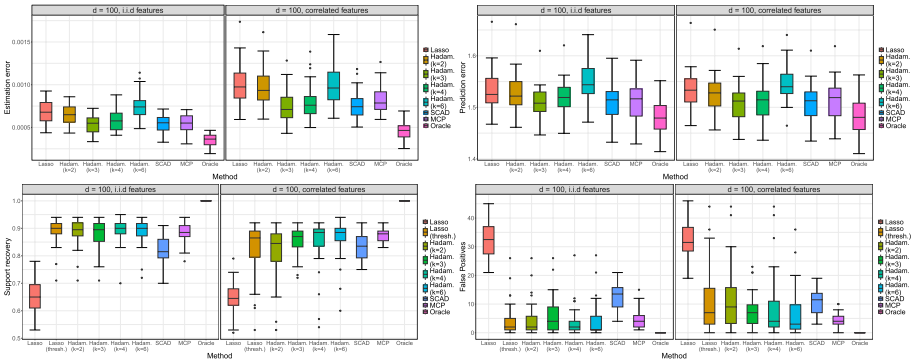
Fig. 14 Simulation results for different true signal choice. Each subfigure contains the following plots: Left: standardized estimation error (top row) and support recovery accuracy (bottom row) for different $\mathbf{\Sigma}$ settings (columns) of our approach for increasing factorization depths, compared against standard implementations of convex ℓ_1 and non-convex SCAD and MCP penalties. Right: test prediction error (top row) and false positives (bottom row) for different settings of $\mathbf{\Sigma}$ (columns)

is a flattened (28, 28) image and the softmax output has 10 units. All trainable weights and biases are overparametrized. Figure 16 shows additional experiments for the HPP_k applied to a ResNet-20 (≈ 0.27m param.) trained on CIFAR10 (He et al., 2016). As for the MLP, the computational overhead increases with depth and decreases with batch size. The recommended batch size of 256 results in a $< 5\%$ increase for $k = 8$.

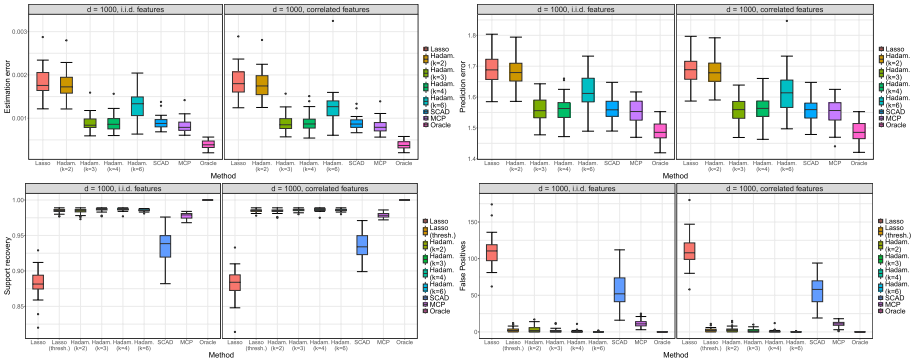
Appendix C Details on Geometric Intuition

C.1 Difference between HPP and HDP: $2\|\beta\|_1$ vs $\|\beta\|_1$

In the literature, both the HPP and HDP are used almost interchangeably to induce ℓ_1 regularization in some way. However, there are subtle differences between the two, result-



(a) Low-dimensional setting ($d = 100$)



(b) High-dimensional setting ($d = 1000$)

Fig. 15 Simulation results for different ground-truth sparsity. Each subfigure contains the following plots: Left: standardized estimation error (top row) and support recovery accuracy (bottom row) for different $\mathbf{\Sigma}$ settings (columns) of our approach for increasing factorization depths, compared against standard implementations of convex ℓ_1 and non-convex SCAD and MCP penalties. Right: test prediction error (top row) and false positives (bottom row) for different settings of $\mathbf{\Sigma}$ (columns)

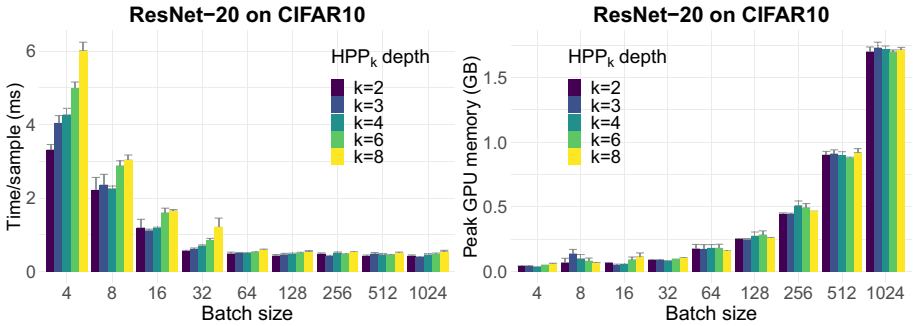


Fig. 16 Time per sample (training) for ResNet20 on CIFAR10. Left: training time per sample for different factorization depths k. Right: peak GPU memory utilization. Means and standard errors over four runs are displayed.

ing in different regularization strengths. Both parametrizations define hyperbolic paraboloids for each $j = 1, \ldots, d$ where the HPP can be transformed into the HDP via rotation and scaling operations. To see this, consider the scalar case $\beta \in \mathbb{R}$. Defining the $-45°$ rotation of any point (u, v) on the Cartesian plane as $(u, v) \mapsto (\frac{u+v}{\sqrt{2}}, \frac{v-u}{\sqrt{2}}) \triangleq \mathrm{rot(u, v)}$, and recalling that the coordinate change from HPP to HDP is $(\gamma, \delta) = (u + v, v - u)$, it follows that $\mathrm{HDP}(\gamma, \delta) = \mathrm{HPP}(\sqrt{2}\,\mathrm{rot}(u, v)) = 2 \cdot \mathrm{HPP}(\mathrm{rot}(u, v))$. This means the HDP is the HPP but with all points rotated by $-45°$ and its output scaled by 2 (or equivalently, its arguments scaled by $\sqrt{2}$.) Since the surrogate ℓ_2 regularizer is the same for both parametrizations, this scaling results in a decreased gap between $\mathcal{K}$ and the surrogate regularizer (cf. Figs. 17 and 4a). Geometrically speaking, the fibers $\mathcal{K}^{-1}(\beta)$ of the HDP extend closer to the origin than for the HPP, allowing minimum-norm points with smaller norm. As derived in Sects. 3.1 and 3.2, for $\beta > 0$, the minimum-norm points for the HPP are $\pm(\sqrt{\beta}, \sqrt{\beta})$, and $(\pm\sqrt{\beta}, 0)$ for the HDP. Evaluating the surrogate ℓ_2 penalty at those points, we find an induced regularizer of $\sqrt{\beta}^2 + \sqrt{\beta}^2 = 2|\beta|$ for the HPP, but only $0^2 + \sqrt{\beta}^2 = |\beta|$ for the HDP.

C.2 Geometric Intuition for HPP_k in Three Dimensions

With ℓ_2 regularization, factorizing a scalar parameter $\beta \in \mathbb{R}$ using $\mathcal{K}(\boldsymbol{u}) = \mathcal{K}(u_1, u_2, u_3) = u_1 u_2 u_3$ yields a minimal constrained ℓ_2 penalty of $\mathcal{R}_{\boldsymbol{\beta}}(u_1, u_2, u_3) = 3|u_1 u_2 u_3|^{2/3}$ over $\mathcal{K}^{-1}(\beta)$, or $\mathcal{R}_{\boldsymbol{\beta}}(\beta) = 3|\beta|^{2/3}$ in terms of β, inducing differentiable sparse $\ell_{2/3}$ regularization (cf. Fig. 18).

C.3 Curvature-Inducing Effects on Optimization Landscape

The parametrizations considered by us (cf. Table 1) have a significant impact on the loss landscape caused by a change in curvature induced by the multiplicative nature of the parametrization. Powerpropagation (36), $\mathcal{K}(\boldsymbol{v}) = \boldsymbol{v} \odot |\boldsymbol{v}|^{\circ(k-1)}$, as a bijective map, allows disentangling the curvature effect from overparametrization, i.e., the curvature is modified in the same base parameter space. The left panel of Fig. 19a shows that for increasing factorization depths $k \in \{2, 4, 6\}$, increasingly sharp transitions at $v \in \{-1, 1\}$

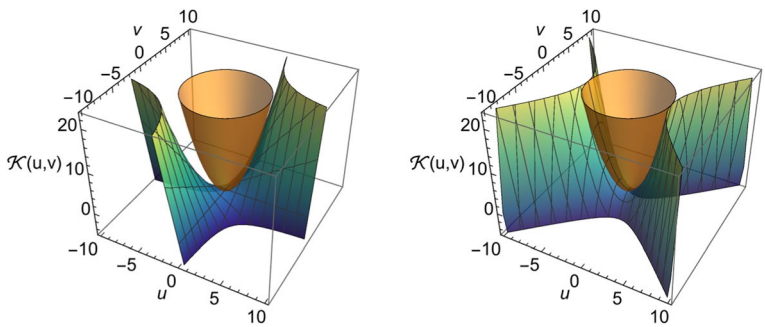


Fig. 17 Left: HPP (blue/green) and the surrogate ℓ_2 regularization $u_j^2 + v_j^2$ (orange). Right: HDP (blue/green) and surrogate ℓ_2 regularization

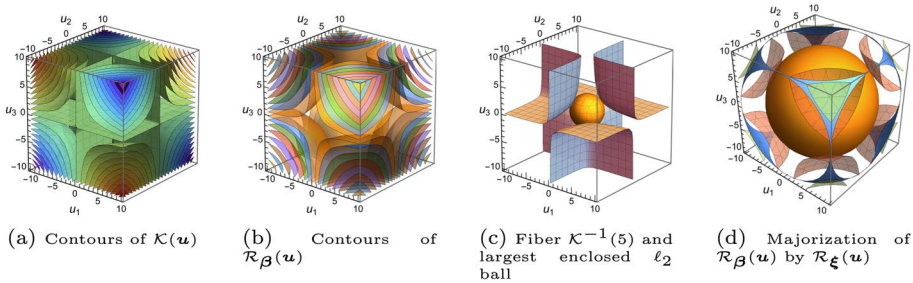


Fig. 18 Visualization of HPP_k, and **a** minimum constrained ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\beta}}(u_1, u_2, u_3)$. **b**HPP_k with $\mathcal{K}(u_1, u_2, u_3) = u_1 u_2 u_3$ and minimum constrained ℓ_2 regularization term $\mathcal{R}_{\boldsymbol{\beta}}(u_1, u_2, u_3) = 3 \cdot |u_1 u_2 u_3|^{2/3}$. **c** fiber of $\mathcal{K}$ at $\beta = 5$, illustrating the location of the 4 points with minimal distance to the origin at the vertices $(\hat{u}_1, \hat{u}_2, \hat{u}_3)$ of the hyperbolic smooth manifold. The vertices lie tangential to the largest enclosed ℓ_2-ball having a radius of $\sqrt{3 \cdot 5^{2/3}}$. Right: two contours each of surrogate ℓ_2 penalty $\mathcal{R}_{\boldsymbol{\xi}}$ (solid shapes) and $\mathcal{R}_{\boldsymbol{\beta}}$ at the same levels (opaque). For each value $\beta = u_1 u_2 u_3 \neq 0$, there are 8 points where the surrogate attains minimum ℓ_2 distance; however, only half are solutions of the SVF by restriction to orthants that respect the sign of β und
er $\mathcal{K}$

are induced. Given an unregularized base objective, the right panel of Fig. 19a shows corresponding equivalent surrogate objectives applying Powerpropagation without surrogate regularization. The left panel of Fig. 19b displays the same base objective with non-smooth, and in parts non-convex, ℓ_q regularization where $q = 2/k$. The right plot depicts the corresponding equivalent surrogates obtained from our optimization transfer.

Appendix D Derivation of Gradient and Hessian of Smooth Surrogate $\mathcal{Q}$

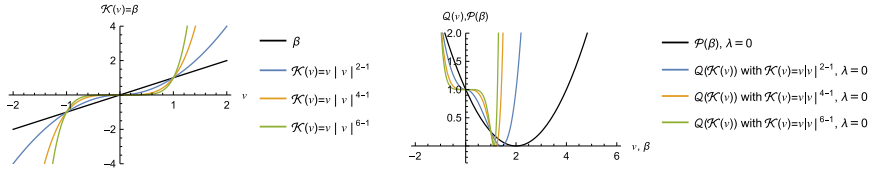
For simplicity, we assume no unregularized parameters $\boldsymbol{\psi}$ so that the objective function $\mathcal{Q} : \mathbb{R}^{d_\xi} \to \mathbb{R}_0^+$ is defined as $\mathcal{Q}(\boldsymbol{\xi}) = \mathcal{L}(\mathcal{K}(\boldsymbol{\xi})) + \lambda \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, where $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}_0^+$ and $\mathcal{K} : \mathbb{R}^{d_\xi} \to \mathbb{R}^d$ are $\mathcal{C}^2$-smooth functions, $\mathcal{R}_{\boldsymbol{\xi}} : \mathbb{R}^{d_\xi} \to \mathbb{R}_0^+$ is a strongly convex ℓ_2 regularization term, and $\lambda \geqslant 0$ a scalar. The gradient of $\mathcal{Q}$ with respect to $\boldsymbol{\xi}$ is given by

$$\nabla_{\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{\xi}) = \mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}^{\top}(\boldsymbol{\xi}) \nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi})) + \lambda \nabla_{\boldsymbol{\xi}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi}),$$

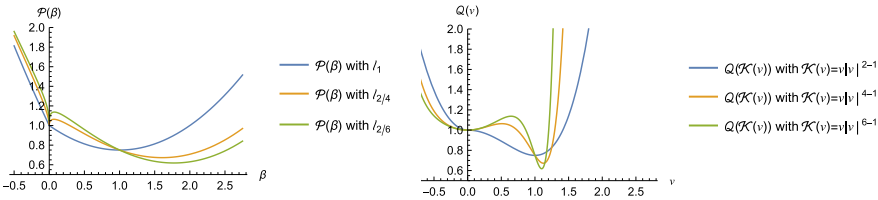
where $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi})$ is the $d \times d_\xi$-dimensional Jacobian of $\mathcal{K}$ at $\boldsymbol{\xi}$, and the gradients $\nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi}))$ and $\nabla_{\boldsymbol{\xi}} \mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ are vectors with d and d_ξ entries. The Hessian of $\mathcal{Q}$ at $\boldsymbol{\xi}$ is then obtained as

$$\mathcal{H}_{Q(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathcal{H}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) \nabla_{\mathcal{K}} \mathcal{L}(\mathcal{K}(\boldsymbol{\xi})) + \mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}^{\top}(\boldsymbol{\xi}) \mathcal{H}_{\mathcal{L}(\mathcal{K}(\boldsymbol{\xi}))}(\boldsymbol{\xi}) \mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) + \lambda \mathcal{H}_{\mathcal{R}_{\boldsymbol{\xi}}}(\boldsymbol{\xi}),$$

where $\mathcal{H}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi})$ is a third-order $d_\xi \times d_\xi \times d$-dimensional tensor, and $\mathcal{H}_{\mathcal{L}(\mathcal{K}(\boldsymbol{\xi}))}(\boldsymbol{\xi})$ and $\mathcal{H}_{\mathcal{R}_{\boldsymbol{\xi}}}(\boldsymbol{\xi})$ are Hessians of dimensions $d \times d$ and $d_\xi \times d_\xi$. From this representation, we can see that $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathcal{H}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathbf{0}$ imply $\mathcal{H}_{\mathcal{Q}(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathbf{0}$ if $\lambda = 0$. For $\lambda > 0$, $\mathcal{J}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathcal{H}_{\mathcal{K}(\boldsymbol{\xi})}(\boldsymbol{\xi}) = \mathbf{0}$ imply that $\mathbf{0}$ is a local minimizer of $\mathcal{Q}(\boldsymbol{\xi})$ due to the positive definiteness of the $\mathcal{H}_{\mathcal{R}_{\boldsymbol{\xi}}}(\boldsymbol{\xi})$ implied by the strong convexity of $\mathcal{R}_{\boldsymbol{\xi}}(\boldsymbol{\xi})$.



(a) **Left**: Powerpropagation $\mathcal{K}(v) = \beta$ (36) for $k \in \{2, 4, 6\}$. **Right**: unregularized objective $\mathcal{P}(\beta) = (1 - \frac{1}{2}\beta)^2$ (black) and equivalent smooth surrogates $\mathcal{Q}(v) = (1 - \frac{1}{2}(v|v|^{k-1}))^2$. Note the additional saddle at $v = 0$.



(b) **Left**: non-smooth $\ell_{2/k}$ regularized base objectives $\mathcal{P}(\beta)$ with $\lambda = \frac{1}{2}$. **Right**: smooth surrogates $\mathcal{Q}(v) = (1 - \frac{1}{2}(v|v|^{k-1}))^2 + \lambda v^2$ equivalent to $\mathcal{P}(\beta)$ on left.

Fig. 19 Visualization of how smooth optimization transfer transforms the loss landscape using Powerprop. (36) on base objectives $\mathcal{P}(\beta) \triangleq (1 - \frac{1}{2}\beta)^2 + \lambda|\beta|^{2/k}$, $k \in \{2, 4, 6\}$

Author Contributions C.K. conducted the theoretical and empirical analyses and created a first draft of the manuscript, as well as implemented the revisions. C.M. proposed the original idea and contributed to the conceptualization of our framework. B.B. helped to steer and fund the project. D.R. contributed to the conceptualization, critically revised the manuscript, and supervised the project. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets used in this study (MNIST, Fashion-MNIST, and CIFAR-10) are publicly available benchmark datasets accessible through common deep learning libraries.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Andriushchenko, M., Varre, A.V., Pillaud-Vivien, L., & Flammarion, N. (2023). Sgd with large step sizes learns sparse features. In: International Conference on Machine Learning, pp. 903–925. PMLR.

Arora, S., Cohen, N., Hu, W., & Luo, Y. (2019). Implicit regularization in deep matrix factorization. Advances in Neural Information Processing Systems 32.

Aubin, J.-P., & Frankowska, H. (2009). Set-valued analysis.
Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning, 4*(1), 1–106.
Balcerzak, M., Strobin, F., & Wachowicz, A. (2013). Bilinear mappings-selected properties and problems. In: Filipczak M., Wagner-Bojakowska E. (red.), Traditional and present-day topics in real analysis. Dedicated to Professor Jan Stanisław Lipiński, Wydawnictwo Uniwersytetu Lódzkiego, Lódź
Balcerzak, M., Behrends, E., & Strobin, F. (2016). On certain uniformly open multilinear mappings. *Banach Journal of Mathematical Analysis, 10*(3), 482–494.
Benning, M., & Burger, M. (2018). Modern regularization methods for inverse problems. *Acta Numerica, 27*, 1–111.
Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2020). What is the state of neural network pruning? *Proceedings of Machine Learning and Systems, 2*, 129–146.
Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics, 5*(1), 232.
Chartrand, R., & Yin, W. (2008). Iteratively reweighted algorithms for compressive sensing. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3869–3872. IEEE.
Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters, 14*(10), 707–710.
Chartrand, R., & Staneva, V. (2008). Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems, 24*(3), Article 035020.
Cheltsov, I., Cornalba, F., Poon, C., & Shardlow, T. (2024). Hadamard langevin dynamics for sampling sparse priors. arXiv preprint arXiv:2411.11403
Chen, S., & Donoho, D. (1994) Basis pursuit. In: Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 41–44. IEEE.
Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review, 43*(1), 129–159.
Chen, Y., Ge, D., Wang, M., Wang, Z., Ye, Y., & Yin, H. (2017). Strong np-hardness for sparse optimization with concave penalty functions. In: International Conference on Machine Learning, pp. 740–747. PMLR.
Chen, F., Kunin, D., Yamamura, A., & Ganguli, S. (2024). Stochastic collapse: How gradient noise attracts SGD dynamics towards simpler subnetworks. Advances in Neural Information Processing Systems 36.
Chou, H.-H., Maly, J., & Rauhut, H. (2023). More is less: Inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA, 12*(3), 1437.
Chou, H.-H., Maly, J., & Verdun, C.M. (2022). Non-negative least squares via overparametrization. arXiv preprint arXiv:2207.08437
Combettes, P. L., & Müller, C. L. (2018). Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications, 457*(2), 1283–1306.
Combettes, P. L., & Müller, C. L. (2020). Perspective maximum likelihood-type estimation via proximal decomposition. *Electronic Journal of Statistics, 14*, 207–238.
Dai, Z., Karzand, M., & Srebro, N. (2021) Representation costs of linear neural networks: Analysis and design. Advances in Neural Information Processing Systems 34.
Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine, 29*(6), 141–142.
Donoho, D. L., & Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences, 100*(5), 2197–2202.
Du, S.S., Jin, C., Lee, J.D., Jordan, M.I., Singh, A., & Poczos, B. (2017). Gradient descent can take exponential time to escape saddle points. Advances in Neural Information Processing Systems 30.
Ergen, T., & Pilanci, M. (2021). Revealing the structure of deep neural networks via convex duality. In: International Conference on Machine Learning, pp. 3004–3014. PMLR.
Ergen, T., & Pilanci, M. (2023). Path regularization: A convexity and sparsity inducing regularization for parallel relu networks. *Advances in Neural Information Processing Systems, 36*, 59761–59786.
Even, M., Pesme, S., Gunasekar, S., & Flammarion, N. (2023). (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems, 36*, 29406–29448.
Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.
Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 109–135.
Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review, 90*(1), 118–145.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics, 7*(3), 397–416.

Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics, 28*(5), 1356–1378.

Ge, D., Jiang, X., & Ye, Y. (2011). A note on the complexity of ℓ_p minimization. *Mathematical Programming, 129*(2), 285–299.

Ge, R., Huang, F., Jin, C., & Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In: Conference on Learning Theory, pp. 797–842. PMLR.

Gissin, D., Shalev-Shwartz, S., & Daniely, A. (2019). The implicit bias of depth: How incremental learning drives generalization. In: International Conference on Learning Representations.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In: ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998 8, pp. 201–206. Springer.

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In: International Conference on Machine Learning, pp. 1832–1841. PMLR.

Gunasekar, S., Lee, J.D., Soudry, D., & Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. Advances in Neural Information Processing Systems 31.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. Advances in neural information processing systems 28.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hirsch, M.W. (2012). Differential topology.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research, 22*(241), 1–124.

Hoff, P. D. (2017). Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis, 115*, 186–198.

Horowitz, C. (1975). An elementary counterexample to the open mapping principle for bilinear maps. *Proceedings of the American Mathematical Society, 53*(2), 293–294.

Huang, J., Zhang, T., &Metaxas, D. (2009). Learning with structured sparsity. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 417–424.

Hu, Y., Li, C., Meng, K., Qin, J., & Yang, X. (2017). Group sparse optimization via $\ell_{p,q}$ regularization. *The Journal of Machine Learning Research, 18*(1), 960–1011.

Hunter, D. R., & Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics, 33*(4), 1617.

Jacobs, T., & Burkholz, R. (2025). Mask in the mirror: Implicit sparsification. In: The Thirteenth International Conference on Learning Representations.

Jacobs, T., Zhou, C., & Burkholz, R. (2025). Mirror, mirror of the flow: How does regularization shape implicit bias? In: Forty-second International Conference on Machine Learning.

Jacot, A., Golikov, E., Hongler, C., & Gabriel, F. (2022). Feature learning in l_2-regularized DNNS: Attraction/repulsion and sparsity. In: Advances in Neural Information Processing Systems.

Jagadeesan, M., Razenshteyn, I., & Gunasekar, S. (2022). Inductive bias of multi-channel linear convolutional networks with bounded weight norm. In: Conference on Learning Theory, pp. 2276–2325. PMLR.

Jenatton, R., Audibert, J.-Y., & Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research, 12*, 2777–2824.

Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., & Jordan, M.I. (2017). How to escape saddle points efficiently. In: International Conference on Machine Learning, pp. 1724–1732. PMLR.

Kaushik, C., Romberg, J., & Muthukumar, V. (2024) Precise asymptotics of reweighted least-squares algorithms for linear diagonal networks. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Kawaguchi, K. (2016). Deep learning without poor local minima. Advances in Neural Information Processing Systems 29.

Kim, D., Lee, S., & Kwon, S. (2021). A unified algorithm for the non-convex penalized estimation: The NCPEN package. *The R Journal, 12*(2), 43–60.

Kolb, C., Bischl, B., & Rügamer, D. (2025). Differentiable attention sparsity via structured d-gating. In: ICLR workshop on sparsity in LLMs (SLLM): Deep dive into mixture of experts, quantization, hardware, and inference.

Kolb, C., Frost, L., Bischl, B., & Rügamer, D. (2025). Differentiable sparsity via d-gating: Simple and versatile structured penalization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Kolb, C., Weber, T., Bischl, B., & Rügamer, D. (2025). Deep weight factorization: Sparse learning through the lens of artificial symmetries. In: The Thirteenth International Conference on Learning Representations.

Labarrière, H., Molinari, C., Rosasco, L., Villa, S., & Vega, C. (2024). Optimization insights into deep diagonal linear networks. arXiv preprint arXiv:2412.16765

Lange, K., Hunter, D. R., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics, 9*(1), 1–20.

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. Advances in Neural Information Processing Systems 2.

Lee, J.D., Simchowitz, M., Jordan, M.I., & Recht, B. (2016). Gradient descent only converges to minimizers. In: Conference on Learning Theory, pp. 1246–1257. PMLR.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., & Recht, B. (2019). First-order methods almost always avoid strict saddle points. *Mathematical Programming, 176*, 311–337.

Levin, E. (2020). Towards optimization on varieties. Undergraduate senior thesis, Princeton University.

Levin, E., Kileel, J., & Boumal, N. (2024). The effect of smooth parametrizations on nonconvex optimization landscapes. Mathematical Programming, 1–49.

Li, Y., & Lin, Q. (2024). Improving adaptivity via over-parameterization in sequence models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Li, J., Nguyen, T., Hegde, C., & Wong, K.W. (2021). Implicit sparse regularization: The impact of depth and early stopping. Advances in Neural Information Processing Systems 34.

Li, J., Nguyen, T.V., Hegde, C., & Wong, R.K. (2023). Implicit regularization for group sparsity. In: The Eleventh International Conference on Learning Representations.

Li, G., Li, S., Li, D., & Ma, C. (2023). The tail-hadamard product parametrization algorithm for compressed sensing. *Signal Processing, 205*, Article 108853.

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2736–2744.

Loh, P.-L., & Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research, 16*(1), 559–616.

Loh, B.P.-L., & Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics, 45*(6), 2455–2482.

Lyu, Q., Lin, Z., She, Y., & Zhang, C. (2013). A comparison of typical ℓ_p minimization algorithms. *Neurocomputing, 119*, 413–424.

Ma, J., & Fattahi, S. (2022). Blessing of nonconvexity in deep linear models: Depth flattens the optimization landscape around the true solution. arXiv preprint arXiv:2207.07612.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics, 34*(3), 1436–1462.

Micchelli, C. A., Morales, J. M., & Pontil, M. (2013). Regularizers for structured sparsity. *Advances in Computational Mathematics, 38*, 455–489.

Moroshko, E., Woodworth, B. E., Gunasekar, S., Lee, J. D., Srebro, N., & Soudry, D. (2020). Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in Neural Information Processing Systems, 33*, 22182–22193.

Nacson, M.S., Ravichandran, K., Srebro, N., & Soudry, D. (2022). Implicit bias of the step size in linear diagonal neural networks. In: International Conference on Machine Learning, pp. 16270–16295. PMLR.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing, 24*(2), 227–234.

Neyshabur, B., Tomioka, R., & Srebro, N. (2015) In search of the real inductive bias: On the role of implicit regularization in deep learning. In: ICLR (Workshop).

Neyshabur, B., Tomioka, R., & Srebro, N. (2015). Norm-based capacity control in neural networks. In: Conference on Learning Theory, pp. 1376–1401. PMLR.

Nouiehed, M., & Razaviyayn, M. (2022). Learning deep models: Critical points and local openness. *INFORMS Journal on Optimization, 4*(2), 148–173.

Olsson, C., Carlsson, M., Andersson, F., & Larsson, V. (2017) Non-convex rank/sparsity regularization and local minima. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 332–340.

Ouyang, W., Liu, Y., Pong, T. K., & Wang, H. (2025). Kurdyka-łojasiewicz exponent via hadamard parametrization. *SIAM Journal on Optimization, 35*(1), 62–91.

Parhi, R., & Nowak, R. D. (2023). Deep learning meets sparse regularization: A signal processing perspective. *IEEE Signal Processing Magazine, 40*(6), 63–74.

Pesme, S., Pillaud-Vivien, L., & Flammarion, N. (2021). Implicit bias of SGD for diagonal linear networks: A provable benefit of stochasticity. *Advances in Neural Information Processing Systems, 34*, 29218–29230.

Pilanci, M., & Ergen, T. (2020). Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In: International Conference on Machine Learning, pp. 7695–7705. PMLR.

Poon, C., & Peyré, G. (2021). Smooth bilevel programming for sparse regularization. Advances in Neural Information Processing Systems 34.

Poon, C., & Peyré, G. (2023). Smooth over-parameterized solvers for non-smooth structured optimization. Mathematical Programming, 1–56.

Ramlau, R., & Zarzer, C. A. (2012). On the minimization of a Tikhonov functional with a non-convex sparsity constraint. *Electronic Transactions on Numerical Analysis, 39*, 476–507.

Rudin, W. (1991). *Functional analysis* (2nd ed.). International Series in Pure and Applied Mathematics. McGraw-Hill Inc.

Schmidt, M., Fung, G., & Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In: European Conference on Machine Learning, pp. 286–297. Springer.

Schwarz, J., Jayakumar, S., Pascanu, R., Latham, P., & Teh, Y. (2021). Powerpropagation: A sparsity inducing weight reparameterisation. Advances in Neural Information Processing Systems 34.

Shin, M., & Liu, J. S. (2022). Neuronized priors for Bayesian sparse linear regression. *Journal of the American Statistical Association, 117*(540), 1695–1710.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group LASSO. *Journal of computational and graphical statistics, 22*(2), 231–245.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Tibshirani, R. (2021). Equivalences between sparse models and neural networks. Working Notes.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory, 52*(3), 1030–1051.

Vaskevicius, T., Kanade, V., & Rebeschini, P. (2019). Implicit regularization for optimal sparse recovery. Advances in Neural Information Processing Systems 32.

Vivien, L.P., Reygner, J., & Flammarion, N. (2022). Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In: Conference on Learning Theory, pp. 2127–2159. PMLR.

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory, 55*(12), 5728–5741.

Wang, Z., & Jacot, A. (2024). Implicit bias of sgd in l2-regularized linear dnns: One-way jumps from high to low rank. In: 12th International Conference on Learning Representations, ICLR 2024.

Wen, W., Wu, C., Wang, Y., Chen, Y., & Li, H. (2016). Learning structured sparsity in deep neural networks. In: Advances in Neural Information Processing Systems, vol. 29.

Wen, F., Chu, L., Liu, P., & Qiu, R. C. (2018). A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access, 6*, 69883–69906.

Woodworth, B., Gunasekar, S., Lee, J.D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., & Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In: Conference on Learning Theory, pp. 3635–3673. PMLR.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

Xu, Z., Zhang, H., Wang, Y., Chang, X., & Liang, Y. (2010). L1/2 regularization. *Science China Information Sciences, 53*(6), 1159–1169.

Xu, Z., Chang, X., Xu, F., & Zhang, H. (2012). L1/2 regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems, 23*(7), 1013–1027.

Yang, L., Zhang, J., Shenouda, J., Papailiopoulos, D., Lee, K., & Nowak, R.D. (2022). A better way to decay: Proximal gradient training algorithms for neural nets. In: OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop).

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhang, C.-H., & Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics, 36*(4), 1567–1594.

Zhang, S., & Xin, J. (2018). Minimization of transformed l 1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing. *Mathematical Programming, 169*(1), 307–336.

Zhao, P., & Yu, B. (2006). On model selection consistency of LASSO. *The Journal of Machine Learning Research, 7*, 2541–2563.

Zhao, P., Yang, Y., & He, Q.-C. (2022). High-dimensional linear regression via implicit regularization. Biometrika.

Ziyin, L. (2024). Symmetry induces structure and constraint of learning. In: Proceedings of the 41st International Conference on Machine Learning.

Ziyin, L., & Wang, Z. (2023). spred: Solving l1 penalty with SGD. In: International Conference on Machine Learning, pp. 43407–43422. PMLR.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Chris Kolb[1,2] · Christian L. Müller[1,2,3,4] · Bernd Bischl[1,2] · David Rügamer[1,2]

✉ Chris Kolb
chris.kolb@stat.uni-muenchen.de

Christian L. Müller
christian.mueller@helmholtz-munich.de

Bernd Bischl
bernd.bischl@stat.uni-muenchen.de

David Rügamer
david.ruegamer@stat.uni-muenchen.de

1 Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany

2 Munich Center for Machine Learning (MCML), Munich, Germany

3 Institute of Computational Biology, Helmholtz Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

4 Center for Computational Mathematics, Flatiron Institute, 162 5th Ave, New York NY 10010, USA