Efficient Fine-Tuning for Structured Sparsity Under Group Repartitioning

Diyang Li
Cornell University
diyang01@cs.cornell.edu

Abstract

Structured sparsity extends classical sparse learning by encouraging sparsity at the level of predefined groups rather than individual parameters. This paradigm is widely adopted in modern machine learning and is typically enforced by sparsity-inducing norms. In real pipelines, the optimal group partition is often unknown a priori and does not remain static after initial training. Parameter groupings are often modified as domain knowledge or upstream representation evolves. Standard approaches require retraining the model from scratch under the new target partition, which discards the optimization effort already invested and becomes computationally prohibitive at scale. In this work, we focus on the group-wise $\ell_{p,1}$-norm, a flexible regularization family that encompasses classical group Lasso and other sparse variants. We develop a principled algorithmic framework that leverages the pre-trained optimum to fine-tune model weights toward a new desired group structure without full retraining. Technically, we decompose arbitrary group-wise transitions into a sequence of tractable primitives and derive the closed-form learning dynamics for each sub-action. Extensive experiments on real-world datasets demonstrate that our method achieves significant acceleration while provably recovering the correct optimal solution.

1 Introduction

Sparsity has long been a workhorse of modern machine learning , enabling model compression, interpretability, and improved generalization in a wide range of regimes [Jordan and Mitchell, 2015, Alpaydin, 2020]. Classical sparse learning typically treats parameters as exchangeable units, encouraging individual entries to become exactly zero independently. In many learning problems, however, parameters naturally exhibit structured patterns beyond element-level selection. Examples include neurons within a hidden layer [Wen et al., 2016], attention heads in a transformer block [Dai et al., 2023], semantic groups in tabular data [Singh and Bedathur, 2023], or correlated signals from the same sensor family [Ahmed et al., 2022]. When the underlying semantics are group-wise structured, selecting or removing parameters via isolated coordinates can be suboptimal, both statistically and computationally [Maurer and Pontil, 2012].

Structured sparsity [Huang et al., 2009, Jenatton et al., 2011] provides a principled framework to exploit such organization by promoting zeros at the level of predefined blocks. One common way to impose this concept is through sparsity-inducing norms [Bach et al., 2012, 2025], which serve as tractable surrogates for the combinatorially hard problem of discrete structure selection. In the group-wise setting, these norms couple parameters within each pre-selected group so that entire groups tend to be kept or eliminated together. The benefits of structured sparsity are well documented [Wainwright, 2014, Bottou et al., 2018], as the resulting models exhibit sparsity patterns respect the intended structure, leading to improved interpretability and more efficient storage and inference.

Despite this maturity, the effectiveness of structured sparsity hinges on the choice of group partition [Bach, 2010]. A well-chosen partition captures the true underlying structure and leads to superior

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

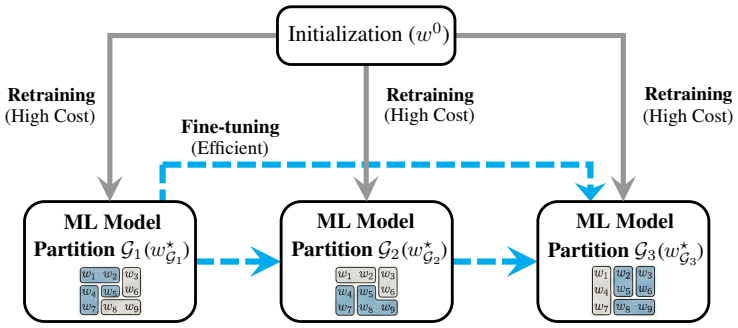

Figure 1: Illustration of partition adaptation setting. Given candidate groupings $\mathcal{G}_1$-$\mathcal{G}_3$, the standard approach (gray arrows) requires retraining from the initial point for each structure. Our fine-tuning strategy (blue arrows) instead leverages the pre-trained $w^\star_{\mathcal{G}_i}$ to efficiently calculate the optimum under new partition $\mathcal{G}_j$.

theoretical and empirical performance. Meanwhile, a key practical complication is that the group partition itself is rarely known a priori, only partially specified, or simply not static after an initial training run [Shervashidze and Bach, 2015, Masoomi et al., 2020]. The space of possible partitions grows super exponentially with the number of parameters. For instance, a 30-dimensional problem admits approximately 8.5×10^{23}, while a 50-dimensional problem yields roughly 1.9×10^{47} candidates [Rota, 1964]. In general, identifying optimal group structure(s) is NP-hard even for moderate parameter dimensions. As a result, practitioners typically rely on domain heuristics or default configurations, later discovering that revisions are necessary when experimenting with alternative groupings [Frecon et al., 2018, Hu et al., 2024].

Even when a group structure has been determined, the partition rarely persists unchanged throughout the model lifecycle, and changing system requirements often necessitate adjustments [Bródka et al., 2011, Lai et al., 2025]. Group definitions may be merged, split, or redefined as domain knowledge evolves [Wang et al., 2020], feature pipelines are refactored [Huang et al., 2009], or upstream representations change [Masoomi et al., 2020]. For example, an engineering team may combine correlated feature templates into a single block [Jia et al., 2023], revise hierarchical taxonomies, or change the granularity at which structured pruning is desired [Hu et al., 2020]. When the target grouping changes, the standard practice is to retrain the model from scratch under the new structure [Sprechmann et al., 2012, Qiao et al., 2017]. This approach discards the computational effort already invested and solves the optimization problem anew. While conceptually straightforward, this approach could be wasteful [Mairal et al., 2010]. The original optimization may have required substantial computational resources, and retraining ignores the valuable information encoded in the pre-trained optimum. For large-scale models or iterative partition refinement workflows, the cumulative cost of repeated retraining becomes computationally expensive [Zhao et al., 2020, Chen et al., 2024].

The above observations reveal a significant gap in the current community: while substantial effort has been devoted to solving structured sparse problems for a fixed partition [Hu et al., 2017, Zhang et al., 2022], comparatively little attention has been paid to what happens when the partition itself is changeable. This work provides an attempt to this exposed question. We study the problem of adapting a pre-trained structured sparse model to new group partitions via efficient fine-tuning (transfer) rather than costly retraining. A schematic illustration of this diagram is provided in Figure 1. Technically, we focus on groupwise sparsity induced by the $\ell_{p,1}$ family of regularizers, a flexible class that includes the classical group Lasso [Yuan and Lin, 2006], logistic group Lasso [Meier et al., 2008], and a range of other structured sparse variants.

To achieve this goal, the underlying challenges are substantial, which perhaps explains why this setting has received little attention in prior work. The main technical difficulty is that the repartitioning changes the geometry of the nonsmooth regularizer and is not a mild perturbation of a trained objective even when the data and loss remain fixed [Blazère et al., 2014, Yang and Lozano, 2017]. This creates a non-trivial circular dependency: to know which groups should be updated, one needs information about the new optimality structure, but to obtain that structure efficiently, one needs to know which groups are relevant [Kwemou, 2016]. Unlike classical warm-start scenarios such as varying the regularization strength [Hu et al., 2017], changing the partition is a discrete structural jump with abrupt support changes even for seemingly simple edits. Consequently, naively running

proximal gradient steps from the pretrained optimum provides limited benefit and offers no guarantee of matching the correct solution. Moreover, the effective accelerations for structured sparse learning such as screening rules [Ndiaye et al., 2016], are tightly tied to the exact group structure and its dual-norm geometry. Once the partition changes, the bounds and certificates are no longer valid and generally cannot be reused without rebuilding, which largely erases the computational advantage of warm starting.

Our Contributions. Our contributions are three-fold. First, on the theoretical side, we develop a mathematical framework for analyzing group-wise partition transitions under $\ell_{p,1}$-regularization. Technically, we identify a minimal set of expressive primitives that act locally on groups and rigorously show that any transition between two arbitrary group partitions can be represented as a finite composition of these 4 primitives. Building on this decomposition, we derive the continuous-time exact evolution flows for the optimizer along a deformation parameter within each primitive. Second, on the algorithmic side, we leverage the above theoretical results to construct an algorithmic framework for structure transfer by fine tuning. The framework integrates the 4 primitives into a unified procedure, where we derive explicit threshold conditions that certify the activation and deactivation of groups during deformation. Computationally, the resulting update reduces to numerically solving a closed-form Volterra integral equation, yielding the equivalently adapted model parameters without running a full iterative retraining procedure. Third, we empirically validate the proposed approach on real-world datasets. Across tasks and partitions, our method consistently recovers the correct optimum and sparsity patterns as retraining baselines while delivering substantial speedups.

Related Works. To the best of our knowledge, no prior work has directly addressed the partition-to-partition model adaptation. The closest early perspective is Huang et al. [2009]. It introduces the notion of dynamic group sparsity in compressive sensing and proposes greedy recovery procedures that can adaptively infer clustered support structures. A similar line of work aims to learn the structure itself from data, where Shervashidze and Bach [2015] studies structured sparsity in a Bayesian framework to infer structure-related parameters, and Frecon et al. [2018] addresses group-wise selection via bilevel optimization for $\ell_{2,1}$-norm, but still targets structure discovery rather than parameter fine-tuning.

2 Preliminaries

Notation. For a positive integer d, let $[d] \coloneqq \{1, \dots, d\}$. For a set $\mathcal{A}$, we denote its cardinality by $|\mathcal{A}|$. For a vector $v \in \mathbb{R}^d$, we write $|v|$ for the elementwise absolute value, and for an index set $\mathcal{A} \subseteq [d]$ we write $v_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ for the subvector restricted to coordinates in $\mathcal{A}$. Similarly, for a differentiable function f we write $\nabla_{\mathcal{A}} f$ for the corresponding subvector of the gradient. We use the standard inner product $\langle a, b\rangle = a^\top b$. For $p \in [1, \infty]$, the $\|\cdot\|_p$ denotes the ℓ_p norm and p^* denotes its dual exponent satisfying $\frac{1}{p} + \frac{1}{p^*} = 1$, with the conventions $1^* = \infty$ and $\infty^* = 1$.

Setup. We are given a dataset $\mathcal{D}$ consisting of n data points, represented by a data matrix $X \in \mathbb{R}^{n\times m}$ and an associated target vector y when available. In the supervised setting we observe $y \in \mathbb{R}^n$, in the semi-supervised setting we observe only a subset of targets $y \in \mathbb{R}^{n_\ell}$ with $n_\ell \le n$, and in the unsupervised setting $y = \varnothing$. Unless otherwise stated, we use the uniform notation $\mathcal{D}$ that does not depend on the supervision regime. Suppose the model weight $w \in \mathbb{R}^d$ to be learned is d-dimensional with $d \ge m$ in typical applications. A group structure (or partition) is a family

$$\mathcal{G} = \{b_1, \dots, b_G\},\ b_g \subseteq [d],$$

forming a disjoint partition of a subset of indices into G groups. That is, for all $i, j \in [G]$ with $i \neq j$ we have $b_i \cap b_j = \varnothing \neq b_i$ and $\bigcup_{g=1}^G b_g \subseteq [d]$. We use $\mathcal{G}^{\complement} = [d] \setminus \bigcup_{g=1}^G b_g$ to collect non-penalized indices, which allows for an unpenalized intercept or other designated parameters. We write $d_g \coloneqq |b_g|$ for the size of group g. Given $\mathcal{G}$ and $p, q \ge 1$, the group-wise $\ell_{p,q}$-norm is defined by $\|w\|_{p,q} \coloneqq (\sum_{g=1}^G \|w_{b_g}\|_p^q)^{1/q}$.

2.1 Learning Objective & Repartitioning Setting

In this paper we focus on the sparsity-inducing case $q = 1$. Let $\mathbf{1}_g \in \mathbb{R}^{d_g}$ denote the all-ones vector on group g. Our scope is the $\ell_{p,1}$-regularized problem

$$\min_{w\in\mathbb{R}^d} \mathbb{F}_{\mathcal{G}}(w) := f(\mathcal{D}; w) + \lambda \sum_{g=1}^{G} \|\mathbf{1}_g\|_{p^*} \left\|w_{b_g}\right\|_p , \tag{1}$$

where $f(\mathcal{D}; \cdot)$ is a data-fitting term and $\lambda > 0$ is a regularization parameter. We call $\|\mathbf{1}_g\|_{p^*} = d_g^{1-1/p}$ in (1) as the (intended) group-wise coefficient. When $p = 2$, the (1) recovers the classic group Lasso penalty $\sum_{g=1}^{G} \sqrt{d_g}\, \|w_{b_g}\|_2$ in Yuan and Lin [2006]. Given a group partition $\mathcal{G}$, let $\mathcal{W}^\star_{\mathcal{G}} := \arg\min_{w\in\mathbb{R}^d} \mathbb{F}_{\mathcal{G}}(w)$ denote the solution set, and let $w^\star_{\mathcal{G}} \in \mathcal{W}^\star_{\mathcal{G}}$ be a chosen optimizer under the group structure $\mathcal{G}$. As illustrated in Figure 1, the central setting of this work is *repartitioning*: after training under an initial structure $\mathcal{G}$, the target structure may change to a different group partition as time evolves or the system state updates, yielding a new structure $\mathcal{G}' = \{b'_1, \ldots, b'_{G'}\}$ consisting of G' groups. Our goal is to compute an optimizer $w^\star_{\mathcal{G}'} \in \mathcal{W}^\star_{\mathcal{G}'}$ by leveraging information from the known $w^\star_{\mathcal{G}}$, rather than solving (1) from scratch under different $\mathcal{G}'$.

Assumption 2.1 (Well-posedness)**.** *Each group in $\mathcal{G}$ and $\mathcal{G}'$ is non-empty. For any partition considered, the objective (1) admits at least one minimizer so $\mathcal{W}^\star_{\mathcal{G}} \neq \varnothing$.*

Throughout, we impose Assumption 2.1. To simplify the presentation, we suppose that $f(\mathcal{D}; \cdot)$ is smooth on $\mathbb{R}^d$ to focus on the nonsmooth geometry induced by the $\ell_{p,1}$ regularizer. In principle, several components extend to non-differentiable losses via Clarke subdifferential calculus [Rockafellar, 1979]. We denote by $\mathcal{A}$ the active set (support) of coordinate-level indices with nonzero optimal weights, and we state the corresponding first-order conditions in Lemma 2.2. Besides, in the nonconvex regime (e.g., neural networks), practical training procedures (SGD, Adam, etc.) typically return approximate stationary points without certificates of global optimality or exact composite-stationarity, even though zero training loss may be theoretically achievable [Jordan and Mitchell, 2015]. Our theory assumes access to an exact stationary point satisfying the strict optimality system in Lemma 2.2; accordingly, our equivalence guarantees should be interpreted at the level of composite stationarity, meaning the fine-tuned solution $w^\star_{\mathcal{G}'}$ satisfies the target stationarity in Lemma 2.2.

Lemma 2.2 (Optimality Conditions [Hu et al., 2017])**.** *If $w^\star$ is a minimizer of (1), then for non-penalized indices we have $\nabla_i f(\mathcal{D}; w^\star) = 0$ for all $i \in \mathcal{G}^{\complement}$. For each penalized group b_g with $g \in [G]$, exactly one of the following two cases holds. If $b_g \nsubseteq \mathcal{A}$ (i.e. inactive group), then $\left\|\nabla_{b_g} f(\mathcal{D}; w^\star)\right\|_{p^*} \leq \lambda \|\mathbf{1}_g\|_{p^*}$. If $b_g \subseteq \mathcal{A}$ (i.e. active group), then*

$$\begin{cases} \left\|\nabla_{b_g} f(\mathcal{D}; w^\star)\right\|_{p^*} = \lambda \|\mathbf{1}_g\|_{p^*} \\ \nabla_{b_g} f(\mathcal{D}; w^\star)^\top w^\star_{b_g} = -\lambda \|\mathbf{1}_g\|_{p^*} \left\|w^\star_{b_g}\right\|_p \end{cases}.$$

3 Theory

This section studies optimization problem (1) under group repartitioning from theory.

3.1 State Primitives

We represent group structures as elements of a discrete state space equipped with a finite set of atomic transitions, which provides a canonical way to factorize an arbitrary repartitioning into local edits that can be analyzed in isolation. We therefore introduce persistent group identifiers and encode a group structure as a labeled state.

Definition 3.1 (Group-structure State)**.** *Fix a finite set of penalized group identifiers $[G+2] \subset \mathbb{N}$, and reserve identifier 0 for the unpenalized group. Given any group-wise coefficient c_g, a group-structure state is a pair $S = (\mathcal{G}, c)$: $\mathcal{G} = \{b_1, \ldots, b_G\}$, $c = (c_1, \ldots, c_G) \in (0, \infty)^G$.*

In (1), $c_g = \|\mathbf{1}_g\|_{p^*} = |b_g|^{1-1/p}$ is the group weight. Repartitioning changes these weights, which enter Algorithm 1 through the coefficient interpolation in Appendix A. This state representation has two key properties. First, at any state, group g is identified with its current membership set b_g

and coefficient c_g. Second, group identities are independent of the current group contents. This is precisely what we need to define a small set of atomic operations that transform one group structure into another. We now define primitive operations, or minimal transformation of $(\mathcal{G}, c)$. Formally, a primitive is a map $\mathcal{P} : \mathcal{S} \to \mathcal{S}$, where $\mathcal{S}$ is the set of all admissible states, and for every $S \in \mathcal{S}$, the successor state $\widetilde{S} = \mathcal{P}(S)$ satisfies

$$\Big|\{\forall g : j \in [d_g] : b_g(j) \neq \tilde{b}_g(j)\}\Big| + \Big|\{\forall g : c_g \neq \tilde{c}_g\}\Big| \leq 1.$$

That is, a primitive $\mathcal{P}$ modifies only one elementary component of $(\mathcal{G}, c)$. There are exactly four primitives, grouped into three semantic categories.

Structural primitive This is the only primitive that changes the combinatorial assignment of coordinates to blocks. The primitive $\texttt{MOVE_FEATURE}(j, g_{\text{dest}})$ takes as input an index $j \in [d]$ and a destination group index $g_{\text{dest}} \in \{0, 1, \ldots, G\}$, where $g_{\text{dest}} = 0$ corresponds to the unpenalized set $\mathcal{G}^{\complement}$. The source group index g_{src} is determined by the current location of j.

Primitive 1 (MOVE_FEATURE)**.** *Action: (i) remove j from its current location, i.e., $b_{g_{src}} \leftarrow b_{g_{src}} \setminus \{j\}$ if $g_{src} \geq 1$, or $\mathcal{G}^{\complement} \leftarrow \mathcal{G}^{\complement} \setminus \{j\}$ if $g_{src} = 0$; (ii) insert j into the destination, i.e., $b_{g_{dest}} \leftarrow b_{g_{dest}} \cup \{j\}$ if $g_{dest} \geq 1$, or $\mathcal{G}^{\complement} \leftarrow \mathcal{G}^{\complement} \cup \{j\}$ if $g_{dest} = 0$; (iii) the coefficients $c_1, \ldots, c_G$ are unchanged.*

Parametric primitive This primitive modifies only the weight associated with a single penalized group while keeping memberships fixed. The primitive $\texttt{ADJUST_COEFF}(g, c_{\text{new}})$ takes as input a group index $g \in [G]$ together with a new coefficient $c_{\text{new}} > 0$.

Primitive 2 (ADJUST_COEFF)**.** *Action: (i) $c_g \leftarrow c_{new}$ while $c_{g'}$ is unchanged for all $g' \neq g$; (ii) the partition $\{b_1, \ldots, b_G\}$ remains unchanged.*

Administrative primitive These primitives manage the identifier set itself and do not reassign any coordinate. They exist to make the transition system closed under changes in the number of blocks, while allowing us to introduce or retire group labels in a controlled way.

The primitive $\texttt{CREATE_GROUP}(c_{\text{new}})$ takes as input a coefficient $c_{\text{new}} > 0$, and similarly, $\texttt{DELETE_GROUP}(g)$ takes as input a group index $g \in [G]$.

Primitive 3 (CREATE_GROUP)**.** *Action: (i) introduce a new group index $G + 1$ with $b_{G+1} \leftarrow \varnothing$ and $c_{G+1} \leftarrow c_{new}$; (ii) the number of groups becomes $G \leftarrow G + 1$.*

Primitive 4 (DELETE_GROUP)**.** *Action: (i) remove b_g and its coefficient c_g from $\mathcal{G}$ and c; (ii) reorganize the identifiers of the remaining groups.*

In the intermediate states, we allow $b_g = \varnothing$ for $g \neq 0$ and interpret such blocks as empty containers. When the group g is empty, for any w we denote $\|w_{b_g}\|_p = 0$.

Definition 3.2 (Reachability)**.** *A state S' is* reachable *from S if there exist $K \in \mathbb{N}$ and a sequence of primitives $S^{(0)} := S$, $S^{(k)} := \mathcal{P}_k\big(S^{(k-1)}\big)$ for $k = 1, \ldots, K$, such that $S^{(K)} = S'$, where each $\mathcal{P}_k$ is an instance of one of Primitives 1-4.*

Theorem 3.3 (Expressivity and Minimality)**.** *For every pair of states $S, S' \in \mathcal{S}$, the state S' is reachable from S using Primitives 1-4. For any subset of the primitive families obtained by removing exactly one of Primitives 1, 2, 3, or 4, there exist states S and S' such that S' is not reachable from S using only the remaining primitives.*

Theorem 3.3 implies that the four primitives form a minimal generating set for arbitrary changes of group structures and coefficients. The state space $\mathcal{S}$ is thus fully connected under the induced reachability relation. In other words, the Primitives 1-4 are sufficient to factorize any change of $\mathcal{G}$ into a finite sequence of atomic steps.

3.2 Group-Structure Distance

To reason about how repartitioning event affect the fine-tuning, we need a quantitative notion of how "far" two group structures are. Let $\mathbf{1}_b \in \mathbb{R}^d$ denote the indicator vector of a block $b \subseteq [d]$, defined by $(\mathbf{1}_b)_i = 1$ if $i \in b$ and 0 otherwise. We set $\widetilde{\mathcal{G}} := \mathcal{G} \cup \{\mathcal{G}^{\complement}\}$ and $\widetilde{\mathcal{G}}' := \mathcal{G}' \cup \{(\mathcal{G}')^{\complement}\}$. Motivated by discrete optimal transport theory [Perrot et al., 2016], we propose a distance that captures the cost of optimally transporting one partition into another.

Definition 3.4 (Group-structure Distance). *For every pair $(b, b') \in \widetilde{\mathcal{G}} \times \widetilde{\mathcal{G}}'$ with $|b| + |b'| > 0$, denote the single block-to-block cost by $c_{p^*}(b, b') := \| |b|^{-\frac{1}{p^*}} \mathbf{1}_b - |b'|^{-\frac{1}{p^*}} \mathbf{1}_{b'} \|_{p^*}$. Let $\mathcal{G}$ and $\mathcal{G}'$ be two admissible group structures on $[d]$. We define their distance as*

$$\operatorname{dist}(\mathcal{G}, \mathcal{G}') := \min_{\{\pi(b,b')\}} \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \pi(b, b')\, c_{p^*}(b, b'), \tag{2}$$

where the penalty-tie relation $\pi(\cdot, \cdot)$ satisfies

$$\sum_{b' \in \widetilde{\mathcal{G}}'} \pi(b, b') = \frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}}, \quad \sum_{b \in \widetilde{\mathcal{G}}} \pi(b, b') = \frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}}.$$

Definition 3.4 can be viewed as an optimal matching between the extended block collections: the coupling π specifies how the normalized mass associated with each block in $\widetilde{\mathcal{G}}$ is distributed across blocks in $\widetilde{\mathcal{G}}'$, and the objective accumulates the corresponding block-to-block costs.

Theorem 3.5. *The $\operatorname{dist}(\cdot, \cdot)$ is a (real-valued) metric on the set of partitions. It is finite for all $\mathcal{G}, \mathcal{G}'$, non-negative, symmetric, and satisfies the identity of indiscernibles $\operatorname{dist}(\mathcal{G}, \mathcal{G}') = 0 \Leftrightarrow \widetilde{\mathcal{G}} = \widetilde{\mathcal{G}}'$, as well as the triangle inequality $\operatorname{dist}(\mathcal{G}, \mathcal{G}'') \le \operatorname{dist}(\mathcal{G}, \mathcal{G}') + \operatorname{dist}(\mathcal{G}', \mathcal{G}'')$ for all $\mathcal{G}, \mathcal{G}', \mathcal{G}''$.*

Theorem 3.5 ensures that $\operatorname{dist}(\mathcal{G}, \mathcal{G}')$ is well-defined for arbitrary admissible $\mathcal{G}$ and $\mathcal{G}'$ and is invariant to relabelings of groups. Beyond its metric properties, the distance function also controls the deviation between solution sets under a standard sharpness condition of objective. To build intuition, we illustrate this stability in a motivating example.

Proposition 3.6 (Solution-set Stability). *Assume $\mathcal{G}^{\complement} = (\mathcal{G}')^{\complement}$ and that, for some $\alpha \ge 1$, $\kappa > 0$, and common bounded region $\mathcal{U} \supseteq \mathcal{W}^\star_{\mathcal{G}} \cup \mathcal{W}^\star_{\mathcal{G}'}$, every $\mathcal{H} \in \{\mathcal{G}, \mathcal{G}'\}$ and $w \in \mathcal{U}$ satisfy $\mathbb{F}_{\mathcal{H}}(w) - \min \mathbb{F}_{\mathcal{H}} \ge \kappa \big(\inf_{u \in \mathcal{W}^\star_{\mathcal{H}}} \|w - u\|_p \big)^{\alpha}$. Then the minimizer sets satisfy*

$$\sup_{w \in \mathcal{W}^\star_{\mathcal{G}}} \inf_{v \in \mathcal{W}^\star_{\mathcal{G}'}} \|w - v\|_p \le \left[\frac{2\lambda R}{\kappa} \Big(d^{2 - \frac{1}{p^*}} + 2 d^{\frac{2}{p^*}} Z \Big) \operatorname{dist}(\mathcal{G}, \mathcal{G}') \right]^{1/\alpha},$$

where $R < \infty$ is a constant such that for all $w \in \mathcal{W}^\star_{\mathcal{G}} \cup \mathcal{W}^\star_{\mathcal{G}'}$, we have $\|w\|_p \le R$, and $Z := \max \left\{ \sum_{b \in \widetilde{\mathcal{G}}} |b|^{1/p^*}, \sum_{b' \in \widetilde{\mathcal{G}}'} |b'|^{1/p^*} \right\}$.

The same bound holds with $\mathcal{G}$ and $\mathcal{G}'$ swapped, and therefore it also bounds the Hausdorff distance between $\mathcal{W}^\star_{\mathcal{G}}$ and $\mathcal{W}^\star_{\mathcal{G}'}$. For example, with $p = 2$, quadratic regression $f(\mathcal{D}; w) = \|Xw - y\|_2^2/(2n)$ and $X^\top X/n \succeq \mu I$ for $\mu > 0$ satisfy the condition with $\alpha = 2$ and $\kappa = \mu/2$, since the group penalty preserves strong convexity. Proposition 3.6 provides the key conceptual justification for our fine-tuning approach: if the repartitioning event is small in the metric $\operatorname{dist}(\mathcal{G}, \mathcal{G}')$, then the target optimizer lies in a controlled neighborhood of the pretrained optimizer, suggesting that one can traverse between them via a carefully designed procedure rather than a full retraining run.

3.3 Primitive-induced Dynamics

This subsection isolates the component of group-structure adaptation by analyzing the learning dynamics induced by each primitive. Suppose that the group structure changes from $\mathcal{G}$ to $\mathcal{G}'$ through a single application of Primitive 1 or 2. The key insight is that under appropriate regularity conditions, the induced primitive affects the optimality system in a rank-structured manner, which allows an exact Volterra integral characterization of its evolution. While these primitives correspond to discrete edits of the state, we analyze its effect on the optimizer through a continuation (quantitative) viewpoint. We view the resulting displacement of $w^\star$ as the path integral of the associated infinitesimal response along an auxiliary (one-dimensional) deformation parameter t.

Theorem 3.7 (Integral Characterization of Primitive 1). *Assume that the sparsity pattern remains invariant under Primitive 1. Fix a `MOVE_FEATURE` primitive instance that transforms $\mathcal{G}$ into $\mathcal{G}'$. Then post-primitive optimum $w^\star_{\mathcal{G}'}$ satisfies that its restriction to $\mathcal{A}$ equals the unique solution of the closed-form Volterra integral equation*

$$w^\star_{\mathcal{G}', \mathcal{A}} = w^\star_{\mathcal{G}, \mathcal{A}} - \int_0^1 \dot{\mathfrak{R}}(\mathcal{P}_{[1]})(w^\star_{\mathcal{G}}, t)\, dt, \tag{3}$$

where the integrand $\grave{\Re}(\mathcal{P}_{[1]})$ is the infinitesimal response at t whose full expression is given in Equation (4) in the appendix. And the resulting $w^\star_{\mathcal{G}'}$ satisfies the optimality conditions in Lemma 2.2.

In Theorem 3.7, the (3) expresses the new optimum as the original optimum plus a *correction term* that accumulates the infinitesimal perturbation of the group membership along the deformation. Since the update is derived from the exact optimality system rather than an approximate gradient step, the computed $w^\star_{\mathcal{G}'}$ is certified to satisfy the optimality conditions (Lemma 2.2) of the post-primitive objective. In practice, the integral of correction term is evaluated numerically (e.g., via quadrature over t), which is substantially cheaper than running iterative batch training until convergence. An analogous result holds for the Primitive 2.

Theorem 3.8 (Integral Characterization of Primitive 2)**.** *Assume the sparsity pattern remains invariant and fix a `ADJUST_COEFF` primitive instance. Then post-primitive optimum $w^\star_{\mathcal{G}'}$ restricted to $\mathcal{A}$ equals the unique solution of the closed-form Volterra integral equation*

$$w^\star_{\mathcal{G}',\mathcal{A}} = w^\star_{\mathcal{G},\mathcal{A}} - \int_0^1 \grave{\Re}(\mathcal{P}_{[2]})(w^\star_{\mathcal{G}}, t)\,dt,$$

where the full expression of integrand $\grave{\Re}(\mathcal{P}_{[2]})$ is in Equation (5) in the appendix, and the resulting $w^\star_{\mathcal{G}'}$ satisfies the first-order optimality conditions stated in Lemma 2.2.

The response $\grave{\Re}(\mathcal{P}_{[2]})$ has a simpler structure than its counterpart $\grave{\Re}(\mathcal{P}_{[1]})$ since Primitive 2 does not alter the combinatorial partition. The fixed-pattern behavior holds locally whenever the primitive perturbation is sufficiently small in the group-structure metric from Definition 3.4, whose condition will be relaxed in Section 4.

Remark 3.9. *We treat the administrative Primitive 3-4 separately, since they induce no change in the $w^\star$ and will only be used for bookkeeping and compositional expressivity.*

3.4 Composite Dynamics

This subsection collapses the singleton dynamics of primitives into a unified integral characterization. Suppose that converting $\mathcal{G}$ into $\mathcal{G}'$ can be achieved using $K \in \mathbb{N}$ applications of Primitive 1-2.

Theorem 3.10 (Commutativity)**.** *Let $\mathcal{P}_1, \ldots, \mathcal{P}_K$ be a sequence of primitives such that every $\mathcal{P}_k$ is either `MOVE_FEATURE`$(j_k, g^{(k)}_{dest})$ with $j_k \in [d]$ and $g^{(k)}_{dest} \in \{0, 1, \ldots, G\}$, or `ADJUST_COEFF`$(g_k, c^{(k)}_{new})$ with $g_k \in [G]$ and $c^{(k)}_{new} > 0$. Define the trajectory $S^{(0)} = S$, $S^{(k)} = \mathcal{P}_k\big(S^{(k-1)}\big)$, $S^{(K)} = \mathcal{P}_K \circ \cdots \circ \mathcal{P}_1(S)$. Fix any $r \in [K-1]$ such that $\mathcal{P}_r$ and $\mathcal{P}_{r+1}$ are of different types. Let $\widetilde{\mathcal{P}}_1, \ldots, \widetilde{\mathcal{P}}_K$ be the sequence obtained by swapping $\mathcal{P}_r$ and $\mathcal{P}_{r+1}$, and define $\widetilde{S}^{(0)} = S$, $\widetilde{S}^{(k)} = \widetilde{\mathcal{P}}_k\big(\widetilde{S}^{(k-1)}\big)$, $\widetilde{S}^{(K)} = \widetilde{\mathcal{P}}_K \circ \cdots \circ \widetilde{\mathcal{P}}_1(S)$. Then $\widetilde{S}^{(K)}$ is well-defined and $\widetilde{S}^{(K)} = S^{(K)}$.*

The Theorem 3.10 formalizes a key property that when two adjacent primitives are of different types, swapping their order does not affect the terminal group-structure state. We introduce auxiliary variables $t_1, \ldots, t_K \in [0, 1]$, where t_k represents the completion level of the k-th primitive application. Collect them into $\mathbf{t} = (t_1, \ldots, t_K)^\top \in [0, 1]^K$ whose parameter space is the K-dimensional unit hypercube. The commutativity ensures that the accumulated correction is independent of the traversal order across different primitive types. We may therefore compute the total correction by

$$w^\star_{\mathcal{G}',\mathcal{A}} = w^\star_{\mathcal{G},\mathcal{A}} - \sum_{i=1}^{K} \int_0^1 \grave{\Re}(\mathcal{P}^{(i)}_{[1 \text{ or } 2]})(w^\star_{\mathcal{G}}, t_i)\,dt_i,$$

where $\mathcal{P}^{(i)}_{[1 \text{ or } 2]}$ is the i-th application of Primitive 1 or 2, and the sum of K line integrals can be interpreted as a contour integral over the edges of the hypercube. Let $u \coloneqq \big(1, 1, \ldots, 1\big)^\top \in \mathbb{R}^K$, $v \coloneqq \frac{u}{\|u\|_2}$. Choose an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{K \times K}$ such that $\boldsymbol{Q}v = e_1$, equivalently, $\boldsymbol{Q}u = \sqrt{K}\, e_1$, where $e_1 = (1, 0, \ldots, 0)^\top$. Define rotated coordinates $\mathbf{s} \coloneqq \boldsymbol{Q}\mathbf{t}$. Under this rotation, the diagonal direction u becomes aligned with the first axis, and the endpoint $\mathbf{t} = \mathbf{1}$ is mapped to $\mathbf{s} = (\sqrt{K}, 0, \ldots, 0)^\top$. A convenient scalar re-parameterization is therefore $\mathbf{t}(t) = \frac{t}{\sqrt{K}}\, u$, $t_k(t) = \frac{t}{\sqrt{K}}$ for $k = 1, \ldots, K$. The corresponding differentials satisfy $d\mathbf{t} = \frac{1}{\sqrt{K}} u\, dt = v\, dt$ and $d\mathbf{s} = \boldsymbol{Q}\, d\mathbf{t} = (\boldsymbol{Q}v)\, dt = e_1\, dt$, which is the sense that the original direction $(1, 1, \ldots, 1)$ is

mapped to a single-axis direction after rotation. This change of coordinates yields a single *aggregated dynamics* $\grave{\Re}(\mathcal{P})(w^\star_{\mathcal{G}}, t)$ that encodes all K primitives simultaneously. Its full expression is given in Equation (6) due to space limit, which provides the theoretical foundation for the algorithmic developments in the next section.

Corollary 3.11 (Fine-tuning Equation)**.** *Assume that the sparsity pattern remains invariant under the Primitive 1, 2. Then the post-transition optimum $w^\star_{\mathcal{G}'}$ has the exact identity*

$$w^\star_{\mathcal{G}'} = \begin{bmatrix} w^\star_{\mathcal{G}',\mathcal{A}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} w^\star_{\mathcal{G},\mathcal{A}} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \int_0^{\sqrt{K}} \grave{\Re}(\mathcal{P})\big(w^\star_{\mathcal{G}}, t\big)dt \\ \mathbf{0} \end{bmatrix}.$$

4 Algorithmic Solution

This section presents a practical fine-tuning procedure in Algorithm 1 that computes the target optimizer $w^\star_{\mathcal{G}'}$ numerically. Algorithm 1 directly applies the aggregated Volterra update from Corollary 3.11 without explicitly enumerating the primitive sequence. In practice, the sparsity pattern may change and we address this by monitoring dual feasibility conditions and updating the active set when violations are detected. For each inactive block g, maintaining $w_{b_g} = \mathbf{0}$ requires the corresponding dual constraint is satisfied. Along the integration trajectory, this condition takes the form

$$\big(\Phi_g(t)\big)^{1/p^*} \le 1, \;\; t \in (0, \sqrt{K}),$$

where $\Phi_g(t)$ is defined in Equation (7) in the appendix. We monitor the shifted quantity $\hat{\Phi}_g(t) := \big(\Phi_g(t)\big)^{1/p^*} - 1$. If $\hat{\Phi}_g(t) \le 0$, the block g is safely inactive. If $\hat{\Phi}_g(t) > 0$, the dual feasibility is violated and the block g must enter the active set.

In implementation, the condition $\hat{\Phi}_g(t) = 0$ is enforced by specifying boundary constraints within the numerical integrator. Each integration step requires evaluating $\grave{\Re}(\mathcal{P})$, which involves solving a linear system of dimension $|\mathcal{A}|$. The procedure is lightweight since all expensive computations are restricted to the current active coordinates, and inactive groups are monitored only through dual-feasibility certificates, yielding an overall complexity that scales favorably compared to solving the optimization problem from scratch. The numerical integration with prescribed boundary conditions falls within the well-studied framework in applied mathematics and numerical optimization [Davis and Rabinowitz, 2007, Brunner, 2017]. Such integrators that incorporate automatic step-size control and boundary detection have been extensively developed and analyzed in the literature [Yang et al., 2020], and they provide mature and computationally efficient machinery for our approach.

Algorithm 1 Partition-Adaptive Fine-Tuning

Input: Pretrained optimum $w^\star_{\mathcal{G}}$, group partitions $\mathcal{G}, \mathcal{G}'$, machine precision $\epsilon > 0$.

$w \leftarrow w^\star_{\mathcal{G}}$, $\mathcal{A} \leftarrow \{i \in [d] : w_i \neq 0\}$, $t \leftarrow 0$.
while $t < \sqrt{K}$ **do**
  Compute the Volterra update

$$w_{\mathcal{A}} \leftarrow w_{\mathcal{A}} - \int_t^{t+h} \grave{\Re}(\mathcal{P})\big(w^\star_{\mathcal{G}}(s), s\big)\, ds,$$

  advance t adaptively with event detection.
  if $\exists\, g : b_g \not\subset \mathcal{A}$ s.t. $\hat{\Phi}_g(t) \ge 0$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \cup b_g$.
  else if $\exists\, g : b_g \subseteq \mathcal{A}$ s.t. $\|w_{b_g}\|_p < \epsilon$ **then**
    $w_{b_g} \leftarrow \mathbf{0}$, $\mathcal{A} \leftarrow \mathcal{A} \setminus b_g$.
  end if
  Refresh $\grave{\Re}(\mathcal{P})$ and Φ_g by $\mathcal{A}$.
end while

Output: Fine-tuned optimum $w^\star_{\mathcal{G}'}$.

5 Empirical Study

In this section, we empirically validate the computational efficiency and theoretical correctness of the proposed framework across diverse real-world benchmarks.

Models and Datasets We conduct experiments on 20 real-world datasets spanning diverse domains, sourced from the OpenML Repository [Vanschoren et al., 2014] and the UCI Repository [Asuncion et al., 2007]. These datasets contain 1,994–58,310 instances and 119–800 input features; see Table 2. All features are standardized to have zero mean and unit variance, and occasional missing values are imputed with zeros. Each dataset is randomly partitioned into 70% for training and 30% for

Table 1: Comparison of mean test loss between batch retraining and Alg. 1 across different repartitioning rounds. Each round isolates a specific structural transition: *Round 1* (splitting: one group into multiple subgroups), *Round 2* (merging: multiple groups into one), *Round 3* (creation: introducing new groups), *Round 4* (deletion: removing existing groups), and *Rounds 5-6* (arbitrary: fully random structural changes). This design ensures coverage of all primitive operations and their compositions. Standard deviations are omitted as they are negligible ($< 10^{-4}$). The mark ✓ indicates that the sparsity pattern of Alg. 1 exactly matches that of batch retraining under the target new structure $\mathcal{G}'$.

Dataset	Norm	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
		Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match
Base Model: Quadratic Loss Regression																			
mtp	$p=3$	6.998	6.998	✓	4.356	4.356	✓	5.641	5.641	✓	7.203	7.203	✓	4.917	4.917	✓	6.482	6.482	✓
($\lambda=0.020$)	$p=4$	7.284	7.284	✓	4.528	4.528	✓	5.872	5.872	✓	7.512	7.512	✓	5.127	5.127	✓	6.741	6.741	✓
fabert	$p=3$	14.07	14.07	✓	10.93	10.93	✓	12.34	12.34	✓	9.876	9.876	✓	15.21	15.21	✓	11.58	11.58	✓
($\lambda=0.015$)	$p=4$	14.55	14.55	✓	11.29	11.29	✓	12.79	12.79	✓	10.24	10.24	✓	15.72	15.72	✓	12.03	12.03	✓
mtp	$p=3$	7.412	7.412	✓	4.681	4.681	✓	6.022	6.022	✓	7.845	7.845	✓	5.231	5.231	✓	6.911	6.911	✓
($\lambda=0.035$)	$p=4$	7.722	7.722	✓	4.855	4.855	✓	6.261	6.261	✓	8.169	8.169	✓	5.447	5.447	✓	7.192	7.192	✓
Base Model: Logistic (Bernoulli) Regression																			
philippine	$p=2$	2.845	2.845	✓	3.912	3.912	✓	2.137	2.137	✓	4.268	4.268	✓	3.054	3.054	✓	2.709	2.709	✓
($\lambda=0.010$)	$p=4$	2.963	2.963	✓	4.071	4.071	✓	2.224	2.224	✓	4.463	4.463	✓	3.176	3.176	✓	2.824	2.824	✓
scm1d	$p=2$	5.766	5.766	✓	4.138	4.138	✓	3.642	3.642	✓	5.104	5.104	✓	3.917	3.917	✓	4.582	4.582	✓
($\lambda=0.012$)	$p=4$	6.012	6.012	✓	4.314	4.314	✓	3.789	3.789	✓	5.321	5.321	✓	4.083	4.083	✓	4.781	4.781	✓
philippine	$p=2$	3.021	3.021	✓	4.104	4.104	✓	2.291	2.291	✓	4.612	4.612	✓	3.246	3.246	✓	2.858	2.858	✓
($\lambda=0.015$)	$p=4$	3.149	3.149	✓	4.283	4.283	✓	2.392	2.392	✓	4.828	4.828	✓	3.388	3.388	✓	2.999	2.999	✓
Base Model: Gamma Regression																			
USPS	$p=2$	9.417	9.417	✓	7.863	7.863	✓	11.52	11.52	✓	8.904	8.904	✓	12.37	12.37	✓	9.108	9.108	✓
($\lambda=0.008$)	$p=3$	9.734	9.734	✓	8.121	8.121	✓	11.89	11.89	✓	9.206	9.206	✓	12.81	12.81	✓	9.401	9.401	✓
gas-drift	$p=2$	18.96	18.96	✓	13.87	13.87	✓	17.33	17.33	✓	12.74	12.74	✓	16.42	16.42	✓	14.51	14.51	✓
($\lambda=0.006$)	$p=5$	19.72	19.72	✓	14.39	14.39	✓	18.04	18.04	✓	13.21	13.21	✓	17.08	17.08	✓	15.06	15.06	✓
gas-drift	$p=2$	18.11	18.11	✓	13.20	13.20	✓	16.52	16.52	✓	12.14	12.14	✓	15.64	15.64	✓	13.85	13.85	✓
($\lambda=0.008$)	$p=4$	18.92	18.92	✓	13.72	13.72	✓	17.23	17.23	✓	12.60	12.60	✓	16.31	16.31	✓	14.39	14.39	✓

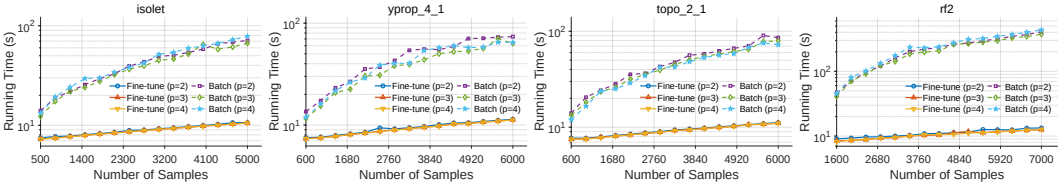

Figure 2: Runtime versus sample size across three models: quadratic loss (left), logistic regression (2 center panels), and Gamma regression (right). Each subplot is labeled with its dataset name in the top-right corner.

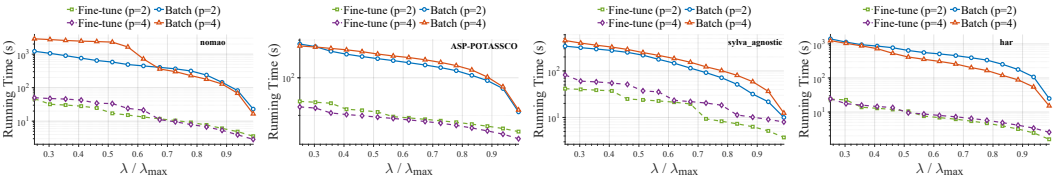

Figure 3: Runtime versus regularization parameter λ across three models: quadratic loss (left), logistic regression (2 center panels), and Gamma regression (right). Each subplot is labeled with its dataset name in the top-right corner.

evaluation. We instantiate the framework with three commonly used losses, i.e., quadratic-loss regression, logistic (Bernoulli) regression, and Gamma regression. Importantly, all three settings fit into the composite objective studied throughout the paper, allowing us to evaluate whether the proposed dynamics accurately track the optimizer under structural changes.

In each structural update (referred to as a *round*), we generate a random initial group structure by sampling both the number of groups and the feature-to-group assignments uniformly at random, producing an initial $\mathcal{G}$. We emphasize that the random regrouping protocol is a *controlled stress test* that isolates the effect of partition edits without confounding application-specific semantics, rather than a literal model of how partitions arise in practice. We first obtain an optimizer $w^\star_{\mathcal{G}}$ by solving the corresponding group-regularized problem. We then generate an updated group structure $\mathcal{G}'$ (again fully randomized, and therefore typically distinct from $\mathcal{G}$), and the task is to obtain the corresponding post-update optimizer $w^\star_{\mathcal{G}'}$. This randomized protocol intentionally creates a wide range of group configurations, ensuring that the evaluation is not tied to a particular hand-designed partitioning scheme.

To our knowledge, there is no established baseline designed specifically for efficiently updating group-structured solutions under changing partitions. Therefore, we compare directly to the natural reference method: naive batch retraining, i.e., solving for $w^\star_{\mathcal{G}'}$ from scratch using the same underlying optimization routine and stopping criterion. This baseline represents the only currently viable

alternative and serves as an upper bound on computational cost. We do *not* warm-start the batch baseline since the repartitioning shifts the loss landscape and active set $\mathcal{A}$, often placing the pretrained iterate in a different basin; see Appendix J for a controlled comparison.

Results and Discussion We first validate the theoretical derivations through controlled numerical experiments. As shown in Table 1, each reported value represents the average over 20 independent runs. The results confirm that our fine-tuning procedure consistently recovers the correct solution under dynamic group structures, achieving test error statistically indistinguishable from that of the batch retraining baseline. These findings provide direct evidence that the integral characterization from Corollary 3.11 accurately captures the optimizer displacement induced by group-structure changes, and that the resulting updated solution satisfies the target optimality system in practice. Having established correctness, we turn to computational efficiency. Figure 2 illustrates the runtime comparison between our method and the batch retraining baseline. The proposed approach yields substantial speedups across datasets, and the advantage becomes more pronounced as the sample size grows. This scaling trend is consistent with the motivation of our framework: rather than repeatedly resolving the full optimization problem after each structural change, we update the optimizer by leveraging the local structure of the optimality conditions and the typically low-dimensional active subspace induced by group sparsity.

We further examine the performance across varying regularization strengths λ, which controls the overall sparsity level of the model. As shown in Figure 3, our algorithm maintains its computational advantage across a wide range of sparsity patterns. As λ increases and the model weights approach the all-zero solution, the computational gains naturally diminish since both methods converge to trivial solutions with minimal computation. Outside this corner case, however, the update remains consistently more efficient than naive retraining. Additional experimental results are provided in Appendix J due to the space limit.

Acknowledgment

I would like to thank Yahya Sattar for insightful discussions.

References

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Steve Dai, Hasan Genc, Rangharajan Venkatesan, and Brucek Khailany. Efficient transformer inference with statically structured sparse attention. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023.

Rajat Singh and Srikanta Bedathur. Embeddings for tabular data: A survey. *arXiv preprint arXiv:2302.11777*, 2023.

Irfan Ahmed, Amaad Khalil, Ishtiaque Ahmed, and Jaroslav Frnda. Sparse signal representation, sampling, and recovery in compressive sensing frameworks. *IEEE Access*, 10:85002–85018, 2022.

Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13:671–690, 2012.

Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 417–424, 2009.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. 2012.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2025.

Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1(1):233–253, 2014.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Francis Bach. Structured sparsity-inducing norms through submodular functions. *Advances in Neural Information Processing Systems*, 23, 2010.

Nino Shervashidze and Francis Bach. Learning the structure for structured sparsity. *IEEE Transactions on Signal Processing*, 63(18):4894–4902, 2015.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. *Advances in Neural Information Processing Systems*, 33:13374–13386, 2020.

Gian-Carlo Rota. The number of partitions of a set. *The American Mathematical Monthly*, 71(5): 498–504, 1964.

Jordan Frecon, Saverio Salzo, and Massimiliano Pontil. Bilevel learning of the group lasso structure. *Advances in neural information processing systems*, 31, 2018.

Yaohua Hu, Xinlin Hu, Carisa Kwok Wai Yu, and Jing Qin. Joint sparse optimization: lower-order regularization method and application in cell fate conversion. *Inverse Problems*, 40(9):095003, 2024.

Piotr Bródka, Stanislaw Saganowski, and Przemyslaw Kazienko. Group evolution discovery in social networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 247–253. IEEE, 2011.

Guannan Lai, Yujie Li, Xiangkun Wang, Junbo Zhang, Tianrui Li, and Xin Yang. Order-robust class incremental learning: Graph-driven dynamic similarity grouping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4894–4904, 2025.

Yingjie Wang, Hong Chen, Feng Zheng, Chen Xu, Tieliang Gong, and Yanhong Chen. Multi-task additive models for robust estimation and automatic structure discovery. *Advances in Neural Information Processing Systems*, 33:11744–11755, 2020.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

Xinlin Hu, Yaohua Hu, Fanjie Wu, Ricky Wai Tak Leung, and Jing Qin. Integration of single-cell multi-omics for gene regulatory network inference. *Computational and structural biotechnology journal*, 18:1925–1938, 2020.

Pablo Sprechmann, Alex Bronstein, and Guillermo Sapiro. Learning efficient structured sparse models. In *Proceedings of the 29th International Conference on Machine Learning*, pages 219–226, 2012.

Lin-bo Qiao, Bo-feng Zhang, Jin-shu Su, and Xi-cheng Lu. A systematic review of structured sparse learning. *Frontiers of Information Technology & Electronic Engineering*, 18(4):445–463, 2017.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

Peilin Zhao, Dayong Wang, Pengcheng Wu, and Steven CH Hoi. A unified framework for sparse online learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(5):1–20, 2020.

Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. Cost-sensitive sparse group online learning for imbalanced data streams. *Machine Learning*, 113(7):4407–4444, 2024.

Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via lp, q regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.

Xiaoqin Zhang, Jingjing Zheng, Di Wang, Guiying Tang, Zhengyuan Zhou, and Zhouchen Lin. Structured sparsity optimization with non-convex surrogates of $\ell_{2,0}$-norm: A unified algorithmic framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6386–6402, 2022.

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

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):53–71, 2008.

Mélanie Blazère, Jean-Michel Loubes, and Fabrice Gamboa. Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory*, 60(4):2303–2318, 2014.

Eunho Yang and Aurélie C Lozano. Sparse+ group-sparse dirty models: Statistical guarantees without unreasonable conditions and a case for non-convexity. In *International Conference on Machine Learning*, pages 3911–3920. PMLR, 2017.

Marius Kwemou. Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. *ESAIM: Probability and Statistics*, 20:309–331, 2016.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. *Advances in neural information processing systems*, 29, 2016.

R Tyrrell Rockafellar. Directionally lipschitzian functions and subdifferential calculus. *Proceedings of the London Mathematical Society*, 3(2):331–355, 1979.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

Hermann Brunner. *Volterra integral equations: an introduction to theory and applications*, volume 30. Cambridge University Press, 2017.

Won Y Yang, Wenwu Cao, Jaekwon Kim, Kyung W Park, Ho-Hyun Park, Jingon Joung, Jong-Suk Ro, Han L Lee, Cheol-Ho Hong, and Taeho Im. *Applied numerical methods using MATLAB*. John Wiley & Sons, 2020.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Reinhold Meise and Dietmar Vogt. *Introduction to functional analysis*. Clarendon press, 1997.

Jingnan Chen, Gengling Dai, and Ning Zhang. An application of sparse-group lasso regularization to equity portfolio optimization and sector selection. *Annals of Operations Research*, 284(1): 243–262, 2020.

Simon B Eickhoff, BT Thomas Yeo, and Sarah Genon. Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686, 2018.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. Kegg: biological systems database as a model of the real world. *Nucleic acids research*, 53(D1): D672–D677, 2025.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1 (3):127–239, 2014.

Appendix

The appendix is organized as follows.

A Formal Definitions **14**

A.1 Definitions of $\grave{\mathfrak{R}}(\mathcal{P}_{[1]})$ and $\grave{\mathfrak{R}}(\mathcal{P}_{[2]})$ 14

A.2 Definition of $\grave{\mathfrak{R}}(\mathcal{P})$ 15

A.3 Threshold Condition 17

B Limitations **18**

C Proof of Theorem 3.3 **18**

C.1 Expressivity 18

C.2 Minimality 22

D Proof of Theorem 3.5 **25**

D.1 Non-negativity 26

D.2 Finiteness 27

D.3 Symmetry 28

D.4 Identity of Indiscernibles 29

D.5 Triangle Inequality 31

E Proof of Proposition 3.6 **34**

F Proof of Theorem 3.7 **38**

G Proof of Theorem 3.8 **42**

H Proof of Theorem 3.10 **47**

H.1 More Motivating Scenarios 49

I Implementation Details **50**

J Additional Results **50**

A Formal Definitions

A.1 Definitions of $\grave{\mathfrak{R}}(\mathcal{P}_{[1]})$ and $\grave{\mathfrak{R}}(\mathcal{P}_{[2]})$

For simplicity, we omit the superscript * from w when there is no ambiguity. We define the two primitive norm-values as

$$r_{\text{src}}(t) := \left\| \left[w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}},\ (1-t) w_j \right] \right\|_p, \quad r_{\text{dest}}(t) := \left\| \left[w_{b_{g_{\text{dest}}} \cap \mathcal{A}},\ t w_j \right] \right\|_p \qquad (g_{\text{dest}} \geq 1).$$

We also define the two numerator vectors, indexed by the corresponding active coordinates and embedded into $\mathbb{R}^{|\mathcal{A}|}$ by zero-filling off the shown indices:

$$a_{\text{src}}(t) := \begin{bmatrix} \left| w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}} \right|^{p-2} \odot w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}} \\ (1-t)^p \left| w_j \right|^{p-2} w_j \\ \mathbf{0} \end{bmatrix}_{|\mathcal{A}| \times 1}, \ a_{\text{dest}}(t) := \begin{bmatrix} \left| w_{b_{g_{\text{dest}}} \cap \mathcal{A}} \right|^{p-2} \odot w_{b_{g_{\text{dest}}} \cap \mathcal{A}} \\ t^p \left| w_j \right|^{p-2} w_j \\ \mathbf{0} \end{bmatrix}_{|\mathcal{A}| \times 1} \qquad (g_{\text{dest}} \geq 1).$$

The complete expression for $\grave{\mathfrak{R}}(\mathcal{P}_{[1]})$ is given by

$$\begin{aligned}
\grave{\mathfrak{R}}(\mathcal{P}_{[1]}) = \Bigg(& \lambda \sum_{\substack{g=1 \\ g \neq g_{\text{src}}, g_{\text{dest}}}}^{G} c_g \\
& \begin{bmatrix} \frac{p}{p^*} \left(\frac{\text{Diag}\left(\left| w_{b_g \cap \mathcal{A}} \right|^{p-2} \right)}{\left\| w_{b_g \cap \mathcal{A}} \right\|_p^{p-1}} - \frac{\left(\left| w_{b_g \cap \mathcal{A}} \right|^{p-2} \odot w_{b_g \cap \mathcal{A}} \right) \left(\left| w_{b_g \cap \mathcal{A}} \right|^{p-2} \odot w_{b_g \cap \mathcal{A}} \right)^\top}{\left\| w_{b_g \cap \mathcal{A}} \right\|_p^{2p-1}} \right) & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_{\mathcal{A} \setminus b_g} \end{bmatrix} \\
& + \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f(\mathcal{D}; w)|_{w=w^\star} + \lambda c_{g_{\text{src}}} \\
& \begin{bmatrix} \frac{p}{p^*} \left(\frac{\text{Diag}\left(\left| w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}} \right|^{p-2}, (1-t)^p \left| w_j \right|^{p-2} \right)}{\left(r_{\text{src}}(t) \right)^{p-1}} - \frac{a_{\text{src}}(t) a_{\text{src}}(t)^\top}{\left(r_{\text{src}}(t) \right)^{2p-1}} \right) & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_{\mathcal{A} \setminus b_{g_{\text{src}}}} \end{bmatrix} \\
& + \lambda c_{g_{\text{dest}}} \begin{bmatrix} \frac{p}{p^*} \left(\frac{\text{Diag}\left(\left| w_{b_{g_{\text{dest}}} \cap \mathcal{A}} \right|^{p-2}, t^p \left| w_j \right|^{p-2} \right)}{\left(r_{\text{dest}}(t) \right)^{p-1}} - \frac{a_{\text{dest}}(t) a_{\text{dest}}(t)^\top}{\left(r_{\text{dest}}(t) \right)^{2p-1}} \right) & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_{\mathcal{A} \setminus \left(b_{g_{\text{dest}}} \cup \{j\} \right)} \end{bmatrix} \Bigg)^{-1} \\
& \cdot \left(\lambda c_{g_{\text{src}}} \left(\frac{p}{p^*} (1-t)^{p-1} \frac{|w_j|^p}{\left(r_{\text{src}}(t) \right)^{2p-1}} a_{\text{src}}(t) - p(1-t)^{p-1} \frac{|w_j|^{p-2} w_j}{\left(r_{\text{src}}(t) \right)^{p-1}} e_j \right) \right. \\
& \left. + \lambda c_{g_{\text{dest}}} \left(p t^{p-1} \frac{|w_j|^{p-2} w_j}{\left(r_{\text{dest}}(t) \right)^{p-1}} e_j - \frac{p}{p^*} t^{p-1} \frac{|w_j|^p}{\left(r_{\text{dest}}(t) \right)^{2p-1}} a_{\text{dest}}(t) \right) \right),
\end{aligned} \tag{4}$$

where the brown terms can be removed if $g_{\text{dest}} = 0$, $\mathbf{O}$ denotes a zero matrix whose subscript indicates its size corresponding to the associated index set, and $\text{Diag}(\cdot)$ denotes the operator that constructs a diagonal matrix from a vector. The complete expression for $\grave{\mathfrak{R}}(\mathcal{P}_{[2]})$ is given by

$$\grave{\mathfrak{R}}(\mathcal{P}_{[2]}) = \lambda (c_{\text{new}} - c_g) \left[\lambda \left(c_g - t c_g + t c_{\text{new}} \right) \mathring{\text{A}} + \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f(\mathcal{D}; w)|_{w=w^\star} + \lambda \sum_{\substack{g'=1 \\ g' \neq g}}^{G} c_{g'} \mathring{\text{A}} \right]^{-1} \begin{bmatrix} \mathbf{0} \\ \vdots \\ \frac{w_{\mathcal{A} \cap b_g} \odot \left| w_{\mathcal{A} \cap b_g} \right|^{p-2}}{\left\| w_{\mathcal{A} \cap b_g} \right\|_p^{p-1}} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \tag{5}$$

where for each g' the matrix $\mathring{\mathrm{A}} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ has entries

$$
\left[\,\mathring{\mathrm{A}}\,\right]_{ij} = \begin{cases} \dfrac{p}{p^*}\left(\dfrac{|w_i|^{p-2}}{\left\|w_{b_{g'}}\right\|_p^{p-1}} - \dfrac{\left(|w_i|^{p-2}\,w_i\right)^2}{\left\|w_{b_{g'}}\right\|_p^{2p-1}}\right), & i=j\in\mathcal{A}\cap b_{g'}, \\ -\dfrac{p}{p^*}\dfrac{\left(|w_i|^{p-2}\,w_i\right)\left(|w_j|^{p-2}\,w_j\right)}{\left\|w_{b_{g'}}\right\|_p^{2p-1}}, & i\neq j,\;\; i,j\in\mathcal{A}\cap b_{g'}, \\ 0, & \text{otherwise.} \end{cases}
$$

On the invertibility in $\grave{\Re}(\mathcal{P}_{[i]})$. The response $\grave{\Re}(\mathcal{P})$ in Theorems 3.7 and 3.8 and Corollary 3.11 requires solving a linear system on the active subspace, and our derivations implicitly assume the active-set-restricted Jacobian is non-singular. For quadratic loss with $p = 2$, this is **automatic** whenever $w^\star$ is a locally unique minimizer on a fixed active manifold. For general losses and arbitrary p, nonsingularity does not follow automatically and is imposed as an explicit **regularity condition**. Across thousands of simulation runs we did *not* encounter numerical issues from near-singularity, suggesting this regularity is very mild in practice. In the rare degenerate case, we replace the inverse with the Moore–Penrose pseudoinverse, which yields the minimum-norm least-squares solution and remains consistent with Lemma 2.2, since null-space components are absorbed by the active-set monitoring in Algorithm 1.

A.2 Definition of $\grave{\Re}(\mathcal{P})$

For a vector $\mathbf{v} \in \mathbb{R}^d$, we define its second-order tensor power as $\mathbf{v}^{\otimes 2} \triangleq \mathbf{v}\otimes\mathbf{v}$. Fix a finite persistent identifier pool $[\bar{G}] \subset \mathbb{N}$ that is large enough to contain all penalized groups appearing before and after the K primitives. New and deleted groups are represented by allowing some $b_g = \emptyset$ or $b'_g = \emptyset$.

Fix $t \in [0, \sqrt{K}]$. The full expression for $\grave{\Re}(\mathcal{P})$ takes the following form

$$
\begin{aligned}
\grave{\Re}(\mathcal{P})\,(w^\star_{\mathcal{G}}, t) = \Bigg[& \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}}\, f(\mathcal{D}; w)\big|_{w=w^\star} + \lambda \sum_{g=1}^{\bar{G}} \frac{c_g p}{p^*} \Bigg(\Bigg(\sum_{j\in\mathcal{A}\cap\left(b_g\cap b'_g\right)} |w_j(t)|^p \\
& + \left(1-\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b_g\setminus b'_g\right)} |w_j(t)|^p + \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b'_g\setminus b_g\right)} |w_j(t)|^p \Bigg)^{-1/p^*} \\
& \cdot \Bigg(\left(1-\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b_g\setminus b'_g\right)} |w_j(t)|^{p-2}\, e_j e_j^\top + \sum_{j\in\mathcal{A}\cap\left(b_g\cap b'_g\right)} |w_j(t)|^{p-2}\, e_j e_j^\top \\
& + \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b'_g\setminus b_g\right)} |w_j(t)|^{p-2}\, e_j e_j^\top \Bigg) - \Bigg(\sum_{j\in\mathcal{A}\cap\left(b_g\cap b'_g\right)} |w_j(t)|^p \\
& + \left(1-\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b_g\setminus b'_g\right)} |w_j(t)|^p + \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b'_g\setminus b_g\right)} |w_j(t)|^p \Bigg)^{-1/p^*-1} \\
& \cdot \Bigg\langle \sum_{j\in\mathcal{A}\cap\left(b_g\cap b'_g\right)} |w_j(t)|^{p-2}\cdot\; w_j(t) e_j + \left(1-\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b_g\setminus b'_g\right)} |w_j(t)|^{p-2}\, w_j(t) e_j \\
& + \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j\in\mathcal{A}\cap\left(b'_g\setminus b_g\right)} |w_j(t)|^{p-2}\, w_j(t) e_j \Bigg\rangle^{\otimes 2} \Bigg) \left(\frac{t}{\sqrt{K}} \left(\frac{c'_g}{c_g} - 1 \right) + 1 \right) \Bigg]^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \lambda \sum_{g=1}^{\bar{G}} \Bigg[\frac{c'_g - c_g}{\sqrt{K}} \Bigg(\sum_{j \in \mathcal{A} \cap \left(b_g \cap b'_g\right)} |w_j(t)|^p + \left(1 - \frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^p \\
&+ \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^p \Bigg)^{-1/p^*} \cdot \Bigg(\left(1 - \frac{t}{\sqrt{K}}\right)^p \cdot \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^{p-2} w_j(t) e_j \\
&+ \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^{p-2} w_j(t) e_j + \sum_{j \in \mathcal{A} \cap \left(b_g \cap b'_g\right)} |w_j(t)|^{p-2} w_j(t) e_j \Bigg) \\
&+ \Bigg(\frac{t\left(c'_g - c_g\right)}{\sqrt{K}} \quad + c_g \Bigg) \Bigg(-\frac{1}{p^*} \Bigg(\left(\frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^p \\
&+ \left(1 - \frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^p + \sum_{j \in \mathcal{A} \cap \left(b_g \cap b'_g\right)} |w_j(t)|^p \Bigg)^{-1/p^* - 1} \\
&\cdot \Bigg(\left(\frac{t}{\sqrt{K}}\right)^{p-1} \cdot \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^p - \left(1 - \frac{t}{\sqrt{K}}\right)^{p-1} \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^p \Bigg) \\
&\cdot \frac{p}{\sqrt{K}} \Bigg(\sum_{j \in \mathcal{A} \cap \left(b_g \cap b'_g\right)} |w_j(t)|^{p-2} w_j(t) e_j + \left(1 - \frac{t}{\sqrt{K}}\right)^p \\
&\cdot \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^{p-2} w_j(t) e_j + \left(\frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^{p-2} w_j(t) e_j \Bigg) \\
&+ \frac{p}{\sqrt{K}} \cdot \Bigg(\left(\frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^p + \left(1 - \frac{t}{\sqrt{K}}\right)^p \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^p \\
&+ \sum_{j \in \mathcal{A} \cap \left(b_g \cap b'_g\right)} |w_j(t)|^p \Bigg)^{-1/p^*} \Bigg(\left(\frac{t}{\sqrt{K}}\right)^{p-1} \sum_{j \in \mathcal{A} \cap \left(b'_g \setminus b_g\right)} |w_j(t)|^{p-2} w_j(t) e_j \\
&- \left(1 - \frac{t}{\sqrt{K}}\right)^{p-1} \sum_{j \in \mathcal{A} \cap \left(b_g \setminus b'_g\right)} |w_j(t)|^{p-2} w_j(t) e_j \Bigg) \Bigg) \Bigg] . \qquad (6)
\end{aligned}
$$

Vectorization. We can compress the expression in (6) into a more organized block-matrix form. To this end, we adopt the notation $c_g = |b_g|^{1/p^*}$ and[1] $c'_g = \left|b'_g\right|^{1/p^*}$. We define the $|\mathcal{A}| \times \bar{G}$ membership-weight matrix $\mathbf{M}(t)$ by its entries

$$
\mathbf{M}_{j,g}(t) = \mathbf{1}_{\left\{j \in b_g \cap b'_g\right\}} + \left(1 - \frac{t}{\sqrt{K}}\right)^p \mathbf{1}_{\left\{j \in b_g \setminus b'_g\right\}} + \left(\frac{t}{\sqrt{K}}\right)^p \mathbf{1}_{\left\{j \in b'_g \setminus b_g\right\}}, \quad j \in \mathcal{A}, g \in [\bar{G}],
$$

and we define $\widehat{\mathbf{M}}(t)$ by its entries

$$
\widehat{\mathbf{M}}_{j,g}(t) = -\frac{p}{\sqrt{K}} \left(1 - \frac{t}{\sqrt{K}}\right)^{p-1} \mathbf{1}_{\left\{j \in b_g \setminus b'_g\right\}} + \frac{p}{\sqrt{K}} \left(\frac{t}{\sqrt{K}}\right)^{p-1} \mathbf{1}_{\left\{j \in b'_g \setminus b_g\right\}}.
$$

With these definitions, we have the vectorized form of (6) as

$$
\begin{aligned}
&\dot{\Re}(\mathcal{P})\left(w^\star_{\mathcal{G}}, t\right) \\
&= \lambda \operatorname{Diag}\left(|w_{\mathcal{A}}(t)|^{p-2} \odot w_{\mathcal{A}}(t)\right) \left[\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f(\mathcal{D}; w)\right|_{w = w^\star}
\end{aligned}
$$

[1] Observe that if $|b_g| = \left|b'_g\right|$ then $c'_g - c_g = 0$ automatically.

$$
\begin{aligned}
&+\lambda\frac{p}{p^*}\left[\operatorname{Diag}\left(|w_{\mathcal{A}}(t)|^{p-2}\odot\left(\mathbf{M}(t)\left(\left[\begin{array}{c} c_1+\frac{t}{\sqrt{K}}(c_1'-c_1)\\ \vdots\\ c_{\bar{G}}+\frac{t}{\sqrt{K}}(c_{\bar{G}}'-c_{\bar{G}})\end{array}\right]\right.\right.\right.\right.\\
&\left.\left.\left.\left.\odot\left(\mathbf{M}(t)^\top|w_{\mathcal{A}}(t)|^p\right)^{-1/p^*}\right)\right)\right)-\operatorname{Diag}\left(|w_{\mathcal{A}}(t)|^{p-2}\odot w_{\mathcal{A}}(t)\right)\mathbf{M}(t)\operatorname{Diag}\\
&\left(\left(\left[\begin{array}{c} c_1+\frac{t}{\sqrt{K}}(c_1'-c_1)\\ \vdots\\ c_{\bar{G}}+\frac{t}{\sqrt{K}}(c_{\bar{G}}'-c_{\bar{G}})\end{array}\right]\odot\left(\mathbf{M}(t)^\top|w_{\mathcal{A}}(t)|^p\right)^{-1/p^*-1}\right)\right.\\
&\left.\mathbf{M}(t)^\top\operatorname{Diag}\left(|w_{\mathcal{A}}(t)|^{p-2}\odot w_{\mathcal{A}}(t)\right)\right]\Bigg]^{-1}\\
&\left[\mathbf{M}(t)\left(\frac{1}{\sqrt{K}}\left[\begin{array}{c} c_1'-c_1\\ \vdots\\ c_{\bar{G}}'-c_{\bar{G}}\end{array}\right]\odot\left(\mathbf{M}(t)^\top|w_{\mathcal{A}}(t)|^p\right)^{-1/p^*}\right)\right.\\
&+\widehat{\mathbf{M}}(t)\left(\left[\begin{array}{c} c_1+\frac{t}{\sqrt{K}}(c_1'-c_1)\\ \vdots\\ c_{\bar{G}}+\frac{t}{\sqrt{K}}(c_{\bar{G}}'-c_{\bar{G}})\end{array}\right]\odot\left(\mathbf{M}(t)^\top|w_{\mathcal{A}}(t)|^p\right)^{-1/p^*}\right)\\
&\left.-\frac{1}{p^*}\mathbf{M}(t)\left(\left[\begin{array}{c} c_1+\frac{t}{\sqrt{K}}(c_1'-c_1)\\ \vdots\\ c_{\bar{G}}+\frac{t}{\sqrt{K}}(c_{\bar{G}}'-c_{\bar{G}})\end{array}\right]\odot\left(\mathbf{M}(t)^\top|w_{\mathcal{A}}(t)|^p\right)^{-1/p^*-1}\odot\left(\widehat{\mathbf{M}}(t)^\top|w_{\mathcal{A}}(t)|^p\right)\right)\right].
\end{aligned}
$$

A.3 Threshold Condition

$$
\begin{aligned}
\Phi_g(t)=&\sum_{j\in b_g'\setminus b_g}\left|\frac{\Psi_j(t)}{\left(c_g+\frac{t}{\sqrt{K}}\left(c_g'-c_g\right)\right)\left(\frac{t\lambda}{\sqrt{K}}\right)}\right|^{p^*}\\
&+\sum_{j\in b_g\setminus b_g'}\left|\frac{\Psi_j(t)}{\lambda\left(c_g+\frac{t}{\sqrt{K}}\left(c_g'-c_g\right)\right)\left(1-\frac{t}{\sqrt{K}}\right)}\right|^{p^*}\\
&+\sum_{j\in b_g\cap b_g'}\left|\frac{\Psi_j(t)}{\lambda\left(\frac{t}{\sqrt{K}}\left(c_g'-c_g\right)+c_g\right)}\right|^{p^*},
\end{aligned}
\tag{7}
$$

where

$$
\begin{aligned}
\Psi(t):&=\nabla f(\mathcal{D};w(t))+\lambda\sum_{\{h:\Lambda_h(w,t)>0\}}\left(c_h+\frac{t}{\sqrt{K}}\left(c_h'-c_h\right)\right)\Theta_h,\\
\Lambda_g(w,t):&=\left(\sum_{j\in b_g\cap b_g'}|w_j|^p+\left(1-\frac{t}{\sqrt{K}}\right)^p\sum_{j\in b_g\setminus b_g'}|w_j|^p+\left(\frac{t}{\sqrt{K}}\right)^p\sum_{j\in b_g'\setminus b_g}|w_j|^p\right)^{1/p},\\
[\Theta_g]_j:&=\frac{|w_j|^{p-2}w_j}{(\Lambda_g(w,t))^{p-1}}\left[\mathbf{1}_{\left\{j\in b_g\cap b_g'\right\}}+\left(1-\frac{t}{\sqrt{K}}\right)^p\mathbf{1}_{\left\{j\in b_g\setminus b_g'\right\}}+\left(\frac{t}{\sqrt{K}}\right)^p\mathbf{1}_{\left\{j\in b_g'\setminus b_g\right\}}\right].
\end{aligned}
$$

B Limitations

Our framework has several limitations. The analysis is restricted to the $\ell_{p,1}$ family with disjoint groups and does not directly cover overlapping, hierarchical, or matrix-valued penalties. The smoothness assumption on f excludes nonsmooth losses such as hinge or absolute deviation, which would require a Clarke-subdifferential treatment, and in nonconvex regimes our guarantees hold only at the level of composite stationarity. Finally, we focus on a single partition-to-partition transition, leaving streaming or online approximate repartitioning to future work.

Future work. A promising direction is to relax the non-overlapping group assumption, thereby accommodating a broader class of real-world applications such as hierarchical or graph-structured sparsity. Another avenue is to extend the framework to nonconvex models, e.g., deep networks, by leveraging mode connectivity and loss landscape geometry to characterize fine-tuning dynamics between local minima.

C Proof of Theorem 3.3

C.1 Expressivity

We first prove the expressivity, namely that for every pair of states $S, S' \in \mathcal{S}$, the state S' is reachable from S using Primitives 1-4.

Proof. Fix arbitrary states

$$S = (\mathcal{G}, c), \quad \mathcal{G} = \{b_1, \dots, b_G\}, c \in (0,\infty)^G$$

and

$$S' = (\mathcal{G}', c'), \quad \mathcal{G}' = \{b'_1, \dots, b'_{G'}\}, c' \in (0,\infty)^{G'}.$$

To keep track of everything, it is convenient to introduce a few auxiliary quantities. For a state $S = (\mathcal{G}, c)$ with $\mathcal{G} = \{b_1, \dots, b_G\}$, we define membership indicators for $g \in \{0, 1, \dots, G\}$ and $j \in [d]$ as

$$m_{g,j}(S) := \begin{cases} 1, & \text{if } g \geq 1 \text{ and } j \in b_g \\ 1, & \text{if } g = 0 \text{ and } j \in \mathcal{G}^{\complement} \\ 0, & \text{otherwise} \end{cases}$$

so that $\forall j \in [d] : \sum_{g=0}^{G} m_{g,j}(S) = 1$. We also define group sizes as $d_g(S) := |b_g| = \sum_{j=1}^{d} m_{g,j}(S)$ for $g \geq 1$, and for notational convenience we set $d_0(S) := \left|\mathcal{G}^{\complement}\right|$, although group 0 is unpenalized. The penalty part of the objective can then be written in equivalent form:

$$\begin{aligned} R_S(w) := \mathbb{F}_S(w) - f(\mathcal{D}; w) &= \lambda \sum_{g=1}^{G} d_g(S)^{1-1/p} \left\|w_{b_g}\right\|_p \\ &= \lambda \sum_{g=1}^{G} \left(\sum_{j=1}^{d} m_{g,j}(S)\right)^{1-1/p} \left(\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p\right)^{1/p} \\ &= \lambda \sum_{g=1}^{G} \left(\sum_{j=1}^{d} m_{g,j}(S)\right)^{1-1/p} \left(\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p\right)^{1/p} \cdot \frac{\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p}{\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p} \\ &= \lambda \sum_{g=1}^{G} \left(\sum_{j=1}^{d} m_{g,j}(S)\right)^{1-1/p} \left(\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p\right)^{1/p} \\ &\quad \underbrace{\frac{\sum_{j=1}^{d} m_{g,j}(S)}{\sum_{j=1}^{d} m_{g,j}(S)}}_{=1} \underbrace{\frac{\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p}{\sum_{j=1}^{d} m_{g,j}(S) |w_j|^p}}_{=1}, \end{aligned}$$

with the understanding that groups with $d_g(S) = 0$ contribute 0 to the sum because both numerator and denominator vanish while the corresponding $\left\|w_{b_g}\right\|_p$ is zero. Using the bound $\sum_{j=1}^{d} m_{g,j}(S) \leq d$ together with

$$\left(\sum_{j=1}^{d} m_{g,j}(S)\,|w_j|^p\right)^{1/p} \leq d^{1/p} \max_{1\leq j\leq d} |w_j|\,,$$

we obtain the global uniform bound

$$\begin{aligned} R_S(w) &\leq \lambda \sum_{g=1}^{G} d_g(S)^{1-1/p} \cdot d_g(S)^{1/p} \max_{1\leq j\leq d} |w_j| \\ &= \lambda \left(\sum_{g=1}^{G} d_g(S)\right) \max_{1\leq j\leq d} |w_j| \leq \lambda d \max_{1\leq j\leq d} |w_j|\,. \end{aligned}$$

We now proceed to construct an explicit path from S to S' in three stages. We begin by equalizing the number of groups. Write $G := |\mathcal{G}|$ and $G' := |\mathcal{G}'|$. We first transform S into a state $\widetilde{S}^{(1)}$ that has exactly G' groups by distinguishing two cases. If $G \leq G'$, we apply only `CREATE` operations. Set $\Delta := G' - G \geq 0$. If $\Delta = 0$, we skip this substep. Otherwise, for $k = 1, \ldots, \Delta$, we choose arbitrary coefficients $c_{\text{new}}^{(k)} > 0$ and apply

$$S^{(0)} := S, \quad S^{(k)} := \texttt{CREATE_GROUP}\left(c_{\text{new}}^{(k)}\right)\left(S^{(k-1)}\right), \quad k = 1, \ldots, \Delta.$$

Define $\widetilde{S}^{(1)} := S^{(\Delta)}$. By construction we have

$$\widetilde{S}^{(1)} = \left(\tilde{\mathcal{G}}^{(1)}, \tilde{c}^{(1)}\right), \quad \tilde{\mathcal{G}}^{(1)} = \left\{\tilde{b}_1^{(1)}, \ldots, \tilde{b}_{G'}^{(1)}\right\},$$

with

$$\tilde{b}_g^{(1)} = b_g \ (1 \leq g \leq G), \quad \tilde{b}_g^{(1)} = \varnothing \ (G < g \leq G')\,,$$

and

$$\tilde{c}_g^{(1)} = c_g \ (1 \leq g \leq G), \quad \tilde{c}_g^{(1)} = c_{\text{new}}^{(g-G)} \ (G < g \leq G')\,.$$

Each `CREATE_GROUP` does not change the membership indicators $m_{g,j}(S)$ for $g \leq G$ and $j \in [d]$, and simply appends a row of zeros to the membership matrix plus a positive entry to c. For any fixed w we have

$$R_{S^{(k)}}(w) - R_{S^{(k-1)}}(w) = \lambda c_{\text{new}}^{(k)}\, d_{G+k-1}\left(S^{(k)}\right)^{1-1/p} \left\|w_{\tilde{b}_{G+k-1}^{(1)}}\right\|_p = 0$$

because $\tilde{b}_{G+k-1}^{(1)} = \varnothing$. The objective is therefore unchanged along this substep:

$$\mathbb{F}_{S^{(k)}}(w) = \mathbb{F}_S(w) \quad \forall k \in \{0, \ldots, \Delta\}, \forall w.$$

If instead $G > G'$, we apply `MOVE` followed by `DELETE` from the top. Set $\Delta := G - G' > 0$. We will eliminate groups with indices $G, G-1, \ldots, G'+1$ by moving all their features to their desired destinations in S' and then deleting the now-empty group. We first precompute a target index map for each feature:

$$\phi(j) := \begin{cases} g & \text{if } j \in b'_g \text{ for some } g \in \{1, \ldots, G'\}\,, \\ 0 & \text{if } j \notin \bigcup_{g=1}^{G'} b'_g, \end{cases} \qquad j \in [d].$$

Thus $\phi(j) \in \{0, 1, \ldots, G'\}$. We now start from $S^{(0)} := S$, and for $g = G, G-1, \ldots, G'+1$ we proceed as follows. For each $j \in b_g$ at that moment, we apply

$$S^{(\text{before } j)} \mapsto S^{(\text{after } j)} := \texttt{MOVE_FEATURE}\,(j, \phi(j))\left(S^{(\text{before } j)}\right).$$

As a consequence,

$$j \notin b_g^{\text{new}}\,, \quad j \in b_{\phi(j)}^{\text{new}} \text{ if } \phi(j) \geq 1, \quad \text{or} \quad j \in (\mathcal{G}^{\text{new}})^{\complement} \text{ if } \phi(j) = 0.$$

After all such operations for all $j \in b_g$, the group b_g becomes empty. Once $b_g = \varnothing$, we apply DELETE_GROUP(g). This removes group g and its coefficient and reindexes groups with index $> g$ downwards by one. Because we proceed in descending order, all groups with index $> g$ have already been removed, so only groups $\{1, \dots, g-1\}$ remain and they are unaffected by the reindexing. At the end of this loop we obtain a state $\widetilde{S}^{(1)} = \left(\tilde{\mathcal{G}}^{(1)}, \tilde{c}^{(1)}\right)$ with exactly G' groups.

We now analyze the effect on the regularizer of a single MOVE_FEATURE. Suppose we move j from group g_{src} to group $g_{\text{dest}} := \phi(j)$, where in the subcase $g_{\text{dest}} = 0$ we interpret group 0 as the unpenalized container. Denote the state before the move by S_{old} and after the move by S_{new}. Then the membership sets satisfy

$$b^{\text{new}}_{g_{\text{src}}} = b^{\text{old}}_{g_{\text{src}}} \setminus \{j\}, \quad b^{\text{new}}_{g_{\text{dest}}} = b^{\text{old}}_{g_{\text{dest}}} \cup \{j\},$$

with all other b_g unchanged. For group sizes we have

$$d^{\text{new}}_{g_{\text{src}}} = d^{\text{old}}_{g_{\text{src}}} - 1, \quad d^{\text{new}}_{g_{\text{dest}}} = d^{\text{old}}_{g_{\text{dest}}} + 1,$$

with the convention that if $g_{\text{dest}} = 0$ then d_0 is unpenalized and does not appear in the sum. For the norms, if both $g_{\text{src}}, g_{\text{dest}} \geq 1$, we have

$$\begin{aligned}
\left\|w_{g^{\text{old}}_{\text{src}}}\right\|_p^p &= \sum_{k \in b^{\text{old}}_{g_{\text{src}}}} |w_k|^p = \sum_{k \in b^{\text{new}}_{g_{\text{src}}}} |w_k|^p + |w_j|^p \,, \\
\left\|w_{g^{\text{new}}_{\text{grc}}}\right\|_p^p &= \sum_{k \in b^{\text{new}}_{\text{grsc}}} |w_k|^p \,, \\
\left\|w_{g^{\text{old}}_{\text{dest}}}\right\|_p^p &= \sum_{k \in g^{\text{old}}_{g_{\text{dest}}}} |w_k|^p \,, \\
\left\|w_{g^{\text{new}}_{\text{dest}}}\right\|_p^p &= \sum_{k \in g^{\text{new}}_{g_{\text{dest}}}} |w_k|^p = \sum_{k \in g^{\text{old}}_{g_{\text{dest}}}} |w_k|^p + |w_j|^p \,.
\end{aligned}$$

Thus the change in the regularizer is

$$\begin{aligned}
&R_{S_{\text{new}}}(w) - R_{S_{\text{old}}}(w) \\
&= \lambda \Bigg[c_{g_{\text{src}}} \left(d^{\text{new}}_{g_{\text{src}}}\right)^{1-1/p} \left\|w_{g^{\text{new}}_{\text{src}}}\right\|_p - c_{g_{\text{src}}} \left(d^{\text{old}}_{g_{\text{src}}}\right)^{1-1/p} \left\|w_{b^{\text{old}}_{\text{src}}}\right\|_p \\
&\quad + c_{g_{\text{dest}}} \left(d^{\text{new}}_{g_{\text{dest}}}\right)^{1-1/p} \left\|w_{g^{\text{new}}_{\text{dest}}}\right\|_p - c_{g_{\text{dest}}} \left(d^{\text{old}}_{g_{\text{dest}}}\right)^{1-1/p} \left\|w_{b^{\text{old}}_{g_{\text{dest}}}}\right\|_p \Bigg] \\
&= \lambda \Bigg[c_{g_{\text{src}}} \left(\left(d^{\text{old}}_{g_{\text{src}}} - 1\right)^{1-1/p} \left(\sum_{k \in b^{\text{old}}_{g_{\text{src}}} \setminus \{j\}} |w_k|^p \right)^{1/p} - \left(d^{\text{old}}_{g_{\text{src}}}\right)^{1-1/p} \left(\sum_{k \in b^{\text{old}}_{g_{\text{src}}} \setminus \{j\}} |w_k|^p + |w_j|^p \right)^{1/p} \right) \\
&\quad + c_{g_{\text{dest}}} \left(\left(d^{\text{old}}_{g_{\text{dest}}} + 1\right)^{1-1/p} \left(\sum_{k \in b^{\text{old}}_{g_{\text{dest}}}} |w_k|^p + |w_j|^p \right)^{1/p} \right. \\
&\quad \left. - \left(d^{\text{old}}_{g_{\text{dest}}}\right)^{1-1/p} \left(\sum_{k \in b^{\text{old}}_{g_{\text{dest}}}} |w_k|^p \right)^{1/p} \right) \Bigg] .
\end{aligned}$$

This quantity is finite and can be bounded via the triangle and Hölder inequalities:

$$\left| \left\|w_{b^{\text{old}}_{g_{\text{src}}}}\right\|_p - \left\|w_{b^{\text{new}}_{g_{\text{src}}}}\right\|_p \right| \leq \left\|w_{b^{\text{old}}_{g_{\text{src}}} \setminus b^{\text{new}}_{g_{\text{src}}}}\right\|_p = |w_j| \,.$$

A similar bound holds for the g_{dest} part. Overall we obtain

$$|R_{S_{\text{new}}}(w) - R_{S_{\text{old}}}(w)| \leq \lambda \left(c_{g_{\text{src}}} + c_{g_{\text{dest}}}\right) C_{p,d} \max_{1 \leq j \leq d} |w_j|$$

for some finite constant $C_{p,d}$ depending only on p, d, and the group sizes. Note that one can make $C_{p,d}$ explicit, but this is not needed here. When the loop over g finishes, we have a state $\widetilde{S}^{(1)} = \left(\tilde{\mathcal{G}}^{(1)}, \tilde{c}^{(1)}\right)$ with exactly G' groups. Thus in both

cases we end the first stage with some state

$$\widetilde{S}^{(1)} = \left(\tilde{\mathcal{G}}^{(1)}, \tilde{c}^{(1)}\right), \quad \tilde{\mathcal{G}}^{(1)} = \left\{\tilde{b}_1^{(1)}, \ldots, \tilde{b}_{G'}^{(1)}\right\},$$

satisfying $\left|\tilde{\mathcal{G}}^{(1)}\right| = G'$. We next turn to aligning the group memberships. We now have $\widetilde{S}^{(1)}$ with exactly G' groups and we edit the membership so that $\tilde{\mathcal{G}}^{(1)}$ becomes exactly $\mathcal{G}'$. Define for $\widetilde{S}^{(1)} = \left(\widetilde{\mathcal{G}}^{(1)}, \tilde{c}^{(1)}\right)$ the map

$$\psi(j) := \begin{cases} g & \text{if } j \in \tilde{b}_g^{(1)} \text{ for some } g \in \{1, \ldots, G'\}, \\ 0 & \text{if } j \notin \bigcup_{g=1}^{G'} \tilde{b}_g^{(1)}, \end{cases} \qquad j \in [d],$$

so that $\psi(j) \in \{0, \ldots, G'\}$ is the current group indicator of feature j in $\widetilde{S}^{(1)}$. Recall that $\phi(j)$ is the target group indicator in S'. We now perform at most one move for each $j \in [d]$:

$$\widetilde{S}^{(1)} \mapsto \widetilde{S}^{(2)} \quad \text{by} \quad \text{if } \psi(j) \neq \phi(j) \text{ then apply } \widetilde{S}^{(1)} \leftarrow \texttt{MOVE_FEATURE}(j, \phi(j)) \left(\widetilde{S}^{(1)}\right),$$

in any order over $j \in [d]$. After this procedure, for every $j \in [d]$ we have by construction

$$j \in \tilde{b}_g^{(2)} \Leftrightarrow \phi(j) = g, \quad j \in \left(\tilde{\mathcal{G}}^{(2)}\right)^{\complement} \Leftrightarrow \phi(j) = 0.$$

By definition of ϕ, we also have in S' the relations

$$j \in b'_g \Leftrightarrow \phi(j) = g, \quad j \in \left(\mathcal{G}'\right)^{\complement} \Leftrightarrow \phi(j) = 0.$$

Therefore $\tilde{b}_g^{(2)} = b'_g$ for all $g = 1, \ldots, G'$, and the complements also match:

$$\left(\tilde{\mathcal{G}}^{(2)}\right)^{\complement} = [d] \setminus \bigcup_{g=1}^{G'} \tilde{b}_g^{(2)} = [d] \setminus \bigcup_{g=1}^{G'} b'_g = \left(\mathcal{G}'\right)^{\complement}.$$

After this second stage we thus have $\widetilde{\mathcal{G}}^{(2)} = \mathcal{G}'$. We can therefore write $\widetilde{S}^{(2)} = (\mathcal{G}', \hat{c})$ for some coefficient vector $\hat{c} \in (0, \infty)^{G'}$ representing whatever coefficients we have at this point. As in the first stage, each individual move has a regularizer change given by the expression above. Summing over at most d such moves, we obtain

$$|R_{\tilde{S}^{(2)}}(w) - R_{\tilde{S}^{(1)}}(w)| \leq \lambda C \max_{1 \leq j \leq d} |w_j|$$

for some finite C depending on the intermediate coefficients and group sizes but independent of w. Finally, we align the coefficients. At this point we have a state $\widetilde{S}^{(2)} = (\mathcal{G}', \hat{c})$ with the correct group structure $\mathcal{G}'$ but possibly incorrect coefficients. We now change the coefficients to match exactly c'. For each $g \in \{1, \ldots, G'\}$, we apply the parametric primitive

$$\widetilde{S}^{(2)} \mapsto \widetilde{S}^{(3)} \quad \text{via} \quad \widetilde{S}^{(2)} \leftarrow \texttt{ADJUST_COEFF}\left(g, c'_g\right) \left(\widetilde{S}^{(2)}\right)$$

one by one for all g. The effect of such an operation on the regularizer is straightforward. Denoting the coefficients before and after by $\widehat{c}_g$ and c'_g respectively while keeping the group structure fixed to $\mathcal{G}'$, we have for fixed w

$$\begin{aligned} &R_{\text{after}}(w) - R_{\text{before}}(w) \\ &\quad = \lambda \left(c'_g d_g^{1-1/p} \left\| w_{b'_g} \right\|_p + \sum_{h \neq g} \hat{c}_h d_h^{1-1/p} \left\| w_{b'_h} \right\|_p - \sum_{h=1}^{G'} \hat{c}_h d_h^{1-1/p} \left\| w_{b'_h} \right\|_p \right) \\ &= \lambda \left(c'_g - \hat{c}_g\right) d_g^{1-1/p} \left\| w_{b'_g} \right\|_p \\ &\leq \lambda \left|c'_g - \hat{c}_g\right| d_g^{1-1/p} \cdot d_g^{1/p} \max_{1 \leq j \leq d} |w_j| = \lambda \left|c'_g - \hat{c}_g\right| d_g \max_{1 \leq j \leq d} |w_j|. \end{aligned}$$

Summing over the finitely many g, the total change in $R_S(w)$ is bounded in magnitude by a finite multiple of $\max_j |w_j|$. After we have applied ADJUST_COEFF $\left(g, c'_g\right)$ for all $g \in \{1, \ldots, G'\}$, we obtain $\widetilde{S}^{(3)} = (\mathcal{G}', c')$ and hence $\widetilde{S}^{(3)} = S'$.

Combining the three stages, we have explicitly produced a finite sequence of primitives consisting of some number of CREATE_GROUP operations, at most $\left(d + \max\left\{G, G' - \min\{G, G'\}\right\}\right)$ instances of MOVE_FEATURE, some number of DELETE_GROUP operations, and exactly G' instances of ADJUST_COEFF, such that

$$S' = \widetilde{S}^{(3)} = \mathcal{P}_K \circ \cdots \circ \mathcal{P}_1(S)$$

for suitable primitives $\mathcal{P}_1, \ldots, \mathcal{P}_K$ of types Primitive 1, 2, 3, 4. Equivalently, for every pair of states $S, S' \in \mathcal{S}$, the state S' is reachable from S using Primitive 1, 2, 3, 4. □

C.2 Minimality

We now prove the minimality by showing that removing any single primitive from the set destroys expressivity.

Proof. We argue separately for removing each primitive. For a state $S = (\mathcal{G}, c)$ with $\mathcal{G} = \{b_1, \ldots, b_G\}$, we define the group-assignment map

$$\gamma_S : [d] \to \{0, 1, \ldots, G\}, \quad \gamma_S(j) = \begin{cases} g, & j \in b_g, g \in \{1, \ldots, G\}, \\ 0, & j \in \mathcal{G}^{\complement}. \end{cases}$$

We also write $G(S) := G$ for the number of penalized groups. Given a trajectory $S^{(0)}, S^{(1)}, \ldots, S^{(K)} \in \mathcal{S}$ with $S^{(k+1)} = \mathcal{P}_{k+1}\left(S^{(k)}\right)$, we adopt the shorthand

$$S^{(k)} = \left(\mathcal{G}^{(k)}, c^{(k)}\right), \quad \mathcal{G}^{(k)} = \left\{b_1^{(k)}, \ldots, b_{G^{(k)}}^{(k)}\right\}, \quad c^{(k)} = \left(c_1^{(k)}, \ldots, c_{G^{(k)}}^{(k)}\right).$$

For demonstration, we may rewrite the regularizer in expanded form as

$$\begin{aligned} R_S(w) := \mathbb{F}_S(w) - f(\mathcal{D}; w) &= \lambda \sum_{g=1}^{G(S)} d_g(S)^{1-1/p} \left\|w_{b_g}\right\|_p \\ &= \lambda \sum_{g=1}^{G(S)} \left(\sum_{j=1}^{d} \mathbf{1}\{j \in b_g\}\right)^{1-1/p} \left(\sum_{j=1}^{d} \mathbf{1}\{j \in b_g\} |w_j|^p\right)^{1/p}, \end{aligned}$$

so that in particular, for any sequence $S^{(0)}, \ldots, S^{(K)}$,

$$\begin{aligned} R_{S^{(K)}}(w) - R_{S^{(0)}}(w) &= \sum_{k=0}^{K-1} \left(R_{S^{(k+1)}}(w) - R_{S^{(k)}}(w)\right) \\ = \lambda \sum_{k=0}^{K-1} \left(\sum_{g=1}^{G^{(k+1)}} \left(d_g\left(S^{(k+1)}\right)^{1-1/p} \left\|w_{b_g^{(k+1)}}\right\|_p\right) - \sum_{g=1}^{G^{(k)}} \left(d_g\left(S^{(k)}\right)^{1-1/p} \left\|w_{b_g^{(k)}}\right\|_p\right)\right). \end{aligned}$$

Case I: Removing MOVE_FEATURE. In this case the available primitives are ADJUST_COEFF, CREATE_GROUP, and DELETE_GROUP. We claim that any trajectory using only these primitives preserves the entire group-assignment map γ_S and hence cannot alter the combinatorial structure of the partition. Indeed, let $S^{(0)}, S^{(1)}, \ldots, S^{(K)}$ be such a trajectory. For each k and each feature index $j \in [d]$, consider the difference $\gamma_{S^{(k+1)}}(j) - \gamma_{S^{(k)}}(j)$.

If $S^{(k+1)}$ is obtained from $S^{(k)}$ via ADJUST_COEFF(g, c_{new}), then by definition the partition $\mathcal{G}^{(k)}$ is unchanged, so $b^{(k+1)} = b_g^{(k)}$ for all g, and hence $\gamma_{S^{(k+1)}}(j) = \gamma_{S^{(k)}}(j)$ for all j. If $S^{(k+1)}$ is obtained via CREATE_GROUP(c_{new}), then again $b_g^{(k+1)} = b_g^{(k)}$ for all preexisting groups, and we merely append an empty set $b_{G^{(k)}+1}^{(k+1)} = \varnothing$. Thus $\gamma_{S^{(k+1)}}(j) = \gamma_{S^{(k)}}(j)$ for all j since no feature index belongs to

the newly created group. Finally, if $S^{(k+1)}$ is obtained by DELETE_GROUP(g), then we remove an empty group $b_g^{(k)} = \varnothing$ and possibly reindex the remaining groups to keep labels $\{1, \dots, G^{(k)} - 1\}$. However, as sets, the collection $\left\{b_1^{(k)}, \dots, b_{G^{(k)}}^{(k)}, \left(\mathcal{G}^{(k)}\right)^{\complement}\right\}$ is the same partition of $[d]$ as $\left\{b_1^{(k+1)}, \dots, b_{G^{(k+1)}}^{(k+1)}, \left(\mathcal{G}^{(k+1)}\right)^{\complement}\right\}$, so every feature j remains in the same subset of the partition. Consequently, $\gamma_{S^{(k+1)}}(j) = \gamma_{S^{(k)}}(j)$ for all $j \in [d]$. By induction, we obtain

$$\gamma_{S^{(k)}}(j) = \gamma_{S^{(0)}}(j) \quad \forall j \in [d], \forall k \in \{0, \dots, K\}.$$

Hence the telescoping sum

$$\begin{aligned}
\sum_{j=1}^{d} \left|\gamma_{S^{(K)}}(j) - \gamma_{S^{(0)}}(j)\right| &= \sum_{j=1}^{d} |\gamma_{S^{(K)}}(j) - \gamma_{S^{(K-1)}}(j) \\
&\quad + \gamma_{S^{(K-1)}}(j) - \gamma_{S^{(K-2)}}(j) + \cdots + \gamma_{S^{(1)}}(j) - \gamma_{S^{(0)}}(j)\Big| \\
&\leq \sum_{j=1}^{d} \sum_{k=0}^{K-1} \left|\gamma_{S^{(k+1)}}(j) - \gamma_{S^{(k)}}(j)\right| \\
&= \sum_{k=0}^{K-1} \sum_{j=1}^{d} \mathbf{1}_{\{\gamma_{S^{(k+1)}}(j) \neq \gamma_{S^{(k)}}(j)\}} = 0
\end{aligned}$$

shows that $\gamma_{S^{(K)}} = \gamma_{S^{(0)}}$ exactly. We now choose explicit states that differ in their group assignment. Take $d \geq 1$ and define

$$S = (\mathcal{G}, c), \quad \mathcal{G} = \{b_1\}, \quad b_1 = \{1\}, \quad \mathcal{G}^{\complement} = \{2, \dots, d\}, \quad c_1 = 1,$$
$$S' = (\mathcal{G}', c'), \quad \mathcal{G}' = \{b_1'\}, \quad b_1' = \varnothing, \quad (\mathcal{G}')^{\complement} = [d], \quad c_1' = 1.$$

Then $\gamma_S(1) = 1$ and $\gamma_{S'}(1) = 0$, so $\gamma_S \neq \gamma_{S'}$. If S' were reachable from S using only the remaining primitives, we would have a trajectory with $S^{(0)} = S$ and $S^{(K)} = S'$, and therefore $\gamma_{S^{(K)}} = \gamma_{S^{(0)}}$, contradicting $\gamma_{S'} \neq \gamma_S$. Hence removing MOVE_FEATURE destroys expressivity.

Case II: Removing ADJUST_COEFF. Now the available primitives are MOVE_FEATURE, CREATE_GROUP, and DELETE_GROUP. Fix a trajectory

$$\begin{aligned}
&S^{(0)}, S^{(1)}, \dots, S^{(K)}, \quad S^{(k+1)} \\
&\quad = \mathcal{P}_{k+1}\left(S^{(k)}\right), \quad \mathcal{P}_{k+1} \in \{\texttt{MOVE_FEATURE}, \texttt{CREATE_GROUP}, \texttt{DELETE_GROUP}\}.
\end{aligned}$$

For a specific group identifier $g_0 \in \left\{1, \dots, G^{(0)}\right\}$, we introduce the alive indicator

$$a_{g_0}^{(k)} := \begin{cases} 1, & \text{if group } g_0 \text{ is present in } \mathcal{G}^{(k)}, \\ 0, & \text{otherwise.} \end{cases}$$

Since we assume persistent pool of identities and no reuse of identifiers, the evolution of $a_{g_0}^{(k)}$ is straightforward. The primitives MOVE_FEATURE and CREATE_GROUP do not remove any existing identifier, hence $a_{g_0}^{(k+1)} = a_{g_0}^{(k)}$ whenever $\mathcal{P}_{k+1} \in \{\texttt{MOVE_FEATURE}, \texttt{CREATE_GROUP}\}$. Meanwhile DELETE_GROUP$(g_0)$ sets $a_{g_0}^{(k+1)} = 0$ permanently. For any other deletion DELETE_GROUP(g) with $g \neq g_0$, we have $a_{g_0}^{(k+1)} = a_{g_0}^{(k)}$. Thus for every step k,

$$a_{g_0}^{(k+1)} = a_{g_0}^{(k)} - \mathbf{1}_{\{\mathcal{P}_{k+1} = \texttt{DELETE_GROUP}(g_0)\}},$$

and summing from $k = 0$ to $K - 1$ yields

$$\begin{aligned}
a_{g_0}^{(K)} - a_{g_0}^{(0)} &= \sum_{k=0}^{K-1} \left(a_{g_0}^{(k+1)} - a_{g_0}^{(k)}\right) \\
&= -\sum_{k=0}^{K-1} \mathbf{1}_{\{\mathcal{P}_{k+1} = \texttt{DELETE_GROUP}(g_0)\}} \leq 0.
\end{aligned}$$

In particular, if a trajectory begins and ends in states where group g_0 is present so that $a_{g_0}^{(0)} = 1$ and $a_{g_0}^{(K)} = 1$, then we must have $\sum_{k=0}^{K-1} \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{DELETE_GROUP}(g_0)\}} = 0$. This means $\texttt{DELETE_GROUP}(g_0)$ is never called, and hence $a_{g_0}^{(k)} = 1$ for all k. Now consider the coefficient $c_{g_0}^{(k)}$ attached to group g_0 at each step. Because the remaining primitives are MOVE_FEATURE, CREATE_GROUP, and $\texttt{DELETE_GROUP}(g \neq g_0)$, we have

$$c_{g_0}^{(k+1)} = \begin{cases} c_{g_0}^{(k)}, & \mathcal{P}_{k+1} = \texttt{MOVE_FEATURE}(\cdot,\cdot), \\ c_{g_0}^{(k)}, & \mathcal{P}_{k+1} = \texttt{CREATE_GROUP}\left(c_{\text{new}}\right), \\ c_{g_0}^{(k)}, & \mathcal{P}_{k+1} = \texttt{DELETE_GROUP}(g),\ g \neq g_0, \end{cases}$$

whenever $a_{g_0}^{(k)} = 1$. That is,

$$\begin{aligned} c_{g_0}^{(k+1)} - c_{g_0}^{(k)} = 0 \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{MOVE_FEATURE}\}} + 0 \\ \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{CREATE_GROUP}\}} + 0 \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{DELETE_GROUP}(g\neq g_0)\}}, \end{aligned}$$

and therefore

$$\begin{aligned} c_{g_0}^{(K)} - c_{g_0}^{(0)} &= \sum_{k=0}^{K-1} \left(c_{g_0}^{(k+1)} - c_{g_0}^{(k)} \right) \\ &= \sum_{k=0}^{K-1} \Big[0 \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{MOVE_FEATURE}\}} + 0 \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{CREATE_GROUP}\}} \\ &\qquad + 0 \cdot \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{DELETE_GROUP}(g\neq g_0)\}} \Big] = 0. \end{aligned}$$

Thus $c_{g_0}^{(K)} = c_{g_0}^{(0)}$ along any trajectory that keeps group g_0 alive. We now consider states in which group $g_0 = 1$ is present at both endpoints but has different coefficients. Take $d \geq 1$ and let

$$S = (\mathcal{G}, c), \quad \mathcal{G} = \{b_1\}, \quad b_1 = \{1\}, \quad \mathcal{G}^{\complement} = \{2, \ldots, d\}, \quad c_1 = 1,$$

$$S' = (\mathcal{G}', c'), \quad \mathcal{G}' = \{b_1'\}, \quad b_1' = \{1\}, \quad (\mathcal{G}')^{\complement} = \{2, \ldots, d\}, \quad c_1' = 2.$$

The partitions $\mathcal{G}$ and $\mathcal{G}'$ coincide exactly, and the only difference between S and S' is the coefficient of the group identifier 1. Suppose for contradiction that S' were reachable from S using only MOVE_FEATURE, CREATE_GROUP, and DELETE_GROUP. Then there would exist a trajectory with $S^{(0)} = S$ and $S^{(K)} = S'$ for which group 1 is present at both endpoints. By the argument above we would have

$$c_1(S') = c_1^{(K)} = c_1^{(0)} = c_1(S),$$

that is,

$$2 = c_1' = c_1(S') = c_1(S) = 1,$$

which is a contradiction. Thus without ADJUST_COEFF, there exist $S, S' \in \mathcal{S}$ such that S' is not reachable from S.

Case III: Removing CREATE_GROUP. In this case the allowed primitives are MOVE_FEATURE, ADJUST_COEFF, and DELETE_GROUP. We claim that along any trajectory using only these primitives the number of penalized groups $G(S)$ is nonincreasing, hence we cannot reach states with strictly more groups than we started with. Consider a trajectory $S^{(0)}, S^{(1)}, \ldots, S^{(K)}$ with $S^{(k+1)} = \mathcal{P}_{k+1}\left(S^{(k)}\right)$. Define $G^{(k)} := G\left(S^{(k)}\right)$. By the definitions of the primitives, at each step we have

$$G^{(k+1)} = \begin{cases} G^{(k)}, & \mathcal{P}_{k+1} = \texttt{MOVE_FEATURE}, \\ G^{(k)}, & \mathcal{P}_{k+1} = \texttt{ADJUST_COEFF}, \\ G^{(k)} - 1, & \mathcal{P}_{k+1} = \texttt{DELETE_GROUP}(\cdot). \end{cases}$$

Equivalently, $G^{(k+1)} - G^{(k)} = -\mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{DELETE_GROUP}\ (\cdot)\}}$. Summing from $k = 0$ to $K - 1$ yields

$$\begin{aligned} G^{(K)} - G^{(0)} &= \sum_{k=0}^{K-1} \left(G^{(k+1)} - G^{(k)} \right) \\ &= -\sum_{k=0}^{K-1} \mathbf{1}_{\{\mathcal{P}_{k+1}=\texttt{DELETE_GROUP}(\cdot)\}} \leq 0. \end{aligned}$$

Therefore $G\left(S^{(K)}\right) = G^{(K)} \leq G^{(0)} = G\left(S^{(0)}\right)$, meaning the number of penalized groups cannot increase along such a trajectory. We now choose S, S' with $G(S') > G(S)$. Let $d \geq 2$ and define

$$S = (\mathcal{G}, c), \quad \mathcal{G} = \{b_1\}, \quad b_1 = [d], \quad \mathcal{G}^{\complement} = \varnothing, \quad c_1 = 1,$$

so $G(S) = 1$. Define $S' = (\mathcal{G}', c')$ with $\mathcal{G}' = \{b'_1, b'_2\}$ where

$$b'_1 = \left\{1, \ldots, \left\lfloor \frac{d}{2} \right\rfloor\right\}, \quad b'_2 = \left\{\left\lfloor \frac{d}{2} \right\rfloor + 1, \ldots, d\right\}, \quad (\mathcal{G}')^{\complement} = \varnothing, \quad c'_1 = 1, \quad c'_2 = 1,$$

so $G(S') = 2$. If S' were reachable from S using only the remaining primitives, there would exist a trajectory with $S^{(0)} = S$ and $S^{(K)} = S'$, hence

$$G(S') = G\left(S^{(K)}\right) \leq G\left(S^{(0)}\right) = G(S) = 1,$$

yielding $2 = G(S') \leq 1$, which is a contradiction. Consequently, without `CREATE_GROUP` we cannot reach all states from a given initial state, so this primitive is necessary.

Case IV: Removing `DELETE_GROUP`. Similar to Case III, we first show that the evolution of the group count is monotone nondecreasing. As before, let $G(S)$ denote the number of penalized groups in S. For a sequence $S^{(0)}, S^{(1)}, \ldots, S^{(K)}$, define

$$\Delta_k := G\left(S^{(k+1)}\right) - G\left(S^{(k)}\right).$$

For $\mathcal{P}_k =$`MOVE_FEATURE`, there is no change in the number of groups so $\Delta_k = 0$. For $\mathcal{P}_k =$ `ADJUST_COEFF` we also have $\Delta_k = 0$. For $\mathcal{P}_k =$ `CREATE_GROUP` (c_{new}) we have $G\left(S^{(k+1)}\right) = G\left(S^{(k)}\right) + 1$, hence $\Delta_k = +1$. Thus $\Delta_k \in \{0, +1\}$. Summing yields

$$\begin{aligned} G\left(S^{(K)}\right) &= G\left(S^{(0)}\right) + \sum_{k=0}^{K-1} \Delta_k \\ &= G\left(S^{(0)}\right) + \sum_{k:\mathcal{P}_k=\texttt{MOVE_FEATURE}} 0 + \sum_{k:\mathcal{P}_k=\texttt{ADJUST_COEFF}} 0 + \sum_{k:\mathcal{P}_k=\texttt{CREATE_GROUP}} 1 \\ &= G\left(S^{(0)}\right) + N_{\text{cre}}, \end{aligned}$$

where $N_{\text{cre}} := \#\{k : \mathcal{P}_k = \texttt{CREATE_GROUP}\} \geq 0$. Therefore

$$G\left(S^{(K)}\right) = G\left(S^{(0)}\right) + N_{\text{cre}} \geq G\left(S^{(0)}\right),$$

meaning $G\left(S^{(K)}\right) \geq G\left(S^{(0)}\right)$ for any sequence without `DELETE_GROUP`. We then choose S, S' with $G(S') < G(S)$. We take S with one group given by

$$\mathcal{G} = \{b_1\}, \quad b_1 = \{1\}, \quad b_0 = [d] \backslash \{1\}, \quad c_1 = 1, \quad G(S) = 1,$$

and S' with no penalized groups given by

$$\mathcal{G}' = \varnothing, \quad b'_0 = [d], \quad c' = \text{empty}, \quad G(S') = 0.$$

If S' were reachable from S using only `MOVE_FEATURE`, `ADJUST_COEFF`, and `CREATE_GROUP`, then along some sequence with $S^{(0)} = S$ and $S^{(K)} = S'$ we would have

$$G(S') = G\left(S^{(K)}\right) \geq G\left(S^{(0)}\right) = G(S),$$

yielding $0 = G(S') \geq 1 = G(S)$, which is a contradiction. Therefore `DELETE_GROUP` is necessary, and without it we cannot reduce the number of groups and expressivity fails. □

D Proof of Theorem 3.5

We decompose Theorem 3.5 into five parts: (**i**) the function $\text{dist}(\cdot, \cdot)$ is non-negative; (**ii**) $\text{dist}(\cdot, \cdot)$ is finite for all $\mathcal{G}, \mathcal{G}'$; (**iii**) $\text{dist}(\cdot, \cdot)$ is symmetric for all $\mathcal{G}, \mathcal{G}'$; (**iv**) $\text{dist}(\cdot, \cdot)$ satisfies the identity of indiscernibles; (**v**) $\text{dist}(\cdot, \cdot)$ satisfies the triangle inequality. We demonstrate each condition in turn.

D.1 Non-negativity

Proof. For $1 \le p^* < \infty$ and every $(b, b') \in \widetilde{\mathcal{G}} \times \widetilde{\mathcal{G}}'$, the block-to-block cost is

$$\begin{aligned}
[c_{p^*}(b,b')]^{p^*} &= \sum_{i=1}^{d} \left| |b|^{-\frac{1}{p^*}} (\mathbf{1}_b)_i - |b'|^{-\frac{1}{p^*}} (\mathbf{1}_{b'})_i \right|^{p^*} \\
&= \sum_{i \in b \setminus b'} \left| |b|^{-\frac{1}{p^*}} - 0 \right|^{p^*} + \sum_{i \in b' \setminus b} \left| 0 - |b'|^{-\frac{1}{p^*}} \right|^{p^*} + \sum_{i \in b \cap b'} \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \\
&= \sum_{i \in b \setminus b'} |b|^{-1} + \sum_{i \in b' \setminus b} |b'|^{-1} + \sum_{i \in b \cap b'} \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \\
&= |b \setminus b'| \cdot |b|^{-1} + |b' \setminus b| \cdot |b'|^{-1} + |b \cap b'| \cdot \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \ge 0.
\end{aligned} \tag{8}$$

Hence $c_{p^*}(b, b') \ge 0$. If $p^* = \infty$, then

$$c_\infty(b,b') = \max_{i \in [d]} \left| |b|^{-0} (\mathbf{1}_b)_i - |b'|^{-0} (\mathbf{1}_{b'})_i \right| = \max_{i \in [d]} |(\mathbf{1}_b)_i - (\mathbf{1}_{b'})_i| \ge 0$$

as a maximum of absolute values. We next verify that the feasible set in the definition of $\mathrm{dist}(\cdot,\cdot)$ is nonempty. For every fixed $b \in \widetilde{\mathcal{G}}$ we have

$$\begin{aligned}
\sum_{b' \in \widetilde{\mathcal{G}}'} \pi(b,b') = \frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} &= \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \cdot 1 \\
&= \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \frac{\sum_{b' \in \widetilde{\mathcal{G}}'} |b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \\
&= \sum_{b' \in \widetilde{\mathcal{G}}'} \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \right),
\end{aligned}$$

and similarly for every fixed $b' \in \widetilde{\mathcal{G}}'$ we have

$$\begin{aligned}
\sum_{b \in \widetilde{\mathcal{G}}} \pi(b,b') = \frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} &= \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \right) \cdot 1 \\
&= \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \right) \frac{\sum_{b \in \widetilde{\mathcal{G}}} |b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \\
&= \sum_{b \in \widetilde{\mathcal{G}}} \left(\frac{|b|^{p^*}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \right).
\end{aligned} \tag{9}$$

Thus $\pi(\cdot,\cdot)$ is feasible as

$$\pi(b,b') := \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a' \in \widetilde{\mathcal{G}}'} |a'|^{\frac{1}{p^*}}} \right), \quad b \in \widetilde{\mathcal{G}}, b' \in \widetilde{\mathcal{G}}'. \tag{10}$$

Now consider any feasible $\pi(\cdot,\cdot)$. Because $\pi(b,b') \ge 0$ and $c_{p^*}(b,b') \ge 0$, we have the pointwise inequality

$$\pi(b,b')\, c_{p^*}(b,b') \ge 0, \quad \forall b \in \widetilde{\mathcal{G}}, \forall b' \in \widetilde{\mathcal{G}}'.$$

Taking the minimum over a nonempty feasible set preserves non-negativity, so

$$\mathrm{dist}(\mathcal{G}, \mathcal{G}') = \min_{\{\pi(b,b')\}} \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \pi(b,b')\, c_{p^*}(b,b') \ge \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} 0 = 0.$$

□

D.2 Finiteness

Proof. Fix $b \in \widetilde{\mathcal{G}}, b' \in \widetilde{\mathcal{G}}'$. For $1 \leq p^* < \infty$ we have

$$\begin{aligned}
&[c_{p^*}(b,b')]^{p^*} \overset{(8)}{=} \frac{|b\backslash b'|}{|b|} + \frac{|b'\backslash b|}{|b'|} + |b\cap b'| \cdot \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \\
&\leq \frac{|b\backslash b'|}{|b|} + \frac{|b'\backslash b|}{|b'|} + |b\cap b'| \cdot \left(|b|^{-\frac{1}{p^*}} + |b'|^{-\frac{1}{p^*}} \right)^{p^*} \\
&\leq \frac{|b\backslash b'|}{|b|} + \frac{|b'\backslash b|}{|b'|} + |b\cap b'| \cdot 2^{p^*-1} \left(\left(|b|^{-\frac{1}{p^*}} \right)^{p^*} + \left(|b'|^{-\frac{1}{p^*}} \right)^{p^*} \right) \\
&= \frac{|b\backslash b'|}{|b|} + \frac{|b'\backslash b|}{|b'|} + 2^{p^*-1} \left(\frac{|b\cap b'|}{|b|} + \frac{|b\cap b'|}{|b'|} \right) \\
&\leq \frac{|b|}{|b|} + \frac{|b'|}{|b'|} + 2^{p^*-1} \left(\frac{|b|}{|b|} + \frac{|b'|}{|b'|} \right) \\
&= 1 + 1 + 2^{p^*-1}(1+1) = 2 + 2^{p^*}.
\end{aligned}$$

Here the second inequality uses the bound $(x+y)^r \leq 2^{r-1}(x^r + y^r)$ valid for $r \geq 1$ and $x, y \geq 0$. The third inequality uses the set-theoretic cardinality bounds

$$|b\backslash b'| \leq |b|, \quad |b'\backslash b| \leq |b'|, \quad |b\cap b'| \leq |b|, \quad |b\cap b'| \leq |b'|.$$

Taking p^*-th roots gives the finite uniform bound

$$c_{p^*}(b,b') \leq \left(2 + 2^{p^*}\right)^{\frac{1}{p^*}} < \infty.$$

For $p^* = \infty$, similar technical analysis yields

$$\begin{aligned}
c_\infty(b,b') = \left\| |b|^{-\frac{1}{\infty}} \mathbf{1}_b - |b'|^{-\frac{1}{\infty}} \mathbf{1}_{b'} \right\|_\infty &= \|\mathbf{1}_b - \mathbf{1}_{b'}\|_\infty \\
&= \max_{i\in[d]} |(\mathbf{1}_b)_i - (\mathbf{1}_{b'})_i| \leq \max_{i\in[d]} \left(|(\mathbf{1}_b)_i| + |(\mathbf{1}_{b'})_i| \right) \\
&\leq \max_{i\in[d]} (1+1) = 2.
\end{aligned}$$

Thus for all $p^* \in [1,\infty]$ and all (b,b') we have $0 \leq c_{p^*}(b,b') \leq 2 + 2^{p^*} < \infty$. Using the feasible $\pi(b,b')$ constructed above, we bound

$$\begin{aligned}
&\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b')\, c_{p^*}(b,b') \\
&= \sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b')^{\frac{1}{p}} \left(\pi(b,b')^{\frac{1}{p^*}} c_{p^*}(b,b') \right) \\
&\leq \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \left(\pi(b,b')^{\frac{1}{p}} \right)^p \right)^{\frac{1}{p}} \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \left(\pi(b,b')^{\frac{1}{p^*}} c_{p^*}(b,b') \right)^{p^*} \right)^{\frac{1}{p^*}} \\
&= \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b') \right)^{\frac{1}{p}} \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b')\, c_{p^*}(b,b')^{p^*} \right)^{\frac{1}{p^*}} \\
&\leq \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b') \right)^{\frac{1}{p}} \left(2^{p^*+1} \sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \pi(b,b') \right)^{\frac{1}{p^*}} \\
&\overset{(10)}{=} 2^{\frac{1}{p^*}+1} \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}} |a'|^{\frac{1}{p^*}}} \right) \right)^{\frac{1}{p}} \\
&\qquad \left(\sum_{b\in\widetilde{\mathcal{G}}} \sum_{b'\in\widetilde{\mathcal{G}}'} \left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}} |a'|^{\frac{1}{p^*}}} \right) \left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} \right) \right)^{\frac{1}{p^*}}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2^{\frac{1}{p^*}+1}\left(\sum_{b\in\widetilde{\mathcal{G}}}\left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)\sum_{b'\in\widetilde{\mathcal{G}}'}\left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}}|a'|^{\frac{1}{p^*}}}\right)\right)^{\frac{1}{p}}\\
&\quad\left(\sum_{b'\in\widetilde{\mathcal{G}}'}\left(\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}}|a'|^{\frac{1}{p^*}}}\right)\sum_{b\in\widetilde{\mathcal{G}}}\left(\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)\right)^{\frac{1}{p^*}}\\
&= 2^{\frac{1}{p^*}+1}\left(\sum_{b\in\widetilde{\mathcal{G}}}\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)^{\frac{1}{p^*}}\left(\sum_{b'\in\widetilde{\mathcal{G}}'}\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}'}|a'|^{\frac{1}{p^*}}}\right)^{\frac{1}{p^*}}\\
&\quad\left(\sum_{b\in\widetilde{\mathcal{G}}}\frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)^{\frac{1}{p}}\left(\sum_{b'\in\widetilde{\mathcal{G}}'}\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}'}|a'|^{\frac{1}{p^*}}}\right)^{\frac{1}{p}}\\
&= 2^{\frac{1}{p^*}+1}\left(\frac{\sum_{b\in\widetilde{\mathcal{G}}}|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)^{\frac{1}{p}}\left(\frac{\sum_{b\in\widetilde{\mathcal{G}}}|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}\right)^{\frac{1}{p^*}}\left(\frac{\sum_{b'\in\widetilde{\mathcal{G}}'}|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}'}|a'|^{\frac{1}{p^*}}}\right)^{\frac{1}{p^*}+\frac{1}{p}}\\
&= 2^{\frac{1}{p^*}+1}\cdot(1)^{\frac{1}{p}+\frac{1}{p^*}}\cdot 1\\
&= 2^{\frac{1}{p^*}+1}\leq 4.
\end{aligned}
$$

In particular, the specific feasible π constructed above has objective value at most 4, so the minimum over all feasible π is also at most 4. Therefore $0 \leq \operatorname{dist}(\mathcal{G},\mathcal{G}') = \min_{\{\pi(b,b')\}}\sum_b\sum_{b'}\pi(b,b')\,c_{p^*}(b,b') \leq 4 < \infty$. □

D.3 Symmetry

Proof. For any blocks $b\in\widetilde{\mathcal{G}}, b'\in\widetilde{\mathcal{G}}'$ we have

$$
\begin{aligned}
c_{p^*}(b,b') = \left\||b|^{-\frac{1}{p^*}}\mathbf{1}_b - |b'|^{-\frac{1}{p^*}}\mathbf{1}_{b'}\right\|_{p^*} &= \left\|-\left(|b'|^{-\frac{1}{p^*}}\mathbf{1}_{b'} - |b|^{-\frac{1}{p^*}}\mathbf{1}_b\right)\right\|_{p^*}\\
&= \left\||b'|^{-\frac{1}{p^*}}\mathbf{1}_{b'} - |b|^{-\frac{1}{p^*}}\mathbf{1}_b\right\|_{p^*} = c_{p^*}(b',b),
\end{aligned}
\tag{11}
$$

where we used the identity $\|u\|_{p^*} = \|-u\|_{p^*}$. Hence the block cost is symmetric. Consider any feasible family $\{\pi(b,b')\}_{b\in\widetilde{\mathcal{G}},b'\in\widetilde{\mathcal{G}}'}$. Define a swapped-index family[2] $\pi^\top(b',b) := \pi(b,b')$ for $b\in\widetilde{\mathcal{G}}, b'\in\widetilde{\mathcal{G}}'$. Then for every fixed $b'\in\widetilde{\mathcal{G}}'$ we have

$$
\sum_{b\in\widetilde{\mathcal{G}}}\pi^\top(b',b) = \sum_{b\in\widetilde{\mathcal{G}}}\pi(b,b') = \frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'\in\widetilde{\mathcal{G}}'}|a'|^{\frac{1}{p^*}}},
$$

which is exactly the row-sum constraint required when $\tilde{\mathcal{G}}'$ plays the role of $\widetilde{\mathcal{G}}$. Likewise, for every fixed $b\in\widetilde{\mathcal{G}}$ we have

$$
\sum_{b'\in\widetilde{\mathcal{G}}'}\pi^\top(b',b) = \sum_{b'\in\widetilde{\mathcal{G}}'}\pi(b,b') = \frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}},
$$

which is exactly the column-sum constraint required in the swapped problem. Moreover $\pi^\top(b',b)\geq 0$ since $\pi(b,b')\geq 0$. Hence $\pi^\top$ is feasible for $\operatorname{dist}(\mathcal{G}',\mathcal{G})$, meaning that swapping a feasible π preserves feasibility with swapped marginals. Using (11) we obtain

$$
\begin{aligned}
\sum_{b'\in\widetilde{\mathcal{G}}'}\sum_{b\in\widetilde{\mathcal{G}}}\pi^\top(b',b)\,c_{p^*}(b',b) &= \sum_{b'\in\widetilde{\mathcal{G}}'}\sum_{b\in\widetilde{\mathcal{G}}}\pi(b,b')\,c_{p^*}(b',b)\\
&= \sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'}\pi(b,b')\,c_{p^*}(b',b) = \sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'}\pi(b,b')\,c_{p^*}(b,b').
\end{aligned}
$$

[2]This represents the same numerical array viewed with reversed indices.

Since for every feasible π of $\mathrm{dist}\,(\mathcal{G},\mathcal{G}')$ we constructed a feasible $\pi^\top$ for $\mathrm{dist}\,(\mathcal{G}',\mathcal{G})$ with the same objective value, we have

$$\begin{aligned}\mathrm{dist}\,(\mathcal{G}',\mathcal{G}) &= \min_{\{\rho(b',b)\}} \sum_{b'\in\widetilde{\mathcal{G}}'}\sum_{b\in\widetilde{\mathcal{G}}} \rho\,(b',b)\,c_{p^*}\,(b',b) \\ &\le \sum_{b'\in\widetilde{\mathcal{G}}'}\sum_{b\in\widetilde{\mathcal{G}}} \pi^\top\,(b',b)\,c_{p^*}\,(b',b) = \sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'} \pi\,(b,b')\,c_{p^*}\,(b,b')\,.\end{aligned} \tag{12}$$

Since this holds for every feasible π in the $(\mathcal{G},\mathcal{G}')$ problem, we may take the minimum over π on the right-hand side to obtain

$$\mathrm{dist}\,(\mathcal{G}',\mathcal{G}) \le \mathrm{dist}\,(\mathcal{G},\mathcal{G}')\,.$$

Applying the same argument with $\mathcal{G}$ and $\mathcal{G}'$ swapped yields the reverse inequality $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') \le \mathrm{dist}\,(\mathcal{G}',\mathcal{G})$, as in (12). Therefore $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') = \mathrm{dist}\,(\mathcal{G}',\mathcal{G})$. □

D.4 Identity of Indiscernibles

Proof. We first show that if $\widetilde{\mathcal{G}} = \widetilde{\mathcal{G}}'$ then $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') \equiv 0$. Assume $\widetilde{\mathcal{G}} = \widetilde{\mathcal{G}}'$ as collections of subsets of $[d]$. We define $\hat{\pi}\,(b,b')$ for $b \in \widetilde{\mathcal{G}}, b' \in \widetilde{\mathcal{G}}' = \widetilde{\mathcal{G}}$ by

$$\hat{\pi}\,(b,b') := \begin{cases} \frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}}, & b = b' \\ 0, & b \ne b' \end{cases}.$$

Then for every $b \in \widetilde{\mathcal{G}}$ we have

$$\sum_{b'\in\widetilde{\mathcal{G}}'} \hat{\pi}\,(b,b') = \sum_{b'\in\widetilde{\mathcal{G}}} \hat{\pi}\,(b,b') = \hat{\pi}(b,b) = \frac{|b|^{\frac{1}{p^*}}}{\sum\limits_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}},$$

and similarly for every $b' \in \widetilde{\mathcal{G}}'$ we have

$$\sum_{b\in\widetilde{\mathcal{G}}} \hat{\pi}\,(b,b') = \hat{\pi}\,(b',b') = \frac{|b'|^{\frac{1}{p^*}}}{\sum\limits_{a'\in\widetilde{\mathcal{G}}'}|a'|^{\frac{1}{p^*}}} = \frac{|b'|^{\frac{1}{p^*}}}{\sum\limits_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}},$$

so $\hat{\pi}$ is feasible. The objective value is

$$\begin{aligned}\mathrm{dist}(\mathcal{G},\mathcal{G}') = \min_{\{\pi(b,b')\}} \sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'} \pi\,(b,b')\,c_{p^*}(b,b') &\le \sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'} \hat{\pi}\,(b,b')\,c_{p^*}\,(b,b') \\ &= \sum_{b\in\widetilde{\mathcal{G}}} \hat{\pi}(b,b) c_{p^*}(b,b) \\ &= \sum_{b\in\widetilde{\mathcal{G}}} \frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}} \left\| |b|^{-\frac{1}{p^*}}\mathbf{1}_b - |b|^{-\frac{1}{p^*}}\mathbf{1}_b \right\|_{p^*} \\ &= \sum_{b\in\widetilde{\mathcal{G}}} \frac{|b|^{\frac{1}{p^*}}}{\sum_{a\in\widetilde{\mathcal{G}}}|a|^{\frac{1}{p^*}}} \cdot 0 = 0.\end{aligned}$$

Hence $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') \le 0$. Since we have already established non-negativity, we conclude $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') \equiv 0$. We now show that $\mathrm{dist}\,(\mathcal{G},\mathcal{G}') = 0$ implies $\widetilde{\mathcal{G}} \equiv \widetilde{\mathcal{G}}'$. Let $\hat{\pi}$ be an optimal feasible plan attaining

$\operatorname{dist}(\mathcal{G}, \mathcal{G}') = 0$ and let Δ denote the symmetric difference operator. For $1 \leq p^* < \infty$ we have

$$
\begin{aligned}
0 &= \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b')\, c_{p^*}(b, b') \\
&\geq \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') \left(c_{p^*}(b, b')^{p^*} \right)^{\frac{1}{p^*}} \\
&= \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') \left(\sum_{i=1}^{d} \left| |b|^{-\frac{1}{p^*}} (\mathbf{1}_b)_i - |b'|^{-\frac{1}{p^*}} (\mathbf{1}_{b'})_i \right|^{p^*} \right)^{\frac{1}{p^*}} \\
&= \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') \left(\sum_{i \in b \setminus b'} \left| |b|^{-\frac{1}{p^*}} \right|^{p^*} + \sum_{i \in b' \setminus b} \left| |b'|^{-\frac{1}{p^*}} \right|^{p^*} + \sum_{i \in b \cap b'} \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \right)^{\frac{1}{p^*}} \\
&= \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') \left(\frac{|b \setminus b'|}{|b|} + \frac{|b' \setminus b|}{|b'|} + |b \cap b'| \cdot \left| |b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}} \right|^{p^*} \right)^{\frac{1}{p^*}} \\
&\geq \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') \left(\frac{|b \setminus b'|}{|b|} + \frac{|b' \setminus b|}{|b'|} \right)^{\frac{1}{p^*}} \\
&\geq \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{|b \setminus b'|}{\max\{|b|, |b'|\}} + \frac{|b' \setminus b|}{\max\{|b|, |b'|\}} \right)^{\frac{1}{p^*}} \\
&= \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{|b \setminus b'| + |b' \setminus b|}{\max\{|b|, |b'|\}} \right)^{\frac{1}{p^*}} \\
&= \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{|(b \setminus b') \cup (b' \setminus b)|}{\max\{|b|, |b'|\}} \right)^{\frac{1}{p^*}} \\
&= \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{|b \Delta b'|}{\max\{|b|, |b'|\}} \right)^{\frac{1}{p^*}} \\
&\geq \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{\mathbf{1}\{b \Delta b' \neq \varnothing\}}{\max\{|b|, |b'|\}} \right)^{\frac{1}{p^*}} \\
&\geq \sum_{b} \sum_{b'} \hat{\pi}(b, b') \left(\frac{\mathbf{1}\{b \neq b'\}}{d} \right)^{\frac{1}{p^*}} \\
&= d^{-\frac{1}{p^*}} \sum_{b \in \widetilde{\mathcal{G}}} \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b')\, \mathbf{1}\{b \neq b'\} = d^{-\frac{1}{p^*}} \sum_{b} \sum_{\substack{b' \\ b' \neq b}} \hat{\pi}(b, b'),
\end{aligned}
\tag{13}
$$

where the penultimate inequality uses $\max\{|b|, |b'|\} \leq d$. As a result, $\sum_{b \in \widetilde{\mathcal{G}}} \sum_{\substack{b' \in \widetilde{\mathcal{G}}' \\ b' \neq b}} \hat{\pi}(b, b') = 0$, which gives

$$
\forall b \in \widetilde{\mathcal{G}}, \forall b' \in \widetilde{\mathcal{G}}' \, (b' \neq b) : \hat{\pi}(b, b') = 0 \tag{14}
$$

because every term is nonnegative. Now we apply the row-marginal constraints. For each fixed $b \in \widetilde{\mathcal{G}}$ we have

$$
\frac{|b|^{\frac{1}{p^*}}}{\sum_{a \in \widetilde{\mathcal{G}}} |a|^{\frac{1}{p^*}}} = \sum_{b' \in \widetilde{\mathcal{G}}'} \hat{\pi}(b, b') = \pi(b, b) \mathbf{1}\left\{ b \in \widetilde{\mathcal{G}}' \right\} + \sum_{\substack{b' \in \widetilde{\mathcal{G}}' \\ b' \neq b}} \hat{\pi}(b, b') \overset{(14)}{=} \hat{\pi}(b, b) \mathbf{1}\left\{ b \in \widetilde{\mathcal{G}}' \right\} + 0. \tag{15}
$$

Since every $b \in \widetilde{\mathcal{G}}$ is nonempty, we have $|b|^{1/p^*} > 0$ and the denominator is positive, so the left-hand side of (15) is strictly positive. Therefore $\mathbf{1}\left\{ b \in \widetilde{\mathcal{G}}' \right\} = 1$, and we conclude that

$$
\forall b \in \widetilde{\mathcal{G}} : \quad b \in \widetilde{\mathcal{G}}' \;\Rightarrow\; \widetilde{\mathcal{G}} \subseteq \widetilde{\mathcal{G}}'. \tag{16}
$$

Swapping $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{G}}'$ and repeating the exact same argument yields $\widetilde{\mathcal{G}}' \subseteq \widetilde{\mathcal{G}}$. Combining with (16) gives $\widetilde{\mathcal{G}} \equiv \widetilde{\mathcal{G}}'$. The edge case $p^* = \infty$ admits a simpler analysis. Since $|b|^{-1/p^*} = 1$, we have

$$c_\infty(b, b') = \|\mathbf{1}_b - \mathbf{1}_{b'}\|_\infty = \mathbf{1}\{b \neq b'\},$$

so (13) simplifies to

$$0 = \sum_{b,b'} \hat{\pi}(b, b')\, c_\infty(b, b') = \sum_{b,b'} \hat{\pi}(b, b')\, \mathbf{1}\{b \neq b'\},$$

and the remaining steps are identical. This completes the proof. □

D.5 Triangle Inequality

Proof. For the general case, we decompose $[d]$ into disjoint regions determined by membership in b, b', b'' as

$$\begin{aligned}[d] = &\left(b \cap b' \cap b''\right) \dot\cup \left(b \cap b' \backslash b''\right) \dot\cup \left(b \cap b'' \backslash b'\right) \dot\cup \left(b' \cap b'' \backslash b\right) \\ &\dot\cup \left(b \backslash (b' \cup b'')\right) \dot\cup \left(b' \backslash (b \cup b'')\right) \dot\cup \left(b'' \backslash (b \cup b')\right) \\ &\dot\cup \left([d] \backslash (b \cup b' \cup b'')\right),\end{aligned} \tag{17}$$

where $\dot\cup$ denotes disjoint union. Fix $b, b', b'' \subseteq [d]$, all nonempty. Define the two vectors

$$r := |b|^{-\frac{1}{p^*}} \mathbf{1}_b - |b'|^{-\frac{1}{p^*}} \mathbf{1}_{b'}, \quad s := |b'|^{-\frac{1}{p^*}} \mathbf{1}_{b'} - |b''|^{-\frac{1}{p^*}} \mathbf{1}_{b''}.$$

Then $r + s = |b|^{-\frac{1}{p^*}} \mathbf{1}_b - |b''|^{-\frac{1}{p^*}} \mathbf{1}_{b''}$. On each of the eight regions in (17), the coordinate value of $r + s$ is constant, hence

$$\begin{aligned}
\|r+s\|_{p^*}^{p^*} &= \sum_{i=1}^{d} |(r+s)_i|^{p^*} \\
&= |b \cap b' \cap b''| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b \cap b' \backslash b''| \left||b|^{-\frac{1}{p^*}}\right|^{p^*} \\
&\quad + |b \cap b'' \backslash b'| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} \\
&+ |b' \cap b'' \backslash b| \left||b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b \backslash (b' \cup b'')| \left||b|^{-\frac{1}{p^*}}\right|^{p^*} + \left|b' \backslash (b \cup b'')\right| \cdot 0^{p^*} \\
&+ |b'' \backslash (b \cup b')| \left||b''|^{-\frac{1}{p^*}}\right|^{p^*} + \left|[d] \backslash (b \cup b' \cup b'')\right| \cdot 0^{p^*} \\
&= |b \cap b' \cap b''| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b' \cap b'' \backslash b|\, |b''|^{-1} + |b'' \backslash (b \cup b')|\, |b''|^{-1} \\
&+ |b \cap b'' \backslash b'| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b \cap b' \backslash b''|\, |b|^{-1} + |b \backslash (b' \cup b'')|\, |b|^{-1}.
\end{aligned} \tag{18}$$

The cost itself is the p^*-root of (18):

$$\begin{aligned}
c_{p^*}(b, b'') &= \left(\|r+s\|_{p^*}^{p^*}\right)^{\frac{1}{p^*}} \\
&= \Bigg[|b \cap b' \cap b''| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b' \cap b'' \backslash b|\, |b''|^{-1} + |b'' \backslash (b \cup b')|\, |b''|^{-1} \\
&\qquad + |b \cap b'' \backslash b'| \left||b|^{-\frac{1}{p^*}} - |b''|^{-\frac{1}{p^*}}\right|^{p^*} + |b \cap b' \backslash b''|\, |b|^{-1} + |b \backslash (b' \cup b'')|\, |b|^{-1}\Bigg]^{\frac{1}{p^*}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
c_{p^*}(b, b') &= \left(\|r\|_{p^*}^{p^*}\right)^{\frac{1}{p^*}} \\
&= \Bigg[|b \cap b' \cap b''| \left||b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}}\right|^{p^*} + |b \cap b' \backslash b''| \left||b|^{-\frac{1}{p^*}} - |b'|^{-\frac{1}{p^*}}\right|^{p^*} + |b \cap b'' \backslash b'|\, |b|^{-1} \\
&\qquad + |b \backslash (b' \cup b'')|\, |b|^{-1} + |b' \cap b'' \backslash b|\, |b'|^{-1} + |b' \backslash (b \cup b'')|\, |b'|^{-1}\Bigg]^{\frac{1}{p^*}}
\end{aligned}$$

and

$$
\begin{aligned}
&c_{p^*}\left(b', b''\right) = \left(\|s\|_{p^*}^{p^*}\right)^{\frac{1}{p^*}} \\
&= \Bigg[\left|b \cap b' \cap b''\right| \left| \left|b'\right|^{-\frac{1}{p^*}} - \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*} + \left|b' \cap b'' \backslash b\right| \left| \left|b'\right|^{-\frac{1}{p^*}} - \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*} \\
&\qquad + \left|b \cap b' \backslash b''\right| \left|b'\right|^{-1} + \left|b' \backslash \left(b \cup b''\right)\right| \left|b'\right|^{-1} \\
&\qquad + \left|b \cap b'' \backslash b'\right| \left|b''\right|^{-1} + \left|b'' \backslash \left(b \cup b'\right)\right| \left|b''\right|^{-1} \Bigg]^{\frac{1}{p^*}}.
\end{aligned}
$$

For $p^* = 1$, we have $\|r+s\|_1 = \sum_i |r_i + s_i| \leq \sum_i \left(|r_i| + |s_i|\right) = \|r\|_1 + \|s\|_1$. For $p^* = \infty$, we have $\|r+s\|_\infty = \max_i |r_i + s_i| \leq \max_i \left(|r_i| + |s_i|\right) \leq \|r\|_\infty + \|s\|_\infty$. We now assume $1 < p^* < \infty$. Let $\operatorname{sgn}(t) \in \{-1, 0, 1\}$ denote the sign of t. Then

$$
\begin{aligned}
&\|r+s\|_{p^*}^{p^*} = \sum_{i=1}^d |r_i + s_i|^{p^*} \\
&= \sum_{i=1}^d |r_i + s_i|^{p^*-1} \operatorname{sgn}\left(r_i + s_i\right)\left(r_i + s_i\right) \\
&= \sum_{i=1}^d |r_i + s_i|^{p^*-1} \operatorname{sgn}\left(r_i + s_i\right) r_i + \sum_{i=1}^d |r_i + s_i|^{p^*-1} \operatorname{sgn}\left(r_i + s_i\right) s_i \\
&\leq \sum_{i=1}^d |r_i + s_i|^{p^*-1} |r_i| + \sum_{i=1}^d |r_i + s_i|^{p^*-1} |s_i| \\
&\leq \left(\sum_{i=1}^d |r_i|^{p^*}\right)^{\frac{1}{p^*}} \left(\sum_{i=1}^d \left(|r_i + s_i|^{p^*-1}\right)^{\frac{p^*}{p^*-1}}\right)^{\frac{p^*-1}{p^*}} \\
&\qquad + \left(\sum_{i=1}^d |s_i|^{p^*}\right)^{\frac{1}{p^*}} \left(\sum_{i=1}^d \left(|r_i + s_i|^{p^*-1}\right)^{\frac{p^*}{p^*-1}}\right)^{\frac{p^*-1}{p^*}} \\
&= \|r\|_{p^*} \left(\sum_{i=1}^d |r_i + s_i|^{p^*}\right)^{\frac{p^*-1}{p^*}} + \|s\|_{p^*} \left(\sum_{i=1}^d |r_i + s_i|^{p^*}\right)^{\frac{p^*-1}{p^*}} \\
&= \left(\|r\|_{p^*} + \|s\|_{p^*}\right) \|r+s\|_{p^*}^{p^*-1}, \qquad (19)
\end{aligned}
$$

where the second equality uses the identity $|x|^{p^*} = |x|^{p^*-1} \operatorname{sgn}(x) x$ and the second inequality follows from Hölder's inequality. Since $|r_i + s_i|$ and $|r_i|$ are constant on each region in (18), we obtain the exact equality

$$
\begin{aligned}
&\sum_{i=1}^d |(r+s)_i|^{p^*-1} |r_i| \\
&= \left|b \cap b' \cap b''\right| \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b'\right|^{-\frac{1}{p^*}} \right| \\
&\qquad + \left|b \cap b' \backslash b''\right| \left| \left|b\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b'\right|^{-\frac{1}{p^*}} \right| \\
&+ \left|b \cap b'' \backslash b'\right| \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b\right|^{-\frac{1}{p^*}} \right| + \left|b' \cap b'' \backslash b\right| \left| \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b'\right|^{-\frac{1}{p^*}} \right| \\
&+ \left|b \backslash \left(b' \cup b''\right)\right| \left| \left|b\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b\right|^{-\frac{1}{p^*}} \right| + \left|b' \backslash \left(b \cup b''\right)\right| \cdot 0^{p^*-1} \cdot \left| \left|b'\right|^{-\frac{1}{p^*}} \right| \\
&+ \left|b'' \backslash \left(b \cup b'\right)\right| \left| \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \cdot 0 + \left|[d] \backslash \left(b \cup b' \cup b''\right)\right| \cdot 0^{p^*-1} \cdot 0 \\
&= \left|b \cap b' \cap b''\right| \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b''\right|^{-\frac{1}{p^*}} \right|^{p^*-1} \left| \left|b\right|^{-\frac{1}{p^*}} - \left|b'\right|^{-\frac{1}{p^*}} \right|
\end{aligned}
$$

$$
\begin{aligned}
&+ |b\cap b'\backslash b''|\,|b|^{-\frac{p^*-1}{p^*}}\left||b|^{-\frac{1}{p^*}}-|b'|^{-\frac{1}{p^*}}\right| \\
&+ |b\cap b''\backslash b'|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*-1}|b|^{-\frac{1}{p^*}} \\
&+ |b'\cap b''\backslash b|\,|b''|^{-\frac{p^*-1}{p^*}}|b'|^{-\frac{1}{p^*}}+|b\backslash(b'\cup b'')|\,|b|^{-\frac{p^*}{p^*}} \\
&\overset{(19)}{\leq}\left[|b\cap b'\cap b''|\left||b|^{-\frac{1}{p^*}}-|b'|^{-\frac{1}{p^*}}\right|^{p^*}+|b\cap b'\backslash b''|\left||b|^{-\frac{1}{p^*}}-|b'|^{-\frac{1}{p^*}}\right|^{p^*}\right. \\
&\quad+|b\cap b''\backslash b'|\,|b|^{-1}+|b\backslash(b'\cup b'')|\,|b|^{-1} \\
&\quad\left.+|b'\cap b''\backslash b|\,|b'|^{-1}+|b'\backslash(b\cup b'')|\,|b'|^{-1}\right]^{\frac{1}{p^*}} \\
&\quad\cdot\left[|b\cap b'\cap b''|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}+|b'\cap b''\backslash b|\,|b''|^{-1}\right. \\
&\quad+|b''\backslash(b\cup b')|\,|b''|^{-1}+|b\cap b''\backslash b'|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*} \\
&\quad\left.+|b\cap b'\backslash b''|\,|b|^{-1}+|b\backslash(b'\cup b'')|\,|b|^{-1}\right]^{\frac{p^*-1}{p^*}}. \qquad (20)
\end{aligned}
$$

Likewise, we have

$$
\begin{aligned}
&\sum_{i=1}^{d}|(r+s)_i|^{p^*-1}|s_i| \\
&=|b\cap b'\cap b''|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*-1}\left||b'|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|+|b\cap b'\backslash b''|\,|b|^{-\frac{p^*-1}{p^*}}|b'|^{-\frac{1}{p^*}} \\
&+|b\cap b''\backslash b'|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*-1}|b''|^{-\frac{1}{p^*}}+|b'\cap b''\backslash b|\,|b''|^{-\frac{p^*-1}{p^*}}\left||b'|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right| \\
&+|b''\backslash(b\cup b')|\,|b''|^{-\frac{p^*}{p^*}} \\
&\overset{(19)}{\leq}\left[|b\cap b'\cap b''|\left||b'|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}+|b'\cap b''\backslash b|\left||b'|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}\right. \\
&\quad+|b\cap b'\backslash b''|\,|b'|^{-1}+|b'\backslash(b\cup b'')|\,|b'|^{-1} \\
&\quad\left.+|b\cap b''\backslash b'|\,|b''|^{-1}+|b''\backslash(b\cup b')|\,|b''|^{-1}\right]^{\frac{1}{p^*}}\cdot\left[|b\cap b'\cap b''|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}\right. \\
&\quad+|b'\cap b''\backslash b|\,|b''|^{-1} \\
&+|b''\backslash(b\cup b')|\,|b''|^{-1}+|b\cap b''\backslash b'|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*} \\
&\quad\left.+|b\cap b'\backslash b''|\,|b|^{-1}+|b\backslash(b'\cup b'')|\,|b|^{-1}\right]^{\frac{p^*-1}{p^*}}. \qquad (21)
\end{aligned}
$$

We now establish the triangle inequality:

$$
\begin{aligned}
&\operatorname{dist}(\mathcal{G},\mathcal{G}'')=\min_{\pi}\sum_{b\in\widetilde{\mathcal{G}}}\sum_{b''\in\widetilde{\mathcal{G}}''}\pi(b,b'')\,c_{p^*}(b,b'') \\
&\leq\sum_{b\in\widetilde{\mathcal{G}}}\sum_{b''\in\widetilde{\mathcal{G}}''}\pi_{13}(b,b'')\left[|b\cap b'\cap b''|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}\right. \\
&\quad+|b'\cap b''\backslash b|\,|b''|^{-1}+|b''\backslash(b\cup b')|\,|b''|^{-1} \\
&\quad\left.+|b\cap b''\backslash b'|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}+|b\cap b'\backslash b''|\,|b|^{-1}+|b\backslash(b'\cup b'')|\,|b|^{-1}\right]^{\frac{1}{p^*}} \\
&\overset{(9)}{\leq}\sum_{b}\sum_{b''}\left(\sum_{b'}\pi_{12}(b,b')\,\frac{\pi_{23}(b',b'')}{\frac{|b'|^{\frac{1}{p^*}}}{\sum_{a'}|a'|^{\frac{1}{p^*}}}}\right)
\end{aligned}
$$

$$
\left[\left(|b\cap b''|\left||b|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}+\frac{|b\backslash b''|}{|b|}+\frac{|b''\backslash b|}{|b''|}\right)^{\frac{1}{p^*}}\right]
$$

$$
\begin{aligned}
&\leq \sum_{b}\sum_{b'}\sum_{b''}\pi_{12}\left(b,b'\right)\pi_{23}\left(b',b''\right)\frac{\sum_{a'}|a'|^{\frac{1}{p^*}}}{|b'|^{\frac{1}{p^*}}}\\
&\quad\left[\left(|b\cap b'|\left||b|^{-\frac{1}{p^*}}-|b'|^{-\frac{1}{p^*}}\right|^{p^*}+\frac{|b\backslash b'|}{|b|}+\frac{|b'\backslash b|}{|b'|}\right)^{\frac{1}{p^*}}\right.\\
&\qquad\left.+\left(|b'\cap b''|\left||b'|^{-\frac{1}{p^*}}-|b''|^{-\frac{1}{p^*}}\right|^{p^*}+\frac{|b'\backslash b''|}{|b'|}+\frac{|b''\backslash b'|}{|b''|}\right)^{\frac{1}{p^*}}\right]\\
&\overset{(20),(21)}{\leq}\sum_{b}\sum_{b'}\pi_{12}\left(b,b'\right)\left(\|r\|_{p^*}^{p^*}\right)^{\frac{1}{p^*}}\\
&\quad+\sum_{b'}\sum_{b''}\pi_{23}\left(b',b''\right)\left(\|s\|_{p^*}^{p^*}\right)^{\frac{1}{p^*}}=\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)+\operatorname{dist}\left(\mathcal{G}',\mathcal{G}''\right).
\end{aligned}
$$

□

E Proof of Proposition 3.6

Proof. Sums below omit empty blocks. The norm and transport bounds are written for $1<p<\infty$ and extend to $p=1,\infty$ by continuity. Define

$$
Z_{\max}:=Z\triangleq\max\left\{\underbrace{\sum_{b\in\widetilde{\mathcal{G}}}|b|^{1/p^*}}_{\triangleq Z_{\mathcal{G}}},\ \underbrace{\sum_{b'\in\widetilde{\mathcal{G}}'}|b'|^{1/p^*}}_{\triangleq Z_{\mathcal{G}'}}\right\}.
$$

Since (2) is a finite-dimensional linear program with nonempty feasible set, we fix an optimal transport plan π attaining $\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)$. By definition we have

$$
\sum_{b\in\widetilde{\mathcal{G}}}\sum_{b'\in\widetilde{\mathcal{G}}'}\pi\left(b,b'\right)c_{p^*}\left(b,b'\right)=\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)
$$

with marginals

$$
\sum_{b'}\pi\left(b,b'\right)=\frac{|b|^{1/p^*}}{Z_{\mathcal{G}}},\quad\sum_{b}\pi\left(b,b'\right)=\frac{|b'|^{1/p^*}}{Z_{\mathcal{G}'}}.
$$

Starting from the definition and applying the Jensen inequality for norms with weights $\pi\left(b,b'\right)\geq 0$ satisfying $\sum_{b,b'}\pi\left(b,b'\right)=1$, we obtain

$$
\begin{aligned}
\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)&=\sum_{b,b'}\pi\left(b,b'\right)\left\||b|^{-1/p^*}\mathbf{1}_b-|b'|^{-1/p^*}\mathbf{1}_{b'}\right\|_{p^*}\\
&\geq\left\|\sum_{b,b'}\pi\left(b,b'\right)\left(|b|^{-1/p^*}\mathbf{1}_b-|b'|^{-1/p^*}\mathbf{1}_{b'}\right)\right\|_{p^*}\\
&=\left\|\sum_{b,b'}\pi\left(b,b'\right)|b|^{-1/p^*}\mathbf{1}_b-\sum_{b,b'}\pi\left(b,b'\right)|b'|^{-1/p^*}\mathbf{1}_{b'}\right\|_{p^*}\\
&=\left\|\sum_{b}\left(\sum_{b'}\pi\left(b,b'\right)\right)|b|^{-1/p^*}\mathbf{1}_b-\sum_{b'}\left(\sum_{b}\pi\left(b,b'\right)\right)|b'|^{-1/p^*}\mathbf{1}_{b'}\right\|_{p^*}\\
&=\left\|\sum_{b}\frac{|b|^{1/p^*}}{Z_{\mathcal{G}}}|b|^{-1/p^*}\mathbf{1}_b-\sum_{b'}\frac{|b'|^{1/p^*}}{Z_{\mathcal{G}'}}|b'|^{-1/p^*}\mathbf{1}_{b'}\right\|_{p^*}
\end{aligned}
$$

$$
\begin{aligned}
&= \left\| \frac{1}{Z_{\mathcal{G}}} \sum_{b} \mathbf{1}_b - \frac{1}{Z_{\mathcal{G}'}} \sum_{b'} \mathbf{1}_{b'} \right\|_{p^*} \\
&\geq \left(\sum_{i=1}^{d} \left| \left(\frac{1}{Z_{\mathcal{G}}} - \frac{1}{Z_{\mathcal{G}'}} \right) \cdot 1 \right|^{p^*} \right)^{1/p^*} \\
&= \left(d \cdot \left| \frac{1}{Z_{\mathcal{G}}} - \frac{1}{Z_{\mathcal{G}'}} \right|^{p^*} \right)^{1/p^*} = d^{1/p^*} \left| \frac{1}{Z_{\mathcal{G}}} - \frac{1}{Z_{\mathcal{G}'}} \right| .
\end{aligned}
$$

Thus

$$
\left| \frac{1}{Z_{\mathcal{G}}} - \frac{1}{Z_{\mathcal{G}'}} \right| \leq d^{-1/p^*} \operatorname{dist}(\mathcal{G}, \mathcal{G}') . \tag{22}
$$

We now expand $|Z_{\mathcal{G}} - Z_{\mathcal{G}'}|$ via reciprocals to obtain

$$
\begin{aligned}
|Z_{\mathcal{G}} - Z_{\mathcal{G}'}| &= |Z_{\mathcal{G}} - Z_{\mathcal{G}'}| \cdot \frac{Z_{\mathcal{G}} Z_{\mathcal{G}'}}{Z_{\mathcal{G}} Z_{\mathcal{G}'}} \\
&= Z_{\mathcal{G}} Z_{\mathcal{G}'} \left| \frac{1}{Z_{\mathcal{G}}} - \frac{1}{Z_{\mathcal{G}'}} \right| \overset{(22)}{\leq} Z_{\mathcal{G}} Z_{\mathcal{G}'} \cdot d^{-1/p^*} \operatorname{dist}(\mathcal{G}, \mathcal{G}') .
\end{aligned} \tag{23}
$$

We bound $Z_{\mathcal{G}} \leq d$ and $Z_{\mathcal{G}'} \leq d$ using a partition-counting argument. Observe that

$$
Z_{\mathcal{G}} = \sum_{b \in \widetilde{\mathcal{G}}} |b|^{1/p^*} \leq \sum_{b \in \widetilde{\mathcal{G}}} |b| = \sum_{b \in \widetilde{\mathcal{G}}} \sum_{i=1}^{d} (\mathbf{1}_b)_i = \sum_{i=1}^{d} \sum_{b \in \widetilde{\mathcal{G}}} (\mathbf{1}_b)_i = \sum_{i=1}^{d} 1 = d,
$$

where the inequality $|b|^{1/p^*} \leq |b|$ follows from $|b| \geq 1$ and $1/p^* \leq 1$, and the identity $\sum_{b \in \widetilde{\mathcal{G}}} (\mathbf{1}_b)_i = 1$ holds since $\widetilde{\mathcal{G}}$ partitions $[d]$. The same reasoning yields $Z_{\mathcal{G}'} \leq d$, hence $Z_{\mathcal{G}} Z_{\mathcal{G}'} \leq d^2$. Substituting into (23) gives

$$
|Z_{\mathcal{G}} - Z_{\mathcal{G}'}| \leq d^2 \cdot d^{-1/p^*} \operatorname{dist}(\mathcal{G}, \mathcal{G}') = d^{2 - \frac{1}{p^*}} \operatorname{dist}(\mathcal{G}, \mathcal{G}') . \tag{24}
$$

Fix arbitrary $w \in \mathbb{R}^d$ and nonempty $b, b' \subseteq [d]$. By rewriting block norms and inserting the normalization used in c_{p^*}, we have

$$
\begin{aligned}
\|w_b\|_p &= \left(\sum_{i=1}^{d} |w_i|^p (\mathbf{1}_b)_i \right)^{1/p} \\
&= \left(\sum_{i=1}^{d} |w_i|^p \left(|b|^{1/p^*} \cdot |b|^{-1/p^*} (\mathbf{1}_b)_i \right)^p \right)^{1/p} \\
&= |b|^{1/p^*} \left(\sum_{i=1}^{d} |w_i|^p \left(|b|^{-1/p^*} (\mathbf{1}_b)_i \right)^p \right)^{1/p} = |b|^{1/p^*} \left\| \left(w_i \cdot |b|^{-1/p^*} (\mathbf{1}_b)_i \right)_{i=1}^{d} \right\|_p ,
\end{aligned}
$$

and similarly

$$
\|w_{b'}\|_p = |b'|^{1/p^*} \left\| \left(w_i \cdot |b'|^{-1/p^*} (\mathbf{1}_{b'})_i \right)_{i=1}^{d} \right\|_p .
$$

Hence

$$
\begin{aligned}
&\left| \|w_b\|_p - \|w_{b'}\|_p \right| \\
&= \left| |b|^{1/p^*} \left\| \left(w_i \cdot |b|^{-1/p^*} (\mathbf{1}_b)_i \right)_{i=1}^{d} \right\|_p - |b'|^{1/p^*} \left\| \left(w_i \cdot |b'|^{-1/p^*} (\mathbf{1}_{b'})_i \right)_{i=1}^{d} \right\|_p \right| \\
&\leq \underbrace{|b|^{1/p^*} \left| \left\| \left(w_i \cdot |b|^{-1/p^*} (\mathbf{1}_b)_i \right)_i \right\|_p - \left\| \left(w_i \cdot |b'|^{-1/p^*} (\mathbf{1}_{b'})_i \right)_i \right\|_p \right|}_{\text{(I)}} \\
&\quad + \underbrace{\left| |b|^{1/p^*} - |b'|^{1/p^*} \right| \cdot \left\| \left(w_i \cdot |b'|^{-1/p^*} (\mathbf{1}_{b'})_i \right)_i \right\|_p}_{\text{(II)}} .
\end{aligned}
$$

For term (I), we use $|\|x\|_p - \|y\|_p| \le \|x-y\|_p$, then expand $\|\cdot\|_p$ and apply $\|\cdot\|_\infty \le \|\cdot\|_{p^*}$ to get

$$\begin{aligned}
\text{(I)} &\le |b|^{1/p^*} \left\| \left(w_i \cdot |b|^{-1/p^*} \left(\mathbf{1}_b\right)_i \right)_i - \left(w_i \cdot |b'|^{-1/p^*} \left(\mathbf{1}_{b'}\right)_i \right)_i \right\|_p \\
&= |b|^{1/p^*} \left(\sum_{i=1}^d |w_i|^p \cdot \left| |b|^{-1/p^*} \left(\mathbf{1}_b\right)_i - |b'|^{-1/p^*} \left(\mathbf{1}_{b'}\right)_i \right|^p \right)^{1/p} \\
&\le |b|^{1/p^*} \left(\sum_{i=1}^d |w_i|^p \cdot \left(\max_{1\le j\le d} \left| |b|^{-1/p^*} \left(\mathbf{1}_b\right)_j - |b'|^{-1/p^*} \left(\mathbf{1}_{b'}\right)_j \right| \right)^p \right)^{1/p} \\
&= |b|^{1/p^*} \left(\max_{1\le j\le d} \left| |b|^{-1/p^*} \left(\mathbf{1}_b\right)_j - |b'|^{-1/p^*} \left(\mathbf{1}_{b'}\right)_j \right| \right) \left(\sum_{i=1}^d |w_i|^p \right)^{1/p} \\
&\le |b|^{1/p^*} \left\| |b|^{-1/p^*} \mathbf{1}_b - |b'|^{-1/p^*} \mathbf{1}_{b'} \right\|_{p^*} \|w\|_p \\
&= |b|^{1/p^*} c_{p^*}(b,b') \|w\|_p \le d^{1/p^*} c_{p^*}(b,b') \|w\|_p. \qquad (25)
\end{aligned}$$

For term (II), we first bound the p-norm term by $\|w\|_p$ and then bound the size difference. Calculating the norm term yields

$$\begin{aligned}
&\left\| \left(w_i \cdot |b'|^{-1/p^*} \left(\mathbf{1}_{b'}\right)_i \right)_i \right\|_p \\
&= \left(\sum_{i\in b'} |w_i|^p \cdot |b'|^{-p/p^*} \right)^{1/p} \\
&= |b'|^{-1/p^*} \left(\sum_{i\in b'} |w_i|^p \right)^{1/p} \\
&= |b'|^{-1/p^*} \|w_{b'}\|_p \\
&\le |b'|^{-1/p^*} \|w\|_p \le \|w\|_p.
\end{aligned}$$

For the size term, if $b \cap b' \neq \varnothing$ we pick $i \in b \cap b'$ and observe that

$$\begin{aligned}
c_{p^*}(b,b') &= \left\| |b|^{-1/p^*} \mathbf{1}_b - |b'|^{-1/p^*} \mathbf{1}_{b'} \right\|_{p^*} \\
&\ge \left| \left(|b|^{-1/p^*} \mathbf{1}_b - |b'|^{-1/p^*} \mathbf{1}_{b'} \right)_i \right| \qquad (26) \\
&= \left| |b|^{-1/p^*} - |b'|^{-1/p^*} \right|,
\end{aligned}$$

from which we deduce

$$\left| |b|^{1/p^*} - |b'|^{1/p^*} \right| \overset{(26)}{\le} |b|^{1/p^*} |b'|^{1/p^*} \cdot c_{p^*}(b,b') \le d^{2/p^*} c_{p^*}(b,b').$$

If instead $b \cap b' = \varnothing$, then $c_{p^*}(b,b') = 2^{1/p^*} \ge 1$ while $\left| |b|^{1/p^*} - |b'|^{1/p^*} \right| \le d^{1/p^*} \le d^{2/p^*} c_{p^*}(b,b')$. Hence in all cases we have

$$\left| |b|^{1/p^*} - |b'|^{1/p^*} \right| \le d^{2/p^*} c_{p^*}(b,b').$$

This yields

$$\text{(II)} \le d^{2/p^*} c_{p^*}(b,b') \cdot \|w\|_p. \qquad (27)$$

Combining (25) and (27), we obtain

$$\begin{aligned}
\left| \|w_b\|_p - \|w_{b'}\|_p \right| &\le d^{1/p^*} \|w\|_p c_{p^*}(b,b') + d^{2/p^*} \|w\|_p c_{p^*}(b,b') \\
&= \left(d^{1/p^*} + d^{2/p^*} \right) \|w\|_p c_{p^*}(b,b') \qquad (28) \\
&\le 2 d^{2/p^*} \|w\|_p c_{p^*}(b,b').
\end{aligned}$$

By definition, adding the common complement block does not affect the difference, so

$$\begin{aligned}\left|\mathbb{F}_{\mathcal{G}}(w)-\mathbb{F}_{\mathcal{G}'}(w)\right| &= \lambda\left|\sum_{b\in\widetilde{\mathcal{G}}}|b|^{1/p^*}\|w_b\|_p-\sum_{b'\in\widetilde{\mathcal{G}}'}|b'|^{1/p^*}\|w_{b'}\|_p\right| \\ &= \lambda\left|Z_{\mathcal{G}}\sum_{b\in\widetilde{\mathcal{G}}}\left(\sum_{b'\in\widetilde{\mathcal{G}}'}\pi\left(b,b'\right)\right)\|w_b\|_p-Z_{\mathcal{G}'}\sum_{b'\in\widetilde{\mathcal{G}}'}\left(\sum_{b\in\widetilde{\mathcal{G}}}\pi\left(b,b'\right)\right)\|w_{b'}\|_p\right| \\ &= \lambda\left|Z_{\mathcal{G}}\sum_{b,b'}\pi\left(b,b'\right)\|w_b\|_p-Z_{\mathcal{G}'}\sum_{b,b'}\pi\left(b,b'\right)\|w_{b'}\|_p\right|.\end{aligned}$$

Adding and subtracting $Z_{\mathcal{G}'}\sum_{b,b'}\pi\left(b,b'\right)\|w_b\|_p$ and applying the triangle inequality, we get

$$\begin{aligned}&\left|\mathbb{F}_{\mathcal{G}}(w)-\mathbb{F}_{\mathcal{G}'}(w)\right| = \lambda\left|\left(Z_{\mathcal{G}}-Z_{\mathcal{G}'}\right)\sum_{b,b'}\pi\left(b,b'\right)\|w_b\|_p+Z_{\mathcal{G}'}\sum_{b,b'}\pi\left(b,b'\right)\left(\|w_b\|_p-\|w_{b'}\|_p\right)\right| \\ &\leq \lambda\left(|Z_{\mathcal{G}}-Z_{\mathcal{G}'}|\sum_{b,b'}\pi\left(b,b'\right)\|w_b\|_p+Z_{\mathcal{G}'}\sum_{b,b'}\pi\left(b,b'\right)\left|\|w_b\|_p-\|w_{b'}\|_p\right|\right) \\ &\overset{(28)}{\leq} \lambda\left(|Z_{\mathcal{G}}-Z_{\mathcal{G}'}|\sum_{b}\frac{|b|^{1/p^*}}{Z_{\mathcal{G}}}\|w_b\|_p+Z_{\mathcal{G}'}\sum_{b,b'}\pi\left(b,b'\right)\left(2d^{2/p^*}\|w\|_p c_{p^*}\left(b,b'\right)\right)\right) \\ &\leq \lambda\left(|Z_{\mathcal{G}}-Z_{\mathcal{G}'}|\sum_{b}\frac{|b|^{1/p^*}}{Z_{\mathcal{G}}}\|w\|_p+Z_{\mathcal{G}'}2d^{2/p^*}\|w\|_p\sum_{b,b'}\pi\left(b,b'\right)c_{p^*}\left(b,b'\right)\right) \\ &= \lambda\left(\|w\|_p\cdot\frac{1}{Z_{\mathcal{G}}}\sum_{b}|b|^{1/p^*}\cdot|Z_{\mathcal{G}}-Z_{\mathcal{G}'}|+Z_{\mathcal{G}'}\cdot 2d^{2/p^*}\|w\|_p\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)\right) \\ &\overset{(24)}{\leq} \lambda\|w\|_p\left(d^{2-\frac{1}{p^*}}\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)+2\max\left\{Z_{\mathcal{G}}d^{2/p^*},Z_{\mathcal{G}'}d^{2/p^*}\right\}\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right)\right) \\ &= \lambda\|w\|_p\left(d^{2-\frac{1}{p^*}}+2d^{2/p^*}\max\left\{Z_{\mathcal{G}},Z_{\mathcal{G}'}\right\}\right)\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right).\end{aligned}$$

We now fix an arbitrary minimizer $w\in\mathcal{W}_{\mathcal{G}}^{\star}$, so that $\mathbb{F}_{\mathcal{G}}(w)=\min\mathbb{F}_{\mathcal{G}}$ and $\|w\|_p\leq R$. Since $\min\mathbb{F}_{\mathcal{G}'}\leq\mathbb{F}_{\mathcal{G}'}(w)$ and $\min\mathbb{F}_{\mathcal{G}}=\mathbb{F}_{\mathcal{G}}(w)$, we have

$$\begin{aligned}\min\mathbb{F}_{\mathcal{G}'}-\min\mathbb{F}_{\mathcal{G}} &\leq \mathbb{F}_{\mathcal{G}'}(w)-\mathbb{F}_{\mathcal{G}}(w) \\ &\leq \left|\mathbb{F}_{\mathcal{G}}(w)-\mathbb{F}_{\mathcal{G}'}(w)\right| \\ &\leq \lambda R\left(d^{2-\frac{1}{p^*}}+2d^{2/p^*}Z_{\max}\right)\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right).\end{aligned}\tag{29}$$

Let $v\in\mathcal{W}_{\mathcal{G}'}^{\star}$ be any element, which exists by Assumption 2.1. Then $\min\mathbb{F}_{\mathcal{G}'}=\mathbb{F}_{\mathcal{G}'}(v)$ and $\min\mathbb{F}_{\mathcal{G}}\leq\mathbb{F}_{\mathcal{G}}(v)$, so

$$\begin{aligned}\min\mathbb{F}_{\mathcal{G}}-\min\mathbb{F}_{\mathcal{G}'} &\leq \left|\mathbb{F}_{\mathcal{G}}(v)-\mathbb{F}_{\mathcal{G}'}(v)\right| \\ &\leq \lambda\|v\|_p\left(d^{2-\frac{1}{p^*}}+2d^{2/p^*}Z_{\max}\right)\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right) \\ &\leq \lambda R\left(d^{2-\frac{1}{p^*}}+2d^{2/p^*}Z_{\max}\right)\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right).\end{aligned}\tag{30}$$

Combining (29) and (30) yields the absolute difference bound

$$\left|\min\mathbb{F}_{\mathcal{G}}-\min\mathbb{F}_{\mathcal{G}'}\right|\leq\lambda R\left(d^{2-\frac{1}{p^*}}+2d^{2/p^*}Z_{\max}\right)\operatorname{dist}\left(\mathcal{G},\mathcal{G}'\right).\tag{31}$$

Applying the order-α growth condition with $\mathcal{H}=\mathcal{G}'$ at $w\in\mathcal{U}$ gives

$$\mathbb{F}_{\mathcal{G}'}(w)-\min\mathbb{F}_{\mathcal{G}'}\geq\kappa\left(\inf_{u\in\mathcal{W}_{\mathcal{G}'}^{\star}}\|w-u\|_p\right)^{\alpha}.\tag{32}$$

Meanwhile, since $\mathbb{F}_{\mathcal{G}}(w) = \min \mathbb{F}_{\mathcal{G}}$ we have

$$\begin{aligned}
&\mathbb{F}_{\mathcal{G}'}(w) - \min \mathbb{F}_{\mathcal{G}'} = (\mathbb{F}_{\mathcal{G}'}(w) - \min \mathbb{F}_{\mathcal{G}}) + (\min \mathbb{F}_{\mathcal{G}} - \min \mathbb{F}_{\mathcal{G}'}) \\
&\leq \left|\mathbb{F}_{\mathcal{G}'}(w) - \mathbb{F}_{\mathcal{G}}(w)\right| + \left|\min \mathbb{F}_{\mathcal{G}} - \min \mathbb{F}_{\mathcal{G}'}\right| \\
&\overset{(31)}{\leq} \lambda R \left(d^{2-\frac{1}{p^*}} + 2d^{2/p^*} Z_{\max}\right) \operatorname{dist}(\mathcal{G}, \mathcal{G}') + \lambda R \left(d^{2-\frac{1}{p^*}} + 2d^{2/p^*} Z_{\max}\right) \operatorname{dist}(\mathcal{G}, \mathcal{G}') \\
&= 2\lambda R \left(d^{2-\frac{1}{p^*}} + 2d^{2/p^*} Z\right) \operatorname{dist}(\mathcal{G}, \mathcal{G}'). \qquad (33)
\end{aligned}$$

Combining (33) with (32) gives

$$\kappa \left(\inf_{u \in \mathcal{W}^\star_{\mathcal{G}'}} \|w - u\|_p\right)^\alpha \leq 2\lambda R \left(d^{2-\frac{1}{p^*}} + 2d^{2/p^*} Z\right) \operatorname{dist}(\mathcal{G}, \mathcal{G}'). \qquad (34)$$

Dividing (34) by κ, taking the $1/\alpha$ power, and then taking the supremum over $w \in \mathcal{W}^\star_{\mathcal{G}}$ yields

$$\sup_{w \in \mathcal{W}^\star_{\mathcal{G}}} \inf_{u \in \mathcal{W}^\star_{\mathcal{G}'}} \|w - u\|_p \leq \left[\frac{2\lambda R}{\kappa} \left(d^{2-\frac{1}{p^*}} + 2d^{2/p^*} Z\right) \operatorname{dist}(\mathcal{G}, \mathcal{G}')\right]^{1/\alpha},$$

which holds for every $w \in \mathcal{W}^\star_{\mathcal{G}}$. Interchanging $\mathcal{G}$ and $\mathcal{G}'$ gives the reverse directed bound. □

F Proof of Theorem 3.7

Proof. We work on the linear subspace $\left\{w \in \mathbb{R}^d : w_{\overline{\mathcal{A}}} = 0\right\}$. We introduce an auxiliary deformation parameter $t \in [0, 1]$ and consider $w^\star$ as a function of t. Fix $t \in [0, 1]$ and $\delta \neq 0$ such that $t + \delta \in [0, 1]$. We define a one-parameter family of objectives $\{w^\star(t)\}_{t \in [0,1]}$ that continuously deforms the pre-change objective into the post-change objective. We also define the deterministic state schedule $S^\urcorner(w, t) \in \mathbb{R}^{|\mathcal{A}|}$ for $w_{\overline{\mathcal{A}}} = \mathbf{0}$ by

$$\begin{aligned}
S^\urcorner(w, t) := \nabla_{w_{\mathcal{A}}} f(\mathcal{D}; w) \qquad &+ \lambda \sum_{\substack{g=1 \\ g \neq g_{\text{src}}, g_{\text{dest}}}}^{G} c_g \nabla_{w_{\mathcal{A}}} \left\|w_{b_g}\right\|_p \\
&+ \lambda c_{g_{\text{src}}} \nabla_{w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p \\
&+ \lambda c_{g_{\text{dest}}} \nabla_{w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{dest}}}}, t w_j\right)\right\|_p.
\end{aligned}$$

The minimizer $w^\star(t)$ satisfies stationarity $\mathbf{0} \in S^\urcorner\left(w^\star(t), t\right)$ for all $t \in [0, 1]$. Hence

$$0 \in S^\urcorner\left(w^\star(t + \delta), t + \delta\right) - S^\urcorner\left(w^\star(t), t\right).$$

For clarity, we rewrite the norms to expose the exact dependence as

$$\begin{aligned}
\left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p^p &= \sum_{i \in (b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}} |w_i|^p + (1-t)^p |w_j|^p, \\
\left\|\left(w_{b_{g_{\text{dest}}}}, t w_j\right)\right\|_p^p &= \sum_{i \in b_{g_{\text{dest}}} \cap \mathcal{A}} |w_i|^p + t^p |w_j|^p.
\end{aligned}$$

For nonzero norms, we obtain the explicit gradient blocks

$$\begin{aligned}
\left(\nabla_{w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p\right)_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}} &= \frac{\left|w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}}\right|^{p-2} \odot w_{(b_{g_{\text{src}}} \setminus \{j\}) \cap \mathcal{A}}}{\left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p^{p-1}}, \\
\left(\nabla_{w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p\right)_j &= (1-t)^p \frac{|w_j|^{p-2} w_j}{\left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p^{p-1}}, \\
\left(\nabla_{w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}} \setminus \{j\}}, (1-t) w_j\right)\right\|_p\right)_{\mathcal{A} \setminus b_{g_{\text{src}}}} &= \mathbf{0},
\end{aligned}$$

$$\left(\nabla_{w_{\mathcal{A}}}\left\|\left(w_{b_{g_{\text{dest}}}}, tw_j\right)\right\|_p\right)_{b_{g_{\text{dest}}}\cap\mathcal{A}} = \frac{\left|w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\right|^{p-2}\odot w_{b_{g_{\text{dest}}}\cap\mathcal{A}}}{\left\|\left(w_{b_{g_{\text{dest}}}}, tw_j\right)\right\|_p^{p-1}},$$

$$\left(\nabla_{w_{\mathcal{A}}}\left\|\left(w_{b_{g_{\text{dest}}}}, tw_j\right)\right\|_p\right)_j = t^p\frac{|w_j|^{p-2}w_j}{\left\|\left(w_{b_{g_{\text{dest}}}}, tw_j\right)\right\|_p^{p-1}},$$

$$\left(\nabla_{w_{\mathcal{A}}}\left\|\left(w_{b_{g_{\text{dest}}}}, tw_j\right)\right\|_p\right)_{\mathcal{A}\setminus\left(b_{g_{\text{dest}}}\cup\{j\}\right)} = \mathbf{0}.$$

Define $\Delta_\delta(t) := w^\star(t+\delta) - w^\star(t)$ and consider the joint affine path $\theta \mapsto (w^\star(t) + \theta\Delta_\delta(t), t+\theta\delta)$. Applying the fundamental theorem of calculus in Banach spaces yields

$$\begin{aligned}0 &= \int_0^1 \frac{d}{d\theta}S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)d\theta\\ &= \int_0^1 \left(\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)\Delta_\delta(t) + \partial_t S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)\delta\right)d\theta.\end{aligned}$$

Dividing by $\delta \neq 0$ and rearranging, we obtain

$$\left(\int_0^1 \nabla_w S\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)d\theta\right)\frac{\Delta_\delta(t)}{\delta} = -\int_0^1 \partial_t S\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)d\theta. \quad (35)$$

We rewrite the averaged Jacobian in (35) as a perturbation of $\nabla_w S^\urcorner$:

$$\begin{aligned}&\int_0^1 \nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)d\theta\\ &= \nabla_w S^\urcorner\left(w^\star(t), t\right) + \int_0^1 \left(\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right)d\theta\\ &= \nabla_w S^\urcorner\left(w^\star(t), t\right)\Bigg[I + \left(\nabla_w S^\urcorner\left(w^\star(t), t\right)\right)^{-1}\int_0^1\\ &\qquad \left(\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right)d\theta\Bigg]. \quad (36)\end{aligned}$$

Define the dimensionless perturbation operator

$$\Xi_\delta(t) := \left(\nabla_w S^\urcorner\left(w^\star(t), t\right)\right)^{-1}\int_0^1 \left(\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right)d\theta.$$

Then (36) reads $\nabla_w S^\urcorner\left(w^\star(t), t\right)\left[I + \Xi_\delta(t)\right] = \int_0^1 \nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right)d\theta$. On the smooth branch, the ℓ_p gradient satisfies $\nabla\|u\|_p = \dfrac{|u|^{p-2}\odot u}{\|u\|_p^{p-1}}$ for any nonzero vector u, where for every $g \neq g_{\text{src}}, g_{\text{dest}}$ with $\left\|w_{b_g\cap\mathcal{A}}\right\|_p \neq 0$ we have

$$\left(\nabla_{w_{\mathcal{A}}}\left\|w_{b_g}\right\|_p\right)_{b_g\cap\mathcal{A}} = \frac{\left|w_{b_g\cap\mathcal{A}}\right|^{p-2}\odot w_{b_g\cap\mathcal{A}}}{\left\|w_{b_g\cap\mathcal{A}}\right\|_p^{p-1}}, \qquad \left(\nabla_{w_{\mathcal{A}}}\left\|w_{b_g}\right\|_p\right)_{\mathcal{A}\setminus b_g} = \mathbf{0}.$$

Since $S^\urcorner \in C^1$ near $(w^\star(t), t)$ and $\Delta_\delta(t) \to 0$ as $\delta \to 0$ by local continuity of $w^\star$ on the smooth branch, applying the implicit function theorem to $S^\urcorner(w, t) = 0$ gives

$$\lim_{\delta\to 0}\sup_{\theta\in[0,1]}\left\|\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right\| = 0.$$

It follows that

$$\begin{aligned}\|\Xi_\delta(t)\| &\le \left\|\left(\nabla_w S^\urcorner\left(w^\star(t), t\right)\right)^{-1}\right\|\int_0^1\\ &\qquad \left\|\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right\|d\theta\\ &\le \left\|\left(\nabla_w S^\urcorner\left(w^\star(t), t\right)\right)^{-1}\right\|\\ &\qquad \cdot \lim_{\delta\to 0}\int_0^1 \left\|\nabla_w S^\urcorner\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \nabla_w S^\urcorner\left(w^\star(t), t\right)\right\|d\theta = 0. \quad (37)\end{aligned}$$

Therefore, for $|\delta|$ small enough we have $\|\Xi_\delta(t)\| \le 1$, so $I + \Xi_\delta(t)$ is invertible and admits the Neumann expansion [Meise and Vogt, 1997]

$$[I+\Xi_\delta(t)]^{-1} = \sum_{k=0}^{\infty} (-\Xi_\delta(t))^k, \quad \left\|[I+\Xi_\delta(t)]^{-1} - I\right\|$$
$$= \left\|\sum_{k=1}^{\infty} (-\Xi_\delta(t))^k\right\| \le \sum_{k=1}^{\infty} \|\Xi_\delta(t)\|^k = \frac{\|\Xi_\delta(t)\|}{1-\|\Xi_\delta(t)\|}. \tag{38}$$

Consequently,

$$\left(\int_0^1 \nabla_w S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta\right)^{-1} = [I+\Xi_\delta(t)]^{-1}\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}.$$

We can solve (35) as

$$\begin{aligned}
&\frac{\Delta_\delta(t)}{\delta} \\
&= -\left[I+\Xi_\delta(t)\right]^{-1}\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\left(\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta\right) \\
&= -\left[I+\Xi_\delta(t)\right]^{-1}\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1} \\
&\quad \left(\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta + \partial_t S^\intercal\left(w^\star(t),t\right) - \partial_t S^\intercal\left(w^\star(t),t\right)\right) \\
&= -\left[I+\Xi_\delta(t)\right]^{-1}\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\left(\partial_t S^\intercal\left(w^\star(t),t\right)\right. \\
&\quad \left.+\int_0^1 \left(\partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \partial_t S^\intercal\left(w^\star(t),t\right)\right) d\theta\right).
\end{aligned} \tag{39}$$

Decomposing the averaged forcing around $\partial_t S^\intercal\left(w^\star(t),t\right)$, we have

$$\begin{aligned}
&\left\|\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta - \partial_t S^\intercal\left(w^\star(t),t\right)\right\| \\
&\le \sup_{\theta\in[0,1]} \left\|\partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) - \partial_t S^\intercal\left(w^\star(t),t\right)\right\| \xrightarrow[\delta\to 0]{} 0
\end{aligned}$$

by continuity of $\partial_t S^\intercal$ and the fact that $\Delta_\delta(t) \to 0$. Subtracting the candidate limit from (39) and expanding the error exactly, we obtain

$$\begin{aligned}
&\left\|\frac{\Delta_\delta(t)}{\delta} + \left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\partial_t S^\intercal\left(w^\star(t),t\right)\right\| \\
&= \left\|-\left([I+\Xi_\delta(t)]^{-1} - I\right)\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\right. \\
&\quad \left(\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta\right) - \left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\cdot \\
&\quad \left.\left(\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta - \partial_t S^\intercal\left(w^\star(t),t\right)\right)\right\| \\
&\overset{(38)}{\le} \frac{\|\Xi_\delta(t)\|}{1-\|\Xi_\delta(t)\|}\cdot\left\|\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\right\|\cdot\left\|\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta\right\| \\
&\quad + \left\|\left(\nabla_w S^\intercal\left(w^\star(t),t\right)\right)^{-1}\right\|\cdot\left\|\int_0^1 \partial_t S^\intercal\left(w^\star(t)+\theta\Delta_\delta(t), t+\theta\delta\right) d\theta - \partial_t S^\intercal\left(w^\star(t),t\right)\right\|.
\end{aligned} \tag{40}$$

We now calculate the explicit Jacobian $\nabla_w S$ and the parameter derivative $\partial_t S$. On the smooth branch, for any nonzero u we have

$$\nabla^2\|u\|_p = \frac{p}{p^*}\left(\frac{\mathrm{Diag}\left(|u|^{p-2}\right)}{\|u\|_p^{p-1}} - \frac{\left(|u|^{p-2}\odot u\right)\left(|u|^{p-2}\odot u\right)^\top}{\|u\|_p^{2p-1}}\right).$$

Therefore, for $g \neq g_{\text{src}}, g_{\text{dest}}$ the contribution to $\nabla_w S$ is block-diagonal over $b_g \cap \mathcal{A}$:

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|w_{b_g}\right\|_p\right)_{b_g\cap\mathcal{A}, b_g\cap\mathcal{A}} = \frac{p}{p^*}\left(\frac{\operatorname{Diag}\left(\left|w_{b_g\cap\mathcal{A}}\right|^{p-2}\right)}{\left\|w_{b_g\cap\mathcal{A}}\right\|_p^{p-1}} - \frac{\left(\left|w_{b_g\cap\mathcal{A}}\right|^{p-2}\odot w_{b_g\cap\mathcal{A}}\right)\left(\left|w_{b_g\cap\mathcal{A}}\right|^{p-2}\odot w_{b_g\cap\mathcal{A}}\right)^\top}{\left\|w_{b_g\cap\mathcal{A}}\right\|_p^{2p-1}}\right),$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|w_{b_g}\right\|_p\right)_{\mathcal{A}\setminus b_g, \mathcal{A}} = \mathbf{0},$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|w_{b_g}\right\|_p\right)_{\mathcal{A}, \mathcal{A}\setminus b_g} = \mathbf{0}.$$

By careful observation, the source `MOVE` term contributes a Hessian supported on $b_{g_{\text{src}}} \cap \mathcal{A}$, and the destination `MOVE` term contributes one supported on $(b_{g_{\text{dest}}} \cup \{j\}) \cap \mathcal{A}$. Formally, we have

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}}\setminus\{j\}},\ (1-t)w_j\right)\right\|_p\right)_{b_{g_{\text{src}}}\cap\mathcal{A},\ b_{g_{\text{src}}}\cap\mathcal{A}} = \frac{p}{p^*}\left(\frac{\operatorname{Diag}\left(\left|w_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}\right|^{p-2},\ (1-t)^p\left|w_j\right|^{p-2}\right)}{\left\|\left(w_{b_{g_{\text{src}}}\setminus\{j\}},\ (1-t)w_j\right)\right\|_p^{p-1}} - \frac{\begin{bmatrix}\left|w_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}\right|^{p-2}\odot w_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}\\ (1-t)^p\left|w_j\right|^{p-2}w_j\end{bmatrix}\begin{bmatrix}\left|w_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}\right|^{p-2}\odot w_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}\\ (1-t)^p\left|w_j\right|^{p-2}w_j\end{bmatrix}^\top}{\left\|\left(w_{b_{g_{\text{src}}}\setminus\{j\}},\ (1-t)w_j\right)\right\|_p^{2p-1}}\right),$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}}\setminus\{j\}},\ (1-t)w_j\right)\right\|_p\right)_{\mathcal{A}\setminus b_{g_{\text{src}}},\ \mathcal{A}} = \mathbf{0},$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{src}}}\setminus\{j\}},\ (1-t)w_j\right)\right\|_p\right)_{\mathcal{A},\ \mathcal{A}\setminus b_{g_{\text{src}}}} = \mathbf{0},$$

and

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{dest}}}},\ tw_j\right)\right\|_p\right)_{\left(b_{g_{\text{dest}}}\cup\{j\}\right)\cap\mathcal{A},\ \left(b_{g_{\text{dest}}}\cup\{j\}\right)\cap\mathcal{A}} = \frac{p}{p^*}\left(\frac{\operatorname{Diag}\left(\left|w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\right|^{p-2},\ t^p\left|w_j\right|^{p-2}\right)}{\left\|\left(w_{b_{g_{\text{dest}}}},\ tw_j\right)\right\|_p^{p-1}} - \frac{\begin{bmatrix}\left|w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\right|^{p-2}\odot w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\\ t^p\left|w_j\right|^{p-2}w_j\end{bmatrix}\cdot\begin{bmatrix}\left|w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\right|^{p-2}\odot w_{b_{g_{\text{dest}}}\cap\mathcal{A}}\\ t^p\left|w_j\right|^{p-2}w_j\end{bmatrix}^\top}{\left\|\left(w_{b_{g_{\text{dest}}}},\ tw_j\right)\right\|_p^{2p-1}}\right),$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{dest}}}},\ tw_j\right)\right\|_p\right)_{\mathcal{A}\setminus\left(b_{g_{\text{dest}}}\cup\{j\}\right),\ \mathcal{A}} = \mathbf{0},$$

$$\left(\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\|\left(w_{b_{g_{\text{dest}}}},\ tw_j\right)\right\|_p\right)_{\mathcal{A},\ \mathcal{A}\setminus\left(b_{g_{\text{dest}}}\cup\{j\}\right)} = \mathbf{0}.$$

Because $\|\Xi_\delta(t)\| \to 0$ by (37) and the forcing average is bounded in a neighborhood of $(w^\star(t), t)$ by continuity on a compact neighborhood, the right-hand side of (40) tends to zero. Therefore,

$$\lim_{\delta\to 0}\frac{w^\star(t+\delta)-w^\star(t)}{\delta} = -\left(\nabla_w S^\top\left(w^\star(t),t\right)\right)^{-1}\partial_t S^\top\left(w^\star(t),t\right).$$

Since $w^\star_{\mathcal{A}}(\cdot)$ is differentiable on $[0,1]$ and its derivative is integrable, we have $w^\star_{\mathcal{A}}(1) - w^\star_{\mathcal{A}}(0) = \int_0^1 \frac{d}{dt} w^\star_{\mathcal{A}}(t)dt$. By construction of the MOVE_FEATURE homotopy, $t=0$ corresponds to $\mathcal{G}$ and $t=1$ corresponds to $\mathcal{G}'$, so we have $w^\star_{\mathcal{A}}(0) = w^\star_{\mathcal{G},\mathcal{A}}$ and $w^\star_{\mathcal{A}}(1) = w^\star_{\mathcal{G}',\mathcal{A}}$. Evaluating at $w = w^\star(t)$ and using the definitions of $r_{\text{src}}(t)$ and $r_{\text{dest}}(t)$, we obtain

$$
\begin{aligned}
w^\star_{\mathcal{G}',\mathcal{A}} &= w^\star_{\mathcal{G},\mathcal{A}} + \int_0^1 \frac{d}{dt} w^\star_{\mathcal{A}}(t)dt \\
&= w^\star_{\mathcal{G},\mathcal{A}} - \int_0^1 \left(\nabla_{w_{\mathcal{A}}} S^\intercal\left(w^\star(t),t\right)\right)^{-1} \partial_t S^\intercal\left(w^\star(t),t\right) dt \qquad (41)\\
&= w^\star_{\mathcal{G},\mathcal{A}} - \int_0^1 \left(\nabla_{w_{\mathcal{A}}} S^\intercal\left(w^\star(t),t\right)\right)^{-1} \left(\lambda c_{g_{\text{src}}} \frac{\partial}{\partial t} \nabla_{w_{\mathcal{A}}} r_{\text{src}}(t) + \lambda c_{g_{\text{dest}}} \frac{\partial}{\partial t} \nabla_{w_{\mathcal{A}}} r_{\text{dest}}(t)\right) dt.
\end{aligned}
$$

Note that $\partial_t S^\intercal$ only touches the two MOVE terms. Holding w fixed and differentiating in t, we obtain the precise block expressions

$$
\begin{aligned}
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{src}}(t)\right)_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}} &= \frac{p}{p^*}(1 \\
&\quad - t)^{p-1} \frac{\left|w^\star_j(t)\right|^p}{\left(r_{\text{src}}(t)\right)^{2p-1}} \left(\left|w^\star_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}(t)\right|^{p-2} \odot w^\star_{(b_{g_{\text{src}}}\setminus\{j\})\cap\mathcal{A}}(t)\right), \\
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{src}}(t)\right)_j &= -p(1-t)^{p-1} \frac{\left|w^\star_j(t)\right|^{p-2} w^\star_j(t)}{\left(r_{\text{src}}(t)\right)^{p-1}} + \frac{p}{p^*}(1 \\
&\quad - t)^{p-1} \frac{\left|w^\star_j(t)\right|^p}{\left(r_{\text{src}}(t)\right)^{2p-1}} \cdot \left((1-t)^p \left|w^\star_j(t)\right|^{p-2} w^\star_j(t)\right), \\
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{src}}(t)\right)_{\mathcal{A}\setminus b_{g_{\text{src}}}} &= \mathbf{0},
\end{aligned}
$$

and

$$
\begin{aligned}
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{dest}}(t)\right)_{b_{g_{\text{dest}}}\cap\mathcal{A}} &= -\frac{p}{p^*} t^{p-1} \frac{\left|w^\star_j(t)\right|^p}{\left(r_{\text{dest}}(t)\right)^{2p-1}} \left(\left|w^\star_{b_{g_{\text{dest}}}\cap\mathcal{A}}(t)\right|^{p-2} \odot w^\star_{b_{g_{\text{dest}}}\cap\mathcal{A}}(t)\right), \\
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{dest}}(t)\right)_j &= pt^{p-1} \frac{\left|w^\star_j(t)\right|^{p-2} w^\star_j(t)}{\left(r_{\text{dest}}(t)\right)^{p-1}} \\
&\quad - \frac{p}{p^*} t^{p-1} \frac{\left|w^\star_j(t)\right|^p}{\left(r_{\text{dest}}(t)\right)^{2p-1}} \cdot \left(t^p \left|w^\star_j(t)\right|^{p-2} w^\star_j(t)\right), \\
\left(\frac{\partial}{\partial t}\nabla_{w_{\mathcal{A}}} r_{\text{dest}}(t)\right)_{\mathcal{A}\setminus\left(b_{g_{\text{dest}}}\cup\{j\}\right)} &= \mathbf{0},
\end{aligned}
$$

which is exactly the claimed dynamic system (4).

Remark F.1 (Case $g_{\text{dest}} = 0$)**.** *If* MOVE_FEATURE$(j,0)$ *is applied, we could delete every destination term in $S^\intercal(w,t)$. The secant identity (35), as well as the Neumanna and resolvent limit (36), (39) remain unchanged.*

□

G Proof of Theorem 3.8

Proof. Primitive 2 ADJUST_COEFF(g, c_{new}) changes only the coefficient of the selected group g. We encode the primitive as a deterministic coefficient trajectory $c_{g'}(t)$ on $t \in [0,1]$ with endpoints $c_g(0) = c_g$ and $c_g(1) = c_{\text{new}}$, while $c_{g'}(t) = c_{g'}$ for all $g' \neq g$. For the group g, we enforce the intermediate coefficients $c_g(t)$ to satisfy the constant-velocity difference identity

$$
\forall t \in [0,1],\ \forall h \text{ with } t, t+h \in [0,1]: \quad \frac{c_g(t+h) - c_g(t)}{h} \equiv c_g - c_{\text{new}}.
$$

Following the analysis in Appendix F, for each $t \in [0,1]$ we define $w^\star(t)$ to be the unique minimizer of

$$\min_{w: w_{\mathcal{A}} = \mathbf{0}} f(\mathcal{D}; w) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \left\| w_{b_{g'}} \right\|_p . \tag{42}$$

For any $i \in b_{g'}$, we compute

$$\begin{aligned}
\frac{\partial}{\partial w_i} \left\| w_{b_{g'}} \right\|_p &= \frac{1}{p} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{1}{p}-1} \cdot \frac{\partial}{\partial w_i} \left(\sum_{k \in b_{g'}} |w_k|^p \right) \\
&= \frac{1}{p} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{1}{p}-1} \cdot p \, |w_i|^{p-2} w_i \\
&= \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} |w_i|^{p-2} w_i .
\end{aligned}$$

If $i \notin b_{g'}$, the derivative vanishes. Hence for each $i \in \mathcal{A}$ we have

$$\left[\nabla_{w_{\mathcal{A}}} \left\| w_{b_{g'}} \right\|_p \right]_i = \begin{cases} \dfrac{|w_i|^{p-2} w_i}{\left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{p-1}{p}}}, & i \in b_{g'}, \\ 0, & i \notin b_{g'}. \end{cases} \tag{43}$$

For $j \in b_{g'}$, we first note the identity valid for $w_i \neq 0$:

$$\begin{aligned}
\frac{\partial}{\partial w_j} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} &= -\frac{p-1}{p} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}-1} \cdot \frac{\partial}{\partial w_j} \left(\sum_{k \in b_{g'}} |w_k|^p \right) \\
&= -\frac{p}{p^*} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{2p-1}{p}} |w_j|^{p-2} w_j .
\end{aligned} \tag{44}$$

For $i = j \in b_{g'}$, we obtain

$$\begin{aligned}
\frac{\partial^2}{\partial w_i^2} \left\| w_{b_{g'}} \right\|_p &= \frac{\partial}{\partial w_i} \left[\left(|w_i|^{p-2} w_i \right) \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} \right] \\
&= \left(\frac{\partial}{\partial w_i} |w_i|^{p-2} w_i \right) \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} + \left(|w_i|^{p-2} w_i \right) \frac{\partial}{\partial w_i} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} \\
&\overset{(44)}{=} \frac{p}{p^*} \left[\frac{|w_i|^{p-2}}{\left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{p-1}{p}}} - \frac{|w_i|^{2p-2}}{\left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{2p-1}{p}}} \right] .
\end{aligned} \tag{45}$$

Similarly, for $i \neq j$ with $i, j \in b_{g'}$ we have

$$\begin{aligned}
\frac{\partial^2}{\partial w_j \partial w_i} \left\| w_{b_{g'}} \right\|_p &= \frac{\partial}{\partial w_j} \left[\left(|w_i|^{p-2} w_i \right) \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{p-1}{p}} \right] \\
&= \left(|w_i|^{p-2} w_i \right) \left(-\frac{p}{p^*} \left(\sum_{k \in b_{g'}} |w_k|^p \right)^{-\frac{2p-1}{p}} |w_j|^{p-2} w_j \right) \\
&= -\frac{p}{p^*} \frac{\left(|w_i|^{p-2} w_i \right) \left(|w_j|^{p-2} w_j \right)}{\left(\sum_{k \in b_{g'}} |w_k|^p \right)^{\frac{2p-1}{p}}}.
\end{aligned} \tag{46}$$

Since $w^\star(t)$ minimizes (42) on $\{w : w_{\overline{\mathcal{A}}} = \mathbf{0}\}$, for each $i \in \mathcal{A}$ the optimality condition reads

$$0 \in \frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \begin{cases} \dfrac{|w_i^\star(t)|^{p-2} w_i^\star(t)}{\left(\sum_{k \in b_{g'}} |w_k^\star(t)|^p \right)^{\frac{p-1}{p}}}, & i \in b_{g'}, \\ 0, & i \notin b_{g'}. \end{cases} \tag{47}$$

Likewise, at $t + h$ we have

$$0 \in \frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t+h)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t+h) \begin{cases} \dfrac{|w_i^\star(t+h)|^{p-2} w_i^\star(t+h)}{\left(\sum_{k \in b_{g'}} |w_k^\star(t+h)|^p \right)^{\frac{p-1}{p}}}, & i \in b_{g'}, \\ 0, & i \notin b_{g'}. \end{cases} \tag{48}$$

Subtracting (47) from (48) and using $c_{g'}(t+h) = c_{g'}(t)$ for $g' \neq g$, we obtain

$$\begin{aligned}
0 = &\left(\frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t+h)\right) - \frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t)\right) \right) \\
&+ \lambda \sum_{\substack{g'=1 \\ g' \neq g}}^{G} c_{g'}(t) \left(\begin{cases} \dfrac{|w_i^\star(t+h)|^{p-2} w_i^\star(t+h)}{\left(\sum_{k \in b_{g'}} |w_k^\star(t+h)|^p \right)^{\frac{p-1}{p}}}, & i \in b_{g'}, \\ 0, & i \notin b_{g'}, \end{cases} \right. \\
&\left. - \begin{cases} \dfrac{|w_i^\star(t)|^{p-2} w_i^\star(t)}{\left(\sum_{k \in b_{g'}} |w_k^\star(t)|^p \right)^{\frac{p-1}{p}}}, & i \in b_{g'}, \\ 0, & i \notin b_{g'}, \end{cases} \right) \\
&+ \lambda \left(c_g(t+h) - c_g(t) \right) \begin{cases} \dfrac{|w_i^\star(t)|^{p-2} w_i^\star(t)}{\left(\sum_{k \in b_g} |w_k^\star(t)|^p \right)^{\frac{p-1}{p}}}, & i \in b_g, \\ 0, & i \notin b_g, \end{cases} \\
&+ \lambda c_g(t+h) \left(\begin{cases} \dfrac{|w_i^\star(t+h)|^{p-2} w_i^\star(t+h)}{\left(\sum_{k \in b_g} |w_k^\star(t+h)|^p \right)^{\frac{p-1}{p}}}, & i \in b_g, \\ 0, & i \notin b_g, \end{cases} \right. \\
&\left. - \begin{cases} \dfrac{|w_i^\star(t)|^{p-2} w_i^\star(t)}{\left(\sum_{k \in b_g} |w_k^\star(t)|^p \right)^{\frac{p-1}{p}}}, & i \in b_g, \\ 0, & i \notin b_g. \end{cases} \right)
\end{aligned} \tag{49}$$

For each $i \in \mathcal{A}$, applying the multivariate fundamental theorem of calculus along the segment $w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)$ yields

$$\begin{aligned}
&\frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t+h)\right) - \frac{\partial}{\partial w_i} f\left(\mathcal{D}; w^\star(t)\right) \\
&= \sum_{j \in \mathcal{A}} \left(\int_0^1 \frac{\partial^2}{\partial w_i \partial w_j} f\left(\mathcal{D}; w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)\right) ds \right) \left(w_j^\star(t+h) - w_j^\star(t)\right), \qquad (50) \\
\mathbf{0} &= \left(w^\star(t+h) - w^\star(t)\right)_{\bar{\mathcal{A}}}.
\end{aligned}$$

We can expand every group-gradient difference entrywise using the ℓ_p Hessian. For each g' and each $i \in \mathcal{A}$ this gives

$$\left[\nabla_{w_\mathcal{A}} \left\|w_{b'}^\star(t+h)\right\|_p\right]_i - \left[\nabla_{w_\mathcal{A}} \left\|w_{b_{g'}}^\star(t)\right\|_p\right]_i = \sum_{j \in \mathcal{A}} \left(\int_0^1 \left[\, \breve{\breve{A}} \,\right]_{ij} ds \right) \left(w_j^\star(t+h) - w_j^\star(t)\right), \qquad (51)$$

where

$$\breve{\breve{A}}_{ij} = \begin{cases}
\dfrac{p}{p^*} \left[\dfrac{\left|w_i^\star(t) + s\left(w_i^\star(t+h) - w_i^\star(t)\right)\right|^{p-2}}{\left(\sum\limits_{k \in b_{g'}} \left|w_k^\star(t) + s\left(w_k^\star(t+h) - w_k^\star(t)\right)\right|^p \right)^{\frac{p-1}{p}}} \right. \\
\qquad \left. - \dfrac{\left|w_i^\star(t) + s\left(w_i^\star(t+h) - w_i^\star(t)\right)\right|^{2p-2}}{\left(\sum\limits_{k \in b_{g'}} \left|w_k^\star(t) + s\left(w_k^\star(t+h) - w_k^\star(t)\right)\right|^p \right)^{\frac{2p-1}{p}}} \right], \\
\qquad i = j,\ i \in b_{g'}, \\
\left(\left|w_i^\star(t) + s\left(w_i^\star(t+h) - w_i^\star(t)\right)\right|^{p-2} \left(w_i^\star(t) + s\left(w_i^\star(t+h) - w_i^\star(t)\right)\right)\right) \\
- \dfrac{p}{p^*} \dfrac{\left(\left|w_j^\star(t) + s\left(w_j^\star(t+h) - w_j^\star(t)\right)\right|^{p-2} \left(w_j^\star(t) + s\left(w_j^\star(t+h) - w_j^\star(t)\right)\right)\right)}{\left(\sum\limits_{k \in b_{g'}} \left|w_k^\star(t) + s\left(w_k^\star(t+h) - w_k^\star(t)\right)\right|^p \right)^{\frac{2p-1}{p}}}, \\
\qquad i \neq j,\ i, j \in b_{g'}, \\
0, \quad \text{otherwise.}
\end{cases}$$

We now substitute the Taylor expansions into the difference equation to obtain the difference-quotient linear system. Plugging (50) and (51) into (49) and dividing by $h \neq 0$, we get for each $i \in \mathcal{A}$,

$$\begin{aligned}
0 &= \sum_{j \in \mathcal{A}} \left(\int_0^1 \frac{\partial^2}{\partial w_i \partial w_j} f\left(\mathcal{D}; w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)\right) ds \right) \frac{w_j^\star(t+h) - w_j^\star(t)}{h} \\
&+ \lambda \sum_{\substack{g'=1 \\ g' \neq g}}^{G} c_{g'}(t) \sum_{j \in \mathcal{A}} \left(\int_0^1 \left[\, \breve{\breve{A}} \,\right]_{ij} ds \right) \frac{w_j^\star(t+h) - w_j^\star(t)}{h} + \lambda \frac{c_g(t+h) - c_g(t)}{h} \left[\nabla_{w_\mathcal{A}} \|w_{b_g}^\star(t)\|_p\right]_i \\
&+ \lambda c_g(t+h) \sum_{j \in \mathcal{A}} \left(\int_0^1 \left[\nabla^2_{w_\mathcal{A} w_\mathcal{A}} \left\|\left(w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)\right)_{b_g}\right\|_p\right]_{ij} ds \right) \frac{w_j^\star(t+h) - w_j^\star(t)}{h}.
\end{aligned}$$

Invoking the Primitive 2 identity $-\dfrac{c_g(t+h) - c_g(t)}{h} = c_{\text{new}} - c_g$ and rearranging, we obtain for each $i \in \mathcal{A}$,

$$\begin{aligned}
&\lambda\left(c_{\text{new}} - c_g\right) \begin{cases} \dfrac{\left|w_i^\star(t)\right|^{p-2} w_i^\star(t)}{\left(\sum_{k \in b_g} \left|w_k^\star(t)\right|^p\right)^{\frac{p-1}{p}}}, & i \in b_g, \\ 0, & i \notin b_g, \end{cases} \\
&= \lambda c_g(t+h) \sum_{j \in \mathcal{A}}
\end{aligned}$$

$$\left(\int_0^1 \left[\nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \left\|(w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right))_{b_g}\right\|_p\right]_{ij} ds\right) \frac{w_j^\star(t+h)-w_j^\star(t)}{h}$$
$$+\lambda \sum_{g'=1 g'\neq g}^{G} c_{g'}(t) \sum_{j\in\mathcal{A}} \left(\int_0^1 \left[\breve{A}\right]_{ij} ds\right) \cdot \frac{w_j^\star(t+h)-w_j^\star(t)}{h}$$
$$+\sum_{j\in\mathcal{A}} \left(\int_0^1 \frac{\partial^2}{\partial w_i \partial w_j} f\left(\mathcal{D}; w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right)\right) ds\right) \frac{w_j^\star(t+h)-w_j^\star(t)}{h}.$$

For $|h|$ sufficiently small, this implies the vector identity

$$\frac{w_{\mathcal{A}}^\star(t+h)-w_{\mathcal{A}}^\star(t)}{h}$$
$$= \lambda\left(c_{\text{new}} - c_g\right)\left[\int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right)\right) ds + \lambda \sum_{\substack{g'=1\\ g'\neq g}}^{G} c_{g'}(t) \int_0^1 \breve{\breve{A}} ds\right.$$
$$\left.+\lambda c_g(t+h) \int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \left\|(w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right))_{b_g}\right\|_p ds\right]^{-1} \nabla_{w_{\mathcal{A}}} \|w^\star_{b_g}(t)\|_p.$$

Note that for all $t\in[0,1]$ we have $w_i^\star(t)\neq 0$ for all $i\in\mathcal{A}$ and $w_i^\star(t)=0$ for all $i\in\bar{\mathcal{A}}$. By compactness of $[0,1]$, there exists $\varepsilon_+>0$ such that $\min_{t\in[0,1]}\min_{i\in\mathcal{A}}|w_i^\star(t)|\geq\varepsilon_+$, and therefore every denominator in (43), (45), and (46) stays bounded away from zero. We next use the exact resolvent identity, i.e., for any two invertible matrices U and V, we have $U^{-1}-V^{-1}=U^{-1}(V-U)V^{-1}$. For any fixed t and small h, this yields

$$\left[\int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right)\right) ds\right.$$
$$+\lambda c_g(t+h) \int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \left\|(w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right))_{b_g}\right\|_p ds$$
$$\left.+\lambda \sum_{\substack{g'=1\\ g'\neq g}}^{G} c_{g'}(t) \int_0^1 \breve{\breve{A}} ds\right]^{-1}$$
$$-\left[\nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \|w^\star_{b_{g'}}(t)\|_p\right]^{-1}$$
$$= \left[\int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right)\right) ds + \lambda \sum_{\substack{g'=1\\ g'\neq g}}^{G} c_{g'}(t) \int_0^1 \breve{\breve{A}} ds + \lambda c_g(t+h)\right.$$
$$\left.\cdot \int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \left\|(w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right))_{b_g}\right\|_p ds\right]^{-1}$$
$$\left(\nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \|w^\star_{b_{g'}}(t)\|_p\right.$$
$$-\left[\int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right)\right) ds\right.$$
$$\left.+\lambda c_g(t+h) \int_0^1 \nabla^2_{w_{\mathcal{A}}w_{\mathcal{A}}} \left\|(w^\star(t)+s\left(w^\star(t+h)-w^\star(t)\right))_{b_g}\right\|_p ds\right.$$

$$+\lambda \sum_{\substack{g'=1 \\ g' \neq g}}^{G} c_{g'}(t) \int_0^1 \check{\check{A}} ds \Bigg] \Bigg) \cdot \left[\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \|w^\star_{b_{g'}}(t)\|_p \right]^{-1}. \tag{52}$$

By continuity of (45) and of the explicit ℓ_p Hessian entries (46), the middle difference in (52) converges to zero entrywise as $h \rightsquigarrow 0$. Together with Assumption 2.1 and compactness, both inverse factors remain bounded in operator norm, hence the entire left side tends to $\mathbf{0}$. Letting $h \rightsquigarrow 0$, we therefore obtain

$$\begin{aligned}
&\Bigg[\int_0^1 \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)\right) ds + \lambda \sum_{\substack{g'=1 \\ g' \neq g}}^{G} c_{g'}(t) \int_0^1 \check{\check{A}} ds \\
&\qquad + \lambda c_g(t+h) \int_0^1 \nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} \left\| \left(w^\star(t) + s\left(w^\star(t+h) - w^\star(t)\right)\right)_{b_g} \right\|_p ds \Bigg]^{-1} \\
&\rightsquigarrow \left[\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f\left(\mathcal{D}; w^\star(t)\right) + \lambda \sum_{g'=1}^{G} c_{g'}(t) \nabla^2_{\mathcal{A}\mathcal{A}} \|w^\star_{b_{g'}}(t)\|_p \right]^{-1}.
\end{aligned}$$

Following the analysis in (41) of Appendix F, for each $i \in \mathcal{A}$ we conclude that

$$\begin{aligned}
w^\star_{\mathcal{G}',i} &= w^\star_{\mathcal{G},i} + \int_0^1 \frac{d}{dt} w^\star_i(t) dt \\
&= w^\star_{\mathcal{G},i} - \lambda\left(c_{\text{new}} - c_g\right) \int_0^1 \sum_{j \in \mathcal{A}} \left(\left[\nabla^2_{w_{\mathcal{A}} w_{\mathcal{A}}} f(\mathcal{D}; w)|_{w=w^\star} + \lambda \sum_{g'=1}^{G} c_{g'}(t) \nabla^2 \|w^\star_{b'}(t)\|_p \right]^{-1} \right)_{ij} \\
&\quad \cdot \begin{cases} \dfrac{|w^\star_j(t)|^{p-2} w^\star_j(t)}{\left(\sum_{k \in b_g} |w^\star_k(t)|^p\right)^{\frac{p-1}{p}}}, & j \in b_g, \\ 0, & j \notin b_g, \end{cases} dt.
\end{aligned} \tag{53}$$

Reorganizing (53) yields (5). □

H Proof of Theorem 3.10

Proof. Write $S^{(k)} = (\mathcal{G}^{(k)}, c^{(k)})$ with $\mathcal{G}^{(k)} = \{b_1^{(k)}, \ldots, b_G^{(k)}\}$ and $c^{(k)} = (c_1^{(k)}, \ldots, c_G^{(k)})$, noting that the number of penalized groups remains G throughout since only Primitives 1-2 are used. For notational uniformity, set $b_0^{(k)} := \left(\mathcal{G}^{(k)}\right)^{\complement}$ for every k. Hence $\{b_0^{(k)}, b_1^{(k)}, \ldots, b_G^{(k)}\}$ forms a partition of $[d]$ for every k, and in particular for every k and every $i \in [d]$ we have the identity

$$\sum_{h=0}^{G} \mathbf{1}\{i \in b_h^{(k)}\} = 1, \quad \mathbf{1}\{i \in b_h^{(k)}\}\, \mathbf{1}\{i \in b_{h'}^{(k)}\} = 0 \ \text{ for } h \neq h'.$$

We first establish a *local commutation* identity at an arbitrary intermediate state. Fix any admissible state $\bar{S} = (\bar{\mathcal{G}}, \bar{c})$ with $\bar{\mathcal{G}} = \{\bar{b}_1, \ldots, \bar{b}_G\}$ and define $\bar{b}_0 := \bar{\mathcal{G}}^{\complement}$. Fix any instantiations $\mathcal{M} := \texttt{MOVE_FEATURE}(j, g_{\text{dest}})$ and $\mathcal{Z} := \texttt{ADJUST_COEFF}(g, c_{\text{new}})$ with $j \in [d]$, $g_{\text{dest}} \in \{0, 1, \ldots, G\}$, $g \in \{1, \ldots, G\}$, and $c_{\text{new}} > 0$. Let $g_{\text{src}} \in \{0, 1, \ldots, G\}$ be the unique index satisfying $j \in \bar{b}_{g_{\text{src}}}$, where uniqueness follows from the partition property of $\{\bar{b}_h\}_{h=0}^{G}$. Define

$$\bar{S}^{\text{MA}} := \mathcal{Z}\big(\mathcal{M}(\bar{S})\big), \quad \bar{S}^{\text{AM}} := \mathcal{M}\big(\mathcal{Z}(\bar{S})\big),$$

and write

$$\bar{S}^{\text{MA}} = (\bar{\mathcal{G}}^{\text{MA}}, \bar{c}^{\text{MA}}), \qquad \bar{S}^{\text{AM}} = (\bar{\mathcal{G}}^{\text{AM}}, \bar{c}^{\text{AM}}),$$

with $\bar{\mathcal{G}}^{\mathrm{MA}} = \{\bar{b}_1^{\mathrm{MA}}, \ldots, \bar{b}_G^{\mathrm{MA}}\}$ and similarly for $\bar{\mathcal{G}}^{\mathrm{AM}}$. We also define $\bar{b}_0^{\mathrm{MA}} := (\bar{\mathcal{G}}^{\mathrm{MA}})^{\complement}$ and $\bar{b}_0^{\mathrm{AM}} := (\bar{\mathcal{G}}^{\mathrm{AM}})^{\complement}$. By Primitive 1, the move $\mathcal{M}$ changes only the membership of the single index j while leaving all other indices fixed. This admits the exact indicator-level representation, i.e., for every $h \in \{0, 1, \ldots, G\}$ and every $i \in [d]$,

$$\mathbf{1}\{i \in \bar{b}_h^{\mathcal{M}}\} = \mathbf{1}\{i \in \bar{b}_h\} + \Big(\mathbf{1}\{h = g_{\text{dest}}\} - \mathbf{1}\{h = g_{\text{src}}\}\Big)\mathbf{1}\{i = j\}, \tag{54}$$

where $\bar{b}_h^{\mathcal{M}}$ denotes the h-th block after applying $\mathcal{M}$ to $\bar{S}$ and $\bar{b}_0^{\mathcal{M}}$ denotes its complement. Identity (54) is an exact rewriting of the two-step update that removes j from its unique source block and inserts it into the destination block, and it relies only on the fact that g_{src} is uniquely defined. On the other hand, by Primitive 2, the adjust $\mathcal{Z}$ never changes membership sets

$$\bar{b}_h^{\mathcal{Z}} = \bar{b}_h, \quad \forall h \in \{0, 1, \ldots, G\}. \tag{55}$$

In particular, the source block index g_{src} computed from membership of j is invariant under $\mathcal{Z}$, namely $j \in \bar{b}_{g_{\text{src}}} \Leftrightarrow j \in \bar{b}_{g_{\text{src}}}^{\mathcal{Z}}$. Consequently, the same g_{src} is used by $\mathcal{M}$ regardless of the order of composition. We now compute membership indicators for the two compositions. Since $\mathcal{Z}$ does not change membership, we have $\mathbf{1}\{i \in \bar{b}_h^{\mathrm{MA}}\} = \mathbf{1}\{i \in \bar{b}_h^{\mathcal{M}}\}$. Using (54), for all h and i,

$$\mathbf{1}\{i \in \bar{b}_h^{\mathrm{MA}}\} = \mathbf{1}\{i \in \bar{b}_h\} + \Big(\mathbf{1}\{h = g_{\text{dest}}\} - \mathbf{1}\{h = g_{\text{src}}\}\Big)\mathbf{1}\{i = j\}. \tag{56}$$

For the reverse order, $\mathcal{Z}$ leaves $\bar{b}_h$ unchanged by (55), and $\mathcal{M}$ uses the same g_{src}. Hence applying (54) with $\bar{b}_h^{\mathcal{Z}} = \bar{b}_h$ yields, for all h and i,

$$\mathbf{1}\{i \in \bar{b}_h^{\mathrm{AM}}\} = \mathbf{1}\{i \in \bar{b}_h\} + \Big(\mathbf{1}\{h = g_{\text{dest}}\} - \mathbf{1}\{h = g_{\text{src}}\}\Big)\mathbf{1}\{i = j\}. \tag{57}$$

Comparing (56) and (57) gives the pointwise identity

$$\mathbf{1}\{i \in \bar{b}_h^{\mathrm{MA}}\} = \mathbf{1}\{i \in \bar{b}_h^{\mathrm{AM}}\} \quad \forall i \in [d],\ \forall h \in \{0, 1, \ldots, G\}. \tag{58}$$

Since two subsets of $[d]$ are equal if and only if their membership indicators coincide at every index, (58) implies

$$\bar{b}_h^{\mathrm{MA}} = \bar{b}_h^{\mathrm{AM}} \quad \forall h \in \{0, 1, \ldots, G\}. \tag{59}$$

The induced equalities on group sizes and block norms follow as explicit consequences of (58). Summing over i yields for every $h \in \{0, 1, \ldots, G\}$,

$$\begin{aligned}
|\bar{b}_h^{\mathrm{MA}}| &= \sum_{i=1}^{d} \mathbf{1}\{i \in \bar{b}_h^{\mathrm{MA}}\} = \sum_{i=1}^{d} \mathbf{1}\{i \in \bar{b}_h^{\mathrm{AM}}\} = |\bar{b}_h^{\mathrm{AM}}|, \\
\|w_{\bar{b}_h^{\mathrm{MA}}}\|_p^p &= \sum_{i=1}^{d} |w_i|^p \mathbf{1}\{i \in \bar{b}_h^{\mathrm{MA}}\} = \sum_{i=1}^{d} |w_i|^p \mathbf{1}\{i \in \bar{b}_h^{\mathrm{AM}}\} = \|w_{\bar{b}_h^{\mathrm{AM}}}\|_p^p, \\
\|w_{\bar{b}_h^{\mathrm{MA}}}\|_p &= \Big(\|w_{\bar{b}_h^{\mathrm{MA}}}\|_p^p\Big)^{1/p} = \Big(\|w_{\bar{b}_h^{\mathrm{AM}}}\|_p^p\Big)^{1/p} = \|w_{\bar{b}_h^{\mathrm{AM}}}\|_p.
\end{aligned} \tag{60}$$

By Primitive 1, $\mathcal{M}$ never modifies coefficients, so $\bar{c}_h^{\mathcal{M}} = \bar{c}_h$ for all $h \in \{1, \ldots, G\}$. By Primitive 2, $\mathcal{Z}$ modifies only the g-th coefficient via $\bar{c}_h^{\mathcal{Z}} = \bar{c}_h + (c_{\text{new}} - \bar{c}_h)\mathbf{1}\{h = g\}$ for all $h \in \{1, \ldots, G\}$. Therefore, composing in the two orders gives for every $h \in \{1, \ldots, G\}$,

$$\bar{c}_h^{\mathrm{MA}} = \bar{c}_h^{\mathcal{Z} \circ \mathcal{M}} = \bar{c}_h^{\mathcal{M}} + (c_{\text{new}} - \bar{c}_h^{\mathcal{M}})\mathbf{1}\{h = g\} = \bar{c}_h + (c_{\text{new}} - \bar{c}_h)\mathbf{1}\{h = g\}, \tag{61}$$

$$\bar{c}_h^{\mathrm{AM}} = \bar{c}_h^{\mathcal{M} \circ \mathcal{Z}} = \bar{c}_h^{\mathcal{Z}} = \bar{c}_h + (c_{\text{new}} - \bar{c}_h)\mathbf{1}\{h = g\}. \tag{62}$$

Comparing (61) and (62) yields

$$\bar{c}^{\mathrm{MA}} = \bar{c}^{\mathrm{AM}}. \tag{63}$$

Equations (59) and (63) together imply that the two composed states coincide:

$$\mathcal{Z}\big(\mathcal{M}(\bar{S})\big) = \mathcal{M}\big(\mathcal{Z}(\bar{S})\big). \tag{64}$$

This equality represents *state-level* commutativity in the sense that it matches the semantics of primitives with group identifiers and exact membership sets, rather than merely equality of objective

values. For completeness, we also verify that (64) implies equality of the full objective functional for all $w \in \mathbb{R}^d$. Using (59), (63), and the definition $F_S(w) \triangleq f(\mathcal{D}; w) + \lambda \sum_{h=1}^{G} \bar{c}_h \, \|w_{\bar{b}_h}\|_p$ with the appropriate $d_h^{1-1/p}$ factors absorbed into $\bar{c}_h$ if desired, we obtain

$$
\begin{aligned}
& F_{\bar{S}^{\mathrm{MA}}}(w) - F_{\bar{S}^{\mathrm{AM}}}(w) \\
&= \lambda \sum_{h=1}^{G} \bar{c}_h^{\mathrm{MA}} \|w_{\bar{b}_h^{\mathrm{MA}}}\|_p \qquad - \lambda \sum_{h=1}^{G} \bar{c}_h^{\mathrm{AM}} \|w_{\bar{b}_h^{\mathrm{AM}}}\|_p \\
&= \lambda \sum_{h=1}^{G} \Big(\bar{c}_h^{\mathrm{MA}} - \bar{c}_h^{\mathrm{AM}} \Big) \|w_{\bar{b}_h^{\mathrm{MA}}}\|_p + \lambda \sum_{h=1}^{G} \bar{c}_h^{\mathrm{AM}} \Big(\|w_{\bar{b}_h^{\mathrm{MA}}}\|_p - \|w_{\bar{b}_h^{\mathrm{AM}}}\|_p \Big) \\
&= \lambda \sum_{h=1}^{G} \underbrace{\Big(\bar{c}_h + (c_{\mathrm{new}} - \bar{c}_h)\mathbf{1}\{h = g\} - \bar{c}_h - (c_{\mathrm{new}} - \bar{c}_h)\mathbf{1}\{h = g\} \Big)}_{=0} \|w_{\bar{b}_h^{\mathrm{MA}}}\|_p \\
&\quad + \lambda \sum_{h=1}^{G} \bar{c}_h^{\mathrm{AM}} \underbrace{\Big(\|w_{\bar{b}_h^{\mathrm{MA}}}\|_p - \|w_{\bar{b}_h^{\mathrm{AM}}}\|_p \Big)}_{=0 \text{ by } (60)} = \; 0.
\end{aligned}
$$

We now turn to the setting with the length-K sequence $\mathcal{P}_1, \ldots, \mathcal{P}_K$ and the swapped sequence $\widetilde{\mathcal{P}}_1, \ldots, \widetilde{\mathcal{P}}_K$ at positions $(r, r+1)$. Since the two sequences share the same prefix up to $r-1$, we have

$$
\widetilde{S}^{(r-1)} = \widetilde{\mathcal{P}}_{r-1} \circ \cdots \circ \widetilde{\mathcal{P}}_1(S) = \mathcal{P}_{r-1} \circ \cdots \circ \mathcal{P}_1(S) = S^{(r-1)}.
$$

Write $\bar{S} := S^{(r-1)} = \widetilde{S}^{(r-1)}$. By hypothesis, $\mathcal{P}_r$ and $\mathcal{P}_{r+1}$ consist of one move and one adjust. Thus $\widetilde{\mathcal{P}}_r = \mathcal{P}_{r+1}$ and $\widetilde{\mathcal{P}}_{r+1} = \mathcal{P}_r$ correspond exactly to swapping an instance of $\mathcal{M}$ and $\mathcal{Z}$ at the same intermediate state $\bar{S}$. Applying (64) at $\bar{S}$ therefore gives the key identity

$$
\widetilde{S}^{(r+1)} = \widetilde{\mathcal{P}}_{r+1}\big(\widetilde{\mathcal{P}}_r(\bar{S})\big) = \mathcal{P}_r\big(\mathcal{P}_{r+1}(\bar{S})\big) = \mathcal{P}_{r+1}\big(\mathcal{P}_r(\bar{S})\big) = S^{(r+1)}. \tag{65}
$$

Now observe that the two sequences also share the same suffix from $r+2$ to K, and that composition is associative. Hence,

$$
\begin{aligned}
\widetilde{S}^{(K)} = \widetilde{\mathcal{P}}_K \circ \cdots \circ \widetilde{\mathcal{P}}_{r+2}\big(\widetilde{S}^{(r+1)}\big) &= \mathcal{P}_K \circ \cdots \circ \mathcal{P}_{r+2}\big(\widetilde{S}^{(r+1)}\big) \\
&\overset{(65)}{=} \mathcal{P}_K \circ \cdots \circ \mathcal{P}_{r+2}\big(S^{(r+1)}\big) = S^{(K)},
\end{aligned} \tag{66}
$$

which proves the one-swap claim of the theorem. It remains to prove the invariance under any finite number of adjacent swaps. If a sequence $\widehat{\mathcal{P}}_1, \ldots, \widehat{\mathcal{P}}_K$ is obtained from $\mathcal{P}_1, \ldots, \mathcal{P}_K$ by a finite chain of adjacent swaps of neighboring `MOVE_FEATURE` and `ADJUST_COEFF` pairs, then applying (66) repeatedly along that chain shows that the final state remains unchanged at each swap and hence unchanged overall. This establishes the stated invariance under any finite number of such swaps. In other words, the resulting state depends only on the two internal orders, i.e., the order of `MOVE_FEATURE` calls among themselves and the order of `ADJUST_COEFF` calls among themselves. □

H.1 More Motivating Scenarios

We provide three real-world scenarios in which a structured-sparse model trained under one group partition $\mathcal{G}$ must be adapted to a revised partition $\mathcal{G}'$, illustrating that repartitioning is a recurring rather than exceptional event.

Finance. In high-frequency trading and sector-based portfolio optimization, assets are grouped by industry sectors and regularized via sparse group Lasso to identify a small number of predictive sectors [Chen et al., 2020]. The Global Industry Classification Standard (GICS) is revised periodically: the 2018 revision alone reassigned roughly 10% of the S&P 500 market capitalization across sectors. Each revision alters $\mathcal{G}$ over the same asset universe and forces a reoptimization of the existing portfolio model.

Neuroimaging. In fMRI-based clinical prediction, voxels are grouped by regions of interest (ROIs) defined by a brain atlas, and group Lasso is used to select predictive regions [Eickhoff et al., 2018]. When the analysis transitions between atlases (e.g., AAL-116 to Glasser-360) or when an atlas itself is updated to incorporate new parcellations, the partition over voxels changes and previously trained models must be refit under the new $\mathcal{G}'$.

Genomics. In pathway-guided gene selection, genes are grouped by biological pathways and group-sparsity penalties select predictive pathways for downstream phenotypes [Kanehisa et al., 2025]. Pathway databases such as KEGG are continuously curated, with multiple releases per year that reassign genes, merge pathways, or split them into finer subunits. Each curation step changes $\mathcal{G}$ and requires the model to be re-trained or fine-tuned to remain consistent with the latest annotation.

I Implementation Details

All datasets used in this study are publicly available from the OpenML repository[3] and the UCI Machine Learning Repository[4]. A summary of the dataset characteristics is provided in Table 2. For brevity, we abbreviate ASP-POTASSCO-regression as ASP-POTASSCO.

All experiments are implemented in MATLAB R2021a and executed on a workstation running Windows 10 with 32GB RAM and an Intel Core i7 CPU. No GPU acceleration is employed, ensuring that the reported timing comparisons reflect pure algorithmic efficiency rather than hardware-specific optimizations.

Table 2: Details of datasets used in our experiments. # Inst. denotes the number of instances, and # Feat. denotes the number of features.

Dataset	# Inst.	# Feat.	Dataset	# Inst.	# Feat.	Dataset	# Inst.	# Feat.	Dataset	# Inst.	# Feat.
mtp	4,450	203	isolet	7,797	617	madeline	3,140	260	us_crime	1,994	128
clean2	6,598	169	fabert	8,237	800	topo_2_1	8,885	267	Speech	3,686	401
yprop_4_1	8,885	252	Indian_pines	9,144	220	rf2	9,125	584	philippine	5,832	309
scm1d	9,803	296	USPS	9,298	256	ASP-POTASSCO	14,234	143	har	10,299	562
sylva_agnostic	14,395	217	gas-drift	13,910	129	nomao	34,465	119	volkert	58,310	181

Batch Retraining Baseline. We compare against a full retraining baseline that resolves (1) independently for every group structure $\mathcal{G}$. We use a standard composite first-order method, namely monotone accelerated proximal gradient with backtracking step sizes, following the MFISTA template [Beck and Teboulle, 2009] and the proximal-operator framework in [Parikh et al., 2014]. At each iteration, we perform one forward step using the gradient of the smooth loss $f(\mathcal{D}; w)$ and one backward step using the proximal map of the $\ell_{p,1}$ group regularizer.

The backward step is computed groupwise because the regularizer is block separable across disjoint groups. For $p = 2$ we apply the closed-form group soft-thresholding operator. For integer $p \geq 3$ we compute the group proximal map to high accuracy by solving its first-order optimality conditions using a globally convergent scalar root-finding routine, implemented with safeguarded Newton updates and bisection. Each retraining run is terminated using a first-order KKT residual that aggregates the stationarity conditions on unpenalized coordinates, stationarity equalities on active groups, and dual feasibility inequalities on inactive groups. Unless stated otherwise, all runs are solved to the same KKT tolerance, with default threshold 10^{-8}. Batch retraining is cold-started, except in the explicitly identified warm-start comparisons in Appendix J.

J Additional Results

Following the similar experimental setup in the main body, we provide additional results along with analysis in Table 3, Figure 4 and Figure 5. We further benchmark against a different baseline, block coordinate descent (BCD), in Figure 7 on Nomao and Gas-drift across square and Gamma losses with warm-starting enabled.

[3] https://www.openml.org/search?type=data

[4] https://archive.ics.uci.edu/datasets

Table 3: Testing loss of batch retraining vs. our Algorithm 1 on Poisson generalized regression model.

Dataset	Norm	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
		Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match	Batch	Alg. 1	Match
Base Model: Poisson Generalized Regression																			
us_crime	$p=2$	5.318	5.318	✓	4.672	4.672	✓	6.041	6.041	✓	4.955	4.955	✓	5.889	5.889	✓	4.203	4.203	✓
($\lambda=0.014$)	$p=4$	5.533	5.533	✓	4.861	4.861	✓	6.287	6.287	✓	5.152	5.152	✓	6.126	6.126	✓	4.362	4.362	✓
clean2	$p=2$	8.764	8.764	✓	10.21	10.21	✓	7.593	7.593	✓	9.846	9.846	✓	8.107	8.107	✓	10.79	10.79	✓
($\lambda=0.011$)	$p=4$	9.128	9.128	✓	10.63	10.63	✓	7.918	7.918	✓	10.25	10.25	✓	8.458	8.458	✓	11.20	11.20	✓
Indian_pines	$p=2$	12.46	12.46	✓	9.831	9.831	✓	15.28	15.28	✓	11.08	11.08	✓	10.42	10.42	✓	14.55	14.55	✓
($\lambda=0.010$)	$p=4$	12.98	12.98	✓	10.21	10.21	✓	15.86	15.86	✓	11.51	11.51	✓	10.82	10.82	✓	15.13	15.13	✓
volkert	$p=2$	17.24	17.24	✓	15.88	15.88	✓	18.96	18.96	✓	16.42	16.42	✓	14.73	14.73	✓	19.51	19.51	✓
($\lambda=0.007$)	$p=4$	17.86	17.86	✓	16.45	16.45	✓	19.58	19.58	✓	17.02	17.02	✓	15.29	15.29	✓	19.98	19.98	✓

Why warm-starting fails to accelerate. Unlike classical regularization-path settings where consecutive problems differ by a smooth perturbation of the objective, repartitioning alters the geometry of the nonsmooth $\ell_{p,1}$ penalty itself: the admissible sparsity patterns, the dual-norm constraints, and the active set $\mathcal{A}$ all change discretely. The pretrained iterate $w^\star_{\mathcal{G}}$ may therefore violate the new dual feasibility on multiple groups, lie in a different basin of $\mathbb{F}_{\mathcal{G}'}$, or carry a support pattern incompatible with $\mathcal{G}'$, so that the proximal-gradient solver spends most of its budget reconfiguring the active set rather than refining a nearly-correct solution. Figure 4 reports learning curves of warm-started vs. cold-started batch baselines on yprop_4_1, USPS, madeline, and topo_2_1 under Poisson regression, and Figure 6 compares Algorithm 1 against warm-started batch trainings; in both, warm-starting yields no consistent acceleration. Occasionally a modest speedup, often a match or a slowdown, with large variance across runs, which is why we do not adopt it as the default in the main text and why exploiting $w^\star_{\mathcal{G}}$ requires the structured fine-tuning developed in Section 4 rather than naive iterative reuse.

Threshold sensitivity. We vary the deactivation threshold in logistic group Lasso ($p = 2$), using repartitionings as in Rounds 5–6 of Table 1. We use 20 trials with fixed partitions and seeds across thresholds and a verified deactivation in each trial. Table 4 compares target group patterns against batch retraining and reports mean runtime and mean final KKT residual.

Table 4: Sensitivity to ϵ (✓: 20/20 matches).

ϵ	Match	Time (s)	KKT residual
10^{-14}	✓	18.72	2.3×10^{-10}
10^{-12}	✓	18.69	2.5×10^{-10}
10^{-10}	✓	18.74	3.1×10^{-10}
10^{-8}	✓	18.81	8.9×10^{-9}
10^{-6}	✓	18.95	1.4×10^{-7}
10^{-4}	18/20	18.62	4.2×10^{-5}

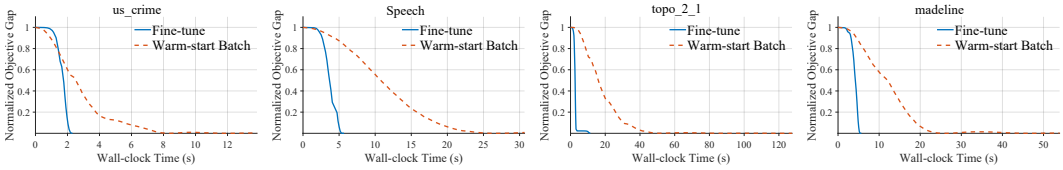

Figure 4: Learning curves during training. Batch retraining is warm-started from the pretrained solution, resulting in identical initial points across methods. Despite this favorable initialization, warm-starting provides negligible acceleration in practice; hence we do not adopt it as the default in the main-text experiments.

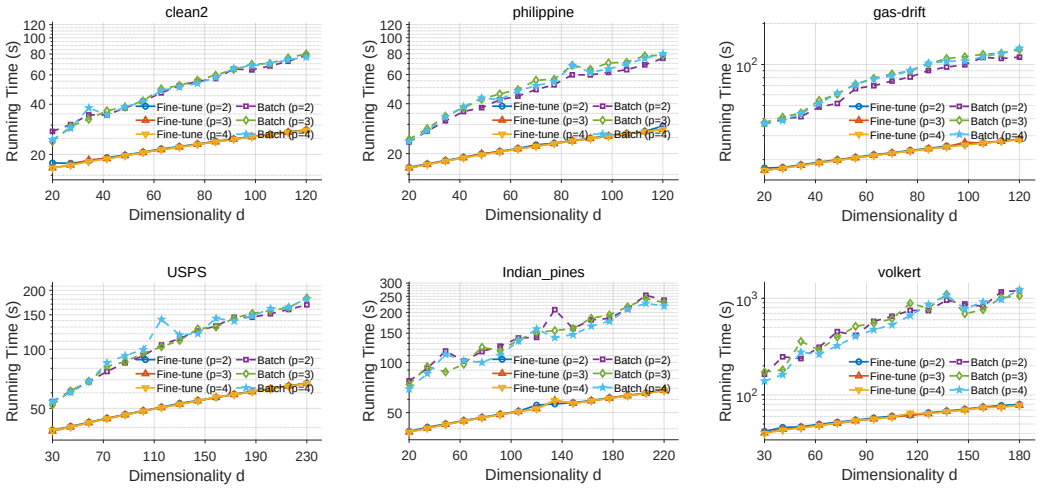

Figure 5: Computational time as a function of feature dimension d across three models: quadratic loss regression (first two), logistic regression (middle two), and Gamma regression (last two). Each model is evaluated on two datasets of varying scales.

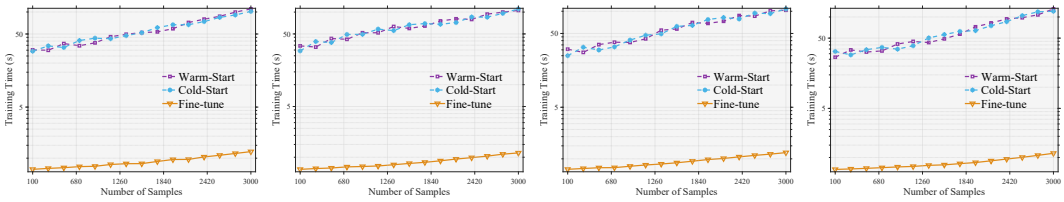

Figure 6: Comparison of our framework against the baseline with warm start and the baseline without warm start. The experimental settings and mechanism follow a similar setup as in Figure 2. Poisson regression with $\lambda = 0.01$ on four datasets: yprop_4_1, USPS, madeline, and topo_2_1. In our extensive simulations, warm-start does **not** provide consistent acceleration under the repartitioning setting, with speedup being inconsistent across runs.

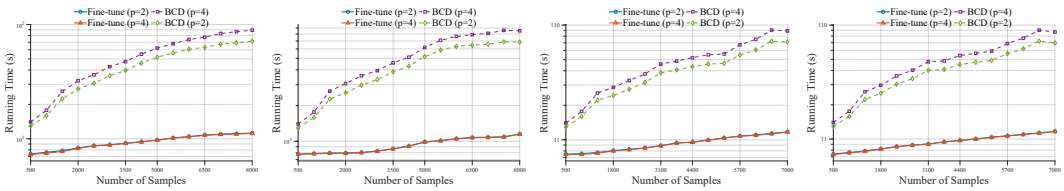

Figure 7: Computational time of our framework vs. Block Coordinate Descent (BCD) baseline. The experimental settings and mechanism follow a similar setup as in Figure 2. First two subfigures are square loss with $\lambda = 0.02$ and $\lambda = 0.03$; last two are Gamma loss with $\lambda = 0.025$ and $\lambda = 0.04$; first two results are on the Nomao dataset and the last two are on the Gas-drift dataset. Note that when $p \neq 2$, each block subproblem is solved numerically to high accuracy. The warm-start strategy is also adapted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims we made are accurate.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are in main body. The verified proofs appear in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The contribution is primarily a new algorithm. See Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See our supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 5 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational research and not tied to particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 5 and Appendix I. Licenses are available referring to the provided citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.