COLOSSEUM OF MECHANISMS: A UNIFIED FRAMEWORK FOR UNSUPERVISED MULTI-MECHANISM LEARNING VIA QUASI-GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

An unsupervised multi-mechanism model has $|J|$ mechanisms and $|I|$ observations. Fitting it optimally means an intractable search over $|J|^{|I|}$ assignments. The 'divide-and-conquer' (DaC) approach assumes a segmentation form and searches only the assignments that form expresses: tractable, but broken when the assumption is incorrect. We propose the Colosseum of Mechanisms (CoM), a 'conquer-then-divide' framework: it fits the parameters first, then reads off the segmentation. At every observation, CoM stages a competition among all $|J|$ mechanisms, then aggregates only the winners. CoM assumes no segmentation form and therefore accommodates any. Optimizing the CoM objective ℓ solves the combinatorial search without traversing those assignments. An algebraic identity makes this possible. This identity assumes nothing about the noise, nor the correctness of the supplied mechanism forms. But the combinatorial difficulty does not disappear; it reappears as nonconvexity and nonsmoothness in ℓ. Our *Quasi*-Gradient Descent algorithm guarantees ε-Clarke stationarity in $\mathcal{O}(\varepsilon^{-2})$ iterations, exactly what standard gradient descent achieves on a single smooth function. Even the simplest CoM case is NP-hard to minimize. We therefore give a necessary condition for global optimality. In DaC's best case, CoM matches its accuracy about $1000\times$ faster. In a non-ideal DaC case, CoM's accuracy and speed are unchanged, while DaC loses eleven accuracy points and runs slower still. On U.S. GNP data, CoM halves the out-of-sample forecast error of a published benchmark. CoM also reconstructs entangled high-frequency dynamics in EEG.

1 INTRODUCTION

Applying one monolithic model to a whole dataset is the default in statistics and machine learning, yet it oversimplifies the behavior of many real systems. A multi-mechanism model instead lets different mechanisms $\{f_j\}_{j\in J}$ govern different regions of the data, indexed by I,

$$\boldsymbol{y}_i = \begin{cases} f_1(\boldsymbol{x}_i, \boldsymbol{\theta}_1, \varepsilon_{1i}), & \text{under condition 1;} \\ \cdots & \\ f_{|J|}(\boldsymbol{x}_i, \boldsymbol{\theta}_2, \varepsilon_{2i}), & \text{under condition } |J|; \end{cases} \qquad i \in I. \tag{1}$$

We rarely observe these conditions in practice. They may be gates choosing among experts (Jacobs et al., 1991), membership in a latent subpopulation (McQueen, 1967), unknown change points in a sequence (Hinkley, 1970), latent states driving a switch (Goldfeld & Quandt, 1973), thresholds on a lagged value (Chan, 1993), or outcomes of an operator (Prentice et al., 1978).

For parametric methods the forms of the f_j are known, or at least assumed. How well a candidate f_j explains $\boldsymbol{y}_i$ is measured by a discrepancy, typically a loss or a regularized loss, which we write as

$$\bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i) := \bar{\ell}_j(\boldsymbol{y}_i, f_j(\boldsymbol{x}_i, \boldsymbol{\theta}_j)). \tag{2}$$

What separates the available routes is whether a segmentation form is assumed, and how. The first route assumes none. Let $\mathcal{P} := \{P = (I_j)_{j\in J} : \sqcup_{j\in J} I_j = I\}$ collect every assignment of I into $|J|$ disjoint, possibly empty blocks, so that $|\mathcal{P}| = |J|^{|I|}$. The oracle problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, P\in\mathcal{P}} R(\boldsymbol{\theta}; P), \qquad R(\boldsymbol{\theta}; P) := \sum\nolimits_{j\in J} \sum\nolimits_{i\in I_j} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i), \tag{3}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

optimizes the region-wise objective jointly over $\boldsymbol{\Theta} \times \mathcal{P}$, where $\boldsymbol{\theta}^\top := (\boldsymbol{\theta}_j^\top)_{j \in J}$ concatenates the mechanism parameters and $\boldsymbol{\Theta} := \prod_{j \in J} \boldsymbol{\Theta}_j$. It is well posed, since for a fixed P the inner problem $\operatorname{argmin}_{\boldsymbol{\theta}} \; R(\boldsymbol{\theta}; P)$ separates across j and is solved blockwise by

$$\boldsymbol{\theta}_j^P := \operatorname{argmin}_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \sum\nolimits_{i \in I_j} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i), \tag{4}$$

leaving only a comparison of $R(\boldsymbol{\theta}^P; P)$ across a finite $\mathcal{P}$. But finiteness is not tractability: $|\mathcal{P}| = |J|^{|I|}$ grows exponentially; a mere 210 observations and 3 mechanisms give $3^{210} > 10^{100}$, more than the estimated number of atoms in the observable universe.

The second route, divide-and-conquer (DaC), makes that search affordable by confining P to a subfamily of assumed segmentation forms, such as a threshold scanned by grid search (Hansen, 2000), a set of change points (Killick et al., 2012), or a dendrogram cut (Loh, 2011), with (4) evaluated after finding the best member of that subfamily. Two costs follow. First, the fit is only as good as the assumed form, and it breaks whenever the true condition has a different shape. Second, the assumption cannot be checked, since the real conditions are never observed in practice; checking it honestly requires traversing all $|J|^{|I|}$ assignments, the search DaC was introduced to avoid.

Our Colosseum of Mechanisms (CoM) introduces the third route: conquer-then-divide, an opposite of DaC. At every i, all $|J|$ candidates are evaluated and only the best fit contributes, so the CoM objective is a sum of pointwise minima,

$$\ell(\boldsymbol{\theta}) := \sum\nolimits_{i \in I} \min\nolimits_{j \in J} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i), \tag{5}$$

in which no segmentation appears. Whatever the true condition is, ℓ is the same function. A rule the objective never names is a rule it cannot get wrong. So it is no surprise that earlier work has arrived at this form under particular assumed conditions: mixtures give K-means in unconventional form (Pollard, 1982); hyperplanes give K-plane regression (Manwani & Sastry, 2015).

Discarding the segmentation costs nothing at the optimum. Theorem 2.1 identifies ℓ with the oracle problem (3) profiled over its $|J|^{|I|}$ assignments: the two share their optimal value and minimizers, so minimizing ℓ alone solves the intractable combinatorial search, and the assignment is read off afterwards from which candidate fits best at $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$. The identity is algebraic, holding at every $\boldsymbol{\theta}$ before any optimization and without even assuming the supplied f_j and $|J|$ are correct; when they are not, $\boldsymbol{\theta}^*$ is still the best fit of the forms supplied over all $|J|^{|I|}$ segmentations (Remark 2.3).

But the combinatorial difficulty is not removed, only redescribed: the pointwise $\min$ introduces *concave kinks*, which destroy convexity and smoothness however well behaved each $\bar{\ell}_j$ is. Section 3 meets it in that form. Writing ∂^C for the Clarke subdifferential (Clarke, 1990), we derive its minimum-norm element in closed form at each observation, sum those pieces into a descent direction, and prove a descent lemma for it. The resulting algorithm, *Quasi*-gradient descent, reaches an ε-Clarke stationary point in $\mathcal{O}(\varepsilon^{-2})$ iterations, at the same step size and with the same constants of vanilla gradient descent on a single smooth objective (Theorem 3.7). The overall computation is linear in $|I|$, against $\mathcal{O}(|I|^2)$ to $\mathcal{O}(|I|^3)$ for the DaC baselines (Remark 3.10). Section 5.1 measures that gap: when the true segmentation form is exactly what these baselines assume, CoM matches their fit about $1000\times$ faster; when it is not, CoM keeps both its accuracy and efficiency, while they lose eleven accuracy points and run slower still.

The $\mathcal{O}(\varepsilon^{-2})$ rate invites comparison with complexity theory for nonsmooth nonconvex optimization, although ε measures a different quantity in each result so rates are not directly comparable. Proximal and model-based methods attain $\mathcal{O}(\varepsilon^{-4})$ from noisy gradients of the Moreau envelope and $\mathcal{O}(\varepsilon^{-2})$ from exact ones (Davis & Drusvyatskiy, 2019; Drusvyatskiy & Paquette, 2019), but require weak convexity, which the concave kinks of ℓ rule out. Methods built on the Goldstein subdifferential, the convex hull of Clarke subdifferentials over a δ-ball, attain $\widetilde{\mathcal{O}}(\delta^{-1}\varepsilon^{-3})$ (Zhang et al., 2020): a slower rate blowing up as $\delta \to 0$, against a criterion weaker than Clarke stationarity at every $\delta > 0$. The stochastic subgradient method (Davis et al., 2020) and gradient sampling (Burke et al., 2005) do apply to ℓ as it stands, but converge with no finite-time rate. Only *Quasi*-GD does all three: applies to ℓ as it stands, matching the fastest rate above, against the strictest criterion.

But stationarity is not optimality, and on a non-convex landscape spurious optima are real: finding the global optimum is NP-hard even for K-means (Dasgupta, 2008), the simplest CoM instance.

Practices like random restarts, careful seeding (Arthur & Vassilvitskii, 2006), or pilot estimates (Hsu & Kakade, 2013) shape better initializations, but certify nothing about the point returned. Theorem 4.1 works on that endpoint instead, a necessary condition for global optimality testable by a simple re-estimation, which also sharpens what *Quasi*-GD stops on, from ε- to exact Clarke stationarity. Beyond the synthetic studies, on real data CoM halves the out-of-sample forecast error of a tuned four-regime benchmark for quarterly U.S. GNP growth, and reconstructs scalp EEG signals, whose entangled high-frequency dynamics switch on conditions no recording reveals.

2 Establishment of the CoM Framework

We begin with a structural break regression (Hansen, 1999) with no noise,

$$y = \begin{cases} 0.6x + 3, & x \leqslant 18, \\ -0.5x + 12, & x > 18, \end{cases} \tag{6}$$

and suppose for now that everything is known except for the break location. A DaC procedure, grid search being the canonical instance, scans a candidate threshold r to minimize

$$\text{obj} = \sum\nolimits_i [y_i - (0.6x_i + 3)]^2 \mathbb{1}\{x_i \leqslant r\} + \sum\nolimits_i [y_i - (-0.5x_i + 12)]^2 \mathbb{1}\{x_i > r\} \geqslant 0, \tag{7}$$

whose global minimum 0 is attained only when $r = 18$ is correctly identified. CoM reads the same model differently: every observation comes from one of the two mechanisms, so wherever the break sits, at least one of the two residuals vanishes at every i, that is, $\min\{[y_i - (0.6x_i + 3)]^2, [y_i - (-0.5x_i + 12)]^2\} \equiv 0$. Aggregating these pointwise minima gives an objective that is **free of** r and still attains its global minimum:

$$\text{obj} = \sum\nolimits_i \min\{[y_i - (0.6x_i + 3)]^2, [y_i - (-0.5x_i + 12)]^2\} \equiv 0. \tag{8}$$

The gap widens once the break condition becomes less tractable. Replacing the threshold in (6) by an oblique condition on y itself, or by a condition on a lagged response, gives

$$y = \begin{cases} 0.6x + 3, & \text{if } y > 1.5x - 15, \\ -0.5x + 12, & \text{if } y \leqslant 1.5x - 15, \end{cases} \tag{9}$$

$$y_i = \begin{cases} 0.6x_i + 3, & \text{if } y_{i-3} > 7, \\ -0.5x_i + 12, & \text{if } y_{i-3} \leqslant 7. \end{cases} \tag{10}$$

Each variant forces a different indicator into (7), and the quality of the resulting fit is determined by how closely that indicator matches the truth, which cannot be checked when the splitting rule is unobserved, as it is in most applications. In contrast, objective (8) is identical in all three cases, and in any other, retaining its global minimum 0 unconditionally.

However, assuming noiseless mechanisms with known parameters is a theoretical luxury. Returning to the general problem (1), each observation must originate from exactly one candidate. That is,

$$\boldsymbol{y}_i \in \{f_j(\boldsymbol{x}_i, \boldsymbol{\theta}_j, \varepsilon_{ji})\}_{j \in J}, \qquad i \in I,$$

where J is a finite nonempty mechanism index set and $\{\varepsilon_{ji}\}_j$ denotes homo- or heteroscedastic noises. No noise distribution is assumed here, since CoM never reads $\bar{\ell}_j$ as a negative log-likelihood: discrepancies are design choices, and the noise models they imply, if any, are read off afterwards. Many survive that reading, the squared loss being the Gaussian case, but for others the implied noise model is contrived or nonexistent: the Huber loss matches a likelihood only under a nonstandard profile, a density Gaussian in the center and Laplace in the tails (Huber, 1964), and the check loss (Koenker & Bassett, 1978) only under an asymmetric Laplace (Yu & Moyeed, 2001). The Tukey biweight (Beaton & Tukey, 1974), Geman–McClure loss (Green, 1990), and Welsch loss (Barron, 2019) correspond to no density at all since their exponentiated negatives are non-integrable. Treating each $\bar{\ell}_j$ as a purely geometric measure of fit keeps all of them admissible.

Staging a competition among all $|J|$ candidates at every observation and aggregating the winners over I yields the CoM objective (5). The form is unfamiliar; the criterion behind it is not. The next theorem identifies it with the oracle problem (3), profiled over the segmentation.

Theorem 2.1. *Let J be finite and nonempty. For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\min_{P \in \mathcal{P}} R(\boldsymbol{\theta}; P) = \min_{P \in \mathcal{P}} \sum\nolimits_{j \in J} \sum\nolimits_{i \in I_j} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i) \;=\; \sum\nolimits_{i \in I} \min_{j \in J} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i) = \ell(\boldsymbol{\theta}), \tag{11}$$

and consequently

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \min_{P \in \mathcal{P}} R(\boldsymbol{\theta}; P) \;=\; \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}), \tag{12}$$

with $\boldsymbol{\theta}^*$ attaining the left-hand infimum if and only if it attains the right-hand one.

Identity (11) is algebraic: it holds pointwise in $\boldsymbol{\theta}$, before any optimization is performed, for any fixed dataset, and with no assumption on the noise, on separation between mechanisms, or on the supplied f_j and $|J|$ being correct. Inner minimization $\min_{P\in\mathcal{P}}$ is always attained because $\mathcal{P}$ is finite. So segmentation agnosticism costs nothing at the optimum: CoM and the brute-force search share both their optimal value and their optimizers. Attainment in $\boldsymbol{\theta}$ is a separate question from (11) and is settled by a simple condition.

Corollary 2.2. *Suppose that for every $j \in J$ and every $S \subseteq I$ the minimum of $\sum_{i\in S} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i)$ over $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$ is attainable. Then ℓ attains its minimum on $\boldsymbol{\Theta}$, and $\boldsymbol{\theta}^* := (\boldsymbol{\theta}^*_j)_{j\in J} \in \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$ induces a disjoint partition $P^* := \sqcup_{j\in J} I^*_j$ under any consistent tie-breaking on*

$$I^*_j := \big\{ i \in I : \bar{\ell}_j(\boldsymbol{\theta}^*_j; \boldsymbol{x}_i, \boldsymbol{y}_i) = \min_{k\in J} \bar{\ell}_k(\boldsymbol{\theta}^*_k; \boldsymbol{x}_i, \boldsymbol{y}_i) \big\}, \quad j \in J. \tag{13}$$

The pair $(\boldsymbol{\theta}^*, P^*)$ then satisfies

$$(\boldsymbol{\theta}^*, P^*) \in \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\ P\in\mathcal{P}} R(\boldsymbol{\theta}; P), \tag{14}$$

that is, minimizing ℓ alone solves the oracle problem: none of its $|J|^{|I|}$ segmentations is enumerated, yet $(\boldsymbol{\theta}^*, P^*)$ is optimal among all of them. Moreover,

$$\boldsymbol{\theta}^*_j \in \operatorname{argmin}_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j} \sum\nolimits_{i\in I^*_j} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i), \quad j \in J. \tag{15}$$

The hypothesis binds each $\bar{\ell}_j$ alone, never the overall ℓ, so most practical losses meet it. The quantifier over every $S \subseteq I$ is necessary, and attainment on $S = I$ alone does not suffice (Remark B.2). Beyond attainment, (15) transfers structure to $\boldsymbol{\theta}^*$. On the induced segmentation, $\boldsymbol{\theta}^*_j$ is the *segment-wise M-estimator* that DaC would report had it been handed I^*_j, and $\boldsymbol{\theta}^*$ is just the concatenation $(\boldsymbol{\theta}^*_j)_{j\in J}$. The properties of $\boldsymbol{\theta}^*$ therefore follow from $R(\boldsymbol{\theta}; P^*)$, a fixed-partition problem simpler than the pointwise minimum defining ℓ. The randomness of I^* is still what a full asymptotic treatment of $\boldsymbol{\theta}^*$ must account for, but its mis-assignment decays as the signal-to-noise ratio grows, as quantified for the K-means case by Lu & Zhou (2016).

Remark 2.3 (Misspecification). *Theorem 2.1 and Corollary 2.2 together fix what $\boldsymbol{\theta}^*$ means when the forms of f_j or the cardinality of J are misspecified: $\boldsymbol{\theta}^*$ is the best fit of the supplied forms over the best of all $|J|^{|I|}$ segmentations of $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i\in I}$. This is the familiar picture from misspecified M-estimation, where the estimator still targets a well-defined best approximation inside the assumed family (Huber, 1967; White, 1982), except that here the target is pinned exactly rather than as an asymptotic limit. Two analysts who supply the same f_j and $\bar{\ell}_j$ therefore target the same estimand under CoM, whereas under DaC the estimand itself depends on the segmentation heuristic, so different change-point or clustering rules define different targets.*

3 Convergence Analysis

As established in Section 2, the oracle combinatorial problem and the CoM problem differ only in form. What changes is the cost: the former ranges over $|J|^{|I|}$ segmentations, the latter over just $|I|$ independent comparisons of $|J|$ numbers. The search is neither relaxed nor approximated, so the difficulty reappears in the $\min_{j\in J}$ operator, which makes $\ell(\boldsymbol{\theta})$ non-convex and non-smooth.

To accommodate these irregularities, we adopt the Clarke subdifferential (Clarke, 1990), which generalizes both the gradient and the convex subdifferential to non-smooth, non-convex objectives. For the analysis that follows, each $\bar{\ell}_j$ is embedded into the global parameter space $\boldsymbol{\Theta} = \mathbb{R}^d$ by zero-padding, with $d := \sum_{j\in J} \dim(\boldsymbol{\theta}_j)$. Writing $\boldsymbol{\theta}$ in place of $\boldsymbol{\theta}_j$ changes no value but places every component in a common space,

$$\ell_j(i, \boldsymbol{\theta}) := \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i) = \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i) + \sum\nolimits_{m\neq j} \mathbf{0}^\top \boldsymbol{\theta}_m, \qquad \ell(\boldsymbol{\theta}) = \sum\nolimits_{i\in I} \min_{j\in J} \ell_j(i, \boldsymbol{\theta}), \tag{16}$$

where i denotes $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for brevity. Then $\nabla_{\boldsymbol{\theta}} \ell_j$ is supported on the j-th block alone:

$$\nabla^\top_{\boldsymbol{\theta}} \ell_j(i, \boldsymbol{\theta}) = (\cdots, \mathbf{0}^\top, \nabla^\top_{\boldsymbol{\theta}_j} \ell_j(i, \boldsymbol{\theta}), \mathbf{0}^\top, \cdots). \tag{17}$$

Throughout, we make three mild assumptions:

(A1) J is finite and nonempty.

(A2) For every $j \in J$ and every $i \in I$, $\ell_j(i, \boldsymbol{\theta})$ is L_0-smooth in $\boldsymbol{\theta}$. Write $L := |I| \cdot L_0$.

(A3) $\ell^\star := \inf_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) > -\infty$, satisfied if each $\ell_j(i, \boldsymbol{\theta})$ is bounded below, as most practical losses are.

3.1 Clarke Subdifferential on Pointwise Minimum

Definition 3.1 (Clarke subdifferential). *Assume $\varphi : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz in $\boldsymbol{\theta}$. Its Clarke subdifferential (Clarke, 1975, Definition 1.1) is the convex hull*

$$\partial^C \varphi(\boldsymbol{\theta}) = \mathbf{conv}\big\{\boldsymbol{s} \in \mathbb{R}^d : \exists\, \boldsymbol{\theta}^m \to \boldsymbol{\theta},\ \nabla\varphi(\boldsymbol{\theta}^m) \text{ exists, and } \nabla\varphi(\boldsymbol{\theta}^m) \to \boldsymbol{s}\big\}. \tag{18}$$

The form of the next theorem is then unsurprising, though its derivation is not immediate, running through *Danskin's Theorem* (Danskin, 1967; Clarke, 1990) as Section C.1 shows.

Theorem 3.2. *Under Assumptions (A1-2), for every $i \in I$ and every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\partial^C(\min_{j\in J} \ell_j(i,\boldsymbol{\theta})) = \mathbf{conv}\{\nabla\ell_k(i,\boldsymbol{\theta}) : k \in J_{min}(i,\boldsymbol{\theta})\}, \tag{19}$$

where $J_{min}(i,\boldsymbol{\theta}) := \operatorname{argmin}_{j\in J} \ell_j(i,\boldsymbol{\theta})$ is the set of active mechanisms at i.

At a kink point this convex hull is a set rather than a vector, and we use only its element of smallest Euclidean norm, ∂_s^C, whose negative is the steepest descent direction (Theorem C.3), as is standard in convex cases (Wright & Recht, 2022, §9.1). For a general locally Lipschitz φ, $\partial_s^C\varphi(\boldsymbol{\theta}) := \operatorname{argmin}_{\boldsymbol{v}\in\partial^C\varphi(\boldsymbol{\theta})} \|\boldsymbol{v}\|_2$, which exists and is unique because $\partial^C\varphi(\boldsymbol{\theta})$ is nonempty (by the *Hahn-Banach Theorem*), convex, and compact (by *Alaoglu's Theorem*). At a non-kink point $J_{min}(i,\boldsymbol{\theta})$ is a singleton $\{j\}$ so (19) collapses to $\nabla\ell_j(i,\boldsymbol{\theta})$, which is then ∂_s^C too.

Computing ∂_s^C at a kink point is a *linearly constrained quadratic program* (LCQP, Section C.2), which in general has no closed form and must be solved iteratively (Wolfe, 1976). Here it has one: the gradients are orthogonal across j by (17), so the Gram matrix is diagonal and the weights fall out as inverse squared norms, supported on the active set.

Corollary 3.3 (Orthogonality gives a closed form). *Under Assumptions (A1-2), for every $i \in I$ and every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\partial_s^C\Big(\min_{j\in J} \ell_j(i,\boldsymbol{\theta})\Big) = \sum_{j\in J_{min}(i,\boldsymbol{\theta})} \frac{\nabla_{\boldsymbol{\theta}}\ell_j(i,\boldsymbol{\theta}) / \left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i,\boldsymbol{\theta})\right\|_2^2}{\sum_{j\in J_{min}(i,\boldsymbol{\theta})} 1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i,\boldsymbol{\theta})\right\|_2^2}. \tag{20}$$

3.2 From Individual to Collective

Obtaining $\partial^C\ell(\boldsymbol{\theta})$ is not easy, let alone $\partial_s^C\ell(\boldsymbol{\theta})$. For the first, ∂^C is subadditive rather than additive, giving only the inclusion $\partial^C(\sum_i \varphi_i) \subseteq \sum_i \partial^C\varphi_i$, with equality when each φ_i is Clarke regular,[1] which $\min_{j\in J} \ell_j(i,\boldsymbol{\theta})$ is not. Equality nonetheless holds, by a max-min argument. All derivations in this subsection are in Sections D and E.

Theorem 3.4. *Under (A1-2), $\partial^C\ell(\boldsymbol{\theta})$ is nonempty and*

$$\partial^C\ell(\boldsymbol{\theta}) = \partial^C\Big(\sum\nolimits_{i\in I} \min_{j\in J} \ell_j(i,\boldsymbol{\theta})\Big) = \sum\nolimits_{i\in I} \partial^C(\min_{j\in J} \ell_j(i,\boldsymbol{\theta})). \tag{21}$$

Accordingly, at each iterate $\boldsymbol{\theta}^{(k)}$ we collect the points where mechanism j is the sole minimizer and those where it shares the minimum,

$$k_j := \Big\{i \in I : J_{min}(i,\boldsymbol{\theta}^{(k)}) = \{j\}\Big\}, \quad k_{j,\text{kink}} := \Big\{i \in I : j \in J_{min}(i,\boldsymbol{\theta}^{(k)}),\ \Big|J_{min}(i,\boldsymbol{\theta}^{(k)})\Big| \geqslant 2\Big\}.$$

Summing the pointwise minimum-norm elements (20) then defines a special direction

$$\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)}) := \sum_{j\in J}\sum_{i\in k_j} \nabla_{\boldsymbol{\theta}}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{j\in J}\sum_{i'\in k_{j,\text{kink}}} \frac{\nabla_{\boldsymbol{\theta}}\ell_j(i',\boldsymbol{\theta}^{(k)}) / \left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})} 1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}. \tag{22}$$

Since $\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)}) = \sum_{i\in I} \partial_s^C(\min_{j\in J} \ell_j(i,\boldsymbol{\theta}^{(k)})) \in \sum_{i\in I} \partial^C(\min_{j\in J} \ell_j(i,\boldsymbol{\theta}^{(k)}))$, this direction is a genuine element of $\partial^C\ell(\boldsymbol{\theta}^{(k)})$ by Theorem 3.4. Because $\ell_m(i',\boldsymbol{\theta}^{(k)}) = \min_{j\in J} \ell_j(i',\boldsymbol{\theta}^{(k)})$ for every $m \in J_{min}(i',\boldsymbol{\theta}^{(k)})$, the same weights decompose the objective itself,

$$\ell(\boldsymbol{\theta}^{(k)}) = \sum_{j\in J}\sum_{i\in k_j} \ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{j\in J}\sum_{i'\in k_{j,\text{kink}}} \frac{\ell_j(i',\boldsymbol{\theta}^{(k)}) / \left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})} 1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}, \tag{23}$$

[1] φ is Clarke regular at $\boldsymbol{\theta}$ if its one-sided directional derivative exists and agrees with the Clarke directional derivative in every direction (Clarke, 1990).

which suggests defining a mechanism-specific decomposition,

$$\ell_j^{(k)}(\boldsymbol{\theta}) := \sum_{i \in k_j} \ell_j(i, \boldsymbol{\theta}) + \sum_{i' \in k_{j,\text{kink}}} \frac{\ell_j(i', \boldsymbol{\theta}) / \left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i', \boldsymbol{\theta}^{(k)}) \right\|_2^2}{\sum_{j \in J_{min}(i', \boldsymbol{\theta}^{(k)})} 1 / \left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i', \boldsymbol{\theta}^{(k)}) \right\|_2^2}, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{24}$$

The interpretation of (24) is clear: for non-kink points $i \in k_j$, mechanism j is uniquely activated and thus contributes in full; for kink points $i' \in k_{j,\text{kink}}$, the activation is shared with other mechanisms, resulting in only a partial contribution.

3.3 *Quasi*-GD Algorithm and Its Convergence

Algorithm 1 *Quasi*-Gradient Descent

1: **Input:** Dataset $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i \in I}$, initialization $\boldsymbol{\theta}^{(0)}$, tolerance level ε, step size $\alpha \in (0, 1/L]$
2: **Set** $k \leftarrow 0$
3: **repeat**
4: Update $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \alpha \cdot \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$
5: $k \leftarrow k + 1$
6: **until** $\|\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})\|_2 < \varepsilon$
7: **Return** $\boldsymbol{\theta}^{(k)}$

We now propose our novel ***Quasi*-gradient descent (*Quasi*-GD)** algorithm 1, whose core is a gradient-descent-style update rule

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \alpha \cdot \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}). \tag{25}$$

This update applies not to ℓ but to $\sum_{j \in J} \ell_j^{(k)}$, a differentiable majorant meeting ℓ at $\boldsymbol{\theta}^{(k)}$ with gradient $\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$ there, as the next lemma shows.

Lemma 3.5. *Equations* (16) *and* (22)–(24) *imply that, for every* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and every* $j \in J$,

$$\sum\nolimits_{j \in J} \ell_j^{(k)}(\boldsymbol{\theta}) \geqslant \ell(\boldsymbol{\theta}), \text{ with equality at } \boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}, \tag{26}$$

$$\sum\nolimits_{j \in J} \nabla_{\boldsymbol{\theta}} \ell_j^{(k)}(\boldsymbol{\theta}^{(k)}) = \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}), \tag{27}$$

$$\langle \nabla_{\boldsymbol{\theta}} \ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) \rangle = \left\| \nabla_{\boldsymbol{\theta}} \ell_j^{(k)}(\boldsymbol{\theta}^{(k)}) \right\|_2^2, \tag{28}$$

$$\sum\nolimits_{j \in J} \langle \nabla_{\boldsymbol{\theta}} \ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) \rangle = \left\| \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) \right\|_2^2. \tag{29}$$

The equality in (26) is special to $\boldsymbol{\theta}^{(k)}$: elsewhere the two sides assign points under different parameters, $\boldsymbol{\theta}^{(k)}$ on the left by (24), $\boldsymbol{\theta}$ on the right, so the majorant sits strictly above. Bounding $\sum_{j \in J} \ell_j^{(k)}$ at $\boldsymbol{\theta}^{(k+1)}$ therefore bounds ℓ there, which is the first inequality below.

Lemma 3.6 (*Quasi* Descent Lemma). *Under Assumptions (A1-2), for any step size* $\alpha \in (0, 1/L]$,

$$\ell(\boldsymbol{\theta}^{(k+1)}) \leqslant \sum\nolimits_{j \in J} \ell_j^{(k)}(\boldsymbol{\theta}^{(k+1)}) \leqslant \ell(\boldsymbol{\theta}^{(k)}) - \frac{\alpha}{2} \left\| \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) \right\|_2^2. \tag{30}$$

Theorem 3.7 (*Quasi*-Gradient Descent Convergence). *Under Assumptions (A1-3), the Quasi-GD update rule* (25) *with step size* $\alpha \in (0, 1/L]$ *bounds the convergence rate by*

$$\min_{0 \leqslant k \leqslant K} \left\| \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) \right\|_2^2 \leqslant \frac{2\big(\ell(\boldsymbol{\theta}^{(0)}) - \ell^*\big)}{\alpha(K+1)}. \tag{31}$$

Corollary 3.8. *The* ε*-Clarke stationarity* $\operatorname{dist}(0, \partial^C \ell(\boldsymbol{\theta}^{(k)})) < \varepsilon$ *is achieved in* $\mathcal{O}\big(\varepsilon^{-2}\big)$ *steps.*

This is the textbook gradient descent guarantee (Wright & Recht, 2022, Chapter 3), item for item: the same step-size range $(0, 1/L]$, the same $\mathcal{O}(1/K)$ decay with the same constants, the same $\mathcal{O}(\varepsilon^{-2})$ efficiency, just with $\widetilde{\nabla}_s \ell$ standing where $\nabla \ell$ cannot. The CoM objective is built from $|J|$ competing mechanisms, but **the competition is free**: nothing in the rate is charged for the $|J|$ mechanisms, for the non-convexity and non-smoothness their competition creates, or for the $|J|^{|I|}$ segmentations the problem ranges over. All the difficulty is absorbed instead into the explicit derivation of (22) and the construction of (24); $|J|$ enters the arithmetic of a single iteration and $|I|$ enters through L, but neither reaches the rate.

Remark 3.9 (**Choosing α without L**). *The admissible range $(0, 1/L]$ needs L, which is often inaccessible. A standard way is backtracking: start from some α_0 and halve until the trial point satisfies* (30)*, taking at most $\lceil \log_2(\alpha_0 L) \rceil$ steps since every $\alpha \leqslant 1/L$ passes.*

Remark 3.10 (**Scaling**). *Detailed in Section F.3, the overall computational complexity of Quasi-GD is $\mathcal{O}(|I| \cdot |J| \cdot \bar{d}L/\varepsilon)$ if $\alpha = 1/L$. This $\mathcal{O}(|I| \cdot |J|)$ scaling bypasses the scaling of grid search $\mathcal{O}(|I|^{|J|-1})$ or $\mathcal{O}(|I|^2)$ (Bai & Perron, 1998), change-point detection $\mathcal{O}(|I|^2 \cdot |J|)$ (Truong et al., 2020), hierarchical clustering $\mathcal{O}(|I|^3)$ (Hastie, 2009), or the intractable $\mathcal{O}(|J|^{|I|})$ for brute-force search (Späth, 1979). And these baselines still reach only the segmentation forms they assume.*

4 Local Optimum Characterization

Everything Section 2 promises is a property of the global minimizer. What Section 3 delivers is an ε-Clarke stationary point. Closing that gap with a sufficient condition is not to be expected, since even K-means, the simplest CoM case, is NP-hard to solve globally (Dasgupta, 2008; Mahajan et al., 2012). A necessary condition costs just a region-wise re-estimation.

Theorem 4.1 (**Necessary condition for global optimality**). *Let $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ be any estimator and P^* the segmentation it induces through* (13)*. Re-estimating region-wise gives $\boldsymbol{\theta}^{**} := (\boldsymbol{\theta}_j^{**})_{j \in J}$, where $\boldsymbol{\theta}_j^{**} := \operatorname{argmin}_{\boldsymbol{\theta}_j} \sum_{i \in I_j^*} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, \boldsymbol{y}_i)$. If $\boldsymbol{\theta}^*$ is a global minimizer of $\ell(\boldsymbol{\theta})$, then $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^*)$.*

The check is one sweep of *alternating minimization* (Csiszár & Tusnády, 1984) on the oracle problem: inducing P^* minimizes $R(\boldsymbol{\theta}^*; \cdot)$ over $\mathcal{P}$ by (11), and re-estimating to $\boldsymbol{\theta}^{**}$ minimizes $R(\cdot; P^*)$ over $\boldsymbol{\Theta}$. The condition it tests is (15), so nothing beyond Section 2 is assumed. The argument uses only the segmentation $\boldsymbol{\theta}^*$ induces, so the test is indifferent to provenance: $\boldsymbol{\theta}^*$ may come from *Quasi*-GD or any other method.

Three consequences make it more than a diagnostic. It is constructive: $\ell(\boldsymbol{\theta}^{**}) \leqslant \ell(\boldsymbol{\theta}^*)$ always, so a failed check hands back $\boldsymbol{\theta}^{**}$, a better initialization for restart. It sharpens the guarantee *Quasi*-GD stops on, from ε- to exact Clarke stationarity. And it disqualifies spurious local optima that the stopping rule accepts, as in Table 1, though it cannot certify the ones that pass.

Proposition 4.2 (**Exact stationarity**). *If $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^*)$, then $0 \in \partial^C \ell(\boldsymbol{\theta}^*)$.*

The table runs the check on the two-regime quadratic $y = [x^2 + 0.5\varepsilon]\mathbb{1}\{x \leqslant 0\} + [x^2 - 10 + \varepsilon]\mathbb{1}\{x > 0\}$, fitted by CoM $\ell(\boldsymbol{\theta}) = \sum_i \min\{[y_i - a_1 x_i^2 - b_1]^2, [y_i - a_2 x_i^2 - b_2]^2\}$, from two random initializations. Both halt with the stationarity measure below tolerance. The test $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^*)$ is equivalent to $\boldsymbol{\theta}^{**} = \boldsymbol{\theta}^*$ whenever the region-wise minimizer is unique (Proposition G.1), as it is under the squared loss here, so the parameter columns can be read as the check itself. For the first, the re-estimation cuts ℓ sevenfold and lands closer to the truth, ruling $\boldsymbol{\theta}^*$ out as global. The second reproduces itself to every reported digit and sits within 0.06 of the truth.

Table 1: Verification of the necessary condition. Local Optimum 1 fails the check: re-estimation strictly lowers ℓ, so it is not global. Local Optimum 2 passes, reproducing both $\boldsymbol{\theta}^*$ and $\ell(\boldsymbol{\theta}^*)$.

$\boldsymbol{\theta}$	Local Optimum 1					Local Optimum 2				
	$y = x^2 + 0.5\varepsilon$		$y = x^2 - 10 + \varepsilon$		$\ell(\boldsymbol{\theta})$	$y = x^2 + 0.5\varepsilon$		$y = x^2 - 10 + \varepsilon$		$\ell(\boldsymbol{\theta})$
	a_1	b_1	a_2	b_2		a_1	b_1	a_2	b_2	
True Parameter	1.00	0.00	1.00	−10.00	1868.2	1.00	0.00	1.00	−10.00	1868.2
$\boldsymbol{\theta}^*$	0.88	0.73	0.84	−11.05	21 942.1	1.00	0.02	1.01	−10.06	1866.5
$\boldsymbol{\theta}^{**}$	0.93	0.41	0.98	−9.91	3022.3	1.00	0.02	1.01	−10.06	1866.5

5 Numerical Experiments

The studies below are ordered by how little an assumed segmentation form can do: a vertical split, an oblique one, regimes too brief to locate, and unobservable conditions.

5.1 Structural Break Regression under Different Segmentation Forms

Over 500 independent trials with heteroscedastic noise, we compare CoM against three mainstream DaC baselines, each on its established R package: grid search (GS, `chngpt` (Fong et al., 2017)),

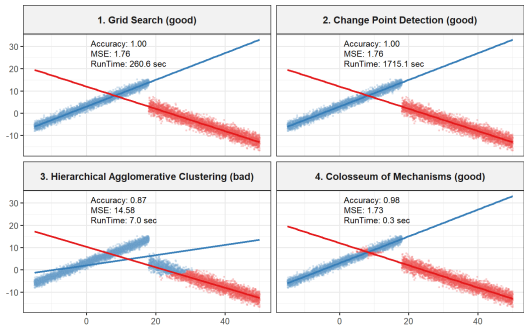

Figure 1: Vertical split at $x = 18$.

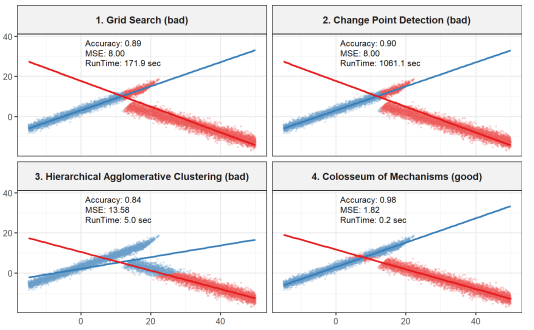

Figure 2: Oblique split at $y = 1.5x - 15$.

Table 2: Structural break regression, mean $\pm$ sd over 500 trials.

Split	**Metric**	**GS**	**CPD**	**HAC**	**CoM**(ours)
Vertical	Accuracy	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.07	0.98 ± 0.00
	MSE	1.76 ± 0.03	1.76 ± 0.03	5.22 ± 6.58	$\mathbf{1.73 \pm 0.03}$
	RunTime (sec)	132.02 ± 45.60	880.43 ± 272.35	4.26 ± 1.48	$\mathbf{0.14 \pm 0.05}$
Oblique	Accuracy	0.89 ± 0.01	0.89 ± 0.01	0.84 ± 0.08	$\mathbf{0.98 \pm 0.00}$
	MSE	7.89 ± 0.14	7.89 ± 0.14	14.46 ± 6.99	$\mathbf{1.76 \pm 0.03}$
	RunTime (sec)	157.05 ± 65.43	1031.31 ± 326.71	4.98 ± 1.86	$\mathbf{0.14 \pm 0.05}$

change-point detection (CPD, `strucchange` (Zeileis et al., 2002)), and hierarchical agglomerative clustering (HAC, `hclust` (Ward, 1963)). The first regression splits vertically,

$$y = \begin{cases} 0.6x + 3 + 1.0\,\varepsilon_1, & x \le 18, \\ -0.5x + 12 + 1.6\,\varepsilon_2, & x > 18, \end{cases}$$

the best case for GS and CPD (Figure 1), whose built-in assumption of a vertical split in x is exactly correct. Even so, they need 132 and 880 seconds against CoM's 0.14; HAC is fast but unstable, 382 of 500 trials reaching accuracy 1, the remaining 118 collapsing to ≈ 0.85 with MSE ≈ 16.

The second regression changes only the switching condition, to $y > 1.5x - 15$ (Figure 2), breaking the built-in assumption. GS and CPD then lose 11 accuracy points and more than quadruple their MSE, whereas CoM is unchanged to two decimals.

Both settings are deliberately simple, so the baselines can run at all. Elsewhere they cannot, neither on the real data below nor on the synthetic studies of Section A, where fast switching violates the DaC premise of contiguous partitions.

5.2 GNP Forecasting under Regime Switching

Threshold-type regime switching has long been a workhorse for output growth and business cycles (Tong & Lim, 1980; Tsay, 1998; Hansen, 2011). Aside from the synthetic TAR studies in Section A.1, we fit a four-regime TAR to U.S. quarterly GNP growth by minimizing

$$\ell(\boldsymbol{\theta}) = \sum\nolimits_t \min_{j \in J} \left(y_t - \mu_j - \sum\nolimits_{k=1}^{p_j} \phi_{jk} y_{t-k} \right)^2,$$

training on 1947:Q1 to 1991:Q1 to match the split of Tiao & Tsay (1994) and forecasting on 1991:Q2 to 2025:Q3. The regimes are too short-lived for the baselines of Section 5.1, so we compare against that study's fitted model, unchanged. CoM fits better on the training period, and the margin widens out of sample (Table 3), so the gain is generalization, not a closer fit to history.

Table 3: Mean squared error for CoM and the benchmark model (Tiao & Tsay, 1994). Both take the same mechanism form, so what differs is how the regimes are found, not the specification.

	In-sample (1947:Q1–1991:Q1)	Out-of-sample (1991:Q2–2025:Q3)
Published benchmark	0.0184	0.0224
CoM	0.0031	0.0095

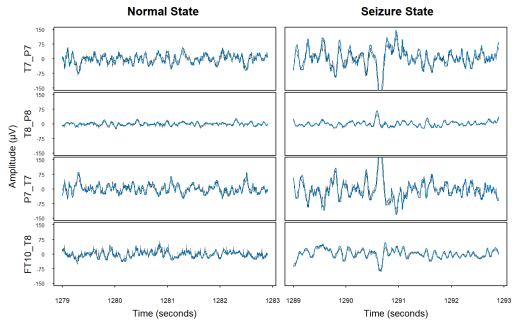

Figure 3: Case 21, CoM recovery (blue) over the recording (black).

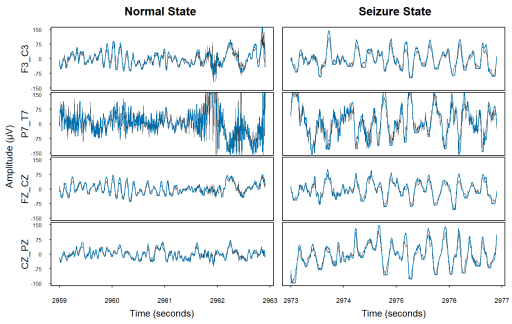

Figure 4: Case 02, same layout.

5.3 EEG Signal Reconstruction

The World Health Organization estimates that epilepsy affects approximately 50 million people globally (WHO, 2026). Seizures manifest as transient, stochastic, patient-specific disruptions of brain activity; while scalp electroencephalogram (EEG) records them readily, separating the underlying components from one or several recordings remains an open challenge.

Let $X_1(t), \ldots, X_N(t)$ be source signals fixed in advance, like the delta, theta, alpha, beta, and gamma rhythms that cortical recordings are conventionally decomposed into (Buzsáki & Draguhn, 2004), or components from a preliminary analysis; failing both, a penalized overcomplete basis serves instead (Section A.4). What changes between a normal and a seizure phase is not the rhythms themselves but how strongly each contributes, so channel m is mixed by one of $|J|$ weight signals at any time, under a condition that no recording reveals:

$$y_t^m = \sum\nolimits_{j \in J} \left[\sum\nolimits_{n=1}^{N} h_{jn}^m X_n(t) + c_j^m + \varepsilon_t \right] \mathbb{1}\{t \in T_j\},$$

where $(T_j)_{j \in J}$ is the unknown recording partition. This is (1) with an unobservable condition, the case CoM is built for, and the CoM objective follows from

$$\ell^m(\boldsymbol{\theta}) = \sum\nolimits_t \min_{j \in J} \left[y_t^m - \sum\nolimits_{n=1}^{N} h_{jn}^m X_n(t) - c_j^m \right]^2.$$

$\ell^m(\boldsymbol{\theta})$ is fitted independently to all 23 channels of Case 21 from the open-access CHB-MIT database (Guttag, 2010). Figure 3 shows the four channels of each whose recovered weights h and c change most between the normal and the seizure phase; the others are omitted to keep the figures legible. CoM reaches RMSE $12.12\,\mu$V in the normal phase and $20.16\,\mu$V in the seizure phase on this case, against 19.43 and 41.69 for a single-mechanism fit of the same form; the margin over the single mechanism widens from $1.60\times$ in the normal phase to $2.07\times$ in the seizure phase. The same tracking holds on Case 02 (Figure 4), whose $P7_T7$ channel carries dense high-frequency activity in the normal phase. We explicitly note that this demonstration illustrates algorithmic capacity only, leaving clinical interpretation and validity for future applied research.

6 Discussion

DaC needs two assumptions, the segmentation form (divide) and the forms of the mechanisms (conquer); CoM needs only the second. The difference is one of reach: a method that assumes a segmentation form can recover only the segmentations that form expresses, while one that assumes none reaches all $|J|^{|I|}$ of them. Where the assumption is correct, DaC is the better tool for inference, though far slower in the cases we ran (Section 5.1); where it is not, CoM is unaffected.

The forms of f_j are the ordinary burden of any parametric model; when unknown, they can be approximated by an overcomplete basis, with a penalty to keep the fit well-posed (Section A.4). When $|J|$ is unknown, no criterion settles it, despite decades of proposals from likelihood-ratio tests to BIC (McLachlan & Peel, 2000). Our recipe is to overshoot: run $|J| = 8$ where 5 is plausible. The surplus mechanisms then claim a negligible share of I and are discarded, as in overfitted Bayesian mixtures (Rousseau & Mengersen, 2011). Since a run typically takes seconds, several values of $|J|$ can be compared. When the forms or $|J|$ are still wrong, Remark 2.3 fixes what $\boldsymbol{\theta}^*$ means.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing: grammar, word choice, and text we had already drafted. AI was also used in plotting code generation and code review; but the core algorithm was written manually. We did not use generative AI tools to generate research ideas, to derive or prove any theoretical result, to write the accompanying code, or to produce or analyze any experimental result. We have reviewed all AI-assisted work: every suggested edit was checked against the intended technical meaning and then accepted, revised, or rejected sentence by sentence, and no claim, citation, or numerical value in this paper originated from a generative AI tool. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

CoM is a theoretical framework, and its output depends on the correctness of the assumed mechanism forms, the quality of the data, and the initialization. One failure mode runs through every application: convergence to a spurious local optimum, at which observations are assigned to the wrong mechanism or a mechanism is recovered with biased parameters. Theorem 4.1 disqualifies some such points but certifies none that pass, so the framework returns no confidence measure that would flag either. What such an error costs depends on the deployment: overlooked or spurious events in monitoring and diagnosis, misspecified regimes in trading and control, and incorrect segmentations propagating into downstream scientific inference. CoM outputs therefore require review by a domain specialist before any decision or application, especially in clinical decision-making or autonomous operation.

The signal reconstruction in Section 5.3 uses the CHB-MIT Scalp EEG Database v1.0.0 (Guttag, 2010), collected at Boston Children's Hospital and hosted on PhysioNet. It is released under the Open Data Commons Attribution License v1.0, which permits use, modification, and redistribution subject to attribution. The database is credited in Section 5.3 and in the references; the recordings are de-identified and openly accessible, and no separate access agreement is required. We use them to illustrate algorithmic capacity only, and draw no clinical conclusion from the reconstructions.

REPRODUCIBILITY STATEMENT

The contribution of this paper is primarily theoretical. Assumptions (A1-A3) are stated in Section 3 and every result names the subset of them it uses; the one further hypothesis, the attainability condition behind Corollary 2.2, is stated in the corollary itself. Every claim is proved in full in the appendix.

The code that produces every reported number is included in the supplementary material, anonymized for review. The data are public: U.S. quarterly GNP growth for Section 5.2, with the training and forecasting periods given there, and the CHB-MIT Scalp EEG Database (Guttag, 2010) for Section 5.3, whose case and channels are named in the text. The baselines of Section 5.1 are the published R packages cited there.

REFERENCES

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, pp. 47–78, 1998.

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4331–4339, 2019.

David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, 2011.

Albert E. Beaton and John W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974. doi: 10.1080/00401706.1974.10489171.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *Science*, 304 (5679):1926–1929, 2004. doi: 10.1126/science.1099745.

K. S. Chan. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21(1):520–533, 1993. doi: 10.1214/aos/1176349040.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics & Decisions*, 1(Supplement Issue):205–237, 1984.

John M Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 1967.

Sanjoy Dasgupta. The hardness of k-means clustering. 2008. URL `https://books.google.com/books?id=riJuAQAACAAJ`.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps: D. drusvyatskiy, c. paquette. *Mathematical Programming*, 178(1): 503–558, 2019.

Youyi Fong, Ying Huang, Peter B Gilbert, and Sallie R Permar. chngpt: Threshold regression model estimation and inference. *BMC bioinformatics*, 18(1):454, 2017.

Stephen M. Goldfeld and Richard E. Quandt. A Markov model for switching regressions. *Journal of Econometrics*, 1(1):3–15, 1973.

Peter J Green. Bayesian reconstructions from emission tomography data using a modified em algorithm. *IEEE Transactions on Medical Imaging*, 9(1):84–93, 1990.

Ruosi Guo, Chunming Zhang, and Zhengjun Zhang. Maximum independent component analysis with application to eeg data. *Statistical Science*, 35(1):145–157, 2020.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, June 2010. doi: 10.13026/C2K01R. URL `https://doi.org/10.13026/C2K01R`. Version 1.0.0.

Bruce E Hansen. Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of econometrics*, 93(2):345–368, 1999.

Bruce E Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

Bruce E Hansen. Threshold autoregression in economics. *Statistics and its Interface*, 4(2):123–127, 2011.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

David V Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 1970.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 11–20, 2013.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 221–233. University of California Press, Berkeley, 1967.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. *Advances in neural information processing systems*, 29, 2016.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Dong Li and Shiqing Ling. On the least squares estimation of multiple-regime threshold autoregressive models. *Journal of Econometrics*, 167(1):240–253, 2012.

Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

Yu Lu and Harrison H. Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical computer science*, 442:13–21, 2012.

Naresh Manwani and PS Sastry. K-plane regression. *Information Sciences*, 292:39–56, 2015.

Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2000.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 93–102. IEEE, 2010.

David Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4): 919–926, 1982.

R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson, N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34:541–554, 1978.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011. doi: 10.1111/j.1467-9868.2011.00781.x.

Myung Hwan Seo. Unit root test in a threshold autoregression: asymptotic theory and residual-based block bootstrap. *Econometric Theory*, 24(6):1699–1716, 2008.

Helmuth Späth. Algorithm 39 clusterwise linear regression. *Computing*, 22(4):367–373, 1979.

George C Tiao and Ruey S Tsay. Some advances in non-linear and adaptive modelling in time-series. *Journal of forecasting*, 13(2):109–131, 1994.

Howell Tong and Keng S Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):245–268, 1980.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

Ruey S Tsay. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association*, 93(443):1188–1202, 1998.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. doi: 10.2307/1912526.

WHO. Epilepsy. `https://www.who.int/news-room/fact-sheets/detail/epilepsy`, 2026. Accessed: 2026-01-27.

Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

Chun Shan Wong and Wai Keung Li. On a mixture autoregressive model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(1):95–115, 2000.

Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

Chun Yip Yau, Chong Man Tang, and Thomas CM Lee. Estimation of multiple-regime threshold autoregressive models with structural breaks. *Journal of the American Statistical Association*, 110(511):1175–1186, 2015.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54 (4):437–447, 2001. doi: 10.1016/S0167-7152(01)00124-9.

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.

Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange: An r package for testing for structural change in linear regression models. *Journal of statistical software*, 7:1–38, 2002.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

A SYNTHETIC SIMULATIONS

A.1 SIMULATION 1: TAR MODELS

Originating from (Tong & Lim, 1980), the Threshold AutoRegressive (TAR) model is a cornerstone of non-linear time series analysis in finance and economics (Yau et al., 2015; Tsay, 1998), spanning across interest rates, stock returns, prices, output growth (Hansen, 2011). However, inference in TAR settings is notoriously difficult: the data exhibit *strong serial dependence* (violating i.i.d. assumptions), and the search for lags and thresholds suffers from the curse of dimensionality (Li & Ling, 2012).

A.1.1 SIMULATION 1.1: STANDARD TAR MODEL

We start with a standard TAR model featuring a single threshold $r = -0.5$ and time lag $d = 1$:

$$y_t = \begin{cases} -1.7 - 0.7y_{t-1} + 0.3\varepsilon_t, & y_{t-1} \leqslant -0.5, \\ 0.1 - 0.9y_{t-1} + 0.3\varepsilon_t, & y_{t-1} > -0.5. \end{cases} \tag{32}$$

The strong dependence on past information violates the i.i.d. assumption many estimators are built on, but the CoM objective below is unaffected, since it assumes nothing about the threshold r or the delay d.

$$\ell(\boldsymbol{\theta}) = \sum_t \min\Big\{[y_t - \phi_{1,0} - \phi_{1,1}y_{t-1}]^2, [y_t - \phi_{2,0} - \phi_{2,1}y_{t-1}]^2\Big\} \tag{33}$$

Table 4 compares the true parameters, the *Quasi*-GD estimates, and the oracle LSE. Furthermore, the estimated threshold derived from the *Quasi*-GD results is accurately identified as $y_{t-1} = -0.50$.

Table 4: Estimation comparison table for the standard TAR model (32).

	Mech 1			Mech 2		
	$\phi_{1,0}$	$\phi_{1,1}$	σ_1	$\phi_{2,0}$	$\phi_{2,1}$	σ_2
true para	-1.70	-0.70	0.30	0.10	-0.90	0.30
Quasi-GD estimation	-1.69	-0.69	0.30	0.10	-0.89	0.30
sd of estimations	0.02	0.02	0.00	0.01	0.02	0.00

A.1.2 SIMULATION 1.2: COMPLEX TAR MODEL

In this subsection CoM is evaluated on a four-regime TAR model with three autoregressive lags, heteroscedastic noise $(\sigma_1, \ldots, \sigma_4) = (0.1, 0.2, 0.1, 0.2)$, three thresholds $(r_1, r_2, r_3) = (-2.5, 0, 2.5)$, and delay $d = 2$. A segmentation-aware search here ranges over the three thresholds, the delay, and the lag order jointly; multiple-regime threshold models of this kind are treated in (Wong & Li, 2000; Seo, 2008).

$$y_t = \begin{cases} \mu_1 + \phi_{11}y_{t-1} + \phi_{12}y_{t-2} + \phi_{13}y_{t-3} + \sigma_1\varepsilon_t, & \text{if } y_{t-2} \leq r_1, \\ \mu_2 + \phi_{21}y_{t-1} + \phi_{22}y_{t-2} + \phi_{23}y_{t-3} + \sigma_2\varepsilon_t, & \text{if } r_1 < y_{t-2} \leq r_2, \\ \mu_3 + \phi_{31}y_{t-1} + \phi_{32}y_{t-2} + \phi_{33}y_{t-3} + \sigma_3\varepsilon_t, & \text{if } r_2 < y_{t-2} \leq r_3, \\ \mu_4 + \phi_{41}y_{t-1} + \phi_{42}y_{t-2} + \phi_{43}y_{t-3} + \sigma_4\varepsilon_t, & \text{if } y_{t-2} > r_3. \end{cases} \tag{34}$$

The CoM objective names none of the thresholds, the delay, or the regime count, and takes the form

$$\ell(\boldsymbol{\theta}) = \sum_t \min_{j\in[4]} \frac{1}{\sigma_j^2}\left[y_t - \mu_j - \sum_{k=1}^{3}\phi_{jk}y_{t-k}\right]^2. \tag{35}$$

Table 5 reports the estimates. Every autoregressive coefficient is recovered to within 0.02 and every noise scale to within 0.003; the intercepts of Mechanisms 1 and 3 are looser, 1.41 against a true 1.30 and 3.74 against 3.80. The segmentation is read off the estimate rather than searched for, giving $(\hat{r}_1, \hat{r}_2, \hat{r}_3) = (-2.5, 0, 2.6)$ against a true $(-2.5, 0, 2.5)$.

Table 5: Estimation comparison for the complex TAR model (34)

	Mech 1					Mech 2					Mech 3					Mech 4				
	μ_1	ϕ_{11}	ϕ_{12}	ϕ_{13}	σ_1	μ_2	ϕ_{21}	ϕ_{22}	ϕ_{23}	σ_2	μ_3	ϕ_{31}	ϕ_{32}	ϕ_{33}	σ_3	μ_4	ϕ_{41}	ϕ_{42}	ϕ_{43}	σ_4
true para	1.30	0.20	−0.10	0.10	0.06	−3.60	0.40	−0.20	0.15	0.08	3.80	0.20	0.25	−0.20	0.07	−3.10	0.30	0.20	−0.05	0.05
Quasi-GD	1.41	0.21	−0.09	0.08	0.06	−3.59	0.40	−0.20	0.15	0.08	3.74	0.21	0.24	−0.20	0.07	−3.10	0.30	0.20	−0.05	0.05
sd of est	0.33	0.02	0.05	0.05	0.01	0.11	0.03	0.02	0.03	0.00	0.22	0.04	0.04	0.03	0.00	0.09	0.02	0.02	0.01	0.00

Remark A.1. *In cases where the precise model specification is unknown, our method remains robust. Accurate estimation can be achieved by assuming a sufficiently large K for an alternative saturated model and incorporating appropriate* ***customized penalty term(s)****:*

$$\ell(\boldsymbol{\theta}) = \sum_t \min_{j\in J}\left\{\left[y_t - \mu_j - \sum_{k=1}^{K}\phi_{j,k}y_{t-k}\right]^2 + \lambda_j P_j(\boldsymbol{\phi}_j)\right\} + \lambda P(\boldsymbol{\phi}), \tag{36}$$

where P_j and P denote customized penalty functions. The formulation allows for $\lambda_j, \lambda = 0$, implying that unpenalized estimation for the corresponding $\boldsymbol{\phi}_j$ or $\boldsymbol{\phi}$ is a special case.

A.2 Simulation 2: MaxICA and StochMaxICA

As a substantial generalization of ICA model (Hyvärinen & Oja, 2000), MaxICA model (Guo et al., 2020) preserves the interpretability of linear formulations while enabling complex, flexible expressions of independent components derived from ICA models (Hyvärinen & Oja, 2000). The operator '$\max\{\cdot\}$' serves as a mathematical characterization of natural competition dynamics, where only the dominant signal is observed at any given time t. We consider a simplified model of the form

$$Y_t = \max_{j\in J}\{h_j Z_j(t) + c_j\}, \tag{37}$$

where $J = \{1, 2, 3\}$ and each component $Z_j(t)$ is known already, or estimated from previous steps, which follows a Fourier structure:

$$\begin{aligned}
Z_1(t) =& 7.8\sin(0.3t + 1) + 6.3\sin(1.5t + 2) + 11.7\sin(2.4t + 0.7),\\
Z_2(t) =& 4.8\sin(1.1t + 1.3) + 6.1\sin(1.8t + 2.2),\\
Z_3(t) =& 11.3\sin(0.6t + 0.6) + 13.5\sin(0.8t + 0.4) + 11.3\sin(1.7t + 1.5).
\end{aligned}$$

Remark A.2. *In contrast to structural break regressions in Section 5.1, segmentation in MaxICA is not forced by an artificial threshold. Rather, it is intrinsically split by the interplay between the generating mechanisms $\{h_j Z_j(t) + c_j\}_j$ and the observing mechanism $\max\{\cdot\}$.*

The MaxICA model provides an excellent case study for distinguishing our proposed CoM objective from DaC strategies. First, the max-operator generates frequent, irregular, and non-artificial segments. Some contain fewer than ten data points, which is too few for a reliable DaC estimate. Second, at each time point t, only the maximum value is observed, while information regarding all remaining components is lost. This results in a **missing information ratio** of $(|J| - 1)/|J|$, here $2/3$, rising toward 1 as $|J|$ grows. Inferring the global structure from data of such missing ratio is analogous to determining the geometry of a submerged iceberg solely from its tip.

A.2.1 Simulation 2.1: MaxICA Model

We now investigate the MaxICA model with an additive noise term

$$Y_t = \max_{j\in[3]}\{h_j Z_j(t) + c_j\} + \varepsilon_t. \tag{38}$$

The CoM estimate in Table 6 is obtained from the objective

$$\ell(\boldsymbol{\theta}) = \sum\nolimits_t \min_{j\in J}[Y_t - h_j Z_j(t) - c_j]^2. \tag{39}$$

Table 6: Estimation comparison for the MaxICA model (38).

	Mech 1		Mech 2		Mech 3		σ
	h_1	c_1	h_2	c_2	h_3	c_3	
true para	3.700	16.773	6.139	11.172	1.299	23.225	4.797
Quasi-GD estimation	3.632	17.932	5.925	13.168	1.268	24.051	4.057
sd of estimations	0.013	0.173	0.037	0.253	0.008	0.126	0.032

Table 7: Estimation comparison for the StochMaxICA model (40).

	Mech 1			Mech 2			Mech 3		
	h_1	c_1	σ_1	h_2	c_2	σ_2	h_3	c_3	σ_3
true para	3.700	16.773	2.797	6.139	11.172	3.752	1.299	23.225	1.834
Quasi-GD est	3.667	17.285	2.718	5.966	12.854	3.367	1.283	21.652	1.874
sd of est	0.009	0.116	0.040	0.036	0.258	0.069	0.004	0.062	0.036

A.2.2 Simulation 2.2: StochMaxICA Model

The StochMaxICA model, a variant that incorporates the noise term inside the $\max\{\cdot\}$ operator, allows for heterogeneous variances. The change is small to write down, and it moves the competition from the signals to the signals plus their noise:

$$Y_t = \max_{j\in[3]}\{h_j Z_j(t) + c_j + \varepsilon_{jt}\}. \tag{40}$$

Here, Y_t is **inseparable** from the noise, meaning that the competitive selection is driven by both signal strength and stochastic fluctuations.

The primary challenge in this setting is that the observed regime may differ from the underlying deterministic maximum. For instance, even if $\operatorname{argmax}_j\{h_j Z_j(t) + c_j\} = 1$, it is very possible that $\operatorname{argmax}_j\{h_j Z_j(t) + c_j + \sigma_j\varepsilon_t\} \neq 1$ due to the influence of $\{\sigma_j\varepsilon_t\}_j$. As illustrated by Table 7, the CoM objective (39) remains effective.

Remark A.3. *The preceding statement might suggest that the specific functional form of $Z_j(t)$ must be known a priori. This is not the case: as in Remark A.1, a saturated model with a penalty suffices, here the overcomplete Fourier basis for a sufficiently large $K \in \mathbb{N}$:*

$$\{1; \sin\omega t, \cos\omega t; \sin 2\omega t, \cos 2\omega t; \cdots; \sin K\omega t, \cos K\omega t\}. \tag{41}$$

The corresponding penalized objective is given by:

$$\ell(\boldsymbol{\theta}) = \sum\nolimits_t \min_{j\in J}\left\{\left[Y_t - \sum_{k=1}^{K}(a_{kj}\sin k\omega t + b_{kj}\cos k\omega t) - c_j\right]^2 + \lambda_j P_j\big(\{a_{kj}, b_{kj}\}_k\big)\right\}, \tag{42}$$

where P_j is recommended to be the Group Lasso penalty (Yuan & Lin, 2006)

$$P_j\big(\{a_{kj}, b_{kj}\}_k\big) := \sum\nolimits_k \sqrt{a_{kj}^2 + b_{kj}^2}. \tag{43}$$

This framework is versatile: similar penalized CoM objectives follow from alternative bases, such as wavelets or splines, depending on the characteristics of the signals.

A.3 Simulation 3: Chaotic ICA

We observe that the CoM objective function is defined as a summation of pointwise best matches. It is therefore **invariant under permutations of the observations**, rendering the sequential order of segmentation intrinsically irrelevant. Thus there might be no segmentation pattern, in which the most extreme case is randomly assigned.

To rigorously test this property, we explore a 'chaotic' setting devoid of any deterministic selection mechanism. Instead, Y_t is selected randomly from the generating mechanisms at each time with some probability, unknown to the analyst:

$$\begin{aligned} Y_t &= \operatorname*{random}_{j\in[3]}\{h_j Z_j(t) + c_j\} + \sigma\varepsilon_t; \\ \mathbb{P}(j = 1, 2, 3) &= (0.3, 0.4, 0.3). \end{aligned} \tag{44}$$

This model is intractable by boundary-specific methods, as there is no segmentation pattern. But the CoM objective (39) is unaffected since only the switching style changes. It remains efficient and accurate, as presented in Table 8. Placing the noise inside the braces instead, as in (40), leaves the same objective applicable; those estimates are not reported here.

Table 8: Estimation comparison for the chaotic MaxICA model (44).

	Mech 1		Mech 2		Mech 3		σ
	h_1	c_1	h_2	c_2	h_3	c_3	
true para	3.700	16.773	6.139	11.172	1.299	23.225	4.797
Quasi-GD estimation	3.703	16.783	6.161	11.157	1.298	23.226	4.448
sd of estimations	0.008	0.096	0.002	0.082	0.007	0.099	0.036

It is worth noting that under this random selection generation, CoM obtains more accurate estimation compared to the deterministic max-operator in Section A.2. This can be attributed to the fact that under the max rule, mechanisms producing lower values are systematically underrepresented or entirely masked by those with higher values, leaving fewer effective observations from those mechanisms and degrading their estimation quality accordingly. In contrast, the random selection probabilities $\mathbb{P}(j = 1, 2, 3) = (0.3, 0.4, 0.3)$ ensure a more balanced observing of each underlying mechanism.

A.4 Applications on General Cases

Inspired by the results of simulations above, we propose a **generalized framework for recovering mixtures of multiple unknown signals**. We assume that the observation is a realization derived from a set of latent signals $\{S_j(t)\}_{j\in J}$ via an unknown selection mechanism $R(\cdot)$:

$$Y_t = R\Big(\{S_j(t)\}_j, \varepsilon_t\Big). \tag{45}$$

Although the specific selection mechanism $R(\cdot)$ is unknown, the CoM framework posits that at any given time t, Y_t is primarily generated by one of the underlying signals $S_j(t)$. Consequently, we can bypass the explicit modeling of $R(\cdot)$ through directly minimizing the component-wise loss:

$$\ell(\boldsymbol{\theta}) = \sum_t \min_{j\in J} \; loss(Y_t, S_j(t)),$$

where the loss function is customized according to the data type. The functional forms of the signals $S_j(t)$'s are unknown, but under mild assumptions they can be approximated, for example, via Fourier expansion (or other semi-parametric or non-parametric approximations):

$$S_j(t) \approx \sum_{k\in[K_j]} (a_{kj}\sin k\omega t + b_{kj}\cos k\omega t), \quad \forall j \in J.$$

Given that the true model order K_j is typically unknown, we employ the overcomplete Fourier basis (41) with truncation $K > \max_j K_j$ and the Group Lasso penalty (43), giving (42) with the loss customized to the data type:

$$\ell(\boldsymbol{\theta}) = \sum_t \min_{j\in J}\left\{ \text{loss}\Big(Y_t, \sum_{k\in[K]} [a_{kj}\sin(k\omega t) + b_{kj}\cos(k\omega t)]\Big) + \lambda_j P_j\Big(\{a_{kj}, b_{kj}\}_{k\in[K]}\Big)\right\}.$$

B Supplementary Material for the Profiling Identity

J is finite and nonempty, $\mathcal{P} := \{P : \sqcup_{j\in J} I_j = I\}$ is the set of all assignments of I into $|J|$ disjoint, possibly empty blocks, and R is the region-wise objective defined in (3). Nothing below uses continuity, convexity, differentiability, a noise model, or a correct $|J|$: both proofs are finite bookkeeping over I and J.

One fact carries both proofs: a segmentation and an assignment map are the same object, and reading a segmentation as a map turns a double sum over blocks into a single sum over observations.

Lemma B.1 (**A segmentation is an assignment map**). *Let J^I denote the set of maps $\sigma : I \to J$. Then*

$$\sigma \longmapsto \big(\sigma^{-1}(j)\big)_{j\in J}, \qquad \sigma^{-1}(j) := \{i \in I : \sigma(i) = j\}, \tag{46}$$

is a bijection from J^I onto $\mathcal{P}$, so that $|\mathcal{P}| = |J|^{|I|}$, and for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$R\Big(\boldsymbol{\theta}; \big(\sigma^{-1}(j)\big)_{j\in J}\Big) = \sum\nolimits_{i\in I} \bar{\ell}_{\sigma(i)}\big(\boldsymbol{\theta}_{\sigma(i)}; \boldsymbol{x}_i, y_i\big). \tag{47}$$

Proof: The preimages $\sigma^{-1}(j)$, $j \in J$, are pairwise disjoint and cover I, so (46) lands in $\mathcal{P}$; empty blocks are permitted, which is exactly what allows σ to omit a value of J. Conversely, given $P \in \mathcal{P}$, each $i \in I$ lies in exactly one block, and setting $\sigma(i)$ to be that block index defines a map $I \to J$. The two constructions invert each other, and $|J^I| = |J|^{|I|}$. For (47), the double sum $\sum_{j\in J}\sum_{i\in\sigma^{-1}(j)}$ visits each $i \in I$ exactly once, at $j = \sigma(i)$, so regrouping it by i is legitimate; the sums are finite, so no convergence question arises. □

Proof of Theorem 2.1. Fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and write $\bar{\ell}_j(i) := \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, y_i)$ for brevity. Since J is nonempty, so is $\mathcal{P}$, and since J is finite, Lemma B.1 gives $|\mathcal{P}| = |J|^{|I|} < \infty$; the outer minimum in (11) is therefore taken over a finite nonempty set and is attained.

Lower bound. Let $P \in \mathcal{P}$ be arbitrary and let σ be the map Lemma B.1 associates with it. By (47) and the fact that $\sigma(i) \in J$ for every i,

$$R(\boldsymbol{\theta}; P) = \sum\nolimits_{i\in I} \bar{\ell}_{\sigma(i)}(i) \geqslant \sum\nolimits_{i\in I} \min_{j\in J} \bar{\ell}_j(i).$$

Minimizing the left-hand side over $\mathcal{P}$ preserves the inequality.

Upper bound. Because J is finite and nonempty, $\operatorname{argmin}_{j\in J} \bar{\ell}_j(i)$ is nonempty for every $i \in I$; pick $\sigma^{\boldsymbol{\theta}}(i)$ in it, arbitrarily where the minimum is attained by more than one mechanism. The segmentation $((\sigma^{\boldsymbol{\theta}})^{-1}(j))_{j\in J}$ lies in $\mathcal{P}$ and, by (47),

$$\min_{P\in\mathcal{P}} R(\boldsymbol{\theta}; P) \leqslant R\Big(\boldsymbol{\theta}; \big((\sigma^{\boldsymbol{\theta}})^{-1}(j)\big)_{j\in J}\Big) = \sum\nolimits_{i\in I} \bar{\ell}_{\sigma^{\boldsymbol{\theta}}(i)}(i) = \sum\nolimits_{i\in I} \min_{j\in J} \bar{\ell}_j(i).$$

The two bounds give the first equality in (11). The second is the definition (5) of ℓ.

Identity (11) holds at every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, so the two functions of $\boldsymbol{\theta}$ it equates are the same function. They therefore share their infimum over $\boldsymbol{\Theta}$, which is (12), and share their set of minimizers, which is the stated equivalence. □

Attainment in $\boldsymbol{\theta}$ is a separate matter, and (11) says nothing about it: equal functions share an infimum, but that infimum may be attained by neither. The blockwise hypothesis of Corollary 2.2 supplies attainment, and nothing more is needed.

Proof of Corollary 2.2. Write $m := \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\ P\in\mathcal{P}} R(\boldsymbol{\theta}; P)$ for the value of the oracle problem. Iterating the infimum and then using that $\mathcal{P}$ is finite and (11),

$$m = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \inf_{P\in\mathcal{P}} R(\boldsymbol{\theta}; P) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min_{P\in\mathcal{P}} R(\boldsymbol{\theta}; P) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}), \tag{48}$$

so the joint infimum over $\boldsymbol{\Theta} \times \mathcal{P}$ and both infima in (12) are the same number.

Attainment. Fix $P \in \mathcal{P}$. The region-wise objective separates across mechanisms,

$$R(\boldsymbol{\theta}; P) = \sum\nolimits_{j\in J} g_j(\boldsymbol{\theta}_j), \qquad g_j(\boldsymbol{\theta}_j) := \sum\nolimits_{i\in I_j} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, y_i),$$

with the j-th summand depending on $\boldsymbol{\theta}_j$ alone. Applying the hypothesis with $S = I_j$, each g_j attains its infimum over $\boldsymbol{\Theta}_j$. Since $\boldsymbol{\Theta} = \prod_{j\in J} \boldsymbol{\Theta}_j$ is a product, the blocks may be chosen independently, so the concatenation of these blockwise minimizers minimizes $R(\cdot; P)$ over $\boldsymbol{\Theta}$, and $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\boldsymbol{\theta}; P)$ exists for each of the $|J|^{|I|}$ segmentations. As $\mathcal{P}$ is finite, the smallest of these finitely many numbers is attained, say by $\boldsymbol{\theta}^\circ$ together with P°, so $m = R(\boldsymbol{\theta}^\circ; P^\circ)$. Then by (11) and (48),

$$\ell(\boldsymbol{\theta}^\circ) = \min_{P\in\mathcal{P}} R(\boldsymbol{\theta}^\circ; P) \leqslant R(\boldsymbol{\theta}^\circ; P^\circ) = m = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}) \leqslant \ell(\boldsymbol{\theta}^\circ),$$

so every inequality is an equality: $\boldsymbol{\theta}^\circ$ attains $\inf_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, hence ℓ attains its minimum on $\boldsymbol{\Theta}$ and with it both infima in (12). In particular $\operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$ is nonempty, so the remaining claims are not vacuous.

Joint optimality (14)**.** Let $\boldsymbol{\theta}^*$ be any CoM minimizer. A consistent tie-breaking is a map $\sigma^* : I \to J$ with $\sigma^*(i) \in \operatorname{argmin}_{j\in J} \bar{\ell}_j(\boldsymbol{\theta}_j^*; \boldsymbol{x}_i, y_i)$ for every i; the blocks it induces through Lemma B.1 are exactly the I_j^* of (13), which are therefore disjoint and cover I, so $P^* \in \mathcal{P}$. By (47), the defining property of σ^*, and (11),

$$R(\boldsymbol{\theta}^*; P^*) = \sum\nolimits_{i\in I} \bar{\ell}_{\sigma^*(i)}\Big(\boldsymbol{\theta}^*_{\sigma^*(i)}; \boldsymbol{x}_i, y_i\Big) = \sum\nolimits_{i\in I} \min_{j\in J} \bar{\ell}_j\big(\boldsymbol{\theta}_j^*; \boldsymbol{x}_i, y_i\big) = \ell(\boldsymbol{\theta}^*) = m,$$

the last equality by (48) and the optimality of $\boldsymbol{\theta}^*$. So $(\boldsymbol{\theta}^*, P^*)$ attains the joint minimum over $\boldsymbol{\Theta} \times \mathcal{P}$. The argument used no property of σ^* beyond selecting a pointwise minimizer, so the conclusion holds for every consistent tie-breaking, and each such choice yields a joint minimizer of the same value m.

Segment-wise form (15)**.** Fix $j_0 \in J$ and let $\boldsymbol{\theta}'_{j_0} \in \boldsymbol{\Theta}_{j_0}$ be arbitrary. Let $\boldsymbol{\theta}'$ agree with $\boldsymbol{\theta}^*$ on every block $j \neq j_0$ and carry $\boldsymbol{\theta}'_{j_0}$ on block j_0. The segmentation P^* is held fixed here, so the separation used above applies with the same blocks and all terms with $j \neq j_0$ cancel:

$$R(\boldsymbol{\theta}'; P^*) - R(\boldsymbol{\theta}^*; P^*) = \sum\nolimits_{i\in I^*_{j_0}} \bar{\ell}_{j_0}\big(\boldsymbol{\theta}'_{j_0}; \boldsymbol{x}_i, y_i\big) - \sum\nolimits_{i\in I^*_{j_0}} \bar{\ell}_{j_0}\big(\boldsymbol{\theta}^*_{j_0}; \boldsymbol{x}_i, y_i\big).$$

By (14) the left-hand side is nonnegative, since $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ and $P^* \in \mathcal{P}$. As $\boldsymbol{\theta}'_{j_0}$ was arbitrary, $\boldsymbol{\theta}^*_{j_0}$ minimizes $\sum_{i\in I^*_{j_0}} \bar{\ell}_{j_0}(\cdot; \boldsymbol{x}_i, y_i)$ over $\boldsymbol{\Theta}_{j_0}$, which is (15). Repeating for each $j_0 \in J$ completes the proof. □

Remark B.2 (**Blockwise attainment**)**.** *Quantifying over every $S \subseteq I$ is necessary, not cautionary: the blocks P^* that the optimum actually uses are known only after $\boldsymbol{\theta}^*$ has been found. Attainment on $S = I$ alone does not suffice. Take the squared loss with $f_j(\boldsymbol{x}; a, b) = ax_1 - e^b x_2$ and the three observations $((1,1),0)$, $((1,0),0)$, $((0,1),-1)$: over I the minimum is attained at $(a, e^b) = (1/3, 2/3)$, whereas over $S = \{1,2\}$ the objective $(a - e^b)^2 + a^2$ has infimum 0, approached only as $b \to -\infty$.*

C Theoretical Preliminaries

To rigorously analyze CoM, we first study its fundamental building block, the summand $\min_{j\in J} \ell_j(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$, which is intrinsically non-smooth and non-convex in $\boldsymbol{\theta}$ (except for the degenerated case $\min_{j\in J} \ell_j(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \ell_k(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$). To convey optimization techniques onto these irregular structures, we employ *Clarke's notations*.

C.1 Clarke's Notations and Pointwise Minimum

Let $x \in \mathbb{R}^d$ be a d-dimensional vector, $\Phi : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function (which automatically holds true for smooth function). *Rademacher's Theorem* asserts that Φ is differentiable almost everywhere in the sense of Lebesgue measure (Rockafellar & Wets, 1998, Theorem 9.60). However, Φ may fail to admit even one-sided directional derivative

$$\Phi'(x; v) \triangleq \lim_{\lambda\downarrow 0} \frac{\Phi(x + \lambda v) - \Phi(x)}{\lambda}, \quad \text{for direction } v \in \mathbb{R}^d$$

at some points, let alone differential. To mitigate this, we introduce instead the **Clarke directional derivative** as

$$\Phi^\circ(x; v) \triangleq \limsup_{\substack{y\to x \\ \lambda\downarrow 0}} \frac{\Phi(y + \lambda v) - \Phi(y)}{\lambda}, \quad \text{for direction } v \in \mathbb{R}^d. \tag{49}$$

By taking the upper limit, the existence of $\Phi^\circ(x; v)$ is confirmed. Since the difference quotient is limited above in light of locally Lipschitz condition, $\Phi^\circ(x; v)$ is a well-defined finite quantity. In

fact, $\Phi^\circ(x;v)$ is positively homogeneous and subadditive, which allows us to define the **Clarke subdifferential** $\partial^C\Phi(x)$, the *generalized gradient* of Φ at x, as follows:

$$\partial^C\Phi(x) \triangleq \big\{\zeta \in \mathbb{R}^d : \Phi^\circ(x;v) \geqslant \langle \zeta, v\rangle \text{ for all } v \in \mathbb{R}^d\big\}. \tag{50}$$

Then $\partial^C\Phi(x)$ is a nonempty (by *Hahn-Banach Theorem*) convex compact (by *Alaoglu's Theorem*) subset of $\mathbb{R}^d$. Further, it can be shown that knowing Φ° is equivalent to knowing $\partial^C\Phi$, since for any v,

$$\Phi^\circ(x;v) = \max\big\{\langle \zeta, v\rangle : \zeta \in \partial^C\Phi(x)\big\}. \tag{51}$$

A more detailed discussion on Clarke's notations and his ideas can be found his marvelous book *Optimization and Nonsmooth Analysis, Frank H. Clarke* Clarke (1990).

Theorem C.1. *Let J be finite, and let $\big\{\phi_j : \mathbb{R}^d \to \mathbb{R}\big\}_{j\in J}$ be a collection of locally Lipschitz functions. Define the pointwise minimum function $\Phi : \mathbb{R}^d \to \mathbb{R}$ as*

$$\Phi(x) = \min_{j\in J} \phi_j(x),$$

then we have

1. *$\Phi(x)$ is also locally Lipschitz, and its Clarke directional derivative is given by*

$$\Phi^\circ(x;v) = \max_{j\in J_{min}(x)} \phi_j^\circ(x;v), \tag{52}$$

where nonempty set $J_{min}(x)$ is the set of active function indices at x, that is,

$$J_{min}(x) \triangleq \operatorname*{argmin}_{j\in J} \phi_j(x) = \left\{j_* : \phi_{j_*}(x) = \min_{j\in J} \phi_j(x)\right\}.$$

2. *The Clarke subdifferential of $\Phi(x)$ is contained in the convex hull*

$$\begin{aligned}\partial^C\Phi(x) \subset\ &\textbf{\textit{conv}}\left\{\bigcup_{j\in J_{min}(x)} \partial^C\phi_j(x)\right\}\\ :=&\left\{\sum_{j\in J_{min}(x)} \lambda_j\partial^C\phi_j(x) : \lambda_j \geqslant 0, \sum_{j\in J_{min}(x)} \lambda_j = 1\right\}.\end{aligned}$$

If in addition, for each $j \in J_{min}(x)$, ϕ_j is strictly differentiable at x (in particular, if each ϕ_j is continuously differentiable), then the inclusion above becomes equality:

$$\partial^C\Phi(x) =\left\{\sum_{j\in J_{min}(x)} \lambda_j\partial^C\phi_j(x) : \lambda_j \geqslant 0, \sum_{j\in J_{min}(x)} \lambda_j = 1\right\}. \tag{53}$$

Interestingly, the appearance of a maximum operator in $\Phi^\circ(x;v)$ may seem counterintuitive, since $\Phi(x)$ itself is defined via a minimum. This phenomenon can be naturally explained since the limiting supremum in (49) forces a maximum symbol. In fact, by an argument analogous to Proposition 4.5.1 in Bertsekas et al. (2003), one can recover the 'expected' minimum expression in the classical directional derivative

$$\Phi'(x;v) = \min_{j\in J_{min}(x)} \phi_j'(x;v). \tag{54}$$

Yet $\Phi'(x;v)$ generally fails to exist for most families ϕ_j,[2] which implies that $\Phi(x)$ is typically **irregular**.[3] The distinction between (52) and (54) highlights precisely why Clarke's generalized differentiability framework is indispensable in our analysis.

[2] In contrast, the Clarke derivative $\Phi^\circ(x;v)$ is always well defined.

[3] A function g is said to be **regular** at x if (i) $g'(x;v)$ exists for all v and (ii) $g'(x;v) = g^\circ(x;v)$ for all v.

Proof of Theorem C.1. 1. $\Phi(x)$ is locally Lipschitz since each ϕ_j is. Since J is finite and nonempty, $\min_{j\in J}\phi_j(x)$ is a minimum over finitely many reals and is attained, so $J_{min}(x)$ is nonempty.

Consider arbitrary sequences $\{y_k\}$ and $\{\alpha_k\}$ with $y_k \to x$ and $\alpha_k \downarrow 0$. For each k, let j_k be a vector in $J_{min}(y_k)$. We now have for sufficiently small α_k,

$$\frac{\Phi(y_k+\alpha_k v)-\Phi(y_k)}{\alpha_k} = \frac{\Phi(y_k+\alpha_k v)-\phi_{j_k}(y_k)}{\alpha_k} = \frac{\phi_{j_k}(y_k+\alpha_k v)-\phi_{j_k}(y_k)}{\alpha_k},$$

where the right equality holds since locally Lipschitz implies continuity. Since $\{j_k\}$ takes values in the finite set J, some $j_* \in J$ recurs infinitely often; passing to that subsequence, $j_k = j_*$ for every k. Since $y_k \to x$, continuity of ϕ_{j_*} gives

$$\phi_{j_k}(y_k) \to \phi_{j_*}(x).$$

Also, by the continuity of Φ, we have

$$\Phi(y_k) \to \Phi(x).$$

According to the definition of j_k, it holds that $\Phi(y_k) = \phi_{j_k}(y_k)$. Thus

$$\phi_{j_*}(x) = \Phi(x),$$

from which we conclude that $j_* \in J_{min}(x)$.

By taking $k \to \infty$, limsup is achieved by the worst-case (largest) Clarke directional derivative among active function indices, that is,

$$\Phi^\circ(x;v) = \max_{j_*\in J_{min}(x)} \phi^\circ_{j_*}(x;v) = \max_{j\in J_{min}(x)} \phi^\circ_j(x;v).$$

2. We know from literature (cf., Danskin (1967), (Clarke, 1990, Proposition 2.3.12)) that for $g(x) \triangleq \max_{j\in J} g_j(x)$ with each ℓ_j locally Lipschitz, $\partial^C g(x) \subset \textbf{conv}\{\partial^C g_j(x) : j \in J_{max}(x)\}$, with $J_{max}(x)$ the set of active maximum function indices. And if g_j is regular at x for each $j \in J_{max}(x)$, the equality holds .

Let $g_j = -\phi_j$, then g_j is locally Lipschitz. And we have

$$g(x) = \max_{j\in J} -\phi_j(x) = -\min_{j\in J}\phi_j(x) = -\Phi(x).$$

In view of the identity $\partial^C(-g) = -\partial^C g$ for $\forall g$, we obtain that

$$\begin{aligned}\partial^C\Phi(x) = -\partial^C g(x) &\subset -\textbf{conv}\{\partial^C(-\phi_j)(x) : j \in J_{max}(x)\}\\ &= \textbf{conv}\{\partial^C\phi_j(x) : j \in J_{min}(x)\},\end{aligned}$$

by noticing that the maximum of $\{-\phi_j\}$ is exactly the minimum of $\{\phi_j\}$.

If ϕ_j is regular, $-\phi_j$ is generally not guaranteed regular. But if ϕ_j is strictly differentiable, so is $-\phi_j$, thus regular, which ensures the equality between $\partial^C\Phi(x)$ and the convex hull.

□

C.2 Steepest Descent

By *Hahn-Banach Theorem*, $\partial^C\Phi(x)$ is nonempty. Let $\partial^C_s\Phi(x)$ denote the element with the smallest Euclidean norm in $\partial^C\Phi(x)$, which is analogous to that in subdifferential:

$$\partial^C_s\Phi(x) \triangleq \operatorname*{argmin}_{v\in\partial^C\Phi(x)} \|v\|_2, \tag{55}$$

For a non-kink point x, $J_{min}(x)$ is a singleton, say $J_{min}(x) = \{j\}$. If in addition, ϕ_j is differentiable at x, Then the Clarke subdifferential reduces to a singleton $\{\nabla\phi_j(x)\}$ by Theorem C.1. Thus

$$\partial^C_s\Phi(x) = \nabla\phi_j(x), \quad x \text{ is a non-kink point.} \tag{56}$$

At a kink point x', however, $\partial^C\Phi(x')$ becomes a complicated convex hull (53). We then claim that the opposite direction, $-\partial^C_s\Phi(x')$, defines the **steepest descent direction**, which is the primary concern for optimization, supported by two key results:

Proposition C.2. *(Clarke, 1990, Proposition 2.3.2) For any locally Lipschitz $\Phi(x)$, if it attains a local minimum or maximum (Clarke stationary point) at x, then $0 \in \partial^C \Phi(x)$. And thus $\partial_s^C \Phi(x) = 0$ by definition.*

Theorem C.3. *For any locally Lipschitz function $\Phi(x)$, assume that $x \in \operatorname{dom} \Phi$ and $0 \notin \partial^C \Phi(x)$, then the vector $-\partial_s^C \Phi(x)$ is the direction of steepest descent at x, in the sense that*

$$\min_{\|w\|_2 \leqslant 1} \Phi^\circ(x; w) = -\left\|\partial_s^C \Phi(x)\right\|_2, \tag{57}$$

with equality holds when $w = -\partial_s^C \Phi(x) / \left\|\partial_s^C \Phi(x)\right\|_2$.

Proof: Note that for all $v \in \partial^C \Phi(x)$ and all $t \in [0, 1]$, we have

$$\left\|\partial_s^C \Phi(x) + t(v - \partial_s^C \Phi(x))\right\|_2^2 \geqslant \left\|\partial_s^C \Phi(x)\right\|_2^2.$$

Expanding the LHS, we get

$$\langle \partial_s^C \Phi(x), v - \partial_s^C \Phi(x) \rangle \geqslant 0, \quad \text{for all } v \in \partial^C \Phi(x).$$

That is, $\langle \partial_s^C \Phi(x), v \rangle \geqslant \left\|\partial_s^C \Phi(x)\right\|_2^2$ for all $v \in \partial^C \Phi(x)$, so that

$$\Phi^\circ(x; -\partial_s^C \Phi(x)) = \max_{v \in \partial^C \Phi(x)} \langle v, -\partial_s^C \Phi(x) \rangle = -\min_{v \in \partial^C \Phi(x)} \langle v, \partial_s^C \Phi(x) \rangle = -\left\|\partial_s^C \Phi(x)\right\|_2^2 < 0.$$

The last inequality holds since $0 \notin \partial^C \Phi(x)$. Thus $-\partial_s^C \Phi(x)$ is a descent direction. Note that

$$\begin{aligned} \min_{\|w\|_2 \leqslant 1} \Phi^\circ(x; w) &= \min_{\|w\|_2 \leqslant 1} \max_{v \in \partial^C \Phi(x)} \langle v, w \rangle \\ &\geqslant \max_{v \in \partial^C \Phi(x)} \min_{\|w\|_2 \leqslant 1} \langle v, w \rangle = \max_{v \in \partial^C \Phi(x)} -\|v\|_2 = -\left\|\partial_s^C \Phi(x)\right\|_2. \end{aligned}$$

We know that $-\partial_s^C \Phi(x)$ is the steepest descent direction. In fact, the equality is attained when $w = -\partial_s^C \Phi(x) / \left\|\partial_s^C \Phi(x)\right\|_2$. Further, we can obtain 'min max = max min' instead of '$\geqslant$' directly in our case owning to *Sion's Minimax Theorem*. □

So for any $x \in \operatorname{dom} \Phi$, whether it is a local extremum or not [4], $\partial_s^C \Phi(x)$ is the most informative direction in $\partial^C \Phi(x)$ and it's of primary importance for optimization purpose.

However, despite its seemingly simple definition (55), the closed form of $\partial_s^C \Phi(x')$ is fundamentally difficult to obtain. Noting that in $\Phi(x) = \min_{j \in J} \phi_j(x)$, for any kink point x', each $\partial^C \phi_j(x')$ is generally a dense set rather than a finite collection, obtaining $\partial_s^C \Phi(x')$ requires solving an inherently complex optimization problem (58) with no closed-form solution, by Theorem C.1 and Theorem C.3:

$$\min_{\lambda} \left\| \sum_{j \in J_{min}(x')} \sum_{\zeta_{jk} \in \partial^C \phi_j(x')} \lambda_{jk} \zeta_{jk} \right\|_2^2 \qquad s.t. \qquad \lambda_{jk} \geqslant 0, \quad \sum\nolimits_{j,k} \lambda_{jk} = 1. \tag{58}$$

What's worse, even when each $\phi_j(x')$ is differentiable, the computation still involves solving

$$\min_{\lambda} \left\| \sum_{j \in J_{min}(x')} \lambda_j \nabla \phi_j(x') \right\|_2^2 \qquad s.t. \qquad \lambda_j \geqslant 0, \quad \sum \lambda_j = 1, \tag{59}$$

which is a *linearly constrained quadratic program (LCQP)* problem, which in general **lacks a closed-form** solution and is left to an iterative solver except in special cases.

[4] Local maxima can be excluded by examining the neighborhood.

C.3 Proof of Closed-form Steepest Descent for Pointwise Minimum

The zero-padding in (16) directly induces a block-level one-hot structure in the gradient of each ℓ_j:

$$\nabla_{\boldsymbol{\theta}}^T \ell_j(i,\boldsymbol{\theta}) = \Big(\cdots, \mathbf{0}^T, \nabla_{\boldsymbol{\theta}_j}^T \bar{\ell}_j(i,\boldsymbol{\theta}_j), \mathbf{0}^T, \cdots\Big) = \Big(\cdots, \mathbf{0}^T, \nabla_{\boldsymbol{\theta}_j}^T \ell_j(i,\boldsymbol{\theta}), \mathbf{0}^T, \cdots\Big). \tag{60}$$

where $\mathbf{0}$ denotes zero vectors of appropriate dimensions. Consequently, the gradients $\{\nabla_{\boldsymbol{\theta}} \ell_j(i,\boldsymbol{\theta})\}_{j\in J}$ have disjoint support and are thus **mutually orthogonal** across j. This orthogonality plays a crucial role in deriving the closed-form $\partial_s^C(\min_{j\in J} \ell_j(i',\boldsymbol{\theta}))$.

Proof of Corollary 3.3. By (59), ∂_s^C at a kink index i' solves an LCQP over $\lambda_j \geqslant 0$ with $\sum_j \lambda_j = 1$, which in general admits no closed form. Orthogonality diagonalizes its objective, so the Karush–Kuhn–Tucker conditions give

$$\begin{aligned}\min_{\lambda} \left\| \sum_{j\in J_{min}(i',\boldsymbol{\theta})} \lambda_j \nabla_{\boldsymbol{\theta}} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2 &\overset{\text{orthogonal}}{=\!=} \min_{\lambda} \sum_{j\in J_{min}(i',\boldsymbol{\theta})} \lambda_j^2 \left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2 \\ &= \frac{1}{\sum_{j\in J_{min}(i',\boldsymbol{\theta})} 1/\left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}\end{aligned}$$

with the last equality holds if $\lambda_j = \frac{1/\left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta})} 1/\left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}$. Thereby we obtain $\partial_s^C(\min_{j\in J} \ell_j(i',\boldsymbol{\theta}))$ with a closed form

$$\begin{aligned}\partial_s^C \left(\min_{j\in J} \ell_j(i',\boldsymbol{\theta}) \right) &= \left(\frac{\nabla_{\boldsymbol{\theta}_j}^T \ell_j(i',\boldsymbol{\theta}) / \left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta})} 1/\left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2} \right)^T_{j\in J_{min}} \\ &\overset{(60)}{=\!=} \sum_{j\in J_{min}(i',\boldsymbol{\theta})} \frac{\nabla_{\boldsymbol{\theta}} \ell_j(i',\boldsymbol{\theta}) / \left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta})} 1/\left\| \nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) \right\|_2^2}.\end{aligned} \tag{61}$$

At a non-kink index $J_{min}(i',\boldsymbol{\theta})$ is a singleton $\{j\}$, and the right-hand side collapses to $\nabla_{\boldsymbol{\theta}} \ell_j(i',\boldsymbol{\theta})$, so (61) applies at every index without a case split.

The derivation presumes every active gradient is nonzero. If $\nabla_{\boldsymbol{\theta}_j} \ell_j(i',\boldsymbol{\theta}) = \mathbf{0}$ for some $j \in J_{min}(i',\boldsymbol{\theta})$, then $\mathbf{0}$ is one of the generators of the hull (53) and $\partial_s^C(\min_{j\in J} \ell_j(i',\boldsymbol{\theta})) = \mathbf{0}$. □

D Proof of the Sum Rule

We use four standard properties of the Clarke subdifferential of locally Lipschitz functions (Clarke, 1990):

- (C1) $\partial^C(\mu\varphi)(\boldsymbol{\theta}) = \mu\, \partial^C\varphi(\boldsymbol{\theta})$ for every $\mu \in \mathbb{R}$; in particular $\partial^C(-\varphi) = -\partial^C\varphi$.
- (C2) A continuously differentiable function is regular.
- (C3) A pointwise maximum of finitely many functions regular at $\boldsymbol{\theta}$ is regular at $\boldsymbol{\theta}$.
- (C4) If $\varphi_1, \ldots, \varphi_n$ are regular at $\boldsymbol{\theta}$, then $\sum_m \varphi_m$ is regular at $\boldsymbol{\theta}$ and $\partial^C(\sum_m \varphi_m)(\boldsymbol{\theta}) = \sum_m \partial^C\varphi_m(\boldsymbol{\theta})$, with equality rather than inclusion.

Proof of Theorem 3.4. Write $m_i(\boldsymbol{\theta}) := \min_{j\in J} \ell_j(i,\boldsymbol{\theta})$. Since $-\min_{j\in J} a_j = \max_{j\in J}(-a_j)$ for any finite family,

$$-\ell(\boldsymbol{\theta}) = \sum\nolimits_{i\in I}(-m_i(\boldsymbol{\theta})) = \sum\nolimits_{i\in I} \max_{j\in J} (-\ell_j(i,\boldsymbol{\theta})). \tag{62}$$

Each negated summand is regular: by (A2) each $\ell_j(i,\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$, so is $-\ell_j(i,\boldsymbol{\theta})$, therefore regular by (C2). By (C3) $-m_i = \max_{j\in J}(-\ell_j)$ is regular at every $\boldsymbol{\theta}$.

So the sum rule (C4) applies and gives equality,

$$\partial^C_{\boldsymbol{\theta}}(-\ell)(\boldsymbol{\theta}) = \sum\nolimits_{i\in I} \partial^C_{\boldsymbol{\theta}}(-m_i)(\boldsymbol{\theta}). \tag{63}$$

Negating back, (C1) gives $\partial^C(-\ell) = -\partial^C \ell$ and $\partial^C(-m_i) = -\partial^C m_i$. Substituting both into (63),

$$-\partial^C_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum\nolimits_{i\in I}\big(-\partial^C_{\boldsymbol{\theta}} m_i(\boldsymbol{\theta})\big) = -\sum\nolimits_{i\in I}\partial^C_{\boldsymbol{\theta}} m_i(\boldsymbol{\theta}),$$

the last step because negation distributes over a Minkowski sum. Multiplying both sides by -1 gives $\partial^C_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum_{i\in I}\partial^C_{\boldsymbol{\theta}} m_i(\boldsymbol{\theta})$, which is the claim. □

E CONSTRUCTION OF $\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$

We define $\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$ as the sum of pointwise minimum-norm elements (61)

$$\begin{aligned}
\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)}) :=& \sum_{i\in I} \partial^C_s\left(\min_{j\in J}\ell_j(i,\boldsymbol{\theta}^{(k)})\right)\\
=& \sum_{i\in \text{ non-kink points}} \partial^C_s\left(\min_{j\in J}\ell_j(i,\boldsymbol{\theta}^{(k)})\right) + \sum_{i'\in \text{ kink points}} \partial^C_s\left(\min_{j\in J}\ell_j(i',\boldsymbol{\theta}^{(k)})\right)\\
\overset{(61)}{=\!=}& \sum_{j\in J}\sum_{i\in k_j}\nabla_{\boldsymbol{\theta}}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{i'\in\text{kink points}}\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}\frac{\nabla_{\boldsymbol{\theta}}\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}\\
=& \sum_{j\in J}\sum_{i\in k_j}\nabla_{\boldsymbol{\theta}}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{j\in J}\sum_{i'\in k_{j,kink}}\frac{\nabla_{\boldsymbol{\theta}}\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}.
\end{aligned} \tag{64}$$

Noticing that at any kink point i', $\min_{j\in J}\ell_j(i',\boldsymbol{\theta}^{(k)}) = \ell_j(i',\boldsymbol{\theta}^{(k)})$ holds for any $j\in J_{min}(i',\boldsymbol{\theta}^{(k)})$, the following decomposition is inspired to be

$$\begin{aligned}
\min_{j\in J}\ell_j(i',\boldsymbol{\theta}^{(k)}) =& \sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}\frac{\min_{j\in J}\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}\\
=& \sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}\frac{\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2},
\end{aligned} \tag{65}$$

which induces a decomposition of $\ell(\boldsymbol{\theta}^{(k)})$ by

$$\begin{aligned}
\ell(\boldsymbol{\theta}^{(k)}) =& \sum_{i\in \text{ non-kink points}}\min_{j\in J}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{i'\in \text{ kink points}}\min_{j\in J}\ell_j(i',\boldsymbol{\theta}^{(k)})\\
\overset{(65)}{=\!=}& \sum_{j\in J}\sum_{i\in k_j}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{i'\in\text{kink points}}\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}\frac{\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}\\
=& \sum_{j\in J}\sum_{i\in k_j}\ell_j(i,\boldsymbol{\theta}^{(k)}) + \sum_{j\in J}\sum_{i'\in k_{j,kink}}\frac{\ell_j(i',\boldsymbol{\theta}^{(k)})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}.
\end{aligned} \tag{66}$$

Then a natural definition $\ell_j^{(k)}$ is thus introduced as

$$\ell_j^{(k)}(\boldsymbol{\theta}) \triangleq \sum_{i\in k_j}\ell_j(i,\boldsymbol{\theta}) + \sum_{i'\in k_{j,\text{kink}}}\frac{\ell_j(i',\boldsymbol{\theta})/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i',\boldsymbol{\theta}^{(k)})}1/\left\|\nabla_{\boldsymbol{\theta}_j}\ell_j(i',\boldsymbol{\theta}^{(k)})\right\|_2^2} \tag{67}$$

for any $\boldsymbol{\theta}\in\mathbb{R}^d$. The superscript k in $\ell_j^{(k)}$ indicates that it strictly follows the assignment k_j and $k_{j,kink}$ under $\boldsymbol{\theta}^{(k)}$.

F Supplementary Material for Convergence Analysis

F.1 Proof of Lemma 3.5

Lemma 3.5 collects three facts about the majorant $\sum_{j\in J}\ell_j^{(k)}$, proved here in turn.

Lemma F.1.

$$\sum_{j\in J}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}) = \ell(\boldsymbol{\theta}^{(k)}), \tag{68}$$

$$\sum_{j\in J}\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}) = \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)}). \tag{69}$$

Proof: Both identities follow from regrouping (67) by observation. At a non-kink index i exactly one j has $i \in k_j$, and $\ell_j(i,\boldsymbol{\theta}^{(k)})$ enters in full; at a kink index i' the weights over $J_{min}(i',\boldsymbol{\theta}^{(k)})$ sum to one. Either way the terms at that index contribute $\min_{j\in J}\ell_j(i,\boldsymbol{\theta}^{(k)})$, and summing over $i \in I$ gives (68). Differentiating the same regrouping at $\boldsymbol{\theta}^{(k)}$ gives (69), since $\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})$ is by definition the sum of the pointwise minimum-norm elements. It is not the minimum-norm element of $\partial^C\ell(\boldsymbol{\theta}^{(k)})$; Theorem 3.4 is what places it in that set. □

Lemma F.2.

$$\sum_{j\in J}\ell_j^{(k)}(\boldsymbol{\theta}) \geqslant \ell(\boldsymbol{\theta}), \quad \forall\boldsymbol{\theta}, k. \tag{70}$$

Proof: Since by definition, at each point $i \in I$, $\ell(\boldsymbol{\theta})$ chooses the mechanism index that minimize $\ell_j(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$ among $j \in J$. However, for the same $\boldsymbol{\theta}$ and i, $\sum_{j\in J}\ell_j^{(k)}(\boldsymbol{\theta})$ still keeps the assigned index chosen under $\boldsymbol{\theta}^{(k)}$, which can never beat $\ell(\boldsymbol{\theta})$'s choice. □

Lemma F.3.

$$\langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle = \left\|\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\right\|_2^2, \tag{71}$$

$$\sum_{j\in J}\langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle = \left\|\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\right\|_2^2 = \sum_{j\in J}\left\|\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\right\|_2^2. \tag{72}$$

Proof: Owning to (60), we notice that for any $\boldsymbol{\theta}$, orthogonality holds:

$$\nabla_{\boldsymbol{\theta}}\ell_j(\boldsymbol{\theta}) \perp \nabla_{\boldsymbol{\theta}}\ell_{j'}(\boldsymbol{\theta}), \qquad \forall j \neq j'. \tag{73}$$

Thus

$$\begin{aligned}\langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle =& \langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \sum_{j\in J}\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\rangle \\ =& \langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\rangle = \left\|\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\right\|_2^2,\end{aligned}$$

and

$$\left\|\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\right\|_2^2 = \left\|\sum_{j\in J}\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\right\|_2^2 = \sum_{j\in J}\left\|\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)})\right\|_2^2,$$

and

$$\begin{aligned}\sum_{j\in J}\langle\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle =& \langle\sum_{j\in J}\nabla_{\boldsymbol{\theta}}\ell_j^{(k)}(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle \\ =& \langle\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)}), \widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\rangle = \left\|\widetilde{\nabla}_s\ell(\boldsymbol{\theta}^{(k)})\right\|_2^2.\end{aligned}$$

□

F.2 *Quasi*-GD Algorithm and Its Convergence

Proof of Lemma 3.6. Write $g := \widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$ and let $m^{(k)} := \sum_{j\in J} \ell_j^{(k)}$ denote the majorant.

Exchanging the order of summation in (67) regroups the majorant by observation,

$$m^{(k)}(\boldsymbol{\theta}) = \sum\nolimits_{i\in I} \sum\nolimits_{j\in J_{min}(i,\boldsymbol{\theta}^{(k)})} w_{ij}^{(k)}\, \ell_j(i,\boldsymbol{\theta}), \qquad w_{ij}^{(k)} := \frac{1/\left\|\nabla_{\boldsymbol{\theta}_j} \ell_j(i,\boldsymbol{\theta}^{(k)})\right\|_2^2}{\sum_{j\in J_{min}(i,\boldsymbol{\theta}^{(k)})} 1/\left\|\nabla_{\boldsymbol{\theta}_j} \ell_j(i,\boldsymbol{\theta}^{(k)})\right\|_2^2}, \tag{74}$$

a nonnegative combination of maps that are L_0-smooth by (A2); at a non-kink index the inner sum is the single term $w_{ij}^{(k)} = 1$. Lipschitz moduli add rather than maximize under such a combination: if each ∇f_m is L_0-Lipschitz and $c_m \geqslant 0$, then $f = \sum_m c_m f_m$ satisfies

$$\|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\|_2 = \left\|\sum\nolimits_m c_m(\nabla f_m(\boldsymbol{u}) - \nabla f_m(\boldsymbol{v}))\right\|_2 \leqslant \Big(\sum\nolimits_m c_m\Big) L_0 \|\boldsymbol{u} - \boldsymbol{v}\|_2 .$$

The weights in (74) sum to one at each i, so their total mass is $\sum_{i\in I} 1 = |I|$ and $m^{(k)}$ is $|I|L_0 = L$-smooth. The count is over observations, not mechanisms, so no factor of $|J|$ enters and L is the constant fixed in (A2), the same at every iterate.

By (69), $\nabla_{\boldsymbol{\theta}} m^{(k)}(\boldsymbol{\theta}^{(k)}) = g$, so the rule (25) reads $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha \nabla_{\boldsymbol{\theta}} m^{(k)}(\boldsymbol{\theta}^{(k)})$: an ordinary gradient step on an L-smooth function, for which the classical descent lemma (Wright & Recht, 2022, Chapter 3) applies verbatim.

That lemma, with $m^{(k)}(\boldsymbol{\theta}^{(k)}) = \ell(\boldsymbol{\theta}^{(k)})$ from (68), gives

$$m^{(k)}(\boldsymbol{\theta}^{(k+1)}) \leqslant \ell(\boldsymbol{\theta}^{(k)}) - \alpha \|g\|_2^2 + \frac{L\alpha^2}{2}\|g\|_2^2 = \ell(\boldsymbol{\theta}^{(k)}) - \alpha\left(1 - \frac{L\alpha}{2}\right)\|g\|_2^2 \leqslant \ell(\boldsymbol{\theta}^{(k)}) - \frac{\alpha}{2}\|g\|_2^2,$$

the last step because $\alpha \leqslant 1/L$ keeps the bracket at least $\frac{1}{2}$. Majorization (70) supplies $\ell(\boldsymbol{\theta}^{(k+1)}) \leqslant m^{(k)}(\boldsymbol{\theta}^{(k+1)})$, which is the first inequality of (30). □

Proof of Theorem 3.7. Summing Lemma 3.6 over $k = 0, \ldots, K$ telescopes the objective values,

$$\frac{\alpha}{2}\sum\nolimits_{k=0}^{K} \left\|\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})\right\|_2^2 \;\leqslant\; \ell(\boldsymbol{\theta}^{(0)}) - \ell(\boldsymbol{\theta}^{(K+1)}) \;\leqslant\; \ell(\boldsymbol{\theta}^{(0)}) - \ell^\star,$$

finite by (A3). The smallest of the $K+1$ summands is at most their average, which is the stated bound. □

Proof of Corollary 3.8. Theorem 3.4 places $\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})$ in $\partial^C \ell(\boldsymbol{\theta}^{(k)})$, so $dist(0, \partial^C \ell(\boldsymbol{\theta}^{(k)})) \leqslant \left\|\widetilde{\nabla}_s \ell(\boldsymbol{\theta}^{(k)})\right\|_2$ at every iterate. By Theorem 3.7 the right-hand side falls below ε once $2\big(\ell(\boldsymbol{\theta}^{(0)}) - \ell^\star\big)/[\alpha(K+1)] < \varepsilon^2$, that is after $\mathcal{O}(\varepsilon^{-2})$ iterations. □

F.3 Computational Complexity Analysis

Let $dim(\theta) = \sum_{j\in J} dim(\theta_j) = |J|\bar{d}$, with $\bar{d}$ the average length of parameters.

Per-Iteration Complexity: The time complexity for a single iteration is bounded by $\mathcal{O}(|J|\bar{d} \cdot |I|)$.

- Calculating $\ell_j(\theta^{(k)}; \boldsymbol{x}_i, y_i)$ for all $j \in J$ and $i \in I$ requires $\mathcal{O}(|J|\bar{d} \cdot |I|)$ operations.
- Updating the segmentation $k_j, k_{j,kink}$ requires $\mathcal{O}(|J| \cdot |I|)$ operations.
- Quasi-GD update requires $\mathcal{O}(\bar{d} \cdot |I|)$ operations.

Overall Time Complexity: To achieve an ε-approximate stationary point, CoM requires $\mathcal{O}(\frac{1}{\alpha\varepsilon})$ total iterations (Corollary3.8). Therefore, the overall computational complexity of the algorithm is $\mathcal{O}(\frac{|J|\bar{d}\cdot|I|}{\alpha\varepsilon})$. If $\alpha = \frac{1}{L}$, the overall complexity would be $\mathcal{O}(\frac{|I|\cdot|J|\bar{d}L}{\varepsilon})$, which is a $\mathcal{O}(|I|)$ linear scaling.

G Characterizing Spurious Local Optima

The convergence rates established for CoM above are non-asymptotic, what the optimization literature calls 'global' convergence, but the word describes the starting point rather than the destination: *Quasi*-GD ensures approximation to a critical point only, and like any method on a non-convex objective it cannot promise a global optimum. What it leaves open is a gap between computation and statistics Bartholomew et al. (2011), since most existing inference for parametric models is predicated strictly on the global optimum.

Bridging this gap has proven difficult. Jin et al. (2016) proved that even for a K-component GMM, among the simplest of multi-mechanism models, classical EM with random initialization converges to a bad critical point with probability at least $1-e^{-\Omega(K)}$. To compensate, practitioners run the algorithm repeatedly from diverse random initializations to obtain a nominally robust solution McLachlan & Peel (2000), multiplying the computational cost by the number of restarts.

Strategic initialization spends the effort before optimization instead, on spatially distributed seeding such as the K-Means++ algorithm Arthur & Vassilvitskii (2006), or on pilot estimators based on the method of moments Moitra & Valiant (2010); Hsu & Kakade (2013). Each of them acts on the start, and none of them tells whether the estimate finally returned is globally optimal. Since the start cannot be certified, Theorem 4.1 certifies the outcome instead, through a necessary condition for global optimality.

Proof of Theorem 4.1. Since $\boldsymbol{\theta}_j^{**} := \operatorname{argmin}_{\boldsymbol{\theta}_j} \sum_{i\in I_j^*} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, y_i)$ for every $j \in J$, minimizing block by block is equivalent to minimizing the sum jointly,

$$\boldsymbol{\theta}^{**} := \operatorname{argmin}_{\boldsymbol{\theta}} \sum\nolimits_{j\in J} \sum\nolimits_{i\in I_j^*} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, y_i), \tag{75}$$

so $\boldsymbol{\theta}^{**}$ minimizes $R(\cdot; P^*)$ over $\boldsymbol{\Theta}$. Three comparisons then chain:

$$\ell(\boldsymbol{\theta}^{**}) \;\leqslant\; R(\boldsymbol{\theta}^{**}; P^*) \;\leqslant\; R(\boldsymbol{\theta}^{*}; P^*) \;=\; \ell(\boldsymbol{\theta}^{*}). \tag{76}$$

The first holds because $\ell = \min_{P\in\mathcal{P}} R(\cdot; P)$ by (11); the second because $\boldsymbol{\theta}^{**}$ minimizes that objective; the third because P^* is the segmentation $\boldsymbol{\theta}^*$ induces, so the minimum in (11) is attained at P^*. Since $\boldsymbol{\theta}^*$ is a global minimizer of ℓ, the left end is no smaller than the right, and every relation in (76) is an equality. In particular $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^{*})$.

□

Proposition G.1 (A unique refit turns the value test into a vector test). *Let $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ induce P^* through* (13)*, and suppose that for every $j \in J$ the minimizer of $\sum_{i\in I_j^*} \bar{\ell}_j(\cdot; \boldsymbol{x}_i, y_i)$ over $\boldsymbol{\Theta}_j$ is unique. Then*

$$\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^{*}) \quad\Longleftrightarrow\quad \boldsymbol{\theta}^{**} = \boldsymbol{\theta}^{*}.$$

Proof: One direction is immediate. For the other, $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^{*})$ collapses (76) to equalities, so $\boldsymbol{\theta}^*$ attains $\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}; P^*)$ alongside $\boldsymbol{\theta}^{**}$. That minimization separates across blocks, so $\boldsymbol{\theta}_j^*$ and $\boldsymbol{\theta}_j^{**}$ both attain $\min_{\boldsymbol{\theta}_j} \sum_{i\in I_j^*} \bar{\ell}_j(\boldsymbol{\theta}_j; \boldsymbol{x}_i, y_i)$, and uniqueness identifies them.

One might suspect that $\boldsymbol{\theta}^{**}$ and $\boldsymbol{\theta}^*$ agree only up to a rotational equivalence or a permutation of labels. They cannot. Definition (13) reads I_j^* off mechanism j at $\boldsymbol{\theta}_j^*$, and $\boldsymbol{\theta}_j^{**}$ is refitted on that same I_j^*, so $\boldsymbol{\theta}_j^* \mapsto I_j^* \mapsto \boldsymbol{\theta}_j^{**}$ never leaves block j. □

Uniqueness holds under the squared loss whenever each I_j^* carries a full-rank design, as in Table 1. It fails for the redescending losses admitted in Section 2, and for an empty I_j^*, where every $\boldsymbol{\theta}_j$ minimizes an empty sum; there only the value form survives.

Proof of Proposition 4.2. Passing the test means $\ell(\boldsymbol{\theta}^{**}) = \ell(\boldsymbol{\theta}^{*})$, which forces every relation in (76) to be an equality, so $\boldsymbol{\theta}^*$ attains $\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}; P^*)$. That minimization separates across blocks, so $\boldsymbol{\theta}_j^*$ minimizes $\sum_{i\in I_j^*} \bar{\ell}_j(\cdot; \boldsymbol{x}_i, y_i) = \sum_{i\in I_j^*} \ell_j(\cdot; \boldsymbol{x}_i, y_i)$ over the whole of $\boldsymbol{\Theta}_j$, which Section 3 leaves

unrestricted, and which is differentiable by (A2). The first-order condition holds with nothing further assumed:

$$\sum\nolimits_{i\in I_j^*} \nabla_{\boldsymbol{\theta}_j} \ell_j(\boldsymbol{\theta}_j^*; \boldsymbol{x}_i, y_i) = 0, \qquad j \in J. \tag{77}$$

Write $\sigma^*(i)$ for the mechanism to which P^* assigns i, so that $I_j^* = \{i \in I : \sigma^*(i) = j\}$, and collect the selected gradients into

$$\boldsymbol{v} := \sum\nolimits_{i\in I} \nabla_{\boldsymbol{\theta}} \ell_{\sigma^*(i)}(i, \boldsymbol{\theta}^*).$$

Each $\nabla_{\boldsymbol{\theta}} \ell_j(i, \boldsymbol{\theta}^*)$ is supported on block j alone by the one-hot structure (60), so $\boldsymbol{v}$ splits blockwise and its j-th block is exactly the left side of (77):

$$\boldsymbol{v} = \left(\sum\nolimits_{i\in I_j^*} \nabla_{\boldsymbol{\theta}_j} \ell_j(\boldsymbol{\theta}_j^*; \boldsymbol{x}_i, y_i) \right)_{j\in J} = 0.$$

It remains to place $\boldsymbol{v}$ inside $\partial^C \ell(\boldsymbol{\theta}^*)$. Definition (13) assigns i only to a mechanism attaining the pointwise minimum, so $\sigma^*(i) \in J_{min}(i, \boldsymbol{\theta}^*)$ for every $i \in I$, and Theorem 3.2 puts $\nabla_{\boldsymbol{\theta}} \ell_{\sigma^*(i)}(i, \boldsymbol{\theta}^*)$ in $\partial^C(\min_{j\in J} \ell_j(i, \boldsymbol{\theta}^*))$ as a vertex of that convex hull. Summing over I and applying Theorem 3.4 gives $0 = \boldsymbol{v} \in \partial^C \ell(\boldsymbol{\theta}^*)$.

The argument does not require $\boldsymbol{\theta}^*$ to be a *Quasi*-GD output: only (13) and (77) enter it, and neither records how $\boldsymbol{\theta}^*$ was produced. The conclusion therefore holds for any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ passing the test, on the same provenance-indifferent footing as Theorem 4.1. □