Proceedings of Machine Learning Research vol :1–42, 2026 39th Annual Conference on Learning Theory

Convergence of Continual Learning in Homogeneous Deep Networks

Matan Schliserman SCHLISERMAN@MAIL.TAU.AC.IL
Blavatnik School of Computer Science and AI, Tel Aviv University

Gon Buzaglo GON.BUZAGLO@PRINCETON.EDU
Princeton University

Itay Evron ITAY@EVRON.ME

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Department of Electrical and Computing Engineering, Technion

Editors: Steve Hanneke and Tor Lattimore

Abstract

We characterize weakly regularized continual classification in homogeneous models as sequential projections onto task margin sets. This result generalizes prior analyses restricted to either stationary (single-task) deep models or continual linear models. We show that global convergence generally fails, even for simple models linear in data but nonlinear in parameters. Nevertheless, by leveraging results from nonconvex projection theory, we identify regularity properties of homogeneous deep networks that guarantee local linear convergence under random and cyclic task sequences. Finally, we extend our analysis to continual regression, unifying the framework for homogeneous models.
Keywords: Continual learning, Lifelong learning, Algorithmic bias, Projection algorithms.

1. Introduction

Continual learning focuses on training models on a sequence of tasks to incrementally accumulate expertise. This paradigm has gained traction with large-scale foundation models, where full retraining is often prohibitive due to computational costs, privacy regulations, or data retention constraints. In these settings, models must be finetuned sequentially as data arrives, while avoiding catastrophic forgetting (Robins, 1995; van de Ven et al., 2025) and maintaining plasticity for future learning (Dohare et al., 2024). Alongside the development of practical algorithms (*e.g.,* Qiao et al. 2024; Behrouz et al. 2025; see Yang et al. 2025 for a survey), deepening our theoretical understanding of the underlying dynamics remains essential. We address this by developing analytical tools specifically for continual learning in homogeneous deep neural networks (DNNs).

Theoretical literature has largely centered on continual linear regression due to its tractability, shedding light on critical aspects of the field: convergence under various task orderings (Evron et al., 2022, 2026; Kong et al., 2023; Attia et al., 2025), task similarity (e.g., Asanuma et al., 2021; Lin et al., 2023; Li and Hiratani, 2025; Tsipory et al., 2025), regularization (e.g., Zhao et al., 2024; Levinstein et al., 2025; Karpel et al., 2026), overparameterization (Goldfarb and Hand, 2023; Goldfarb et al., 2024), and algorithmic effects (Doan et al., 2021; Peng and Risteski, 2022; Peng et al., 2023). Other notable works study continual linear classification, primarily deriving convergence guarantees under explicit or implicit regularization (Evron et al., 2023; Jung et al., 2025). While illuminating, linear analysis cannot fully account for the complex, nonlinear dynamics of modern DNNs.

Some recent works analyze continual learning in *nonlinear* models, largely focusing on simplified settings, such as two-layer models that enable precise analyses (Lee et al., 2021, 2022; Li et al., 2025; Taheri et al., 2025). Others leverage perturbative, mean-field, or scaling-limit techniques, including regimes that interpolate between lazy and feature-learning dynamics, but are typically restricted to shallow architectures or infinite-width limits (Shan et al., 2024; Graldi et al., 2025). As a result, depth influences forgetting primarily through static feature statistics or last-layer adaptation. In contrast, our work analyzes continual learning in homogeneous deep networks of arbitrary depth, with all layers trained and without relying on lazy-training or infinite-width assumptions, thereby capturing fully nonlinear feature evolution across layers.

A popular approach to stabilizing continual models and preventing forgetting is the use of explicit regularization in the parameter space (e.g., Kirkpatrick et al., 2017). These methods aim to bias the optimization toward previous iterates via:

$$\Theta_t^{(\lambda)} \leftarrow \operatorname{argmin}_{\Theta}\left\{\mathcal{L}_t(\Theta) + \lambda\left\|\Theta - \Theta_{t-1}^{(\lambda)}\right\|_{\mathbf{B}_t}^2\right\}, \tag{1}$$

where $\mathbf{B}_t \succeq 0$ skews the regularization toward directions deemed "important" for previous tasks—*e.g.,* based on their Fisher information (see Benzing, 2022). Interestingly, even isotropic regularization ($\mathbf{B}_t = \mathbf{I}$) has proven beneficial both practically (e.g., Lubana et al., 2021; Smith et al., 2023) and theoretically (Li et al., 2023; Levinstein et al., 2025; Karpel et al., 2026). Our work adopts a standard analytical framework with *weak* isotropic regularization in the limit as $\lambda \downarrow 0$, as detailed next.

The limit of *weak* regularization is an influential regime for analyzing algorithmic biases—extending beyond continual learning—due to its tractability and connections to minimum-norm and max-margin solutions (e.g., Hastie et al., 2022; Aubin et al., 2020). Specifically, in "traditional" stationary classification, Rosset et al. (2004) proved that weakly regularized *linear* models converge[1] to their max-margin counterparts. That is, linear models trained with margin-based losses—*e.g.,* the logistic loss—on separable data recover the Hard-Margin SVM solution as regularization vanishes:

$$\boldsymbol{\theta}^{(\lambda)} = \operatorname*{argmin}_{\boldsymbol{\theta}}\left\{\mathcal{L}(\boldsymbol{\theta}) + \lambda\left\|\boldsymbol{\theta}\right\|^2\right\} \xrightarrow[\lambda\downarrow 0]{\text{in direction}} \operatorname*{argmax}_{\|\bar{\boldsymbol{\theta}}\|\le 1}\min_i y_i \bar{\boldsymbol{\theta}}^\top \mathbf{x}_i = \operatorname*{argmin}_{y_i\bar{\boldsymbol{\theta}}^\top\mathbf{x}_i \ge 1,\, \forall i}\|\bar{\boldsymbol{\theta}}\|. \tag{2}$$

A later work by Wei et al. (2019) generalized this to DNNs, showing that the normalized margin of a weakly-regularized homogeneous model f converges to the max margin:

$$\min_i y_i f(\mathbf{x}_i; \tfrac{\Theta^{(\lambda)}}{\|\Theta^{(\lambda)}\|}) \xrightarrow[\lambda\downarrow 0]{} \max_{\|\Theta\|\le 1}\min_i y_i f(\mathbf{x}_i;\Theta). \tag{3}$$

In continual learning, Evron et al. (2023) used weak regularization to study jointly-separable linear classification. They showed an equivalence between the iterates of weakly regularized continual linear classification and those of a sequential margin-separating projection algorithm. That is, defining a convex "margin set" $\mathcal{C}_t \triangleq \{\bar{\boldsymbol{\theta}} \mid y_i^{(t)}\bar{\boldsymbol{\theta}}^\top\mathbf{x}_i^{(t)} \ge 1,\ \forall i\}$, they showed:

$$\left(\boldsymbol{\theta}_t^{(\lambda)} = \operatorname*{argmin}_{\boldsymbol{\theta}}\left\{\mathcal{L}_t(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}^{(\lambda)}\|^2\right\}\right)_t \xrightarrow[\lambda\downarrow 0]{\text{in direction}} \left(\bar{\boldsymbol{\theta}}_t = \operatorname*{argmin}_{\bar{\boldsymbol{\theta}}\in\mathcal{C}_t}\left\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{t-1}\right\|^2\right)_t. \tag{4}$$

We complete the analysis landscape by proving that weakly-regularized continual learning with homogeneous DNNs projects onto *nonconvex* margin sets, $\mathcal{C}_t \triangleq \{\bar{\Theta} \mid y_i^{(t)} f(\mathbf{x}_i^{(t)};\bar{\Theta}) \ge 1,\ \forall i\}$:

$$\left(\Theta_t^{(\lambda)} = \operatorname*{argmin}_{\Theta}\left\{\mathcal{L}_t(\Theta) + \lambda\|\Theta - \Theta_{t-1}^{(\lambda)}\|^2\right\}\right)_t \xrightarrow[\lambda\downarrow 0]{\text{in direction}} \left(\bar{\Theta}_t \in \operatorname*{argmin}_{\bar{\Theta}\in\mathcal{C}_t}\left\|\bar{\Theta} - \bar{\Theta}_{t-1}\right\|^2\right)_t. \tag{5}$$

Table 1: Landscape of weak-regularization analysis across learning regimes and model classes.

	Linear models	**Homogeneous DNNs**
Stationary setting	Rosset et al. (2004)	Wei et al. (2019)
Continual learning	Evron et al. (2023)	Our work (2026)

After establishing an equivalence between continual classification in homogeneous models and sequential projections, we first show that, unlike previous results for linear models, global convergence is generally not guaranteed in homogeneous models. Encouragingly, we prove *local* convergence guarantees for these models through tools from the literature on nonconvex projections (e.g., Lewis and Malick, 2008; Dao and Phan, 2019). Lastly, we demonstrate the generality of these tools by extending the analysis from classification to regression. Overall, these tools provide a principled framework for understanding continual learning in deep neural networks.

Summary of contributions. The main contributions of this paper are:

1. We prove that weakly-regularized continual learning with homogeneous DNNs implicitly performs sequential margin projections.
2. We show that, unlike in prior work on linear models, these projection sets are not necessarily convex, leading to qualitative differences between settings.
3. We demonstrate that global convergence is not guaranteed; consequently, forgetting can be *catastrophic*, even for a simple homogeneous model with only 4 parameters and 2 tasks.
4. Bridging to the nonconvex projection literature, we establish local convergence guarantees for random and cyclic task orderings, with rates depending on model Lipschitzness and smoothness.
5. Finally, we extend our results to continual regression in homogeneous models, providing a unified framework for both classification and regression.

2. Continual Classification in Homogeneous DNNs as Sequential Projections

Notation. We denote $[n] \triangleq \{1, 2, \dots, n\}$, $[z]_+ \triangleq \max\{0, z\}$, and $\mathcal{B}_\delta(\mathbf{u}) \triangleq \{\mathbf{v} \mid \|\mathbf{v} - \mathbf{u}\| \leq \delta\}$. We call a function $f : \mathbb{R}^p \to \mathbb{R}$ *positively homogeneous* of degree $r \in \mathbb{R}$ if $f(c\mathbf{z}) = c^r f(\mathbf{z}), \forall c > 0$.

Models. Throughout the paper we consider r-positively-homogeneous models parametrized by weights $\Theta \in \mathbb{R}^p$. For example, a fully connected neural network with $L \geq 1$ layers, no bias terms, and a positively homogeneous activation function of degree 1 (such as ReLU or leaky ReLU) is L-positively-homogeneous in its parameters $\Theta = (\mathbf{W}_1, \dots, \mathbf{W}_L)$:

$$f(\mathbf{x}; \Theta) = \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \sigma(\cdots \sigma(\mathbf{W}_1 \mathbf{x}))) \quad \Longrightarrow \quad f(\mathbf{x}; c\Theta) = c^L f(\cdot; \Theta), \quad \forall c > 0 \,. \tag{6}$$

In this section, we focus on binary classification to establish our fundamental result: weakly regularized homogeneous models in continual learning converge to a sequential projection algorithm. We extend these results to regression in Section 4. In both settings, the projection perspective bridges our analysis to existing literature on nonconvex projections, facilitating rigorous convergence results.

1. In this introduction, we say that $\boldsymbol{\theta}^{(\lambda)}$ converges in direction to $\bar{\boldsymbol{\theta}}$ if $\lim_{\lambda \downarrow 0} \frac{\boldsymbol{\theta}^{(\lambda)}}{\|\boldsymbol{\theta}^{(\lambda)}\|} = \frac{\bar{\boldsymbol{\theta}}}{\|\bar{\boldsymbol{\theta}}\|}$.

2.1. Setting: Continual Classification

Learning setup. We consider a setup with M classification tasks $\left(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}\right), \ldots, \left(\mathbf{X}^{(M)}, \mathbf{y}^{(M)}\right)$ consisting of feature matrices $\mathbf{X}^{(m)} \in \mathbb{R}^{n_m \times d}$ and binary labels $\mathbf{y}^{(m)} \in \{-1, 1\}^{n_m}$. The learner is exposed to tasks sequentially in $k \in \mathbb{N}^+$ iterations according to a *task ordering* $\tau : [k] \to [M]$, which are commonly assumed to be cyclic or random, to capture adversarial vs. nonadversarial behaviors (Evron et al., 2022, 2023; Kong et al., 2023; Levinstein et al., 2025). That is, for iteration $t \in [k]$,

$$\tau_{\text{cyc}}(t) = ((t-1) \bmod M) + 1, \qquad \tau_{\text{iid}}(t) \overset{\text{i.i.d.}}{\sim} \text{Unif}\{1, \ldots, M\}\,. \tag{7}$$

Our analysis assumes the model can separate all M training sets. This assumption follows prior theoretical work on continual linear classification (Evron et al., 2023; Jung et al., 2025), where it is justified by standard teacher assumptions or high-dimensional randomness (e.g., Cover, 1965). For the deeper architectures considered here, joint separability is a significantly milder requirement; sufficiently expressive networks can interpolate finite datasets under general conditions, a phenomenon rooted in universal approximation results (Cybenko, 1989).

Assumption A (Joint Separability) *Define the individual feasible sets*

$$\mathcal{C}_m \triangleq \left\{ \bar{\Theta} \;\middle|\; y_i^{(m)} f(\mathbf{x}_i^{(m)}; \bar{\Theta}) \geq 1,\ \forall i \in [n_m] \right\}, \qquad \forall m \in [M]\ . \tag{8}$$

We assume the intersection, i.e., the joint feasible set, is nonempty: $\mathcal{C}^\star \triangleq \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$.

Metrics. We follow the projection analysis for continual classification in Evron et al. (2023) and derive results for three quantities of interest:

1. **Distance to joint feasible set:** $d(\bar{\Theta}, \mathcal{C}^\star) = \min_{\bar{\Theta}^\star \in \mathcal{C}^\star} \|\bar{\Theta} - \bar{\Theta}^\star\|$.
2. **Distance to individual feasible set:** $\forall m \in [M]\,,\ d(\bar{\Theta}, \mathcal{C}_m) = \min_{\bar{\Theta}' \in \mathcal{C}_m} \|\bar{\Theta} - \bar{\Theta}'\|$.
3. **Forgetting:** At iteration t, we define the forgetting of a task previously seen at iteration $t' \leq t$ as the maximal hinge loss over its training points: $F_{\tau(t')}(\bar{\Theta}_t) = \max_i \left[1 - y_i f(\mathbf{x}_i^{\tau(t')}; \bar{\Theta}_t)\right]_+$.

As noted by Evron et al. (2023), the hinge loss possesses favorable properties for analyzing continual learning. First, we will show that immediately after learning task m, the iterate resides in the feasible set, $\bar{\Theta} \in \mathcal{C}_m$; consequently, the forgetting of the most recent task is zero. More importantly, lower forgetting—quantified here by a lower hinge loss—implies improved training margins, which are often tied to better generalization.

2.2. Fundamental Result: Continual Classification to Sequential Margin Projections

Learning algorithm. We consider a learner that minimizes the margin-based logistic loss,[2] defined for each task $m \in [M]$ as

$$\mathcal{L}_m(\Theta) \triangleq \frac{1}{n_m} \sum_{i=1}^{n_m} \log\left(1 + \exp(-\,y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta))\right). \tag{9}$$

2. While we measure forgetting using the hinge loss—which can reach zero once the margin is sufficiently large—the learner optimizes the logistic loss. This choice aligns our model with standard deep learning practices and is made for simplicity (*i.e.,* our analysis can be extended to other common loss functions).

In Algorithm 1 we adopt a common strategy in continual learning by regularizing parameter shifts to prevent catastrophic forgetting. Such methods effectively trade off model plasticity for increased stability. While several prominent works utilize weighted L_2 regularization (often guided by Fisher information) to protect task-specific parameters (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Benzing, 2022), we focus on isotropic L_2 regularization. This variant offers a lower memory footprint and greater analytical tractability while remaining empirically competitive (Hsu et al., 2018; Lubana et al., 2021; Smith et al., 2023). Furthermore, isotropic regularization has become a focal point of recent theoretical inquiries into model consolidation (e.g., Li et al., 2023; Levinstein et al., 2025; Karpel et al., 2026).

Formally, in Algorithm 1 we study the regularization at the limit as $\lambda \downarrow 0$ and show that its iterates converge to the sequential projections of Algorithm 2.

Algorithm 1 Regularized Continual Learning

Initialization: $\Theta_0^{(\lambda)} = \mathbf{0}$
For each iteration $t = 1, \ldots, k$:
$\Theta_t^{(\lambda)} \leftarrow \underset{\Theta}{\operatorname{argmin}} \left\{ \mathcal{L}_{\tau(t)}(\Theta) + \lambda \left\| \Theta - \Theta_{t-1}^{(\lambda)} \right\|^2 \right\}$
Output: $\Theta_k^{(\lambda)}$

Algorithm 2 Sequential Margin Projections

Initialization: $\bar{\Theta}_0 = \mathbf{0}$
For each iteration $t = 1, \ldots, k$:
$\bar{\Theta}_t \leftarrow \mathbf{\Pi}_{\tau(t)}(\bar{\Theta}_{t-1}) \triangleq \underset{\bar{\Theta} \in \mathcal{C}_{\tau(t)}}{\operatorname{argmin}} \left\| \bar{\Theta} - \bar{\Theta}_{t-1} \right\|^2$
Output: $\bar{\Theta}_k$

Theorem 1 (Weakly-Regularized CL $\to$ Sequential Margin Projections) *Consider an r-positively-homogeneous model $f(\cdot;\Theta) : \mathcal{X} \to \mathbb{R}$, i.e., $f(x; c\Theta) = c^r f(x;\Theta)$ for all $c > 0$. Assume individual separability, i.e., nonempty feasible sets $\mathcal{C}_1, \ldots, \mathcal{C}_M$ (implied by Assumption A). Then, as $\lambda \downarrow 0$, Algorithm 1 trained with the logistic loss aligns with Algorithm 2. That is, for every iteration $t \in [k]$, any sequence $\lambda \downarrow 0$ admits a subsequence (λ_j) and a point $\bar{\Theta}_t \in \mathbf{\Pi}_{\tau(t)}(\bar{\Theta}_{t-1}) \triangleq \operatorname{argmin}_{\bar{\Theta} \in \mathcal{C}_{\tau(t)}} \left\| \bar{\Theta} - \bar{\Theta}_{t-1} \right\|^2$ such that*

$$\frac{\Theta_t^{(\lambda_j)}}{\left\|\Theta_t^{(\lambda_j)}\right\|} \xrightarrow{j \to \infty} \frac{\bar{\Theta}_t}{\left\|\bar{\Theta}_t\right\|}.$$

The full proof is provided in Appendix A. Below, we describe our proof techniques, contrast them with prior work, and outline the core ideas behind the derivation.

Comparison to prior work. As discussed in the introduction, the limit of weak regularization is a standard analytical tool. In stationary classification—equivalent to the *first* iteration of Algorithms 1 and 2—weakly-regularized solutions have been linked to max-margin solutions in both linear (Rosset et al., 2004) and homogeneous models (Wei et al., 2019). More recently, Evron et al. (2023) extended this framework to *continual* linear classification. Similar to our approach, they adjust the projection reference point to capture how expertise accumulates across tasks. Our work completes this landscape by analyzing weakly regularized continual learning in the broader class of nonlinear, homogeneous DNNs (see Table 1).

Technically, our proof of Theorem 1 differs from previous approaches. For instance, while Evron et al. (2023) rely on convex geometry and KKT optimality, these tools do not readily extend to our setting because the feasible sets induced by DNNs are nonconvex. Furthermore, Wei et al. (2019) rely on the pointwise convergence of the loss to a fixed limit. This notion of convergence is insufficient for the continual setting, where the regularization term is dynamic and defined relative to the shifting limit of the previous task.

To address these challenges, we employ the framework of Γ-convergence. This allows us to handle nonconvex geometry by analyzing the global minimizers of a sequence of evolving functionals.

Definition 2 (Γ-convergence; Braides, 2006, Theorem 2.1) *Let $G_n : \mathbb{R}^p \to (-\infty, +\infty]$ and $G : \mathbb{R}^p \to (-\infty, +\infty]$. We say that G_n Γ-converges to G if:*

(a) For every $\Theta \in \mathbb{R}^p$ and every sequence $\Theta_n \to \Theta$, $G(\Theta) \leq \liminf_{n\to\infty} G_n(\Theta_n)$.

(b) For every $\Theta \in \mathbb{R}^p$ there exists a sequence $\Theta_n \to \Theta$ such that $G(\Theta) \geq \limsup_{n\to\infty} G_n(\Theta_n)$.

The key property of Γ-convergence is that the Γ-limit of a sequence of functionals ensures the convergence of their respective minimizers to a minimizer of the limit, as formalized next.

Lemma 3 (Fundamental Theorem of Γ-convergence; Braides, 2006, Theorem 2.10) *Let $(G_n)_{n\in\mathbb{N}}$ be a family of functionals $G_n : \mathbb{R}^p \to (-\infty, +\infty]$. Assume that $(G_n)_{n\in\mathbb{N}}$ is equicoercive, i.e., that for every $C \in \mathbb{R}$, $\{\Theta \in \mathbb{R}^p : G_n(\Theta) \leq C\} \subset K_C$ for some compact set K_C. Then, if G_n Γ-converges to a functional $G : \mathbb{R}^p \to (-\infty, +\infty]$ as $n \to \infty$, the following hold:*

(a) ***Convergence of minimal values:*** $\lim_{n\to\infty} \inf_{\Theta\in\mathbb{R}^p} G_\lambda(\Theta) = \min_{\Theta\in\mathbb{R}^p} G(\Theta)$.

(b) ***Convergence of minimizers:*** *Every sequence (Θ_n) such that $\Theta_n \in \operatorname{argmin} G_n$ admits a convergent subsequence (Θ_{n_j}) such that $\lim_{j\to\infty} \Theta_{n_j} = \Theta^\star$, for $\Theta^\star \in \operatorname{argmin} G$. If the minimizer $\Theta^\star$ of G is unique, then the entire sequence (Θ_n) converges to $\Theta^\star$.*

Having established these preliminaries, we now detail the key ideas of our proof.

Proof sketch of our Theorem 1. Let (λ_n) satisfy $\lambda_n \to 0$. For $\lambda > 0$ and $t \in [k]$, define $c_{\lambda,r} = (\log \frac{1}{\lambda})^{1/r}$ and $\hat{\Theta}_t^{(\lambda)} = \Theta_t^{(\lambda)}/c_{\lambda,r}$. To establish convergence in direction, we first show convergence of the scaled iterates $\hat{\Theta}_t^{(\lambda)}$. For this, we need to prove that there exists a subsequence (λ_ℓ) such that $\hat{\Theta}_t^{(\lambda_\ell)} \to \bar{\Theta}_t$ for some $\bar{\Theta}_t \in \operatorname{argmin}_{\bar{\Theta}\in\mathcal{C}_{\tau(t)}} \|\bar{\Theta} - \bar{\Theta}_{t-1}\|^2$.

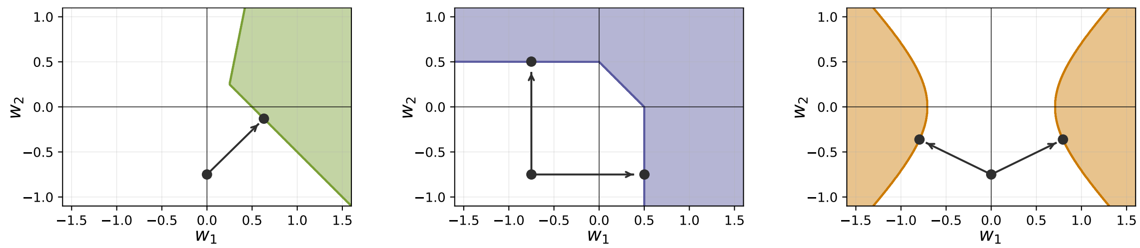
We proceed by induction on t. Assume that for some subsequence (λ_j), $\hat{\Theta}_{t-1}^{(\lambda_j)} \to \bar{\Theta}_{t-1} \in \operatorname{argmin}_{\bar{\Theta}\in\mathcal{C}_{\tau(t-1)}} \|\bar{\Theta} - \bar{\Theta}_{t-2}\|^2$, we define

$$G_t^{(\lambda)}(\Theta) \triangleq \frac{\mathcal{L}_{\tau(t)}(c_{\lambda,r}\Theta)}{c_{\lambda,r}^2 \lambda} + \|\Theta - \hat{\Theta}_{t-1}^{(\lambda)}\|^2.$$

A change of variables yields

$$\Theta_t^{(\lambda)} \in \operatorname{argmin}\{\mathcal{L}_{\tau(t)}(\Theta) + \lambda\|\Theta - \Theta_{t-1}^{(\lambda)}\|^2\} \quad\iff\quad \hat{\Theta}_t^{(\lambda)} \in \operatorname{argmin} G_t^{(\lambda)}(\Theta).$$

Then, we show that, along (λ_j), the functionals $G_t^{(\lambda_j)}$ are equicoercive and Γ-converge to $G_t(\Theta) = \mathbf{1}_{\mathcal{C}_{\tau(t)}}(\Theta) + \|\Theta - \bar{\Theta}_{t-1}\|^2$, where $\mathbf{1}_{\mathcal{C}}(\Theta) = 0$ for $\Theta \in \mathcal{C}$ and ∞ otherwise. By Lemma 3, any sequence of minimizers $\hat{\Theta}_t^{(\lambda_j)}$ admits a convergent subsequence whose limit is a global minimizer of G_t, establishing the inductive step. Since this holds for any (λ_n) with $\lambda_n \to 0$, and since the directions of the scaled iterates $\hat{\Theta}_t^{(\lambda)}$ and the original iterates $\Theta_t^{(\lambda)}$ are identical, the theorem follows.



(a) **Linear model.** Consider two 2D data points $\mathbf{x}_1 = [5, -1]^\top$, $\mathbf{x}_2 = [2, 2]^\top$, $y_1 = y_2 = +1$, and a linear model $f(\mathbf{x}; w_1, w_2) = w_1 x_1 + w_2 x_2$. As explained in Evron et al. (2023), the resulting feasible set is a closed and convex affine polyhedral cone of the form $\left[\begin{smallmatrix}5 & -1\\ 2 & 2\end{smallmatrix}\right]\left[\begin{smallmatrix}w_1\\ w_2\end{smallmatrix}\right] \geq \left[\begin{smallmatrix}1\\ 1\end{smallmatrix}\right]$.

(b) **Simplified ReLU.** Consider a single 1D data point $x = 2, y = +1$, and a 1-positively-homogeneous model $f(x; w_1, w_2) = [w_1 x]_+ + [w_2 x]_+$. Here, the induced feasible set $f(2; w_1, w_2) \geq 1$ is nonconvex. Consequently, a single point can have multiple valid projections.

(c) **Squared linear model.** Consider a single 1D data point $x = 2, y = +1$, and a 2-positively-homogeneous model $f(x; w_1, w_2) = w_1^2 x - w_2^2 x$ as in Woodworth et al. (2020). While comprised of two convex regions, the resulting feasible set $f(2; w_1, w_2) \geq 1$ is nonconvex.

Figure 1: **Feasible sets under homogeneous models are not necessarily convex.** In the two-parameter spaces depicted, only the linear model yields convex feasible sets and, consequently, unique projections.

3. Convergence Analysis from a Projection Perspective

The established projection perspective (Theorem 1) facilitates analyzing continual learning through the lens of projection theory. Indeed, prior work has used classical results from alternating projections and Projections Onto Convex Sets (POCS) to analyze the convergence of continual *linear* models (Evron et al., 2022, 2023). However, these tools are not directly applicable in our nonlinear setting as feasible sets are no longer convex. While this raises several challenges—detailed below—the projection viewpoint remains informative for structuring our analysis of the induced dynamics.

3.1. Challenges: Projections onto Nonconvex Sets

Extending projection-based analysis beyond convexity breaks classical analytical frameworks.

Nonunique projections. As illustrated in Figure 1, linear models induce convex feasible sets for which projections are uniquely defined—but general homogeneous models yield nonconvex feasible sets where a projection $\mathbf{\Pi}_{\mathcal{C}}(\mathbf{u}) \in \operatorname{argmin}_{\mathbf{v}\in\mathcal{C}} \|\mathbf{u}-\mathbf{v}\|_2$ may admit multiple minimizers. This nonuniqueness is critical: Algorithm 2 may select an arbitrary minimizer at each step, potentially inducing qualitatively distinct dynamics and branching into trajectories with diverging behaviors. Such branching complicates any uniform analysis that must account for all admissible selection rules.

Breakdown of projection properties. Another distinction is that convex projections monotonically approach the *joint* feasible set, *i.e.*, $d(\mathbf{\Pi}_{\mathcal{C}_m}(\bar{\boldsymbol{\theta}}), \mathcal{C}^\star) \leq d(\bar{\boldsymbol{\theta}}, \mathcal{C}^\star)$, due to operator nonexpansiveness, *i.e.*, $\|\mathbf{\Pi}_{\mathcal{C}}(\mathbf{u}) - \mathbf{\Pi}_{\mathcal{C}}(\mathbf{v})\| \leq \|\mathbf{u} - \mathbf{v}\|$ (Evron et al., 2023, Lemma 4.5 and Corollary D.1). While crucial for convergence, these properties fail in the nonconvex regime; for instance, nonuniqueness in Figure 1b implies that it is even possible that, roughly, $\|\mathbf{\Pi}_{\mathcal{C}}(\mathbf{u} + \boldsymbol{\varepsilon}) - \mathbf{\Pi}_{\mathcal{C}}(\mathbf{u})\| \gg 0$ while $\|\mathbf{u} + \boldsymbol{\varepsilon} - \mathbf{u}\| = \varepsilon \to 0$. Overall, this removes primary analytical mechanisms used in prior work.

In Section 3.2, we show how this property breakdown precludes global convergence, while Section 3.3 establishes local convergence for models initialized near the intersection.

3.2. Lower Bound: Forgetting Can Be Catastrophic

Evron et al. (2022) proposed that forgetting is truly "catastrophic" only when it fails to converge to zero in the infinite-task-sequence limit. In their continual linear regression setting, they demonstrated that cyclic or random task orderings eliminate forgetting as the number of iterations $k \to \infty$. Similar convergence was later established for continual linear classification (Evron et al., 2023).

While significant, these results are expected from a projection perspective. Both works prove an equivalence to sequential projections onto *convex* sets—specifically affine subspaces in regression and polyhedral cones in classification (see Figure 1a). Such processes are well-known to converge to the intersection under cyclic or random orderings (Deutsch and Hundal, 2006; Nedić, 2010). We now ask: *Can forgetting be catastrophic in homogeneous models whose projections are nonconvex?*

We answer in the affirmative. We show that the sequential margin projections of Algorithm 2 may fail to converge even as $k \to \infty$. This means that continual learning on homogeneous models can forget *catastrophically*, *i.e.,* never reaching a configuration that satisfies all tasks simultaneously.

Technically, we employ a standard 2-positively-homogeneous model that is linear in the data but nonlinear in its parameters (e.g., as in Woodworth et al., 2020), as illustrated in Figure 1c.

Definition 4 (Squared Linear Model) *For inputs $\mathbf{x} \in \mathbb{R}^d$, we define a 2-positively-homogeneous model parameterized by $\bar{\Theta} = (\mathbf{u}, \mathbf{v})$ where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$:*

$$f(\mathbf{x}; \mathbf{u}, \mathbf{v}) \;=\; \langle \mathbf{u}^2, \mathbf{x} \rangle - \langle \mathbf{v}^2, \mathbf{x} \rangle \;=\; \sum_{i=1}^{d} \left(u[i]^2 - v[i]^2 \right) x[i] \,.$$

We construct a specific case exhibiting catastrophic forgetting with as few as $M = 2$ tasks and $p = 4$ learnable parameters. This highlights a major qualitative gap between convex and nonconvex projection interpretations of continual learning: whereas for linear models, catastrophic forgetting arises only in the limit of $M \to \infty$ tasks (Evron et al., 2022, 2023), our nonconvex setting may fail with minimal complexity. Moreover, unlike the convex case—where task repetition drives iterates toward joint feasibility—alternating between $M = 2$ tasks here does *not* lead to convergence.

Theorem 5 (Catastrophic Forgetting in Squared Linear Models) *Let $f(\mathbf{x}; \bar{\Theta})$ be a squared linear model in $d = 2$ parameterized by $\bar{\Theta} = (\mathbf{u}, \mathbf{v})$ where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$. Fix $\varepsilon \in (0, 0.01)$ and define two tasks*

$$\mathbf{X}^{(1)} = \left[\mathbf{x}_1^\top\right] = \left[\left(\tfrac{1}{9} + \varepsilon,\; 1 - \varepsilon\right)\right], \qquad \mathbf{X}^{(2)} = \left[\mathbf{x}_2^\top\right] = \left[\left(\tfrac{1}{9} - \varepsilon,\; -1 - \varepsilon\right)\right], \qquad \mathbf{y}^{(1)} = \mathbf{y}^{(2)} = (1) \;.$$

(a) The two tasks are jointly separable, i.e., $\mathcal{C}^\star \triangleq \mathcal{C}_1 \cap \mathcal{C}_2 \neq \varnothing$ where the feasible sets are $\mathcal{C}_1 = \{\bar{\Theta} \in \mathbb{R}^4 \mid f(\mathbf{x}_1; \bar{\Theta}) \geq 1\}$, $\mathcal{C}_2 = \{\bar{\Theta} \in \mathbb{R}^4 \mid f(\mathbf{x}_2; \bar{\Theta}) \geq 1\}$.

(b) Algorithm 2 forgets catastrophically: $d(\bar{\Theta}_t, \mathcal{C}^\star) \geq 2$, $\max_{m \in \{1,2\}} F_m(\bar{\Theta}_t) \geq 1.5$, $\forall t \geq 0$.

The proof, provided in Appendix B, relies on the fact that for the parameters $\bar{\Theta} = (\mathbf{u}, \mathbf{v})$ to lie in the intersection $\mathcal{C}_1 \cap \mathcal{C}_2$, it must hold that $u[1] \geq 2$. However, we show that the sequential projection dynamics, when initialized at the origin $\bar{\Theta}_0 = \mathbf{0}$, are restricted to the subspace defined by $u[1] = 0$ and $v[1] = 0$. Namely, under the specific geometry of our construction, each projection step modifies only $u[2]$ and $v[2]$, leaving the other components at zero. Consequently, the iterates $\bar{\Theta}_t$ are effectively trapped in a subspace bounded away from the joint feasible set, *i.e.,* $\mathcal{C}_1 \cap \mathcal{C}_2$.

Importantly, the failure in Theorem 5 is not an artifact of a "degenerate" dataset or an adversarial task ordering: it persists over a nonzero measure of datasets under any ordering. Thus, global convergence guarantees require further assumptions in this nonconvex setting, as introduced next.

Moreover, the lower-bound construction in Theorem 5 extends beyond the case of zero initialization. In particular, a similar phenomenon can arise for arbitrary (and in particular random) initializations within a bounded ball. By appropriately adjusting the ratio between the coordinates in the construction, one can ensure that the distance to the jointly feasible set $\mathcal{C}^\star$ is arbitrarily larger than the initialization radius. In this regime, the sequential projection dynamics behave similarly to the zero-initialization case: only a subset of the coordinates is updated, even though reaching a global solution requires movement in other directions.

3.3. Upper Bounds via Local Convergence Analysis

As we established, global convergence is generally unattainable in the nonconvex setting, even when individual tasks are solved to optimality. Thus, we focus on local convergence analysis, as is common in nonconvex optimization, where global guarantees are elusive (e.g., Bottou et al., 2016). A practical motivation is the sequential finetuning of large pretrained foundation models. These models often exhibit small zero-shot loss, implying an initialization near a joint solution. Our analysis shows that local convergence *can* be guaranteed in this regime: if the initialization is sufficiently close to the joint feasible set $\mathcal{C}^\star$, the sequence of projections converges to $\mathcal{C}^\star$ at a linear rate.

We establish this result when f is β-smooth[3] and G-Lipschitz[4] for every data point $\mathbf{x}$. In particular, both assumptions hold for DNNs with *smooth* activations (*e.g.,* squared ReLU), where the Lipschitz constant may depend on architectural properties such as the network depth and homogeneity degree, as well as the magnitude of the data. We prove the following local convergence theorem.

Theorem 6 (Convergence Rate for Lipschitz, Smooth Homogeneous Models) *Consider a G-Lipschitz, β-smooth, r-positively-homogeneous model $f(\cdot;\bar{\Theta}):\mathcal{X}\to\mathbb{R}$. Assume joint separability, i.e., nonempty $\mathcal{C}^\star=\cap_m\mathcal{C}_m\neq\varnothing$ (Assumption A). Let ε hold $\frac{1}{(1-\varepsilon)^{M-1}}=1+\frac{r^2}{2(M-1)G^2\|\Theta^\star\|^2}$ and $\delta=\varepsilon r/\beta\|\Theta^\star\|$. Then, there exists $\Theta^\star\in\mathcal{C}^\star$ such that if $\bar{\Theta}_0\in\mathcal{B}_\delta(\Theta^\star)$, the sequential projections of Algorithm 2 converge linearly with a rate depending on the task ordering, i.e., $\forall k\in\mathbb{N}^+$,*

- ***Random:*** $$\frac{1}{G}\mathop{\mathbb{E}}_{\tau_{iid}}\max_{m\in[M]}F_m(\bar{\Theta}_k)\ \le\ \mathop{\mathbb{E}}_{\tau_{iid}}d(\bar{\Theta}_k,\mathcal{C}^\star)\ \le\ \exp\left(-\frac{k}{M}\frac{r^2}{4G^2\|\Theta^\star\|^2}\right)\mathop{\mathbb{E}}_{\tau_{iid}}d(\bar{\Theta}_0,\mathcal{C}^\star).$$
- ***Cyclic** ($M\mid k$)**:*** $$\frac{1}{G}\max_{m\in[M]}F_m(\bar{\Theta}_k)\ \le\ d(\bar{\Theta}_k,\mathcal{C}^\star)\ \le\ \exp\left(-\frac{k}{M^2}\frac{r^2}{4G^2\|\Theta^\star\|^2}\right)d(\bar{\Theta}_0,\mathcal{C}^\star).$$

The proof is provided in Appendix C. Here, we make a few remarks and then outline the proof idea.

Remark 7 (Comparison to Linear Model) *For a linear model, the local convergence rates above recover the global rates of Evron et al. (2023) for both cyclic and random task orderings.*

Remark 8 (Nonzero Initialization) *Local convergence requires the initialization of Algorithm 2 to lie within a neighborhood $\mathcal{B}_\delta(\Theta^\star)$. While we prove Theorem 1 for the case of zero initialization, this does not necessitate that $\Theta^\star$ is near the origin. Indeed, by using a technical modification, it is possible to extend the analysis to other fixed initializations where the initialization of Algorithm 2 scales according to the weak regularization's λ. This could be the case, for instance, when Algorithm 2 is initialized with a model pretrained using weight decay* with the same order of λ, *a change that does not alter the underlying geometric intuition.*

3. $f(\mathbf{x};\cdot)$ is β-smooth if $\forall\mathbf{u},\mathbf{w}$, $\left|f(\mathbf{x};\mathbf{u})-f(\mathbf{x};\mathbf{w})-\langle\nabla f(\mathbf{x};\mathbf{w}),\mathbf{u}-\mathbf{w}\rangle\right|\le\frac{\beta}{2}\left\|\mathbf{u}-\mathbf{w}\right\|^2$.
4. $f(\mathbf{x};\cdot)$ is G-Lipschitz if $\forall\mathbf{u},\mathbf{w}$, $\left|f(\mathbf{x};\mathbf{u})-f(\mathbf{x};\mathbf{w})\right|\le G\left\|\mathbf{u}-\mathbf{w}\right\|$.

Remark 9 (Order of Limits) *The number of iterations k may be arbitrarily large to guarantee convergence, potentially requiring task repetitions. Following (Evron et al., 2023, Remark 4.6), we clarify that our analysis takes the limit $\lambda \downarrow 0$* after *fixing k.*

Proof sketch. As in the convex case where proofs are based on properties of POCS (Evron et al., 2023), our proof of Theorem 6 leverages properties of projections on *nonconvex* sets. However, as discussed, nonconvex projections can be ill-behaved; they are often nonunique and tiny perturbations can cause large "jumps". As a result, nonconvex projections are typically analyzed under regularity conditions that rule out pathological geometry and ensure the projection is locally stable, which is essential for any convergence guarantees (e.g., Lewis and Malick, 2008; Lewis et al., 2009; Dao and Phan, 2019). The resulting proof consists of two steps.

1. **First step: Smoothness and Lipschitz continuity imply regularity.** We show that a smooth and G-Lipschitz homogeneous model satisfies the regularity properties required for local convergence. We rely on tools from variational analysis and set regularity (e.g., Rockafellar and Wets, 1998) to assert that the feasible sets induced by homogeneous DNNs satisfy the needed local regularity conditions. In particular, we establish the following two conditions.

   Definition 10 ((ε, δ)-Regularity) *Let $\mathcal{C}$ be a nonempty subset of $\mathbb{R}^p$, $\mathbf{w} \in \mathbb{R}^p$, $\varepsilon \geq 0$ and $\delta > 0$. Let $\mathbf{\Pi}_{\mathcal{C}}(\mathbf{w})$ project $\mathbf{w}$ onto $\mathcal{C}$ and define $N_{\mathcal{C}}^{\text{prox}}(\mathbf{w}) \triangleq \left\{\lambda(\mathbf{z} - \mathbf{w}) \;\middle|\; \mathbf{z} \in \mathbf{\Pi}_{\mathcal{C}}^{-1}(\mathbf{w}),\ \lambda \geq 0\right\}$. We say that $\mathcal{C}$ is (ε, δ)-regular at $\mathbf{w}$ if*

   $$\mathbf{x}, \mathbf{y} \in \mathcal{C} \cap \mathbb{B}(\mathbf{w}; \delta),\ \mathbf{u} \in N_{\mathcal{C}}^{\text{prox}}(\mathbf{x}) \quad \Longrightarrow \quad \langle \mathbf{u}, \mathbf{x} - \mathbf{y} \rangle \leq \varepsilon \|\mathbf{u}\| \cdot \|\mathbf{x} - \mathbf{y}\|.$$

   Definition 11 (κ-Linear Regularity of Set Collection) *A system $\{\mathcal{C}_i\}_{i \in I}$ is κ-linearly regular on a subset $U \subseteq X$ if, for all $x \in U$, $d(x, \cap_{i \in I} \mathcal{C}_i) \leq \kappa \max_{i \in I} d(x, \mathcal{C}_i)$.*

   In Lemmas C.10 and C.11 in Appendix C we show that, in the smooth and Lipschitz case, homogeneous models satisfy the conditions given in Definitions 10 and 11.

2. **Second step: Invoke convergence under regularity.** We invoke a local linear convergence result for nonconvex projections. In particular, the following proposition shows that under the conditions of Definitions 10 and 11, the iterates of Algorithm 2 exhibit local convergence to the intersection $\mathcal{C}^\star$ both in random and cyclic orderings. For random ordering, the proof appears in Appendix C, while the cyclic case follows directly from Dao and Phan (2019, Corollary 5.10).

   Proposition 12 (Stepwise decrease under random and cyclic projections) *Consider closed sets $\mathcal{C}_1, \ldots, \mathcal{C}_M \subset \mathbb{R}^p$ with nonempty intersection $\mathcal{C}^\star \triangleq \bigcap_{m=1}^M \mathcal{C}_m \neq \varnothing$, and let $\Theta^\star \in \mathcal{C}^\star$. Assume that there exist constants $\kappa > 0$, $\varepsilon \in [0, 1)$, and $\delta > 0$ such that:*

   (a) Each set $\mathcal{C}_m$ is (ε, δ)-regular *at $\Theta^\star$.*
   (b) The collection $\{\mathcal{C}_m\}_{m=1}^M$ is κ-linearly regular *on the ball $\mathcal{B}_{\delta/2}(\Theta^\star)$.*

   Then, we get one-step decrease depending on the task ordering,

   (a) ***Random ordering.*** *Let $\tau = \tau_{iid}, t \geq 1$. If $\bar{\Theta}_{t-1} \in \mathcal{B}_{\delta/2}(\Theta^\star)$ and $\rho_{iid} \triangleq \frac{1}{1-\varepsilon} - \frac{1}{M\kappa^2} < 1$, it holds that $\mathbb{E}_{\tau(t)}\left[d^2(\bar{\Theta}_t, \mathcal{C}^\star) \middle| \bar{\Theta}_{t-1}\right] \leq \rho_{iid}\, d^2(\bar{\Theta}_{t-1}, \mathcal{C}^\star)$.*
   (b) ***Cyclic ordering.*** *Let $\tau = \tau_{cyc}$ and t such that $M \mid t$. If, $\bar{\Theta}_t \in \mathcal{B}_{\delta/2}(\Theta^\star)$ and $\rho_c \triangleq \left(\frac{1}{(1-\varepsilon)^{M-1}} - \frac{1}{(M-1)\kappa^2}\right)^{\frac{1}{2(M-1)}} < 1$, then, $d(\bar{\Theta}_{t+M}, \mathcal{C}) \leq \rho_{cyc}\, d(\bar{\Theta}_t, \mathcal{C})$.*

4. Extension: Continual Regression in Homogeneous DNNs as Sequential Projections

We now show that the techniques developed above are general enough to extend naturally to continual regression. This extension requires only minor adjustments to the definitions.

Notational adjustments. For a task $m \in [M]$, the label vector is no longer binary, but rather $\mathbf{y}^{(m)} \in \mathbb{R}^{n_m}$. Consequently, the feasible sets are now defined as,

$$\mathcal{C}_m \triangleq \left\{ \Theta \in \mathbb{R}^p \;\middle|\; f(\mathbf{x}_i^{(m)}; \Theta) = y_i^{(m)},\ \forall i \in [n_m] \right\}, \qquad \forall m \in [M] \,. \tag{10}$$

Accordingly, the joint separability assumption (Assumption A) is now a joint *realizability* assumption, still requiring $\mathcal{C}^\star \triangleq \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$.

Furthermore, instead of the logistic loss, we now optimize the mean squared loss, *i.e.,*

$$\mathcal{L}_m(\Theta) \triangleq \frac{1}{n_m} \sum_{i=1}^{n_m} \left(f(\mathbf{x}_i^{(m)}; \Theta) - y_i^{(m)} \right)^2, \qquad \forall m \in [M] \,. \tag{11}$$

We still consider $d(\cdot, \mathcal{C}_m)$ and $d(\cdot, \mathcal{C}^\star)$ (as in Section 2.1), but redefine the *forgetting* with respect to the squared loss instead of the hinge loss, *i.e.,*

$$F_{\tau(t')}(\bar{\Theta}_t) = \max_i \left(f(\mathbf{x}_i^{\tau(t')}; \bar{\Theta}_t) - y_i^{\tau(t')} \right)^2, \qquad \forall t \in [k] \,, \forall t' \leq t \,. \tag{12}$$

Result. Analogously to our classification result, we establish the following theorem for the convergence of Algorithm 1 to Algorithm 2 as $\lambda \downarrow 0$ for the regression case.

Theorem 13 (Weakly-Regularized Continual Regression → Sequential Projections) *Consider a model $f(\cdot; \Theta) : \mathcal{X} \to \mathbb{R}$. Assume individual realizability, i.e., nonempty feasible sets $\mathcal{C}_1, \ldots, \mathcal{C}_M$ (Eq. (10)). Then, as $\lambda \downarrow 0$, Algorithm 1 trained with the squared loss aligns with Algorithm 2. That is, for every iteration $t \in [k]$, any sequence $\lambda \downarrow 0$ admits a subsequence (λ_j) and a point $\bar{\Theta}_t \in \mathbf{\Pi}_{\tau(t)}(\bar{\Theta}_{t-1}) \triangleq \operatorname{argmin}_{\bar{\Theta} \in \mathcal{C}_{\tau(t)}} \left\| \bar{\Theta} - \bar{\Theta}_{t-1} \right\|^2$ such that $\Theta_t^{(\lambda_j)} \xrightarrow{j \to \infty} \bar{\Theta}_t \in \mathbf{\Pi}(\bar{\Theta}_{t-1})$.*

Proofs for this section appear in Appendix D; as with classification, the analysis uses Γ-convergence.

4.1. Convergence for Overparameterized Homogeneous Regression Models

Mirroring the classification case (Theorem 5), we first show that the 2-positively-homogeneous squared model of Definition 4 may suffer from *catastrophic* forgetting.

Lemma 14 (Catastrophic Forgetting in Squared Models: Simplified Version) *Under the same task construction of Theorem 5, the tasks are jointly separable, yet Algorithm 2 forgets catastrophically. That is, $\mathcal{C}^\star \neq \varnothing$ but $d\big(\bar{\Theta}_t, \mathcal{C}^\star\big) \;\geq\; 2$ and $\max_m F_m(\bar{\Theta}_t) \geq 3$, for every iteration $t \geq 0$.*

The formal statement and its proof appear in Appendix D. Specifically, we show that in the *classification* case (Theorem 5), the current iterate $\bar{\Theta}_t$ always lies on the boundary of the latest feasible set; thus, the dynamics of Algorithm 2 is identical for regression and classification.

After establishing that global convergence is not guaranteed in the general homogeneous regression setting, we now show that local convergence can still be achieved. In contrast to our classification result, the local convergence result for regression requires the additional assumption $y_{\min} = \min_{m,i} |y_i^{(m)}| \neq 0$. This assumption is imposed to ensure that the (ε, δ)-regularity condition holds (see Lemma D.3 for more details).

Theorem 15 (Convergence Rate for Lipschitz, Smooth Homogeneous Models) *Consider a G-Lipschitz, β-smooth, r-positively-homogeneous model $f(\cdot;\bar{\Theta}) : \mathcal{X} \to \mathbb{R}$. Assume joint realizability, i.e., nonempty intersection $\mathcal{C}^\star = \cap_m \mathcal{C}_m \neq \varnothing$. Let $y_{\min} = \min_{m,i} |y_i^{(m)}| \neq 0$. Let ε hold $\frac{1}{(1-\varepsilon)^{M-1}} = 1 + \frac{r^2}{2(M-1)G^2\|\Theta^\star\|^2}$ and $\delta = \frac{y_{\min}\varepsilon r}{\beta\|\Theta^\star\|}$. Then, there exists $\Theta^\star \in \mathcal{C}^\star$ s.t. if $\bar{\Theta}_0 \in \mathcal{B}_\delta(\Theta^\star)$, the sequential projections of Algorithm 2 converge linearly with a rate depending on the task ordering, i.e., $\forall k \in \mathbb{N}^+$,*

- ***Random:*** $$\frac{1}{G}\mathop{\mathbb{E}}_{\tau_{iid}} \max_{m\in[M]} F_m(\bar{\Theta}_k) \;\le\; \mathop{\mathbb{E}}_{\tau_{iid}} d^2(\bar{\Theta}_k, \mathcal{C}^\star) \;\le\; \exp\left(-\frac{k}{M^2}\frac{r^2 y_{\min}^2}{2G^2\|\Theta^\star\|^2}\right)\mathop{\mathbb{E}}_{\tau_{iid}} d^2(\bar{\Theta}_0, \mathcal{C}^\star).$$
- ***Cyclic** ($M \mid k$):* $$\frac{1}{G}\max_{m\in[M]} F_m(\bar{\Theta}_k) \;\le\; d^2(\bar{\Theta}_k, \mathcal{C}^\star) \;\le\; \exp\left(-\frac{k}{M^2}\frac{r^2 y_{\min}^2}{2G^2\|\Theta^\star\|^2}\right) d^2(\bar{\Theta}_0, \mathcal{C}^\star).$$

5. Discussion

Our results frame continual learning in homogeneous deep networks as a geometric process governed by sequential projections onto task-induced margin sets, unifying several empirical and theoretical observations. Below, we discuss additional connections to prior work and highlight promising directions for further inquiry.

Geometry as the source of catastrophic forgetting. Our results suggest that *catastrophic* forgetting is a fundamental geometric consequence of nonconvexity. Unlike linear models with convex feasible sets (Evron et al., 2023), homogeneous DNNs induce nonconvex sets where even exact sequential projections may fail to reach a point in the joint feasible set. Our lower bound in Section 3.2 demonstrates that this occurs even in low-dimensional models under benign task orderings, proving that nonconvex geometry alone is sufficient to preclude global convergence.

The effect of depth. The convergence rates in Theorems 6 and 15 are governed by the ratio r/G, where r is the homogeneity degree and G is the Lipschitz constant. While depth increases r (see Eq. (6)), the Lipschitz constant G typically grows exponentially with the number of layers (e.g., Szegedy et al., 2014; Virmaux and Scaman, 2018). Consequently, increased depth tends to decrease the r/G ratio, thereby *degrading* the local convergence rate. This provides a theoretical mechanism for the findings of Guha and Lakshman (2024), who observe that deeper networks suffer from exacerbated forgetting (see also Mirzadeh et al., 2022; Ramasesh et al., 2020). A compelling avenue for future work is to characterize how specific architectures influence the geometry of margin sets (Eq. (8)) and their corresponding regularity moduli.

Characterizing convergence in deep continual models. Empirical studies have shown that task repetition can mitigate forgetting even without explicit algorithmic intervention (Lesort et al., 2023; Hemati et al., 2025). While recent analytical work has explained this in linear models through cyclic or random task orderings (Evron et al., 2022, 2023, 2026; Kong et al., 2023; Levinstein et al., 2025; Attia et al., 2025), such analyses often rely on the NTK regime to approximate deep models. Our framework more directly reflects the intrinsic nonconvex dynamics of deep architectures. We hope the analytical foundation laid here facilitates a deeper understanding of failure modes, the benefits of pretraining, and the conditions for local stability in deep continual learning.

Acknowledgments

The research of MS was Funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 2549/19 and 3174/23), a grant from the Tel Aviv University Center for AI and Data Science (TAD) and from the Len Blavatnik and the Blavatnik Family foundation.

GB thanks Elad Hazan and Princeton University for the financial support.

The research of DS was Funded by the European Union (ERC, A-B-C-Deep, 101039436). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (ERCEA). Neither the European Union nor the granting authority can be held responsible for them.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, Oct 2021.

Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Frederik Benzing. Unifying regularisation methods for continual learning. *AISTATS*, 2022.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. 2016. ISSN 0036-1445. doi: 10.1137/16M1080173.

Andrea Braides. A handbook of Γ-convergence. In *Handbook of Differential Equations: stationary partial differential equations*, volume 3, pages 101–213. Elsevier, 2006.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Minh N Dao and Hung M Phan. Linear convergence of projection algorithms. *Mathematics of Operations Research*, 44(2):715–738, 2019.

Frank Deutsch and Hein Hundal. The rate of convergence for the cyclic projections algorithm i: Angles between convex sets. *Journal of Approximation Theory*, 142(1):36–55, 2006.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1072–1080, 2021.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, pages 4028–4079. PMLR, 2022.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9440–9484. PMLR, 23–29 Jul 2023.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to sgd and back: Better rates for random task orderings in continual linear models. In *37th International Conference on Algorithmic Learning Theory*, 2026.

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting - an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024.

Jacopo Graldi, Alessandro Breccia, Giulia Lanzillotta, Thomas Hofmann, and Lorenzo Noci. The importance of being lazy: Scaling limits of continual learning. In *Forty-second International Conference on Machine Learning*, 2025.

Etash Kumar Guha and Vihan Lakshman. On the diminishing returns of width for continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Hamed Hemati, Lorenzo Pellegrini, Xiaotian Duan, Zixuan Zhao, Fangfang Xia, Marc Masana, Benedikt Tscheschner, Eduardo Veas, Yuxiang Zheng, Shiji Zhao, Shao-Yuan Li, Sheng-Jun Huang, Vincenzo Lomonaco, and Gido M. van de Ven. Continual learning in the presence of repetition. *Neural Networks*, 183:106920, 2025.

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gilad Karpel, Edward Moroshko, Ran Levinstein, Ron Meir, Daniel Soudry, and Itay Evron. Optimal l2 regularization in high-dimensional continual linear regression. In *37th International Conference on Algorithmic Learning Theory*, 2026.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Mark Kong, William Swartworth, Halyun Jeong, Deanna Needell, and Rachel Ward. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021.

Sebastian Lee, Stefano Sarao Mannelli, Claudia Clopath, Sebastian Goldt, and Andrew Saxe. Maslow's hammer in catastrophic forgetting: Node re-use vs. node activation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12455–12477. PMLR, 17–23 Jul 2022.

Timothée Lesort, Oleksiy Ostapenko, Pau Rodríguez, Diganta Misra, Md Rifat Arefin, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pages 43–65. PMLR, 2023.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

Adrian S Lewis and Jérôme Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

Adrian S Lewis, D Russell Luke, and Jérôme Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485–513, 2009.

Boqi Li, Youjun Wang, and Weiwei Liu. Towards understanding catastrophic forgetting in two-layer convolutional neural networks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 36057–36095. PMLR, 13–19 Jul 2025.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup, editors, *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 513–533. PMLR, 22–25 Aug 2023.

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. In *Forty-second International Conference on Machine Learning*, 2025.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21078–21100. PMLR, 23–29 Jul 2023.

Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *ICML Workshop on Theory and Foundations of Continual Learning*, 2021.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Seyed Iman Mirzadeh, Arslan Chaudhry, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. *ICML*, 2022.

Angelia Nedić. Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7655–7660. IEEE, 2010.

Binghui Peng and Andrej Risteski. Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Liangzu Peng, Paris Giampouras, and René Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, and Yuan Xie. Prompt gradient projection for continual learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2020.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2): 123–146, 1995.

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

Saharon Rosset, Ji Zhu, and Trevor J Hastie. Margin maximizing loss functions. In *Advances in neural information processing systems*, pages 1237–1244, 2004.

Haozhe Shan, Qianyi Li, and Haim Sompolinsky. Order parameters and phase transitions of continual learning in deep neural networks. *arXiv preprint arXiv:2407.10315*, 2024.

James Seale Smith, Junjiao Tian, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *CVPR Workshop on Continual Learning in Computer Vision*, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Hossein Taheri, Avishek Ghosh, and Arya Mazumdar. On the theory of continual learning with gradient descent for neural networks. *arXiv preprint arXiv:2510.05573*, 2025.

Matan Tsipory, Ran Levinstein, Itay Evron, Mark Kong, Deanna Needell, and Daniel Soudry. Are greedy task orderings better than random in continual linear regression? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 1.09 - continual learning and catastrophic forgetting. In John Wixted, editor, *Learning and Memory: A Comprehensive Reference (Third Edition)*, pages 153–168. Academic Press, Oxford, third edition edition, 2025. ISBN 978-0-443-15755-4.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5):1–38, 2025.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

Appendix Contents

Appendix A. Proofs for Section 2.2

Proof of Theorem 1. For any $\lambda \in (0,1)$, we define $c_{\lambda,r} = \left(\log \frac{1}{\lambda}\right)^{\frac{1}{r}}$. We prove in induction that for every t, every sequence $\lambda \downarrow 0$ admits a subsequence (λ_j) and a point $\bar{\Theta}_t \in \mathcal{C}_{\tau(t)}$ such that

$$\frac{\Theta_t^{(\lambda_j)}}{c_{\lambda_j,r}} \longrightarrow \bar{\Theta}_t \qquad \text{for} \qquad \bar{\Theta}_t \in \underset{\Theta \in \mathcal{C}_{\tau(t)}}{\operatorname{argmin}} \left\|\Theta - \bar{\Theta}_{t-1}\right\|^2.$$

The base case, $t = 0$ holds trivially for the constant sequence $\Theta_0^\lambda = 0 = \bar{\Theta}_0$. For the step, let (λ) be a sequence that converges to zero. We assume that there exists a subsequence (λ_j) and a point $\bar{\Theta}_{t-1} \in \mathcal{C}_{\tau(t-1)}$ that holds

$$\frac{\Theta_{t-1}^{(\lambda_j)}}{c_{\lambda_j,r}} \longrightarrow \bar{\Theta}_{t-1} \qquad \text{for} \qquad \bar{\Theta}_{t-1} \in \underset{\Theta \in \mathcal{C}_{\tau(t-1)}}{\operatorname{argmin}} \left\|\Theta - \bar{\Theta}_{t-2}\right\|^2.$$

For every λ, we define

$$\hat{\Theta}_t^{(\lambda)} \triangleq \frac{\Theta_t^{(\lambda)}}{c_{\lambda,r}}, \quad G_t^{(\lambda)}(\Theta) \triangleq \frac{\mathcal{L}_{\tau(t)}(c_{\lambda,r}\Theta)}{c_{\lambda,r}^2 \lambda} + \left\|\Theta - \hat{\Theta}_{t-1}^{(\lambda)}\right\|^2.$$

Then, for every $\lambda > 0$,

$$\begin{aligned}
&\Theta_t^{(\lambda)} \in \arg\min \mathcal{L}_{\tau(t)}(\Theta) + \lambda\|\Theta - \Theta_{t-1}^{(\lambda)}\|^2 \\
&\iff \Theta_t^{(\lambda)} \in \arg\min \frac{1}{\lambda}\mathcal{L}_{\tau(t)}(\Theta) + \|\Theta - \Theta_{t-1}^{(\lambda)}\|^2 \\
&\iff \hat{\Theta}_t^{(\lambda)} \in \arg\min \frac{1}{\lambda}\mathcal{L}_{\tau(t)}(c_{\lambda,r}\Theta) + \|c_{\lambda,r}\Theta - \Theta_{t-1}^{(\lambda)}\|^2 \\
&\iff \hat{\Theta}_t^{(\lambda)} \in \arg\min \frac{1}{\lambda c_{\lambda,r}^2}\mathcal{L}_{\tau(t)}(c_{\lambda,r}\Theta) + \|\Theta - \hat{\Theta}_{t-1}^{(\lambda)}\|^2 \\
&\iff \hat{\Theta}_t^{(\lambda)} \in \arg\min G_t^{(\lambda)}(\Theta).
\end{aligned}$$

Now, we look at the sequence $\hat{\Theta}_t^{(\lambda_j)}$. For proving the required, it is sufficient to show a sub-sequence of $\hat{\Theta}_t^{(\lambda_j)}$ that converges to $\bar{\Theta}_t \in \operatorname{argmin}_{\Theta \in \mathcal{C}_{\tau(t)}} \left\|\Theta - \bar{\Theta}_{t-1}\right\|^2$. Thus, by, Lemma 3, it is sufficient to prove that the sequence $G_t^{(\lambda_j)}$ is equi-coercive and that this sequence Γ-converges to some function G_t with $\bar{\Theta}_t \in \arg\min G_t$ when j goes to ∞.

Γ-convergence. For the Γ-convergence, let $G_t(\Theta) = \mathbf{1}_{\mathcal{C}_{\tau(t)}}(\Theta) + \|\Theta - \bar{\Theta}_{t-1}\|^2$, where $1_{\mathcal{C}_{\tau(t)}}(\Theta) = 0$ if $\Theta \in \mathcal{C}_{\tau(t)}$ and ∞ otherwise.

- For the *liminf* property, let $\bar{\Theta}$ and a sequence $\hat{\Theta}^{(\lambda_j)} \to \bar{\Theta}$. We need to prove that

$$G_t(\bar{\Theta}) \leq \liminf_{j\to\infty} G_t^{(\lambda_j)}(\hat{\Theta}^{(\lambda_j)}).$$

For the second term of the function, by the induction hypothesis and continuity of norm, $\left\|\hat{\Theta}_t^{(\lambda_j)} - \hat{\Theta}_{t-1}^{(\lambda_j)}\right\|^2 \to \|\bar{\Theta} - \bar{\Theta}_{t-1}\|^2$. For the first term of the function, since $\mathcal{L}_{\tau(t)} \geq 0$,

$$\liminf_{j\to\infty} \frac{\mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\hat{\Theta}^{(\lambda_j)})}{\lambda_j \, c_{\lambda_j,r}^2} \ \geq \ 0.$$

Combining both together we get that if $\bar{\Theta} \in \mathcal{C}_{\tau(t)}$ then

$$\begin{aligned} G_t(\bar{\Theta}) &= \mathbf{1}_{\mathcal{C}_{\tau(t)}}(\bar{\Theta}) + \|\bar{\Theta} - \bar{\Theta}_{t-1}\|^2 = \|\bar{\Theta} - \bar{\Theta}_{t-1}\|^2 = \liminf_{j\to\infty} \left\|\hat{\Theta}_t^{(\lambda_j)} - \hat{\Theta}_{t-1}^{(\lambda_j)}\right\|^2 \\ &\leq \liminf_{j\to\infty} \frac{\mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\hat{\Theta}^{(\lambda_j)})}{\lambda_j\, c^2_{\lambda_j,r}} + \left\|\hat{\Theta}^{(\lambda_j)} - \hat{\Theta}_{t-1}^{(\lambda_j)}\right\|^2 \\ &= \liminf_{j\to\infty} G_t^{(\lambda_j)}(\hat{\Theta}^{(\lambda_j)})\,. \end{aligned}$$

If $\bar{\Theta} \notin \mathcal{C}_{\tau(t)}$, let $u_{\lambda,i} := c^r_{\lambda,r}\, y_i^{(\tau(t))} f(x_i^{(\tau(t))}; \hat{\Theta}^{(\lambda)})$ and consider the index realizing $\min_i u_{\lambda,i}$. Since $\bar{\Theta} \notin \mathcal{C}_{\tau(t)}$, we know that this index i satisfies $u_{\lambda,i} < c^r_{\lambda,r}$. First consider the case where $u_{\lambda_j,i} \leq 0$ infinitely often in the sequence. Then by using the property that if $u \leq 0$ then $\log(1+e^{-u}) \geq \log 2$, it holds for this sub-sequence that

$$\mathcal{L}_{\tau(t)}(c_{\lambda_j,a}\hat{\Theta}^{(\lambda_j)}) \geq (\log 2)/n_{\tau(t)}\,,$$

and therefore, as for $\bar{\Theta} \notin \mathcal{C}_{\tau(t)}$ we have that $G_t(\bar{\Theta}) = \infty$, we get that

$$\liminf_{j\to\infty} G_t^{(\lambda_j)}(\hat{\Theta}^{(\lambda_j)}) \geq \mathcal{L}_{\tau(t)}(c_{\lambda,r}\hat{\Theta}^{(\lambda_j)})/(\lambda_j c^2_{\lambda_j,r}) \to \infty = G_t(\bar{\Theta})\,.$$

For the case where it does not hold that $u_{\lambda_j,i} \leq 0$ infinitely often in the sequence, there exists $\delta \in (0,1)$ s.t. $0 < u_{\lambda_j,i} \leq c^r_{\lambda_j,r}(1-\delta)$. As a result, since $e^{-c^r_{\lambda_j,r}} = e^{-\log(1/\lambda_j)} = \lambda_j$, we get

$$\begin{aligned} &\mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\hat{\Theta}^{(\lambda_j)}) \geq \frac{1}{2n_{\tau(t)}} e^{-u_{\lambda_j,i}} \geq \frac{1}{2n_{\tau(t)}} e^{-c^r_{\lambda_j,r}(1-\delta)} \geq \frac{1}{2n_{\tau(t)}} \lambda_j^{1-\delta} \\ \implies &\frac{\mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\hat{\Theta}^{(\lambda_j)})}{(\lambda c^2_{\lambda_j,r})} \geq \frac{1}{2n_{\tau(t)}} \frac{\lambda_j^{-\delta}}{c^2_{\lambda_j,r}} \to \infty. \end{aligned}$$

- For the *Limsup*, let $\bar{\Theta}$. If $\bar{\Theta} \in \mathcal{C}_{\tau(t)}$, let $\hat{\Theta}^{(\lambda_j)} = \bar{\Theta}$ for every j (constant sequence). Then, for every i, $y_i^{(\tau(t))} f(\mathbf{x}_i^{(\tau(t))}; \bar{\Theta}) \geq 1$, and,

$$\begin{aligned} \mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\bar{\Theta}) &= \frac{1}{n_{\tau(t)}} \sum_i \log\left(1 + e^{-c^r_{\lambda_j,r}\, y_i^{(m)} f(\mathbf{x}_i^{(m)};\Theta)}\right) \\ &\leq \frac{1}{n_{\tau(t)}} \sum_i e^{-c^r_{\lambda_j,r}\, y_i^{(m)} f(\mathbf{x}_i^{(m)};\Theta)} \leq e^{-c^r_{\lambda_j,r}} = \lambda_j\,. \end{aligned}$$

  Hence $\mathcal{L}_{\tau(t)}(c_{\lambda_j,r}\bar{\Theta})/(\lambda_j c^2_{\lambda_j,r}) \leq 1/c^2_{\lambda_j,r} \to 0$ as $j \to \infty$. Then, by the induction hypothesis and continuity of norm, $G_t^{(\lambda_j)}(\bar{\Theta}) \to \|\bar{\Theta} - \bar{\Theta}_{t-1}\|^2 = G_t(\bar{\Theta})$. Otherwise, $\bar{\Theta} \notin \mathcal{C}_{\tau(t)}$. Then $G_t(\bar{\Theta}) = +\infty$, so the Γ-limsup property is trivial.

Boundedness of level sets. By the positivity of $\mathcal{L}_{\tau(t)}$, for all $\bar{\Theta}$,

$$G_t^{(\lambda_j)}(\bar{\Theta}) \;\geq\; \left\|\bar{\Theta} - \hat{\Theta}_{t-1}^{(\lambda_j)}\right\|^2.$$

Thus every sublevel set $\{\bar{\Theta}:\ G_t^{(\lambda_j)}(\bar{\Theta}) \leq C\}$ is contained in the closed ball centered at $\hat{\Theta}_{t-1}^{(\lambda_j)}$ of radius $\sqrt{C}$. These balls are compact sets and the equi-coercivity follows.

By Lemma 3, there is a subsequence $\hat{\Theta}_t^{(\lambda_{j_\ell})} \to \bar{\Theta}_t$. For getting the convergence in direction, note that $\bar{\Theta}_t \in \mathbf{\Pi}_{\tau(t)}(\bar{\Theta}_{t-1})$ (as a minimizer of G_t), which implies $y_i^{\tau(t)} f(\mathbf{x}_i^{(\tau(t))}; \bar{\Theta}_t) \geq 1$ for all i. Therefore, $\bar{\Theta}_t \neq 0$ (since by homogeneity $f(\cdot; 0) = 0$). Since the limit is nonzero, the mapping $u \mapsto u/\|u\|$ is continuous at $\bar{\Theta}_t$. As a result,

$$\frac{\Theta_t^{(\lambda_{j_\ell})}}{\left\|\Theta_t^{(\lambda_{j_\ell})}\right\|} = \frac{c_{\lambda_{j_\ell}} \hat{\Theta}_t^{(\lambda_{j_\ell})}}{\left\|c_{\lambda_{j_\ell}} \hat{\Theta}_t^{(\lambda_{j_\ell})}\right\|} = \frac{\hat{\Theta}_t^{(\lambda_{j_\ell})}}{\left\|\hat{\Theta}_t^{(\lambda_{j_\ell})}\right\|} \xrightarrow{j\to\infty} \frac{\bar{\Theta}_t}{\left\|\bar{\Theta}_t\right\|}.$$

This concludes the proof. ■

Appendix B. Proofs for Section 3.2

In this section, we prove the lemmas used for the proof of Theorem 5.

The proof is based on an analysis of the dynamics of Algorithm 2 in the construction given in Theorem 5. We begin with several lemmas.

Proposition 1 (Projection onto $\mathcal{C}_1$ from the origin) *Let $a_1, b > 0$ with $b > a_1$. Consider*

$$\min_{x,y\in\mathbb{R}} \; x^2 + y^2 \quad \textit{s.t.} \quad a_1x^2 + by^2 \geq 1.$$

Then every minimizer satisfies $x = 0$ and $y = \pm b^{-1/2}$. In particular, the minimum value is $1/b$.

Proof. Let (x, y) be feasible. First we show that any minimizer must satisfy

$$a_1x^2 + by^2 = 1.$$

Assume in contradiction that $r = a_1x^2 + by^2 > 1$ and set $\alpha := \frac{1}{\sqrt{r}} \in (0, 1)$. Then $(\alpha x, \alpha y)$ satisfies the constraint with equality:

$$a_1(\alpha x)^2 + b(\alpha y)^2 = \alpha^2(a_1x^2 + by^2) = 1,$$

and the objective strictly decreases:

$$(\alpha x)^2 + (\alpha y)^2 = \alpha^2(x^2 + y^2) < x^2 + y^2,$$

implying that $r = 1$.

Second,on the boundary $a_1x^2 + by^2 = 1$ we have

$$y^2 = \frac{1 - a_1x^2}{b}, \qquad \text{with } 0 \leq a_1x^2 \leq 1 \iff 0 \leq x^2 \leq \frac{1}{a_1}.$$

Substituting into the objective gives

$$x^2 + y^2 = x^2 + \frac{1 - a_1x^2}{b} = \frac{1}{b} + x^2\Big(1 - \frac{a_1}{b}\Big).$$

Since $b > a_1$, we have $1 - \frac{a_1}{b} > 0$, so the right-hand side is strictly increasing in x^2. Therefore it is minimized when x^2 is minimal, i.e. when $x^2 = 0$. With this choice of $x = 0$, the active constraint implies $by^2 = 1$, hence $y = \pm b^{-1/2}$. This proves the claim. ∎

Proposition 2 (Projection onto $\mathcal{C}_2$ from the y–axis) *Let $a_2, b, c > 0$ with $c > a_2$ and let $u = (0, b^{-1/2}, 0)$. Consider the optimization problem*

$$\min_{(x,y,z)\in\mathbb{R}^3} F(x, y, z) := x^2 + (y - b^{-1/2})^2 + z^2 \quad \textit{s.t.} \quad a_2x^2 - cy^2 + cz^2 \geq 1.$$

Then every minimizer satisfies

$$x = 0, \qquad y = \pm\frac{1}{2\sqrt{b}}, \qquad z = \pm\sqrt{\frac{1}{4b} + \frac{1}{c}}.$$

In particular, the set of minimizers is

$$B := \left\{ \left(0,\ \pm\frac{1}{2\sqrt{b}},\ \pm\sqrt{\frac{1}{4b}+\frac{1}{c}}\right) \right\},$$

and every $\Theta^{(2)} = (0, y_0, z_0) \in B$ *satisfies*

$$z_0^2 - y_0^2 = \frac{1}{c}.$$

Proof. First, we show that for any solution of the problem, the constraint is active. Let (x, y, z) be feasible with $a_2x^2 - cy^2 + cz^2 > 1$. Scaling $(x, y, z) \mapsto \alpha(x, y, z)$ with

$$\alpha := \frac{1}{\sqrt{a_2x^2 - cy^2 + cz^2}} \in (0, 1)$$

preserves feasibility with equality and strictly decreases F since $F(\alpha\bar{\Theta}) = \alpha^2\bar{\Theta}$ for every $\bar{\Theta}$. Thus every minimizer satisfies

$$a_2x^2 - cy^2 + cz^2 = 1.$$

Next, from feasibility we have $cz^2 \geq 1 + cy^2 - a_2x^2$. Substituting into F yields

$$F(x, y, z) \geq x^2 + (y - b^{-1/2})^2 + \frac{1 + cy^2 - a_2x^2}{c} = \left(1 - \frac{a_2}{c}\right)x^2 + (y - b^{-1/2})^2 + y^2 + \frac{1}{c}.$$

Since function $(y - b^{-1/2})^2 + y^2$ is strictly convex and is minimized at $y = \frac{1}{2}b^{-1/2}$, where it takes value $\frac{1}{2b}$. Hence every feasible point satisfies

$$F(x, y, z) \geq \left(1 - \frac{a_2}{c}\right)x^2 + \frac{1}{2b} + \frac{1}{c}.$$

Equality is achieved when $x = 0$, $y = \frac{1}{2}b^{-1/2}$, and

$$cz^2 = 1 + cy^2 \quad \Longleftrightarrow \quad z^2 = \frac{1}{4b} + \frac{1}{c}.$$

This yields the claimed minimizers.

The final identity $z_0^2 - y_0^2 = \frac{1}{c}$ follows by direct substitution. ■

Proposition 3 (Projections onto $\mathcal{C}_1$) *Let* $\mathcal{C}_1 := \left\{(x, y, z) \in \mathbb{R}^3 :\ a_1x^2 + by^2 - bz^2 \geq 1\right\}, \mathcal{C}_2 := \left\{(x, y, z) \in \mathbb{R}^3 :\ a_2x^2 - cy^2 + cz^2 \geq 1\right\}$ *for* $a_1 := \frac{1}{9} + \varepsilon, \quad b := 1 - \varepsilon, a_2 := \frac{1}{9} - \varepsilon, \quad c := 1 + \varepsilon.$ *Fix* $\bar{\Theta}_t = (0, y_0, z_0) \in \mathcal{C}_2$ *and consider the projection problem*

$$\min_{(x,y,z)\in\mathbb{R}^3} \ F(x, y, z) := x^2 + (y - y_0)^2 + (z - z_0)^2 \quad \text{s.t.} \quad g(x, y, z) := a_1x^2 + by^2 - bz^2 - 1 \ \geq\ 0.$$

Then any global minimizer $(x^\star, y^\star, z^\star)$ *satisfies* $x^\star = 0$.

Proof. The objective F and the constraint function g are C^∞. Since $u \notin C_1$ and C_1 is closed, a minimizer exists and is not equal to u. Then, every minimizer satisfies $g(x^\star, y^\star, z^\star) = 0$. On $g = 0$, $\nabla g(x, y, z) = (2a_1 x, 2by, -2bz) \neq 0$ (otherwise $x = y = z = 0$ would contradict $g(0, 0, 0) = -1$). So LICQ holds and KKT is necessary.

Next, we calculate the KKT points. We use the Lagrangian.

$$\mathcal{L}(x, y, z, \lambda) := F(x, y, z) - \lambda g(x, y, z), \qquad \lambda \geq 0.$$

Stationarity gives

$$\begin{aligned} \partial_x \mathcal{L} = 2x - 2\lambda a_1 x = 0 \quad &\iff \quad x(1 - \lambda a_1) = 0, \\ \partial_y \mathcal{L} = 2(y - y_0) - 2\lambda b y = 0 \quad &\iff \quad y(1 - \lambda b) = y_0, \\ \partial_z \mathcal{L} = 2(z - z_0) + 2\lambda b z = 0 \quad &\iff \quad z(1 + \lambda b) = z_0. \end{aligned}$$

Primal feasibility and complementary slackness imply $g(x^\star, y^\star, z^\star) = 0$, thus, since $x^\star(1 - \lambda a_1) = 0$, either $x^\star = 0$ or $\lambda = 1/a_1$.

Branch I corresponds to $x = 0$. In this case, $y = \frac{y_0}{1 - \lambda b}, z = \frac{z_0}{1 + \lambda b}$, and $g = 0$ becomes $b(y^2 - z^2) = 1$.

Branch II corresponds to $\lambda = 1/a_1$. In this case, $y = \frac{y_0}{1 - b/a_1}, z = \frac{z_0}{1 + b/a_1}$, and $g = 0$ yields $x^2 = \frac{1 - b(y^2 - z^2)}{a_1}$.

Finally, it remains to show that any global minimizer will be in the first branch ($x^\star = 0$).

Branch II yields an objective value of

$$\|(x, y, z) - \bar{\Theta}_t\|^2 \geq \frac{1}{a_1},$$

Since $a_1 = \frac{1}{9} + \varepsilon \leq \frac{1}{9} + 0.01 < 0.13$,

$$\frac{1}{a_1} > 7.6.$$

In addition, branch I contains feasible point with small objective. Consider the feasible point with $x = 0$, $z = z_0$, and

$$y = \operatorname{sign}(y_0) \sqrt{z_0^2 + \frac{1}{b}},$$

which satisfies $by^2 - bz^2 = 1$ (hence lies on the boundary of $\mathcal{C}_1$). Then

$$\|(x, y, z) - \bar{\Theta}_t\|^2 = (\sqrt{z_0^2 + 1/b} - |y_0|)^2.$$

Using $z_0^2 \geq y_0^2 + 1/c$ and $1/b \leq 1/c + 2\varepsilon$, one gets

$$\sqrt{z_0^2 + 1/b} - |y_0| \;\leq\; \sqrt{y_0^2 + 1/c + 1/b} - |y_0| \;\leq\; \sqrt{2/b} \;\leq\; \sqrt{2.02},$$

hence

$$\|(x, y, z) - \bar{\Theta}_t\|^2 \;\leq\; 2.02.$$

Therefore Branch I attains objective ≤ 2.02, whereas Branch II is $\geq 1/a_1 > 7.6$. So every global minimizer lies in Branch I, and in particular has $x = 0$.

■

Proposition 4 (Projections onto $\mathcal{C}_2$) *Let* $\mathcal{C}_1 := \left\{(x,y,z)\in\mathbb{R}^3:\ a_1x^2+by^2-bz^2\geq 1\right\}, \mathcal{C}_2 := \left\{(x,y,z)\in\mathbb{R}^3:\ a_2x^2-cy^2+cz^2\geq 1\right\}$ *for* $a_1 := \frac{1}{9}+\varepsilon, \quad b := 1-\varepsilon, a_2 := \frac{1}{9}-\varepsilon, \quad c := 1+\varepsilon.$ *Fix* $\bar{\Theta}_t = (0, y_0, z_0) \in \mathcal{C}_2$ *and consider the projection problem onto* $\mathcal{C}_2$*:*

$$\min_{(x,y,z)\in\mathbb{R}^3} \ F(x,y,z) := x^2+(y-y_0)^2+(z-z_0)^2 \quad \text{s.t.} \quad g(x,y,z) := a_2x^2-cy^2+cz^2-1\geq 0.$$

Then any global minimizer $(x^\star, y^\star, z^\star)$ *satisfies* $x^\star = 0$*.*

Proof. The objective F and the constraint function g are C^∞. Since $\bar{\Theta}_t \in \mathcal{C}_1$, we have $by_0^2 - bz_0^2 \geq 1$, which implies $y_0^2 > z_0^2$. However, points in $\mathcal{C}_2$ with $x = 0$ satisfy $cz^2 - cy^2 \geq 1$, implying $z^2 > y^2$. Thus $\bar{\Theta}_t \notin \mathcal{C}_2$. Since $\mathcal{C}_2$ is closed, a minimizer exists. Then, every minimizer satisfies $g(x^\star, y^\star, z^\star) = 0$. On $g = 0$, $\nabla g(x,y,z) = (2a_2x, -2cy, 2cz) \neq 0$ (otherwise $x = y = z = 0$ would contradict $g(0,0,0) = -1$). So LICQ holds and KKT is necessary.

Next, we calculate the KKT points. We use the Lagrangian.

$$\mathcal{L}(x,y,z,\lambda) := F(x,y,z) - \lambda g(x,y,z), \qquad \lambda \geq 0.$$

Stationarity gives

$$\begin{aligned}
\partial_x\mathcal{L} &= 2x - 2\lambda a_2 x = 0 \quad \iff \quad x(1-\lambda a_2) = 0,\\
\partial_y\mathcal{L} &= 2(y-y_0) + 2\lambda c y = 0 \quad \iff \quad y(1+\lambda c) = y_0,\\
\partial_z\mathcal{L} &= 2(z-z_0) - 2\lambda c z = 0 \quad \iff \quad z(1-\lambda c) = z_0.
\end{aligned}$$

Primal feasibility and complementary slackness imply $g(x^\star, y^\star, z^\star) = 0$, thus, since $x^\star(1-\lambda a_2) = 0$, either $x^\star = 0$ or $\lambda = 1/a_2$.

Branch I corresponds to $x = 0$. In this case, $y = \frac{y_0}{1+\lambda c}, z = \frac{z_0}{1-\lambda c}$, and $g = 0$ becomes $c(z^2 - y^2) = 1$.

Branch II corresponds to $\lambda = 1/a_2$. In this case, $y = \frac{y_0}{1+c/a_2}, z = \frac{z_0}{1-c/a_2}$, and $g = 0$ yields $x^2 = \frac{1-c(z^2-y^2)}{a_2}$.

Finally, it remains to show that any global minimizer will be in the first branch ($x^\star = 0$). Branch II yields an objective value of

$$\|(x,y,z) - \bar{\Theta}_t\|^2 \geq \frac{1}{a_2}.$$

Since $a_2 = \frac{1}{9} - \varepsilon$,

$$\frac{1}{a_2} > 9.$$

In addition, branch I contains feasible point with small objective. Consider the feasible point with $x = 0$, $y = y_0$, and

$$z = \operatorname{sign}(z_0)\sqrt{y_0^2 + \frac{1}{c}},$$

which satisfies $cz^2 - cy^2 = 1$ (hence lies on the boundary of $\mathcal{C}_2$). Then

$$\|(x,y,z) - \bar{\Theta}_t\|^2 = (\sqrt{y_0^2 + 1/c} - |z_0|)^2.$$

Using $y_0^2 \geq z_0^2 + 1/b$ (from $\bar{\Theta}_t \in \mathcal{C}_1$), the term $\sqrt{y_0^2 + 1/c} - |z_0|$ is maximized when $z_0 = 0$ and $y_0^2 = 1/b$. Thus

$$\sqrt{y_0^2 + 1/c} - |z_0| \;\leq\; \sqrt{1/b + 1/c} \;\leq\; \sqrt{2.02},$$

hence

$$\|(x, y, z) - \bar{\Theta}_t\|^2 \;\leq\; 2.02.$$

Therefore Branch I attains objective ≤ 2.02, whereas Branch II is $\geq 1/a_2 > 9$. So every global minimizer lies in Branch I, and in particular has $x = 0$. ■

Now given the technical lemmas above we can finally prove Theorem 5.

Proof of Theorem 5. First for Part (a), let $a_1 := \frac{1}{9} + \varepsilon, b := 1 - \varepsilon, a_2 := \frac{1}{9} - \varepsilon, c := 1 + \varepsilon$. In this proof we refer to $\bar{\Theta} = (u, v)$ as a vector in $\mathbb{R}^4$ where its first and third entries are u and the two other entries are v. It holds that

$$\mathcal{C}_1 = \{\bar{\Theta} \in \mathbb{R}^4 \mid a_1\bar{\Theta}[1]^2 - a_1\bar{\Theta}[2]^2 + b\bar{\Theta}[3]^2 - b\bar{\Theta}[4]^2 \geq 1\},$$

$$\mathcal{C}_2 = \{\bar{\Theta} \in \mathbb{R}^4 \mid a_2\bar{\Theta}[1]^2 - a_2\bar{\Theta}[2]^2 - c\bar{\Theta}[3]^2 + c\bar{\Theta}_4^2 \geq 1\}.$$

Then, since $\varepsilon < 0.01$, $\bar{\Theta} = (4, 0, 0, 0) \in \mathcal{C}_1 \cap \mathcal{C}_2$.

For part (b), we first show that that for any sequence of projections, for every iteration t, $\bar{\Theta}_t[2] = 0$ by induction of (t). The basis of $t = 0$ follows by initialization in the origin. For the step assume that $\bar{\Theta}_t[2]$ and assume in contradiction that $\bar{\Theta}_{t+1}[2] = \delta \neq 0$. Let $\tilde{\Theta} = \bar{\Theta}_{t+1}[2] - \delta e_2$. It holds that $\tilde{\Theta}_2 = 0$, and for every $m \in 1, 2$

$$f(\tilde{\Theta}, \mathbf{x}_m) \geq \bar{\Theta}_{t+1} \geq 1,$$

and,

$$\|\tilde{\Theta} - \bar{\Theta}_t\| < \|\bar{\Theta}_t - \bar{\Theta}_{t+1}\|,$$

in a contradiction to the fact that $\bar{\Theta}_{t+1}$ is the projection of $\bar{\Theta}_{t+1}$ on $\mathcal{C}_{\tau_t}$.

Since $\bar{\Theta}_2$ remains zero during all iterations, we identify each $\bar{\Theta}$ with a vector $(x, y, z) \in \mathbb{R}^3$, where $\bar{\Theta}_t[1] = x, \bar{\Theta}_t[3] = y, \bar{\Theta}_t[4] = z$ and analyze the projection sequence in this 3-dimensional space, where

$$\mathcal{C}_1 := \Big\{(x, y, z) \in \mathbb{R}^3 : \; a_1x^2 + by^2 - bz^2 \geq 1\Big\}, \mathcal{C}_2 := \Big\{(x, y, z) \in \mathbb{R}^3 : \; a_2x^2 - cy^2 + cz^2 \geq 1\Big\}.$$

This will not change the dynamics of the projections.

For simplicity, we assume that $\tau_1 = 1$, the case where $\tau_1 = 2$ is analogous. Under this choice, we prove that $d\big(\bar{\Theta}_t, \mathcal{C}^\star\big) \geq 2.99$. For the first step, we minimize

$$x^2 + y^2 + z^2 \quad \text{s.t.} \quad a_1x^2 + by^2 - bz^2 \geq 1.$$

At any minimizer one must have $z = 0$: setting $z = 0$ strictly decreases the objective and increases feasibility. Thus we reduce to

$$\min_{x,y} x^2 + y^2 \quad \text{s.t.} \quad a_1x^2 + by^2 \geq 1.$$

Since $b > a_1 > 0$, by Proposition 1, the minimum is attained at $x = 0$ and $y^2 = 1/b$, giving

$$\bar{\Theta}_1 \in A := \{(0, \pm b^{-1/2}, 0)\}.$$

For the next projection, which is onto $\mathcal{C}_2$, by symmetry it suffices to project $\bar{\Theta} = (0, b^{-1/2}, 0)$ (the second case is analogous). We minimize

$$F(x, y, z) = x^2 + (y - b^{-1/2})^2 + z^2 \quad \text{s.t.} \quad a_2 x^2 - cy^2 + cz^2 \geq 1.$$

By Proposition 2, it follows that

$$\bar{\Theta}_2 \in B := \left\{ \left(0,\ \pm\frac{1}{2\sqrt{b}},\ \pm\sqrt{\frac{1}{4b} + \frac{1}{c}} \right) \right\}.$$

In particular, every $\bar{\Theta}_2 = (0, y_0, z_0) \in B$ satisfies

$$z_0^2 - y_0^2 = \frac{1}{c} \quad \Longrightarrow \quad z_0^2 \geq y_0^2 + \frac{1}{c}.$$

We continue that for all $t \geq 2$, $\bar{\Theta}_t[1] = 0$ by induction in t. Assume $\bar{\Theta}_t = (0, y_0, z_0)$ If $\bar{\Theta}_t \in \mathcal{C}_2$ and the projection is onto $\mathcal{C}_1$, then, $a_2 \cdot 0^2 - cy_0^2 + cz_0^2 \geq 1$, i.e. $z_0^2 \geq y_0^2 + \frac{1}{c}$. Consider projecting $\bar{\Theta}_t$ onto $\mathcal{C}_1$, i.e.

$$\min_{(x,y,z)} \ \|(x, y, z) - \bar{\Theta}_t\|^2 \quad \text{s.t.} \quad a_1 x^2 + by^2 - bz^2 \geq 1.$$

In Proposition 3, we show that any minimizer of this optimization problem satisfies $x = 0$.

The case where $\bar{\Theta}_t \in \mathcal{C}_1$ and the projection is onto $\mathcal{C}_2$, is analogous and is proved in Proposition 4.

As a result, we proved that for every t, $\bar{\Theta}_t[1] = 0$. It is left to bound from below the distance between $\bar{\Theta}_t$ and the intersection $\mathcal{C}^\star$.

Let $\bar{\Theta} \in \mathcal{C}^\star$. Multiply the $\mathcal{C}_1$ constraint by c and the $\mathcal{C}_2$ constraint by b and add:

$$c(a_1\bar{\Theta}[1]^2 - a_1\bar{\Theta}[2]^2 + b\bar{\Theta}[3]^2 - b\bar{\Theta}[4]^2) \ + \ b(a_2\bar{\Theta}[1]^2 - a_2\bar{\Theta}[2]^2 - c\bar{\Theta}[3]^2 + c\bar{\Theta}[4]^2) \ \geq \ c + b.$$

This gives,

$$(\frac{2}{9} + 2\varepsilon^2)\bar{\Theta}[1]^2 = (ca_1 + ba_2)\bar{\Theta}[1]^2 \geq (ca_1 + ba_2)\bar{\Theta}[1]^2 - (ca_1 + ba_2)\bar{\Theta}[2]^2 \ \geq \ b + c = 2.$$

Thus,

$$\bar{\Theta}[1]^2 \ \geq \ \frac{2}{\frac{2}{9} + 2\varepsilon^2} = \frac{1}{\frac{1}{9} + \varepsilon^2}.$$

But for all $t \geq 0$ we proved $\bar{\Theta}_t[1] = 0$, so for any $\bar{\Theta} \in \mathcal{C}^\star$,

$$\|\bar{\Theta}_t - \bar{\Theta}\|^2 \ \geq \ \bar{\Theta}[1]^2 \ \geq \ \frac{1}{\frac{1}{9} + \varepsilon^2}.$$

Therefore

$$d(\bar{\Theta}_t, \mathcal{C}^\star) \ \geq \ \sqrt{\frac{1}{\frac{1}{9} + \varepsilon^2}} > 2$$

as claimed. For the forgetting, we notice that for every t, there exists $m \in 1, 2$ such that $\bar{\Theta}_t \in \mathcal{C}_m$.

If $m = 1$, since $\bar{\Theta}_t$ is always on the boundary, it holds that

$$1 = y_1 f(\mathbf{x}_1; \bar{\Theta}_t) = \langle u^2, \mathbf{x}_1 \rangle - \langle v^2, \mathbf{x}_1 \rangle = (1 - \varepsilon) u[2]^2 - (1 - \varepsilon) v[2]^2.$$

This implies, $v[2]^2 = -\frac{1}{1-\varepsilon} + u[2]^2$. Then, for the other task, $m = 2$, it holds that,

$$\begin{aligned} y_2 f(\mathbf{x}_2; \bar{\Theta}_t) &= \langle u^2, \mathbf{x}_2 \rangle - \langle v^2, \mathbf{x}_2 \rangle \\ &= (-1 - \varepsilon) u[2]^2 - (-1 - \varepsilon) v[2]^2 \\ &= (-1 - \varepsilon) u[2]^2 + (1 + \varepsilon)(-\frac{1}{1 - \varepsilon} + u[2]^2) \\ &= -\frac{1 + \varepsilon}{1 - \varepsilon} \\ &\leq -1, \end{aligned}$$

and, $1 - y_2 f(\mathbf{x}_2; \bar{\Theta}_t) \geq 2$. This implies,

$$F_2(\bar{\Theta}_t) \geq (1 - y_2 f(\mathbf{x}_2; \bar{\Theta}_t)) \geq 2.$$

If $m = 2$, similarly, it holds that

$$1 = y_2 f(\mathbf{x}_2; \bar{\Theta}_t) = \langle u^2, \mathbf{x}_2 \rangle - \langle v^2, \mathbf{x}_2 \rangle = (-1 - \varepsilon) u[2]^2 - (-1 - \varepsilon) v[2]^2.$$

This implies, $u[2]^2 = -\frac{1}{1+\varepsilon} + v[2]^2$. Then, for the other task, $m = 1$, it holds that,

$$\begin{aligned} y_1 f(\mathbf{x}_1; \bar{\Theta}_t) &= \langle u^2, \mathbf{x}_1 \rangle - \langle v^2, \mathbf{x}_1 \rangle \\ &= (1 - \varepsilon) u[2]^2 - (1 - \varepsilon) v[2]^2 \\ &= -(1 - \varepsilon) v[2]^2 + (1 - \varepsilon)(-\frac{1}{1 + \varepsilon} + v[2]^2) \\ &= -\frac{1 - \varepsilon}{1 + \varepsilon} \\ &\leq -0.9, \end{aligned}$$

and, $1 - y_1 f(\mathbf{x}_1; \bar{\Theta}_t) \geq 1.9$. then,

$$F_1(\bar{\Theta}_t) \geq (1 - y_1 f(\mathbf{x}_1; \bar{\Theta}_t)) \geq 1.5.$$

■

Appendix C. Proofs for Section 3.3

In this section, we prove the lemmas used for the proof of Theorem 6.

C.1. Additional Regularity Conditions

We begin with defining more regularity condition that will be used to prove the required conditions and proving that several conditions imply other conditions. We begin with the following definition of MFCQ condition.

Definition 1 (MFCQ condition) *(e.g. Lyu and Li (2019)] Given m functions $h_j : \mathbb{R}^n \to \mathbb{R}$ and a Set $B = \{h_j \geq 0, \forall j \in [m]\} \subseteq \mathbb{R}^n$, we say the system defining set B satisfies the MFCQ at a feasible point $\bar{\mathbf{x}} \in B$ if there exists a vector $\mathbf{d} \in \mathbb{R}^n$ such that for all active constraints $j \in I_B(\mathbf{x}) = \{j \mid h_j(\bar{\mathbf{x}}) = 0\}$,*

$$\langle \nabla h_j(\bar{\mathbf{x}}), \mathbf{d} \rangle > 0.$$

In addition, if for all active constraints, $\langle \nabla h_j(\bar{\mathbf{x}}), \mathbf{d} \rangle > \gamma > 0$, we say that B satisfies MFCQ with margin γ.

Then, we define another regularity condition, named metric regularity and show that it is implied by MFCQ (Definition 1).

Definition 2 (Metric regularity, Example 9.44 of Rockafellar and Wets (1998)) *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable, let $D \subseteq \mathbb{R}^m$ be a closed convex set, and let $\bar{\mathbf{x}} \in F^{-1}(D)$. In addition,*

$$N_D(F(\bar{\mathbf{x}})) = \{\mathbf{v} \in \mathbb{R}^n \mid \langle \mathbf{v}, \mathbf{y} - F(\bar{\mathbf{x}}) \rangle \leq 0, \quad \forall \mathbf{y} \in D\}.$$

The constraint system $\{\mathbf{x} \mid F(\mathbf{x}) \in D\}$ is said to be metrically regular at $\bar{\mathbf{x}}$ if

$$\left[\boldsymbol{\lambda} \in N_D(F(\bar{\mathbf{x}})) \text{ and } \nabla F(\bar{\mathbf{x}})^\top \boldsymbol{\lambda} = 0 \right] \implies \boldsymbol{\lambda} = 0. \tag{13}$$

Lemma C.3 (MFCQ with margin implies Metric Regularity) *Let $\mathcal{C} := \{\mathbf{x} \in \mathbb{R}^n \mid h(\mathbf{x}) \geq 0\}$, where $h : \mathbb{R}^n \to \mathbb{R}^k$ is a smooth vector-valued function $h(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_k(\mathbf{x}))$. Assume that $\mathcal{C}$ satisfies the MFCQ with margin γ. Then, it satisfies also metric regularity.*

Proof. Let D be the k-dimensional positive orthant. Suppose $\boldsymbol{\lambda}$ satisfies the condition in (13) for D and h. Let d the vector from Definition 1. Then, $\boldsymbol{\lambda}[j] \leq 0$ for all j and

$$0 = \mathbf{d}^\top \nabla h(\bar{\mathbf{x}})^\top \boldsymbol{\lambda} = \sum_{j \in [k]} \boldsymbol{\lambda}[j] \nabla h_j(\bar{\mathbf{x}})^\top \mathbf{d}.$$

Using the MFCQ condition ($\nabla h_j(\bar{\mathbf{x}})^\top \mathbf{d} \geq \gamma$) and since $\boldsymbol{\lambda}[j] \leq 0$ (by the fact that $\boldsymbol{\lambda} \in N_D(h(\bar{\mathbf{x}}))$, it holds that

$$0 \leq \gamma \sum_{j \in [k]} \boldsymbol{\lambda}[j] \leq 0.$$

Then, it follows that $\boldsymbol{\lambda} = 0$. ■

Now, we prove that MFCQ and metric regularity of each set implies κ-linear regularity with respect to the intersection of the sets.

For the proof we use the following claim from Rockafellar and Wets (1998).

Proposition 4 (Property of metric regularity) *(Example 9.44 in Rockafellar and Wets (1998)) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a β smooth mapping and let $D \subseteq \mathbb{R}^m$ be a closed set. Define the constraint set $\mathcal{C} = F^{-1}(D) = \{\mathbf{x} \mid F(\mathbf{x}) \in D\}$. Let $\bar{\mathbf{x}} \in \mathcal{C}$. If the constraint set satisfies metric regularity at $\bar{\mathbf{x}}$ and MFCQ with margin γ at $\bar{\mathbf{x}}$. Then, for $\mu = \max_{\mathbf{y} \in N_D(F(\bar{\mathbf{x}})), \|\mathbf{y}\|=1} \frac{1}{\nabla F(\bar{\mathbf{x}})\mathbf{y}}$, for any $\mathbf{x}$ near $\bar{\mathbf{x}}$ in holds that*

$$d(\mathbf{x}, \mathcal{C}) \leq \mu d(F(\mathbf{x}), D)$$

Lemma C.5 (Metric regularity implies Linear Regularity) *Let M. For each $m \in M$, let the set $\mathcal{C}_m$ be defined by a system of inequalities, $\mathcal{C}_m := \{\mathbf{x} \in \mathbb{R}^n \mid h_m(\mathbf{x}) \geq 0\}$, where $h_m : \mathbb{R}^n \to \mathbb{R}^{n_m}$ is a smooth vector-valued function $h_m(\mathbf{x}) = (h_{m,1}(\mathbf{x}), \ldots, h_{m,n_m}(\mathbf{x}))$. Define the global feasible set as the intersection: $\mathcal{C}^\star := \bigcap_{m \in M} \mathcal{C}_m$. Fix $\bar{\mathbf{x}} \in \mathcal{C}^\star$. If every $h_{m,j}$ is G-Lipschitz and every $\mathcal{C}_m$ satisfies metric regularity at $\bar{\mathbf{x}}$. Then the collection $\{\mathcal{C}_m\}_{m \in M}$ is κ linearly regular at $\bar{\mathbf{x}}$ with $\kappa = \frac{G\|\mathbf{v}\|}{\gamma}$.*

Proof. For any unit vector $\boldsymbol{\lambda} \in N_D(h_m(\bar{\mathbf{x}}))$, where D is the n_m-dimensional positive orthant, we have $\boldsymbol{\lambda}[j] \leq 0$. Then, by the fact that $\|\mathbf{u}\| = \sup_{\mathbf{w} \neq 0} \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|}$, and MFCQ with margin γ and vector v,

$$\|\nabla h_m(\bar{\mathbf{x}})^\top \boldsymbol{\lambda}\| \geq \frac{|\langle \mathbf{v}, \nabla h_m(\bar{\mathbf{x}})^\top \boldsymbol{\lambda} \rangle|}{\|\mathbf{v}\|} = \frac{|\sum_j \boldsymbol{\lambda}[j] \langle \nabla h_{m,j}(\bar{\mathbf{x}}), \mathbf{v} \rangle|}{\|\mathbf{v}\|} \geq \frac{\gamma \sum |\boldsymbol{\lambda}[j]|}{\|\mathbf{v}\|} \geq \frac{\gamma}{\|\mathbf{v}\|}.$$

Thus, by Proposition 4, $\mu = \sup \frac{1}{\|\nabla h_m(\bar{\mathbf{x}})^\top \boldsymbol{\lambda}\|} \leq \frac{\|\mathbf{v}\|}{\gamma}$. Now let h be the function which is components are all of the constraints $\{h_{m,i}\}$. Than, since $\mathcal{C}^\star = h^{-1}(D)$, by Proposition 4, it holds that

$$d(\mathbf{x}, \mathcal{C}^\star) \leq \mu d(h(\mathbf{x}), D) = \mu \max_{(m,j)} \max\{0, -h_{m,j}(\mathbf{x})\} = \frac{\|\mathbf{v}\|}{\gamma} \max_{(m,j)} \max\{0, -h_{t,j}(\mathbf{x})\}. \tag{14}$$

Now, it holds that,

$$\max_{(m,j)} \max\{0, -h_{m,j}(\mathbf{x})\} = \max_{m \in M} \left(\max_{j=1\ldots n_m} \max\{0, -h_{m,j}(\mathbf{x})\} \right),$$

thus, for $a_m(\mathbf{x}) := \max_j \max\{0, -h_{m,j}(\mathbf{x})\}$, we get that,

$$d(\mathbf{x}, \mathcal{C}^\star) \leq \frac{\|\mathbf{v}\|}{\gamma} \max_{m \in M} a_m(\mathbf{x}).$$

Now, Let $m \in M$ and $\mathbf{x}$. If $\mathbf{x} \in \mathcal{C}_m$, then $a_m(\mathbf{x}) = 0$. Otherwise, let $\mathbf{p} = \boldsymbol{\Pi}_{\mathcal{C}_m}(\mathbf{x})$. By definition, $\mathbf{p} \in \mathcal{C}_m$, so $h_{m,j}(\mathbf{p}) \geq 0$ for all $j = 1 \ldots n_m$. For any specific component j:

$$-h_{m,j}(\mathbf{x}) \leq h_{m,j}(\mathbf{p}) - h_{m,j}(\mathbf{x}) \leq G\|\mathbf{p} - \mathbf{x}\| = Gd(\mathbf{x}, \mathcal{C}_m).$$

Since this holds for every j, it holds for the maximum:

$$a_m(\mathbf{x}) = \max_j \max\{0, -h_{m,j}(\mathbf{x})\} \leq Gd(\mathbf{x}, \mathcal{C}_m).$$

Combining all together, we get that,

$$d(\mathbf{x}, \mathcal{C}^\star) \leq \frac{\|\mathbf{v}\|}{\gamma} \max_m \left(G d(\mathbf{x}, \mathcal{C}_m)\right) = \frac{G\|\mathbf{v}\|}{\gamma} \max_m d(\mathbf{x}, \mathcal{C}_m).$$

■

Now we prove that metric regularity implies (ε, δ) regularity. In fact, we prove that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that (ε, δ) regularity is implied. In the proof, we use the following claim from Rockafellar and Wets (1998).

Proposition 6 (Theorem 6.31 in Rockafellar and Wets (1998)) *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a smooth vector valued mapping and let $D \subseteq \mathbb{R}^m$ be a closed set. Define for every set A*

$$N_A^{\text{prox}}(\mathbf{u}) = \{\mathbf{v} \in \mathbb{R}^n \mid \exists\, \sigma > 0 \text{ s.t. } \mathbf{u} \in \Pi_A(\mathbf{u} + \sigma \mathbf{v}).\}$$

Define the constraint set $\mathcal{C} = F^{-1}(D) = \{\mathbf{x} \mid F(\mathbf{x}) \in D\}$. Then for any $\bar{\mathbf{x}} \in \mathcal{C}$, if metric regularity holds at $\bar{\mathbf{x}}$, and D is convex, then,

$$N_{\mathcal{C}}^{\text{prox}}(\bar{\mathbf{x}}) = \{\nabla F(\bar{\mathbf{x}})^\top \boldsymbol{\lambda} \mid \boldsymbol{\lambda} \in N_D^{\text{prox}}(F(\bar{\mathbf{x}}))\}. \tag{15}$$

Lemma C.7 (Metric regularity implies (ε, δ)-Regularity) *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a smooth vector valued mapping, and let $D \subseteq \mathbb{R}^m$ be a closed convex set. Define the feasible set $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid F(\mathbf{x}) \in D\}$. Assume that metric regularity (Definition 2) and MFCQ holds at $\bar{\mathbf{x}} \in \mathcal{C}$ with margin γ. Then, for every $\varepsilon > 0$, $\mathcal{C}$ is $\left((\varepsilon, \frac{\varepsilon r}{\beta\|\bar{\mathbf{x}}\|}\right)$-regular at $\bar{\mathbf{x}}$.*

Proof. Let $\varepsilon > 0$ be given. We need to find $\delta > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{C} \cap \mathcal{B}_\delta(\bar{\mathbf{x}})$ and any proximal normal $\mathbf{v} \in N_{\mathcal{C}}^{\text{prox}}(\mathbf{x})$, the inequality $\langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle \leq \varepsilon \|\mathbf{v}\| \|\mathbf{y} - \mathbf{x}\|$ holds.

By Proposition 6, it holds that for $\mathbf{x}$ near $\bar{\mathbf{x}}$, any proximal normal $v \in N_{\mathcal{C}}^{\text{prox}}(\mathbf{x})$ can be represented as:

$$\mathbf{v} = \nabla F(\mathbf{x})^\top \boldsymbol{\lambda} \quad \text{with} \quad \boldsymbol{\lambda} \in N_D^{\text{prox}}(F(\mathbf{x})).$$

Since F is smooth vector valued mapping we have β' such that,

$$\|F(\mathbf{y}) - F(\mathbf{x}) - \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x})\| \leq \frac{\beta'}{2} \|\mathbf{y} - \mathbf{x}\|^2$$

By Proposition 6, and using the Cauchy-Schwarz inequality on the error term, it holds that:

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle &= \langle \nabla F(\mathbf{x})^\top \boldsymbol{\lambda}, \mathbf{y} - \mathbf{x} \rangle \\
&= \langle \boldsymbol{\lambda}, \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x}) \rangle \\
&= \langle \boldsymbol{\lambda}, F(\mathbf{y}) - F(\mathbf{x}) - (F(\mathbf{y}) - F(\mathbf{x}) - \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x})) \rangle \\
&= \langle \boldsymbol{\lambda}, F(\mathbf{y}) - F(\mathbf{x}) \rangle - \langle \boldsymbol{\lambda}, F(\mathbf{y}) - F(\mathbf{x}) - \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x}) \rangle \\
&\leq \underbrace{\langle \boldsymbol{\lambda}, F(\mathbf{y}) - F(\mathbf{x}) \rangle}_{\leq 0} + \|\boldsymbol{\lambda}\| \, \|F(\mathbf{y}) - F(\mathbf{x}) - \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x})\| \\
&\leq \frac{\beta' \|\boldsymbol{\lambda}\|}{2} \|\mathbf{y} - \mathbf{x}\|^2.
\end{aligned}$$

The first term is ≤ 0 because D is convex (and thus, $N_D((F(\mathbf{x})) = N_D^{\text{prox}}(F(\mathbf{x}))$, $\boldsymbol{\lambda} \in N_D(F(\mathbf{x}))$, and $F(\mathbf{y}) \in D$. For the second term,by MFCQ with margin γ and vector $\mathbf{u} = \bar{\mathbf{x}}$,

$$\|\mathbf{v}\| = \|\nabla F(\bar{\mathbf{x}})^\top \boldsymbol{\lambda}\| \geq \frac{|\langle \mathbf{u}, \nabla F(\bar{\mathbf{x}})^\top \boldsymbol{\lambda}\rangle|}{\|\mathbf{u}\|} = \frac{|\sum_j \boldsymbol{\lambda}[j] \langle \nabla F_j(\bar{\mathbf{x}}), \mathbf{u}\rangle|}{\|\mathbf{u}\|} \geq \frac{\gamma \sum |\boldsymbol{\lambda}[j]|}{\|\mathbf{u}\|} \geq \|\boldsymbol{\lambda}\| \frac{\gamma}{\|\mathbf{u}\|}.$$

This implies,

$$\|\boldsymbol{\lambda}\| \leq \frac{\|\mathbf{v}\| \|\mathbf{u}\|}{\gamma}.$$

Thus, combining all together, we get,

$$\langle \mathbf{v}, \mathbf{y} - \mathbf{x}\rangle \leq \frac{\|\mathbf{v}\| \|\mathbf{u}\|}{\gamma} \left(\frac{\beta'}{2} \|\mathbf{y} - \mathbf{x}\|^2 \right).$$

Thus, to satisfy the (ε, δ)-regularity condition, we denote $\kappa = \frac{\|\mathbf{u}\|}{\gamma} = \frac{\|\bar{\mathbf{x}}\|}{\gamma}$. We need to choose δ such that for every $\mathbf{x}, \mathbf{y} \in \mathcal{B}_\delta(\bar{\mathbf{x}})$

$$\frac{\kappa \beta'}{2} \|\mathbf{v}\| \|\mathbf{y} - \mathbf{x}\|^2 \leq \varepsilon \|\mathbf{v}\| \|\mathbf{y} - \mathbf{x}\|.$$

Dividing by $\|\mathbf{v}\| \|\mathbf{y} - \mathbf{x}\|$ (assuming nonzero, otherwise the inequality holds trivially), this requires $\frac{\kappa \beta'}{2} \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon$. Since $\mathbf{x}, \mathbf{y} \in \mathcal{B}_\delta(\bar{\mathbf{x}})$, we have $\|\mathbf{y} - \mathbf{x}\| \leq 2\delta$. Thus, it suffices to choose δ such that:

$$\frac{\kappa \beta'}{2}(2\delta) \leq \varepsilon \implies \delta \leq \frac{\varepsilon}{\kappa \beta'}.$$

With this δ, the set is (ε, δ)-regular. ∎

C.2. Proof of Theorem 6

In this section we prove Theorem 6. We begin with the proof of Proposition 12.

Proof of Proposition 12. The statement for cyclic ordering follows directly from Corollary 5.10 in Dao and Phan (2019). Here we prove the statement for random ordering. Fix $\mathbf{x}_{t-1} \in \mathcal{B}_{\delta/2}(\Theta^\star)$ and let $\bar{\mathbf{x}} \in \boldsymbol{\Pi}_{\mathcal{C}^\star}(\mathbf{x}_{t-1})$. Since $\Theta^\star \in \mathcal{C}^\star$, we have

$$\|\mathbf{x}_{t-1} - \bar{\mathbf{x}}\| = d(\mathbf{x}_{t-1}, \mathcal{C}^\star) \leq d(\mathbf{x}_{t-1}, \Theta^\star) \leq \delta/2\,.$$

By the triangle inequality,

$$d(\bar{\mathbf{x}}, \Theta^\star) \leq \|\mathbf{x}_{t-1} - \bar{\mathbf{x}}\| + d(\mathbf{x}_{t-1}, \Theta^\star) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta\,.$$

Thus, both $\mathbf{x}_{t-1}$ and $\bar{\mathbf{x}}$ lie in $\mathcal{B}_\delta(\Theta^\star)$, ensuring the local regularity assumptions apply.

Let $\tau(t) \in \{1, \ldots, M\}$ be arbitrary and let $\mathbf{x}_t \in \boldsymbol{\Pi}_{\mathcal{C}_{\tau(t)}}(\mathbf{x}_{t-1})$. By (ε, δ)–regularity and Proposition 3.5 of Dao and Phan (2019) (with $\boldsymbol{\lambda} = 1$), the exact projector $\boldsymbol{\Pi}_{\mathcal{C}_{\tau(t)}}$ is $(\Omega_{\tau(t)}, \gamma, \beta)$–quasi firmly Fejér monotone on $\mathcal{B}_{\delta/2}(\Theta^\star)$ with

$$\gamma = \frac{1}{1 - \varepsilon}, \qquad \beta = 1, \qquad \Omega_{\tau(t)} \triangleq \mathcal{C}_{\tau(t)} \cap \mathcal{B}_\delta(\Theta^\star)\,.$$

Since $\bar{\mathbf{x}} \in \mathcal{C}^\star \subseteq \mathcal{C}_{\tau(t)}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\delta(\Theta^\star)$, we have $\bar{\mathbf{x}} \in \Omega_{\tau(t)}$, and hence the quasi firm Fejér inequality yields:

$$\|\mathbf{x}_t - \bar{\mathbf{x}}\|^2 + \|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 \;\leq\; \gamma\|\mathbf{x}_{t-1} - \bar{\mathbf{x}}\|^2 = \gamma d^2(\mathbf{x}_{t-1}, \mathcal{C}^\star)\,.$$

Using $d(\mathbf{x}_t, \mathcal{C}^\star) \leq \|\mathbf{x}_t - \bar{\mathbf{x}}\|$ and $\|\mathbf{x}_{t-1} - \mathbf{x}_t\| = d(\mathbf{x}_{t-1}, \mathcal{C}_{\tau(t)})$ (exact projection), we obtain

$$d^2(\mathbf{x}_t, \mathcal{C}^\star) \;\leq\; \gamma d^2(\mathbf{x}_{t-1}, \mathcal{C}^\star) - d^2(\mathbf{x}_{t-1}, \mathcal{C}_{\tau(t)})\,.$$

Taking expectation with respect to $\tau(t)$ yields

$$\mathbb{E}_{\tau(t)}[d^2(\mathbf{x}_t, \mathcal{C}^\star)] \;\leq\; \gamma d^2(\mathbf{x}_{t-1}, \mathcal{C}^\star) - \mathbb{E}_{\tau(t)}[d^2(\mathbf{x}_{t-1}, \mathcal{C}_{\tau(t)})]\,.$$

By linear regularity,

$$\max_{1\leq m\leq M} d(\mathbf{x}_{t-1}, \mathcal{C}_m) \geq \frac{1}{\kappa} d(\mathbf{x}_{t-1}, \mathcal{C}^\star)\,,$$

and therefore

$$\mathbb{E}_{\tau(t)}[d^2(\mathbf{x}_{t-1}, \mathcal{C}_{\tau(t)})] = \frac{1}{M}\sum_{m=1}^{M} d^2(\mathbf{x}_{t-1}, \mathcal{C}_m) \;\geq\; \frac{1}{M\kappa^2} d^2(\mathbf{x}_{t-1}, \mathcal{C}^\star)\,.$$

Overall, we showed,

$$\mathbb{E}_{\tau(t)}[d^2(\mathbf{x}_t, \mathcal{C}^\star)] \;\leq\; \left(\gamma - \frac{1}{M\kappa^2}\right) d^2(\mathbf{x}_{t-1}, \mathcal{C}^\star)\,.$$

Substituting $\gamma = \frac{1}{1-\varepsilon}$ completes the proof. ■

Lemma C.8 *Let $f(\mathbf{x};\Theta)$ and $\mathcal{C}_1, \ldots \mathcal{C}_M; \mathcal{C}^\star$ as in Section 2. Let δ and Θ be such that $d(\Theta, \mathcal{C}^\star) \leq \delta$. Let $\Theta^\star = \mathbf{\Pi}_{\mathcal{C}^\star}(\Theta)$, and assume that $f(\mathbf{x}, \cdot)$ is G-Lipschitz in $\mathcal{B}_\delta(\Theta^\star) = \{\Theta \in \mathbb{R}^p \mid \|\Theta - \Theta^\star\| \leq \delta\}$. Then, it holds that,*

$$\max_m F_m(\Theta) \leq G d(\Theta, \mathcal{C}^\star)$$

Proof. Let $m \in [M]$ and $(\mathbf{x}_i^{(m)}, y_i^{(m)})$. By the fact that for every such data point, it holds that $y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta^\star) \geq 1$ and G-Lipschitzness, it holds that,

$$1 - y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta) \leq y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta^\star) - y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta) \leq G|y|\|\Theta^\star - \Theta\| = G d(\Theta, \mathcal{C}^\star)$$

Since this holds for any data point, it holds also for the maximal, thus,

$$\max_m F_m(\Theta) \triangleq \max_{i,m} \max\{0,\, 1 - y_i^{(m)} f(\mathbf{x}_i^{(m)}; \Theta)\} \leq G d(\Theta, \mathcal{C}^\star).$$

■

Now, we use the lemmas form Section C.1 to show that if for every $\mathbf{x}$, $f(\mathbf{x};\Theta)$ is G-Lipschitz, smooth and positively homogeneous functions to show that the regularity conditions mentioned in Proposition 12 hold.

We begin with the following lemma that shows that for homogeneous models with degree r, each feasible set satisfies the MFCQ condition at $\Theta^\star$ given in Definition 1 with margin r such the vector d in this definition is $\Theta^\star$.

Lemma C.9 (MFCQ of homogeneous models)
Consider a r-positively-homogeneous model $f(\cdot;\bar{\Theta}) : \mathcal{X} \to \mathbb{R}$ that is G-Lipschitz and β-smooth. Assume joint separability, i.e., nonempty intersection $\mathcal{C}^\star = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$. Let $\Theta^\star \in \mathcal{C}^\star = \bigcap_m \mathcal{C}_m$. Then, each $\mathcal{C}_m$ satisfies the MFCQ condition at $\Theta^\star$ with margin r for $d = \Theta^\star$. In addition, also $\mathcal{C}^\star$ satisfies the MFCQ condition at $\Theta^\star$ with margin r for $\mathbf{d} = \Theta^\star$.

Proof. Let m and let $I_m := \{i : y_i^{(m)} f(\Theta^\star, \mathbf{x}_i^{(m)}) = 1\} = \{i : y_i^{(m)} f(\Theta^\star, \mathbf{x}_i^{(m)}) = 1\}$. If I_m is empty, the definition holds trivially. Otherwise, since $f(x;\Theta)$ is smooth and positively homogeneous of degree r, Euler's theorem for homogeneous functions yield, for every $\mathbf{x}$, $\langle \nabla_\Theta f(\mathbf{x};\Theta^\star), \Theta^\star \rangle = r f(\mathbf{x};\Theta^\star)$. Then, for every $i \in I_m$, $h_{m,i} = y_i^{(m)} f(\Theta^\star, \mathbf{x}_i^{(m)}) - 1 = 0$, satisfies,

$$\langle \nabla h_{m,i}(\Theta^\star), \Theta^\star \rangle = y_i^{(m)} \langle \nabla_\Theta f(\Theta^\star, \mathbf{x}_i^{(m)}), \Theta^\star \rangle = y_i^{(m)} r f(\Theta^\star, \mathbf{x}_i^{(m)}) = r > 0.$$

Then, the MFCQ condition holds for $\mathbf{d} = \Theta^\star$ and margin r. The proof for $\mathcal{C}^\star$ is identical. ■

Now we turn to prove (ε, δ) regularity.

Lemma C.10 ((ε, δ) regularity of homogeneous models) *Consider a r-positively-homogeneous model $f(\cdot;\bar{\Theta}) : \mathcal{X} \to \mathbb{R}$ that is G-Lipschitz and β-smooth. Assume joint separability, i.e., nonempty intersection $\mathcal{C}^\star = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$. Let $\Theta^\star \in \mathcal{C}^\star = \bigcap_t \mathcal{C}_m$. Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that every set $\mathcal{C}_m$ is (ε, δ)-regular at $\Theta^\star$.*

Proof. Let $m \in M$ By Lemma C.9 MFCQ condition holds with margin r. In addition, by Lemma C.3 metric regularity also holds. Then, by Lemma C.7, the lemma follows. ■

Lemma C.11 (κ-linear regularity of homogeneous models)
Consider a r-positively-homogeneous model $f(\cdot;\bar{\Theta}) : \mathcal{X} \to \mathbb{R}$ that is G-Lipschitz and β-smooth. Assume joint separability, i.e., nonempty intersection $\mathcal{C}^\star = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$. Let $\Theta^\star \in \mathcal{C}^\star = \bigcap_m \mathcal{C}_m$. Then, if f is G-Lipschitz, the collection $\{\mathcal{C}_m\}_{m=1}^M$ satisfy κ-linear regularity for $\kappa = \frac{G\|\Theta^\star\|}{r}$ around $\Theta^\star$.

Proof. Let $m \in M$ By Lemma C.9, MFCQ condition holds for the vector $\|\Theta^\star\|$ with margin r. In addition, by Lemma C.3 metric regularity also holds. Then, by Lemma C.5, the lemma follows. ■

Now we can turn to the proof of Theorem 6.

Proof of Theorem 6. For the cyclic ordering, let ε be such that $\frac{1}{(1-\varepsilon)^{M-1}} = 1 + \frac{1}{2(M-1)\kappa^2}$. Then by Lemma C.10 each set $\mathcal{C}_m$ is (ε, δ) regular. In addition, by Lemma C.11, the collection $\{\mathcal{C}_m\}_{m=1}^M$ satisfy κ-linear regularity for $\kappa = \frac{G\|\Theta^\star\|}{r}$ in a around $\Theta^\star$. As a result, the conditions of Proposition 12 holds in a δ-neighborhood of $\Theta^\star$. By the choice of ε we get, for the cyclic order, that for any t, $d(\bar{\Theta}_{t+M}, \mathcal{C}^\star) \;\le\; \rho\, d(\bar{\Theta}_t, \mathcal{C}^\star)$

$$\begin{aligned}
\rho &= \left(1 - \frac{1}{2(M-1)\kappa^2}\right)^{1/2(M-1)} \\
&= \left(1 - \frac{1}{2(M-1)(G\|\Theta^\star\|/r)^2}\right)^{1/2(M-1)} \\
&= \left(1 - \frac{r^2}{2(M-1)G^2\|\Theta^\star\|^2}\right)^{1/2(k-1)} \\
&\le \left(\exp\left(-\frac{r^2}{2MG^2\|\Theta^\star\|^2}\right)\right)^{1/2M} \\
&= \exp\left(-\frac{r^2}{4M^2G^2\|\Theta^\star\|^2}\right),
\end{aligned}$$

Thus, for $M \mid k$

$$d(\bar{\Theta}_k, \mathcal{C}^\star) \le \exp\left(-\frac{kr^2}{4MG^2\|\Theta^\star\|^2}\right) d(\bar{\Theta}_0, \mathcal{C}^\star)$$

For the random order, using Proposition 12, we get that for ε such that such that $\frac{1}{1-\varepsilon} = 1 + \frac{1}{2M\kappa^2}$,

$$\mathbb{E}_{\tau(t)}\left[d^2(\bar{\Theta}_t, \mathcal{C}^\star)\right] \;\le\; \left(1 - \frac{1}{2M\kappa^2}\right) d^2(\bar{\Theta}_{t-1}, \mathcal{C}^\star).$$

Thus, for $k = cM$,

$$\begin{aligned}
\mathbb{E}_{\tau}\left[d^2(\bar{\Theta}_k, \mathcal{C}^\star)\right] \;&\le\; \left(1 - \frac{1}{2M\kappa^2}\right)^k d^2(\bar{\Theta}_0, \mathcal{C}^\star) \\
&\le \exp\left(-\frac{k}{2M\kappa^2}\right) d^2(\bar{\Theta}_0, \mathcal{C}^\star) \\
&= \exp\left(-\frac{kr^2}{2MG^2\|\Theta^\star\|^2}\right) d^2(\bar{\Theta}_0, \mathcal{C}^\star)
\end{aligned}$$

By Jensen inequality,

$$\mathbb{E}_{\tau}\left[d(\bar{\Theta}_k, \mathcal{C}^\star)\right] \;\le\; \sqrt{\mathbb{E}_{\tau}\left[d^2(\bar{\Theta}_k, \mathcal{C}^\star)\right]} \le \exp\left(-\frac{kr^2}{4MG^2\|\Theta^\star\|^2}\right) d(\bar{\Theta}_0, \mathcal{C}^\star)$$

For the forgetting, the statement follows by Lemma C.8. ■

Appendix D. Proofs for Section 4

D.1. Proof of Theorem 13

Proof of Theorem 13. As in classification, we employ the theory of Γ-convergence. Let $G_t^{(\lambda)}(\Theta) := \frac{1}{\lambda}\mathcal{L}_{\tau(t)}(\Theta) + \|\Theta - \Theta_{t-1}\|^2$. It holds that $\Theta \in \operatorname{argmin}_{\Theta \in \mathcal{C}_{\tau(t)}} \mathcal{L}_{\tau(t)}(\Theta) + \lambda\|\Theta - \Theta_{t-1}\|^2$ if and only if $\Theta \in \operatorname{argmin}_{\Theta \in \mathcal{C}_{\tau(t)}} G_t^{(\lambda)}(\Theta)$. We show that $G_t^{(\lambda)}$ Γ-converges to the following funcation as $\lambda \downarrow 0$:

$$G_t(\Theta) = \begin{cases} \|\Theta - \Theta_{t-1}\|^2 & \text{if } \Theta \in \mathcal{C}_{\tau(t)} \\ +\infty & \text{otherwise.} \end{cases}$$

First, for the Liminf inequality, let $\Theta_t^{(\lambda)} \to \Theta$. By the definition of the feasible set, if $\Theta \notin \mathcal{C}_{\tau(t)}$, then $\mathcal{L}_m(\Theta) > 0$. Let $\mathcal{L}_m(\Theta) = \delta > 0$. Since f is continuous, $\mathcal{L}_m$ is continuous.

Therefore, there exists a neighborhood around Θ where $\mathcal{L}_m(\cdot) > \delta/2$. Since $\Theta_t^{(\lambda)} \to \Theta$, for sufficiently small λ, $\Theta_t^{(\lambda)}$ lies in this neighborhood. Thus:

$$G_t^{(\lambda)}(\Theta_t^{(\lambda)}) = \frac{1}{\lambda}\mathcal{L}_m(\Theta_t^{(\lambda)}) + \left\|\Theta_t^{(\lambda)} - \Theta_{t-1}\right\|^2 \geq \frac{\delta}{2\lambda} + 0.$$

Taking the limit as $\lambda \downarrow 0$, we have $\lim G_t^{(\lambda)}(\Theta_t^{(\lambda)}) = +\infty$. Since $G_t(\Theta) = +\infty$, the inequality holds.

Otherwise, if $\Theta \in \mathcal{C}_{\tau(t)}$. Since $\mathcal{L}_m(\cdot) \geq 0$. we have:

$$G_t^{(\lambda)}(\Theta_t^{(\lambda)}) \geq \left\|\Theta_t^{(\lambda)} - \Theta_{t-1}\right\|^2.$$

Thus, by continuity, as $\Theta_t^{(\lambda)} \to \Theta$, the term $\|\Theta_t^{(\lambda)} - \Theta_{t-1}\|^2 \to \|\Theta - \Theta_{t-1}\|^2$ and $\liminf_{\lambda \downarrow 0} G_t^{(\lambda)}(\Theta_t^{(\lambda)}) \geq \|\Theta - \Theta_{t-1}\|^2 = G_t(\Theta)$.

For the Limsup property, We need to show that for any Θ, there exists a sequence $\Theta_t^{(\lambda)} \to \Theta$ such that $\limsup_{\lambda \downarrow 0} G_t^{(\lambda)}(\Theta_t^{(\lambda)}) \leq G_t(\Theta)$. We choose the constant sequence $\Theta_t^{(\lambda)} = \Theta$ for all λ. If $\Theta \notin \mathcal{C}_{\tau(t)}$, $G_t(\Theta) = +\infty$, so the inequality $\limsup G_t^{(\lambda)}(\Theta_t^{(\lambda)}) \leq \infty$ is trivially satisfied. Otherwise, if $\Theta \in \mathcal{C}_{\tau(t)}$, then $\mathcal{L}_m(\Theta) = 0$ and $G_t^{(\lambda)}(\Theta) = G_t(\Theta)$.

Finally for the boundedness of level sets, the function $G_t^{(\lambda)}(\Theta)$ is bounded below by $\|\Theta - \Theta_{t-1}\|^2$. Since the sublevel sets of this lower bound are compact, the sequence of functionals is equi-coercive.

The theorem follows by Part (b) of Lemma 3 and the fact that Θ_t is a minimizer of G_t. ■

D.2. Proof of Lemma 14

Lemma D.1 (Catastrophic Forgetting in Squared Models) *Consider a squared model $f(\mathbf{x};\bar{\Theta})$ in $d=2$ parameterized by $\bar{\Theta}=(\mathbf{u},\mathbf{v})$ where $\mathbf{u},\mathbf{v}\in\mathbb{R}^2$. Fix $\varepsilon\in(0,0.01)$ and define two tasks*

$$\mathbf{X}^{(1)}=\left[\mathbf{x}_1^\top\right]=\left[\left(\tfrac{1}{9}+\varepsilon,\ 1-\varepsilon\right)\right],\qquad \mathbf{X}^{(2)}=\left[\mathbf{x}_2^\top\right]=\left[\left(\tfrac{1}{9}-\varepsilon,\ -1-\varepsilon\right)\right],\qquad \mathbf{y}^{(1)}=\mathbf{y}^{(2)}=(1)\ .$$

Then,

(a) The induced feasible sets $\mathcal{C}_1=\{\bar{\Theta}\in\mathbb{R}^4 \mid f(\mathbf{x}_1;\bar{\Theta})=y_1\}$, $\mathcal{C}_2=\{\bar{\Theta}\in\mathbb{R}^4 \mid f(\mathbf{x}_2;\bar{\Theta})=y_2\}$ have a nonempty intersection, i.e., $\mathcal{C}^\star:=\mathcal{C}_1\cap\mathcal{C}_2\neq\varnothing$.

(b) Algorithm 2 satisfies, for all $t\geq 0$, $d\big(\bar{\Theta}_t,\mathcal{C}^\star\big)\ \geq\ 2$ and $\max_m F_m(\bar{\Theta}_t)\geq 3$.

Proof of Lemma D.1. We now consider the dynamics where the task sets are defined by the same constraints as Theorem 5, except there is equality instead of inequality. However, we observe that the sequential projections onto these equality sets generate the exact same trajectory as the projections onto the inequality sets $\mathcal{C}_m=\{\bar{\Theta}\mid f(x_m;\bar{\Theta})\geq 1\}$.

This follows immediately from Proposition 1, Proposition 2, and Proposition 3 in Appendix B. Specifically, those lemmas prove that for every projection step involved in the inequality-constrained dynamics, the unique minimizer satisfies the constraint with strict equality (*i.e.,* $f(\mathbf{x}_m;\bar{\Theta})=1$). Since the solution to the relaxed problem (≥ 1) lies on the boundary, it is necessarily the solution to the equality-constrained problem. Thus, the sequence of iterates $\{\bar{\Theta}_t\}$ is identical in both settings. By those lemmas we get that for every t it holds for $\bar{\Theta}_t=(\mathbf{u}_t,\mathbf{v}_t)$ that $\mathbf{u}_t=0$.

Now, let $\bar{\Theta}=(u,v)\in\mathcal{C}^\star$. Multiply the $\mathcal{C}_1$ constraint by c and the $\mathcal{C}_2$ constraint by b and add:

$$\begin{aligned}&c(a_1\Theta^\star[1]^2-a_1\Theta^\star[2]^2+b\Theta^\star[3]^2-b\bar{\Theta}[4]^2)\\&+\ b(a_2\Theta^\star[1]^2-a_2\Theta^\star[2]^2-c\Theta^\star[3]^2+c\bar{\Theta}[4]^2)\\&=c+b=2.\end{aligned}$$

This gives,

$$\left(\frac{2}{9}+2\varepsilon^2\right)(\Theta^\star[1]^2-\Theta^\star[2]^2)=(ca_1+ba_2)\Theta^\star[1]^2-(ca_1+ba_2)\Theta^\star[2]^2=2.$$

In particular,

$$\Theta^\star[1]^2\geq\Theta^\star[1]^2-\Theta^\star[2]^2=\frac{1}{\frac{1}{9}+\varepsilon^2}.$$

But for all $t\geq 0$ we proved $\bar{\Theta}_t[1]=0$, so for any $\bar{\Theta}\in\mathcal{C}^\star$,

$$\|\bar{\Theta}_t-\bar{\Theta}\|^2\ \geq\ \Theta^\star[1]^2\ \geq\ \frac{1}{\frac{1}{9}+\varepsilon^2}.$$

Therefore,

$$d(\bar{\Theta}_t,\mathcal{C}^\star)\ \geq\ \sqrt{\frac{1}{\frac{1}{9}+\varepsilon^2}}>2$$

as claimed. For the loss, we notice that for every t, there exists $m\in 1,2$ such that $\bar{\Theta}_t\in\mathcal{C}_m$.

If $m = 1$, since $\bar{\Theta}_t$ is always on the boundary, it holds that

$$1 = y_1 = f(\mathbf{x}_1; \bar{\Theta}_t) = \langle \mathbf{u}_t^2, \mathbf{x}_1 \rangle - \langle \mathbf{v}_t^2, \mathbf{x}_1 \rangle = (1-\varepsilon)\mathbf{u}_t[2]^2 - (1-\varepsilon)\mathbf{v}_t[2]^2.$$

This implies, $\mathbf{v}_t[2]^2 = -\frac{1}{1-\varepsilon} + \mathbf{u}_t[2]^2$. Then, for the other task, $m = 2$, it holds that,

$$\begin{aligned} f(\mathbf{x}_2; \bar{\Theta}_t) &= (-1-\varepsilon)\mathbf{u}_t[2]^2 - (-1-\varepsilon)\mathbf{v}_t[2]^2 \\ &= (-1-\varepsilon)\mathbf{u}_t[2]^2 + (1+\varepsilon)(-\frac{1}{1-\varepsilon} + \mathbf{u}_t[2]^2) \\ &= -\frac{1+\varepsilon}{1-\varepsilon} \\ &\le -1\,, \end{aligned}$$

This implies $F_2(\bar{\Theta}_t) = |f(\mathbf{x}_2; \bar{\Theta}_t) - y_2|^2 \ge 4$.

If $m = 2$, similarly, it holds that

$$1 = y_2 = f(\mathbf{x}_2; \bar{\Theta}_t) = \langle u^2, \mathbf{x}_2 \rangle - \langle v^2, \mathbf{x}_2 \rangle = (-1-\varepsilon)\mathbf{u}[2]^2 - (-1-\varepsilon)\mathbf{v}[2]^2.$$

This implies, $\mathbf{u}[2]^2 = -\frac{1}{1+\varepsilon} + \mathbf{v}[2]^2$. Then, for the other task, $m = 1$, it holds that,

$$\begin{aligned} f(\mathbf{x}_1; \bar{\Theta}_t) &= (1-\varepsilon)\mathbf{u}_t[2]^2 - (1-\varepsilon)\mathbf{v}_t[2]^2 \\ &= -(1-\varepsilon)\mathbf{v}_t[2]^2 + (1-\varepsilon)(-\frac{1}{1+\varepsilon} + \mathbf{v}_t[2]^2) \\ &= -\frac{1-\varepsilon}{1+\varepsilon} \\ &\le -0.9\,, \end{aligned}$$

and, $F_1(\bar{\Theta}_t) = |f(\mathbf{x}_1; \bar{\Theta}_t) - y_1|^2 \ge 3$. ■

D.3. Proof of Theorem 15

In this section, we analyze Algorithm 2 for the multi-task regression problem. Each task m imposes a set of equality constraints:

$$h_{i,m}(\Theta) = f(\mathbf{x}_i^{(m)}; \Theta) - y_i^{(m)} = 0.$$

Let $\mathcal{C}_{m,i} = \{\Theta \in \mathbb{R}^p \mid h_{i,m}(\Theta) = 0\}$ be the feasible set for a single data point. The global solution set is the intersection $\mathcal{C}^\star = \bigcap_{m,i} \mathcal{C}_{m,i}$. We discuss the β-smooth and G-Lipschitz case.

Now we turn to prove κ-linear regularity for homogeneous models with inequality constraints.

Lemma D.2 (κ-Linear Regularity of Homogeneous Models) *Consider a r-positively-homogeneous model $f(\cdot; \bar{\Theta}) : \mathcal{X} \to \mathbb{R}$ that is G-Lipschitz and β-smooth. Assume joint separability, i.e., nonempty intersection $\mathcal{C}^\star = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$. Let $y_{\min} = \min_{m,i} |y_i^{(m)}| \neq 0$ and $\Theta^\star \in \mathcal{C}^\star$. Let ε holding $\frac{1}{(1-\varepsilon)^{M-1}} = 1 + \frac{r^2}{2(M-1)G^2\|\Theta^\star\|^2}$. Then, for $\delta = \frac{\varepsilon r y_{\min}}{\beta\|\Theta^\star\|}$, the collection $\{\mathcal{C}_m\}_{m=1}^M$ satisfies κ-linear-regularity on $\mathcal{B}_\delta(\Theta^\star)$ with constant $\kappa = \frac{G\|\Theta^\star\|}{r y_{\min}}$.*

Proof. We define the constraint functions as $h_{i,m}(\Theta) = f(\mathbf{x}_i^{(m)}; \Theta) - y_i^{(m)} = 0$. By Euler's Homogeneous Function Theorem,

$$\langle \nabla h_{i,m}(\Theta^\star), \Theta^\star \rangle = \langle \nabla_\Theta f(\mathbf{x}_i^{(m)}; \Theta^\star), \Theta^\star \rangle = r f(\mathbf{x}_i^{(m)})(\Theta^\star)\Theta^\star = r y_i^{(m)} \geq r y_{\min}.$$

and, if h_m is the function $h_m = (h_{1,m} \ldots, h_{n_m,m})$,

$$\|\nabla h_m(\Theta^\star)^\top \boldsymbol{\lambda}\| \geq \frac{|\langle \Theta^\star, \nabla h_m(\Theta^\star)^\top \boldsymbol{\lambda} \rangle|}{\|\Theta^\star\|} = \frac{|\sum_j \boldsymbol{\lambda}[j] \langle \nabla h_j(\Theta^\star), \Theta^\star \rangle|}{\|\Theta^\star\|} \geq \frac{r y_{\min} \sum |\boldsymbol{\lambda}[j]|}{\|\Theta^\star\|}.$$

Thus, if $\nabla h_m(\Theta^\star)^\top \boldsymbol{\lambda} = 0$, this implies $\boldsymbol{\lambda} = 0$, thus, metric regularity holds. As a result, by Proposition 4, for x near $\Theta^\star$ it holds for

$$\mu = \max_{\boldsymbol{\lambda} \in N_D(h_m(\Theta^\star)), \|\boldsymbol{\lambda}\|=1} \frac{1}{\nabla h_m(\Theta^\star)\boldsymbol{\lambda}} \leq \frac{\|\Theta^\star\|}{y_{\min} r},$$

that

$$d(\Theta, \mathcal{C}^\star) \leq \mu \max_{i,m} |h_{i,m}(\Theta)| \leq \frac{\|\Theta^\star\|}{r y_{\min}} \max_{i,m} |h_{i,m}(\Theta)|.$$

Now, by local G-Lipschitzness, for every i, m it holds that,

$$|h_{i,m}(\Theta)| = |h_{i,m}(\Theta) - h_{i,m}(\boldsymbol{\Pi}_{\mathcal{C}_m}(\Theta)| \leq G d(\Theta, \mathcal{C}_m)$$

Since this holds also for the maximal example, we get,

$$d(\Theta, \mathcal{C}^\star) \leq \frac{G\|\Theta^\star\|}{r y_{\min}} \max_{m \in M} d(\Theta, \mathcal{C}_m).$$

■

For (ε, δ)-regularity, we have the following lemma.

Lemma D.3 (**(ε,δ)-Regularity of Homogeneous Models)** *Consider an r-positively-homogeneous model $f(\cdot;\bar{\Theta}) : \mathcal{X} \to \mathbb{R}$ that is G-Lipschitz and β-smooth. Assume joint separability, i.e., nonempty intersection $\mathcal{C}^\star = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M \neq \varnothing$. Let $y_{\min} = \min_{m,i} |y_i^{(m)}| \neq 0$ and $\Theta^\star \in \mathcal{C}^\star$. Then, for every $\varepsilon > 0$ and $\delta = \frac{\varepsilon r y_{\min}}{\beta \|\Theta^\star\|}$, every feasible set $\mathcal{C}_m$ is (ε,δ)-regular at $\Theta^\star$.*

Proof. Let $\mathcal{C}_m = \{\Theta \mid h_m(\Theta) = 0\}$. Since h_m is smooth, its gradient $\nabla h_m(\Theta)$ is continuous.

To prove (ε,δ)-regularity, we need to show that for any $\mathbf{x},\mathbf{y} \in \mathcal{C}_m \cap \mathcal{B}_\delta(\Theta^\star)$ and any proximal normal vector $\mathbf{v} \in N^{prox}_{\mathcal{C}_m}(\Theta^\star)$, the following holds:

$$\langle \mathbf{v}, \mathbf{y}-\mathbf{x}\rangle \leq \varepsilon \|\mathbf{v}\|\|\mathbf{y}-\mathbf{x}\|.$$

Let $\mathbf{v} \in N^{prox}_{\mathcal{C}_m}(\Theta^\star)$ be such a vector. By Euler's Homogeneous Function Theorem,

$$\langle \nabla h_{i,m}(\Theta^\star), \Theta^\star\rangle = \langle \nabla_\Theta f(\mathbf{x}_i^{(m)};\Theta^\star), \Theta^\star\rangle = r f(\mathbf{x}_i^{(m)})(\Theta^\star)\Theta^\star = r y_i^{(m)} \geq r y_{\min}. \tag{16}$$

If h_m is the function $h_m = (h_{1,m}\dots, h_{n_m,m})$,

$$\|\nabla h_m(\Theta^\star)^\top \boldsymbol{\lambda}\| \geq \frac{|\sum_j \boldsymbol{\lambda}[j]\langle \nabla h_{j,m}(\Theta^\star), \Theta^\star\rangle|}{\|\Theta^\star\|} \geq \frac{r y_{\min}\sum|\boldsymbol{\lambda}[j]|}{\|\Theta^\star\|} \geq \|\boldsymbol{\lambda}\|\frac{r y_{\min}}{\|\Theta^\star\|}. \tag{17}$$

Since h_m is smooth vector valued mapping we have β' such that,

$$\|h_m(\mathbf{y}) - h_m(\mathbf{x}) - \nabla h_m(\mathbf{x})(\mathbf{y}-\mathbf{x})\| \leq \frac{\beta'}{2}\|\mathbf{y}-\mathbf{x}\|^2$$

By Proposition 6, and using the Cauchy-Schwarz inequality on the error term, it holds that:

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{y}-\mathbf{x}\rangle &= \langle \nabla h_m(\mathbf{x})^\top \boldsymbol{\lambda}, \mathbf{y}-\mathbf{x}\rangle \\
&= \langle \boldsymbol{\lambda}, \nabla h_m(\mathbf{x})(\mathbf{y}-\mathbf{x})\rangle \\
&= \langle \boldsymbol{\lambda}, h_m(\mathbf{y}) - h_m(\mathbf{x}) - (h_m(\mathbf{y}) - h_m(\mathbf{x}) - \nabla h_m(\mathbf{x})(\mathbf{y}-\mathbf{x}))\rangle \\
&= \langle \boldsymbol{\lambda}, h_m(\mathbf{y}) - h_m(\mathbf{x})\rangle - \langle \boldsymbol{\lambda}, h_m(\mathbf{y}) - h_m(\mathbf{x}) - \nabla h_m(\mathbf{x})(\mathbf{y}-\mathbf{x})\rangle \\
&\leq \langle \boldsymbol{\lambda}, h_m(\mathbf{y}) - h_m(\mathbf{x})\rangle + \|\boldsymbol{\lambda}\|\,\|h_m(\mathbf{y}) - h_m(\mathbf{x}) - \nabla h_m(\mathbf{x})(\mathbf{y}-\mathbf{x})\| \\
&= \frac{\beta'\|\boldsymbol{\lambda}\|}{2}\|\mathbf{y}-\mathbf{x}\|^2.
\end{aligned}$$

Thus, by Eq. (17), to satisfy the (ε,δ)-regularity condition, we denote $\kappa = \frac{\|\Theta^\star\|}{r y_{\min}}$. We need to choose δ such that for every $\mathbf{x},\mathbf{y} \in \mathcal{B}_\delta(\bar{\mathbf{x}})$

$$\frac{\kappa\beta'}{2}\|\mathbf{v}\|\|\mathbf{y}-\mathbf{x}\|^2 \leq \varepsilon\|\mathbf{v}\|\|\mathbf{y}-\mathbf{x}\|.$$

Dividing by $\|\mathbf{v}\|\|\mathbf{y}-\mathbf{x}\|$ (assuming nonzero, otherwise the inequality holds trivially), this requires $\frac{\kappa\beta'}{2}\|\mathbf{y}-\mathbf{x}\| \leq \varepsilon$. Since $\mathbf{x},\mathbf{y} \in \mathcal{B}_\delta(\bar{\mathbf{x}})$, we have $\|\mathbf{y}-\mathbf{x}\| \leq 2\delta$. Thus, it suffices to choose δ such that:

$$\frac{\kappa\beta'}{2}(2\delta) \leq \varepsilon \implies \delta \leq \frac{\varepsilon}{\kappa\beta'}.$$

With this δ, the set is (ε,δ)-regular. ∎

Proof of Theorem 15. The proof follows identically to the classification case, utilizing the established (ε, δ)-regularity and κ-linear regularity for the equality-constrained sets. For the forgetting, the statement follows by the fact that, for every m,

$$\begin{aligned} F_m(\bar{\Theta}_t) &= \max_i \left| f(\mathbf{x}_i^{(m)}; \Theta_t^\star) - y_i^{(m)} \right|^2 = \max_i \left| f(\mathbf{x}_i^{(m)}; \bar{\Theta}_t) - f(\mathbf{x}_i^{(m)}; \mathbf{\Pi}_{\mathcal{C}^\star}(\Theta^\star) \right|^2 \\ &\leq G^2 d^2(\bar{\Theta}_t, \mathcal{C}^\star). \end{aligned}$$

■