Task Similarity Matters: Greedy Orderings in Continual Linear Regression

Anonymous Author(s)
Affiliation
Address
email

Abstract

We analyze task ordering strategies in continual learning for realizable linear regression. We focus on task orderings that greedily maximize dissimilarity between consecutive tasks, a concept briefly explored in prior work but still surrounded by open questions. Using tools from the Kaczmarz method literature, we formalize these orderings and develop both geometric and algebraic intuitions around them. We show empirically that, under random data, greedy orderings lead to faster convergence of the loss compared to random orderings. In a simplified setting, we prove bounds on the loss and establish optimality guarantees for greedy orderings. However, we also construct an adversarial task sequence that exploits high dimensionality to induce maximal forgetting under greedy orderings—an effect to which random orderings are notably more robust. Altogether, our findings advance the theoretical understanding of task orderings in continual learning, offer new insights into Kaczmarz methods, and provide a foundation for future research.

1 Introduction

Continual learning is a subfield of machine learning in which a learner is exposed to tasks or datasets sequentially. In such setups, only a single task is typically accessible at any given time—due to, for instance, data retention or privacy constraints, computational limitations, or the temporal nature of the environment. While much work in continual learning focuses on mitigating forgetting or improving transfer, the role of the *task ordering* remains underexplored.

Understanding how task order affects learning—and what characterizes optimal orderings—is important for both theoretical and practical reasons. Such understanding can illuminate failure modes, clarify the interplay between forgetting and transfer, and guide the design of continual environments and algorithms. Furthermore, it can enable active control over task sequences in settings that permit it, situating the problem at the intersection of continual learning, multitask learning, curriculum learning, and active learning. This line of inquiry raises compelling questions with significant computational and financial implications in the era of large language models and foundation models:

- *What constitutes an "optimal" task ordering?*
- *Is it better to learn when adjacent tasks are similar or dissimilar?*
- *Should we hope to outperform random orderings?*
- *What are the failure modes of "greedy" orderings?*

One compelling direction in the continual learning literature is the design of task orderings informed by task similarity. This idea has appeared in several earlier works, with varying degrees of emphasis and differing motivations [e.g., 34, 43, 48, 54, 55, 61, 68, 69]. Most closely related to our work is Bell and Lawrence [10], who were among the first to explicitly and systematically examine such orderings

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

in continual learning. They hypothesized that optimal performance would arise when adjacent tasks are *similar*. Surprisingly, they empirically found the opposite—orderings with *dissimilar* adjacent tasks led to better performance. More recently, Li and Hiratani [47] reached a similar conclusion and further proposed arranging tasks from the least to the most "typical". While these studies are thought-provoking, they are either empirical [10, 54, 55], based on restrictive data assumptions [47, 48], or focused solely on task-incremental settings [61], with some of their findings appearing inconclusive or contradictory. This underscores the need for a more rigorous *theoretical* understanding.

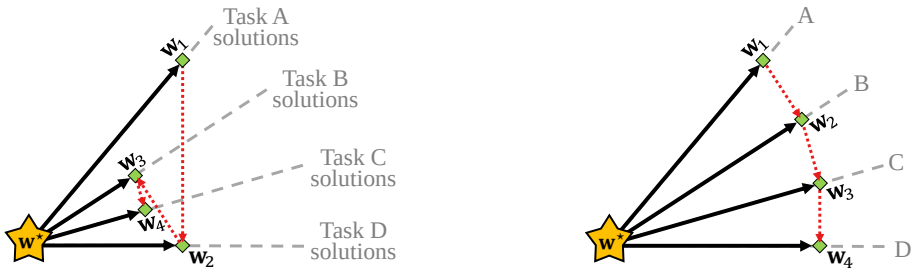
To this end, we aim to formalize "similarity-guided" orderings through *greedy* task selection, leveraging tools and formulations from related fields. We begin with a projection-based perspective on continual learning, following prior work [24, 25]. We then introduce two greedy orderings—Maximum Distance (MD) and Maximum Residual (MR)—commonly studied in the Kaczmarz [56, 57] and projection onto convex sets literatures [2, 31]. Using these orderings, we develop geometric, analytical, and empirical insights into the advantages of greediness, and derive motivating guarantees in a special case. The resulting intuition is illustrated in Figure 1 below.

Focusing on single-pass task orderings (with no repetitions), we present an adversarial task collection where greediness fails due to the problem's dimensionality, in stark contrast to random orderings. Surprisingly, we find that this does *not* extend to greedy orderings with repetition—proving a dimensionality-independent upper bound on their forgetting. Moreover, in a slight contrast to the common wisdom in random orderings—where with-replacement orderings usually perform better than ones without replacement—we show that in greedy orderings, repetition empirically performs better on simple data. Finally, we present a hybrid scheme combining greedy and random orderings, demonstrating some of its empirical and analytical benefits.

We hope that the theoretical foundations—perspectives, tools, and findings—laid out in this paper will inspire future practical and theoretical work on similarity-guided task orderings.

Summary of our contributions.

1. We formalize similarity-guided orderings in continual linear regression via greedy strategies, drawing on tools and intuitions from projection and Kaczmarz literature (Section 3).
2. In experiments on randomly-generated isotropic data and highly-correlated data, we show that greedy orderings converge faster than random orderings (Section 4.1).
3. We prove optimality and convergence guarantees for high-rank tasks (Section 4.2).
4. For general-rank data in high dimensions, we construct an adversarial failure mode where greedy orderings provably induce maximal forgetting (Section 5.1).
5. In contrast, greedy orderings *with repetition* provably converge, regardless of dimensionality (Section 5.2).
6. We combine greedy and random orderings into a hybrid strategy that performs well empirically and inherits the bounds of random orderings, avoiding greedy failure modes (Section 5.3).



(a) A greedy ordering with *dissimilar* adjacent tasks.

(b) A greedy ordering with *similar* adjacent tasks.

Figure 1: **Intuition.** Consider a collection of jointly-realizable linear regression tasks (*e.g.,* `A,B,C,D`). Each task has an affine solution space (*e.g.,* where $\mathbf{X}_{\mathtt{A}}\mathbf{w} = \mathbf{y}_{\mathtt{A}}$), and $\mathbf{w}_\star$ is an "offline solution" at the intersection of all tasks. Employing a projection perspective on learning in continual models [24, 25], we see that transitions between *dissimilar* tasks (e.g., `A`→`D`→`B`→`C`) intuitively lead to faster convergence toward the intersection compared to transitions between *similar* tasks (e.g., `A`→`B`→`C`→`D`).

2 Setting: Continual linear regression

We focus on continual linear regression, common in theoretical continual learning [e.g., 5, 21, 24, 26, 28, 36, 48, 60]. This setting, though simple, already gives rise to key continual learning phenomena, such as complex interactions between forgetting, task similarity, and overparameterization [see 29].

Notations. We reserve bold symbols for matrices and vectors, *e.g.,* $\mathbf{X}, \mathbf{w}$. We use $\|\cdot\|$ to denote the Euclidean norm of vectors and the spectral (L2) norm of matrices. $\mathbf{X}^+$ denotes the Moore–Penrose pseudoinverse of a matrix. Finally, we denote $[n] = 1, \ldots, n$.

Formally, the learner is given access to a *task collection* of T linear regression tasks, *i.e.,* $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_T, \mathbf{y}_T)$ where $\mathbf{X}_m \in \mathbb{R}^{n_t \times d}, \mathbf{y}_m \in \mathbb{R}^{n_t}$. We denote the data "radius" by $R \triangleq \max_{m\in[T]} \|\mathbf{X}_m\|$. For k iterations, the learner sequentially learns the tasks according to a *task ordering* $\tau : [k] \to [T]$, which—as this paper shows—can be crucial in continual learning.

Scheme 1 Continual linear regression (to convergence)

Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
  $\mathbf{w}_t \leftarrow$ Start from $\mathbf{w}_{t-1}$ and minimize the current task's loss $\mathcal{L}_{\tau(t)}(\mathbf{w}) \triangleq \left\|\mathbf{X}_{\tau(t)}\mathbf{w} - \mathbf{y}_{\tau(t)}\right\|^2$ with (S)GD to convergence
Output $\mathbf{w}_k$

We assume throughout the paper that there exist *offline solutions* that perfectly solve all T tasks *jointly*. This is a common assumption[1] in many theoretical continual learning papers, which facilitates the analysis [e.g., 24, 25, 26, 29, 40, 42]. Moreover, it is a reasonable assumption in highly overparameterized models and is thus linked to the linear dynamics of deep neural networks in the neural tangent kernel (NTK) regime [see 16, 38].

Assumption 2.1 (Joint Linear Realizability)**.** Assume the intersection of *all* individual task solution subspaces is nonempty, *i.e.,* $\mathcal{W}_\star \triangleq \bigcap_{m=1}^T \mathcal{W}_m \triangleq \bigcap_{m=1}^T \left\{\mathbf{w} \in \mathbb{R}^d \,\middle|\, \mathbf{X}_m\mathbf{w} = \mathbf{y}_m\right\} \neq \varnothing$.

We focus on the offline solution with the minimum norm, often linked to improved generalization.

Definition 2.2 (Minimum-Norm Offline Solution)**.** Denote specifically $\mathbf{w}_\star \triangleq \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_\star} \|\mathbf{w}\|$.

We follow previous prominent theoretical work [e.g., 21, 24, 25, 26, 29] and study the model's ability to not "forget" previously seen *training* data (as opposed to generalization performance). This focus isolates continual dynamics from statistical effects that also arise in non-continual, stationary settings.

Definition 2.3 (Average loss)**.** The average (training) loss of an individual task $m \in [T]$ is defined as $\mathcal{L}_m(\mathbf{w}) \triangleq \|\mathbf{X}_m\mathbf{w} - \mathbf{y}_m\|^2$. The training loss we analyze is the average across all T tasks. In our realizable setting, it takes the following form:

$$\mathcal{L}(\mathbf{w}_k) \triangleq \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2} \cdot \frac{1}{T}\sum_{m=1}^T \mathcal{L}_m(\mathbf{w}) = \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2} \cdot \frac{1}{T}\sum_{m=1}^T \|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2 \,,$$

where we also normalize by the generally unavoidable scaling factors $\|\mathbf{w}_\star\|$ and $R \triangleq \max_{m\in[T]} \|\mathbf{X}_m\|$.

Remark 2.4 (Forgetting vs. loss)*.* Another common quantity in the theoretical continual learning literature is the forgetting, defined as the loss *degradation* at iteration k across *only* previously seen tasks, *i.e.,* $\frac{1}{k}\sum_{t=1}^k \left(\mathcal{L}_{\tau(t)}(\mathbf{w}_k) - \mathcal{L}_{\tau(t)}(\mathbf{w}_t)\right)$. In our realizable setting it reduces to $\frac{1}{k}\sum_{t=1}^k \mathcal{L}_{\tau(t)}(\mathbf{w}_k)$, or $\frac{1}{k}\sum_{t=1}^k \left\|\mathbf{X}_{\tau(t)}\mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2$. Since we mostly focus on single-pass orderings, where each task is seen exactly once, forgetting coincides with average loss. Thus, to ease presentation, we analyze only the average loss, though our analysis still applies to forgetting at the end of the task sequence.

[1] Another trend in continual learning theory is to assume an underlying linear model, like we do, but allow an additive label noise [e.g., 20, 28, 45, 46, 48, 85]. However, this comes at the cost of strong assumptions on the *features—e.g.,* commutable covariance matrices or i.i.d. features across tasks. To some extent, the analysis in Section 5.1 of Evron et al. [24] suggests that, under such assumptions, task ordering has limited impact. Thus, it may not be a suitable starting point for studying similarity-guided orderings, in contrast to our assumption.

Another insightful quantity is the distance to $\mathbf{w}_\star$.

Definition 2.5 (Distance to the offline solution)**.** After k iterations, the (squared) distance is,

$$D^2(\mathbf{w}_k) = \frac{1}{\|\mathbf{w}_\star\|^2} \cdot \|\mathbf{w}_k - \mathbf{w}_\star\|^2 \ .$$

This distance upper bounds the loss, as can be shown using simple norm inequalities.

Proposition 2.6 (Linking the Quantities)**.** *After k iterations of Scheme 1 on jointly realizable tasks, the loss is upper bounded by the distance to the offline solution.*

$$\mathcal{L}(\mathbf{w}_k) = \frac{1}{\|\mathbf{w}_\star\|^2 R^2} \cdot \frac{1}{T} \sum_{m=1}^{T} \|\mathbf{X}_m (\mathbf{w}_k - \mathbf{w}_\star)\|^2 \leq \frac{1}{\|\mathbf{w}_\star\|^2} \cdot \|\mathbf{w}_k - \mathbf{w}_\star\|^2 = D^2(\mathbf{w}_k) \,.$$

In some cases, the distance can remain large while the loss (and forgetting) vanishes, showing that converging to $\mathbf{w}_\star$ is not mandatory for continual learning [24]. Focusing on the loss paves the way to universal convergence, independent of the problem's complexity, *e.g.,* its condition number [24, 65].

Geometric interpretation to learning. In each iteration of Scheme 1, the learner minimizes the squared loss of the current task to convergence.[2] Each iterate $\mathbf{w}_t$ of this scheme above is known [24] to implicitly follow the following closed-form update rule,

$$\mathbf{w}_t = \mathbf{X}^+_{\tau(t)} \mathbf{y}_{\tau(t)} + \left(\mathbf{I}_d - \mathbf{X}^+_{\tau(t)} \mathbf{X}_{\tau(t)}\right) \mathbf{w}_{t-1} \,. \tag{1}$$

Conveniently, in our realizable setting, this update rule admits an intuitive geometric interpretation.

Evron et al. [24] identified the orthogonal projection operator,

$$\mathbf{P}_m \triangleq \mathbf{I}_d - \mathbf{X}^+_m \mathbf{X}_m$$

which we use for mathematical purposes only (Scheme 1 never explicitly computes pseudoinverses or SVDs). Under the realizability assumption $\mathbf{y}_{\tau(t)} = \mathbf{X}_{\tau(t)} \mathbf{w}_\star$. We plug it into Eq. (1) and obtain:

$$\begin{aligned} \mathbf{w}_t &= \mathbf{X}^+_{\tau(t)} \mathbf{X}_{\tau(t)} \mathbf{w}_\star + \left(\mathbf{I}_d - \mathbf{X}^+_{\tau(t)} \mathbf{X}_{\tau(t)}\right) \mathbf{w}_{t-1} \\ \mathbf{w}_t - \mathbf{w}_\star &= \mathbf{P}_{\tau(t)} \left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right) \,. \end{aligned} \tag{2}$$

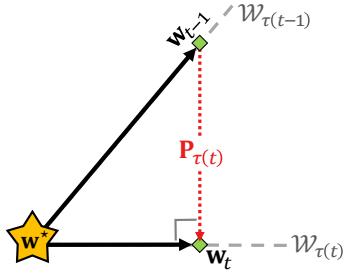


Figure 2: **Projection dynamics.**

Geometrically, $\mathbf{w}_{t-1}$ is projected by an affine projection onto the solution space of task $\tau(t)$.

This projection-based perspective has proven useful in prior theoretical work on continual learning [24, 25]. In the next section, we adopt this viewpoint to build intuition about greedy orderings.

3 Greedy task orderings: A formal approach and intuition

As discussed in the introduction (Section 1), the learning order plays a crucial role in the dynamics of many machine learning settings. This phenomenon has also been observed in continual learning, both analytically and empirically. Several works have proposed leveraging "similarity-aware" task orderings, in which dissimilar tasks are placed consecutively. However, the existing literature still lacks the rigor and analytical tools needed to fully understand such orderings. To address this gap, this section draws on connections between continual linear regression and other research areas to formalize greedy task orderings and develop the mathematical tools necessary to study them.

Geometric intuition. As illustrated in Figure 2, the projection perspective allows us to decompose $\|\mathbf{w}_t - \mathbf{w}_\star\|^2$ using projection properties and the Pythagorean theorem as:

$$\begin{aligned} \|\mathbf{w}_t - \mathbf{w}_\star\|^2 &= \|\mathbf{w}_{t-1} - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 \\ &= \|\mathbf{w}_{t-1} - \mathbf{w}_\star\|^2 - \left\|(\mathbf{I} - \mathbf{P}_{\tau(t)})(\mathbf{w}_{t-1} - \mathbf{w}_\star)\right\|^2 . \end{aligned} \tag{3}$$

Thus, to try and minimize $\|\mathbf{w}_t - \mathbf{w}_\star\|^2$, one could greedily maximize $\left\|(\mathbf{I} - \mathbf{P}_{\tau(t)})(\mathbf{w}_{t-1} - \mathbf{w}_\star)\right\|^2$.

[2] This simplifies the analysis; other analytical approaches exist, *e.g.,* a fixed number of steps per task [40].

This has inspired a myriad of studies on Kaczmarz[3] or projection methods [e.g., 2, 12, 22, 57] that employed a greedy ordering policy in the following spirit.

Definition 3.1 (Maximum Distance Ordering)**.** Greedily maximize the distance between iterates:

$$\tau_{\text{MD}}(t) = \text{argmax}_{m\in[T]\setminus\tau_{\text{MD}}(1:t-1)} \left\|\left(\mathbf{I}-\mathbf{P}_m\right)\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2, \quad \forall t\in[T] \ ,$$

where $\tau_{\text{MD}}\left(1:t-1\right) \triangleq \{\tau_{\text{MD}}\left(1\right),\ldots,\tau_{\text{MD}}\left(t-1\right)\}$.[4]

Our earlier Figure 1a illustrates the MD ordering and how it leads to faster convergence to $\mathbf{w}_\star$.

Distance and task similarity. The distance between iterates $\mathbf{w}_{\tau(t-1)}$ and $\mathbf{w}_{\tau(t)}$ reflects some angle between the affine solution subspaces of their corresponding tasks—and more generally—relates to the principal angles between these subspaces [24]. These angles can be used to quantify task similarity, as illustrated in the setting of Section 4.2 and Figure 1.

An alternative greedy ordering found in the literature is the Maximum Residual ordering [e.g., 2, 30, 57, 82]. This rule is easier to compute in full, or to estimate using a small validation set.

Definition 3.2 (Maximum Residual Ordering)**.** Greedily select the task exhibiting the greatest error:

$$\tau_{\text{MR}}(t) \ = \text{argmax}_{m\in[T]\setminus\tau_{\text{MR}}(1:t-1)} \left\|\mathbf{X}_m\mathbf{w}_{t-1}-\mathbf{y}_m\right\|^2, \quad \forall t\in[T] \ .$$

Notice that the MD and MR orderings are related since $\mathbf{X}_m = \mathbf{X}_m\mathbf{X}_m^+\mathbf{X}_m = \mathbf{X}_m\left(\mathbf{I}-\mathbf{P}_m\right)$, and,

$$\left\|\mathbf{X}_m\mathbf{w}_{t-1}-\mathbf{y}_m\right\|^2 = \left\|\mathbf{X}_m\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2 \leq \left\|\mathbf{X}_m\right\|^2\left\|\left(\mathbf{I}-\mathbf{P}_m\right)\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2 .$$

Single-pass greedy orderings. Throughout most of this paper, we focus on "single-pass" greedy orderings, where each task is encountered exactly once. Although disallowing repetitions departs slightly from the motivating literature on Kaczmarz and projection methods, it can be seen as more natural in continual learning settings [see 47]. Moreover, even in curriculum or multitask learning scenarios, restricting each task to a single pass may help reduce training costs. In Section 5.2, we discuss and empirically compare the effect of repetitions under different orderings.

Computational tractability of greedy policies. As explained above, the benefits of greedy orderings are quite intuitive. The cost of computing the greedy rules in Definition 3.1 and Eq. (4), of course, introduces a tradeoff between convergence rate and overall computational cost. Before continuing our investigation of these orderings, we briefly address their computational feasibility.

(i) **Estimation:** Greedy rules can often be estimated efficiently in practical scenarios. For example, the maximum residual rule (Definition 3.2) requires the current loss of each available task. This quantity can be estimated using a small validation set or approximated via dimensionality reduction techniques, as done in the Kaczmarz literature [22]. In deep networks, computing that rule requires only forward passes and may reduce the number of gradient steps—thereby lowering overall time and memory costs by limiting costly backward passes [37].

(ii) **Heuristics**: The greedy rules in our paper rely on residuals to quantify the similarity between the current task and the remaining ones. This approach is exemplified in Figure 1 and Eq. (5) of Section 4.2, and is related to principal angles between subspaces [see 24]. Alternatively, one could utilize heuristic notions of task similarity. Such "metrics" can be predefined [43] or computed using hessians [10], zero-shot performance [47], or task embeddings [1, 54].

(iii) **Structured tasks**: If each step updates relatively few residuals (*e.g.,* in a Kaczmarz setting with sparse columns and rows, or more generally with many orthogonal pairs of rows), only few residuals must be recomputed, reducing the overall cost [57].

(iv) **A theoretical tool:** We employ greedy orderings as an "ideal" *proxy* for understanding optimal and similarity-guided task orderings. This allows us not only to derive convergence results, but also to explore failure modes and examine key aspects of such strategies.

[3]The Kaczmarz method [23, 41], further explained in Section 6, iteratively solves a linear system of equations.

[4]In practice, the MD rule is easy to compute for rank-1 tasks, since it reduces to $\frac{1}{\|\mathbf{x}_m\|^2}\left\|\mathbf{x}_m^\top\mathbf{w}_{t-1}-y_m\right\|^2$. In higher ranks, this rule is harder to compute—but the MR rule, presented next, is feasible.

4 Benefits of greedy orderings

Existing rates. As discussed, greedy strategies have a long-standing history in related areas. They have been employed in the Kaczmarz method [57, 58] and its block variants [52, 56, 79, 82, 84], using deterministic [57] or probabilistic [7, 8, 74, 83] selection rules. These works—like much of the Kaczmarz literature—primarily analyze the distance to the solution $\mathbf{w}_\star$ (Definition 2.5). In contrast, our focus, and that of related continual learning literature [e.g., 24, 25, 40, 42], centers on convergence of the loss (Definition 2.3). Nevertheless, existing analyses and convergence rates for greedy Kaczmarz methods *already* illustrate the potential advantages of greedy selection, particularly in light of the relationship between distance and loss (Proposition 2.6).

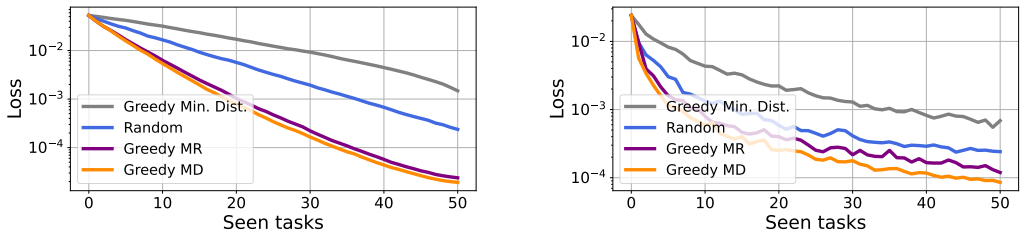
A natural competitor to greedy strategies is the random strategy, uniformly sampling tasks (rows or blocks in the Kaczmarz context) from the task collection $[T]$. That is,

$$\tau_{\text{Unif}}(1), \ldots, \tau_{\text{Unif}}(k) \sim \text{Uniform}\left([T]\right) . \tag{4}$$

In the aforementioned literature, the greedy orderings provably achieve better upper bounds on the distance to $\mathbf{w}_\star$, compared to random orderings, across many regimes.

4.1 Illustrative example: Randomly generated tasks

Next, we compare the performance of different task ordering strategies on synthetic data. The feature matrices, *i.e.,* $\mathbf{X}_1, \ldots, \mathbf{X}_T$, are drawn from either an isotropic Gaussian distribution or from an anisotropic Gaussian distribution with a diagonal covariance matrix and exponentially decaying eigenvalues. We compare the two "dissimilarity-maximizing" greedy strategies (MD, MR) to the random ordering (Eq. (4)) and a complementary, minimum distance, strategy. Our results show that transitioning between dissimilar tasks consistently outperforms both random and similar transitions across the presented settings and additional data-generating parameter regimes in App. B.



(a) Isotropic Data. (b) High-Correlation Data.

Figure 3: **Comparison of orderings under random data.** Sampled $T = 50$ tasks of rank $r = 10$ in $d = 100$ dimensions, from Gaussian distributions with (a) identical and (b) exponentially-decaying eigenvalues. The Maximum Distance and Residual strategies (MD, MR) outperform the random and similarity-maximizing strategies. Full details, including more combinations of T, d, r, are provided in App. B, showing these conclusions extend to other parameter regimes.
Similarly, *all* figures in the main body are complemented by supplemental figures in the appendices, covering broader regimes of T, d, and r.

4.2 Provable benefits for high rank, "nearly determined" tasks

To further motivate greedy orderings, we analyze a simple setup where each task's data matrix is of nearly full rank, *i.e.,* $\text{rank}(\mathbf{X}_m) = d - 1$, $\forall m \in [T]$. Even in this setup ($d = 2$), it has been shown that arbitrary orderings of $T \to \infty$ may lead to *catastrophic* forgetting, or maximal losses [24].

In this setup, each projection operator $\mathbf{P}_m = \mathbf{I}_d - \mathbf{X}_m^+ \mathbf{X}_m$ is rank 1 and can be expressed as $\mathbf{P}_m = \mathbf{v}_m \mathbf{v}_m^\top$ for some unit vector $\mathbf{v}_m \in \mathbb{R}^d$. Then, the Maximum Distance rule (Def. 3.1) can be rewritten (see Eq. (6) in App. C) as,

$$\tau_{\text{MD}}(t) = \text{argmin}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \left(\mathbf{v}_m^\top \mathbf{v}_{\tau(t-1)}\right)^2 , \tag{5}$$

where we define $\mathbf{v}_{\tau(0)} \triangleq \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|} (\mathbf{w}_0 - \mathbf{w}_\star)$. Notice that $\left(\mathbf{v}_m^\top \mathbf{v}_{\tau(t-1)}\right)^2 = \cos^2(\theta_{m,\tau(t-1)})$ quantifies similarity between task m and the former task $\tau(t-1)$.

Optimality of Greedy Orderings. Earlier in Eq. (3), we motivated the MD ordering as greedily maximizing the decrease in $\|\mathbf{w}_t - \mathbf{w}_\star\|$. Does this guarantee a minimal distance $\|\mathbf{w}_T - \mathbf{w}_\star\|$ at the end of the sequence? Even in our simple, "nearly determined" setting, finding an *optimal* task ordering is (1) computationally hard, as it reduces to the maximum-weight Hamiltonian path problem,[5] and (2) challenging to analyze and discuss. Nonetheless, we prove that the MD ordering yields a square-root approximation of the optimal distance at the end of learning.

Lemma 4.1 (Optimality guarantee when $r = d - 1$)**.** *Let $\mathbf{w}_T^{\tau_{\mathrm{MD}}}$ and $\mathbf{w}_T^{\tau_\star}$ be the iterates after learning T jointly realizable tasks of rank $d-1$ under the Maximum Distance ordering and a minimum-distance ordering (respectively). Then, their distances to the offline solution hold,*

$$0 \le D^2(\mathbf{w}_T^{\tau_\star}) \le D^2(\mathbf{w}_T^{\tau_{\mathrm{MD}}}) \triangleq \frac{\|\mathbf{w}_T^{\tau_{\mathrm{MD}}} - \mathbf{w}_\star\|^2}{\|\mathbf{w}_\star\|^2} \le \frac{\|\mathbf{w}_T^{\tau_\star} - \mathbf{w}_\star\|}{\|\mathbf{w}_\star\|} \triangleq D(\mathbf{w}_T^{\tau_\star}) \le 1\,.$$

The full proofs for this section are given in App. C.

What about the loss? The optimality of the distance does not imply optimality of the average loss, as exemplified in Figure 7 in the discussion. Instead, we now derive an upper bound for the loss.

Lemma 4.2 (Loss bound when $r = d - 1$)**.** *Under the Maximum Distance greedy ordering over T jointly-realizable tasks of rank $d-1$, the loss of Scheme 1 after T iterations is upper bounded as,*

$$\mathcal{L}(\mathbf{w}_T) = \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2} \cdot \frac{1}{T} \sum_{m=1}^{T} \|\mathbf{X}_m \mathbf{w}_T - \mathbf{y}_m\|^2 \le \frac{1}{eT}\,.$$

Question. *Do the favorable guarantees on distance and loss extend to tasks of general rank?*

5 Failure modes and surprises in greedy orderings

5.1 Greedy orderings can fail where random ones do not

Under random orderings, with or without replacement, Evron et al. [26] proved a universal, dimensionality-independent rate,

$$\mathbb{E}_{\tau_{\mathrm{Unif}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\mathrm{Unif}}})\right] \le \frac{14}{\sqrt[4]{k}}\,.$$

In contrast, we present an adversarial construction where the greedy ordering fails to learn on a task collection of T tasks in $d = T + 1$ dimensions, exploiting the dimensionality to undermine the greedy ordering. The full construction details and proof are provided in App. D.

Theorem 5.1 (Greedy lower bound)**.** *For any $d \ge 30$, there exists an adversarial task collection with $T = d - 1$ jointly-realizable tasks of different rank such that both greedy orderings (MD, MR) forget* catastrophically. *That is, the loss at the end of the sequence is, $\mathcal{L}(\mathbf{w}_T^{\tau_{\mathrm{MD}}}), \mathcal{L}(\mathbf{w}_T^{\tau_{\mathrm{MR}}}) \ge \frac{1}{8} - \frac{1}{4d}$.*

We demonstrate the behavior of an adversarial task collection using $T = 999$ tasks in $d = 1000$ dimensions. Our constructed collection "tricks" the greedy orderings: slowly increasing not only the loss on *all* tasks, but also the forgetting of *previous* tasks. The model is thus unable to accumulate knowledge, and practically forgets everything it learns.

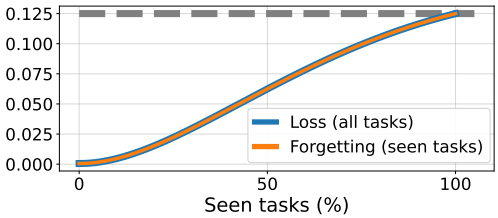


Figure 4: **Learning the adversarial construction.**

[5] Continual learning papers closely related to ours [10, 47] have also discussed Hamiltonian paths in the context of "optimal" task orderings. We show how greedy orderings approximate them.

5.2 Single-pass vs. repetition in greedy orderings

So far, we have focused on *single-pass* greedy orderings, in which each task is learned exactly once. These are conceptually related to without-replacement sampling and (re)shuffling techniques in SGD and the Kaczmarz method. In those areas, such repetition-free strategies often yield faster convergence than with-replacement sampling, both empirically [13, 58, 75] and in theory [11, 32, 33, 39, 53; but see 19, 64]. We ask: *Does the advantage of orderings without repetition extend to greedy orderings?*

Next, we derive a bound which—though possibly loose—already illustrates that repetition in greedy orderings avoids the failure mode seen in single-pass orderings (Theorem 5.1).

Proposition 5.2 (Dimensionality-independent bound for greedy orderings with repetition)**.** *For any task collection of T jointly realizable tasks, the loss under greedy maximum distance (MD) ordering* with repetition*, i.e., $\tau_{\text{MD-R}}$, after $k \geq 2$ iterations, is upper bounded as $\mathcal{L}(\mathbf{w}_k^{\tau_{\text{MD-R}}}) = \mathcal{O}(1/\log k)$.*

We evaluate the effect of repetition across orderings under random data. As in prior work, random sampling *without* replacement outperforms *with* replacement. In contrast, repetition benefits greedy orderings, likely due to *larger* updates and faster convergence to $\mathbf{w}_\star$. The slowdown in the single-pass case likely reflects the exhaustion of high dissimilarities.
Full details, experiments, and proof appear in App. E.

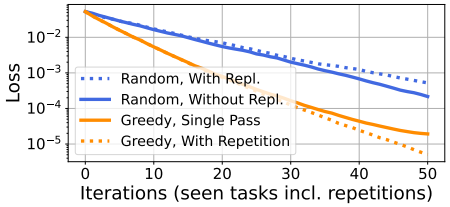


Figure 5: **The effect of repetitions.**

Intuitively, restricting replacement in *random* orderings exposes the learner to more data, while repetition in *greedy* selection allows the learner to consider all tasks at each step.

5.3 Hybrid task orderings: The best of both worlds

Motivated by the success of greedy Kaczmarz and importance sampling methods [3, 57], as well as convergence bounds for random orderings in continual learning [24, 26], we introduce a "hybrid" strategy. In this approach, tasks are selected greedily as long as the decrements $\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2$ (see Eq. (3)) remain above a threshold; afterward, selection switches to random sampling. Hybrid schemes have also been explored in Kaczmarz [18, 57], coordinate descent [27], and Schwarz [30] methods.

Scheme 2 Hybrid ordering (τ_{H})

Input: $\beta \in \left[0,\ \|\mathbf{w}_0 - \mathbf{w}_\star\|^2\right]$
For each iteration $t = 1, \ldots, k$: # Use greedy selection as long as the threshold is met
  $m' \leftarrow \operatorname{argmax}_{m \in [T] \setminus \tau_{\text{H}}(1:t-1)} \|(\mathbf{I} - \mathbf{P}_m)(\mathbf{w}_{t-1} - \mathbf{w}_\star)\|^2$ # Compute greedy selection
  If $\|(\mathbf{I} - \mathbf{P}_{m'})(\mathbf{w}_{t-1} - \mathbf{w}_\star)\|^2 \geq \beta$ **Then** $\tau_{\text{H}}(t) \leftarrow m'$ **Else Break**
$\tau_{\text{H}}(t:k) \sim \text{Unif}\left([T] \setminus \tau_{\text{H}}(1:t-1)\right)$ # Choose remaining tasks randomly without replacement

Empirically, the hybrid ordering performs better than random but worse than greedy. This matches our intuition from Eq. (3) and Figure 1a: greedy selection takes larger "steps" (or projections), particularly early on, when most tasks are still available. Once these projections diminish, we switch to the random ordering, which—unlike the greedy approach—cannot be adversarially "tricked" into failure (see Section 5.1).
Further details and experiments appear in App. F.1.

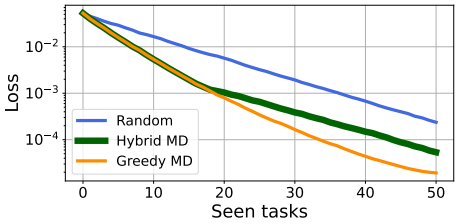


Figure 6: **Hybrid ordering experiment.**

Analytically, any bound for without-replacement random orderings, *e.g.,* an $\mathcal{O}\left(1/k^{1/4}\right)$ bound [26], can extend to our hybrid Scheme 2, showing again that it avoids the failure mode of Section 5.1.

Theorem 5.3 (informal)**.** *Assume any bound of the form $\mathbb{E}_{\tau_{\text{Unif}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\text{Unif}}})\right] = \mathcal{O}\left(1/k^\alpha\right)$, $\alpha \in (0, 1]$, established for the without-replacement τ_{Unif}. Then, setting a threshold of $\beta = \Omega\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T^{1-\alpha}}\right)$, guarantees a similar bound $\mathbb{E}_{\tau_{\text{H}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\text{H}}})\right] = \mathcal{O}\left(1/k^\alpha\right)$ for the hybrid ordering τ_{H}.*

The exact theorem and its proof are given in App. F.2. While our analysis sets β using $\|\mathbf{w}_0 - \mathbf{w}_\star\|$, the hybrid method remains useful, *e.g.,* with a heuristic β.

6 Discussion and related work

Throughout the paper, we have extensively discussed connections to other literatures, with a focus on continual learning and the Kaczmarz method. Below, we outline additional ideas and connections. Due to space constraints, further related work is deferred to App. A.

Task orderings in continual learning theory. Continual learning theory often treats task orderings as arbitrary. However, several analytical works [e.g., 15, 24, 25, 26, 40, 42] have shown that certain orderings—typically cyclic or random—can mitigate forgetting. Lin et al. [48] also explored the role of task similarity and reached conclusions similar to ours, though key differences remain: (1) their generalization analysis relied on restrictive assumptions requiring i.i.d. features across all tasks; (2) they assumed a separate *teacher* per task, unlike our setting; and (3) task orderings were not their primary focus.

Task typicality at the end of learning. Li and Hiratani [47] suggested that tasks should be arranged from least to most "typical". While we did not focus on this aspect of orderings, our geometric interpretation can illustrate it. Our motivation was to minimize the distance $\|\mathbf{w}_k - \mathbf{w}_\star\|^2$, which upper bounds the *loss* $\frac{1}{T}\sum_{m=1}^{T}\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2$. However, this bound can be loose, and minimizing the distance does not guarantee the lowest loss. For example, in the figure, although $\|\mathbf{w}_{\mathtt{A}} - \mathbf{w}_\star\|^2 = \|\mathbf{w}_{\mathtt{C}} - \mathbf{w}_\star\|^2$, the point $\mathbf{w}_{\mathtt{C}}$ is a better *ending point* than $\mathbf{w}_{\mathtt{A}}$, inducing a lower loss (the arrows represent the residuals). This happens because task C is more typical—*i.e.,* more similar to other tasks—than task A.

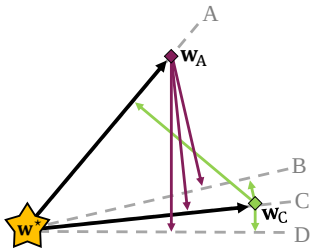


Figure 7: **Task typicality.**

Regret today or loss tomorrow? In Section 3, we motivated the use of greedy orderings to minimize the distance to the offline solution $\|\mathbf{w}_k - \mathbf{w}_\star\|^2$, which in turn upper bounds the average loss over *all* tasks: $\frac{1}{T}\sum_{m=1}^{T}\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2$. This objective is related, but not identical, to the notion of *regret* which quantifies the loss along the optimization *path* on *consecutive* tasks, *i.e.,* $\frac{1}{k}\sum_{t=1}^{k}\|\mathbf{X}_{\tau(t)}(\mathbf{w}_{t-1} - \mathbf{w}_\star)\|^2$. From this definition and Figure 1, we observe that regret—though also upper bounded by the distances $\|\mathbf{w}_{t-1} - \mathbf{w}_\star\|^2$—can often benefit from transitions between *similar* tasks rather than *dissimilar* ones. In other words, when the goal is to make accurate predictions *during* learning—*e.g.,* in decision-making—transitioning between *similar* tasks may be preferable. Conversely, when the objective is to minimize average loss over *all* tasks—*e.g.,* in curriculum or multitask learning—our findings suggest that transitioning between *dissimilar* tasks is preferable.

Other continual setups. The specific continual learning setup can dramatically influence the behavior of task orderings. Our work considers a "domain-incremental" setting, where the model learns the same problem across different domains—*i.e.,* $\mathcal{P}(X)$ changes while $\mathcal{P}(Y|X)$ is fixed [77].

Alternatively, one could consider a "task-incremental" setup, where distinct tasks—with possibly different $\mathcal{P}(Y|X)$—are learned, and the task identity is known at both train *and* test time. In this setting, [55, 61] trained a *separate* linear model per task and found that *similarity-maximizing* orderings prevailed, seemingly contradicting our findings (*e.g.,* in Figure 1). However, in such scenarios forgetting is *less* of a concern, and the focus shifts to inter-task *transfer*, which benefits from similar consecutive tasks (see also earlier discussion on regret). Hence, their results complement ours.

Others have studied "class-incremental" learning (CIL), where each task introduces new objects or classes, aiming for strong overall performance (*e.g.,* in popular split benchmarks [76]). However, comparing this setting to ours is challenging for two reasons: (1) softmax layers are hard to analyze in continual settings, with limited theoretical understanding to date; (2) in most CIL work, another major factor—beyond *inter-task similarity*—plays a central role, as discussed next.

The majority of studies on task ordering in continual learning support our conclusion that sequential task *dissimilarity* is beneficial [10, 24, 48, 50, 54, 63, 67, 70]. Some CIL papers suggest that adjacent task *similarity* is preferable [34, 51]. However, a closer look reveals that these studies modify the class composition *within* each task, assembling tasks with high *intra-task heterogeneity* [6, 34]. This likely

leads to wider minima and stronger "transferability" to other tasks, thus explaining their improved results. Such configurations resemble curriculum learning more than continual learning.[6] We found *one* CIL study contradicting our conclusions is [81], perhaps due to their empirical setup.[7] Finally, we remark that the questions and effects discussed here are related to the *interleaving effect* examined in educational psychology [59, 66].

Future work. One could extend our findings to other settings, such as class- and task-incremental, discussed earlier. Moreover, our realizability assumption could be relaxed (allowing label noise) or removed entirely (with nonlinear models), perhaps borrowing tools from Kaczmarz literature [9, 82].

References

[1] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. C. Fowlkes, S. Soatto, and P. Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019.

[2] S. Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6: 382–392, 1954.

[3] G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

[4] V. Alvarez, S. Mazuelas, and J. A. Lozano. Supervised learning with evolving tasks and performance guarantees. *Journal of Machine Learning Research*, 26(17):1–59, 2025.

[5] H. Asanuma, S. Takagi, Y. Nagano, Y. Yoshida, Y. Igarashi, and M. 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.

[6] N. Ashtekar, J. Zhu, and V. G. Honavar. Class incremental learning from first principles: A review. *Transactions on Machine Learning Research*, 2025.

[7] Z.-Z. Bai and W.-T. Wu. On greedy randomized kaczmarz method for solving large sparse linear systems. *SIAM Journal on Scientific Computing*, 40(1):A592–A606, 2018.

[8] Z.-Z. Bai and W.-T. Wu. On relaxed greedy randomized kaczmarz methods for solving large sparse linear systems. *Applied Mathematics Letters*, 83:21–26, 2018.

[9] Z.-Z. Bai and W.-T. Wu. On greedy randomized augmented kaczmarz method for solving large sparse inconsistent linear systems. *SIAM Journal on Scientific Computing*, 43(6):A3892–A3911, 2021.

[10] S. J. Bell and N. D. Lawrence. The effect of task ordering in continual learning. *arXiv preprint arXiv:2205.13323*, 2022.

[11] P. Beneventano. On the trajectories of sgd without replacement. *arXiv preprint arXiv:2312.16143*, 2023.

[12] P. A. Borodin and E. Kopecká. Alternating projections, remotest projections, and greedy approximation. *Journal of Approximation Theory*, 260:105486, 2020. ISSN 0021-9045.

[13] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633. Citeseer, 2009.

[14] S. Braun, D. Neil, and S.-C. Liu. A curriculum learning method for improved noise robustness in automatic speech recognition. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 548–552, 2017. doi: 10.23919/EUSIPCO.2017.8081267.

[6] The learner controls the internal composition of tasks to create "easier" tasks, as in curriculum learning [78].

[7] They construct the first task using a random half of the classes. This strong "pretraining" leads to low initial loss, as the model already learns *half* the classes. This resembles the failure mode discussed in Section 5.1.

[15] X. Cai and J. Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. 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. Curran Associates, Inc., 2019.

[17] R. Das, X. Chen, B. Ieong, P. Bansal, and sujay sanghavi. Understanding the training speedup from sampling with approximate losses. In *Forty-first International Conference on Machine Learning*, 2024.

[18] J. A. De Loera, J. Haddock, and D. Needell. A sampling kaczmarz–motzkin algorithm for linear feasibility. *SIAM Journal on Scientific Computing*, 39(5):S66–S87, 2017.

[19] C. M. De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] M. Ding, K. Ji, D. Wang, and J. Xu. Understanding forgetting in continual learning with linear regression. In *Forty-first International Conference on Machine Learning*, 2024.

[21] T. Doan, M. Abbana Bennani, B. Mazoure, G. Rabusseau, and P. 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.

[22] Y. C. Eldar and D. Needell. Acceleration of randomized kaczmarz method via the johnson–lindenstrauss lemma. *Numerical Algorithms*, 58:163–177, 2011.

[23] T. Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35(1):1–12, 1980.

[24] I. Evron, E. Moroshko, R. Ward, N. Srebro, and D. Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, pages 4028–4079. PMLR, 2022.

[25] I. Evron, E. Moroshko, G. Buzaglo, M. Khriesh, B. Marjieh, N. Srebro, and D. 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.

[26] I. Evron, R. Levinstein, M. Schliserman, U. Sherman, T. Koren, D. Soudry, and N. Srebro. Better rates for random task orderings in continual linear models. *arXiv preprint arXiv:2504.04579*, 2025.

[27] H. Fang, G. Fang, T. Yu, and P. Li. Efficient greedy coordinate descent via variable partitioning. In C. de Campos and M. H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 547–557. PMLR, 27–30 Jul 2021.

[28] D. Goldfarb and P. 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.

[29] D. Goldfarb, I. Evron, N. Weinberger, D. Soudry, and P. Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting - an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] M. Griebel and P. Oswald. Greedy and randomized versions of the multiplicative schwarz method. *Linear Algebra and its Applications*, 437(7):1596–1610, 2012.

[31] L. Gubin, B. T. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[32] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, Mar 2021. ISSN 1436-4646. doi: 10.1007/s10107-019-01440-w. URL https://doi.org/10.1007/s10107-019-01440-w.

[33] D. Han and J. Xie. A simple linear convergence analysis of the reshuffling kaczmarz method. *arXiv preprint arXiv:2410.01140*, 2024.

[34] C. He, R. Wang, and X. Chen. Rethinking class orders and transferability in class incremental learning. *Pattern Recognition Letters*, 161:67–73, 2022. ISSN 0167-8655.

[35] H. Hemati, L. Pellegrini, X. Duan, Z. Zhao, F. Xia, M. Masana, B. Tscheschner, E. Veas, Y. Zheng, S. Zhao, et al. Continual learning in the presence of repetition. In *CVPR Workshop on Continual Learning in Computer Vision*, 2024.

[36] N. Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[37] E. Hoffer, B. Weinstein, I. Hubara, S. Gofman, and D. Soudry. Infer2train: leveraging inference for better training of deep networks. In *NeurIPS 2018 Workshop on Systems for ML*, page 40, 2018.

[38] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[39] H. Jeong and D. Needell. Linear convergence of reshuffling kaczmarz methods with sparse constraints. *SIAM Journal on Scientific Computing*, 2025. to appear.

[40] H. Jung, H. Cho, and C. Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] S. Kaczmarz. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937.

[42] M. Kong, W. Swartworth, H. Jeong, D. Needell, and R. Ward. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[43] A. Lad, R. Ghani, Y. Yang, and B. Kisiel. Toward optimal ordering of prediction tasks. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 884–893. SIAM, 2009.

[44] T. Lesort, O. Ostapenko, P. Rodríguez, D. Misra, M. R. Arefin, L. Charlin, and I. Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pages 43–65. PMLR, 2023.

[45] H. Li, J. Wu, and V. Braverman. Fixed design analysis of regularization-based continual learning. In S. Chandar, R. Pascanu, H. Sedghi, and D. 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.

[46] H. Li, J. Wu, and V. Braverman. Memory-statistics tradeoff in continual learning with structural regularization. *arXiv preprint arXiv:2504.04039*, 2025.

[47] Z. Li and N. Hiratani. Optimal task order for continual learning of multiple tasks. *arXiv preprint arXiv:2502.03350*, 2025.

[48] S. Lin, P. Ju, Y. Liang, and N. 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.

[49] Y. Lu, S. Y. Meng, and C. De Sa. A general analysis of example-selection for stochastic gradient descent. In *International Conference on Learning Representations (ICLR)*, volume 10, 2022.

[50] G. Mantione-Holmes, J. Leo, and J. Kalita. Utilizing priming to identify optimal class ordering to alleviate catastrophic forgetting. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 57–64. IEEE, 2023.

[51] M. Masana, B. Twardowski, and J. Van de Weijer. On class orderings for incremental learning. *arXiv preprint arXiv:2007.02145*, 2020.

[52] C.-Q. Miao and W.-T. Wu. On greedy randomized average block kaczmarz method for solving large linear systems. *Journal of Computational and Applied Mathematics*, 413:114372, 2022.

[53] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[54] C. V. Nguyen, A. Achille, M. Lam, T. Hassner, V. Mahadevan, and S. Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv (verify) preprint arXiv:1908.01091*, 2019.

[55] T. Nguyen, C. N. Nguyen, Q. Pham, B. T. Nguyen, S. Ramasamy, X. Li, and C. V. Nguyen. Sequence transferability and task order selection in continual learning. *arXiv preprint arXiv:2502.06544*, 2025.

[56] Y.-Q. Niu and B. Zheng. A greedy block kaczmarz algorithm for solving large-scale linear systems. *Applied Mathematics Letters*, 104:106294, 2020.

[57] J. Nutini, B. Sepehry, I. Laradji, M. Schmidt, H. Koepke, and A. Virani. Convergence rates for greedy kaczmarz algorithms, and faster randomized kaczmarz rules using the orthogonality graph. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 547–556, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

[58] P. Oswald and W. Zhou. Convergence analysis for kaczmarz-type methods in a hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015.

[59] S. C. Pan. The interleaving effect: mixing it up boosts learning. *Scientific American*, 313(2), 2015.

[60] L. Peng, P. Giampouras, and R. Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023.

[61] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, 2015.

[62] S. Rajput, K. Lee, and D. Papailiopoulos. Permutation-based SGD: Is random optimal? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=YiBa9HKTyXE`.

[63] V. V. Ramasesh, E. Dyer, and M. Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2020.

[64] B. Recht and C. Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory (COLT)*, 2012.

[65] S. Reich and R. Zalas. Polynomial estimates for the method of cyclic projections in hilbert spaces. *Numerical Algorithms*, pages 1–26, 2023.

[66] D. Rohrer, R. F. Dedrick, and S. Stershic. Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, 107(3):900, 2015.

[67] P. Ruvolo and E. Eaton. Active task selection for lifelong machine learning. In *Twenty-seventh AAAI conference on artificial intelligence*, 2013.

[68] H. Sajjad, N. Durrani, F. Dalvi, Y. Belinkov, and S. Vogel. Neural machine translation training in a multi-domain scenario. *arXiv preprint arXiv:1708.08712*, 2017.

[69] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2608–2615, 2017.

[70] C. Schouten. *Investigating Task Order in Online Class-Incremental Learning*. PhD thesis, Master's thesis, Department of Mathematics and Computer Science, AutoML . . . , 2024.

[71] H. Shan, Q. Li, and H. Sompolinsky. Order parameters and phase transitions of continual learning in deep neural networks. *arXiv preprint arXiv:2407.10315*, 2024.

[72] A. Shrivastava, A. K. Gupta, and R. B. Girshick. Training region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016. URL `https://api.semanticscholar.org/CorpusID:2843566`.

[73] S. Stojanov, S. Mishra, N. A. Thai, N. Dhanda, A. Humayun, C. Yu, L. B. Smith, and J. M. Rehg. Incremental object learning from contiguous views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8777–8786, 2019.

[74] Y. Su, D. Han, Y. Zeng, and J. Xie. On the convergence analysis of the greedy randomized kaczmarz method. *arXiv preprint arXiv:2307.01988*, 2023.

[75] R.-Y. Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

[76] S. Swaroop, C. V. Nguyen, T. D. Bui, and R. E. Turner. Improving and understanding variational continual learning. *arXiv preprint arXiv:1905.02099*, 2019.

[77] G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[78] X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

[79] A.-Q. Xiao, J.-F. Yin, and N. Zheng. On fast greedy block kaczmarz methods for solving large consistent linear systems. *Computational and Applied Mathematics*, 42(3):119, 2023.

[80] Y. Xu and B. Mirzasoleiman. Ordering for non-replacement sgd. *arXiv preprint arXiv:2306.15848*, 2023.

[81] Z. Yang and H. Li. Task ordering matters for incremental learning. In *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6, 2021.

[82] J. Zhang, Y. Wang, and J. Zhao. On maximum residual nonlinear kaczmarz-type algorithms for large nonlinear systems of equations. *Journal of Computational and Applied Mathematics*, 425: 115065, 2023.

[83] J.-J. Zhang. A new greedy kaczmarz algorithm for the solution of very large linear systems. *Applied Mathematics Letters*, 91:207–212, 2019.

[84] Y. Zhang and H. Li. Greedy motzkin–kaczmarz methods for solving linear systems. *Numerical Linear Algebra with Applications*, 29(4):e2429, 2022.

[85] X. Zhao, H. Wang, W. Huang, and W. Lin. A statistical theory of regularization-based continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

A Further related work

Alternative viewpoint: The Kaczmarz method. The continual linear regression scheme described in this work maps directly to the Kaczmarz method [23, 41], a classical iterative algorithm for solving linear systems of equations. In our context, the solved system is, $\mathbf{X}\mathbf{w} = \mathbf{y}$, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \in \mathbb{R}^{N \times d}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} \in \mathbb{R}^{N}, \quad \text{where } N = \sum_{m=1}^{T} n_m.$$

Evron et al. [24] pointed out that in each iteration, the Kaczmarz method solves the current "block" system $\mathbf{X}_{\tau(t)}\mathbf{w} = \mathbf{y}_{\tau(t)}$ using an update rule equivalent to our continual update in Eq. (1). Owing to this equivalence, all the observations and results in our paper extend naturally to the greedy Kaczmarz method. However, whereas the Kaczmarz literature typically analyzes convergence in terms of the distance to the intersection $\mathbf{w}_\star$ (see Definition 2.5), our focus is on the *loss*—that is, the residuals (see Definition 2.3).

Task orderings in continual learning theory. Continual learning theory often treats task orderings as arbitrary. However, several analytical works [e.g., 15, 24, 25, 26, 40, 42] have shown that certain orderings—typically cyclic or random—can mitigate forgetting. While some works downplayed ordering effects—arguing they are often minor—and deferred their study to future work [71], others designed continual learning algorithms specifically for *evolving sequences*, where adjacent tasks are highly *similar* [4]. We follow a different line of work, cited throughout this paper and expanded upon here, that investigates how pairwise task similarities or *dissimilarities* influence common continual learning algorithms.

A particularly relevant work discussed throughout our paper is Bell and Lawrence [10], which advocated for pairwise task dissimilarities as a guiding principle for task ordering. Their study was among the first to empirically investigate orderings that transition between similar or dissimilar tasks. Tasks were represented as vertices in a complete graph, with edge weights corresponding to a predefined distance between tasks; in this framework, each Hamiltonian path defines a possible task sequence. While they hypothesized that a minimum-weight path (favoring similar tasks in succession) would yield the best continual performance, their empirical findings on simple neural networks indicated the opposite: maximum-weight paths, which place *dissimilar* tasks adjacently, often led to improved performance. However, these results were not always statistically significant (see the error bars in their Figure 5). Their findings motivated our work to revisit the question of task ordering from a more analytically grounded perspective, using formal definitions and theoretical tools to better understand and justify similarity-guided orderings.

Li and Hiratani [47] conducted a deeper investigation into similarity-guided task orderings. They also found that adjacent tasks should be *dissimilar*, and further explored the notion of task "typicality" (discussed in Section 6). Their empirical results—also obtained using neural networks—are more statistically robust than those of Bell and Lawrence [10]. In addition, they derived analytical results for a linear regression model that support their empirical observations. However, their theoretical analysis relies on a restrictive random data model in which all task features are drawn from a simplified distribution. In contrast, our analysis accommodates *arbitrary* feature matrices, allowing for richer and more realistic forms of task similarity. Like Bell and Lawrence [10], their work focuses primarily on the role of pairwise task similarities in continual learning. By contrast, we draw on tools from the optimization literature on the Kaczmarz and projection methods, to formalize and study *greedy* task orderings specifically—both as a practical approach and as a proxy for optimal orderings.

Ruvolo and Eaton [67] proposed an "information maximization" approach to task ordering, using a diversity-based heuristic closely related to our greedy maximum residual (MR) strategy (Definition 3.2). While they demonstrated improved performance over random orderings, their model choice likely limited the potential for rigorous theoretical analysis, which we provide in this work.

Lin et al. [48] also examined the role of task similarity and arrived at conclusions broadly aligned with ours. While their work is influential, several key differences set it apart from our approach. First, their generalization analysis relies on restrictive assumptions, such as i.i.d. features across all tasks. They also assume a distinct *teacher* model per task, in contrast to our setting, where all tasks are explained by a single overparameterized model—an assumption more reflective of modern deep learning and common in domain-incremental learning. As a result, their notion of

task similarity is based on the similarity between teachers, rather than more practical measures such as similarities between feature matrices or residuals (see Definitions 3.1 and 3.2). Moreover, their analysis becomes vacuous in highly overparameterized regimes (see their Figure 1(c)), whereas ours remains informative. Crucially, task ordering was not the primary focus of their study. Although they observed ordering effects in their generalization bounds for regression and supported this with brief experiments on classification, our work offers a substantially more comprehensive treatment of similarity-guided task orderings. We provide formal definitions, geometric and algebraic intuitions, greedy strategies, optimality results, computational considerations, empirical validation, and an analysis of failure modes and repetition.

SGD and example selection. Evron et al. [26] showed that continual linear regression trained to convergence (our Scheme 1) reduces to Incremental Gradient Descent (IGD). Specifically, learning an *entire* task is equivalent to taking a *single* large gradient step with respect to a modified objective. While they used this reduction to analyze *random* task orderings via last-iterate SGD analysis, we leverage it here to draw connections between *greedy* task orderings and greedy example selection strategies in SGD.

Most of the literature on example selection in SGD assumes multi-epoch settings, where each sample is seen multiple times. In such regimes, it is common to randomly shuffle the dataset once, or reshuffle it at the start of each epoch [e.g., 32, 53]. Although widely used, random permutations are not necessarily *optimal* [62]. For instance, Lu et al. [49] showed that greedy permutations—computed at the beginning of each epoch—can yield faster convergence than random ones. However, their analysis relies on (1) multiple epochs and (2) very small step sizes, making it inapplicable to single-pass settings like ours.

Das et al. [17] demonstrated that a selection rule akin to our maximum residual strategy (Definition 3.2) accelerates early-stage convergence of the average-iterate loss, but may underperform random orderings asymptotically. This finding further motivates our hybrid approach (Scheme 2) and aligns with our experimental results in Figure 6. They also analyzed an approximate selection rule, supporting our observations on computational feasibility in Section 3. Finally, it is also possible to select greedily by gradient magnitude instead of loss minimization [80], or to "mine" hard examples, *i.e.,* those with high loss, at the mini-batch level [72].

B Appendix to Section 4.1: Experiments comparing ordering methods

All figures report averages over 10 repeated experiments, where the same task collections are used for the different ordering strategies. Shaded regions (see App. E.1 and F.1) indicate ± 1 standard error intervals, even when not visually discernible. In App. B.3 we further discuss the statistical significance of our experiments.

Compute resources. All experiments—including those not shown—were completed within 4 hours on a home PC equipped with an Intel i5-9400F CPU and 16GB of RAM.

B.1 Isotropic data

Figures 8 and 9 extend the previous experiment on isotropic data (Figure 3a) to varying dimensions d, ranks r and task counts T. Results confirm consistent patterns: greedy (dissimilarity maximizing) methods outperform random, and MD is better than MR across all settings (sometimes only slightly).

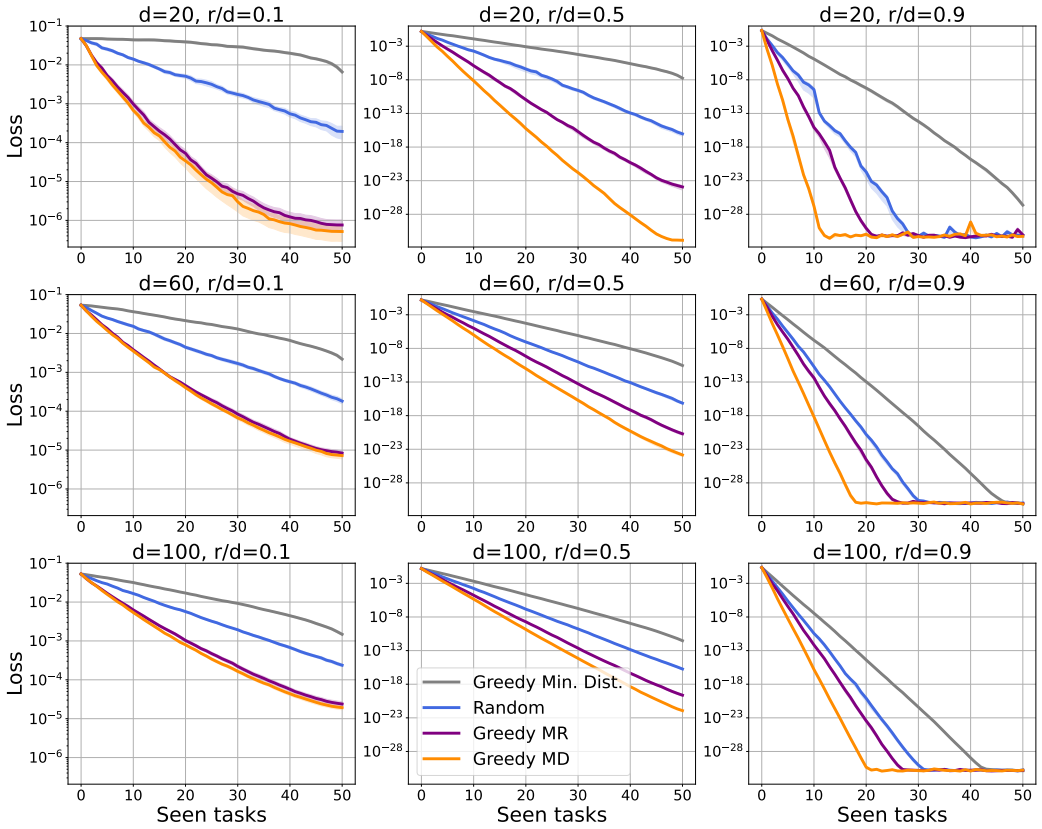


Figure 8: **Comparing orderings for varying dimensions d and ranks r of the data matrices, for isotropic data.** $T = 50$. We observe that, for such isotropic data, the random ordering performance is determined solely by the ratio r/d. In contrast, greedy orderings that prioritize *dissimilarity* benefit from a lower dimension when r/d is fixed (to see that, focus on single columns in the grid). We hypothesize that this is because an increased *task "density"* in lower dimensions: when r/d is fixed, increasing d increases $d - r$, expanding the set of possible task projections (see Eq. (2)). As a result, a fixed number of tasks T covers this space more sparsely in higher dimensions. In lower dimensions, the same T tasks yield denser coverage, increasing the likelihood that greedy dissimilarity-based selection identifies tasks with large projections.
In all strategies, higher task rank consistently yields improved performance (focus on single rows). This is because the solution subspaces are of rank $d - r$, so increasing r (with fixed d) lowers the subspace rank, increasing the distances between them and resulting in larger projections.

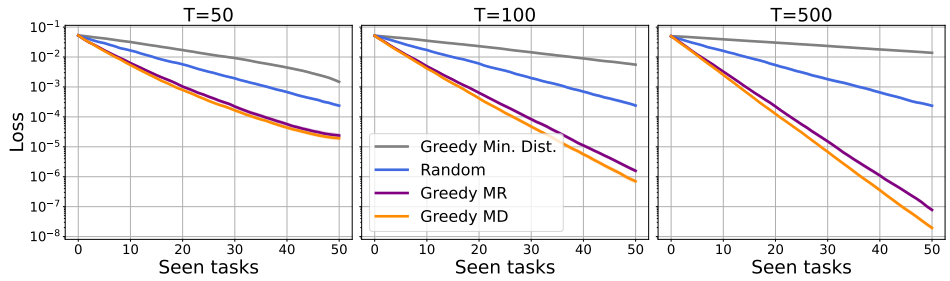


Figure 9: **Comparing orderings for varying task count T, for isotropic data.** $d = 100,\ r = 10$. Dissimilarity-based greedy strategies become more effective as the number of tasks increases. This is since in an isotropic setting, where task directions are sampled uniformly, increasing the number of tasks increases the *coverage* of the unit sphere. This results in a higher probability of encountering task pairs with large angular separation between their solution subspaces, which greedy ordering utilizes.

B.2 Anisotropic data

The anisotropic data in Figure 3b was sampled from a Gaussian distribution with exponentially decaying eigenvalues, as detailed in Scheme 3, resulting in high task correlation. This arises because tasks tend to align with the dominant eigen-directions, leading to strong pairwise similarity.

Scheme 3 Generating tasks with high correlation

Require: Input dimension d, task rank r, number of tasks T, edge eigenvalues $\lambda_1 = 10^{-3}$, $\lambda_d = 10^3$

1: Sample $\mathbf{A} \sim \mathcal{N}(0,1)^{d\times d}$ and symmetrize: $\mathbf{A}_{\text{sym}} \leftarrow \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$
2: Compute SVD: $\mathbf{A}_{\text{sym}} = \mathbf{USU}^\top$
3: Define diagonal spectrum: $\mathbf{\Lambda} \leftarrow \operatorname{diag}\left(\lambda_1 \exp\left(\ln\left(\lambda_d/\lambda_1\right)\frac{i}{d-1}\right)\right)_{i=0}^{d-1}$
4: Construct covariance: $\mathbf{\Sigma} \leftarrow \mathbf{U\Lambda U}^\top$
5: **for** $t = 1$ to T **do**
6: Sample $\mathbf{Z}_t \sim \mathcal{N}\left(0,1\right)^{r\times d}$
7: Set $\mathbf{X}_t \leftarrow \mathbf{Z}_t\mathbf{\Sigma}^{1/2}$
8: **end for**
9: **Output:** $\{\mathbf{X}_t\}_{t=1}^T$

Figures 10 and 11 below extend the experiment in Figure 3b, revealing some interesting trends compared to the isotropic case.

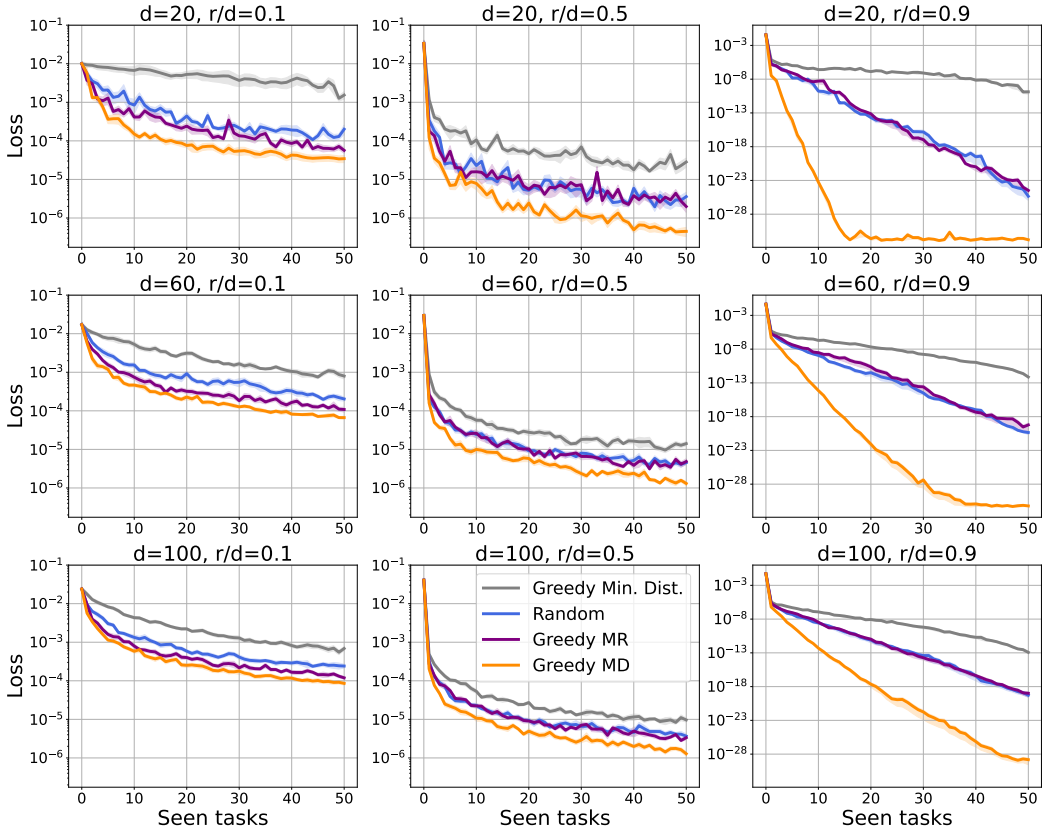


Figure 10: **Comparing orderings for varying dimensions d and ranks r of the data matrices, for anisotropic data.** $T = 50$. Compared to the isotropic case (Figure 8), we observe slower rates for all strategies. This is easily explained by all pairwise distances between task solution subspaces becoming smaller, due to the higher correlation in the anisotropic case.
Interestingly, as rank increases (focusing on a single row in the grid), the Maximum Residual (MR; Definition 3.2) ordering underperforms and seemingly aligns with the random one. This may stem from the combination of high rank and strong *intra*-task correlation, which leads to *ill-conditioned* data matrices (for each task). In such a case, small perturbations, or steps, in the solution space may cause disproportionately large changes in residuals. As a result, MR is misled into selecting tasks with large residuals that advance the iterate only marginally toward the intersection ($\mathbf{w}_\star$).

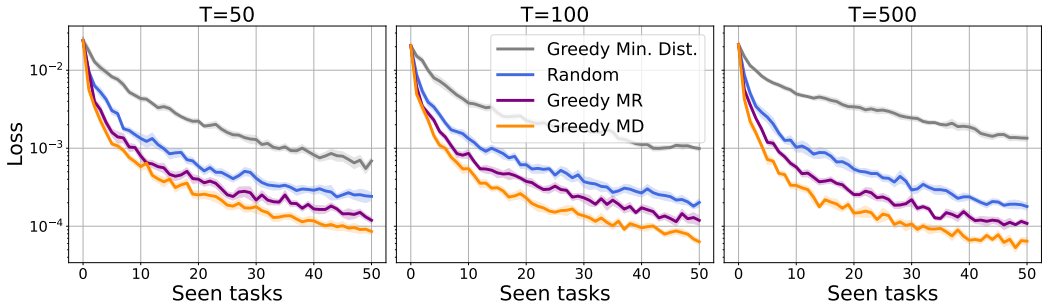


Figure 11: **Comparing orderings for varying task count T, for anisotropic data.** $d = 100,\ r = 10$. Unlike in the isotropic case (Figure 9), greedy orderings do not significantly benefit from increasing the number of tasks T. This is likely since, in the anisotropic case, a large number of tasks must be added to induce the substantial "angles" that greedy orderings can exploit. Put differently, under our anisotropic distribution, the probability that any set of 50 tasks are mutually orthogonal—and thus beneficial to greedy orderings—is extremely small for any reasonable number of tasks T.

B.3 A note on statistical significance

All appendix figures include confidence intervals of ± 1 standard error, although these are often too narrow to be visible. While different task collections introduce slight variations in outcomes, the overall trends are highly consistent. This is illustrated in the following figure, where we replicate the plot from Figure 3a, overlaying individual runs from all 10 repeated experiments. Despite some run-to-run variability, the standard error remains small, reinforcing the robustness of our qualitative conclusions.

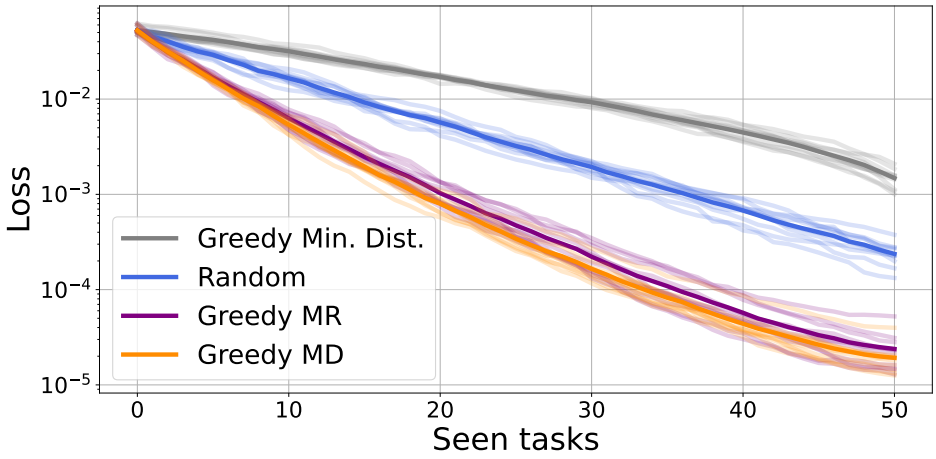


Figure 12: **Same as Figure 3a, with shaded plots for each individual experiment.** While minor variations exist across experiments, the low standard error confirms the consistency of the results.

C Proofs for Section 4.2: Greedy orderings of "nearly determined" tasks

Recall Lemma 4.1. Let $\mathbf{w}_T^{\tau_{\text{MD}}}$ and $\mathbf{w}_T^{\tau_\star}$ be the iterates after learning T jointly realizable tasks of rank $d-1$ under the Maximum Distance ordering and a minimum-distance ordering (respectively). Then, their distances to the offline solution hold,

$$0 \le D^2(\mathbf{w}_T^{\tau_\star}) \le D^2(\mathbf{w}_T^{\tau_{\text{MD}}}) \triangleq \frac{\left\|\mathbf{w}_T^{\tau_{\text{MD}}} - \mathbf{w}_\star\right\|^2}{\|\mathbf{w}_\star\|^2} \le \frac{\left\|\mathbf{w}_T^{\tau_\star} - \mathbf{w}_\star\right\|}{\|\mathbf{w}_\star\|} \triangleq D(\mathbf{w}_T^{\tau_\star}) \le 1\,.$$

Proof. The distance at the end of an ordering τ is

$$\begin{aligned} D^2\left(\mathbf{w}_T^{\tau}\right) &\triangleq \frac{\left\|\mathbf{w}_T^{\tau} - \mathbf{w}_\star\right\|^2}{\|\mathbf{w}_\star\|^2} = \frac{1}{\|\mathbf{w}_\star\|^2}\left\|\mathbf{v}_{\tau(T)}\mathbf{v}_{\tau(T)}^\top \cdots \mathbf{v}_{\tau(1)}\mathbf{v}_{\tau(1)}^\top\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\right\|^2 \\ &= \frac{1}{\|\mathbf{w}_\star\|^2}\left(\mathbf{v}_{\tau(1)}^\top\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\right)^2 \cdot \prod_{t=1}^{T-1}\left(\mathbf{v}_{\tau(t)}^\top \mathbf{v}_{\tau(t+1)}\right)^2 . \end{aligned}$$

Let $\tau = \tau_{\text{MD}}, \tau_\star$ be the greedy MD ordering and an optimal ordering leading to the minimal distance (respectively). Denote for simplicity $c\left(i,j\right) = \begin{cases} \frac{1}{\|\mathbf{w}_\star\|^2}\left(\mathbf{v}_j^\top\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right)^2 & i = 0, j \in [T] \\ \left(\mathbf{v}_i^\top \mathbf{v}_j\right)^2 & i,j \in [T] \end{cases}$.

Then, we have,

$$\begin{aligned} D^2\left(\mathbf{w}_T^{\tau_\star}\right) &= \frac{1}{\|\mathbf{w}_\star\|^2}\left(\mathbf{v}_{\tau_\star(1)}^\top\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right)^2 \cdot \prod_{t=1}^{T-1}\left(\mathbf{v}_{\tau_\star(t)}^\top \mathbf{v}_{\tau_\star(t+1)}\right)^2 \\ &= c\left(0, \tau_\star\left(1\right)\right)\prod_{t=1}^{T-1} c\left(\tau_\star\left(t\right), \tau_\star\left(t+1\right)\right) \\ &= c\left(0, \tau_\star\left(1\right)\right)\prod_{t\in\mathcal{C}} c\Big(\tau\left(\tau^{-1}\left(\tau_\star\left(t\right)\right)\right), \tau_\star\left(t+1\right)\Big)\cdot\prod_{t\notin\mathcal{C}} c\Big(\tau_\star\left(t\right), \tau\left(\tau^{-1}\left(\tau_\star\left(t+1\right)\right)\right)\Big), \end{aligned}$$

where we define the index set $\mathcal{C} = \left\{t \mid 1 \le t \le T-1,\ \ \tau^{-1}\left(\tau_\star\left(t\right)\right) < \tau^{-1}\left(\tau_\star\left(t+1\right)\right)\right\}$.

Employing greediness, we get

$$\begin{aligned} D^2\left(\mathbf{w}_T^{\tau_\star}\right) \ge c\left(0,\tau\left(1\right)\right) &\underbrace{\prod_{t\in\mathcal{C}} c\Big(\tau\left(\tau^{-1}\left(\tau_\star\left(t\right)\right)\right), \tau\left(1+\tau^{-1}\left(\tau_\star\left(t\right)\right)\right)\Big)}_{\text{here, } \tau^{-1}(\tau_\star(t))<T}\cdot \\ &\cdot\underbrace{\prod_{t\notin\mathcal{C}} c\Big(\tau\left(1+\tau^{-1}\left(\tau_\star\left(t+1\right)\right)\right), \tau\left(\tau^{-1}\left(\tau_\star\left(t+1\right)\right)\right)\Big)}_{\text{here, } \tau^{-1}(\tau_\star(t+1))<T}. \end{aligned}$$

Then, since $\tau^{-1}\left(\tau_\star\left(\cdot\right)\right)$ "covers" $[T]$ and $c\left(i,j\right) \le 1$, iterating over the entire $1,\ldots,T-1$ will simply add elements to the product and make it smaller. That is,

$$\begin{aligned} D^2\left(\mathbf{w}_T^{\tau_\star}\right) &\ge c\left(0,\tau\left(1\right)\right)\cdot\prod_{\ell=1}^{T-1} c\left(\tau\left(\ell\right),\tau\left(1+\ell\right)\right)\cdot\prod_{\ell=1}^{T-1} c\left(\tau\left(1+\ell\right),\tau\left(\ell\right)\right) \\ &\ge \left(c\left(0,\tau\left(1\right)\right)\prod_{\ell=1}^{T-1} c\left(\tau\left(\ell\right),\tau\left(1+\ell\right)\right)\right)^2 \\ &= \left(\frac{1}{\|\mathbf{w}_\star\|^2}\left(\mathbf{v}_{\tau(1)}^\top\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right)^2\prod_{t=1}^{T-1}\left(\mathbf{v}_{\tau(t)}^\top \mathbf{v}_{\tau(t+1)}\right)^2\right)^2 = \left(D^2\left(\mathbf{w}_T^{\tau}\right)\right)^2 \end{aligned}$$

$$\Rightarrow 1 \ge D\left(\mathbf{w}_T^{\tau_\star}\right) \ge D^2\left(\mathbf{w}_T^{\tau_{\text{MD}}}\right) \ge D^2\left(\mathbf{w}_T^{\tau_\star}\right) \ge 0\,.$$

□

Recall Lemma 4.2. Under the Maximum Distance greedy ordering over T jointly-realizable tasks of rank $d-1$, the loss of Scheme 1 after T iterations is upper bounded as,

$$\mathcal{L}(\mathbf{w}_T) = \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2} \cdot \frac{1}{T} \sum_{m=1}^{T} \|\mathbf{X}_m \mathbf{w}_T - \mathbf{y}_m\|^2 \leq \frac{1}{eT}\,.$$

Proof. We aim to bound the average loss using projection matrices,

$$\begin{aligned}
\mathcal{L}_{\tau_{\text{MD}}}(\mathbf{w}_T) &= \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2 T} \sum_{m=1}^{T} \|\mathbf{X}_m \mathbf{w}_T - \mathbf{y}_m\|^2 = \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2 T} \sum_{m=1}^{T} \|\mathbf{X}_m \left(\mathbf{w}_T - \mathbf{w}_\star\right)\|^2 \\
&= \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2 T} \sum_{m=1}^{T} \|\mathbf{X}_m \mathbf{X}_m^{+} \mathbf{X}_m \left(\mathbf{w}_T - \mathbf{w}_\star\right)\|^2 \\
&\leq \tfrac{1}{\|\mathbf{w}_\star\|^2 R^2 T} \sum_{m=1}^{T} \|\mathbf{X}_m\|^2 \|\left(\mathbf{I} - \mathbf{P}_m\right)\left(\mathbf{w}_T - \mathbf{w}_\star\right)\|^2 \\
\left[\text{Eq. (2)}\right] &\leq \tfrac{1}{\|\mathbf{w}_\star\|^2 T} \sum_{t=1}^{T} \left\| \left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right) \prod_{s=1}^{T} \mathbf{P}_{\tau(s)} (\mathbf{w}_0 - \mathbf{w}_\star) \right\|^2 .
\end{aligned}$$

Since each task matrix $\mathbf{X}_i$ has rank $d-1$, each projection $\mathbf{P}_i$ is rank 1 and can be written as $\mathbf{P}_i = \mathbf{v}_i \mathbf{v}_i^\top$ for a unit vector $\mathbf{v}_i$. Substituting this and $\mathbf{v}_{\tau(0)} = \frac{1}{\|\mathbf{w}_\star\|}\left(\mathbf{w}_0 - \mathbf{w}_\star\right)$, the bound becomes:

$$\begin{aligned}
\mathcal{L}_{\tau_{\text{MD}}}(\mathbf{w}_T) &\leq \frac{1}{T} \sum_{t=1}^{T} \left\| \left(\mathbf{I} - \mathbf{v}_{\tau(t)} \mathbf{v}_{\tau(t)}^\top\right) \mathbf{v}_{\tau(T)} \mathbf{v}_{\tau(T)}^\top \cdots \mathbf{v}_{\tau(1)} \mathbf{v}_{\tau(1)}^\top \mathbf{v}_{\tau(0)} \right\|^2 \\
&\leq \underbrace{\left(\mathbf{v}_{\tau(1)}^\top \mathbf{v}_{\tau(0)}\right)^2}_{\leq 1} \frac{1}{T} \sum_{s=1}^{T} \left\| \left(\mathbf{I} - \mathbf{v}_{\tau(t)} \mathbf{v}_{\tau(t)}^\top\right) \mathbf{v}_{\tau(T)} \right\|^2 \prod_{s=1}^{T-1} \left(\mathbf{v}_{\tau(s+1)}^\top \mathbf{v}_{\tau(s)}\right)^2 \\
\left[\text{projection properties}\right] &\leq \left(1 - \frac{1}{T} \sum_{s=1}^{T} \left(\mathbf{v}_{\tau(T)}^\top \mathbf{v}_{\tau(s)}\right)^2\right) \prod_{s=1}^{T-1} \left(\mathbf{v}_{\tau(s+1)}^\top \mathbf{v}_{\tau(s)}\right)^2 .
\end{aligned}$$

Then, we use algebraic and projection properties to rewrite the greedy ordering as:

$$\begin{aligned}
\tau_{\text{MD}}(t) &= \operatorname*{argmax}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \|\left(\mathbf{I} - \mathbf{P}_m\right)\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\|^2 \\
&= \operatorname*{argmax}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \left(\|\mathbf{w}_{t-1} - \mathbf{w}_\star\|^2 - \|\mathbf{P}_m \left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\|^2\right) \\
&= \operatorname*{argmin}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \|\mathbf{P}_m \left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\|^2 \\
&= \operatorname*{argmin}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \|\mathbf{v}_m \mathbf{v}_m^\top \mathbf{v}_{\tau(t-1)} \mathbf{v}_{\tau(t-1)}^\top \left(\mathbf{w}_{t-2} - \mathbf{w}_\star\right)\|^2 \\
&= \operatorname*{argmin}_{m \in [T] \setminus \tau_{\text{MD}}(1:t-1)} \left(\mathbf{v}_m^\top \mathbf{v}_{\tau(t-1)}\right)^2 . \qquad (6)
\end{aligned}$$

Then, employing greediness as reformulated above and inequality of arithmetic and geometric mean, we obtain:

$$\prod_{s=1}^{T-1} \left(\mathbf{v}_{\tau(s+1)}^\top \mathbf{v}_{\tau(s)}\right)^2 \leq \prod_{s=1}^{T} \left(\mathbf{v}_{\tau(T)}^\top \mathbf{v}_{\tau(s)}\right)^2 \leq \left(\frac{1}{T} \sum_{s=1}^{T} \left(\mathbf{v}_{\tau(T)}^\top \mathbf{v}_{\tau(s)}\right)^2\right)^T .$$

Substituting back into the forgetting, it is now bounded as,

$$\mathcal{L}_{\tau_{\text{MD}}}(\mathbf{w}_T) \leq \left(1 - \frac{1}{T} \sum_{s=1}^{T} \left(\mathbf{v}_{\tau(T)}^\top \mathbf{v}_{\tau(s)}\right)^2\right) \left(\frac{1}{T} \sum_{s=1}^{T} \left(\mathbf{v}_{\tau(T)}^\top \mathbf{v}_{\tau(s)}\right)^2\right)^T \leq \frac{1}{eT}\,,$$

where we invoked an algebraic property that $(1-x)x^T \leq \frac{1}{eT}, \forall x \in [0,1]$. □

D Lower bound proof (Theorem 5.1)

Recall Theorem 5.1. For any $d \geq 30$, there exists an adversarial task collection with $T = d - 1$ jointly-realizable tasks of different rank such that both greedy orderings (MD, MR) forget *catastrophically*. That is, the loss at the end of the sequence is, $\mathcal{L}(\mathbf{w}_T^{\tau_{\mathrm{MD}}}), \mathcal{L}(\mathbf{w}_T^{\tau_{\mathrm{MR}}}) \geq \frac{1}{8} - \frac{1}{4d}$.

Proof outline. For a given dimension d, we construct a sequence of d iterates $(\mathbf{w}_t)_{t=1}^d$, corresponding to $T = d - 1$ tasks $(\mathbf{X}_t)_{t=2}^d$ of decreasing rank, which are jointly-realizable with $\mathbf{w}_\star = \mathbf{0}$ (*i.e.,* $\forall t \in \{2...T\}$, $\mathbf{y}_t = \mathbf{0}$), and show that:

1. Given this specific choice of tasks and matching iterates, the loss (or forgetting) is catastrophic as mentioned in Theorem 5.1. We start with this part as motivation.
2. The chosen iterates are a valid ordering of iterates under the chosen tasks.
3. The chosen ordering adheres to greedy selection rules, both MD and MR, under the chosen tasks. *This part is quite lengthy.*

In the construction we start the iterates from $t = 1$ and tasks from $t = 2$, contrary to other parts of the paper, for no particular reason other than ease of notation. For this same reason we chose $\mathbf{w}_\star = \mathbf{0}$, and the iterates starting with $\mathbf{w}_1 = \mathbf{e}_1$. The same construction holds for a shifted frame of reference where all iterates (and $\mathbf{w}_\star$) are shifted by $-\mathbf{e}_1$.

D.1 Construction details

We first construct the *iterates* as follows:

$$\mathbf{w}_1 = \mathbf{e}_1 = \left[1, \underbrace{0, \ldots, 0}_{d-1 \text{ times}}\right]^\top ,$$

$$\forall t \in \{2...d\} : \mathbf{w}_t = \left[\frac{(\mathbf{w}_{t-1})_1 + \sqrt{(\mathbf{w}_{t-1})_1^2 - 4\beta_t}}{2}, \underbrace{c^{t-2}\frac{1}{\sqrt{d}}, \ldots, c^{t-2}\frac{1}{\sqrt{d}}}_{t-1 \text{ times}}, 0, \ldots, 0\right]^\top ,$$

where $c \triangleq 2^{-1/d}$ and $\beta_t \triangleq \frac{((t-1)c-(t-2))c^{2t-5}}{d}$.

We denote $x_t \triangleq (\mathbf{w}_t)_1$, defined recursively by $x_1 = 1$, $x_t = \frac{x_{t-1}+\sqrt{x_{t-1}^2-4\beta_t}}{2}$, $\forall t \in \{2...d\}$.

Since $\mathbf{w}_t \neq \mathbf{w}_{t-1}$, we are free to define the unit vector

$$\mathbf{u}_t = \frac{\mathbf{w}_t - \mathbf{w}_{t-1}}{\|\mathbf{w}_t - \mathbf{w}_{t-1}\|} \in \text{span}\left(\mathbf{e}_1, \ldots, \mathbf{e}_t\right) .$$

We now construct the tasks:

$$\mathbf{X}_t = \begin{bmatrix} -\mathbf{u}_t^\top - \\ -\mathbf{e}_{t+1}^\top - \\ \vdots \\ -\mathbf{e}_d^\top - \end{bmatrix} = \begin{bmatrix} -\mathbf{u}_t^\top - \\ \mathbf{I}_{t+1:d} \end{bmatrix} \in \mathbb{R}^{(d-t+1)\times d}, \forall t \in \{2...d\} .$$

Then, it is easy to see that $\mathbf{P}_t \triangleq \mathbf{I}_d - \mathbf{X}_t^+ \mathbf{X}_t = \mathbf{I}_d - \mathbf{I}_{t+1:d} - \mathbf{u}_t\mathbf{u}_t^\top = \underbrace{\mathbf{I}_t}_{\text{rank } t} - \mathbf{u}_t\mathbf{u}_t^\top$.

D.2 Showing lower bound for the loss

For each task $\mathbf{X}_m$, its individual loss at time $t = d$ is given by:

$$\begin{aligned}
\mathcal{L}_m\left(\mathbf{w}_d\right) &\triangleq \left\|\mathbf{X}_m \mathbf{w}_d\right\|^2 = \left\| \begin{bmatrix} -\mathbf{u}_m^\top - \\ \mathbf{I}_{m+1:d} \end{bmatrix} \mathbf{w}_d \right\|^2 = \left(\mathbf{u}_m^\top \mathbf{w}_d\right)^2 + \left\|\mathbf{I}_{m+1:d} \mathbf{w}_d\right\|^2 \\
&\geq \left\|\mathbf{I}_{m+1:d} \mathbf{w}_d\right\|^2 = \sum_{j=m+1}^{d} \left(\mathbf{w}_d\right)_j^2 \\
[j \geq 2] &= (d-m) \frac{c^{2d-4}}{d} = \left(1 - \frac{m}{d}\right) c^{2d-4} = \left(1 - \frac{m}{d}\right) 2^{-(2d-4)/d} \\
&= \frac{1}{4}\left(1 - \frac{m}{d}\right) 2^{4/d} \geq \frac{1}{4}\left(1 - \frac{m}{d}\right) .
\end{aligned}$$

So the average loss after all iterates, which coincides with the forgetting (see Remark 2.4) is:

$$\begin{aligned}
\mathcal{L}\left(\mathbf{w}_d\right) &= \frac{1}{T} \sum_{m \in \{2 \ldots d\}} \mathcal{L}_m\left(\mathbf{w}_d\right) = \frac{1}{d-1} \sum_{m=2}^{d} \mathcal{L}_m\left(\mathbf{w}_d\right) \\
&\geq \frac{1}{4\left(d-1\right)} \sum_{m=2}^{d} \left(1 - \frac{m}{d}\right) = \frac{1}{4\left(d-1\right)} \left(d - 1 - \frac{\sum_{m=2}^{d} m}{d}\right) \\
&= \frac{1}{4} - \frac{d+2}{8d} = \frac{1}{8} - \frac{1}{4d} .
\end{aligned}$$

D.3 Proving that the iterates can be formed from projections of the given tasks

As a sanity check, we notice that $\mathbf{P}_t$ is a real symmetric matrix, and assert its idempotence,

$$\begin{aligned}
\mathbf{P}_t^2 &= \left(\mathbf{I}_t - \mathbf{u}_t \mathbf{u}_t^\top\right)^2 = \mathbf{I}_t^2 - \mathbf{u}_t \mathbf{u}_t^\top \mathbf{I}_t - \mathbf{I}_t \mathbf{u}_t \mathbf{u}_t^\top + \mathbf{u}_t \mathbf{u}_t^\top \mathbf{u}_t \mathbf{u}_t^\top \\
&= \mathbf{I}_t - \mathbf{u}_t \mathbf{u}_t^\top - \mathbf{u}_t \mathbf{u}_t^\top + \mathbf{u}_t \mathbf{u}_t^\top = \mathbf{I}_t - \mathbf{u}_t \mathbf{u}_t^\top = \mathbf{P}_t .
\end{aligned}$$

Firstly we show that, as required from projections, $\mathbf{w}_t^\top \left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) = 0$:

$$\begin{aligned}
\mathbf{w}_t^\top \left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) &= \sum_{i=1}^{d} \left(\mathbf{w}_t\right)_i^2 - \sum_{i=1}^{d} \left(\mathbf{w}_t\right)_i \left(\mathbf{w}_{t-1}\right)_i = \left(\mathbf{w}_t\right)_1^2 + \sum_{i=2}^{t} \left(\mathbf{w}_t\right)_i^2 - \sum_{i=1}^{t-1} \left(\mathbf{w}_t\right)_i \left(\mathbf{w}_{t-1}\right)_i \\
&= \left(\mathbf{w}_t\right)_1^2 - \left(\mathbf{w}_t\right)_1 \left(\mathbf{w}_{t-1}\right)_1 + \sum_{i=2}^{t} \frac{c^{2t-4}}{d} - \sum_{i=2}^{t-1} \frac{c^{t-2} c^{t-3}}{d} \\
&= \left(\mathbf{w}_t\right)_1^2 - \left(\mathbf{w}_t\right)_1 \left(\mathbf{w}_{t-1}\right)_1 + \frac{\left(t-1\right) c^{2t-4} - \left(t-2\right) c^{2t-5}}{d} \\
&= \left(\mathbf{w}_t\right)_1^2 - \left(\mathbf{w}_t\right)_1 \left(\mathbf{w}_{t-1}\right)_1 + \underbrace{\frac{\left(\left(t-1\right) c - \left(t-2\right)\right) c^{2t-5}}{d}}_{=\beta_t} \\
&= \left(\mathbf{w}_t\right)_1^2 - \left(\mathbf{w}_t\right)_1 \left(\mathbf{w}_{t-1}\right)_1 + \beta_t ,
\end{aligned}$$

and it is readily seen that our construction choice of $\left(\mathbf{w}_t\right)_1 = \frac{\left(\mathbf{w}_{t-1}\right)_1 + \sqrt{\left(\mathbf{w}_{t-1}\right)_1^2 - 4\beta_t}}{2}$ implies

$$\mathbf{w}_t^\top \left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) = 0 .$$

Finally, we show that the iterates are indeed a sequence the corresponding projections:

$$\begin{aligned}
\mathbf{P}_t\mathbf{w}_{t-1} &= \left(\mathbf{I}_t - \mathbf{u}_t\mathbf{u}_t^\top\right)\mathbf{w}_{t-1} = \mathbf{I}_t\mathbf{w}_{t-1} - \mathbf{u}_t\mathbf{u}_t^\top\mathbf{w}_{t-1} \\
&= \mathbf{w}_{t-1} - \left(\frac{\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)^\top}{\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|}\mathbf{w}_{t-1}\right)\mathbf{u}_t = \mathbf{w}_{t-1} - \frac{\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)^\top\mathbf{w}_{t-1}}{\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2}\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) \\
&= \mathbf{w}_{t-1} - \frac{\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)^\top\mathbf{w}_{t-1} - \overbrace{\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)^\top\mathbf{w}_t}^{=0}}{\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2}\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) \\
&= \mathbf{w}_{t-1} + \frac{\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)^\top\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right)}{\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|^2}\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) \\
&= \mathbf{w}_{t-1} + \left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) = \mathbf{w}_t\,.
\end{aligned}$$

D.4 Proving that the iterates adhere to greedy ordering rules

D.4.1 Maximum Distance (MD)

We wish to prove that the greedy MD rule agrees with the ordering we chose. That is,

$$\tau_t \triangleq \operatorname{argmax}_{t'\in[T]\setminus\{\tau_2,\ldots,\tau_{t-1}\}}\left\|\left(\mathbf{I} - \mathbf{P}_{t'}\right)\mathbf{w}_{t-1}\right\|^2 = t\,.$$

By induction on the validity of the greediness for $\tau_2,\ldots,\tau_{t-1}$, the step is (and the induction base for $t=2$ is shown exactly the same):

$$\begin{aligned}
\tau_t &\triangleq \operatorname{argmax}_{t'\in[T]\setminus\{\tau_2,\ldots,\tau_{t-1}\}}\left\|\left(\mathbf{I} - \mathbf{P}_{t'}\right)\mathbf{w}_{t-1}\right\|^2 \\
[\text{induction assumption}] &= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left\|\left(\mathbf{I}_d - \mathbf{P}_{t'}\right)\mathbf{w}_{t-1}\right\|^2 \\
&= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left\|\left(\mathbf{I}_d - \mathbf{I}_{t'} + \mathbf{u}_{t'}\mathbf{u}_{t'}^\top\right)\mathbf{w}_{t-1}\right\|^2 \\
[t' > t-1] &= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left\|\cancel{\left(\mathbf{I}_d - \mathbf{I}_{t'}\right)\mathbf{w}_{t-1}} + \mathbf{u}_{t'}\mathbf{u}_{t'}^\top\mathbf{w}_{t-1}\right\|^2 \\
\left[\left\|\mathbf{u}_{t'}\right\|^2 = 1\right] &= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left(\mathbf{u}_{t'}^\top\mathbf{w}_{t-1}\right)^2 \\
&= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left(\frac{\left(\mathbf{w}_{t'} - \mathbf{w}_{t'-1}\right)^\top}{\left\|\mathbf{w}_{t'} - \mathbf{w}_{t'-1}\right\|}\mathbf{w}_{t-1}\right)^2\,.
\end{aligned}$$

D.4.2 Maximum Residual (MR)

We wish to prove that the greedy MR rule agrees with the ordering we chose. That is,

$$\tau_t \triangleq \operatorname{argmax}_{t'\in[T]\setminus\{\tau_2,\ldots,\tau_{t-1}\}}\left\|\mathbf{X}_{t'}\mathbf{w}_{t-1}\right\|^2 = t\,.$$

By induction on the validity of the greediness for $\tau_2,\ldots,\tau_{t-1}$, the step is (and the induction base for $t=2$ is shown exactly the same):

$$\begin{aligned}
\tau_t &\triangleq \operatorname{argmax}_{t'\in[T]\setminus\{\tau_2,\ldots,\tau_{t-1}\}}\left\|\mathbf{X}_{t'}\mathbf{w}_{t-1}\right\|^2 \\
[\text{induction assumption}] &= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left\|\mathbf{X}_{t'}\mathbf{w}_{t-1}\right\|^2 = \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left\|\begin{bmatrix} -\mathbf{u}_{t'}^\top - \\ \mathbf{I}_{t'+1:d}\end{bmatrix}\mathbf{w}_{t-1}\right\|^2 \\
[t' > t-1] &= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left(\left(\mathbf{u}_{t'}^\top\mathbf{w}_{t-1}\right)^2 + \left\|\cancel{\mathbf{I}_{t'+1:d}\mathbf{w}_{t-1}}\right\|^2\right) \\
&= \operatorname{argmax}_{t'\in\{t,\ldots,T\}}\left(\frac{\left(\mathbf{w}_{t'} - \mathbf{w}_{t'-1}\right)^\top}{\left\|\mathbf{w}_{t'} - \mathbf{w}_{t'-1}\right\|}\mathbf{w}_{t-1}\right)^2\,.
\end{aligned}$$

We get that the MR and MD rules coincide in this case.

D.4.3 How we prove greediness holds: Delta positivity

We wish to show monotonous decrease (w.r.t. $k \geq t$) of $\left(\frac{\left((\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} \right)^2}{\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2} \right)_k$.

The difference between consecutive iterates is

$$\mathbf{w}_{k-1} - \mathbf{w}_k = \Bigg[x_{k-1} - x_k, \underbrace{\frac{c^{k-3}(1-c)}{\sqrt{d}}, \ldots, \frac{c^{k-3}(1-c)}{\sqrt{d}}}_{k-2 \text{ times}}, -\frac{c^{k-2}}{\sqrt{d}}, 0, \ldots, 0 \Bigg] .$$

We notice that $\forall k \geq t$ the term $(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1}$ is **positive** since,

$$(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} = \underbrace{(x_{k-1} - x_k)}_{>0, \text{ from G.6}} \underbrace{x_{t-1}}_{>0} + (t-2) \underbrace{\frac{c^{k-3}(1-c)}{\sqrt{d}} \frac{c^{t-3}}{\sqrt{d}}}_{>0} > 0 .$$

This means that we can alternatively show monotonous decrease $\forall k \geq t$ for

$$\left(\frac{(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1}}{\|\mathbf{w}_{k-1} - \mathbf{w}_k\|} \right)_k .$$

To this end, we wish to show that the next quantity is **positive** $\forall t \in \{2 \ldots d-1\}$ (we are reminded that the first step is at $t = 2$ due to our choice, and that at the last step there is only one choice), $\forall k \in \{t \ldots d-1\}$:

$$\begin{aligned} &\frac{(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1}}{\|\mathbf{w}_{k-1} - \mathbf{w}_k\|} - \frac{(\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1}}{\|\mathbf{w}_k - \mathbf{w}_{k+1}\|} \\ &\propto \|\mathbf{w}_k - \mathbf{w}_{k+1}\| (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| (\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1} \triangleq \Delta_{t,k} . \end{aligned}$$

We are going to show this holds numerically for low dimensions ($d < 25{,}000$), and prove it analytically $\forall d \geq 25{,}000$.

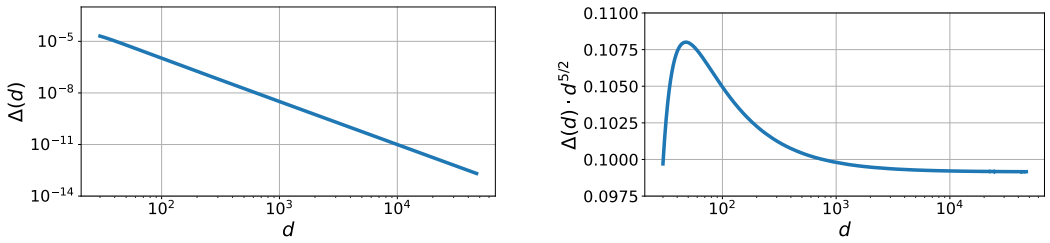
D.4.4 Showing delta positivity numerically for low dimensions

We use the following facts to write code that verifies $\Delta_{t,k} > 0 \; \forall k \geq t, \; \forall d < 25{,}000$:

$$\begin{aligned} \|\mathbf{w}_{k-1} - \mathbf{w}_k\| &= \sqrt{(x_{k-1} - x_k)^2 + (k-2)\left(\frac{c^{k-3}(1-c)}{\sqrt{d}}\right)^2 + \left(\frac{c^{k-3}}{\sqrt{d}}\right)^2} \\ &= \sqrt{(x_{k-1} - x_k)^2 + \frac{k-2}{d} c^{2k-6}(1-c)^2 + \frac{1}{d} c^{2k-6}} , \\ (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} &= (x_{k-1} - x_k) x_{t-1} + (t-2) \frac{c^{k-3}(1-c)}{\sqrt{d}} \frac{c^{t-3}}{\sqrt{d}} . \end{aligned}$$

For each value of dimension d, we calculated the series $(x)_k$ using its recursive definition, and calculated $\Delta(d) \triangleq \min_{\{t,k \,|\, t \in \{2 \ldots d-1\}, \, k \in \{t \ldots d-1\}\}} \Delta_{t,k}$ using these formulas. As shown in Figure 13, we found $\Delta(d)$ remains positive $\forall d \in \{30 \ldots 47{,}000\}$ (for completeness, any dimension above $25{,}000$ is redundant here). In addition, as will be seen analytically (Eq. (7)), we have that $\Delta(d)$ should correlate with $d^{-\frac{5}{2}}$, and for completeness we show this holds numerically for the lower dimensions as well, by showing $\Delta(d) \cdot d^{\frac{5}{2}}$ is approximately constant.

Compute resources. This numerical validation took 4 days to run on a home PC with i5-9400F CPU and 16GB RAM.



(a) $\Delta(d)$ Positivity

(b) Correlation with $d^{-\frac{5}{2}}$

Figure 13: **Numerical positivity of** $\Delta(d) \triangleq \min_{\{t,k \,|\, t\in\{2\ldots d-1\},\, k\in\{t\ldots d-1\}\}} \Delta_{t,k}$

D.4.5 Showing delta positivity analytically for high dimensions

Due to the length of this part we defer it to App. G, where we prove that $\forall k \geq t,\ \forall d \geq 25{,}000$,

$$\Delta_{t,k} \triangleq \|\mathbf{w}_k - \mathbf{w}_{k+1}\| \left(\mathbf{w}_{k-1} - \mathbf{w}_k\right)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\mathbf{w}_k - \mathbf{w}_{k+1}\right)^\top \mathbf{w}_{t-1} > 0\,.$$

Conclusion Together with the numerical verification, we have established that $\Delta_{t,k} > 0$ for all $k \geq t$ and all $d \geq 30$. This completes the proof of the iterates' adherence to the greedy ordering rules, and thereby concludes the overall proof of the adversarial construction that yields a lower bound on the loss under single-pass greedy orderings.

E Appendix for Section 5.2: Single-pass vs. repetition

E.1 Experiments on single-pass vs. repetition

Figure 5 was produced using the same data and settings as Figure 3a: $d = 100$, $r = 10$, $T = 50$.

In this section, the "Greedy" orderings use the Maximum Distance rule (Definition 3.1).

We extend the experiment on the effect of repetitions by exploring varying data settings.

Isotropic data. The conclusions of Section 5.2 extend to more regimes: repetitions are beneficial in greedy ordering while replacement harms random ordering.

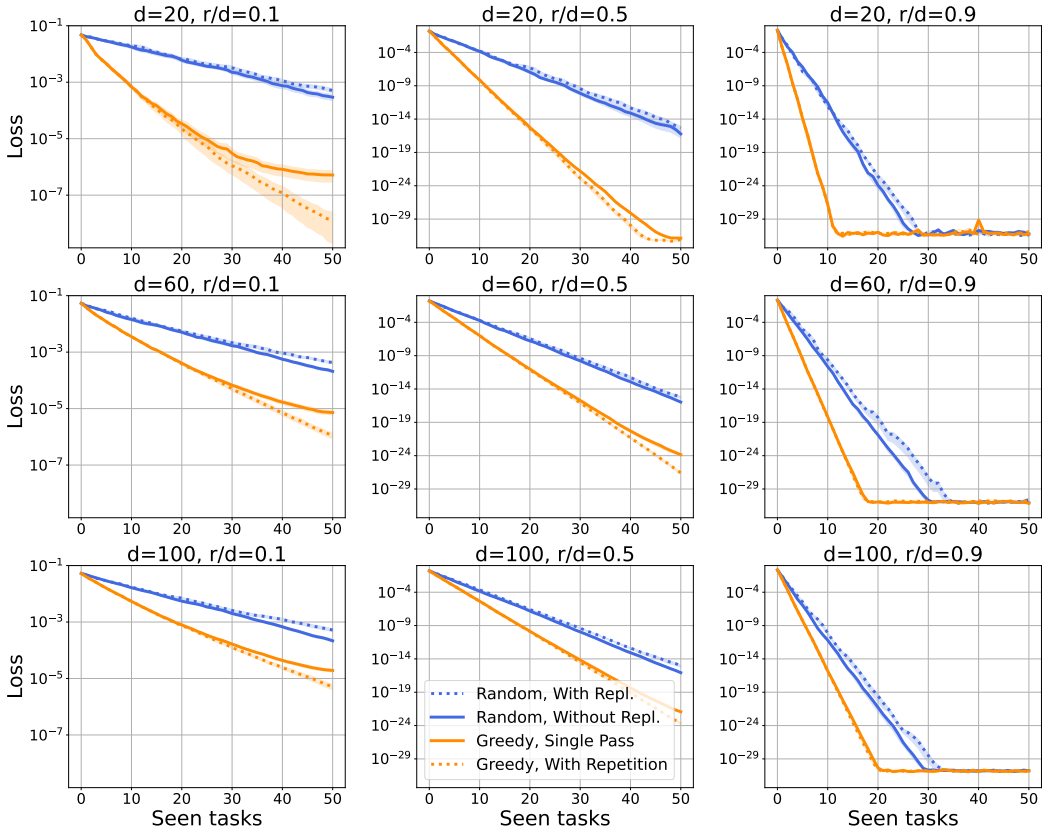


Figure 14: **The effect of repetitions for varying dimensions d and ranks r of the data matrices, for isotropic data.** $T = 50$. Random orderings without-replacement consistently outperform their with-replacement counterparts. In contrast, greedy orderings benefit from repetition: allowing repeated tasks yields better performance than the single-pass variant. As we explained in Section 5.2, repetition in greedy orderings outperforms no repetition because they enable taking larger steps (and converge faster to the offline solution $\mathbf{w}_\star$).

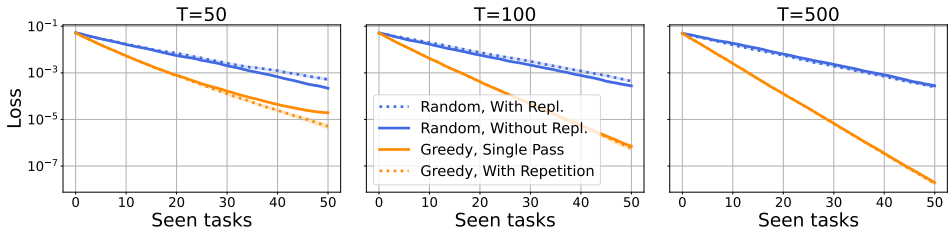


Figure 15: **The effect of repetitions for varying task count T, for isotropic data.** $d = 100,\ r = 10$. As task count increases, the differences between with and without repetition diminish. Notice, however, that in all subplots we only learn the first 50 tasks. It is readily observed in the left subplot that the effect of repetition becomes pronounced in the latter parts of the task sequences. As can be expected, repetition offers less benefit when many diverse, unexplored tasks remain.

Anisotropic data. Next, we observe that the effect of repetitions diminishes for correlated data.

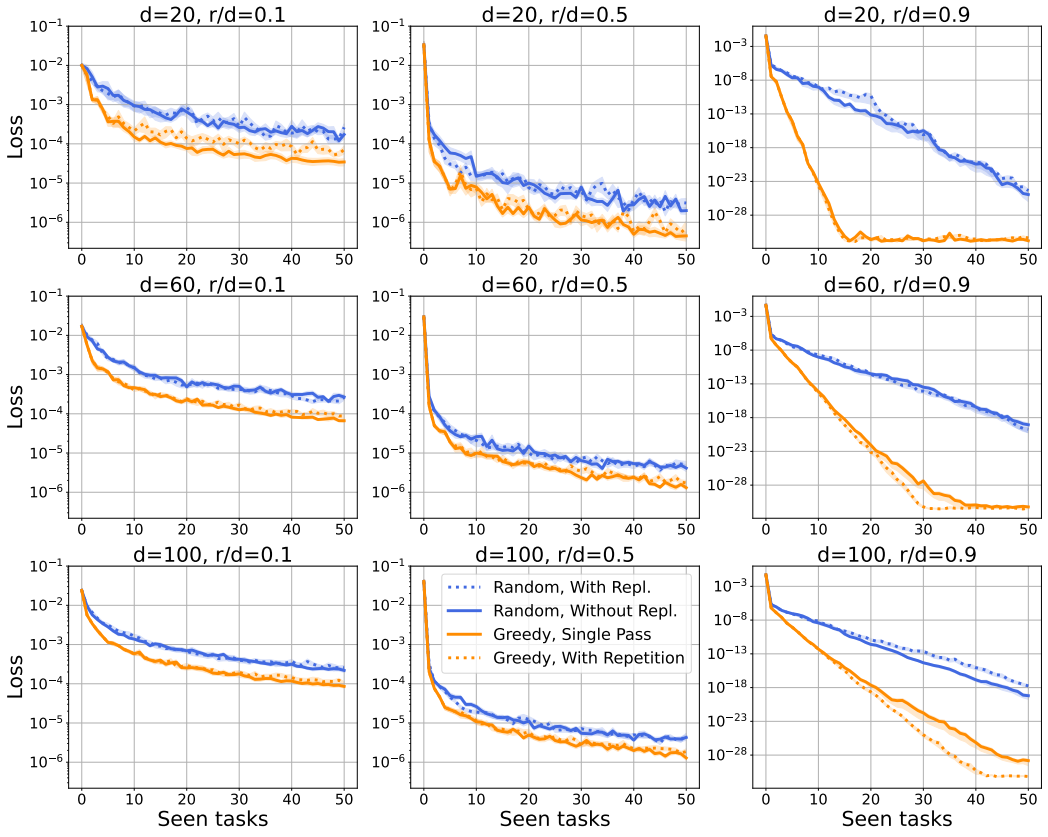


Figure 16: **The effect of repetitions for varying dimensions d and ranks r of the data matrices, for anisotropic data.** $T = 50$. In highly correlated settings, repetitions become less impactful due to the inherent similarity between tasks. Interestingly, in low-rank settings (left column), task repetition can slightly hinder the performance of greedy strategies. We hypothesize that repetition causes greedy orderings to alternate between a small subset of tasks with relatively large mutual angles, while neglecting others. In this regime, tasks are highly similar, and convergence toward the offline solution $\mathbf{w}_\star$ is inherently slow, reducing its utility as an upper bound on the loss (Proposition 2.6). As a result, neglecting some tasks (which are of low rank) may harm the average loss, even if it improves proximity to $\mathbf{w}_\star$.

Remark. We omit the figure for the corresponding experiment with varying number of tasks T, as it offers no additional insights beyond those shown in Figure 15.

E.2 Proof of upper bound for loss of greedy orderings with repetition

Recall Proposition 5.2. For any task collection of T jointly realizable tasks, the loss under greedy maximum distance (MD) ordering *with repetition*, *i.e.,* $\tau_{\text{MD-R}}$, after $k \geq 2$ iterations, is upper bounded as $\mathcal{L}(\mathbf{w}_k^{\tau_{\text{MD-R}}}) = \mathcal{O}(1/\log k)$.

In order to prove Proposition 5.2, we first prove the following propositions:

Proposition E.1. *Under greedy MD ordering, either single-pass or with repetition, we have:*

$$\|\mathbf{w}_0 - \mathbf{w}_t\|^2 \leq t \|\mathbf{w}_0 - \mathbf{w}_1\|^2 .$$

Proof. $t = 1$ is trivial. Consider $t \geq 2$,

$$\begin{aligned}
\|\mathbf{w}_0 - \mathbf{w}_t\|^2 &= \|(\mathbf{w}_0 - \mathbf{w}_\star) - (\mathbf{w}_t - \mathbf{w}_\star)\|^2 \\
&= \left\|\left(\mathbf{w}_0 - \mathbf{w}_\star - \mathbf{P}_{\tau(t)}\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right) - \left(\mathbf{w}_t - \mathbf{w}_\star - \mathbf{P}_{\tau(t)}\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right)\right\|^2 \\
\text{[Eq. (2)]} &= \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right) - \mathbf{P}_{\tau(t)}\left(\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right) - \left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right)\right\|^2 \\
\text{[orthogonal proj.]} &= \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2 + \left\|\mathbf{P}_{\tau(t)}\left(\mathbf{w}_0 - \mathbf{w}_{t-1}\right)\right\|^2 \\
\text{[contraction]} &\leq \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2 + \|\mathbf{w}_0 - \mathbf{w}_{t-1}\|^2 \\
\text{[recursively]} &\leq \sum_{i=2}^{t} \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(i)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2 + \|\mathbf{w}_0 - \mathbf{w}_1\|^2 \\
&= \sum_{i=1}^{t} \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(i)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2 ,
\end{aligned}$$

and specifically, under the greedy policy, **either single-pass or with repetition**, we get,

$$\|\mathbf{w}_0 - \mathbf{w}_t\|^2 \leq t \left\|\left(\mathbf{I} - \mathbf{P}_{\tau(1)}\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2 = t \|\mathbf{w}_0 - \mathbf{w}_1\|^2 .$$

□

Proposition E.2. *Under greedy MD ordering, either single-pass or with repetition, we have:*

$$\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 \leq 2t \|\mathbf{w}_0 - \mathbf{w}_1\|^2 .$$

Proof. $t = 1$ is trivial. Consider $t \geq 2$,

$$\begin{aligned}
\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 &= \|(\mathbf{w}_{t-1} - \mathbf{w}_\star) - (\mathbf{w}_t - \mathbf{w}_\star)\|^2 \\
\text{[Eq. (2)]} &= \|(\mathbf{w}_{t-1} - \mathbf{w}_\star) - \mathbf{P}_t\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\|^2 \\
\text{[projection]} &\leq \|(\mathbf{w}_{t-1} - \mathbf{w}_\star) - \mathbf{P}_t\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\|^2 \\
&= \|(\mathbf{w}_{t-1} - \mathbf{w}_\star) - (\mathbf{w}_0 - \mathbf{w}_\star) - (\mathbf{P}_t\left(\mathbf{w}_0 - \mathbf{w}_\star\right) - (\mathbf{w}_0 - \mathbf{w}_\star))\|^2 \\
&\leq 2\left(\|\mathbf{w}_{t-1} - \mathbf{w}_0\|^2 + \|(\mathbf{I} - \mathbf{P}_t)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\|^2\right) \\
\text{[greedy+above]} &\leq 2\left((t-1)\|\mathbf{w}_0 - \mathbf{w}_1\|^2 + \|\mathbf{w}_0 - \mathbf{w}_1\|^2\right) \\
&= 2t \|\mathbf{w}_0 - \mathbf{w}_1\|^2 .
\end{aligned}$$

□

Proposition E.3. *Under greedy MD ordering, either single-pass or with repetition, we have* $\forall k \geq 2$*:*

$$\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 < \frac{2\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\ln k}\,.$$

Proof. We showed $\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 \leq 2t\|\mathbf{w}_0 - \mathbf{w}_1\|^2$, and thus $\forall k > t$,

$$\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 \leq 2(k-t+1)\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2$$
$$\frac{1}{2(k-t+1)}\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 \leq \|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2\,.$$

From the Pythagorean theorem we have,

$$\|\mathbf{w}_k - \mathbf{w}_\star\|^2 = \|\mathbf{w}_{k-1} - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 = \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=1}^{k}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2\,.$$

Combining, we get,

$$\begin{aligned}
\|\mathbf{w}_k - \mathbf{w}_\star\|^2 &= \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=1}^{k}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 \\
&\leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=1}^{k}\frac{1}{2(k-t+1)}\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 \\
&= \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2\frac{1}{2}\sum_{i=1}^{k}\frac{1}{i} \\
&\leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2\frac{\ln k}{2}\,.
\end{aligned}$$

And finally, from Proposition 2.6,

$$\begin{aligned}
&0 \leq \mathcal{L}(\mathbf{w}_k) \leq \tfrac{1}{\|\mathbf{w}_\star\|^2}\|\mathbf{w}_k - \mathbf{w}_\star\|^2 \leq \tfrac{1}{\|\mathbf{w}_\star\|^2}\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2\frac{\ln k}{2}\right) \\
\implies &\|\mathbf{w}_{k-1} - \mathbf{w}_k\|^2 \leq \frac{2\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\ln k}\,.
\end{aligned}$$

□

We are now ready prove Proposition 5.2:

Proof. Under greedy MD ordering with repetitions we have:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_k) &= \tfrac{1}{\|\mathbf{w}_\star\|^2R^2T}\sum_{m=1}^{T}\|\mathbf{X}_m\mathbf{w}_k - \mathbf{y}_m\|^2 = \tfrac{1}{\|\mathbf{w}_\star\|^2R^2T}\sum_{m=1}^{T}\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2 \\
&\leq \tfrac{1}{\|\mathbf{w}_\star\|^2R^2T}\sum_{m=1}^{T}\|\mathbf{X}_m\|^2\|(\mathbf{I} - \mathbf{P}_m)(\mathbf{w}_k - \mathbf{w}_\star)\|^2 \\
&\leq \tfrac{1}{\|\mathbf{w}_\star\|^2T}\sum_{m=1}^{T}\|(\mathbf{I} - \mathbf{P}_m)(\mathbf{w}_k - \mathbf{w}_\star)\|^2 \\
\text{[greedy+repetitions]} &\leq \tfrac{1}{\|\mathbf{w}_\star\|^2}\|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2 \\
\text{[above, } \mathbf{w}_0 = \mathbf{0}\text{]} &\leq \tfrac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}\frac{2\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\ln(k+1)} = \frac{2}{\ln(k+1)}.
\end{aligned}$$

□

F Appendix for Section 5.3: Hybrid task ordering

F.1 Hybrid ordering experiments

Figure 6 was acquired using the same data as Figure 3a, and using the dimension and rank-dependent upper bound of $2\,(d-r)/k$ from Evron et al. [26] to set β, since the universal bound of $14/k^{1/4}$ requires more than 50 iterations to be effective. The hybrid method results with intermediate performance between random and greedy. The figures demonstrate that the hybrid approach combines trends we have seen earlier (App. B) for random and greedy MD, in terms of the effect of dimension, rank, task count and task correlation on the performance.

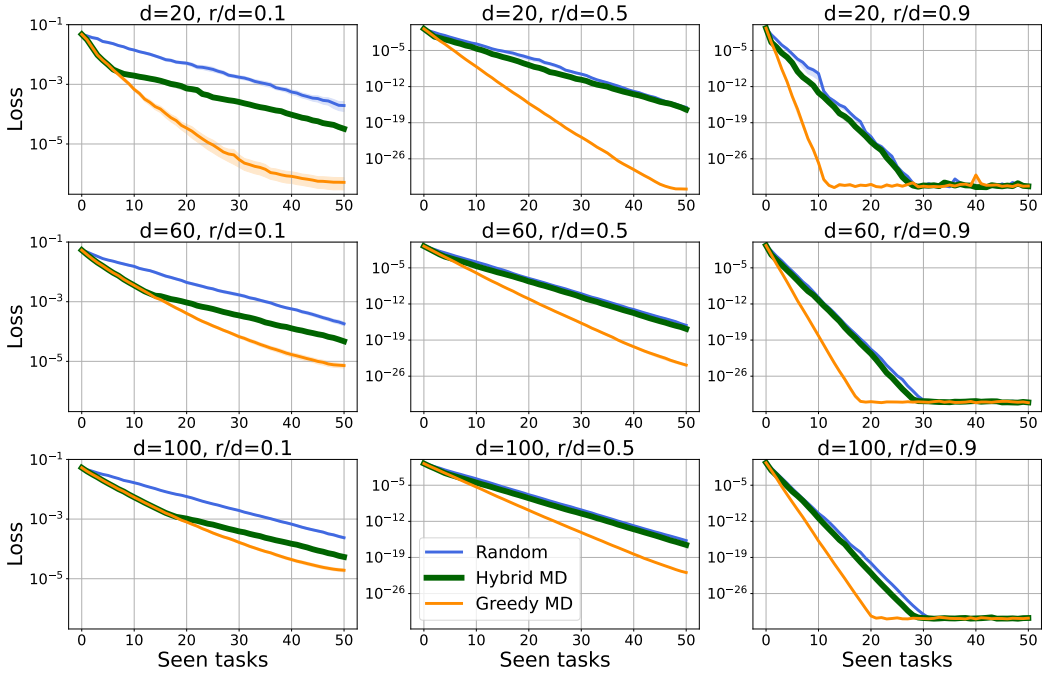


Figure 17: **Hybrid performance for varying dimensions d and ranks r of the data matrices, for isotropic data.** $T = 50$. In high-rank and/or low-dimensional settings, the rank-dependent upper bound employed by the hybrid strategy in this case is lower, prompting an earlier transition from the greedy to the random phase. Interestingly, the performance of the random phase within the hybrid method is slightly inferior to that of fully random ordering—possibly because the initial greedy steps deplete the set of "extreme" tasks that would otherwise drive greater progress.

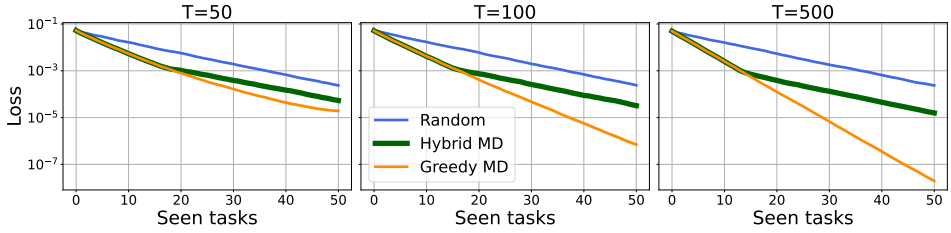


Figure 18: **Hybrid performance for varying task count T, for isotropic data.** $d = 100,\ r = 10$. We see similar trends. Note that the previously observed slight drop in performance of the random iterates following the greedy phase is less pronounced with higher task counts, possibly since more extreme tasks remain available for selection.

Anisotropic data. Similar trends were observed under anisotropic data, and we therefore omit the corresponding figures for brevity.

F.2 Proof of the hybrid upper bound

Recall Theorem 5.3 (informal). Assume any bound of the form $\mathbb{E}_{\tau_{\mathrm{Unif}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\mathrm{Unif}}})\right] = \mathcal{O}\left(1/k^{\alpha}\right)$, $\alpha \in (0,1]$, established for the without-replacement τ_{Unif}. Then, setting a threshold of $\beta = \Omega\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T^{1-\alpha}}\right)$, guarantees a similar bound $\mathbb{E}_{\tau_{\mathrm{H}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\mathrm{H}}})\right] = \mathcal{O}\left(1/k^{\alpha}\right)$ for the hybrid ordering τ_{H}.

In more exact terms, we will show the following holds:

Full version of Theorem 5.3. Given a known upper bound for the expected normalized loss (Definition 2.3) in random ordering without replacement of T jointly-realizable tasks, of the form $\mathbb{E}_{\tau_{\mathrm{Unif}}}\left[\mathcal{L}(\mathbf{w}_k^{\tau_{\mathrm{Unif}}})\right] \leq \frac{C}{k^{\alpha}}$ with $C > 0$ and $0 < \alpha \leq 1$, for T such that $\frac{C}{T^{\alpha}} \leq \frac{1}{2-\alpha}$, Scheme 2 is sure to give a lower upper bound on the expected loss when $\beta \geq \beta_{\min} = \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 \frac{T^{\alpha} - C(1-\alpha)}{CT}$.

Proof. We denote $\beta = \tilde{\beta}\|\mathbf{w}_0 - \mathbf{w}_\star\|^2$. The last step t for which $\max_{m\in[T]\setminus\tau(1:t-1)} \|(\mathbf{I} - \mathbf{P}_m)(\mathbf{w}_{t-1} - \mathbf{w}_\star)\|^2 \geq \tilde{\beta}\|\mathbf{w}_0 - \mathbf{w}_\star\|^2$ *consecutively* holds is some $t = s$, where $0 \leq s \leq k$. The following holds:

$$\|\mathbf{w}_s - \mathbf{w}_\star\|^2 = \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=1}^{s}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2\left(1 - \tilde{\beta}s\right) .$$

We are reminded of the definition for the normalized loss for a solution vector $\mathbf{w}$ with a task collection of T tasks $[T]$, starting from some starting point $\mathbf{w}_0$ and having a minimum norm offline joint solution $\mathbf{w}_\star$:

$$\mathcal{L}^{([T],\mathbf{w}_0)}[\mathbf{w}] \triangleq \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\sum_{m\in[T]}\|\mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star)\|^2 .$$

If we perform k iterations of this algorithm, where $0 \leq s \leq k \leq T$ (unless $k = T$, then $s \leq T-1$ since there is no meaning to the ordering in the last step when there is only one task), then:

$$\begin{aligned}
&\mathbb{E}_\tau \mathcal{L}_\tau^{([T],\mathbf{w}_0)}[\mathbf{w}_k] = \frac{1}{\|\mathbf{w_0} - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\sum_{m=1}^{T}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right] \\
&= \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\left[\sum_{t=1}^{s}\mathbb{E}\left[\left\|\mathbf{X}_{\tau(t)}(\mathbf{w}_k - \mathbf{w}_\star)\right\|^2\right] + \sum_{m\in[T]\setminus\tau(1:s)}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right]\right] \\
&\leq \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\left[R^2\sum_{t=1}^{s}\mathbb{E}\left[\|\mathbf{w}_k - \mathbf{w}_\star\|^2\right] + \sum_{m\in[T]\setminus\tau(1:s)}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right]\right] \\
&\overset{(1)}{\leq} \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\left[R^2\sum_{t=1}^{s}\mathbb{E}\left[\|\mathbf{w}_s - \mathbf{w}_\star\|^2\right] + \sum_{m\in[T]\setminus\tau(1:s)}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right]\right] \\
&\overset{(2)}{=} \frac{1}{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 R^2}\frac{1}{T}\left[R^2 s\|\mathbf{w}_s - \mathbf{w}_\star\|^2 + \sum_{m\in[T]\setminus\tau(1:s)}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right]\right] \\
&= \frac{\|\mathbf{w}_s - \mathbf{w}_\star\|^2}{T\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}\left(s + (T-s)\left(\frac{1}{\|\mathbf{w}_s - \mathbf{w}_\star\|^2 R^2}\frac{1}{T-s}\sum_{m\in[T]\setminus\tau(1:s)}\mathbb{E}\left[\|\mathbf{X}_m(\mathbf{w}_k - \mathbf{w}_\star)\|^2\right]\right)\right) \\
&= \frac{\|\mathbf{w}_s - \mathbf{w}_\star\|^2}{T\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}\left(s + (T-s)\,\mathbb{E}_\tau \mathcal{L}_\tau^{([T]\setminus\tau(1:s),\mathbf{w}_s)}[\mathbf{w}_k]\right) \\
&\leq \frac{1 - \tilde{\beta}s}{T}\left(s + (T-s)\,\mathbb{E}_\tau \mathcal{L}_\tau^{([T]\setminus\tau(1:s),\mathbf{w}_s)}[\mathbf{w}_k]\right) .
\end{aligned}$$

Where (1) is since $s \leq k$, and (2) is since $\mathbf{w}_s$ is deterministic. This means we can plug in any upper bound for the expected normalized loss of the random ordering, for the collection of $T - s$

tasks $[T] \setminus \tau(1:s)$ with the starting point $\mathbf{w}_s$, replacing dependence on k with $k-s$. If we have an upper bound for the expected normalized loss of random ordering of $f(k)$, which is a positive and decreasing function of k, we have:

$$\mathbb{E}_\tau \mathcal{L}_\tau^{([T],\mathbf{w}_0)}[\mathbf{w}_k] \leq \frac{1-\tilde{\beta}s}{T}\left(s+(T-s)f(k-s)\right) .$$

Plugging in $s=0$, we get no greedy iterates and thus the bound is exactly what you get for random ordering.

We want a condition on $\tilde{\beta}$ for which continuing with greedy iterates as long as the condition from Scheme 2 holds, necessarily improves the bound. This means we want the bound to decrease with s. Thus we demand $\forall s \in [k]: \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1-\tilde{\beta}s}{T}\left(s+(T-s)f(k-s)\right)\right) \leq 0$:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1-\tilde{\beta}s}{T}\left(s+(T-s)f(k-s)\right)\right) \\
&= \frac{1}{T}\left(-\tilde{\beta}\left(s+(T-s)f(k-s)\right)+\left(1-\tilde{\beta}s\right)\left(1+\left(-f(k-s)-(T-s)f'(k-s)\right)\right)\right) \\
&= \frac{1}{T}\Big(-\tilde{\beta}s-\tilde{\beta}Tf(k-s)+\tilde{\beta}sf(k-s)+1-f(k-s)-(T-s)f'(k-s)-\tilde{\beta}s \\
&\qquad +\tilde{\beta}sf(k-s)+\tilde{\beta}s(T-s)f'(k-s)\Big) \\
&= \frac{1}{T}\left(1-2\tilde{\beta}s-\left(1+\tilde{\beta}T-2\tilde{\beta}s\right)f(k-s)-\left(1-\tilde{\beta}s\right)(T-s)f'(k-s)\right) ,
\end{aligned}$$

and when demanding this to be ≤ 0 we get:

$$\begin{aligned}
&\tilde{\beta}\left(-2s-(T-2s)f(k-s)+s(T-s)f'(k-s)\right) \leq -1+f(k-s)+(T-s)f'(k-s) \\
&\tilde{\beta} \geq \frac{1-f(k-s)-(T-s)f'(k-s)}{Tf(k-s)+2s\left(1-f(k-s)\right)-s(T-s)f'(k-s)} .
\end{aligned}$$

Note that when $f(k-s) \leq 1$, which is the only interesting case for upper bounds, and since $f'(k-s)$ is negative, both the numerator and denominator are positive.

Continuing:

$$\begin{aligned}
\tilde{\beta} &\geq \frac{1-f(k-s)-(T-s)f'(k-s)}{Tf(k-s)+2s\left(1-f(k-s)\right)-s(T-s)f'(k-s)} \\
&= \frac{1-f(k-s)-(T-s)f'(k-s)}{s\left(1-f(k-s)-(T-s)f'(k-s)\right)-s+Tf(k-s)+s-sf(k-s)} \\
&= \left(s+\frac{(T-s)f(k-s)}{1-f(k-s)-(T-s)f'(k-s)}\right)^{-1} \\
\tilde{\beta}^{-1} &\leq s+\frac{(T-s)f(k-s)}{1-f(k-s)-(T-s)f'(k-s)} .
\end{aligned}$$

We demand this holds $\forall s \in [k]$. If we further assume convexity of f, which is the common case for such upper bounds, we can notice that this expression decreases with k, so we can get a stronger bound which doesn't depend on our choice of k if we demand:

$$\tilde{\beta}^{-1} \leq s+\frac{(T-s)f(T-s)}{1-f(T-s)-(T-s)f'(T-s)} .$$

Moreover, if we assume a polynomial bound of the form $f(k)=\frac{C}{k^\alpha}$ where $0<\alpha\leq 1$, we get that $f'(k)=-\frac{\alpha C}{k^{\alpha+1}}$, and thus:

$$\begin{aligned}
\tilde{\beta}^{-1} &\leq s+\frac{(T-s)\frac{C}{(T-s)^\alpha}}{1-\frac{C}{(T-s)^\alpha}+(T-s)\frac{\alpha C}{(T-s)^{\alpha+1}}} = s+\frac{C(T-s)^{1-\alpha}}{1-\frac{C}{(T-s)^\alpha}+\frac{\alpha C}{(T-s)^\alpha}} \\
&= s+\frac{C(T-s)}{(T-s)^\alpha\left(1-\frac{C(1-\alpha)}{(T-s)^\alpha}\right)} = s+\frac{C(T-s)}{(T-s)^\alpha-C(1-\alpha)} \triangleq g(s) .
\end{aligned}$$

We are looking for an upper bound on $\tilde{\beta}^{-1}$ that will hold for all values of s. We can show $g(s)$ increases with s:

$$\begin{aligned}
\frac{\mathrm{d}g(s)}{\mathrm{d}s} &= 1 + \frac{-C\left((T-s)^{\alpha} - C(1-\alpha)\right) - C(T-s)\left(-\alpha(T-s)^{\alpha-1}\right)}{\left((T-s)^{\alpha} - C(1-\alpha)\right)^2} \\
&= 1 - C\frac{-\alpha(T-s)^{\alpha} + (T-s)^{\alpha} - C(1-\alpha)}{\left((T-s)^{\alpha} - C(1-\alpha)\right)^2} \\
&= 1 + \alpha C\frac{(T-s)^{\alpha}}{\left((T-s)^{\alpha} - C(1-\alpha)\right)^2} - C\frac{1}{(T-s)^{\alpha} - C(1-\alpha)} \\
&\geq 1 + \alpha C\frac{(T-s)^{\alpha} - C(1-\alpha)}{\left((T-s)^{\alpha} - C(1-\alpha)\right)^2} - C\frac{1}{(T-s)^{\alpha} - C(1-\alpha)} \\
&= 1 - \frac{C(1-\alpha)}{(T-s)^{\alpha} - C(1-\alpha)} = \frac{(T-s)^{\alpha} - 2C(1-\alpha)}{(T-s)^{\alpha} - C(1-\alpha)}\,.
\end{aligned}$$

This derivative is positive when $(T-s)^{\alpha} \geq 2C(1-\alpha)$.

We note that if $(T-s)^{\alpha} \leq C$, the upper bound on the loss is better if we don't switch to random ordering at all (if we ever get to such a large value of s). This means we assume $(T-s)^{\alpha} > C > C(1-\alpha)$. Moreover, even if the derivative switches sign, we can see that the upper bound on $\tilde{\beta}^{-1}$ for $s_{\max} = T - C^{1/\alpha}$ will still be larger than the upper bound for $s = 0$:

$$\begin{aligned}
g\left(T - C^{1/\alpha}\right) &= T - C^{1/\alpha} + \frac{C \cdot C^{1/\alpha}}{C - C(1-\alpha)} = T - C^{1/\alpha} + \frac{C^{1/\alpha}}{\alpha} \\
&= T + C^{1/\alpha}\left(\alpha^{-1} - 1\right) \geq T \\
g(0) &= \frac{CT}{T^{\alpha} - C(1-\alpha)} \\
g\left(T - C^{1/\alpha}\right) - g(0) &\geq T\left(1 - \frac{C}{T^{\alpha} - C(1-\alpha)}\right) = T\left(\frac{T^{\alpha} - C(2-\alpha)}{T^{\alpha} - C(1-\alpha)}\right)\,.
\end{aligned}$$

This can only be negative when $T < \left(C(2-\alpha)\right)^{1/\alpha}$, for which the bound $f(T) = \frac{C}{T^{\alpha}} > \frac{1}{2-\alpha}$. If we only care about values of T such that $f(T) \leq \frac{1}{2-\alpha}$, since the bound is quite useless if it is larger than $\frac{1}{2}$ anyway, it is guaranteed that the lowest upper bound for $\tilde{\beta}^{-1}$ is for $s = 0$, and we get:

$$\begin{aligned}
\tilde{\beta}^{-1} &\leq \frac{CT}{T^{\alpha} - C(1-\alpha)} \\
\tilde{\beta} &\geq \tilde{\beta}_{\min} = \frac{T^{\alpha} - C(1-\alpha)}{CT}\,.
\end{aligned}$$

□

G Delta positivity proof

This section supplements App. D, we recommend reviewing it beforehand if you have not already done so.

Reminder. In this section we prove that

$$\Delta_{t,k} \triangleq \|\mathbf{w}_k - \mathbf{w}_{k+1}\| \left(\mathbf{w}_{k-1} - \mathbf{w}_k\right)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\mathbf{w}_k - \mathbf{w}_{k+1}\right)^\top \mathbf{w}_{t-1} > 0,$$

$\forall k \geq t, \forall d \geq 25{,}000$.

In some places in our proofs, we will need a closed-form approximation of the first coordinates $x_k \triangleq (\mathbf{w}_k)_1$ which we obtain recursively. Let us propose such an approximation:

$$\tilde{x}_k = \sqrt{1 - \frac{1}{\ln 4} + 4^{-\frac{k}{d}} \left(\frac{1}{\ln 4} - \frac{k}{d}\right)}.$$

This will be formalized and proven in App. H. In addition this gives us a lower bound $x_k \geq 0.45,\ \forall k \in [d]$ when $d \geq 25{,}000$ (Corollary H.2).

G.1 Proof outline

The proof is straightforward - we decompose $\Delta_{t,k}$ to smaller parts, and attempt to lower bound each of these parts. We then combine all of these lower bounds to achieve a lower bound on $\Delta_{t,k}$ and find a sufficient condition on d for which this lower bound is positive. This condition, revealed in Eq. (7), is already satisfied when $d \geq 25{,}000$, concluding the proof. We begin by bounding some intermediate quantities that appear later in the derivation, and starting in App. G.3.6 we decompose and lower bound $\Delta_{t,k}$.

G.2 Auxiliary: Algebraic inequalities

Claim G.1. $\forall d \in \mathbb{N}$ and $1 \leq n \leq d$, it holds that $1 - c^n \triangleq 1 - 2^{-n/d} \in \left[\frac{n\ln(2)}{d} - \frac{n^2\ln^2(2)}{2d^2},\ \frac{n\ln(2)}{d}\right]$. Particularly, this shows $1 - c \in \left[\frac{\ln(2)}{d} - \frac{\ln^2(2)}{2d^2},\ \frac{\ln(2)}{d}\right]$.

Proof. To show the upper bound, we define $\alpha = n/d \in (0, 1]$ and $f(\alpha) = 1 - 2^{-\alpha} - \alpha\ln(2)$, and notice that f is *decreasing* in $(0, 1]$ since

$$f'(\alpha) = \left(2^{-\alpha} - 1\right)\ln(2) \propto 2^{-\alpha} - 1 < 0, \quad \forall \alpha \in (0, 1]\ .$$

Then, this means $f(\alpha) = 1 - 2^{-n/d} - \frac{n\ln(2)}{d} \leq \lim_{\alpha\to 0^+} f(\alpha) = 0$ as required.

Conversely, we get the lower bound by showing that the function $g\left(\alpha = \frac{n}{d}\right) = 1 - 2^{-n/d} - \left(\frac{n\ln(2)}{d} - \frac{n^2\ln^2(2)}{2d^2}\right)$ is *increasing* in $(0, 1]$,

$$\begin{aligned} g(\alpha) &= 1 - 2^{-\alpha} - \left(\alpha\ln(2) - \frac{\alpha^2\ln^2(2)}{2}\right),\ \lim_{\alpha\to 0^+} g(0) = 1 - 2^{-0} - 0 = 0\,, \\ g'(\alpha) &= \ln(2)\left(2^{-\alpha} + \alpha\ln(2) - 1\right) \propto 2^{-\alpha} + \alpha\ln(2) - 1 = -f(\alpha) + 1 \\ &\geq -\lim_{\alpha\to 0^+} f(\alpha) + 1 = -\left(1 - 2^{-0} - 0\right) + 1 = 1 > 0\,. \end{aligned}$$

□

Claim G.2. For $\forall d, n, m \in \mathbb{N}$ and $k \in [d]$, we have $c^{nk-m} \geq 2^{-n}$.

Proof. Notice that $c^z = 2^{-z/d}$ is *decreasing* with z. Plugging in $z = nk - m \leq nd$, we get $c^z \geq c^{nd} = 2^{-n}$. □

Claim G.3. $\forall k \in [1, d]$ it holds that $1 - (1 - c)(k - 1) = ((k - 1)c - (k - 2)) \in [0, 1]$.

Proof. It is clear that $(1-c)(k-1) \triangleq \left(1-2^{-1/d}\right)(k-1) \geq 0$. Then, we can simply show that from Claim G.1:

$$\underbrace{(1-c)}_{\geq 0}(k-1) \leq (1-c)(d-1) \leq \frac{\ln(2)}{d}(d-1) < \ln(2) < 1\,.$$

□

Claim G.4. $\forall k \in [1,d]$ it holds that $kc-(k-1) > 0$.

Proof.

$$kc-(k-1) \geq k\left(1-\frac{\ln 2}{d}\right) - k + 1 = 1 - \ln 2\frac{k}{d} \geq 1 - \ln 2 > 0\,.$$

□

Claim G.5. $\forall k \in [d]$ it holds that $\beta_k \in \left[\frac{0.3c^{2k-5}}{d},\ \frac{c^{2k-5}}{d}\right]$.

Proof.

$$\beta_k = \frac{((k-1)c-(k-2))c^{2k-5}}{d} = \Big(1 - \underbrace{(1-c)(k-1)}_{\in[0,\ln(2)]\subset[0,0.7]}\Big)\cdot\frac{c^{2k-5}}{d} \in \left[\frac{0.3c^{2k-5}}{d},\ \frac{c^{2k-5}}{d}\right].$$

□

Claim G.6. x_k is decreasing and $\forall k \in [d]\,,\ x_k \leq 1$.

Proof. Decreasing follows immediately from positivity of β_k (see Claim G.5) and the construction, and since $x_1 = 1$ we get $\forall k \in [d]\,,\ x_k \leq 1$. □

Claim G.7. $\forall k \in [2,d]$ it holds that $\beta_k \leq \frac{1}{cd}$.

Proof.

$$\beta_k \leq \frac{c^{2k-5}}{d} \leq \frac{c^{2\cdot 2-5}}{d} = \frac{1}{cd}\,.$$

□

Claim G.8. $\forall a > 0, b \in \mathbb{R}\setminus\{0\}$ such that $a+b \geq 0$, it holds that $\sqrt{a+b} < \sqrt{a} + \frac{b}{2\sqrt{a}}$.

Proof.

$$\begin{aligned} 0 < b^2 &\iff 4a(a+b) < 4a^2 + 4ab + b^2 \\ &\iff 2\sqrt{a(a+b)} < 2a + b \iff \sqrt{a+b} < \sqrt{a} + \frac{b}{2\sqrt{a}}\,. \end{aligned}$$

□

Claim G.9. $\forall d \geq 1:\ 2^{1/d} \geq 1 + \frac{\ln 2}{d}$

Proof. Using Taylor's expansion:

$$2^{1/d} = e^{\frac{\ln 2}{d}} = 1 + \frac{\ln 2}{d} + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\frac{\ln 2}{d}\right)^i \geq 1 + \frac{\ln 2}{d}$$

□

Claim G.10. If $|x_k - \tilde{x}_k| \leq \epsilon$ and $x_k \geq 0$, $\left|x_k^2 - \tilde{x}_k^2\right| \leq 2x_k\epsilon_d + \epsilon_d^2$.

Proof. Defining $r = \tilde{x}_k - x_k$, we have,

$$\begin{aligned} \left|x_k^2 - \tilde{x}_k^2\right| &= \left|x_k^2 - (x_k+r)^2\right| = \left|2x_k r - r^2\right| \leq |2x_k r| + r^2 = 2\left|x_k(\tilde{x}_k - x_k)\right| + (\tilde{x}_k - x_k)^2 \\ &\leq 2x_k\epsilon + \epsilon^2\,. \end{aligned}$$

□

G.3 Proof body

G.3.1 Analyzing $x_{k-1}-x_k$, $(x_{k-1}-x_k)^2$, $\frac{x_k}{x_{k-1}}$

Proposition G.11. *For any $k \geq 2$, it holds that,*

$$x_{k-1}-x_k \triangleq f_{x_{k-1}}(\beta_k) \in \left[\frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3}+\frac{2\beta_k^3}{x_{k-1}^5},\ \frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3}+\frac{2\beta_k^3}{\left(x_{k-1}^2-4\beta_k\right)^{5/2}}\right]$$
$$[\textit{when } d \geq 25{,}000] \subseteq \left[\frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3},\ \frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3}+\frac{113c^{6k-15}}{d^3}\right]$$

Proof. By construction, we have

$$x_{k-1}-x_k = \frac{x_{k-1}-\sqrt{x_{k-1}^2-4\beta_k}}{2}.$$

Define $f_z(x)=\frac{1}{2}\left(z-\sqrt{z^2-4x}\right)$ for $z\in[0.45,1]$ (see Claim G.6, Corollary H.2) and $z^2 \gg x > 0$.
Expand with Taylor:

$$f_z(0) = f_z(0)$$

$$f_z^{(1)}(x) = -\frac{1}{4}\frac{-4}{\sqrt{z^2-4x}} = \frac{1}{\sqrt{z^2-4x}} \qquad f_z^{(1)}(0) = \frac{1}{z}$$

$$f_z^{(2)}(x) = 2\left(z^2-4x\right)^{-3/2} \qquad f_z^{(2)}(0) = \frac{2}{z^3}$$

$$f_z^{(3)}(x) = 2\frac{3}{2}\cdot 4\left(z^2-4x\right)^{-5/2} = 12\left(z^2-4x\right)^{-5/2} \qquad f_z^{(3)}(0) = \frac{12}{z^5}$$

And notice that generally $\forall z^2 \gg x > 0$ we have $f_z^{(n)}(x) > 0$.

Then, by Lagrange's form of the remainder, the error of the quadratic approximation (around $x=0$) is given by

$$\begin{aligned} f_z(x) &= \frac{f(0)}{0!}x^0+\frac{f^{(1)}(0)}{1!}x^1+\frac{f^{(2)}(0)}{2!}x^2+R_2(x) \\ &= 0+\frac{x}{z}+\frac{2x^2}{2z^3}+R_2(x) = \frac{x}{z}+\frac{x^2}{z^3}+R_2(x), \end{aligned}$$

where

$$R_2(x) = \frac{f^{(3)}(x_0)}{3!}(x-0)^3 = \frac{12\left(z^2-4x_0\right)^{-5/2}}{6}x^3 \in \left[\frac{2x^3}{z^5},\frac{2x^3}{\left(z^2-4x\right)^{5/2}}\right].$$

since $x_0 \in [0,x]$.

We get that

$$x_{k-1}-x_k = f_{x_{k-1}}(\beta_k) \in \left[\frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3}+\frac{2\beta_k^3}{x_{k-1}^5},\ \frac{\beta_k}{x_{k-1}}+\frac{\beta_k^2}{x_{k-1}^3}+\frac{2\beta_k^3}{\left(x_{k-1}^2-4\beta_k\right)^{5/2}}\right].$$

Finally, since $\beta_k \le \frac{c^{2k-5}}{d}$ and $x_{k-1} \in [0.45, 1]$ we have

$$\begin{aligned}
\frac{2\beta_k^3}{\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}} &\le \frac{2\left(\frac{c^{2k-5}}{d}\right)^3}{\left(0.45^2 - 4\frac{c^{2k-5}}{d}\right)^{5/2}} \le \frac{2c^{6k-15}}{d^3\left(0.45^2 - 4\frac{1}{cd}\right)^{5/2}} \\
&\le \frac{2c^{6k-15}}{d^{1/2}\left(0.2d - 4\cdot 2^{1/d}\right)^{5/2}} \\
\left[d \ge 10{,}000 \Rightarrow 4\cdot 2^{1/d} \le 0.00041d\right] &\le \frac{2c^{6k-15}}{d^{1/2}\left(0.2d - 0.00041d\right)^{5/2}} \le \frac{113c^{6k-15}}{d^3}
\end{aligned}$$

□

Proposition G.12. *For any $k \ge 2$ (and $d \ge 25{,}000$), it holds that,*

$$\frac{x_k}{x_{k-1}} \in \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1 + \frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4},\ 1 - \frac{\beta_k}{x_{k-1}^2}\right).$$

Proof. We employ the bounds we found for $x_{k-1} - x_k$:

$$\begin{aligned}
\frac{x_k}{x_{k-1}} &= \frac{1}{x_{k-1}}\left(x_k - x_{k-1} + x_{k-1}\right) = 1 - \frac{1}{x_{k-1}}\left(x_{k-1} - x_k\right) \\
&\in \left[1 - \frac{\beta_k}{x_{k-1}^2} - \frac{\beta_k^2}{x_{k-1}^4} - \frac{2\beta_k^3}{x_{k-1}\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}},\ 1 - \frac{\beta_k}{x_{k-1}^2} - \frac{\beta_k^2}{x_{k-1}^4} - \frac{2\beta_k^3}{x_{k-1}^6}\right] \\
&\subset \left[1 - \frac{\beta_k}{x_{k-1}^2} - \frac{\beta_k^2}{x_{k-1}^4} - \frac{2\beta_k^3}{x_{k-1}\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}},\ 1 - \frac{\beta_k}{x_{k-1}^2}\right].
\end{aligned}$$

Notice that from the bounds on β_k, x_{k-1}, we have:

$$\begin{aligned}
\frac{2\beta_k^3}{x_{k-1}\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}} &= \frac{2\beta_k^2\beta_k x_{k-1}^3}{x_{k-1}^4\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}} \le \frac{2\beta_k^2\frac{c^{2k-5}}{d}1^3}{x_{k-1}^4\left(0.45^2 - 4\frac{c^{2k-5}}{d}\right)^{5/2}} \\
&\le \frac{2\beta_k^2\frac{1}{d}}{x_{k-1}^4\left(0.45^2 - 4\frac{1}{cd}\right)^{5/2}} = \frac{2\beta_k^2}{d\cdot 0.45^2 x_{k-1}^4\left(1 - \frac{4}{0.45^2}\frac{1}{2^{-1/d}d}\right)^{5/2}} \\
[d \ge 10{,}000] &\le \frac{2\beta_k^3}{d\cdot 0.45^2 x_{k-1}^4\left(1 - \frac{4}{0.45^2}\frac{1}{2^{-1/10000}\cdot 10000}\right)^{5/2}} \le \frac{\beta_k^3}{x_{k-1}^4}\cdot\frac{10}{d}
\end{aligned}$$

Since $d \ge 10{,}000$, we obtain $\frac{2\beta_k^3}{x_{k-1}\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}} \le \frac{10}{d}\frac{\beta_k^2}{x_{k-1}^4}$. Overall, we get

$$\frac{x_k}{x_{k-1}} \in \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1 + \frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4},\ 1 - \frac{\beta_k}{x_{k-1}^2}\right).$$

□

Proposition G.13. *For $k \geq 2$ (and $d \geq 25{,}000$), it holds that,*

$$(x_{k-1} - x_k)^2 = \left(\frac{x_{k-1} - \sqrt{x_{k-1}^2 - 4\beta_k}}{2} \right)^2 \in \left[\frac{\beta_k^2}{x_{k-1}^2},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{113c^{6k-15}}{d^3} \right].$$

Proof. We exploit the Taylor expansion of the following function (for $z^2 \gg x > 0$),

$$f(x) = \left(\frac{z - \sqrt{z^2 - 4x}}{2} \right)^2$$
$$f^{(1)}(x) = \frac{z}{\sqrt{z^2 - 4x}} - 1,\ f^{(2)}(x) = \frac{2z}{(z^2 - 4x)^{3/2}},\ f^{(3)}(x) = \frac{12z}{(z^2 - 4x)^{5/2}}$$

Then, by Lagrange's form of the remainder, the error of the quadratic approximation (around $x = 0$) is given by

$$\begin{aligned} f(x) &= \frac{f(0)}{0!}x^0 + \frac{f^{(1)}(0)}{1!}x^1 + \frac{f^{(2)}(0)}{2!}x^2 + R_2(x) \\ &= 0 + 0 \cdot x^1 + \frac{2}{2z^2}x^2 + R_2(x) = \frac{x^2}{z^2} + R_2(x)\ , \end{aligned}$$

where

$$R_2(x) = \frac{f^{(3)}(x_0)}{3!}(x-0)^3 = \frac{12z}{6(z^2 - 4x_0)^{5/2}}x^3 = \frac{2z}{(z^2 - 4x_0)^{5/2}}x^3 \in \left[\frac{2x^3}{z^4}, \frac{2z \cdot x^3}{(z^2 - 4x)^{5/2}} \right],$$

since $x_0 \in [0, x]$.

Then, setting $z = x_{k-1} \in [0.45, 1]$, we can now conclude that,

$$\begin{aligned} (x_{k-1} - x_k)^2 &= \left(\frac{x_{k-1} - \sqrt{x_{k-1}^2 - 4\beta_k}}{2} \right)^2 \triangleq f(\beta_k) \\ &\in \left[\frac{\beta_k^2}{x_{k-1}^2} + \frac{2\beta_k^3}{x_{k-1}^4},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{2\beta_k^3 x_{k-1}}{\left(x_{k-1}^2 - 4\beta_k\right)^{5/2}} \right]. \end{aligned}$$

Finally, since $\beta_k \leq \frac{c^{2k-5}}{d}$ and $x_{k-1} \in [0.45, 1]$ we have

$$\begin{aligned} (x_{k-1} - x_k)^2 &\in \left[\frac{\beta_k^2}{x_{k-1}^2} + \frac{2c^{6k-15}}{x_{k-1}^4 d^3},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{2c^{6k-15}x_{k-1}}{d^3\left(x_{k-1}^2 - 4\frac{c^{2k-5}}{d}\right)^{5/2}} \right] \\ &\subseteq \left[\frac{\beta_k^2}{x_{k-1}^2},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{2 \cdot c^{6k-15}}{d^3\left(0.45^2 - 4\frac{1}{cd}\right)^{5/2}} \right] \\ &\subseteq \left[\frac{\beta_k^2}{x_{k-1}^2},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{2c^{6k-15}}{d^{1/2}\left(0.2d - 4 \cdot 2^{1/d}\right)^{5/2}} \right] \\ \left[d \geq 10{,}000 \Rightarrow 4 \cdot 2^{1/d} \leq 0.00041d\right] &\subseteq \left[\frac{\beta_k^2}{x_{k-1}^2},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{2c^{6k-15}}{d^{1/2}\left(0.2d - 0.00041d\right)^{5/2}} \right] \\ &\subseteq \left[\frac{\beta_k^2}{x_{k-1}^2},\ \frac{\beta_k^2}{x_{k-1}^2} + \frac{113c^{6k-15}}{d^3} \right]. \end{aligned}$$

□

G.3.2 Expanding the inner product $(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1}$

Proposition G.14. *Let $t \in [d]$ and $t < k \leq d$ (and $d \geq 25{,}000$). Then,*

$$(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} \in \left[\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c)\,,\right.$$
$$\left.\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c) + \frac{113c^{6k-15}}{d^3}\right]$$

Proof. We use the expanded form of the inner product, that is,

$$0 < (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} = (x_{k-1} - x_k)\, x_{t-1} + (t-2)\,\frac{c^{k-3}(1-c)}{\sqrt{d}}\frac{c^{t-3}}{\sqrt{d}}$$
$$= (x_{k-1} - x_k)\, x_{t-1} + \frac{t-2}{d}c^{k+t-6}(1-c)\,.$$

Since we already showed $x_{k-1} - x_k \in \left[\frac{\beta_k}{x_{k-1}} + \frac{\beta_k^2}{x_{k-1}^3},\ \frac{\beta_k}{x_{k-1}} + \frac{\beta_k^2}{x_{k-1}^3} + \frac{113c^{6k-15}}{d^3}\right]$, we now have,

$$(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} \in \left[x_{t-1}\left(\frac{\beta_k}{x_{k-1}} + \frac{\beta_k^2}{x_{k-1}^3}\right) + \frac{t-2}{d}c^{k+t-6}(1-c)\,,\right.$$
$$\left. x_{t-1}\left(\frac{\beta_k}{x_{k-1}} + \frac{\beta_k^2}{x_{k-1}^3} + \frac{113c^{6k-15}}{d^3}\right) + \frac{t-2}{d}c^{k+t-6}(1-c)\right]$$
$$= \left[\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c)\,,\right.$$
$$\left.\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c) + \frac{113c^{6k-15}x_{t-1}}{d^3}\right]$$
$$[x_{t-1} \leq 1] \subseteq \left[\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c)\,,\right.$$
$$\left.\frac{x_{t-1}}{x_{k-1}}\beta_k + \frac{x_{t-1}}{x_{k-1}^3}\beta_k^2 + \frac{t-2}{d}c^{k+t-6}(1-c) + \frac{113c^{6k-15}}{d^3}\right]$$

□

G.3.3 Bounding $h(k) \triangleq \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d} c^{2k-6} (1-c)^2 + \frac{1}{d} c^{2k-6}}$

Proposition G.15. *For any $k \geq 2$ (when $d \geq 25{,}000$), $h(k) \in \left[\frac{c^{k-3}}{\sqrt{d}}, \frac{c^{k-3}}{\sqrt{d}} + \frac{5.42}{d^{3/2}}\right]$.*

Proof. The lower bound is easy to obtain:

$$h(k) = \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d} c^{2k-6} (1-c)^2 + \frac{1}{d} c^{2k-6}} > \sqrt{\frac{1}{d} c^{2k-6}} \geq \frac{c^{k-3}}{\sqrt{d}} .$$

To get the upper bound, we employ the inequality $(1-c) \triangleq 1 - 2^{-1/d} \leq \frac{\ln(2)}{d}$, and get,

$$\begin{aligned}
h(k) &= \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d} c^{2k-6} (1-c)^2 + \frac{c^{2k-6}}{d}} \\
&\leq \sqrt{\frac{\beta_k^2}{(0.45)^2} + (1-c)^2 + \frac{c^{2k-6}}{d}} \\
&\leq \sqrt{\frac{\beta_k^2}{(0.45)^2} + \frac{\ln^2(2)}{d^2} + \frac{c^{2k-6}}{d}} \\
\left[\beta_k \leq \frac{1}{cd}\right] &\leq \sqrt{\frac{1}{(0.45)^2 c^2 d^2} + \frac{\ln^2(2)}{d^2} + \frac{c^{2k-6}}{d}} \\
\left[d \geq 10{,}000 \rightarrow c^2 \geq 0.99986\right] &\leq \sqrt{\frac{c^{2k-6}}{d} + \frac{5.42}{d^2}} \\
[\text{G.8}] &< \sqrt{\frac{c^{2k-6}}{d}} + \frac{1}{2\sqrt{\frac{c^{2k-6}}{d}}} \cdot \frac{5.42}{d^2} = \frac{c^{k-3}}{\sqrt{d}} + \frac{1}{2c^{k-3}} \cdot \frac{5.42}{d^{3/2}} \\
\left[c^{k-m} \geq 2^{-1}\right] &\leq \frac{c^{k-3}}{\sqrt{d}} + \frac{5.42}{d^{3/2}} .
\end{aligned}$$

□

G.3.4 Bounding $\frac{h(k+1)}{h(k)}$

Proposition G.16. *For any $k \geq 2$ (when $d \geq 500$),*

$$\frac{h\left(k+1\right)}{h\left(k\right)} \in \left[c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2},\, c + \frac{2.44}{x_k^4 c^{2k-3} d^2}\right].$$

Proof. We start by expanding the expression in a way that will be useful for both the upper and the lower bounds,

$$\begin{aligned}
\frac{h\left(k+1\right)}{h\left(k\right)} &= \sqrt{\frac{\frac{\beta_{k+1}^2}{x_k^2} + \frac{k-1}{d}c^{2k-4}\left(1-c\right)^2 + \frac{1}{d}c^{2k-4}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}} \\
&= \sqrt{\frac{\frac{\beta_{k+1}^2}{x_k^2} + \boldsymbol{\frac{k-2}{d}}c^{2k-4}\left(1-c\right)^2 + \frac{1}{d}c^{2k-4}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}} + \frac{\boldsymbol{\frac{1}{d}}c^{2k-4}\left(1-c\right)^2}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}} \\
&= \sqrt{c^2 + \frac{\frac{\beta_{k+1}^2}{x_k^2} - \frac{c^2\beta_k^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}} + \frac{\frac{1}{d}c^{2k-4}\left(1-c\right)^2}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}}.
\end{aligned}$$

For the upper bound. We show that,

$$\begin{aligned}
&\frac{h\left(k+1\right)}{h\left(k\right)} \\
&= \sqrt{c^2 + \frac{\frac{\beta_{k+1}^2}{x_k^2} - \frac{c^2\beta_k^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}} + \frac{\frac{1}{d}c^{2k-4}\left(1-c\right)^2}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}} \\
&\overset{(1)}{\leq} \sqrt{c^2 + \boldsymbol{\beta_k^2}\frac{\frac{1}{x_k^2} - \frac{c^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}} + \frac{\frac{1}{d}\left(1-c\right)^2}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}} \\
&\overset{(2)}{\leq} \sqrt{c^2 + \frac{\mathbf{1}}{\boldsymbol{c^2d^2}}\frac{\frac{1}{x_k^2} - \frac{c^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}} + \frac{\frac{1}{d}\left(\boldsymbol{\frac{\ln(2)}{d}}\right)^2}{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}\left(1-c\right)^2 + \frac{1}{d}c^{2k-6}}} \\
&\leq \sqrt{c^2 + \frac{1}{c^2d^2}\frac{\frac{1}{x_k^2} - \frac{c^2}{x_{k-1}^2}}{\frac{1}{d}c^{2k-6}} + \frac{\frac{\ln^2(2)}{d^3}}{\frac{1}{d}c^{2k-6}}} = \sqrt{c^2 + \frac{x_{k-1}^2 - c^2x_k^2}{x_{k-1}^2x_k^2c^{2k-4}d} + \frac{\ln^2\left(2\right)}{c^{2k-6}d^2}} \\
&\overset{(3)}{\leq} \sqrt{c^2 + \frac{x_{k-1}^2 - c^2x_k^2}{x_k^4c^{2k-4}d} + \frac{\ln^2\left(2\right)}{c^{2k-6}d^2}} = \sqrt{c^2 + \frac{x_{k-1}^2 - c^2x_{k-1}^2}{x_k^4c^{2k-4}d} + \frac{c^2x_{k-1}^2 - c^2x_k^2}{x_k^4c^{2k-4}d} + \frac{\ln^2\left(2\right)}{c^{2k-6}d^2}} \\
&\overset{(4)}{\leq} \sqrt{c^2 + \left(1-c^2\right)\frac{1}{x_k^4c^{2k-4}d} + c^2\frac{x_{k-1}^2 - x_k^2}{x_k^4c^{2k-4}d} + \frac{\ln^2\left(2\right)}{c^{2k-6}d^2}},
\end{aligned}$$

where (1) is since $\beta_{k+1} < \beta_k,\ c < 1$; (2) is since $\beta_k < \frac{1}{cd},\ 1 - c \leq \frac{\ln 2}{d}$; (3) is since $x_k \leq x_{k-1}$; and (4) is since $x_{k-1} \leq 1$. To upper bound $x_{k-1}^2 - x_k^2$ we use the recursive formula of x_k, showing

that

$$
\begin{aligned}
x_{k-1}^2 - x_k^2 &= x_{k-1}^2 - \frac{x_{k-1}^2 + 2x_{k-1}\sqrt{x_{k-1}^2 - 4\beta_k} + x_{k-1}^2 - 4\beta_k}{4} \\
&= x_{k-1}^2 - \frac{x_{k-1}^2 + x_{k-1}\sqrt{x_{k-1}^2 - 4\beta_k} - 2\beta_k}{2} \\
&= \frac{x_{k-1}^2}{2} - \frac{x_{k-1}\sqrt{x_{k-1}^2 - 4\beta_k} - 2\beta_k}{2} = \frac{x_{k-1}^2}{2} - \frac{x_{k-1}^2\sqrt{1 - 4\frac{\beta_k}{x_{k-1}^2}} - 2\beta_k}{2} \\
\left[1 - z \le \sqrt{1-z}\right] &\le \frac{x_{k-1}^2}{2} - \frac{x_{k-1}^2\left(1 - 4\frac{\beta_k}{x_{k-1}^2}\right) - 2\beta_k}{2} = \frac{4\beta_k + 2\beta_k}{2} = 3\beta_k \,.
\end{aligned}
$$

Back to our expression,

$$
\begin{aligned}
\frac{h\left(k+1\right)}{h\left(k\right)} &\le \sqrt{c^2 + \left(1-c^2\right)\frac{1}{x_k^4 c^{2k-4} d} + \frac{3\beta_k}{x_k^4 c^{2k-6} d} + \frac{\ln^2\left(2\right)}{c^{2k-6} d^2}} \\
\left[\beta_k \le \frac{1}{cd}\right] &\le \sqrt{c^2 + \left(1-c^2\right)\frac{1}{x_k^4 c^{2k-4} d} + \frac{3}{x_k^4 c^{2k-5} d^2} + \frac{\ln^2\left(2\right)}{c^{2k-6} d^2}} \\
\left[1 - c^2 \le \frac{\ln\left(4\right)}{d}\right] &\le \sqrt{c^2 + \frac{\ln\left(4\right)}{x_k^4 c^{2k-4} d^2} + \frac{3}{x_k^4 c^{2k-5} d^2} + \frac{\ln^2\left(2\right)}{c^{2k-6} d^2}} \\
\left[c < 1\right] &\le \sqrt{c^2 + \frac{\ln\left(4\right)}{x_k^4 c^{2k-4} d^2} + \frac{3}{x_k^4 \boldsymbol{c^{2k-4}} d^2} + \frac{\ln^2\left(2\right)}{\boldsymbol{c^{2k-4}} d^2}} \le \sqrt{c^2 + \frac{4.39 + 0.49 x_k^4}{x_k^4 c^{2k-6} d^2}} \\
\left[x_k \le 1\right] &\le \sqrt{c^2 + \frac{4.88}{x_k^4 c^{2k-4} d^2}} = c\sqrt{1 + \frac{4.88}{x_k^4 c^{2k-2} d^2}} \le c + \frac{2.44}{x_k^4 c^{2k-3} d^2} \,,
\end{aligned}
$$

where in the last inequality we used the fact that $\forall z > 0$, $\sqrt{1+z} \le 1 + \frac{z}{2}$ (since $\left(1 + \frac{z}{2}\right)^2 = 1 + z + \frac{z^2}{4} \ge 1 + z = \left(\sqrt{1+z}\right)^2$).

878 **For the lower bound.** We show that, □

$$
\begin{aligned}
&\frac{h\left(k+1\right)}{h\left(k\right)}\\
&=\sqrt{c^2+\frac{\frac{\beta_{k+1}^2}{x_k^2}-\frac{c^2\beta_k^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2}+\frac{k-2}{d}c^{2k-6}\left(1-c\right)^2+\frac{1}{d}c^{2k-6}}+\frac{\frac{1}{d}c^{2k-4}\left(1-c\right)^2}{\frac{\beta_k^2}{x_{k-1}^2}+\frac{k-2}{d}c^{2k-6}\left(1-c\right)^2+\frac{1}{d}c^{2k-6}}}\\
&\geq\sqrt{c^2+\frac{\frac{\beta_{k+1}^2}{x_k^2}-\frac{c^2\beta_k^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2}+\frac{k-2}{d}c^{2k-6}\left(1-c\right)^2+\frac{1}{d}c^{2k-6}}}\\
&\geq c\sqrt{1+\frac{\frac{\beta_{k+1}^2}{x_k^2}-\frac{\beta_k^2}{x_{k-1}^2}}{\frac{\beta_k^2}{x_{k-1}^2}+\frac{k-2}{d}c^{2k-6}\left(1-c\right)^2+\frac{1}{d}c^{2k-6}}}\\
&\overset{(1)}{\geq} c\sqrt{1+\frac{1}{x_{k-1}^2}\frac{\beta_{k+1}^2-\beta_k^2}{\frac{\beta_k^2}{x_{k-1}^2}+\frac{k-2}{d}c^{2k-6}\left(1-c\right)^2+\frac{1}{d}c^{2k-6}}},
\end{aligned}
$$

879 where (1) is since $x_k \leq x_{k-1}$. Since $(\beta_k)_k$ is positive and decreasing, $\beta_{k+1}^2-\beta_k^2<0$, and so we
880 can simplify the expression using the fact that $\sqrt{1-z}\geq 1-z$, $\forall z\in(0,1)$:

$$
\frac{h\left(k+1\right)}{h\left(k\right)}\geq c\sqrt{1-\frac{\left|\beta_k^2-\beta_{k+1}^2\right|}{\frac{1}{d}c^{2k-6}x_{k-1}^2}}\geq c-\frac{\left|\beta_k^2-\beta_{k+1}^2\right|}{\frac{1}{d}c^{2k-5}x_{k-1}^2}.
$$

881 Focusing on $\frac{\left|\beta_{k+1}^2-\beta_k^2\right|}{\frac{1}{d}c^{2k-5}}$, and since $1-\left(1-c\right)\left(k-1\right)=\left(\left(k-1\right)c-\left(k-2\right)\right)$,

$$
\begin{aligned}
&\frac{\left|\beta_{k+1}^2-\beta_k^2\right|}{\frac{1}{d}c^{2k-5}}\\
&=\frac{\beta_k^2-\beta_{k+1}^2}{\frac{1}{d}c^{2k-5}}=\frac{\left(\frac{\left(\left(k-1\right)c-\left(k-2\right)\right)c^{2k-5}}{d}\right)^2-\left(\frac{\left(kc-\left(k-1\right)\right)c^{2k-3}}{d}\right)^2}{\frac{1}{d}c^{2k-5}}\\
&=\frac{c^{2k-5}}{d}\left(\left(1-\left(1-c\right)\left(k-1\right)\right)^2-\left(1-\left(1-c\right)k\right)^2c^4\right)\\
&=\frac{c^{2k-5}}{d}\left(\left(1-c^4\right)-2k\underbrace{\left(1-c\right)\left(1-c^4\right)}_{\geq 0}+k^2\left(1-c\right)^2\left(1-c^4\right)+\right.\\
&\qquad\left.+2\left(1-c\right)+\underbrace{\left(1-c\right)^2}_{\geq 0}\left(-2k+1\right)\right)\\
&\leq\frac{c^{2k-5}}{d}\left(\left(1-c^4\right)+k^2\left(1-c\right)^2\left(1-c^4\right)+2\left(1-c\right)+\left(1-c\right)^2\right)
\end{aligned}
$$

Using the previously derived bounds of $1-c \in \left[\frac{\ln(2)}{d} - \frac{\ln^2(2)}{2d^2}, \frac{\ln(2)}{d}\right]$, we can get,

$$\begin{aligned}
\frac{\left|\beta_{k+1}^2 - \beta_k^2\right|}{\frac{1}{d}c^{2k-5}} &\leq \frac{c^{2k-5}}{d}\left(\left(1-c^4\right) + k^2\frac{\ln^2(2)}{d^2}\left(1-c^4\right) + 2\frac{\ln(2)}{d} + \frac{\ln^2(2)}{d^2}\right) \\
\left[d \geq 500 \geq 1000\ln^2(2)\right] &\leq \frac{c^{2k-5}}{d}\left(\left(1-c^4\right) + k^2\frac{\ln^2(2)}{d^2}\left(1-c^4\right) + 2\frac{\ln(2)}{d} + \frac{0.001}{d}\right) \\
[k \leq d] &\leq \frac{c^{2k-5}}{d}\left(\left(1-c^4\right) + \ln^2(2)\left(1-c^4\right) + \frac{\ln(4)+0.001}{d}\right) \\
&\leq \frac{c^{2k-5}}{d}\left(\left(1+\ln^2(2)\right)\left(1-c^4\right) + \frac{\ln(4)+0.001}{d}\right).
\end{aligned}$$

Notice that we can use the previously derived bound of $1-c^n \leq \frac{n\ln(2)}{d}$, thus obtaining

$$\frac{\left|\beta_{k+1}^2 - \beta_k^2\right|}{\frac{1}{d}c^{2k-5}} \leq \frac{c^{2k-5}}{d}\left(\left(1+\ln^2(2)\right)\frac{4\ln(2)}{d} + \frac{\ln(4)+0.001}{d}\right) \leq 5.5\frac{c^{2k-5}}{d^2}.$$

Finally, we get,

$$\frac{h(k+1)}{h(k)} \geq c - \frac{\left|\beta_k^2 - \beta_{k+1}^2\right|}{\frac{1}{d}c^{2k-5}x_{k-1}^2} \geq c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}.$$

G.3.5 Expanding the norm

Proposition G.17. *For any $k \geq 2$, $\|\mathbf{w}_{k-1} - \mathbf{w}_k\| \in \left[h(k),\, h(k) + \frac{56.5c^{6k-15}}{c^{k-3}d^{5/2}}\right]$, where $h(k) \triangleq \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}(1-c)^2 + \frac{1}{d}c^{2k-6}}$.*

Proof. By construction we have

$$\begin{aligned}\|\mathbf{w}_{k-1} - \mathbf{w}_k\| &= \sqrt{(x_{k-1} - x_k)^2 + (k-2)\left(\frac{c^{k-3}(1-c)}{\sqrt{d}}\right)^2 + \left(\frac{c^{k-3}}{\sqrt{d}}\right)^2} \\ &= \sqrt{(x_{k-1} - x_k)^2 + \frac{k-2}{d}c^{2k-6}(1-c)^2 + \frac{1}{d}c^{2k-6}}\,.\end{aligned}$$

Before, we proved that $(x_{k-1} - x_k)^2 \in \left[\frac{\beta_k^2}{x_{k-1}^2},\, \frac{\beta_k^2}{x_{k-1}^2} + \frac{113c^{6k-15}}{d^3}\right]$. Now, we show the resulting bounds for $\|\mathbf{w}_{k-1} - \mathbf{w}_k\|$ which employ that bound.

The lower bound is immediate, since

$$\begin{aligned}\|\mathbf{w}_{k-1} - \mathbf{w}_k\| &= \sqrt{(x_{k-1} - x_k)^2 + \frac{k-2}{d}c^{2k-6}(1-c)^2 + \frac{1}{d}c^{2k-6}} \\ &\geq \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}(1-c)^2 + \frac{1}{d}c^{2k-6}} \triangleq h(k)\,.\end{aligned}$$

The upper bound requires an additional algebraic inequality of $\forall a, b > 0:\ \sqrt{a+b} < \sqrt{a} + \frac{b}{2\sqrt{a}}$ and the inequality of $h(k) \geq \frac{1}{\sqrt{d}}c^{k-3}$, i.e.,

$$\begin{aligned}\|\mathbf{w}_{k-1} - \mathbf{w}_k\| &\leq \sqrt{\frac{\beta_k^2}{x_{k-1}^2} + \frac{k-2}{d}c^{2k-6}(1-c)^2 + \frac{1}{d}c^{2k-6} + \frac{113c^{6k-15}}{d^3}} \\ &= \sqrt{h^2(k) + \frac{113c^{6k-15}}{d^3}} \leq h(k) + \frac{113c^{6k-15}}{2h(k)\,d^3} \leq h(k) + \frac{56.5c^{6k-15}}{c^{k-3}d^{5/2}}\,.\end{aligned}$$

□

G.3.6 Combining the expansions

Proposition G.18. *When $d \geq 25{,}000$,*

$$\begin{aligned}
&\|\mathbf{w}_k - \mathbf{w}_{k+1}\| \left(\mathbf{w}_{k-1} - \mathbf{w}_k\right)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\mathbf{w}_k - \mathbf{w}_{k+1}\right)^\top \mathbf{w}_{t-1} \\
&\geq A_1(k) + A_2(k) + A_3(k) \\
&\quad - \frac{1}{d^{7/2}}\left(\frac{96c^{6k-15}}{x_k} + 113c^{7k-18}\right) - \frac{1}{d^{9/2}}\left(\frac{56.5c^{9k-18}}{x_k^3} + 614c^{6k-30}\right),
\end{aligned}$$

where $A_1(k) \triangleq x_{t-1}\left(\frac{h(k+1)}{x_{k-1}}\beta_k - \frac{h(k)}{x_k}\beta_{k+1}\right)$, $A_2(k) \triangleq x_{t-1}\left(\frac{h(k+1)}{x_{k-1}^3}\beta_k^2 - \frac{h(k)}{x_k^3}\beta_{k+1}^2\right)$, $A_3(k) \triangleq \frac{t-2}{d}(1-c)\,c^{k+t-6}\left(h(k+1) - ch(k)\right)$.

Proof. Keeping in mind that we wish to bound $\|\mathbf{w}_k - \mathbf{w}_{k+1}\| \left(\mathbf{w}_{k-1} - \mathbf{w}_k\right)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\mathbf{w}_k - \mathbf{w}_{k+1}\right)^\top \mathbf{w}_{t-1}$, we start lower bounding the right expression. Using the bounds for $\left(\mathbf{w}_{k-1} - \mathbf{w}_k\right)^\top \mathbf{w}_{t-1}$ and $\|\mathbf{w}_{k-1} - \mathbf{w}_k\|$ we derived above, we get,

$$\begin{aligned}
&- \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\mathbf{w}_k - \mathbf{w}_{k+1}\right)^\top \mathbf{w}_{t-1} \\
&\geq - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{(t-2)\,c^{k+t-5}(1-c)}{d} + \frac{113c^{6k-15}}{d^3}\right) \\
&\geq - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{(t-2)\,c^{k+t-5}(1-c)}{d}\right) \\
&\quad - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \frac{113c^{6k-15}}{d^3} \\
&\geq \underbrace{- \left(h(k) + \frac{56.5c^{6k-15}}{c^{k-3}d^{5/2}}\right)\left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{t-2}{d}c^{k+t-5}(1-c)\right)}_{\triangleq a(k)} \\
&\quad \underbrace{- \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \frac{113c^{6k-15}}{d^3}}_{\triangleq b(k)}.
\end{aligned}$$

The right function is easily bounded as,

$$\begin{aligned}
b(k) &= - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| \frac{113c^{6k-15}}{d^3} \\
&\geq - \left(h(k) + \frac{56.5c^{6k-15}}{c^{k-3}d^{5/2}}\right)\frac{113c^{6k-15}}{d^3} = -\frac{113c^{6k-15}}{d^3}h(k) - \frac{6384.5c^{12k-30}}{d^{11/2}c^{k-3}} \\
&\geq -\frac{113c^{6k-15}}{d^3}\left(\frac{\sqrt{c^{2k-6}}}{\sqrt{d}} + \frac{5.42}{d^{3/2}}\right) - \frac{6384.5c^{12k-30}}{d^{11/2}c^{k-3}} \\
&= -\frac{113c^{6k-15}c^{k-3}}{d^{7/2}} - \frac{612.46c^{6k-15}}{d^{9/2}} - \frac{6384.5c^{12k-30}}{d^{11/2}c^{k-3}} \\
\left[c^{k-3} \geq c^d = 0.5\right] &\geq -\frac{113c^{7k-18}}{d^{7/2}} - \frac{612.46c^{6k-15}}{d^{9/2}} - \frac{12769c^{12k-30}}{d^{11/2}}
\end{aligned}$$

The left function is further decomposed as,

$$a(k) = \underbrace{-h(k)\left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{t-2}{d}c^{k+t-5}(1-c)\right)}_{\triangleq a_1(k)} \underbrace{-\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{t-2}{d}c^{k+t-5}(1-c)\right)}_{\triangleq a_2(k)}.$$

Then,

$$\begin{aligned}
a_2(k) &= -\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{x_{t-1}}{x_k}\beta_{k+1} + \frac{x_{t-1}}{x_k^3}\beta_{k+1}^2 + \frac{t-2}{d}c^{k+t-5}(1-c)\right) \\
\left[x_{t-1} \le 1, \frac{t-2}{d} < 1\right] &\ge -\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{1}{x_k}\beta_{k+1} + \frac{1}{x_k^3}\beta_{k+1}^2 + c^{k+t-5}(1-c)\right) \\
\left[\beta_{k+1} \le \frac{c^{2k-3}}{d}\right] &\ge -\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{c^{2k-3}}{x_k d} + \frac{c^{4k-6}}{x_k^3 d^2} + c^{k+t-5}(1-c)\right) \\
&\ge -\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{c^{2k-3}}{x_k d} + \frac{c^{4k-6}}{x_k^3 d^2} + \frac{\ln(2)}{d}c^{k+t-5}\right) \\
[x_k \le 1] &\ge -\frac{56.5c^{5k-12}}{d^{5/2}}\left(\frac{c^{2k-3} + \ln(2)c^{k+t-5}}{x_k d} + \frac{c^{4k-6}}{x_k^3 d^2}\right) \\
&= -\frac{56.5c^{6k-15}}{d^{5/2}\cdot x_k d}\left(c^k + \ln(2)c^{t-2} + \frac{c^{3k-3}}{x_k^2 d}\right) \\
[c < 1] &\ge -\frac{56.5c^{6k-15}}{d^{7/2}\cdot x_k}\left(1 + \ln(2) + \frac{c^{3k-3}}{x_k^2 d}\right) \\
&\ge -\frac{96c^{6k-15}}{d^{7/2}\cdot x_k} - \frac{56.5c^{9k-18}}{d^{9/2}\cdot x_k^3}.
\end{aligned}$$

Overall we got,

$$
\begin{aligned}
&\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}-\left\|\mathbf{w}_{k-1}-\mathbf{w}_k\right\|\left(\mathbf{w}_k-\mathbf{w}_{k+1}\right)^\top\mathbf{w}_{t-1},\\
&\geq\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}+a_1\left(k\right)+a_2\left(k\right)+b\left(k\right)\\
&\geq\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}+a_1\left(k\right)\\
&\quad-\frac{96c^{6k-15}}{d^{7/2}\cdot x_k}-\frac{56.5c^{9k-18}}{d^{9/2}\cdot x_k^3}-\frac{113c^{7k-18}}{d^{7/2}}-\frac{612.46c^{6k-15}}{d^{9/2}}-\frac{12769c^{12k-30}}{d^{11/2}}\\
&\geq\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}+a_1\left(k\right)\\
&\quad-\frac{96c^{6k-15}}{d^{7/2}\cdot x_k}-\frac{56.5c^{9k-18}}{d^{9/2}\cdot x_k^3}-\frac{113c^{7k-18}}{d^{7/2}}-\frac{612.46c^{6k-30}}{d^{9/2}}-\frac{12769c^{6k-30}}{d^{11/2}}\\
&\overset{(1)}{\geq}\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}\\
&\quad+a_1\left(k\right)-\frac{96c^{6k-15}}{d^{7/2}\cdot x_k}-\frac{56.5c^{9k-18}}{d^{9/2}\cdot x_k^3}-\frac{113c^{7k-18}}{d^{7/2}}-\frac{614c^{6k-30}}{d^{9/2}}\\
&\geq\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}\\
&\quad+a_1\left(k\right)-\frac{1}{d^{7/2}}\left(\frac{96c^{6k-15}}{x_k}+113c^{7k-18}\right)-\frac{1}{d^{9/2}}\left(\frac{56.5c^{9k-18}}{x_k^3}+614c^{6k-30}\right),
\end{aligned}
$$

where (1) is since $d\geq 10{,}000$. Focusing on the left terms, we get the **overall** expression, which we need to show is **positive**. We again use previously-derived inequalities, to show,

$$
\begin{aligned}
&\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}+a_1\left(k\right)\\
&=\left\|\mathbf{w}_k-\mathbf{w}_{k+1}\right\|\left(\mathbf{w}_{k-1}-\mathbf{w}_k\right)^\top\mathbf{w}_{t-1}\\
&\quad-h\left(k\right)\left(\frac{x_{t-1}}{x_k}\beta_{k+1}+\frac{x_{t-1}}{x_k^3}\beta_{k+1}^2+\frac{t-2}{d}c^{k+t-5}\left(1-c\right)\right)\\
&\geq h\left(k+1\right)\left(\frac{x_{t-1}}{x_{k-1}}\beta_k+\frac{x_{t-1}}{x_{k-1}^3}\beta_k^2+\frac{t-2}{d}c^{k+t-6}\left(1-c\right)\right)\\
&\quad-h\left(k\right)\left(\frac{x_{t-1}}{x_k}\beta_{k+1}+\frac{x_{t-1}}{x_k^3}\beta_{k+1}^2+\frac{t-2}{d}c^{k+t-5}\left(1-c\right)\right)\\
&=\underbrace{x_{t-1}\left(\frac{h\left(k+1\right)}{x_{k-1}}\beta_k-\frac{h\left(k\right)}{x_k}\beta_{k+1}\right)}_{\triangleq A_1(k)}+\underbrace{x_{t-1}\left(\frac{h\left(k+1\right)}{x_{k-1}^3}\beta_k^2-\frac{h\left(k\right)}{x_k^3}\beta_{k+1}^2\right)}_{\triangleq A_2(k)}\\
&\quad+\underbrace{\frac{t-2}{d}\left(1-c\right)c^{k+t-6}\left(h\left(k+1\right)-ch\left(k\right)\right)}_{\triangleq A_3(k)},
\end{aligned}
$$

which we will bound separately below. □

G.3.7 The second term, $A_2(k)$, is insignificant $\mathcal{O}\left(\frac{1}{d^{7/2}}\right)$

Proposition G.19. *When $d \geq 25{,}000$,*

$$A_2(k) = x_{t-1}\left(\frac{h(k+1)}{x_{k-1}^3}\beta_k^2 - \frac{h(k)}{x_k^3}\beta_{k+1}^2\right) \geq -\frac{14.88 x_{t-1} c^{3k-12}}{x_k^3 d^{7/2}}\,.$$

Proof. We start from,

$$A_2(k) = x_{t-1}\left(\frac{h(k+1)}{x_{k-1}^3}\beta_k^2 - \frac{h(k)}{x_k^3}\beta_{k+1}^2\right) = \frac{x_{t-1}}{x_k^3}h(k)\beta_{k+1}^2 \underbrace{\left(\frac{h(k+1)}{h(k)}\frac{x_k^3}{x_{k-1}^3}\frac{\beta_k^2}{\beta_{k+1}^2} - 1\right)}_{\triangleq a(k)}\,.$$

Dissecting the terms in $a(k)$,

$$\frac{\beta_k}{\beta_{k+1}} = \frac{\frac{((k-1)c-(k-2))c^{2k-5}}{d}}{\frac{(kc-(k-1))c^{2k-3}}{d}} = \frac{1}{c^2} + \underbrace{\frac{1-c}{(kc-(k-1))}}_{>0,\ \text{from G.4}}c^2 \geq \frac{1}{c^2}$$

$$\frac{\beta_k^2}{\beta_{k+1}^2} \geq \frac{1}{c^4}\,.$$

We already showed that $\frac{x_k}{x_{k-1}} \in \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4},\ 1 - \frac{\beta_k}{x_{k-1}^2}\right)$, and we simplify it even further

$$\frac{x_k}{x_{k-1}} \geq 1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4} \overset{\beta_k \leq \frac{1}{d}}{\geq} 1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k}{x_{k-1}^4 d}$$

$$[x_{k-1} \geq 0.45] \geq 1 - \beta_k\left(\frac{1}{(0.45)^2} + \frac{\left(1+\frac{10}{d}\right)}{(0.45)^4 d}\right) \overset{d \geq 10{,}000}{\geq} 1 - 4.95\beta_k\,.$$

Now, using the algebraic inequality that $\forall z \in (0,1)\,,\ (1-z)^3 = 1 - 3z + 3z^2 - z^3 > 1 - 3z$, we get,

$$\frac{x_k^3}{x_{k-1}^3} > 1 - 14.85\beta_k\,.$$

Moreover, recall that we already showed that $\frac{h(k+1)}{h(k)} \geq c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}$. Now, focusing on $a(k)$,

$$\begin{aligned}
a(k) &= \frac{h(k+1)}{h(k)} \cdot \frac{x_k^3}{x_{k-1}^3}\frac{\beta_k^2}{\beta_{k+1}^2} - 1 \geq \left(c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}\right)(1 - 14.85\beta_k)\frac{1}{c^4} - 1 \\
&= \left(\frac{1}{c^3} - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}\right)(1 - 14.85\beta_k) - 1 \geq \left(1 - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}\right)(1 - 14.85\beta_k) - 1 \\
&= -14.85\beta_k - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2} + \frac{81.675c^{2k-9}}{x_{k-1}^2 d^2}\beta_k \geq -14.85\beta_k - \frac{5.5c^{2k-9}}{x_{k-1}^2 d^2} \\
\left[\beta_k \leq \frac{c^{2k-5}}{d}\right] &\geq -\frac{14.85c^{2k-5}}{d} - \frac{5.5c^{2k-9}}{x_{k-1}^2 d^2} \geq -\frac{14.85c^{2k-5}}{d} - \frac{5.5c^{2k-9}}{(0.45)^2 d^2} \\
&\geq -\frac{14.85c^{2k-5}}{d} - \frac{27.17c^{2k-9}}{d^2} \geq -\frac{14.85c^{2k-9}}{d} - \frac{27.17c^{2k-9}}{d^2} \\
[d \geq 10{,}000] &\geq -\frac{14.86c^{2k-9}}{d}\,.
\end{aligned}$$

And finally,

$$\begin{aligned}
\frac{1}{x_{t-1}}A_2(k) &= \frac{1}{x_k^3}h(k)\beta_{k+1}^2 \cdot a(k) \geq -\frac{1}{x_k^3}h(k)\beta_{k+1}^2 \cdot \frac{14.86c^{2k-9}}{d} \\
\left[\beta_{k+\mathbf{1}} \leq \frac{1}{d}, h(k) \leq \frac{c^{k-3}}{\sqrt{d}} + \frac{5.42}{d^{3/2}}\right] &\geq -\frac{1}{x_k^3}\frac{\frac{c^{k-3}}{\sqrt{d}} + \frac{5.42}{d^{3/2}}}{d^2}\frac{14.86c^{2k-9}}{d} \\
&= -\frac{1}{x_k^3}\frac{c^{3k-12}}{d^{5/2}}\frac{14.86}{d} - \frac{1}{x_k^3}\frac{5.42}{d^{7/2}}\frac{14.86c^{2k-9}}{d} \\
[c<1] &\geq -\frac{14.86c^{3k-12}}{x_k^3 d^{7/2}} - \frac{80.55c^{2k-12}}{x_k^3 d^{9/2}} \\
\left[d \geq 10{,}000, c^{k-3} \geq c^d = 0.5\right] &\geq -\frac{14.86c^{3k-12}}{x_k^3 d^{7/2}} - \frac{0.0081 \cdot c^{3k-12}}{x_k^3 d^{7/2} c} \\
&\geq -\frac{\left(14.86 + \frac{0.0081}{0.5}\right)c^{3k-12}}{x_k^3 d^{7/2}} \\
A_2(k) &\geq -\frac{14.88x_{t-1}c^{3k-12}}{x_k^3 d^{7/2}},
\end{aligned}$$

thus concluding this part. □

G.3.8 The third term, $A_3(k)$, is insignificant $\mathcal{O}\left(\frac{1}{d^{7/2}}\right)$

Proposition G.20. *When $d \geq 25{,}000$,*

$$|A_3(k)| = \left|\frac{t-2}{d}(1-c)\,c^{k+t-6}\left(h(k+1)-ch(k)\right)\right| \leq \frac{6.77c^{k-6}}{x_k^4}\left(\frac{c^{k-3}}{d^{7/2}}+\frac{5.42}{d^{9/2}}\right).$$

Proof. Notice that,

$$\begin{aligned}
|A_3(k)| &= \left|\frac{t-2}{d}(1-c)\,c^{k+t-6}\left(h(k+1)-ch(k)\right)\right| \\
&= \frac{t-2}{d}(1-c)\,c^{k+t-6}\left|h(k+1)-ch(k)\right| \leq (1-c)\,c^{k+t-6}\left|h(k+1)-ch(k)\right| \\
&\leq \ln(2)\,c^{k+t-6}\frac{h(k)}{d}\left|\frac{h(k+1)}{h(k)}-c\right| \leq \ln(2)\,c^{k+t-6}\left(\frac{c^{k-3}}{d^{3/2}}+\frac{5.42}{d^{5/2}}\right)\left|\frac{h(k+1)}{h(k)}-c\right|,
\end{aligned}$$

where we used the facts that $1-c \leq \frac{\ln 2}{d}$ and $h(k) \leq \frac{c^{k-3}}{\sqrt{d}}+\frac{5.42}{d^{3/2}}$.

Using $\frac{h(k+1)}{h(k)} \in \left[c-\frac{5.5c^{2k-5}}{x_{k-1}^2 d^2},\, c+\frac{2.44}{x_k^4 c^{2k-3}d^2}\right]$, we finally get,

$$\begin{aligned}
|A_3(k)| &\leq \ln(2)\,c^{k+t-6}\left(\frac{c^{k-3}}{d^{3/2}}+\frac{5.42}{d^{5/2}}\right)\left|\frac{h(k+1)}{h(k)}-c\right| \\
&\leq \ln(2)\,c^{k+t-6}\left(\frac{c^{k-3}}{d^{3/2}}+\frac{5.42}{d^{5/2}}\right)\max\left(\frac{5.5c^{2k-5}}{x_{k-1}^2 d^2},\frac{2.44}{x_k^4 c^{2k-3}d^2}\right) \\
[x_k < x_{k-1}] &\leq \ln(2)\,c^{k+t-6}\left(\frac{c^{k-3}}{d^{7/2}}+\frac{5.42}{d^{9/2}}\right)\max\left(\frac{5.5c^{2k-5}}{x_k^4},\frac{2.44}{x_k^4 c^{2k-3}}\right) \\
[c<1] &\leq \frac{\ln(2)\,c^{k+t-6}}{x_k^4}\left(\frac{c^{k-3}}{d^{7/2}}+\frac{5.42}{d^{9/2}}\right)\max\left(5.5c^{2k-5},\frac{2.44}{c^{2k}}\right) \\
\left[c^{nk-m} \geq 2^{-n},\ k\geq 2,\ c<1\right] &\leq \frac{\ln(2)\,c^{k+t-6}}{x_k^4}\left(\frac{1}{cd^{7/2}}+\frac{5.42}{d^{9/2}}\right)\max\left(\frac{5.5}{c},9.76\right) \\
\left[c \geq 2^{-1/10000} \geq 0.9999\right] &\leq \frac{\ln(2)\,c^{k+t-6}}{x_k^4}\left(\frac{1}{0.9999d^{7/2}}+\frac{5.42}{d^{9/2}}\right)\max\left(\frac{5.5}{0.9999},9.76\right) \\
&\leq \frac{6.77c^{k+t-6}}{x_k^4}\left(\frac{1}{d^{7/2}}+\frac{5.42}{d^{9/2}}\right).
\end{aligned}$$

□

G.3.9 Back to the first term, $A_1(k)$

Proposition G.21. *When $d \geq 25{,}000$,*

$$A_1(k) = x_{t-1}\left(\frac{h(k+1)}{x_{k-1}}\beta_k - \frac{h(k)}{x_k}\beta_{k+1}\right) \geq x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}}\right)$$

Proof. We have

$$A_1(k) = x_{t-1}\left(\frac{h(k+1)}{x_{k-1}}\beta_k - \frac{h(k)}{x_k}\beta_{k+1}\right) = \underbrace{\frac{x_{t-1}}{x_k}h(k)\beta_{k+1}}_{=\Theta\left(d^{-3/2}\right)}\underbrace{\left(\frac{x_k}{x_{k-1}}\frac{h(k+1)}{h(k)}\frac{\beta_k}{\beta_{k+1}} - 1\right)}_{\triangleq a(k)}.$$

We are going to use the previously-derived lower bounds of $\frac{x_k}{x_{k-1}} \geq 1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4}$ and $\frac{h(k+1)}{h(k)} \geq c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}$. To lower bound $\frac{\beta_k}{\beta_{k+1}} = \frac{1}{c^2} + \frac{1-c}{(1-k(1-c))c^2}$, we need a slightly stronger bound than before. Specifically, notice that for any $z \in (0,1)$, $\frac{z}{1-z} \geq z$. Then, since $1-c \in \left[\frac{\ln(2)}{d} - \frac{\ln^2(2)}{2d^2}, \frac{\ln(2)}{d}\right] \Longrightarrow k(1-c) \in \left[\frac{k}{d}\ln(2) - \frac{k}{2d^2}\ln^2(2), \frac{k}{d}\ln(2)\right] \subseteq (0,1)$, and

$$\frac{1-c}{(1-k(1-c))c^2} = \frac{1}{c^2 k}\frac{\boldsymbol{k(1-c)}}{(1-\boldsymbol{k(1-c)})} \geq \frac{1}{c^2 k}k(1-c) = \frac{1-c}{c^2}.$$

We now get,

$$\frac{\beta_k}{\beta_{k+1}} \geq \frac{1}{c^2} + \frac{1-c}{c^2} = \frac{2-c}{c^2}.$$

We are now ready to lower bound $a(k)$ as,

$$\begin{aligned}
a(k) &= \frac{x_k}{x_{k-1}}\frac{h(k+1)}{h(k)}\frac{\beta_k}{\beta_{k+1}} - 1 \\
&\geq \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4}\right)\left(c - \frac{5.5c^{2k-5}}{x_{k-1}^2 d^2}\right)\left(\frac{2-c}{c^2}\right) - 1 \\
&= \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4}\right)\left(1 - \frac{5.5c^{2k-6}}{x_{k-1}^2 d^2}\right)\left(\frac{2-c}{c}\right) - 1
\end{aligned}$$

Using Claim G.9: $\frac{2-c}{c} = \frac{2}{c} - 1 = 2\cdot 2^{1/d} - 1 \geq 2\left(1+\frac{\ln(2)}{d}\right) - 1 = 1 + \frac{\ln(4)}{d}$, we get:

$$\begin{aligned}
a(k) &\geq \left(1 - \frac{\beta_k}{x_{k-1}^2} - \left(1+\frac{10}{d}\right)\frac{\beta_k^2}{x_{k-1}^4}\right)\left(1 - \frac{5.5c^{2k-6}}{x_{k-1}^2 d^2}\right)\left(1 + \frac{\ln(4)}{d}\right) - 1 \\
&= \underbrace{\frac{\ln(4)}{d} - \frac{\beta_k}{x_{k-1}^2}}_{\mathcal{O}(d^{-1})} - \underbrace{\left(\frac{5.5c^{2k-6}}{x_{k-1}^2 d^2} + \frac{\left(1+\frac{10}{d}\right)\beta_k^2}{x_{k-1}^4} + \frac{\ln(4)\beta_k}{x_{k-1}^2 d}\right)}_{\Theta(d^{-2})} \\
&\quad + \underbrace{\frac{5.5c^{2k-6}\beta_k}{x_{k-1}^4 d^2} - \frac{5.5\ln(4)c^{2k-6}}{x_{k-1}^2 d^3} - \frac{\left(1+\frac{10}{d}\right)\ln(4)\beta_k^2}{x_{k-1}^4 d}}_{\mathcal{O}(d^{-3})} \\
&\quad + \underbrace{\frac{5.5\left(1+\frac{10}{d}\right)c^{2k-6}\beta_k^2}{x_{k-1}^6 d^2} + \frac{5.5\ln(4)c^{2k-6}}{x_{k-1}^2 d^3}\left(\frac{\beta_k}{x_{k-1}^2} + \frac{\left(1+\frac{10}{d}\right)\beta_k^2}{x_{k-1}^4}\right)}_{\mathcal{O}(d^{-4})}.
\end{aligned}$$

Lower bounding negligible positive terms by 0, we get,

$$a(k) \geq \underbrace{\frac{\ln(4)}{d} - \frac{\beta_k}{x_{k-1}^2}}_{\mathcal{O}(d^{-1})} - \underbrace{\left(\frac{5.5c^{2k-6}}{x_{k-1}^2 d^2} + \frac{\left(1+\frac{10}{d}\right)\beta_k^2}{x_{k-1}^4} + \frac{\ln(4)\,\beta_k}{x_{k-1}^2 d}\right)}_{\Theta(d^{-2})}$$
$$- \underbrace{\left(\frac{5.5\ln(4)\,c^{2k-6}}{x_{k-1}^2 d^3} + \frac{\left(1+\frac{10}{d}\right)\ln(4)\,\beta_k^2}{x_{k-1}^4 d}\right)}_{\Theta(d^{-3})}.$$

We will now simplify the least significant terms above further. We start from an *upper* bound to the $\Theta\left(d^{-2}\right)$ term (since its sign is negative in the expression above),

$$\begin{aligned}
\frac{5.5c^{2k-6}}{x_{k-1}^2 d^2} + \frac{\left(1+\frac{10}{d}\right)\beta_k^2}{x_{k-1}^4} + \frac{\ln(4)\,\beta_k}{x_{k-1}^2 d} &\leq \frac{5.5c^{2k-6}}{x_{k-1}^2 d^2} + \frac{1.001\beta_k^2}{x_{k-1}^2 x_{k-1}^2} + \frac{\ln(4)\,\beta_k}{x_{k-1}^2 d} \\
\left[\beta_k \leq \frac{c^{2k-5}}{d} \leq \frac{1}{cd},\ x_{k-1} \geq 0.45\right] &\leq \frac{5.5c^{2k-6}}{x_{k-1}^2 d^2} + \frac{1.001c^{4k-10}}{x_{k-1}^2 d^2 0.45^2} + \frac{\ln(4)\,c^{2k-5}}{x_{k-1}^2 d^2} \\
&\leq \frac{c^{2k-6}}{x_{k-1}^2 d^2}\left(5.5 + 4.95c^{2k-4} + \ln 4 \cdot c\right) \\
[k \geq 2,\ c \leq 1] &\leq \frac{c^{2k-6}}{x_{k-1}^2 d^2}\left(5.5 + 4.95 + \ln 4\right) \\
&\leq \frac{11.84c^{2k-6}}{x_{k-1}^2 d^2}.
\end{aligned}$$

Similarly, for the $\Theta\left(d^{-3}\right)$ term, we again employ the upper bound $\beta_k \leq \frac{c^{2k-5}}{d} \leq \frac{1}{cd}$, and obtain,

$$\begin{aligned}
\frac{5.5\ln(4)\,c^{2k-6}}{x_{k-1}^2 d^3} + \frac{\left(1+\frac{10}{d}\right)\ln(4)\,\beta_k^2}{x_{k-1}^4 d} &\leq \frac{5.5\ln(4)\,c^{2k-6}}{x_{k-1}^2 d^3} + \frac{1.001\ln(4)\,c^{2k-5}}{x_{k-1}^4 d^3 c^3} \\
&\leq \frac{c^{2k-6}}{x_{k-1}^4 d^3}\left(5.5\ln(4) + 1.001\ln(4)\,c^{-2}\right) \\
&\leq \frac{c^{2k-8}}{x_{k-1}^4 d^3}\left(5.5\ln(4) + 1.001\ln(4)\right) \leq \frac{9.02c^{2k-8}}{x_{k-1}^4 d^3}.
\end{aligned}$$

And so, we get the following lower bound,

$$\begin{aligned}
a(k) &\geq \frac{\ln(4)}{d} - \frac{\beta_k}{x_{k-1}^2} - \left(\frac{11.84c^{2k-6}}{x_{k-1}^2 d^2} + \frac{9.02c^{2k-8}}{x_{k-1}^4 d^3}\right) \\
[x_k \leq x_{k-1}] &\geq \frac{\ln(4)}{d} - \frac{\beta_k}{x_k^2} - \frac{c^{2k-6}}{x_k^2 d^2}\left(11.84 + \frac{9.02c^{-2}}{x_k^2 d}\right).
\end{aligned}$$

Back to the overall term we are trying to lower bound,

$$\begin{aligned}
\frac{1}{x_{t-1}}A_1(k) &= \frac{h(k+1)}{x_{k-1}}\beta_k - \frac{h(k)}{x_k}\beta_{k+1} = \underbrace{\frac{1}{x_k}h(k)\beta_{k+1}}_{=\Theta\left(d^{-3/2}\right)}\underbrace{\left(\frac{x_k}{x_{k-1}}\frac{h(k+1)}{h(k)}\frac{\beta_k}{\beta_{k+1}}-1\right)}_{\triangleq a(k)} \\
&\geq \frac{1}{x_k}h(k)\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}-\frac{c^{2k-6}}{x_k^2d^2}\left(11.84+\frac{9.02c^{-2}}{x_k^2d}\right)\right) \\
&= \frac{1}{x_k}h(k)\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{1}{x_k}h(k)\beta_{k+1}\frac{c^{2k-6}}{x_k^2d^2}\left(11.84+\frac{9.02c^{-2}}{x_k^2d}\right) \\
&\overset{(1)}{\geq} \frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{1}{x_k}\frac{\frac{c^{k-3}}{\sqrt{d}}+\frac{5.42}{d^{3/2}}}{d}\frac{c^{2k-6}}{x_k^2d^2}\left(11.84+\frac{9.02c^{-2}}{x_k^2d}\right) \\
&\overset{(2)}{\geq} \frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{c^{3k-9}}{x_k^3d^{7/2}}\left(11.84+\frac{9.02\cdot 2^{2/10000}}{0.45^2d}\right) \\
&\quad -\frac{5.42c^{2k-6}}{x_k^3d^{9/2}}\left(11.84+\frac{9.02\cdot 2^{2/10000}}{0.45^2d}\right) \\
&\geq \frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{11.84c^{3k-9}}{x_k^3d^{7/2}}-\frac{44.55c^{3k-9}}{x_k^3d^{9/2}} \\
&\quad -\frac{64.18c^{2k-6}}{x_k^3d^{9/2}}-\frac{241.46c^{2k-6}}{x_k^3d^{11/2}} \\
&\geq \frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{11.84c^{3k-9}}{x_k^3d^{7/2}}-\frac{44.55c^{2k-9}}{x_k^3d^{9/2}} \\
&\quad -\frac{64.18c^{2k-9}}{x_k^3d^{9/2}}-\frac{241.46c^{2k-9}}{x_k^3d^{11/2}} \\
&\overset{(3)}{\geq} \frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{11.84c^{3k-9}}{x_k^3d^{7/2}}-\frac{109c^{2k-9}}{x_k^3d^{9/2}} \\
A_1(k) &\geq \frac{x_{t-1}}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)-\frac{11.84x_{t-1}c^{3k-9}}{x_k^3d^{7/2}}-\frac{109x_{t-1}c^{2k-9}}{x_k^3d^{9/2}},
\end{aligned}$$

where (1) is since $h(k)\in\left[\frac{c^{k-3}}{\sqrt{d}},\frac{c^{k-3}}{\sqrt{d}}+\frac{5.42}{d^{3/2}}\right]$, $\beta_{k+1}\leq\frac{1}{d}$; (2) is since $d\geq 10{,}000$ and $x_k\geq 0.45$; and (3) is since $d\geq 10{,}000$. It remains to get a lower bound for $\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)$. First, we show

$$\begin{aligned}
b(k)\triangleq\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2} &= \frac{\ln(4)}{d}-\frac{\left(1-(1-c)(k-1)\right)c^{2k-5}}{x_k^2d} \\
&= \frac{\ln(4)}{d}-\frac{c^{2k-5}}{x_k^2d}+\frac{1}{x_k^2c^3}\cdot(1-c)\frac{k-1}{d}c^{2(k-1)} \\
&= \frac{\ln(4)}{d}-\frac{c^{2k-5}}{x_k^2d}+\frac{1}{x_k^2c^3}\cdot(1-c)\left(\frac{k-1}{d}\right)4^{-\frac{k-1}{d}} \\
&\geq \frac{\ln(4)}{d}-\frac{c^{2k-5}}{x_k^2d}+\frac{1}{x_k^2c^3}\cdot\frac{1-c}{4}\left(\frac{k-1}{d}\right),
\end{aligned}$$

where we used an algebraic property that $4^{-z} \geq \frac{1}{4}, \forall z \in [0,1]$. Continuing,

$$\begin{aligned}
b(k) &\geq \frac{\ln(4)}{d} - \frac{c^{2k-5}}{x_k^2 d} + \frac{1}{4x_k^2 c^3} \cdot \left(\frac{\ln(2)}{d} - \frac{\ln^2(2)}{2d^2}\right)\left(\frac{k-1}{d}\right) \\
&= \frac{\ln(4)}{d} - \frac{c^{2k-5}}{x_k^2 d} + \frac{\ln(2)}{4x_k^2 c^3 d}\left(\frac{k-1}{d}\right) - \frac{\ln^2(2)}{8x_k^2 c^3 d^2}\left(\frac{k-1}{d}\right) \\
[c \leq 1] &\geq \frac{\ln(4)}{d} - \frac{c^{2k-5}}{x_k^2 d} + \frac{\ln(2)}{4x_k^2 d}\left(\frac{k-1}{d}\right) - \frac{\ln^2(2)}{8x_k^2 c^3 d^2}\left(\frac{k-1}{d}\right) \\
&\geq \frac{\ln(4)}{d} - \frac{c^{2k-5}}{x_k^2 d} + \frac{\ln(2)}{4x_k^2 d}\left(\frac{k-1}{d}\right) - \frac{\ln^2(2)}{8(0.45)^2 c^3 d^2} \cdot 1 \\
&\geq \frac{\ln(4)}{d} - \frac{c^{2k-5}}{x_k^2 d} + \frac{\ln(2)}{4x_k^2 d}\left(\frac{k-1}{d}\right) - \frac{0.3}{c^3 d^2} \\
&\geq \frac{\ln(4)}{d} + \frac{\ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} - \frac{0.3}{c^3 d^2}.
\end{aligned}$$

Below, we are going to use the closed-form approximation of x_k, for which we have established $|x_k - \tilde{x}_k| \leq \frac{170.4}{d} = \epsilon$ (Lemma H.1), and also note that $\left|x_k^2 - \tilde{x}_k^2\right| \leq 2x_k\epsilon + \epsilon^2$ (Claim G.10).

Reminder: $\tilde{x}_k = \sqrt{1 - \frac{1}{\ln 4} + 4^{-\frac{k}{d}}\left(\frac{1}{\ln 4} - \frac{k}{d}\right)} = \sqrt{1 - \frac{1}{\ln 4} + c^{2k}\left(\frac{1}{\ln 4} - \frac{k}{d}\right)}$.

$$\begin{aligned}
b(k) + \frac{0.3}{c^3 d^2} &\geq \frac{\ln(4)}{d} + \frac{\ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} \\
&= \frac{4x_k^2 \ln(4) + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} \\
&\geq \frac{4\tilde{x}_k^2 \ln(4) + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} - \frac{4\ln(4)}{4x_k^2 d}\left|x_k^2 - \tilde{x}_k^2\right| \\
&\geq \frac{4\tilde{x}_k^2 \ln(4) + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} - \frac{\ln(4)}{x_k^2 d}\left(2x_k\epsilon + \epsilon^2\right) \\
&= \frac{4\left(\ln(4) - 1 + c^{2k}\left(1 - \frac{k}{d}\ln(4)\right)\right) + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k-5}}{4x_k^2 d} - \frac{\epsilon^2}{x_k^2 d} - \frac{\ln(16)\,\epsilon}{x_k d} \\
&\geq \frac{1.545 + \ln(2)\left(\frac{k-1}{d}\right) + 4c^{2k}\left(1 - \frac{k}{d}\ln(4)\right) - 4c^{2k-5}}{4x_k^2 d} - \frac{\epsilon^2}{x_k^2 d} - \frac{\ln(16)\,\epsilon}{x_k d} \\
&= \frac{1.545 + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k}\ln(4)\frac{k}{d}}{4x_k^2 d} - \frac{c^{2k-5}\left(1 - c^5\right)}{x_k^2 d} - \frac{\epsilon^2}{x_k^2 d} - \frac{\ln(16)\,\epsilon}{x_k d} \\
[\text{G.1}] &\geq \frac{1.545 + \ln(2)\left(\frac{k-1}{d}\right) - 4c^{2k}\ln(4)\frac{k}{d}}{4x_k^2 d} - \frac{5\ln(2)\,c^{2k-5}}{x_k^2 d^2} - \frac{\epsilon^2}{x_k^2 d} - \frac{\ln(16)\,\epsilon}{x_k d}.
\end{aligned}$$

Focusing on the left nominator,

$$1.545+\ln(2)\left(\frac{k-1}{d}\right)-4c^{2k}\ln(4)\frac{k}{d}=1.545+\underbrace{\ln(2)}_{>0}\left(\frac{1}{d}(k-1)-8c^{2k}\frac{k}{d}\right)$$
$$=1.545+\ln(2)\left(\left(1-8c^{2k}\right)\frac{k}{d}-\frac{1}{d}\right)1.545-\ln(2)\left(\left(8c^{2k}-1\right)\frac{k}{d}+\frac{1}{d}\right)$$
$$=1.545-\ln(2)\left(\left(8\cdot4^{-\frac{k}{d}}-1\right)\frac{k}{d}+\frac{1}{d}\right).$$

To upper bound $g(x)=8x\cdot4^{-x}$ (inside $x\in[0,1]$), we show that

$$0\overset{!}{=}g'(x)=8\cdot4^{-x}-8x\ln(4)4^{-x}=8\cdot4^{-x}(1-x\ln(4)),$$

solved by $x=\frac{1}{\ln(4)}$, which falls inside $x\in[0,1]$, meaning it is a global optimum.

The second derivative is

$$g''(x)=\left(8\cdot4^{-x}-8\ln(4)\cdot4^{-x}x\right)'=-4^{2-x}\ln(2)-8\ln(4)\cdot4^{-x}(1-x\ln(4))$$
$$g''\left(\frac{1}{\ln(4)}\right)=-4^{2-\frac{1}{\ln(4)}}\ln(2)=-\frac{16\ln(2)}{e}<0,$$

meaning that the $x=\frac{1}{\ln(4)}$ is the global **maximum.** Also note: $\left(4^{\frac{1}{\ln 4}}\right)^{\ln 4}=4\Rightarrow4^{\frac{1}{\ln 4}}=e$

So overall, we get,

$$1.545+\ln(2)\left(\frac{k-1}{d}\right)-4c^{2k}\ln(4)\frac{k}{d}\geq1.545-\ln(2)\left(\left(8\cdot4^{-\frac{k}{d}}-1\right)\frac{k}{d}+\frac{1}{d}\right)$$
$$\geq1.545-\ln(2)\left(\left(8\cdot4^{-\frac{1}{\ln 4}}-1\right)\frac{1}{\ln 4}+\frac{1}{d}\right)=1.545-\frac{\ln(2)}{d}-\frac{\ln(2)}{2\ln(2)}\left(\frac{8}{e}-1\right)$$
$$\geq1.545-0.972-\frac{0.7}{d}$$
$$\geq0.573-\frac{0.7}{d}.$$

Finally,

$$b(k)+\frac{0.3}{c^3d^2}+\frac{\epsilon^2}{x_k^2d}+\frac{\ln(16)\epsilon}{x_kd}+\frac{5\ln(2)c^{2k-5}}{x_k^2d^2}\geq\frac{1.545+\ln(2)\left(\frac{k-1}{d}\right)-4c^{2k}\ln(4)\frac{k}{d}}{4x_k^2d}$$
$$\geq\frac{0.573-\frac{0.7}{d}}{4x_k^2d}\geq\frac{\mathbf{0.14325}}{\boldsymbol{x_k^2d}}-\frac{0.175}{x_k^2d^2}.$$

Going back to $\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)$, we have

$$\beta_{k+1}\left(\frac{\ln(4)}{d}-\frac{\beta_k}{x_k^2}\right)=\beta_{k+1}b(k)$$
$$\geq\beta_{k+1}\left(\frac{0.14325}{x_k^2d}-\left(\frac{0.3}{c^3d^2}+\frac{\epsilon^2}{x_k^2d}+\frac{\epsilon\ln(16)}{x_kd}+\frac{5\ln(2)c^{2k-5}}{x_k^2d^2}+\frac{0.175}{x_k^2d^2}\right)\right)$$
$$[c<1,\ k\geq2]\geq\beta_{k+1}\left(\frac{0.14325}{x_k^2d}-\left(\frac{0.3}{c^3d^2}+\frac{\epsilon^2}{x_k^2d}+\frac{\epsilon\ln(16)}{x_kd}+\frac{5\ln(2)}{cx_k^2d^2}+\frac{0.175}{x_k^2d^2}\right)\right)$$
$$=\beta_{k+1}\left(\frac{0.14325}{x_k^2d}-\frac{1}{x_k}\left(\frac{0.3x_k}{c^3d^2}+\frac{\epsilon^2}{x_kd}+\frac{\epsilon\ln(16)}{d}+\frac{5\ln(2)}{cx_kd^2}+\frac{0.175}{x_kd^2}\right)\right)$$
$$[0.45\leq x_k\leq1]\geq\beta_{k+1}\left(\frac{0.14325}{x_k^2d}-\frac{1}{x_k}\left(\frac{0.3}{c^3d^2}+\frac{\epsilon\ln(16)}{d}+\frac{5\ln(2)}{c\cdot0.45d^2}+\frac{0.175}{0.45d^2}+\frac{\epsilon^2}{0.45d}\right)\right)$$
$$\geq\beta_{k+1}\left(\frac{0.14325}{x_k^2d}-\frac{1}{x_k}\left(\frac{0.3}{c^3d^2}+\frac{2.78\epsilon}{d}+\frac{7.71}{cd^2}+\frac{0.39}{d^2}+\frac{2.3\epsilon^2}{d}\right)\right)$$

and since $c = 2^{-1/d} \geq 0.9999, \forall d \geq 10{,}000$, and plugging in $\epsilon = \frac{170.4}{d}$:

$$\begin{aligned}
&\beta_{k+1}\left(\frac{\ln(4)}{d} - \frac{\beta_k}{x_k^2}\right) \\
&\geq \beta_{k+1}\left(\frac{0.14325}{x_k^2 d} - \frac{1}{x_k}\left(\frac{0.3}{0.9999^3 d^2} + \frac{2.78 \cdot 170.4}{d^2} + \frac{7.71}{0.9999 d^2} + \frac{0.39}{d^2} + \frac{2.3 \cdot 170.4^2}{d^3}\right)\right) \\
&\geq \beta_{k+1}\left(\frac{0.14325}{x_k^2 d} - \frac{1}{x_k}\left(\frac{482.2}{d^2} + \frac{66783.17}{d^3}\right)\right) \\
&\overset{(1)}{\geq} \beta_{k+1}\left(\frac{0.14325}{x_k^2 d} - \frac{1}{x_k}\left(\frac{488.88}{d^2}\right)\right) \geq \frac{\beta_{k+1}}{x_k d}\left(\frac{0.14325}{x_k} - \frac{488.88}{d}\right),
\end{aligned}$$

where (1) is since $d \geq 10{,}000$. The inside of the parenthesis is **positive** $\forall d \geq \left\lceil \frac{488.88}{0.14325} \right\rceil = 3413$, so we can bound the expression by lower bounding $\frac{\beta_{k+1}}{x_k^2}$.

$$\begin{aligned}
\beta_{k+1}\left(\frac{\ln(4)}{d} - \frac{\beta_k}{x_k^2}\right) &\geq \frac{\beta_{k+1}}{x_k d}\left(\frac{0.14325}{x_k} - \frac{488.88}{d}\right) \geq \frac{0.3c^{2k-3}}{d^2}\left(\frac{0.14325}{x_k^2} - \frac{488.88}{x_k d}\right) \\
&\geq c^{2k-3}\left(\frac{0.0429}{x_k^2 d^2} - \frac{146.7}{x_k d^3}\right).
\end{aligned}$$

And then,

$$\begin{aligned}
A_1(x) &\geq x_{t-1}\left(\frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}\beta_{k+1}\left(\frac{\ln(4)}{d} - \frac{\beta_k}{x_k^2}\right) - \frac{11.84c^{3k-9}}{x_k^3 d^{7/2}} - \frac{109c^{2k-9}}{x_k^3 d^{9/2}}\right) \\
&\geq x_{t-1}\left(\frac{1}{x_k}\frac{c^{k-3}}{\sqrt{d}}c^{2k-3}\left(\frac{0.0429}{x_k^2 d^2} - \frac{146.7}{x_k d^3}\right) - \frac{11.84c^{3k-9}}{x_k^3 d^{7/2}} - \frac{109c^{2k-9}}{x_k^3 d^{9/2}}\right) \\
&= x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{146.7c^{3k-6}}{x_k^2 d^{7/2}} - \frac{11.84c^{3k-9}}{x_k^3 d^{7/2}} - \frac{109c^{2k-9}}{x_k^3 d^{9/2}}\right) \\
[c < 1] &\geq x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{146.7c^{3k-9}}{x_k^2 d^{7/2}} - \frac{11.84c^{3k-9}}{x_k^3 d^{7/2}} - \frac{109c^{3k-9}}{c^k x_k^3 d^{9/2}}\right) \\
&\overset{(1)}{\geq} x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{146.7c^{3k-9}}{x_k^2 d^{7/2}} - \frac{11.84c^{3k-9}}{0.45x_k^2 d^{7/2}} - \frac{109c^{3k-9}}{0.5 \cdot 0.45 x_k^2 d^{9/2}}\right) \\
&\geq x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{173.02c^{3k-9}}{x_k^2 d^{7/2}} - \frac{484.45c^{3k-9}}{x_k^2 d^{9/2}}\right) \\
[d \geq 10{,}000] &\geq x_{t-1}\left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}}\right),
\end{aligned}$$

where (1) is since $x_k \geq 0.45,\ c^k \geq c^d = 0.5$.

□

G.4 Conclusion

We are reminded that we want to show positivity of $\Delta_{t,k} \triangleq \|\mathbf{w}_k - \mathbf{w}_{k+1}\| (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| (\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1}$. Applying Proposition G.18, we show $\forall k \geq t$:

$$\begin{aligned}
\Delta_{t,k} &= \|\mathbf{w}_k - \mathbf{w}_{k+1}\| (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| (\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1} \\
&\geq A_1(k) + A_2(k) + A_3(k) \\
&\quad - \frac{1}{d^{7/2}} \left(\frac{96c^{6k-15}}{x_k} + 113c^{7k-18} \right) - \frac{1}{d^{9/2}} \left(\frac{56.5c^{9k-18}}{x_k^3} + 614c^{6k-30} \right) \\
&\overset{(1)}{\geq} A_1(k) + A_2(k) + A_3(k) \\
&\quad - \frac{1}{d^{7/2}} \left(\frac{96c^{6k-15}}{x_k} + 113c^{7k-18} \right) - \frac{1}{d^{9/2}} \left(\frac{56.5c^{7k-18}}{x_k^3} + 614c^{7k-18}c^{-k-12} \right) \\
&\overset{(2)}{\geq} A_1(k) + A_2(k) + A_3(k) \\
&\quad - \frac{1}{d^{7/2}} \left(\frac{96c^{6k-15}}{x_k} + 113c^{7k-18} \right) - \frac{c^{7k-18}}{d^{9/2}} \left(\frac{56.5}{x_k^3} + 614 \cdot 1.00007^{12} \cdot 2 \right) \\
&\overset{(3)}{\geq} A_1(k) + A_2(k) + A_3(k) \\
&\quad - \frac{1}{d^{7/2}} \left(\frac{96c^{6k-15}}{0.45} + 113c^{7k-18} \right) - \frac{c^{7k-18}}{d^{9/2}} \left(\frac{56.5}{0.45^3} + 1238.36 \right) \\
&\geq A_1(k) + A_2(k) + A_3(k) - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113c^{7k-18} \right) - \frac{1858.39c^{7k-18}}{d^{9/2}} \\
&\overset{(4)}{\geq} A_1(k) + A_2(k) + A_3(k) - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right),
\end{aligned}$$

where (1) is since $c < 1,\ k \geq 2$; (2) is since $2^{1/10000} \leq 1.00007,\ c^{-k} \leq c^{-d} = 2$; (3) is since $x_{k-1} \geq 0.45$; and (4) is since $d \geq 10{,}000$. Plugging in the results of Propositions G.19, G.20 and G.21, we derive

$$\begin{aligned}
&\Delta_{t,k} = \|\mathbf{w}_k - \mathbf{w}_{k+1}\| (\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1} - \|\mathbf{w}_{k-1} - \mathbf{w}_k\| (\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1} \\
&\geq \underbrace{x_{t-1} \left(\frac{0.0429c^{3k-6}}{x_k^3 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}} \right)}_{A_1(k)} \underbrace{- \frac{14.88x_{t-1}c^{3k-12}}{x_k^3 d^{7/2}}}_{A_2(k)} \underbrace{- \frac{6.77c^{k+t-6}}{x_k^4} \left(\frac{1}{d^{7/2}} + \frac{5.42}{d^{9/2}} \right)}_{A_3(k)} \\
&\quad - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right) \\
&= \frac{0.0429x_{t-1}c^{3k-6}}{x_k^3 d^{5/2}} - \frac{173.07x_{t-1}c^{3k-9}}{x_k^2 d^{7/2}} - \frac{14.88x_{t-1}c^{3k-12}}{x_k^3 d^{7/2}} - \frac{6.77c^{k+t-6}}{x_k^4} \left(\frac{1}{d^{7/2}} + \frac{5.42}{d^{9/2}} \right) \\
&\quad - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right) \\
&\overset{(1)}{\geq} \frac{0.0429c^{3k-6}}{x_k^2 d^{5/2}} - \frac{173.07x_{t-1}c^{3k-9}}{x_k^2 d^{7/2}} - \frac{14.88x_{t-1}c^{3k-12}}{x_k^3 d^{7/2}} - \frac{6.77c^{k+t-6}}{x_k^4} \left(\frac{1}{d^{7/2}} + \frac{5.42}{d^{9/2}} \right) \\
&\quad - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right) \\
&\overset{(2)}{\geq} \frac{0.0429c^{3k-6}}{x_k^2 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}} - \frac{14.88c^{3k-12}}{x_k^3 d^{7/2}} - \frac{6.77c^{k+t-6}}{x_k^4} \left(\frac{1}{d^{7/2}} + \frac{5.42}{d^{9/2}} \right) \\
&\quad - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right) \\
&\overset{(3)}{\geq} \frac{0.0429c^{3k-6}}{x_k^2 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}} - \frac{14.88c^{3k-12}}{x_k^3 d^{7/2}} - \frac{6.78c^{k+t-6}}{x_k^4 d^{7/2}} \\
&\quad - \frac{1}{d^{7/2}} \left(213.34c^{6k-15} + 113.19c^{7k-18} \right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(4)}{\geq} \frac{0.0429c^{3k-6}}{x_k^2 d^{5/2}} - \frac{173.07c^{3k-9}}{x_k^2 d^{7/2}} - \frac{14.88c^{3k-12}}{x_k^3 d^{7/2}} - \frac{6.78c^{k+t-6}}{x_k^4 d^{7/2}} \\
&\qquad - \frac{213.34c^{6k-15} + 113.19c^{7k-18}}{x_k^2 d^{7/2}} \\
&= \frac{1}{x_k^2 d^{5/2}} \left(0.0429c^{3k-6} - \frac{1}{d} \Bigg(173.07c^{3k-9} \right. \\
&\qquad\qquad \left. + \frac{14.88c^{3k-12}}{x_k} + \frac{6.78c^{k+t-6}}{x_k^2} + 213.34c^{6k-15} + 113.19c^{7k-18} \Bigg) \right) \\
&\overset{(5)}{\geq} \frac{1}{x_k^2 d^{5/2}} \left(0.0429c^{3k-6} - \frac{1}{d} \Bigg(173.07c^{3k-9} \right. \\
&\qquad\qquad \left. + \frac{14.88c^{3k-12}}{0.45} + \frac{6.78c^{k+t-6}}{0.45^2} + 213.34c^{6k-15} + 113.19c^{7k-18} \Bigg) \right) \\
&\geq \frac{1}{x_k^2 d^{5/2}} \left(0.0429c^{3k-6} - \frac{1}{d} \left(173.07c^{3k-9} \right. \right. \\
&\qquad\qquad \left. \left. + 33.07c^{3k-12} + 33.49c^{k+t-6} + 213.34c^{6k-15} + 113.19c^{7k-18} \right) \right) \\
&= \frac{c^{3k-6}}{x_k^2 d^{5/2}} \left(0.0429 - \frac{1}{d} \left(173.07c^{-3} \right. \right. \\
&\qquad\qquad \left. \left. + 33.07c^{-6} + 33.49c^{-2k+t} + 213.34c^{3k-9} + 113.19c^{4k-12} \right) \right) \\
&\overset{(6)}{\geq} \frac{c^{3k-6}}{x_k^2 d^{5/2}} \left(0.0429 - \frac{1}{d} \left(173.07c^{-3} + 33.07c^{-6} + 33.49c^{-2k} + 213.34c^{-3} + 113.19c^{-4} \right) \right) \\
&\overset{(7)}{\geq} \frac{c^{3k-6}}{x_k^2 d^{5/2}} \left(0.0429 - \frac{1}{d} \left(173.07c^{-3} + 33.07c^{-6} + 33.49 \cdot 4 + 213.34c^{-3} + 113.19c^{-4} \right) \right) \\
&\overset{(8)}{\geq} \frac{c^{3k-6}}{x_k^2 d^{5/2}} \left(0.0429 - \frac{1}{d} \left(173.07 \cdot 0.9999^{-3} \right. \right. \\
&\qquad\qquad \left. \left. + 33.07 \cdot 0.9999^{-6} + 33.49 \cdot 4 + 213.34 \cdot 0.9999^{-3} + 113.19 \cdot 0.9999^{-4} \right) \right)
\end{aligned}
$$

$$
\Rightarrow \Delta_{t,k} \geq \frac{c^{3k-6}}{x_k^2 d^{5/2}} \left(0.0429 - \frac{666.82}{d} \right), \tag{7}
$$

where (1) is since $x_{t-1} > x_k$; (2) is since $x_{t-1} \leq 1$; (3) is since $d \geq 10{,}000$; (4) is since $x_k \leq 1$; (5) is since $x_k \geq 0.45$; (6) is since $c < 1,\ k \geq 2$; (7) is since $c^{-2k} = 4^{k/d} \leq 4$; and (8) is since $d \geq 10{,}000 \Rightarrow c \geq 0.9999$. And so, a sufficient condition for $\frac{(\mathbf{w}_{k-1} - \mathbf{w}_k)^\top \mathbf{w}_{t-1}}{\|\mathbf{w}_{k-1} - \mathbf{w}_k\|} - \frac{(\mathbf{w}_k - \mathbf{w}_{k+1})^\top \mathbf{w}_{t-1}}{\|\mathbf{w}_k - \mathbf{w}_{k+1}\|}$ to be positive and monotonicity to hold, is that $d \geq \left\lceil \frac{666.82}{0.0429} \right\rceil = 15{,}544$. Since this is smaller than $25{,}000$, this concludes our proof of positivity of $\Delta_{t,k}$.

H Approximation of greedy construction first iterate

This section supplements App. G, we recommend reviewing it beforehand if you have not already done so.

We prove the following lemma in this section:

Lemma H.1. *Given the series* $(x)_k$ *recursively defined by* $x_1 = 1$, $x_t = \frac{x_{t-1}+\sqrt{x_{t-1}^2-4\beta_t}}{2}$, $\forall t \in \{2,\dots,d\}$ *where* $c \triangleq 2^{-1/d}$ *and* $\beta_t \triangleq \frac{((t-1)c-(t-2))c^{2t-5}}{d}$, *and the series* $\tilde{x}_k = \sqrt{1-\frac{1}{\ln 4}+4^{-\frac{k}{d}}\left(\frac{1}{\ln 4}-\frac{k}{d}\right)}$, *we have* $\forall d \geq 30$, $\forall k \in [d]$*:*

$$|x_k - \tilde{x}_k| \leq \frac{170.4}{d}\,.$$

Before proving this lemma, we note the following will immediately hold:

Corollary H.2. $\forall d \geq 25{,}000$, $\forall k \in [d]$*:* $x_k \geq 0.45$.

This is since x_k is decreasing (Claim G.5), so $\forall k \in [d]$:

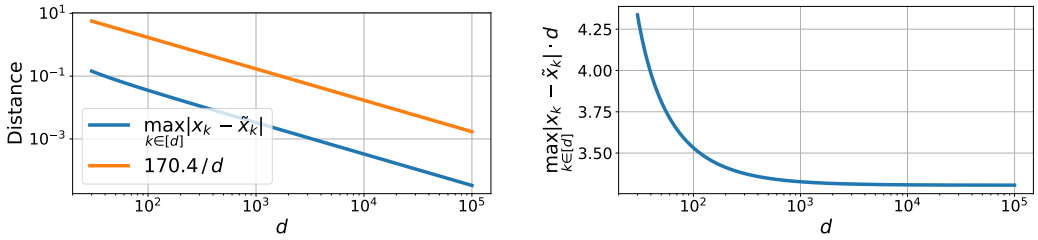
$$x_k \geq x_d \geq \tilde{x}_d - \frac{170.4}{d} = \sqrt{1-\frac{1}{\ln 4}+4^{-1}\left(\frac{1}{\ln 4}-1\right)} - \frac{170.4}{d}$$
$$[d \geq 25{,}000] \geq 0.45\,.$$

This corollary is very useful for the proof in App. G.

H.1 Proof outline

Firstly, we show this holds numerically for $30 \leq d < 100{,}000$, as can be seen in Figure 19. We then prove analytically for $d \geq 100{,}000$, using Euler's method.

Compute resources The numerical validation took 6 hours to run on a home PC with i5-9400F CPU and 16GB RAM.



(a) $|x_k - \tilde{x}_k| \leq \frac{170.4}{d}$. This is a loose, analytically derived upper bound.

(b) Actual upper bound is $< \frac{4.5}{d}$

Figure 19: **Numerical proof of Lemma H.1 for d<100,000.** Using the recursive definition of x_k, we calculated the series for each value of d, $\forall k \in [d]$, and compared with $\tilde{x}_k$.

H.1.1 Euler's method construction

Define

$$f(\tau, x) = d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\,,$$
$$\beta(\tau) = \frac{\left((d\tau-1)\,2^{-1/d} - (d\tau-2)\right)2^{(5-2d\tau)/d}}{d}\,.$$

Then using step size of $h = \frac{1}{d}$ in Euler's method we have the iterates

$$\begin{aligned} x_{k+1} &= x_k + h \cdot f(\tau_k, x_k) \,, \\ \tau_{k+1} &= \tau_k + h \,, \end{aligned}$$

and thus

$$x_{k+1} = x_k + \frac{\sqrt{x_k^2 - 4\beta\left(\frac{k+1}{d}\right)} - x_k}{2} = \frac{x_k + \sqrt{x_k^2 - 4\beta_{k+1}}}{2} \,,$$

which are exactly the iterates we want to solve for.

These are the Euler's iterates for the differential equation

$$\begin{aligned} x'(\tau) &= f(\tau, x(\tau)) \,, \\ x(0) &= 1 \,. \end{aligned}$$

While it's hard to find an exact solution to this equation, we have managed (see Proposition H.17) to prove that for $d \geq 100{,}000$:

$$|x(\tau) - \tilde{x}(\tau)| \leq \frac{38.9822}{d} \,,$$

where we defined the *function*

$$\tilde{x}(\tau) = \sqrt{1 - \frac{1}{\ln 4} + 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)} \,,$$

such that $\tilde{x}_k = \tilde{x}\left(\frac{k}{d}\right)$. We now proceed to bound the iterates using the global truncation error of Euler's method.

$$\left| x_k - x\left(\frac{k}{d}\right) \right| \leq \frac{hM}{2L}\left(\exp\left(L\left(\tfrac{k}{d} - 0\right)\right) - 1\right) \leq \frac{M}{2Ld}\left(e^L - 1\right) \,,$$

where

$$\begin{aligned} L &= \max_{x,\tau \in [0,1]} \left| \frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x) \right| \text{ (where } \tau \text{ is treated as a constant)} \,, \\ M &= \max_{\tau \in [0,1]} \left| \frac{\mathrm{d}^2}{\mathrm{d}\tau^2} x(\tau) \right| = \left| \frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau, x(\tau)) \right| \,. \end{aligned}$$

For $d \geq 100{,}000$ we have $L \leq 4.7955$ from Proposition H.19, and Proposition H.20 gives $M \leq 10.5027$.

So in total

$$\left| x_k - x\left(\frac{k}{d}\right) \right| \leq \frac{10.5027}{2 \cdot 4.7955 d} \cdot \left(e^{4.7955} - 1\right) \leq \frac{131.3685}{d} \,.$$

Now we combine this with the above and get

$$\begin{aligned} |x_k - \tilde{x}_k| = \left| x_k - \tilde{x}\left(\frac{k}{d}\right) \right| &= \left| x_k - x\left(\frac{k}{d}\right) + x\left(\frac{k}{d}\right) - \tilde{x}\left(\frac{k}{d}\right) \right| \\ &\leq \left| x_k - x\left(\frac{k}{d}\right) \right| + \left| x\left(\frac{k}{d}\right) - \tilde{x}\left(\frac{k}{d}\right) \right| \\ &\leq \frac{131.3685}{d} + \frac{38.9822}{d} \leq \frac{170.4}{d} \,. \end{aligned}$$

H.2 Claims used to prove Lemma H.1

Remark H.3. The solution to the ODE $x(\tau)\,x'(\tau) = -f(x)\,,\ x(0) = 1$ is

$$x(\tau) = \sqrt{1 - 2\int_0^\tau f(s)\,ds}.$$

Claim H.4. $\forall 0 \le a \le M \le 1:\ 1 - \frac{1-\sqrt{1-M}}{M}a \le \sqrt{1-a} \le 1 - \frac{a}{2}$.

Proof. The right side inequality is trivial: $\left(1-\frac{a}{2}\right)^2 = 1 - a + \frac{a^2}{4} \ge 1 - a = \left(\sqrt{1-a}\right)^2$.

For the left side: denote $f(a) = \sqrt{1-a}$. f is concave: $f'(a) = -\frac{1}{2\sqrt{1-a}},\ f''(a) = \frac{\frac{-2}{2\sqrt{1-a}}}{4(1-a)} \le 0$. So we have $\forall 0 \le a \le M \le 1$:

$$\begin{aligned}\sqrt{1-a} = f(a) &\ge \frac{(M-a)\,f(0) + a f(M)}{M} \\ &= 1 - \frac{a}{M} + \frac{a}{M}\sqrt{1-M} = 1 - \frac{1-\sqrt{1-M}}{M}a\,.\end{aligned}$$

□

Proposition H.5. *Assuming* $\forall\tau \in [0,1] : 0 \le \frac{g(\tau)}{x^2} \le 1$, *the solution of* $x(0) = 1,\ x'(\tau) = d\frac{\sqrt{x^2 - g(\tau)} - x}{2}$ *obeys*

$$x(\tau) \in \left[\sqrt{1 - d\frac{\left(1-\sqrt{1-M}\right)}{M}\int_0^\tau g(s)\,\mathrm{d}s},\ \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s}\right]$$

for $M := \max_{s\in[0,1]} \frac{g(s)}{x(s)^2}$.

Proof. If $0 \le g(0) \le 1$, we know from continuity that $g(\tau) \le x^2(\tau)$, at least until some $\tau = \beta > 0$. Note that the solution does not exist when $g(\tau) > x^2(\tau)$. We only care about solutions that exist until at least $\tau = 1$, so we assume there exists $0 \le M \le 1$ such that $\forall\tau \in [0,1] : 0 \le \frac{g(\tau)}{x^2} \le M \le 1$.

From Claim H.4 we have:

$$\begin{aligned}
1 - \frac{1-\sqrt{1-M}}{M}\frac{g(\tau)}{x^2} &\le \sqrt{1 - \frac{g(\tau)}{x^2}} \le 1 - \frac{g(\tau)}{2x^2} \\
\left(1 - \frac{1-\sqrt{1-M}}{M}\frac{g(\tau)}{x^2}\right)x &\le x\sqrt{1 - \frac{g(\tau)}{x^2}} \le \left(1 - \frac{g(\tau)}{2x^2}\right)x \\
x - \frac{1-\sqrt{1-M}}{M}\frac{g(\tau)}{x} &\le \sqrt{x^2 - g(\tau)} \le x - \frac{g(\tau)}{2x} \\
-\frac{1-\sqrt{1-M}}{M}\frac{g(\tau)}{x} &\le \sqrt{x^2 - g(\tau)} - x \le -\frac{g(\tau)}{2x} \\
-d\frac{1-\sqrt{1-M}}{2M}\frac{g(\tau)}{x} &\le d\frac{\sqrt{x^2 - g(\tau)} - x}{2} \le -\frac{dg(\tau)}{4x} \\
\frac{\left(\sqrt{1-M} - 1\right)dg(\tau)}{2Mx(\tau)} &\le x'(\tau) \le \frac{-dg(\tau)}{4x(\tau)} \\
\frac{\left(\sqrt{1-M} - 1\right)dg(\tau)}{2M} &\le x'(\tau)\,x(\tau) \le \frac{-dg(\tau)}{4}\,.
\end{aligned}$$

From Remark H.3, we know that the solution to the ODE $x(\tau)\,x'(\tau) = -f(x),\ x(0) = 1$ is $x(\tau) = \sqrt{1 - 2\int_0^\tau f(s)\,ds}$. substituting we get:

$$\frac{\left(\sqrt{1-M}-1\right)dg(\tau)}{2M} \le -f(x) \le \frac{-dg(\tau)}{4}$$
$$\int_0^\tau \frac{\left(\sqrt{1-M}-1\right)dg(s)}{2M}\mathrm{d}s \le -\int_0^\tau f(s)\,\mathrm{d}s \le \int_0^\tau \frac{-dg(s)}{4}\mathrm{d}s$$
$$\sqrt{1-2\int_0^\tau \frac{d}{2}\frac{\left(1-\sqrt{1-M}\right)}{M}g(s)\,\mathrm{d}s} \le \sqrt{1-2\int_0^\tau f(s)\,\mathrm{d}s} \le \sqrt{1-2\int_0^\tau \frac{d}{4}g(s)\,\mathrm{d}s}\,.$$

From this follows:

$$\sqrt{1-2\int_0^\tau \frac{d}{2}\frac{\left(1-\sqrt{1-M}\right)}{M}g(s)\,\mathrm{d}s} \le x(\tau) \le \sqrt{1-2\int_0^\tau \frac{d}{4}g(s)\,\mathrm{d}s}$$
$$\sqrt{1-d\frac{\left(1-\sqrt{1-M}\right)}{M}\int_0^\tau g(s)\,\mathrm{d}s} \le x(\tau) \le \sqrt{1-\frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s}\,.$$

□

Claim H.6. $f(d) \triangleq -d\left(1-2^{-1/d}\right)$ is decreasing $\forall d \ge 1$.

Proof. $f'(d) = -\left(1-2^{-1/d}\right) - d\left((-1)\cdot\frac{1}{d^2}\ln 2\cdot 2^{-1/d}\right) = -\left(1-2^{-1/d}\right) + \frac{\ln 2}{d}2^{-1/d} = -1 - \left(1-\frac{\ln 2}{d}\right)2^{-1/d} < 0$. □

Claim H.7. $\forall d \ge 1:\ d\left(2^{1/d}-1\right) \ge \ln 2$.

Proof. Using Taylor's expansion:

$$2^{1/d} = e^{\frac{\ln 2}{d}} = 1 + \frac{\ln 2}{d} + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\frac{\ln 2}{d}\right)^i$$
$$\Rightarrow d\left(2^{1/d}-1\right) = \ln 2 + \sum_{i=2}^{\infty}\frac{1}{i!}\frac{(\ln 2)^i}{d^{i-1}} \ge \ln 2\,.$$

□

Claim H.8. $-d\left(1-2^{-1/d}\right) \ge -\ln 2$ (alternatively: $2^{-1/d} \ge 1 - \frac{\ln 2}{d}$).

Proof. From Claim H.6 we know that $-d\left(1-2^{-1/d}\right)$ is decreasing with d, so we have $-d\left(1-2^{-1/d}\right) \ge \lim_{d\to\infty} -d\left(1-2^{-1/d}\right)$:

$$\lim_{d\to\infty} -d\left(1-2^{-1/d}\right) = \lim_{h\to 0^+}\frac{2^{-h}-1}{h}$$
$$= \lim_{h\to 0^+}\frac{2^{-h}-2^0}{h}\,.$$

We recognize this as the definition of the derivative of 2^{-x} for $x = 0^+$,so we have:

$$\lim_{d\to\infty} -d\left(1-2^{-1/d}\right) = \frac{\mathrm{d}\left(2^{-x}\right)}{\mathrm{d}x}\left(x = 0^+\right) = -\ln 2\cdot 2^0 = -\ln 2\,.$$

□

Claim H.9. $\beta(\tau) \geq 0$, decreasing and convex for $\tau \leq \frac{1}{\ln 2}$.

Proof. Reminder that $\beta(\tau) = \frac{\left((d\tau-1)2^{-1/d}-(d\tau-2)\right)2^{(5-2d\tau)/d}}{d}$, and $d \geq 1$.

Denote $\beta(\tau) = \frac{1}{d} f(\tau) g(\tau)$, where $f(\tau) = (d\tau-1)2^{-1/d} - (d\tau-2)$ and $g(\tau) = 2^{\frac{5-2d\tau}{d}}$.

We have $\forall \tau,\ g(\tau) > 0$.

Note that from Claim H.8 $1 - \frac{\ln 2}{d} \leq 2^{-1/d} \leq 1$, so:

$$\begin{aligned} f(\tau) &= 2^{-1/d} d\tau - 2^{-1/d} - d\tau + 2 \\ &\geq \left(1 - \frac{\ln 2}{d}\right) d\tau - 1 - d\tau + 2 \\ &= -\tau \ln 2 + 1\,, \end{aligned}$$

so $f(\tau) \geq 0$ for $\tau \leq \frac{1}{\ln 2}$. Thus $\beta(\tau) \geq 0$ for $\tau \leq \frac{1}{\ln 2}$. Now we note that $f'(\tau) = d\left(2^{-1/d} - 1\right) < 0,\ \forall d \geq 1,\ \forall \tau$ and $g'(\tau) = -2\ln 2 \cdot 2^{\frac{5-2d\tau}{d}} < 0,\ \forall d,\ \forall \tau$

So:

$$\beta'(\tau) = \frac{1}{d}\left(f'(\tau) g(\tau) + g'(\tau) f(\tau)\right) < 0\,,$$

as long as $g(\tau) > 0$ and $f(\tau) \geq 0$ - which we get for $\tau \leq \frac{1}{\ln 2}$.

Now note $f''(\tau) = 0$ and $g''(\tau) = 4\ln^2 2 \cdot 2^{\frac{5-2d\tau}{d}} > 0$, so:

$$\begin{aligned} \beta''(\tau) &= \frac{1}{d}\left(f''(\tau) g(\tau) + f'(\tau) g'(\tau) + g''(\tau) f(\tau) + f'(\tau) g'(\tau)\right) \\ &= \frac{1}{d}\left(2f'(\tau) g'(\tau) + g''(\tau) f(\tau)\right) > 0\,, \end{aligned}$$

as long as $f(\tau) \geq 0$ - which we get for $\tau \leq \frac{1}{\ln 2}$. □

Claim H.10. For $x(\tau)$, $g(\tau) = 4\beta\left(\tau + \frac{1}{d}\right)$. We also have $\forall d \geq 3$, $\max_{s\in[0,1]} g(s) = g(0) = 4\frac{2^{3/d}}{d}$.

Proof. Substituting $x'(\tau) = d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)}-x}{2}$ in $x'(\tau) = d\frac{\sqrt{x^2-g(\tau)}-x}{2}$ we get $g(\tau) = 4\beta\left(\tau + \frac{1}{d}\right)$.

For $\tau \in [0,1]$ and $d \geq 3$, $\tau + \frac{1}{d} \leq \frac{1}{\ln 2}$. We get from Claim H.9 that β is decreasing, so:

$$\begin{aligned} \max_{s\in[0,1]} g(s) = g(0) &= 4\beta\left(\frac{1}{d}\right) \\ &= 4\frac{\left((1-1)2^{-1/d} - (1-2)\right)2^{(5-2)/d}}{d} = 4\frac{2^{3/d}}{d} \end{aligned}$$

□

Proposition H.11. *For $d \geq 100{,}000$ and $\tau \in [0,1]$, $0 \leq d\int_0^\tau g(s)\,\mathrm{d}s \leq 1.5821$.*

Proof. For $\tau \in [0,1]$ and $d \geq 3$, $\tau + \frac{1}{d} \leq \frac{1}{\ln 2}$. We get from Claim H.9 that β is positive and thus $d\int_0^\tau g(s)\,\mathrm{d}s = 4d\int_0^\tau \beta\left(s+\frac{1}{d}\right)\mathrm{d}s \geq 0$. For the right side inequality, we have:

$$\begin{aligned}
d\int_0^\tau g(s)\,\mathrm{d}s &= 4d\int_0^\tau \beta\left(s+\frac{1}{d}\right)\mathrm{d}s \\
[\beta \geq 0] &\leq 4d\int_0^1 \beta\left(s+\frac{1}{d}\right)\mathrm{d}s \\
&= 4d\int_0^1 \left(\frac{\left((ds+1-1)\,2^{-1/d}-(ds+1-2)\right)2^{(5-2ds-2)/d}}{d}\right)\mathrm{d}s \\
&= 4\int_0^1 \left(ds2^{-1/d}-ds+1\right)2^{3/d}2^{-2s}\mathrm{d}s \\
&= 4\cdot 2^{3/d}\left[\int_0^1 2^{-2s}\mathrm{d}s - d\left(1-2^{-1/d}\right)\int_0^1 s2^{-2s}\mathrm{d}s\right] \\
&= 4\cdot 2^{3/d}\left[\left[-\frac{2^{-2s}}{\ln 4}\right]\Bigg|_0^1 - d\left(1-2^{-1/d}\right)\left[-\frac{2^{-2s}\left(s\ln 4+1\right)}{\ln^2 4}\right]\Bigg|_0^1\right] \\
&= 4\cdot 2^{2/d}\left[2^{1/d}\frac{3}{4\ln 4} - d\left(2^{1/d}-1\right)\left[\frac{4-(\ln 4+1)}{4\ln^2 4}\right]\right]
\end{aligned}$$

From Claim H.7 we know that $d\left(2^{1/d}-1\right) \geq \ln 2$, so:

$$\begin{aligned}
d\int_0^\tau g(s)\,\mathrm{d}s &\leq 4d\int_0^1 \beta\left(s+\frac{1}{d}\right)\mathrm{d}s \leq 4\cdot 2^{2/d}\left[2^{1/d}\frac{3}{4\ln 4} - \ln 2\left[\frac{3-\ln 4}{4\ln^2 4}\right]\right] \\
[d \geq 100{,}000] &\leq 4\cdot 2^{2/100000}\left[2^{1/100000}\frac{3}{4\ln 4} - \ln 2\left[\frac{3-\ln 4}{4\ln^2 4}\right]\right] \leq 1.5821
\end{aligned}$$

□

Claim H.12. $\frac{1-\sqrt{1-x}}{x} - \frac{1}{2} \leq \frac{x}{2}$ for $x \in (0,1]$

Proof.

$$a(x) \triangleq \frac{1-\sqrt{1-x}}{x} = \frac{1-\sqrt{1-x}}{x}\frac{1+\sqrt{1-x}}{1+\sqrt{1-x}} = \frac{x}{x\left(1+\sqrt{1-x}\right)} = \frac{1}{1+\sqrt{1-x}}.$$

This function is monotonically increasing and convex for $x \in (0,1]$:

$$\begin{aligned}
a'(x) &= \frac{\frac{1}{2\sqrt{1-x}}}{\left(1+\sqrt{1-x}\right)^2}, \\
a''(x) &= \frac{-1\cdot\left(2\frac{-1}{2\sqrt{1-x}}\left(1+\sqrt{1-x}\right)^2 + 2\left(1+\sqrt{1-x}\right)\frac{-1}{2\sqrt{1-x}}\cdot 2\sqrt{1-x}\right)}{4(1-x)\left(1+\sqrt{1-x}\right)^4} \\
&= \frac{\frac{1}{\sqrt{1-x}}\left(1+\sqrt{1-x}\right)^2 + 2\left(1+\sqrt{1-x}\right)}{4(1-x)\left(1+\sqrt{1-x}\right)^4} \geq 0
\end{aligned}$$

Note that in $x=0$ there is a removable discontinuity, and:

$$\lim_{x\to 0}\frac{1-\sqrt{1-x}}{x} = \frac{1}{2}$$

So we have:

$$\frac{1-\sqrt{1-x}}{x} = a(x) \le (1-x)\,a(0) + xa(1) = (1-x)\frac{1}{2} + x = \frac{1}{2} + \frac{1}{2}x$$
$$\implies 2\left(\frac{1-\sqrt{1-x}}{x} - \frac{1}{2}\right) \le x\,.$$

□

Proposition H.13. *For* $d \ge 100{,}000,\ M \triangleq \max_{s\in[0,1]} \frac{g(s)}{x(s)^2} \le \frac{19.158}{d}$

Proof. We first note that the solution of $x(0) = 1$, $x'(\tau) = d\frac{\sqrt{x^2 - g(\tau)} - x}{2}$ is decreasing from $\tau = 0$ and as long as $g(\tau) \ge 0$, which we know from Claim H.9 is the case for $\tau \in [0,1]$ and $d \ge 3$, since $\tau + \frac{1}{d} \le \frac{1}{\ln 2}$. This means the minimum of $x(\tau)$ in $[0,1]$ is $x(1)$. In addition, since we assume $g(\tau) \le x^2(\tau)$, and for $\tau \in [0,1]$ we have $g(\tau) > 0$, we know $x(\tau) > 0$ there and thus the minimum of $x(\tau)^2$ is also $x(1)^2$, and that $x(1)^2 \le x(0)^2 = 1$.

So we know, applying Claim H.10:

$$M \triangleq \max_{s\in[0,1]} \frac{g(s)}{x(s)^2} \le \frac{\max_{s\in[0,1]} g(s)}{\min_{s\in[0,1]} x(s)^2} = \frac{4\frac{2^{3/d}}{d}}{x(1)^2}$$

From Proposition H.5 we know that:

$$\sqrt{1 - d\frac{\left(1-\sqrt{1-M}\right)}{M}\int_0^1 g(s)\,ds} \le x(1)$$

Substituting and denoting $A = d\int_0^1 g(s)\,\mathrm{d}s$ we get:

$$\sqrt{1 - \frac{\left(1-\sqrt{1-\frac{4\frac{2^{3/d}}{d}}{x(1)^2}}\right)}{\frac{4\frac{2^{3/d}}{d}}{x(1)^2}}A} \le x(1)$$
$$1 - \frac{\left(1-\sqrt{1-\frac{4\frac{2^{3/d}}{d}}{x(1)^2}}\right)}{\frac{4\frac{2^{3/d}}{d}}{x(1)^2}}A \le x(1)^2$$
$$\frac{4\frac{2^{3/d}}{d}}{x(1)^2} - A + A\sqrt{1 - \frac{4\frac{2^{3/d}}{d}}{x(1)^2}} \le 4\frac{2^{3/d}}{d}$$
$$A\sqrt{1 - \frac{4\frac{2^{3/d}}{d}}{x(1)^2}} \le -4\frac{2^{3/d}}{d}\left(\frac{1}{x(1)^2} - 1\right) + A$$

For simplicity denote $z = \frac{1}{x(1)^2},\ r = 4\frac{2^{3/d}}{d}$. We are reminded that $z \ge 1$, and we are looking for an upper bound for it, so we can have a lower bound for $x(1)^2$. We have:

$$A^2(1-rz) \le (-r(z-1) + A)^2 = r^2(z-1)^2 - 2Ar(z-1) + A^2$$
$$-A^2rz \le r^2z^2 - 2r^2z + r^2 - 2Arz + 2Ar$$
$$0 \le rz^2 + \left(A^2 - 2A - 2r\right)z + 2A + r$$

Finding the roots:

$$z_{1,2} = \frac{2r + A(2-A) \pm \sqrt{(2r + A(2-A))^2 - 4r(2A + r)}}{2r}$$

Since we are looking for an upper bound, we care about the smaller root:

$$\begin{aligned} z &\leq \frac{2r + A(2-A) - \sqrt{4r^2 + 4rA(2-A) + A^2(2-A)^2 - 8rA - 4r^2}}{2r} \\ &= \frac{2r + A(2-A) - \sqrt{-4rA^2 + A^2(2-A)^2}}{2r} = \frac{2r + A(2-A) - A\sqrt{(2-A)^2 - 4r}}{2r} \\ &= 1 + \frac{A(2-A)}{2r}\left(1 - \sqrt{1 - \frac{4r}{(2-A)^2}}\right) \end{aligned}$$

For $d \geq 100{,}000$, $\frac{4r}{(2-A)^2} = \frac{4 \cdot 4 \frac{2^{3/d}}{d}}{(2-A)^2} \leq \frac{4 \cdot 4 \cdot \frac{2^{3/100000}}{100000}}{(2-1.5821)^2} \leq 10^{-3} < 1$, (we used Proposition H.11), so we can apply Claim H.12:

$$\begin{aligned} z &\leq 1 + \frac{A(2-A)}{2r}\left(\frac{4r}{2(2-A)^2}\left(\frac{4r}{(2-A)^2} + 1\right)\right) \\ &= 1 + \frac{A}{(2-A)}\left(\frac{4r}{(2-A)^2} + 1\right) = 1 + \frac{A}{2-A} + \frac{4Ar}{(2-A)^3} \\ \left[d \geq 100{,}000 \Rightarrow r \leq 4.1 \cdot 10^{-5}\right] &\leq 1 + \frac{A}{2-A} + 4 \cdot 4.1 \cdot 10^{-5}\frac{A}{(2-A)^3} \\ [A \leq 1.5821,\ \text{H.11}] &\leq 1 + \frac{1.5821}{0.4179} + 4 \cdot 4.1 \cdot 10^{-5} \cdot \frac{1.5821}{0.4179^3} \leq 4.7894 \end{aligned}$$

So we have For $d \geq 100{,}000$:

$$\begin{aligned} &\frac{1}{x(1)^2} \leq 4.7894 \\ &\Longrightarrow M \leq 4.7894 \cdot 4\frac{2^{3/d}}{d} \leq 19.1576 \cdot \frac{2^{3/100000}}{d} \leq \frac{19.158}{d}. \end{aligned}$$

□

Proposition H.14. *For $d \geq 100{,}000$, we have* $\left|x(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s}\right| \leq \frac{33.1539}{d}$

Proof. Let's denote $A = d\int_0^\tau g(s)\,\mathrm{d}s$. We know that

$$\begin{aligned}
\sqrt{1 - \frac{\left(1-\sqrt{1-M}\right)}{M}A} \leq x(\tau) \leq \sqrt{1-\frac{1}{2}A} \\
\left|x(\tau) - \sqrt{1-\frac{1}{2}A}\right| &\leq \sqrt{1-\frac{1}{2}A} - \sqrt{1-\frac{\left(1-\sqrt{1-M}\right)}{M}A} \\
&\leq \frac{1-\frac{1}{2}A - 1 + \frac{\left(1-\sqrt{1-M}\right)}{M}A}{\sqrt{1-\frac{1}{2}A} + \sqrt{1-\frac{\left(1-\sqrt{1-M}\right)}{M}A}} \\
&\leq \frac{\left(\frac{\left(1-\sqrt{1-M}\right)}{M} - \frac{1}{2}\right)A}{\sqrt{1-\frac{1}{2}A} + \sqrt{1-\frac{\left(1-\sqrt{1-M}\right)}{M}A}} \\
&\leq \frac{A}{\sqrt{1-\frac{1}{2}A}}\left(\frac{\left(1-\sqrt{1-M}\right)}{M} - \frac{1}{2}\right).
\end{aligned}$$

From the last claim we have for $d \geq 100{,}000$ that $A \leq 1.5821$, then $\frac{A}{\sqrt{1-\frac{1}{2}A}} \leq \frac{1.5821}{\sqrt{1-\frac{1.5821}{2}}} \leq 3.4611$.

We further know from Claim H.12 that $\frac{1-\sqrt{1-x}}{x} - \frac{1}{2} \leq \frac{x}{2}$ for $x \in [0,1]$.

So we have, applying Proposition H.13:

$$\left|x(\tau) - \sqrt{1-\frac{1}{2}A}\right| \leq 3.4611\frac{M}{2} \leq \frac{3.4611 \cdot 19.158}{2d} \leq \frac{33.1539}{d}$$

□

Claim H.15. $\forall x \in [0,1]:\ 2^x \leq 1 + x$.

Proof. 2^x is convex, so we get in $[0,1]$:

$$2^x \leq (1-x)\,2^0 + x2^1 = 1 - x + 2x = 1 + x\,.$$

□

Proposition H.16. *For $d \geq 100{,}000$, We have* $\left|\tilde{x}(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s}\right| \leq \frac{5.8283}{d}$.

Proof. We have

$$\tilde{x}(\tau) = \sqrt{1 - \frac{1}{\ln 4} + 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)}$$

Now

$$\begin{aligned}
\beta(\tau) &= \frac{\left((d\tau - 1)\,2^{-1/d} - (d\tau - 2)\right)2^{(5-2d\tau)/d}}{d} \\
g(\tau) &= 4\beta\left(\tau + \frac{1}{d}\right) = 4\frac{\left(d\tau 2^{-1/d} - (d\tau - 1)\right)2^{(3-2d\tau)/d}}{d}
\end{aligned}$$

Let's define $A(\tau) = \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s$, $B(\tau) = \frac{1}{\ln 4} - 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)$

We have $A(0) = B(0) = 0$. and $A(\tau) - B(\tau)$ is non negative and increasing for $\tau \in [0,1]$:

$$\begin{aligned}
\frac{\mathrm{d}\left(A(\tau) - B(\tau)\right)}{\mathrm{d}\tau} &= \frac{d}{2} g(\tau) - 4^{-\tau}\left(2 - \tau \ln 4\right) \\
&= \frac{d}{2} 4 \frac{\left(d\tau 2^{-1/d} - (d\tau - 1)\right) 2^{(3-2d\tau)/d}}{d} - 4^{-\tau}\left(2 - \tau \ln 4\right) \\
&= 4^{-\tau}\left(2\left(2^{3/d} - 1\right) + \tau\left(\ln 4 - 2 \cdot 2^{3/d} d\left(1 - 2^{-1/d}\right)\right)\right)
\end{aligned}$$

If we assume $\ln 4 \geq 2 \cdot 2^{3/d} d\left(1 - 2^{-1/d}\right)$, then the derivative is in fact positive and we are done. If we assume the opposite we have:

$$\begin{aligned}
\frac{\mathrm{d}\left(A(\tau) - B(\tau)\right)}{\mathrm{d}\tau} &= 4^{-\tau}\left(2\left(2^{3/d} - 1\right) - \tau\left(2 \cdot 2^{3/d} d\left(1 - 2^{-1/d}\right) - \ln 4\right)\right) \\
[\tau \in [0,1]] &\geq 4^{-\tau}\left(2\left(2^{3/d} - 1\right) - \left(2 \cdot 2^{3/d} d\left(1 - 2^{-1/d}\right) - \ln 4\right)\right) \\
&= 4^{-\tau}\left(2 \cdot 2^{3/d}\left(1 - d\left(1 - 2^{-1/d}\right)\right) - 2 + \ln 4\right) \\
[\text{H.8}] &\geq 4^{-\tau}\left(2 \cdot 2^{3/d}(1 - \ln 2) - 2 + \ln 4\right) \\
&\geq 4^{-\tau}\left(2(1 - \ln 2) - 2 + \ln 4\right) \\
&= 4^{-\tau}\left(2 - 2\ln 2 - 2 + \ln 4\right) = 0
\end{aligned}$$

This means that

$$0 \leq A(\tau) - B(\tau) \leq A(1) - B(1)$$

In Proposition H.11 we saw that:

$$\begin{aligned}
d\int_0^1 g(s)\,\mathrm{d}s &\leq 4 \cdot 2^{2/d}\left[2^{1/d}\frac{3}{4\ln 4} - \ln 2\left[\frac{3 - \ln 4}{4\ln^2 4}\right]\right] \\
\Rightarrow A(1) &\leq 2 \cdot 2^{2/d}\left[2^{1/d}\frac{3}{4\ln 4} - \ln 2\left[\frac{3 - \ln 4}{4\ln^2 4}\right]\right] \\
[\text{H.15}, d \geq 2,\ \ln 4 = 2\ln 2] &\leq 2\left(1 + \frac{2}{d}\right)\left[\frac{3\left(1 + \frac{1}{d}\right)}{8\ln 2} - \ln 2\left[\frac{3 - \ln 4}{16\ln^2 2}\right]\right] \\
&= 2\left(1 + \frac{2}{d}\right)\left[\frac{6 + \frac{6}{d}}{16\ln 2} - \frac{3 - \ln 4}{16\ln 2}\right] = \left(1 + \frac{2}{d}\right)\left[\frac{3 + 2\ln 2 + \frac{6}{d}}{8\ln 2}\right].
\end{aligned}$$

So:

$$\begin{aligned}
A(1) - B(1) &\leq \left(1 + \frac{2}{d}\right)\left[\frac{3 + 2\ln 2 + \frac{6}{d}}{8\ln 2}\right] - \left(\frac{1}{\ln 4} - \frac{1}{4}\left(\frac{1}{\ln 4} - 1\right)\right) \\
&= \left(1 + \frac{2}{d}\right)\left[\frac{3 + 2\ln 2 + \frac{6}{d}}{8\ln 2}\right] - \left(\frac{4}{4\ln 4} - \frac{1}{4\ln 4} + \frac{\ln 4}{4\ln 4}\right) \\
&= \left(1 + \frac{2}{d}\right)\left[\frac{3 + 2\ln 2 + \frac{6}{d}}{8\ln 2}\right] - \frac{3 + 2\ln 2}{8\ln 2} \\
&= \frac{6}{d \cdot 8\ln 2} + \frac{2}{d}\left[\frac{3 + 2\ln 2 + \frac{6}{d}}{8\ln 2}\right] \\
[d \geq 100{,}000] &\leq \frac{2.6641}{d}
\end{aligned}$$

So we have:

$$0 \leq A(\tau) - B(\tau) \leq \frac{2.6641}{d}$$

Now note that

$$\tilde{x}(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s} = \sqrt{1 - B(\tau)} - \sqrt{1 - A(\tau)} = \frac{1 - B(\tau) - (1 - A(\tau))}{\sqrt{1 - B(\tau)} + \sqrt{1 - A(\tau)}}$$
$$= \frac{A(\tau) - B(\tau)}{\sqrt{1 - A(\tau)} + \sqrt{1 - B(\tau)}}$$

Since $0 \leq A(\tau) - B(\tau)$ we have $\tilde{x}(\tau) \geq \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,\mathrm{d}s}$.

Now note that

$$\begin{aligned}\sqrt{1 - A(\tau)} + \sqrt{1 - B(\tau)} &\geq \sqrt{1 - B(\tau)} \\ &= \sqrt{1 - \frac{1}{\ln 4} + 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)} \\ &\geq \sqrt{1 - \frac{1}{\ln 4} + 4^{-1}\left(\frac{1}{\ln 4} - 1\right)} \\ &\geq 0.4571\end{aligned}$$

So

$$0 \leq \tilde{x}(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,ds} \leq \frac{2.6641}{0.4571d} \leq \frac{5.8283}{d}$$

□

Proposition H.17. *For $d \geq 100{,}000$, we have $|\tilde{x}(\tau) - x(\tau)| \leq \frac{38.9822}{d}$.*

Proof. From Proposition H.16 and Proposition H.14:

$$\begin{aligned}|\tilde{x}(\tau) - x(\tau)| &= \left|\tilde{x}(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,ds} + \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,ds} - x(\tau)\right| \\ &\leq \left|\tilde{x}(\tau) - \sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,ds}\right| + \left|\sqrt{1 - \frac{d}{2}\int_0^\tau g(s)\,ds} - x(\tau)\right| \\ &\leq \frac{5.8283}{d} + \frac{33.1539}{d} = \frac{38.9822}{d}\end{aligned}$$

□

Corollary H.18. *For $d \geq 100{,}000$, we have $x(\tau) \geq 0.4567,\ \forall \tau \in [0,1]$.*

Proof. Note that $\tilde{x}(\tau) = \sqrt{1 - \frac{1}{\ln 4} + 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)}$ is decreasing for $\tau \in [0,1]$:

$$\tilde{x}'(\tau) = \frac{-4^{-\tau}(2 - \tau \ln 4)}{2\sqrt{1 - \frac{1}{\ln 4} + 4^{-\tau}\left(\frac{1}{\ln 4} - \tau\right)}} \leq 0$$

So it is lowest at $\tau = 1$. Combined with $|x(\tau) - \tilde{x}(\tau)| \leq \frac{38.9822}{d}$, we get:

$$\begin{aligned}x &\geq \sqrt{1 - \frac{1}{\ln 4} + \frac{1}{4}\left(\frac{1}{\ln 4} - 1\right)} - \frac{38.9822}{d} \\ &\geq 0.4571 - \frac{38.9822}{d} \\ [d \geq 100{,}000] &\geq 0.4571 - \frac{38.9822}{100{,}000} \geq 0.4567\end{aligned}$$

□

Proposition H.19. *For $d \geq 100{,}000$, $L \triangleq \max_{x,\tau\in[0,1]} \left|\frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x)\right| \leq 4.7955$.*

Proof. L, the Lipschitz constant of f, is given by

$$L \triangleq \max_{x,\tau\in[0,1]} \left|\frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x)\right| .$$

We have:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau + \frac{1}{d}\right)} - x}{2}\right] = \frac{d}{2}\left(\frac{x}{\sqrt{x^2 - 4\beta\left(\tau + \frac{1}{d}\right)}} - 1\right) .$$

Assume that $x \geq x_{\min}$, $\tau \in [0, 1]$. from $d \geq 3$, $\tau + \frac{1}{d} \leq \frac{1}{\ln 2}$. From Claim H.9, we get $\beta\left(\tau + \frac{1}{d}\right) \geq 0$.This means that $\frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x) \geq 0$. So

$$L = \max_{x,\tau\in[0,1]} \frac{\mathrm{d}}{\mathrm{d}x} f(\tau, x) .$$

For any fixed x, the maximum $\beta\left(\tau + \frac{1}{d}\right)$ will maximize L. From Claim H.9,we know that β is decreasing with τ, so to maximize L, $\tau = 0$. To maximize $\frac{d}{2}\left(\frac{x}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}} - 1\right)$, note that this function is decreasing with respect to x:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{d}{2}\left(\frac{x}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}} - 1\right)\right] &= \frac{d}{2}\left(\frac{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)} - x\frac{x}{\sqrt{x^2 - 4\beta\left(\tau + \frac{1}{d}\right)}}}{x^2 - 4\beta\left(\frac{1}{d}\right)}\right) \\ &= \frac{d}{2}\left(\frac{x^2 - 4\beta\left(\frac{1}{d}\right) - x^2}{\left(x^2 - 4\beta\left(\frac{1}{d}\right)\right)^{\frac{3}{2}}}\right) = \frac{d}{2}\left(\frac{-4\beta\left(\frac{1}{d}\right)}{\left(x^2 - 4\beta\left(\frac{1}{d}\right)\right)^{\frac{3}{2}}}\right) \leq 0\end{aligned}$$

So the optimal x, is $x_{\min}$. So

$$L = \frac{d}{2}\left(\frac{x_{\min}}{\sqrt{x_{\min}^2 - 4\beta\left(\frac{1}{d}\right)}} - 1\right)$$

Now

$$4\beta\left(\frac{1}{d}\right) = 4\frac{\left((1-1)\,2^{-1/d} - (1-2)\right)2^{(5-2)/d}}{d} = 4\frac{2^{3/d}}{d}$$

And applying Corollary H.18 we get:

$$\begin{aligned}L &\leq \frac{d}{2}\left(\frac{0.4567}{\sqrt{0.4567^2 - 4\frac{2^{3/d}}{d}}} - 1\right) \\ [d \geq 100{,}000] &\leq \frac{d}{2}\left(\frac{0.4567}{\sqrt{0.4567^2 - 4\frac{2^{3/100000}}{d}}} - 1\right) \leq \frac{d}{2}\left(\frac{0.4567}{\sqrt{0.4567^2 - \frac{4.0001}{d}}} - 1\right) \\ &\leq \frac{d}{2}\left(\frac{1}{\sqrt{1 - \frac{19.1783}{d}}} - 1\right) = \frac{d}{2}\left(\frac{1}{\sqrt{1 - \frac{19.1783}{d}}} - 1\right)\frac{\frac{1}{\sqrt{1-\frac{19.1783}{d}}} + 1}{\frac{1}{\sqrt{1-\frac{19.1783}{d}}} + 1} \\ &= \frac{d}{2}\left(\frac{\frac{1}{1-\frac{19.1783}{d}} - 1}{\frac{1}{\sqrt{1-\frac{19.1783}{d}}} + 1}\right) \leq \frac{d}{2}\left(\frac{\frac{\frac{19.1783}{d}}{1-\frac{19.1783}{d}}}{1+1}\right) = \frac{19.1783}{4}\frac{1}{1 - \frac{19.1783}{d}} \\ &\leq \frac{19.1783}{4}\frac{1}{1 - \frac{19.1783}{100000}} \leq 4.7955\end{aligned}$$

□

Proposition H.20. *From $d \geq 100{,}000$, $M \triangleq \max_{\tau\in[0,1]} \left|\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} x(\tau)\right| \leq 10.5027$*

Proof. M is defined as an upper bound on the second derivative (absolute value) of $x(\tau)$ in the relevant interval:

$$M \triangleq \max_{\tau\in[0,1]} \left|\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} x(\tau)\right| = \max_{\tau\in[0,1]} \left|\frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau, x(\tau))\right| .$$

We have:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau, x(\tau)) \\
&= \frac{\partial}{\partial\tau}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\right] + x'(\tau)\frac{\partial}{\partial x}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\right] \\
&= \frac{\partial}{\partial\tau}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\right] + f(\tau, x(\tau))\frac{\partial}{\partial x}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\right] .
\end{aligned}$$

For the first term:

$$\begin{aligned}
\frac{\partial}{\partial\tau}\left[d\frac{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)} - x}{2}\right] &= \frac{d}{2}\left(\frac{-4}{2\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)}}\right)\frac{\partial}{\partial\tau}\beta\left(\tau+\frac{1}{d}\right) \\
&= -\frac{d}{\sqrt{x^2 - 4\beta\left(\tau+\frac{1}{d}\right)}}\frac{\partial}{\partial\tau}\beta\left(\tau+\frac{1}{d}\right)
\end{aligned}$$

From Claim H.9 we know β is positive and decreasing, so $\frac{d}{\sqrt{x^2-4\beta\left(\tau+\frac{1}{d}\right)}}$ is maximized at $\tau = 0$; In addition $\frac{\partial}{\partial\tau}\beta\left(\tau+\frac{1}{d}\right) \leq 0$, and thus the entire expression is non negative. We know β is convex, so the absolute value of the negative $\frac{\partial}{\partial\tau}\beta\left(\tau+\frac{1}{d}\right)$ is also maximized at $\tau = 0$. All in all, the entire expression is maximized at $\tau = 0$, and is bounded by:

$$\begin{aligned}
0 \leq &-\frac{d}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}}\frac{\partial}{\partial\tau}\beta\left(\tau+\frac{1}{d}\right)\Big|_{\tau=0} \\
&= -\frac{d}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}}\frac{1}{d}\Big(d\left(2^{-1/d}-1\right)2^{\frac{5-2d(1/d)}{d}} \\
&\qquad\qquad -2\ln 2\cdot 2^{\frac{5-2d(1/d)}{d}}\left((d(1/d)-1)2^{-1/d} - (d(1/d)-2)\right)\Big) \\
&= \frac{2^{3/d}}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}}\left(2\ln 2 + d\left(1-2^{-1/d}\right)\right) \\
&= \frac{2^{2/d}}{\sqrt{x^2 - 4\beta\left(\frac{1}{d}\right)}}\left(2^{1/d}\cdot 2\ln 2 - d\left(2^{1/d}-1\right)\right) \\
[\text{H.7}] \leq &\frac{2^{2/d}}{\sqrt{x^2 - 4\frac{2^{3/d}}{d}}}\left(2^{1/d}\cdot 2\ln 2 - \ln 2\right) \\
&\leq \frac{2^{2/100000}}{\sqrt{x^2 - 4\frac{2^{3/100000}}{100000}}}\left(2^{1/100000}\cdot 2\ln 2 - \ln 2\right) \\
&\leq \frac{0.6932}{\sqrt{x^2 - 4.1\cdot 10^{-5}}} \\
[\text{H.18}] \leq &\frac{0.6932}{\sqrt{0.4567^2 - 4.1\cdot 10^{-5}}} \leq 1.518
\end{aligned}$$

Now from Proposition H.19, we have

$$0 \leq \frac{\mathrm{d}}{\mathrm{d}x}\left[d\frac{\sqrt{x^2-4\beta\left(\tau+\frac{1}{d}\right)}-x}{2}\right] \leq L = 4.7955$$

Now we need to bound $f(\tau, x(\tau))$.

$$f(\tau,x) = d\frac{\sqrt{x^2-4\beta\left(\tau+\frac{1}{d}\right)}-x}{2}$$

This is always negative.

From Claim H.9 we know β is positive and decreasing, so its maximum, which minimizes this and thus maximizes its absolute value, is received at $\tau = 0$.

We further know $f(0,x)$ is increasing with x (see the beginning of the proof for Proposition H.19), so its absolute value decreases with x.

So we have:

$$\begin{aligned}
0 \geq d\frac{\sqrt{x^2-4\beta\left(\tau+\frac{1}{d}\right)}-x}{2} &\geq d\frac{\sqrt{x^2-4\beta\left(\frac{1}{d}\right)}-x}{2} \\
[d \geq 100{,}000 \Rightarrow x \geq 0.4567] &\geq d\frac{\sqrt{0.4567^2-4\frac{2^{3/d}}{d}}-0.4567}{2} \\
&\geq d\frac{\sqrt{0.4567^2-4\frac{2^{3/100{,}000}}{d}}-0.4567}{2} \\
&\geq \frac{0.4567}{2}\frac{\sqrt{1-\frac{19.1782}{d}}-1}{\frac{1}{d}} \\
&= -\frac{0.4567}{2}\cdot 19.1782\frac{1-\sqrt{1-\frac{19.1782}{d}}}{\frac{19.1782}{d}} \\
[\text{H.12}] &\geq -\frac{0.4567}{2}\cdot 19.1782\left(\frac{1}{2}+\frac{19.1782}{2d}\right) \\
[d \geq 100{,}000] &\geq -\frac{0.4567}{2}\cdot 19.1782\left(\frac{1}{2}+\frac{19.1782}{2\cdot 100000}\right) \geq -2.1901
\end{aligned}$$

We get

$$-2.1901 \leq f(\tau,x) \leq 0$$

So

$$\frac{\mathrm{d}}{\mathrm{d}\tau}f(\tau,x(\tau)) \in [0,1.518] + [-2.1901,0]\cdot[0,4.7955] \subseteq [-10.5027,1.518]$$

So

$$M \leq 10.5027\,.$$

□