Accelerating SGD for Highly Ill-Conditioned Huge-Scale Online Matrix Completion

Gavin Zhang
University of Illinois at Urbana–Champaign
jialun2@illinois.edu

Hong-Ming Chiu
University of Illinois at Urbana–Champaign
hmchiu2@illinois.edu

Richard Y. Zhang
University of Illinois at Urbana–Champaign
ryz@illinois.edu

Abstract

The matrix completion problem seeks to recover a $d \times d$ ground truth matrix of low rank $r \ll d$ from observations of its individual elements. Real-world matrix completion is often a huge-scale optimization problem, with d so large that even the simplest full-dimension vector operations with $O(d)$ time complexity become prohibitively expensive. Stochastic gradient descent (SGD) is one of the few algorithms capable of solving matrix completion on a huge scale, and can also naturally handle streaming data over an evolving ground truth. Unfortunately, SGD experiences a dramatic slow-down when the underlying ground truth is ill-conditioned; it requires at least $O(\kappa \log(1/\epsilon))$ iterations to get ϵ-close to ground truth matrix with condition number κ. In this paper, we propose a preconditioned version of SGD that preserves all the favorable practical qualities of SGD for huge-scale online optimization while also making it agnostic to κ. For a symmetric ground truth and the Root Mean Square Error (RMSE) loss, we prove that the preconditioned SGD converges to ϵ-accuracy in $O(\log(1/\epsilon))$ iterations, with a rapid linear convergence rate as if the ground truth were perfectly conditioned with $\kappa = 1$. In our numerical experiments, we observe a similar acceleration for ill-conditioned matrix completion under the root mean square error (RMSE) loss, Euclidean distance matrix (EDM) completion under pairwise square loss, and collaborative filtering under the Bayesian Personalized Ranking (BPR) loss.

1 Introduction

The matrix completion problem seeks to recover an underlying $d \times d$ ground truth matrix M of low rank $r \ll d$ from observations of its individual matrix elements $M_{i,j}$. The problem appears most prominently in the context of collaborative filtering and recommendation system, but also numerous other applications. In this paper, we focus on the *symmetric* and *positive semidefinite* variant of the problem, in which the underlying matrix M can be factored as $M = ZZ^T$ where the factor matrix Z is $d \times r$, though that our methods have natural extensions to the nonsymmetric case. We note that the symmetric positive semidefinite variant is actually far more common in collaborative filtering, due to the prevalence of *item-item* models, which enjoy better data (most platforms contain several orders of magnitude more users than items) and more stable recommendations (the similarity between items tends to change slowly over time) than user-user and user-item models.

For the full-scale, online instances of matrix completion that arise in real-world collaborative filtering, stochastic gradient descent or SGD is the only viable algorithm for learning the underlying matrix M. The basic idea is to formulate a candidate matrix of the form XX^T with respect to a

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

learned factor matrix $X \in \mathbb{R}^{d\times r}$, and to minimize a cost function of the form $\phi(XX^T - M)$. Earlier work used the root mean square error (RMSE) loss $\|XX^T - M\|_F^2 = \sum_{i,j}(XX^T - M)_{i,j}^2$, though later work have focused on pairwise losses like the BPR [1] that optimize for ordering and therefore give better recommendations. For the RMSE loss, the corresponding SGD iterations with (rescaled) learning rate $\alpha > 0$ reads

$$x_{i,+} = x_i - \alpha \cdot \left(x_i^T x_j - M_{ij}\right) x_j, \quad x_{j,+} = x_j - \alpha \cdot \left(x_i^T x_j - M_{ij}\right) x_i, \tag{1}$$

where M_{ij} is the sampled (i,j)-th element of the ground truth matrix M, and x_i, x_j and $x_{i,+}, x_{j,+}$ denote the i-th and j-th rows of the current iterate X_t and new iterate X_{t+1}. Pairwise losses like the BPR can be shown to have a similar update equation over three rows of X [1]. Given that only two or three rows of X are accessed and updated at any time, SGD is readily accessible to massive parallelization and distributed computing. For very large values of d, the update equation (1) can be run by multiple workers in parallel without locks, with vanishing probability of collision [2]. The blocks of X that are more frequently accessed together can be stored on the same node in a distributed memory system.

Unfortunately, the convergence rate of SGD can sometimes be extremely slow. One possible explanation, as many recent authors have pointed out [3–6], is that matrix factorization models are very sensitive to ill-conditioning of the ground truth matrix M. The number of SGD iterations grows at least linearly the condition number κ, which here is defined as the ratio between the largest and the r-th largest singular values of M. Ill-conditioning causes particular concern because most real-world data are ill-conditioned. In one widely cited study [7], it was found that the dominant singular value accounts for only ≈80% prediction accuracy, with diversity of individual preferences making up the remainder ill-conditioned singular values. Cloninger et al. [8] notes that there are certain applications of matrix completion that have condition numbers as high as $\kappa = 10^{15}$.

This paper is inspired by a recent full-batch gradient method called ScaledGD [4, 9] and a closely related algorithm PrecGD [5] in which gradient descent is made immune to ill-conditioning in the ground truth by right-rescaling the full-batch gradient by the matrix $(X^TX)^{-1}$. Applying this same strategy to the SGD update equation (1) yields the row-wise updates

$$x_{i,+} = x_i - \alpha \cdot \left(x_i^T x_j - M_{ij}\right) P x_j, \quad x_{j,+} = x_j - \alpha \cdot \left(x_i^T x_j - M_{ij}\right) P x_i, \tag{2a}$$

in which we *precompute and cache* the preconditioner $P = (X^TX)^{-1}$ ahead of time[1], and update it after the iteration as

$$P_+ = (P^{-1} + x_{i,+}x_{i,+}^T + x_{j,+}x_{j,+}^T - x_i x_i^T - x_j x_j^T)^{-1} \tag{2b}$$

by making four calls to the Sherman–Morrison rank-1 update formula

$$(P^{-1} + uu^T)^{-1} = P - \frac{Puu^TP}{1 + u^TPu}, \qquad (P^{-1} - uu^T)^{-1} = P + \frac{Puu^TP}{1 - u^TPu}.$$

This way, the rescaled update equations use just $O(r^2)$ arithmetic operations, which for modest values of r is only marginally more than the $O(r)$ cost of the unscaled update equations (1). Indeed, the nearest-neighbor algorithms inside most collaborative filters have exponential complexity with respect to the latent dimensionality r, and so are often implemented with r small enough for (1) and (2) to have essentially the same runtime. Here, we observe that the rescaled update equations (2) preserve essentially all of the practical advantages of SGD for huge-scale, online optimization: it can also be run by multiple workers in parallel without locks, and it can also be easily implemented over distributed memory. The only minor difference is that separate copies of P should be maintained by each worker, and resynchronized once differences grow large.

Contributions In this paper, we provide a rigorous proof that the rescaled update equations (1), which we name ScaledSGD, become *immune* to the effects of ill-conditioning in the underlying ground truth matrix. For symmetric matrix completion under the root mean squared error (RMSE) loss function, regular SGD is known to have an iteration count of $O(\kappa^4 \cdot dr \log(d/\epsilon))$ within a local neighborhood of the ground truth [10]. This figure is optimal in the dimension d, the rank

[1] For an initialization, if the d rows of X_0 are selected from the unit Gaussian as in $x_1, \ldots, x_d \sim \mathcal{N}(0, \sigma^2 I_r)$, then we can simply set $P_0 = \sigma^2 I$ without incurring the $O(d)$ time needed in explicitly computing $P_0 = (X_0^T X_0)^{-1}$.

r, and the final accuracy ϵ, but suboptimal by four exponents with respect to condition number κ. In contrast, we prove for the same setting that ScaledSGD attains an optimal convergence rate, converging to ϵ-accuracy in $O(dr\log(d/\epsilon))$ iterations for all values of the condition number κ. In fact, our theoretical result predicts that ScaledSGD converges as if the ground truth matrix is perfectly conditioned, with a condition number of $\kappa = 1$.

At first sight, it appears quite natural that applying the ScaledGD preconditioner to SGD should result in accelerated convergence. However, the core challenge of stochastic algorithms like SGD is that each iteration can have substantial *variance* that "drown out" the expected progress made in the iteration. In the case of ScaledSGD, a rough analysis would suggest that the highly ill-conditioned preconditioner should improve convergence in expectation, but at the cost of dramatically worsening the variance.

Surprisingly, we find in this paper that the specific scaling $(X^TX)^{-1}$ used in ScaledSGD not only does not worsen the variance, but in fact *improves* it. Our key insight and main theoretical contribution is Lemma 4, which shows that the same mechanism that allows ScaledGD to converge faster (compared to regular GD) also allows ScaledSGD to enjoy reduced variance (compared to regular SGD). In fact, it is this effect of variance reduction that is responsible for most (κ^3 out of κ^4) of our improvement over the previous state-of-the-art. It turns out that a careful choice of preconditioner can be used as a mechanism for *variance reduction*, while at the same time also fulfilling its usual, classical purpose, which is to accelerate convergence in expectation.

Related work Earlier work on matrix completion analyzed a convex relaxation of the original problem, showing that nuclear norm minimization can recover the ground truth from a few incoherent measurements [11–15]. This approach enjoys a near optimal sample complexity but incurs an $O(d^3)$ per-iteration computational cost, which is prohibitive for a even moderately large d. More recent work has focused more on a nonconvex formulation based on Burer and Monteiro [16], which factors the optimization variable as $M = XX^T$ where $X \in \mathbb{R}^{d\times r}$ and applies a local search method such as alternating-minimization [17–20], projected gradient descent [21, 22] and regular gradient descent [23–26]. A separate line of work [27–35] focused on global properties of nonconvex matrix recovery problems, showing that the problem has no spurious local minima if sampling operator satisfies certain regularity conditions such as incoherence or restricted isometry.

The convergence rate of SGD has been well-studied for general classes of functions [36–39]. For matrix completion in particular, Jin et al. [10] proved that SGD converges towards an ϵ-accurate solution in $O(d\kappa^4\log(1/\epsilon))$ iterations where κ is the condition number of M. Unfortunately, this quartic dependence on κ makes SGD extremely slow and impractical for huge-scale applications.

This dramatic slow down of gradient descent and its variants caused by ill-conditioning has become well-known in recent years. Several recent papers have proposed full-batch algorithms to overcome this issue [9, 40, 41], but these methods cannot be used in the huge-scale optimization setting where d is so large that even full-vector operations with $O(d)$ time complexity are too expensive. As a deterministic full-batch method, ScaledGD [9] requires a projection onto the set of incoherent matrices at every iteration in order to maintain rapid convergence. Instead our key finding here is that the stochasticity of SGD alone is enough to keep the iterates as incoherent as the ground truth, which allows for rapid progress to be made. The second-order method proposed in [41] costs at least $O(d)$ per-iteration and has no straightforward stochastic analog. PrecGD [5] only applies to matrices that satisfies matrices satisfying the restricted isometry property, which does not hold for matrix completion.

2 Background: Linear convergence of SGD

In our theoretical analysis, we restrict our attention to symmetric matrix completion under the root mean squared error (RMSE) loss function. Our goal is to solve the following nonconvex optimization

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) \stackrel{\text{def}}{=} \|XX^T - ZZ^T\|_F^2 \qquad \text{where } Z = [z_1, z_2, \ldots, z_n]^T \in \mathbb{R}^{d\times r} \tag{3}$$

in which we assume that the $d \times d$ ground truth $ZZ^T \succeq 0$ matrix is exactly rank-r, with a finite *condition number*

$$\kappa \stackrel{\text{def}}{=} \lambda_{\max}(ZZ^T)/\lambda_r(ZZ^T) = \lambda_{\max}(Z^TZ)/\lambda_{\min}(Z^TZ) < \infty. \tag{4}$$

In order to be able to reconstruct ZZ^T from a small number of measurements, we will also need to assume that the ground truth has small *coherence* [42]

$$\mu \stackrel{\text{def}}{=} \frac{d}{r} \cdot \max_{1 \le i \le d} \|e_i^T Z (Z^T Z)^{-1/2}\|^2. \tag{5}$$

Recall that μ takes on a value from 1 to d/r, with the smallest achieved by dense, orthonormal choices of Z whose rows all have magnitudes of $1/\sqrt{d}$, and the largest achieved by a ground truth ZZ^T containing a single nonzero element. Assuming incoherence $\mu = O(1)$ with respect to d, it is a well-known result that all d^2 matrix elements of ZZ^T can be perfectly reconstructed from just $O(dr \log d)$ random samples of its matrix elements [12, 43].

This paper considers solving (3) in the huge-scale, online optimization setting, in which individual matrix elements of the ground truth $(ZZ^T)_{i,j} = z_i^T z_j$ are revealed one-at-a-time, uniformly at random with replacement, and that a current iterate X is continuously updated to streaming data. We note that this is a reasonably accurate model for how recommendation engines are tuned to user preferences in practice, although the uniformity of random sampling is admittedly an assumption made to ease theoretical analysis. Define the stochastic gradient operator as

$$SG(X) = 2d^2 \cdot (x_i^T x_j - z_i^T z_j)(e_i x_j^T + e_j x_i^T) \quad \text{where } (i,j) \sim \text{Unif}([d] \times [d]),$$

where $x_i, x_j \in \mathbb{R}^r$ are the i-th and j-th rows of X, and the scaling d^2 is chosen that, over the randomness of the sampled index (i,j), we have exactly $\mathbb{E}[SG(X)] = \nabla f(X)$. Then, the classical online SGD algorithm can be written as

$$X_{t+1} = X_t - \alpha SG(X_t) \qquad \text{where } \alpha > 0. \tag{SGD}$$

Here, we observe that a single iteration of SGD coincides with full-batch gradient descent in expectation, as in $\mathbb{E}[X_{t+1}|X_t] = X_t - \alpha \nabla f(X_t)$. Therefore, assuming that bounded deviations and bounded variances, it follows from standard arguments that the behavior of many iterations of SGD should concentrate about that of full-batch gradient descent $X_{t+1} = X_t - \alpha \nabla f(X_t)$.

Within a region sufficiently close to the ground truth, full-batch gradient descent is well-known to converge at a linear rate to the ground truth [23, 44]. Within this same region, Jin et al. [10] proved that SGD also converges linearly. For an incoherent ground truth with $\mu = O(1)$, they proved that SGD with an aggressive choice of step-size is able to recover the ground truth to ϵ-accuracy $O(\kappa^4 dr \log(d/\epsilon))$ iterations, with each iteration costing $O(r)$ arithmetic operations and selecting 1 random sample. This iteration count is optimal with respect to d, r, and $1/\epsilon$, although its dependence on κ is a cubic factor (i.e. a factor of κ^3) worse than full-batch gradient descent's figure of $O(\kappa \log(1/\epsilon))$, which is itself already quite bad, given that κ in practice can readily take on values of 10^3 to 10^6.

Theorem 1 (Jin, Kakade, and Netrapalli [10]). *For $Z \in \mathbb{R}^{d \times r}$ with $\sigma_{\max}(Z) = 1$ and $f(X) = \|XX^T - ZZ^T\|_F^2$ and $h_i(X) = \|e_i^T X\|^2$, define the following*

$$f_{\max} \stackrel{\text{def}}{=} \left(\frac{1}{10\kappa}\right)^2, \qquad h_{\max} \stackrel{\text{def}}{=} 20 \cdot \kappa^2 \cdot \frac{\mu r}{d}.$$

For an initial point $X_0 \in \mathbb{R}^{d \times r}$ that satisfies $f(X_0) \le \frac{1}{2} f_{\max}$ and $\max_i h_i(X_0) \le \frac{1}{2} h_{\max}$, there exists some constant c such that for any learning rate $\alpha < c \cdot (\kappa \cdot h_{\max} \cdot d^2 \log d)^{-1}$, with probability at least $1 - T/d^{10}$, we will have for all $t \le T$ iterations of SGD that

$$f(X_t) \le \left(1 - \frac{\alpha}{2 \cdot \kappa}\right)^t \cdot f_{\max}, \qquad \max_i h_i(X_t) \le h_{\max}.$$

The reason for Theorem 1's additional κ^3 dependence beyond full-batch gradient descent is due to its need to maintain *incoherence* in its iterates. Using standard techniques on martingale concentration, one can readily show that SGD replicates a single iteration of full-batch gradient descent over an epoch of d^2 iterations. This results in an iteration count $O(\kappa \cdot d^2 \log(1/\epsilon))$ with an optimal dependence on κ, but the entire matrix is already fully observed after collecting d^2 samples. Instead, Jin et al. [10] noted that the *variance* of SGD iterations is controlled by the step-size α times the maximum coherence $\mu_X = \frac{d}{r} \cdot \max_{i,t} \|e_i^T X_t\|^2$ over the iterates $X_t, X_{t-1}, \ldots, X_0$. If the iterates can be kept incoherent with $\mu_X = O(1)$, then SGD with a more aggressive step-size will reproduce an iteration of full-batch gradient descent after an epoch of just $O(dr \log d)$ iterations.

The main finding in Jin et al. [10]'s proof of Theorem 1 is that the stochasticity of SGD is enough to keep the iterates incoherent. This contrasts with full-batch methods at the time, which required an added regularizer [20, 30, 45] or an explicit projection step [9]. (As pointed out by a reviewer, it was later shown by Ma et al. [46] that full-batch gradient descent is also able to maintain incoherence without a regularizer nor a projection.) Unfortunately, maintaining incoherence requires shrinking the step-size by a factor of κ, and the actual value of μ_X that results is also a factor of κ^2 worse than the original coherence μ of the ground truth Z. The resulting iteration count $O(\kappa^4 \cdot dr \log(d/\epsilon))$ is made optimal with respect to d, r, and $1/\epsilon$, but only at the cost of worsening its the dependence on the condition number κ by another *three exponents*.

Finally, the quality of the initial point X_0 also has a dependence on the condition number κ. In order to guarantee linear convergence, Theorem 1 requires X_0 to lie in the neighborhood $\|X_0X_0^T - ZZ^T\|_F < \lambda_{\min}(Z^TZ) = O(\kappa^{-1})$. This dependence on κ is optimal, because full-batch gradient descent must lose its ability to converge linearly in the limit $\kappa \to \infty$ [5, 6]. However, the leading constant can be very pessmistic, because the theorem must formally exclude spurious critical points X_{spur} that have $\nabla f(X_{\text{spur}}) = 0$ but $f(X_{\text{spur}}) > 0$ in order to be provably correct. In practice, it is commonly observed that SGD converges globally, starting from an arbitrary, possibly random initialization [30], at a linear rate that is consistent with local convergence theorems like Theorem 1. It is now commonly argued that gradient methods can escape saddle points with high probability [47], and so their performance is primarily dictated by local convergence behavior [48, 49].

3 Proposed algorithm and main result

Inspired by a recent full-batch gradient method called ScaledGD [4, 9] and a closely related algorithm PrecGD [5], we proposed the following algorithm

$$X_{t+1} = X_t - \alpha SG(X_t)(X_t^TX_t)^{-1} \qquad \text{where } \alpha > 0. \tag{ScaledSGD}$$

As we mentioned in the introduction, the preconditioner $P = (X^TX)^{-1}$ can be precomputed and cached in a practical implementation, and afterwards efficiently updated using the Sherman–Morrison formula. The per-iteration cost of ScaledSGD is $O(r^2)$ arithmetic operations and 1 random sample, which for modest values of r is only marginally more than the cost of SGD.

Our main result in this paper is that, with a region sufficiently close to the ground truth, this simple rescaling allows ScaledSGD to converge linearly to ϵ-accuracy $O(dr\log(d/\epsilon))$ iterations, with no further dependence on the condition number κ. This iteration count is optimal with respect to d, r, $1/\epsilon$, and κ, and in fact matches SGD with a *perfectly conditioned* ground truth $\kappa = 1$. In our numerical experiments, we observe that ScaledSGD converges globally from a random initialization at the same rate as SGD as if $\kappa = 1$.

Theorem 2 (Main)**.** *For $Z \in \mathbb{R}^{d\times r}$ with $\sigma_{\max}(Z) = 1$ and $f(X) = \|XX^T - ZZ^T\|_F^2$ and $g_i(X) = e_i^TX(X^TX)^{-1}X^Te_i$, select a radius $\rho < 1/2$ and set*

$$f_{\max} \stackrel{\text{def}}{=} \left(\frac{\rho}{\kappa}\right)^2, \qquad g_{\max} \stackrel{\text{def}}{=} \frac{2^4}{(1-2\rho)^2} \cdot \frac{\mu r}{d}.$$

For an initial point $X_0 \in \mathbb{R}^{d\times r}$ that satifies $f(X_0) \le \frac{1}{2}f_{\max}$ and $\max_i g_i(X_0) \le \frac{1}{2}g_{\max}$, there exists some constant c such that for any learning rate $\alpha < c \cdot [(g_{\max} + \rho) \cdot d^2 \log d]^{-1}$, with probability at least $1 - T/d^{10}$, we will have for all $t \le T$ iterations of ScaledSGD that:

$$f(X_t) \le \left(1 - \frac{\alpha}{2}\right)^t \cdot f_{\max}, \qquad \max_i g_i(X_t) \le g_{\max}.$$

Theorem 2 eliminates all dependencies on the condition number κ in Theorem 1 except for the quality of the initial point, which we had already noted earlier as being optimal. Our main finding is that it is possible to maintain incoherence while making aggressive step-sizes towards a highly ill-conditioned ground truth ZZ^T. In fact, Theorem 2 says that, with high probability, the maximum coherence μ_X over of any iterate X_t will only be a *mild constant factor* of ≈ 16 times worse than the coherence μ of the ground truth ZZ^T. This is particularly surprising in view of the fact that every iteration of ScaledSGD involves inverting a potentially highly ill-conditioned matrix $(X^TX)^{-1}$. In

contrast, even without inverting matrices, Theorem 1 says that SGD is only able to keep μ_X within a factor of κ^2 of μ, and only by shrinking the step-size α by another factor of κ.

However, the price we pay for maintaining incoherence is that the quality of the initial point X_0 now gains a dependence on dimension d, in addition to the condition number κ. In order to guarantee fast linear convergence independent of κ, Theorem 2 requires X_0 to lie in the neighborhood $\|X_0X_0^T - ZZ^T\|_F < \mu r \lambda_{\min}(Z^TZ)/d = (\kappa d)^{-1}$, so that ρ can be set to be the same order of magnitude as $g_{\max}$. In essence, the "effective" condition number of the ground truth has been worsened by another factor of d. This shrinks the size of our local neighborhood by a factor of d, but has no impact on the convergence rate of the resulting iterations.

In the limit that $\kappa \to \infty$ and the search rank r becomes *overparameterized* with respect to the true rank $r^\star < r$ of ZZ^T, both full-batch gradient descent and SGD slows down to a *sublinear* convergence rate, in theory and in practice [5, 6]. While Theorem 2 is no longer applicable, we observe in our numerical experiments that ScaledSGD nevertheless maintains its fast linear convergence rate as if $\kappa = 1$. Following PrecGD [5], we believe that introducing a small identity perturbation to the scaling matrix of ScaledSGD, as in $(X^TX + \eta I)^{-1}$ for some $\eta \approx \sqrt{f(X)}$, should be enough to rigorously extend Theorem 2 to the overparameterized regime. We leave this extension as future work.

4 Key ideas for the proof

We begin by explaining the mechanism by which SGD slows down when converging towards an ill-conditioned ground truth. Recall that

$$\mathbb{E}[SG(X)] = \mathbb{E}[2d^2 \cdot (XX^T - ZZ^T)_{i,j} \cdot (e_ie_j^T + e_je_i^T)X] = 4(XX^T - ZZ^T)X = \nabla f(X).$$

As XX^T converges towards an ill-conditioned ground truth ZZ^T, the factor matrix X must become progressively ill-conditioned, with

$$\lambda_{\min}(X^TX) = \lambda_r(XX^T) \le \lambda_r(ZZ^T) + \|XX^T - ZZ^T\|_F \le \frac{1+\rho}{\kappa}.$$

Therefore, it is possible for components of the error vector $XX^T - ZZ^T$ to become "invisible" by aligning within the ill-conditioned subspaces of X. As SGD progresses towards the solution, these ill-conditioned subspaces of X become the slowest components of the error vector to converge to zero. On the other hand, the maximum step-size that can be taken is controlled by the most well-conditioned subspaces of X. A simple idea, therefore, is to rescale the ill-conditioned components of the gradient $\nabla f(X)$ in order to make the ill-conditioned subspaces of X more "visible".

More concretely, define the local norm of the gradient as $\|\nabla f(X)\|_X = \|\nabla f(X)(X^TX)^{1/2}\|_F$ and its corresponding dual norm as $\|\nabla f(X)\|_X^* = \|\nabla f(X)(X^TX)^{-1/2}\|_F$. It has long been known (see e.g. [23, 44]) that rescaling the gradient yields

$$\|\nabla f(X)\|_X^* \stackrel{\text{def}}{=} \|4(XX^T - ZZ^T)X(X^TX)^{-1/2}\|_F = 4\cos\theta \cdot \|XX^T - ZZ^T\|_F,$$

where θ is the angle between the error vector $XX^T - ZZ^T$ and the linear subspace $\{XY^T + YX^T : Y \in \mathbb{R}^{d\times r}\}$. This insight immediately suggests an iteration like $X_+ = X - \alpha\nabla f(X)(X^TX)^{-1}$. In fact, the gradients of f have some Lipschitz constant L, so

$$\begin{aligned} f(X_+) &\le f(X) - \alpha\langle \nabla f(X), \nabla f(X)(X^TX)^{-1}\rangle + \frac{L}{2}\alpha^2\|\nabla f(X)(X^TX)^{-1}\|_F^2, \\ &\le f(X) - \alpha(\|\nabla f(X)\|_X^*)^2 + \frac{L_X}{2}\alpha^2(\|\nabla f(X)\|_X^*)^2, \\ &\le \left[1 - \alpha\cdot 8\cos^2\theta\right] f(X) \qquad \text{for } \alpha \le 1/L_X. \end{aligned}$$

However, a naive analysis finds that $L_X = L/\lambda_{\min}(X^TX) \approx L\cdot\kappa$, and this causes the step-size to shrink by a factor of κ. The main motivating insight behind ScaledGD [4, 9] and later PrecGD [5] is that, with a finer analysis, it is possible to prove Lipschitz continuity under a local change of norm.

Lemma 3 (Function descent)**.** *Let $X, Z \in \mathbb{R}^{n\times r}$ satisfy $\|XX^T - ZZ^T\|_F \le \rho\lambda_{\min}(Z^TZ)$ where $\rho < 1/2$. Then, the function $f(X) = \|XX^T - ZZ^T\|_F^2$ satisfies*

$$f(X+V) \le f(X) + \langle \nabla f(X), V\rangle + \frac{L_X}{2}\|V\|_X^2, \qquad (\|\nabla f(X)\|_X^*)^2 \ge 13\cdot f(X)$$

for all $\|V\|_X \le C \cdot \sqrt{f(X)}$ *with* $L_X = 6 + 8C + 2C^2 = O(1 + C^2)$.

This same idea can be "stochastified" in a straightforward manner. Conditioning on the current iterate X, then the new iterate $X_+ = X - \alpha SG(X)(X^TX)^{-1}$ has expectation

$$\mathbb{E}[f(X_+)] \le f(X) - \alpha\langle \nabla f(X), \mathbb{E}[SG(X)(X^TX)^{-1}]\rangle + \alpha\frac{L_X}{2}\mathbb{E}[(\|SG(X)\|_X^*)^2].$$

The linear term evaluates as $\mathbb{E}[SG(X)(X^TX)^{-1}] = \nabla f(X)(X^TX)^{-1}$, while the quadratic term is

$$\mathbb{E}[(\|SG(X)\|_X^*)^2] \le \sum_{i,j} 4d^2 \cdot (XX^T - ZZ^T)_{i,j}^2 \cdot 4\max_i(\|e_i^TX\|_X^*)^2 = 16 \cdot f(X) \cdot \max_i g_i(X),$$

where $g_i(X) = e_i^TX(X^TX)^{-1}X^Te_i = (\|e_i^TX\|_X^*)^2$. Combined, we obtain geometric convergence

$$\mathbb{E}[f(X_+)] \le \left(1 - \alpha \cdot 8\cos^2\theta\right) f(X) \qquad \text{for } \alpha = O(g_{\max}^{-1} \cdot d^{-2}). \tag{6}$$

We see that the step-size depends crucially on the incoherence $g_i(X) \le g_{\max}$ of the current iterate. If the current iterate X is incoherent with $g_{\max} = O(1/d)$, then a step-size of $\alpha = O(1/d)$ is possible, resulting in convergence in $O(dr\log(d/\epsilon))$ iterations, which can be shown using standard martingale techniques [10]. But if the current iterate is $g_{\max} = O(1)$, then only a step-size of $\alpha = O(1/d^2)$ is possible, which forces us to compute d^2 iterations, thereby obviating the need to complete the matrix in the first place.

Therefore, in order for prove rapid linear convergence, we need to additionally show that with high probability, the coherence $g_k(X)$=$(\|e_k^TX\|_X^*)^2$ remains $O(1)$ throughout ScaledGD iterations. This is the most challenging part of our proof. Previous methods that applied a similar scaling to full-batch GD [9] required an explicit projection onto the set of incoherent matrices at each iteration. Applying a similar projection to ScaledSGD will take $O(d)$ time, which destroys the scalability of our method. On the other hand, Jin et al. [10] showed that the randomness in SGD is enough to keep the coherence of the iterates within a factor of κ^2 times worse than the coherence of the ground truth, and only by a step-size of at most $\alpha = O(\kappa^{-1})$.

Surprisingly, here we show that the randomness in ScaledSGD is enough to keep the coherence of the iterates with a *constant factor* of the coherence the ground truth, using a step-size with no dependence on κ. The following key lemma is the crucial insight of our proof. First, it says that function $g_k(X)$ satisfies a "descent lemma" with respect to the local norm $\|\cdot\|_X^*$. Second, and much more importantly, it says that descending $g_k(X)$ along the scaled gradient direction $\nabla f(X)(X^TX)^{-1}$ incurs a linear decrement $\frac{1-2\rho}{1-\rho}g_k(X)$ with no dependence of the condition number κ. This is in direct analogy to the function value decrement in (6), which has no dependence on κ, and in direct contrast to the proof of Jin et al. [10], which is only able to achieve a decrement of $(8/\kappa)g_k(X)$ due to the lack of rescaling by $(X^TX)^{-1}$.

Lemma 4 (Coherence descent)**.** *Let* $g_k(X) = e_k^TX(X^TX)^{-1}X^Te_k$. *Under the same conditions as Lemma 3, we have*

$$g_k(X+V) \le g_k(X) + \langle V, \nabla g_k(X)\rangle + \frac{5(\|V\|_X^*)^2}{1 - 2\|V\|_X^*},$$

$$\langle \nabla g_k(X), \nabla f(X)(X^TX)^{-1}\rangle \ge \left[\frac{1-2\rho}{1-\rho}g_k(X) - \frac{1}{1-\rho}\sqrt{g_k(X)g_k(Z)}\right].$$

Conditioning on X, we have for the search direction $V = SG(X)(X^TX)^{-1}$ and $X_+ = X + V$

$$\begin{aligned}
\mathbb{E}[g_k(X_+)] &\le g_k(X) - \alpha\langle \nabla g_k(X), \mathbb{E}[V]\rangle + \alpha^2 \cdot \mathbb{E}\left[\frac{(\|V\|_X^*)^2}{1 - 2\|V\|_X^*}\right] \\
&\le \left(1 - \frac{1-2\rho}{1-\rho}\alpha\right) g_k(X) + \alpha \cdot \frac{1}{1-\rho} \cdot \sqrt{g_k(X)g_k(Z)} + \alpha^2 \cdot \frac{\mathbb{E}\left[(\|V\|_X^*)^2\right]}{1 - 2\|V\|_X^*} \\
&\le \left(1 - \frac{1-2\rho}{1-\rho}\alpha\right) g_k(X) + \alpha \cdot \frac{\sqrt{\mu/g_{\max}}}{1-\rho} \cdot g_{\max} + \alpha^2 \cdot \frac{O(d^2 \cdot g_{\max} \cdot \rho^2)}{1 - O(g_{\max}^{1/2} \cdot \rho)} \\
&\le (1 - \zeta\alpha)\, g_k(X) + \alpha \cdot \frac{\zeta}{2} g_{\max} \qquad \text{for } \alpha = O(\rho^{-1}d^{-2}).
\end{aligned} \tag{7}$$

It then follows that $g_k(X_+)$ converges geometrically towards $\frac{1}{2}g_{\max}$ in expectation, with a convergence rate $(1-\zeta\alpha)$ that is independent of the condition number κ:

$$\mathbb{E}[g_k(X_+) - \frac{1}{2}g_{\max}] \leq \left[(1-\zeta\alpha)\, g_k(X) + \alpha \cdot \frac{\zeta}{2} g_{\max}\right] - \frac{1}{2}g_{\max} \leq (1-\zeta\alpha)\left[g_k(X) - \frac{1}{2}g_{\max}\right].$$

The proof of Theorem 2 then follows from standard techniques, by making the two decrement conditions (6) and (7) into supermartingales and applying a standard concentration inequality. We defer the rigorous proof to appendix E.

5 Experimental validation

In this section we compare the practical performance of ScaledSGD and SGD for the RMSE loss function in Theorem 2 and two real-world loss functions: the pairwise RMSE loss used to complete Euclidean Distance Matrices (EDM) in wireless communication networks; and the Bayesian Personalized Ranking (BRP) loss used to generate personalized item recommendation in collaborative filtering. In each case, ScaledSGD remains highly efficient since it only updates two or three rows at a time, and the preconditioner P can be computed through low-rank updates, for a per-iteration cost of $O(r^2)$. All of our experiments use random Gaussian initializations and an initial $P = \sigma^2 I$. To be able to accurately measure and report the effects of ill-conditioning on ScaledSGD and SGD, we focus on small-scale synthetic datasets in the first two experiments, for which the ground truth is explicitly known, and where the condition numbers can be finely controlled. In addition, to gauge the scalability of ScaledSGD on large-scale real-world datasets, in the third experiment, we apply ScaledSGD to generate personalized item recommendation using MovieLens25M dataset [50], for which the underlying item-item matrix has more than 62,000 items and 100 million pairwise samples are used during training. (Due to space constraints, we defer the details on the experimental setup, mathematical formulations, and the actual update equations to Appendix A.) The code for all experiments are available at `https://github.com/Hong-Ming/ScaledSGD`.

Matrix completion with RMSE loss. The problem formulation is discussed in Section 3. Figure 1 plots the error $f(X) = \|XX^T - M\|_F^2$ as the number of epochs increases. As expected, in the well-conditioned case, both ScaledSGD and SGD converges to machine error at roughly the same linear rate. However, in the ill-conditioned case, SGD slows down significantly while ScaledSGD converges at almost exactly the same rate as in the well-conditioned case.

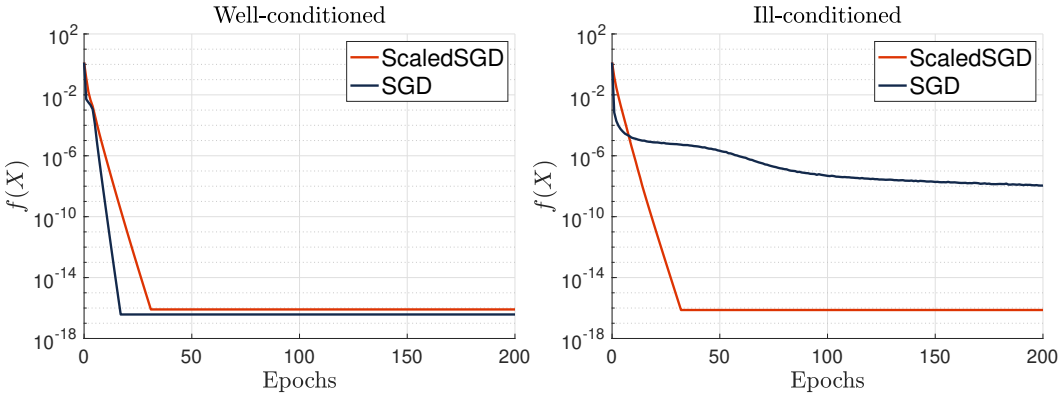

Figure 1: **Matrix Completion with RMSE loss.** We compare the convergence rate of ScaledSGD and SGD for a well-conditioned and ill-conditioned ground truth matrix of size 30×30 and rank 3. (**Left**) Well-conditoned M, $\kappa(M) = 1$. Step-size $\alpha = 0.3$. Both ScaledGD and SGD converges quickly to the ground truth. (**Right**) Ill-conditoned M, $\kappa(M) = 10^4$. Step-size $\alpha = 0.3$. SGD stagnates while ScaledSGD retains the same convergence rate as the well-conditioned case.

Euclidean distance matrix (EDM) completion. The Euclidean distance matrix (EDM) is a matrix of pairwise distance between d points in Euclidean space [51]. In applications such as wireless sensor networks, estimation of unknown distances, i.e., completing the EDM is often required. We

emphasize that this loss function is a pairwise loss, meaning that each measurement indexes multiple elements of the ground truth matrix.

To demonstrate the efficacy of ScaledSGD, we conduct two experiments where D is well-conditioned and ill-conditioned respectively: **Experiment 1.** We uniformly sample 30 points in a cube center at origin with side length 2, and use them to compute the ground truth EDM D. In this case, each row $x_i \in \mathbb{R}^3$ corresponds to the coordinates of the i-th sample. The corresponding matrix $X \in \mathbb{R}^{30\times 3}$ is well-conditioned because of the uniform sampling. **Experiment 2.** The ground truth EDM is generated with 25 samples lie in the same cube in experiment 1, and 5 samples lie far away from the the cube. These five outliers make the corresponding X become ill-conditioned.

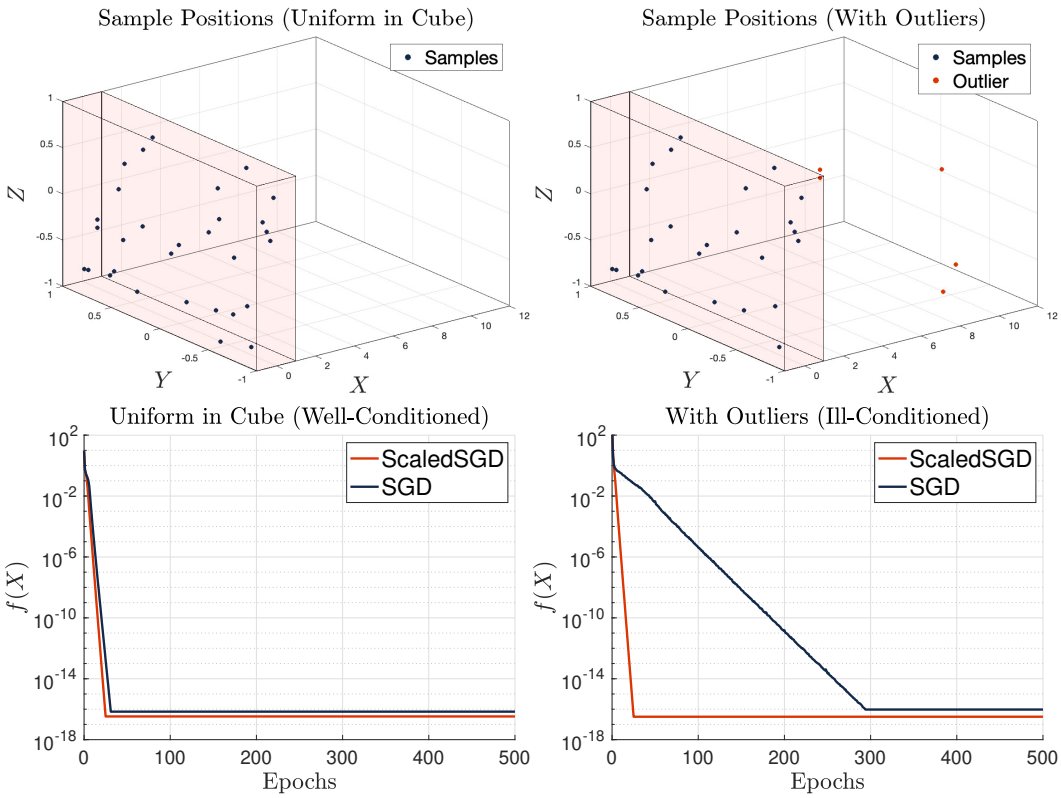

Figure 2: **Euclidean distance matrix (EDM) completion.** We compare the convergence rate of ScaledSGD and SGD for EDM completion for two set of samples. (**Upper right**) 30 samples are uniformly distributed in the pink cube center at origin. (**Upper left**) 25 samples (in blue) are uniformly distributed in the cube, 5 outlier samples (in orange) are outside of the cube. (**Lower left**) Sample uniformly in cube. (**Lower right**) Sample with outliers.

Item-item collaborative filtering (CF). In the task of item-item collaborative filtering (CF), the ground truth M is an $d \times d$ matrix where d is the number of items we wish to rank and the i, j-th of M is a similarity measure between the items. Our goal is to learn a low-rank matrix that preserves the ranking of similarity between the items. For instance, given a pairwise sample (i, j, k), if item i is more similar to item j than item k, then $M_{ij} > M_{ik}$. We want to learn a low-rank matrix that also has this property, i.e., the i, j-th entry is greater than the i, k-th entry.

To gauge the scalability of ScaledSGD on a large-scale real-world dataset, we perform simulation on item-item collaborative filtering using a $62,000 \times 62,000$ item-item matrix M obtained from MovieLens25M dataset. The CF model is trained using Bayesian Personalized Ranking (BRP) loss [1] on a training set, which consists of 100 million pairwise samples in M. The performance of CF model is evaluated using Area Under the Curve (AUC) score [1] on a test set, which consists of 10 million pairwise samples in M.

Figure 3 plots the training BPR loss and testing AUC score within the first epoch (filled with red) and the second epoch (filled with blue). In order to measure the efficacy of ScaledSGD, we compare its testing AUC score against a standard baseline NP-Maximum: the upper bound on the best achievable AUC score by non-personalized ranking methods, in which the same ranking is used for all items. Based on the results in Figure 3, we emphasize two important points in the figure.

First, the percentage of training samples needed for ScaledSGD to achieve the same testing AUC scores as NP-Maximum is roughly 4 times smaller than SGD. Though both ScaledSGD and SGD are able to achieve higher AUC score than NP-Maximum before finishing the first epoch, ScaledSGD achieve the same AUC score as NP-Maximum after training on 11% of training samples while SGD requires 46% of them. We note that in this experiment, the size of the training set is 100 million, this means that SGD would require 35 million more iterations than ScaledSGD before it can reach NP-Maximum.

Second, the percentage of training samples needed for ScaledSGD to converge after the first epoch is also roughly 4 times smaller than SGD. Given that both ScaledSGD and SGD converge to AUC score at around 0.9 within the second epoch (area filled with blue), we indicate the percentage of training samples when both algorithms reach 0.9 AUC score in Figure 3. As expected, ScaledSGD is able to converge using fewer samples than SGD, with only 16% of training samples. SGD, on the other hand, requires 72% training samples.

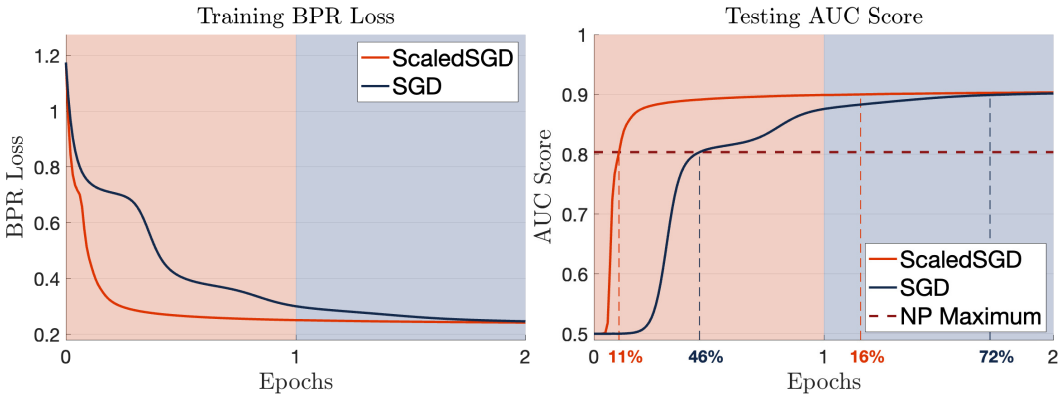

Figure 3: **Item-item collaborative filtering (CF).** We compare the training BPR loss and testing AUC score of ScaledSGD and SGD. (**Left**) Training BPR loss on a training set consists of 100 million samples. (**Right**) Testing AUC score on a test set consists of 10 million samples.

6 Conclusions

We propose an algorithm called ScaledSGD for huge scale online matrix completion. For the non-convex approach to solving matrix completion, ill-conditioning in the ground truth causes SGD to slow down significantly. ScaledSGD preserves all the favorable qualities of SGD while making it immune to ill-conditioning. For the RMSE loss, we prove that with an initial point close to the ground truth, ScaledSGD converges to an ϵ-accurate solution in $O(\log(1/\epsilon))$ iterations, independent of the condition number κ. We also run numerical experiments on a wide range of other loss functions commonly used in applications such as collaborative filtering, distance matrix recovery, etc. We find that ScaledSGD achieves similar acceleration on these losses, which means that it is widely applicable to many real problems. It remains future work to provide rigorous justification for these observations.

References

[1] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[2] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

[3] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[4] Tian Tong, Cong Ma, Ashley Prater-Bennette, Erin Tripp, and Yuejie Chi. Scaling and scalability: Provable nonconvex low-rank tensor completion. In *International Conference on Artificial Intelligence and Statistics*, pages 2607–2617. PMLR, 2022.

[5] Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

[6] Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

[7] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[8] Alexander Cloninger, Wojciech Czaja, Ruiliang Bai, and Peter J Basser. Solving 2d fredholm integral from incomplete measurements using compressive sensing. *SIAM journal on imaging sciences*, 7(3):1775–1798, 2014.

[9] Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

[10] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

[11] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[12] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[13] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[14] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[15] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1): 1665–1697, 2012.

[16] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[17] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

[18] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Conference on learning theory*, pages 638–678. PMLR, 2014.

[19] Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.

[20] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[21] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[22] Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. In *Conference on Learning Theory*, pages 1007–1034. PMLR, 2015.

[23] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

[24] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016.

[25] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[26] Jianhao Ma and Salar Fattahi. Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 2021.

[27] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[28] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[29] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[30] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[31] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[32] Ji Chen and Xiaodong Li. Memory-efficient kernel pca via partial matrix sampling and non-convex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

[33] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20(114):1–34, 2019.

[34] Richard Y Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

[35] Cédric Josz and Lexiao Lai. Nonsmooth rank-one matrix factorization landscape. *Optimization Letters*, pages 1–21, 2021.

[36] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[37] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[38] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[39] Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1475–1485. PMLR, 2020.

[40] Tian Tong, Cong Ma, and Yuejie Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021.

[41] Christian Kümmerle and Claudio M Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *International Conference on Machine Learning*, pages 5872–5883. PMLR, 2021.

[42] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[43] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[44] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.

[45] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[46] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[47] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

[48] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[49] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[50] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

[51] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[52] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[53] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[54] Brent Smith and Greg Linden. Two decades of recommender systems at amazon. com. *Ieee internet computing*, 21(3):12–18, 2017.

[55] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

A Supplemental Details on Experiments in Main Paper

A.1 Experimental setup and datasets used

Simulation environment We implement both ScaledSGD and SGD in MATLAB (version R2021a). All simulations in this paper are performed on a computer with Apple silicon M1 pro chip with 10-core CPU, 16-core GPU, and 32GB of RAM.

Datasets The datasets we use for the experiments in the main paper are described below.

- **Matrix completion with RMSE loss**: In the simulation result shown in Figure 1, we synthetically generate both the well-conditioned and ill-conditioned ground truth matrix M. We fix both M to be a rank-3 matrix of size 30×30. To generate M, we sample a random orthonormal matrix $U \in \mathbb{R}^{30\times 3}$ and set $M = USU^T$. For well-conditioned case, we set $S = \mathrm{diag}(2, 2, 2)$, thus M is perfectly conditioned with $\kappa = 1$. For ill-conditioned case, we let $S = \mathrm{diag}(10, 10^{-1}, 10^{-3})$, so that M is ill-conditioned with $\kappa = 10^4$.
- **Euclidean distance matrix completion**: In this simulation shown in Figure 2, the ground truth Euclidean distance matrix D for experiments 1 and 2 are generated with respect to their sample matrix X as $D_{ij} = \|x_i - x_j\|_2^2$. For the sample points in **Experiment 1**, we randomly sample (without replacement) 30 points in 3-dimensional cube centered at origin with side length 2, and the corresponding sample matrix X has conditioned number $\kappa = 1.3908$. For the sample points in **Experiment 2**, we take the first 5 sample points in experiment 1 and perturb its x-coordinate by 10, and keep the rest of the 25 samples intact. The corresponding sample matrix X has conditioned number $\kappa = 8.0828$.
- **Item-item collaborative filtering**: In this simulation shown in Figure 3, we use the MovieLens25M dataset [50], which is a standard benchmark for algorithms for recommendation systems.[2] This dataset consists of 25 million ratings over 62,000 movies by 162,000 users, the ratings are stored in an user-item matrix G whose (i, j)-th entry is the rating that the i-th user gives the j-th movie. The rating is from 1 to 5, where a higher score indicates a stronger preference. If the i, j-th entry is 0, then no rating is given. For the simulation of item-item collaborative filtering, the (i, j)-th entry of the ground truth item-item matrix M contains the *similarity score* between the item i and j, and can be computed by measuring cosine similarity between the i-th and j-th column of G. The details are described in A.4.

Hyperparameter and initialization We start SGD and PrecSGD at the same initial point in each simulation. The initial points for each simulation are drawn from the standard Gaussian distribution.

- **Matrix completion with RMSE loss**: The step-size for both ScaledSGD and SGD are set to be $\alpha = 0.3$. The search rank for both ScaledSGD and SGD are set to be $r = 3$.
- **Euclidean distance matrix completion**: Since SGD is only stable for small step-size in EDM completion problem, while ScaledSGD can tolerance larger step-sizes, we pick the largest possible step-size for ScaledSGD and SGD in both experiments. **Experiment 1**: step-size for ScaledSGD $\alpha = 0.2$, step-size for SGD $\alpha = 0.02$. **Experiment 2**: step-size for ScaledSGD $\alpha = 0.2$, step-size for SGD $\alpha = 0.002$. The search rank for both ScaledSGD and SGD are set to be $r = 3$.
- **Item-item collaborative filtering**: Similar to EDM completion problem, SGD is only stable for small step-size, we pick the largest possible step-size for both ScaledSGD and SGD. The step-size for ScaledSGD $\alpha = 1,000$, step-size for SGD $\alpha = 0.01$. The search rank for both ScaledSGD and SGD are set to be $r = 3$.

A.2 Matrix completion with RMSE loss

We now turn to the practical aspects of implementing ScaledSGD for RMSE loss function. In practical setting, suppose we are given a set $\Omega = \{(i, j)\}$ that contains indices for which we know the value of M_{ij}, our goal is to recover the missing elements in M by solving the following nonconvex

[2] The MovieLens25M dataset is accessible at https://grouplens.org/datasets/movielens/25m/

optimization

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) = \frac{1}{2|\Omega|}\sum_{(i,j)\in\Omega}\left(x_i^T x_j - M_{ij}\right)^2.$$

The gradient of $f(X)$ is

$$\nabla f(X) = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\left(x_i^T x_j - M_{ij}\right)\left(e_i e_j^T + e_j e_i^T\right)X.$$

ScaledGD update equations for RMSE loss Each iteration of ScaledSGD samples one element $(i,j)\in\Omega$ uniformly. The resulting iteration updates only two rows of X

$$x_{i,+} = x_i - \alpha\cdot\left(x_i^T x_j - M_{ij}\right)Px_j, \quad x_{j,+} = x_j - \alpha\cdot\left(x_i^T x_j - M_{ij}\right)Px_i.$$

The update on P is low-rank

$$P_+ = (P^{-1} + x_{i,+}x_{i,+}^T + x_{j,+}x_{j,+}^T - x_i x_i^T - x_j x_j^T)^{-1},$$

and can be computed by calling four times of rank-1 Sherman–Morrison–Woodbury (SMW) update formula in $O(r^2)$ time

$$(P^{-1} + uu^T)^{-1} = P - \frac{Puu^TP}{1+u^TPu}, \qquad (P^{-1} - uu^T)^{-1} = P + \frac{Puu^TP}{1-u^TPu}. \tag{8}$$

In practice, this low-rank update can be "pushed" onto a centralized storage of the preconditioner P. Heuristically, independent copies of P can be maintained by separate, distributed workers, and a centralized dispatcher can later merge the updates to P by simply adding the cumulative low-rank updates onto the existing centralized copy.

A.3 Euclidean distance matrix (EDM) completion

Suppose that we have d points $x_1,\ldots,x_d\in\mathbb{R}^r$ in r dimensional space, the Euclidean distance matrix $D\in\mathbb{R}^{d\times d}$ is a matrix of pairwise squared distance between d points in Euclidean space, namely, $D_{ij} = \|x_i - x_j\|^2$. Many applications, such as wireless sensor networks, communication and machine learning, require Euclidean distance matrix to provide necessary services. However, in practical scenario, entries in D that correspond to points far apart are often missing due to high uncertainty or equipment limitations in distance measurement. The task of Euclidean distance matrix completion is to recover the missing entries in D from a set of available measurement, and this problem can be formulated as a rank r matrix completion problem with respect to pairwise square loss function. Specifically, let $X\in\mathbb{R}^{d\times r}$ be a matrix containing $x_1,\ldots,x_d$ in its row and let $M = XX^T$ be the Grammian of X. Each entry of D can be written in terms of three entries in M

$$D_{ij} = \|x_i - x_j\|^2 = x_i^T x_i - 2x_i^T x_j + x_j^T x_j = M_{ii} - 2M_{ij} + M_{jj}.$$

Hence, given a set of sample $\Omega = \{(i,j)\}$ in D, the pairwise square loss function for EDM completion reads

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) = \frac{1}{4|\Omega|}\sum_{(i,j)\in\Omega}\left(x_i^T x_i - 2x_i^T x_j + x_j^T x_j - D_{ij}\right)^2.$$

The gradient of $f(X)$ is

$$\nabla f(X) = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\left(x_i^T x_i - 2x_i^T x_j + x_j^T x_j - D_{ij}\right)\left[\left(e_i e_i^T + e_j e_j^T\right) - \left(e_i e_j^T + e_j e_i^T\right)\right]X.$$

ScaledGD update equations for EDM completion Each iteration of ScaledSGD samples one element $(i,j)\in\Omega$ uniformly. The resulting iteration updates only two rows of X

$$x_{i,+} = x_i - \alpha\cdot\left(x_i^T x_i - 2x_i^T x_j + x_j^T x_j - D_{ij}\right)P(x_i - x_j),$$
$$x_{j,+} = x_j - \alpha\cdot\left(x_i^T x_i - 2x_i^T x_j + x_j^T x_j - D_{ij}\right)P(x_j - x_i)$$

Similarly, the update on P is low-rank and can be computed by calling four times of equation (8).

A.4 Item-item collaborative filtering

In the task of item-item collaborative filtering, the ground truth M is an $d \times d$ matrix where d is the number of items we wish to rank and the i, j-th of M is a similarity measure between the items. Our goal is to learn a low-rank matrix that preserves the ranking of similarity between the items. For instance, suppose that item i is more similar to item j than item k, then $M_{ij} > M_{ik}$, we want to learn a low-rank matrix XX^T that also has this property, i.e., $x_i^T x_j \geq x_i^T x_k$ where x_i is the i-th row of X.

Similarity score An important building block of item-item recommendation systems is the so-called item-item similarity matrix [52–54], which we denote by M. The i, j-th entry of this matrix is the pairwise *similarity scores* of items i and j. There are various measures of similarity. In our experiments we adopt a common similarity measure known as cosine similarity [53]. As a result, the item-item matrix can be computed from the user-item matrix. In particular, let g_i, g_j denote the i-th and j-th columns of the user-item matrix G, corresponding to the ratings given by all users to the i-th and j-th items. Then the (i, j)-th element of the item-item matrix M is set to

$$M_{ij} = g_i^T g_j / (\|g_i\| \|g_j\|).$$

In general, the item-item matrix computed this way will be very sparse and not capable of generating good recommendations. Our goal is to complete the missing entries of this matrix, assuming that that M is low-rank. As we will see, we can formulate this completion problem as an optimization problem over the set of rank-r matrices.

Pairwise entropy loss (BPR loss) The Bayesian Personalized Ranking (BRP) loss [1] is a widely used loss function in the context of collaborative filtering. For the task of predicting a personalized ranking of a set of items (videos, products, etc.), BRP loss often outperforms RMSE loss because it is directly optimized for ranking; most collaborative filtering models that use RMSE loss are essentially scoring each individual item based on user implicit feedbacks, in applications that only positive feedbacks are available, the models will not be able to learn to distinguish between negative feedbacks and missing entries.

The BPR loss in this context can be defined as follows. Let $\Omega = \{(i, j, k)\}$ denote a set of indices for which we observe the ranking of similarity between items i, j, k. Our observations are of the form $Y_{ijk} = 1$ if $M_{ij} > M_{ik}$ and $Y_{ijk} = 0$ otherwise. In other words, $Y_{ijk} = 1$ if item i is more similar to item j than to item k. We form a candidate matrix of the form XX^T, where $X \in \mathbb{R}^{d \times r}$. Our hope is that XX^T preserves the ranking between the items. The BPR loss function is designed to enforce this property.

Let x_i denote the i-th row of X and set $z_{ijk} = (XX^T)_{ij} - (XX^T)_{ik} = x_i^T (x_j - x_k)$. The BPR loss attempts to preserve the ranking of samples in each row of M by minimizing the logistic loss with respect to $(Y_{ijk}, \sigma(z_{ijk}))$, where $\sigma(\cdot)$ is the sigmoid function:

$$\min_{X \in \mathbb{R}^{n \times r}} f(X) = \frac{1}{|\Omega|} \sum_{(i,j,k) \in \Omega} -Y_{ijk} \log\left(\sigma(z_{ijk})\right) - (1 - Y_{ijk}) \log\left(1 - \sigma(z_{ijk})\right).$$

Then the gradient of $f(X)$ is

$$\nabla f(X) = \frac{1}{|\Omega|} \sum_{(i,j,k) \in \Omega} \left(\sigma(z_{ijk}) - Y_{ijk}\right) \left[\left(e_i e_j^T + e_j e_i^T\right) - \left(e_i e_k^T + e_k e_i^T\right)\right] X.$$

Similarly to the previous section, each iteration of ScaledSGD samples one element $(i, j, k) \in \Omega$ uniformly. The resulting iteration updates only three rows of X, as in

$$x_{i,+} = x_i - \alpha \cdot \left(\sigma(z_{ijk}) - Y_{ijk}\right) P(x_j - x_k), \quad x_{j,+} = x_j - \alpha \cdot \left(\sigma(z_{ijk}) - Y_{ijk}\right) P x_i,$$
$$x_{k,+} = x_k + \alpha \cdot \left(\sigma(z_{ijk}) - Y_{ijk}\right) P x_i$$

Similar to before, the preconditioner P can be updated via six call of equation (8) in $O(r^2)$ time.

The AUC score The AUC score [1] is a popular evaluation metric for recommendation system. Roughly speaking, the AUC score of a candidate matrix XX^T is the percentage of ranking of the

entries of M that is preserved by XX^T. Specifically, for each sample $(i,j,k) \in \Omega$, we define a indicator variable δ_{ijk} as

$$\delta_{ijk} = \begin{cases} 1 & \text{if } z_{ijk} > 0 \text{ and } Y_{ijk} = 1 \\ 1 & \text{if } z_{ijk} \leq 0 \text{ and } Y_{ijk} = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where we recall that Y_{ijk} is our observation and $z_{ijk} = (XX^T)_{ij} - (XX^T)_{ik}$. In other words, $\delta_{ijk} = 1$ only if the ranking between M_{ij} and M_{ik} is preserved by $(XX^T)_{ij}$ and $(XX^T)_{ik}$. The AUC score is then defined as the ratio

$$\text{AUC} = \frac{1}{|\Omega|} \sum_{(i,j,k)\in\Omega} \delta_{ijk}.$$

Thus, a higher AUC score indicates that the candidate matrix XX^T perserves a larger percentage of the pairwise comparisons in $|\Omega|$.

CF model training for Figure 3 We precompute a dataset of item-item pairwise comparisons M from the MovieLens25M datasets of user-item ratings, and then run ScaledSGD and SGD over 2 epochs on this dataset. Let $d = 62,000$ denote the number of items and let $\Omega \subseteq [d]^3$ denote the set of observations, i.e., the set of entries (i,j,k) where we observe $Y_{ijk} = 1$ if $M_{ij} > M_{ik}$ and $Y_{ijk} = 0$ if $M_{ij} < M_{ik}$. We construct Ω by sampling 110 million pairwise measurements uniformly at random without replacement from $(i,j,k) \sim [d]^3$ that have either $M_{ij} > M_{ik}$ or $M_{ij} < M_{ik}$. We do this because the item-item matrix M remains highly sparse, and there are many pairs of (i,j) and (i,k) for which $M_{ij} = M_{ik} = 0$.

To ensure independence between training and testing, we divide the set Ω into two disjoint sets Ω_{train} and Ω_{test}. The first set Ω_{train} consists of 100 million of all observations, which we use to fit our model. The second set Ω_{test} consists of 10 million samples for which we use to calculate the AUC score of our model on new data.

B Additional Experiments on pointwise cross-entropy loss

This problem is also known as 1-bit matrix completion [55]. Here our goal is to recover a rank-r matrix M through *binary* measurements. Specifically, we are allowed to take independent measurements on every entry M_{ij}, which we denote by Y_{ij}. Let $\sigma(\cdot)$ denote the sigmoid function, then $Y_{ij} = 1$ with probability $\sigma(M_{ij})$ and $Y_{ij} = 0$ otherwise. After a number of measurements are taken on each entry in the set Ω, let y_{ij} denote the percentage of measurements on the i,j-th entry that is equal to 1. The plan is to find the maximum likelihood estimator for M by minimizing a cross-entropy loss defined as follow

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} -y_{ij} \log\left(\sigma(x_i^T x_j)\right) - (1-y_{ij}) \log\left(1-\sigma(x_i^T x_j)\right).$$

We assume an ideal case where the number of measurements is large enough so that $y_{ij} = \sigma(M_{ij})$ and the entries are fully observed. The gradient of $f(X)$ is

$$\nabla f(X) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \left(\sigma(x_i^T x_j) - y_{ij}\right)\left(e_i e_j^T + e_j e_i^T\right) X.$$

ScaledGD update equations for pointwise cross-entropy loss. Each iteration of ScaledSGD samples one element $(i,j) \in \Omega$ uniformly. The resulting iteration updates only two rows of X

$$x_{i,+} = x_i - \alpha \cdot \left(\sigma(x_i^T x_j) - y_{ij}\right) P x_j, \quad x_{j,+} = x_j - \alpha \cdot \left(\sigma(x_i^T x_j) - y_{ij}\right) P x_i.$$

The preconditioner P can be updated by calling four times of equation (8) as in RMSE loss.

Matrix completion with pointwise cross-entropy loss. We apply ScaledSGD to perform matrix completion through minimizing pointwise cross-entropy loss. In this experiment, the well-conditioned and ill-conditioned ground truth matrix M is the same as those in Figure 1, and the

process of data generation are described in A.1. The learning rate for both ScaledSGD and SGD are set to be $\alpha = 1$. The search rank for both ScaledSGD and SGD are set to be $r = 3$.

Figure 4 plots the error $f(X) = \|XX^T - M\|_F^2$ against the number of epochs. Observe that the results shown in Figure 4 are almost identical to that of the RMSE loss shown in Figure 1. Ill-conditioning causes SGD to slow down significantly while ScaledSGD is unaffected.

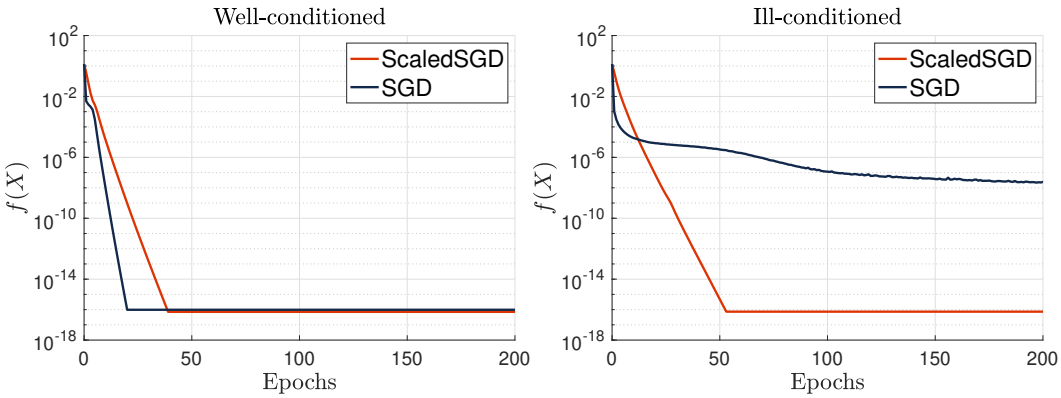

Figure 4: **Matrix Completion with pointwise cross entropy loss.** We compare the convergence rate of ScaledSGD and SGD for a well-conditioned and ill-conditioned ground truth matrix M used in Figure 1. (**Left**) Well-conditoned M. (**Right**) Ill-conditoned M.

C Additional Experiments with Noise

To mimic the real-world datasets, we corrupt each entry of ground truth matrix M by white Gaussian noise. We first generate a noiseless well-conditioned and ill-conditioned matrix $\tilde{M}$ following same procedure as the one described in A.1. For well-conditioned case, we set the singular value as $S = \text{diag}(10, 10, 10)$. For the ill-conditioned case, we set $S = \text{diag}(10, 10^{-1}, 10^{-3})$. To obtain a noisy ground truth, we generate a matrix of white Gaussian noise W corresponding to a fixed signal to noise ratio (SNR), which is defined as SNR $= 20\log_{10}(\|\tilde{M}\|_F/\|W\|_F)$. Finally, we set $M = \tilde{M} + W$. For the experiments in this section, we set SNR $= 15$dB. For the case of well-conditioned $\tilde{M}$, the resulting $M = \tilde{M} + W$ is full-rank with condition number $\kappa = 310.72$. For the case of ill-conditioned $\tilde{M}$, the resulting M is full-rank with condition number $\kappa = 423.5022$.

Matrix completion with RMSE loss on noisy datasets We plot the convergence rate of ScaledSGD and SGD under the noisy setting in Figure 5. In this experiment, we pick a larger search rank $r = 5$ to accommodate the noisy ground truth. Observe that SGD slows down in both the well-conditioned and ill-conditioned case due to the addition of white Gaussian noise and the larger search rank r, while ScaledSGD converge linearly toward the noise floor.

We also plot the noise floor, which can be computed as follows. First we take the eigendecomposition of $M = Q\Lambda Q^T$, where Q is an orthonormal matrix and Λ is a diagonal matrix containing the eigenvalues of M sorted in descending order in its diagonal entries. Let Λ' be a diagonal matrix such that $\Lambda'_{ii} = \Lambda_{ii}$ if $i \leq r$, and $\Lambda'_{ii} = 0$ otherwise, then the noise floor is defined as the RMSE between M and its best rank-r approximation $M' = Q\Lambda' Q^T$, which is equal to $\frac{1}{2|\Omega|}\sum_{(i,j)\in\Omega}\left(M'_{ij} - M_{ij}\right)^2$.

The step-sizes in the simulation are set to be the largest possible step-sizes for which ScaledSGD and SGD can converge to the noise floor. For ScaleSGD, the step-size is set to be $\alpha = 0.15$. For SGD, the step-size is set to be $\alpha = 0.01$. SNR $= 20\log_{10}(\|\tilde{M}\|_F/\|W\|_F) = 15$dB.

Matrix completion with pointwise cross-entropy loss on noisy datasets We plot the convergence rate of ScaledSGD and SGD under the noisy setting in Figure 6. Similar to RMSE loss in noisy setting, SGD show down in both well-conditioned and ill-conditioned case, while ScaledSGD converge linearly toward the noise floor. In this simulation, the search rank is set to be $r = 5$. The step-size are set to be the largest possible step-sizes for which ScaledSGD and SGD can converge

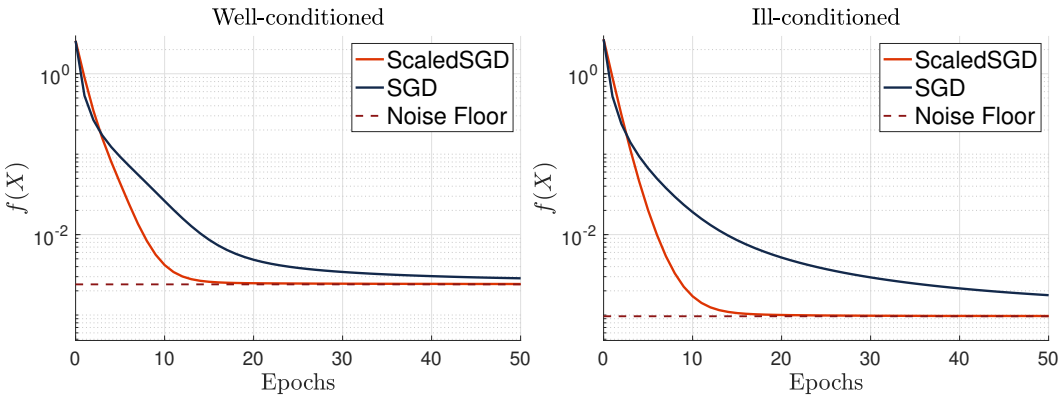

Figure 5: **Matrix Completion with RMSE loss in noisy setting.** We compare the convergence rate of ScaledSGD and SGD for noisy ground truth matrix $M = \tilde{M} + W$ computed with respect to a well-conditioned and ill-conditioned $\tilde{M}$ and white Gaussian noise W. (**Left**) Well-conditioned $\tilde{M}$. (**Right**) Ill-conditioned $\tilde{M}$.

to the noise floor. For ScaleSGD, the step-size is set to be $\alpha = 0.15$. For SGD, the step-size is set to be $\alpha = 0.01$. SNR $= 15$dB.

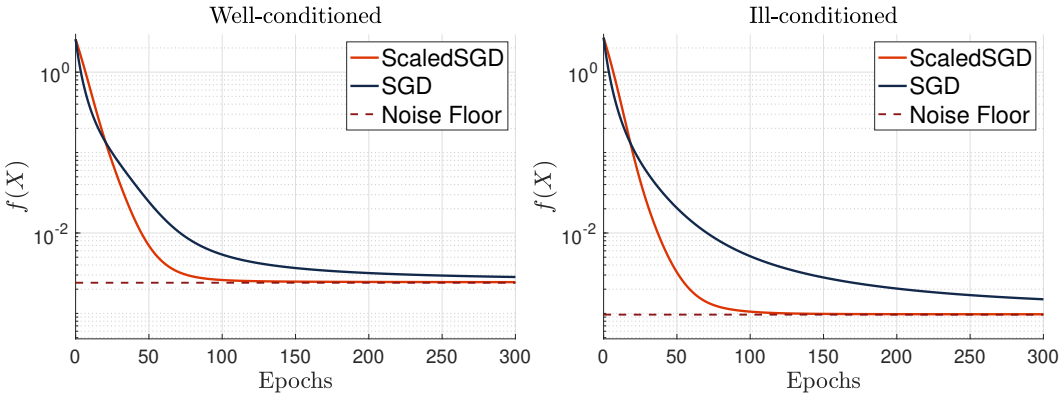

Figure 6: **Matrix Completion with pointwise cross-entropy loss in the noisy setting.** We compare the convergence rate of ScaledSGD and SGD for noisy ground truth matrix $M = \tilde{M} + W$ computed with respect to a well-conditioned and ill-conditioned $\tilde{M}$ and white Gaussian noise W. (**Left**) Well-conditioned $\tilde{M}$. (**Right**) Ill-conditioned $\tilde{M}$.

D Additional simulation on item-item collaborative filtering

Finally, we perform additional experiments on item-item collaborative filtering in order to compare the ability of ScaledSGD and SGD to generate good recommendations using matrix factorization.

Dataset For additional simulations on item-time collaborative filtering, we use the Small-scale and Full-scale MovieLens Latest datasets [50] in order to gauge the performance of our algorithm on different scales. First, we run experiments on the Small-scale MovieLens Latest dataset that has 100,000 ratings over 9,000 movies by 600 users. Second, we also run experiments on the Full-scale MovieLens Latest dataset with 27 million total ratings over 58,000 movies by 280,000 users. [3]

Experimental Setup The process of training a collaborative filtering model is described in A. The hyperparameters and initializations for experiments in this section are described below

[3] Both datasets are accessible at `https://grouplens.org/datasets/movielens/latest/`

- **Small-scale dataset** We sample $m = |\Omega| = 500{,}000$ pairwise observations. We set our search rank to be $r = 10$, so the optimization variable X is of size 4180×10. Both ScaledGD and SGD are initialized using a random Gaussian initial point. For ScaledSGD the step-size is 10^2 and for SGD the step-size is 10^{-1}.
- **Full-scale dataset** We sample $m = |\Omega| = 2.7$ million pairwise observations in a first experiment, and $m = |\Omega| = 10$ million pairwise observations in a second experiment. In both cases, we set our search rank to be $r = 4$, so the optimization variable X is of size 59047×4 and 4321441×4 respectively. For ScaledSGD the step-size is 10^2 and for SGD the step-size is 10^{-1}.

Difference in learning rates between SGD and ScaledSGD In our experiments, the step-sizes for both ScaledSGD and SGD are set using a hyperparameter search. We observe that under identical conditions, ScaledSGD is able to tolerate a much larger step-size than SGD. Thus, we set the step-size to be the largest magnitude possible for both methods.

Results The results of our experiments for SGD and ScaledSGD are plotted in Figures 7, 8, and 9. In all three cases, we see both algorithms converging towards a good fit before apparently overfitting. This is reflected in the fact that both algorithms decrease the training loss towards some minimum value before increasing the training loss again. Both algorithms have a comparable rate of convergence, although SGD is actually a little bit faster in the smaller examples. Crucially, we observe in all three cases that ScaledSGD attains a better asymptote than SGD, which we suspect is due to its ability to better minimize the training loss.

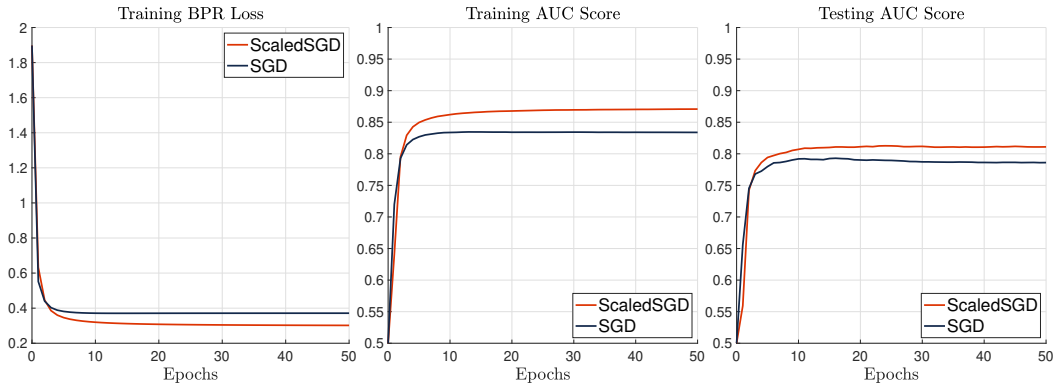

Figure 7: **Collaborative filtering with pairwise entropy loss (BPR loss).** (Small Scale with $m = 500,000$ pairwise measurements). We compare the BPR loss and AUC score of ScaledSGD and SGD using both the training set Ω_{train} and the test set Ω_{test}.

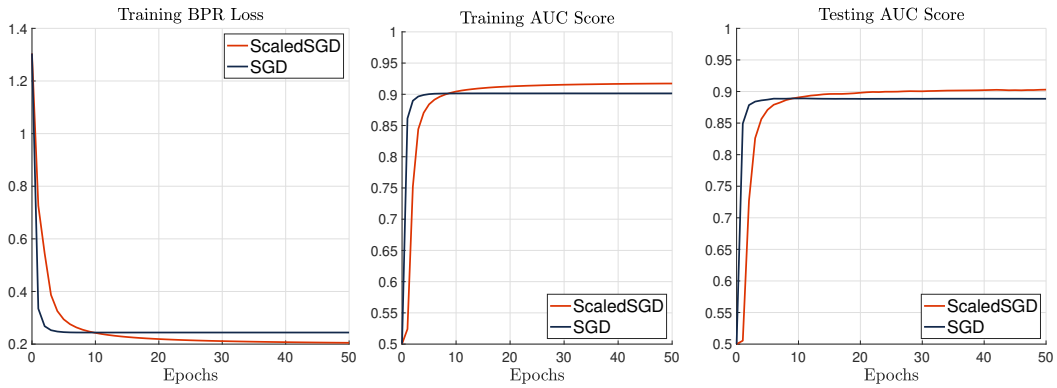

Figure 8: **Collaborative filtering with pairwise entropy loss (BPR loss).** (Full Scale with $m = 2.7$ million pairwise measurements). We compare the BPR loss and AUC score of ScaledSGD and SGD using both the training set Ω_{train} and the test set Ω_{test}.

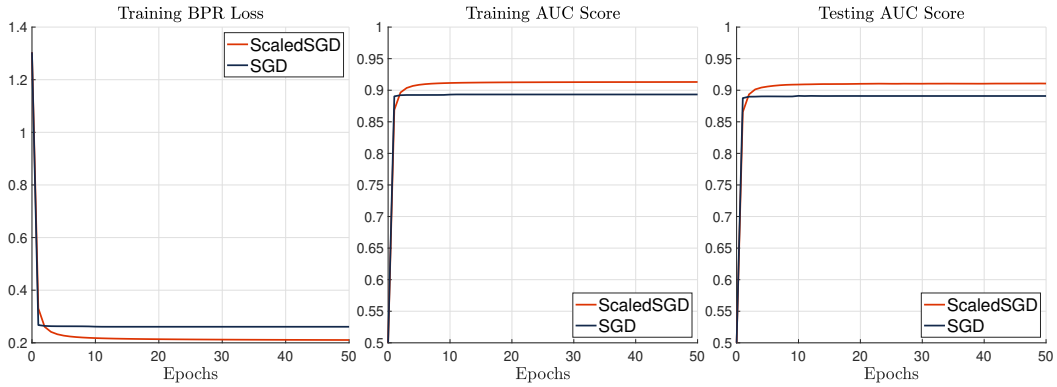

Figure 9: **Collaborative filtering with pairwise entropy loss (BPR loss).** (Full Scale with $m = 10$ million pairwise measurements). We compare the BPR loss and AUC score of ScaledSGD and SGD using both the training set Ω_{train} and the test set Ω_{test}.

E Proof of the theoretical results

In this section, we show that, in expectation, the search direction $V = SG(X)(X^T X^{-1})$ makes a geometric decrement to both the function value f and the incoherence g. A key idea is to show that the size of the decrement in f is controlled by the coherence $g_{\max} \geq g_k(X)$ of the current iterate, and this motivates the need to decrement g_k in order to keep the iterates incoherent. Our key result is that both decrements are independent of the condition number κ.

E.1 Preliminaries

We define the inner product between two matrices as $\langle X, Y \rangle \overset{\text{def}}{=} \operatorname{tr}(X^T Y)$, which induces the Frobenius norm as $\|X\|_F = \sqrt{\langle X, X \rangle}$. The vectorization $\operatorname{vec}(X)$ is the column-stacking operation that turns an $m \times n$ matrix into a length-mn vector; it preserves the matrix inner product $\langle X, Y \rangle = \operatorname{vec}(X)^T \operatorname{vec}(Y)$ and the Frobenius norm $\|\operatorname{vec}(X)\| = \|X\|_F$.

We denote $\lambda_i(M)$ and $\sigma_i(M)$ as the i-th eigenvalue and singular value of a symmetric matrix $M = M^T$, ordered from the most positive to the most negative. We will often write $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to index the most positive and most negative eigenvalues, and $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ for the largest and smallest singular values.

Recall for any matrix $V \in \mathbb{R}^{d \times r}$, we define its local norm with respect to $X \in \mathbb{R}^{d \times r}$ as

$$\|V\|_X = \|V(X^T X)^{1/2}\|_F, \qquad \|V\|_X^* = \|V(X^T X)^{-1/2}\|_F.$$

Also recall that we have defined the stochastic gradient operator

$$SG(X) = 2d^2 \cdot (XX^T - ZZ^T)_{i,j} \cdot (e_i e_j^T + e_j e_i^T) X \tag{9}$$

where $(i, j) \sim [d]^2$ is selected uniformly at random. This way SGD is written $X_+ = X - \alpha SG(X)$ and ScaledSGD is written $X_+ = X - \alpha SG(X)(X^T X)^{-1}$ for step-size $\alpha > 0$.

E.2 Function value convergence

Recall that Theorem 1, due to Jin et al. [10], says that SGD converges to ϵ accuracy in $O(\kappa^4 \log(1/\epsilon))$ iterations, with a four-orders-of-magnitude dependence on the condition number κ. By comparison, our main result Theorem 2 says that ScaledSGD converges to ϵ accuracy in $O(\log(1/\epsilon))$ iterations, completely independence on the condition number κ. In this section, we explain how the first two factors of κ are eliminated, by considering the full-batch counterparts of these two algorithms.

First, consider full-batch gradient descent on the function $f(X) = \|XX^T - ZZ^T\|_F^2$. It follows from the local Lipschitz continuity of f that

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) - \alpha\langle \nabla f(X_t), \nabla f(X_t)\rangle + \alpha \cdot (L/2) \cdot \|\nabla f(X_t)\|_F^2 \\ &= f(X_t) - \underbrace{\alpha\|\nabla f(X_t)\|_F^2}_{\text{linear progress}} + \alpha^2 \cdot (L/2) \cdot \underbrace{\|\nabla f(X_t)\|_F^2}_{\text{inverse step-size}} \end{aligned} \tag{10}$$

where $X_{t+1} = X_t - \alpha\nabla f(X_t)$. Here, the linear progress term determines the amount of progress that can proportionally be made with a sufficiently small step-size α, whereas the inverse step-size term basically controls how large the step-size can be. In the case of full-batch gradient descent, it is long known that an X_t that is sufficiently close to Z will satisfy the following

$$8\lambda_{\min}(Z^TZ) \cdot f(X_t) \leq \|\nabla f(X_t)\|_F^2 \leq 16\lambda_{\max}(Z^TZ) \cdot f(X_t)$$

and therefore, taking $\lambda_{\max}(Z^TZ) = 1$ and $\lambda_{\min}(Z^TZ) = \kappa^{-1}$ where κ is the condition number, we have linear convergence

$$f(X_{t+1}) \leq \left(1 - \alpha \cdot 8\kappa^{-1} + \alpha^2 \cdot 8L\right) f(X_t) \leq \left(1 - \alpha \cdot 4\kappa^{-1}\right) f(X_t)$$

for step-sizes of $\alpha \leq 2/(\kappa L)$. Therefore, it follows from this analysis that full-batch gradient descent takes $T = O(\kappa^2 \log(1/\epsilon))$ iterations to converge to ϵ-accuracy. In this iteration count, one factor of κ arises from the linear progress term, which shrinks as $O(\kappa^{-1})$ as κ grows large. The second factor of κ arises because the inverse step-size term is a factor of κ larger than the linear progress term, which restricts the maximum step-size to be no more than $O(\kappa^{-1})$.

The following lemma, restated from the main text, shows that an analogous analysis for full-batch ScaledSGD proves an iteration count of $T = O(\log(1/\epsilon))$ with no dependence on the condition number κ. In fact, it proves that full-batch ScaledSGD converges like full-batch gradient descent with a perfect condition number $\kappa = 1$.

Lemma 5 (Function descent, Lemma 3 restated)**.** *Let $X, Z \in \mathbb{R}^{n\times r}$ satisfy $\|XX^T - ZZ^T\|_F \leq \rho\lambda_{\min}(Z^TZ)$ where $\rho < 1/2$. Then, the function $f(X) = \|XX^T - ZZ^T\|_F^2$ satisfies*

$$|f(X+V) - f(X) - \langle \nabla f(X), V\rangle| \leq \frac{L_X}{2} \cdot \|V\|_X^2, \tag{11}$$

$$13 \cdot f(X) \leq (\|\nabla f(X)\|_X^*)^2 \leq 16 \cdot f(X), \tag{12}$$

for all $\|V\|_X \leq C \cdot \sqrt{f(X)}$ with $L_X = 6 + 8C + 2C^2 = O(1 + C^2)$.

It follows that the iteration $X_{t+1} = X_t - \alpha\nabla f(X_t)(X_t^TX_t)^{-1}$ yields

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) - \alpha\langle \nabla f(X_t), \nabla f(X_t)(X_t^TX_t)^{-1}\rangle + \alpha \cdot (L_X/2) \cdot \|\nabla f(X_t)(X_t^TX_t)^{-1}\|_X^2 \\ &= f(X_t) - \alpha(\|\nabla f(X_t)\|_X^*)^2 + \alpha^2 \cdot (L_X/2) \cdot (\|\nabla f(X_t)\|_X^*)^2 && (13) \\ &\leq \left(1 - \alpha \cdot 8 + \alpha^2 \cdot 8L_X\right) f(X_t) \leq (1 - \alpha \cdot 4)\, f(X_t) && (14) \end{aligned}$$

for step-sizes of $\alpha \leq 2/L_X$, where $L_X = 6 + 8(4) + 2(4)^2$. Therefore, we conclude that full-batch ScaledGD takes $T = O(\log(1/\epsilon))$ iterations to converge to ϵ-accuracy, as if the condition number were perfectly $\kappa = 1$.

Note that Lemma 5 has been proved in both Tong et al. [9] and Zhang et al. [5]. For completeness, we give a proof inspired by Zhang et al. [5].

Proof of Lemma 5. We prove (11) via a direct expansion of the quadratic

$$\begin{aligned} \underbrace{\|(X+V)(X+V)^T - ZZ^T\|_F^2}_{f(X+V)} &= \underbrace{\|XX^T - ZZ^T\|_F^2}_{f(X)} + \underbrace{2\langle XX^T - ZZ^T, XV^T + VX^T\rangle}_{\langle \nabla f(X), V\rangle} \\ &+ \underbrace{2\langle XX^T - ZZ^T, VV^T\rangle + \|XV^T + VX^T\|_F^2}_{\frac{1}{2}\langle \nabla^2 f(X)[V], V\rangle} + \underbrace{2\langle VX^T + XV^T, VV^T\rangle}_{\frac{1}{6}\langle \nabla^3 f(X)[V,V], V\rangle} + \underbrace{\|VV^T\|_F^2}_{\frac{1}{24}\langle \nabla^4 f(X)[V,V,V], V\rangle} \end{aligned}$$

and it then follows by simple counting that

$$|f(X+V) - f(X) - \langle \nabla f(X), V\rangle| \le \frac{L_2}{2}\|V\|_X^2 + \frac{L_3}{6}\|V\|_X^3 + \frac{L_4}{24}\|V\|_X^4,$$
$$L_2 = 4 + 2\frac{\|XX^T - ZZ^T\|_F}{\lambda_{\min}(X^TX)}, \quad L_3 = \frac{24}{\lambda_{\min}(X^TX)}, \quad L_4 = \frac{24}{\lambda_{\min}^2(X^TX)}.$$

Now, from Weyl's inequality that

$$\lambda_{\min}(X^TX) = \lambda_r(XX^T) \ge \lambda_r(ZZ^T) - \|XX^T - ZZ^T\|_F \ge (1-\rho)\cdot\lambda_{\min}(Z^TZ)$$

and therefore $\|XX^T - ZZ^T\|_F/\lambda_{\min}(X^TX) \le \rho/(1-\rho) \le 1$ because $\rho < 1/2$. It follows that $L_2 \le 6$. If $\|V\|_X \le C\cdot\sqrt{f(X)}$, then $\frac{L_3}{3}\|V\|_X \le 8C$ and $\frac{L_4}{12}\|V\|_X^2 \le 2C^2$.

For the upper-bound in (12), we have simply

$$\|\nabla f(X)\|_X^* = 4\|(XX^T - ZZ^T)X(X^TX)^{-1/2}\|_F \le 4\|XX^T - ZZ^T\|_F.$$

For the lower-bound in (12), we evoke Zhang et al. [5, Lemma 12] with RIP constant $\delta = 0$ and regularization parameter $\eta = 0$ to yield[4]

$$\|\nabla f(X)\|_X^* = \max_{\|Y\|_X=1} 2\langle XY^T + YX^T, XX^T - ZZ^T\rangle = 4\|XX^T - ZZ^T\|_F\cdot\cos\theta$$

in which $\cos\theta$ is defined between $XX^T - ZZ^T$ and the set $\{XY^T + YX^T : Y \in \mathbb{R}^{d\times r}\}$, as in

$$\cos\theta = \max_{Y\in\mathbb{R}^{d\times r}} \frac{\langle XY^T + YX^T, XX^T - ZZ^T\rangle}{\|XY^T + YX^T\|_F\cdot\|XX^T - ZZ^T\|_F}.$$

It follows from Zhang et al. [5, Lemma 13] that

$$\sin\theta = \frac{\|(I - XX^\dagger)(XX^T - ZZ^T)(I - XX^\dagger)\|_F}{\|XX^T - ZZ^T\|} \le \frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}}.$$

Hence, for $\rho < 1/2$, we have

$$\frac{(\|\nabla f(X)\|_X^*)^2}{\|XX - ZZ^T\|_F^2} = 16\cos^2\theta \ge 16\left(1 - \frac{1}{2}\frac{\rho^2}{1-\rho^2}\right) \ge 16\left(1 - \frac{1}{6}\right) = \frac{40}{3} > 13.$$

□

E.3 Coherence convergence

We now explain that ScaledSGD eliminates the last two factors of κ from SGD because it is able to keep its iterates a factor of κ^2 more incoherent. First, consider regular SGD on the function $f(X) = \|XX^T - ZZ^T\|_F^2$. Conditioning on the current iterate, we have via the local Lipschitz continuity of f:

$$\begin{aligned}\mathbb{E}[f(X_+)] &\le f(X) - \alpha\langle\nabla f(X), \mathbb{E}[SG(X)]\rangle + \alpha\cdot(L/2)\cdot\mathbb{E}[\|SG(X)\|_F^2] \\ &= f(X) - \underbrace{\alpha\|\nabla f(X)\|_F^2}_{\text{linear progress}} + \alpha^2\cdot(L/2)\cdot\underbrace{\mathbb{E}[\|SG(X)\|_F^2]}_{\text{inverse step-size}} \end{aligned} \tag{15}$$

where $X_+ = X - \alpha SG(X)$. In expectation, the linear progress term of SGD coincides with that of full-batch gradient descent in (10). The inverse step-size term, however, is up to a factor of d^2 times larger. To see this, observe that

$$\begin{aligned}\mathbb{E}[\|SG(X)\|_F^2] &= \frac{1}{d^2}\sum_{i=1}^d\sum_{j=1}^d \|2d^2\cdot(XX^T - ZZ^T)_{i,j}\cdot(e_ie_j^T + e_je_i^T)X\|_F^2 \\ &= 4d^2\cdot\sum_{i=1}^d\sum_{j=1}^d (XX^T - ZZ^T)_{i,j}^2\cdot\|(e_ie_j^T + e_je_i^T)X\|_F^2 \\ &= 4d^2\cdot f(X)\cdot\|(e_ie_j^T + e_je_i^T)X\|_F^2 \le 16d^2\cdot f(X)\cdot\max_i\|e_i^TX\|_F^2.\end{aligned}$$

[4] Here, we correct for a factor-of-two error in Zhang et al. [5, Lemma 12].

In a coarse analysis, we can simply bound $\max_i \|e_i^T X\|_F^2 \le \lambda_{\max}(X^T X) = O(1)$ to yield

$$f(X_+) \le \left(1 - \alpha \cdot 8\kappa^{-1} + 16d^2 \cdot \alpha^2\right) f(X) \le \left(1 - \alpha \cdot 4\kappa^{-1}\right) f(X)$$

for step-sizes of $\alpha \le 4/(\kappa d^2)$. Hence, we conclude that it takes $T = O(\kappa^2 d^2 \log(1/\epsilon))$ iterations to converge to ϵ-accuracy, with an epoch of d^2 iterations of SGD essentially recreating a single iteration of full-batch gradient descent. Unfortunately, the matrix is already fully observed after d^2 iterations, and so this result is essentially vacuous.

Here, Jin et al. [10] pointed out that the term $h_{\max} = \max_i \|e_i^T X\|_F^2$ measures the *coherence* of the $d \times r$ iterate X, and can be as small as $O(1/d)$ for small values of rank $r = O(1)$. Conditioned on the current iterate X, they observed that the function $h_i(X) = \|e_i^T X\|_F^2$ converges towards a finite value in expectation

$$\begin{aligned}\mathbb{E}[h_i(X_+)] &\le \left(1 - \alpha \cdot 8\kappa^{-1}\right) h_i(X) + \alpha \cdot 8\sqrt{h_i(X) h_i(Z)} + \alpha^2/2 \cdot \mathbb{E}[\|e_i^T SG(X)\|_F^2] \\ &\le \left(1 - \alpha \cdot 8\kappa^{-1}\right) h_i(X) + \alpha \cdot 8\sqrt{h_i(X) h_i(Z)} + \alpha^2 \cdot O(d^2 h_{\max}^2).\end{aligned}$$

Let us define γ as the *ratio* between the coherences of the ground truth Z and the iterate X:

$$\gamma = \frac{\max_i \|e_i^T X\|_F^2}{\max_j \|e_j^T Z\|_F^2} = \frac{\max_i h_i(Z)}{\max_j h_j(Z)} \qquad \Longleftrightarrow \qquad \max_j \|e_j^T Z\|_F^2 \le \gamma^{-1} \cdot h_{\max}.$$

Crucially, we require $\gamma = \kappa^2$ in order for $h_i(X)$ to converge towards $\frac{1}{2} h_{\max}$ in expectation:

$$\begin{aligned}\mathbb{E}[h_i(X_+) - \frac{1}{2} h_{\max}] &\le \left(1 - \alpha \cdot 8\kappa^{-1}\right) h_i(X) + \alpha \cdot 4\gamma^{-1/2} h_{\max} - \frac{1}{2} h_{\max} \\ &\le \left(1 - \alpha \cdot 8\kappa^{-1}\right) \left[h_i(X) - \left(\frac{1 - \alpha \cdot 8\gamma^{-1/2}}{1 - \alpha \cdot 8\kappa^{-1}}\right) \frac{1}{2} h_{\max}\right].\end{aligned}$$

As a consequence, we conclude that, while SGD is able to keep its iterates X incoherent, their actual coherence $h_{\max} = \max_i \|e_i^T X\|_F^2$ is up to a factor of κ^2 worse than the coherence $\max_j \|e_j^T Z\|_F^2$ of the ground truth Z.

Using a standard supermartingale argument, Jin et al. [10] extended the analysis above to prove that if the ground truth Z has coherence $\max_j \|e_j^T Z\|_F^2 = O(1/d)$, then the SGD generates iterates X that have coherence $\max_i \|e_i^T X\|_F^2 \le h_{\max} = O((\kappa^2/d) \log d)$, which is two factors worse in κ as expected. Combined, this proves that SGD converges to ϵ accuracy in $T = O(\kappa^4 dr \log(d/\epsilon))$ iterations with the step-size of $\alpha = O(\kappa^{-1} d^{-1} h_{\max}^{-1})$ and iterate coherence $h_{\max} = O((\kappa^2/d) \log d)$, which is another two factors of κ worse than full-batch gradient descent.

The following lemma, restated from the main text, shows that an analogous analysis for ScaledSGD proves that the algorithm maintains iterates X whose coherences have no dependence on κ. Here, we need to define a different incoherence function $g_i(X) = \|e_i X (X^T X)^{-1/2}\|^2 \equiv (\|e_i X\|_X^*)^2$ in order to "stochastify" our previous analysis for full-batch ScaledGD. Surprisingly, the factors of $(X^T X)^{-1}$ in both the new definition of $g_i(X)$ and the search direction $SG(X)(X^T X)^{-1}$ do not hurt incoherence, but in fact improves it.

Lemma 6 (Coherence descent, Lemma 4 restated)**.** *Let $X, Z \in \mathbb{R}^{n \times r}$ satisfy $\|XX^T - ZZ^T\|_F \le \rho \lambda_{\min}(Z^T Z)$ where $\rho < 1/2$. Then, the functions $f(X) = \|XX^T - ZZ^T\|_F^2$ and $g_k(X) = e_k^T X (X^T X)^{-1} X^T e_k$ satisfy*

$$\begin{aligned}|g_k(X + V) - g_k(X) - \langle V, \nabla g_k(X)\rangle| &\le \frac{5(\|V\|_X^*)^2}{1 - 2\|V\|_X^*}, \\ \langle \nabla g_k(X), \nabla f(X)(X^T X)^{-1}\rangle &\ge \left[\frac{1 - 2\rho}{1 - \rho} g_k(X) - \frac{1}{1 - \rho}\sqrt{g_k(X) g_k(Z)}\right].\end{aligned}$$

Conditioning on X, we have for the search direction $V = SG(X)(X^T X)^{-1}$ and $X_+ = X + V$

$$\begin{aligned}\mathbb{E} g_k(X_+) &\le g_k(X) - \alpha \langle \nabla g_k(X), \mathbb{E}[V]\rangle + \alpha^2 \cdot \mathbb{E}\left[\frac{(\|V\|_X^*)^2}{1 - 2\|V\|_X^*}\right] \\ &\le (1 - \zeta\alpha)\, g_k(X) + \alpha \cdot \frac{\zeta}{2} g_{\max} \qquad \text{for } \alpha = O(\rho^{-1} d^{-2}) \qquad (16)\end{aligned}$$

where $\zeta = \frac{1-2\rho}{1-\rho}$. It then follows that $g_k(X_+)$ converges geometrically towards $\frac{1}{2}g_{\max}$ in expectation, with a convergence rate $(1-\zeta\alpha)$ that is independent of the condition number κ:

$$\mathbb{E}\left[g_k(X_+) - \frac{1}{2}g_{\max}\right] \leq \left[(1-\zeta\alpha)\, g_k(X) + \alpha \cdot \frac{\zeta}{2} g_{\max}\right] - \frac{1}{2}g_{\max} \leq (1-\zeta\alpha)\left[g_k(X) - \frac{1}{2}g_{\max}\right].$$

Before we prove Lemma 6, we first need to prove a simple claim.

Lemma 7 (Change of norm)**.** *The local norm* $\|V\|_X^* = \|V(X^TX)^{-1/2}\|$ *satisfies*

$$\frac{(\|V\|_X^*)^2}{1 + 2\|Y-X\|_X^* + (\|Y-X\|_X^*)^2} \leq (\|V\|_Y^*)^2 \leq \frac{(\|V\|_X^*)^2}{1 - 2\|Y-X\|_X^*}.$$

Proof. The upper-bound follows because

$$\operatorname{tr}(VP_YV^T) = \operatorname{tr}(VP_X^{1/2}[P_X^{-1/2}P_YP_X^{-1/2}]P_X^{1/2}V^T) \leq \operatorname{tr}(VP_XV^T)/\lambda_{\min}[P_X^{1/2}P_Y^{-1}P_X^{1/2}]$$

where $P_Y = (Y^TY)^{-1}$ and $P_X = (X^TX)^{-1}$ and therefore

$$P_X^{1/2}P_Y^{-1}P_X^{1/2} \succeq I + P_X^{1/2}[X^T(Y-X) + (Y-X)^TX]P_X^{1/2}$$

and $\sigma_{\max}[P_X^{1/2}[X^T(Y-X)P_X^{1/2}] \leq \|Y-X\|_X$ because $XP_X^{1/2}$ is orthonormal. The lower-bound follows similarly. □

We are ready to prove Lemma 6.

Proof of Lemma 6. It follows from the intermediate value version of Taylor's theorem that there exists some $\tilde{X} = X + tV$ with $t \in [0,1]$ that

$$g_i(X+V) - g_i(X) - \langle \nabla g_i(X), V\rangle = \frac{1}{2}\langle \nabla^2 g_i(\tilde{X})[V], V\rangle.$$

Let $P = (X^TX)^{-1}$ and $U = e_ie_i^T(I - XPX^T)V$ and $G = VPX^Te_ie_i^T$. By direct computation, we have

$$\frac{1}{2}\langle \nabla g_i(X), V\rangle = \langle (I - XPX^T)e_ie_i^TXP, V\rangle = \langle U, XP\rangle = \langle I - XPX^T, G\rangle,$$

$$\frac{1}{2}\langle \nabla^2 g_i(X)[V], V\rangle = \langle UP - XP(U^TX + X^TU)P, V\rangle - \langle (I - XPX^T)(G + G^T)XP, V\rangle,$$

by differentiating XP and XPX^T respectively. A coarse count yields $\frac{1}{2}\langle \nabla^2 g_i(X)[V], V\rangle \leq 5(\|V\|_X^*)^2$ and therefore

$$|g_i(X+V) - g_i(X) - \langle \nabla g_i(X), V\rangle| \leq 5\|V\|_{X+tV}^2 \leq \frac{5(\|V\|_X^*)^2}{1 - 2t\|V\|_X^*} \leq \frac{5(\|V\|_X^*)^2}{1 - 2\|V\|_X^*},$$

which is the first claim. Now, observe that the two functions have gradient

$$\nabla g_i(X) = 2[I - X(X^TX)^{-1}X^T]e_ie_i^TX(X^TX)^{-1}, \qquad \nabla f(X) = 4(XX^T - ZZ^T)X.$$

Directly substituting yields

$$\begin{aligned}\frac{1}{8}\langle \nabla g_i(X), \nabla f(X)(X^TX)^{-1}\rangle &= \langle [I - X(X^TX)^{-1}X^T]e_ie_i^TX(X^TX)^{-1}, (XX^T - ZZ^T)X(X^TX)^{-1}\rangle \\ &= e_i^TX(X^TX)^{-2}X^TZZ^TX(X^TX)^{-1}X^Te_i - e_i^TZZ^TX(X^TX)^{-2}X^Te_i\end{aligned}$$

where the second line follows from the fact that

$$\langle [I - X(X^TX)^{-1}X^T]e_ie_i^TX(X^TX)^{-1}, XX^TX(X^TX)^{-1}\rangle = 0.$$

The second claim follows from the following three identities

$$\lambda_{\min}(X^TX) \geq (1-\rho)\lambda_{\min}(Z^TZ) \tag{17}$$

$$e_i^TZZ^TX(X^TX)^{-2}X^Te_i \leq \frac{1}{1-\rho}\|e_i^TX\|_X^* \cdot \|e_i^TZ\|_Z^* \tag{18}$$

$$e_i^TX(X^TX)^{-2}X^TZZ^TX(X^TX)^{-1}X^Te_i \geq \frac{1-2\rho}{1-\rho} \cdot (\|e_i^TX\|_X^*)^2 \tag{19}$$

We have (17) via Weyl's inequality:

$$\lambda_{\min}(X^TX) = \lambda_r(XX^T) = \lambda_r(ZZ^T + XX^T - ZZ^T) \geq \lambda_r(ZZ^T) - \|XX^T - ZZ^T\|_F.$$

We have (18) by rewriting

$$\begin{aligned} e_i^T ZZ^T X(X^TX)^{-2}X^T e_i &= (e_i^T P)(P^T ZZ^T X(X^TX)^{-2}X^T Q)(Q^T e_i) \\ &\leq \|e_i^T P\| \|ZZ^T X(X^TX)^{-2}X^T\| \|e_i^T Q\| \end{aligned}$$

and rewriting $ZZ^T = XX^T - E$ where $E = XX^T - ZZ^T$ and evoking (17) as in

$$\|ZZ^T X(X^TX)^{-2}X^T\| \leq \underbrace{\|XX^T X(X^TX)^{-2}X^T\|}_{=1} + \underbrace{\|E\| \cdot \|X(X^TX)^{-2}X^T\|}_{\leq \rho/(1-\rho)}$$

and noting that $1 + \frac{\rho}{1-\rho} = \frac{1}{1-\rho}$. We have (19) again by substituting $ZZ^T = XX^T - E$

$$\begin{aligned} e_i^T X(X^TX)^{-2}X^T ZZ^T X(X^TX)^{-1}X^T e_i &= e_i^T X(X^TX)^{-1}X^T e_i - e_i^T X(X^TX)^{-2}X^T E X(X^TX)^{-1}X^T e_i \\ &\geq e_i^T X(X^TX)^{-1}X^T e_i \cdot (1 - \underbrace{\|E\| \cdot \|X(X^TX)^{-2}X^T\|}_{\leq \rho/(1-\rho)}) \end{aligned}$$

and then noting that $1 - \frac{\rho}{1-\rho} = \frac{1-2\rho}{1-\rho}$. □

E.4 Proof of the main result

In the previous two subsections, we showed that when conditioned on the current iterate X_t, a single step of ScaledSGD $X_{t+1} = X_t - \alpha SG(X_t)(X_t^TX_t)^{-1}$ is expected to geometrically converge both the loss function f and each of the incoherence functions g_i, as in

$$\mathbb{E}[f(X_{t+1})] \leq (1-\alpha)f(X_t), \quad \mathbb{E}[g_i(X_{t+1}) - \frac{1}{2}g_{\max}] \leq \left(1 - \frac{1-2\rho}{1-\rho}\alpha\right)\left[g_i(X_t) - \frac{1}{2}g_{\max}\right].$$

In this section, we will extend this geometric convergence to T iterations of ScaledSGD. Our key challenge is to verify that the *variances* and *maximum deviations* of the sequences $f(X_0), f(X_1), \ldots, f(X_T)$ and $g_i(X_0), g_i(X_1), \ldots, g_i(X_T)$ have the right dependence on the dimension d, the radius ρ, the condition number κ, the maximum coherence $g_{\max}$, and the iteration count t, so that T iterations of ScaledSGD with a step-size of $\alpha \leq c/[(g_{\max} + \rho)d^2 \log d]$ results in no more than a multiplicative factor of 2 deviation from expectation. Crucially, we must check that the cumulated deviation over T iterations does not grow with the iteration count T, and that the convergence rate is independent of the condition number κ. We emphasize that the actual approach of our proof via the Azuma–Bernstein inequality is textbook; to facilitate a direct comparison with SGD, we organize this section to closely mirror Jin et al. [10]'s proof of Theorem 1.

Let $f_{\max} = \rho^2 \cdot \lambda_{\min}^2(Z^TZ)$ and $g_{\max} = \frac{16}{(1-2\rho)^2}\max_i g_i(Z)$. Our goal is to show that the following event happens with probability $1 - T/d^{10}$:

$$\mathfrak{E}_t \equiv \left\{ f(X) \leq \left(1 - \frac{\alpha}{2}\right)^t \cdot f_{\max}, \quad \max_i g_i(X_\tau) \leq g_{\max} \quad \text{for all } \tau \leq t \right\}, \tag{20}$$

Equivalently, conditioned on event $\mathfrak{E}_t$, we want to prove that the probability of failure at time $t+1$ is $\delta \equiv 1/d^{10}$. We split this failure event into a probability of $\frac{\delta}{2}$ that the function value clause fails to hold, as in $f(X_{t+1}) > (1-\alpha/2)^t \cdot f_{\max}$, and a probability of $\frac{\delta}{2d}$ that any one of the d incoherence caluses fails to hold, as in $g_i(X_{t+1}) > g_{\max}$. Then, cumulated over T steps, the total probability of failure would be $T \cdot \delta = T/d^{10}$ as desired.

We begin by setting up a supermartingale on the loss function f. Our goal is to show that the variance and the maximum deviation of this supermartingale have the right dependence on $\alpha, d, \rho, \kappa, g_{\max}$, so that a step-size of $\alpha \leq c/(g_{\max}d^2 \log d)$ with a sufficiently small $c > 0$ will keep the cumulative deviations over T iterations within a factor of 2. Note that, by our careful choice of the coherence function g_i, the following statement for ScaledSGD match the equivalent statements for SGD with a perfect condition number $\kappa = 1$; see Jin et al. [10, Section B.2].

Lemma 8 (Function value supermartingale)**.** *Let $f(X) = \|XX^T - ZZ^T\|_F^2$. Define $f_{\max} = \rho^2 \cdot \lambda_{\min}^2(Z^TZ)$ and $g_{\max} = \frac{16}{(1-2\rho)^2}\max_i g_i(Z)$. For a sufficiently small $c > 0$, the following with learning rate $\alpha \le c/(g_{\max} d^2 \log d)$ is a supermartingale*

$$F_t = (1-\alpha)^{-t} f(X_t) \cdot 1_{\mathfrak{E}_t},$$

meaning that $\mathbb{E}[F_{t+1}|X_t, \dots, X_0] \le F_t$ holds for all $t \in \{0, 1, 2, \dots\}$. Moreover, there exist sufficiently large constants $C_{\text{dev}}, C_{\text{var}} > 0$ such that the following holds with probability one:

$$\mathbb{E}[F_t|X_{t-1}, \dots, X_0] - F_t \le C_{\text{dev}} \cdot \alpha \cdot d^2 \cdot g_{\max} \cdot (1-\alpha)^{-t}\left(1 - \frac{\alpha}{2}\right)^t f_{\max},$$
$$\mathbf{Var}[F_t|X_{t-1}, \dots, X_0] \le C_{\text{var}} \cdot \alpha^2 \cdot d^2 \cdot g_{\max} \cdot (1-\alpha)^{-2t}\left(1 - \frac{\alpha}{2}\right)^{2t} f_{\max}^2.$$

Proof. The proof is technical but straightforward; it is deferred to Section E.5. □

Lemma 9 (Function value concentration)**.** *Let the initial point satisfy $f(X_0) \le \frac{1}{2} f_{\max}$. Then, there exists a sufficiently small constant $c > 0$ such that for all learning rates $\alpha < c/(g_{\max} d^2 \log d)$, we have*

$$\mathbf{Pr}\left(f_i(X_{t+1})1_{\mathfrak{E}_t} > \left(1 - \frac{\alpha}{2}\right)^t f_{\max}\right) = \mathbf{Pr}\left(\mathfrak{E}_t \cap \left\{f_i(X_{t+1}) > \left(1 - \frac{\alpha}{2}\right)^t f_{\max}\right\}\right) \le \frac{1}{2d^{10}}.$$

Proof. Let $\sigma^2 = \sum_{\tau=1}^t \mathbf{Var}[F_\tau|X_{\tau-1}, \dots, X_0]$ and let R satisfy $\mathbb{E}[F_\tau|X_{\tau-1}, \dots, X_0] \le X_\tau + R$ almost surely for all $\tau \in \{1, 2, \dots, t\}$. Recall via the standard Azuma–Bernstein concetration inequality for supermartingales that $\mathbf{Pr}(F_t \ge F_0 + s) \le \exp\left(-\frac{s^2/2}{\sigma^2 + Rs/3}\right)$. Equivalently, there exists a large enough constant $C > 0$ in $s = C \cdot (1-\alpha)^t \left[\sqrt{\sigma^2 \log d} + R \log d\right]$ such that the following is true

$$\mathbf{Pr}\left(f(X_{t+1})1_{\mathfrak{E}_t} \ge (1-\alpha)^t f(X_0) + s\right) \le \frac{1}{2d^{10}}.$$

Given that $f(X_0) \le \frac{1}{2} f_{\max}$ and therefore $(1-\alpha)^t f(X_0) \le \frac{1}{2}\left(1 - \frac{\alpha}{2}\right)^t \cdot f_{\max}$ holds by hypothesis, the desired claim is true if we can show that $s \le \frac{1}{2}\left(1 - \frac{\alpha}{2}\right)^t \cdot f_{\max}$. Crucially, we observe that the variance term in s does not blow-up with time t

$$\begin{aligned}
(1-\alpha)^{2t} \cdot \sigma^2 &\le f_{\max}^2 \cdot C_{\text{var}} \cdot d^2 \cdot g_{\max} \cdot \alpha^2 \cdot \sum_{\tau=1}^t (1-\alpha)^{2t-2\tau}\left(1 - \frac{\alpha}{2}\right)^{2\tau} \\
&= \left(1 - \frac{\alpha}{2}\right)^{2t} f_{\max}^2 \cdot C_{\text{var}} \cdot d^2 \cdot g_{\max} \cdot \alpha^2 \cdot \sum_{\tau=1}^t \left(\frac{1-\alpha}{1-\alpha/2}\right)^{2t-2\tau} \\
&\le \left(1 - \frac{\alpha}{2}\right)^{2t} f_{\max}^2 \cdot C_{\text{var}} \cdot d^2 \cdot g_{\max} \cdot \alpha
\end{aligned}$$

due to the geometric series expansion $\sum_{\tau=0}^t \beta^{t-\tau} = (1-\beta^{t+1})/(1-\beta)$. Substituting the deviations term, choosing a step-size $\alpha \le c/(\rho d^2 \log d)$ for sufficiently small c yields

$$\begin{aligned}
s &= \left(1 - \frac{\alpha}{2}\right)^t \cdot \left[\sqrt{C_{\text{var}} \cdot d^2 \cdot g_{\max} \cdot \alpha \cdot \log d} + C_{\text{dev}} \cdot d^2 \cdot g_{\max} \cdot \alpha \cdot \log d\right] \cdot f_{\max} \\
&\le \frac{1}{2}\left(1 - \frac{\alpha}{2}\right)^t f_{\max}.
\end{aligned}$$

□

We now set up a supermartingale on each of the incoherence functions g_i. Again, our goal is to show that the variance and the maximum deviation of this supermartingale have the right dependence on $\alpha, d, \rho, \kappa, g_{\max}$, so that a step-size of $\alpha \le c/(\rho d^2 \log d)$ with a sufficiently small $c > 0$ will keep the cumulative deviations over T iterations within a factor of 2. Note that Jin et al. [10, Section B.2]'s proof tracks a different function $h_i(X) = e_i^T XX^T e_i$ that is substantially simpler, but pays a penalty of two to three factors of the condition number κ.

Lemma 10 (Incoherence supermartingale)**.** *Let $g_i(X) = e_i^T X(X^T X)^{-1} X^T e_i$. Define $g_{\max} = \frac{16}{(1-2\rho)^2} \max_i g_i(Z)$. For a fixed $i \in [n]$ with sufficiently small $c > 0$, the following with learning rate $\alpha < c/(\rho d^2 \log d)$ is a supermartingale*

$$G_{it} = (1 - \zeta \cdot \alpha)^{-t} \left(g(X_t) \cdot 1_{\mathfrak{E}_{t-1}} - \frac{\zeta}{2} g_{\max} \right) \text{ where } \zeta = \frac{1-2\rho}{1-\rho} < 1,$$

meaning that $\mathbb{E}[G_{i(t+1)} | X_t, \ldots, X_0] \leq G_{it}$ holds for all $t \in \{0, 1, 2, \ldots\}$. Moreover, there exist sufficiently large constants $C_{dev}, C_{var} > 0$ with no dependence on $g_{\max}, n, t$ such that

$$\mathbb{E}[G_{it} | X_{t-1}, \ldots, X_0] - G_{it} \leq C_{\text{dev}} \cdot \alpha \cdot d^2 \cdot \rho \cdot (1 - \zeta \cdot \alpha)^{-t} g_{\max},$$
$$\mathbf{Var}[G_{it} | X_{t-1}, \ldots, X_0] \leq C_{\text{var}} \cdot \alpha^2 \cdot d^2 \cdot \rho^2 \cdot (1 - \zeta \cdot \alpha)^{-2t} g_{\max}^2.$$

Proof. The proof is long but straightforward; it is deferred to Section E.5. □

Lemma 11 (Incoherence concentration)**.** *Let the initial point satisfy $\max_i g_i(X_0) \leq \frac{1}{2} g_{\max}$. Then, there exists a sufficiently small constant $c > 0$ such that for all learning rates $\alpha < c/(\rho d^2 \log d)$, we have*

$$\mathbf{Pr}(g_i(X_{t+1}) 1_{\mathfrak{E}_t} > g_{\max}) = \mathbf{Pr}(\mathfrak{E}_t \cap \{g_i(X_{t+1}) > g_{\max}\}) \leq \frac{1}{2d^{11}}. \tag{21}$$

Proof. Let $\sigma^2 = \sum_{\tau=1}^t \mathbf{Var}[G_{i\tau} | X_{\tau-1}, \ldots, X_0]$ and let R satisfy $\mathbb{E}[G_{i\tau} | X_{\tau-1}, \ldots, X_0] \leq X_\tau + R$ almost surely for all $\tau \in \{1, 2, \ldots, t\}$. Recall via the standard Azuma–Bernstein concetration inequality for supermartingales that

$$\mathbf{Pr}\left(G_{it} \geq G_{i0} + s\right) \leq \exp\left(-\frac{s^2/2}{\sigma^2 + Rs/3} \right).$$

Equivalently, there exists a large enough constant $C > 0$ such that the following is true

$$\mathbf{Pr}\left(g_i(X_{t+1}) 1_{\mathfrak{E}_t} \geq \frac{1}{2} g_{\max} + (1 - \zeta \cdot \alpha)^t \left(g(X_0) - \frac{1}{2} g_{\max} \right) + s' \right) \leq \frac{1}{2d^{11}}$$
$$\text{where } s' = C \cdot (1 - \zeta \cdot \alpha)^t \cdot \left[\sqrt{\sigma^2 \log d} + R \log d \right].$$

Given that $g(X_0) \leq \frac{1}{2} g_{\max}$ holds by hypothesis, the desired claim is true if we can show that $s' \leq \frac{1}{2} g_{\max}$. Crucially, we observe that the variance term in s' does not blow-up with time t

$$\begin{aligned}(1 - \zeta \cdot \alpha)^{2t} \sigma^2 &\leq g_{\max}^2 \cdot C_{\text{var}} \cdot d^2 \cdot \rho^2 \cdot \alpha^2 \sum_{\tau=1}^t (1 - \zeta \cdot \alpha)^{2t-2\tau} \\ &\leq g_{\max}^2 \cdot C_{\text{var}} \cdot d^2 \cdot \rho^2 \cdot \alpha\end{aligned}$$

due to the geometric series expansion $\sum_{\tau=0}^t \beta^{t-\tau} = (1 - \beta^{t+1})/(1 - \beta)$. Substituting the deviations term, choosing a step-size $\alpha \leq c/(\rho d^2 \log d)$ for sufficiently small c yields

$$s' = O\left(\sqrt{\alpha \cdot d^2 \cdot \rho^2 \cdot g_{\max}^2 \cdot \log d} \right) + O\left(\alpha \cdot d^2 \cdot \rho \cdot g_{\max} \cdot \log d \right) = \frac{g_{\max}}{2}.$$

□

In summary, Lemma 9 requires a step-size of $\alpha \leq c/(g_{\max} d^2 \log d)$ to keep deviations on f small, while Lemma 11 requires a step-size of $\alpha \leq c/(\rho d^2 \log d)$ to keep deviations on g_i small. Therefore, it follows that a step-size $\alpha \leq c/((g_{\max} + \rho) d^2 \log d)$ will keep both deviations small.

Proof of Theorem 2. For a step-size $\alpha \leq c/((g_{\max} + \rho) d^2 \log d)$ with sufficiently small $c > 0$, both concentration bounds Lemma 9 and Lemma 11 are valid. Combined, we take the trivial union bound to determine the probability of *failure* at the $(t+1)$-th step, after succeeding after t steps:

$$\begin{aligned}\mathbf{Pr}(\mathfrak{E}_t \cap \overline{\mathfrak{E}}_{t+1}) &= \sum_{i=1}^d \mathbf{Pr}(\mathfrak{E}_t \cap \{g_i(X_{t+1}) \geq g_{\max}\}) + \mathbf{Pr}(\mathfrak{E}_t \cap \{f(X_{t+1}) \geq (1 - \frac{\alpha}{2})^{t+1} f_{\max}). \\ &\leq d \cdot \frac{1}{2d^{11}} + \frac{1}{2d^{10}} = \frac{1}{d^{10}}.\end{aligned}$$

Here, $\overline{\mathfrak{E}}_{t+1}$ denotes the complement of $\mathfrak{E}_{t+1}$. The probability of failure at the T-th step is then the cummulative probability of failing at the $(t+1)$-th step, after succeeding after t steps, over all $t \leq T$:

$$\mathbf{Pr}(\overline{\mathfrak{E}}_T) \leq \sum_{t=1}^{T} \mathbf{Pr}(\mathfrak{E}_{t-1} \cap \overline{\mathfrak{E}}_t) \leq \frac{T}{d^{10}}$$

and this proves that $\mathfrak{E}_T$ happens with probability $1 - T/d^{10}$ as desired. □

E.5 Proofs of supermartingale deviations and variances

We will now verify the supermartingales and their deviations and variances in detail. We first begin by proving the following bounds on the size of the stochastic gradient.

Lemma 12. *Let $X, Z \in \mathbb{R}^{n\times r}$ satisfy $\|XX^T - ZZ^T\|_F \leq \rho \cdot \lambda_{\min}(Z^TZ)$ with $\rho < 1/2$ and $\max_i e_i^T X(X^TX)^{-1}Xe_i \leq g_{\max}$ and $\max_i e_i^T Z(Z^TZ)^{-1}Ze_i \leq g_{\max}$. Then, with respect to the randomness of the following*

$$SG(X) = 2d^2 \cdot (XX^T - ZZ^T)_{i,j} \cdot (e_ie_j^T + e_je_i^T)X$$

where $(i,j) \sim [d]^2$ is selected uniformly at random, we have:

1. $\|SG(X)\|_X^* \leq 8d^2 \cdot g_{\max}^{1/2} \cdot \|XX^T - ZZ^T\|_F$.
2. $\|SG(X)(X^TX)^{-1}\|_X^* \leq 16d^2 \cdot g_{\max}^{1/2} \cdot \rho$.
3. $\mathbb{E}(\|SG(X)\|_X^*)^p \leq 2^{2p} \cdot d^{2(p-1)} \cdot g_{\max}^{p/2} \cdot \|XX^T - ZZ^T\|_F^p$.
4. $\mathbb{E}(\|SG(X)(X^TX)^{-1}\|_X^*)^p \leq 2^{3p} \cdot d^{2(p-1)} \cdot g_{\max}^{p/2} \cdot \rho^p$.

Proof. Let us write $E = XX^T - ZZ^T$. To prove (i) we have

$$\|SG(X)\|_X^* = 2d^2 \cdot E_{i,j} \cdot (\|(e_ie_j^T + e_je_i^T)X\|_X^*) \leq 4d^2 \cdot E_{i,j} \cdot \max_i \|e_i^TX\|_X^*$$

and if we write $Q_X = X(X^TX)^{-1/2}$ and $Q_Z = Z(Z^TZ)^{-1/2}$ we have

$$\begin{aligned} E_{i,j} = e_i^TEe_j &= e_i^TQ_XQ_X^T(XX^T - ZZ^T)e_j - e_i^T(I - Q_XQ_X^T)ZZe_j \\ &\leq \|e_k^TQ_X\|\|Q_X^T(XX^T - ZZ^T)e_j\| + \|e_i^T(I - Q_XQ_X^T)ZZ^TQ_Z\|\|Q_Z^Te_j\|, \\ &\leq g_{\max}^{1/2} \cdot \|XX^T - ZZ^T\|_F + \|XX^T - ZZ^T\|_F \cdot g_{\max}^{1/2}. \end{aligned}$$

we use the fact that $(I - Q_XQ_X^T)(XX^T - ZZ^T) = -(I - Q_XQ_X^T)ZZ^T$ in the first and last lines. To prove (ii) we have

$$\|SG(X)(X^TX)^{-1}\|_X^* = \frac{\|SG(X)\|_X^*}{\lambda_{\min}(X^TX)} \leq \frac{8d^2 \cdot g_{\max}^{1/2} \cdot \rho \cdot \lambda_{\min}(Z^TZ)}{(1-\rho) \cdot \lambda_{\min}(Z^TZ)} \leq 16d^2 \cdot g_{\max}^{1/2} \cdot \rho$$

where we used Weyl's inequality $\lambda_r(XX^T) \geq \lambda_r(ZZ^T) - \|XX^T - ZZ^T\|_F$. To prove (iii) we have

$$\mathbb{E}(\|SG(X)\|_X^*)^p = \frac{1}{d^2}\sum_{i,j} d^{2p} \cdot (2E_{i,j})^p \cdot (\|(e_ie_j^T + e_je_i^T)X\|_X^*)^p \leq 2^{2p} \cdot d^{2(p-1)} \cdot \|E\|_F^p \cdot \max_i(\|e_i^TX\|_X^*)^p$$

where we used $(\sum_i x_i^2)^{1/2} \geq (\sum_i x_i^p)^{1/p}$ for any $p \geq 2$. The proof of (iv) follows identically by applying the proof of (ii) to the proof of (iii). □

We now prove the properties of the function value supermartingale F_t.

Proof of Lemma 8. Conditioning on the current iterate X_t and the event $\mathfrak{E}_t$, the new iterate $X_{t+1} = X_t - \alpha SG(X_t)(X_t^TX_t)^{-1}$ has expectation

$$\mathbb{E}[f(X_{t+1})] \leq f(X_t) - \alpha\langle\nabla f(X_t), \mathbb{E}[SG(X_t)(X_t^TX_t)^{-1}]\rangle + \frac{L_X}{2}\alpha^2 \cdot \mathbb{E}[(\|SG(X_t)\|_X^*)^2]$$

with $L_X = O(1)$ by evoking Lemma 3 noting that $\|\alpha SG(X_t)\|_X^* = O(1)\cdot\sqrt{f(X_t)}$ for the step-size $\alpha \le c/(g_{\max} d^2 \log d)$, since

$$\begin{aligned}\|\alpha SG(X_t)\|_X^* &= \alpha \cdot 2d^2 \cdot (X_t X_t^T - ZZ^T)_{i,j} \|(e_i e_j^T + e_j e_i^T) X_t\|_X^* \\ &\le \alpha \cdot 2d^2 \cdot |X_t X_t^T - ZZ^T|_\infty \cdot 2\sqrt{g_{\max}} \\ &\le \frac{c}{g_{\max} d^2 \log d} \cdot 2d^2 \cdot \sqrt{f(X_t) g_{\max}} \cdot 2\sqrt{g_{\max}} = \frac{4c}{\log d}\sqrt{f(X_t)}.\end{aligned}$$

The linear term evaluates simply as $\mathbb{E}[SG(X)(X^T X)^{-1}] = \nabla f(X)(X^T X)^{-1}$, while the quadratic term evaluates

$$\begin{aligned}\mathbb{E}[(\|SG(X)\|_X^*)^2] &= \frac{1}{d^2}\sum_{i,j} 4d^4 \cdot (XX^T - ZZ^T)_{i,j}^2 (\|(e_i e_j^T + e_j e_i^T) X\|_X^*)^2 \\ &\le \sum_{i,j} 4d^2 \cdot (XX^T - ZZ^T)_{i,j}^2 \cdot 4 g_{\max} = 16 \cdot g_{\max} \cdot d^2 \cdot f(X)\end{aligned}$$

Combined, substituting $(\|\nabla f(X)\|_X^*)^2 \ge 13 \cdot f(X)$, it follows that we have geometric convergence

$$\begin{aligned}\mathbb{E}[f(X_{t+1})] &\le f(X_t) - \alpha \langle \nabla f(X_t), \mathbb{E}[SG(X_t)(X_t^T X_t)^{-1}]\rangle + \frac{L_X}{2}\alpha^2 \cdot \mathbb{E}[(\|SG(X_t)\|_X^*)^2] \\ &\le (1-2\alpha) f(X_t) + L_X \cdot \alpha^2 \cdot 8 \cdot g_{\max} \cdot d^2 \cdot f(X_t) \le (1-\alpha) f(X_t)\end{aligned}$$

where we observe that we can pick a small enough constant c in the step-size $\alpha \le c/(g_{\max} d^2 \log d)$ so that

$$L_X \cdot \alpha^2 \cdot 8 \cdot g_{\max} \cdot d^2 \cdot f(X_t) = \frac{c \cdot L_X \cdot 8 \cdot g_{\max} \cdot d^2}{g_{\max} d^2 \log d}\alpha f(X_t) \le \alpha f(X_t).$$

Now, to confirm that F_t is a martingale, it remains to see that

$$\begin{aligned}\mathbb{E}[F_{t+1}|X_t] &= (1-\alpha)^{-(t+1)}\,\mathbb{E}[f(X_{t+1}) \cdot 1_{\mathfrak{E}_t}|X_t] \le (1-\alpha)^{-(t+1)}\,(1-\alpha)\, f(X_t) 1_{\mathfrak{E}_t} \\ &\le (1-\alpha)^{-t} f(X_t) 1_{\mathfrak{E}_{t-1}} = F_t,\end{aligned}$$

where the last inequality follows from $1_{\mathfrak{E}_t} \le 1_{\mathfrak{E}_{t-1}}$.

We now bound the deviations on F_t. Conditioning on the previous iterates $X_t, \dots, X_0$, we obseve that the $f(X_t)$ terms cancel:

$$\begin{aligned}f(X_{t+1}) \cdot 1_{\mathfrak{E}_t} - \mathbb{E}[f(X_{t+1}) \cdot 1_{\mathfrak{E}_t}] \le \Big[&-\alpha\langle \nabla f(X_t), [SG(X_t) - \mathbb{E}SG(X_t)](X_t^T X_t)^{-1}\rangle \\ &+ \frac{\alpha^2 \cdot L_X}{2}(\|SG(X_t)\|_X^*)^2 + \mathbb{E}(\|SG(X_t)\|_X^*)^2\Big] \cdot 1_{\mathfrak{E}_t}.\end{aligned} \tag{22}$$

Here we have for the linear term

$$\begin{aligned}\langle \nabla f(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle \cdot 1_{\mathfrak{E}_t} &\le \|\nabla f(X_t)\|_X^* \|SG(X_t)\|_X^* \cdot 1_{\mathfrak{E}_t} \\ &\le 4\sqrt{f(X_t)} \cdot 4d^2\sqrt{f(X_t)} \cdot g_{\max} \cdot 1_{\mathfrak{E}_t} = O(d^2 g_{\max}) f(X_t) \cdot 1_{\mathfrak{E}_t}\end{aligned}$$

and the quadratic term

$$(\|SG(X_t)\|_X^*)^2 \cdot 1_{\mathfrak{E}_t} \le d^4 \cdot f(X_t) \cdot g_{\max}^2 \cdot 1_{\mathfrak{E}_t} = O(d^4 g_{\max}^2) f(X_t) \cdot 1_{\mathfrak{E}_t}.$$

Therefore, using the maximum value to bound the expectation, we have

$$\begin{aligned}F_{t+1} - \mathbb{E}[F_{t+1}|X_t, \dots, X_0] &\le \alpha\,(1-\alpha)^{-t} \cdot \big[O(d^2 g_{\max}) f(X_t) + \alpha O(d^4 g_{\max}^2) f(X_t)\big] \cdot 1_{\mathfrak{E}_t} \\ &\le \alpha\,(1-\alpha)^{-t} \cdot O(d^2 g_{\max}) f(X_t) \cdot 1_{\mathfrak{E}_t} \\ &\le C_{\text{dev}} \cdot \alpha\,(1-\alpha)^{-t}\left(1 - \frac{\alpha}{2}\right)^t f_{\max} \cdot d^2 \cdot g_{\max} \cdot 1_{\mathfrak{E}_t}\end{aligned}$$

where again we observe that a step-size like $\alpha \le c/(g_{\max} d^2 \log d) = O(d^{-2} g_{\max}^{-1})$ yields the cancellation of exponents $\alpha \cdot O(d^4 g_{\max}^2) = O(d^2 g_{\max})$.

Finally, we bound the variance. Conditioned on all previous iterates $X_t, \ldots, X_0$ we have

$$\begin{aligned}\mathbf{Var}(\langle \nabla f(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle \cdot 1_{\mathfrak{E}_t}) &\leq \mathbb{E}[\langle \nabla f(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle^2 \cdot 1_{\mathfrak{E}_t}] \\ &\leq \; (\|\nabla f(X_t)\|_{X_t}^*)^2 \cdot \mathbb{E}[(\|SG(X_t)\|_X^*)^2] \cdot 1_{\mathfrak{E}_t} \leq O(d^2 g_{\max}) \cdot f(X_t)^2 \cdot 1_{\mathfrak{E}_t},\end{aligned}$$

and also

$$\mathbf{Var}((\|SG(X_t)\|_X^*)^2 \cdot 1_{\mathfrak{E}_t}) \leq \mathbb{E}[(\|SG(X_t)\|_X^*)^4 \cdot 1_{\mathfrak{E}_t}] = O(d^6 g_{\max}^3) f(X_t)^2 \cdot 1_{\mathfrak{E}_t}.$$

By the same expansion in (22) we have

$$\begin{aligned}\mathbf{Var}(F_{t+1}|X_t, \ldots, X_0) &\leq \alpha^2 (1-\alpha)^{-2t} \cdot \left[O(d^2 g_{\max}) \cdot f(X_t)^2 + \alpha^2 O(d^6 g_{\max}^3) f(X_t)^2\right] \cdot 1_{\mathfrak{E}_t} \\ &\leq \alpha^2 (1-\alpha)^{-2t} \cdot O(d^2 g_{\max}) \cdot f(X_t)^2 \cdot 1_{\mathfrak{E}_t} \\ &\leq C_{\mathrm{var}} \cdot \alpha^2 (1-\alpha)^{-2t} \left(1 - \frac{\alpha}{2}\right)^{2t} f_{\max}^2 \cdot d^2 \cdot g_{\max} \cdot 1_{\mathfrak{E}_t}\end{aligned}$$

where again we observe that a step-size like $\alpha \leq c/(g_{\max} d^2 \log d) = O(d^{-2} g_{\max}^{-1})$ yields the cancellation of exponents $\alpha^2 \cdot (d^6 g_{\max}^3) = O(d^2 g_{\max})$. □

We now prove properties of the incoherence martingale.

Proof of Lemma 10. Conditioning on X_t and the event $\mathfrak{E}_t$, we have for $V = SG(X_t)(X_t^T X_t)^{-1}$

$$\begin{aligned}\mathbb{E}[g_i(X_{t+1})] &\leq g_k(X_t) - \alpha \langle \nabla g_i(X_t), \mathbb{E}[V]\rangle + \alpha^2 \cdot \mathbb{E}\left[\frac{(\|V\|_X^*)^2}{1 - 2\|V\|_X^*}\right] \\ &\leq \left(1 - \frac{1-2\rho}{1-\rho}\alpha\right) g_i(X_t) + \alpha \cdot \frac{1}{1-\rho} \cdot \sqrt{g_i(X) g_i(Z)} + \alpha^2 \cdot \frac{\mathbb{E}\left[(\|V\|_X^*)^2\right]}{1 - 2\|V\|_X^*} \\ &\leq \left(1 - \frac{1-2\rho}{1-\rho}\alpha\right) g_i(X_t) + \alpha \cdot \frac{\sqrt{g_i(Z)/g_{\max}}}{1-\rho} \cdot g_{\max} + \alpha^2 \cdot \frac{O(d^2 \cdot g_{\max} \cdot \rho^2)}{1 - O(g_{\max}^{1/2} \cdot \rho)} \\ &\leq (1 - \zeta\alpha)\, g_i(X_t) + \alpha \cdot \frac{\zeta}{2} g_{\max} \qquad \text{for } \alpha = O(\rho^{-1} d^{-2}).\end{aligned}$$

Here we note that we have carefully chosen $g_{\max}$ so that the ratio $\max_i g_i(Z)/g_{\max} = [(1-\rho) \cdot \zeta/4]^2$. It then follows that the following is a supermartingale

$$G_{it} = (1 - \zeta\alpha)^{-t} \left(g_i(X_t) \cdot 1_{\mathfrak{E}_{t-1}} - \frac{\zeta}{2} g_{\max}\right).$$

Indeed, we have

$$\begin{aligned}\mathbb{E}[G_{i(t+1)}|X_t, \ldots, X_0] &= (1 - \zeta\alpha)^{-(t+1)} \left(\mathbb{E}[g_i(X_{t+1}) \cdot 1_{\mathfrak{E}_t}|X_t, \ldots, X_0] - \frac{\zeta}{2} g_{\max}\right) \\ &\leq (1 - \zeta\alpha)^{-(t+1)} \left[(1 - \zeta\alpha)\, g_i(X_t) \cdot 1_{\mathfrak{E}_t} + \alpha \cdot \frac{\zeta}{2} g_{\max} \cdot 1_{\mathfrak{E}_t} - \frac{\zeta}{2} g_{\max}\right] \\ &\leq (1 - \zeta\alpha)^{-t} \left[g_i(X_t) \cdot 1_{\mathfrak{E}_{t-1}} - \left(\frac{1-\alpha}{1-\alpha}\right) \cdot \frac{\zeta}{2} g_{\max}\right] = G_{it}\end{aligned}$$

where the final line uses $1_{\mathfrak{E}_t} \leq 1_{\mathfrak{E}_{t-1}} \leq 1$.

We now bound the deviations on G_{it}. Conditioning on the previous iterates $X_t, \ldots, X_0$, we obseve that the $g_i(X_t)$ terms cancel:

$$\begin{aligned}g_i(X_{t+1}) \cdot 1_{\mathfrak{E}_t} - \mathbb{E}[g_i(X_{t+1}) \cdot 1_{\mathfrak{E}_t}] \leq \Big[&-\alpha \langle \nabla g_i(X_t), [SG(X_t) - \mathbb{E}SG(X_t)](X_t^T X_t)^{-1}\rangle \\ &+ 5\alpha^2 \cdot \frac{(\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^2 + \mathbb{E}(\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^2}{1 - \alpha\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*}\Big] \cdot 1_{\mathfrak{E}_t}. \quad (23)\end{aligned}$$

Here we have for the linear term

$$\|\nabla g(X)\| \cdot 1_{\mathfrak{E}_t} \leq \|2[I - X(X^T X)^{-1} X^T] e_i e_i^T X (X^T X)^{-1}\| \cdot 1_{\mathfrak{E}_t} \leq O(\kappa^{1/2} \cdot g_{\max}^{1/2})$$

and $\|SG(X)\|_X^* \cdot 1_{\mathfrak{E}_t} \leq O(d^2 \cdot \sqrt{g_{\max} f(X)}) = O(d^2\sqrt{g_{\max}}\rho/\kappa)$ noting that $f_{\max} = \rho^2/\kappa^2$ and hence

$$\begin{aligned}|\langle \nabla g(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle| &\leq \|\nabla g(X_t)\| \cdot \|SG(X_t)\|_X^* \\ &\leq O(\sqrt{\kappa g_{\max}}) \cdot O(d^2\sqrt{g_{\max}}\rho/\kappa) = O(g_{\max} \cdot d^2\rho) \cdot 1_{\mathfrak{E}_t}.\end{aligned}$$

We have for the quadratic term

$$\frac{5(\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^2}{1-\alpha\|SG(X_t)(X^T X)^{-1}\|_X^*} \cdot 1_{\mathfrak{E}_t} \leq \frac{O(d^4 \cdot g_{\max} \cdot \rho^2)}{1-\alpha \cdot O(d^2 \cdot g_{\max}^{1/2} \cdot \rho)} \cdot 1_{\mathfrak{E}_t} = O(g_{\max} \cdot d^4\rho^2) \cdot 1_{\mathfrak{E}_t}.$$

Therefore, using the maximum value to bound the expectation, we have

$$\begin{aligned}G_{i(t+1)} - \mathbb{E}[G_{i(t+1)}|X_t, \ldots, X_0] &\leq \alpha\,(1-\alpha)^{-t}\, g_{\max} \cdot \left[O(d^2\rho) + \alpha O(d^4\rho^2) f(X_t)\right] \cdot 1_{\mathfrak{E}_t} \\ &\leq \alpha\,(1-\alpha)^{-t}\, g_{\max} \cdot O(d^2\rho) \cdot 1_{\mathfrak{E}_t}\end{aligned}$$

where again we observe that a step-size like $\alpha \leq c/(\rho d^2 \log d) = O(d^{-2}\rho^{-1})$ yields the cancellation of exponents $\alpha \cdot O(d^4\rho^2) = O(d^2\rho)$.

Finally, we bound the variance. Conditioned on all previous iterates $X_t, \ldots, X_0$ we have

$$\begin{aligned}\mathbf{Var}(\langle \nabla g_i(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle \cdot 1_{\mathfrak{E}_t}) &\leq \mathbb{E}[\langle \nabla g_i(X_t), SG(X_t)(X_t^T X_t)^{-1}\rangle^2 \cdot 1_{\mathfrak{E}_t}] \\ &\leq \ (\|\nabla g_i(X_t)\|)^2 \cdot \mathbb{E}[(\|SG(X_t)\|_X^*)^2] \cdot 1_{\mathfrak{E}_t} \leq O(g_{\max}^2 \cdot d^2\rho^2) \cdot 1_{\mathfrak{E}_t},\end{aligned}$$

and also

$$\begin{aligned}\mathbf{Var}\left(\frac{5(\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^2}{1-\alpha\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*} \cdot 1_{\mathfrak{E}_t}\right) &\leq \frac{25 \cdot \mathbb{E}[(\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^4]}{(1-\alpha\|SG(X_t)(X_t^T X_t)^{-1}\|_X^*)^2} \cdot 1_{\mathfrak{E}_t} \\ &\leq \frac{O(g_{\max}^2 \cdot d^6\rho^4)}{1-\alpha \cdot O(d^2 \cdot g_{\max}^{1/2} \cdot \rho)} \cdot 1_{\mathfrak{E}_t} = O(g_{\max}^2 \cdot d^6\rho^4) \cdot 1_{\mathfrak{E}_t}.\end{aligned}$$

By the same expansion in (23) we have

$$\begin{aligned}\mathbf{Var}(G_{i(t+1)}|X_t, \ldots, X_0) &\leq \alpha^2\,(1-\alpha)^{-2t} \cdot \left[O(g_{\max}^2 \cdot d^2\rho^2) \cdot 1_{\mathfrak{E}_t} + \alpha^2 O(g_{\max}^2 \cdot d^6\rho^4) \cdot 1_{\mathfrak{E}_t},\right] \cdot 1_{\mathfrak{E}_t} \\ &\leq \alpha^2\,(1-\alpha)^{-2t}\, g_{\max}^2 \cdot O(d^2\rho^2) \cdot 1_{\mathfrak{E}_t}\end{aligned}$$

where again we observe that a step-size like $\alpha \leq c/(\rho d^2 \log d) = O(d^{-2}\rho^{-1})$ yields the cancellation of exponents $\alpha^2 \cdot (d^6\rho^4) = O(d^2\rho^2)$. □