Perfect Parallelization in Mini-Batch SGD with Classical Momentum Acceleration

Sachin Garg
University of Michigan
sachg@umich.edu

Michał Dereziński
University of Michigan
derezin@umich.edu

Abstract

Accelerating stochastic gradient methods with classical momentum schemes, such as Polyak's heavy ball, has proven highly successful in training large-scale machine learning models, particularly when combined with the hardware acceleration of large mini-batch computations. Yet, the effect of classical momentum on stochastic mini-batch optimization has been poorly understood theoretically, with prior works requiring strong noise assumptions and extremely large mini-batches. In this work, we develop a general theory of stochastic momentum acceleration for optimizing over quadratics in the interpolation regime, a popular abstraction for studying deep learning dynamics which also includes classical methods such as randomized Kaczmarz and coordinate descent. Our framework encompasses both heavy ball and Nesterov-style momentum, allows for arbitrary mini-batch sizes, and makes minimal assumptions on the stochastic noise. In particular, we show that acceleration from classical momentum is directly proportional to the gradient mini-batch size (up to a natural saturation point), thereby enabling perfect parallelization of mini-batch computations. Our theory also provides a simple choice for the momentum parameter, which is shown to be effective empirically.

1 Introduction

Consider the following optimization task which naturally arises in training machine learning models:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathbb{E}_{\xi\sim\mathcal{D}}\big[f(\mathbf{w};\xi)\big], \tag{1}$$

where $\mathbf{w}$ represents the model parameters and $\xi \sim \mathcal{D}$ represents the stochastic noise. Problem (1) provides a powerful abstraction for designing and analyzing algorithms for model training, particularly in the context of deep learning and large language models [31, 23, 13, 50].

When given oracle access to the full gradient $\mathbf{g}(\mathbf{w}) = \nabla F(\mathbf{w})$ of the function F, problem (1) is very well understood. Here, the simplest strategy is to iteratively follow the direction of steepest descent,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\,\mathbf{g}(\mathbf{w}_t). \tag{2}$$

When F is smooth and strongly convex, this Gradient Descent algorithm (GD) attains linear convergence, i.e., it reaches an ϵ-approximate solution after $O(\kappa_F \log(1/\epsilon))$ iterations, where κ_F is an appropriate condition number of the deterministic function F. In 1964, Polyak [42] showed that the convergence rate of GD can be improved by introducing momentum into its descent trajectory:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\,\mathbf{g}(\mathbf{w}_t) + \beta\,(\mathbf{w}_t - \mathbf{w}_{t-1}). \tag{3}$$

With an appropriate choice of step size α and momentum parameter β, this Heavy Ball Momentum algorithm (HBM) reaches an ϵ-approximate solution after $O(\sqrt{\kappa_F}\log(1/\epsilon))$ iterations. In 1983, Nesterov [39] proposed a refinement of this scheme, known as Nesterov Accelerated Gradient (NAG), which essentially incorporates one additional gradient step to both $\mathbf{w}_t$ and $\mathbf{w}_{t-1}$ in the momentum.

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

For large-scale problems, instead of the full gradient we often only have access to stochastic gradients $\nabla f(\mathbf{w}, \xi)$. By aggregating m such gradients into the vector $\mathbf{g}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \nabla f(\mathbf{w}; \xi_i)$, the GD update (2) turns into the popular mini-batch Stochastic Gradient Descent algorithm (SGD, [43]), which has become a workhorse of modern machine learning. The success of SGD in deep learning has been partly attributed to the fact that it is particularly effective in the *interpolation regime* [4, 37, 3], i.e., when all of the functions $f(\cdot, \xi)$ can be simultaneously minimized (the model fits perfectly to the training data). In this setting, for smooth and strongly convex objectives, SGD attains linear convergence with $O(\kappa_f \log(1/\epsilon))$ iteration complexity, where κ_f is a corresponding condition number of the stochastic function f. The use of large mini-batch size m allows us to leverage parallel computing architectures while reducing the variance of the stochastic gradient [33], however its effect on the iteration complexity of vanilla SGD is limited [7] and not fully understood in practice [17].

A natural strategy for improving the performance of mini-batch SGD is to incorporate the classical momentum update (3) of Polyak, or its Nesterov counterpart. This approach has proven highly successful empirically, particularly when paired with large mini-batches [46, 23, 44, 15], and it inspired state-of-the-art training algorithms such as Adam [30] as well as many other more elaborate acceleration schemes for SGD [21, 52, 28]. However, this has led to a fundamental theory-practice gap: While there is significant theoretical progress on accelerating SGD through carefully designed schemes [11, 22, 1, 2, 27, 26, 48, 34, 6, 54, 51], the SGD algorithms used in practice still largely rely on the simple and classical momentum based on HBM, whose remarkable effectiveness remains poorly understood. Initial efforts in that direction were done by Loizou and Richtarik [36], who showed linear convergence of SGD with HBM, but without attaining any acceleration. Bollapragada, Chen and Ward [5] showed that HBM-accelerated mini-batch SGD can exhibit faster convergence on quadratic tasks, with related and follow-up works by [32, 40, 47, 56]. However, in order to attain this acceleration, all of these prior works require either strong additional assumptions on the stochastic noise or prohibitively large mini-batch sizes that depend heavily on problem parameters.

1.1 Our contributions

In this work, we address the above theory-practice gap by developing a general framework for analyzing the convergence of momentum-based stochastic processes, which encompasses mini-batch SGD with both HBM and NAG acceleration for stochastic quadratic minimization in the interpolation regime. Our analysis leads to a strikingly simple conclusion about the role of classical momentum in large-scale stochastic optimization:

Classical momentum enables perfect parallelization for mini-batch SGD in the interpolation regime.

Main result. Concretely, our main technical result (Theorem 1) shows that properly tuned SGD using mini-batch size m and either HBM or NAG momentum attains $O\big((\kappa_f/m + \sqrt{\kappa_f})\log(\kappa_f/\epsilon)\big)$ iteration complexity in the interpolation regime. Note that for $m = 1$ (no mini-batching) this result essentially recovers the previously mentioned linear convergence for SGD, but as we increase m, the iteration complexity improves proportionally to $1/m$ until it saturates at an accelerated $\sqrt{\kappa_f}$ rate. Thus, up to the point of saturation, the overall work (number of stochastic gradients computed) remains constant, allowing us to fully exploit the embarrassingly parallel mini-batch computations. Such perfect parallelization does not in general hold for mini-batch SGD without acceleration, even over quadratics in the interpolation regime. Thus, our analysis provides an explanation for the success of momentum acceleration in allowing deep learning optimizers such as Adam to benefit from large mini-batches and highly parallelizable GPU architectures.

Applications. Our analysis framework can also be directly applied to accelerate several popular SGD algorithms for solving large-scale regression tasks, such as randomized Kaczmarz [45] and coordinate descent [19]. In particular, we use our theory to develop a simple momentum-accelerated block coordinate descent algorithm, and we evaluate it on several kernel ridge regression tasks. Our numerical results confirm the perfect parallelization phenomenon, and they also show that a simple choice of $\beta = 1 - 1/m$, as motivated by our theory, already leads to meaningful acceleration. We also compare classical momentum with an optimally tuned state-of-the-art acceleration scheme for block coordinate descent [11], showing that, surprisingly, in certain large mini-batch regimes classical momentum yields substantially better performance.

1.2 Overview of our techniques

Limitations of prior approaches. The convergence behavior of a momentum-accelerated stochastic method follows a dynamical system that is characterized by a product of non-symmetric random matrices. Existing matrix concentration results can be used to bound the spectral norm of a product of random matrices in terms of the product of the spectral norms of the expected values of the matrices [25]. Unfortunately, this strategy proves vacuous for momentum acceleration, as the product of norms diverges to infinity even when the norm of the product converges to zero, because of the lack of symmetry. To address this, [5] diagonalize the matrices so that they can use matrix product concentration effectively. However, passing from the original matrices to the diagonalized form incurs a penalty factor that appears both in the convergence bound and in the required mini-batch size, and can blow up independently of the condition number of the problem. In fact, the penalty can even be infinite, as the underlying matrices may not be diagonalizable. To get around this, [40] diagonalize the matrices to a Jordan form, however this still does not avoid a potentially large penalty factor in both the convergence bound and the mini-batch size.

Our approach. In order to get around these issues, we abandon both the diagonalization step and matrix product concentration. Instead, we use the Schur decomposition together with a matrix recursion to bound the original matrix product directly. Our analysis consists of two key components, following a bias-variance decomposition of the expected error.

1. Bias: *Convergence of expected iterates via Schur decomposition*. For the expected iterates, it suffices to analyze the product of certain *deterministic* non-symmetric matrices. Here, we start by rearranging the matrices into a block-diagonal form, and carefully analyze both their real and complex eigenvalues in order to bound the spectral radius. Next, we turn this into a spectral norm bound for the matrix product. In this step, instead of using diagonalization as done in prior work, we rely on the Schur decomposition [24], thereby avoiding condition number factors in our bound.

2. Variance: *Convergence of expected error via matrix recursion*. In order to control the random iterates, we must bound the stochastic error incurred at each iteration of the process. Instead of relying on black-box matrix concentration results, we do this by deriving a recursion for the expectation of the product of the random matrices, which relates its spectral norm to all monomials of the individual block matrices from the block-diagonal representations derived in the bias analysis. Unfolding this recursion proves quite technical as it requires separately handling real and complex eigenvalues of the diagonal blocks.

Our analysis is not specific to one algorithm, but rather provides a general framework for effectively characterizing the convergence behavior of stochastic momentum methods. We illustrate this by stating our results for a class of stochastic processes which includes both HBM and NAG as special cases, and can be instantiated to provide convergence guarantees for a variety of SGD-type algorithms.

2 Stochastic momentum acceleration framework

Notation. We represent vectors by lowercase boldface letters and matrices by uppercase boldface letters. For symmetric matrices, we use Loewner ordering and say $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is a positive semidefinite (psd) matrix. $\|\mathbf{A}\|$ denotes spectral norm, $\|\mathbf{A}\|_F$ denotes Frobenius norm of matrix $\mathbf{A}$. For a psd matrix $\mathbf{A}$, $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$ denotes the Mahalanobis norm of $\mathbf{x}$. We let $\mathbf{A}^\dagger$ denote the Moore-Penrose pseudoinverse of $\mathbf{A}$. We use $\tilde{O}$, $\tilde{\Omega}$ notation to hide logarithmic factors.

Stochastic contraction process. We start by defining a class of stochastic dynamical systems called stochastic contraction processes [9]. This framework can be instantiated to represent the dynamics of many stochastic algorithms, as described below, and thus serves as a useful abstraction for our theory.

Definition 1. *For a random psd matrix $\mathbf{\Pi}$ such that $\mathbf{0} \preceq \mathbf{\Pi} \preceq \mathbf{I}$ and scalar $\alpha \in [0, 1]$, we let $\mathcal{A}(\mathbf{\Delta}; \alpha) := (\mathbf{I} - \alpha\,\mathbf{\Pi})\mathbf{\Delta}$ define a stochastic contraction with average rate $\bar{\mathbf{\Pi}} = \mathbb{E}[\mathbf{\Pi}]$. We call $\mathbf{\Pi}$ the stochastic rate matrix of $\mathcal{A}$. Given a sequence of independent $\mathcal{A}_t$ with the same $\bar{\mathbf{\Pi}}$ and α, we call $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$ a stochastic contraction process.*

To see how this definition corresponds to an SGD algorithm, consider the minimization problem (1) over $F(\mathbf{w}) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{w};\xi)]$, where for every ξ, function $f(\mathbf{w};\xi)$ is a convex quadratic with

respect to $\mathbf{w}$ that is minimized at some $\mathbf{w}^* \in \mathbb{R}^d$. As long as function $f(\cdot;\xi)$ is L-smooth, i.e., $\|\nabla f(\mathbf{w},\xi) - \nabla f(\mathbf{v},\xi)\| \leq L\|\mathbf{w}-\mathbf{v}\|$, then the evolution of the error vector $\mathbf{\Delta}_t = \mathbf{w}_t - \mathbf{w}^*$ resulting from the SGD update (2) with $\mathbf{g}(\mathbf{w}_t) = \frac{1}{L}\nabla f(\mathbf{w}_t;\xi_t)$ follows a stochastic contraction process:

$$\mathbf{\Delta}_{t+1} = \mathcal{A}_t(\mathbf{\Delta}_t;\alpha), \qquad \text{where} \qquad \mathcal{A}_t(\mathbf{\Delta};\alpha) = \Big(\mathbf{I} - \frac{\alpha}{L}\nabla^2 f(\mathbf{w}_t;\xi_t)\Big)\mathbf{\Delta}. \tag{4}$$

The stochastic rate matrix defining the stochastic contraction process $\mathcal{A}_t$ is $\mathbf{\Pi}_t = \frac{1}{L}\nabla^2 f(\mathbf{w}_t;\xi_t)$, and its average rate is $\bar{\mathbf{\Pi}} = \frac{1}{L}\nabla^2 F(\mathbf{w}^*)$. Note that the Hessian is the same everywhere since both $f(\cdot;\xi_t)$ and F are assumed to be quadratic.

Stochastic momentum process. We next show how to introduce momentum into this process by incorporating information from two preceding steps. We define these dynamics in sufficient generality to capture both HBM and NAG momentum as special cases.

Definition 2. *Given a stochastic contraction process $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$, initial vectors $\mathbf{\Delta}_{-1}, \mathbf{\Delta}_0$, and parameters $\beta, \omega \geq 0$, define the corresponding stochastic momentum process recursively as:*

$$\mathbf{\Delta}_{t+1} = \mathcal{A}_t(\mathbf{\Delta}_t;\alpha) + \beta\Big[\mathcal{A}_t(\mathbf{\Delta}_t;\omega) - \mathcal{A}_{t-1}(\mathbf{\Delta}_{t-1};\omega)\Big]. \tag{5}$$

Note that if we set the momentum parameter β to zero, then we recover the usual SGD dynamics of a stochastic contraction process. On the other hand, if we set $\beta > 0$ but $\omega = 0$, then the momentum term simplifies to $\beta[\mathbf{\Delta}_t - \mathbf{\Delta}_{t-1}]$, which precisely corresponds to Polyak's heavy-ball momentum (HBM). On the other hand, by choosing $\omega = \alpha$ we introduce one more SGD step to the momentum term, thereby recovering Nesterov's momentum (NAG). For example, we can use this framework to model HBM acceleration for the SGD process $\{\mathcal{A}_t\}_{t\geq 0}$ from (4) by letting

$$\mathbf{\Delta}_t = \mathbf{w}_t - \mathbf{w}^*, \quad \text{and} \quad \mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\alpha}{L}\nabla f(\mathbf{w}_t;\xi_t) + \beta(\mathbf{w}_t - \mathbf{w}_{t-1}).$$

It is easy to verify that such a sequence $\{\mathbf{\Delta}_t\}_{t\geq 0}$ is a stochastic momentum process associated with $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$, momentum parameter β, and $\omega = 0$. A similar derivation follows immediately for Nesterov's momentum by setting $\omega = \alpha$.

Example 1: Randomized Kaczmarz. As a concrete example of an algorithm that falls directly under our framework, we consider the randomized Kaczmarz (RK) algorithm [45], along with its block versions [38]. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$ define a linear regression problem with the loss function $F(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2$. Also, let $\mathcal{D}$ be a distribution over subsets $S \subseteq \{1, ..., n\}$, and consider the following update:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}:\, \mathbf{A}_{S_t}\mathbf{w} = \mathbf{y}_{S_t}} \|\mathbf{w} - \mathbf{w}_t\|^2 = \mathbf{w}_t - \mathbf{A}_{S_t}^\dagger(\mathbf{A}_{S_t}\mathbf{w}_t - \mathbf{y}_{S_t}), \qquad S_t \sim \mathcal{D}, \tag{6}$$

where $\mathbf{A}_{S_t}$ denotes the submatrix consisting of rows indexed by S_t. In particular, when $S \sim \mathcal{D}$ generates single-sample sets such that $\Pr[S = \{i\}] \propto \|\mathbf{A}_{i,:}\|^2$, then this recovers classical randomized Kaczmarz. When there exists $\mathbf{w}^*$ such that $\mathbf{A}\mathbf{w}^* = \mathbf{y}$ (i.e., in the interpolation regime), the update (6) can be cast as a stochastic contraction process $(\{\mathcal{A}_t\}_{t\geq 0}, 1)$ by letting $\mathbf{\Pi}_t = \mathbf{A}_{S_t}^\dagger \mathbf{A}_{S_t}$ and $\mathbf{\Delta}_t = \mathbf{w}_t - \mathbf{w}^*$. In particular, for randomized Kaczmarz, a simple calculation shows that this process has an average rate $\bar{\mathbf{\Pi}} = \mathbb{E}[\mathbf{\Pi}_t] = \mathbf{A}^\top\mathbf{A}/\|\mathbf{A}\|_F^2$. Similar or better rates have been derived for block versions of Kaczmarz [11], and other extensions such as sketch-and-project [12]. By applying momentum via Definition 2 to $(\{\mathcal{A}_t\}_{t\geq 0}, 1)$, we can precisely recover convergence dynamics of the error vector of the accelerated Kaczmarz algorithm considered by [5].

Example 2: Coordinate Descent. Another application of our framework comes in the context of solving positive semidefinite linear systems, such as in kernel ridge regression. For a positive definite matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ and $\mathbf{y} \in \mathbb{R}^n$, consider the (block) coordinate descent (CD) update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{I}_{S_t}^\top(\mathbf{K}_{S_t,S_t})^\dagger(\mathbf{K}\mathbf{w}_t - \mathbf{y})_{S_t}, \qquad S_t \sim \mathcal{D}, \tag{7}$$

which only updates the coordinates of $\mathbf{w}_t$ indexed by S_t. Letting $\mathbf{w}^*$ denote the solution of the linear system $\mathbf{K}\mathbf{w} = \mathbf{y}$, we can cast (7) as a stochastic contraction process $(\{\mathcal{A}_t\}_{t\geq 0}, 1)$ by choosing $\mathbf{\Delta}_t = \mathbf{K}^{1/2}(\mathbf{w}_t - \mathbf{w}^*)$ and $\mathbf{\Pi}_t = \mathbf{K}_{S_t}^{1/2}(\mathbf{K}_{S_t,S_t})^\dagger\mathbf{K}_{S_t}^{1/2}$. Here, the canonical distribution $\mathcal{D}$ for single-sample coordinate descent is $\Pr[S = \{i\}] \propto K_{i,i}$, which yields average rate $\bar{\mathbf{\Pi}} = \mathbf{K}/\mathrm{tr}(\mathbf{K})$.

Convergence rate of a stochastic contraction process. The above examples are of base algorithms, i.e., without momentum acceleration. Before we analyze their accelerated variants through the stochastic momentum framework, we establish their base convergence rate via our general formalism.

Proposition 1. *Given a stochastic contraction process* $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$ *with average rate* $\bar{\mathbf{\Pi}}$*, and an initial vector* $\mathbf{\Delta}_0 \in \mathrm{range}(\bar{\mathbf{\Pi}})$*, consider the sequence* $\mathbf{\Delta}_{t+1} = \mathcal{A}_t(\mathbf{\Delta}_t; \alpha)$*. Then, we have:*

$$\mathbb{E}\,\|\mathbf{\Delta}_t\|^2 \leq \left(1 - \frac{1}{\kappa}\right)^t \|\mathbf{\Delta}_0\|^2, \qquad \textit{where} \quad \kappa = 1/\alpha\lambda^+_{\min}(\bar{\mathbf{\Pi}}).$$

We will refer to κ as the stochastic condition number associated with the process $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$. For instance, by plugging in our concrete examples of (single-sample) randomized Kaczmarz and coordinate descent [45], we recover the classical convergence guarantees for these methods, with stochastic condition numbers $\kappa_{\mathrm{RK}} = \|\mathbf{A}\|_F^2\|\mathbf{A}^\dagger\|^2$ and $\kappa_{\mathrm{CD}} = \mathrm{tr}(\mathbf{K})\|\mathbf{K}^\dagger\|$.

Effect of mini-batching. The stochastic contraction framework is naturally amenable to mini-batching of the updates via simple averaging of the contractions. It is thus natural to ask how mini-batching affects the convergence dynamics.

Definition 3. *Given a stochastic contraction* $\mathcal{A}$ *with average rate* $\bar{\mathbf{\Pi}}$*, let* $\mathcal{A}^{[m]}$ *denote the average of* m *independent copies of* $\mathcal{A}$*, i.e.,*

$$\mathcal{A}^{[m]}(\mathbf{\Delta}; \alpha) := \frac{1}{m}\sum_{i=1}^m \mathcal{A}^{(i)}(\mathbf{\Delta}; \alpha), \qquad \textit{where} \quad \mathcal{A}^{(i)} \sim \mathcal{A}. \tag{8}$$

For example, if $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$ is a stochastic contraction process for the SGD algorithm given by (4), then $(\{\mathcal{A}_t^{[m]}\}_{t\geq 0}, \alpha)$ corresponds to mini-batch SGD where $\mathbf{g}(\mathbf{w}_t) = \frac{1}{mL}\sum_{i=1}^m \nabla f(\mathbf{w}_t, \xi_t^{(i)})$. Note that mini-batching is different from choosing a larger block set S_t in randomized Kaczmarz or coordinate descent, since those schemes do more than just averaging the samples (such as inverting the matrices $\mathbf{A}_{S_t}$ and $\mathbf{K}_{S_t,S_t}$). In fact, combining blocking and mini-batching in RK and CD is a very natural strategy for making these methods scalable to parallel architectures, as discussed later.

Proposition 2. *If* $\mathcal{A}$ *is a stochastic contraction with average rate* $\bar{\mathbf{\Pi}}$*, then* $\mathcal{A}^{[m]}$ *is also a stochastic contraction with the same average rate* $\bar{\mathbf{\Pi}}$*. Moreover, if* $\mathbf{\Pi}^{[m]}$ *is the stochastic rate matrix associated with* $\mathcal{A}^{[m]}$*, then it follows that:*

$$\mathbb{E}\Big[\big(\mathbf{\Pi}^{[m]} - \bar{\mathbf{\Pi}}\big)^2\Big] \preceq \frac{1}{m}\cdot\bar{\mathbf{\Pi}}(\mathbf{I} - \bar{\mathbf{\Pi}}). \tag{9}$$

Proposition 2, together with Proposition 1, illustrates why mini-batching of SGD does not in general lead to faster convergence rates, since the average rate matrix $\bar{\mathbf{\Pi}}$ (and thus also the associated stochastic condition number) remains the same regardless of m. However, mini-batching does reduce the variance of the stochastic noise matrix, as shown in (9), which will be crucial for the stochastic momentum dynamics. We defer the proofs of Propositions 1 and 2 to Appendix E.

3 Main result

We are now ready to state our main result, which characterizes the convergence dynamics of a stochastic momentum process (Definition 2) corresponding to stochastic contractions with mini-batching (Definitions 1 and 3). Crucially, our result makes no assumptions on the mini-batch size m or the stochastic noise, and applies to both heavy-ball and Nesterov momentum.

Theorem 1. *Consider a stochastic contraction process* $(\{\mathcal{A}_t\}_{t\geq 0}, \alpha)$ *with* $\alpha \in [0, 1]$*, average rate* $\bar{\mathbf{\Pi}}$*, and condition number* $\kappa = 1/\alpha\lambda^+_{\min}(\bar{\mathbf{\Pi}})$*. For* $m \geq 1$*,* $\mathbf{\Delta}_0 \in \mathrm{range}(\bar{\mathbf{\Pi}})$*, let* $\mathbf{\Delta}_{-1} = \mathbf{0}$*,* $\{\mathbf{\Delta}_t\}_{t\geq 0}$ *be a stochastic momentum process for* $(\{\mathcal{A}_t^{[m]}\}_{t\geq 0}, \alpha)$ *with* $\omega \in [0, \alpha]$ *and momentum parameter*

$$\beta = \begin{cases} 0 & \textit{if} \quad m \leq c_1, \\ 1 - \frac{c_1}{2m} & \textit{if} \quad m \in [c_1, c_2\sqrt{\kappa}], \\ 1 - \frac{c_1}{2c_2\sqrt{\kappa}} & \textit{if} \quad m \geq c_2\sqrt{\kappa}, \end{cases}$$

where $c_1, c_2 > 0$ *are absolute constants. Then, there are absolute constants* $C, c > 0$ *such that:*

$$\mathbb{E}\,\|\mathbf{\Delta}_t\|^2 \leq Ct^3\left(1 - c\min\left\{\frac{m}{\kappa}, \frac{1}{\sqrt{\kappa}}\right\}\right)^t \|\mathbf{\Delta}_0\|^2 \qquad \textit{for any } t \geq 1.$$

Remark 1. *In particular, we get* $\mathbb{E}\,\|\mathbf{\Delta}_t\|^2 \le \epsilon\|\mathbf{\Delta}_0\|^2$ *after* $t = O((\kappa/m + \sqrt{\kappa})\log(\kappa/\epsilon))$ *iterations, compared to* $t = O(\kappa\log(1/\epsilon))$ *without momentum. Thus, for* $m \in [1, \Theta(\sqrt{\kappa})]$*, momentum enables perfect parallelization up to a logarithmic factor (which is caused by the* t^3 *term in the bound).*

Remark 2. *In the regime* $m \in [1, \Theta(\sqrt{\kappa})]$*, our result yields a simple problem-independent momentum parameter* $\beta = 1 - \Theta(1/m)$*. Numerical experiments confirm that this choice works well in practice.*

Corollary (Mini-batch SGD with HBM for quadratic interpolation). Let $F(\mathbf{w})$ be as defined in (1), where for every ξ, $f(\cdot;\xi)$ is a convex quadratic with L-Lipschitz gradient, and suppose there exists $\mathbf{w}^\star \in \mathbb{R}^d$ such that $\nabla f(\mathbf{w}^\star;\xi) = 0$ almost surely. Let F be μ-strongly convex and $\kappa_f = L/\mu$. At every iteration, draw an independent mini-batch $\xi_t^{(i)}$ for $i = 1, \dots, m$ and let

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \beta(\mathbf{w}_t - \mathbf{w}_{t-1}) - \frac{1}{mL}\sum_{i=1}^{m} \nabla f\big(\mathbf{w}_t; \xi_t^{(i)}\big).$$

With $\mathbf{w}_{-1} = \mathbf{w}_0 = 0$, choose the momentum parameter β as in Theorem 1, with $\kappa = \kappa_f$. Then there exist absolute constants $C, c > 0$ such that, for every $t \ge 1$,

$$\mathbb{E}\|\mathbf{w}_t - \mathbf{w}^\star\|^2 \le Ct^3\left(1 - c\min\left\{\frac{m}{\kappa_f}, \frac{1}{\sqrt{\kappa_f}}\right\}\right)^t \|\mathbf{w}^\star\|^2.$$

3.1 Proof sketch

First, without loss of generality we can assume that $\alpha = 1$ (otherwise, we can simply rescale $\bar{\mathbf{\Pi}}$ appropriately). Also, note that the regime of $1 \le m \le c_1$ follows immediately from Proposition 1. Next, let $\{\mathbf{\Pi}_t^{[m]}\}_{t\ge 0}$ be the stochastic noise matrices associated with $\{\mathcal{A}_t^{[m]}\}_{t\ge 0}$. From Proposition 2, these matrices satisfy the variance bound (9). In fact, this is the only property we need out of the mini-batching, so to simplify the notation, we will drop the superscript and consider a sequence $\{\mathbf{\Pi}_t\}_{t\ge 0}$ of matrices such that $\mathbb{E}\,(\mathbf{\Pi}_t - \bar{\mathbf{\Pi}})^2 \preceq \frac{1}{m}\bar{\mathbf{\Pi}}(\mathbf{I} - \bar{\mathbf{\Pi}})$. With these conventions, after simple calculations the stochastic momentum process (5) can be expressed as the following recursion:

$$\mathbf{\Delta}_{t+1} = \big((1+\beta)\mathbf{I} - (1+\beta\omega)\,\mathbf{\Pi}_t\big)\mathbf{\Delta}_t - \beta\big(\mathbf{I} - \omega\,\mathbf{\Pi}_{t-1}\big)\mathbf{\Delta}_{t-1}. \tag{10}$$

For notational inconvenience, we restrict ourselves to NAG acceleration in this proof sketch (i.e. $\omega = \alpha = 1$). The detailed proof in the appendix covers HBM and NAG momentum simultaneously. Let $\bar{\mathbf{\Pi}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the eigendecomposition of $\bar{\mathbf{\Pi}}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$. Let $\mathrm{rank}(\bar{\mathbf{\Pi}}) = r \le d$, so that $\lambda_i = 0$ for $i > r$. Following prior works [26, 5, 40], we study the dynamics of stochastic process $\mathbf{\Delta}_t$ in (10) via the following two-step matrix transition rule:

$$\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t+1}\\ \mathbf{V}^\top\mathbf{u}_{t+1}\end{bmatrix} = \underbrace{\begin{bmatrix}(1+\beta)(\mathbf{I} - \mathbf{V}^\top\mathbf{\Pi}_t\,\mathbf{V}) & -\beta\mathbf{I}\\ \mathbf{I} - \mathbf{V}^\top\mathbf{\Pi}_t\,\mathbf{V} & \mathbf{0}\end{bmatrix}}_{:=\mathbf{Y}_t} \cdot \begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_t\\ \mathbf{V}^\top\mathbf{u}_t\end{bmatrix}, \tag{11}$$

where $\mathbf{u}_{t+1} = (\mathbf{I} - \mathbf{\Pi}_t)\mathbf{\Delta}_t$ and $\mathbf{u}_0 = 0$. We break down the proof of Theorem 1 into two main components, starting with the convergence of the expected iterates (bias), and then analyzing the effect of the stochastic noise (variance).

3.1.1 Bias: Convergence of $\|\mathbb{E}\mathbf{\Delta}_t\|$

The convergence behavior of the expected iterates is characterized by a simpler transition rule:

$$\mathbb{E}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t+1}\\ \mathbf{V}^\top\mathbf{u}_{t+1}\end{bmatrix} = \mathbf{T}^{t+1}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_0\\ \mathbf{V}^\top\mathbf{u}_0\end{bmatrix}, \quad \text{where } \mathbf{T} := \mathbb{E}[\mathbf{Y}_t] = \begin{bmatrix}(1+\beta)(\mathbf{I} - \mathbf{\Lambda}) & -\beta\mathbf{I}\\ \mathbf{I} - \mathbf{\Lambda} & \mathbf{0}\end{bmatrix}. \tag{12}$$

The first thing that comes to mind is to use $\big\|\mathbf{T}^t\| \le \|\mathbf{T}\|^t$, but, unfortunately, for non-zero β the spectral norm of $\mathbf{T}$ is greater than 1, making this naive idea ineffective. However, the spectral radius of $\mathbf{T}$ is less than 1 and this paves the way ahead, suggesting a two step approach:

Step 1: Bound the spectral radius of $\mathbf{T}^t$, denoted as $\rho(\mathbf{T}^t)$;

Step 2: Using structural properties of $\mathbf{T}$, bound $\big\|\mathbf{T}^t\|$ in terms of $\rho(\mathbf{T}^t)$.

Step 1. By carefully permuting rows and columns of $\mathbf{T}$, we can transform it into a block-diagonal matrix with 2 by 2 blocks on the diagonal. In particular, the following lemma holds:

Lemma 1 (Block-diagonalization)**.** *There exists a permutation matrix* $\mathbf{P}$ *such that,*

$$\mathbf{P}\mathbf{T}\mathbf{P}^\top = \mathrm{diag}(\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_d), \quad \textit{where } \mathbf{T}_i := \begin{bmatrix} (1+\beta)(1-\lambda_i) & -\beta \\ 1-\lambda_i & 0 \end{bmatrix}.$$

Due to Lemma 1, taking maximum over the spectral radius of each $\mathbf{T}_i$ will upper bound $\rho(\mathbf{T})$. Interestingly, the eigenvalues of $\mathbf{T}_i$ depend on interaction between λ_i and β, and can be either real or complex depending upon whether $\lambda_i \leq \frac{(1-\beta)^2}{(1+\beta)^2}$. Let γ_i denote the eigenvalue of $\mathbf{T}_i$ with the larger magnitude, so that $|\gamma_i|$ equals the spectral radius of $\mathbf{T}_i$. We prove the following result:

Lemma 2 (Spectral radius of $\mathbf{T}_i$)**.** *Let* $\beta = 1 - \frac{1}{\phi}$ *with* $\phi \geq 2$*. Then,*

$$|\gamma_i|^2 \leq \begin{cases} \big(1-\frac{1}{\phi}\big)(1-\lambda_i), & \textit{if } \lambda_i \geq \frac{(1-\beta)^2}{(1+\beta)^2} \textit{ (i.e., when } \gamma_i \textit{ is complex)}, \\ \big(1-\frac{\phi\lambda_i}{2}\big)^2(1-\lambda_i)^2, & \textit{otherwise (i.e., when } \gamma_i \textit{ is real).} \end{cases}$$

It is easy to see that the maximum is attained at $i = r$, thus yielding an upper bound on $\rho(\mathbf{T})$.

Step 2. To go from $\rho(\mathbf{T}^t)$ to $\big\|\mathbf{T}^t\big\|$, previous work [5] assumes that $\mathbf{T}$ is diagonalizable, i.e., there exists an invertible matrix $\mathbf{C}$ such that $\mathbf{C}\mathbf{T}\mathbf{C}^{-1}$ is a diagonal matrix. This assumption implies,

$$\big\|\mathbf{T}^t\big\| \leq \big\|\mathbf{C}\big\| \cdot \big\|\mathbf{C}^{-1}\big\| \cdot \rho(\mathbf{T}^t).$$

Unfortunately, $\mathbf{T}$ need not be diagonalizable. In fact, if for any i, $\lambda_i = \frac{(1-\beta)^2}{(1+\beta)^2}$, then $\mathbf{T}$ is not diagonalizable. More problematically, it can be shown that

$$\|\mathbf{C}\| \cdot \|\mathbf{C}^{-1}\| = \Omega\Big(\frac{\phi}{\delta}\Big), \quad \text{where } \delta = \min_i \Big\{\big|\lambda_i - \frac{(1-\beta)^2}{(1+\beta)^2}\big|\Big\}.$$

To avoid this blow up, we abandon the diagonalization approach and rely on a fundamental matrix decomposition result in linear algebra, known as Schur's decomposition.

Lemma 3 (Schur's decomposition [24])**.** *For every* i*, there exists a unitary matrix* $\mathbf{U}_i$ *such that*

$$\mathbf{U}_i^H \mathbf{T}_i \mathbf{U}_i = \begin{bmatrix} \gamma_{i1} & x_i \\ 0 & \gamma_{i2} \end{bmatrix},$$

where γ_{i1} *and* γ_{i2} *are the two eigenvalues of* $\mathbf{T}_i$*,* $1 \leq |x_i| \leq 3$ *and* $\mathbf{U}_i^H$ *denotes the hermitian of* $\mathbf{U}_i$*.*

Using Schur's decomposition it readily follows that for any i,

$$\|\mathbf{T}_i^t\| = \|\mathbf{U}_i^H \mathbf{T}_i^t \mathbf{U}_i\| \leq O(t) \cdot |\gamma_i|^{t-1}.$$

Using the above in conjunction with Lemma 2 we show that:

Theorem 2 (Convergence of $\|\mathbb{E}\boldsymbol{\Delta}_t\|$)**.** *For* $\rho = \big(1-\frac{\phi}{2\kappa}\big)\big(1-\frac{1}{\kappa}\big)$*, the transition rule (11) satisfies,*

$$\|\mathbb{E}\boldsymbol{\Delta}_{t+1}\|^2 \leq O(t^2) \cdot \rho^t \|\boldsymbol{\Delta}_0\|^2.$$

3.1.2 Variance: Convergence of $\mathbb{E}\|\boldsymbol{\Delta}_t\|^2$

Note that (11) implies,

$$\mathbf{P}\begin{bmatrix} \mathbf{V}^\top \boldsymbol{\Delta}_t \\ \mathbf{V}^\top \mathbf{u}_t \end{bmatrix} = \underbrace{\Big(\prod_{j=0}^{t-1} \mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\Big)}_{\boldsymbol{\Sigma}_t} \cdot \mathbf{P}\begin{bmatrix} \mathbf{V}^\top \boldsymbol{\Delta}_0 \\ \mathbf{V}^\top \mathbf{u}_0 \end{bmatrix}.$$

To establish convergence of $\mathbb{E}\|\boldsymbol{\Delta}_t\|^2$, it suffices to bound $\|\mathbb{E}[\boldsymbol{\Sigma}_t^\top \boldsymbol{\Sigma}_t]\|$. Since $\boldsymbol{\Sigma}_t$ is a product of non-symmetric random matrices, upper bounding $\big\|\mathbb{E}[\boldsymbol{\Sigma}_t^\top \boldsymbol{\Sigma}_t]\big\|$ is challenging. Prior work [5] uses matrix product concentration [25] and require careful tuning of the momentum parameter to keep $\mathbf{T}$ diagonalizable while controlling $\|\mathbf{C}\| \cdot \|\mathbf{C}^{-1}\|$. This leads to mini-batch requirements scaling as $\|\mathbf{C}\|^2 \cdot \|\mathbf{C}^{-1}\|^2$, which can become arbitrarily large near non-diagonalizability, and introduces additional condition number dependence in m. Consequently, their guarantees degrade rapidly unless the mini-batch size is sufficiently large. These limitations stem from the reliance on product concentration results and diagonalization. In contrast, we avoid both and directly bound $\|\mathbb{E}[\boldsymbol{\Sigma}_t^\top \boldsymbol{\Sigma}_t]\|$.

Recursive approach to bound $\|\mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t]\|$. Let $\bar{\mathbf{\Sigma}}_t := \mathbb{E}[\mathbf{\Sigma}_t]$ denote the matrix mean and $\mathbf{M}_t := \mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t] - \bar{\mathbf{\Sigma}}_t^\top \bar{\mathbf{\Sigma}}_t$ denote the matrix variance of $\mathbf{\Sigma}_t$. Moreover, let $\bar{\mathbf{Y}} = \mathbb{E}[\mathbf{P}\mathbf{Y}_t\mathbf{P}^\top]$ and $\mathbf{R}_t := \mathbf{P}(\mathbf{Y}_t - \mathbb{E}[\mathbf{Y}_t])\mathbf{P}^\top$. Then, for any $p_{t-1} \geq \left\|\mathbb{E}[\mathbf{\Sigma}_{t-1}^\top \mathbf{\Sigma}_{t-1}]\right\|$, we have the following matrix recursion:

$$\mathbf{M}_t \preceq p_{t-1} \cdot \mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t] + \bar{\mathbf{Y}}^\top \mathbf{M}_{t-1} \bar{\mathbf{Y}}. \tag{13}$$

Note that, $\bar{\mathbf{Y}}$ and $\bar{\mathbf{\Sigma}}_t$ are block-diagonal due to Lemma 1. If one can show that $\mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t]$ is upper bounded by block-diagonal matrices, then via the matrix recursion, so is $\mathbf{M}_t$. We show this in the following lemma.

Lemma 4. *Let $\mathbf{P}$ be the permutation matrix from Lemma 1. Then we have,*

$$\mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t] \preceq 5 \cdot \operatorname{diag}(\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_n), \qquad \textit{where} \qquad \mathbf{W}_i = \operatorname{diag}\left(\frac{\lambda_i(1-\lambda_i)}{m}, 0\right).$$

Using Lemma 4 and the repeated use of (13), it follows that both $\mathbf{M}_t$ and $\mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t]$ are also upper bounded by block-diagonal matrices. By taking a maximum over those blocks, we prove the following recursive expression for $\left\|\mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t]\right\|$.

Lemma 5. *For any $t \geq 1$, if there exist $p_0, p_1, p_2, \cdots, p_{t-1}$ such that $\left\|\mathbb{E}[\mathbf{\Sigma}_k^\top \mathbf{\Sigma}_k]\right\| \leq p_k$ for $0 \leq k \leq t-1$, then there exists p_t such that for $q_i = 5\lambda_i(1-\lambda_i)/m$,*

$$\left\|\mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t]\right\| \leq p_t \leq \max_{1 \leq i \leq r} \left(\left\|(\mathbf{T}_i^\top)^t \mathbf{T}_i^t\right\| + q_i \cdot \sum_{j=0}^{t-1} p_j \cdot \left\|(\mathbf{T}_i^\top)^{t-1-j} \mathbf{T}_i^{t-1-j}\right\| \right).$$

Unfolding the recursion. The next step is to unfold the recursive expression for p_t. By coupling with a simpler recursion, we show that, defining $\rho_t := \max_{1 \leq i \leq r} \left(|\gamma_i| + q_i \cdot \ell_i(t)\right)$,

$$p_t \leq \rho_t^t + \sum_{j=0}^{t-1} \rho_t^j \cdot \max_{1 \leq i \leq r} \left\|(\mathbf{T}_i^\top)^{t-j} \mathbf{T}_i^{t-j}\right\|, \tag{14}$$

for a carefully defined function $\ell_i(t)$, see (25). We can now use the bias analysis (Lemmas 2 and 3) to obtain tight bounds for the terms $\left\|(\mathbf{T}_i^\top)^{t-j} \mathbf{T}_i^{t-j}\right\|$. The final challenging task is to bound ρ_t.

Upper bounding ρ_t. We provide a fine-grained analysis of ρ_t by carefully analyzing $|\gamma_i| + q_i \cdot \ell_i(t)$ for every single $i \leq r$. Recall that, depending upon the interaction between λ_i and β, the eigenvalue γ_i can be real or complex (Lemma 2).

Case 1: $\lambda_i - \frac{(1-\beta)^2}{(1+\beta)^2}$ is sufficiently positive. This is the easy case where γ_i is complex, and we get

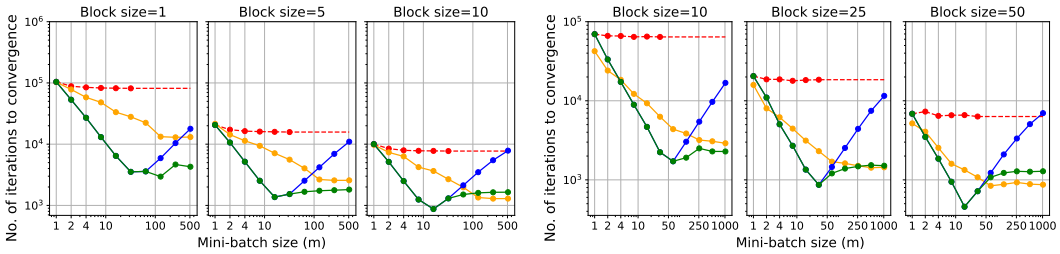
$$|\gamma_i| + q_i \cdot \ell_i(t) < 1 - \frac{1}{32\phi}, \qquad \text{for } \phi \leq \frac{m}{C}. \tag{15}$$

Case 2: $\lambda_i - \frac{(1-\beta)^2}{(1+\beta)^2}$ is close to zero. In this case, we enter the regime of near non-diagonalizability of $\mathbf{T}$. A crude analysis here would lead to a blow up factor in m. However, we are able to avoid this by exploiting the specific form of $\ell_i(t)$ and combine it with $\lambda_i \approx \frac{(1-\beta)^2}{(1+\beta)^2}$ to show that (15) still holds.

Case 3: $\lambda_i - \frac{(1-\beta)^2}{(1+\beta)^2}$ is sufficiently negative. Previous works such as [5, 40, 32] select large values for the momentum parameter β and for the mini-batch size m, because they enforce the condition that $\lambda_i \geq \frac{(1-\beta)^2}{(1+\beta)^2}$. Our crucial insight is that, for smaller values of β, the quantity $\lambda_i - \frac{(1-\beta)^2}{(1+\beta)^2}$ can be negative and sufficiently bounded away from 0 for some i. In fact, this is the case that determines the convergence rate for $\mathbb{E}\|\mathbf{\Delta}_t\|^2$ and allows us to handle SGD with classical momentum even under small mini-batches (see details in Appendix C). In particular, in this case we show that for $\phi \leq \frac{m}{C}$,

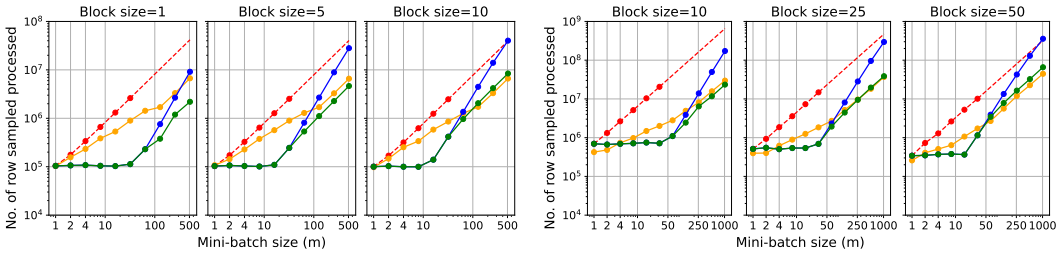
$$|\gamma_i| + q_i \cdot \ell_i(t) \leq 1 - \frac{\phi \lambda_i}{4}. \tag{16}$$

Combining (15), (16) and maximizing over i, bounds ρ_t. Substituting in (14) and bounding $\left\|(\mathbf{T}_i^\top)^{t-j} \mathbf{T}_i^{t-j}\right\|$ bounds p_t. Noting the relation between m and ϕ finishes the proof of Theorem 1.



(a) Covtype: iterations vs mini-batch size

(b) California Housing: iterations vs mini-batch size



(c) Covtype: row samples vs mini-batch size

(d) California Housing: row samples vs mini-batch size

CD CDpp CD+NAG ($\beta = 1 - 1/m$) CD+NAG ($\beta =$ adaptive)

Figure 1: Convergence of block coordinate descent with NAG momentum and baselines. Top row: iterations to reach error 10^{-7}. Bottom row: total work, measured by the number of sampled rows.

4 Experiments

For our experiments, we consider kernel ridge regression on four benchmark datasets sourced from Scikit-learn [41] and OpenML [49] (more details and results in Appendix D). For each dataset, we solve the resulting psd linear system $(\mathbf{K} + \lambda \mathbf{I})\mathbf{w} = \mathbf{y}$, where $\mathbf{K}$ is the kernel matrix and $\lambda \geq 0$ is the ridge parameter, using randomized block coordinate descent (7) with uniform sampling as the base update. We then incorporate mini-batching by averaging m independently generated update vectors in each iteration. Note that this mini-batching is separate from block sampling, which is less readily parallelizable due to the block inverse step. Each of the m update vectors in the mini-batch is produced using an independently sampled random block. Finally, we consider four different configurations of the algorithm, depending on the type of momentum acceleration:

1. CD: Vanilla block coordinate descent (7) using mini-batches of size m and varying block sizes.

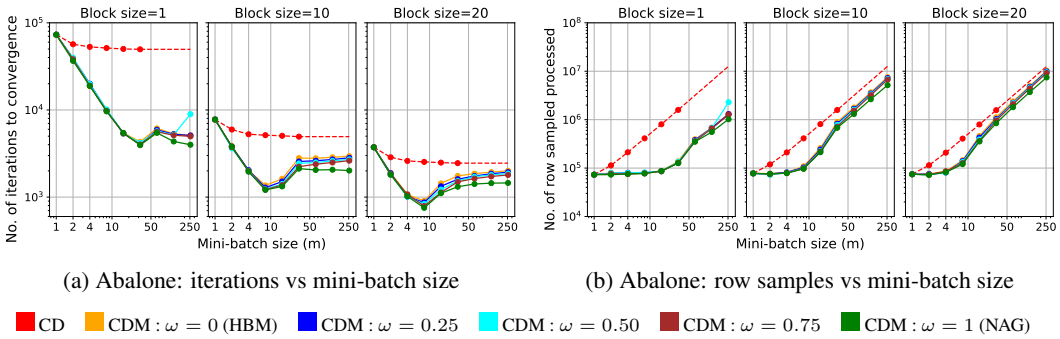
2. CDpp: For comparison, we include a tuned state-of-the-art acceleration scheme for CD (i.e., not classical momentum) based on a Lyapunov-based approach with two auxiliary sequences [18, 10, 11].

3. CD+NAG ($\beta = 1 - 1/m$): CD with classical NAG momentum (i.e., $\omega = 1$ in our framework), using a simple choice of momentum parameter motivated by our theory.

4. CD+NAG ($\beta =$ adaptive): We implement a heuristic strategy that learns the near-optimal β for any m by executing an adaptive binary search procedure at runtime (see Appendix D).

We also separately compare several variants of classical momentum by interpolating between **CD+HBM** ($\omega = 0$) and **CD+NAG** ($\omega = 1$), see Fig. 2. Although our theory (Theorem 1) guarantees identical convergence rates for all $\omega \in [0, 1]$, empirical performance can depend on the choice of ω, with larger values often yielding improved results. This observation is also noted in the remark following Lemma 7. Furthermore, as shown in Fig. 2, all variants of classical momentum perform very similarly; nevertheless, $\omega = 1$ consistently achieves the best results, which is why we chose NAG momentum in our main experiments.

In Fig. 1(a)-(b) we can see that, as expected, mini-batching by itself does not improve the convergence of **CD**. (Due to this phenomenon, we ran **CD** only up to mini-batch size $m = 32$, and extrapolated for larger values of m.) However, incorporating momentum into the update unlocks significant speed-up from mini-batching. Remarkably, while the specialized acceleration scheme of **CDpp** tends to attain better acceleration than **CD+NAG** for $m = 1$, the classical momentum does a better job of exploiting



(a) Abalone: iterations vs mini-batch size

(b) Abalone: row samples vs mini-batch size

Figure 2: Convergence behavior for CDM with different ω across block sizes. Left column: iterations to reach error 10^{-10}. Right column: total work, measured by the number of sampled rows.

large mini-batches. This can be seen even more clearly in Fig. 1(c)-(d), which shows total work performed by the algorithm. For **CD+NAG**, the total work stays constant as we increase m, until reaching a saturation point, which allows for perfect parallelization of mini-batches. Notably, this perfect parallelization phenomenon is not observed for **CDpp**.

Next, comparing **CD+NAG** with $\beta = 1 - 1/m$ to its adaptive variant, we observe that in the perfect parallelization regime the simple choice of momentum parameter is essentially optimal. For larger m, as suggested by our theory, β should be capped at $1 - \Theta(1/\sqrt{\kappa})$, which is why the adaptive selection procedure becomes more effective.

5 Conclusions

We develop a theoretical framework for quantifying the effect of classical momentum schemes like heavy ball on the convergence of mini-batch SGD. Our results demonstrate that momentum enables perfect parallelization of mini-batch computations, which is confirmed by numerical experiments.

Acknowledgments and Disclosure of Funding

This work was supported in part by NSF CAREER Grant CCF-233865 and a Google ML and Systems Junior Faculty Award. The work was done in part while MD was visiting the Simons Institute for the Theory of Computing.

References

[1] Seth J Alderman, Roan W Luikart, and Nicholas F Marshall. Randomized kaczmarz with geometrically smoothed momentum. *SIAM Journal on Matrix Analysis and Applications*, 45(4):2287–2313, 2024.

[2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

[3] Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. *arXiv preprint arXiv:2507.11274*, 2025.

[4] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

[5] Raghu Bollapragada, Tyler Chen, and Rachel Ward. On the fast convergence of minibatch heavy ball momentum. *IMA Journal of Numerical Analysis*, 45(3):1397–1424, 2025.

[6] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[7] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

[8] Bugra Can, Mert Gurbuzbalaban, and Lingjiong Zhu. Accelerated linear convergence of stochastic momentum methods in wasserstein distances. In *International Conference on Machine Learning*, pages 891–901. PMLR, 2019.

[9] Michał Dereziński and Xiaoyu Dong. Last-iterate convergence of randomized kaczmarz and sgd with greedy step size. *arXiv preprint arXiv:2604.09909*, 2026.

[10] Michal Dereziński, Daniel LeJeune, Deanna Needell, and Elizaveta Rebrova. Fine-grained analysis and faster algorithms for iteratively solving linear systems. *Journal of Machine Learning Research*, 26(144):1–49, 2025.

[11] Michał Dereziński, Deanna Needell, Elizaveta Rebrova, and Jiaming Yang. Randomized kaczmarz methods with beyond-krylov convergence. *SIAM Journal on Matrix Analysis and Applications*, 46(4):2558–2588, 2025.

[12] Michał Dereziński and Elizaveta Rebrova. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science*, 6(1):127–153, 2024.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[14] Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on learning theory*, pages 658–695. PMLR, 2015.

[15] Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and why momentum accelerates sgd: An empirical study. *arXiv preprint arXiv:2306.09000*, 2023.

[16] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. 2018.

[17] Noah Golmant, Nikita Vemuri, Zhewei Yao, Vladimir Feinberg, Amir Gholami, Kai Rothauge, Michael W Mahoney, and Joseph Gonzalez. On the computational inefficiency of large batch sizes for stochastic gradient descent. *arXiv preprint arXiv:1811.12941*, 2018.

[18] Robert Gower, Filip Hanzely, Peter Richtárik, and Sebastian U Stich. Accelerated stochastic matrix inversion: general theory and speeding up bfgs rules for faster second-order optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[19] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[20] Kanan Gupta, Jonathan W Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. *Advances in Neural Information Processing Systems*, 37:20694–20744, 2024.

[21] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[22] Deren Han and Jiaxin Xie. On pseudoinverse-free randomized methods for linear systems: Unified framework and acceleration. *Optimization Methods and Software*, 41(1):82–117, 2026.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2013.

[25] De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. Matrix concentration for products. *Foundations of Computational Mathematics*, 22(6):1767–1799, 2022.

[26] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

[27] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on learning theory*, pages 1042–1085. PMLR, 2018.

[28] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[29] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[32] Kiwon Lee, Andrew Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of mini-batch momentum: batch size saturation and convergence in high dimensions. *Advances in Neural Information Processing Systems*, 35:36944–36957, 2022.

[33] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

[34] Ji Liu and Stephen Wright. An accelerated randomized kaczmarz algorithm. *Mathematics of Computation*, 85(297):153–178, 2016.

[35] Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.

[36] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[37] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[38] Deanna Needell and Joel A Tropp. Paved with good intentions: analysis of a randomized block kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.

[39] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

[40] Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. *arXiv preprint arXiv:2312.14567*, 2023.

[41] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[42] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[43] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[45] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[46] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. pmlr, 2013.

[47] Kejie Tang, Weidong Liu, Yichen Zhang, and Xi Chen. Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023.

[48] Bryan Van Scoy, Randy A Freeman, and Kevin M Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2017.

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

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] 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.

[52] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

[53] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*, 2023.

[54] Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in neural information processing systems*, 34:7333–7345, 2021.

[55] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.

[56] Yun Zeng, Deren Han, Yansheng Su, and Jiaxin Xie. On adaptive stochastic heavy ball momentum for solving linear systems. *SIAM Journal on Matrix Analysis and Applications*, 45(3):1259–1286, 2024.

A Further discussion of related work

In addition to the classical HBM and NAG momentum, a variety of more sophisticated momentum-based methods have been proposed in the literature [11, 22, 1, 2, 27, 26, 48, 34, 6, 54, 51]. Concurrently, substantial theoretical progress has been made toward understanding the convergence properties of stochastic gradient descent (SGD) equipped with various forms of momentum.

Alongside these developments, a consistent empirical finding is the accelerated convergence achieved by incorporating classical momentum schemes such as HBM or NAG into mini-batch SGD [46, 23, 44, 15]. Notwithstanding this strong empirical evidence, the theoretical understanding of the mechanisms underlying such acceleration remains incomplete. In particular, even in the setting of quadratic objectives, a comprehensive explanation of when and why these classical momentum methods yield acceleration is still lacking.

A substantial body of work has analyzed the behavior of mini-batch SGD with HBM and NAG [20, 56, 40, 47, 5, 36, 8, 55, 16, 29, 14]. Many of these studies establish that, in general, SGD with $O(1)$ size mini-batches and with such momentum does not admit provable improvements over its non-momentum counterpart [29, 35, 26, 53], attributing observed empirical gains primarily to variance reduction induced by large mini-batches. Despite these negative results, more recent works have revisited this question under restrictive settings, including large mini-batches, carefully tuned step sizes and momentum parameters, and specific assumptions on the data and noise [5, 47, 40, 56, 32]. For instance, [56] demonstrate accelerated convergence of randomized linear solvers with HBM via adaptive parameter tuning; however, their analysis relies on distributional assumptions on the sketching matrices that may not hold for commonly used subsampling schemes, and their guarantees do not explicitly quantify the improvement in convergence rates. Similarly, [32] establish acceleration for mini-batch SGD, but they assume independent homoscedastic noise and orthogonally invariant data, and their results hold only asymptotically with mini-batch sizes proportional to the data size.

The works most closely related to ours are [5, 40, 47]. In particular, [5] establish accelerated convergence of mini-batch SGD with HBM under the requirement that the batch size satisfies $m = \tilde{\Omega}(\kappa_f \sqrt{\kappa_F})$, along with additional stable diagonalization assumptions. In a different setting, [47] prove acceleration for inconsistent linear systems, but require a significantly larger batch size, specifically $m = \tilde{\Omega}(\kappa_f^3)$. More recently, [40] show that mini-batch SGD with HBM has iteration complexity $\tilde{O}\big(\sqrt{\kappa} + \frac{d}{m\epsilon}\big)$ to reach ϵ-accuracy, showing sublinear convergence that improves with mini-batch size. They do not consider the interpolation regime, but rather, rely on other assumptions such as anisotropic gradient noise and carefully designed step-size schedules. Since we are focused on the interpolation regime and attaining linear convergence, their results are largely incomparable to ours.

B Convergence analysis for expected iterates

We start by recalling the matrix transition rule (11),

$$\begin{bmatrix} \mathbf{V}^\top \mathbf{\Delta}_{t+1} \\ \mathbf{V}^\top \mathbf{u}_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta)\mathbf{I} - (1+\beta\omega)\mathbf{V}^\top \mathbf{\Pi}_t \mathbf{V} & -\beta \mathbf{I} \\ \mathbf{V}^\top(\mathbf{I} - \omega\, \mathbf{\Pi}_t)\mathbf{V} & \mathbf{0} \end{bmatrix}}_{=\mathbf{Y}_t} \cdot \begin{bmatrix} \mathbf{V}^\top \mathbf{\Delta}_t \\ \mathbf{V}^\top \mathbf{u}_t \end{bmatrix}$$

where $\omega \in [0, 1]$, $\mathbf{u}_{t+1} = (\mathbf{I} - \omega\, \mathbf{\Pi}_t)\mathbf{\Delta}_t$ and $\mathbf{u}_0 = 0$. As range$(\mathbf{\Pi}_t) \subseteq$ range$(\bar{\mathbf{\Pi}})$, we note that if $\mathbf{\Delta}_0 \in$ range$(\bar{\mathbf{\Pi}})$, then for all t, $\mathbf{\Delta}_t, \mathbf{u}_t \in$ range$(\bar{\mathbf{\Pi}})$. In this section, we analyze the update rule (11) in expectation. Let $\mathbf{T}$ denote $\mathbb{E}\mathbf{Y}_t$. We have,

$$\mathbf{T} = \begin{bmatrix} (1+\beta)\mathbf{I} - (1+\beta\omega)\mathbf{V}^\top \bar{\mathbf{\Pi}} \mathbf{V} & -\beta \mathbf{I} \\ \mathbf{I} - \omega \mathbf{V}^\top \bar{\mathbf{\Pi}} \mathbf{V} & \mathbf{0} \end{bmatrix}.$$

It is useful to transform $\mathbf{T}$ into a block-diagonal form. The following lemma guarantees that a permutation matrix $\mathbf{P}$ exists such that $\mathbf{T}$ can be transformed into a block-diagonal matrix.

Lemma 6 (Block-diagonalization via permutation)**.** *There exists a permutation matrix* $\mathbf{P}$ *such that,*

$$\mathbf{P}\mathbf{T}\mathbf{P}^\top = \mathrm{diag}(\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_d),$$

where $\mathbf{T}_i = \begin{bmatrix} (1+\beta) - (1+\beta\omega)\lambda_i & -\beta \\ 1-\omega\lambda_i & 0 \end{bmatrix}$, *and*

$$\mathbf{P}\begin{bmatrix} \mathbf{V}^\top \mathbf{\Delta}_{t+1} \\ \mathbf{V}^\top \mathbf{u}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^\top \mathbf{\Delta}_{t+1} \\ \mathbf{v}_1^\top \mathbf{u}_{t+1} \\ \cdots \\ \mathbf{v}_d^\top \mathbf{\Delta}_{t+1} \\ \mathbf{v}_d^\top \mathbf{u}_{t+1} \end{bmatrix}.$$

The resulting block-diagonal structure helps us throughout the analysis. As $(\mathbf{PTP}^\top)^t = \mathbf{PT}^t\mathbf{P}^\top$ and $\|\mathbf{PT}^t\mathbf{P}^\top\| = \|\mathbf{T}^t\|$, we need to upper bound $\|(\mathbf{PTP}^\top)^t\|$. Due to the block-diagonal form of $\mathbf{PTP}^\top$, we have $\|(\mathbf{PTP}^\top)^t\| = \max_i \|\mathbf{T}_i^t\|$. For any fixed i, let $\gamma_{i1}, \gamma_{i2} \in \mathbb{C}$ denote the two eigenvalues of $\mathbf{T}_i$. We always take γ_{i1} to be the eigenvalue with larger magnitude. In the next lemma, we provide a tight upper bound on the magnitude of eigenvalues of $\mathbf{T}_i$.

Lemma 7 (Bounding spectral radius of transition matrix)**.** *Let* $\omega \in [0,1]$ *and* $\beta = 1 - \frac{1}{\phi}$ *for some* $\phi \geq 2$*. Then, the following holds:*

If $\mathbf{T}_i$ *has complex eigenvalues then,*

$$|\gamma_{i2}|^2 = |\gamma_{i1}|^2 = \Big(1 - \frac{1}{\phi}\Big)(1 - \omega\lambda_i).$$

If $\mathbf{T}_i$ *has real and equal eigenvalues then,*

$$|\gamma_{i2}|^2 = |\gamma_{i1}|^2 = \Big(1 - \frac{1}{\phi}\Big)(1 - \omega\lambda_i).$$

If $\mathbf{T}_i$ *has real and distinct eigenvalues then,*

$$|\gamma_{i2}|^2 < |\gamma_{i1}|^2 < \Big(1 - \frac{\phi\lambda_i}{2}\Big)^2 (1 - \omega\lambda_i)^2.$$

Proof. We have,

$$\mathbf{T}_i = \begin{bmatrix} (1+\beta) - (1+\beta\omega)\lambda_i & -\beta \\ 1-\omega\lambda_i & 0 \end{bmatrix}$$

It is easy to verify that the eigenvalues of $\mathbf{T}_i$, i.e., γ_{i1} and γ_{i2} are determined by the following expression.

$$(\gamma_{i1}, \gamma_{i2}) = \frac{1}{2}\Big((1-\lambda_i) + \beta(1-\omega\lambda_i) \pm \sqrt{\big((1-\lambda_i) + \beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i)}\Big).$$

The expression inside the square root can be negative, zero, or positive. So we consider three cases:

1. $\big((1-\lambda_i) + \beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i) < 0$: In this case, both γ_{i1} and γ_{i2} are complex and have the same magnitude. We get,

$$|\gamma_{i2}|^2 = |\gamma_{i1}|^2 = \beta(1-\omega\lambda_i) = \Big(1 - \frac{1}{\phi}\Big)(1 - \omega\lambda_i). \tag{17}$$

2. $\big((1-\lambda_i) + \beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i) = 0$. In this case, $\gamma_{i1}^2 = \gamma_{i2}^2 = \frac{1}{4}\Big((1-\lambda_i) + \beta(1-\omega\lambda_i)\Big)^2 = \beta(1-\omega\lambda_i)$. Therefore,

$$\gamma_{i2}^2 = \gamma_{i1}^2 = \Big(1 - \frac{1}{\phi}\Big)(1 - \omega\lambda_i). \tag{18}$$

3. $\big((1-\lambda_i) + \beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i) > 0$. The larger eigenvalue is given as,

$$\gamma_{i1} = \frac{1}{2}\Big((1-\lambda_i) + \beta(1-\omega\lambda_i) + \sqrt{\big((1-\lambda_i) + \beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i)}\Big). \tag{19}$$

Let $u_i = \frac{1-\lambda_i}{1-\omega\lambda_i}$,

$$\begin{aligned}
\gamma_{i1} &= \frac{1}{2}\Big(u_i + \beta + \sqrt{\big(u_i+\beta\big)^2 - \frac{4\beta}{1-\omega\lambda_i}}\Big)(1-\omega\lambda_i) \\
&= \frac{1}{2}\Big(u_i + \beta + \sqrt{u_i^2 + \beta^2 + \frac{2\beta(1-\lambda_i)}{1-\omega\lambda_i} - \frac{4\beta}{1-\omega\lambda_i}}\Big)(1-\omega\lambda_i) \\
&= \frac{1}{2}\Big(u_i + \beta + \sqrt{u_i^2 + \beta^2 - \frac{2\beta(1+\lambda_i)}{1-\omega\lambda_i}}\Big)(1-\omega\lambda_i) \\
&= \frac{1}{2}\Big(u_i + \beta + \sqrt{u_i^2 + \beta^2 - \frac{2\beta(1-\lambda_i)}{1-\omega\lambda_i} - \frac{4\beta\lambda_i}{1-\omega\lambda_i}}\Big)(1-\omega\lambda_i) \\
&= \frac{1}{2}\Big(u_i + \beta + \sqrt{(u_i-\beta)^2 - \frac{4\beta\lambda_i}{1-\omega\lambda_i}}\Big)(1-\omega\lambda_i)
\end{aligned}$$

Let $x = u_i - \beta$, $y = \sqrt{\frac{4\beta\lambda_i}{1-\omega\lambda_i}}$, then we have $x = \frac{yr}{2}$, where $r = \frac{2x}{y} > 2$. Now, if $x, y \in \mathbb{R}$, $x > y > 0$ and $x \le \frac{yr}{2}$ for some $r > 2$, then $\sqrt{x^2 - y^2} \le x - \frac{y}{r}$. Therefore,

$$\begin{aligned}
\gamma_{i1} &< \frac{1}{2}\Big(u_i + \beta + u_i - \beta - \frac{y}{r}\Big)(1-\omega\lambda_i) \\
&= \Big(u_i - \frac{y^2}{4x}\Big)(1-\omega\lambda_i) \\
&= \Big(1 - \frac{y^2}{4(1-\beta)}\Big)(1-\omega\lambda_i).
\end{aligned}$$

The last inequality holds as $u_i \le 1$ and $x \le 1-\beta$. Therefore, we get

$$\gamma_{i1} < \Big(1 - \frac{\phi\beta\lambda_i}{1-\omega\lambda_i}\Big)(1-\omega\lambda_i)$$

Now for $\beta \ge 0.5$, we get

$$\gamma_{i1} < \Big(1 - \frac{\phi\lambda_i}{2}\Big)(1-\omega\lambda_i) \tag{20}$$

Combining the relations (17), (18) and (20) we finish the proof. ■

Remark 3. *Lemma 7 implies that NAG momentum can perform slightly better than HBM as $\omega = 1$ for NAG and $\omega = 0$ for HBM. We demonstrate this phenomenon empirically in Fig 2, where we compare iteration complexity of block coordinate descent with momentum for various values of ω.*

We are now ready for the analysis of expected iterates. As $\mathbf{\Delta}_t, \mathbf{u}_t \in \text{range}(\bar{\mathbf{\Pi}})$ for all t, if for some i, $\mathbf{v}_i \in \text{Null}(\bar{\mathbf{\Pi}})$, then $\mathbf{v}_i^\top \mathbf{\Delta}_t = 0$ and $\mathbf{v}_i^\top \mathbf{u}_t = 0$. Let r denote the rank of $\bar{\mathbf{\Pi}}$ and note that $\mathbf{v}_i^\top \mathbf{\Delta}_t = \mathbf{v}_i^\top \mathbf{u}_t = 0$ for all t and $i > r$. Let ϕ be any fixed constant satisfying $2 \le \phi \le \min\big\{\frac{m}{C}, \frac{1}{3\sqrt{\lambda_r}}\big\}$, where C is some absolute constant (to be determined later) and m is as it appears in Proposition 2. Let the momentum parameter $\beta = 1 - \frac{1}{\phi}$. We prove the following result:

Theorem 3. *Let $\omega \in [0,1]$. Then, the matrix transition rule (11) satisfies,*

$$\|\mathbb{E}\mathbf{\Delta}_{t+1}\|^2 \le (3t+5)^2 \cdot \rho^t . \|\mathbf{\Delta}_0\|^2,$$

where $\rho = \big(1 - \frac{\phi\lambda_r}{2}\big)(1-\omega\lambda_r)$.

We will need the following auxiliary lemmas to finish the proof of Theorem 3.

Lemma 8 (Two-step matrix transition rule)**.** *We have*

$$\|\mathbb{E}\mathbf{\Delta}_{t+1}\|^2 \le \max_{1\le i\le r} \|\mathbf{T}_i^{t+1}\|^2 \cdot \|\mathbf{\Delta}_0\|^2.$$

Proof. Let $t \geq 0$. The update rule (11) satisfies,

$$\mathbb{E}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top \boldsymbol{\Delta}_{t+1} \\ \mathbf{V}^\top \mathbf{u}_{t+1}\end{bmatrix} = \mathbf{P}\mathbf{T}\mathbf{P}^\top \cdot \mathbb{E}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top \boldsymbol{\Delta}_{t} \\ \mathbf{V}^\top \mathbf{u}_{t}\end{bmatrix}.$$

Due to the block diagonal structure of $\mathbf{P}\mathbf{T}\mathbf{P}^\top$, we get

$$\mathbb{E}\begin{bmatrix}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t+1} \\ \mathbf{v}_i^\top \mathbf{u}_{t+1}\end{bmatrix} = \mathbf{T}_i \cdot \mathbb{E}\begin{bmatrix}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t} \\ \mathbf{v}_i^\top \mathbf{u}_{t}\end{bmatrix}$$

Recursively applying the above relation we get

$$\mathbb{E}\begin{bmatrix}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t+1} \\ \mathbf{v}_i^\top \mathbf{u}_{t+1}\end{bmatrix} = \mathbf{T}_i^{t+1} \cdot \mathbb{E}\begin{bmatrix}\mathbf{v}_i^\top \boldsymbol{\Delta}_{0} \\ \mathbf{v}_i^\top \mathbf{u}_{0}\end{bmatrix}$$

Therefore,

$$\|\mathbb{E}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t+1}\|^2 \leq \|\mathbb{E}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t+1}\|^2 + \|\mathbb{E}\mathbf{v}_i^\top \mathbf{u}_{t+1}\|^2 \leq \|\mathbf{T}_i^{t+1}\|^2 \cdot \left(\|\mathbb{E}\mathbf{v}_i^\top \boldsymbol{\Delta}_0\|^2 + \|\mathbf{v}_i^\top \mathbf{u}_0\|^2\right). \quad (21)$$

As $\mathbf{u}_0 = 0$, we get

$$\|\mathbb{E}\mathbf{v}_i^\top \boldsymbol{\Delta}_{t+1}\|^2 \leq \|\mathbf{T}_i^{t+1}\|^2 \cdot \|\mathbf{v}_i^\top \boldsymbol{\Delta}_0\|^2.$$

Now summing over i and noting that $\mathbf{v}_i^\top \boldsymbol{\Delta}_t = 0$ for all $i > r$, we get

$$\|\mathbb{E}\mathbf{V}^\top \boldsymbol{\Delta}_{t+1}\|^2 \leq \max_{1\leq i\leq r} \|\mathbf{T}_i^{t+1}\|^2 \cdot \|\mathbf{V}^\top \boldsymbol{\Delta}_0\|^2.$$

Noting that $\mathbf{V}$ has orthonormal columns finishes the proof. ■

To complete the proof of Theorem 3, we need to bound the spectral norm of $\mathbf{T}_i^t$. Note that Lemma 7 bounds the spectral radius of $\mathbf{T}_i^t$, not the spectral norm. For bounding the spectral norm, we use the following fundamental matrix decomposition result. For any complex valued matrix $\mathbf{U}$, let $\mathbf{U}^H$ denote the conjugate transpose of $\mathbf{U}$.

Lemma 9 (Schur's decomposition, Theorem 2.3.1 in [24])**.** *For every i, there exists a unitary matrix $\mathbf{U}_i$ such that*

$$\mathbf{U}_i^H \mathbf{T}_i \mathbf{U}_i = \begin{bmatrix}\gamma_{i1} & x_i \\ 0 & \gamma_{i2}\end{bmatrix},$$

where γ_{i1} and γ_{i2} are the two eigenvalues of $\mathbf{T}_i$ and x_i is a constant such that $|x_i| \leq 3$. The unitary matrix, $\mathbf{U}_i$ is explicitly given as

$$\mathbf{U}_i = \frac{1}{\sqrt{(1-\omega\lambda_i)^2 + |\gamma_{i1}|^2}} \cdot \begin{bmatrix}\gamma_{i1} & -(1-\omega\lambda_i) \\ (1-\omega\lambda_i) & \bar{\gamma}_{i1}\end{bmatrix},$$

where $\bar{\gamma}_{i1}$ denotes complex conjugate of γ_{i1}. Furthermore, for any $k \geq 1$,

$$\mathbf{U}_i^H \mathbf{T}_i^k \mathbf{U}_i = \begin{bmatrix}\gamma_{i1}^k & x_i \sum_{j=0}^{k-1} \gamma_{i2}^j \gamma_{i1}^{k-1-j} \\ 0 & \gamma_{i2}^k\end{bmatrix}$$

Corollary 1. *If $\gamma_{i1} \neq \gamma_{i2}$, then $x_i \sum_{j=0}^{k-1} \gamma_{i2}^j \gamma_{i1}^{k-1-j} = x_i \gamma_{i1}^{k-1} \cdot \frac{1-(\gamma_{i2}/\gamma_{i1})^k}{1-(\gamma_{i2}/\gamma_{i1})}$, implying*

$$|x_i \sum_{j=0}^{k-1} \gamma_{i2}^j \gamma_{i1}^{k-1-j}| \leq |x_i| \cdot |\gamma_{i1}|^k \cdot \frac{3}{|\gamma_{i1} - \gamma_{i2}|}.$$

Corollary 2. *If $\gamma_{i1} = \gamma_{i2}$, then $x_i \sum_{j=0}^{k-1} \gamma_{i2}^j \gamma_{i1}^{k-1-j} = x_i \gamma_{i1}^{k-1} \cdot k$, implying,*

$$|x_i \sum_{j=0}^{k-1} \gamma_{i2}^j \gamma_{i1}^{k-1-j}| = |x_i| \cdot |\gamma_{i1}|^k \cdot \frac{k}{|\gamma_{i1}|}.$$

Corollary 3. *For any i,*

$$\|\mathbf{T}_i^k\| = \|\mathbf{U}_i^H \mathbf{T}_i^k \mathbf{U}_i\| \leq (3k+2) \cdot |\gamma_{i1}|^{k-1}.$$

Finishing proof of Theorem 3. Using Corollary 3 we get $\|\mathbf{T}_i^{t+1}\|^2 \le (3t+5)^2 \cdot |\gamma_{i1}|^{2t}$. Furthermore, if λ_r is such that $\big((1-\lambda_r)+\beta(1-\omega\lambda_r)\big)^2 - 4\beta(1-\omega\lambda_r) \le 0$, then we have $|\gamma_{i1}|^2 \le \Big(1-\frac{1}{2\phi}\Big)(1-\omega\lambda_r)$ for $i \le r$. On the other hand if, $\big((1-\lambda_i)+\beta(1-\omega\lambda_i)\big)^2 - 4\beta(1-\omega\lambda_i) > 0$, then $|\gamma_{i1}|^2 < \Big(1-\frac{\phi\lambda_r}{2}\Big)(1-\omega\lambda_r)$ for $i \le r$. Now the statement follows by noting that $\phi < \frac{1}{\sqrt{\lambda_r}}$.

C Convergence in ℓ_2-norm with mini-batch averaging

In this section, we aim to recover a result similar to Theorem 3 but in ℓ_2 norm. Recall the transformed update rule,

$$\mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t+1}\\ \mathbf{V}^\top\mathbf{u}_{t+1}\end{bmatrix} = \mathbf{P}\mathbf{Y}_t\mathbf{P}^\top \cdot \mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t}\\ \mathbf{V}^\top\mathbf{u}_{t}\end{bmatrix},$$

As $\mathbf{v}_i^\top\mathbf{\Delta}_t = \mathbf{v}_i^\top\mathbf{u}_t = 0$ for $i > r$, we can restrict the dynamics to the top-$2r$ dimensional space of the lifted state. Let $\mathbf{Q} \in \mathbb{R}^{2r\times 2d}$ consist of first $2r$ rows of $2d\times 2d$ identity matrix, then $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_{2r}$, while $\mathbf{Q}^\top\mathbf{Q}$ is the orthogonal projector onto the first $2r$ coordinates. We get,

$$\mathbf{Q}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t+1}\\ \mathbf{V}^\top\mathbf{u}_{t+1}\end{bmatrix} = \mathbf{Q}\mathbf{P}\mathbf{Y}_t\mathbf{P}^\top\mathbf{Q}^\top \cdot \mathbf{Q}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t}\\ \mathbf{V}^\top\mathbf{u}_{t}\end{bmatrix}$$

Recursively we get,

$$\mathbf{Q}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{t+1}\\ \mathbf{V}^\top\mathbf{u}_{t+1}\end{bmatrix} = \Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\mathbf{Q}^\top\Big) \cdot \mathbf{Q}\mathbf{P}\begin{bmatrix}\mathbf{V}^\top\mathbf{\Delta}_{0}\\ \mathbf{V}^\top\mathbf{u}_{0}\end{bmatrix}.$$

Therefore,

$$\mathbb{E}\|\mathbf{\Delta}_{t+1}\|^2 \le \Big\|\mathbb{E}\Big[\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_j^\top\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big]\Big\| \cdot \|\mathbf{\Delta}_0\|^2.$$

Our task is to bound $\Big\|\mathbb{E}\Big[\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_j^\top\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big]\Big\|$. We prove the following theorem, which is only a restatement of Theorem 1.

Theorem 4 (General result)**.** *Let ϕ satisfy $2 \le \phi \le \min\big\{\frac{m}{C}, \frac{1}{3\sqrt{\lambda_r}}\big\}$ for $C > 96000$. Then, update rule (11) with $\beta = 1-\frac{1}{\phi}$ satisfies,*

$$\mathbb{E}\|\mathbf{\Delta}_{t+1}\|^2 \le 50(t+1)^3 \cdot \rho^t \cdot \|\mathbf{\Delta}_0\|^2,$$

where $\rho = 1-\frac{\phi\lambda_r}{4}$.

Note that Theorem 1 follows from Theorem 4 by taking $c_1 = 2\cdot 96000$, $c_2 = c_1/3$.

C.1 Proof of Theorem 4

We prove a series of results, building towards the proof of Theorem 4. Let $\mathbf{Z}_t = \mathbf{V}^\top\mathbf{\Pi}_t\mathbf{V}$, and $\bar{\mathbf{Z}} = \mathbb{E}\mathbf{Z}_t$. Note that $\bar{\mathbf{Z}} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots \lambda_d)$. We have,

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_t - \bar{\mathbf{Z}})^2 = \mathbb{E}(\mathbf{Z}_t^2) - \bar{\mathbf{Z}}^2 &= \mathbb{E}(\mathbf{Z}_t^2) - \mathbf{\Lambda}^2\\ &= \mathbf{V}^\top\mathbb{E}(\mathbf{\Pi}_t^2)\mathbf{V} - \mathbf{\Lambda}^2\end{aligned}$$

As by Definition 1, $\mathbf{\Pi}_t \preceq \mathbf{I}$, we always have $\mathbb{E}(\mathbf{\Pi}_t^2) \preceq \bar{\mathbf{\Pi}}$, implying $\mathbb{E}(\mathbf{Z}_t - \bar{\mathbf{Z}})^2 \preceq \mathbf{\Lambda} - \mathbf{\Lambda}^2$. However, upper bounding $\mathbb{E}(\mathbf{\Pi}_t^2)$ by $\mathbb{E}\,\mathbf{\Pi}_t$ can be quite suboptimal, for instance, when $\mathbf{\Pi}_t$ is itself an average of multiple independent contractions. For proving a tighter bound on the second moment of $\mathbf{Z}_t$, we use Proposition 2. Due to Proposition 2, we have,

$$\mathbb{E}\big(\mathbf{Z}_t - \bar{\mathbf{Z}}\big)^2 = \mathbf{V}^\top\mathbb{E}\big(\mathbf{\Pi}_t - \bar{\mathbf{\Pi}}\big)^2\mathbf{V} \preceq \frac{1}{m}\cdot\mathbf{V}^\top\bar{\mathbf{\Pi}}(\mathbf{I}-\bar{\mathbf{\Pi}})\mathbf{V} = \frac{1}{m}\cdot\mathbf{\Lambda}(\mathbf{I}-\mathbf{\Lambda}) \tag{22}$$

Lemma 10. *Let* $\mathbf{P}$ *be the permutation matrix as in Lemma 6. Then, we have,*

$$\mathbf{P}\cdot\mathbb{E}\big[(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)^\top(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)\big]\cdot\mathbf{P}^\top\preceq 5\cdot\mathrm{diag}(\mathbf{W}_1,\mathbf{W}_2,\cdots,\mathbf{W}_d),$$

where $\mathbf{W}_i=\mathrm{diag}\left(\frac{\lambda_i(1-\lambda_i)}{m},0\right)$.

Proof. With $\mathbf{Z}_t=\mathbf{V}^\top\mathbf{\Pi}_t\mathbf{V}$ and $\bar{\mathbf{Z}}=\mathbb{E}\mathbf{Z}_t$, we have,

$$\mathbf{Y}_t=\begin{bmatrix}(1+\beta)\mathbf{I}-(1+\beta\omega)\mathbf{Z}_t & -\beta\mathbf{I}\\ \mathbf{I}-\omega\mathbf{Z}_t & \mathbf{0}\end{bmatrix},\ \mathbb{E}\mathbf{Y}_t=\begin{bmatrix}(1+\beta)\mathbf{I}-(1+\beta\omega)\bar{\mathbf{Z}} & -\beta\mathbf{I}\\ \mathbf{I}-\omega\bar{\mathbf{Z}} & \mathbf{0}\end{bmatrix},$$

and,

$$\begin{aligned}\mathbb{E}[(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)^\top(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)]&=\mathbb{E}\begin{bmatrix}(1+\beta\omega)(\bar{\mathbf{Z}}-\mathbf{Z}_t) & \omega(\bar{\mathbf{Z}}-\mathbf{Z}_t)\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\cdot\begin{bmatrix}(1+\beta\omega)(\bar{\mathbf{Z}}-\mathbf{Z}_t) & \mathbf{0}\\ \omega(\bar{\mathbf{Z}}-\mathbf{Z}_t) & \mathbf{0}\end{bmatrix}\\&=\begin{bmatrix}\big((1+\beta\omega)^2+\omega^2\big)\cdot\mathbb{E}\big(\mathbf{Z}_t-\bar{\mathbf{Z}}\big)^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}.\end{aligned}$$

Using relation (22) we get,

$$\mathbb{E}[(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)^\top(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)]\preceq\frac{5}{m}\cdot\begin{bmatrix}\mathbf{\Lambda}(\mathbf{I}-\mathbf{\Lambda}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

The last step is to observe that permuting the rows and columns using permutation matrix $\mathbf{P}$ yields the result. ■

In the next result, we construct a recursive framework that will help us in upper bounding the spectral norm of the product of random matrices $\mathbf{Y}_t$. As already exploited in Lemma 7 and Lemma 8, it is convenient to do the analysis on the transformed matrices (by using permutation $\mathbf{P}$). Consider the following notations:

$$\begin{aligned}\mathbf{\Sigma}_{t+1}=\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\mathbf{Q}^\top,\ \ \bar{\mathbf{\Sigma}}_{t+1}=\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbb{E}\mathbf{Y}_t\mathbf{P}^\top\mathbf{Q}^\top,\\ \mathbf{R}_t=\mathbf{Q}\mathbf{P}(\mathbf{Y}_t-\mathbb{E}\mathbf{Y}_t)\mathbf{P}^\top\mathbf{Q}^\top,\ \bar{\mathbf{Y}}=\mathbf{Q}\mathbf{P}\mathbb{E}\mathbf{Y}_t\mathbf{P}^\top\mathbf{Q}^\top,\end{aligned}$$

where $\mathbf{P}$ is the permutation matrix such that $\mathbf{P}\mathbb{E}\mathbf{Y}_t\mathbf{P}^\top=\mathrm{diag}(\mathbf{T}_1,\mathbf{T}_2,\cdots,\mathbf{T}_d)$. We prove the following lemma:

Lemma 11. *For any* $t\geq 1$*, there exist* $p_0,p_1,...,p_t,p_{t+1}$ *such that* $p_0=1$ *and,*

$$\big\|\mathbb{E}[\mathbf{\Sigma}_{t+1}^\top\mathbf{\Sigma}_{t+1}]\big\|\leq p_{t+1}=\max_{1\leq i\leq r}\left(\big\|(\mathbf{T}_i^\top)^{t+1}\mathbf{T}_i^{t+1}\big\|+\frac{5\lambda_i(1-\lambda_i)}{m}\cdot\sum_{j=0}^{t}p_j\cdot\big\|(\mathbf{T}_i^\top)^{t-j}\mathbf{T}_i^{t-j}\big\|\right).$$

Proof. We provide a proof by doing induction over t. Note that $t=0$ forms the base case, and as $\mathbf{\Sigma}_0=\mathbf{I}$, the base case hold for $p_0=1$. Now for $k\leq t$, let $\big\|\mathbb{E}[\mathbf{\Sigma}_k^\top\mathbf{\Sigma}_k]\big\|\leq p_k$, for p_k satisfying the the inequality in statement of Lemma 11, and let $\mathbf{M}_k=\mathbb{E}[\mathbf{\Sigma}_k^\top\mathbf{\Sigma}_k]-\bar{\mathbf{\Sigma}}_k^\top\bar{\mathbf{\Sigma}}_k$. For $\mathbf{\Sigma}_{t+1}$ we have,

$$\begin{aligned}\mathbb{E}[\mathbf{\Sigma}_{t+1}^\top\mathbf{\Sigma}_{t+1}]&=\mathbb{E}\big[\mathbf{Q}\mathbf{P}\mathbf{Y}_t^\top\mathbf{P}^\top\mathbf{Q}^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{Q}\mathbf{P}\mathbf{Y}_t\mathbf{P}^\top\mathbf{Q}^\top\big]\\&=\mathbb{E}\big[\mathbf{Q}\mathbf{P}\mathbb{E}\mathbf{Y}_t^\top\mathbf{P}^\top\mathbf{Q}^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{Q}\mathbf{P}\mathbb{E}\mathbf{Y}_t\mathbf{P}^\top\mathbf{Q}^\top\big]+\mathbb{E}\big[\mathbf{R}_t^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{R}_t\big]\\&=\mathbb{E}\big[\bar{\mathbf{Y}}^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\bar{\mathbf{Y}}\big]+\mathbb{E}\big[\mathbf{R}_t^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{R}_t\big]\\&=\bar{\mathbf{\Sigma}}_{t+1}^\top\bar{\mathbf{\Sigma}}_{t+1}+\bar{\mathbf{Y}}^\top\big(\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]-\bar{\mathbf{\Sigma}}_t^\top\bar{\mathbf{\Sigma}}_t\big)\bar{\mathbf{Y}}+\mathbb{E}\big[\mathbf{R}_t^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{R}_t\big]\end{aligned}$$

Rearranging the terms, we get,

$$\underbrace{\mathbb{E}[\mathbf{\Sigma}_{t+1}^\top\mathbf{\Sigma}_{t+1}]-\bar{\mathbf{\Sigma}}_{t+1}^\top\bar{\mathbf{\Sigma}}_{t+1}}_{\mathbf{M}_{t+1}}=\mathbb{E}\big[\mathbf{R}_t^\top\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]\mathbf{R}_t\big]+\bar{\mathbf{Y}}^\top\big(\mathbb{E}[\mathbf{\Sigma}_t^\top\mathbf{\Sigma}_t]-\bar{\mathbf{\Sigma}}_t^\top\bar{\mathbf{\Sigma}}_t\big)\bar{\mathbf{Y}}.$$

By induction hypothesis $\left\|\mathbb{E}[\mathbf{\Sigma}_t^\top \mathbf{\Sigma}_t]\right\| \leq p_t$, we get,

$$\mathbf{M}_{t+1} \preceq p_t \cdot \mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t] + \bar{\mathbf{Y}}^\top \mathbf{M}_t \bar{\mathbf{Y}}, \tag{23}$$

$$\mathbb{E}[\mathbf{\Sigma}_{t+1}^\top \mathbf{\Sigma}_{t+1}] \preceq \bar{\mathbf{\Sigma}}_{t+1}^\top \bar{\mathbf{\Sigma}}_{t+1} + p_t \cdot \mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t] + \bar{\mathbf{Y}}^\top \mathbf{M}_t \bar{\mathbf{Y}}, \tag{24}$$

Note that $\mathbb{E}[\mathbf{R}_t^\top \mathbf{R}_t] = \mathbf{M}_1$. Using Lemma 10, it follows that $\mathbf{M}_1$ is upper bounded by a block-diagonal matrix. Let us define $\mathbf{D}_1 := 5 \cdot \mathrm{diag}(\mathbf{W}_1, \mathbf{W}_2, \cdots \mathbf{W}_r)$ where $\mathbf{W}_i's$ are as given in Lemma 10, and for $k \geq 1$, $\mathbf{D}_{k+1} := p_k \cdot \mathbf{D}_1 + \bar{\mathbf{Y}}^\top \mathbf{D}_k \bar{\mathbf{Y}}$. Note that $\mathbf{D}_1$ is block-diagonal and $\mathbf{M}_1 \preceq \mathbf{D}_1$. Let's proceed inductively and assume that $\mathbf{D}_k$ is block-diagonal and $\mathbf{M}_k \preceq \mathbf{D}_k$. By definition of $\mathbf{D}_{k+1}$, we get $\mathbf{D}_{k+1}$ is block-diagonal as $\bar{\mathbf{Y}}$ is block-diagonal and by induction hypothesis $\mathbf{D}_k$ is block-diagonal. Furthermore, by (23), $\mathbf{M}_{k+1} \preceq \mathbf{D}_{k+1}$ as $\mathbf{M}_1 \preceq \mathbf{D}_1$ and $\mathbf{M}_k \preceq \mathbf{D}_k$ by induction hypothesis. Therefore, by induction, $\forall k,\ \mathbf{M}_k \preceq \mathbf{D}_k$, and $\mathbf{D}_k's$ are block-diagonal satisfying $\mathbf{D}_{k+1} = p_k \cdot \mathbf{D}_1 + \bar{\mathbf{Y}}^\top \mathbf{D}_k \bar{\mathbf{Y}}$. We now write the recursive expression for diagonal blocks in $\mathbf{D}_t$. Let the ith diagonal block of $\bar{\mathbf{Y}}$ be $\mathbf{T}_i$, and i^{th} diagonal block of $\mathbf{D}_t$ be $\mathbf{D}_{t,i}$. This yields,

$$\mathbf{D}_{t+1,i} = p_t \cdot \mathbf{D}_{1,i} + \mathbf{T}_i^\top \mathbf{D}_{t,i} \mathbf{T}_i.$$

Unfolding the above recursion, we get,

$$\begin{aligned}\mathbf{D}_{t+1,i} &= p_t \cdot \mathbf{D}_{1,i} + \mathbf{T}_i^\top (p_{t-1} \cdot \mathbf{D}_{1,i} + \mathbf{T}_i^\top \mathbf{D}_{t-1,i} \mathbf{T}_i) \mathbf{T}_i \\ &= p_t \cdot \mathbf{D}_{1,i} + p_{t-1} \cdot \mathbf{T}_i^\top \mathbf{D}_{1,i} \mathbf{T}_i + (\mathbf{T}_i^\top)^2 \mathbf{D}_{t-1,i} \mathbf{T}_i^2.\end{aligned}$$

Continuing the recursion, we get

$$\mathbf{D}_{t+1,i} = \sum_{j=0}^{t} p_j \cdot (\mathbf{T}_i^\top)^{t-j} \mathbf{D}_{1,i} \mathbf{T}_i^{t-j}, \text{or}$$

$$\mathbf{D}_{t,i} = \sum_{j=0}^{t-1} p_j \cdot (\mathbf{T}_i^\top)^{t-1-j} \mathbf{D}_{1,i} \mathbf{T}_i^{t-1-j}.$$

Substituting in (24), we bound $\left\|\mathbb{E}\mathbf{\Sigma}_{t+1}^\top \mathbf{\Sigma}_{t+1}\right\|$ as

$$\begin{aligned}\left\|\mathbb{E}\mathbf{\Sigma}_{t+1}^\top \mathbf{\Sigma}_{t+1}\right\| &\preceq \max_{1 \leq i \leq r} \left\| (\mathbf{T}_i^\top)^{t+1} \mathbf{T}_i^{t+1} + p_t \cdot \mathbf{D}_{1,i} + \sum_{j=0}^{t-1} p_j \cdot (\mathbf{T}_i^\top)^{t-j} \mathbf{D}_{1,i} \mathbf{T}_i^{t-j} \right\| \\ &\preceq \max_{1 \leq i \leq r} \left\| (\mathbf{T}_i^\top)^{t+1} \mathbf{T}_i^{t+1} + \sum_{j=0}^{t} p_j \cdot (\mathbf{T}_i^\top)^{t-j} \mathbf{D}_{1,i} \mathbf{T}_i^{t-j} \right\|.\end{aligned}$$

Using triangle inequality and norm submultiplicativity, we get,

$$\left\|\mathbb{E}\mathbf{\Sigma}_{t+1}^\top \mathbf{\Sigma}_{t+1}\right\| \leq p_{t+1} = \max_{1 \leq i \leq r} \left(\left\| (\mathbf{T}_i^\top)^{t+1} \mathbf{T}_i^{t+1} \right\| + \|\mathbf{D}_{1,i}\| \cdot \sum_{j=0}^{t} p_j \cdot \left\| (\mathbf{T}_i^\top)^{t-j} \mathbf{T}_i^{t-j} \right\| \right).$$

The proof follows by substituting for $\|\mathbf{D}_{1,i}\|$ using Lemma 10. ■

Now the task is to upper bound the recursive expression for p_t in Lemma 11. In the next lemma, we explicitly bound the spectral norm of $(\mathbf{T}_i^\top)^k \mathbf{T}_i^k$ for any i and any $k \geq 1$.

Lemma 12. *For any integer $k > 0$,*

$$\left\| (\mathbf{T}_i^\top)^k \mathbf{T}_i^k \right\| \leq 11 \cdot |\gamma_{i1}|^{2k} \cdot h_i^2(k).$$

where $h_i(k) = \min\left\{ \frac{k}{|\gamma_{i1}|}, \frac{2}{|\gamma_{i1} - \gamma_{i2}|} \right\}$ if $\gamma_{i1} \neq \gamma_{i2}$, and $h_i(k) = \frac{k}{|\gamma_{i1}|}$ if $\gamma_{i1} = \gamma_{i2}$.

Proof. We use Lemma 9 to upper bound $\|\mathbf{T}_i^k\|_F^2$. Note that,

$$\begin{aligned}\left\| (\mathbf{T}_i^\top)^k \mathbf{T}_i^k \right\| &\leq \|\mathbf{T}_i^k\|_F^2 \\ &= \left(|\gamma_{i1}|^{2k} + |x_i|^2 |g_i(k)|^2 + |\gamma_{i2}|^{2k} \right),\end{aligned}$$

where $g_i(k) = \sum_{j=0}^{k-1} \gamma_{i1}^j \gamma_{i2}^{k-1-j}$. When $\gamma_{i1} \neq \gamma_{i2}$, then, $|g_i(k)| = |\gamma_{i1}|^k \cdot \min\left\{\frac{k}{|\gamma_{i1}|}, \frac{|1-(\gamma_{i2}/\gamma_{i1})^k|}{|\gamma_{i1}-\gamma_{i2}|}\right\} \leq |\gamma_{i1}|^k \cdot h_i(k)$. Also, when $\gamma_{i1} = \gamma_{i2}$ then, $|g_i(k)| = k|\gamma_{i1}|^{k-1} \leq |\gamma_{i1}|^k \cdot h_i(k)$. Note that $h_i(k) > 1$, and as $|x_i| \leq 3$, we get,

$$\left\|(\mathbf{T}_i^\top)^k \mathbf{T}_i^k\right\| \leq 11 \cdot |\gamma_{i1}|^{2k} \cdot h_i^2(k).$$

■

Now suppose $t > 0$ and our aim is to upper bound p_{t+1}. In the following analysis, we construct an upper bound for p_{t+1}, by defining f_{t+1}, which itself has a recursive definition, but a simpler one compared to p_{t+1}. We will use Lemma 11 and Lemma 12 to build this recursion. For convenience, consider the following notations:

$$\alpha_i(t) := \left\|(\mathbf{T}_i^\top)^t \mathbf{T}_i^t\right\|, \;\; q_i := \frac{5\lambda_i(1-\lambda_i)}{m}, \;\; \ell_i(t) := \max_{0 \leq k \leq t} \left\{11 \cdot |\gamma_{i1}|^k \cdot h_i^2(k)\right\}, \tag{25}$$

where $h_i(k)$ is defined in Lemma 12 for $k > 0$, and for notational convenience we take $h_i(0) = 1$. We have $p_0 = 1$, and

$$\begin{aligned}
p_{t+1} &= \max_{1 \leq i \leq r} \left\{\left\|(\mathbf{T}_i^\top)^{t+1} \mathbf{T}_i^{t+1}\right\| + q_i \cdot \sum_{j=0}^{t} p_j \cdot \left\|(\mathbf{T}_i^\top)^{t-j} \mathbf{T}_i^{t-j}\right\|\right\} \\
&\leq \max_{1 \leq i \leq r} \left\{\alpha_i(t+1) + q_i \cdot \sum_{j=0}^{t} p_j \cdot |\gamma_{i1}|^{t-j} \cdot \ell_i(t-j)\right\} \\
&\leq \max_{1 \leq i \leq r} \left\{\alpha_i(t+1) + q_i \cdot \ell_i(t) \cdot \sum_{j=0}^{t} p_j \cdot |\gamma_{i1}|^{t-j}\right\}.
\end{aligned}$$

Informed by the above relations, we construct upper bounds for p_k for $0 \leq k \leq t+1$, denoted by f_k and defined as follows:

$$\begin{aligned}
&f_0 = 1, \\
&f_{k+1} = \max_{1 \leq i \leq r} \left(\alpha_i(k+1) + q_i \cdot \ell_i(t) \cdot \sum_{j=0}^{k} f_j \cdot |\gamma_{i1}|^{k-j}\right) \text{ for } 0 \leq k \leq t.
\end{aligned}$$

Note that $p_{k+1} \leq f_{k+1}$ for all $0 \leq k \leq t$. For any fixed k such that $0 \leq k \leq t$,

$$\begin{aligned}
f_{k+1} &= \max_{1 \leq i \leq r} \left(\alpha_i(k+1) + q_i \cdot \ell_i(t) \cdot \sum_{j=0}^{k} f_j \cdot |\gamma_{i1}|^{k-j}\right) \\
&= \max_{1 \leq i \leq r} \Bigg(\alpha_i(k+1) + q_i \cdot \ell_i(t) \cdot f_k + |\gamma_{i1}| \cdot \underbrace{q_i \cdot \ell_i(t) \cdot \sum_{j=0}^{k-1} f_j \cdot |\gamma_{i1}|^{k-1-j}}_{*}\Bigg)
\end{aligned}$$

Now note, irrespective of the index i, the term marked as $*$ is always upper bounded by f_k. Therefore,

$$\begin{aligned}
f_{k+1} &\leq \max_{1 \leq i \leq r} \left(\alpha_i(k+1) + \left(|\gamma_{i1}| + q_i \cdot \ell_i(t)\right) \cdot f_k\right) \\
&\leq \max_{1 \leq i \leq r} \left(|\gamma_{i1}| + q_i \cdot \ell_i(t)\right) \cdot f_k + \max_{1 \leq i \leq r} \alpha_i(k+1) \\
&\leq \rho \cdot f_k + \max_{1 \leq i \leq r} \alpha_i(k+1).
\end{aligned}$$

where we let $\rho := \max_{1 \leq i \leq r} \left(|\gamma_{i1}| + q_i \cdot \ell_i(t)\right)$. Now, using the above recursive relation we can upper bound f_{t+1} as,

$$\begin{aligned}
f_{t+1} &\leq \rho \cdot f_t + \max_{1 \leq i \leq r} \alpha_i(t+1) \\
&\leq \rho^2 \cdot f_{t-1} + \rho \cdot \max_{1 \leq i \leq r} \alpha_i(t) + \max_{1 \leq i \leq r} \alpha_i(t+1) \\
&\leq \rho^{t+1} + \sum_{j=0}^{t} \rho^j \cdot \max_{1 \leq i \leq r} \alpha_i(t+1-j).
\end{aligned} \tag{26}$$

We need to bound ρ. In the following lemma, we provide a bound on the term $|\gamma_{i1}| + q_i \cdot \ell_i(t)$ for any i. For simplicity, we provide the proof for $\omega \in \{0, 1\}$ (which covers both HBM and NAG). The proof for $\omega \in (0, 1)$ is completely analogous and follows by using a more general version of Lemma 14.

Lemma 13. *Let* $2 \leq \phi \leq \min\left\{\frac{m}{C}, \frac{1}{3\sqrt{\lambda_r}}\right\}$ *for* $C > 96000$ *and* $\omega \in \{0, 1\}$. *Then for any* $1 \leq i \leq r$,

$$|\gamma_{i1}| + q_i \cdot \ell_i(t) \leq 1 - \frac{\phi\lambda_r}{4}.$$

Proof. As we are proving the result for a fixed i, we drop the subscript i for notational convenience. The proof proceeds in cases:

1. *Case 1:* γ_1 *is complex.* We know from Lemma 7 that $|\gamma_1|^2 = \left(1 - \frac{1}{\phi}\right)(1 - \omega\lambda)$. We get

$$|\gamma_1| + q \cdot \ell(t) < \left(1 - \frac{1}{2\phi}\right) + \frac{5\lambda(1-\lambda)}{m} \cdot \ell(t). \tag{27}$$

Recall $\ell(t) = \max_{0 \leq k \leq t}\left\{11 \cdot |\gamma_1|^k \cdot h^2(k)\right\}$ and $h(k) = \min\left\{\frac{k}{|\gamma_1|}, \frac{2}{|\gamma_1 - \gamma_2|}\right\}$. Let k^* be the index such that $\ell(t) = 11 \cdot |\gamma_1|^{k^*} \cdot h^2(k^*)$. We get,

$$|\gamma_1| + q \cdot \ell(t) < \left(1 - \frac{1}{2\phi}\right) + \frac{55\lambda(1-\lambda)}{m} \cdot h^2(k^*) \cdot |\gamma_1|^{k^*}.$$

Now look at $h^2(k^*)$,

$$h^2(k^*) = \min\left\{\frac{(k^*)^2}{|\gamma_1|^2}, \frac{4}{|\gamma_1 - \gamma_2|^2}\right\}.$$

Due to Lemma 7, we know

$$\begin{aligned}|\gamma_1 - \gamma_2|^2 &= \Big(4\beta(1 - \omega\lambda) - \big((1-\lambda) + \beta(1-\omega\lambda)\big)^2\Big) \\ &= (1-\lambda) \cdot \big(4\beta\delta - (1+\beta\delta)^2(1-\lambda)\big),\end{aligned}$$

where $\delta = \frac{1-\omega\lambda}{1-\lambda}$. Also, as $\beta = 1 - \frac{1}{\phi} \geq 0.5$, we have $\frac{1}{|\gamma_1|^2} \leq \frac{2}{1-\omega\lambda} \leq \frac{2}{1-\lambda}$. Therefore,

$$|\gamma_1| + q \cdot \ell(t) < \left(1 - \frac{1}{2\phi}\right) + \frac{55\lambda}{m} \cdot \min\left\{2(k^*)^2, \frac{4}{4\beta\delta - (1+\beta\delta)^2(1-\lambda)}\right\} \cdot |\gamma_1|^{k^*} \tag{28}$$

Let $\epsilon = 4\beta\delta - (1+\beta\delta)^2(1-\lambda)$. We get,

$$\begin{aligned}1 - \lambda &= \frac{4\beta\delta}{(1+\beta\delta)^2} - \frac{\epsilon}{(1+\beta\delta)^2} \\ \lambda &= 1 - \frac{4\beta\delta}{(1+\beta\delta)^2} + \frac{\epsilon}{(1+\beta\delta)^2} \\ &= \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2} + \frac{\epsilon}{(1+\beta\delta)^2}\end{aligned} \tag{29}$$

We consider two subcases:

(a) *Subcase-1:* $\epsilon > \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$. Using (29) we get,

$$\frac{\lambda}{4\beta\delta - (1+\beta\delta)^2(1-\lambda)} < 1 + \frac{1}{(1+\beta\delta)^2} < 2. \tag{30}$$

(b) *Subcase-2:* $\epsilon \leq \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$. Using (29) we have $\lambda \leq \left(1 + \frac{1}{(1+\beta\delta)^2}\right)\frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$. Note that $\delta \in \{1, \frac{1}{1-\lambda}\}$ and $\beta \geq 0.5$. Using item 1 from Lemma 14 with $x = \delta$ we get,

$$\lambda \leq \left(1 + \frac{1}{(1+\beta)^2}\right)\frac{(1-\beta)^2}{(1+\beta)^2} < \frac{3}{2(2\phi-1)^2}, \quad \text{(as } \beta = 1 - \frac{1}{\phi}\text{)}.$$

We get,

$$\lambda \cdot (k^*)^2 \cdot |\gamma_1|^{k^*} \leq \frac{3}{2(2\phi-1)^2} \cdot (k^*)^2 \cdot |\gamma_1|^{k^*}.$$

The quantity $(k^*)^2 \cdot |\gamma_1|^{k^*}$ is maximized at $k^* = \frac{-2}{\log|\gamma_1|}$, implying

$$\lambda \cdot (k^*)^2 \cdot |\gamma_1|^{k^*} \leq \frac{3}{2(2\phi-1)^2} \cdot \frac{4}{(1-|\gamma_1|)^2} \cdot \frac{1}{e^2},$$

where in the last inequality we used the relation $\frac{-2}{\log|\gamma_1|} < \frac{2}{1-|\gamma_1|}$ and $|\gamma_1|^{k^*} = e^{-2}$, with e being Euler's number. As $|\gamma_1| < 1 - \frac{1}{2\phi}$, we have $\frac{1}{(1-|\gamma_1|)^2} < 4\phi^2$. Therefore,

$$\lambda \cdot (k^*)^2 \cdot |\gamma_1|^{k^*} \leq \frac{3}{2(2\phi-1)^2} \cdot \frac{16\phi^2}{e^2} \leq \frac{3}{4\phi^2} \cdot \frac{16\phi^2}{e^2} < 2. \tag{31}$$

Depending on the above two subcases, we substitute (30) and (31) in (28) to prove

$$|\gamma_1| + q \cdot \ell(t) < \left(1 - \frac{1}{2\phi}\right) + \frac{55 \cdot 8}{m}.$$

If $\phi \leq \frac{m}{1760}$, we get,

$$|\gamma_1| + q \cdot \ell(t) < 1 - \frac{1}{4\phi}. \tag{32}$$

2. *Case 2:* γ_1 is real and $\gamma_1 = \gamma_2$. In this case we have,

$$\lambda = \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2} \leq \frac{(1-\beta)^2}{(1+\beta)^2} = \frac{1}{(2\phi-1)^2},$$

where the last inequality follows from item 2 from Lemma 14. The proof is exactly same to subcase-2 of the first case (complex γ_1). Using Lemma 7 we have, $\gamma_1^2 = \left(1 - \frac{1}{\phi}\right)(1-\omega\lambda)$. Therefore,

$$|\gamma_1| + q \cdot \ell(t) < \left(1 - \frac{1}{2\phi}\right) + \frac{5\lambda(1-\lambda)}{m} \cdot \ell(t).$$

Recall $\ell(t) = \max_{0 \leq k \leq t} \left\{11 \cdot |\gamma_1|^k \cdot h^2(k)\right\}$ and $h(k) = \frac{k}{|\gamma_1|}$. Let k^* be the index such that $\ell(t) = 11 \cdot |\gamma_1|^{k^*} \cdot h^2(k^*)$. We get,

$$\begin{aligned}
|\gamma_1| + q \cdot \ell(t) &< \left(1 - \frac{1}{2\phi}\right) + \frac{55\lambda(1-\lambda)}{m} \cdot \frac{(k^*)^2}{|\gamma_1|^2} \cdot |\gamma_1|^{k^*} \\
&< \left(1 - \frac{1}{2\phi}\right) + \frac{110\lambda}{m} \cdot (k^*)^2 \cdot |\gamma_1|^{k^*} \qquad \left(\text{as } \gamma_1^2 \geq \frac{1}{2}(1-\lambda)\right) \\
&< \left(1 - \frac{1}{2\phi}\right) + \frac{110}{m} \cdot \frac{1}{(2\phi-1)^2} \cdot \frac{16\phi^2}{e^2} \\
&< \left(1 - \frac{1}{2\phi}\right) + \frac{125}{m}.
\end{aligned}$$

Therefore, for $\phi < \frac{m}{500}$, we get

$$|\gamma_1| + q \cdot \ell(t) < 1 - \frac{1}{4\phi}. \tag{33}$$

3. *Case 3:* γ_1 is real with $\gamma_1 > \gamma_2$. Using Lemma 7, we have $\gamma_1 < \left(1 - \frac{\phi\lambda}{2}\right)(1-\omega\lambda)$. Therefore,

$$\gamma_1 + q \cdot \ell(t) < \left(1 - \frac{\phi\lambda}{2}\right) + \frac{5\lambda(1-\lambda)}{m} \cdot \ell(t).$$

Here again, $\ell(t) = \max_{0\le k\le t}\left\{11\cdot|\gamma_1|^k\cdot h^2(k)\right\}$ and $h(k) = \min\left\{\frac{k}{\gamma_1}, \frac{2}{\gamma_1-\gamma_2}\right\}$. Let k^* be the index such that $\ell(t) = 11\cdot|\gamma_1|^{k^*}\cdot h^2(k^*)$. We get,

$$\gamma_1 + q\cdot\ell(t) < \Big(1-\frac{\phi\lambda}{2}\Big) + \frac{55\lambda(1-\lambda)}{m}\cdot h^2(k^*)\cdot\gamma_1^{k^*}, \tag{34}$$

where,

$$h^2(k^*) = \min\Big\{\frac{(k^*)^2}{\gamma_1^2}, \frac{4}{(\gamma_1-\gamma_2)^2}\Big\}.$$

We have

$$\begin{aligned}(\gamma_1-\gamma_2)^2 &= \Big(\big((1-\lambda)+\beta(1-\omega\lambda)\big)^2 - 4\beta(1-\omega\lambda)\Big)\\ &= (1-\lambda)\Big((1+\beta\delta)^2(1-\lambda) - 4\beta\delta\Big),\end{aligned}$$

where $\delta = \frac{1-\omega\lambda}{1-\lambda}$. Let $\epsilon = (1+\beta\delta)^2(1-\lambda) - 4\beta\delta$. Note that this yields

$$\lambda = \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2} - \frac{\epsilon}{(1+\beta\delta)^2}.$$

Moreover, let $a(\lambda) = \lambda - \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$ and let λ' denote the zero of $a(\lambda)$ in $(0,1)$. It turns out that $a(\lambda)$ is an increasing function for all λ satisfying $a(\lambda) < 0$. As λ in our case satisfies $a(\lambda) < 0$, we get

$$\lambda < \lambda' < \frac{(1-\beta)^2}{(1+\beta)^2} = \frac{1}{(2\phi-1)^2} \le \frac{1}{9}. \tag{35}$$

where we used item 2 from Lemma 14 again for the second last inequality. The last inequality used $\phi \ge 2$. Consider the following two subcases:

(a) *Subcase 1:* $0 < \epsilon \le \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$. We get,

$$\lambda \ge \Big(1-\frac{1}{(1+\beta\delta)^2}\Big)\frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}.$$

Using item 3 from Lemma 14 we get

$$\lambda > \frac{(1-\beta)^2}{2(1+\beta)^2} = \frac{1}{2(2\phi-1)^2} > \frac{1}{8\phi^2}. \tag{36}$$

We know that (see proof of Lemma 7),

$$\gamma_1 > \frac{1}{2}\Big((1-\lambda)+\beta(1-\omega\lambda)\Big) > \frac{1}{2}(1+\beta)(1-\lambda) > \frac{3}{4}(1-\lambda) > \frac{2}{3} \quad \text{(by (35))}$$

Therefore,

$$\begin{aligned}\gamma_1 + q\cdot\ell(t) &< \Big(1-\frac{\phi\lambda}{2}\Big) + \frac{55\lambda(1-\lambda)}{m}\cdot\frac{(k^*)^2}{\gamma_1^2}\cdot\gamma_1^{k^*}\\ &< \Big(1-\frac{\phi\lambda}{2}\Big) + \frac{125\lambda}{m}\cdot(k^*)^2\cdot\gamma_1^{k^*} \quad (\text{as } \gamma_1 > \frac{2}{3}).\end{aligned}$$

As $(k^*)^2\cdot\gamma_1^{k^*}$ is maximized at $k^* = \frac{-2}{\log\gamma_1}$ and $\gamma_1 < 1-\frac{\phi\lambda}{2}$, we get

$$\gamma_1 + q\cdot\ell(t) < \Big(1-\frac{\phi\lambda}{2}\Big) + \frac{125\lambda}{m}\cdot\frac{3}{\phi^2\lambda^2}.$$

Finally using (36) we get,

$$\begin{aligned}\gamma_1 + q\cdot\ell(t) &< \Big(1-\frac{1}{16\phi}\Big) + \frac{3000}{m}\\ &< 1-\frac{1}{32\phi},\end{aligned} \tag{37}$$

where last inequality follows for $\phi \le \frac{m}{96000}$.

(b) *Subcase 2:* $\epsilon > \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}$. In this case we can simply lower bound $(\gamma_1 - \gamma_2)^2$. Note that,

$$\lambda < \Big(1 - \frac{1}{(1+\beta\delta)^2}\Big) \cdot \frac{(1-\beta\delta)^2}{(1+\beta\delta)^2}.$$

Using item 4 from Lemma 14 with $x = \delta$ we get $\frac{(1-\beta\delta)^2}{(1+\beta\delta)^2} \geq \frac{1}{2} \cdot \frac{(1-\beta)^2}{(1+\beta)^2}$. Therefore,

$$\begin{aligned}(\gamma_1 - \gamma_2)^2 &= (1-\lambda)\cdot\epsilon \\ &> (1-\lambda)\cdot\frac{(1-\beta\delta)^2}{(1+\beta\delta)^2} \\ &\geq (1-\lambda)\cdot\frac{(1-\beta)^2}{(1+\beta)^2} = (1-\lambda)\cdot\frac{1}{(2\phi-1)^2}.\end{aligned}$$

Substituting the above in (34) we get,

$$\begin{aligned}\gamma_1 + q\cdot\ell(t) &< 1 - \frac{\phi\lambda}{2} + \frac{220\lambda}{m}\cdot(2\phi-1)^2 \\ &< 1 - \frac{\phi\lambda}{2} + \frac{880\lambda\phi^2}{m}\end{aligned}$$

For $\phi < \frac{m}{3520}$ we get,

$$\gamma_1 + q\cdot\ell(t) < 1 - \frac{\phi\lambda}{4}. \tag{38}$$

Combining the inequalities (32), (33), (37) and (38) we get for $\phi \leq \min\{\frac{m}{C}, \frac{1}{3\sqrt{\lambda_r}}\}$ and $C > 96000$,

$$|\gamma_1| + q\cdot\ell(t) \leq 1 - \frac{\phi\lambda_r}{4}.$$

This finishes the proof. ■

Finishing the proof of Theorem 4. We are now ready to complete the proof of Theorem 4. Recall the bound on p_{t+1} in (26). We had,

$$p_{t+1} \leq f_{t+1} < \rho^{t+1} + \sum_{j=0}^{t} \rho^j \cdot \max_i \alpha_i(t+1-j)$$

Due to Lemma 13 we know,

$$\rho \leq 1 - \frac{\phi\lambda_r}{4}.$$

Furthermore, $\alpha_i(t+1-j) = \big\|(\mathbf{T}_i^\top)^{t+1-j}\mathbf{T}_i^{t+1-j}\big\|$, and we know from Lemma 9 that

$$\begin{aligned}\big\|(\mathbf{T}_i^\top)^{t+1-j}\mathbf{T}_i^{t+1-j}\big\| &\leq \big(3(t+1-j)+2\big)^2\cdot|\gamma_{i1}|^{2(t-j)} \\ &\leq \big(3t+5\big)^2\cdot|\gamma_{i1}|^{2(t-j)} \\ &\leq \big(3t+5\big)^2\cdot\rho^{t-j}.\end{aligned}$$

Therefore,

$$\begin{aligned}p_{t+1} &< \rho^{t+1} + (t+1)\cdot\big(3t+5\big)^2\cdot\rho^t \\ &< 2(t+1)\cdot\big(3t+5\big)^2\cdot\rho^t \\ &< 50(t+1)^3\cdot\rho^t.\end{aligned}$$

Note that p_{t+1} upper bounds $\Big\|\mathbb{E}\Big[\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_j^\top\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big(\prod_{j=0}^{t}\mathbf{Q}\mathbf{P}\mathbf{Y}_{t-j}\mathbf{P}^\top\mathbf{Q}^\top\Big)\Big]\Big\|$ and this finishes the proof of Theorem 4.

The following helper lemma is used in the proof of Lemma 13 for $\omega = 0$. A more general variant can be used to prove Lemma 13 for any $\omega \in (0,1)$. The proof of Lemma 14 follows by a routine calculus exercise.

Lemma 14. *Let $x \geq 1$, $\frac{1}{2} \leq \beta < 1$, and define*

$$f(x) = \left(1 + \frac{1}{(1+\beta x)^2}\right)\frac{(1-\beta x)^2}{(1+\beta x)^2}, \quad g(x) = \left(1 - \frac{1}{(1+\beta x)^2}\right)\frac{(1-\beta x)^2}{(1+\beta x)^2}$$

1. *Over the set of all $x \geq 1$ satisfying $1 - \frac{1}{x} \leq f(x)$, the function $f(x)$ is maximized at $x = 1$. In particular, $f(x) < \frac{3}{2} \cdot \frac{(1-\beta)^2}{(1+\beta)^2}$.*
2. *Over the set of all $x \geq 1$ satisfying $1 - \frac{1}{x} = \frac{(1-\beta x)^2}{(1+\beta x)^2}$, we have $\frac{(1-\beta x)^2}{(1+\beta x)^2} \leq \frac{(1-\beta)^2}{(1+\beta)^2}$.*
3. *Over the set of all $x \geq 1$ satisfying $1 - \frac{1}{x} < \frac{(1-\beta x)^2}{(1+\beta x)^2}$, $\inf_x g(x) > \frac{4}{5} \cdot \frac{(1-\beta)^2}{(1+\beta)^2}$.*
4. *Over all the set of all $x \geq 1$ satisfying $1 - \frac{1}{x} < g(x)$, $\inf_x \frac{(1-\beta x)^2}{(1+\beta x)^2} > \frac{1}{2} \cdot \frac{(1-\beta)^2}{(1+\beta)^2}$.*

D Additional experiments

In this section, we provide additional details about our experimental setup, and include numerical results on two datasets to complement the results in Section 4. Finally, we also explore the dependence of classical momentum on the ω parameter, which interpolates between HBM and NAG.

D.1 Experimental setup

We transform the data matrices through `StandardScaler` from `sklearn.preprocessing`. In particular, after sampling $n = 2048$ rows from the dataset, we perform the following transformation,

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
K = rbf_kernel(X_scaled, gamma=0.1)
```

We consider the following values for λ and ϵ tolerance in our experiment:

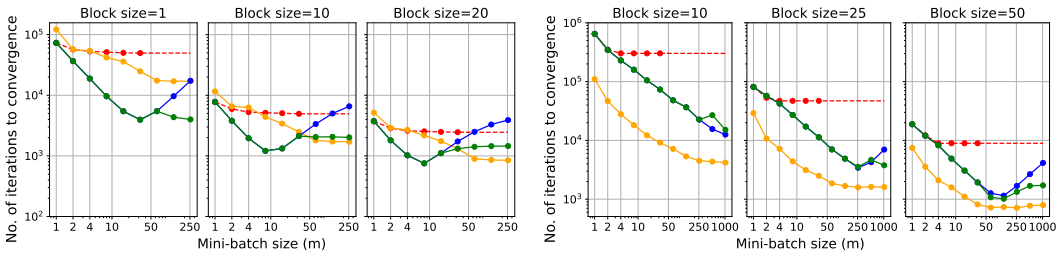
1. `Covtype`, $\lambda = 0.5$, $\epsilon = 10^{-7}$.
2. `California Housing`, $\lambda = 0.16$, $\epsilon = 10^{-7}$.
3. `Abalone`, $\lambda = 20$, $\epsilon = 10^{-10}$.
4. `Phoneme`, $\lambda = 10^{-10}$, $\epsilon = 10^{-6}$.

All experiments were conducted on a MacBook Air (M3) with 16 GB of memory.

D.2 Plots on `Abalone` and `Phoneme` datasets

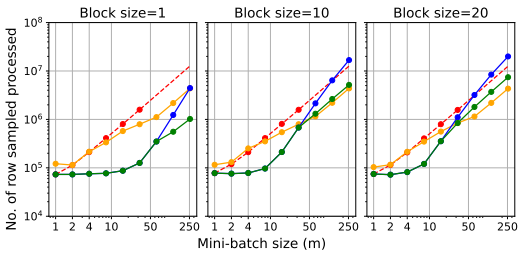
In Fig 3, we compare, **CD**, **CDpp**, **CD+NAG** ($\beta = 1 - 1/m$), and **CD+NAG** ($\beta =$ adaptive) on two additional datasets. The results on `Abalone` dataset follow the same pattern as observed in Fig 1 for `California Housing` and `Covtype` datasets. However, the results on `Phoneme` dataset look different, as **CDpp** performs better than **CD+NAG**. We provide insights into this phenomenon in the following paragraph.

Comparison between CDpp and CD+NAG. Under certain incoherence assumptions on the data matrix $\mathbf{K}$, the condition number of $\bar{\mathbf{\Pi}}$ can be bounded as $O\big(\frac{n\bar{\kappa}_k}{k}\big)$, where k is the block-size and $\bar{\kappa}_k$ denotes a certain stochastic condition number quantity (see [11] for details). In this case, the iteration complexity of **CDpp** with mini-batch size m can be shown to scale as $\tilde{O}\big(\sqrt{\frac{\bar{\kappa}_k}{m}} \cdot \frac{n}{k} + \sqrt{\bar{\kappa}_k \frac{n}{k}}\big)$, whereas the iteration complexity of **CD+NAG** scales as $\tilde{O}\big(\frac{\bar{\kappa}_k}{m} \cdot \frac{n}{k} + \sqrt{\bar{\kappa}_k \frac{n}{k}}\big)$. Therefore, when $m < \bar{\kappa}_k$, **CDpp** can converge faster than **CD+NAG**, but when $m \geq \bar{\kappa}_k$, the perfect parallelization phenomenon takes over and **CD+NAG** becomes faster. While this discussion requires some additional assumptions on the data, it provides some intuition for why one type of acceleration may be preferable over the other, depending on the conditioning properties of the dataset.

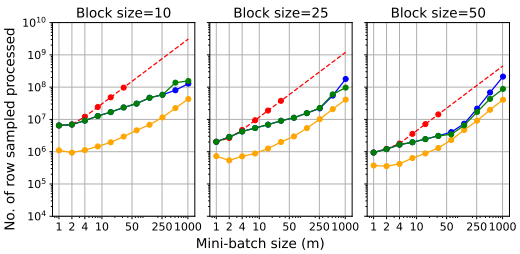


(a) Abalone: iterations vs mini-batch size

(b) Phoneme: iterations vs mini-batch size



(c) Abalone: row samples vs mini-batch size



(d) Phoneme: row samples vs mini-batch size

CD CDpp CD + NAG ($\beta = 1 - 1/m$) CD + NAG ($\beta =$ adaptive)

Figure 3: Convergence of block coordinate descent with NAG momentum and baselines. Top row: iterations to reach error 10^{-10} for Abalone and 10^{-6} for Phoneme. Bottom row: total work, measured by the number of sampled rows.

D.3 Adaptive momentum

Next, we describe the implementation of **CD+NAG** ($\beta =$ adaptive), where we select the momentum parameter β adaptively during the course of the optimization. Consider solving the positive definite linear system $\mathbf{K}\mathbf{w} = \mathbf{y}$ using block coordinate descent with NAG momentum, and with mini-batch size m (here, for simplicity we absorb the regularizer λ into $\mathbf{K}$). Our aim is to adaptively find $\phi^* \in [2, m]$ such that $\beta^* = 1 - \frac{1}{\phi}$ is a near-optimal momentum parameter for mini-batch size m.

We provide the full heuristic logic for adaptive momentum tuning in Algorithm 1. The matrix $\tilde{\mathbf{K}}$ and vector $\tilde{\mathbf{b}}$ are simply a few randomly sampled rows from $\mathbf{K}$ and $\mathbf{b}$, used to approximate the true residuals. In all our experiments, we sampled 100 rows randomly from the data matrices for approximating the true residual.

E Proofs omitted from Section 2

Proof of Proposition 1 As $\mathbf{\Delta}_{t+1} = \mathcal{A}_t(\mathbf{\Delta}_t; \alpha) = (\mathbf{I} - \alpha\,\mathbf{\Pi}_t)\mathbf{\Delta}_t$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{\Delta}_{t+1}\|^2 &= \mathbf{\Delta}_t^\top \mathbb{E}[(\mathbf{I} - \alpha\,\mathbf{\Pi}_t)^2]\mathbf{\Delta}_t \\
&\le \mathbf{\Delta}_t^\top \mathbb{E}[(\mathbf{I} - \alpha\,\mathbf{\Pi}_t)]\mathbf{\Delta}_t \qquad \text{(since } \mathbf{0} \preceq \mathbf{\Pi}_t \preceq \mathbf{I} \text{ and } 0 < \alpha \le 1) \\
&= \mathbf{\Delta}_t^\top (\mathbf{I} - \alpha\bar{\mathbf{\Pi}})\mathbf{\Delta}_t.
\end{aligned}$$

As $\mathbf{\Delta}_0 \in \text{range}(\bar{\mathbf{\Pi}})$, all $\mathbf{\Delta}_t \in \text{range}(\bar{\mathbf{\Pi}})$ with probability 1. Therefore, we get

$$\mathbb{E}\|\mathbf{\Delta}_{t+1}\|^2 \le \left(1 - \frac{1}{\kappa}\right)\|\mathbf{\Delta}_0\|^2,$$

where $\kappa = 1/\alpha\lambda_{\min}^+(\bar{\mathbf{\Pi}})$.

Proof of Proposition 2 It is easy to see that

$$\mathbf{\Pi}^{[m]} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{\Pi}^{(i)},$$

Algorithm 1 Adaptive momentum tuning for **CD+NAG**

Require: $\mathbf{K}$, $\tilde{\mathbf{K}}$, $\mathbf{b}$, $\tilde{\mathbf{b}}$, m, maximum iterations T
Require: Warm-up length T_{warm}, check interval T_{check}
Initialize $\mathbf{w}_{-1} = 0$, $\mathbf{x}_{-1} = 0$
Set initial momentum $\beta \leftarrow 1 - \frac{1}{m}$
▷ Warm-up phase
for $t = 0, \ldots, T_{\mathrm{warm}} - 1$ **do**
  $\mathbf{x}_{t+1} \leftarrow$ **mini-batch CD step**$(\mathbf{w}_t, m)$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{x}_{t+1} + \beta(\mathbf{x}_{t+1} - \mathbf{x}_t)$
end for
▷ Adaptive phase
$\beta_+ = 1 - \frac{1}{m}$
$\beta_- = 1 - \frac{1}{2}$
$(\mathbf{w}_{+,t}, \mathbf{x}_{+,t}) \leftarrow (\mathbf{w}_t, \mathbf{x}_t)$
$(\mathbf{w}_{-,t}, \mathbf{x}_{-,t}) \leftarrow (\mathbf{w}_t, \mathbf{x}_t)$
while $t < T$ and $\frac{1-\beta_-}{1-\beta_+} > 2$ **do**
  for $i \in \{+, -\}$ **do**
    $\mathbf{x}_{i,t+1} \leftarrow$ **mini-batch CD step**$(\mathbf{w}_{i,t}, m)$
    $\mathbf{w}_{i,t+1} \leftarrow \mathbf{x}_{i,t+1} + \beta_i(\mathbf{x}_{i,t+1} - \mathbf{x}_{i,t})$
    if $t \bmod T_{\mathrm{check}} = 0$ **then**
      $r_i \leftarrow \|\mathbf{K}\mathbf{x}_{i,t+1} - \tilde{\mathbf{b}}\| / \|\tilde{\mathbf{b}}\|$
      $R \leftarrow r_+ / r_-$
      if $R < 1$ **then** $p \leftarrow p + 1$
      else $p \leftarrow 0$
      end if
      if $R \geq 3$ **then** ▷ Large momentum is too aggressive
        $\phi_+ \leftarrow \frac{1}{2}\frac{1}{1-\beta_+}$, $\beta_+ \leftarrow 1 - \frac{1}{\phi_+}$
        $\mathbf{w}_{+,t+1} \leftarrow \mathbf{w}_{-,t+1}$, $\mathbf{x}_{+,t+1} \leftarrow \mathbf{x}_{-,t+1}$
        $p \leftarrow 0$
      else if $R \leq 1/2$ or $p > 20$ **then** ▷ Small momentum is too conservative
        $\phi_- \leftarrow \frac{2}{1-\beta_-}$, $\beta_- \leftarrow 1 - \frac{1}{\phi_-}$
        $\mathbf{w}_{-,t+1} \leftarrow \mathbf{w}_{+,t+1}$, $\mathbf{x}_{-,t+1} \leftarrow \mathbf{x}_{+,t+1}$
        $p \leftarrow 0$
      end if
    end if
  end for
  $t \leftarrow t + 1$
end while
$\beta \leftarrow \beta_+$
$(\mathbf{w}_t, \mathbf{x}_t) \leftarrow (\mathbf{w}_{+,t}, \mathbf{x}_{+,t})$
while $t < T$ **do** ▷ **CD+NAG** with optimal β
  $\mathbf{x}_{t+1} \leftarrow$ **mini-batch CD step**$(\mathbf{w}_t, m)$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{x}_{t+1} + \beta(\mathbf{x}_{t+1} - \mathbf{x}_t)$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}_t$

where $\mathbf{\Pi}^{(i)}$ is the stochastic rate matrix associated with $\mathcal{A}^{(i)}$. Therefore,

$$\begin{aligned}
\mathbb{E}\Big[(\mathbf{\Pi}^{[m]} - \bar{\mathbf{\Pi}})^2\Big] &= (\mathbf{\Pi}^{[m]})^2 + \bar{\mathbf{\Pi}}^2 - \mathbb{E}[\mathbf{\Pi}^{[m]}] \cdot \bar{\mathbf{\Pi}} - \bar{\mathbf{\Pi}} \cdot \mathbb{E}[\mathbf{\Pi}^{[m]}] \\
&= \big(\mathbf{\Pi}^{[m]}\big)^2 - \bar{\mathbf{\Pi}}^2 \\
&= \frac{1}{m^2} \cdot \sum_{i=1}^{m} \mathbb{E}\big[\big(\mathbf{\Pi}^{(i)}\big)^2\big] + \frac{1}{m^2} \cdot \sum_{i,j=1, i \neq j}^{m} \mathbb{E}\big[\mathbf{\Pi}^{(i)}\,\mathbf{\Pi}^{(j)}\big] - \bar{\mathbf{\Pi}}^2 \\
&= \frac{1}{m} \cdot \mathbb{E}\big[\big(\mathbf{\Pi}^{(1)}\big)^2\big] - \frac{1}{m} \cdot \bar{\mathbf{\Pi}}^2 \\
&\preceq \frac{1}{m} \cdot \bar{\mathbf{\Pi}}(\mathbf{I} - \bar{\mathbf{\Pi}}).
\end{aligned}$$

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We provide theoretical results in Section 3 establishing all the claims made in abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Section 2 we describe that the scope of our theoretical framework is limited to quadratics in the interpolation regime.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Our main result Theorem 1 holds under Definition 1, 2 and 3.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We disclose our pseudocode and data preparation details in Appendix D

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The datasets used in our experiments are publicly available through Scikit-learn [41] and OpenML [49]. We include the code in the supplementary, and plan to provide a link to a public repository upon publication.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Appendix D

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: We perform just 3 independent trials for every experiment and report the median performance. In fact, just 1 trial also leads to similar looking experimental plots, making error bars unnecessary for problem at hand.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix D

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: [N/A]

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification:[N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite Scikit-learn [41] and OpenML [49] as our data sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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