Understanding Progressive Training Through the Framework of Randomized Coordinate Descent

Rafał Szlendak
AI Initiative
KAUST, Saudi Arabia

Elnur Gasanov
AI Initiative
KAUST, Saudi Arabia

Peter Richtárik
AI Initiative
KAUST, Saudi Arabia

Abstract

We propose a Randomized Progressive Training algorithm (RPT) – a stochastic proxy for the well-known Progressive Training method (PT) [Karras et al. 2018]. Originally designed to train GANs (Goodfellow et al., 2014), PT was proposed as a heuristic, with no convergence analysis even for the simplest objective functions. On the contrary, to the best of our knowledge, RPT is the first PT-type algorithm with rigorous and sound theoretical guarantees for general smooth objective functions. We cast our method into the established framework of Randomized Coordinate Descent (RCD) [Nesterov 2012; Richtárik and Takáč 2014], for which (as a by-product of our investigations) we also propose a novel, simple and general convergence analysis encapsulating strongly-convex, convex and nonconvex objectives. We then use this framework to establish a convergence theory for RPT. Finally, we validate the effectiveness of our method through extensive computational experiments.

1 INTRODUCTION

The modern practice of supervised learning typically relies on training huge-dimensional, overparametrized models using first-order gradient methods. However, in most cases, using the full gradient information in the training procedure is computationally infeasible. The gold-standard approach of bypassing this difficulty is provided by the family of Stochastic Gradient Descent (SGD) algorithms which use only a partial information about the gradients for the reward of reduced compute time. Progressive Training (PT) [Karras et al. 2018], albeit a non-stochastic method, relies on a similar idea. In its original form, it was developed as a method for training generative adversarial networks (GANs) (Goodfellow et al. 2014) by progressively growing the neural network that we are training. Specifically, for a fixed period of time we only train the first layer of our GAN while keeping the weights corresponding to the other layers intact; then, we perform updates with respect to the first two layers, and so on.

Karras et al. (2018) justified two main advantages of this approach: an increased stability of the training (since the training procedure starts from a simpler, reduced model), and reduced computation cost. The method was proposed as a heuristic, with no convergence guarantees even for the simplest objectives, such as quadratics or convex functions.

Observe that in the PT procedure, the domain of the parameters which we are trying to learn is partitioned into blocks which correspond to the respective layers of the neural network that PT aims to train. Motivated by this, in this work we are going to be concerned with minimizing the general objective functions $f$ whose domain will be written as $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}$. For example, $\mathbb{R}^{d_i}$ might represent the parameters of the $i$th layer of the neural network in question.

Algorithm 1 offers a formal description of the PT procedure applied to the general differentiable function $f : \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B} \to \mathbb{R}$. The training procedure is divided into $B$ epochs; in each epoch $b \in [B]$ we want to update the first $b$ coordinate blocks. Let $G_b f(x) \subseteq \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}$ denote the vector $$(\nabla_1 f(x), \ldots, \nabla_b f(x), 0, \ldots, 0),$$ where $\nabla_i f$ is the gradient of $f$ with respect to the $i$th coordinate block. In each epoch $b$ we make the gradient update $$x \leftarrow x - \gamma_b G_b f(x)$$ for $T_b$ iterations, where $\gamma_b > 0$ is a suitably chosen
stepsizes which is fixed for the entire epoch $b$.

**Algorithm 1 Progressive Training**

1: **Input:** initial iterate $x \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}$; stepsizes $\gamma_1, \gamma_2, \ldots, \gamma_n > 0$; integers $T_1, T_2, \ldots, T_n \geq 0$
2: for $b = 1, \ldots, B$ do
3: for $t = 0, \ldots, T_b - 1$ do
4: $x \leftarrow x - \gamma_b G_b f(x)$
5: end for
6: end for
7: **Output:** $x$

1.1 Existing literature

Wang et al. (2022) attempt to analyze the federated version of PT called ProgFed. Unfortunately, as argued in Appendix A, their analysis is vacuous.

Recently, a novel and related research thread emerged around the idea of independent subnet training (IST) which relies on partitioning the neural network in question into sparser subnets, training them in parallel across the devices and periodically aggregating the model. Yuan et al. (2022) introduce IST and provide its analysis. However, the authors make rather strong assumptions on the objective function, such as Lipschitz continuity of the objective function and certain assumptions on the stochastic gradient error. Analogous technique is applied by Dun et al. (2022) for training ResNet (He et al., 2016). It is important to note that the IST framework differs from RPT in several key components. Indeed, IST partitions the neural network in question into sparser submodels which are then trained across the devices in parallel. The model is then aggregated and this procedure repeats. In contrast, RPT performs a single update at a time of a random subnetwork chosen in a very particular way (which simulates the deterministic PT). Moreover, IST is a distributed method; in contrast, in our paper we focus on the single-node setup.

Notably, none of the aforementioned papers contains mathematically sound analysis for general smooth functions.

1.2 Our contributions and organization of the paper

As we have argued, there is still a considerable gap in theoretical understanding of the PT mechanism. The key objective of this paper is fixing this gap by proposing a randomized proxy for PT which mimics its behavior and has simple and general analysis with mild assumptions. The contributions of our work can be summarized as follows:

- In Section 2 we develop a simple and general analysis of Randomized Coordinate Descent (RCD) for strongly-convex, convex and nonconvex objectives. Our analysis recovers previous convergence rates of RCD and allows us to fully characterize its iteration complexity in strongly-convex, convex and nonconvex regimes via a single quantity $L_P$.

- In Sections 3 and 4 we propose a method called RPT which mimics the behavior of PT, and, as we argue, is the first PT-type method with sound theoretical guarantees for smooth objectives. Instead of progressively expanding the model which we are trying to optimize, in each iteration RPT makes a choice of the submodel according to a carefully chosen randomness, and then makes a gradient update of this model. We then view RPT as a particular instance of RCD, which immediately establishes its convergence. We finally identify regimes in which RPT can be advantageous over vanilla gradient descent (GD) through a novel and general notion of total computation complexity.

- In Section 5 we present a series of empirical experiments that demonstrate the advantages of our proposed method. These experiments are divided into three sub-sections. The first sub-section employs synthetic data to demonstrate that, when appropriately generated, our proposed Randomized Progressive Training (RPT) method can converge at a faster rate than Gradient Descent while also incurring less computation cost. In the second and third sub-sections, we utilize real-world datasets to evaluate the performance of the RPT method, including testing on a ridge regression problem across various datasets and on an image classification task on the CIFAR10 dataset. The results demonstrate that the RPT method exhibits comparable or even superior performance in comparison to other methods.

1.3 Assumptions

In this section we lay out the assumptions that we are going to make throughout the whole work. In an attempt to make the analysis as general as possible, in this paper we focus on the problem of minimizing an arbitrary differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, where $\mathbb{R}^d$ is decomposed as $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}$ for some pre-specified $d_1, \ldots, d_B$.

We first recall the notion of matrix smoothness (Qu et al., 2016; Safaryan et al., 2021) which generalizes the standard $L$-smoothness assumption and allows one to encapsulate more curvature information of the objective function.
\textbf{Definition 1.} Let $f : \mathbb{R}^d \to \mathbb{R}$ and $L$ be a positive semidefinite matrix. We say that $f$ is $L$-smooth if
\[ f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{\|x - y\|^2}{2} \]
for all $x, y \in \mathbb{R}^d$, where $\|v\|_L \defeq \sqrt{v^\top LV}$.\

We make the following assumption throughout the whole paper.
\textbf{Assumption 1.} $f$ is $L$-smooth.

With Assumption 1 taken for granted, we analyse RPT in each of the three standard settings:
\textbf{Assumption 2.} $f$ is $\mu$-strongly convex, that is, there exists a positive constant $\mu$ such that
\[ f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu \|x - y\|^2}{2} \]
for all $x, y \in \mathbb{R}^d$.
\textbf{Assumption 3.} $f$ is convex and has a global minimizer $x^\ast$.
\textbf{Assumption 4.} $f$ is nonconvex and lower-bounded by $f^\ast \in \mathbb{R}$.

Under assumptions 2, 3 and 4 respectively we will be interesting in minimizing the following quantities:
\[ \mathbb{E} \left[ \|x^t - x^\ast\|^2 \right], \mathbb{E} \left[ f(x^t) - f(x^\ast) \right], \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right]. \]

\section{NEW ANALYSIS OF RANDOMIZED COORDINATE DESCENT}

\subsection{SkGD}
Randomized Coordinate Descent (RCD) is a celebrated technique of minimizing high-dimensional objective functions $f : \mathbb{R}^d \to \mathbb{R}$, where updating all gradient coordinates is prohibitively expensive. In its original form \cite{Nesterov2012, RichtarikTakac2014}, RCD used a biased randomized estimator of the gradient; in each iteration it chooses a random sub-block of coordinates and performs a gradient step with respect to that sub-block alone.

Recently, \cite{Safaryan2021} developed an unbiased version of RCD (called SkGD) which updates the iterates through unbiased sketches. Specifically, in each iteration $t$, SkGD starts from choosing a random set of coordinates $S^t \subset [d]$. Denote $p_j = \text{Prob}(j \in S^t)$ and assume $p_j > 0$. Then SkGD performs an update of form
\[ x^{t+1} = x^t - \gamma C^t \nabla f(x^t), \]
where
\[ C^t = \text{Diag}(c_1, \ldots, c_d), \quad c_j = \begin{cases} 1/p_j & \text{if } j \in S^t \\ 0 & \text{otherwise} \end{cases}. \]
It is then readily verifiable that $C^t \nabla f(x^t)$ is an unbiased gradient estimator, that is
\[ \mathbb{E} \left[ C^t \nabla f(x^t) \right] = \nabla f(x^t). \]

In the remainder of this paper we will always denote $P \defeq \text{Diag}(p_1, \ldots, p_d)$.

\subsection{Key quantity: $L_P$}
We now define the most important component of our analysis. In the rest of the paper $\lambda_{\text{max}}(M)$ will denote the largest eigenvalue of matrix $M$.
\textbf{Definition 2.} Let $C$ be an unbiased sketch operator and $P$ its associated probability matrix. Denote the smoothness matrix of the objective function $f$ by $L$. Then
\[ L_P \defeq \lambda_{\text{max}} \left( P^{-1/2}LP^{-1/2} \right). \]

It turns out that this quantity describes the convergence of RCD in all of the regimes in question (assumptions 2, 3, 4). Therefore, by estimating it for a particular choice of sampling distribution $P$ one obtains a unified analysis of RCD in the aforementioned regimes. Moreover, in the following lemma we provide sharp estimates for $L_P$.
\textbf{Lemma 1.} Let $L$ be any positive semidefinite matrix, $C$ an unbiased sketch operator, and $P$ its probability matrix. Then
\[ \lambda_{\text{max}}(\mathbb{E}[CLC]) \leq L_P \leq \lambda_{\text{max}}(L)\lambda_{\text{max}}(P^{-1}). \]

The first bound will be useful in proving the lemma on the convergence of RCD for nonconvex function. The second one quantifies how large can $L_P$ be in terms of the smoothness constant of $f$ and $P$. In practice, however, $L_P$ can be much smaller.

\subsection{Convergence results}
We begin with a preliminary lemma:
\textbf{Lemma 2.} Let $W = \text{Diag}(w_1, \ldots, w_d)$, where $w_1, \ldots, w_d$ are any positive entries and assume that $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth. Then
\[ \|\nabla f(x)\|_W \leq 2A(f(x) - f(x^\ast)), \]
where $A = \lambda_{\text{max}} \left( W^{1/2}LW^{1/2} \right)$.

The following estimate of a second moment of the gradient estimator is then an easy corollary of Lemma 2.
Lemma 3. Let $C : \mathbb{R}^d \to \mathbb{R}^d$ be an unbiased sketch operator. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and denote $g(x) = CVf(x)$. Then

$$E \left[ \|g(x)\|^2 \right] = \|\nabla f(x)\|_{P^{-1}} \leq 2L_P(f(x) - f^*) \cdot$$

The convergence result for $f$ satisfying Assumption 2 is now an immediate consequence of Corollary A.1 of (Gorbunov et al. 2020), which provides a unified analysis of SGD under strongly convex.

Theorem 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth and $\mu$-strongly convex. Let $C$ be an unbiased sketch with associated probability matrix $P$. Then the iterates $x^t$ of SkGD with stepsize $0 < \gamma \leq 1/L_P$ and update rule $x^{t+1} = x^t - \gamma CVf(x^t)$ satisfy

$$E \left[ \|x^t - x^*\|^2 \right] \leq (1 - \mu\gamma)^t \|x^0 - x^*\|^2 \cdot$$

Under assumptions 2 and 4, there also exist unified analyses of SGD (respectively Khaled et al. 2023 and Khaled and Richtárik 2023) where Lemma 2 is applicable. However, as we argue in Appendix D, applying these generic results to our setting leads to suboptimal convergence rates. Nevertheless, in Mishchenko et al. 2023, the analysis of the DIANA algorithm has $\mu = 1$ and treating the compressor as a sketch results in an asymptotic convergence rate for the non-convex setting that coincides with the rate proposed in our study. However, this analysis is based on the assumption of constant smoothness, in contrast to the matrix smoothness adopted in our study. We instead develop a novel analysis of SkGD in convex and non-convex regimes and present it as Theorems 2 and 3 respectively.

Theorem 2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth, convex, and lower-bounded. Furthermore, suppose that $f$ attains its minimum at $x^*$. Let $(x^t)_{t=0}^\infty$ be the iterates of the SkGD with unbiased sketch $C$ and stepsize $0 < \gamma \leq \frac{1}{2L_P}$.

Denote $\bar{x}^T = \frac{1}{T+1} \sum_{t=0}^T x^t$. Then

$$E \left[ f(\bar{x}^T) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2}{\gamma(T + 1)} \cdot$$

In particular, for $\gamma = 1/(2L_P)$ and $T + 1 \geq \frac{2L_P\|x^0 - x^*\|^2}{\epsilon}$, we have that

$$E \left[ f(\bar{x}^T) - f(x^*) \right] \leq \epsilon \cdot$$

3 RANDOMIZED PROGRESSIVE TRAINING

3.1 Problem formulation

For the sake of consistency with the PT framework, we shall decompose the domain of the objective function $f$ into $B$ disjoint subblocks. Thus, we will view the domain of $f$ as $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}$, where $\sum_{i=1}^B d_i = d$. We will further assume that $f$ is $L$-smooth, where $L$ is a symmetric, positive definite matrix of size $d \times d$. We let $L_i$ be the $i$th diagonal subblock of $L$ of size $d_i \times d_i$. We put $L_i \text{def} = \lambda_{\max}(L_i)$. We will be interested in minimizing this function under assumptions 2 or 4 i.e. finding

$$\text{argmin}_{x \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B}} f(x) \cdot$$

3.2 Randomized Progressive Training

We now introduce the $PT$-sketch operator, the key building block of the Randomized Progressive Training (RPT) algorithm.

Definition 3. Let $1 = p_1 \geq p_2 \geq \cdots \geq p_B > 0$. We additionally put $p_{B+1} = 0$. Let $\sigma = (\sigma_1, \ldots, \sigma_B)$ be any permutation of set $[B]$. For each $i \in [B]$ define the matrix $C_i \in \mathbb{R}^{d_i \times d}$ as follows:

$$C_i = \text{Diag} \left( (\xi_{i1}^j d_1, \ldots, \xi_{iB}^j d_B) \right),$$

where $\text{Diag}(v)$ denotes the diagonal matrix with $v$ as its diagonal, $(x)_m \text{def} = (x, \ldots, x) \in \mathbb{R}^m$, and

$$\xi_{ij}^j = \begin{cases} 
1/p_j & \text{if } j \leq i \\
0 & \text{otherwise.}
\end{cases}$$

We finally define a random matrix $C$ which is equal to $C_i$ with probability $p_i - p_{i+1}$ and call it a $PT$-sketch.
With this definition, we are now ready to formulate the RPT algorithm.

Algorithm 2 Randomized Progressive Training

1: Input: initial iterates \( x^0 \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B} \); step-size \( \gamma \); integer \( T \geq 0 \)
2: for \( t = 1, \ldots, T \) do
3: Draw the PT-sketch \( P^t \)
4: \( x^{t+1} = x^t - \gamma P^t \nabla f(x^t) \)
5: end for
6: Output: \( x^T \)

RPT is an instance of SkGD with probability matrix

\[
P = \text{Diag} \left( (p_{\sigma_1})_{d_1}, \ldots, (p_{\sigma_B})_{d_B} \right).
\]

Hence, theorems [1] [2] [3] immediately apply to it. Our task is now to choose \( P \) in a way that minimizes the total computation complexity.

Remark 1. There are two main differences between RPT and classical approaches to progressive training (Karras et al., 2018; Wang et al., 2022). Firstly, our method chooses the size of coordinate updates in a randomized fashion. PT, however, is a deterministic method which grows the coordinate updates according to a pre-set schedule.

Moreover, PT assumes a particular order of importance of the coordinate blocks – it always updates the first coordinate block, then includes the second block, and so on. However, we note that in general there is no particular reason to prioritize the first coordinate block, and RPT allows us to decide on the order of importance by choosing the permutation \( \sigma \) in Definition [3].

Remark 2. The formulation of RPT is stemming from the problem formulation described in Section [3.1] which assumes a particular division of the domain elements into \( B \) coordinate blocks. While our analysis holds for any such division, we do not specify how to decide on it, noting that it is often problem-specific (for instance, a coordinate block might be a layer of a neural network (Karras et al., 2018; Wang et al., 2022)).

4 ANALYSIS OF RPT

4.1 Cost model

In this section we will define the computation cost of optimizing the objective function from Section 3.1. By construction, in the RPT procedure the variable \( x^t \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_B} \) is divided into \( B \) coordinate blocks and updated with respect to a certain subset of these blocks. In our discussion we will assume that by carrying out an update with respect to block \( i \) the algorithm suffers a penalty \( c_i \). Hence, if \( C^t = C_i \), then cost \( t \), the cost incurred in the iteration \( t \) is equal to \( \sum_{j=1}^{i} c_j \). Moreover, without loss of generality we make a simplifying normalizing assumption on the costs \( c_i \) for the sake of presentation.

Assumption 5. \( \sum_{i=1}^{B} c_i = 1 \).

Now, recall that \( C = C_i \) with probability \( p_i - p_{i+1} \). Therefore, the expected cost of the entire optimization procedure is

\[
\mathbb{E} \left[ \text{cost}(T) \right] = \mathbb{E} \left[ \sum_{t=1}^{T} \text{cost}_t \right] = T \sum_{i=1}^{B} (p_i - p_{i+1}) \sum_{j=1}^{c_i} c_j = T \sum_{i=1}^{B} p_i c_i.
\]

Previous convergence results show that in order to reach the \( \varepsilon \)-accurate solution in all regimes described by assumptions [2] [3] and [4] one needs \( T \) to be proportional to \( L_P \). Thus,

\[
\mathbb{E} \left[ \text{cost}(T) \right] \propto c(P) = L_P \sum_{i=1}^{B} p_i c_i,
\]

regardless of the regime in question. Thus, choosing the optimal RPT sketch is equivalent to solving the minimization problem

\[
\min_{\sigma} \min_{1 = p_1 \geq \cdots \geq p_B > 0} \sum_{i=1}^{B} L_P \sum_{i=1}^{B} p_i c_i = \left( \begin{array}{c}
\min_{\sigma} \\
\min_{1 \geq p_1 \geq \cdots \geq p_B> 0}
\end{array} \right) \sum_{i=1}^{B} L_P \sum_{i=1}^{B} p_i c_i.
\]

4.2 Choosing the probabilities \( p_i \)

Finding the global minimum of the quantity \( c(P) \) in general is a challenging task when \( L \) is simply any positive definite matrix; \( c(P) \) is a non-convex function that can intricately depend on the structure of matrix \( L \). However, note that taking \( p_i = 1 \) for all \( i \in \{B\} \) recovers gradient descent, i.e. the optimal \( c(P) \) is at most \( L \). Thus, the solution of Problem (1) is not worse than gradient descent. Furthermore, one can find a closed-form minimizer for the tight upper bound for \( L_P \). We first invoke Lemma 1 from Nesterov (2012).

Lemma 4. Consider an arbitrary positive semidefinite matrix \( M \in \mathbb{R}^{d \times d} \) with diagonal subblocks \( M_1, \ldots, M_B \), where each \( M_i \in \mathbb{R}^{d_i \times d_i} \). Put \( M_i = \lambda_{\max}(M_i) \) and \( M = \lambda_{\max}(M) \). Then

\[
\max_{i} M_i \leq M \leq \sum_{i=1}^{B} M_i.
\]
In Appendix C we argue that these bounds are tight. We then have
\[ c(P) \leq \sum_{i=1}^{B} \frac{L_{\sigma_i}}{p_i} \cdot \sum_{i=1}^{B} p_i c_{\sigma_i}. \]

The below lemma describes the global minimizer of this upper bound.

**Lemma 5.**
\[ \sum_{i=1}^{B} \frac{L_{\sigma_i}}{p_i} \cdot \sum_{i=1}^{B} p_i c_{\sigma_i} \geq \left( \sum_{i=1}^{B} \sqrt{L_i c_i} \right)^2. \]

Moreover, this bound is met by taking \(\sigma\) to be such that
\[ \sqrt{\frac{L_{\sigma_1}}{c_{\sigma_1}}} \geq \cdots \geq \sqrt{\frac{L_{\sigma_B}}{c_{\sigma_B}}}. \]

and
\[ p_i = \frac{\sqrt{L_{\sigma_i}/c_{\sigma_i}}} {\max_j \sqrt{L_j/c_j}}. \]

This choice of probabilities is similar to the one proposed by Zhu et al. (2016) (which, contrary to our paper, treats accelerated version of RCD) where the probabilities \(p_i\) are proportional to \(\sqrt{L_i}\). However, Zhu et al. (2016) consider the setup of non-overlapping blocks and analysis iteration complexity only. On the contrary, our work, to the best of our knowledge, is the first one to consider such choice of probabilities in a non-accelerated setting and motivates this choice through the total complexity analysis.

We summarize the discussion with the corollary:

**Corollary 1.** Let \(P\) be the optimal solution of Equation (1). Then,
\[ c(P) \leq \min \left\{ L_i \left( \sum_{i=1}^{B} \sqrt{L_i c_i} \right)^2 \right\}. \]

## 5 EXPERIMENTS

We conduct several experiments to evaluate the performance of Randomized Progressive Training in comparison to other methods on various target functions and datasets.

### 5.1 Quadratic functions on synthetic data

In the first experiment, we show how fast Randomized Progressive Training can be compared to vanilla Gradient Descent. When running RPT with probabilities proposed in Lemma 5 theory suggests that total computation cost of
\[ O \left( \frac{\left( \sum_{i=1}^{B} \sqrt{L_i c_i} \right)^2}{\varepsilon} \right) \]

is required to achieve \(\varepsilon\)-accuracy. At the same time, standard theory for GD under Assumption 5 (meaning that a cost of one GD iteration is one) suggests \(O \left( \frac{\varepsilon}{L} \right)\) total cumulative cost. We vary block smoothness constants \(L_i\) and computation costs \(c_i\) so that the ratio \(\left( \sum_{i=1}^{B} \sqrt{L_i c_i} \right)^2/L\) is relatively large. In this vein, we generate nine quadratic functions \(f(x) = \frac{1}{2} x^T A x\) with varying dimensions and properties of the diagonal matrix \(A\). The number of blocks is three in all experiments. We consider the computation cost \(c_i\) of block \(i\) to be proportional to its size. As shown on Figure 1 when the size of the block with the smallest smoothness constant is much larger than one of the largest, RPT shows significant speedup over GD. Table 1 compares speedups predicted by theory with actual ones.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Theory speedup</th>
<th>Actual speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>dims</td>
<td>10 10 10</td>
<td>1.1 1.9</td>
</tr>
<tr>
<td>Ls</td>
<td>272 53.3 11</td>
<td>3.4 19.8</td>
</tr>
<tr>
<td>dims</td>
<td>10 50 250</td>
<td>5.7 26.9</td>
</tr>
<tr>
<td>Ls</td>
<td>270.5 55 11</td>
<td>12.7 36.6</td>
</tr>
<tr>
<td>dims</td>
<td>10 100 1000</td>
<td>1.5 2.2</td>
</tr>
<tr>
<td>Ls</td>
<td>256.7 54.8 11</td>
<td>6.4 17.6</td>
</tr>
<tr>
<td>dims</td>
<td>10 50 2500</td>
<td>12.3 64.9</td>
</tr>
<tr>
<td>Ls</td>
<td>274.8 55 11</td>
<td>1093.9 109.7 11</td>
</tr>
<tr>
<td>dims</td>
<td>10 100 1000</td>
<td>36.6 128.1</td>
</tr>
<tr>
<td>Ls</td>
<td>1080.4 109.6 11</td>
<td>1066.7 110 11</td>
</tr>
<tr>
<td>dims</td>
<td>10 100 1000</td>
<td>2.2 2.8</td>
</tr>
<tr>
<td>Ls</td>
<td>27414 547 11</td>
<td>27166 550 11</td>
</tr>
<tr>
<td>dims</td>
<td>10 50 2500</td>
<td>15.4 20</td>
</tr>
<tr>
<td>Ls</td>
<td>26020 549 11</td>
<td>10 100 1000</td>
</tr>
<tr>
<td>dims</td>
<td>10 500 25000</td>
<td>40.1 62.4</td>
</tr>
<tr>
<td>Ls</td>
<td>27363 550 11</td>
<td>10 100 1000</td>
</tr>
</tbody>
</table>

Table 1: Comparison of theoretical and actual speedup of Randomized Progressive Training versus Gradient Descent with quadratic functions. In the "Setup" section, "dims" refers to the dimension of a block and "Ls" denotes the corresponding smoothness constant.
5.2 Ridge regression on real datasets

In our second set of experiments, we assess the convergence of Randomized Progressive Training, Gradient Descent, and Cyclic Block Coordinate Descent (CBCD) (Sun and Hong, 2015) on the ridge regression problem

$$\min f(x) = \frac{1}{n} \sum_{i=1}^{n} (a_i, x) - b_i)^2 + \lambda \|x\|_2^2,$$

where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ and $\lambda$ is a regularization parameter. We consider four regression datasets: California housing, Los Angeles ozone (Hastie et al., 2009), prostate cancer, and white wine quality (Cortez et al., 2009). Each dataset feature is normalized, and a bias vector is added. Then, heuristically, features are permuted so that the diagonal of the Hessian matrix is sorted in descending order. As in the previous experiment, we consider $c_i$ proportional to the corresponding block size.

As can be seen from Figure 2, although the theoretical speedup of RPT over GD, according to Table 2, is close to 1, the actual speedup is still noteworthy. However, both algorithms are slower than Cyclic Block Coordinate Descent.

5.3 Image classification on CIFAR10

In our final computational experiment, we trained a Deep Residual Neural Network (He et al., 2016) with 18 layers (ResNet18) for the task of image classification on the CIFAR10 dataset (Krizhevsky, 2009).

Table 2: Comparison of theoretical and actual speedup of Randomized Progressive Training versus Gradient Descent for ridge regression problems.
Implementation. The CIFAR10 dataset comprises 50’000 training samples and 10’000 test samples, which are distributed evenly among 10 classes. The samples are RGB images of size $32 \times 32$. To facilitate hyperparameter tuning, we further divided the training dataset into training and validation datasets with a ratio $9 : 1$. During the training process, we applied a padding of 4, randomly cropped to size $32 \times 32$, and randomly flipped images horizontally. We trained the neural network with three algorithms: Stochastic Gradient Descent, Progressive Training ([Wang et al., 2022]), and our own method. The batch size remains fixed at 128 for each method. We implemented all settings using the Python packages JAX ([Bradbury et al., 2018]), Flax ([Heek et al., 2020]), and Optax.

Computation cost. The computation cost associated with a convolutional layer in ResNet18 is estimated by calculating the ratio of the total number of elements in the kernel to the total number of parameters present in ResNet18. Similarly, the computation cost attributed to the fully connected layer is determined by dividing the total number of weights within that layer by the overall number of parameters. It is noted for simplicity that the computation cost associated with the batch normalization layer is disregarded and considered as zero. These computations reflect the costs incurred for processing the entire batch.

Fine-tuning. Fine-tuning is a crucial step in the process of training neural networks, as the lack of smoothness and convexity properties inherent in these models often renders the application of theoretical results ineffective ([Szegedy et al., 2014]). To address this, a common approach for hyperparameter tuning is to utilize a grid search method. In our case, for the Randomized Progressive Training method, we fine-tuned the learning rate and sampling probabilities. A grid search was conducted for the learning rate over the values $[1, 0.5, 0.1, 0.05, 0.01, 0.001]$, and for sampling probabilities, we evaluated two different strategies. The first strategy used exponentially decaying probabilities with rates $[0.999, 0.995, 0.99, 0.95, 0.9, 0.85]$, e.g., for a rate of 0.9, the probabilities were calculated as $(p_1, \ldots, p_{22}) = (1, 0.9, 0.9^2, \ldots, 0.9^{21})$. The second strategy used evenly spaced probabilities between $p_1 = 1$ and $p_{22}$, where $p_{22}$ ran over the list $[0.7, 0.6, 0.5, 0.4, 0.3, 0.2]$. The best results were obtained with a learning rate of 1 and exponentially decaying sampling probabilities with a rate of 0.9. For Progressive Training (PT), we only fine-tuned the learning rate and utilized the round sizes $T_1, \ldots, T_{22}$ suggested in ([Wang et al., 2022]) as the best strategy. Specifically, we set $T_1 = \cdots = T_{21} = \frac{T}{44}$, and $T_{22} = \frac{23T}{44}$, where $T$ is the total number of steps. The grid search for the learning rate was the same as for RPT. Additionally, Stochastic Gradient Descent parameters such as the learning rate, weight decay, and momentum were set according to the guidelines in ([Moreau et al., 2022]). The batch size in all three settings was 128.

Results. The results of our image classification task are summarized in Figure 3. While it may appear that RPT converges more slowly and has a lower test accuracy than SGD, based on Figures 3c and 3d it is important to note that this comparison is not entirely fair as one RPT step can take significantly less time than one SGD step. To more accurately compare the performance of RPT to its competitors, we have evaluated the performance versus computation cost, as shown in Figures 3a and 3b. From these figures, it is clear that RPT outperforms both competitors.

Remark 3. In our practical deployment of Algorithms 1 and 2, at every iteration, we initially compute the entire gradient, retaining only those elements as dictated by the algorithm. This procedure accurately emulates algorithms but proves to be impractical for real-world application due to the necessity of comput-
Figure 3: Comparison of Randomized Progressive Training (RPT), Stochastic Gradient Descent (SGD), and Progressive Training (PT) on an image classification task using a ResNet18 model. The block computation cost $c_i$ is proportional to the number of elements in the kernel for convolution layers, the number of weights in the fully connected layer, and is disregarded (assigned a value of zero) for the batch normalization layer. Normalization is implemented in such a manner that the cumulative computation cost across all blocks for a single batch equals 1. Consequently, the computation cost for one iteration of Stochastic Gradient Descent over a batch is also one.

ing the full gradient, effectively nullifying any theoretical advantages. Consequently, comparing the efficiency of Randomized Progressive Training against a robustly implemented Stochastic Gradient Descent, as included in the standard Optax package, particularly regarding time efficiency, proves to be an inequitable evaluation.

6 Limitations

The primary constraint of this study resides in the computational challenges associated with the determination of the constants denoted as $L_i$ within a real-world context, which forms the foundation for the theoretical derivations presented. Furthermore, a critical assumption underpinning this study pertains to the smoothness of the objective function, a premise that may not inherently hold true for neural networks.
References


Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]
A ISSUES WITH THE THEORY FOR PROGFED

We will first briefly outline the notation that Wang et al. (2022) use in their paper.

We denote by \( \mathcal{M} \) the neural network that we aim to train. Put \( \mathcal{M}^s \) to be network \( \mathcal{M} \) truncated to its first \( s \) layers. Let \( \mathcal{L} \) denote loss function of interest and put \( f \) and \( f^s \) to be \( \mathcal{L} \circ \mathcal{M} \) and \( \mathcal{L} \circ \mathcal{M}^s \) respectively. Denote \( x^s \) to be the parameter vector of \( \mathcal{M}^s \). Finally, for \( s \leq i \leq S \) we put \( x_i |_{E_s} \) and \( \nabla f^i(x^i) |_{E_s} \) to be the projections of vectors \( x^i \) and \( \nabla f^i(x^i) \) onto the parameter space of \( \mathcal{M}^s \).

The first result by Wang et al. (2022) proves that

\[
\frac{1}{T} \sum_{t=0}^{T-1} \alpha_t^2 \left\| \nabla f^s(x_t^i) |_{E_s} \right\| < \varepsilon
\]

after \( \mathcal{O} \left( \frac{q^2}{\varepsilon} + \frac{1}{\varepsilon} \right) \cdot LF_0 \) iterations. Here \( x_t^i \) is the \( t \)th iterate of ProgFed,

\[
\alpha_t = \min \left\{ 1, \frac{\langle \nabla f(x_t), \nabla f^s(x_t) |_{E_s} \rangle}{\| \nabla f^s(x_t) \|_2^2} \right\},
\]

\( L \) is the smoothness constant of \( f \), and \( F_0 \) is the function suboptimality. Thus, the numbers \( \alpha_t \) depend on the iterates. No bounds on \( \alpha_t \) are provided. In particular, the theory does not exclude the possibility that \( \alpha_t^2 \) are extremely close to 0. This renders the result meaningless.

The second result that Wang et al. (2022) attempts to bound \( \frac{1}{T} \sum_{t=0}^{T-1} \| \nabla f(x_t) \|_2^2 \). Unfortunately, it suffers from a similar issue. The authors claim that

\[
\mathcal{O} \left( \frac{q^4 \sigma^2}{\varepsilon^2} + \frac{q^2}{\varepsilon} \right) \cdot LF_0,
\]

where \( q := \max_{t \in [T]} \left( q_t := \frac{\| \nabla f(x_t) \|}{\alpha_t^2 \| \nabla f^s(x_t) \|_2} \right) \). Wang et al. (2022) do not provide any bound on \( q \), which does not depend on the iterates. Thus, it is not possible to decide \textit{a priori} how many iterations do we need in order to reach the \( \varepsilon \)-accurate solution, because the theory is dependent on the iterates \textit{produced during the run of the method}. In other words, their second convergence result (in its current state) does not exclude an infinite runtime, so in fact it does not show the convergence at all.

B PROOFS

B.1 Proof of Lemma 1

\textbf{Lemma 1.} Let \( \mathbf{L} \) be any positive semidefinite matrix, \( \mathbf{C} \) - any unbiased sketch operator, and \( \mathbf{P} \) - its probability matrix. Then

\[
\lambda_{\max}(\mathbb{E}[\mathbf{CLC}]) \leq LP \leq \lambda_{\max}(\mathbf{L})\lambda_{\max}(\mathbf{P}^{-1}).
\]

\textit{Proof.} We start from proving the first inequality. Let \( x \in \mathbb{R}^d \). Then

\[
x^\top \mathbf{P}^{1/2} \mathbb{E}[\mathbf{CLC}] \mathbf{P}^{1/2} x = \mathbb{E} \left[ \left\| \mathbf{L}^{1/2} \mathbf{C} \mathbf{P}^{1/2} x \right\|^2 \right] \leq \lambda_{\max}(\mathbf{L}) \mathbb{E} \left[ \left\| \mathbf{C} \mathbf{P}^{1/2} x \right\|^2 \right] = \lambda_{\max}(\mathbf{L}) \mathbb{E} \left[ \sum_{i=1}^{n} \frac{1}{p_i} x_i^2 \chi(C=C_i) \right] = \lambda_{\max}(\mathbf{L}) \sum_{i=1}^{n} \frac{x_i^2}{p_i}
\]

\[
= \lambda_{\max}(\mathbf{L}) \left\| x \right\|^2.
\]
Thus,
\[ \lambda_{\text{max}}(P^{1/2}E[CLC]P^{1/2}) \leq \lambda_{\text{max}}(L), \]
i.e.
\[ \lambda_{\text{max}}(E[CLC]) \leq \lambda_{\text{max}}(P^{-1/2}LP^{-1/2}). \]

For the second inequality, similarly,
\[ x^T P^{-1/2}LP^{-1/2}x = \|L^{1/2}P^{-1/2}x\|^2 \leq \lambda_{\text{max}}(L) \|P^{-1/2}x\|^2 \leq \lambda_{\text{max}}(L)\lambda_{\text{max}}(P^{-1}) \|x\|^2. \]

Thus,
\[ \lambda_{\text{max}}(P^{-1/2}LP^{-1/2}) \leq \lambda_{\text{max}}(L)\lambda_{\text{max}}(P^{-1}). \]

\[ \square \]

B.2 Proof of Lemma 2

Lemma 2. Let \( W = \text{Diag}(w_1, \ldots, w_d) \), where \( w_1, \ldots, w_d \) are any positive entries and assume that \( f : \mathbb{R}^d \to \mathbb{R} \) is \( L \)-smooth. Then
\[ \|\nabla f(x)\|^2_W \leq 2A(f(x) - f(x^*)), \]
where \( A = \lambda_{\text{max}}(W^{1/2}LW^{1/2}) \).

Proof. Fix \( \alpha \in \mathbb{R} \) and \( x \in \mathbb{R}^d \). Since \( x^* \) is a minimizer of \( f \) and since \( f \) is \( L \)-smooth, we have
\[ f(x^*) \leq f(x) - \alpha W\nabla f(x) \]
\[ \leq f(x) - \alpha \langle \nabla f(x), W\nabla f(x) \rangle + \frac{\alpha^2}{2} \|W\nabla f(x)\|^2_L \]
\[ = f(x) - \alpha \langle \nabla f(x), W\nabla f(x) \rangle + \frac{\alpha^2}{2} \|W\nabla f(x)\|^2 \]
By definition of \( A \), we have \( W^{1/2}LW^{1/2} \leq AI_d \), which implies that \( WLW \leq AW \). Plugging this estimate into the above inequality, we get
\[ f(x^*) \leq f(x) - \alpha \|\nabla f(x)\|^2_W + \frac{\alpha^2A}{2} \|\nabla f(x)\|^2 \]
\[ = f(x) - \left( \alpha - \frac{\alpha^2A}{2} \right) \|\nabla f(x)\|^2 \]
Since the above inequality holds for any \( \alpha \in \mathbb{R} \), it also holds for \( \alpha^* = \frac{1}{A} \); this \( \alpha \) maximizes the expression \( \alpha - \frac{\alpha^2A}{2} \).
It only remains to plug \( \alpha^* \) into the above expression and rearrange the inequality.

B.3 Proof of Lemma 3

Lemma 3. Let \( C : \mathbb{R}^d \to \mathbb{R}^d \) be an unbiased sketch operator. Let \( f : \mathbb{R}^d \to \mathbb{R} \) be differentiable and denote \( g(x) = C\nabla f(x) \). Then
\[ E[\|g(x)\|^2] \leq \|\nabla f(x)\|^2_{P^{-1}} \leq 2L_P(f(x) - f^*). \]

Proof.
\[ E[\|g(x)\|^2] = E[\|C\nabla f(x)\|^2] \]
\[ = E[\nabla f(x)^T C^2 \nabla f(x)] \]
\[ = \nabla f(x)^T E[C^2] \nabla f(x) \]
\[ = \nabla f(x)^T P^{-1} \nabla f(x) \]
\[ = \|\nabla f(x)\|^2_{P^{-1}}. \]
The last inequality follows directly from Lemma 2.

B.4 Proof of Theorem 2

**Theorem 2.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{L}$-smooth, convex, and lower-bounded. Furthermore, suppose that $f$ attain its minimum at $x^\ast$. Let $(x^t)_{t=0}^\infty$ be the iterates of the SkGD with unbiased sketch $C$ and stepsize $0 < \gamma \leq \frac{1}{2\mathcal{L}_P}$.

Denote $\bar{x}^T = \frac{1}{T+1} \sum_{t=0}^T x^t$. Then

$$E \left[ f(\bar{x}^T) - f(x^\ast) \right] \leq \frac{\|x^0 - x^\ast\|^2}{\gamma(T+1)}.$$ 

In particular, for $\gamma = 1/(2\mathcal{L}_P)$ and $T+1 \geq \frac{2\mathcal{L}_P\|x^0 - x^\ast\|^2}{\epsilon}$, we have that

$$E \left[ f(\bar{x}^T) - f(x^\ast) \right] \leq \epsilon.$$ 

**Proof.** Taking conditional expectations and expanding the brackets yields

$$E \left[ \|x^{t+1} - x^\ast\|^2 | x^t \right] = \|x^t - x^\ast\|^2 - 2\gamma E \left[ \langle \nabla f(x^t), x^t - x^\ast \rangle | x^t \right] + \gamma^2 E \left[ \|\nabla f(x^t)\|^2 | x^t \right]$$

$$\leq \|x^t - x^\ast\|^2 - 2\gamma \langle \nabla f(x^t), x^t - x^\ast \rangle + 2\gamma^2 \mathcal{L}_P \|f(x^t) - f(x^\ast)\|$$

$$\leq \|x^t - x^\ast\|^2 - 2\gamma \langle f(x^t) - f(x^\ast) \rangle + \gamma \langle f(x^t) - f(x^\ast) \rangle$$

$$= \|x^t - x^\ast\|^2 - \gamma \langle f(x^t) - f(x^\ast) \rangle,$$

where the first inequality follows from Lemma 3 and the second one from convexity of $f$ and $0 < \gamma \leq 1/(2\mathcal{L}_P)$. Rearranging the above inequality, taking the expectation of both sides and using the tower property gives

$$\gamma E \left[ f(x^t) - f(x^\ast) \right] \leq E \left[ \|x^t - x^\ast\|^2 - \|x^{t+1} - x^\ast\|^2 \right].$$

Thus, by convexity of $f$, we finally arrive at

$$E \left[ f(\bar{x}^T) - f(x^\ast) \right] \leq \frac{1}{\gamma(T+1)} \sum_{t=0}^T \langle f(x^t) - f(x^\ast) \rangle$$

$$\leq \frac{1}{\gamma(T+1)} \left( \|x^0 - x^\ast\|^2 - E \left[ \|x^{T+1} - x^\ast\|^2 \right] \right)$$

$$\leq \frac{\|x^0 - x^\ast\|^2}{\gamma(T+1)}.$$

B.5 Proof of Theorem 3

**Theorem 3.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{L}$-smooth and lower bounded by $f^\ast$. Let $(x^t)_{t=0}^\infty$ be the iterates of SkGD with unbiased sketch $C$ and stepsize $0 < \gamma \leq \frac{1}{\mathcal{L}_P}$.

Denote $\bar{x}^T = \text{argmin}_{t=0,\ldots,T} \|\nabla f(x^t)\|^2$. Then

$$E \left[ \|\nabla f(\bar{x}^T)\|^2 \right] \leq \frac{2\delta^0}{\gamma(T+1)},$$

where $\delta^0 = f(x^0) - f^\ast$. In particular, for $\gamma = 1/\mathcal{L}_P$ and $T+1 \geq \frac{2\mathcal{L}_P\delta^0}{\epsilon}$, we have that

$$E \left[ \|\nabla f(\bar{x}^T)\|^2 \right] \leq \epsilon.$$
Before carrying out the proof of Theorem 3 we note that in this theorem, one can replace $L_P$ by

$$L_M \overset{\text{def}}{=} \lambda_{\max} (E[CLC])$$

which is at most $L_P$ by Lemma 1. However, the constant $L_M$ is difficult to interpret qualitatively. For the sake of unified presentation of the RPT analysis in all regimes in question, we therefore chose to replace it by $L_P$.

**Proof.** By $L$-smoothness we have

$$f(x^{t+1}) \leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{\|x^{t+1} - x^t\|^2}{2L}$$

$$= f(x^t) - \gamma \langle \nabla f(x^t), C\nabla f(x^t) \rangle + \frac{\gamma^2}{2} \|C\nabla f(x^t)\|^2$$

$$= f(x^t) - \gamma \langle \nabla f(x^t), C\nabla f(x^t) \rangle + \frac{\gamma^2}{2} \langle \nabla f(x^t), CCLC\nabla f(x^t) \rangle.$$  

Thus, taking expectation conditional on $x^t$ yields

$$E[f(x^{t+1})|x^t] \leq f(x^t) - \gamma \|\nabla f(x^t)\|^2 + \frac{\gamma^2}{2} \|\nabla f(x^t)\|^2 E[CLC] \nabla f(x^t)$$

$$\leq f(x^t) - \gamma \|\nabla f(x^t)\|^2 + \frac{\gamma^2}{2} L_M \|\nabla f(x^t)\|^2$$

$$\leq f(x^t) - \gamma \|\nabla f(x^t)\|^2 + \frac{\gamma^2}{2} L_P \|\nabla f(x^t)\|^2,$$

i.e.

$$\left( \gamma - \frac{\gamma^2 L_P}{2} \right) \|\nabla f(x^t)\|^2 \leq f(x^t) - E[f(x^{t+1})|x^t].$$

By taking expectations of both sides, using tower property and taking the average over $t$, we get

$$\frac{\gamma}{2} \min_{t=0,...,T} \|\nabla f(x^t)\|^2 \leq \frac{\gamma}{2T+1} \sum_{t=0}^T \|\nabla f(x^t)\|^2$$

$$\leq \left( \frac{\gamma}{2} + \frac{\gamma}{2} (1 - \gamma L_P) \right) \frac{1}{T+1} \sum_{t=0}^T \|\nabla f(x^t)\|^2$$

$$= \left( \gamma - \frac{\gamma^2 L_P}{2} \right) \frac{1}{T+1} \sum_{t=0}^T \|\nabla f(x^t)\|^2$$

$$\leq \frac{1}{T+1} E[f(x^0) - f(x^{T+1})]$$

$$\leq \frac{1}{T+1} \delta^0.$$  

Rearranging gives

$$E[\|\nabla f(x^T)\|^2] \leq \frac{2\delta^0}{\gamma(T+1)}.$$

**B.6 Proof of Lemma 4**

**Lemma 4.** Consider an arbitrary positive semidefinite matrix $M \in \mathbb{R}^{d \times d}$ with diagonal subblocks $M_1, \ldots, M_B$, where each $M_i \in \mathbb{R}^{d_i \times d_i}$. Put $M_i = \lambda_{\max}(M_i)$ and $M = \lambda_{\max}(M)$. Then

$$\max_{i=1}^B M_i \leq M \leq \sum_{i=1}^B M_i.$$
Proof. The second inequality is just Lemma 1 from [Nesterov (2012)]. Let us now prove the first inequality. Without loss of generality we may assume that \( \max_i L_i = L_1 \). Denote \( e_i \) to be the \( i \)th standard unit vector of \( \mathbb{R}^d \). Let \( S = \text{Span}(e_1, \ldots, e_{d_1}) \). Now,

\[
L_1 = \sup_{x \in \mathbb{R}^{d_1}, \|x\| = 1} x^\top L_1 x
= \sup_{x \in S, \|x\| = 1} x^\top L x
\leq \sup_{x \in \mathbb{R}^d, \|x\| = 1} x^\top L x.
\]

This concludes the proof.

B.7 Proof of Lemma 5

Lemma 5.

\[
\sum_{i=1}^B \frac{L_{\sigma_i}}{p_i} \cdot \sum_{i=1}^B p_i c_{\sigma_i} \geq \left( \sum_{i=1}^B \sqrt{L_i c_i} \right)^2.
\]

Moreover, this bound is met by taking \( \sigma \) to be such that

\[
\sqrt{\frac{L_{\sigma_1}}{c_{\sigma_1}}} \geq \cdots \geq \sqrt{\frac{L_{\sigma_B}}{c_{\sigma_B}}},
\]

and

\[
p_i = \frac{\sqrt{L_{\sigma_i}/c_{\sigma_i}}}{\max_j \sqrt{L_j/c_j}}.
\]

Proof. By Cauchy-Schwarz inequality,

\[
\sum_{i=1}^B \frac{L_{\sigma_i}}{p_i} \cdot \sum_{i=1}^B p_i c_{\sigma_i} \geq \left( \sum_{i=1}^B \sqrt{L_{\sigma_i}/p_i} \cdot \sqrt{p_i c_{\sigma_i}} \right)^2
= \left( \sum_{i=1}^B \sqrt{L_i c_i} \right)^2.
\]

Since we are free to choose permutation \( \sigma \), we may clearly choose it in such a way that

\[
\sqrt{\frac{L_{\sigma_1}}{c_{\sigma_1}}} \geq \cdots \geq \sqrt{\frac{L_{\sigma_B}}{c_{\sigma_B}}}.
\]

Plugging in this choice of permutation \( \sigma \) and probablities \( p_1, \ldots, p_B \) verifies that the lower bound from Lemma 5 is attainable.
C DISCUSSION ABOUT TIGHTNESS OF LEMMA 4

Here we consider the 2-block case argue that the maximum eigenvalue \( \lambda_{\text{max}}(M) \) can be arbitrarily close to \( \lambda_{\text{max}}(M_1) + \lambda_{\text{max}}(M_2) \). Extension of the below result to an arbitrary number of blocks is very straightforward.

**Proposition 1.** Let \( M_1 \in S_{++}^{d_1}, M_2 \in S_{++}^{d_2} \), and \( C = \sqrt{\lambda_{\text{max}}(M_1)(\lambda_{\text{max}}(M_2) - \varepsilon)} \cdot v u^\top \), where \( \varepsilon \in (0, \lambda_{\text{max}}(M_2)) \), \( u \in \mathbb{R}^{d_1} \) and \( v \in \mathbb{R}^{d_2} \) are unit eigenvectors corresponding to maximum eigenvalues of matrices \( M_1 \) and \( M_2 \), respectively. Then the matrix \( M = \begin{bmatrix} M_1 & C \\ C & M_2 \end{bmatrix} \) is positive definite, and it holds that

\[
\lambda_{\text{max}}(M) \geq \lambda_{\text{max}}(M_1) + \lambda_{\text{max}}(M_2) - \varepsilon.
\]

**Proof.** First, let us check the matrix \( M = \begin{bmatrix} M_1 & C \\ C & M_2 \end{bmatrix} \) is positive definite. Since \( M_1 \) is positive definite, it suffices to check that \( M_2 - CM_1^{-1}C^\top \) is positive definite. The second term can be rewritten as

\[
CM_1^{-1}C^\top = \lambda_{\text{max}}(M_1)(\lambda_{\text{max}}(M_2) - \varepsilon)u^\top u M_1^{-1}u v^\top (\lambda_{\text{max}}(M_2) - \varepsilon)v v^\top.
\]

Now, through eigenvalue decomposition of \( M_2 \), we see that eigenvalues of \( M_2 - CM_1^{-1}C^\top \) are the same as eigenvalues of \( M_2 \) except for the maximum one, which is replaced by \( \varepsilon > 0 \). Thus, \( M_2 - CM_1^{-1}C^\top \) is positive definite.

Let \( M_1 = A_1 \Sigma_1 A_1^\top \) and \( M_2 = A_2 \Sigma_2 A_2^\top \) be eigenvalue decompositions of the matrices, and without the loss of generality assume that the eigenvalues are sorted in a descending order. We also set \( a = \sqrt{\lambda_{\text{max}}(M_1)(\lambda_{\text{max}}(M_2) - \varepsilon)} \). Then

\[
M' \overset{\text{def}}{=} \text{Diag}(A_1, A_2)^\top M \text{Diag}(A_1, A_2) = \begin{bmatrix} \Sigma_A & \text{Diag}(a, 0, \ldots, 0) \\ \text{Diag}(a, 0, \ldots, 0) & \Sigma_2 \end{bmatrix}
\]

has the same set of eigenvalues as \( M \). We now have

\[
\lambda_{\text{max}}(M') = \lambda_{\text{max}}(M_1) + \lambda_{\text{max}}(M_2) - \varepsilon.
\]

It remains to note that \( \lambda_{\text{max}}(M') = \lambda_{\text{max}}(M) \), since the sets of eigenvalues coincide. \( \square \)

D DISCUSSION ABOUT THE UNIFIED ANALYSIS IN NONCONVEX AND CONVEX SETTINGS

D.1 Nonconvex setting

Recent theoretical breakthrough of Khaled and Richtárik (2023) offers generic result on the convergence of SGD under very mild assumptions. However, in this section we argue that the rates promised by this result is suboptimal for the setting of our paper, which is why we derived a separate analysis of RCD in this regime.
We manage to obtain faster rate by exploiting Lemma 4.

for all \( x \in \mathbb{R}^d \).

Theorem 4. Suppose that \( f \) is lower-bounded and \( \mathcal{L} \)-smooth. Let \( g \) be an unbiased gradient estimator satisfying Assumption 6 with \( B = C = 0 \). Let \( 0 < \gamma \leq \frac{1}{\sqrt{LA}} \). Denote \( f(x^0) - f^{\inf} \) by \( \delta_0 \). Then, the iterates \( x^t \) of SGD using the estimator \( g \) and stepsize \( \gamma \) satisfy

\[
\min_{0 \leq t \leq T-1} \mathbb{E} \left[ \| \nabla f(x^t) \|^2 \right] \leq \frac{2(1 + L \gamma^2 A^T \delta_0)}{\gamma T}.
\]

In particular, for

\[
T \geq \frac{144 \delta_0^2 LA}{\varepsilon^2}
\]

we have \( \min_{0 \leq t \leq T-1} \mathbb{E} \left[ \| \nabla f(x^t) \|^2 \right] \leq \varepsilon \).

Notice that for nonconvex smooth problems, by Lemma 3 Assumption 6 is satisfied for \( A = L_P \) and \( B = C = 0 \). Moreover,

\[
\mathbb{E} \left[ \| g(x) \|^2 \right] = \mathbb{E} \left[ \| \nabla f(x) \|^2_{P^{-1}} \right] \leq \lambda_{\max} \left( P^{-1} \right) \| \nabla f(x) \|^2,
\]

so Assumption 6 is also satisfied for \( B = \lambda_{\max} \left( P^{-1} \right) \) and \( A = C = 0 \). There seem to be no other ways in which the RPT estimator can satisfy Assumption 6.

Plugging in both of these cases into Theorem 4 yields suboptimal convergence rates.

Indeed, for \( A = L_P \) and \( B = C = 0 \) the rate of RPT guaranteed by Theorem 4 is \( O \left( 1/\sqrt{T} \right) \), which is suboptimal (we have proved that it is in fact \( O(1/T) \)).

For \( B = \lambda_{\max} \left( P^{-1} \right) \) and \( A = C = 0 \) the rate is \( O(1/T) \) but it is proportional to

\[
L \lambda_{\max} \left( P^{-1} \right) = \lambda_{\max}(L) \lambda_{\max} \left( P^{-1} \right) \geq L_P,
\]

by Lemma 4.

We manage to obtain faster rate by exploiting \( \mathcal{L} \)-smoothness and the fact that \( \mathbb{E} \left[ \| g(x) \|^2_L \right] \) can be expressed as a suitable matrix norm.

D.2 Convex setting

Applying generic result of Khaled et al. (2023) yields somewhat good convergence guarantee \( O(1/T) \) rate is recovered); however, it still suffers from an unnecessary dependency on function suboptimality.

Theorem 5 (Adapted from Khaled et al. (2023)). Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be convex, differentiable and lower bounded with global minimizer \( x^* \). Suppose that an unbiased gradient estimator \( g \) satisfies

\[
\mathbb{E} \left[ \| g(x) \|^2 \right] \leq 2A(f(x) - f(x^*)).
\]

Choose stepsize

\[
\gamma \leq \min \left\{ \frac{1}{4A}, \frac{1}{2L} \right\}.
\]

Then the SGD procedure with gradient estimator \( g \) and stepsize \( \gamma \) satisfies

\[
\mathbb{E} \left[ f(\bar{x}^T) - f(x^*) \right] \leq \frac{2 \gamma \delta^0 + \| x^0 - x^* \|^2}{\gamma T},
\]

where \( \bar{x}^T \overset{def}{=} \frac{1}{T} \sum_{t=0}^{T-1} x^t \) and \( \delta^0 \overset{def}{=} f(x^0) - f(x^*) \).
As discussed previously, the assumption from Theorem 5 is satisfied for $A = L_P \geq L$. Thus, the optimal theoretical stepsize is $\gamma = 1/(4L_P)$. Plugging in this stepsize yields the rate

$$O \left( \frac{\delta^0 + L_P \|x^0 - x^*\|^2}{T} \right),$$

whereas we obtained the rate of

$$O \left( \frac{L_P \|x^0 - x^*\|^2}{T} \right).$$