Escaping Saddle Points with Compressed SGD

June 4, 2021

Abstract

Stochastic gradient descent (SGD) is a prevalent optimization technique for large-scale distributed machine learning. While SGD computation can be efficiently divided between multiple machines, communication typically becomes a bottleneck in the distributed setting. Gradient compression methods can be used to alleviate this problem, and a recent line of work shows that SGD augmented with gradient compression converges to an ε-first-order stationary point. In this paper we extend these results to convergence to an ε-*second*-order stationary point (ε-SOSP), which is to the best of our knowledge the first result of this type. In addition, we show that, when the stochastic gradient is not Lipschitz, compressed SGD with RandomK compressor converges to an ε-SOSP with the same number of iterations as uncompressed SGD [Jin et al., 2021] (JACM), while improving the total communication by a factor of $\tilde{\Theta}(\sqrt{d}\varepsilon^{-3/4})$, where d is the dimension of the optimization problem. We present additional results for the cases when the compressor is arbitrary and when the stochastic gradient is Lipschitz.[1]

1 Introduction

Stochastic Gradient Descent (SGD) and its variants are the main workhorses of modern machine learning. Distributed implementations of SGD on a cluster of machines with a central server and a large number of workers are frequently used in practice due to the massive size of the data. In distributed SGD each machine holds a copy of the model and the computation proceeds in rounds. In every round, each worker finds a stochastic gradient based on its batch of examples, the server averages these stochastic gradients to obtain the gradient of the entire batch, makes an SGD step, and broadcasts the updated model parameters to the workers. With a large number of workers, computation parallelizes efficiently while communication becomes the main bottleneck [Chilimbi et al., 2014, Strom, 2015], since each worker needs to send its gradients to the server and receive the updated model parameters. Common solutions for this problem include: local SGD and its variants, when each machine performs multiple local steps before communication [Stich, 2018]; decentralized architectures which allow pairwise communication between the workers [McMahan et al., 2017] and gradient compression, when a compressed version of the gradient is communicated instead of the full gradient [Bernstein et al., 2018, Stich et al., 2018, Karimireddy et al., 2019]. In this work, we consider the latter approach, which we refer to as *compressed SGD*.

Most machine learning models can be described by a d-dimensional vector of parameters $\mathbf{x}$ and the model quality can be estimated as a function $f(\mathbf{x})$. Hence optimization of the model parameters can be cast a minimization problem $\min_{\mathbf{x}} f(\mathbf{x})$, where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a continuous function, which can be optimized using continuous optimization techniques, such as SGD. Fast convergence of compressed SGD to a first-order stationary point (FOSP, $\|\nabla f(\mathbf{x})\| < \varepsilon$) was shown recently for various gradient compression schemes [Bernstein et al., 2018, Stich et al., 2018, Karimireddy et al., 2019, Ivkin et al.,

[1] The full version and the code can be found at `https://github.com/nips2021123/nips2021`

2019, Alistarh et al., 2017]. However, even an exact FOSP can be either a local minimum, a saddle point or a local maximum. While local minima often correspond to good solutions in machine learning applications [Ge et al., 2016, Sun et al., 2016, Bhojanapalli et al., 2016], saddle points and local maxima are always suboptimal and it is important for an optimization algorithm to avoid converging to them. In particular, Choromanska et al. [2015] show that for neural networks many local minima are almost optimal, but the corresponding loss functions have a combinatorial explosion in the number of saddle points. Furthermore, Dauphin et al. [2014] show that saddle points can significantly slow down SGD convergence and hence it is important to be able to escape from them efficiently.

Since finding a local minimum is NP-hard in general [Anandkumar and Ge, 2016], a common relaxation of this requirement is to find an approximate second-order stationary point (SOSP), i.e. a point with a small gradient norm ($\|\nabla f(\mathbf{x})\| < \varepsilon$) and the smallest (negative) eigenvalue being small in absolute value ($\lambda_{\min}(\nabla^2 f(\mathbf{x})) > -\varepsilon_H$). When f has ρ-Lipschitz Hessian (i.e. $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \rho\|x-y\|$ for all x, y), a standard choice of ε_H is $\sqrt{\rho\varepsilon}$ [Nesterov and Polyak, 2006], and such approximate SOSP is commonly referred as an ε-SOSP. While second-order optimization methods allow one to escape saddle points, such methods are typically substantially more expensive computationally. A line of work originating with the breakthrough of Ge et al. [2015] shows that first-order methods can escape saddle points when perturbations are added at certain iterations. In particular, a follow-up Jin et al. [2021] show that SGD converges to an ε-SOSP in an almost optimal number of iterations.

In this paper, we show that even *compressed* SGD can efficiently converge to an ε-SOSP. To the best of our knowledge, this is the first result showing convergence of compressed methods to a second-order stationary point.

1.1 Related Work

Escaping from saddle points While it is known that gradient descent with random initialization converges to a local minimum almost surely [Lee et al., 2016], existence of saddle points may result in exponential number of steps with non-negligible probability [Du et al., 2017]. Classical approaches for escaping from saddle points assume access to second-order information [Nesterov and Polyak, 2006, Curtis et al., 2014]. Although these algorithms find a second-order stationary point in $O(\varepsilon^{-3/2})$ iterations, each iteration requires computation of the full Hessian matrix, which can be prohibitive for high-dimensional problems in practice. Some approaches relax this requirement, and instead of full Hessian matrix they only require access to a Hessian-vector product oracle [Carmon and Duchi, 2016, Agarwal et al., 2017]. While in certain settings, including training of neural networks, it's possible to compute Hessian-vector products (HVP) efficiently [Pearlmutter, 1994, Schraudolph, 2002], such an oracle might not be available in general. Furthermore, in practice HVP-based approaches are significantly more complex compared to SGD (especially if the workers aren't communicating in every iteration in the distributed setting) and require additional hyperparameter tuning. Moreover, HVP is typically used for approximate an eigenvector computation, which may increase the number of iterations by a logarithmic factor.

Limitations of second-order methods motivate a long line of recent research on escaping from saddle points using first-order algorithms, starting from Ge et al. [2015]. Jin et al. [2017] show that perturbed gradient descent finds ε-SOSP in $\tilde{O}(\varepsilon^{-2})$ iterations. Later, this is improved by a series of accelerated algorithms [Carmon et al., 2016, Agarwal et al., 2017, Carmon et al., 2017, Jin et al., 2018] which achieves $\tilde{O}\left(\varepsilon^{-7/4}\right)$ iteration complexity. There are also a number of algorithms designed for finite sum setting where $f(x) = \sum_{i=1}^n f_i(x)$ [Reddi et al., 2017, Allen-Zhu and Li, 2018, Fang et al., 2018], or in case when only stochastic gradients are available [Tripuraneni et al., 2018, Jin et al., 2021], including variance reduction techniques [Allen-Zhu, 2018, Fang et al., 2018]. The sharpest rates in these settings have been obtained by Fang et al. [2018], Zhou and Gu [2019] and Fang et al. [2019].

Compressed SGD While gradient compression may require a complex communication protocol, from theoretical perspective this process is often treated as a black-box function: a (possibly randomized) function $\mathcal{C}$ is called a μ-compressor if $\mathbb{E}\left[\|\mathbf{x} - \mathcal{C}(\mathbf{x})\|^2\right] < (1-\mu)\|\mathbf{x}\|^2$. In a simplified form, the update step in compressed SGD can be expressed as $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta\mathcal{C}(\nabla f(\mathbf{x}))$[2]. Notable examples of compressors include the following:

Sign function $\mathcal{C}(\mathbf{x}) = \frac{\|x\|_1}{d} sign(x)$ is a $1/d$-compressor [Bernstein et al., 2018]. Representation of $\mathcal{C}(x)$ requires $O(d)$ bits, but it is hard to compute in distributed settings: it's not clear how to find the signs of the coordinates without knowing the full vector, which requires each worker to send all coordinates. A practical solution is for each worker to communicate Sign of its local gradient, and the final sign for each coordinate is selected by majority vote. Unfortunately, the resulting vector is not necessarily a compression of the gradient.

Quantization [Alistarh et al., 2017] uniformly splits segment $[0, \|\mathbf{x}\|]$ into s buckets of the same size. Let $\ell_i = \left\lfloor \frac{|\mathbf{x}_i|}{\|\mathbf{x}\|/s} \right\rfloor$; then $|\mathbf{x}_i|$ is randomly rounded to one of $\ell_i \frac{\|\mathbf{x}\|}{s}$ and $(\ell_i + 1)\frac{\|\mathbf{x}\|}{s}$. The compressor returns non-zero coordinates after rounding. For $s = 1$, Quantization $Q(\mathbf{x})$ can be represented using $\tilde{O}(\sqrt{d})$ bits. While it doesn't fall into the compression framework, since $\|Q(\mathbf{x}) - \mathbf{x}\|$ can be much greater than $\|\mathbf{x}\|$, it has a property $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}$, which allows one to show convergence.

TopK function preserves only k largest (by the absolute value) coordinates of a vector and is a k/d-compressor [Stich et al., 2018]. This compressor can be represented using $\tilde{O}(k)$ bits, but similarly to Sign, it is hard to compute in distributed settings. To address this issue, Alistarh et al. [2018] assume that TopK of the average gradient is close to the average of TopK of local gradients and show that this assumption holds in practice.

Sketch-based TopK [Ivkin et al., 2019] is randomized communication-efficient compressor based on Count Sketch, which recovers top-k coordinates in a distributed setting. It uses the fact that Count Sketch is a linear sketch (and therefore it can be easily combined across multiple machines) and can be used to recover top-k coordinates of the vector with high probability. Therefore, it can be used as an efficient k/d-compressor requiring $\tilde{O}(k)$ communication.

RandomK compressor preserves k random coordinates of a vector. It is a k/d-compressor [Stich et al., 2018] requiring $O(k)$ communication.

While it is known that compressed SGD converges (e.g. Karimireddy et al. [2019] and the works above), the convergence was shown only to a FOSP. The crucial idea to facilitate convergence is to use error-feedback [Stich et al., 2018]: the difference between the true gradient and its compression is propagated to the next iteration.

1.2 Our Contributions

Our main contribution is the analysis showing that perturbed compressed SGD with error-feedback can escape from saddle points efficiently. Moreover, we show faster convergence rate for a certain type of compressors and show that such compressors exist. Inspired by the ideas from Jin et al. [2021] and Stich et al. [2018], we present an algorithm (Algorithm 1) which uses perturbed compressed gradients with error-feedback and converges to an ε-second-order stationary point (see Theorem 3.3). Our main results shows that compressed SGD with RandomK compressor achieves substantial communication improvement:

[2]The actual update equation is more complicated, see Algorithm 1.

Theorem 1.1 (Informal, Theorem 3.4 and Corollary 3.5) *Assume that f has Lipschitz gradient and Lipschitz Hessian. Let $\alpha = 1$ when the stochastic gradient is Lipschitz and $\alpha = d$ otherwise. Then SGD with* RandomK *compressor (which selects k random coordinates) with $k = \frac{d\varepsilon^{3/4}}{\sqrt{\alpha}}$ converges to an ε-SOSP after $\tilde{O}\left(\frac{\alpha}{\varepsilon^4}\right)$ iterations, with $\tilde{O}\left(\frac{d\sqrt{\alpha}}{\varepsilon^{3+1/4}}\right)$ total communication per worker.*

Compared with the uncompressed case, the total communication improves by $\varepsilon^{-1/4}$ when the stochastic gradient is Lipschitz and by $\sqrt{d}\varepsilon^{-3/4}$ otherwise (the sharpest results for SGD are by Fang et al. [2019] and Jin et al. [2021] respectively). In Theorem 1.1, we heavily rely on the following property of RandomK: when its randomness (i.e. sampled k coordinates) is fixed, the compressor becomes a linear function. For other compressors, this property doesn't necessarily hold; in this case, we show convergence with a slower convergence rate:

Theorem 1.2 (Informal, Theorem 3.3 and Corollary 3.6) *Assume that f has Lipschitz gradient and Lipschitz Hessian. Let $\alpha = 1$ when the stochastic gradient is Lipschitz and $\alpha = d$ otherwise. Let $\mathcal{C}$ be a k/d-compressor requiring $\tilde{O}(k)$ communication. Then SGD with compressor $\mathcal{C}$ with $k = \frac{d\sqrt{d}\varepsilon^{3/4}}{\sqrt{\alpha}}$ converges to an ε-SOSP after $\tilde{O}\left(\frac{\alpha}{\varepsilon^4}\right)$ iterations, with $\tilde{O}\left(\frac{d\sqrt{d}\sqrt{\alpha}}{\varepsilon^{3+1/4}}\right)$ total communication per worker.*

Compared with the uncompressed case, the total communication improves by $\frac{\varepsilon^{-1/4}}{\sqrt{d}}$ when the stochastic gradient is Lipschitz (note that this is the only setting where the convergence improvement is conditional, requiring $\varepsilon = o(d^{-2})$) and by $\varepsilon^{-3/4}$ otherwise. Table 1 in Section 3.2 shows communication improvements for various choices of compression parameters. We outline our main techniques and technical contributions in Section 3.3 and give the complete proof in Appendix A and B.

2 Preliminaries

Function Properties For a twice differentiable nonconvex function $f\colon \mathbb{R}^d \to \mathbb{R}$, we consider the unconstrained minimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. We use the following standard [Jin et al., 2021, Fang et al., 2019, Xu et al., 2018, Allen-Zhu, 2018, Zhou et al., 2018] assumptions about the objective function f:

Assumption A *f is $f_{\max}$-bounded, L-smooth and has ρ-Lipschitz Hessian, i.e. for all x, y:*

$$|f(\mathbf{x}) - f(\mathbf{y})| \le f_{\max}, \quad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \le \rho\|\mathbf{x} - \mathbf{y}\|$$

Assumption B *Access to an unbiased stochastic gradient oracle $\nabla F(\mathbf{x}, \theta)$, whose randomness is controlled by a parameter $\theta \sim \mathcal{D}$[a], with bounded variance, i.e. for all $\mathbf{x}$:*

$$\mathbb{E}_{\theta\sim\mathcal{D}}\left[\nabla F(\mathbf{x}, \theta)\right] = \nabla f(\mathbf{x}), \quad \mathbb{E}_{\theta\sim\mathcal{D}}\left[\|\nabla F(\mathbf{x}, \theta) - \nabla f(\mathbf{x})\|^2\right] \le \sigma^2$$

[a] E.g. θ is a minibatch selected at the current iteration

As shown by the above works, smoothness allows one to achieve fast convergence for nonconvex optimization problems (namely, to use the folklore descent lemma). Similarly, Lipschitz Hessian allows one to show fast second-order convergence, since, within a certain radius, the function stays close to its quadratic approximation (see e.g. Boyd et al. [2004, Section 9.5.3]). As common in the literature, in our convergence rates we treat L, ρ and σ^2 as constants. We also consider an additional optional assumption (see [Jin et al., 2021] for a justification):

Assumption C (Lipschitz stochastic gradient, optional) *For any* $\mathbf{x}, \mathbf{y}, \theta$*:*

$$\|\nabla F(\mathbf{x}, \theta) - \nabla F(\mathbf{y}, \theta)\| \leq \tilde{\ell}\|\mathbf{x} - \mathbf{y}\|.$$

From machine learning perspective, Assumption C means that for the same mini-batch, if the initial models are close, their updates are also close. For neural networks, since each network layer is a composition of an activation function and a linear function, the assumption holds when activation functions are Lipschitz (note however that $\tilde{\ell}$ may grow exponentially with the number of layers)

Gradient Compression Our goal is to optimize f in a distributed setting [Dekel et al., 2012, Li et al., 2014]: given $\mathcal{W}$ workers, for each worker i we have a corresponding data distribution $\mathcal{D}_i$. Then the each worker has a corresponding function $f_i(\mathbf{x}) = \mathbb{E}_{\theta \sim \mathcal{D}_i}[F(\mathbf{x}, \theta)]$ and $f = \sum_{i=1}^{\mathcal{W}} f_i$. In a typical distributed SGD setting, each worker computes a stochastic gradient $\nabla F_i(\mathbf{x}, \theta_i)$ and sends it to the coordinator machine. The coordinator machine computes the average of these gradients $\mathbf{v} = \frac{1}{\mathcal{W}} \sum_{i=1}^{\mathcal{W}} \nabla F_i(\mathbf{x}, \theta_i)$ and broadcasts it to the workers, which update the local parameters $\mathbf{x} \leftarrow \mathbf{x} - \eta \mathbf{v}$ (η is the step size).

With this approach, with increase of the number of machines, the computation can be perfectly parallelized. However, with each machine required to send its gradient, communication becomes the main bottleneck [Chilimbi et al., 2014, Strom, 2015]. There exist various solutions to this problem (see Section 1), including gradient compression, when each machine sends an approximation of its gradient. Then coordinator averages these approximations and broadcasts the average to all machines (possibly compressing it again, see discussion on TopK and Sign in Section 1.1).

Depending on the compression method, this protocol provides different gradient approximation and different communication per machine. There is a natural trade-off between approximation and communication, and it's not clear whether having smaller per-iteration communication results in smaller total communication required for convergence. The approximation quality can be formalized using the following definition:

Definition 2.1 (Stich et al. [2018]) *Function* $\mathcal{C}(\mathbf{x}, \tilde{\theta})$*, whose randomness is controlled by a parameter* $\tilde{\theta} \sim \tilde{\mathcal{D}}$[a]*, is a* μ-compressor *if*

$$\mathbb{E}_{\tilde{\theta} \sim \tilde{\mathcal{D}}}\left[\|\mathbf{x} - \mathcal{C}(\mathbf{x}, \tilde{\theta})\|^2\right] < (1 - \mu)\|\mathbf{x}\|^2.$$

[a] E.g. for RandomK, $\tilde{\theta}$ is the set of indices of coordinates.

Section 1.1 provides examples of important compressors. In our analysis, we consider general and linear compressors separately, and in the latter case, we show an improved convergence rate.

Definition 2.2 $\mathcal{C}$ *is a* linear compressor *if* $\mathcal{C}(\cdot, \tilde{\theta})$ *is a linear function for any* $\tilde{\theta}$*.*

One example of a linear compressor is RandomK, which picks k random coordinates of a vector; it's a k/d-compressor [Stich et al., 2018] and can be computed easily in the distributed setting.

Stationary Points The optimization problem of finding a global minimum or even a local minimum is NP-hard for nonconvex objectives [Nesterov, 2000, Anandkumar and Ge, 2016]. Instead, as is standard in the literature, we show convergence to an approximate FOSP or an approximate SOSP, see Section 1.

Definition 2.3 *If f is differentiable then $\mathbf{x}$ is an ε-First-Order Stationary Point if $\|\nabla f(\mathbf{x})\| \leq \varepsilon$.*

An ε-FOSP can be a local maximum, a local minimum or a saddle point. While local minima typically correspond to good solutions, saddle points and local maxima are inherently suboptimal. Assuming non-degeneracy, saddle points and local maxima have escaping directions, corresponding to Hessian's negative eigenvectors. Following Nesterov and Polyak [2006] we refer to points with no escape directions (up to a second-order approximation) as approximate second-order stationary points:

Definition 2.4 (Nesterov and Polyak [2006]) *If f is a twice differentiable ρ-Hessian Lipschitz function then $\mathbf{x}$ is an ε-Second-Order Stationary Point if $\|\nabla f(\mathbf{x})\| \leq \varepsilon$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\varepsilon}$, where $\lambda_{\min}$ is the smallest eigenvalue*[a].

[a] While one can consider two threshold parameters – ε_g for ∇f and ε_H for $\nabla^2 f$ – we follow convention of Nesterov and Polyak [2006] which selects $\varepsilon_H = -\sqrt{\rho\varepsilon}$, which, intuitively, balances first-order and second-order variability.

An important property of points which are not ε-SOSP is that they are unstable: adding a small perturbation allows gradient descent to escape them [Ge et al., 2015] (similar results were shown for e.g. stochastic [Jin et al., 2021] and accelerated [Jin et al., 2018] gradient descent). In this work we show that this property holds even for SGD with gradient compression.

3 Algorithm and Analysis

Algorithm We present our algorithm in Algorithm 1, a compressed stochastic gradient descent approach based on Stich et al. [2018, Algorithm 1]. In order to achieve convergence to a SOSP, similarly to Jin et al. [2021], we add artificial random noise ξ_t to gradient at every iteration, which allows compressed gradient descent to escape saddle points. At every iteration t, we compute the stochastic gradient $\nabla F(\mathbf{x}_t, \theta_t)$. Then we add artificial noise ξ_t, compress the resulting value (Line 10) and update the current iterate $\mathbf{x}_t$ using the compressed value (Line 11). However, the information is not lost during compression: the difference between the computed value and the compressed value (Line 12), $\mathbf{e}_{t+1}$, is added to the gradient in the next iteration. Karimireddy et al. [2019] show that carrying over the error term improves convergence of compressed SGD to a FOSP. Algorithm 1 accepts an additional Boolean parameter $reset\_error$. When this parameter is true, we set $\mathbf{e}_t$ to 0 (Line 7) when conditions in Line 5 hold: either we moved far from the point where the condition was triggered last time (intuitively, the condition indicates that we successfully escaped from a saddle point), or we spent a certain number of iterations since that event (to ensure that the accumulated compression error is sufficiently bounded).

Distributed Setting Algorithm 1 provides a general framework for compressed SGD in distributed settings, with implementation details depending on the choice of the compressor function $\mathcal{C}$. ξ_t can be efficiently shared between machines using shared randomness. Each machine i maintains its own local $\mathbf{e}_t^{(i)}$ which can be computed as $\mathbf{e}_{t+1}^{(i)} \leftarrow \mathbf{e}_t^{(i)} + \nabla F_i(\mathbf{x}_t, \theta_t) + \xi_t - \mathbf{g}_t^{(i)}$. Then $\mathbf{e}_t = \frac{1}{\mathcal{W}} \sum_{i=1}^{\mathcal{W}} \mathbf{e}_t^{(i)}$. Finally, the norm in Line 1 of Algorithm 7 can be efficiently computed within multiplicative approximation using linear sketches.

Algorithm 1: Compressed SGD

parameters: η – step size, T – number of iterations, r^2 – variance of the artificial noise, *reset_error* – flag indicating whether compression error should be periodically reset to zero, $\mathcal{I}$ – the number of iterations required for escaping, $\mathcal{R}$ – escaping radius
input : objective f, compressor function $\mathcal{C}$, starting point $\mathbf{x}_0$
output : ε-SOSP of f

$\mathbf{e}_0 \leftarrow 0^d$
if *reset_error* **then** $t' \leftarrow 0$;
for $t = 0 \dots T-1$ **do**
  // Reset the error after $\mathcal{I}$ iterations or in case we moved far from the initial point
  if $reset\_error$ *and* $(t - t' > \mathcal{I}$ *or* $\|\mathbf{x}_{t'} - (\mathbf{x}_t - \eta \mathbf{e}_t)\| > \mathcal{R})$ **then**
    if $f(\mathbf{x}_t) < f(\mathbf{x}_{t'}) - \mathcal{F}$ **then** $\mathbf{x}_t \leftarrow \mathbf{x}_t - \eta \mathbf{e}_t$ **else** $\mathbf{x}_t \leftarrow \mathbf{x}_{t'}$;
    $t' \leftarrow t, \quad \mathbf{e}_t \leftarrow 0^d$
  end
  Sample $\xi_t \sim \mathcal{N}_d(0^d, r^2), \quad \theta_t \sim \mathcal{D}, \quad \tilde{\theta}_t \sim \tilde{\mathcal{D}}$
  $\mathbf{g}_t \leftarrow \mathcal{C}(\mathbf{e}_t + \nabla F(\mathbf{x}_t, \theta_t) + \xi_t, \tilde{\theta}_t)$ // Compressed gradient
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \mathbf{g}_t$ // Compressed gradient descent step
  $\mathbf{e}_{t+1} \leftarrow \mathbf{e}_t + \nabla F(\mathbf{x}_t, \theta_t) + \xi_t - \mathbf{g}_t$ // Update the error
end
return $\mathbf{x}_T$

3.1 Convergence to an ε-FOSP

In the following statements, $\tilde{O}$ hides polynomial dependence on $L, \rho, f_{\max}, \sigma, \tilde{\ell}$ and polylogarithmic dependence on all parameters. The first result is similar to that of Stich et al. [2018] (after reformulation in terms of ε-FOSP), but is more general: it covers the case when μ is close to 0 and doesn't require any bounds on $\|\nabla F(\mathbf{x}, \theta)\|$ or $\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|$, which are common assumptions in the literature (see Section 1.1). The proof of the theorem is presented in Appendix A.

Theorem 3.1 (Convergence to ε-FOSP) *Let f satisfy Assumptions A and B and let $\mathcal{C}$ be a μ-compressor. Then for Algorithm 1 with $reset\_error = false$ and $\eta = \tilde{O}\left(\min\left(\varepsilon^2, \frac{\mu}{\sqrt{1-\mu}}\varepsilon\right)\right)$, after $T = \tilde{O}(\frac{1}{\varepsilon^2 \eta}) = \tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu \varepsilon^3}\right)$ iterations, at least half of visited points are ε-FOSP.*

Corollary 3.2 *For a $1/d$-compressor with $\tilde{O}(1)$ communication (polylogarithmic on all parameters), the total communication per worker is $\tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{d}{\varepsilon^3}\right)$, which outperforms full SGD communication $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ by a factor of $\min\left(d, \varepsilon^{-1}\right)$.*

3.2 Convergence to an ε-SOSP

In the following statements, $\tilde{O}$ hides polynomial dependence on $L, \rho, f_{\max}, \sigma, \tilde{\ell}$ and polylogarithmic dependence on all parameters. The next two theorems present our main result, namely that compressed SGD converges to an ε-SOSP[3]. The first theorem addresses the general compressor case.

[3] see proof sketch in Section 3.3 and the full proof in Appendix B

Theorem 3.3 (Convergence to ε-SOSP for general compressor) *Let f satisfy Assumptions A and B, let $\mathcal{C}$ be a μ-compressor. Let $\alpha = 1$ when Assumption C holds and $\alpha = d$ otherwise. Then for Algorithm 1 with reset_error = true and $\eta = \tilde{O}\left(\min\left(\frac{\varepsilon^2}{\alpha}, \frac{\mu\varepsilon}{\sqrt{1-\mu}}, \frac{\mu^2\sqrt{\varepsilon}}{(1-\mu)d}\right)\right)$, after $T = \tilde{O}\left(\frac{1}{\varepsilon^2\eta}\right) = \tilde{O}\left(\frac{\alpha}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu\varepsilon^3} + \frac{d(1-\mu)}{\mu^2\varepsilon^2\sqrt{\varepsilon}}\right)$ iterations, at least half of visited points are ε-SOSP.*

In general, convergence to an ε-SOSP is noticeably slower than convergence to an ε-FOSP. The reason for such behavior is that, in the analysis of convergence to a SOSP, compression introduces an error similar to that of the stochastic noise. When the stochastic gradient is not Lipschitz (i.e. Assumption C doesn't hold), the number of iterations decreases by a factor of d due to stochastic noise. Unfortunately, in general the compression is not Lipschitz even for deterministic gradients: consider a TopK compression of a vector where each coordinate is 1 with small perturbations. However, if the compressor is linear (Definition 2.2), we show improved convergence rate: the last term in the number of iterations decreases by the factor of d.

Theorem 3.4 (Convergence to ε-SOSP for linear compressor) *Let f satisfy Assumptions A and B, let $\mathcal{C}$ be a linear compressor. Let $\alpha = 1$ when Assumption C holds and $\alpha = d$ otherwise. Then for Algorithm 1 with reset_error = false and $\eta = \tilde{O}\left(\min\left(\frac{\varepsilon^2}{\alpha}, \frac{\mu\varepsilon}{\sqrt{1-\mu}}, \frac{\mu^2\sqrt{\varepsilon}}{1-\mu}\right)\right)$, after $T = \tilde{O}\left(\frac{1}{\varepsilon^2\eta}\right) = \tilde{O}\left(\frac{\alpha}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu\varepsilon^3} + \frac{1-\mu}{\mu^2\varepsilon^2\sqrt{\varepsilon}}\right)$ iterations, at least half of visited points are ε-SOSP.*

Since RandomK is a linear compressor, by balancing the terms we have:

Corollary 3.5 *For RandomK compressor with $k = \frac{d\varepsilon^{3/4}}{\sqrt{\alpha}}$, the total number of iterations of Algorithm 1 is $\tilde{O}(\frac{\alpha}{\varepsilon^4})$ and the total communication per worker is $\tilde{O}\left(\frac{d\sqrt{\alpha}}{\varepsilon^{3+1/4}}\right)$. When Assumption C holds, the total communication for RandomK decreases by the factor of $\tilde{\Theta}(\varepsilon^{-1/4})$ compared with the unconstrained case [Fang et al., 2019]. Otherwise, the total communication decreases by the factor of $\tilde{\Theta}(\sqrt{d}\varepsilon^{-3/4})$ (compared with [Jin et al., 2021]).*

Compressed SGD in Distributed Settings Below we consider different scenarios to illustrate how convergence depends on the properties of the compressor. Recall that sketch-based TopK is a k/d-compressor which requires $\tilde{O}(k)$ communication. Selecting $\mu = k/d$, with $k \ll d$, by Theorem 3.3 we have $\eta = \tilde{O}\left(\min\left(\frac{\varepsilon^2}{\alpha}, \frac{k\varepsilon}{d}, \frac{k^2\sqrt{\varepsilon}}{d^3}\right)\right)$. Therefore, the total number of iterations is $\tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{d}{k\varepsilon^3} + \frac{d^3}{k^2\varepsilon^2\sqrt{\varepsilon}}\right)$ and the total communication is $\tilde{O}\left(\frac{k}{\varepsilon^4} + \frac{d}{\varepsilon^3} + \frac{d^3}{k\varepsilon^2\sqrt{\varepsilon}}\right)$.

Note that the above reasoning considers a worst-case scenario. However, in practice it's often possible to achieve good compression at a low communication cost due to the fact that gradient coordinates have heavy-hitters, which are easy to recover using TopK. We formulate this beyond worst-case scenario as the following optional assumption:

Assumption D (Optional) *There exists a constant $c < 1$ such that for all t, $\mathcal{C}(\nabla F(\mathbf{x}_t, \theta_t) + \xi_t + \mathbf{e}_t)$ provides a c-compression and requires $\tilde{O}(1)$ bits of communication per worker.*

In other words, for all computed values, $\mathcal{C}$ provides a constant compression and requires a polylogarithmic amount of communication. This assumption can be satisfied under various conditions. For example, some methods may take advantage of the situation when gradients between adjacent iterations are close [Hanzely et al., 2018]. In cases when certain coordinates are much more prominent in the gradient compared to others, TopK compressor will show good performance.

Table 1: Convergence to ε-SOSP with uncompressed SGD, with sketch-based TopK compressor, with RandomK compressor, and with a constant-compressor requiring constant communication (Assumption D, beyond worst-case assumption). We considered two settings depending on whether the stochastic gradient is Lipschitz (i.e. Assumption C holds). For each setting we select the optimal μ based on our bounds. The results show that communication of SGD with RandomK compression outperforms that of the uncompressed SGD by $\tilde{O}(\varepsilon^{-1/4})$ when Assumption C holds and by $\tilde{O}(\sqrt{d}\varepsilon^{-3/4})$ otherwise. Based on our results, since a constant-memory constant-communication compressor is not necessarily linear, depending on d and ε, it may converge slower than RandomK.

	Setting	μ	Iterations	Total comm. per worker	Total comm. improvement
Lipschitz ∇F	Uncompressed [Fang et al., 2019]	1	$\tilde{O}\left(\frac{1}{\varepsilon^{3.5}}\right)$	$\tilde{O}\left(\frac{d}{\varepsilon^{3.5}}\right)$	–
	Sketch-based TopK	$\sqrt{d}\varepsilon^{3/4}$ ($\varepsilon = o(d^{-2})$)	$\tilde{O}\left(\frac{1}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d\sqrt{d}}{\varepsilon^{3+1/4}}\right)$	$\tilde{\Theta}\left(\frac{\varepsilon^{-1/4}}{\sqrt{d}}\right)$
	RandomK	$\varepsilon^{3/4}$	$\tilde{O}\left(\frac{1}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d}{\varepsilon^{3+1/4}}\right)$	$\boldsymbol{\tilde{\Theta}\left(\varepsilon^{-1/4}\right)}$
	Constant-memory c-compressor	$c > 0$	$\tilde{O}\left(\frac{1}{\varepsilon^{4}} + \frac{d}{\varepsilon^{2}\sqrt{\varepsilon}}\right)$	$\tilde{O}\left(\frac{1}{\varepsilon^{4}} + \frac{d}{\varepsilon^{2}\sqrt{\varepsilon}}\right)$	$\tilde{\Theta}\left(\min(d, \frac{1}{\sqrt{\varepsilon}})\right)$
non-Lipschitz ∇F	Uncompressed [Jin et al., 2021]	1	$\tilde{O}\left(\frac{d}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d^2}{\varepsilon^{4}}\right)$	–
	Sketch-based TopK	$\varepsilon^{3/4}$	$\tilde{O}\left(\frac{d}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d^2}{\varepsilon^{3+1/4}}\right)$	$\tilde{\Theta}\left(\varepsilon^{-3/4}\right)$
	RandomK	$\frac{\varepsilon^{3/4}}{\sqrt{d}}$	$\tilde{O}\left(\frac{1}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d\sqrt{d}}{\varepsilon^{3+1/4}}\right)$	$\boldsymbol{\tilde{\Theta}\left(\sqrt{d}\varepsilon^{-3/4}\right)}$
	Constant-memory c-compressor	$c > 0$	$\tilde{O}\left(\frac{d}{\varepsilon^{4}}\right)$	$\tilde{O}\left(\frac{d}{\varepsilon^{4}}\right)$	$\tilde{\Theta}\left(d\right)$

Corollary 3.6 *Algorithm 1 converges to ε-SOSP in a number of settings, as shown in Table 1.*

3.3 Proof Sketch

In this section, we outline the main techniques used to prove Theorems 3.3 and 3.4. A recent breakthrough line of work focused on convergence of first-order methods to ε-SOSP Ge et al. [2015], Carmon and Duchi [2016], Jin et al. [2017], Tripuraneni et al. [2018], Jin et al. [2021] (JACM) has developed a comprehensive set of analytic techniques. We start by outlining Jin et al. [2021], which is the sharpest known SGD analysis in the case when the stochastic gradient is not Lipschitz.

Let $\mathbf{x}_0$ be an iterate such that $\lambda_{\min}(\nabla^2 f(\mathbf{x}_0)) < -\sqrt{\rho\varepsilon}$, and $\mathbf{v}_1$ be the eigenvector corresponding to $\lambda_{\min}$. Consider sequences $\{\mathbf{x}_t\}$ and $\{\mathbf{x}'_t\}$ starting with $\mathbf{x}_0$ which are referred to as *coupling sequences*: their distributions match the distribution of compressed SGD iterates (i.e. both sequences can be produced by Algorithm 1), and they share the same randomness, with an exception that their artificial noise has the opposite sign in the direction $\mathbf{v}_1$. The main idea is that such artificial noise combined with SGD updates ensures that projection of $\mathbf{x}_t - \mathbf{x}'_t$ on $\mathbf{v}_1$ increases exponentially, and therefore at least one of the sequences moves far from $\mathbf{x}$. After that, one can use an "Improve or localize" Lemma which states that, if we move far from the original point, then the objective decreases substantially.

If we have an access to a deterministic gradient oracle and the objective function is quadratic, then gradient descent behaves similarly to the power method, since in this case:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) = \mathbf{x}_t - \eta \nabla^2 f(\mathbf{x}_0)\mathbf{x}_t = (I - \eta \nabla^2 f(\mathbf{x}_0))\mathbf{x}_t.$$

Adding artificial noise guarantees that there is a non-trivial projection of $\mathbf{x}_t - \mathbf{x}'_t$ on direction $\mathbf{v}_1$, and the power method further amplifies this projection. In general, the SGD behavior deviates from power method due to: 1) the difference between f and its quadratic approximation and 2) stochastic noise. Jin et al. [2021] show that the errors introduced by these deviations are dominated by the increase in direction $\mathbf{v}_1$, and therefore SGD successfully escapes saddle points.

Outline of our compressed SGD analysis The analysis above is not applicable to our algorithm due to gradient compression and error-feedback. Moreover, in the case of an arbitrary compressor we change the algorithm even further by periodically setting $\mathbf{e}_t$ to 0.

One of the major changes is that errors introduced by the compression lead to even greater deviation of SGD from the power method, and this deviation can potentially dominate other terms: if the compression error is accumulated from the beginning of the algorithm execution, then the compression error can be arbitrarily large. Let $\mathbf{e}'_t$ be the compression error sequence corresponding to $\mathbf{x}'_t$ such that $\mathbf{e}'_0 = \mathbf{e}_0$. The deviation of SGD from the power method caused by compression can be expressed as:

$$\eta^2 \nabla^2 f(\mathbf{x}_0) \sum_{i=1}^{t-1} (I - \eta \nabla^2 f(\mathbf{x}_0))^{t-1-i} (\mathbf{e}_i - \mathbf{e}'_i). \quad \text{(Proposition B.12 and Lemma B.16)}$$

It remains to bound $\|\mathbf{e}_i - \mathbf{e}'_i\|$ for all i. For $G_t = \mathbf{e}_t + \nabla F(\mathbf{x}_t, \theta_t) + \xi_t$ (with G'_t defined analogously), since $\mathbf{e}_{t+1} = G_t - \mathcal{C}(G_t)$, by linearity of $\mathcal{C}$ we have:

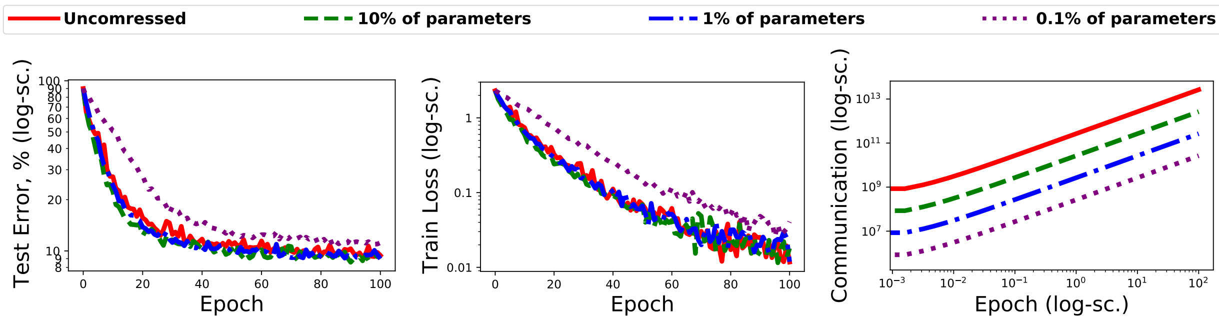
$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{e}_{t+1} - \mathbf{e}'_{t+1}\|^2\right] &= \mathbb{E}\left[\|(G_t - G'_t) - \mathcal{C}(G_t - G'_t, \tilde{\theta}_t)\|^2\right] \\
&\le (1-\mu)\mathbb{E}\left[\|G_t - G'_t\|^2\right] \\
&= (1-\mu)\mathbb{E}\left[\|(\mathbf{e}_t - \mathbf{e}'_t) + (\nabla F(\mathbf{x}_t, \theta_t) - \nabla F(\mathbf{x}'_t, \theta_t)) + (\xi_t - \xi'_t)\|^2\right].
\end{aligned}$$

Since $\mathbf{e}_0 = \mathbf{e}'_0$, after telescoping, $\|\mathbf{e}_{t+1} - \mathbf{e}'_{t+1}\|$ can be bounded using $\|\nabla F(\mathbf{x}_i, \theta_i) - \nabla F(\mathbf{x}'_i, \theta_i)\|$ and $\|\xi_i - \xi'_i\|$ for $i \in [0:t]$ (Lemma B.17). In other words, when escaping from a saddle point, the deviation can bounded based on gradients and noises encountered during escaping. Therefore it is comparable to other terms and can be bounded with an appropriate choice of η.

Unfortunately, for the arbitrary compressor case we don't have a good estimation on $\mathcal{E}_t$, since in general we don't have better bound on $\|\mathbf{e}_i - \mathbf{e}'_i\|$ than $\|\mathbf{e}_i\| + \|\mathbf{e}'_i\|$ (see proof of Lemma B.16). Lemma A.5 bounds the compression error $\mathbf{e}_t$ in terms of $\|\nabla f(\mathbf{x}_0)\|, \ldots, \|\nabla f(\mathbf{x}_t)\|$:

$$\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \le \frac{2(1-\mu)}{\mu} \sum_{i=0}^{t-1} \left(1 - \frac{\mu}{2}\right)^{t-i} \mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2 + \chi^2\right],$$

but the bound depends on all gradients starting from the first iteration. To solve this problem, we periodically set the compression error to 0 (Line 5 of Algorithm 1). Let t' be an iteration such that $\mathbf{e}_{t'}$ is set to 0: then, when escaping from $\mathbf{x}_{t'}$, we can apply Lemma A.5 with i starting from t'. This leads to major difference from the Jin et al. [2021] analysis: we need to consider large- and small-gradient cases separately. When the gradient at $\mathbf{x}_{t'}$ is large (Lemma B.7), we show that nearby gradients are also large, and the objective improves by the Compressed Descent Lemma A.7. Otherwise, we can bound the error norm for the next few iterations (Lemma B.16).



(a) Test Error (b) Train Loss (c) # of communicated parameters

Figure 1: Convergence of distributed SGD ($\eta = 0.1$, batch size is 8 per machine) with RandomK compressor when 100% (full gradient), 10%, 1% and 0.1% of coordinates are used. ResNet34 model is trained on CIFAR-10 distributed across 10 machines: each machine corresponds to a single class. SGD with 10% and 1% compression achieves performance similar to that of uncompressed SGD, while requiring significantly less communication

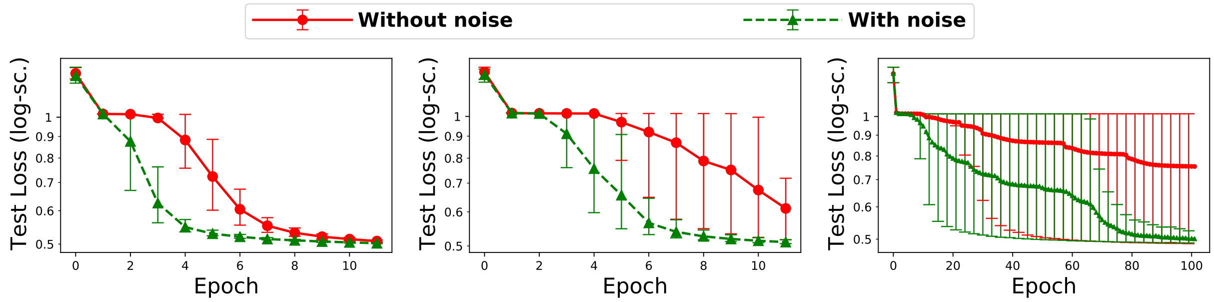
Finally, the analysis uses not only the sequence of iterates $\{\mathbf{x}_t\}$, but also the corrected sequence $\{\mathbf{y}_t\}$ where $\mathbf{y}_t = \mathbf{x}_t - \eta \mathbf{e}_t$ (similarly, $\mathbf{y}'_t = \mathbf{x}'_t - \eta \mathbf{e}'_t$). Intuitively, $\mathbf{e}_t$ accumulates the difference between the communicated and the original gradient, and therefore the goal of $\mathbf{y}_t$ is to offset the compression error. Typically, $\mathbf{x}_t$ is used as an argument of $\nabla f(\cdot)$, while $\mathbf{y}_t$ is used in distances and as an argument of $f(\cdot)$, which noticeably complicates the analysis. In particular, if some property holds for $\mathbf{x}_t$, it doesn't necessarily hold for $\mathbf{y}_t$ and vice versa: for example, since $\mathbf{x}_t$ and $\mathbf{y}_t$ are not necessarily close, bound $\|\mathbf{y}_t - \mathbf{y}'_t\|$ doesn't in general imply bound on $\|\mathbf{x}_t - \mathbf{x}'_t\|$. However, in our analysis, we show that we can bound $\|\mathbf{x}_t - \mathbf{x}'_t\|$, which is required to bound $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}'_t)\|$ in Lemma B.18.

4 Experiments

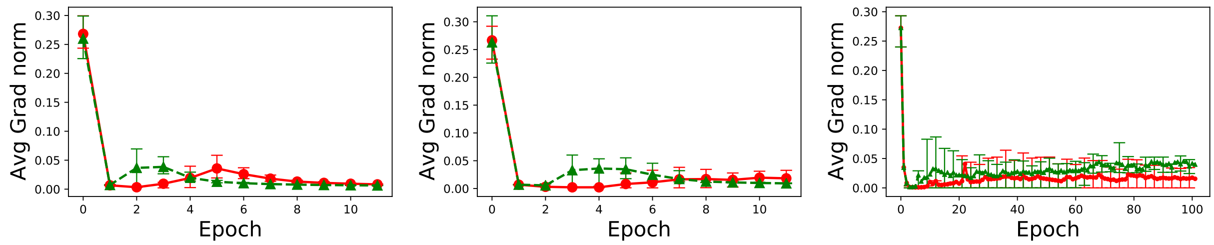
In our experiments, we show that noisy Compressed SGD achieves convergence comparable with full SGD and successfully escapes saddle points. We perform our first set of experiments on ResNet34 model trained using CIFAR-10 dataset with step size 0.1. We distribute the data across 10 machines, such that each machine contains data from a single class. We analyze convergence of compressed SGD with RandomK compressor when 100%, 10%, 1% and 0.1% random gradient coordinates are communicated. Figure 1 shows that SGD with RandomK with 10% or 1% of coordinates compression converges as fast as the full SGD, while requiring substantially smaller communication.

In our second set of experiments, we show that SGD indeed encounters saddle points and noise facilitates escaping from them. We compare uncompressed SGD, SGD with TopK compressor (0.1% of coordinates), and SGD with RandomK compressor (0.1% of coordinates) on deep MNIST autoencoder[4]. In all settings, we compare their convergence rates with and without noise. Figure 2 shows that SGD does encounter saddle points: e.g. in Figure 2a, for SGD without noise, during epochs 1-3, the gradient norm is close to 0 and the objective value doesn't improve. However, compressed SGD escapes from the saddle points, and noise significantly improves the escaping rate.

[4]The encoder is defined using 3 convolutional layers with ReLU activation, with the following parameters: (channels=16, kernel=3, stride=2, padding=1), (channels=32, kernel=3, stride=2, padding=1) and (channels=64, kernel=7, stride=1, padding=0). The decoder is symmetrical.



(a) Uncompressed loss (b) TopK loss (c) Random compressor loss



(d) Uncompressed gradient norm (e) TopK gradient norm (f) RandomK gradient norm

Figure 2: Convergence of SGD ($\eta = 0.1$, batch size is 64) without compression (left), with TopK (0.1% of coordinates) compression (middle), and with RandomK (0.1% of coordinates) compression (right) on MNIST autoencoder dataset without noise (red) and with Gaussian noise (green, $\sigma = 0.01$ for each coordinate). Data points correspond to average over 10 executions and error bars correspond to 10%- and 90%-quantiles. The bottom row shows the norms of the gradients averaged over the last 100 iterations. The figure shows that SGD encounters and escapes saddle points for all compressors, and adding noise facilitates escaping from the saddle points[†].

[†] For the sake of presentation, to ensure that gradient converges to 0, we decrease the magnitude of the artificial noise at later iterations. With a fixed noise magnitude, as our theory predicts, gradient norm converges if a smaller step size is used, but this requires significantly more iterations. Note that this modification only affects the gradient convergence as the objective converges even with fixed noise and a large step size.

5 Conclusion

We give the first result for convergence of compressed SGD to an ε-SOSP, and it's possible that the convergence rate and the total communication can be further improved. When Assumption C holds, it is likely that the communication can be improved by an $\varepsilon^{-1/4}$ factor using techniques from Fang et al. [2019], which achieve $\tilde{O}(\varepsilon^{-3.5})$ convergence rate under Assumption C. Using variance reduction techniques, which achieve $\tilde{O}(\varepsilon^{-3})$ convergence rate, we expect $\varepsilon^{-1/2}$ improvement. Finally, it remains open whether linearity of the compressor is required for Theorem 3.4: similarly to the stochastic gradient case, it may suffice for the compressor to be Lipschitz.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. *arXiv preprint arXiv:1809.10505*, 2018.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In *Advances in neural information processing systems*, pages 2675–2686, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pages 3716–3726, 2018.

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 81–102, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/anandkumar16.html.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3873–3881. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6271-global-optimality-of-local-search-for-low-rank-matrix-recovery.pdf.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yair Carmon and John C. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *ArXiv*, abs/1612.00547, 2016.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Yair Carmon, Oliver Hinder, John C Duchi, and Aaron Sidford. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. *arXiv preprint arXiv:1705.02766*, 2017.

Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 571–582, 2014.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $o(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, 27:2933–2941, 2014.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. *arXiv preprint arXiv:1902.00247*, 2019.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2981–2989, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. Sega: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pages 2082–2093, 2018.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. In *Advances in Neural Information Processing Systems*, pages 13144–13154, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1724–1732. JMLR.org, 2017. URL http://dl.acm.org/citation.cfm?id=3305381.3305559.

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.

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

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *NIPS*, volume 2, pages 1–4, 2014.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Comput.*, 6(1):147–160, January 1994. ISSN 0899-7667. doi: 10.1162/neco.1994.6.1.147. URL https://doi.org/10.1162/neco.1994.6.1.147.

Sashank J Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Francis Bach, Ruslan Salakhutdinov, and Alexander J Smola. A generic approach for escaping saddle points. *arXiv preprint arXiv:1709.01434*, 2017.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

Nikko Strom. Scalable distributed dnn training using commodity gpu cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2379–2383, 2016.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2904–2913, 2018.

Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5530–5540, 2018.

Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. *arXiv preprint arXiv:1901.11518*, 2019.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Finding local minima via stochastic nested variance reduction. *arXiv preprint arXiv:1806.08782*, 2018.

A Convergence to ε-FOSP

In this section we prove Theorem 3.1, showing that Algorithm 1 converges to an approximate first-order stationary point. Results and proofs are inspired by Karimireddy et al. [2019], with the key difference in that we show how to avoid using the bounded gradient assumption: $\mathbb{E}\left[\|\nabla F\|^2\right] \le G^2$ and handle the case of μ-compressors with $\mu \ll 1$. Furthermore, Compressed Descent Lemma (Lemma A.7) is a foundation for showing a second-order convergence.

Definition A.1 (Noise and compression parameters) *We use the following notation:*

- $\zeta_t = \nabla F(\mathbf{x}_t, \theta_t) - \nabla f(\mathbf{x}_t)$ *is stochastic gradient noise. This noise has variance* σ^2
- ξ_t *is artificial Gaussian noise added at every iteration. This noise has variance* r^2
- $\psi_t = \zeta_t + \xi_t$ *is the total noise. This noise has variance* $\chi^2 = \sigma^2 + r^2$*.*
- *We assume that gradients are compressed using a* μ*-compressor* C*.*

For the sake of the analysis, similarly to Karimireddy et al. [2019], we introduce an auxiliary sequence of corrected iterates $\{\mathbf{y}_t\}$, which remove the impact of the compression error.

Definition A.2 (Corrected iterates) *The sequence of* corrected iterates $\{\mathbf{y}_t\}$ *is defined as*

$$\mathbf{y}_t = \mathbf{x}_t - \eta \mathbf{e}_t$$

Proposition A.3 *For the sequence* $\{\mathbf{y}_t\}$*, we have* $\mathbf{y}_{t+1} - \mathbf{y}_t = -\eta(\nabla f(\mathbf{x}_t) + \psi_t)$

Proof : Recall that $\mathbf{e}_{t+1} = \nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t - \mathbf{g}_t$ and $\mathbf{g}_t = \mathcal{C}(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t, \theta_t)$ and thus

$$\mathcal{C}(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t, \tilde{\theta}_t) = \nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t - \mathbf{e}_{t+1}.$$

Substituting this into equation for $\mathbf{y}_{t+1}$:

$$\begin{aligned}
\mathbf{y}_{t+1} &= \mathbf{x}_{t+1} - \eta \mathbf{e}_{t+1} \\
&= \mathbf{x}_t - \eta \mathcal{C}(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t, \tilde{\theta}_t) - \eta \mathbf{e}_{t+1} \qquad (\text{Since } \mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathcal{C}(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t, \tilde{\theta}_t)) \\
&= \mathbf{x}_t - \eta(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t - \mathbf{e}_{t+1}) - \eta \mathbf{e}_{t+1} \\
&= \mathbf{x}_t - \eta(\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t) \\
&= \mathbf{x}_t - \eta \mathbf{e}_t - \eta(\nabla f(\mathbf{x}_t) + \psi_t) \\
&= \mathbf{y}_t - \eta(\nabla f(\mathbf{x}_t) + \psi_t)
\end{aligned}$$

□

In our derivations, we'll often use conditional expectation with respect to current iterates.

Definition A.4 (Conditional expectation w.r.t. iterate) *Let t be an iteration and ξ be a random variable. Then*

$$\mathbb{E}_t\left[\xi\right] := \mathbb{E}_{(\theta_t, \tilde{\theta}_t, \xi_t), (\theta_{t+1}, \tilde{\theta}_{t+1}, \xi_{t+1}), \ldots}\left[\xi \mid \mathbf{x}_t, \mathbf{e}_t\right]$$

A.1 Compression Error Bound

Recall that the compression error terms $\mathbf{e}_t$ in Algorithm 1 represent the difference between the computed gradient and the compressed gradient. Similarly to how stochastic noise increases the number of iterations compared with deterministic gradient descent, compression errors also increase the number of iterations, and therefore it's important to bound $\|\mathbf{e}_t\|$.

Lemma A.5 (Compression Error Bound) *Let $\mathbf{x}_t, \mathbf{e}_t$ be defined as in Algorithm 1 and let χ^2 be as in Definition A.1. Then under Assumptions A and B, for any t we have*

$$\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \leq \frac{2(1-\mu)}{\mu} \sum_{i=0}^{t-1} \left(1 - \frac{\mu}{2}\right)^{t-i} \mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2 + \chi^2\right],$$

In particular, by considering a uniform bound on $\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2\right]$ and taking the sum of the geometric series, we get a result similar to Karimireddy et al. [2019, Lemma 3]:

$$\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \leq \frac{4(1-\mu)}{\mu^2} \left(\max_{i=0}^{t-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2\right] + \chi^2\right)$$

Proof : The proof is similar to the one of Karimireddy et al. [2019, Lemma 3]. The main difference is that we don't rely on the bounded gradient assumption.

By definition of $\mathbf{e}_{t+1}$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}_{t+1}\|^2\right] &= \mathbb{E}\left[\|\mathbf{e}_t + \nabla f(\mathbf{x}_t) + \psi_t - \mathcal{C}(\mathbf{e}_t + \nabla f(\mathbf{x}_t) + \psi_t, \theta_t)\|^2\right] \\ &\leq (1-\mu)\mathbb{E}\left[\|\mathbf{e}_t + \nabla f(\mathbf{x}_t) + \psi_t\|^2\right]\end{aligned}$$

By using inequality $\|a+b\|^2 \leq (1+\nu)\|a\|^2 + (1+\frac{1}{\nu})\|b\|^2$ for any ν:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}_{t+1}\|^2\right] &\leq (1-\mu)((1+\nu)\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] + (1+\frac{1}{\nu})\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) + \psi_t\|^2\right]) \\ &\leq \sum_{i=0}^{t}(1-\mu)^{t-i+1}(1+\nu)^{t-i}(1+\frac{1}{\nu})\mathbb{E}\left[\|\nabla f(\mathbf{x}_i) + \psi_i\|^2\right] \qquad \text{(Telescoping)} \\ &\leq \frac{1}{\nu}\sum_{i=0}^{t}((1-\mu)(1+\nu))^{t-i+1}\,\mathbb{E}\left[\|\nabla f(\mathbf{x}_i) + \psi_i\|^2\right]\end{aligned}$$

By selecting $\nu = \frac{\mu}{2(1-\mu)}$, we have $(1-\mu)(1+\nu) = 1 - \frac{\mu}{2}$. Therefore:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}_{t+1}\|^2\right] &\leq \frac{2(1-\mu)}{\mu}\sum_{i=0}^{t}\left(1-\frac{\mu}{2}\right)^{t-i+1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i) + \psi_i\|^2\right] \\ &= \frac{2(1-\mu)}{\mu}\sum_{i=0}^{t}\left(1-\frac{\mu}{2}\right)^{t-i+1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2 + \chi^2\right] \qquad (\mathbb{E}\left[\chi \mid \mathbf{x}_i\right] = 0)\end{aligned}$$

□

For the sum of $\|\mathbf{e}_t\|^2$, we have the following, simpler expression:

Corollary A.6 *Under assumptions of Lemma A.5, we have*

$$\sum_{\tau=0}^{t}\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right] \le \frac{4(1-\mu)}{\mu^2}\sum_{\tau=0}^{t}(\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right]+\chi^2)$$

Proof :

$$\begin{aligned}\sum_{\tau=0}^{t}\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right] &\le \frac{2(1-\mu)}{\mu}\sum_{\tau=0}^{t}\sum_{i=0}^{\tau}\left(1-\frac{\mu}{2}\right)^{\tau-i+1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2+\chi^2\right] \\ &\le \frac{2(1-\mu)}{\mu}\sum_{i=0}^{t}\left(\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2+\chi^2\right]\sum_{\tau=i}^{t}\left(1-\frac{\mu}{2}\right)^{\tau-i+1}\right)\end{aligned}$$

Bounding $\sum_\tau \left(1-\frac{\mu}{2}\right)^{\tau-i+1}$ with the sum of the geometric series $\frac{2}{\mu}$, we have:

$$\sum_{\tau=0}^{t}\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right] \le \frac{4(1-\mu)}{\mu^2}\sum_{i=0}^{t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2+\chi^2\right]$$

□

A.2 Compressed Descent Lemma

The following descent lemma is the key tool in the analysis as it allows us to bound gradient norms across multiple iterations.

Lemma A.7 (Compressed Descent Lemma) *Let f satisfy Assumptions A and B and χ^2 be as in Definition A.1. For $\eta < \frac{1}{4L}\min(\frac{\mu}{\sqrt{1-\mu}}, 1)$, for any T we have:*

$$\sum_{\tau=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] \le \frac{4(f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right])}{\eta} + \eta T\chi^2\left(2L+\frac{8L^2\eta(1-\mu)}{\mu^2}\right)$$

Using this lemma, we'll later show that for sufficiently large T, multiple visited points have small gradients (note that by dividing the left-hand side by T we obtain an average squared gradient norm), making them ε-FOSP. On the right-hand side the first term is bounded by $4f_{\max}/\eta$, while the other two terms can be bounded by selecting a sufficiently small η. The second term arises from stochastic gradient noise, while the last term stems from the compression error.

Proof : The proof is similar to the one of Karimireddy et al. [2019, Theorem II]. By the folklore descent lemma, using notation $\mathbb{E}_t[\cdot]$ from Definition A.4:

$$\begin{aligned}&\mathbb{E}_t\left[f(\mathbf{y}_{t+1})\right] \\ &\le f(\mathbf{y}_t) + \langle\nabla f(\mathbf{y}_t), \mathbb{E}_t\left[\mathbf{y}_{t+1}-\mathbf{y}_t\right]\rangle + \frac{L}{2}\mathbb{E}_t\left[\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right] \\ &= f(\mathbf{y}_t) - \eta\mathbb{E}_{\theta_t,\tilde{\theta}_t}\left[\langle\nabla f(\mathbf{y}_t), \nabla f(\mathbf{x}_t)+\psi_t\rangle \mid \mathbf{x}_t, \mathbf{e}_t\right] + \frac{L\eta^2}{2}\mathbb{E}_{\theta_t,\tilde{\theta}_t}\left[\|\nabla f(\mathbf{x}_t)+\psi_t\|^2 \mid \mathbf{x}_t, \mathbf{e}_t\right] \quad (\textit{Prop. A.3}) \\ &\le f(\mathbf{y}_t) - \eta\|\nabla f(\mathbf{x}_t)\|^2 - \eta\langle\nabla f(\mathbf{y}_t)-\nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t)\rangle + \frac{L\eta^2}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\eta^2\chi^2}{2} \quad (\mathbb{E}\left[\psi_t\right]=0) \\ &\le f(\mathbf{y}_t) - \eta\left(1-\frac{L\eta}{2}\right)\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\eta^2\chi^2}{2} - \eta\langle\nabla f(\mathbf{y}_t)-\nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t)\rangle\end{aligned}$$

Using inequality $|\langle a, b\rangle| \le \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$ and smoothness, we have:

$$\begin{aligned}
&\mathbb{E}_t\left[f(\mathbf{y}_{t+1}) \mid \mathbf{x}_t, \mathbf{e}_t\right] \\
&\le f(\mathbf{y}_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\eta^2\chi^2}{2} + \frac{\eta}{2}\|\nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}_t)\|^2 + \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\le f(\mathbf{y}_t) - \eta\left(\frac{1}{2} - \frac{L\eta}{2}\right)\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\eta^2\chi^2}{2} + \frac{\eta L^2}{2}\|\mathbf{y}_t - \mathbf{x}_t\|^2 && (L\text{-smoothness}) \\
&\le f(\mathbf{y}_t) - \eta\left(\frac{1}{2} - \frac{L\eta}{2}\right)\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\eta^2\chi^2}{2} + \frac{\eta^3 L^2}{2}\|\mathbf{e}_t\|^2 && (\text{Def. A.2 of } \mathbf{y}_t)
\end{aligned}$$

Using telescoping and taking the expectation, we bound $f(\mathbf{y}_{t+1})$:

$$\mathbb{E}\left[f(\mathbf{y}_{t+1})\right] \le f(\mathbf{y}_0) - \eta\left(\frac{1}{2} - \frac{L\eta}{2}\right)\sum_{\tau=0}^{t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] + \frac{L\chi^2\eta^2(t+1)}{2} + \eta^3 L^2\sum_{\tau=0}^{t}\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right]$$

Bounding $\sum_\tau \|\mathbf{e}_\tau\|^2$ by Corollary A.6, we have:

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{y}_t)\right] \\
&\le f(\mathbf{y}_0) - \eta\left(\frac{1}{2} - \frac{L\eta}{2}\right)\sum_{\tau=0}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] + \frac{L\chi^2\eta^2 t}{2} + \frac{2\eta^3 L^2(1-\mu)}{\mu^2}\sum_{i=0}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2 + \chi^2\right] \\
&\le f(\mathbf{y}_0) - \eta\left(\frac{1}{2} - \frac{L\eta}{2} - \frac{2\eta^2 L^2(1-\mu)}{\mu^2}\right)\sum_{\tau=0}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] + \frac{L\chi^2\eta^2 t}{2} + \frac{2\eta^3 L^2\chi^2(1-\mu)t}{\mu^2}
\end{aligned}$$

Using that $\eta < \frac{1}{4L}\min\left(\frac{\mu}{\sqrt{1-\mu}}, 1\right)$, we bound the coefficient before $\sum_{\tau=0}^{t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right]$ with $\frac{\eta}{4}$:

$$\mathbb{E}\left[f(\mathbf{y}_t)\right] \le f(\mathbf{y}_0) - \frac{\eta}{4}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] + \eta^2\chi^2 t\left(\frac{L}{2} + \frac{2L^2\eta(1-\mu)}{\mu^2}\right)$$

After regrouping the terms, we get the final result:

$$\sum_{\tau=0}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] \le \frac{4(f(\mathbf{y}_0) - \mathbb{E}\left[f(\mathbf{y}_t)\right])}{\eta} + \eta\chi^2 t\left(2L + \frac{8L^2\eta(1-\mu)}{\mu^2}\right)$$

□

When showing convergence to SOSP, we'll need a generalization of this Lemma which start tracking communication error from the last iteration when the error was 0:

Corollary A.8 *Let f satisfy Assumptions A and B and χ^2 be as in Definition A.1. If t_0 is an iteration of Algorithm 1 such that $\mathbf{e}_{t_0} = 0$ and $\eta < \frac{1}{4L}\min(\frac{\mu}{\sqrt{1-\mu}}, 1)$, then for any T we have:*

$$\sum_{\tau=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t_0+\tau})\|^2\right] \le \frac{4(f(\mathbf{y}_{t_0}) - \mathbb{E}\left[f(\mathbf{y}_{t_0+T})\right])}{\eta} + \eta T\chi^2\left(2L + \frac{8L^2\eta(1-\mu)}{\mu^2}\right)$$

A.3 Convergence to ε-FOSP

Theorem A.9 (Convergence to ε-FOSP) *Let f satisfy Assumptions A and B. Then for $\eta = \tilde{O}\left(\min\left(\varepsilon^2, \frac{\mu}{\sqrt{1-\mu}}\varepsilon\right)\right)$, after $T = \tilde{\Theta}\left(\frac{1}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu\varepsilon^3}\right)$ iterations, at least half of visited points are ε-FOSP.*

Proof : Proof by contradiction. For $\eta < \frac{1}{4L}\min\left(\frac{\mu}{\sqrt{1-\mu}}, 1\right)$, if less than half points are ε-FOSP, then by Lemma A.7:

$$\frac{T\varepsilon^2}{2} \le \sum_{\tau=0}^{T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|^2\right] \le \frac{4f_{\max}}{\eta} + \eta\chi^2 T\left(2L + \frac{8L^2\eta(1-\mu)}{\mu^2}\right)$$

It suffices to guarantee that all terms on the right-hand side are at most $\frac{T\varepsilon^2}{6}$:

$$\begin{aligned}
2L\eta\chi^2 T \le \frac{T\varepsilon^2}{6} &\iff \eta \le \frac{\varepsilon^2}{12L\chi^2} && = \tilde{\Theta}(\varepsilon^2) \\
\frac{8L^2\chi^2\eta^2 T(1-\mu)}{\mu^2} \le \frac{T\varepsilon^2}{6} &\iff \eta \le \frac{\mu\varepsilon}{\sqrt{1-\mu}L\chi\sqrt{48}} && = \tilde{\Theta}\left(\frac{\mu\varepsilon}{\sqrt{1-\mu}}\right) \\
\frac{4f_{\max}}{\eta} \le \frac{T\varepsilon^2}{6} &\iff T \ge \frac{24f_{\max}}{\varepsilon^2\eta} && = \tilde{\Theta}\left(\frac{1}{\eta\varepsilon^2}\right) = \tilde{\Theta}\left(\frac{1}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu\varepsilon^3}\right)
\end{aligned}$$

Therefore, after $\tilde{\Theta}\left(\frac{1}{\varepsilon^4} + \frac{\sqrt{1-\mu}}{\mu\varepsilon^3}\right)$ iterations at least half of the points are ε-FOSP. $\square$

B Convergence to ε-SOSP

By rescaling we can assume that $\varepsilon \le 1$. Recall that $\alpha = 1$ when Assumption C holds and $\alpha = d$ otherwise. We introduce the following auxiliary notation:

Definition B.1 (Step sizes)

$$\max \eta \textit{ for SGD} \quad \eta_\sigma = \frac{\varepsilon^2}{L(1+d\sigma^2)} + \min\left(\frac{\varepsilon^2}{L(1+\sigma^2)}, \frac{\sqrt{\rho\varepsilon}}{\tilde{\ell}^2}\right) = \tilde{O}\left(\frac{\varepsilon^2}{\alpha}\right)$$

max η for compressed SGD:

$$\textit{For a general compressor:} \quad \eta_\mu = \min\left(\frac{\mu\varepsilon}{\sqrt{1-\mu}L\sigma}, \frac{\mu^2\sqrt{\varepsilon}}{(1-\mu)L^2 d}\right) = \tilde{O}\left(\min\left(\frac{\mu\varepsilon}{\sqrt{1-\mu}}, \frac{\mu^2\sqrt{\varepsilon}}{(1-\mu)d}\right)\right)$$

$$\textit{For a linear compressor:} \quad \eta_\mu = \min\left(\frac{\mu\varepsilon}{\sqrt{1-\mu}L\sigma}, \frac{\mu^2\sqrt{\rho\varepsilon}}{(1-\mu)L^2}\right) = \tilde{O}\left(\min\left(\frac{\mu\varepsilon}{\sqrt{1-\mu}}, \frac{\mu^2\sqrt{\varepsilon}}{1-\mu}\right)\right)$$

Intuitively, selecting step size $\eta \le \eta_\sigma$ suffices to show convergence of SGD [Jin et al., 2021]. In addition, selecting $\eta \le \eta_\mu$ allows us to extend the results to compressed SGD.

Definition B.2 *Our choice of parameters is the following ($c_\eta, c_\mathcal{I}, c_\mathcal{R}, c_\mathcal{F}, c_r$ hide polylogarithmic dependence on all parameters, the conditions on them will be specified later):*

$$\begin{aligned}
\textit{Step size} \quad & \eta = c_\eta \min(\eta_\sigma, \eta_\mu) \\
\textit{Iterations required for escaping} \quad & \mathcal{I} = c_\mathcal{I} \frac{1}{\eta\sqrt{\rho\varepsilon}} \\
\textit{Escaping radius} \quad & \mathcal{R} = c_\mathcal{R} \sqrt{\frac{\varepsilon}{\rho}} \\
\textit{Objective change after escaping} \quad & \mathcal{F} = c_\mathcal{F} \sqrt{\frac{\varepsilon^3}{\rho}} \\
\textit{Noise standard deviation} \quad & r = c_r \frac{\varepsilon}{\sqrt{L\eta}}
\end{aligned} \tag{1}$$

Recall that $\chi^2 = \sigma^2 + r^2 = \sigma^2 + \frac{c_r\varepsilon^2}{L\eta}$ by Definition A.1 and $f_{\max} = f(\mathbf{x}_{t_0}) - f(\mathbf{x}^*)$. We will show that after $\mathcal{I}$ iterations the objective decreases by $\mathcal{F}$. Therefore, the objective decreases on average by $\frac{\mathcal{F}}{\mathcal{I}} = \tilde{\Omega}(\varepsilon^2\eta)$ per iteration resulting in $\tilde{O}\left(\frac{f_{\max}}{\varepsilon^2\eta}\right)$ iterations overall. See Table 1 for the number of iterations and total communication in various settings.

Intuitively, the motivation for this choice of parameters is the following. Let $\mathbf{x}$ be a point such that $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < -\sqrt{\rho\varepsilon}$ and $\|\nabla f(\mathbf{x})\| = 0$.

- Our analysis happens inside $B(\mathbf{x}, \mathcal{R})$, and we want $\lambda_{\min}(\nabla^2 f(\mathbf{z})) < -\frac{\sqrt{\rho\varepsilon}}{2}$ for all $\mathbf{z} \in B(\mathbf{x}, \mathcal{R})$. By the Hessian-Lipschitz property, for $\mathbf{z} \in B(\mathbf{x}, \mathcal{R})$ we have $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{z})\| \le \rho\mathcal{R}$. To have $\rho\mathcal{R} \le \frac{\sqrt{\rho\varepsilon}}{2}$, we choose $\mathcal{R} \le \frac{1}{2}\sqrt{\frac{\varepsilon}{\rho}}$.

- Let $-\gamma$ be the smallest eigenvalue and $\mathbf{v}_1$ be the corresponding eigenvector of $\nabla f^2(\mathbf{x})$. Assume that our function is quadratic and, after adding noise, the projection on $\mathbf{v}_1$ is $\tilde{\Theta}(1)$ (it is actually polynomial or reverse-polynomial on all parameters, which doesn't change the idea).

Table 2: Convergence to ε-SOSP for full SGD and for constant-size compression (the choice of parameters is not optimal; see Table 1 for the optimal choice). For any choice of μ and η: $T = \tilde{O}(1/\eta\varepsilon^2)$, $\mathcal{R} = \tilde{O}(\sqrt{\varepsilon})$, $\mathcal{F} = \tilde{O}(\sqrt{\varepsilon^3})$.

Settings	μ	η	$\mathcal{I}$	r
Uncompressed Lipschitz ∇F	0	$\tilde{O}\left(\varepsilon^2\right)$	$\tilde{O}\left(\varepsilon^{3/2}\right)$	$\tilde{O}\left(1\right)$
Compressed Lipschitz ∇F	$\frac{1}{d}$	$\tilde{O}\left(\min\left(\varepsilon^2, \frac{\varepsilon}{d}, \frac{\sqrt{\varepsilon}}{d^3}\right)\right)$	$\tilde{O}\left(\frac{1}{\eta\sqrt{\varepsilon}}\right)$	$\tilde{O}\left(\frac{\varepsilon}{\sqrt{\eta}}\right)$
RandomK Lipschitz ∇F	$\frac{1}{d}$	$\tilde{O}\left(\min\left(\varepsilon^2, \frac{\varepsilon}{d}, \frac{\sqrt{\varepsilon}}{d^2}\right)\right)$	$\tilde{O}\left(\frac{1}{\eta\sqrt{\varepsilon}}\right)$	$\tilde{O}\left(\frac{\varepsilon}{\sqrt{\eta}}\right)$
Uncompressed non-Lipschitz ∇F	0	$\tilde{O}\left(\frac{\varepsilon^2}{d}\right)$	$\tilde{O}\left(d\varepsilon^{3/2}\right)$	$\tilde{O}\left(\sqrt{d}\right)$
Compressed non-Lipschitz ∇F	$\frac{1}{d}$	$\tilde{O}\left(\min\left(\frac{\varepsilon^2}{d}, \frac{\sqrt{\varepsilon}}{d^3}\right)\right)$	$\tilde{O}\left(\frac{1}{\eta\sqrt{\varepsilon}}\right)$	$\tilde{O}\left(\frac{\varepsilon}{\sqrt{\eta}}\right)$
RandomK non-Lipschitz ∇F	$\frac{1}{d}$	$\tilde{O}\left(\min\left(\frac{\varepsilon^2}{d}, \frac{\sqrt{\varepsilon}}{d^2}\right)\right)$	$\tilde{O}\left(\frac{1}{\eta\sqrt{\varepsilon}}\right)$	$\tilde{O}\left(\frac{\varepsilon}{\sqrt{\eta}}\right)$

Then after t iterations, this projection increases by the factor of $(1+\eta\gamma)^t$. For every $1/\eta\gamma$ iterations, the projection increases approximately by the factor of e. Therefore, to reach $\mathcal{R}$ starting from $\Theta(1)$, we need $\tilde{O}(\frac{1}{\eta\gamma})$ iterations, which is at most $\tilde{O}(\frac{1}{\eta\sqrt{\rho\varepsilon}})$

- In a certain sense, the best improvement we can hope to achieve is by moving from $\mathbf{x}$ to $\mathbf{x} + \mathcal{R}\mathbf{v}_1$. If $\nabla f(\mathbf{x}) = 0$ and the objective is quadratic in direction $\mathbf{v}_1$ with eigenvalue γ, the objective decreases by $\gamma\mathcal{R}^2 = \Omega(\sqrt{\frac{\varepsilon^3}{\rho}})$, which motivates the choice of $\mathcal{F}$.
- Bound on r arises from the fact that $\chi^2 \approx r^2$ and that we want to bound the last term in Lemma B.4 with $\mathcal{F}$.

We formalize the first item in the following proposition:

Proposition B.3 *Let $\mathbf{x}$ be a point such that $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < -\sqrt{\rho\varepsilon}$. Then for any $\mathbf{z} \in B(\mathbf{x}, \mathcal{R})$, $\lambda_{\min}(\nabla^2 f(\mathbf{z})) < -\sqrt{\rho\varepsilon}/2$.*

B.1 Proof outline

Our proof is mainly based on the ideas from Jin et al. [2021]. We first introduce "Improve or localize" lemma (Lemma B.4): if after the limited number of iterations the objective doesn't sufficiently improve, we conclude that we didn't move far from the original point. Similarly to Jin et al. [2021], we introduce a notion of coupling sequences: two gradient descent sequences having the same distribution such that, as long as we start from a saddle point, at least one of these sequences escapes, and therefore its objective improves. Since distributions of these sequences match distribution of sequence generated by gradient descent, we conclude that the algorithm sufficiently improves the objective.

Our analysis differs from Jin et al. [2021] in several ways. The first difference is that, aside from $\{\mathbf{x}_{t_0+t}\}$, our equations use another sequence $\{\mathbf{y}_{t_0+t}\}$ ($\mathbf{x}_{t_0+t}$ mainly participate as arguments of $\nabla f(\cdot)$, while $\mathbf{y}_{t_0+t}$ participate as argument of $f(\cdot)$ and in distances). This leads to the following challenge: if some relation holds for $\mathbf{y}_{t_0+t}$, it doesn't necessary holds for $\mathbf{x}_{t_0+t}$. For example, if we have a bound on $\|\mathbf{y}_{t_0+t} - \mathbf{y}'_{t_0+t}\|$, we don't necessarily have a bound on $\|\mathbf{x}_{t_0+t} - \mathbf{x}'_{t_0+t}\|$, and it needs to be established separately.

Another difference is that, for a general compressor, we have to split our analysis into two parts: large gradient case and small gradient case. When our initial gradient is large, then we either escape the saddle points or the nearby gradients are also large, and by Lemma A.7 the objective improves (see details in Lemma B.7). If the gradient is small, we use "Improve or localize" Lemma as described above. In the latter case, similarly to Jin et al. [2021], we have to bound errors which arise from the fact that the function is not quadratic and gradients are not deterministic (see Definition B.11). However, we have an additional error term stemming from gradient compression (see Definition B.11); to bound this term (see Lemma B.16), we need bounded $\|\mathbf{e}_{t_0+t}\|$, and for that we use our assumptions that gradients are small.

B.2 Improve or localize

We first show that, if gradient descent moves far enough from the initial point, then function value sufficiently decreases. The following lemma considers the general case, while Corollary B.5 considers the simplified form, obtained by substituting parameters from Definition B.2.

Lemma B.4 (Improve or localize) *Let f satisfy Assumptions A and B and let $\mathbf{y}_{t_0+t}$ and χ be defined as in Definition A.1. If t_0 is an iteration of Algorithm 1 such that $\mathbf{e}_{t_0} = 0$, then using notation $\mathbb{E}_t[\cdot]$ from Definition A.4, for $\eta < \frac{1}{4L}\min(\frac{\mu}{\sqrt{1-\mu}}, 1)$ we have*

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \frac{\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right]}{8\eta t} - \eta^2\chi^2 t\left(L + \frac{2(1-\mu)L^2\eta}{\mu^2}\right) - \eta\chi^2$$

Proof : Let $\psi_t = \zeta_t + \xi_{t_0+t}$. By Proposition A.3, $\mathbf{y}_{i+1} = \mathbf{y}_{t_0+i} - \eta(\nabla f(\mathbf{x}_{t_0+i}) + \psi_i)$. Since noises are independent:

$$\mathbb{E}_{t_0}\left[\|\sum_{i=0}^{t-1}\psi_{t_0+i}\|^2\right] = \sum_{i=0}^{t-1}\mathbb{E}_{t_0}\left[\|\psi_{t_0+i}\|^2\right] = \sum_{\tau=0}^{t-1}\chi^2 = t\chi^2$$

By Proposition A.3:

$$\begin{aligned}
\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] &= \eta^2\mathbb{E}_{t_0}\left[\|\sum_{i=0}^{t-1}(\nabla f(\mathbf{x}_{t_0+i}) + \psi_{t_0+i})\|^2\right] \\
&\leq 2\eta^2\mathbb{E}_{t_0}\left[\|\sum_{i=0}^{t-1}\nabla f(\mathbf{x}_{t_0+i})\|^2 + \|\sum_{i=0}^{t-1}\psi_{t_0+i}\|^2\right] \\
&\leq 2\eta^2 t\sum_{i=0}^{t-1}\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+i})\|^2\right] + 2\eta^2\chi^2 t
\end{aligned}$$

Since $\eta < \frac{1}{4L}\min(\frac{\mu}{\sqrt{1-\mu}}, 1)$, by Corollary A.8:

$$\begin{aligned}
\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] &\leq 2\eta^2 t\left(\frac{4(f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right])}{\eta} + \eta\chi^2 t\left(2L + \frac{8(1-\mu)L^2\eta}{\mu^2}\right)\right) + 2\eta^2\chi^2 t \\
&\leq 2\eta t\left(4(f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right]) + \eta^2\chi^2 t\left(4L + \frac{8(1-\mu)L^2\eta}{\mu^2}\right) + 4\eta\chi^2\right)
\end{aligned}$$

After regrouping the terms, we have:

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \frac{\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right]}{8\eta t} - \eta^2\chi^2 t\left(L + \frac{2(1-\mu)L^2\eta}{\mu^2}\right) - \eta\chi^2,$$

□

Corollary B.5 *Let t_0 be an iteration of Algorithm 1 such that $\mathbf{e}_{t_0} = 0$. Under Assumptions A and B, for $\mathcal{F}, \mathcal{I}$ chosen as specified in Definition B.2, for any $t \leq \mathcal{I}$ we have:*

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \frac{\sqrt{\rho\varepsilon}}{8c_{\mathcal{I}}}\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] - \mathcal{F}$$

Proof : By Lemma B.4, the first term on the right-hand side stems from $t \leq \mathcal{I} = \frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}$. With our choice of parameters, we can bound negative terms with $\mathcal{F}$ (recall that $\mathcal{F} = c_{\mathcal{F}}\sqrt{\frac{\varepsilon^3}{\rho}}$).

Bounding $\eta\chi^2$.

$$\eta\chi^2 = \eta\sigma^2 + \eta r^2 \leq c_\eta\frac{\varepsilon^2}{L} + c_r^2\frac{\varepsilon^2}{L} = (c_\eta + c_r^2)\frac{\sqrt{\varepsilon^3}}{\sqrt{\rho}} \cdot \frac{\sqrt{\rho\varepsilon}}{L} \leq (c_\eta + c_r^2)\frac{\sqrt{\varepsilon^3}}{\sqrt{\rho}},$$

where we use that $\sqrt{\rho\varepsilon} \leq L$, since otherwise all ε-FOSP are ε-SOSP.

Bounding $\eta^2\chi^2 tL$. Since $\mathcal{I} = c_{\mathcal{I}}\frac{1}{\eta\sqrt{\rho\varepsilon}}$ and $t \leq \mathcal{I}$:

$$\eta^2\chi^2 tL \leq \frac{\eta\chi^2 L}{\sqrt{\rho\varepsilon}} \leq \eta\chi^2,$$

and we use the estimation above.

Bounding $\eta^2\chi^2 t \cdot \frac{2(1-\mu)L^2\eta}{\mu^2}$.

$$\begin{aligned}
\frac{\eta^3\chi^2 t(1-\mu)L^2}{\mu^2} &\leq \frac{c_{\mathcal{I}}\eta^2\chi^2(1-\mu)L^2}{\mu^2\sqrt{\rho\varepsilon}} && (t \leq \mathcal{I} = \frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}) \\
&\leq \frac{c_{\mathcal{I}}\eta^2(1-\mu)L^2\left(\sigma^2 + \frac{c_r\varepsilon^2}{L\eta}\right)}{\mu^2\sqrt{\rho\varepsilon}} && (\chi^2 = \sigma^2 + \frac{c_r\varepsilon^2}{L\eta}) \\
&\leq \frac{c_{\mathcal{I}}(1-\mu)}{\mu^2\sqrt{\rho\varepsilon}}\left(\eta_\mu^2 L^2\sigma^2 + c_r\eta_\mu L\varepsilon^2\right) && (\eta \leq \eta_\mu) \\
&\leq 2c_{\mathcal{I}}c_\eta\frac{\sqrt{\varepsilon^3}}{\sqrt{\rho}} && (\eta_\mu \leq \frac{\mu\varepsilon}{\sqrt{1-\mu}L\sigma} \text{ and } \eta_\mu \leq \frac{\mu^2\sqrt{\varepsilon}}{1-\mu})
\end{aligned}$$

To guarantee that the sum of these terms is at most $\mathcal{F}$, it suffices to select parameters so that $c_\eta + c_r^2 + c_{\mathcal{I}}c_\eta \leq c_{\mathcal{F}}/2$. □

Corollary B.6 *Let t_0 be an iteration of Algorithm 1 such that $\mathbf{e}_{t_0} = 0$. Under Assumptions A and B, for $\mathcal{F}, \mathcal{R}, \mathcal{I}$ chosen as specified in Definition B.2, if there exists $t \in [0, \mathcal{I}]$ such that $\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] > \mathcal{R}^2$, then $f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \mathcal{F}$.*

Proof : By Lemma B.4, since $\mathcal{R} = c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}}$ and $\mathcal{F} = c_{\mathcal{F}}\sqrt{\frac{\varepsilon^3}{\rho}}$:

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \frac{\sqrt{\rho\varepsilon}\mathcal{R}^2}{8c_{\mathcal{I}}} - \mathcal{F} = \frac{c_{\mathcal{R}}^2\varepsilon\eta\sqrt{\rho\varepsilon}}{8c_{\mathcal{I}}\eta\rho} - \mathcal{F} = \left(\frac{c_{\mathcal{R}}^2}{8c_{\mathcal{I}}c_{\mathcal{F}}} - 1\right)\mathcal{F} \geq \mathcal{F},$$

where the last inequality holds when $16c_{\mathcal{I}}c_{\mathcal{F}} \leq c_{\mathcal{R}}^2$. □

B.3 Large gradient case: $\|\nabla f(\mathbf{x}_{t_0})\| \geq 4L\mathcal{R}$

In this section, we consider the case when the gradient is large, and therefore we can make sufficient progress simply by the Compressed Descent Lemma. Note that the results from this section are only required when the compressor is not linear.

Lemma B.7 (Large gradient case) *Let t_0 be an iteration of Algorithm 1 such that $\mathbf{e}_{t_0} = 0$. Under Assumptions A and B, for $\mathcal{F}, \mathcal{R}, \mathcal{I}$ chosen as specified in Definition B.2, if $\|\nabla f(\mathbf{x}_{t_0})\| > 4L\mathcal{R}$, then after at most $\mathcal{I}$ iterations the objective decreases by $\mathcal{F}$.*

Proof : Using notation $\mathbb{E}_t[\cdot]$ from Definition A.4, if there exists $t \leq \mathcal{I}$ such that $\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] > \mathcal{R}^2$, then by Corollary B.6, the objective decreases by at least $\mathcal{F}$.

Consider the case when $\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] \leq \mathcal{R}^2$ for all t. First, to bound the error term, we show by induction that $\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+t})\|^2\right] \leq 4\|\nabla f(\mathbf{x}_{t_0})\|^2$ for all $t \leq \mathcal{I}$.

$$\begin{aligned}
\|\nabla f(\mathbf{x}_{t_0+t})\|^2 &= \|\nabla f(\mathbf{x}_{t_0}) - (\nabla f(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0+t}))\|^2 \\
&\leq 2\|\nabla f(\mathbf{x}_{t_0})\|^2 + 2\|\nabla f(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0+t})\|^2 && (\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)) \\
&\leq 2\|\nabla f(\mathbf{x}_{t_0})\|^2 + 2L^2\|\mathbf{x}_{t_0} - \mathbf{x}_{t_0+t}\|^2 && \text{(Smoothness)} \\
&\leq 2\|\nabla f(\mathbf{x}_{t_0})\|^2 + 4L^2\|\mathbf{y}_{t_0} - \mathbf{y}_{t_0+t}\|^2 + 4L^2\|\mathbf{y}_{t_0+t} - \mathbf{x}_{t_0+t}\|^2 && (\text{Same inequality and } \mathbf{x}_{t_0} = \mathbf{y}_{t_0}) \\
&\leq 2\|\nabla f(\mathbf{x}_{t_0})\|^2 + 4L^2\|\mathbf{y}_{t_0} - \mathbf{y}_{t_0+t}\|^2 + 4L^2\eta^2\|\mathbf{e}_{t_0+t}\|^2 && (\text{Definition A.2 of } \mathbf{y}_{t_0+t})
\end{aligned}$$

By Lemma A.5 and the induction hypothesis, we have:

$$\mathbb{E}_{t_0}\left[\|\mathbf{e}_{t_0+t}\|^2\right] \leq \frac{4(1-\mu)}{\mu^2}\left(\max_{\tau=0}^{t-1}\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+\tau})\|^2\right] + \chi^2\right) \leq \frac{4(1-\mu)}{\mu^2}\left(4\|\nabla f(\mathbf{x}_{t_0})\|^2 + \chi^2\right),$$

and therefore for η chosen as in Definition B.2, $L^2\eta^2\mathbb{E}_{t_0}\left[\|\mathbf{e}_{t_0+t}\|^2\right] \leq \frac{\|\nabla f(\mathbf{x}_{t_0})\|^2}{4}$. By taking the expectation in the equation above, we have:

$$\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+t})\|^2\right] \leq 2\|\nabla f(\mathbf{x}_{t_0})\|^2 + 4L^2\mathcal{R}^2 + \frac{\|\nabla f(\mathbf{x}_{t_0})\|^2}{4} \leq 4\|\nabla f(\mathbf{x}_{t_0})\|^2$$

Given the bound on $\|\mathbf{e}_{t_0+t}\|$, we can give a lower bound on gradient norm:

$$\begin{aligned}
\|\nabla f(\mathbf{x}_{t_0+t})\|^2 &= \|\nabla f(\mathbf{x}_{t_0}) - (\nabla f(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0+t}))\|^2 \\
&\geq \|\nabla f(\mathbf{x}_{t_0})\|^2 + \|\nabla f(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0+t})\|^2 - 2\|\nabla f(\mathbf{x}_{t_0})\| \cdot \|\nabla f(\mathbf{x}_{t_0}) - \nabla f(\mathbf{x}_{t_0+t})\| \\
&\geq \|\nabla f(\mathbf{x}_{t_0})\|(\|\nabla f(\mathbf{x}_{t_0})\| - 2\|\nabla f(\mathbf{y}_{t_0}) - \nabla f(\mathbf{y}_{t_0+t})\| - 2\|\nabla f(\mathbf{y}_{t_0+t}) - \nabla f(\mathbf{x}_{t_0+t})\|)
\end{aligned}$$

By taking expectations and using the fact that $\mathbb{E}_{t_0}\left[\|x\|\right] \leq \sqrt{\mathbb{E}_{t_0}\left[\|x\|^2\right]}$ and bound on $\|\mathbf{e}_{t_0+t}\|$, we have:

$$\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+t})\|^2\right] \geq \nabla f(\mathbf{x}_{t_0})(\|\nabla f(\mathbf{x}_{t_0})\| - 2L\mathcal{R} - \frac{\|\nabla f(\mathbf{x}_{t_0})\|}{4}) \geq 4L^2\mathcal{R}^2$$

By Lemma A.7, we know:

$$\sum_{\tau=0}^{\mathcal{I}} \mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+\tau})\|^2\right] \le \frac{4(f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{\mathcal{I}})\right])}{\eta} + \eta\chi^2\mathcal{I}\left(2L + \frac{8(1-\mu)L^2\eta}{\mu^2}\right)$$

Therefore:

$$\begin{aligned}
f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+\mathcal{I}})\right] &\ge \frac{\eta\mathcal{I}}{4}\left(4L^2\mathcal{R}^2 - \eta\chi^2(2L + \frac{8(1-\mu)L^2\eta}{\mu^2})\right) \\
&\ge \eta\mathcal{I}L^2\mathcal{R}^2 - \mathcal{F} && \text{(See proof of Corollary B.5)} \\
&\ge \frac{c_{\mathcal{I}}\eta}{\eta\sqrt{\rho\varepsilon}}\frac{c_{\mathcal{R}}^2 L^2\varepsilon}{\rho} - \mathcal{F} && (\mathcal{R} = c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}} \text{ and } \mathcal{I} = \frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}) \\
&\ge \frac{c_{\mathcal{I}}}{\sqrt{\rho\varepsilon}} c_{\mathcal{R}}^2\varepsilon^2 - \mathcal{F} && (\text{since } L \ge \sqrt{\rho\varepsilon}) \\
&\ge \mathcal{F},
\end{aligned}$$

where the last inequality holds when $c_{\mathcal{I}}c_{\mathcal{R}}^2 \ge 2c_{\mathcal{F}}$. □

B.4 Small Gradient Case: $\|\nabla f(\mathbf{x}_{t_0})\| < 4L\mathcal{R}$

Coupling Sequences

Let $H = \nabla^2 f(\mathbf{x}_{t_0})$, then $g(\mathbf{x}) = \mathbf{x}^\top H\mathbf{x}$ is a quadratic approximation of f in the vicinity of $\mathbf{x}_{t_0}$. Let $-\gamma$ be the smallest eigenvalue of H and $\mathbf{v}_1$ be the corresponding eigenvector. Then we construct *coupling sequences* $\mathbf{x}_{t_0+t}$ and $\mathbf{x}'_{t_0+t}$ in the following way: $\mathbf{x}_{t_0+t}$ is the sequence from Algorithm 1; $\mathbf{x}'_{t_0+t}$ has the same stochastic randomness θ as $\mathbf{x}_{t_0+t}$, and its artificial noise ξ'_{t_0+t} is the same as ξ_{t_0+t} with exception of the coordinate corresponding to $\mathbf{v}_1$, which has an opposite sign.

Definition B.8 (Coupling sequences) *For iteration t_0, given $\mathbf{x}_{t_0}$ and $\mathbf{e}_{t_0}$, the coupling sequences are defined as follows (note the definition of ξ'_{t_0+t}):*

$$\begin{aligned}
& & \mathbf{e}'_{t_0} &= \mathbf{e}_{t_0} \\
\xi_{t_0+t} &\sim \mathcal{N}(0, r^2) & \xi'_{t_0+t} &= \xi_{t_0+t} - 2\langle \mathbf{v}_1, \xi_{t_0+t}\rangle \mathbf{v}_1 \\
\theta_{t_0+t} &\sim \mathcal{D}, \quad \tilde{\theta}_{t_0+t} \sim \tilde{\mathcal{D}} & \theta'_{t_0+t} &= \theta_{t_0+t}, \quad \tilde{\theta}'_{t_0+t} = \tilde{\theta}_{t_0+t} \\
\mathbf{g}_{t_0+t} &= C(\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) + \xi_{t_0+t} + \mathbf{e}_{t_0+t}, \tilde{\theta}_{t_0+t}) & \mathbf{g}'_{t_0+t} &= C(\nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t}) + \xi'_{t_0+t} + \mathbf{e}'_{t_0+t}, \tilde{\theta}_{t_0+t}) \\
\mathbf{y}_{t_0+t} &= \mathbf{x}_{t_0+t} - \eta\mathbf{e}_{t_0+t} & \mathbf{y}'_{t_0+t} &= \mathbf{x}'_{t_0+t} - \eta\mathbf{e}'_{t_0+t} \\
\mathbf{x}_{t+1} &= \mathbf{x}_{t_0+t} - \eta\mathbf{g}_t & \mathbf{x}'_{t+1} &= \mathbf{x}'_{t_0+t} - \eta\mathbf{g}'_t \\
\mathbf{e}_{t+1} &= \nabla F(\mathbf{x}t_0 + t, \theta_{t_0+t}) + \xi_{t_0+t} + \mathbf{e}_{t_0+t} - \mathbf{g}_{t_0+t} & \mathbf{e}'_{t+1} &= \nabla F(x'_t, \theta_{t_0+t}) + \xi'_{t_0+t} + \mathbf{e}'_{t_0+t} - \mathbf{g}'_{t_0+t}
\end{aligned}$$

A notable fact is that both sequences correspond to the same distribution.

Proposition B.9 *For any t_0 and t, $\mathbf{x}_{t_0+t}$ and $\mathbf{y}_{t_0+t}$ from Definition B.8 have the same distribution as $\mathbf{x}'_{t_0+t}$ and $\mathbf{y}'_{t_0+t}$.*

Proof : By definition of $\mathbf{y}_{t_0+t}$ and $\mathbf{y}'_{t_0+t}$, it suffices show that $\mathbf{x}_{t_0+t}$ and $\mathbf{e}_{t_0+t}$ have the same distributions as $\mathbf{x}'_{t_0+t}$ and $\mathbf{e}'_{t_0+t}$. Proof by Induction with trivial base case $\mathbf{y}_{t_0} = \mathbf{y}'_{t_0} = \mathbf{x}_{t_0} - \eta\mathbf{e}_{t_0}$.

We want to show that if the statement holds for t, then it holds for $t+1$. To show that $\mathbf{x}_{t_0+t+1}$ and $\mathbf{x}'_{t_0+t+1}$ habe the same distribution it remains to show that $\mathbf{g}_t$ and $\mathbf{g}'_t$ have the same distribution:

- Since $\mathbf{x}_{t_0+t}$ and $\mathbf{x}'_{t_0+t}$ have the same distribution, $\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t})$ and $\nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})$ have the same distribution.
- Since $\mathcal{N}(0, r^2)$ is symmetric and ξ'_{t_0+t} is the same as ξ_{t_0+t} with exception of one coordinate, which has an opposite sign, ξ_{t_0+t} and ξ'_{t_0+t} have the same distribution.
- $\mathbf{e}_{t_0+t}$ and $\mathbf{e}'_{t_0+t}$ have the same distribution.

Similarly, $\mathbf{e}_{t+1}$ has the same distribution as $\mathbf{e}'_{t+1}$, since $\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t})$, ξ_{t_0+t}, $\mathbf{e}_{t_0+t}$ and $\mathbf{g}_t$ have the same distribution as $\nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})$, ξ'_{t_0+t}, $\mathbf{e}'_{t_0+t}$ and $\mathbf{g}'_t$. $\square$

Since our sequences have the same distribution, we have $\mathbb{E}_{t_0}\left[f(\mathbf{x}_{t_0+t})\right] = \mathbb{E}_{t_0}\left[f(\mathbf{x}'_{t_0+t})\right]$. We want to show that in a few iterations $\mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0+t}$ becomes sufficiently large and, therefore, at least one of $\mathbf{y}_{t_0+t}$ and $\mathbf{y}'_{t_0+t}$ is far from $\mathbf{x}_{t_0}$. By applying Corollary B.6 we will show that the objective sufficiently decreases.

Difference Between Coupling Sequences

In order to capture the difference between the coupling sequences, we introduce the following notation:

Definition B.10 (Difference between sequences) *Using notation from Definition B.8, we introduce differences between the sequences:*

$$\hat{\mathbf{x}}_{t_0+t} = \mathbf{x}'_{t_0+t} - \mathbf{x}_{t_0+t} \qquad \hat{\mathbf{e}}_{t_0+t} = \mathbf{e}'_{t_0+t} - \mathbf{e}_{t_0+t} \qquad \hat{\zeta}_t = \zeta'_{t_0+t} - \zeta_{t_0+t}$$
$$\hat{\xi}_{t_0+t} = \xi'_{t_0+t} - \xi_{t_0+t} \qquad \hat{\mathbf{y}}_{t_0+t} = \mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0+t}$$

Definition B.11 (Error terms) *Let $\delta_i = \int_0^1 \nabla^2 f(\alpha \mathbf{x}'_{t_0+i} + (1-\alpha)\mathbf{x}_{t_0+i}) d\alpha - H$. Then*

$$\Delta_t = \eta \sum_{i=0}^{t-1} (I - \eta H)^{t-i-1} \delta_i \hat{\mathbf{x}}_{t_0+i}$$
$$\mathcal{E}_t = \eta \sum_{i=0}^{t-1} (I - \eta H)^{t-i-1} (\hat{\mathbf{e}}_{t_0+i} - \hat{\mathbf{e}}_{t_0+i+1})$$
$$Z_t = \eta \sum_{i=0}^{t-1} (I - \eta H)^{t-i-1} \hat{\zeta}_{t_0+i}$$
$$\Xi_t = \eta \sum_{i=0}^{t-1} (I - \eta H)^{t-i-1} \hat{\xi}_{t_0+i},$$

Proposition B.12 *For any t: $\hat{\mathbf{x}}_{t_0+t} = -(\Delta_t + \mathcal{E}_t + Z_t + \Xi_t)$.*

In the simplest case, the objective is quadratic and we have access to an uncompressed deterministic gradient. When it's not the case, the introduced terms show how the actual algorithm behavior is different:

- Δ_t corresponds to quadratic approximation error.
- $\mathcal{E}_t$ corresponds to compression error.
- Z_t corresponds to difference arising from SGD noise.
- Ξ_t corresponds to difference arising from artificial noise.

Intuitively, Ξ_t is a good term, and other terms are negligible ($\|\Delta_t + \mathcal{E}_t + Z_t\| < \frac{1}{2}\|\Xi_t\|$).
Proof :

$$
\begin{aligned}
&\hat{\mathbf{x}}_{t_0+t+1} \\
&= \mathbf{x}'_{t_0+t+1} - \mathbf{x}_{t_0+t+1} \\
&= \mathbf{y}'_{t_0+t+1} + \eta\mathbf{e}'_{t_0+t+1} - (\mathbf{y}_{t_0+t+1} + \eta\mathbf{e}_{t_0+t+1}) && \text{(Def. of } \mathbf{y}_{t_0+t} \text{ and } \mathbf{y}'_{t_0+t}) \\
&= \eta\hat{\mathbf{e}}_{t_0+t+1} + (\mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0+t}) \\
&\quad - \eta\left((\nabla f(x'_{t_0+t}) - \nabla f(x_{t_0+t})) + (\zeta'_{t_0+t} - \zeta_{t_0+t}) + (\xi'_{t_0+t} - \xi_{t_0+t})\right) && \text{(Upd. equation for } \mathbf{y}_{t_0+t}) \\
&= \eta(\hat{\mathbf{e}}_{t+1} - \hat{\mathbf{e}}_{t_0+t}) + \hat{\mathbf{x}}_{t_0+t} - \eta\left((\delta_t + H)\hat{\mathbf{x}}_{t_0+t} + \hat{\zeta}_{t_0+t} + \hat{\xi}_{t_0+t}\right) && \text{(Def. of } \delta_{t_0+t} \text{ and } \mathbf{y}_{t_0+t}) \\
&= \eta(\hat{\mathbf{e}}_{t+1} - \hat{\mathbf{e}}_{t_0+t}) + (I - \eta H)\hat{\mathbf{x}}_{t_0+t} - \eta\left(\delta_t\hat{\mathbf{x}}_{t_0+t} + \hat{\zeta}_{t_0+t} + \hat{\xi}_{t_0+t}\right) \\
&= (I - \eta H)\hat{\mathbf{x}}_{t_0+t} - \eta\left(\delta_t\hat{\mathbf{x}}_{t_0+t} + (\hat{\mathbf{e}}_{t_0+t} - \hat{\mathbf{e}}_{t_0+t+1}) + \hat{\zeta}_{t_0+t} + \hat{\xi}_{t_0+t}\right)
\end{aligned}
$$

Using telescoping, we get the required expression. □

Since $\hat{\mathbf{y}}_{t_0+t} = \hat{\mathbf{x}}_{t_0+t} - \eta\hat{\mathbf{e}}_{t_0+t}$, we have:

$$\hat{\mathbf{x}}_{t_0+t} = -(\Delta_t + \mathcal{E}_t + Z_t + \Xi_t) \iff \hat{\mathbf{y}}_{t_0+t} = -(\Delta_t + (\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}) + Z_t + \Xi_t),$$

and we'll use $\|\hat{\mathbf{y}}_{t_0+t}\|$ in Corollary B.6.

Bounding Accumulated Compression Error

Compared to SGD analysis, an additional term $\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}$ appears. This term corresponds to accumulated error arising from compression, and we have to bound it. Motivated by Jin et al. [2021], we introduce the following quantity:

Definition B.13 *Standard deviation of sum of random variables with standard deviations* $(1+\eta\gamma)^i$, $i = 0, \ldots, t-1$, *is*

$$\beta_t = \sqrt{\sum_{i=0}^{t-1}(1+\eta\gamma)^{2i}}$$

Proposition B.14 (Jin et al. [2021], Lemma 29) *If* $\eta\gamma \in [0,1]$, *then for all* t: $\beta_t \leq \frac{(1+\eta\gamma)^t}{\sqrt{2\eta\gamma}}$, *and for all* $t \geq \frac{2}{\eta\gamma}$: $\beta_t \geq \frac{(1+\eta\gamma)^t}{\sqrt{6\eta\gamma}}$.

Proposition B.15 *For any* $t \leq \mathcal{I}$, *where* $\mathcal{I}$ *is defined in Definition B.2:*

$$\left(\sum_{i=0}^{t-1}(1+\eta\gamma)^{t-1-i}\right)^2 \leq \frac{c_{\mathcal{I}}\beta_t^2}{\eta\sqrt{\rho\varepsilon}}$$

Proof : By Cauchy-Schwarz inequality:

$$\left(\sum_{i=0}^{t-1}(1+\eta\gamma)^{t-1-i}\right)^2 \leq t\sum_{i=0}^{t-1}(1+\eta\gamma)^{2(t-1-i)} \leq \mathcal{I}\beta_t^2 = \frac{c_{\mathcal{I}}\beta_t^2}{\eta\sqrt{\rho\varepsilon}}$$

□

Lemma B.16 (Bounding accumulated compression error) *Let t_0 be an iteration such that $\mathbf{e}_{t_0} = 0$ in Algorithm 1. Under Assumptions A and B, let χ be as in Definition A.1, η and $\mathcal{R}$ as in Definition B.2, $\mathcal{E}_t$ and $\hat{\mathbf{e}}_{t_0+t}$ be as in Definition B.11, and β_t be as in Definition B.13 . Let $-\gamma \leq -\sqrt{\rho\varepsilon}/2$ be the smallest negative eigenvalue of $\nabla^2 f(\mathbf{x}_{t_0})$. If $\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] < \mathcal{R}^2$ for all $t \leq \mathcal{I}$ and $\|\nabla f(\mathbf{x}_{t_0})\| \leq 4L\mathcal{R}$, then using notation $\mathbb{E}_t\left[\cdot\right]$ from Definition A.4, for $t \leq \mathcal{I}$ we have:*

$$\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \leq \frac{20c_{\mathcal{I}}\eta^3(1-\mu)L^2\chi^2\beta_t^2}{\mu^2\sqrt{\rho\varepsilon}}$$

Proof : Expanding sum in $\mathcal{E}_t$ and using that $\hat{\mathbf{e}}_{t_0} = 0$:

$$\begin{aligned}
\mathcal{E}_t &= \eta\sum_{i=0}^{t-1}(I-\eta H)^{t-1-i}(\hat{\mathbf{e}}_{t_0+i} - \hat{\mathbf{e}}_{t_0+i+1}) && \text{(By Definition B.11)} \\
&= \eta(-\hat{\mathbf{e}}_{t_0+t} + \sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}((I-\eta H) - I)\hat{\mathbf{e}}_{t_0+i}) && \text{(By telescoping)} \\
&= -\eta\hat{\mathbf{e}}_{t_0+t} + \eta^2 H\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}\hat{\mathbf{e}}_{t_0+i}
\end{aligned}$$

We will now estimate $\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}^2_{t_0+t}\|\right]$. Since $-\gamma$ is the smallest negative eigenvalue of H, we have $\|I - \eta H\| \leq (1+\eta\gamma)$.

$$\begin{aligned}
&\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \\
&= \mathbb{E}_{t_0}\left[\|\eta^2 H\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}\hat{\mathbf{e}}_{t_0+i}\|^2\right] \\
&\leq \eta^4 L^2\mathbb{E}_{t_0}\left[(\sum_i(1+\eta\gamma)^{t-1-i}\|\hat{\mathbf{e}}_{t_0+i}\|)^2\right] && \text{(By L-smoothness, $\lambda_{\max}(H) \leq L$)} \\
&\leq 2\eta^4 L^2(\sum_i(1+\eta\gamma)^{t-1-i})^2\max_i\mathbb{E}_{t_0}\left[\|\hat{\mathbf{e}}_{t_0+i}\|^2\right] && (\mathbb{E}_{t_0}\left[ab\right] \leq \max(\mathbb{E}_{t_0}\left[a^2\right], \mathbb{E}_{t_0}\left[b^2\right])) \\
&\leq 2\eta^4 L^2 t(\sum_i(1+\eta\gamma)^{t-1-i})^2\max_i\mathbb{E}_{t_0}\left[\|\mathbf{e}'_{t_0+i} - \mathbf{e}_{t_0+i}\|^2\right] && \text{(By definition of $\hat{\mathbf{e}}_{t_0+i}$)} \\
&\leq 4\eta^4 L^2 t(\sum_i(1+\eta\gamma)^{t-1-i})^2\max_i\mathbb{E}_{t_0}\left[\|\mathbf{e}'_{t_0+i}\|^2 + \|\mathbf{e}_{t_0+i}\|^2\right] && \text{(By Cauchy-Schwarz)} \\
&\leq 8\eta^4 L^2 t(\sum_i(1+\eta\gamma)^{t-1-i})^2\max_i\mathbb{E}_{t_0}\left[\|\mathbf{e}_{t_0+i}\|^2\right] && \text{($\mathbf{e}_{t_0+i}$ and $\mathbf{e}'_{t_0+i}$ have the same distribution)}
\end{aligned}$$

Similarly to Lemma B.7, we can show that $\mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+i})\|^2\right] \le 40L^2\mathcal{R}^2$. By Lemma A.5:

$$\begin{aligned}
&\mathbb{E}_{t_0}\left[\|\mathbf{e}_{t_0+t}\|^2\right] \\
&\le \frac{4(1-\mu)}{\mu^2}(\max_i \mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+i})\|^2\right] + \chi^2) && \text{(By Lemma A.5)} \\
&\le \frac{4(1-\mu)}{\mu^2}(40L^2\mathcal{R}^2 + \chi^2) && \text{(By assumption } \mathbb{E}_{t_0}\left[\|\nabla f(\mathbf{x}_{t_0+i})\|^2\right] \le 40L^2\mathcal{R}^2\text{)} \\
&\le \frac{5(1-\mu)\chi^2}{\mu^2} && \text{(Selecting sufficiently small } c_{\mathcal{R}} \text{ in the definition of } \mathcal{R}\text{)}
\end{aligned}$$

Substituting this result into the inequality for $\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|$:

$$\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \le 4\eta^4 L^2 t \left(\sum_i (1+\eta\gamma)^{t-1-i}\right)^2 \frac{5(1-\mu)\chi^2}{\mu^2} \le \frac{20c_{\mathcal{I}}\eta^4(1-\mu)L^2\chi^2\beta_t^2}{\mu^2\eta\sqrt{\rho\varepsilon}},$$

where we bounded the series using Proposition B.15. □

Lemma B.17 (Bounding accumulated compression error for linear compressor) *Under conditions of Lemma B.16, additionally assume that the compressor is linear (Definition 2.2). When $\eta \le \eta_\sigma$, for $t \le \mathcal{I}$:*

$$\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \le \frac{9c_{\mathcal{I}}\eta^3(1-\mu)L^2\beta_t^2 r^2}{\mu^2 d\sqrt{\rho\varepsilon}}$$

Note that, compared with Lemma B.16, the bound is improved by the factor of d.
Proof :

$$\begin{aligned}
\hat{\mathbf{e}}_{t_0+t+1} &= \mathbf{e}_{t_0+t+1} - \mathbf{e}'_{t_0+t+1} \\
&= \nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) + \xi_{t_0+t} + \mathbf{e}_{t_0+t} - \mathcal{C}(\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) + \xi_{t_0+t} + \mathbf{e}_{t_0+t}, \tilde{\theta}_{t_0+t}) \\
&\quad - (\nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t}) + \xi'_{t_0+t} + \mathbf{e}'_{t_0+t} - \mathcal{C}(\nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t}) + \xi'_{t_0+t} + \mathbf{e}'_{t_0+t}, \tilde{\theta}_{t_0+t})) \\
&= (\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + (\xi_{t_0+t} - \xi'_{t_0+t}) + (\mathbf{e}_{t_0+t} - \mathbf{e}'_{t_0+t}) \\
&\quad - \mathcal{C}\left((\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + (\xi_{t_0+t} - \xi'_{t_0+t}) + (\mathbf{e}_{t_0+t} - \mathbf{e}'_{t_0+t}), \tilde{\theta}_{t_0+t}\right) \\
&= (\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + \hat{\xi}_{t_0+t} + \hat{\mathbf{e}}_{t_0+t} \\
&\quad - \mathcal{C}\left((\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + \hat{\xi}_{t_0+t} + \hat{\mathbf{e}}_{t_0+t}, \tilde{\theta}_{t_0+t}\right)
\end{aligned}$$

We estimating the norm of $\hat{\mathbf{e}}_{t_0+t}$ using linearity of $\mathcal{C}$:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\theta}_{t_0+t}}\left[\|\hat{\mathbf{e}}_{t+1}\|^2 \mid \mathbf{x}_{t_0+t}, \mathbf{e}_{t_0+t}, \theta_{t_0+t}\right] \\
&= \mathbb{E}_{\tilde{\theta}_{t_0+t}}\Big[\|(\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + \hat{\xi}_{t_0+t} + \hat{\mathbf{e}}_{t_0+t} \\
&\qquad - \mathcal{C}\left((\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + \hat{\xi}_{t_0+t} + \hat{\mathbf{e}}_{t_0+t}, \tilde{\theta}_{t_0+t}\right)\|^2\Big] \\
&\le (1-\mu)\mathbb{E}_{\tilde{\theta}_{t_0+t}}\left[\|(\nabla F(\mathbf{x}_{t_0+t}, \theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t}, \theta_{t_0+t})) + \hat{\xi}_{t_0+t} + \hat{\mathbf{e}}_{t_0+t}\|^2\right]
\end{aligned}$$

Similarly to the proof of Lemma A.5, for any ν we have:

$$\begin{aligned}
&\mathbb{E}_{t_0}\left[\|\hat{\mathbf{e}}_{t_0+t+1}\|^2\right] \\
&\le (1-\mu)\mathbb{E}_{t_0}\left[(1+\frac{1}{\nu})\|(\nabla F(\mathbf{x}_{t_0+t},\theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t},\theta_{t_0+t})) + \hat{\xi}_{t_0+t}\|^2 + (1+\nu)\|\hat{\mathbf{e}}_{t_0+t}\|^2\right] \\
&\le \frac{1}{\nu}\sum_{i=0}^{t}\left((1-\mu)(1+\nu)\right)^{t-i+1}\mathbb{E}_{t_0}\left[\|(\nabla F(\mathbf{x}_{t_0+t},\theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t},\theta_{t_0+t})) + \hat{\xi}_{t_0+t}\|^2\right]
\end{aligned}$$

By selecting $\nu = \frac{\mu}{2(1-\mu)}$ and computing the sum of a geometric series, we have:

$$\begin{aligned}
\mathbb{E}_{t_0}\left[\|\hat{\mathbf{e}}_{t_0+t+1}\|^2\right] &\le \frac{2(1-\mu)}{\mu}\sum_{i=0}^{t}\left(1-\frac{\mu}{2}\right)^{t-i+1}\mathbb{E}_{t_0}\left[\|(\nabla F(\mathbf{x}_{t_0+t},\theta_{t_0+t}) - \nabla F(\mathbf{x}'_{t_0+t},\theta_{t_0+t})) + \hat{\xi}_{t_0+t}\|^2\right] \\
&\le \frac{2(1-\mu)}{\mu}\sum_{i=0}^{t}\left(1-\frac{\mu}{2}\right)^{t-i+1}\max_i \mathbb{E}_{t_0}\left[\|(\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})) + \hat{\xi}_{t_0+i}\|^2\right] \\
&\le \frac{4(1-\mu)}{\mu^2}\max_{i=0}^{t}\mathbb{E}_{t_0}\left[\|(\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})) + \hat{\xi}_{t_0+i}\|^2\right] \\
&\le \frac{4(1-\mu)}{\mu^2}\max_{i=0}^{t}\mathbb{E}_{t_0}\left[\|\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})\|^2\right] + \frac{r^2}{d} + \|\hat{\mathbf{e}}_{t_0}\|^2
\end{aligned}$$

Substituting this into bound for $\mathcal{E} + \eta\hat{\mathbf{e}}_{t_0+t}$ and bounding the series by Proposition B.15:

$$\begin{aligned}
\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] &\le 2\eta^4L^2\left(\sum_i (1+\eta\gamma)^{t-1-i}\right)^2 \max_{i=0}^{t-1}\mathbb{E}_{t_0}\left[\|\hat{\mathbf{e}}_{t_0+i}\|^2\right] \\
&\le \frac{8c_{\mathcal{I}}\eta^4(1-\mu)L^2\beta_t^2}{\mu^2\eta\sqrt{\rho\varepsilon}}\left(\max_{i=0}^{t-1}\mathbb{E}_{t_0}\left[\|\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})\|^2\right] + \frac{r^2}{d}\right)
\end{aligned}$$

Depending on whether Assumption C holds, we consider the following cases:

- When Assumption C holds, we bound $\mathbb{E}_{t_0}\left[\|\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})\|^2\right]$ with $\tilde{\ell}^2\mathbb{E}_{t_0}\left[\|\hat{\mathbf{x}}\|^2\right] \le 4\tilde{\ell}^2\mathcal{R}^2$. Since $\mathcal{R} = c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}}$ and $r = c_r\frac{\varepsilon}{\sqrt{L\eta}}$, we can select constants $c_{\mathcal{R}}, c_r$ and c_η so that the second term dominates the first one.
- When Assumption C doesn't hold hold, we use the following bound:

  $$\begin{aligned}
  &\mathbb{E}_{t_0}\left[\|\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})\|^2\right] \\
  &\le 3\mathbb{E}_{t_0}\left[\|\nabla F(\mathbf{x}_{t_0+i},\theta_{t_0+i}) - \nabla f(\mathbf{x}_{t_0+i})\|^2 + \|\nabla f(\mathbf{x}_{t_0+i}) - \nabla f(\mathbf{x}'_{t_0+i})\|^2 + \|\nabla f(\mathbf{x}'_{t_0+i}) - \nabla F(\mathbf{x}'_{t_0+i},\theta_{t_0+i})\|^2\right] \\
  &\le 6(\sigma^2 + L^2\mathcal{R}^2)
  \end{aligned}$$

  as $c(\sigma^2 + \mathcal{R}^2)$. Using that $\eta \le \eta_\sigma \le \frac{\varepsilon^2}{d}$, we again can select the constants so that the second term dominates.

As a result, we achieve the required bound:

$$\mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \le \frac{9c_{\mathcal{I}}\eta^3(1-\mu)L^2\beta_t^2r^2}{\mu^2d\sqrt{\rho\varepsilon}}$$

$\square$

Escaping From a Saddle Point

We now show that, if the starting point is a saddle point, we move sufficiently far from it.

Lemma B.18 (Non-localization) *Let t_0 be an iteration such that $\mathbf{e}_{t_0} = 0$ in Algorithm 1. Under Assumptions A and B, let η and r be as in Definition B.2 and β_t be as in Definition B.13. Let $\gamma = -\lambda_{\min}(\nabla^2 f(\mathbf{x}_{t_0})) > \frac{\sqrt{\rho\varepsilon}}{2}$ and $\mathbb{E}_{t_0}\left[\|\mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0}\|^2\right] < \mathcal{R}^2$ for all $t \le \mathcal{I}$. Then for all $t \le \mathcal{I}$, for some constant c:*

$$\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+t}\|^2\right] \ge c\frac{\beta_t^2\eta^2 r^2}{d},$$

where $\hat{\mathbf{y}}_{t_0+t} = \mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0+t}$ as in Definition B.10.

Proof : To simplify the presentation, we use c to denote constants, and it may change its meaning from line to line.

$$\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+t}\|^2\right] \ge (\max(0, \mathbb{E}_{t_0}\left[\|\Xi_t\| - \|\Delta_t\| - \|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\| - \|Z_t\|\right]))^2$$

We show that $\mathbb{E}_{t_0}\left[\Xi_t\right] = \Omega\left(\frac{\beta_t \eta r}{\sqrt{d}}\right)$, and terms aside from Ξ_t are negligible, namely that in expectation $\mathbb{E}_{t_0}\left[\|\Delta_t\|\right], \mathbb{E}_{t_0}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|\right], \mathbb{E}_{t_0}\left[\|Z_t\|\right] \le \frac{1}{10}\mathbb{E}_{t_0}\left[\|\Xi_t\|\right]$.[5] We prove the inequality by induction. The inequality holds for $t = 0$ since all terms are 0.

Estimating Ξ_t. Since Ξ_t is a sum of independent Gaussians with variances $4(1+\eta\gamma)^{2(t-i-1)}\frac{\eta^2 r^2}{d}$, its total variance is

$$\mathbb{E}_{t_0}\left[\|\Xi_t\|^2\right] = 4\frac{\eta^2 r^2}{d}\sum_{i=0}^{t-1}(1+\eta\gamma)^{2i} = 4\frac{\eta^2 r^2}{d}\beta_t^2,$$

And since Ξ_t is a zero-mean Guassian random variable, we know $\mathbb{E}_{t_0}\left[\|\Xi_t\|\right]^2 = \frac{2}{\pi}\mathbb{E}_{t_0}\left[\|\Xi_t\|^2\right]$. Note that from the induction hypothesis it follows that $\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+t}\|^2\right] \le 2\mathbb{E}_{t_0}\left[\|\Xi_t\|^2\right] \le 8\frac{\eta^2 r^2 \beta_t^2}{d}$.

Bounding Δ_i. By the Hessian Lipschitz property, $\mathbb{E}_{t_0}\left[\|\delta_i\|^2\right] \le 4\rho^2\mathcal{R}^2$.Then for $i \le t$ and for η selected as in Definition B.2 (see proofs of Lemmas B.16 and B.17), by the induction hypothesis:

$$\mathbb{E}_{t_0}\left[\|\hat{\mathbf{x}}_{t_0+i}\|^2\right] \le 2\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+i}\|^2\right] + 2\eta^2\mathbb{E}_{t_0}\left[\|\hat{\mathbf{e}}_{t_0+i}\|^2\right] \le c\frac{\eta^2 r^2 \beta_i^2}{d}$$

[5] Most of the proof can go through if we consider $\mathbb{E}_{t_0}\left[\|\cdot\|^2\right]$ instead of $\mathbb{E}_{t_0}\left[\|\cdot\|\right]$. There is only one place in the estimation of $\|\Delta_t\|$ which requires the first momentum.

Therefore:

$$
\begin{aligned}
&\mathbb{E}_{t_0}\left[\|\Delta_t\|\right] \\
&= \mathbb{E}_{t_0}\left[\|\eta\sum_{i=0}^{t-1}(I-\eta H)^{t-i-1}\delta_i\hat{\mathbf{x}}_{t_0+i}\|\right] && \text{(Definition B.11)} \\
&\le \eta\mathbb{E}_{t_0}\left[\sum_{i=0}^{t-1}\|I-\eta H\|^{t-i-1}\cdot\|\delta_i\|\cdot\|\hat{\mathbf{x}}_{t_0+i}\|\right] \\
&\le \eta\sum_{i=0}^{t-1}(1+\eta\gamma)^{t-i-1}\sqrt{\mathbb{E}_{t_0}\left[\|\delta_i\|^2\right]\cdot\mathbb{E}_{t_0}\left[\|\hat{\mathbf{x}}_{t_0+i}\|^2\right]} && \text{(Cauchy-Schwarz)} \\
&\le c\eta\rho\mathcal{R}\frac{\eta r}{\sqrt{d}}\mathbb{E}_{t_0}\left[\sum_{i=0}^{t-1}(1+\eta\gamma)^{t-i-1}\beta_i\right] && \left(\mathbb{E}_{t_0}\left[\|\delta_i\|^2\right]\le 4\rho^2\mathcal{R}^2 \text{ and } \mathbb{E}_{t_0}\left[\|\hat{\mathbf{x}}_{t_0+i}\|^2\right]\le c\frac{\eta^2r^2\beta_t^2}{d}\right) \\
&\le c\eta\rho\mathcal{R}\frac{\eta r}{\sqrt{d}}\mathbb{E}_{t_0}\left[\sum_{i=0}^{t-1}\frac{(1+\eta\gamma)^{t-1}}{\sqrt{\eta\gamma}}\right] && \text{(Proposition B.14)} \\
&\le c\eta\rho\mathcal{R}\frac{\eta r}{\sqrt{d}}\mathcal{I}\beta_t && \text{(Proposition B.14, another direction)} \\
&\le c\frac{\eta r\beta_t}{\sqrt{d}}\left(\eta\rho c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}}\cdot\frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}\right) && \left(\mathcal{R}=c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}} \text{ and } \mathcal{I}=\frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}\right) \\
&\le cc_{\mathcal{R}}c_{\mathcal{I}}\frac{\eta r\beta_t}{\sqrt{d}},
\end{aligned}
$$

and it suffices to choose $cc_{\mathcal{R}}c_{\mathcal{I}} \le \frac{1}{40}$ so that $\mathbb{E}_{t_0}\left[\|\Delta_t\|\right] \le \frac{1}{10}\mathbb{E}_{t_0}\left[\|\Xi_t\|\right]$.

Bounding $\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_{t_0+t}\|$. For a general compressor, by Lemma B.16 we know that

$$
\mathbb{E}_{t_0}\left[\|\mathcal{E}_t+\eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \le \frac{cc_{\mathcal{I}}\eta^3(1-\mu)L^2\chi^2\beta_t^2}{\mu^2\sqrt{\rho\varepsilon}}
$$

Using $\chi \le 2r$, to show that $\mathbb{E}_{t_0}\left[\|\mathcal{E}_t+\eta\hat{\mathbf{e}}_{t_0+t}\|\right]^2 \le \mathbb{E}_{t_0}\left[\|\mathcal{E}_t+\eta\hat{\mathbf{e}}_{t_0+t}\|^2\right] \le \frac{1}{100}\mathbb{E}_{t_0}\left[\|\Xi_t\|\right]^2$, it suffices to guarantee that

$$
\frac{c_{\mathcal{I}}\eta^3(1-\mu)L^2\chi^2\beta_t^2}{\mu^2\sqrt{\rho\varepsilon}} \le c\frac{\beta_t^2\eta^2r^2}{d} \iff \eta \le c\frac{\mu^2\sqrt{\rho\varepsilon}r^2}{c_{\mathcal{I}}d(1-\mu)\chi^2L^2}
$$

Using that $\chi^2 \le 2r^2$ for sufficiently large c_r, we have:

$$
\eta \le c\frac{\mu^2\sqrt{\rho\varepsilon}}{c_{\mathcal{I}}d(1-\mu)L^2}
$$

For a linear compressor, By Lemma B.17 we have:

$$
\frac{c_{\mathcal{I}}\eta^3(1-\mu)L^2\beta_t^2r^2}{\mu^2d\sqrt{\rho\varepsilon}} \le c\frac{\beta_t^2\eta^2r^2}{d} \iff \eta \le c\frac{\mu^2\sqrt{\rho\varepsilon}}{c_{\mathcal{I}}(1-\mu)L^2}
$$

Bounding $\|Z_t\|$**.** First, we consider the case when Assumption C doesn't hold (i.e. $\tilde{\ell} = +\infty$). Since Z_t is the sum of independent random variables:

$$\mathbb{E}_{t_0}\left[\|Z_t\|^2\right] \le \eta^2 \sum_{i=0}^{t-1} (1+\eta\gamma)^{2(t-i-1)} 2\eta^2\sigma^2 \le 2\eta^4\beta_t^2\sigma^2$$

To prove that $\mathbb{E}_{t_0}[\|Z_t\|]^2 \le \mathbb{E}_{t_0}\left[\|Z_t\|^2\right] < \frac{1}{100}\mathbb{E}_{t_0}[\|\Xi_t\|]^2$, it suffices to show that

$$\eta^4\beta_t^2\sigma^2 \le c\frac{\beta_t^2\eta^2 r^2}{d} \iff \sigma\sqrt{d} \le cr \iff \sigma^2 d \le cc_r^2 \frac{\varepsilon^2}{L\eta} \iff \eta \le c\frac{c_r^2\varepsilon^2}{\sigma^2 L d},$$

which holds when $c_\eta \le cc_r^2$, by Definition B.2.

Finally, we consider the case when Assumption C holds (i.e. $\tilde{\ell} < +\infty$). Since stochastic gradient is Lipschitz, we have $\|\hat{\zeta}_i\| \le 2\tilde{\ell}\|\hat{x}_i\|$ and:

$$\begin{aligned}
\mathbb{E}_{t_0}\left[\|Z_t\|^2\right] &= \mathbb{E}_{t_0}\left[\|\eta\sum_{i=0}^{t-1}(I-\eta H)^{t-i-1}\hat{\zeta}_i\|^2\right] && \text{(Definition B.11)} \\
&\le \eta^2 \sum_{i=0}^{t-1} \mathbb{E}_{t_0}\left[\|(I-\eta H)^{t-i-1}\hat{\zeta}_i\|^2\right] && \text{(Noises are independent)} \\
&\le \eta^2 \sum_{i=0}^{t-1} \|(1+\eta\gamma)^{t-i-1}\|^2 \cdot \mathbb{E}_{t_0}\left[\|\hat{\zeta}_i\|^2\right] && \text{(Since } \gamma \text{ is the smallest negative eigenvalue of } H\text{)} \\
&\le \eta^2 \sum_{i=0}^{t-1} \|(1+\eta\gamma)^{t-i-1}\|^2 \cdot \tilde{\ell}^2 \mathbb{E}_{t_0}\left[\|\hat{\mathbf{x}}_{t_0+i}\|^2\right] && \text{(Assumption C)} \\
&\le c\eta^2\mathcal{I}\frac{\eta^2 r^2\beta_t^2}{d} && \text{(See derivation for } \|\Delta_t\| \text{ above)}
\end{aligned}$$

Therefore $\mathbb{E}_{t_0}[\|Z_t\|] \le c\eta\tilde{\ell}\sqrt{\mathcal{I}}\frac{\beta_t\eta r}{\sqrt{d}}$. To guarantee that $\mathbb{E}_{t_0}[\|Z_t\|] \le \frac{1}{10}\mathbb{E}_{t_0}[\|\Xi_t\|]$, it suffices to show that

$$\eta\tilde{\ell}\sqrt{\mathcal{I}}\frac{\beta_t\eta r}{\sqrt{d}} \le c\frac{\eta r\beta_t}{\sqrt{d}} \iff \eta\tilde{\ell}\sqrt{\mathcal{I}} \le c \iff \frac{c_{\mathcal{I}}^2\eta\tilde{\ell}^2}{\sqrt{\rho\varepsilon}} \le c \iff \eta \le c\frac{\sqrt{\rho\varepsilon}}{c_{\mathcal{I}}^2\tilde{\ell}^2},$$

which holds when $c_{\mathcal{I}}^2 c_\eta \le c$. □

Theorem B.19 *Under Assumptions A and B, for η as in Definition B.2, after $T = \tilde{O}\left(\frac{f_{\max}}{\eta\varepsilon^2}\right)$ iterations of Algorithm 1 at least half of points $\mathbf{x}_{t_0}$ such that the condition at Line 5 is triggered at iteration t_0 are ε-SOSP. The condition in Line 5 is triggered at most $\tilde{O}(\frac{f_{\max}}{\mathcal{F}}) = \tilde{O}(\frac{T}{\mathcal{I}}) = \tilde{O}(\varepsilon^{-3/2})$ times.*

Note that the fraction of ε-SOSP can be made arbitrary close to 1: to achieve $1-c$ fraction, we show that $c/2$ fraction doesn't have large gradients and $c/2$ fraction doesn't have escape direction. For simplicity, in the theorem we consider $c = 1/2$.

Proof : As in the previous Lemma, c is used to denote constants and may change its meaning from line to line. The general idea is the following: to show that at least half of points are ε-SOSP, it suffices to show that at most quarter of the points have large gradient, i.e. $\|\nabla f(\mathbf{x}_{t_0})\| \ge \varepsilon$, and we show that at most quarter of the points have escape directions, i.e. $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{t_0})) < -\sqrt{\rho\varepsilon}$. As shown in Appendix A, at most quarter of the points have gradient $\|\nabla f(\mathbf{x}_t)\| \geq \varepsilon$. It remains to show that there is at most quarter of points such that $\lambda_{\min}(\nabla^2(f(\mathbf{x}_t))) \leq -\sqrt{\rho\varepsilon}$. By Line 5 of Algorithm 1, and Proposition B.3, for any t' there exists $t_0 \in [t' - \mathcal{I}, t']$ such that $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{t_0})) < -\sqrt{\rho\varepsilon}/2$ and $\mathbf{e}_{t_0} = 0$. We show that when compressed SGD starts from $\mathbf{x}_{t_0}$, the objective improves.

Proposition B.20 *If $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{t_0})) < -\frac{\sqrt{\rho\varepsilon}}{2}$, then for some $t \leq \mathcal{I}$:*

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+t})\right] \geq \mathcal{F}$$

Proof : By Lemma B.18, we bound $\hat{\mathbf{y}}_{t_0+t} = \mathbf{y}'_{t_0+t} - \mathbf{y}_{t_0+t}$ (Definition B.10):

$$\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+t}\|^2\right] \geq c\frac{\beta_t^2\eta^2 r^2}{d} \geq c\frac{(1+\eta\gamma)^{2t}\eta^2}{d\eta\gamma}\cdot\frac{\varepsilon^2}{L\eta} \geq c\frac{(1+\eta\gamma)^{2t}\varepsilon^2}{d\gamma L}$$

For $t = \mathcal{I}$, we have $(1+\eta\gamma)^{\mathcal{I}} \geq (1+\eta\sqrt{\rho\varepsilon})^{c_{\mathcal{I}}/\eta\sqrt{\rho\varepsilon}} \geq e^{c_{\mathcal{I}}}$. By selecting $c_{\mathcal{I}} \geq c\log\frac{dL\rho 2\mathcal{R}^2}{\sigma\varepsilon}$ for some c, we have $\mathbb{E}_{t_0}\left[\|\hat{\mathbf{y}}_{t_0+t}\|^2\right] \geq 2\mathcal{R}^2$, and therefore:

$$\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}_{\mathcal{I}}\|^2\right] = \max(\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}_{\mathcal{I}}\|^2\right], \mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}'_{\mathcal{I}}\|\right]^2) \geq \frac{1}{2}\mathbb{E}_{t_0}\left[\|\mathbf{y}'_{\mathcal{I}} - \mathbf{y}_{\mathcal{I}}\|^2\right] \geq \mathcal{R}^2$$

Since by Proposition B.9 $\mathbf{y}_{t_0+t}$ and $\mathbf{y}'_{t_0+t}$ have the same distribution, $\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}_{\mathcal{I}}\|\right] = \mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}'_{\mathcal{I}}\|\right]$, and therefore

$$\mathbb{E}_{t_0}\left[\|\mathbf{y}_{t_0} - \mathbf{y}_{\mathcal{I}}\|^2\right] \geq \mathcal{R}^2,$$

and by Corollary B.6:

$$f(\mathbf{y}_{t_0}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_0+\mathcal{I}})\right] \geq \mathcal{F},$$

and therefore the objective decreases by $\mathcal{F}$ after $\mathcal{I}$ iterations. □

Let $t_1 = 0, t_2, \ldots, t_k = T$ be the iterations where the condition at Line 5 of Algorithm 1 is triggered.

Using this proposition, we have:

$$f(\mathbf{y}_0) - f(\mathbf{y}_T) \geq \sum_i (f(\mathbf{y}_{t_i}) - \mathbb{E}_{t_0}\left[f(\mathbf{y}_{t_{i+1}})\right]) \geq \frac{1}{4}\frac{T}{\mathcal{I}}\mathcal{F} \geq \frac{c}{4}\frac{f_{\max}\eta\sqrt{\rho\varepsilon}}{\eta\varepsilon^2}\sqrt{\frac{\varepsilon^3}{\rho}} \geq \frac{c}{4}f_{\max},$$

which is impossible by selecting a sufficiently large constant in the choice of T.

□