OPT2020: 12th Annual Workshop on Optimization for Machine Learning

# Escaping Saddle Points with Compressed SGD

**Dmitrii Avdiukhin** DAVDYUKH@IU.EDU
**Grigory Yaroslavtsev** GRIGORY@GRIGORY.US
*Indiana University, Bloomington*[*]

## Abstract

Stochastic gradient descent (SGD) is a prevalent method for solving smooth nonconvex problems arising in machine learning. Since SGD computation can be efficiently distributed across multiple machines, communication often becomes the main bottleneck in applications. Gradient compression methods can be used to alleviate this problem, and recent line of work shows that SGD with certain compression methods convergence to an $\varepsilon$-first-order stationary point. In this work we extend these result to convergence to an $\varepsilon$-*second*-order stationary point. By using Compressed SGD we show that, compared to the uncompressed case:

- When stochastic gradient is not Lipschitz, total communication decreases by $\tilde{O}(\varepsilon^{-3/4})$,
- When stochastic gradient is Lipschitz and $\varepsilon = o(d^{-2/3})$, total communication decreases by $\tilde{O}(\varepsilon^{-3/4}/\sqrt{d})$.

## 1. Introduction

Escaping from saddle points in nonconvex optimization is a topic of interest in a number of recent optimization papers for machine learning [2, 7, 11, 17, 19]. Remarkably, first-order methods are able to find approximate second-order stationary points in a number of iterations comparable to those required to find first-order stationary points [11].

In practice, for massive machine learning workloads a large number of machines is required to speed up the training process. Communication typically becomes the main bottleneck in training [5, 16] and hence a common solution is to apply gradient compression at every step [1, 9]. In this paper we show that the main workhorse of distributed optimization for deep learning, stochastic gradient descent (SGD), achieves fast guaranteed convergence to an approximate second-order stationary point even when an optimal gradient compression is applied at every step. While it was shown recently [9, 12] that this holds for first-order convergence, ours is the first analysis of second-order convergence. In this sense, our results are similar to the breakthrough work of [7], who were the first to show second-order convergence for uncompressed gradient descent methods.

Our main technical contribution is the analysis showing that compressed SGD can escape from saddle points efficiently. Inspired by the ideas from [11] and [15] we present an algorithm (Algorithm 1) which uses perturbed compressed gradients with error-feedback and converges to an $\varepsilon$-second-order stationary point (see Theorem 6). Table 1 outlines communication improvements for various choices of compression parameters. Unlike stochastic noise, which can't behave adversarially, errors arising from gradient compression can be highly correlated, introducing some amount of slowdown in convergence compared to the optimal uncompressed methods of [11]. Despite this we

[*] Research supported by NSF award CCF-1657477 and Facebook Faculty Research Award.

are able to show substantial improvements in total communication for certain settings. It remains open whether optimal rates of convergence can be achieved with compression.

## 2. Preliminaries

**Function properties** For a twice differentiable nonconvex function $f$, we consider the unconstrained minimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

**Assumption 1** We use the following standard [2, 6, 11, 18, 19] assumptions about the objective function $f$:

**Assumption 1.A** $f$ is $f_{\max}$-bounded, has $L$-Lipschitz gradient and $\rho$-Lipschitz Hessian:

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

Lipschitz gradient is required to achieve fast convergence for nonconvex optimization problems. In addition, Lipschitz Hessian allows one to show fast second-order convergence.

**Assumption 1.B** Unbiased stochastic gradient $\nabla F(\mathbf{x},\theta)$, where $\theta$ is a randomness-controlling parameter (e.g. a minibatch selected at a given iteration), with bounded variance:

$$\mathbb{E}_\theta\left[\nabla F(\mathbf{x},\theta)\right] = \nabla f(\mathbf{x}), \qquad \mathbb{E}_\theta\left[\|\nabla F(\mathbf{x},\theta) - \nabla f(\mathbf{x})\|^2\right] \leq \sigma^2$$

**Assumption 1.C** Lipschitz stochastic gradient. For any $\mathbf{x},\mathbf{y},\theta$:

$$\|\nabla F(\mathbf{x},\theta) - \nabla F(\mathbf{y},\theta)\| \leq \tilde{\ell}\|\mathbf{x}-\mathbf{y}\|, \quad \tilde{\ell}\in[0;+\infty]$$

Note that $\tilde{\ell}$ can be $+\infty$, corresponding to the case when this assumption doesn't hold. From machine learning perspective, Assumption 1.C means that for the same mini-batch, if the initial models are close, their updates are also close. For neural networks, each network layer is a composition of an activation function and a linear function, such assumption holds when each activation function is 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, when we have $\mathcal{W}$ machines, each corresponding to a differentiable function $f_i$ such that $f = \sum_{i=1}^{\mathcal{W}} f_i$. Each machine computes a stochastic gradient $\nabla F_i(\mathbf{x},\theta)$ such that $\mathbb{E}_\theta\left[\nabla F_i(\mathbf{x},\theta)\right] = \nabla f_i(\mathbf{x})$. After that, the gradients are gathered on the coordinator machine which computes the full stochastic gradient $\nabla F(\mathbf{x},\theta) = \frac{1}{\mathcal{W}}\sum_{i=1}^{\mathcal{W}} \nabla F(\mathbf{x},\theta)$. Using this gradient the coordinator can perform a stochastic gradient step: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta\nabla F(\mathbf{x}_t,\theta)$, where $\mathbf{x}_t$ is the previous iterate and $\eta$ denotes the step size.

The key advantage of this approach is that by increasing the number of machines the computation can be parallelized perfectly. However, with each machine required to send its gradient, communication becomes the main bottleneck. A popular solution to this issue is gradient compression: each machine sends only an approximation of its gradient, e.g. a sign of each coordinate [3], $k$ random coordinates [15], the $k$ largest coordinates [15], the compressed difference with the previous gradient [8], gradient quantization [1]. Then coordinator averages these approximations and broadcasts it to all machines (possibly compressing it again).

We are interested in two properties of this protocol: how good the approximation is and how much communication per machine it requires. The first property is formalized in the following definition:

**Definition 1** *A randomized function $\mathcal{C}(\mathbf{x})$ is a* $\lambda$-compressor *if*

$$\mathbb{E}\left[\|\mathbf{x}-\mathcal{C}(\mathbf{x})\|^2\right] < \lambda\|\mathbf{x}\|^2$$

For example, a scaled sign function $\mathcal{C}(\mathbf{x}) = \frac{\|x\|_1}{d} sign(x)$ is a $(1-\frac{1}{d})$-compressor and returning top $k$ coordinates of a vector results in a $(1-\frac{k}{d})$-compressor. However, it's nontrivial to compute compressors efficiently in a distributed setting, given that each machine stores only a part of the input $\mathbf{x}^{(i)}$ and computing $\mathbf{x} = \frac{1}{\mathcal{W}}\sum_{i=1}^{\mathcal{W}} \mathbf{x}^{(i)}$ explicitly would require $O(d)$ communication per machine. Addressing this question, the recent work of [9] uses a communication-efficient compressor with the following property:

**Lemma 2 ([9], Lemma 1, reformulated)** *There exists a $1-k/d$ compressor $\mathcal{C}$ such that $\mathcal{C}(\mathbf{x})$ requires only $\tilde{O}(k)$ bits of communication per worker for any $\mathbf{x}$.*

This compressor returns $k$ largest coordinates of the full stochastic gradient. The key idea is to have the coordinator recover the indices of the $k$ largest coordinates using COUNT SKETCH [4] and then evaluate these coordinates on each machine. COUNT SKETCH allows one to achieve this with $\tilde{O}(k)$ communication per worker.

**Stationary points** While our goal is to find a local minimum, finding it is in general NP-hard [13]. Instead, as is standard in the literature, we can show convergence to an approximate first-order stationary point or an approximate second-order stationary point.

**Definition 3** *For a differentiable function $f$, $\mathbf{x}$ is an $\varepsilon$-first-order stationary point ($\varepsilon$-FOSP) if $\|\nabla f(\mathbf{x})\| \le \varepsilon$.*

An $\varepsilon$-FOSP can be a local maximum, a local minimum or a saddle point. While local minima often 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 [14] we refer to points with no escape directions (up to some approximation) as approximate second-order stationary points:

**Definition 4 ([14])** *For a twice-differentiable, $\rho$-Hessian Lipschitz function $f$, $\mathbf{x}$ is an $\varepsilon$-second-order stationary point ($\varepsilon$-SOSP) if $\|\nabla f(\mathbf{x})\| \le \varepsilon$ and $\nabla^2 f(\mathbf{x}) \succeq -\sqrt{\rho\varepsilon}$.*

While one can consider two threshold parameters – $\varepsilon_g$ for $\nabla f$ and $\varepsilon_H$ for $\nabla^2 f$ – we follow convention of [14] which selects $\varepsilon_H = -\sqrt{\rho\varepsilon}$, intuitively balancing first-order and second-order variability. An important property of points which are not $\varepsilon$-SOSP is that they are unstable: adding a small perturbation allows gradient descent to escape them [7]. Similar results were shown for other gradient descent variations, e.g. stochastic [11] and accelerated [10] gradient descent. In this work we will show that this property holds even for stochastic gradient descent with gradient compression.

## 3. Algorithm and analysis

**Algorithm** We present our algorithm in Algorithm 1. This algorithm is a compressed stochastic gradient descent based on Algorithm 1 from [15]. However, in order to achieve second-order

convergence we add an artificial random noise $\xi_t$ to gradient at every iteration (similarly to [11]). As we show in Appendix B, this modification allows gradient descent to escape saddle points.

At every iteration $t$ the algorithm computes stochastic gradient $\nabla F(\mathbf{x}_t, \theta_t)$ and adds artificial noise $\xi_t$ to it. The resulting value is compressed and we update the current iterate $\mathbf{x}_t$ using this value. However, the error resulting from compression is not ignored: we calculate error feedback $\mathbf{e}_{t+1}$ – the difference between the computed value and the compressed value – and add it to the gradient in the next iteration. [12] shows that carrying over the error term makes a difference for this algorithm's convergence to a first-order stationary point.

**Algorithm 1:** Compressed SGD

**parameters:** step size $\eta$, number of iterations $T$, artificial noise variance $r^2$,
**input :** objective $f$, compressor function $\mathcal{C}$, starting point $\mathbf{x}_0$
**output :** $\varepsilon$-SOSP of $f$
$\mathbf{e}_0 \leftarrow 0^d$;
**for** $t = 0 \dots T-1$ **do**
  $\mathbf{g}_t \leftarrow \mathcal{C}(\nabla F(\mathbf{x}_t, \theta_t) + \xi_t + \mathbf{e}_t), \quad \xi_t \sim \mathcal{N}_d(0^d, r^2)$;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \mathbf{g}_t$ ;
  $\mathbf{e}_{t+1} \leftarrow \mathbf{e}_t + \nabla F(\mathbf{x}_t, \theta_t) + \xi_t - \mathbf{g}_t$;
**end**
**return** $\mathbf{x}_T$

**Analysis** To simplify the presentation, we introduce notation $\delta = {}^{(1-\lambda)}/\sqrt{\lambda}$. In the following statements, $\tilde{O}$ hides polynomial dependence on $L, \rho, f_{\max}, \sigma, \tilde{\ell}$ and polynomial dependence on all parameters. The following result is similar to that of [15], but is slightly more general: it covers the case when $\lambda$ is close to 0 and doesn't require an assumption of bounded gradients. The proof of the Theorem is presented in Appendix A;

**Theorem 5 (Convergence to $\varepsilon$-FOSP)** *Let $f$ satisfy Assumption 1, $\mathcal{C}$ be a $\lambda$-compressor and $\delta = \frac{(1-\lambda)}{\sqrt{\lambda}}$. Then for $\eta = \tilde{O}\left(\min(\varepsilon^2, \delta\varepsilon)\right)$, after $T = \tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{1}{\delta\varepsilon^3}\right)$ iterations, at least half of visited points are $\varepsilon$-FOSP.*

The following theorem is our main result and shows that compressed SGD converges to an $\varepsilon$-SOSP. The proof of the Theorem is presented in Appendix B;

**Theorem 6 (Convergence to $\varepsilon$-SOSP)** *Let $f$ satisfy Assumption 1, $\mathcal{C}$ be a $\lambda$-compressor and $\delta = \frac{(1-\lambda)}{\sqrt{\lambda}}$. Let $\eta_\sigma = \tilde{O}(\varepsilon^2)$ if Assumption C is satisfied and $\eta_\sigma = \tilde{O}\left(\frac{\varepsilon^2}{d}\right)$ otherwise, $\eta_\lambda = \tilde{O}\left(\min\left(\delta\varepsilon, \frac{\delta^2\sqrt{\varepsilon}}{d}\right)\right)$ and $\eta = \min(\eta_\sigma, \eta_\lambda)$. Then after $T = \tilde{O}\left(\frac{1}{\varepsilon^2\eta}\right)$ iterations, at least half of visited points are $\varepsilon$-SOSP.*

Convergence to $\varepsilon$-SOSP requires $\eta \leq \eta_\lambda$, which may result in noticeably slower convergence rate compared to $\varepsilon$-FOSP convergence. The reason for such behavior is that, for convergence to $\varepsilon$-SOSP, compression introduces an error similar to that of the stochastic noise; however, unlike the stochastic error, the compression is not Lipschitz even for deterministic gradients. For example, consider the sign compressor used in [12]: $\mathcal{C}(\mathbf{x}) = \frac{\|\mathbf{x}\|_1}{d} sign(\mathbf{x})$. Two points $\mathbf{x}_1 = (\varepsilon, \dots, \varepsilon)$ and $\mathbf{x}_2 = (-\varepsilon, \dots, -\varepsilon)$ can be arbitrary arbitrarily close for small $\varepsilon$, but the difference between their compressions is constant. See Lemma 22 in Appendix B for more details.

**Compressed SGD in distributed settings** Below we consider different scenarios to illustrate how convergence depends on the properties of the compressor. To estimate the total communication in the compressed case, recall that by Lemma 2 there exists a $(1-\frac{k}{d})$-compressor which requires $\tilde{O}(k)$ communication. By selecting $\lambda = 1-\frac{k}{d}$, where $k = o(d)$, we have $\delta = \Theta\left(\frac{k}{d}\right)$ and $\eta_\lambda = \tilde{O}\left(\min\left(\frac{k\varepsilon}{d}, \frac{k^2\sqrt{\varepsilon}}{d^3}\right)\right)$. 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 Lemma 2 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 gradients often have heavy coordinates, which are easy to recover. We formulate this beyond worst-case scenarios as the following optional assumption:

**Assumption 2** 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 [8]. In cases when certain coordinates are much more prominent in the gradient compared to others, top-$k$ compressors show good performance.

**Corollary 7** *Algorithm 1 converges to $\varepsilon$-SOSP in a number of settings, as shown in Table 1.*

Table 1: Convergence to $\varepsilon$-SOSP in various settings. *Uncompressed* setting corresponds to the standard SGD convergence analysis. *Compressed* setting corresponds to using a compressor (Lemma 2) of an appropriate size. *Constant-size sketch* is our beyond worst-case assumption (Assumption 2) where we assume constant compression with $\tilde{O}(1)$ communication. The last column shows an improvement in total communication compared to the uncompressed case. For Lipschitz stochastic gradients $\nabla F$, compression gives improvement for $\varepsilon = o(d^{-2/3})$

| Setting | $\lambda$ | Iterations | Total communication per worker | Total communication improvement |
|---|---|---|---|---|
| Uncompressed Lipschitz $\nabla F$ | $0$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | – |
| Compressed Lipschitz $\nabla F$ | $1-\sqrt{d}\varepsilon^{3/4}$ $(\varepsilon = o(d^{-2/3}))$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d\sqrt{d}}{\varepsilon^{3+1/4}}\right)$ | $\tilde{O}\left(\frac{1}{\varepsilon^{3/4}\sqrt{d}}\right)$ |
| Constant-size sketch Lipschitz $\nabla F$ | $c < 1$ | $\tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{d}{\varepsilon^3}\right)$ | $\tilde{O}\left(\frac{1}{\varepsilon^4} + \frac{d}{\varepsilon^3}\right)$ | $\tilde{O}\left(\min(d, \frac{1}{\varepsilon})\right)$ |
| Uncompressed non-Lipschitz $\nabla F$ | $0$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d^2}{\varepsilon^4}\right)$ | – |
| Compressed non-Lipschitz $\nabla F$ | $1-\varepsilon^{3/4}$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d^2}{\varepsilon^{3+1/4}}\right)$ | $\tilde{O}(\frac{1}{\varepsilon^{3/4}})$ |
| Constant-size sketch non-Lipschitz $\nabla F$ | $c < 1$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(d\right)$ |

## References


[1] 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.

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

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

[4] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, pages 1–31, 2019.

[12] 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.

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

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

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

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

[17] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2019.

[18] 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.

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

## Appendix A. Convergence to $\varepsilon$-FOSP

In this section we prove Theorem 5, showing convergence to an approximate first-order stationary point. Results and proofs are inspired by [12], with the key difference in that we show how to avoid using the bounded gradient assumption: $\mathbb{E}\left[\|\nabla F\|^2\right] \leq G^2$ and handle the case of $\lambda$-compressors with $\lambda \ll 1$.

**Definition 8** *(Noise and compression parameters) We use the following notation:*

- $\zeta_t = \nabla F(\mathbf{x}_t, \theta_t) - \nabla f(\mathbf{x}_t)$ *is a stochastic gradient noise. This noise has variance* $\sigma^2$
- $\xi_t$ *is an artificial Gaussian noise added at every iteration. This noise has variance* $r^2$
- $\psi_t = \zeta_t + \xi_t$ *is the overall noise. This noise has variance* $\chi^2 = \sigma^2 + r^2$.
- *We assume that compression of the gradients is done using a* $\lambda$*-compressor* $C$*. In order to simplify the derivations we introduce an auxiliary parameter* $\delta = {}^{(1-\lambda)}/_{\sqrt{\lambda}}$

In order to perform the analysis, similarly to [12], we introduce an auxiliary sequence of noisy iterates $\{\mathbf{y}_t\}$ defined below. These iterates allow one to remove the impact of the compression error so that we can analyze it separately from the noise.

**Definition 9** *(Noisy iterates) Let the sequence of noisy iterates* $\{\mathbf{y}_t\}$ *be defined as* $\mathbf{y}_t = \mathbf{x}_t - \eta \mathbf{e}_t$.

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

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

Hence for the $\{\mathbf{y}_t\}$ sequence we have:

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

Thus $\{\mathbf{y}_t\}$ iterates remove the impact of the compression error from the analysis.

### A.1. Compression error estimation

Recall that the compression error terms $\mathbf{e}_t$ in Algorithm 1 represent the difference between stochastic gradient (SG) and compressed SG. Similarly to how SG noise leads to increase in the number of iterations compared to non-stochastic gradient descent, the presence of $\mathbf{e}_t$ also increases the number of iterations, and therefore it's important to bound their norm.

**Lemma 10 (Compression error estimation)** *Let* $\mathbf{x}_t, \mathbf{e}_t$ *be defined as in Algorithm 1 and let* $\chi^2$ *be as in Definition 8. Then under Assumption 1, for any* $t$ *we have*

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

In particular, letting $\delta = {}^{(1-\lambda)}/\sqrt{\lambda}$ we get a result similar to Lemma 3 from [12]:

$$\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \leq \frac{4}{\delta^2}(\max(\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2\right]) + \chi^2)$$

**Proof** The proof is similar to proof of Lemma 3 from [12]. The main difference is that we don't rely on the bounded gradient assumption.

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

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

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

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

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

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

■

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

**Corollary 11** *Under assumptions of Lemma 10, we have*

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

**Proof**

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

Bounding $\frac{\lambda}{1-\lambda}\sum\left(\frac{1+\lambda}{2}\right)^{\tau}$ with $\frac{2\lambda}{(1-\lambda)^2}=\frac{2}{\delta^2}$, we have:

$$\sum_{\tau=0}^{t}\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right]\leq\frac{4}{\delta^2}\leq\frac{2\eta^3L^2}{\delta^2}\sum_{i=0}^{t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2+\chi^2\right]$$

■

### A.2. Descent Lemma

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

**Lemma 12 (Descent lemma)** *Let $f$ satisfy Assumption 1. Let $\chi^2$, $\delta$ be as in Definition 8. For $\eta<\frac{1}{4L}\min(\delta,1)$, for any $T$ we have:*

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

Using this lemma, we'll later show that for sufficiently large $T$, multiple visited points have small gradients (note that the left-hand side divided by $T$ we obtain an average squared gradient norm), making them $\varepsilon$-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 $\eta$. The second term arises from stochastic gradient noise, while the last term appears because of compression.

**Proof** The proof is similar to proof of Theorem II from [12].

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{y}_{t+1})|\mathbf{x}_t,\mathbf{e}_t\right]&\leq f(\mathbf{y}_t)+\langle\nabla f(\mathbf{y}_t),\mathbb{E}\left[\mathbf{y}_{t+1}-\mathbf{y}_t|\mathbf{x}_t,\mathbf{e}_t\right]\rangle+\frac{L}{2}\mathbb{E}\left[\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2|\mathbf{x}_t,\mathbf{e}_t\right]\\
&=f(\mathbf{y}_t)-\eta\langle\nabla f(\mathbf{y}_t),\nabla f(\mathbf{x}_t)\rangle+\frac{L\eta^2}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)+\psi_t\|^2|\mathbf{x}_t,\mathbf{e}_t\right]\\
&\leq f(\mathbf{y}_t)-\eta\|\nabla f(\mathbf{x}_t)\|^2-\eta\langle\nabla f(\mathbf{y}_t)-\nabla f(\mathbf{x}_t),\nabla f(\mathbf{x}_t)\rangle+\frac{L\eta^2}{2}\|\nabla f(\mathbf{x}_t)\|^2+\frac{L\chi^2\eta^2}{2}\\
&\leq f(\mathbf{y}_t)-\eta(1-\frac{L\eta}{2})\|\nabla f(\mathbf{x}_t)\|^2+\frac{L\chi^2\eta^2}{2}-\eta\langle\nabla f(\mathbf{y}_t)-\nabla f(\mathbf{x}_t),\nabla f(\mathbf{x}_t)\rangle
\end{aligned}$$

By using inequality $|\langle a,b\rangle|\leq\frac{\|a\|^2}{2}+\frac{\|b\|^2}{2}$ and Lipschitz gradient assumption, we have:

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

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

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

Bounding sum of $\|\mathbf{e}_t\|^2$ by Corollary 11, we have:

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

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

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

After regrouping the terms, we get the required result:

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

■

### A.3. Convergence to $\varepsilon$-FOSP

**Theorem 13 (Convergence to $\varepsilon$-FOSP)** *Let $f$ satisfy Assumption 1. Then for $\eta = \tilde{O}\left(\min(\varepsilon^2, \delta\varepsilon)\right)$, after $T = \tilde{\Theta}\left(\frac{1}{\varepsilon^4} + \frac{1}{\delta\varepsilon^3}\right)$ iterations, at least half of visited points are $\varepsilon$-FOSP.*

**Proof** Proof by contradiction. For $\eta < \frac{1}{4L}\min(\delta, 1)$, if less than half points are $\varepsilon$-FOSP, then by Lemma 12:

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

It suffices to guarantee that all terms are at most $\frac{T\varepsilon^2}{6}$:

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

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

## Appendix B. Convergence to $\varepsilon$-SOSP

By rescaling we can assume that $\varepsilon \leq 1$. Recall that $\delta = \frac{1-\lambda}{\sqrt{\lambda}}$ by Definition 8. We introduce the following auxiliary notation:

**Definition 14 (Step sizes)**

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

$$\textit{Min. step size for compressed SGD} \quad \eta_\lambda = \min\left(\frac{\delta\varepsilon}{L\sigma}, \frac{\delta^2\sqrt{\varepsilon}}{Ld}\right) = \tilde{O}\left(\min\left(\frac{\delta\varepsilon}{L}, \frac{\delta^2\sqrt{\varepsilon}}{d}\right)\right)$$

Intuitively, selecting step size $\eta \leq \eta_\sigma$ suffices to show convergence of SGD [11]. In addition, selecting $\eta \leq \eta_\lambda$ allows us to extend the results to compressed SGD. When $\tilde{\ell} = +\infty$, $\eta_\sigma = \tilde{O}\left(\frac{\varepsilon^2}{d}\right)$, and when $\tilde{\ell}$ is a constant, $\eta_\sigma = \tilde{O}(\varepsilon^2)$.

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):

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

Table 2: Convergence to $\varepsilon$-SOSP for various settings.

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

Recall that $\chi^2 = \sigma^2 + r^2 = \sigma^2 + \frac{c_r \varepsilon^2}{L\eta}$ by Definition 8 and $f_{\max} = f(\mathbf{x}_0) - f(\mathbf{x}^*)$. 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.

### B.1. Proof outline

Our proof is mainly based on the ideas from [11]. We introduce "Improve or localize" lemma (Lemma 15): 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 [11], 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 [11] in several ways. The first difference is that, aside from $\mathbf{x}_t$, our equations have another sequence $\mathbf{y}_t$ ($\mathbf{x}_t$ mainly participate as arguments of $\nabla f(\cdot)$, while $\mathbf{y}_t$ participate as argument of $f(\cdot)$ and in distances). This introduces the following challenge: if some relation holds for $\mathbf{y}_t$, it doesn't necessary holds for $\mathbf{x}_t$. For example, if we have a bound on $\|\mathbf{y}_t - \mathbf{y}'_t\|$, we don't necessarily have a bound on $\|\mathbf{x}_t - \mathbf{x}'_t\|$, and it needs to be established separately.

Another difference is that 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 nearby gradients are also large, and by Lemma 12 the objective improves (see Lemma 18). Otherwise, we use "Improve or localize" Lemma as described above. In the latter case, similarly to [11], we have to bound errors which arise from the fact that the function is not quadratic and gradients are not deterministic (see Definition 20). However, we have an additional error term stemming from gradient compression (see Definition 20); to bound this term (see Lemma 22), we need bounded $\|\mathbf{e}_t\|$, 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 16 considers the simplified form, obtained by substituting parameters from Equation 1.

**Lemma 15 (Improve or localize)** *Under Assumption 1, for $\eta < \frac{1}{4L}\min(\delta, 1)$, for $\mathbf{y}_t, \chi, \delta$ defined as in Definition 8, we have*

$$f(\mathbf{y}_0) - \mathbb{E}\left[f(\mathbf{y}_T)\right] \geq \frac{\mathbb{E}\left[\|\mathbf{y}_T - \mathbf{y}_0\|^2\right]}{8\eta T} - \eta^2\chi^2 T\left(L + \frac{2L^2\eta}{\delta^2}\right) - \eta\chi^2$$

**Proof** Let $\psi_t = \zeta_t + \xi_t$. Note that $\mathbf{y}_{t+1} = \mathbf{y}_t - \eta(\nabla f(\mathbf{x}_t) + \psi_t)$.

$$\mathbb{E}\left[\|\sum_{t=0}^{T}\psi_t\|^2\right] = \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{E}\left[\|\psi_t\|^2\right]\right] = \sum_{\tau=0}^{T}\chi^2 = T\chi^2$$

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}_T-\mathbf{y}_0\|^2\right] &= \eta^2\mathbb{E}\left[\|\sum_{i=0}^{T-1}(\nabla f(\mathbf{x}_i)+\psi_i)\|^2\right]\\ &\le 2\eta^2\mathbb{E}\left[\|\sum_{i=0}^{T-1}\nabla f(\mathbf{x}_i)\|^2+\|\sum_{i=0}^{T-1}\psi_i\|^2\right]\\ &\le 2\eta^2 T\sum_{i=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_i)\|^2\right]+2\eta^2\chi^2 T\end{aligned}$$

Since $\eta < \frac{1}{4L}\min(\delta,1)$, by Lemma 12:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}_T-\mathbf{y}_0\|^2\right] &\le 2\eta^2 T\left(\frac{4(f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right])}{\eta}+\eta\chi^2 T\left(2L+\frac{8L^2\eta}{\delta^2}\right)\right)+2\eta^2\chi^2 T\\ &\le 2\eta T\left(4(f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right])+\eta^2\chi^2 T\left(2L+\frac{8L^2\eta}{\delta^2}\right)+\eta\chi^2\right)\\ &\le 2\eta T\left(4(f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right])+\eta^2\chi^2 T\left(4L+\frac{8L^2\eta}{\delta^2}\right)+4\eta\chi^2\right)\end{aligned}$$

After regrouping the terms, we have:

$$f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right] \ge \frac{\mathbb{E}\left[\|\mathbf{y}_T-\mathbf{y}_0\|^2\right]}{8\eta T}-\eta^2\chi^2 T\left(L+\frac{2L^2\eta}{\delta^2}\right)-\eta\chi^2,$$

To guarantee that the sum of these terms is at most $\frac{\mathcal{F}}{2}$, it suffices to select parameters so that $c_\eta + c_r^2 + c_{\mathcal{I}}c_\eta \le c_{\mathcal{F}}/4$.

■

**Corollary 16** *Under Assumption 1, for $\mathcal{F},\mathcal{I}$ chosen as specified in Equation 1, for any $T \le \mathcal{I}$ we have:*

$$f(\mathbf{y}_0)-\mathbb{E}\left[f(\mathbf{y}_T)\right] \ge \frac{\sqrt{\rho\varepsilon}}{8c_{\mathcal{I}}}\mathbb{E}\left[\|\mathbf{y}_T-\mathbf{y}_0\|^2\right]-\frac{\mathcal{F}}{2}$$

**Proof** With our choice of parameters, we can bound negative terms on the right-hand side of Lemma 15.

**Bounding $\eta\chi^2$.**

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

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

**Bounding $\eta^2\chi^2 TL$.**

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

which is equal to the term estimated above.

**Bounding $\eta^2\chi^2 T \cdot \frac{2L^2\eta}{\delta^2}$.**

$$\frac{\eta^3\chi^2 TL^2}{\delta^2} \leq \frac{c_{\mathcal{I}}\eta^2\chi^2 L^2}{\delta^2\sqrt{\rho\varepsilon}} \leq \frac{c_{\mathcal{I}}\eta^2 L^2\left(\sigma^2 + \frac{c_r\varepsilon^2}{L\eta}\right)}{\delta^2\sqrt{\rho\varepsilon}} \leq \frac{c_{\mathcal{I}}}{\delta^2\sqrt{\rho\varepsilon}}\left(\eta_\lambda^2 L^2\sigma^2 + c_r\eta_\lambda L\varepsilon^2\right) \leq 2c_{\mathcal{I}}c_\eta\frac{\sqrt{\varepsilon^3}}{\sqrt{\rho}}$$

■

**Corollary 17** *Under Assumption 1, for $\mathcal{F}, \mathcal{R}, \mathcal{I}$ chosen as specified in Equation 1, if there exists $t \in [0, \mathcal{I}]$ such that $\|\mathbf{y}_t - \mathbf{y}_0\| > \mathcal{R}$, then $f(\mathbf{y}_0) - \mathbb{E}\left[f(\mathbf{y}_t)\right] \geq \mathcal{F}$.*

**Proof** By Lemma 15:

$$f(\mathbf{y}_0) - \mathbb{E}\left[f(\mathbf{y}_t)\right] \geq \frac{\mathcal{R}^2}{2\eta\mathcal{I}} - \frac{\mathcal{F}}{2} = \frac{c_{\mathcal{R}}^2\varepsilon\eta\sqrt{\rho\varepsilon}}{2c_{\mathcal{I}}\eta\rho} - \frac{\mathcal{F}}{2} = \left(\frac{c_{\mathcal{R}}^2}{2c_{\mathcal{I}}c_{\mathcal{F}}} - \frac{1}{2}\right)\mathcal{F} \geq \mathcal{F},$$

where the last inequality holds when $3c_{\mathcal{I}}c_{\mathcal{F}} \leq c_{\mathcal{R}}^2$. ■

### B.3. Large gradient case: $\|\nabla f(\mathbf{x}_0)\| \geq 3L\mathcal{R}$

**Lemma 18 (Large gradient case)** *Under Assumption 1 for $\mathcal{F}, \mathcal{R}, \mathcal{I}$ chosen as specified in Equation 1, if $\|\nabla f(\mathbf{x}_0)\| > 3L\mathcal{R}$, then after at most $\mathcal{I}$ iterations the objective decreases by $\mathcal{F}$.*

**Proof** If there exists $t \leq \mathcal{I}$ such that $\|\mathbf{y}_t - \mathbf{y}_0\| > \mathcal{R}$, then by Corollary 17, the objective decreases by at least $\mathcal{F}$.

First we show by induction that $\|\nabla f(\mathbf{x}_t)\| \geq \frac{\|\nabla f(\mathbf{x}_0)\|}{3}$ and $\|\nabla f(\mathbf{x}_t)\| \leq 2\|\nabla f(\mathbf{x}_0)\|$ for all $t \leq \mathcal{I}$.

$$\begin{aligned}
\|\nabla f(\mathbf{x}_t)\| &= \|\nabla f(\mathbf{x}_0) - (\nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t))\| \\
&\geq \|\nabla f(\mathbf{x}_0)\| - \|\nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\| \\
&\geq \|\nabla f(\mathbf{x}_0)\| - L\|\mathbf{x}_0 - \mathbf{x}_t\| \\
&\geq \|\nabla f(\mathbf{x}_0)\| - L\|\mathbf{y}_0 - \mathbf{y}_t\| - L\|\mathbf{y}_t - \mathbf{x}_t\| \\
&\geq \|\nabla f(\mathbf{x}_0)\| - L\mathcal{R} - \eta\|\mathbf{e}_t\|
\end{aligned}$$

In the equation above, we have to bound $\|\mathbf{e}_t\|$

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{e}_{t+1}\|\right] &\leq \sqrt{\lambda}\|\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t\| && \text{(By definition of } \lambda\text{-compressor)} \\
&\leq \sqrt{\lambda}\|\nabla f(\mathbf{x}_t) + \psi_t\| + \sqrt{\lambda}\|\mathbf{e}_t\| && \text{(By triangle inequality)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}\|\nabla f(\mathbf{x}_i) + \psi_i\| && \text{(By telescoping)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}(\|\nabla f(\mathbf{x}_i)\| + \chi) && \text{(By triangle inequality)}
\end{aligned}$$

Since $\|\nabla f(\mathbf{x}_i)\| \leq 2\|\nabla f(\mathbf{x}_0)\|$, for sufficiently small $\eta$, we have $\|\mathbf{e}_t\| \leq \frac{\|\nabla f(\mathbf{x}_0)\|}{3}$, and therefore

$$\|\nabla f(\mathbf{x}_t)\| \geq \|\nabla f(\mathbf{x}_0)\| - L\mathcal{R} - \eta\|\mathbf{e}_t\| \geq \frac{\|\nabla f(\mathbf{x}_0)\|}{3}$$

The upper bound is similar:

$$\|\nabla f(\mathbf{x}_0)\| \leq \|\nabla f(\mathbf{x}_0)\| + L\mathcal{R} + \eta\|\mathbf{e}_t\| \leq 2\|\nabla f(\mathbf{x}_0)\|$$

By Lemma 12, we know:

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

Therefore:

$$\begin{aligned}
f(\mathbf{y}_0) - \mathbb{E}\left[f(\mathbf{y}_\mathcal{I})\right] &\geq \frac{\eta\mathcal{I}}{4}\left(L^2\mathcal{R}^2 - \eta\chi^2(2L + \frac{8L^2\eta}{\delta^2})\right) \\
&\geq \frac{c_\eta\eta}{4\eta\sqrt{\rho\varepsilon}}\left(\frac{c_\mathcal{R}^2 L^2\varepsilon}{\rho} - 10c_\eta\varepsilon^2\right) \\
&\geq \frac{c_\eta\eta}{4\eta\sqrt{\rho\varepsilon}}\left(c_\mathcal{R}^2\varepsilon^2 - 10c_\eta\varepsilon^2\right) && \text{(since } L \geq \sqrt{\rho\varepsilon}\text{)} \\
&\geq \mathcal{F},
\end{aligned}$$

where the last inequality holds when $c_\eta(c_\mathcal{R}^2 - 10c_\eta) \geq c_\mathcal{F}$. ■

### B.4. Small Gradient Case: $\|\nabla f(\mathbf{x}_0)\| < 3L\mathcal{R}$

#### B.4.1. COUPLING SEQUENCES

Let $H = \nabla^2 f(\mathbf{x}_0)$; we use $\mathbf{x}^\top H\mathbf{x}$ as a quadratic approximation of $f$ near $x_0$. Let $v_1$ be the eigenvector corresponding to the smallest eigenvalue $\gamma$ of $H$. Then we construct *coupling sequences* $\mathbf{x}_t$ and $\mathbf{x}'_t$ in the following way: $\mathbf{x}_t$ is constructed as described in Algorithm 1; $\mathbf{x}'_t$ has the same

stochastic randomness $\theta$ as $\mathbf{x}_t$, and its artificial noise $\xi'_t$ is the same as $\xi_t$ with exception of the coordinate corresponding to $v_1$, which has an opposite sign.

$$\begin{aligned} & & \mathbf{e}'_0 &= \mathbf{e}_0 \\ \xi_t &\sim \mathcal{N}(0, r^2) & \xi'_t &= \xi_t - 2\langle v_1, \xi_t \rangle v_1 \\ \mathbf{g}_t &= C(\nabla F(\mathbf{x}_t, \theta_t) + \xi_t + \mathbf{e}_t) & \mathbf{g}'_t &= C(\nabla F(\mathbf{x}'_t, \theta_t) + \xi'_t + \mathbf{e}'_t) \\ \mathbf{y}_t &= \mathbf{x}_t - \eta \mathbf{e}_t & \mathbf{y}'_t &= \mathbf{x}'_t - \eta \mathbf{e}'_t \\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta \mathbf{g}_t & \mathbf{x}'_{t+1} &= \mathbf{x}'_t - \eta \mathbf{g}'_t \\ \mathbf{e}_{t+1} &= \nabla F(x_t, \theta_t) + \xi_t + \mathbf{e}_t - \mathbf{g}_t & \mathbf{e}'_{t+1} &= \nabla F(x'_t, \theta_t) + \xi'_t + \mathbf{e}'_t - g'_t \end{aligned} \tag{2}$$

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

**Lemma 19** *For all $t$, $\mathbf{x}_t$ and $\mathbf{y}_t$ from Equation 2 have the same distribution as $\mathbf{x}'_t$ and $\mathbf{y}'_t$.*

**Proof** By definition of $\mathbf{y}_t$ and $\mathbf{y}'_t$, it suffices show whether $\mathbf{x}_t$ and $\mathbf{e}_t$ have the same distributions as $\mathbf{x}'_t$ and $\mathbf{e}'_t$.

Proof by Induction. $\mathbf{y}_0 = \mathbf{y}'_0 = \mathbf{x}_0 - \eta \mathbf{e}_0$.

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

- Since $\mathbf{x}_t$ and $\mathbf{x}'_t$ have the same distribution, $\nabla F(\mathbf{x}_t, \theta_t)$ and $\nabla F(\mathbf{x}'_t, \theta_t)$ have the same distribution.
- Since $\mathcal{N}(0, r^2)$ is symmetric and $\xi'_t$ is the same as $\xi_t$ with exception of one coordinate, which has an opposite sign, $\xi_t$ and $\xi'_t$ have the same distribution.
- $\mathbf{e}_t$ and $\mathbf{e}'_t$ have the same distribution.

Similarly, $\mathbf{e}_{t+1}$ has the same distribution as $\mathbf{e}'_{t+1}$, since $\nabla F(\mathbf{x}_t, \theta_t)$, $\xi_t$, $\mathbf{e}_t$ and $\mathbf{g}_t$ have the same distribution as $\nabla F(\mathbf{x}'_t, \theta_t)$, $\xi'_t$, $\mathbf{e}'_t$ and $\mathbf{g}'_t$. ■

Since our sequences have the same distribution, we have $\mathbb{E}[f(\mathbf{x}_t)] = \mathbb{E}[f(\mathbf{x}'_t)]$. We want to show that in a few iterations $\mathbf{y}'_t - \mathbf{y}_t$ becomes sufficiently large and, therefore, at least one of $\mathbf{y}_t$ and $\mathbf{y}'_t$ is far from $\mathbf{x}_0$. By applying Lemma 15 we will show that the objective sufficiently decreases.

### B.4.2. EXPRESSING THE DIFFERENCE BETWEEN COUPLING SEQUENCES

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

$$\hat{\mathbf{x}}_t = \mathbf{x}'_t - \mathbf{x}_t \qquad \hat{\mathbf{e}}_t = \mathbf{e}'_t - \mathbf{e}_t \qquad \hat{\zeta}_t = \zeta'_t - \zeta_t \qquad \hat{\xi}_t = \xi'_t - \xi_t$$

We split $\hat{\mathbf{x}}_t$ into 4 terms: $\hat{\mathbf{x}}_t = -(\Delta_t + \mathcal{E}_t + Z_t + \Xi_t)$, corresponding to different sources of approximation error, defined as follows:

**Definition 20** *Let* $\delta_i = \int_0^1 \nabla^2 f(\alpha \mathbf{x}'_i + (1-\alpha)\mathbf{x}_i)d\alpha - H$. *Then*

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

Recall that $\zeta_t$ is an SGD noise, $\xi_t$ is an artificial noise, $\mathbf{e}_t$ is the compression error.

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

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

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

We'll now prove the expansion.

$$\begin{aligned}
\hat{\mathbf{x}}_{t+1} &= \mathbf{x}'_{t+1} - \mathbf{x}_{t+1} \\
&= \mathbf{y}'_{t+1} + \eta \mathbf{e}'_{t+1} - (\mathbf{y}_{t+1} + \eta \mathbf{e}_{t+1}) && \text{(By definition of } \mathbf{y}_t \text{ and } \mathbf{y}'_t) \\
&= \eta \hat{\mathbf{e}}_{t+1} + (\mathbf{y}'_t - \mathbf{y}_t) - \eta \left( (\nabla f(x'_t) - \nabla f(x_t)) + (\zeta'_t - \zeta_t) + (\xi'_t - \xi_t) \right) && \text{(By update equation for } \mathbf{y}_t) \\
&= \eta(\hat{\mathbf{e}}_{t+1} - \hat{\mathbf{e}}_t) + \hat{\mathbf{x}}_t - \eta \left( (\delta_t + H)\hat{\mathbf{x}}_t + \hat{\zeta}_t + \hat{\xi}_t \right) && \text{(By definition of } \delta_t \text{ and } \mathbf{y}_t) \\
&= \eta(\hat{\mathbf{e}}_{t+1} - \hat{\mathbf{e}}_t) + (I - \eta H)\hat{\mathbf{x}}_t - \eta \left( \delta_t \hat{\mathbf{x}}_t + \hat{\zeta}_t + \hat{\xi}_t \right) \\
&= (I - \eta H)\hat{\mathbf{x}}_t - \eta \left( \delta_t \hat{\mathbf{x}}_t + (\hat{\mathbf{e}}_t - \hat{\mathbf{e}}_{t+1}) + \hat{\zeta}_t + \hat{\xi}_t \right)
\end{aligned}$$

Using telescoping, we get the required expression. Since $\mathbf{y}'_t - \mathbf{y}_t = \hat{\mathbf{x}}_t - \eta \hat{\mathbf{e}}_t$:

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

and we'll use $\mathbf{y}'_t - \mathbf{y}_t$ in Corollary 17.

### B.4.3. BOUNDING ACCUMULATED COMPRESSION ERROR

Compared to SGD analysis, an additional term $\mathcal{E}_t + \eta \hat{\mathbf{e}}_t$ appears. This term corresponds to accumulated error arising from compression, and we have to bound its norm.

**Definition 21** *Following [11], we introduce the following term which is proportional to* $\|\mathbf{y}_t\|$*:*

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

In [11] it was shown that

$$\beta(t) = \Theta\left(\frac{(1+\eta\gamma)^t}{\sqrt{\eta\gamma}}\right)$$

**Lemma 22 (Bounding accumulated compression error)** *Under Assumption 1, let* $\chi$ *and* $\delta$ *be as in Definition 20,* $\mathcal{E}_t$ *and* $\hat{\mathbf{e}}_t$ *be as in Definition 20,* $\beta_t$ *be as in Definition 21 and* $\eta$ *and* $\mathcal{R}$ *as in Equation 1. Assume that* $\max(\|\mathbf{y}_t - \mathbf{y}_0\|, \|\mathbf{y}'_t - \mathbf{y}_0\|) < \mathcal{R}$ *for all* $t$ *and* $\|\nabla f(\mathbf{x}_0)\| \leq 2L\mathcal{R}$*. Let* $-\gamma$ *be the smallest negative eigenvalue of* $\nabla^2 f(\mathbf{x}_0)$ *such that* $\gamma \geq \frac{\sqrt{\rho\varepsilon}}{2}$*. Then under Assumptions A, B, D we have:*

$$\mathbb{E}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|\right] \leq \frac{6\eta^{3/2}L\chi\beta_t}{\delta\sqrt{\gamma}}$$

**Proof** Expanding sum in $\mathcal{E}_t$ and using that $\hat{\mathbf{e}}_0 = 0$:

$$\begin{aligned}
\mathcal{E}_t &= \eta\sum_{i=0}^{t-1}(I-\eta H)^{t-1-i}(\hat{\mathbf{e}}_i - \hat{\mathbf{e}}_{i+1}) && \text{(By Definition 20)}\\
&= \eta(-\hat{\mathbf{e}}_t + \sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}((I-\eta H)-I)\hat{\mathbf{e}}_i) && \text{(By telescoping)}\\
&= -\eta\hat{\mathbf{e}}_t + \eta^2 H\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}\hat{\mathbf{e}}_i
\end{aligned}$$

We can now estimate $\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|$. Since $-\gamma$ is the smallest negative eigenvalue of $H$, we have $\|I - \eta H\| \leq (1+\eta\gamma)$.

$$\begin{aligned}
\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\| &= \|\eta^2 H\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}\hat{\mathbf{e}}_i\|\\
&\leq \eta^2 L\sum_i(1+\eta\gamma)^{t-1-i}\|\hat{\mathbf{e}}_i\| && \text{(By gradient Lipschitzness } \lambda_{\max}(H)\leq L)\\
&\leq \eta^2 L\sum_i(1+\eta\gamma)^{t-1-i}\|\mathbf{e}'_i - \mathbf{e}_i\| && \text{(By definition of } \hat{\mathbf{e}}_i)\\
&\leq \eta^2 L\sum_i(1+\eta\gamma)^{t-1-i}(\|\mathbf{e}'_i\| + \|\mathbf{e}_i\|) && \text{(By triangle inequality)}
\end{aligned}$$

in the equation above, we have to bound $\|\mathbf{e}_t\|$:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{e}_{t+1}\|\right] &\leq \sqrt{\lambda}\|\nabla f(\mathbf{x}_t) + \psi_t + \mathbf{e}_t\| && \text{(By definition of } \lambda\text{-compressor)} \\
&\leq \sqrt{\lambda}\|\nabla f(\mathbf{x}_t) + \psi_t\| + \sqrt{\lambda}\|\mathbf{e}_t\| && \text{(By triangle inequality)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}\|\nabla f(\mathbf{x}_i) + \psi_i\| && \text{(By telescoping)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}(\|\nabla f(\mathbf{x}_i)\| + \chi) && \text{(By triangle inequality)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}(\|\nabla f(\mathbf{y}_i)\| + \|\nabla f(\mathbf{y}_i) - \nabla f(\mathbf{x}_i)\| + \chi) && \text{(By triangle inequality)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}(L\mathcal{R} + L\|\mathbf{y}_i - \mathbf{x}_i\| + \chi) && \text{(By theorem assumption and Lipschitz condition)} \\
&\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i}(L\mathcal{R} + \eta\|\mathbf{e}_i\| + \chi) && \text{(By definition of } \mathbf{y}_i\text{)}
\end{aligned}$$

We'll show by induction that $\|\mathbf{e}_t\| \leq \frac{6\chi}{\delta}$. By selecting small enough constant $c_{\mathcal{R}}$ in the definition of $\mathcal{R}$, we have $L\mathcal{R} \leq \chi$. Using the induction hypothesis, $\eta\|\mathbf{e}_i\| \leq 6\chi\frac{\eta}{\delta} \leq \chi$ by selecting a sufficiently small $c_\eta$ in the definition of $\eta$.

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{e}_{t+1}\|\right] &\leq \sum_{i=0}^{t-1} \sqrt{\lambda}^{t-i} 3\chi && \text{(Using bounds on } L\mathcal{R} \text{ and } \eta\mathbf{e}_t\text{)} \\
&\leq \frac{3\sqrt{\lambda}\chi}{1-\sqrt{\lambda}} && \text{(Taking a sum of the geometric series)} \\
&\leq \frac{3\sqrt{\lambda}(1+\sqrt{\lambda})\chi}{1-\lambda} \\
&\leq \frac{6\chi}{\delta}
\end{aligned}$$

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

$$\mathbb{E}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|\right] \leq \frac{6\eta^2 L\chi}{\delta}\sum_i (1+\eta\gamma)^{t-1-i} \leq \frac{6\eta^2 L\chi\beta_t}{\delta\sqrt{\eta\gamma}},$$

where we estimated the series in the following way:

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

■

### B.4.4. ESCAPING FROM A SADDLE POINT

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

**Lemma 23 (Non-localization)** *Under Assumption 1, let $\chi$ and $\delta$ be as in Definition 20, $\beta_t$ be as in Definition 21 and $\eta$ and $r$ as in Equation 1. Then for any $t$*

$$\mathbb{E}\left[\|\mathbf{y}'_t - \mathbf{y}_t\|\right] = \Theta\left(\frac{\beta_t \eta r}{\sqrt{d}}\right)$$

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

$$4\frac{\eta^2 r^2}{d}\sum_{i=0}^{t-1}(1+\eta\gamma)^{2i} = 4\frac{\eta^2 r^2}{d}\beta_t^2$$

and therefore $\mathbb{E}\left[\|\Xi_t\|\right] = \sqrt{\frac{8}{\pi}}\frac{\beta_t \eta r}{\sqrt{d}}$.

We show that terms aside from $\Xi_t$ are negligible, namely that $\|\Delta_t + (\mathcal{E}_t + \eta\hat{\mathbf{e}}_t) + Z_t\| \le \frac{1}{2}\|\Xi_t\|$. We prove the inequality by induction. The inequality holds for $t = 0$ since all terms are 0. Assume that inequality holds for $t$, namely

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_t\|\right] \le 2\mathbb{E}\left[\|\Xi_t\|\right] \le \frac{4\eta r \beta_t}{\sqrt{d}}$$

It suffices to show that each of $\|\Delta_t\|$, $\|Z_t\|$ and $\|\mathcal{E}_t\|$ is less than $\frac{1}{10}\|\Xi_t\|$.

**Bounding $\Delta_i$.** By Hessian Lipschitz property, $\mathbb{E}\left[\|\delta_i\|\right] \le 2\rho R$, and by induction hypothesis $\mathbb{E}\left[\|\hat{\mathbf{x}}_i\|\right] \le 4\frac{\eta r \beta_t}{\sqrt{d}}$ for $i \le t$. Therefore:

$$\begin{aligned}
\mathbb{E}\left[\Delta_t\right] &= \mathbb{E}\left[\eta\sum_{i=0}^{t-1}(I-\eta H)^{t-i-1}\delta_i\hat{\mathbf{x}}_i\right] && \text{(By definition 20)}\\
&\le \eta\sum_{i=0}^{t-1}\|I-\eta H\|^{t-i-1}\cdot\mathbb{E}\left[\|\delta_i\|\cdot\|\hat{\mathbf{x}}_i\|\right] && \text{(Bounding norms of all terms)}\\
&\le 2\eta\sum_{i=0}^{t-1}(1+\eta\gamma)^{t-i-1}\rho\mathcal{R}\mathbb{E}\left[\|\hat{x}_i\|\right] && \text{(Using bound on } \|\delta_i\|)\\
&\le 8\eta\rho\mathcal{R}\mathcal{I}\frac{\beta_t\eta r}{\sqrt{d}} && \text{(By induction hypothesis)}\\
&\le 8\eta\rho c_{\mathcal{R}}\sqrt{\frac{\varepsilon}{\rho}}\frac{c_{\mathcal{I}}}{\eta\sqrt{\rho\varepsilon}}\frac{\eta r\beta_t}{\sqrt{d}} && \text{(Expanding } \mathcal{R} \text{ and } \mathcal{I} \text{ by equation 1)}\\
&= 8c_{\mathcal{I}}c_{\mathcal{R}}\frac{\eta r\beta_t}{\sqrt{d}}
\end{aligned}$$

which is less than $\frac{1}{10}\mathbb{E}\left[\|\Xi_t\|\right]$ when $c_{\mathcal{I}}c_{\mathcal{R}} \le \frac{1}{100}$.

**Bounding $\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|$** By Lemma 22 we know that

$$\mathbb{E}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|\right] \leq \frac{6\eta^{3/2}L\chi\beta_t}{\delta\sqrt{\gamma}}$$

Using $\chi \leq 2r$, to show that $\mathbb{E}\left[\|\mathcal{E}_t + \eta\hat{\mathbf{e}}_t\|\right] \leq \frac{1}{10}\mathbb{E}\left[\|\Xi_t\|\right]$, it suffices to guarantee that

$$\frac{6\eta^{3/2}L\chi\beta_t}{\delta\sqrt{\gamma}} \leq \frac{\beta_t\eta r}{10\sqrt{d}} \iff \sqrt{\eta} \leq \frac{\delta\sqrt{\gamma}r}{60\sqrt{d}\chi L}$$

which holds when (using bounds on $\chi$ and $\gamma$)

$$\eta \leq \frac{\delta^2\sqrt{\rho\varepsilon}}{2\cdot 60^2 dL^2}$$

**Bounding $\|Z_t\|$.** First, we consider the case when Assumption 1.C doesn't hold (i.e. $\tilde{\ell} = +\infty$). Since $\zeta_t|\zeta_0, \ldots, \zeta_{t-1}$ is a Gaussian distribution and $Z_t$ is also a sum of independent random variables:

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

Therefore, $\mathbb{E}\left[\|Z_t\|\right] \leq 2\eta^2\sigma\beta_t$. To prove $\mathbb{E}\left[\|Z_t\|\right] < \frac{1}{10}\mathbb{E}\left[\|\Xi_t\|\right]$, it suffices to show that

$$2\eta\sigma\beta_t \leq \frac{\eta r\beta_t}{10\sqrt{d}} \iff 20\sigma\sqrt{d} \leq r \iff 400\sigma^2 d \leq c_r^2\frac{\varepsilon^2}{L\eta} \iff \eta \leq \frac{c_r^2\varepsilon^2}{400\sigma^2 Ld},$$

which holds when $100c_\eta \leq c_r^2$.

Finally, we consider the case when Assumption 1.C holds (i.e. $\tilde{\ell} < +\infty$). Since stochastic gradient is Lipschitz, we have $\hat{\zeta}_i \leq 2\tilde{\ell}\mathcal{R}$ and:

$$\begin{aligned}
\mathbb{E}\left[\|Z_t\|^2\right] &= \|\eta\sum_{i=0}^{t-1}(I-\eta H)^{t-i-1}\hat{\zeta}_i\|^2 && \text{(By definition 20)}\\
&\leq \eta^2\mathcal{I}\sum_{i=0}^{t-1}\|(I-\eta H)^{t-i-1}\hat{\zeta}_i\|^2 && \text{(By Cauchy-Schwarz)}\\
&\leq \eta^2\mathcal{I}\sum_{i=0}^{t-1}\|(1+\eta\gamma)^{t-i-1}\|^2\cdot\|\hat{\zeta}_i\|^2 && \text{(Since } \gamma \text{ is the smallest negative eigenvalue of } H)
\end{aligned}$$

and therefore $\mathbb{E}\left[\|Z_t\|\right] \leq 2\eta\tilde{\ell}\sqrt{\mathcal{I}}\frac{\beta_t\eta r}{\sqrt{d}}$ To guarantee that $\mathbb{E}\left[\|Z_t\|\right] \leq \frac{1}{10}\mathbb{E}\left[\|\Xi_t\|\right]$, it suffices to show that

$$2\eta\tilde{\ell}\sqrt{\mathcal{I}}\frac{\beta_t\eta r}{\sqrt{d}} \leq \frac{\eta r\beta_t}{10\sqrt{d}} \iff \eta\tilde{\ell}\sqrt{\mathcal{I}} \leq \frac{1}{20} \iff \frac{c_{\mathcal{I}}^2\eta\tilde{\ell}^2}{\sqrt{\rho\varepsilon}} \leq \frac{1}{400} \iff \eta \leq \frac{\sqrt{\rho\varepsilon}}{400c_{\mathcal{I}}^2\tilde{\ell}^2},$$

which holds when $400c_{\mathcal{I}}^2c_\eta \leq 1$. ■

**Theorem 24** *Under Assumption 1, for $\eta$ as in Equation 1, after $\tilde{O}\left(\frac{1}{\eta\varepsilon^2}\right)$ iterations of Algorithm 1, at least half of visited points are $\varepsilon$-SOSP.*

Note that the fraction of $\varepsilon$-SOSP can be made arbitrary large.

**Proof** First we show that, if $\lambda_{\min}(\nabla^2 f(\mathbf{x}_0)) > -\frac{\sqrt{\rho\varepsilon}}{2}$, then for some $t \leq \mathcal{I}$

$$f(\mathbf{x}_0) - \mathbb{E}\left[f(\mathbf{x}_t)\right] \geq \mathcal{F}$$

By Lemma 23:

$$\mathbb{E}\left[\|\mathbf{y}'_t - \mathbf{y}_t\|\right] \geq 4\frac{\beta_t \eta r}{\sqrt{d}} \geq 4\frac{\eta(1+\eta\gamma)^t}{\sqrt{d\eta\gamma}} \cdot \frac{\varepsilon}{\sqrt{L\eta}} \geq 4\frac{\sigma\varepsilon}{\sqrt{Ld\gamma}}(1+\eta\gamma)^t$$

Substituting $t = \mathcal{I}$, 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 $c_{\mathcal{I}} \geq c\log\frac{dL\rho 2\mathcal{R}}{\sigma\varepsilon}$ for some $c$, we have $\mathbb{E}\left[\|\mathbf{y}'_t - \mathbf{y}_t\|\right] \geq 2\mathcal{R}$, and therefore:

$$\max(\mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{y}_{\mathcal{I}}\|\right], \mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{y}'_{\mathcal{I}}\|\right]) \geq \frac{1}{2}\mathbb{E}\left[\|\mathbf{y}'_{\mathcal{I}} - \mathbf{y}_{\mathcal{I}}\|\right] \geq \mathcal{R}$$

Since by Lemma 19 $\mathbf{y}_t$ and $\mathbf{y}'_t$ have the same distribution, $\mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{y}_{\mathcal{I}}\|\right] = \mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{y}'_{\mathcal{I}}\|\right]$, and therefore

$$\mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{y}_{\mathcal{I}}\|\right] \geq \mathcal{R},$$

By Corollary 17:

$$f(\mathbf{x}_0) - \mathbb{E}\left[f(\mathbf{x}_{\mathcal{I}})\right] \geq \mathcal{F},$$

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

We split all iterations into chunks of size $\mathcal{I}$. For each chunk $[s, s+\mathcal{I}]$ we consider the following cases:

- If $\|\nabla f(\mathbf{x}_s)\| \geq 2L\mathcal{R}$, then by Lemma 18 the objective decreases by $\mathcal{F}$, and therefore there are at most $O(\frac{f_{\max}}{\mathcal{F}})$ such chunks.
- If $\lambda_{\min}(\nabla^2 f(\mathbf{x}_s)) \leq -\frac{\sqrt{\rho\varepsilon}}{2}$, then, as shown above, the objective also decreases by $\mathcal{F}$.
- If $\|\nabla f(\mathbf{x}_s)\| \leq 2L\mathcal{R}$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}_s)) \geq -\frac{\sqrt{\rho\varepsilon}}{2}$, then, by Hessian-Lipschitz property, by selecting sufficiently small $\mathcal{R}$ we guarantee that $\mathbf{x}_t$ is an $\varepsilon$-SOSP for all $t \in [s, s+\mathcal{I})$. By Lemma 15, objective increases by at most $\frac{\mathcal{F}}{2}$.

If after $T$ iterations less than half of points are $\varepsilon$-SOSP, then the objective decreases by $\frac{T}{\mathcal{I}}(\mathcal{F} - \frac{\mathcal{F}}{2})$, which is greater than $f_{\max}$ for $T \geq \frac{2f_{\max}}{\mathcal{I}}$. ■

## Appendix C. Choice of parameters

**Proposition 25 (Corollary 7 restated)** *For various settings of Algorithm 1 we have the convergence rate to $\varepsilon$-SOSP as shown in Table 3.*

**Proof**

Table 3: Convergence to $\varepsilon$-SOSP in various settings. Uncompressed setting corresponds to the standard SGD convergence analysis. Compressed setting corresponds to using a compressor (Lemma 2 of an appropriate size). Constant-size sketch is our beyond worst-case assumption (Assumption 2) where we assume constant compression with $\tilde{O}(1)$ communication. The last column shows an improvement in total communication compared to the uncompressed case. For Lipschitz stochastic gradients $\nabla F$, compression provides improvement when $\varepsilon = o(d^{-2/3})$

| Setting | $\lambda$ | Communication per round per worker | Iterations | Total communication per worker | Total communication improvement |
|---|---|---|---|---|---|
| Uncompressed Lipschitz $\nabla F$ | $0$ | $O(d)$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | – |
| Compressed Lipschitz $\nabla F$ | $1-\sqrt{d}\varepsilon^{3/4}$ $(\varepsilon = o(d^{-2/3}))$ | $O(d^{3/2}\varepsilon^{3/4})$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d\sqrt{d}}{\varepsilon^{3+1/4}}\right)$ | $\tilde{O}\left(\frac{1}{\varepsilon^{3/4}\sqrt{d}}\right)$ |
| Constant-size sketch Lipschitz $\nabla F$ | $c<1$ | $\tilde{O}(1)$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}+\frac{d}{\varepsilon^3}\right)$ | $\tilde{O}\left(\frac{1}{\varepsilon^4}+\frac{d}{\varepsilon^3}\right)$ | $\tilde{O}\left(\min(d,\frac{1}{\varepsilon})\right)$ |
| Uncompressed non-Lipschitz $\nabla F$ | $0$ | $O(d)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d^2}{\varepsilon^4}\right)$ | – |
| Compressed non-Lipschitz $\nabla F$ | $1-\varepsilon^{3/4}$ | $O(d\varepsilon^{3/4})$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d^2}{\varepsilon^{3+1/4}}\right)$ | $\tilde{O}(\frac{1}{\varepsilon^{3/4}})$ |
| Constant-size sketch non-Lipschitz $\nabla F$ | $c<1$ | $\tilde{O}(1)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$ | $\tilde{O}\left(d\right)$ |

**Uncompressed case.** To simplify the presentation, let $\alpha = 1$ when Assumption 1.C holds and let $\alpha = d$ otherwise. Therefore, $\eta_\sigma = \tilde{O}(\frac{\varepsilon^2}{\alpha})$.

In uncompressed case, $\mathcal{C}(\mathbf{x}) = \mathbf{x}$ and $\lambda = 0$. In this case, $\eta_\lambda = \infty$, $\eta = \eta_\sigma = \tilde{O}\left(\frac{\varepsilon^2}{\alpha}\right)$ and the number of iterations is $\tilde{O}\left(\frac{\alpha}{\varepsilon^4}\right)$.

Since $\mathbf{x}$ requires $\Theta(d)$ memory, the total communication is $\tilde{O}\left(\frac{\alpha d}{\varepsilon^4}\right)$.

**Compressed, Lipschitz stochastic gradient.** By selecting $\lambda = 1 - \frac{k}{d}$, where $k = o(d)$, we have $\delta = \Theta\left(\frac{k}{d}\right)$ and $\eta_\lambda = \tilde{O}\left(\min\left(\frac{k\varepsilon}{d}, \frac{k^2\sqrt{\varepsilon}}{d^3}\right)\right)$. 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 $\tilde{O}\left(\frac{k}{\varepsilon^4}+\frac{d}{\varepsilon^3}+\frac{d^3}{k\varepsilon^2\sqrt{\varepsilon}}\right)$.

To balance the first and the third terms: $\frac{k}{\varepsilon^4} = \frac{d^3}{\varepsilon^2\sqrt{\varepsilon}}$, we select $k = d^{3/2}\varepsilon^{3/4}$, which results in total communication being $\tilde{O}\left(\frac{d\sqrt{d}}{\varepsilon^{3+1/4}}\right)$. Compared with communication in unconstrained case, namely $\tilde{O}\left(\frac{d}{\varepsilon^4}\right)$, we achieve $\frac{1}{\varepsilon^{3/4}\sqrt{d}}$ improvement, which improves communication when $\varepsilon = o(d^{-2/3})$.

**Compressed, non-Lipschitz stochastic gradient.** The total number of iterations is $\tilde{O}\left(\frac{d}{\varepsilon^4}+\frac{d}{k\varepsilon^3}+\frac{d^3}{k^2\varepsilon^2\sqrt{\varepsilon}}\right)$ and the total communication is $\tilde{O}\left(\frac{kd}{\varepsilon^4}+\frac{d}{\varepsilon^3}+\frac{d^3}{k\varepsilon^2\sqrt{\varepsilon}}\right)$.

Balancing the first and the third terms, we select $k = d\varepsilon^{3/4}$. The total communication is $\tilde{O}\left(\frac{d^2}{\varepsilon^{3+1/4}}\right)$, which gives improvement of $\varepsilon^{-3/4}$ compared with unconstrained case.

**Good sketch case.** Lemma 2 considers the worst case, which doesn't always arise in practice. Assume that we have a a $c$-compressor $\mathcal{C}$ some for constant $c$ which requires $\tilde{O}(1)$ memory of communication. In this case, the total communication is $\tilde{O}\left(\frac{\alpha}{\varepsilon^4} + \frac{d}{\varepsilon^3}\right)$. In non-Lipschitz case, we obtain $d$ improvement, and in Lipschitz case, we obtain $\min(d, \frac{1}{\varepsilon})$ improvement.

■