SIGNPROX: ONE-BIT PROXIMAL ALGORITHM FOR NONCONVEX STOCHASTIC OPTIMIZATION

Xiaojian Xu[1] *, Ulugbek S. Kamilov*[1,2]

[1] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130

[2] Department of Electrical and Systems Engineering, Washington University in St. Louis, MO 63130

ABSTRACT

Stochastic gradient descent (SGD) is one of the most widely used optimization methods for parallel and distributed processing of large datasets. One of the key limitations of distributed SGD is the need to regularly communicate the gradients between different computation nodes. To reduce this communication bottleneck, recent work has considered a one-bit variant of SGD, where only the sign of each gradient element is used in optimization. In this paper, we extend this idea by proposing a stochastic variant of the proximal-gradient method that also uses one-bit per update element. We prove the theoretical convergence of the method for non-convex optimization under a set of explicit assumptions. Our results indicate that the compressed method can match the convergence rate of the uncompressed one, making the proposed method potentially appealing for distributed processing of large datasets.

Index Terms— Proximal-gradient method, forward-backward algorithm, stochastic gradient descent, nonconvex optimization.

1. INTRODUCTION

Efficient processing of large datasets is a fundamental problem in modern signal processing. In many applications, the task can be formulated as an optimization problem of the form

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{arg\,min}} \{f(\boldsymbol{x})\} \quad \text{with} \quad f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x}), \tag{1}$$

where the data-fidelity term d penalizes mismatch to the data and the regularizer r enforces desirable properties in $\boldsymbol{x}$ such as sparsity or positivity. For a differentiable function f, the solution of (1) can be approximated iteratively with the classical gradient method [1, 2]

$$\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \gamma \nabla f(\boldsymbol{x}^{t-1}), \tag{2}$$

where $\gamma > 0$ is the step size. However, when f consists of a large number of component functions

$$f(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} f_k(\boldsymbol{x}), \tag{3}$$

the cost of computing the full gradient ∇f can become prohibitively expensive. In such cases, it is common to rely on the *stochastic gradient descent (SGD)* [3] that approximates the gradient at every iteration either with that of a single component f_k or with an average of B component gradients as

$$\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \gamma \nabla \widehat{f}(\boldsymbol{x}^{t-1}) \quad \text{with} \quad \nabla \widehat{f}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \nabla f_{k_b}(\boldsymbol{x}), \tag{4}$$

where $k_1, \ldots, k_B$ are independent random variables that are distributed uniformly over $[1 \ldots K]$.

A powerful feature of SGD is that it can be easily parallelized by splitting the computation of B gradients over multiple compute nodes [4]. However, distributed SGD suffers from a significant communication overhead due to the frequent gradient updates transmitted between the nodes. As the size of the gradient scales proportionally to the number of optimization parameters, it can reach hundreds of millions of variables for certain large-scale applications such as 3D imaging [5] or deep learning [6]. Motivated by this problem, recent work has considered a compressed SGD, where the algorithm compresses $\nabla \widehat{f}$ during optimization [7–9]. A particularly simple variant of compressed SGD is signSGD [9], which only keeps the sign of the stochastic gradient at every iteration

$$\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \gamma\, \mathsf{sgn}\Big(\nabla \widehat{f}(\boldsymbol{x}^{t-1})\Big), \tag{5}$$

and hence compresses each stochastic gradient to a single bit. Remarkably, it was shown that under some conditions this simple scheme can match the convergence rate of uncompressed SGD [9].

While the current formulation of signSGD is both conceptually elegant and widely applicable, it does not take advantage of the recent progress in the proximal optimization theory [10]. In many applications, the regularizer r in (1) consists of functions r_k with easily computable proximals [11]

$$\mathsf{prox}_{\gamma r_k}(\boldsymbol{y}) \triangleq \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{arg\,min}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \gamma r_k(\boldsymbol{x}) \right\}, \tag{6}$$

which enables efficient optimization with a class of methods known as proximal algorithms. For example, two widely popular methods for large-scale optimization, FISTA [12–16] and ADMM [17–19], are both examples of proximal algorithms.

In this paper, we propose a novel framework for one-bit stochastic optimization based on the proximal-gradient extension of signSGD. Our method, called signProx, is similar to signSGD in the sense that it also uses only one-bit per element of the update. In fact, we will see that under some conditions signProx is exactly equivalent to signSGD. On the other hand, signProx also enables gradient-free optimization, and hence generalizes signSGD to problems with easily computable proximals. One of the key contribution of this paper is the theoretical analysis of signProx for nonconvex optimization under a set of transparent assumptions. Our analysis and simulations reveal that signProx can converge as fast or faster than the noncompressed algorithm, which makes compressed proximal optimization appealing for processing large datasets.

This material is based upon work supported by the National Science Foundation under Grant No. 1813910.

Algorithm 1 SPGM

input: $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\gamma > 0$, and $B \geq 1$
for $t = 1, 2, \ldots$ **do**
  sample a vector $\boldsymbol{k}$ with i.i.d. elements $k_b \sim \boldsymbol{\theta}$
  $\boldsymbol{x}^t \leftarrow \widehat{\mathsf{P}}_{\boldsymbol{k}}(\boldsymbol{x}^{t-1})$
end for

Algorithm 2 signProx

input: $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\gamma > 0$, and $B \geq 1$
for $t = 1, 2, \ldots$ **do**
  sample a vector $\boldsymbol{k}$ with i.i.d. elements $k_b \sim \boldsymbol{\theta}$
  $\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \gamma\, \mathsf{sgn}(\boldsymbol{x}^{t-1} - \widehat{\mathsf{P}}_{\boldsymbol{k}}(\boldsymbol{x}^{t-1}))$
end for

2. MAIN RESULTS

In this section, we present our main results. We first introduce signProx and then follow up by analyzing its convergence.

2.1. Compressed optimization using proximals

The central building block of our algorithm is the following proximal-gradient mapping

$$\mathsf{P}_k(\boldsymbol{x}) \triangleq \mathsf{prox}_{\gamma r_k}(\boldsymbol{x} - \gamma \nabla d(\boldsymbol{x})), \quad k \in [1, \ldots, K] \tag{7}$$

which computes a gradient-step with respect to the function d and then evaluates the proximal with respect to another function r_k, both with a step-size $\gamma > 0$. Throughout this paper, we will assume that d is a smooth, but possibly nonconvex, function. On the other hand, to have a well-defined proximal, we assume that r_k are all closed, proper, and convex functions. We also define the following convex combination of mappings in (7)

$$\mathsf{P}(\boldsymbol{x}) \triangleq \mathbb{E}[\mathsf{P}_k(\boldsymbol{x})] = \sum_{k=1}^{K} \theta_k \mathsf{P}_k(\boldsymbol{x}), \tag{8}$$

where we can express the sum as an expectation since $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k = 1$. It is known that a convex combination of proximals gives another proximal [20, 21], which means that there exists a closed, proper, and convex function r such that

$$\mathsf{P}(\boldsymbol{x}) = \mathsf{prox}_{\gamma r}(\boldsymbol{x} - \gamma \nabla d(\boldsymbol{x})). \tag{9}$$

Algorithm 1 summarizes a stochastic alternative to the traditional proximal-gradient method [22, 23], which we will call SPGM in this paper. Instead of evaluating the full proximal-gradient step (8), which can be costly for large K, it computes an average proximal-gradient over a mini-batch of size B

$$\widehat{\mathsf{P}}_{\boldsymbol{k}}(\boldsymbol{x}) \triangleq \frac{1}{B} \sum_{b=1}^{B} \mathsf{P}_{k_b}(\boldsymbol{x}) = \mathsf{prox}_{\gamma \widehat{r}}(\boldsymbol{x} - \gamma \nabla d(\boldsymbol{x})), \tag{10}$$

where each $k_b \in [1 \ldots K]$ is sampled independently according to the probability distribution $\boldsymbol{\theta}$ from (8). When the dependence of $\widehat{\mathsf{P}}_{\boldsymbol{k}}$ on $\boldsymbol{k}$ is clear from context, we will sometimes omit the subscript $\boldsymbol{k}$ from the notation as in $\widehat{\mathsf{P}}$. The second equality in (10) is due to the fact that $\widehat{\mathsf{P}}$ is a convex combination of proximals, and hence itself a valid proximal of some convex function $\widehat{r}$.

Algorithm 2 summarizes the main contribution of this paper: one-bit compressed version of SPGM. Similarly to signSGD [9], it requires only a single bit for updating each element of the iterate. However, the update direction at iteration t is given by the sign of the quantity $(\boldsymbol{x}^{t-1} - \widehat{\mathsf{P}}(\boldsymbol{x}^{t-1}))$. The choice of this direction is deliberate as it coincides with the gradient-mapping defined as follows.

Definition 1. For an objective $f(\boldsymbol{x}) = d(\boldsymbol{x}) + r(\boldsymbol{x})$ and a step-size $\gamma > 0$, the *gradient mapping* is defined as the operator

$$\mathsf{G}(\boldsymbol{x}) \triangleq \frac{1}{\gamma}(\boldsymbol{x} - \mathsf{P}(\boldsymbol{x})) = \frac{1}{\gamma}(\boldsymbol{x} - \mathsf{prox}_{\gamma r}(\boldsymbol{x} - \gamma \nabla d(\boldsymbol{x}))), \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

It is common to analyze the convergence of proximal algorithms using the gradient mapping, since $\mathsf{G}(\boldsymbol{x}^*) = 0$ *if and only if* $\boldsymbol{x}^*$ is the critical point of f [24]. Hence, signProx simply uses a one-bit approximation for elements of the gradient mapping at every iteration.

We conclude this section by noting that signProx can also be seen as a generalization of signSGD. Let $d = 0$ and define r_k to be a linear approximation of f_k around $\boldsymbol{x}^{t-1}$

$$r_k(\boldsymbol{x}) = f_k(\boldsymbol{x}^{t-1}) + \nabla f_k(\boldsymbol{x}^{t-1})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^{t-1}). \tag{11}$$

Then, one can verify that the stochastic proximal-gradient iteration (10) reduces to the SGD iteration (4)

$$\widehat{\mathsf{P}}(\boldsymbol{x}^{t-1}) = \frac{1}{B} \sum_{b=1}^{B} \mathsf{P}_{k_b}(\boldsymbol{x}^{t-1}) = \boldsymbol{x}^{t-1} - \gamma \nabla \widehat{f}(\boldsymbol{x}^{t-1}). \tag{12}$$

Which means that the update of signProx will reduce to

$$\boldsymbol{x}^t = \boldsymbol{x}^{t-1} - \gamma \mathsf{sgn}(\boldsymbol{x}^{t-1} - \widehat{\mathsf{P}}(\boldsymbol{x}^{t-1})) = \boldsymbol{x}^{t-1} - \gamma \mathsf{sgn}(\nabla \widehat{f}(\boldsymbol{x}^{t-1})).$$

Hence, signSGD can be interpreted as Algorithm 2 applied to a linear approximation of a function.

2.2. Theoretical analysis

We now discuss the convergence of SPGM and signProx. Convergence of the stochastic proximal-gradient algorithms in the convex setting was analyzed by Bertsekas [23]. Here, we focus on the case where d is nonconvex. Our result for signProx extends the analysis of signSGD in [9] using the theory of proximal optimization.

Assumption 1. We analyze SPGM under the following assumptions:

(a) The objective function f has a finite minimum $f^* = f(\boldsymbol{x}^*)$ attained at some $\boldsymbol{x}^* \in \mathbb{R}^n$.

(b) The function d is differentiable and has a Lipschitz continuous gradient with a constant $L > 0$.

(c) All functions r_k are closed, proper, and convex. We also assume that they have Lipschitz continuous gradients with the same constant $L > 0$.

(d) The proximal-gradient mappings have a bounded variance

$$\mathbb{E}\left[\|\mathsf{P}_k(\boldsymbol{x}) - \mathsf{P}(\boldsymbol{x})\|^2\right] \leq \gamma^2 \sigma^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$

for some constant $\sigma > 0$, where $\gamma > 0$ is the step-size.

All these are standard assumptions used in the analysis of stochastic optimization algorithms. The dependence of the variance (d) on γ might seem surprising; however, note that this comes from the dependence of the proximal-gradient on γ, with $\gamma=0$ implying that $\mathsf{P}_k(\boldsymbol{x})=\mathsf{P}(\boldsymbol{x})$ for all $k\in[1\dots K]$ and $\boldsymbol{x}\in\mathbb{R}^n$.

Theorem 1. Run SPGM for T iterations under Assumption 1 with the step $\gamma=1/(L\sqrt{T})$ and the mini-batch size $B=1$. Then, we have that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathsf{G}(\boldsymbol{x}^{t-1})\|_2^2\right]\leq\frac{1}{\sqrt{T}}\left[2L(f(\boldsymbol{x}^0)-f^*)+3\sigma^2\right].$$

The proof is given in Section 5.1. This establishes that SPGM converges to the critical point of the objective f.

Our analysis of signProx will need the following more elaborate set of assumptions.

Assumption 2. We analyze signProx under the following assumptions:

(a) The objective function f has a finite minimum $f^*=f(\boldsymbol{x}^*)$ attained at some $\boldsymbol{x}^*\in\mathbb{R}^n$.

(b) The function d is differentiable and there exists a nonnegative vector $\boldsymbol{L}\triangleq(L_1,\dots,L_n)$ such that

$$|\nabla d(\boldsymbol{x})_i-\nabla d(\boldsymbol{y})_i|\leq L_i|x_i-y_i|,\ \forall i\in[1\dots n],\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$$

(c) All functions r_k are closed, proper, and convex. We additionally assume that they all satisfy

$$|\nabla r_k(\boldsymbol{x})_i-\nabla r_k(\boldsymbol{y})_i|\leq L_i|x_i-y_i|,\ \forall i\in[1\dots n],\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n.$$

(d) The proximal-gradient mappings have a bounded variance

$$\mathbb{E}\left[(\mathsf{P}_k(\boldsymbol{x})_i-\mathsf{P}(\boldsymbol{x})_i)^2\right]\leq\gamma^2\sigma_i^2,\ \forall i\in[1\dots n],\forall\boldsymbol{x}\in\mathbb{R}^n,$$

for a positive $\boldsymbol{\sigma}\triangleq(\sigma_1,\dots,\sigma_n)$, where $\gamma>0$ is the step-size.

Note (b) and (d) lead to the standard assumption of Lipschitz continuity by defining a Lipschitz constant $L\triangleq\|\boldsymbol{L}\|_\infty$. Similarly, the standard variance bound is recovered by setting $\sigma^2=\|\boldsymbol{\sigma}\|_2^2$. Also note that when the mini-batch size is $B>1$, the variance bound is effectively reduced by B for the mini-batch.

Theorem 2. Run signProx for T iterations under Assumption 2 with the step $\gamma=1/(2\|\boldsymbol{L}\|_1\sqrt{T})$ and the mini-batch size $B=T$. Then, we have that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathsf{G}(\boldsymbol{x}^{t-1})\|_1\right]\leq\frac{4}{\sqrt{T}}\left[\|\boldsymbol{L}\|_1(f(\boldsymbol{x}^0)-f^*)+\|\boldsymbol{\sigma}\|_1+1\right].$$

The proof is given in Section 5.2. One can see that signProx has the same ℓ_1-geometry as signSGD, where the convergence rate depends on the ℓ_1-norm of the gradient mapping, the stochasticity via $\boldsymbol{\sigma}$, and the curvature via $\boldsymbol{L}$. Surprisingly, it is possible for signProx to outperform SPGM, when the gradient-mapping is dense but has a sparse set of extremely noisy components (see the detailed discussion for signSGD in [9]). Our simulations in the next section will highlight this situation by comparing the relative performances of SPGM and signProx for nonconvex phase retrieval. Finally, to conclude this section, note that the theoretical analysis here was done for nonconvex functions f. It would be very interesting to see how convexity of d can strengthen these convergence results. Note that the majority vote in signSGD [9] also directly applies to our algorithm; however, we omit its analysis due to the page limitation.

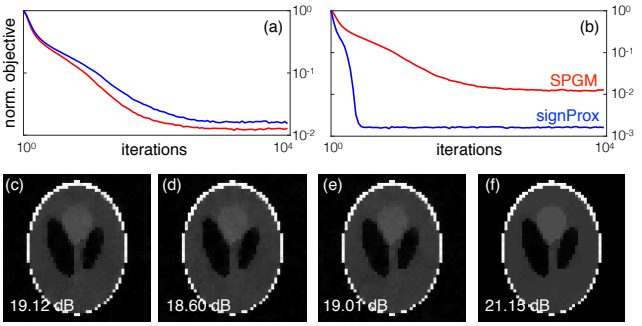

Fig. 1: Comparison of SPGM and signProx on the problem of generalized phase retrieval using TV regularization in two scenarios: with dense (a, c, d) and sparse (b, e, f) stochasticity $\boldsymbol{\sigma}$ of the proximal. The images (c, e) are the outputs of SPGM, while (d, f) are the outputs of signProx. This figure illustrates that in some settings signProx can converge similarly or even faster than SPGM.

3. NUMERICAL ILLUSTRATION

We illustrate the relative performance of SPGM and signProx with a simple example. Consider the problem of generalized phase retrieval that was extensively considered in the literature [25–27]. When the signal is real, the goal is to reconstruct $\boldsymbol{x}\in\mathbb{R}^n$ given a set of nonlinear measurements $\boldsymbol{y}=|\boldsymbol{z}|^2$ with $\boldsymbol{z}=\boldsymbol{H}\boldsymbol{x}\in\mathbb{R}^m$. This problem can be formulated as nonconvex optimization with a data-fidelity term $d(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{y}-|\boldsymbol{H}\boldsymbol{x}|^2\|_2^2$ and a sparsity-preserving regularizer such as total variation (TV) [28–30]. Figure 1 considers the reconstruction of a 50×50 *Shepp-Logan phantom* from $m=3000$ intensity measurements $\boldsymbol{y}$, where the measurement matrix $\boldsymbol{H}$ is random with i.i.d. $\mathcal{N}(0,1/m)$ elements. We obtain a stochastic algorithm by using an unbiased estimate $\widehat{\mathsf{P}}$ for P, obtained by adding a random noise to P distributed i.i.d as $p(e)=\rho\mathcal{G}(e;\sigma_e^2)+(1-\rho)\delta(e)$, where $\mathcal{G}(e;\sigma_e^2)$ denotes the Gaussian pdf of variance σ_e^2 and $\delta(e)$ is the Dirac delta function. By setting $\rho\in(0,1]$, we control the sparsity of the noise in the proximal-gradient, which directly corresponds to shaping the stochasticity $\boldsymbol{\sigma}$ in Assumption 2(d). Intuitively, when there is a sparse set of very noisy updates, the performance of SPGM will be dominated disproportionally by the noise, while the effect on signProx will be reduced as it discards the update amplitude. This is visible in Figure 1, where (a, c, d) correspond to the dense stochasticity scenario ($\rho=1$) and (b, e, f) correspond to the sparse stochasticity scenario ($\rho=0.1$). In both cases, the standard deviation of the noise is kept constant to the product of γ and $\sigma=0.1$. The step-size γ is selected for the best performance. The convergence is quantified with the normalized objective $(f(\boldsymbol{x}^t)-f^*)/(f(\boldsymbol{x}^0)-f^*)$, where f^* was obtained using the full TV reconstruction. One can observe that the convergence rate of SPGM is the same in both settings, while the convergence of signProx is faster when the stochasticity is sparse.

4. CONCLUSION

We have proposed a new signProx algorithm for stochastic optimization. The updates of signProx are compressed as each stochastic update contains a single-bit per element. We have proved the convergence of the method on nonconvex objectives under explicit assumptions. The future work will investigate potential applications and will further strengthen the theoretical analysis presented here.

5. APPENDIX

5.1. Proof of Theorem 1

Consider a single iteration of SPGM $\boldsymbol{x}^+ = \widehat{\mathsf{P}}(\boldsymbol{x}) = \boldsymbol{x} - \gamma \widehat{\mathsf{G}}(\boldsymbol{x})$, where we used the definition of the gradient mapping. Consider also $\widetilde{\boldsymbol{x}} = \mathsf{P}(\boldsymbol{x}) = \boldsymbol{x} - \gamma \mathsf{G}(\boldsymbol{x})$. Note that for $B = 1$, we have that

$$\mathbb{E}[\widehat{\mathsf{P}}(\boldsymbol{x})] = \mathsf{P}(\boldsymbol{x}) \quad \Rightarrow \quad \mathbb{E}[\widehat{\mathsf{G}}(\boldsymbol{x})] = \mathsf{G}(\boldsymbol{x})$$
$$\mathbb{E}[\|\widehat{\mathsf{P}}(\boldsymbol{x}) - \mathsf{P}(\boldsymbol{x})\|_2^2] \leq \gamma^2\sigma^2 \quad \Rightarrow \quad \mathbb{E}[\|\widehat{\mathsf{G}}(\boldsymbol{x}) - \mathsf{G}(\boldsymbol{x})\|_2^2] \leq \sigma^2.$$

We can then obtain the following bound

$$\begin{aligned}
&f(\boldsymbol{x}^+) = d(\boldsymbol{x}^+) + r(\boldsymbol{x}^+) \\
&\leq f(\boldsymbol{x}) + [\nabla d(\boldsymbol{x}) + \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2 \\
&\quad + [\nabla r(\boldsymbol{x}^+) - \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}) \\
&\leq f(\boldsymbol{x}) - \gamma \mathsf{G}(\boldsymbol{x})^\mathsf{T}\widehat{\mathsf{G}}(\boldsymbol{x}) + \frac{\gamma^2 L}{2}\|\widehat{\mathsf{G}}(\boldsymbol{x})\|_2^2 + \gamma^2 L\|\widehat{\mathsf{G}}(\boldsymbol{x}) - \mathsf{G}(\boldsymbol{x})\|_2^2,
\end{aligned}$$

where the first inequality uses the Lipschitz continuity of ∇d and twice the convexity of r, and the second inequality uses the definition of the gradient mappings, Cauchy-Schwarz inequality, and the Lipschitz continuity of ∇r. By taking the conditional expectation and setting $\gamma = 1/(L\sqrt{T})$, we obtain

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}^+) - f(\boldsymbol{x}) \,|\, \boldsymbol{x}] \\
&\leq -\gamma\|\mathsf{G}(\boldsymbol{x})\|_2^2 + \frac{\gamma^2 L}{2}(\|\mathsf{G}(\boldsymbol{x})\|_2^2 + \sigma^2) + \gamma^2 L\sigma^2 \\
&\leq -\frac{1}{L\sqrt{T}}\|\mathsf{G}(\boldsymbol{x})\|_2^2 + \frac{1}{2LT}\|\mathsf{G}(\boldsymbol{x})\|_2^2 + \frac{3\sigma^2}{2LT} \\
&\leq -\frac{1}{2L\sqrt{T}}\|\mathsf{G}(\boldsymbol{x})\|_2^2 + \frac{3\sigma^2}{2LT},
\end{aligned}$$

where the final inequality uses the fact that $1/T \leq 1/\sqrt{T}$ for all $T \geq 1$. By rearranging the terms and summing up the gradient-mapping norms at different iterations, we finally obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathsf{G}(\boldsymbol{x}^{t-1})\|_2^2\right] &\leq \frac{1}{\sqrt{T}}\left[2L(f(\boldsymbol{x}^0) - \mathbb{E}[f(\boldsymbol{x}^T)]) + 3\sigma^2\right] \\
&\leq \frac{1}{\sqrt{T}}\left[2L(f(\boldsymbol{x}^0) - f^*) + 3\sigma^2\right],
\end{aligned}$$

which proves the result.

5.2. Proof of Theorem 2

Consider a single iteration of signProx

$$\boldsymbol{x}^+ = \boldsymbol{x} - \gamma \mathsf{sgn}(\boldsymbol{x} - \widehat{\mathsf{P}}(\boldsymbol{x})) = \boldsymbol{x} - \gamma \mathsf{sgn}(\widehat{\mathsf{G}}(\boldsymbol{x})),$$

and a full proximal-gradient iteration $\widetilde{\boldsymbol{x}} = \mathsf{P}(\boldsymbol{x}) = \boldsymbol{x} - \gamma\mathsf{G}(\boldsymbol{x})$. We can obtain the following bound

$$\begin{aligned}
&f(\boldsymbol{x}^+) = d(\boldsymbol{x}^+) + r(\boldsymbol{x}^+) \\
&\leq d(\boldsymbol{x}) + \nabla d(\boldsymbol{x})^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + \sum_{i=1}^{n}\frac{L_i}{2}(x_i^+ - x_i)^2 \\
&\quad + r(\boldsymbol{x}) + \nabla r(\boldsymbol{x}^+)^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) \\
&= f(\boldsymbol{x}) + [\nabla d(\boldsymbol{x}) + \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + \sum_{i=1}^{n}\frac{L_i}{2}(x_i^+ - x_i)^2 \\
&\quad + [\nabla r(\boldsymbol{x}^+) - \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}).
\end{aligned}$$

We separately bound the last term above as follows

$$\begin{aligned}
&[\nabla r(\boldsymbol{x}^+) - \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) \\
&= [\nabla r(\boldsymbol{x}^+) - \nabla r(\boldsymbol{x})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + [\nabla r(\boldsymbol{x}) - \nabla r(\widetilde{\boldsymbol{x}})]^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) \\
&\leq \sum_{i=1}^{n}\left[|\nabla r(\boldsymbol{x}^+)_i - \nabla r(\boldsymbol{x})_i||x_i^+ - x_i| + |\nabla r(\boldsymbol{x})_i - \nabla r(\widetilde{\boldsymbol{x}})_i||x_i^+ - x_i|\right] \\
&\leq \sum_{i=1}^{n}[L_i(x_i^+ - x_i)^2 + L_i|x_i - \widetilde{x}_i||x_i^+ - x_i|] \\
&= \gamma^2\|\boldsymbol{L}\|_1 + \gamma\sum_{i=1}^{n}L_i|x_i - \widetilde{x}_i| \leq \gamma^2\|\boldsymbol{L}\|_1 + \gamma\|\boldsymbol{L}\|_\infty\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_1 \\
&\leq \gamma^2\|\boldsymbol{L}\|_1 + \gamma^2\|\boldsymbol{L}\|_1\|\mathsf{G}(\boldsymbol{x})\|_1,
\end{aligned}$$

where we used the smoothness assumption on ∇r, the proximal-gradient iterate $\gamma\mathsf{G}(\boldsymbol{x}) = \boldsymbol{x} - \widetilde{\boldsymbol{x}}$, and the fact that $\|\boldsymbol{L}\|_\infty \leq \|\boldsymbol{L}\|_1$. By using this bound in the original inequality, we obtain

$$\begin{aligned}
&f(\boldsymbol{x}^+) - f(\boldsymbol{x}) \leq -\gamma\mathsf{G}(\boldsymbol{x})^\mathsf{T}\mathsf{sgn}(\widehat{\mathsf{G}}) + \frac{3\gamma^2}{2}\|\boldsymbol{L}\|_1 + \gamma^2\|\boldsymbol{L}\|_1\|\mathsf{G}(\boldsymbol{x})\|_1 \\
&= -\gamma\|\mathsf{G}(\boldsymbol{x})\|_1 + \frac{3\gamma^2}{2}\|\boldsymbol{L}\|_1 + \gamma^2\|\boldsymbol{L}\|_1\|\mathsf{G}(\boldsymbol{x})\|_1 \\
&\quad + \gamma\mathsf{G}(\boldsymbol{x})^\mathsf{T}[\mathsf{sgn}(\mathsf{G}(\boldsymbol{x})) - \mathsf{sgn}(\widehat{\mathsf{G}}(\boldsymbol{x}))] \\
&= -\gamma\|\mathsf{G}(\boldsymbol{x})\|_1 + \frac{3\gamma^2}{2}\|\boldsymbol{L}\|_1 + \gamma^2\|\boldsymbol{L}\|_1\|\mathsf{G}(\boldsymbol{x})\|_1 \qquad (13) \\
&\quad + 2\gamma\sum_{i=1}^{n}|\mathsf{G}(\boldsymbol{x})_i|\mathbb{1}[\mathsf{sgn}(\mathsf{G}(\boldsymbol{x})_i) \neq \mathsf{sgn}(\widehat{\mathsf{G}}(\boldsymbol{x})_i)],
\end{aligned}$$

where $\mathbb{1}[\cdot]$ is an indicator function. The expectation of this function can be further bounded as was done in [9]

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}[\mathsf{sgn}(\mathsf{G}(\boldsymbol{x})_i) \neq \mathsf{sgn}(\widehat{\mathsf{G}}(\boldsymbol{x})_i)]] = \mathbb{P}[\mathsf{sgn}(\mathsf{G}(\boldsymbol{x})_i) \neq \mathsf{sgn}(\widehat{\mathsf{G}}(\boldsymbol{x})_i)] \\
&\leq \mathbb{P}[|\widehat{\mathsf{G}}(\boldsymbol{x})_i - \mathsf{G}(\boldsymbol{x})_i| \geq |\mathsf{G}_i(\boldsymbol{x})|] \leq \frac{\mathbb{E}[|\widehat{\mathsf{G}}(\boldsymbol{x})_i - \mathsf{G}(\boldsymbol{x})_i|]}{|\mathsf{G}(\boldsymbol{x})_i|} \\
&\leq \frac{\sqrt{\mathbb{E}[(\widehat{\mathsf{G}}(\boldsymbol{x})_i - \mathsf{G}(\boldsymbol{x})_i)^2]}}{|\mathsf{G}(\boldsymbol{x})_i|} \leq \frac{\sigma_i}{\sqrt{T}|\mathsf{G}(\boldsymbol{x})_i|},
\end{aligned}$$

where in the second row we used probability relaxation and the Markov inequality, in the third we used the Jensen's inequality and the variance bound for the mini-batch of size $B = T$. By plugging this expression back into (13) and taking the conditional expectation

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}^+) - f(\boldsymbol{x})|\boldsymbol{x}] \\
&\leq -\gamma\|\mathsf{G}(\boldsymbol{x})\|_1 + \frac{3\gamma^2}{2}\|\boldsymbol{L}\|_1 + \gamma^2\|\boldsymbol{L}\|_1\|\mathsf{G}(\boldsymbol{x})\|_1 + \frac{2\gamma}{\sqrt{T}}\|\boldsymbol{\sigma}\|_1 \\
&\leq -\frac{\|\mathsf{G}(\boldsymbol{x})\|_1}{2\|\boldsymbol{L}\|_1\sqrt{T}} + \frac{\|\mathsf{G}(\boldsymbol{x})\|_1}{4\|\boldsymbol{L}\|_1 T} + \frac{3}{8\|\boldsymbol{L}\|_1 T} + \frac{\|\boldsymbol{\sigma}\|_1}{\|\boldsymbol{L}\|_1 T} \\
&\leq -\frac{\|\mathsf{G}(\boldsymbol{x})\|_1}{4\|\boldsymbol{L}\|_1\sqrt{T}} + \frac{3}{8\|\boldsymbol{L}\|_1 T} + \frac{\|\boldsymbol{\sigma}\|_1}{\|\boldsymbol{L}\|_1 T},
\end{aligned}$$

where in the second line we set the step-size to $\gamma = 1/(2\|\boldsymbol{L}\|_1\sqrt{T})$. By rearranging the terms and summing up the gradient-mapping norms at different iterations, we finally obtain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathsf{G}(\boldsymbol{x}^{t-1})\|_1\right] \leq \frac{4}{\sqrt{T}}\left[\|\boldsymbol{L}\|_1(f(\boldsymbol{x}^0) - f^*) + \|\boldsymbol{\sigma}\|_1 + 1\right],$$

which completes the proof.

6. REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, 2004.

[2] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, 2004.

[3] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, September 1951.

[4] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server," in *Symposium on Operating Systems Design and Implementation (OSDI-14)*, Broomfield, CO, USA, October 06-08, 2014, pp. 583–598.

[5] U. S. Kamilov, I. N. Papadopoulos, M. H. Shoreh, A. Goy, C. Vonesch, M. Unser, and D. Psaltis, "Optical tomographic image reconstruction based on beam propagation and sparse regularization," *IEEE Trans. Comp. Imag.*, vol. 2, no. 1, pp. 59–70,, March 2016.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 21-26, 2017, pp. 2261–2269.

[7] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its applications to data-parallel distributed training of speech DNNs," in *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, September 14-18, 2014, pp. 1058–1062.

[8] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, December 4-9, 2017.

[9] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGN: Compressed optimization for non-convex problems," in *Proc. 35th Int. Conf. Machine Learning (ICML)*, Stockholm, Sweden, July 2018.

[10] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2014.

[11] J. J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.

[12] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, August 2003.

[13] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, "A ℓ_1-unified variational framework for image restoration," in *Proc. ECCV*, Springer, Ed., New York, 2004, vol. 3024, pp. 1–13.

[14] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, November 2004.

[15] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, December 2007.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[17] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[18] M. K. Ng, P. Weiss, and X. Yuan, "Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2710–2736, August 2010.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[20] H. H. Bauschke, R. Goebel, Y. Lucet, and X. Wang, "The proximal average: Basic theory," *SIAM J. Optim.*, vol. 19, no. 2, pp. 766–785, 2008.

[21] Y. Yu, "Better approximation and faster algorithm using the proximal average," in *Neural Information Processing Systems (NIPS)*, Lake Tahoe, CA, USA, December 5-10, 2013, pp. 458–466.

[22] A. Beck and M. Teboulle, *Convex Optimization in Signal Processing and Communications*, chapter Gradient-Based Algorithms with Applications to Signal Recovery Problems, pp. 42–88, Cambridge, 2009.

[23] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program. Ser. B*, vol. 129, pp. 163–195, 2011.

[24] A. Beck, *First-Order Methods in Optimization*, MOS-SIAM Series on Optimization. SIAM, 2017.

[25] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, no. 1, Feb. 2013.

[26] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, October 2012.

[27] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 87–109, May 2015.

[28] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1–4, pp. 259–268, November 1992.

[29] A. Beck and M. Teboulle, "Fast gradient-based algorithm for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, November 2009.

[30] U. S. Kamilov, "A parallel proximal algorithm for anisotropic total variation minimization," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 539–548, February 2017.