Convergence diagnostic for communication efficient SGD

Jerry Chee and Ping Li

Cognitive Computing Lab, Baidu USA

July 25, 2019

Abstract

Convergence detection in stochastic iterative optimization methods remains more art than science. As training of machine learning algorithms becomes ever more resource intensive with larger data sets and greater computing power, it becomes even more important to determine when such methods should terminate. In this paper we focus on signSGD, a variant of SGD where the gradient values are truncated to just their sign components. There has been much recent work in gradient compression to reduce the communication of distributed SGD implementations. We provide a non-asymptotic convergence analysis to show that signSGD moves from a transient phase where iterates quickly remove their dependence on initial conditions, to a stationary phase of bounded radius around a minimum point. To detect this phase transition we present a principled statistical convergence diagnostic for signSGD. We provide theoretical and empirical evidence supporting that the proposed diagnostic works well.

Keywords: Convergence detection, signSGD, gradient compression

1 Introduction

The future of machine learning will require even larger models to train on even larger data sets. Stochastic gradient descent (SGD) is the workhorse training algorithm. It was originally developed as a serial algorithm, and much recent work has been done to speedup SGD through parallelization. A popular approach has been to use a parameter-server framework to split data amongst workers and individually compute gradient updates which much be gathered, averaged, and then re-distributed. This gradient communication can be prohibitive to increased training speeds. To alleviate this, there has been recent interest in gradient compression to reduce the communication overhead.

One simple way to compress the gradient is to take the sign of each element. These methods have been studied and were called 1-bit methods for speech models, and more recently have been called signSGD. This gradient compression is easy to implement, and does not suffer from the variance explosions of other unbiased gradient compression schemes. However, the sign of the gradient is no longer an unbiased estimate of the true gradient. Bernstein et al. 2018 presented a convergence analysis under a non-convex loss. We focus on the (strongly) convex case; at some point signSGD will enter a local minima which is (strongly) convex. We present a non-asymptotic convergence analysis under (strongly) convex loss. This analysis will show the existence of a transient and stationary phase for signSGD. We focus on a constant learning rate as this is commonly used in practice, and it makes the transition from transient to stationary phase explicitly clear. For constant rate signSGD, in the transient phase it moves quickly towards the minimum point to forget its initial conditions. In the stationary phase, signSGD with constant rate remains in a bounded region around a local minima. Thus once the stationary region has been reached, it is pointless to keep running.

Inspiration is drawn from the work of Chee and Toulis (2018), which present a convergence diagnostic for vanilla SGD. At a high level, the idea is that in the transient phase successive gradients are likely to be pointing in the same direction, and thus the inner product of successive gradients is likely to be positive on average. When the stationary phase is reached, iterates are likely to be oscillating around the minimum, and thus the inner product of successive gradients is likely to be negative on average. Though the gradient information has been compressed to just the sign, we believe that the same intuition holds.

1.1 Related work

Error correction. Maybe say its ok doesn't converge because use constant learning rate?

1.2 Our contributions

Convexity is a key assumption behind our analysis. Intuitively, our proposed diagnostic detects when signSGD is oscillating as a result of the curvature of the loss around a minimum point. Even for non-convex loss, it is assumed that there are convex regions around the minimum points. We present a non-asymptotic convergence analysis for signSGD under (strongly) convex loss. We present a principled statistical convergence diagnostic for signSGD. It keeps a running mean of the inner product between successive gradients, and activates when less than zero.

2 Convergence analysis

Consider the stochastic optimization problem:

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} f(\theta) = \mathbb{E}[\ell(\theta, \xi)] \tag{1}$$

The n-th estimate of $\theta_\star$ is:

$$\theta_n = \theta_{n-1} - \gamma sign(\nabla \ell(\theta_{n-1}, \xi_n)) \tag{2}$$

We consider a class of loss functions slightly more general than convex.

Assumption 2.1. *Let f satisfy for all x, y*

$$\Big(sign(\nabla f(y)) - sign(\nabla f(x))\Big)^\top (y - x) \geq 0$$

At the minimum point $f_\star$, we define $sign(\nabla f_\star) = 0$.

This class of functions has a unique minimum, but does not have to be convex. For example, consider the sin function on the bounded interval $(-1.5\pi, 0.5\pi)$. This function is not convex, and yet it satisfies Assumption 2.1. We can view the sign of the gradient as imposing a convex structure upon functions which have a unique minimum, but are not convex.

Theorem 2.2. *Let f satisfy Assumption 2.1. The distance of the iterates from $\theta_\star$ is bounded by $\|\theta_0 - \theta_\star\|^2$. The learning rate satisfies $\gamma < 1/2\mu_{\theta_0}$ for some $\mu_{\theta_0} > 0$. Then,*

$$\mathbb{E}[\|\theta_n - \theta_\star\|^2] \leq (1 - 2\gamma\mu_{\theta_0})^n \ \mathbb{E}[\|\theta_0 - \theta_\star\|^2] + \frac{d(\gamma + 4)}{2\mu_{\theta_0}}$$

Proof. We follow the approach taken by Bach and Moulines 2011, and adapt it for signSGD. For brevity of notation let $\nabla \ell_{n-1} \equiv \nabla \ell(\theta_{n-1}, \xi_n)$. First derive a recursive relation for $\|\theta_n - \theta_\star\|^2$. Unless otherwise noted, $\|\cdot\|$ represents the L2 norm.

$$\begin{aligned}
\theta_n - \theta_\star &= \theta_{n-1} - \theta_\star - \gamma sign(\nabla \ell_{n-1}) \\
\|\theta_n - \theta_\star\|^2 &= \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma sign(\nabla \ell_{n-1})^\top (\theta_{n-1} - \theta_\star) + \gamma^2 \|sign(\nabla \ell_{n-1})\|^2 \\
\mathbb{E}[\|\theta_n - \theta_\star\|^2] &= \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma \mathbb{E}[sign(\nabla \ell_{n-1})]^\top (\theta_{n-1} - \theta_\star) + \gamma^2 \mathbb{E}[\|sign(\nabla \ell_{n-1})\|^2]
\end{aligned}$$

The third term is equal to $\gamma^2 d$. In Bach and Moulines 2011 this term is bounded with a bound on the variance of the stochastic gradient. However this is not needed for signSGD as the gradients only contain the sign information. To bound the second term, we use a technique from Bernstein et al 2018 to deal with the fact that the sign gradient is now a biased estimate of the true gradient. First decompose the gradient estimate. [Do I use the dot product or element-wise sum for the second term bellow?]

$$\begin{aligned}\mathbb{E}[sign(\nabla \ell_{n-1})]^\top(\theta_{n-1}-\theta_\star) &= sign(\nabla f(\theta_{n-1}))^\top(\theta_{n-1}-\theta_\star)\\ &\quad - 2sign(\nabla f(\theta_{n-1}))^\top\ \mathbb{P}[sign(\nabla \ell_{n-1}) \neq sign(\nabla f(\theta_{n-1}))]\end{aligned}$$

The probability represents a vector where the i-th element is equal to $\mathbb{P}[sign(\nabla \ell_{n-1})_i \neq sign(\nabla f(\theta_{n-1}))_i]$. By Assumption 2.1, there exists constant $\mu_{\theta_0} > 0$ such that $sign(f(\theta_{n-1}))^\top(\theta_{n-1} - \theta_\star) \geq \mu_{\theta_0}\|\theta_{n-1}-\theta_\star\|^2$. We can rewrite the term that captures the amount that the sign of the gradient estimate is incorrect, and use Markov's inequality.

$$\begin{aligned}\mathbb{P}[sign(\nabla \ell_{n-1})_i \neq sign(\nabla f(\theta_{n-1})_i)] &= \mathbb{P}[|sign(\nabla \ell_{n-1})_i - sign(\nabla f(\theta_{n-1})_i)| \geq 1]\\ &\leq \mathbb{E}[|sign(\nabla \ell_{n-1})_i - sign(\nabla f(\theta_{n-1})_i)|]\\ &\leq 2\end{aligned}$$

[Could instead bound with $\sqrt{\mathbb{E}[|sign(\nabla \ell_{n-1})_i - sign(\nabla f(\theta_{n-1})_i)|^2]} \leq \sigma^2$] Combining these bounds and taking expectation on both sides we get the following recursive relation:

$$\begin{aligned}\mathbb{E}[\|\theta_n - \theta_\star\|^2] &\leq (1-2\gamma\mu_{\theta_0})\ \mathbb{E}[\|\theta_{n-1}-\theta_\star\|^2] + d(\gamma^2 + 4\gamma)\\ &= (1-2\gamma\mu_{\theta_0})^n\ \mathbb{E}[\|\theta_0-\theta_\star\|^2] + d(\gamma^2+4\gamma)\sum_{i=0}^{n-1}(1-2\gamma\mu_{\theta_0})^i\\ &\leq (1-2\gamma\mu_{\theta_0})^n\ \mathbb{E}[\|\theta_0-\theta_\star\|^2] + \frac{d(\gamma+4)}{2\mu_{\theta_0}}\end{aligned}$$

□

Remarks. μ_{θ_0} is similar to the strong convexity parameter, however it also depends on the initial conditions. It is reasonable to assume that the distance from $\theta_\star$ as the initial parameter θ_0 begins a fixed distance from $\theta_\star$, and the procedure would only have unbounded iterates if it diverged. This convergence analysis is not sensitive to the lipschitz parameter of the gradients, or to the noise level of the gradients. The dimension in the bound is not prohibitive as the L2 norm is proportional to the dimension d.

Theorem 2.2 supports the existence of the transient and stationary phase for signSGD under functions described by Assumption 2.1, which includes convex functions. The first

Algorithm 1: Convergence diagnostic for signSGD

```
Input : Learning rate $\gamma$, Data-set $D$ with $N$ data points $(x_n, y_n)$, Initial point $\theta$,
        Burnin period burnin, Initial point $\theta_0$.
$\theta \leftarrow \theta_0$
S $\leftarrow 0$
while Not converged do
    for $n$ in 1 to $N$ do
        $\theta \leftarrow \theta - \gamma sign(\nabla\ell(\theta, \xi_n))$
        if burnin done then
            S $\leftarrow S + \langle \nabla\ell(\theta, \xi_n), \nabla\ell(\theta, \xi_{n-1}) \rangle$
            if $S < 0$ then
                return $\theta$
```

term of the bound dominates in the transient phase when the initial conditions are forgotten exponentially fast, this is when the bias dominates. The second term dominates in the stationary phase when signSGD remains trapped in a region of radius $O(\sqrt{d(\gamma + 4)/2\mu_{\theta_0}})$. We also see the tradeoff for the constant learning rate which has been widely observed, such as in Bach and Moulies 2011. A higher learning rate increases the rate at which initial conditions are forgotten. However, this enacts a tradeoff where the radius of the stationary phase is much larger.

While Theorem 2.2 provides theoretical insights into the runtime behavior of signSGD, it cannot be practically used. The data dependent constant μ_{θ_0} is too difficult to estimate reliably.

3 Convergence diagnostic

A theoretical convergence analysis helpful for understanding, but is difficult to use in practical scenarios because of the need to estimate data-dependent constants.

Theorem 3.1. *Let f satisfy Assumption 2.1. [Need to be careful difference between $\|\theta_n - \theta_\star\|^2$ and $\|\theta_{n-1} - \theta_n\|^2$.] Suppose that Theorem 2.2 holds, such that $\mathbb{E}[\|\theta_{n-1} - \theta_n\|^2] \leq \gamma M$ for some positive M and large enough n. Let $\delta_\star > 0$ such that $\mathbb{E}[f(\theta_n) - f(\theta_\star)] \leq \delta_\star$ for large enough n. It holds that $\gamma < (2d - \delta_\star)/M$ where d is the dimension. Then,*

$$\frac{1}{\gamma}\mathbb{E}[sign(\nabla\ell(\theta_n, \xi_{n+1}))]^\top(\theta_{n-1} - \theta_n) < 0$$

Proof. Re-arrange the update in Equation 2 to get $sign(\nabla\ell(\theta_{n-1}, \xi_n)) = \frac{1}{\gamma}(\theta_{n-1} - \theta_n)$. We use this to rewrite the inner product

$$sign(\nabla\ell(\theta_n, \xi_{n+1}))^\top sign(\nabla\ell(\theta_{n-1}, \xi_n)) = \frac{1}{\gamma} sign(\nabla\ell(\theta_n, \xi_{n+1}))^\top(\theta_{n-1} - \theta_n) \tag{3}$$

Apply expectation to both sides of Equation 3, we decompose the expectation into the expected value and offset.

$$\begin{aligned}\frac{1}{\gamma}\mathbb{E}[sign(\nabla\ell(\theta_n,\xi_{n+1}))]^\top(\theta_{n-1}-\theta_n) &= \frac{1}{\gamma}sign(\nabla f(\theta_n))^\top(\theta_{n-1}-\theta_n) \\ &\quad -\frac{2}{\gamma}\sum_{i=1}^{d} sign(\nabla f(\theta_n)_i)\ \mathbb{P}[sign(\nabla\ell(\theta_n,\xi_{n+1}))_i \neq sign(\nabla f(\theta_n)_i)]\end{aligned} \tag{4}$$

From Assumption 2.1 it follows that for all $y, x \in \mathbb{R}^d$, $f(y) \geq f(x) + sign(\nabla f(x))^\top(y - x) - \|y-x\|^2$. In addition, the second term in Equation 4 can be bounded bellow by $-2d/\gamma$. Applying these bounds,

$$\begin{aligned}\frac{1}{\gamma}\mathbb{E}[sign(\nabla\ell(\theta_n,\xi_{n+1}))]^\top(\theta_{n-1}-\theta_n) &\leq \frac{1}{\gamma}[f(\theta_{n-1}) - f(\theta_n) + \|\theta_{n-1}-\theta_n\|^2] - \frac{2d}{\gamma} \\ &\leq \frac{1}{\gamma}[f(\theta_{n-1}) - f(\theta_\star) + \|\theta_{n-1}-\theta_n\|^2] - \frac{2d}{\gamma}\end{aligned}$$

The second inequality is due to the assumption that $f(\theta_\star)$ is the minimum point. Now apply expectation to both sides, and apply the assumption bounds.

$$\begin{aligned}\frac{1}{\gamma}\mathbb{E}[sign(\nabla\ell(\theta_n,\xi_{n+1}))]^\top(\theta_{n-1}-\theta_n) &\leq \frac{1}{\gamma}[\delta_\star + \gamma M - 2d] \\ &< 0\end{aligned}$$

By our condition on γ, the bound is negative.

□

4 Quadratic loss

5 GLM?

6 Experiments

Try it on CIFAR10, see if behaves better with sign gradient.