A Non-asymptotic comparison of SVRG and SGD: tradeoffs between compute and speed

Qingru Zhang, Yuhuai Wu, Fartash Faghri, Tianzong Zhang, Jimmy Ba

Vector Institute

2019-10-6

Contents

1. Introduction
2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model
3. The Dynamic of SGD
4. A Dilemma For SVRG
5. Simulations on Noisy least squares Regression Model
6. Experiments on Benchmark Datasets

Contents

1. Introduction

2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model

3. The Dynamic of SGD

4. A Dilemma For SVRG

5. Simulations on Noisy least squares Regression Model

6. Experiments on Benchmark Datasets

Contribution

Summary of our contributions:

- We show the exact expected loss of SVRG and SGD as a function of iterations and computational cost.
- We discuss the trade-offs between the total computational cost and convergence performance.
- We consider two different training regimes with and without label noise.
  - i Under noisy labels, the analysis suggests SGD only outperforms SVRG under a mild total computational cost.
  - ii However, SGD always exhibits a faster convergence compared to SVRG when there is no label noise.
- Numerical experiments validate our theoretical predictions on both MNIST and CIFAR-10.
  - i In particular, the comparison on underparameterized neural networks closely matches with our noisy least squares model prediction.
  - ii Whereas, the effect of overparameterization is captured by the regression model without label noise.

Contents

1. Introduction
2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model
3. The Dynamic of SGD
4. A Dilemma For SVRG
5. Simulations on Noisy least squares Regression Model
6. Experiments on Benchmark Datasets

SVRG Algorithm

SVRG: inner-outer loop algorithm.

In the outer loop:

- For every T steps, we evaluate a large batch gradient $\bar{\mathbf{g}} = \frac{1}{N}\sum_i^N \nabla_{\boldsymbol{\theta}^{(mT)}} L_i$, where $N \gg b$, and m is the outer loop index.
- We store the parameters at reference points $\boldsymbol{\theta}^{(mT)}$.

In the inner loop:

$$\boldsymbol{\theta}^{(mT+t+1)} = \boldsymbol{\theta}^{(mT+t)} - \alpha^{(t)} \left(\hat{\mathbf{g}}^{(mT+t)} - \tilde{\mathbf{g}}^{(mT+t)} + \bar{\mathbf{g}} \right) \tag{1}$$

where $\hat{\mathbf{g}}^{(mT+t)} = \frac{1}{b}\sum_i^b \nabla_{\boldsymbol{\theta}^{(mT+t)}} L_i$ is the current batch gradient and $\tilde{\mathbf{g}}^{(mT+t)} = \frac{1}{b}\sum_i^b \nabla_{\boldsymbol{\theta}^{(mT)}} L_i$ is the old gradient.

The noisy least squares regression model

Our model:
The input data is d-dimensional, and the output label is generated by a linear teacher model with additive noise,

$$(\boldsymbol{x}_i, \epsilon_i) \sim P_x \times P_\epsilon; \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^* + \epsilon_i,$$

where $\mathbb{E}[\boldsymbol{x}_i] = \boldsymbol{\mu} \in \mathbb{R}^d$ and $\mathrm{Cov}(\boldsymbol{x}_i) = \Sigma$, $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma_y^2$.

Assumptions:

- $\boldsymbol{\mu} = \mathbf{0}$.
- Σ is diagonal.
- $\boldsymbol{\theta}^* = \mathbf{0}$

Objective Function:

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) := \mathbb{E}\left[\frac{1}{2}(\boldsymbol{x}_i^\top \boldsymbol{\theta} - y_i)^2\right]. \qquad (2)$$

The Variance-Bias Decomposition

Under our assumptions, the **expected loss** can be simplified as a function of the second moment of the iterate,

$$L(\boldsymbol{\theta}^{(t)}) = \frac{1}{2}\mathbb{E}\left[\left(\boldsymbol{x}_i^\top \boldsymbol{\theta}^{(t)} - \epsilon_i\right)^2\right] = \frac{1}{2}\left(tr(\Sigma\mathbb{E}[\boldsymbol{\theta}^{(t)}\boldsymbol{\theta}^{(t)\top}]) + \sigma_y^2\right).$$

Hence for the following analysis we mainly focus on deriving the dynamics of the second moment $\mathbb{E}[\boldsymbol{\theta}^{(t)}\boldsymbol{\theta}^{(t)\top}]$, denoted as $\mathrm{M}(\boldsymbol{\theta}^{(t)})$.

When Σ is diagonal, the loss can further be reduced to $\frac{1}{2}\mathrm{diag}(\Sigma)^\top \mathrm{diag}(\mathbb{E}[\boldsymbol{\theta}^{(t)}\boldsymbol{\theta}^{(t)\top}]) + \frac{1}{2}\sigma_y^2$.

We denote $\mathrm{diag}(\mathbb{E}[\boldsymbol{\theta}^{(t)}\boldsymbol{\theta}^{(t)\top}])$ by $\mathbf{m}(\boldsymbol{\theta}^{(t)})$.

Contents

The update rule of SGD

Min-batch Gradient:

$$\hat{\boldsymbol{g}}^{(t)} = \frac{1}{b}\sum_{i}^{b}(\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\theta}^{(t)} - \boldsymbol{x}_i\epsilon_i) = X_bX_b^\top\boldsymbol{\theta}^{(t)} - \frac{1}{\sqrt{b}}X_b\boldsymbol{\epsilon}_b, \tag{3}$$

where $X_b = \frac{1}{\sqrt{b}}[\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots ; \boldsymbol{x}_b] \in \mathbb{R}^{d\times b}$, and the noise vector $\boldsymbol{\epsilon}_b = [\epsilon_1; \epsilon_2; \cdots ; \epsilon_b]^\top \in \mathbb{R}^b$.

SGD Update Rule:

$$\boldsymbol{\theta}^{(t+1)} = (\mathrm{I} - \alpha X_bX_b^\top)\boldsymbol{\theta}^{(t)} + \frac{\alpha}{\sqrt{b}}X_b\boldsymbol{\epsilon}_b.$$

Pre Definition

Definition (Formula for dynamics)

We define the following functions and identities,

$$
\begin{aligned}
&\mathrm{M}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\top], \quad \mathbf{m}(\boldsymbol{\theta}) = \mathrm{diag}(\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]), \\
&\mathrm{C}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\theta}\boldsymbol{\theta}^\top\boldsymbol{x}\boldsymbol{x}^\top] - \Sigma\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]\Sigma, \\
&V = \alpha^2\sigma_y^2\mathrm{diag}(\Sigma), \quad R = (\mathrm{I} - \alpha\Sigma)^2 + \frac{\alpha^2}{b}(\Sigma^2 + \mathrm{diag}(\Sigma)\mathrm{diag}(\Sigma)^\top), \\
&Q = \frac{2\alpha^2}{b}(\Sigma^2 + \mathrm{diag}(\Sigma)\mathrm{diag}(\Sigma)^\top), \quad P = \mathrm{I} - \alpha\Sigma, \\
&F = \frac{2\alpha^2(N+b)}{Nb}(\Sigma^2 + \mathrm{diag}(\Sigma)\mathrm{diag}(\Sigma)^\top).
\end{aligned}
$$

The Dynamic of SGD

The Dynamic of second moment of the iterate:

$$\mathrm{M}(\boldsymbol{\theta}^{(t+1)}) = \underbrace{(\mathrm{I}-\alpha\Sigma)\mathrm{M}(\boldsymbol{\theta}^{(t)})(\mathrm{I}-\alpha\Sigma)}_{\textcircled{1}:\ \text{gradient descent shrinkage}} + \underbrace{\frac{\alpha^2}{b}\mathrm{C}(\boldsymbol{\theta}^{(t)})}_{\textcircled{2}:\ \text{input noise}} + \underbrace{\frac{\alpha^2\sigma_y^2}{b}\Sigma}_{\textcircled{3}:\ \text{label noise}} \tag{4}$$

Analysis:

- The term $\textcircled{1}$ leads to an exponential shrinkage of the loss due to the gradient descent update.
- Since we are using a noisy gradient, the second term $\textcircled{2}$ represents the variance of stochastic gradient caused by the random input X_b.
- The term $\textcircled{3}$ comes from the label noise ϵ_b.

The Expected Second Moment

Theorem (SGD Dynamics and Decay Rate)

Given the noisy linear regression objective function (Eq. 2), under the assumption that $x \sim \mathcal{N}(\mathbf{0}, \Sigma)$ *with* Σ *diagonal and* $\boldsymbol{\theta}^* = 0$*, we can express* $\mathrm{C}(\boldsymbol{\theta})$ *as a function of* $\mathbf{m}(\boldsymbol{\theta})$*:*

$$diag(\mathrm{C}(\boldsymbol{\theta})) = (\Sigma^2 + diag(\Sigma) diag(\Sigma)^\top) \cdot \mathbf{m}(\boldsymbol{\theta}) \tag{5}$$

Then we derive following dynamics of expected second moment of $\boldsymbol{\theta}$*:*

$$\mathbf{m}(\boldsymbol{\theta}^{(t)}) = R^t \left(\mathbf{m}(\boldsymbol{\theta}^{(0)}) - \frac{V}{b(\mathrm{I} - R)} \right) + \frac{V}{b(\mathrm{I} - R)},$$

where

$$V = \alpha^2 \sigma_y^2 diag(\Sigma), \quad R = (\mathrm{I} - \alpha\Sigma)^2 + \frac{\alpha^2}{b} (\Sigma^2 + diag(\Sigma) diag(\Sigma)^\top),$$

Contents

1. Introduction
2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model
3. The Dynamic of SGD
4. A Dilemma For SVRG
5. Simulations on Noisy least squares Regression Model
6. Experiments on Benchmark Datasets

The Update Rule of SVRG

SVRG Update:

$$\boldsymbol{\theta}^{(mT+t+1)} = \boldsymbol{\theta}^{(mT+t)} - \alpha^{(t)} \left(\hat{\mathbf{g}}^{(mT+t)} - \tilde{\mathbf{g}}^{(mT+t)} + \bar{\mathbf{g}} \right) \tag{6}$$

where $\hat{\mathbf{g}}^{(mT+t)} = \frac{1}{b} \sum_i^b \nabla_{\boldsymbol{\theta}^{(mT+t)}} L_i$ is the current batch gradient and $\tilde{\mathbf{g}}^{(mT+t)} = \frac{1}{b} \sum_i^b \nabla_{\boldsymbol{\theta}^{(mT)}} L_i$ is the old gradient.

Thus, we have

$$\begin{aligned} \boldsymbol{\theta}^{(mT+t+1)} = & \left(\mathrm{I} - \alpha X_b X_b^\top \right) \boldsymbol{\theta}^{(mT+t)} + \alpha \left(X_b X_b^\top - X_N X_N^\top \right) \boldsymbol{\theta}^{(mT)} \\ & + \frac{\alpha}{\sqrt{N}} X_N \epsilon_N \end{aligned} \tag{7}$$

The Dilemma for SVRG

Lemma (The Dynamic of SVRG)

The dynamics of the second moment of the iterate following SVRG update rule is given by,

$$\mathrm{M}(\boldsymbol{\theta}^{(mT+t+1)}) = \underbrace{(I-\alpha\Sigma)\mathrm{M}(\boldsymbol{\theta}^{(mT+t)})(\mathrm{I}-\alpha\Sigma)}_{\text{① \textit{gradient descent shrinkage}}} + \underbrace{\frac{\alpha^2}{b}\mathrm{C}(\boldsymbol{\theta}^{(mT+t)})}_{\text{② \textit{input noise}}} \tag{8}$$

$$+ \underbrace{\frac{\alpha^2\sigma_y^2}{N}\Sigma}_{\text{③ \textit{label noise}}} + \underbrace{\alpha^2\frac{N+b}{Nb}\mathrm{C}(\boldsymbol{\theta}^{(mT)})}_{\text{④\textit{variance due to} } \tilde{\mathbf{g}}^{(mT+t)}} \tag{9}$$

$$- \underbrace{\frac{\alpha^2}{b}\left(\mathrm{C}(\boldsymbol{\theta}^{(mT)})\left(\mathrm{I}-\alpha\Sigma\right)^t + \left(\mathrm{I}-\alpha\Sigma\right)^t\mathrm{C}(\boldsymbol{\theta}^{(mT)})\right)}_{\text{⑤ \textit{Variance reduction from control variate}}}.$$

The Delimma of SVRG

Conclusions:

- First notice that terms ①, ②, ③ reappear, contributed by the SGD update.
- The additional terms, ④ and ⑤, are due to the control variate.
- Observe that the variance reduction term ⑤ decays exponentially throughout the inner loop, with decay rate $\mathrm{I} - \alpha\Sigma$, which is the same term that governs the decay rate of the term ①, hence resulting in a conflict between the two.
- If we want to reduce the term ① as fast as possible, we would prefer a large learning rate, i.e. $\alpha \to \frac{1}{\lambda_{\mathsf{max}}(\Sigma)}$. But this will also make the boosts provided by the control variate diminish rapidly, leading to a poor variance reduction.

The Delimma of SVRG

- The term ④ makes things even worse as it will maintain as a constant throughout the inner loop, contributing to an extra variance on top of the variance from standard SGD.
- On the other hand, if one chooses a small learning rate for the variance reduction to take effect, this inevitably will make the decay rate for term ① smaller, resulting in a slower convergence.
- A good news for SVRG is that the label noise (term ③) is scaled by $\frac{b}{N}$, which lets SVRG converge to a lower loss value than SGD – a strict advantage of SVRG compared to SGD.

The Expected Second Moment of SVRG

Theorem (SVRG Dynamics and Decay ate)

Given the noisy linear regression objective function (Eq. 2), under the assumption that $x \sim \mathcal{N}(\mathbf{0}, \Sigma)$ *with* Σ *diagonal and* $\boldsymbol{\theta}^* = 0$*, the dynamics for SVRG in* $\mathbf{m}(\boldsymbol{\theta})$ *is given by:*

$$\mathbf{m}(\boldsymbol{\theta}^{((m+1)T)}) = \lambda(\alpha, b, T, N, \Sigma)\mathbf{m}(\boldsymbol{\theta}^{(mT)}) + \frac{\mathrm{I} - R^T}{\mathrm{I} - R}\frac{V}{N}, \tag{10}$$

$$\lambda(\alpha, b, T, N, \Sigma) = R^T - \Big(\sum_{k=0}^{T-1} R^k (P^{-1})^k\Big) P^{T-1} Q \tag{11}$$

$$+ (\mathrm{I} - R^T)(\mathrm{I} - R)^{-1} F. \tag{12}$$

where

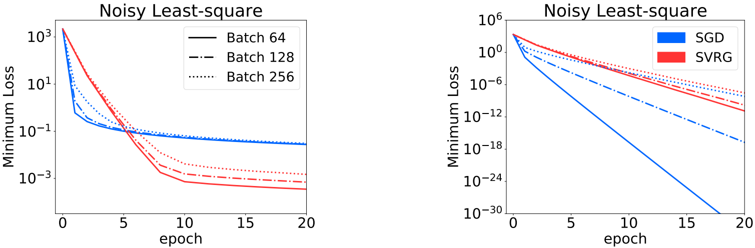
$$Q = \frac{2\alpha^2}{b}(\Sigma^2 + diag(\Sigma)diag(\Sigma)^\top), \quad P = \mathrm{I} - \alpha\Sigma,$$

$$F = \frac{2\alpha^2(N+b)}{Nb}(\Sigma^2 + diag(\Sigma)diag(\Sigma)^\top).$$

Contents

1. Introduction
2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model
3. The Dynamic of SGD
4. A Dilemma For SVRG
5. Simulations on Noisy least squares Regression Model
6. Experiments on Benchmark Datasets

Simulations on Noisy least squares Regression Model



(a) With Label Noise

(b) Without Label Noise

Figure: The minimum loss achieved by following SGD (blue) and SVRG (red) over a set of hyperparameters in a noisy least-square dynamics simulation for cases with and without label noise. The plot suggests that in the presence of label noise, there is a tradeoff between computational cost and convergence speed. In the absence of label noise, SGD strictly dominates SVRG in convergence speed for all computational cost.

Simulations on Noisy least squares Regression Model

Observations from Our Simulations:

- **The case with label noise**: The plot demonstrated an explicit trade-off between computational cost and convergence speed.
  - a crossing point of between SGD and SVRG appear, indicating SGD achieved a faster convergence speed in the first phase of the training, but converged to a higher loss, for all per-iteration compute cost.
  - The per-iteration computational cost does not seem to affect the time crossing point takes place. For all these three costs, the crossing points in the plot are at around the same time: 5.5 epochs.
- **The case of no label noise**: Both methods achieved linear convergence, while SGD achieved a much faster rate than SVRG, showing absolute dominance in this regime.

Contents

1. Introduction

2. Preliminary
   - SVRG Algorithm
   - The noisy least squares regression model

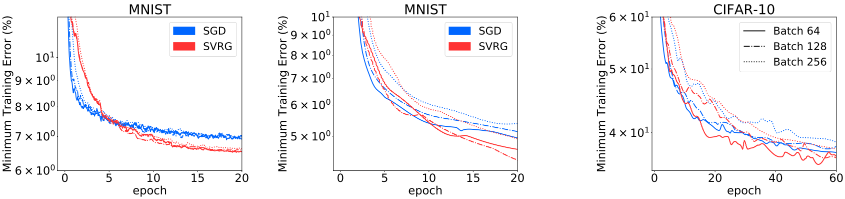
3. The Dynamic of SGD

4. A Dilemma For SVRG

5. Simulations on Noisy least squares Regression Model

6. Experiments on Benchmark Datasets

Underparameterized Setting



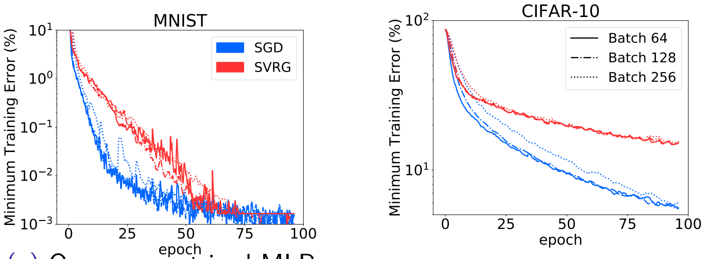
(a) Logistic Regression on MNIST.

(b) underparameterized MLP

(c) underparameterized CNN

Figure: The minimum loss achieved by following SGD (blue) and SVRG (red) over a set of hyperparameters for training on MNIST and CIFAR-10 with underparameterized models. All the results in these plot suggested there is a tradeoff between computational cost and convergence speed when comparing SGD and SVRG.

The overparameterized Setting



(a) Over-paremetrized MLP

(b) overparameterized CNN

Figure: The minimum loss achieved by following SGD (blue) and SVRG (red) over a set of hyperparameters for training on MNIST and CIFAR-10 with overparameterized models. In this setting we observed strict dominance of SGD over SVRG in convergence speed for all computational cost, matching our previous theoretical prediction.

Thanks!