When does stochastic gradient descent work without variance reduction?

Chuan-Zheng Lee (czlee), Hüseyin İnan (hinan1)

1 Introduction

Consider an optimization problem of the form

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad J(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} J_i(\mathbf{x}). \tag{1}$$

Problems of this form are common in machine learning, statistical estimation and other applications. If J is convex, it is well-known that (batch) gradient descent achieves linear convergence, *i.e.* achieves ε-level accuracy in $O(\log(1/\varepsilon))$ time. However, when m is large, evaluating $\nabla J(\mathbf{x})$ is expensive. For this reason, it is common instead to use *stochastic gradient descent* (SGD), which selects an index $i \in \{1,\dots,m\}$ at uniformly random and updates according to

$$\mathbf{x}^{(t+1)} := \mathbf{x}^{(t)} - \mu\nabla J_i(\mathbf{x}^{(t)}),$$

where t is an iteration index and μ is the step size (learning rate). Because SGD updates the current estimate $\mathbf{x}^{(t)}$ after computing each $\nabla J_i(\mathbf{x}^{(t)})$ rather than waiting for the whole $\nabla J(\mathbf{x}^{(t)})$, it often converges more quickly in practice, particularly in early iterations. However, if μ is held constant, SGD does not guarantee convergence: as $\mathbf{x}^{(t)}$ approaches the true minimum $\mathbf{x}^\star \triangleq \arg\min_{\mathbf{x}} J(\mathbf{x})$, the variance of $\nabla J_i(\mathbf{x}^{(t)})$ (over the random index i) can remain large, causing the estimate $\mathbf{x}^{(t)}$ to "jump around" the true minimum in successive iterations t, without getting appreciably closer. To address this problem, there is a large literature of techniques called *variance reduction*. For example, one technique is to reduce the step size $\mu = \mu_t$ as a function of t.

Breaking this mold, some recent works have presented algorithms similar to SGD that—surprisingly—do not require any variance reduction to ensure convergence. These works relate to a class of non-convex problems known as *phase retrieval*, in which an unknown vector $\mathbf{x} \in \mathbb{C}^n$ is to be recovered from m measurements $y_i = |\mathbf{a}_i^*\mathbf{x}|^2, i = 1,\dots,m$, (with $\mathbf{a}_i \in \mathbb{C}^n$ known), and a complex-domain cousin of gradient descent that they call *Wirtinger flow*, after the Wirtinger derivative.

Wirtinger flow was first proposed by Candès *et al.* in [CLS15]; Chen and Candès then presented an improved version called *truncated Wirtinger flow* (TWF) in [CC15] using a different loss function as its objective. An incremental version—analogous to SGD—came in [KÖ16], in which Kolte and Özgür showed that *incremental truncated Wirtinger flow* (ITWF) achieves the same asymptotic convergence as TWF, in practice runs much faster (like SGD), and—unlike SGD—requires no variance reduction in order to achieve this result.

In a similar vein, Zhang *et al.* proposed *reshaped Wirtinger flow* (RWF) and its incremental counterpart (IRWF) in [ZL16], again using a different objective loss function. They showed that RWF achieves the same asymptotic complexity as TWF, and—notably for our purposes—that the incremental version, IRWF, also matches RWF without any need for variance reduction.

Our project thus seeks to understand which attributes of optimization problems of the form (1) allow a *batch* algorithm (like gradient descent or TWF) to be converted to an *incremental* algorithm (like SGD or ITWF), without sacrificing any convergence properties and without any need for variance reduction. While the formidable task of identifying the relevant attributes in the most general terms is ongoing, in this paper we report on the successful application of the methods in [KÖ16] to two well-known classes of optimization problems: least squares and support vector machines.

2 Least squares

Consider the standard least squares problem. We are given a vector $\mathbf{y} \in \mathbb{R}^m$ and a skinny matrix $A \in \mathbb{R}^{m\times n}, m \geq n$, and we wish to find $\mathbf{x} \in \mathbb{R}^n$ such that, approximately, $\mathbf{y} \approx A\mathbf{x}$. More precisely, we wish to minimize the loss function

$$\ell(\mathbf{y}, A\mathbf{x}) = \frac{1}{2m}\|\mathbf{y} - A\mathbf{x}\|^2 = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \mathbf{a}_i^T\mathbf{x})^2,$$

where $A = [\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_m]^T$. In this analysis, we will assume that the rows of A are independent and distributed according to $\mathbf{a}_i \sim \mathcal{N}(0, I)$ for $i = 1,\dots,m$.

We will start with the case where $\mathbf{y} \in \mathcal{R}(A)$, *i.e.*, there exists some $\mathbf{x}^\star$ such that $\mathbf{y} = A\mathbf{x}^\star$. We will show that the standard SGD algorithm does not need any variance reduction method to converge to the optimal value, *i.e.*, that we can use a constant step size in the stochastic update. The corresponding SGD update for this problem is as follows:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \mu(\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} - y_{i_t})\mathbf{a}_{i_t}, \tag{2}$$

where i_t (for each $t = 1, 2, \dots$) is uniformly chosen at random from $\{1, 2, \dots, m\}$.

Before we present the first result for least squares, we state two concentration bounds which shortly prove helpful. Their proofs are in the appendix.

Lemma 1. *Let $\mathbf{a}_i \sim \mathcal{N}(0, I)$ i.i.d. for $i = 1,\dots,m$, and let j be chosen uniformly at random from $\{1,\dots,m\}$. Then for any $\delta > 0$, there exist universal constants $C, c_0, c_1 > 0$ such that if $m \geq c_0 n\delta^{-2}$, then*

$$(1-\delta)\|\mathbf{h}\|^2 \leq \mathbb{E}_j\left[(\mathbf{a}_j^T\mathbf{h})^2\right] \leq (1+\delta)\|\mathbf{h}\|^2 \tag{3}$$

with probability $1 - C\exp(-c_1 m\delta^2)$, simultaneously for all non-zero vectors $\mathbf{h} \in \mathbb{R}^n$, where the expectation is taken over the choice of j, conditional on $\{\mathbf{a}_i\}$.

Lemma 2. *Let $\mathbf{a}_i \sim \mathcal{N}(0, I)$ for $i = 1, \ldots, m$. Then*

$$\|\mathbf{a}_i\|^2 < 6n \tag{4}$$

with probability $1 - m\exp(-25n/8)$, simultaneously for all $i = 1, \ldots, m$.

We now present the convergence result for the case when $\mathbf{y} \in \mathcal{R}(A)$.

Theorem 1. *If the rows of A are independent and distributed according to $\mathbf{a}_i \sim \mathcal{N}(0, I)$, $i = 1, \ldots, m$, and $\mathbf{y} \in \mathcal{R}(A)$, then there exist universal constants $C, c_0, c_1, c_2 > 0$ and $0 < \rho < 1$, such that with probability at least $1 - Cm\exp(-c_1 n)$ and $\mu = c_2/n$, if $m \geq c_0 n$, the iterates of SGD algorithm (2), initialized at $\mathbf{x}^{(0)}$, satisfy*

$$\mathbb{E}_{\{i_0,\ldots,i_{t-1}\}}\left[\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2\right] \leq \left(1 - \frac{\rho}{n}\right)^t \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2, \tag{5}$$

where the expectation is taken over the choices of indices $\{i_0, \ldots, i_{t-1}\}$ (conditional on $\mathbf{a}_1, \ldots, \mathbf{a}_m$).

Proof. Defining $\mathbf{h} = \mathbf{x}^{(t)} - \mathbf{x}^\star$, we have

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t)} - \mu(\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} - y_{i_t})\mathbf{a}_{i_t} - \mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{h} - \mu(\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} - y_{i_t})\mathbf{a}_{i_t}\|^2\right] \\
&= \|\mathbf{h}\|^2 - 2\mu\mathbb{E}_{i_t}\left[(\mathbf{a}_{i_t}^T\mathbf{h})^2\right] + \mu^2\mathbb{E}_{i_t}\left[\|\mathbf{a}_{i_t}\|^2(\mathbf{a}_{i_t}^T\mathbf{h})^2\right].
\end{aligned}$$

We bound the second term above using Lemma 1 and the third term using Lemma 2, to get

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \\
&\leq \|\mathbf{h}\|^2 - 2\mu(1-\delta)\|\mathbf{h}\|^2 + \mu^2 6n(1+\delta)\|\mathbf{h}\|^2 \\
&= \left(1 - 2\mu(1-\delta) + \mu^2 6n(1+\delta)\right)\|\mathbf{h}\|^2.
\end{aligned}$$

Choosing $\mu = \frac{1-\delta}{6n(1+\delta)}$ then tells us that

$$\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \leq \left(1 - \frac{\rho}{n}\right)\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2,$$

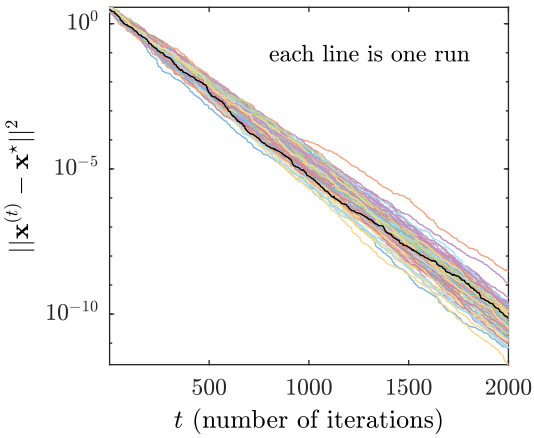
and taking the expectation over i_{t-1} we have

$$\mathbb{E}_{i_t,i_{t-1}}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \leq \left(1 - \frac{\rho}{n}\right)\mathbb{E}_{i_{t-1}}\left[\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2\right]. \tag{6}$$

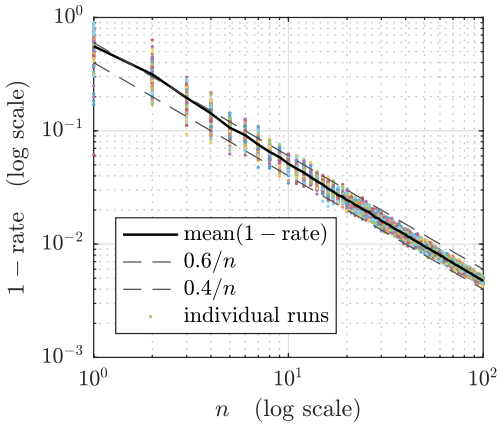
Applying (6) recursively on $i_{t-1}, \ldots, i_0$ then yields the result. □

Theorem 1 states that, when $y_1, \ldots, y_m$ (measurements in sensing applications, or labels in machine learning) are perfectly consistent with $\mathbf{a}_1, \ldots, \mathbf{a}_m$ (the transform in sensor applications, or training examples in machine learning) and $\mathbf{x}^\star$ (the vector to be recovered, or the parameters), the SGD iterates $\mathbf{x}^{(t)}$ do at least as well as linear convergence to the true $\mathbf{x}^\star$ with high probability.

Our simulations indicate that this bound is tight. In Figure 1(a), we ran SGD on randomly generated instances of the least squares problem with $y \in \mathcal{R}(A)$, and plot $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$ against t. We observe that convergence is indeed linear. In



(a) Convergence under Theorem 1 conditions (parameters $m = 300, n = 30, \mu = 5 \times 10^{-3}$)



(b) Relationship between convergence rate and n

Figure 1: Simulation results for Theorem 1

Figure 1(b), we compute the *rate* of convergence as the average gradient of each line in Figure 1(a), but this time we also repeat this for different values of n. We then plot $1 - \text{rate}$ against n, one point per run. Here, we observe that $1 - \text{rate}$ indeed appears to follow a ρ/n curve, as the bound would suggest, for some ρ between 0.4 and 0.6. These plots suggest that the bound in Theorem 1 is probably tight.

We note that Gaussian distribution is not the only example for Theorem 1. Specifically, Lemma 1 holds when the rows of matrix A are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^n$. Lemma 2 could also be obtained by either the Hoeffding-type or Bernstein-type inequality. For instance, if the entries of the rows of matrix A are independent and distributed according to some bounded random variables, we obtain Theorem 1 as well.

In the case where $y \notin \mathcal{R}(A)$, convergence is not linear, but is still bounded by the residual, as follows.

Theorem 2. *Say that the rows of A are independent and distributed according to $\mathbf{a}_i \sim \mathcal{N}(0, I)$, $i = 1, \ldots, m$, let $\mathbf{x}^\star = \arg\min_{\mathbf{x}} \ell(\mathbf{y}, A\mathbf{x})$ and define the residual as $\mathbf{r} = \mathbf{y} - A\mathbf{x}^\star$. Then there exist universal constants $C, c_0, c_1, c_2, c_3 > 0$ and $0 < \rho < 1$ such that with probability at least $1 - Cm\exp(-c_1 n)$ and $\mu = c_2/n$, if $m \geq c_0 n$, then the iterates of SGD algorithm*

(2), *initialized at* $\mathbf{x}^{(0)}$, *satisfy*

$$\mathbb{E}_{\{i_0,\dots,i_{t-1}\}}\left[\|\mathbf{x}^{(t)}-\mathbf{x}^\star\|^2\right] \lesssim \frac{\|\mathbf{r}\|^2}{m} + \left(1-\frac{\rho}{n}\right)^t \|\mathbf{x}^{(0)}-\mathbf{x}^\star\|^2,$$

where the $\lesssim$ *sign indicates a multiplicative constant.*

Proof. We start by using a similar expansion to and the same bounds as those used in Theorem 1. The residue introduces some additional terms in our expansion, but we can use the fact that $\mathbf{r} \in \text{null}(A)$ (so $\mathbf{r}^T A\mathbf{h} = 0\ \forall \mathbf{h} \in \mathbb{R}^n$) to simplify the expression. Note that $r_i = y_i - \mathbf{a}_i^T\mathbf{x}$. Then we have

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{h}-\mu(\mathbf{a}_{i_t}^T\mathbf{x}^{(t)}-y_{i_t})\mathbf{a}_{i_t}\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{h}-\mu(\mathbf{a}_{i_t}^T\mathbf{h}-r_{i_t})\mathbf{a}_{i_t}\|^2\right] \\
&= \|\mathbf{h}\|^2 - 2\mu\mathbb{E}_{i_t}\left[(\mathbf{a}_{i_t}^T\mathbf{h})^2\right] + \mu^2\mathbb{E}_{i_t}\left[\|\mathbf{a}_{i_t}\|^2(\mathbf{a}_{i_t}^T\mathbf{h})^2\right] \\
&\quad + \mu^2\mathbb{E}_{i_t}\left[r_{i_t}^2\|\mathbf{a}_{i_t}\|^2\right] \\
&\le \left(1-2\mu(1-\delta)+\mu^2 6n(1+\delta)\right)\|\mathbf{h}\|^2 + \frac{\mu^2 6n}{m}\|\mathbf{r}\|^2, \qquad (7)
\end{aligned}$$

where in the last step we used the same steps as in the proof of Theorem 1 for the first three terms, and Lemma 2 for the last term.

The rest of the proof then follows similar logic to the two regimes described in Section 6.2 of [KÖ16] and Section 6 of [CC15]. Define *Regime I* to be the case where $\|\mathbf{h}\| \ge \frac{c_3}{\sqrt{m}}\|\mathbf{r}\|$ for some large constant c_3, and define *Regime II* to be the case where $\|\mathbf{h}\| < \frac{c_3}{\sqrt{m}}\|\mathbf{r}\|$. In Regime I, from (7) we get

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^\star\|^2\right] \\
&\le \left[1-2\mu(1-\delta)+\mu^2 6n\left(1+\delta+\frac{1}{c_3^2}\right)\right]\|\mathbf{h}\|^2.
\end{aligned}$$

We may then select

$$\mu = \frac{1-\delta}{6n(1+\delta+1/c_3^2)} \qquad (8)$$

to achieve linear convergence of the form

$$\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^\star\|^2\right] \le \left(1-\frac{\rho}{n}\right)\|\mathbf{h}\|^2, \qquad (9)$$

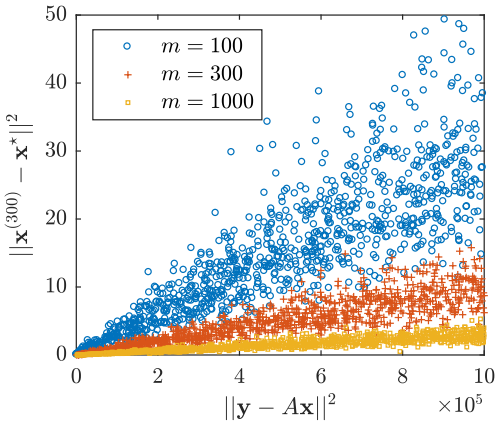
similarly to how we did in Theorem 1. This linear convergence remains until $\|\mathbf{h}\|$ exits Regime I.

In Regime II, from (7) we get

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^\star\|^2\right] \\
&\le \left[\left(1-2\mu(1-\delta)+\mu^2 6n(1+\delta)\right)\frac{c_3^2}{m} + \frac{\mu^2 6n}{m}\right]\|\mathbf{r}\|^2 \\
&< \left[\left[1-\frac{(1-\delta)^2}{6n(1+\delta+1/c_3^2)}\right]\frac{c_3^2}{m} + \frac{1-\delta}{(1+\delta+1/c_3^2)m}\right]\|\mathbf{r}\|^2 \\
&< \left(c_3^2+\frac{1-\delta}{1+\delta}\right)\frac{1}{m}\|\mathbf{r}\|^2. \qquad (10)
\end{aligned}$$

After an iteration in Regime II, may either stay in Regime II or bounce back up to Regime I. In the latter case, $\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2$ may increase in expectation, but the next iterate will still satisfy (10), and having just moved to Regime I it will then begin reducing through linear convergence again. Hence $\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^\star\|^2\right] \lesssim \frac{1}{m}\|\mathbf{r}\|^2$.

The theorem then follows by combining (9) and (10). □



(a) Convergence error under Theorem 2 conditions

Figure 2: Simulation results for Theorem 2

To verify Theorem 2 numerically, we ran simulations with different values of n and different residues, and took note of the error $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$ once it appeared to have stopped linear convergence, which in this case was always by $t = 300$. Figure 2 shows a plot with one point for each such simulation. The bound on the error does indeed appear to be linear in $\|\mathbf{y} - A\mathbf{x}^\star\|^2$, and it also decreases as predicted with increasing m, if holding the residue constant. One might note that in practical scenarios, the residue would not be independent of m, since if the elements of $\mathbf{y} \in \mathbb{R}^m$ are identically distributed, one would expect $\|\mathbf{y} - A\mathbf{x}^\star\|$ to scale with m; Figure 2 is useful mainly for showing the correctness of Theorem 2.

3 Support vector machines

We now consider the support vector machine without regularization:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ J(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m \max\{1-y_i\mathbf{a}_i^T\mathbf{x}, 0\}.$$

We assume that the data is linearly separable, *i.e.*, there exists $\mathbf{x}^\star \in \mathbb{R}^n$ such that $y_i = \text{sign}(\mathbf{a}_i^T\mathbf{x}^\star)$ for $i = 1,\dots,m$. We can further assume that $y_i\mathbf{a}_i^T\mathbf{x}^\star \ge 1+\delta,\ i = 1,\dots,m$ (*i.e.* δ is the minimum excess margin). Note that the latter can be obtained by the first assumption since $y_i\mathbf{a}_i^T\mathbf{x}^\star = |\mathbf{a}_i^T\mathbf{x}^\star|$ so we can scale $\mathbf{x}^\star$ until we satisfy the margin requirement when the data is linearly separable.

We now present the corresponding convergence result in the following theorem:

Theorem 3. *If the rows of* A *are independent and distributed according to* $\mathbf{a}_i \sim \mathcal{N}(0, I),\ i = 1,\dots,m$, *then there exist universal constants* $C, c_0, c_1, c_2 > 0$ *such that with probability at least* $1 - Cm\exp(-c_1 n)$ *and* $\mu = c_2/n$, *if* $m \ge c_0 n$, *the iterates of the SGD algorithm, initialized at* $\mathbf{x}^{(0)}$, *satisfy*

$$\mathbb{E}_{\{i_0,\dots,i_{t-1}\}}\left[\|\mathbf{x}^{(t)}-\mathbf{x}^\star\|^2\right] \le \|\mathbf{x}^{(0)}-\mathbf{x}^\star\|^2 - \frac{t\delta^2}{6mn}, \qquad (11)$$

for as long as $J(\mathbf{x}^{(t)}) > 0$.

Proof. Defining $\mathbf{h} = \mathbf{x}^{(t)} - \mathbf{x}^\star$, we have

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t)} + \mu y_{i_t}\mathbf{a}_{i_t}\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\} - \mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}_{i_t}\left[\|\mathbf{h} + \mu y_{i_t}\mathbf{a}_{i_t}\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}\|^2\right] \\
&= \|\mathbf{h}\|^2 + 2\mu\mathbb{E}_{i_t}\left[y_{i_t}\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}(\mathbf{a}_{i_t}^T\mathbf{h})\right] \\
&\quad + \mu^2\mathbb{E}_{i_t}\left[\|\mathbf{a}_{i_t}\|^2\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}\right].
\end{aligned}$$

We note that we can bound the second term as

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[y_{i_t}\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}(\mathbf{a}_{i_t}^T\mathbf{h})\right] \\
&= \mathbb{E}_{i_t}\left[\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}(y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)})\right] \\
&\quad - \mathbb{E}_{i_t}\left[\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}(y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^\star)\right] \\
&\leq \mathbb{E}_{i_t}\left[\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}\right] - \mathbb{E}_{i_t}\left[\mathbf{1}\{y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^{(t)} < 1\}(y_{i_t}\mathbf{a}_{i_t}^T\mathbf{x}^\star)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\}\left[1 - |\mathbf{a}_i^T\mathbf{x}^\star|\right].
\end{aligned}$$

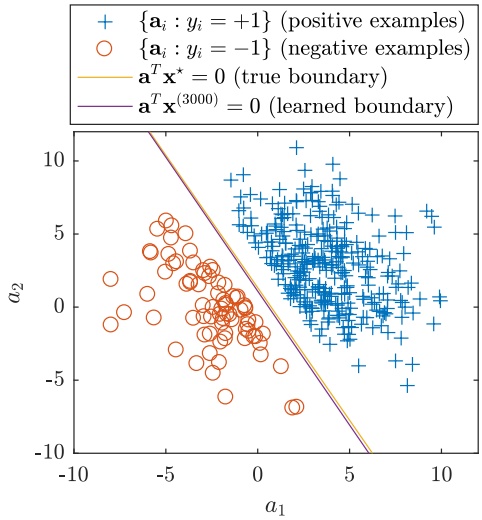
We similarly use Lemma 2 and with the bound on the second term, then

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2\right] \\
&\leq \|\mathbf{h}\|^2 + 2\mu\frac{1}{m}\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\}\left[1 - |\mathbf{a}_i^T\mathbf{x}^\star|\right] \\
&\quad + \mu^2\frac{6n}{m}\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\} \\
&\leq \|\mathbf{h}\|^2 - 2\mu\frac{1}{m}\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\}\delta \\
&\quad + \mu^2\frac{6n}{m}\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\} \\
&= \|\mathbf{h}\|^2 + (-2\mu\frac{1}{m} + \mu^2\frac{6n}{m})\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\}
\end{aligned}$$

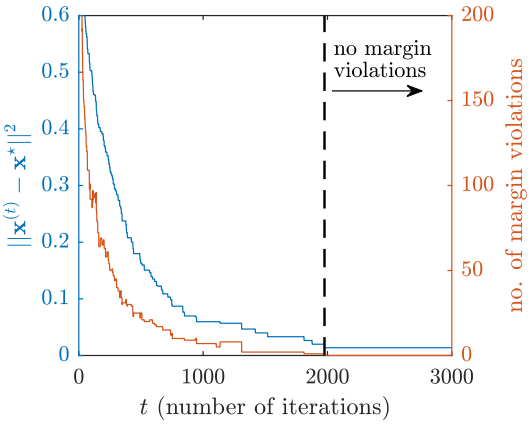
where the second inequality is from the margin assumption. We note that as long as $J(\mathbf{x}^{(t)}) > 0$, we have $\sum_{i=1}^m \mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x}^{(t)} < 1\} \geq 1$. Choosing $\mu = \delta/6n$ and iterating over t concludes the proof. □

The convergence result in this section is slightly different to the least squares result: our result applies only until the estimate $\mathbf{x}^{(t)}$ yields no margin violations. We provide an example in Figure 3 which depicts this phenomenon: it can be seen that the error $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$ decreases until there are no margin violations, after which it never improves, because the gradient $\nabla J_i(\mathbf{x}) = -y_i\mathbf{a}_i\mathbf{1}\{y_i\mathbf{a}_i^T\mathbf{x} < 1\}$ is zero for every i.

We note that the bound provided by Theorem 3 is a weak one, because we bound the improvement by the case where only one margin violation is detected, which is the case only near convergence.



(a) Example of SVM result



(b) Example of SVM convergence

Figure 3: Simulation results for Theorem 3

4 Discussion and future work

In this project, we have applied the methods introduced in [KÖ16] to least squares and support vector machines, showing that where a perfect fit is possible, SGD will converge to the correct solution with high probability. For least squares without perfect fit, we showed a bound on the error in terms of the norm of the residue. Furthermore, for least squares, our numerical experiments provide evidence that our bounds are tight.

These results are valuable steps forward for our broader aim: to understand which attributes of optimization problems allow stochastic gradient descent to converge without variance reduction. Based on our current work, as well as [KÖ16] and [ZL16], we further conjecture that following characteristics may be salient:

- A "perfect fit" being represented by the data to be fitted, as in Theorems 1 and 3, noting that Theorem 2 included a term in the error that does not reduce to zero as t increases.

- The form of the loss function, leading to an update rule

amenable to being bounded by a function of $\left\|\mathbf{x}^{(t)} - \mathbf{x}^\star\right\|^2$, where $\mathbf{x}^\star$ is the true optimal point.

- The assumption that the sensing vectors are i.i.d., leading to the use of concentration bounds to show convergence with high probability.

In particular, our results have confirmed our earlier rejection of the possibility that the truncation step in [KÖ16] (which, informally speaking, rejects measurements/examples that are likely to be outliers) might have been a controlling factor, because neither of the SGD algorithms considered in this paper involve any similar truncation.

We plan to take the following next steps for this work:

- We would like to make our bound for the SVM tighter, and also consider the case where there is the usual regularization term in the objective function with the data being possibly not linearly separable.

- We wish to continue working on other well-known loss functions to see if we can get similar results.

- We would like to understand for precisely which classes of distributions similar convergence results hold and for which ones they do not.

Finally, our ultimate goal is to find the necessary and sufficient conditions for SGD to converge with fixed step size and without variance reduction. We believe this pursuit will further understanding of SGD, elucidate in what sorts of applications it is most likely to be useful without variance reduction, and possibly even lead to a different way of understanding the convergence of SGD and SGD-like algorithms in the literature.

References

[Ver10] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *CoRR* abs/1011.3027 (2010).

[CLS15] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. "Phase Retrieval via Wirtinger Flow: Theory and Algorithms". In: *IEEE Transactions on Information Theory* 61.4 (Apr. 2015), pp. 1985–2007.

[CC15] Yuxin Chen and Emmanuel Candes. "Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems". In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 739–747.

[KÖ16] Ritesh Kolte and Ayfer Özgür. "Phase Retrieval via Incremental Truncated Wirtinger Flow". In: *CoRR* abs/1606.03196 (2016).

[ZL16] Huishuai Zhang and Yingbin Liang. "Reshaped Wirtinger Flow for Solving Quadratic System of Equations". In: *Advances in Neural Information Processing Systems 29*. 2016, pp. 2622–2630.

5 Appendix

5.1 Proof of Lemma 1

We note that we are showing a much stronger result in this lemma: Instead of the case where $\mathbf{h} = \mathbf{x}^{(t)} - \mathbf{x}^\star$, we consider all non-zero vectors $\mathbf{h} \in \mathbb{R}^n$ simultaneously.

The following result follows from Theorem 5.39 in [Ver10]:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T - I \right\| \leq \delta,$$

with probability $1 - C\exp(-c_1 m\delta^2)$ if $m \geq c_0 n\delta^{-2}$ for some universal constants C, c_0, c_1. For any $\mathbf{h} \in \mathbb{R}^n$ and $\mathbf{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, the following holds

$$\begin{aligned} &\left| \mathbf{h}^T \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T - I \right) \mathbf{h} \right| \\ &\leq \|\mathbf{h}\| \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T - I \right\| \|\mathbf{h}\| = \|\mathbf{h}\|^2 \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T - I \right\|. \end{aligned}$$

Therefore,

$$\left| \mathbf{h}^T \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T - I \right) \mathbf{h} \right| \leq \delta \|\mathbf{h}\|^2,$$

and through some algebraic manipulation, we arrive at

$$(1-\delta)\|\mathbf{h}\|^2 \leq \frac{1}{m} \sum_{i=1}^{m} (\mathbf{a}_i^T \mathbf{h})^2 \leq (1+\delta)\|\mathbf{h}\|^2.$$

Since $\mathbb{E}_{i_t}\left[(\mathbf{a}_i^T \mathbf{h})^2\right] = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{a}_i^T \mathbf{h})^2$, this completes the lemma.

5.2 Proof of Lemma 2

To prove this inequality, we use Corollary 3.17 of the STATS 311 lecture notes:

Corollary 1. *Let $X_1, \ldots, X_n$ be independent mean-zero (σ_i^2, b_i)-sub-exponential random variables. Define $b_* := \max_i b_i$. Then for all $t \geq 0$ and all vectors $a \in \mathbb{R}^n$, we have*

$$P\left(\sum_{i=1}^{n} a_i X_i \geq t \right) \leq \exp\left(-\frac{1}{2} \min\left\{ \frac{t^2}{\sum_{i=1}^{n} a_i^2 \sigma_i^2}, \frac{t}{b_* ||a||_\infty} \right\} \right).$$

We note that $a_{ij} \sim \mathcal{N}(0,1)$ and chi-square distribution is sub-exponential with parameters $(4, 4)$. Therefore, we have

$$P(\|\mathbf{a}_i\|^2 \geq 6n) = P\left(\sum_{j=1}^{n} (\mathbf{a}_{ij}^2 - 1) \geq 5n \right) \leq \exp\left(-\frac{25n}{8} \right).$$

Using the union bound, we have simultaneously for all $i = 1, \ldots, m$,

$$\|\mathbf{a}_i\|^2 < 6n$$

with probability $1 - m\exp(-25n/8)$.