Supplement for "Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence"

Yi Xu [1] Qihang Lin [2] Tianbao Yang [1]

1. Proof of Theorem 1

Theorem 1. *Suppose Assumption 1 holds and $F(\mathbf{w})$ obeys the LGC (6). Given $\delta \in (0,1)$, let $\tilde{\delta} = \delta/K$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$ and t be the smallest integer such that $t \geq \max\{9, 1728\log(1/\tilde{\delta})\}\frac{G^2 D_1^2}{\epsilon_0^2}$. Then ASSG-c guarantees that, with a probability $1-\delta$, $F(\mathbf{w}_K) - F_* \leq 2\epsilon$. As a result, the iteration complexity of ASSG-c for achieving an 2ϵ-optimal solution with a high probability $1-\delta$ is $O(c^2 G^2 \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$.*

Proof. Let $\mathbf{w}^\dagger_{k,\epsilon}$ denote the closest point to $\mathbf{w}_k$ in $\mathcal{S}_\epsilon$. Define $\epsilon_k = \frac{\epsilon_0}{2^k}$. Note that $D_k = \frac{D_1}{2^{k-1}} \geq \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}$ and $\eta_k = \frac{\epsilon_{k-1}}{3G^2}$. We will show by induction that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ for $k = 0, 1, \ldots$ with a high probability, which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon_{k-1} + \epsilon$, we will show that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ with a high probability. By Lemma 1, we have

$$\begin{aligned} \|\mathbf{w}^\dagger_{k-1,\epsilon} - \mathbf{w}_{k-1}\|_2 &\leq \frac{c}{\epsilon^{1-\theta}}(F(\mathbf{w}_{k-1}) - F(\mathbf{w}^\dagger_{k-1,\epsilon})) \\ &\leq \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}} \leq D_k. \end{aligned} \tag{1}$$

We apply Lemma 2 to the k-th stage of Algorithm 1 conditioned on randomness in previous stages. With a probability $1 - \tilde{\delta}$ we have

$$\begin{aligned} F(\mathbf{w}_k) - F(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \frac{\eta_k G^2}{2} &+ \frac{\|\mathbf{w}_{k-1} - \mathbf{w}^\dagger_{k-1,\epsilon}\|_2^2}{2\eta_k t} \\ &+ \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}. \end{aligned} \tag{2}$$

We now consider two cases for $\mathbf{w}_{k-1}$. First, we assume $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon$, i.e. $\mathbf{w}_{k-1} \in \mathcal{S}_\epsilon$. Then we have $\mathbf{w}^\dagger_{k-1,\epsilon} = \mathbf{w}_{k-1}$ and

$$\begin{aligned} F(\mathbf{w}_k) - F(\mathbf{w}^\dagger_{k-1,\epsilon}) &\leq \frac{\eta_k G^2}{2} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}} \\ &\leq \frac{\epsilon_k}{3} + \frac{\epsilon_{k-1}}{6} = \frac{2\epsilon_k}{3}. \end{aligned}$$

The second inequality using the fact that $\eta_k = \frac{2\epsilon_k}{3G^2}$ and $t \geq 1728\log(1/\tilde{\delta})\frac{G^2 D_1^2}{\epsilon_0^2}$. As a result,

$$F(\mathbf{w}_k) - F_* \leq F(\mathbf{w}^\dagger_{k-1,\epsilon}) - F_* + \frac{2\epsilon_k}{3} \leq \epsilon + \epsilon_k.$$

Next, we consider $F(\mathbf{w}_{k-1}) - F_* > \epsilon$, i.e. $\mathbf{w}_{k-1} \notin \mathcal{S}_\epsilon$. Then we have $F(\mathbf{w}^\dagger_{k-1,\epsilon}) - F_* = \epsilon$. Combining (1) and (2), we get

$$F(\mathbf{w}_k) - F(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \frac{\eta_k G^2}{2} + \frac{D_k^2}{2\eta_k t} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}.$$

Since $\eta_k = \frac{2\epsilon_k}{3G^2}$ and $t \geq \max\{9, 1728\log(1/\tilde{\delta})\}\frac{G^2 D_1^2}{\epsilon_0^2}$, we have each term in the R.H.S of above inequality bounded by $\epsilon_k/3$. As a result,

$$F(\mathbf{w}_k) - F(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \epsilon_k \Rightarrow F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon.$$

with a probability $1 - \tilde{\delta}$. Therefore by induction, with a probability at least $(1-\tilde{\delta})^K$ we have

$$F(\mathbf{w}_K) - F_* \leq \epsilon_K + \epsilon \leq 2\epsilon.$$

Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1 - \delta$ and we complete the proof. □

2. Proof of Lemma 3

Lemma 3. *For any $t \geq 1$, we have $\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta G$ and $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta G$.*

Proof. By the optimality of $\widehat{\mathbf{w}}_*$, we have for any $\mathbf{w} \in \mathcal{K}$

$$\left(\partial F(\widehat{\mathbf{w}}_*) + \frac{1}{\beta}(\widehat{\mathbf{w}}_* - \mathbf{w}_1)\right)^\top (\mathbf{w} - \widehat{\mathbf{w}}_*) \geq 0.$$

[1] Department of Computer Science, The University of Iowa, Iowa City, IA 52246, USA [2] Department of Management Sciences, The University of Iowa, Iowa City, IA 52246, USA. Correspondence to: Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

Let $\mathbf{w} = \mathbf{w}_1$, we have

$$\partial F(\widehat{\mathbf{w}}_*)^\top(\mathbf{w}_1 - \widehat{\mathbf{w}}_*) \geq \frac{\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2}{\beta}.$$

Because $\|\partial F(\widehat{\mathbf{w}}_*)\|_2 \leq G$ due to $\|\partial f(\mathbf{w};\xi)\|_2 \leq G$, then

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2 \leq \beta G.$$

Next, we bound $\|\mathbf{w}_t - \mathbf{w}_1\|_2$. According to the update of $\mathbf{w}_{t+1}$ we have

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 \\ \leq&\|\mathbf{w}'_{t+1} - \mathbf{w}_1\|_2 \\ =&\| - \eta_t\partial f(\mathbf{w}_t;\xi_t) + (1 - \eta_t/\beta)(\mathbf{w}_t - \mathbf{w}_1)\|_2.\end{aligned}$$

We prove $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta G$ by induction. First, we consider $t = 1$, where $\eta_t = 2\beta$, then

$$\|\mathbf{w}_2 - \mathbf{w}_1\|_2 \leq \|2\beta\partial f(\mathbf{w}_t;\xi_t)\|_2 \leq 2\beta G.$$

Then we consider any $t \geq 2$, where $\eta_t/\beta \leq 1$. Then

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 \\ \leq&\left\|-\frac{\eta_t}{\beta}\beta\partial f(\mathbf{w}_t;\xi_t) + \left(1 - \frac{\eta_t}{\beta}\right)(\mathbf{w}_t - \mathbf{w}_1)\right\|_2 \\ \leq&\frac{\eta_t}{\beta}\beta G + \left(1 - \frac{\eta_t}{\beta}\right)2\beta G \leq 2\beta G.\end{aligned}$$

Therefore

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta G.$$

□

3. Proof of Theorem 2

Theorem 2. *Suppose Assumption 1 holds and $F(\mathbf{w})$ obeys the LGC (6). Given $\delta \in (0, 1/e)$, let $\tilde{\delta} = \delta/K$, $K = \lceil\log_2(\frac{\epsilon_0}{\epsilon})\rceil$, $\beta_1 \geq \frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}}$ and t be the smallest integer such that $t \geq \max\{3, \frac{136\beta_1 G^2(1+\log(4\log t/\tilde{\delta})+\log t)}{\epsilon_0}\}$. Then ASSG-r guarantees that, with a probability $1 - \delta$, $F(\mathbf{w}_K) - F_* \leq 2\epsilon$. As a result, the iteration complexity of ASSG-r for achieving an 2ϵ-optimal solution with a high probability $1 - \delta$ is $O(c^2G^2\log(\epsilon_0/\epsilon)\log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $\beta_1 = O(\frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}})$.*

To prove the theorem, we first show the following results of high probability convergence bound.

Lemma 4. *Given $\mathbf{w}_1 \in \mathcal{K}$, apply T-iterations of (9). For any fixed $\mathbf{w} \in \mathcal{K}$, $\delta \in (0, 1)$, and $T \geq 3$, with a probability at least $1 - \delta$, the following inequality holds*

$$\begin{aligned}F(\widehat{\mathbf{w}}_T) - F(\mathbf{w}) \leq&\frac{\|\mathbf{w} - \mathbf{w}_1\|_2^2}{2\beta} \\ &+ \frac{34\beta G^2\left(1 + \log T + \log(4\log T/\delta)\right)}{T},\end{aligned}$$

where $\widehat{\mathbf{w}}_t = \sum_{\tau=1}^t \mathbf{w}_t/t$.

Proof. Let $\mathbf{g}_t = \partial f(\mathbf{w}_t;\xi_t) + (\mathbf{w}_t - \mathbf{w}_1)/\beta$ and $\partial\widehat{F}(\mathbf{w}_t) = \partial F(\mathbf{w}_t) + (\mathbf{w}_t - \mathbf{w}_1)/\beta$. Note that $\|\mathbf{g}_t\|_2 \leq 3G$. According to the standard analysis for the stochastic gradient method we have

$$\begin{aligned}\mathbf{g}_t^\top(\mathbf{w}_t - \widehat{\mathbf{w}}_*) \leq&\frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 \\ &+ \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2.\end{aligned}$$

Then

$$\begin{aligned}\partial\widehat{F}(\mathbf{w}_t)^\top(\mathbf{w}_t - \widehat{\mathbf{w}}_*) \leq&\frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 \\ &+\frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 + (\partial\widehat{F}(\mathbf{w}_t) - \mathbf{g}_t)^\top(\mathbf{w}_t - \widehat{\mathbf{w}}_*).\end{aligned}$$

By strong convexity of $\widehat{F}$ we have

$$\widehat{F}(\widehat{\mathbf{w}}_*) - \widehat{F}(\mathbf{w}_t) \geq \partial\widehat{F}(\mathbf{w}_t)^\top(\widehat{\mathbf{w}}_* - \mathbf{w}_t) + \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2.$$

Then

$$\begin{aligned}&\widehat{F}(\mathbf{w}_t) - \widehat{F}(\widehat{\mathbf{w}}_*) \\ \leq&\frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 \\ &+ (\partial\widehat{F}(\mathbf{w}_t) - \mathbf{g}_t)^\top(\mathbf{w}_t - \widehat{\mathbf{w}}_*) - \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 \\ \leq&\frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 \\ &+ \underbrace{(\partial F(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t))^\top(\mathbf{w}_t - \widehat{\mathbf{w}}_*)}_{\zeta_t} - \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2.\end{aligned}$$

By summing the above inequalities across $t = 1, \ldots, T$, we have

$$\begin{aligned}&\sum_{t=1}^T(\widehat{F}(\mathbf{w}_t) - \widehat{F}(\widehat{\mathbf{w}}_*)) \qquad (3)\\ \leq&\sum_{t=1}^{T-1}\frac{1}{2}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \frac{1}{2\beta}\right)\|\widehat{\mathbf{w}}_* - \mathbf{w}_{t+1}\|_2^2 \\ &+ \sum_{t=1}^T\zeta_t - \frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 \\ &- \frac{1}{4\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + \frac{1}{2\eta_1}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + \frac{9G^2}{2}\sum_{t=1}^T\eta_t \\ \leq&\sum_{t=1}^T\zeta_t - \frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 + 9\beta G^2(1 + \log T). \qquad (4)\end{aligned}$$

where the last inequality uses $\eta_t = \frac{2\beta}{t}$. Next, we bound R.H.S of the above inequality. We need the following lemma.

Lemma 5. *(Lemma 3 (Kakade & Tewari, 2008)) Suppose $X_1,\ldots,X_T$ is a martingale difference sequence with $|X_t|\le b$. Let*

$$\mathit{Var}_t X_t = \mathit{Var}(X_t|X_1,\ldots,X_{t-1}).$$

where Var denotes the variance. Let $V=\sum_{t=1}^T \mathit{Var}_t X_t$ be the sum of conditional variance of X_t's. Further, let $\sigma=\sqrt{V}$. Then we have for any $\delta<1/e$ and $T\ge 3$,

$$\Pr\left(\sum_{t=1}^T X_t > \max\{2\sigma, 3b\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta)}\right)$$
$$\le 4\delta\log T.$$

To proceed the proof of Lemma 4. We let $X_t=\zeta_t$ and $D_T=\sum_{t=1}^T\|\mathbf{w}_t-\widehat{\mathbf{w}}_*\|_2^2$. Then $X_1,\ldots,X_T$ is a martingale difference sequence. Let $D=3\beta G$. Note that $|\zeta_t|\le 2GD$. By Lemma 5, for any $\delta<1/e$ and $T\ge 3$, with a probability $1-\delta$ we have

$$\sum_{t=1}^T\zeta_t\le$$
$$\max\left\{2\sqrt{\log(\frac{4\log T}{\delta})\sum_{t=1}^T\mathrm{Var}_t\zeta_t},\, 6GD\log(\frac{4\log T}{\delta})\right\}. \tag{5}$$

Note that

$$\sum_{t=1}^T\mathrm{Var}_t\zeta_t\le\sum_{t=1}^T\mathrm{E}_t[\zeta_t^2]\le 4G^2\sum_{t=1}^T\|\mathbf{w}_t-\widehat{\mathbf{w}}_*\|_2^2=4G^2D_T.$$

As a result, with a probability $1-\delta$,

$$\begin{aligned}\sum_{t=1}^T\zeta_t\le&4G\sqrt{\log(4\log T/\delta)}\sqrt{D_T}+6GD\log(4\log T/\delta)\\ \le&16\beta G^2\log(4\log T/\delta)+\frac{1}{4\beta}D_T\\ &+6GD\log(4\log T/\delta).\end{aligned}$$

As a result, with a probability $1-\delta$,

$$\begin{aligned}&\sum_{t=1}^T\zeta_t-\frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_*-\mathbf{w}_t\|_2^2\\ \le&16\beta G^2\log(4\log T/\delta)+6GD\log(4\log T/\delta)\\ =&34\beta G^2\log(4\log T/\delta).\end{aligned}$$

Thus, with a probability $1-\delta$

$$\begin{aligned}&\widehat{F}(\widehat{\mathbf{w}}_T)-\widehat{F}(\widehat{\mathbf{w}}_*)\\ \le&\frac{34\beta G^2\log(4\log T/\delta)}{T}+\frac{9\beta G^2(1+\log T)}{T}\\ \le&\frac{34\beta G^2(1+\log T+\log(4\log T/\delta))}{T}.\end{aligned}$$

Using the facts that $F(\widehat{\mathbf{w}}_T)\le\widehat{F}(\widehat{\mathbf{w}}_T)$ and $\widehat{F}(\widehat{\mathbf{w}}_*)\le\widehat{F}(\mathbf{w})=F(\mathbf{w})+\frac{\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\beta}$, we have

$$\begin{aligned}&F(\widehat{\mathbf{w}}_T)-F(\mathbf{w})-\frac{\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\beta}\\ \le&\frac{34\beta G^2(1+\log T+\log(4\log T/\delta))}{T}.\end{aligned}$$

□

Next, let us start to prove Theorem 2.

Proof of Theorem 2. Let $\mathbf{w}_{k,\epsilon}^\dagger$ denote the closest point to $\mathbf{w}_k$ in the ϵ sublevel set. Define $\epsilon_k\triangleq\frac{\epsilon_0}{2^k}$. First, we note that $\beta_k\ge\frac{2c^2\epsilon_{k-1}}{\epsilon^{2(1-\theta)}}$. We will show by induction that $F(\mathbf{w}_k)-F_*\le\epsilon_k+\epsilon$ for $k=0,1,\ldots$ with a high probability, which leads to our conclusion when $k=K$. The inequality holds obviously for $k=0$. Conditioned on $F(\mathbf{w}_{k-1})-F_*\le\epsilon_{k-1}+\epsilon$, we will show that $F(\mathbf{w}_k)-F_*\le\epsilon_k+\epsilon$ with a high probability. We apply Lemma 4 to the k-th stage of Algorithm 2 conditioned on the randomness in previous stages. With a probability at least $1-\tilde{\delta}$ we have

$$\begin{aligned}&F(\mathbf{w}_k)-F(\mathbf{w}_{k-1,\epsilon}^\dagger)\\ \le&\frac{1}{2\beta_k}\|\mathbf{w}_{k-1,\epsilon}^\dagger-\mathbf{w}_{k-1}\|_2^2\\ &+\frac{34\beta_kG^2(1+\log t+\log(4\log t/\tilde{\delta}))}{t}.\end{aligned}\tag{6}$$

We now consider two cases for $\mathbf{w}_{k-1}$. First, we assume $F(\mathbf{w}_{k-1})-F_*\le\epsilon$, i.e. $\mathbf{w}_{k-1}\in\mathcal{S}_\epsilon$. Then we have $\mathbf{w}_{k-1,\epsilon}^\dagger=\mathbf{w}_{k-1}$ and

$$\begin{aligned}F(\mathbf{w}_k)-F(\mathbf{w}_{k-1,\epsilon}^\dagger)\le&\frac{34\beta_kG^2(1+\log t+\log(4\log t/\tilde{\delta}))}{t}\\ \le&\frac{\epsilon_k}{2}.\end{aligned}$$

The last inequality uses the fact that $t\ge\frac{136\beta_1G^2(1+\log(4\log t/\tilde{\delta})+\log t)}{\epsilon_0}$. As a result,

$$F(\mathbf{w}_k)-F_*\le F(\mathbf{w}_{k-1,\epsilon}^\dagger)-F_*+\frac{\epsilon_k}{2}\le\epsilon+\epsilon_k.$$

Next, we consider $F(\mathbf{w}_{k-1})-F_*>\epsilon$, i.e. $\mathbf{w}_{k-1}\notin\mathcal{S}_\epsilon$. Then we have $F(\mathbf{w}_{k-1,\epsilon}^\dagger)-F_*=\epsilon$. Similar to the proof of Theorem 1, by Lemma 1, we have

$$\|\mathbf{w}_{k-1,\epsilon}^\dagger-\mathbf{w}_{k-1}\|_2\le\frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}.\tag{7}$$

Combining (6) and (7), we have

$$\begin{aligned}&F(\mathbf{w}_k)-F(\mathbf{w}_{k-1,\epsilon}^\dagger)\\ \le&\frac{1}{2\beta_k}\left(\frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}\right)^2+\frac{34\beta_kG^2(1+\log t+\log(4\log t/\tilde{\delta}))}{t}.\end{aligned}$$

Using the fact that $\beta_k \geq \frac{2c^2\epsilon_{k-1}}{\epsilon^{2(1-\theta)}}$ and $t \geq \frac{68\beta_k G^2(1+\log t+\log(4\log t/\tilde{\delta}))}{\epsilon_k} = \frac{136\beta_1 G^2(1+\log t+\log(4\log t/\tilde{\delta}))}{\epsilon_0}$, we get

$$F(\mathbf{w}_k) - F(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \frac{\epsilon_{k-1}}{4} + \frac{\epsilon_k}{2} = \epsilon_k,$$

which together with the fact that $F(\mathbf{w}^\dagger_{k-1,\epsilon}) = F_* + \epsilon$ implies

$$F(\mathbf{w}_k) - F_* \leq \epsilon + \epsilon_k.$$

Therefore by induction, we have with a probability at least $(1-\tilde{\delta})^K$,

$$F(\mathbf{w}_K) - F_* \leq \epsilon_K + \epsilon = \frac{\epsilon_0}{2^K} + \epsilon \leq 2\epsilon,$$

where the last inequality is due to the value of $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil$. Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1-\delta$. □

4. Proof of Theorem 3

Theorem 3 (RASSG with unknown c)**.** *Let $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil$ in Algorithm 3. Suppose $D_1^{(1)}$ is sufficiently large so that there exists $\hat{\epsilon}_1 \in [\epsilon, \epsilon_0/2]$, with which $F(\cdot)$ satisfies a LGC (6) on $\mathcal{S}_{\hat{\epsilon}_1}$ with $\theta \in (0,1)$ and the constant c, and $D_1^{(1)} = \frac{c\epsilon_0}{\hat{\epsilon}_1^{1-\theta}}$. Let $\hat{\delta} = \frac{\delta}{K(K+1)}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(1)}/\epsilon_0\right)^2$. Then with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$ calls of ASSG-c, Algorithm 3 finds a solution $\mathbf{w}^{(S)}$ such that $F(\mathbf{w}^{(S)}) - F_* \leq 2\epsilon$. The total number of iterations of RASSG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O(\lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$.*

Proof. Since $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{\epsilon_0}{\hat{\epsilon}_1})\rceil$, $D_1^{(1)} = \frac{c\epsilon_0}{\hat{\epsilon}_1^{1-\theta}}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(\frac{GD_1^{(1)}}{\epsilon_0}\right)^2$, following the proof of Theorem 1, we can show that with a probability $1 - \frac{\delta}{K+1}$,

$$F(\mathbf{w}^{(1)}) - F_* \leq 2\hat{\epsilon}_1. \tag{8}$$

By running ASSG-c starting from $\mathbf{w}^{(1)}$ which satisfies (8) with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{2\hat{\epsilon}_1}{\hat{\epsilon}_1/2})\rceil$, $D_1^{(2)} = \frac{c\epsilon_0}{(\hat{\epsilon}_1/2)^{1-\theta}} \geq \frac{c2\hat{\epsilon}_1}{(\hat{\epsilon}_1/2)^{1-\theta}}$, and $t_2 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(2)}/\epsilon_0\right)^2$, Theorem 1 ensures that

$$F(\mathbf{w}^{(2)}) - F_* \leq \hat{\epsilon}_1$$

with a probability at least $(1-\delta/(K+1))^2$. By continuing the process, with $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$ we can prove that with a probability at least $(1-\delta/(K+1))^S \geq 1-\delta\frac{S}{K+1} \geq 1-\delta$,

$$F(\mathbf{w}^{(S)}) - F_* \leq 2\hat{\epsilon}_1/2^{S-1} \leq 2\epsilon.$$

The total number of iterations for the S calls of ASSG-c is bounded by

$$\begin{aligned}
T_S =& K\sum_{s=1}^S T_s = K\sum_{s=1}^S t_1 2^{2(s-1)(1-\theta)} \\
=& Kt_1 2^{2(S-1)(1-\theta)} \sum_{s=1}^S \left(1/2^{2(1-\theta)}\right)^{S-s} \\
\leq& Kt_1 2^{2(S-1)(1-\theta)} \frac{1}{1-1/2^{2(1-\theta)}} \\
\leq& O\left(Kt_1 \left(\frac{\hat{\epsilon}_1}{\epsilon}\right)^{2(1-\theta)}\right) \leq \widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)}).
\end{aligned}$$

□

5. Proof of Theorem 4

Theorem 4 (RASSG with unknown θ)**.** *Let $\theta = 0$, $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil$ in Algorithm 3. Assume $D_1^{(1)}$ is sufficiently large so that there exists $\hat{\epsilon}_1 \in [\epsilon, \epsilon_0/2]$ rendering that $D_1^{(1)} = \frac{B_{\hat{\epsilon}_1}\epsilon_0}{\hat{\epsilon}_1}$. Let $\hat{\delta} = \frac{\delta}{K(K+1)}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(1)}/\epsilon_0\right)^2$. Then with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$ calls of ASSG-c, Algorithm 3 finds a solution $\mathbf{w}^{(S)}$ such that $F(\mathbf{w}^{(S)}) - F_* \leq 2\epsilon$. The total number of iterations of RASSG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O\left(\lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \log(1/\delta)\frac{G^2B_{\hat{\epsilon}_1}^2}{\epsilon^2}\right)$.*

Proof. The proof is similar to the proof of Theorem 3, and we reprove it for completeness. It is easy to show that $t_1 \geq \frac{136\beta_1^{(1)}G^2(1+\log(4\log t_1/\hat{\delta})+\log t_1)}{\epsilon_0}$. Following the proof of Theorem 2, we then can show that with a probability $1-\frac{\delta}{S}$,

$$F(\mathbf{w}^{(1)}) - F_* \leq 2\hat{\epsilon}_1 \tag{9}$$

with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{\epsilon_0}{\hat{\epsilon}_1})\rceil$ and $\beta_1^{(1)} = \frac{2c^2\epsilon_0}{\hat{\epsilon}_1^{2(1-\theta)}}$. By running ASSG-r starting from $\mathbf{w}^{(1)}$ which satisfies (9) with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{2\hat{\epsilon}_1}{\hat{\epsilon}_1/2})\rceil$, $t_2 = t_1 2^{2(1-\theta)} \geq \frac{136\beta_1^{(2)}G^2(1+\log(4\log t_2/\hat{\delta})+\log t_2)}{\epsilon_0}$ and $\beta_1^{(2)} = \frac{2c^2\epsilon_0}{(\hat{\epsilon}_1/2)^{2(1-\theta)}} \geq \frac{2c^2\hat{\epsilon}_1/2}{(\hat{\epsilon}_1/2)^{2(1-\theta)}}$, Theorem 2 ensures that

$$F(\mathbf{w}^{(2)}) - F_* \leq \hat{\epsilon}_1$$

with a probality at least $(1-\delta/S)^2$. By continuing the process, with $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$, we can prove that with a probality at least $(1-\delta/S)^S \geq 1-\delta$

$$F(\mathbf{w}^{(S)}) - F_* \leq 2\hat{\epsilon}_1/2^{S-1} \leq 2\epsilon$$

The total number of iterations for the S calls of ASSG-c is bounded by

$$\begin{aligned}
T_S =& K\sum_{s=1}^{S} T_s = K\sum_{s=1}^{S} t_1 2^{2(s-1)(1-\theta)} \\
=& Kt_1 2^{2(S-1)(1-\theta)} \sum_{s=1}^{S}\left(1/2^{2(1-\theta)}\right)^{S-s} \\
\leq& Kt_1 2^{2(S-1)(1-\theta)} \frac{1}{1-1/2^{2(1-\theta)}} \\
\leq& O\left(Kt_1\left(\frac{\hat{\epsilon}_1}{\epsilon}\right)^{2(1-\theta)}\right) \leq \widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)})
\end{aligned}$$

□

6. Monotonicity of B_ϵ/ϵ

Lemma 6. $\frac{B_\epsilon}{\epsilon}$ *is monotonically decreasing in* ϵ.

Proof. Consider $\epsilon' > \epsilon > 0$. Let $\mathbf{x}_{\epsilon'}$ be any point on $\mathcal{L}_{\epsilon'}$ such that $dist(\mathbf{x}_{\epsilon'}, \Omega_*) = B_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ be the closest point to $\mathbf{x}_{\epsilon'}$ in Ω_* so that $\|\mathbf{x}^*_{\epsilon'} - \mathbf{x}_{\epsilon'}\| = B_{\epsilon'}$. We define a new point between $\mathbf{x}_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ as

$$\bar{\mathbf{x}} = \frac{B_\epsilon}{B_{\epsilon'}}\mathbf{x}_{\epsilon'} + \frac{B_{\epsilon'} - B_\epsilon}{B_{\epsilon'}}\mathbf{x}^*_{\epsilon'}.$$

Since $0 < B_\epsilon < B_{\epsilon'}$, $\bar{\mathbf{x}}$ is strictly between $\mathbf{x}_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ and $dist(\bar{\mathbf{x}}, \Omega_*) = \|\mathbf{x}^*_{\epsilon'} - \bar{\mathbf{x}}\| = \frac{B_\epsilon}{B_{\epsilon'}}\|\mathbf{x}^*_{\epsilon'} - \mathbf{x}_{\epsilon'}\| = B_\epsilon$. By the convexity of F, we have

$$\frac{F(\bar{\mathbf{x}}) - F_*}{dist(\bar{\mathbf{x}}, \Omega_*)} \leq \frac{F(\mathbf{x}_{\epsilon'}) - F_*}{dist(\mathbf{x}_{\epsilon'}, \Omega_*)} = \frac{\epsilon'}{B_{\epsilon'}}.$$

Note that we must have $F(\bar{\mathbf{x}}) - F_* \geq \epsilon$ since, otherwise, we can move $\bar{\mathbf{x}}$ towards $\mathbf{x}_{\epsilon'}$ until $F(\bar{\mathbf{x}}) - F_* = \epsilon$ but $dist(\bar{\mathbf{x}}, \Omega_*) > B_\epsilon$, contradicting with the definition of B_ϵ. Then, the proof is completed by applying $F(\bar{\mathbf{x}}) - F_* \geq \epsilon$ and $dist(\bar{\mathbf{x}}, \Omega_*) = B_\epsilon$ to the previous inequality. □

7. Additional Experiments

The datasets used in experiments are from libsvm[1] website. We summarize the basic statistics of datasets in Table 1.

Table 1. Statistics of real datasets

Name	#Training (n)	#Features (d)	Type
covtype.binary	581,012	54	Classification
real-sim	72,309	20,958	Classification
url	2,396,130	3,231,961	Classification
million songs	463,715	90	Regression
E2006-tfidf	16,087	150,360	Regression
E2006-log1p	16,087	4,272,227	Regression

To examine the convergence behavior of ASSG with different values of the iterations in per-stage, we also provide the results of ASSG with very large t and add them into Figure 1. For completeness, we replot all these results in Figure 2. The results show that ASSG with smaller t converges much faster to an ϵ-level set than ASSG with larger t, while ASSG with larger t can converge to a much smaller objective. In some case, ASSG with larger t is not as good as SSG in earlier stages but overall it converges faster to a smaller objective than SSG. We also present the results for $\lambda = 10^{-2}$ in Figure 3, which are similar to that for $\lambda = 10^{-4}$ in Figure 2.

In Figures 2 and 3, we compare RASSG with SVRG++ in terms of running time (cpu time) since SVRG++ computes a full gradient in each outer loop while RASSG only goes one sample in each iteration. Following many previous studies, we also include the results in terms of the number of full gradient pass in Figure 4 both for $\lambda = 10^{-2}$ and $\lambda = 10^{-4}$. Similar trend results can be found in Figure 4, indecating that RASSG converges faster than other three algorithms.

References

Kakade, Sham M. and Tewari, Ambuj. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pp. 801–808, 2008.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

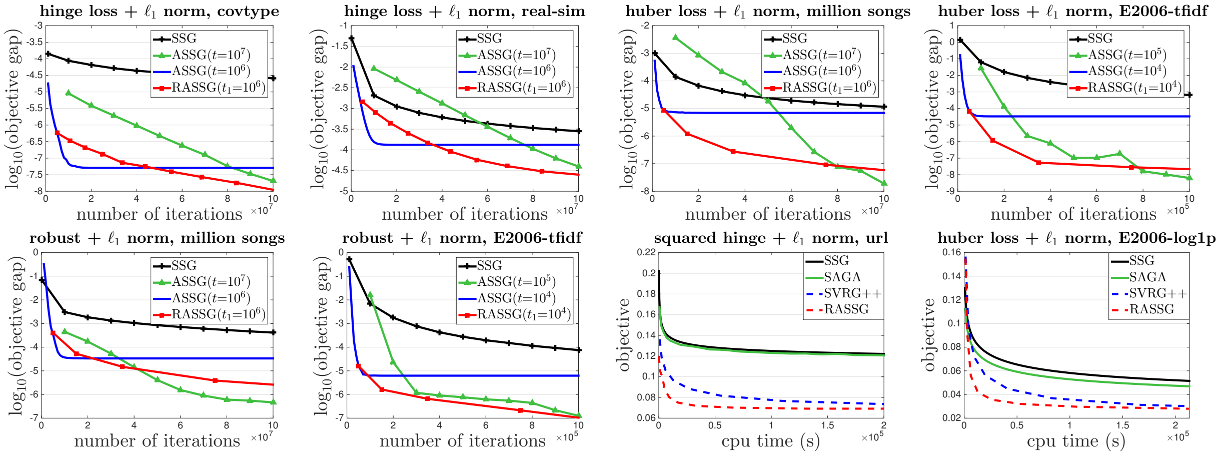


Figure 2. Comparison of different algorithms for solving different problems on different datasets ($\lambda = 10^{-4}$).

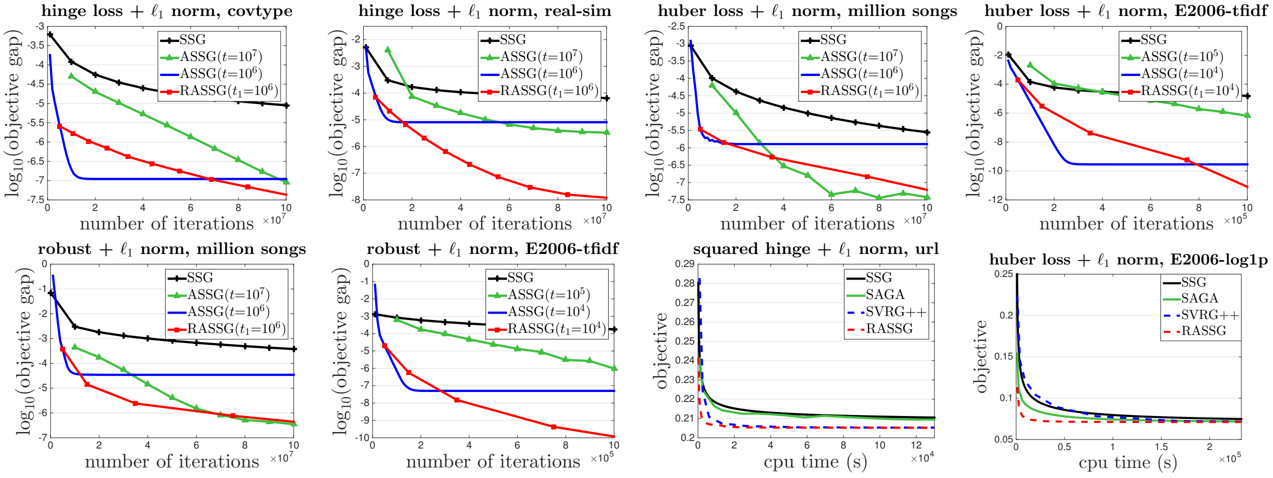
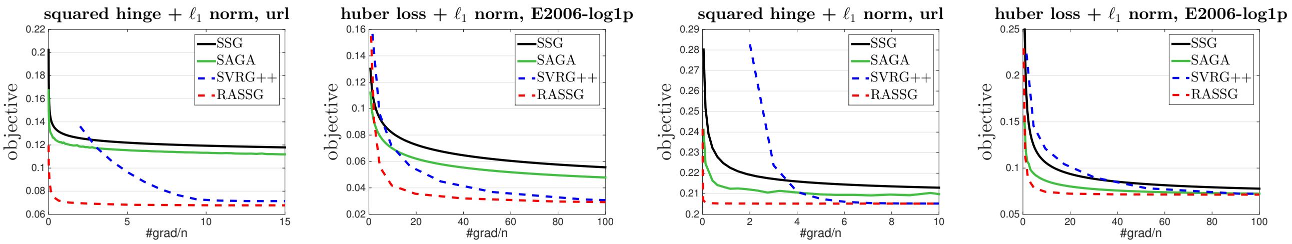


Figure 3. Comparison of different algorithms for solving different problems on different datasets ($\lambda = 10^{-2}$).



(a) $\lambda = 10^{-4}$

(b) $\lambda = 10^{-2}$

Figure 4. Comparison of different algorithms for solving different problems on different datasets by number of gradient pass.