Overview

1. Introduction
2. Current Results
3. Our Results
4. Main Techniques
Consider observations $\xi_1, \ldots, \xi_n$. Convex loss function $f(\cdot, \xi_i) : \mathbb{R}^d \to \mathbb{R}$.

**Empirical Risk Minimization:**

$$x^* = \arg\min_{x \in D} \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) := \arg\min_{x \in D} \nabla \hat{F}(x, \xi_i).$$

- **SGD with replacement (SGD):** fix step size sequence $\alpha_t \geq 0$. Start at $x_0 \in D$. For every time step generate independent random variable $I_t \sim \text{unif}([n])$.

  $$x_{t+1} = x_t - \alpha_t \nabla f(x_t, \xi_{I_t})$$

- Easy to analyze since independence of $I_t$ ensures that $\mathbb{E}_{I_t} \nabla f(x_t, \xi_{I_t}) = \hat{F}(x_t)$.

- Sharp non-asymptotic guarantees available but seldom used in practice.
In practice, the order of data is fixed (say $\xi_1, \ldots, \xi_n$) and the data is selected in this order, one after the other. One such pass is called an epoch. The algorithm is run for $K$ epochs. A randomized version of this ‘gets rid’ of the bad orderings.

- **SGD without Replacement (SGDo)** At the beginning of the $k$th epoch, draw an independent uniformly random permutation $\sigma_k$.

  $$x_{k,i} = x_{k,i-1} - \alpha_{k,i} \nabla f(x_{k,i}; \xi_{\sigma_k(i)})$$

- This is closer to the algorithm implemented in practice.
- Harder to analyze since $\mathbb{E}\nabla f(x_{k,i}; \xi_{\sigma_k(i)}) \neq \mathbb{E}\nabla \hat{F}(x_{k,i})$
Experiments\(^1\) found that on many problems SGDo converges as \(O(1/K^2)\), which is faster than SGD which converges at \(O(1/K)\). (\(K\) = number of epochs)

Theoretically, it wasn’t even shown that SGDo ‘matches’ the rate of SGD for all \(K\).

### Currently Known Bounds

<table>
<thead>
<tr>
<th>Paper</th>
<th>Guarantee</th>
<th>Assumptions</th>
<th>Step Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GÜRBÜZBALABAN ET AL. 2015</td>
<td>$O\left(\frac{C(n,d)}{K^2}\right)$</td>
<td>Lipschitz, Strong convexity, Smoothness, Hessian Lipschitz. $K &gt; \kappa^{1.5\sqrt{n}}$</td>
<td>$\frac{1}{K}$</td>
</tr>
<tr>
<td>HaoChen and Sra 2018</td>
<td>$\tilde{O}\left(\frac{1}{n^2K^2} + \frac{1}{K^3}\right)$</td>
<td>Smoothness, $K &gt; \kappa^2$</td>
<td>$\frac{\log nK}{\mu nK}$</td>
</tr>
<tr>
<td><strong>This Paper</strong></td>
<td>$\tilde{O}\left(\frac{1}{nK^2}\right)$</td>
<td>Lipschitz, Strong convexity, Smoothness, $K &gt; \kappa^2$</td>
<td>$\frac{\log nK}{\mu nK}$</td>
</tr>
<tr>
<td>Shamir 2016</td>
<td>$O\left(\frac{1}{nK}\right)$</td>
<td>Lipschitz, Strong convexity, Smoothness, Generalized Linear Function, $K = 1$</td>
<td>$\frac{1}{\mu nK}$</td>
</tr>
<tr>
<td><strong>This Paper</strong></td>
<td>$O\left(\frac{1}{nK}\right)$</td>
<td>Lipschitz, Strong convexity, Smoothness</td>
<td>$\min\left(\frac{2}{L}, \frac{\log nK}{\mu nK}\right)$</td>
</tr>
<tr>
<td>Shamir 2016</td>
<td>$O\left(\frac{1}{\sqrt{nK}}\right)$</td>
<td>Lipschitz, Generalized Linear Function, $K = 1$</td>
<td>$\frac{1}{\sqrt{nK}}$</td>
</tr>
<tr>
<td><strong>This Paper</strong></td>
<td>$O\left(\frac{1}{\sqrt{nK}}\right)$</td>
<td>Lipschitz, Smoothness</td>
<td>$\min\left(\frac{2}{L}, \frac{1}{\sqrt{nK}}\right)$</td>
</tr>
</tbody>
</table>

*Table 1. Comparison of our results with previously known results in terms of number of functions $n$ and number of epochs $K$. For simplicity, we suppress the dependence on other problem dependent parameters such as Lipschitz constant, strong convexity, smoothness etc.*
Small number of Epochs

- Assumptions: $f(\cdot; \xi_i)$ is $L$ smooth, $\|\nabla f(\cdot; \xi_i)\| \leq G$, diam($\mathcal{W}$) $\leq D$.
- Suboptimality $O\left(\frac{GD}{\sqrt{nK}}\right)$ (leading order, General case)
- Suboptimality $O\left(\frac{G^2 \log nK}{\mu nK}\right)$ (leading order, $\mu$ strongly convex)
- Shamir’s result$^2$ only works for generalized linear functions and when $K = 1$.
- All other “acceleration” results hold only when $K$ is very large.

---

Large number of Epochs

- Assumptions: $f(\cdot; \xi_i)$ is $L$ smooth, $\|\nabla f(\cdot; \xi_i)\| \leq G$ and $\hat{F}$ is $\mu$ strongly convex.
- When $K \gtrsim \kappa^2$, Suboptimality: $O\left(\frac{\kappa^2 G^2 (\log nK)^2}{\mu nK^2}\right)$
- Previous results\(^3\) require Hessian smoothness and $K \geq \kappa^{1.5} \sqrt{n}$ to give suboptimality of $O\left(\frac{\kappa^4}{n^2 K^2} + \frac{\kappa^4}{K^3}\right)$.
- Without smoothness assumption, there can be no acceleration.

---

Main Techniques

- Main bottleneck in analysis: \( \mathbb{E} \nabla f(x_{k,i}; \xi_{\sigma_k(i)}) \neq \mathbb{E} \nabla \hat{F}(x_{k,i}) \).

- If \( \sigma'_k \) is independent of \( \sigma_k \),

\[
\mathbb{E} \nabla f(x_{k,i}; \xi_{\sigma'_k(i)}) = \mathbb{E} \nabla \hat{F}(x_{k,i}).
\]

- Therefore,

\[
\mathbb{E} \nabla f(x_{k,i}; \xi_{\sigma_k(i)}) = \mathbb{E} \nabla \hat{F}(x_{k,i}) + O(d_W(x_{k,i} | \sigma_k(i) = r, x_{k,i}))
\]

- Through coupling arguments: \( d_W(x_{k,i} | \sigma_k(i) = r, x_{k,i}) \lesssim \alpha_{k,0} G \)
For the smooth and strongly convex case,
\[ \nabla \hat{F}(x^*) = 0 = \frac{1}{n} \sum_{i=1}^{n} f(x^*, \xi_{\sigma_k(i)}). \] (Note that this doesn’t hold with independent sampling).

Therefore, when \( x_{k,0} \approx x^* \) we show by coupling arguments that:

\[ 0 \approx \nabla \hat{F}(x_{k,0}) \approx \frac{1}{n} \sum_{i=1}^{n} f(x_{i,k}, \xi_{\sigma_k(i)}). \]

This is similar to the variance reduction as seen in modifications of SGD like SAGA, SVRG etc.
References


Questions?