Stochastic Gradient Descent with Importance Sampling

Rachel Ward
UT Austin

Joint work with
Deanna Needell (Claremont McKenna College)
and Nathan Srebro (TTIC / Technion)
Recall: Gradient descent

Problem: Minimize $F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$, $n$ very large

Gradient descent:

Initialize $x_0$

$$x^{(j+1)} = x^{(j)} - \gamma \nabla F(x^{(j)})$$

$$= x^{(j)} - \gamma \sum_{i=1}^{n} \nabla f_i(x^{(j)})$$

Not practical in huge dimension $n$!
**Stochastic gradient descent**

Minimize $F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

Initialize $x_0$

For $j = 1, 2, \ldots$

\[
dx^{(j+1)} = x^{(j)} - \gamma \nabla f_i(x^{(j)})
\]

Goal: nonasymptotic bounds on $\mathbb{E}\|x^{(j)} - x^*\|^2$
Minimize $F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) = \mathbb{E} f_i(x)$

We begin by assuming

1. $\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_i \|x - y\|^2$

2. $F$ is strongly convex: $\langle x - y, \nabla F(x) - \nabla F(y) \rangle \geq \mu \|x - y\|^2$

Convergence rate of SGD should depend on

1. A condition number,
   \[ \kappa_{av} = \frac{1}{n} \sum_{\mu} L_i, \quad \kappa_{\text{max}} = \max_{\mu} L_i, \quad \kappa_2 = \sqrt{\frac{1}{n} \sum_{\mu} L_i^2} \]

2. Consistency: $\sigma^2 = \mathbb{E} \|\nabla f_i(x^*)\|^2$
SGD - convergence rates

(Needell, Srebro, W’, 2013): Under smoothness and convexity assumptions, the SGD iterates satisfy

$$\mathbb{E} \| x^{(k)} - x^* \|^2 \leq [1 - 2\gamma \mu (1 - \gamma \sup_i L_i)]^k \| x_0 - x^* \|^2 + \frac{\gamma \sigma^2}{\mu (1 - \gamma \sup_i L_i)},$$

Corollary:

$$\mathbb{E} \| x^{(k)} - x^* \|_2^2 \leq \varepsilon \text{ after }$$

$$k = 2 \log(\varepsilon / \varepsilon_0) \left( \frac{\sup_i L_i}{\mu} + \frac{\sigma^2}{\mu^2 \varepsilon} \right)$$

SGD iterations, with optimized constant step-size.
SGD - convergence rates

We showed:
\[ k = 2 \log(\frac{\varepsilon}{\varepsilon_0}) \left( \frac{\sup_i L_i}{\mu} + \frac{\sigma^2}{\mu^2 \varepsilon} \right) \]
SGD iterations suffice for \( \mathbb{E}\|x^{(k)} - x_*\|^2_2 \leq \varepsilon \)

Compare to:

(Bach and Moulines, 2011) :
\[ k = 2 \log(\frac{\varepsilon}{\varepsilon_0}) \left( \left( \frac{\sqrt{\frac{1}{n} \sum_i L_i^2}}{\mu} \right)^2 + \frac{\sigma^2}{\mu^2 \varepsilon} \right) \]
SGD iterations suffice for \( \mathbb{E}\|x^{(k)} - x_*\|^2_2 \leq \varepsilon \)
\[ k \propto \frac{\sup_i L_i}{\mu} \text{ vs. } k \propto \left( \frac{1}{n} \sum_i \frac{L_i^2}{\mu} \right)^2 \text{ steps} \]

Difference in proof is that we use the co-coercivity lemma for smooth functions with Lipschitz gradient

\[ \|x^{(k+1)} - x_*\|^2 = \|x^{(k)} - x_* - \gamma \nabla f_i(x_k)\|^2 \]

\[ \leq \|x^{(k)} - x_*\|^2 - 2\gamma < x^{(k)} - x_*, \nabla f_i(x^{(k)}) > \]

\[ + 2\gamma^2 \|\nabla f_i(x^{(k)}) - \nabla f_i(x_*)\|^2 + 2\gamma^2 \|\nabla f_i(x_*)\|^2 \]

\[ \leq \|x^{(k)} - x_*\|^2 - 2\gamma < x^{(k)} - x_*, \nabla f_i(x^{(k)}) > \]

\[ + 2\gamma^2 L_i < x^{(k)} - x_*, \nabla f_i(x_k) - \nabla f_i(x_*) > + 2\gamma^2 \|\nabla f_i(x_*)\|^2 \]
Consider the least squares case:

\[ F(x) = \frac{1}{2} \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2 \]

\[ = \frac{1}{2} \| Ax - b \|^2 \]

Assume consistency: \( Ax_* = b, \sigma^2 = 0 \)

\[ \frac{\sup_i L_i}{\mu} = (n \sup_i \| a_i \|^2)(\| A^\dagger \|^2) \]
These convergence rates are tight

Consider the system

\[
\begin{pmatrix}
  1 & 0 \\
  0 & 1/\sqrt{n} \\
  0 & 1/\sqrt{n} \\
  \vdots & \vdots \\
  0 & 1/\sqrt{n}
\end{pmatrix}
\begin{pmatrix}
  1 \\
  0 \\
  \vdots \\
  0
\end{pmatrix}
= \begin{pmatrix}
  1 \\
  0 \\
  \vdots \\
  0
\end{pmatrix}
\]

Here, \( \frac{\sup_i L_i}{\mu} = n \sup_i \|a_i\|^2 \|A^\dagger\|^2 = n \)

In this example, we need \( k = n \) steps to get any accuracy
Better convergence rates using weighted sampling strategies?
SGD with weighted sampling

observe:

\[ F(x) = \frac{1}{n} \sum_i f_i(x) = \mathbb{E}(w) \frac{1}{w_i} f_i(x) \]

given weights \( w_i \) such that \( \sum_i w_i = n \).

SGD unbiased update with weighted sampling: Let \( \tilde{f}_i = \frac{1}{w_i} f_i \)

\[
    x^{(j+1)} = x^{(j)} - \gamma \nabla \tilde{f}_{i_k}(x^{(j)}) \\
    = x^{(j)} - \gamma \frac{1}{w_{i_k}} \nabla f_{i_k}(x^{(j)})
\]

where \( \mathbb{P}(i_k = i) = \frac{w_i}{\sum_j w_j} \)
SGD with weighted sampling

\[ F(x) = \frac{1}{n} \sum_i f_i(x) = \mathbb{E}^{(w)} \frac{1}{w_i} f_i(x) \]

given weights \( w_i \) such that \( \sum_i w_i = n \).

Weighted sampling strategies in stochastic optimization not new:

- (Strohmer, Vershynin 2009): randomized Kaczmarz algorithm
- (Nesterov 2010, Lee, Sidford 2013): biased sampling for accelerated stochastic coordinate descent
SGD with weighted sampling

Our previous result: for $F(x) = \mathbb{E} f_i(x)$ and $x^{k+1} = x^k - \gamma \nabla f_i(x^k)$,

$$
\mathbb{E}\|x^k - x_*\|^2_2 \leq \varepsilon
$$

after $k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{L_{\max}[f_i]}{\mu} + \frac{\sigma^2[f_i]}{\mu^2\varepsilon} \right)$ steps

Corollary for weighted sampling:

$$
\mathbb{E}(w)\|x^k - x_*\|^2_2 \leq \varepsilon \text{ after }
$$

$$
k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{L_{\max}\left[\frac{1}{w_i} f_i\right]}{\mu} + \frac{\sigma^2\left[\frac{1}{w_i} f_i\right]}{\mu^2\varepsilon} \right) \text{ steps}
$$
Choice of weights

For $F(x) = \mathbb{E}^*(w) f_i(x)$,

$$\mathbb{E}(w) \|x^k - x_*\|_2^2 \leq \varepsilon \text{ after }$$

$$k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{L_{\text{max}} \left[ \frac{1}{w_i} f_i \right]}{\mu} + \frac{\sigma^2 \left[ \frac{1}{w_i} f_i \right]}{\mu^2 \varepsilon} \right) \text{ steps}$$

If $\sigma^2 = 0$, choose weights to minimize $L_{\text{max}} \left[ \frac{1}{w_i} f_i \right]$:

$$\mathbb{E}(w) \|x^k - x_*\|_2^2 \leq \varepsilon \text{ after }$$

$$k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{1}{n} \sum_i L_i \right) \text{ steps, using weights } w_i = \frac{n L_i}{\sum_i L_i}$$
Improved convergence rate with weighted sampling

Recall the example: \[
\begin{pmatrix}
1 & 0 \\
0 & 1/\sqrt{n} \\
0 & 1/\sqrt{n} \\
\vdots & \vdots \\
0 & 1/\sqrt{n}
\end{pmatrix}
\begin{pmatrix}
1 \\
0 \\
\vdots \\
0
\end{pmatrix}
= \begin{pmatrix}
1 \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

Since this system is consistent, \( \sigma^2 = 0 \) and

\[
\frac{L_{\text{max}}}{\mu}(A) = n \quad \Rightarrow \quad O(n) \text{ steps using uniform sampling}
\]

\[
\frac{\overline{L}_i}{\mu}(A) = 2 \quad \Rightarrow \quad O(1) \text{ steps using biased sampling}
\]
Improved convergence rate with weighted sampling

Recall the example:

\[
\begin{pmatrix}
1 & 0 \\
0 & \frac{1}{\sqrt{n}} \\
0 & \frac{1}{\sqrt{n}} \\
\vdots & \vdots \\
0 & \frac{1}{\sqrt{n}}
\end{pmatrix}
\begin{pmatrix}
1 \\
0 \\
0 \\
\vdots \\
0
\end{pmatrix} =
\begin{pmatrix}
1 \\
0 \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

But what if the system is not consistent?
\[ \mathbb{E}(w) \| x^k - x_* \|_2^2 \leq \varepsilon \text{ after } \]
\[ k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{L_{\text{max}} \left[ \frac{1}{w_i} f_i \right]}{\mu} + \frac{\sigma^2 \left[ \frac{1}{w_i} f_i \right]}{\mu^2 \varepsilon} \right) \text{ steps} \]

Choosing weights \( w_i = \frac{n L_i}{\sum_i L_i} \) gives \( L_{\text{max}} \left[ \frac{1}{w_i} f_i \right] = \frac{1}{n} \sum_i L_i \)

Choosing weights \( w_i = 1 \) gives \( \sigma^2 \left[ \frac{1}{w_i} f_i \right] = \sigma^2 [f_i] \)

**Partially-biased sampling:**
Choosing weights \( w_i = \frac{1}{2} + \frac{1}{2} \frac{n L_i}{\sum_i L_i} \) gives
\[ L_{\text{max}} \left[ \frac{1}{w_i} f_i \right] \leq 2 \frac{1}{n} \sum_i L_i \text{ and } \sigma^2 \left[ \frac{1}{w_i} f_i \right] \leq 2 \sigma^2 [f_i] \]

Partially biased sampling gives strictly better convergence rate, up to a factor of 2
**SGD - convergence rates**

**Uniform sampling:**

(Bach and Moulines, 2011):

\[
k = 2 \log(\varepsilon/\varepsilon_0) \left( \left( \frac{\sqrt{\frac{1}{n} \sum_i L_i^2}}{\mu} \right)^2 + \frac{\sigma^2}{\mu^2 \varepsilon} \right)
\]

SGD iterations suffice for \( \mathbb{E}\|x^{(k)} - \mathbf{x}_*\|_2^2 \leq \varepsilon \)

**Partially biased sampling:**

(Needell, Srebro, W, 2013):

\[
k = 2 \log(\varepsilon/\varepsilon_0) \left( \frac{\sup_i L_i}{\mu} + \frac{\sigma^2}{\mu^2 \varepsilon} \right)
\]

SGD iterations suffice for \( \mathbb{E}\|x^{(k)} - \mathbf{x}_*\|_2^2 \leq \varepsilon \)

\[
k = 4 \log(\varepsilon/\varepsilon_0) \left( \frac{1}{n} \sum_i L_i \right) + \frac{\sigma^2}{\mu^2 \varepsilon}
\]

SGD iterations suffice for \( \mathbb{E}\|x^{(k)} - \mathbf{x}_*\|_2^2 \leq \varepsilon \)
We have been operating in the setting

1. Each $\nabla f_i$ has Lipschitz constant $L_i$:
   \[ \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_i \|x - y\|^2 \]

2. $F$ has strong convexity parameter $\mu$:
   \[ \langle x - y, \nabla F(x) - \nabla F(y) \rangle \geq \mu \|x - y\|^2 \]

Other settings and weaker assumptions.
- Removing strong convexity assumption
- Non-smooth
Smoothness, but no strong convexity:

- Each $\nabla f_i$ has Lipschitz constant $L_i$:
  \[ \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_i \|x - y\|^2 \]

- (Srebro et. al 2010): Number of iterations of SGD:
  \[ k = O \left( \frac{(\sup_i L_i)\|x_*\|^2}{\varepsilon} \frac{F(x_*) + \varepsilon}{\varepsilon} \right) \]

- (Foygel and Srebro 2011): Cannot replace $\sup_i L_i$ with $\frac{1}{n} \sum_i L_i$

Using weights $w_i = \frac{L_i}{\frac{1}{n} \sum_i L_i}$, the number of iterations is:
\[ k = O \left( \frac{\frac{1}{n} \sum_i L_i \|x_*\|^2}{\varepsilon} \frac{F(x_*) + \varepsilon}{\varepsilon} \right) \]
Even less restrictive, we assume now only that

1. Each $f_i$ has Lipschitz constant $G_i$:
   $$\|f_i(x) - f_i(y)\|^2 \leq G_i \|x - y\|^2$$

- (Srebro et. al 2010): Number of iterations of SGD still depends linearly on $\sup_i L_i$.

- (Foygel and Srebro 2011): Dependence cannot be replaced with $\frac{1}{n} \sum_i L_i$ (using uniform sampling)

Using weights $w_i = \frac{L_i}{\frac{1}{n} \sum_i L_i}$, dependence is replaced by $\frac{1}{n} \sum_i L_i$
Less restrictive:

$$\| f_i(x) - f_i(y) \|^2 \leq G_i \| x - y \|^2$$

Here, SGD convergence rate depends linearly on $G_i^2$

Using weights $w_i = \frac{G_i}{\frac{1}{n} \sum_i G_i}$, dependence is reduced to

$$\left( \frac{1}{n} \sum_i G_i \right)^2 \leq \frac{1}{n} \sum_i G_i^2$$

$$\frac{1}{n} \sum_i G_i^2 = \left( \frac{1}{n} \sum_i G_i \right)^2 + \text{Var}[G_i]$$

(Zang, Zhao 2013) also consider importance sampling in this setting.
Future work:

• Strategies for sampling \textit{blocks} of indices at each iteration. Optimal block size?

• Optimal \textit{adaptive} sample strategies given limited or no information about $L_i$?

• Simulations on real data

Thanks!
References