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

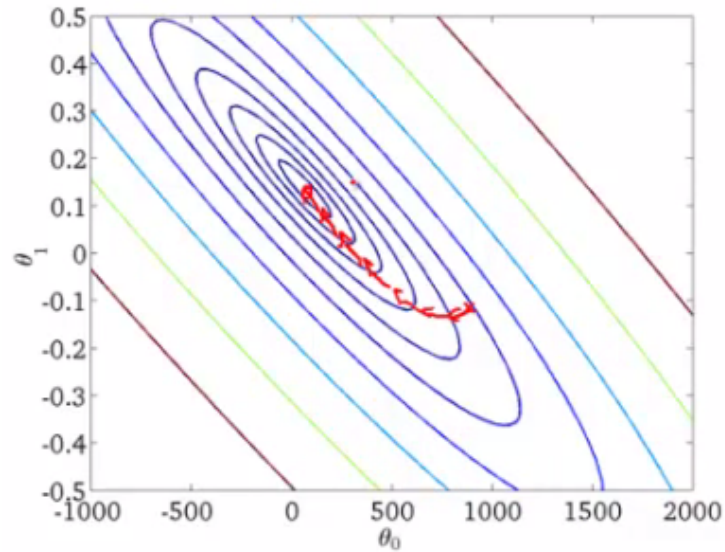
Problem: Minimize $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$, n very large

Gradient descent:

Initialize $\mathbf{x_0}$

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

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



Not practical in huge dimension n!

Stochastic gradient descent

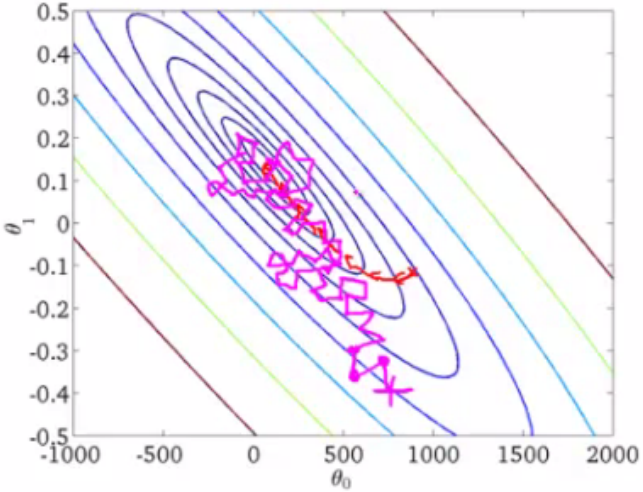
Minimize $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$

Initialize $\mathbf{x_0}$

For $j = 1, 2, \ldots$

draw index $i = i_j$ at random

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



Goal: nonasymptotic bounds on $\mathbb{E}\|\mathbf{x}^{(j)} - \mathbf{x}^*\|^2$

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

We begin by assuming

1. $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq L_i \|\mathbf{x} - \mathbf{y}\|^2$
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*,
   $\kappa_{av} = \frac{\frac{1}{n}\sum_i L_i}{\mu}$, $\kappa_{\max} = \frac{\max_i L_i}{\mu}$, $\kappa_2 = \frac{\sqrt{\frac{1}{n}\sum_i L_i^2}}{\mu}$
2. Consistency: $\sigma^2 = \mathbb{E}\|\nabla f_i(\mathbf{x}^*)\|^2$

SGD - convergence rates

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

$$\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}_*\|^2 \leq \left[1 - 2\gamma\mu(1 - \gamma \sup_i L_i)\right]^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\gamma\sigma^2}{\mu(1-\gamma \sup_i L_i)},$$

Corollary:

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

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

SGD iterations, with optimized constant step-size.

SGD - convergence rates

We showed:

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

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

Compare to:

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

$$\boxed{k \propto \frac{\sup_i L_i}{\mu} \text{ vs. } k \propto \left(\frac{\frac{1}{n}\sum_i L_i^2}{\mu}\right)^2 \text{ steps}}$$

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

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}_*\|^2 = \|\mathbf{x}^{(k)} - \mathbf{x}_* - \gamma \nabla f_i(\mathbf{x}_k)\|^2$$

$$\leq \|\mathbf{x}^{(k)} - \mathbf{x}_*\|^2 - 2\gamma < \mathbf{x}^{(k)} - \mathbf{x}_*, \nabla f_i(\mathbf{x}^{(k)}) >$$
$$+ 2\gamma^2 \|\nabla f_i(\mathbf{x}^{(k)}) - \nabla f_i(\mathbf{x}_*)\|^2 + 2\gamma^2 \|\nabla f_i(\mathbf{x}_*)\|^2$$

$$\leq \|\mathbf{x}^{(k)} - \mathbf{x}_*\|^2 - 2\gamma < \mathbf{x}^{(k)} - \mathbf{x}_*, \nabla f_i(\mathbf{x}^{(k)}) >$$
$$+ 2\gamma^2 L_i < \mathbf{x}^{(k)} - \mathbf{x}_*, \nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_*) > + 2\gamma^2 \|\nabla f_i(\mathbf{x}_*)\|^2$$

These convergence rates are tight

Consider the least squares case:

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}\left(\langle \mathbf{a_i}, \mathbf{x}\rangle - b_i\right)^2$$

$$= \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

$$\mathbf{A} = \begin{pmatrix} - & \mathbf{a_1} & - \\ - & \mathbf{a_2} & - \\ - & \mathbf{a_3} & - \\ & \vdots & \\ & \vdots & \\ & \vdots & \\ - & \mathbf{a_n} & - \end{pmatrix}$$

Assume consistency: $\mathbf{A}\mathbf{x}_* = \mathbf{b}$, $\sigma^2 = 0$

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

These convergence rates are tight

Consider the system

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

Here, $\frac{\sup_i L_i}{\mu} = n \sup_i \|\mathbf{a}_i\|^2 \|\mathbf{A}^\dagger\|^2 = n$

In this example, we need $k = n$ steps to get any accuracy

Better convergence rates using weighted sampling strategies?

SGD with weighted sampling

observe:

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

given weights w_i such that $\sum_i w_i = n$.

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

$$\begin{aligned}\mathbf{x}^{(\mathbf{j}+\mathbf{1})} &= \mathbf{x}^{(\mathbf{j})} - \gamma \nabla \tilde{f}_{i_k}(\mathbf{x}^{(\mathbf{j})}) \\ &= \mathbf{x}^{(\mathbf{j})} - \gamma \frac{1}{w_{i_k}} \nabla f_{i_k}(\mathbf{x}^{(\mathbf{j})})\end{aligned}$$

where $\mathbb{P}(i_k = i) = \frac{w_i}{\sum_j w_j}$

SGD with weighted sampling

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

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 $F(\mathbf{x}) = \mathbb{E} f_i(\mathbf{x})$ and $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma \nabla f_i(\mathbf{x}^k)$,

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

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

Corollary for weighted sampling:

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

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

Choice of weights

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

$$\mathbb{E}^{(w)} \|\mathbf{x}^k - \mathbf{x}_*\|_2^2 \leq \varepsilon \text{ after}$$
$$k = 2\log(\varepsilon/\varepsilon_0)\left(\frac{L_{\max}\left[\frac{1}{w_i} f_i\right]}{\mu} + \frac{\sigma^2\left[\frac{1}{w_i} f_i\right]}{\mu^2 \varepsilon}\right) \text{ steps}$$

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

$$\mathbb{E}^{(w)} \|\mathbf{x}^k - \mathbf{x}_*\|_2^2 \leq \varepsilon \text{ after}$$
$$k = 2\log(\varepsilon/\varepsilon_0)\left(\frac{\frac{1}{n}\sum_i L_i}{\mu}\right) \text{ steps, using weights } w_i = \frac{nL_i}{\sum_i L_i}$$

Improved convergence rate with weighted sampling

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

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

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

$$\frac{\overline{L_i}}{\mu}(\mathbf{A}) = 2 \quad \Rightarrow \quad O(1) \text{ steps using biased sampling}$$

Improved convergence rate with weighted sampling

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

But what if the system is not consistent?

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

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

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

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

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

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

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

(Partially)
biased sampling:

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

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

We have been operating in the setting

1. Each ∇f_i has Lipschitz constant L_i:
   $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq L_i \|\mathbf{x} - \mathbf{y}\|^2$

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

Other settings and weaker assumptions.
- Removing strong convexity assumption
- Non-smooth

Smoothness, but no strong convexity:

- Each ∇f_i has Lipschitz constant L_i:
  $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq L_i \|\mathbf{x} - \mathbf{y}\|^2$

- (Srebro et. al 2010): Number of iterations of SGD:

$$k = O\left(\frac{(\sup_i L_i)\|\mathbf{x}_*\|^2}{\varepsilon} \frac{F(\mathbf{x}_*) + \varepsilon}{\varepsilon}\right)$$

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

Using weights $w_i = \frac{L_i}{\frac{1}{n}\sum_i L_i}$,

$$k = O\left(\frac{(\frac{1}{n}\sum_i L_i)\|\mathbf{x}_*\|^2}{\varepsilon} \frac{F(\mathbf{x}_*) + \varepsilon}{\varepsilon}\right)$$

Even less restrictive, we assume now only that

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

- (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:

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

Here, SGD convergence rate depends linearly on $\overline{G_i^2}$

Using weights $w_i = \frac{G_i}{\frac{1}{n}\sum_i G_i}$, dependence is reduced to $(\frac{1}{n}\sum_i G_i)^2 \leq \frac{1}{n}\sum_i G_i^2$

$$\frac{1}{n}\sum_i G_i^2 = (\frac{1}{n}\sum_i G_i)^2 + \cancel{\mathrm{Var}[G_i]}$$

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

Future work:

- Strategies for sampling *blocks* of indices at each iteration. Optimal block size?
- Optimal *adaptive* sample strategies given limited or no information about L_i?
- Simulations on real data

Thanks!

References

[1] Needell, Srebro, Ward. "Stochastic gradient descent and the randomized Kaczmarz algorithm." arXiv preprint arXiv:1310.5715 (2013).