WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Closing the convergence gap of SGD without replacement

Shashank Rajput, Anant Gupta, Dimitris Papailiopoulos

Stochastic Gradient Descent (SGD)

- Problem : $\min_x F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$
- Algorithm:
  1. At each iteration, sample f_i randomly from $\{f_1, \dots, f_n\}$
  2. $x_{t+1} := x_t - \alpha \nabla f_i(x_t), \alpha$ is the step size
  3. Repeat for T iterations

Sampling with replacement

- SGD with replacement is theoretically well understood

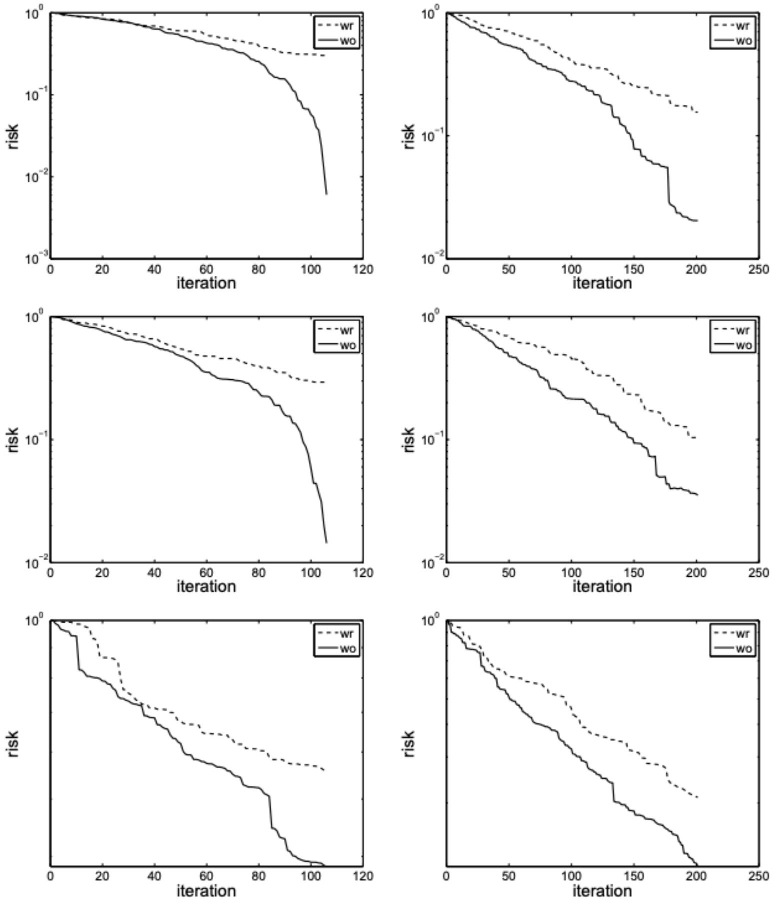
However, in practice we sample without replacement

SGD without replacement (SGDo)

1. Repeat K times
   1. $I = \{f_1, \ldots, f_n\}$
   2. Repeat n times
      1. Sample f_i uniformly at random from I
      2. Remove f_i from I
      3. $x_{t+1} = x_t - \alpha \nabla f_i(x_t)$

} Epoch

Known to be faster in practice! [1]



With v/s Without replacement [2]

[1]: Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. 2009
[2]: Benjamin Recht and Christopher Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. 2012

Why should SGDo be faster?

- Example:
  - Let $f_1(x) = (x+1)^2,\ f_2(x) = (x-1)^2$. Start at $x = 0$.
  - SGDo: Both functions seen in epoch, iterates stay close to 0.
  - SGD: With probability $1/2$, f_1 missed or f_2 missed.

- SGDo : Every function is seen once in n iterations.
- SGD : In n iterations, n/e functions missed.

Variance over an epoch is reduced for SGDo!

SGDo – Theoretically elusive

- Until recently, SGDo eluded theoretical analysis
- Why?
  - SGD: Easy because $\mathbb{E}[\nabla f_i(x_t)] = \nabla F(x_t)$
  - SGDo: Difficult because $\mathbb{E}[\nabla f_i(x_t)] \neq \nabla F(x_t)$

- Error metric: $\mathbb{E}[\|x_T - x^*\|^2]$
- SGD error $= O\left(\frac{1}{T}\right)$

Can SGDo (provably) do better?

Our results

- SGDo error bounds:

n = # functions
K = # epochs
$T = nK$
SGD error $= O(1/T)$

Upper bound [3,4]	$O\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right), O\left(\frac{n}{T^2}\right)$
Lower bound [5]	$\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$
Our upper bound	$O\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$

Neither upper bound is better than the other!

F is strongly convex quadratic

[3]: Jeffery Z HaoChen and Suvrit Sra. Random shuffling beats sgd after finite epochs. 2018
[4]: Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. 2019
[5]: Itay Safran and Ohad Shamir. How good is sgd with random shuffling?

Our results

- SGDo error bounds:

n = # functions
K = # epochs
$T = nK$
SGD error = $O(1/T)$

Upper bound [4]	$O\left(\frac{n}{T^2}\right)$
Lower bound [5]	$\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$
Our lower bound	$\Omega\left(\frac{n}{T^2}\right)$

Surprisingly, lower bound is different for non-quadratics!

F is strongly convex smooth function

[4]: Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. 2019
[5]: Itay Safran and Ohad Shamir. How good is sgd with random shuffling?

Upper bound

Upper bound - Approach

- $x_1 =$ Start of epoch, $x_n =$ End of epoch
- Idea [3]:

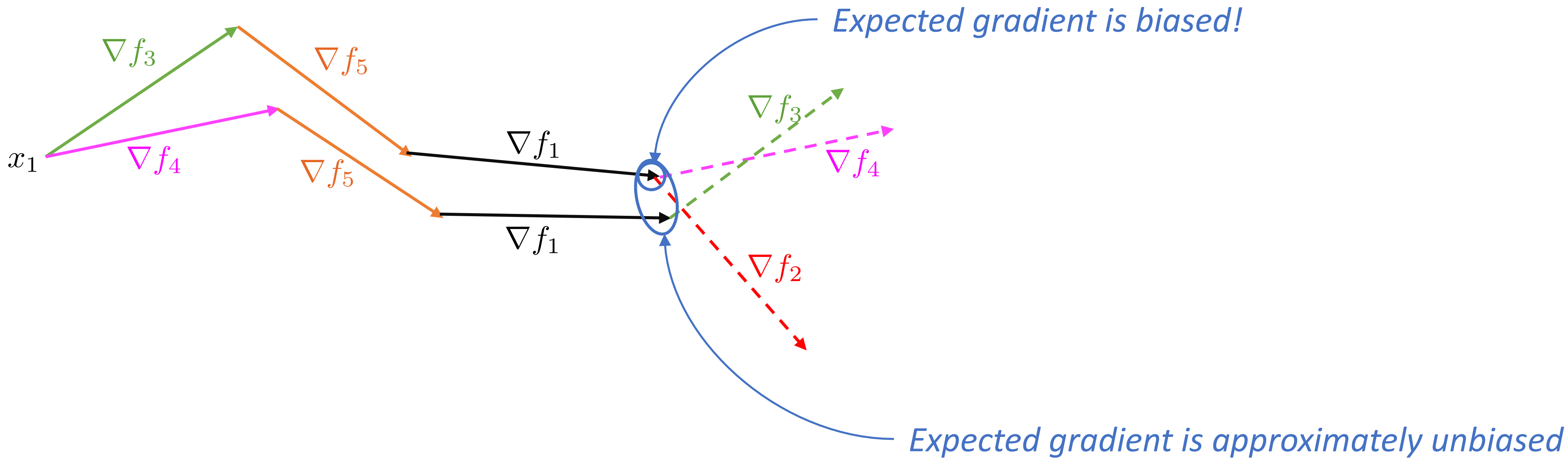
$$x_n - x_1 = \alpha \sum_i \nabla f_{\sigma(i)}(x_i) \approx \alpha \sum_i \nabla f_{\sigma(i)}(x_1) = \underbrace{\alpha n \nabla F(x_1)}_{n \text{ steps of gradient descent!}}$$

- Key lemma [4] : $\mathbb{E}[\|x_i - x_1\|^2] = O(i\alpha^2)$
  - $\|x_i - x_1\|^2$ grows as $i\alpha^2$ instead of $i^2\alpha^2$
  - (Tight!)

[3]: Jeffery Z HaoChen and Suvrit Sra. Random shuffling beats sgd after finite epochs. 2018
[4]: Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. 2019

Iterate coupling

- Assume $n = 5$: $\{f_1, f_2, f_3, f_4, f_5\}$



- Same coupling as [4]

[4]: Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. 2019

Lower bound

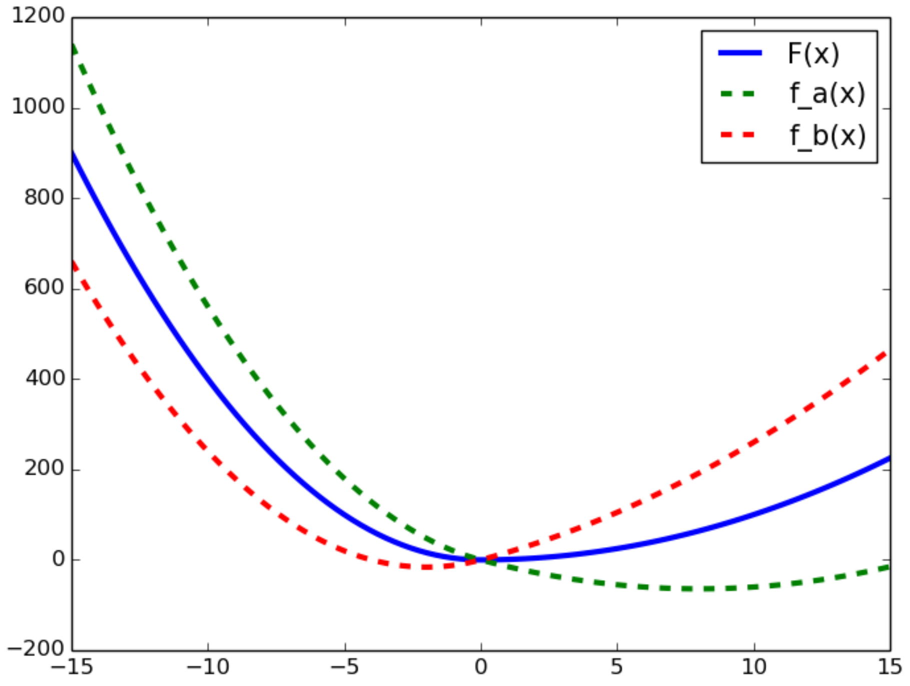
Lower bound - Function

- If F has Lipschitz Hessian, Error $= O\left(\frac{n^3}{T^3} + \frac{1}{T^2}\right)$ [3]
- Need non-Lipschitz Hessian : Piece-wise quadratic!

$$F(x) = \frac{1}{n}\left(\sum_{i=1}^{n/2} f_a(x) + \sum_{i=1}^{n/2} f_b(x)\right)$$

$$\text{where, } f_a(x) = \begin{cases} x^2 + x & x \geq 0 \\ Rx^2 + x & x < 0 \end{cases}$$

$$\text{and, } f_b(x) = \begin{cases} x^2 - x & x \geq 0 \\ Rx^2 - x & x < 0 \end{cases}$$



- Hessian discontinuous at 0, minimizer at 0.

[3]: Jeffery Z HaoChen and Suvrit Sra. Random shuffling beats sgd after finite epochs. 2018

Proof sketch

- Consider the function gradients

$$\nabla f_a(x) = \begin{cases} 2x + 1 & x \geq 0 \\ 2Rx + 1 & x < 0 \end{cases} \qquad \nabla f_b(x) = \begin{cases} 2x - 1 & x \geq 0 \\ 2Rx - 1 & x < 0 \end{cases}$$

- When x is small, the gradient is dominated by the gradients of linear terms
- These are Rademacher variables (but not independent)
- For $i \leq \frac{n}{4}$, $|x_i| \geq C\alpha\sqrt{i}$

Proof sketch

$$\nabla f_a(x) = \begin{cases} 2x + 1 & x \geq 0 \\ 2Rx + 1 & x < 0 \end{cases}$$

$$\nabla f_b(x) = \begin{cases} 2x - 1 & x \geq 0 \\ 2Rx - 1 & x < 0 \end{cases}$$

- $x_n - x_1 = \alpha \sum_{i=1}^{n} \nabla f_{\sigma(i)}(x_i)$
- The sum of gradients from linear terms $= \mathbf{0}$
- The sum of gradients from quadratic terms

$$\sum_{x_i<0} \alpha R x_i + \sum_{x_i \geq 0} \alpha x_i \approx \sum_{x_i \geq 0} \alpha R x_i \quad (\text{assume } R >> 1)$$

Plug the value from previous slide and recurse for K epochs

Conclusion

- In this work, we close the gap in convergence rates of SGDo.
- We discovered an interesting phenomenon :

SGDo converges faster for strongly convex quadratics than general strongly convex smooth functions.

Future Work

- Do there exist "optimal" permutations? Distribution of convergence rates for permutations.
- Can these analyses be extended to algorithms that compress gradients?
- Can we analyze convergence for "system-friendly" shuffling schemes?