Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization



Blake Woodworth
(TTIC)

Jialei Wang
(UChicago→2σ Investments)

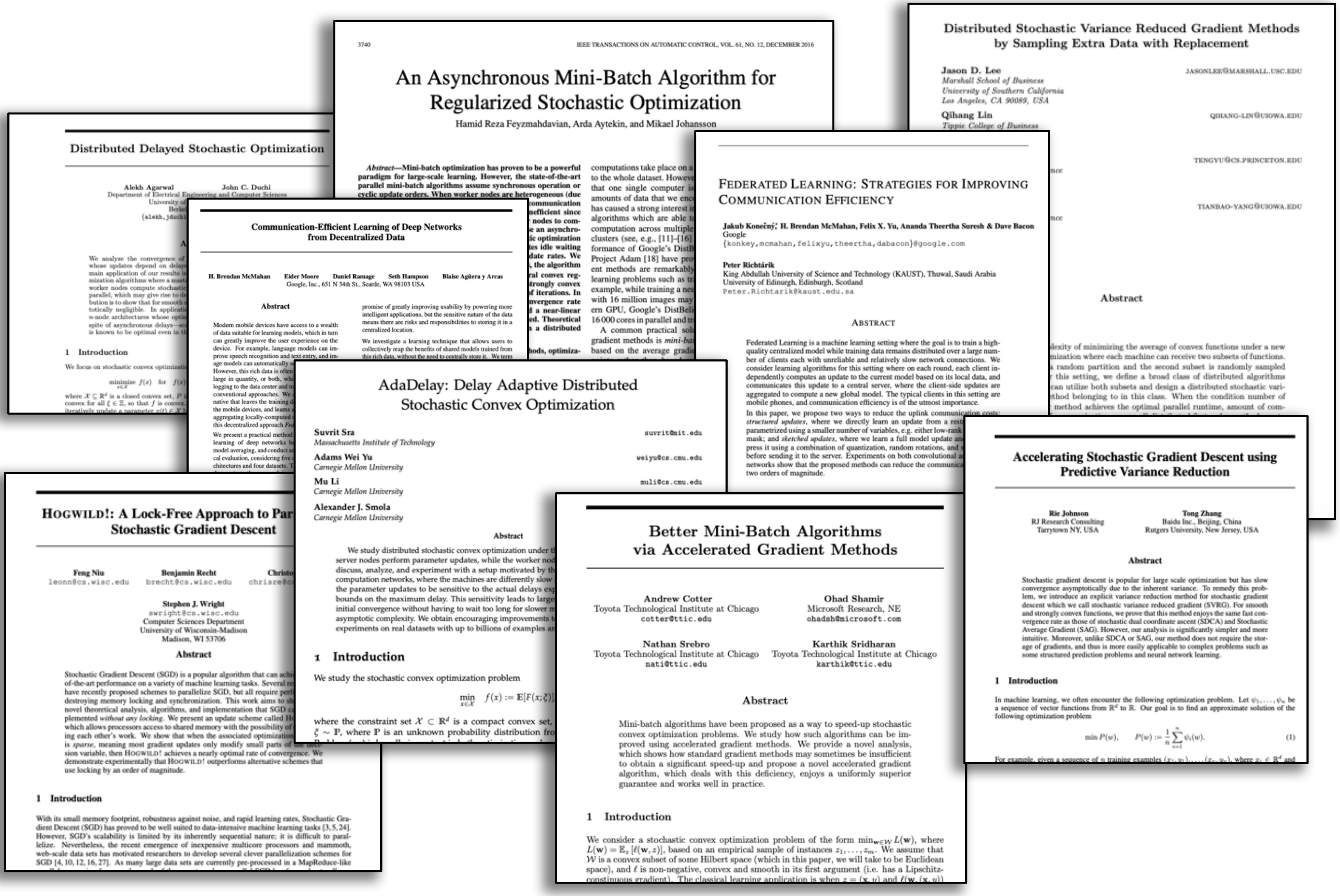
Adam Smith
(Boston University)


H. Brendan McMahan
(Google)

Nati Srebro
(TTIC)

Parallel Stochastic Optimization/Learning

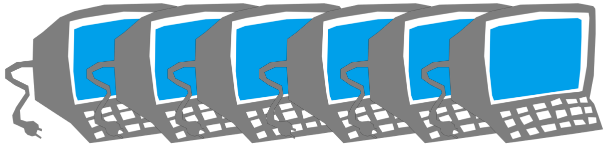
$$\min_x F(x) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [f(x;z)]$$

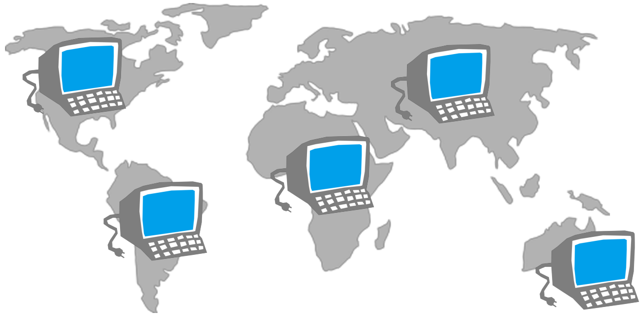


Many parallelization scenarios:

- Synchronous parallelism
- Asynchronous parallelism
- Delayed updates
- Few/many workers
- Infrequent communication
- Federated learning

…

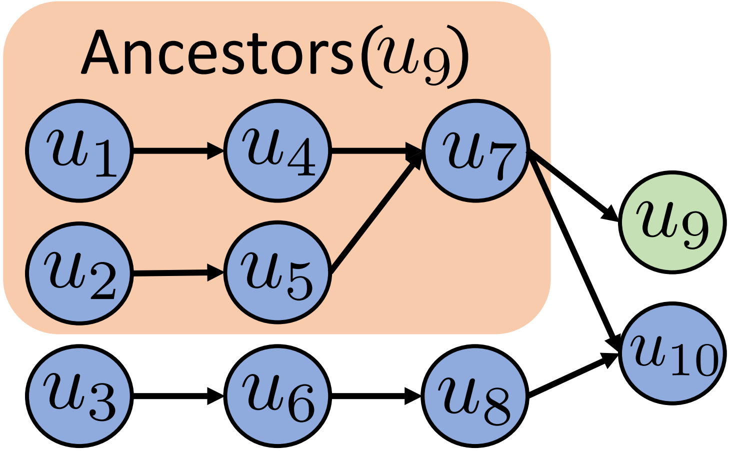




What is the best we can hope for in a given parallelism scenario?

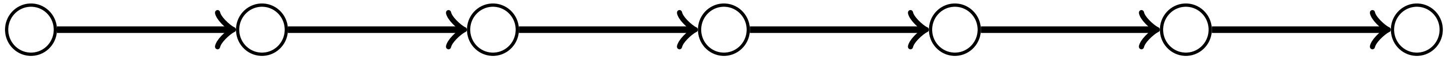
What is the best we can hope for in a given parallelism scenario?

- We formalize the parallelism in terms of a dependency graph:

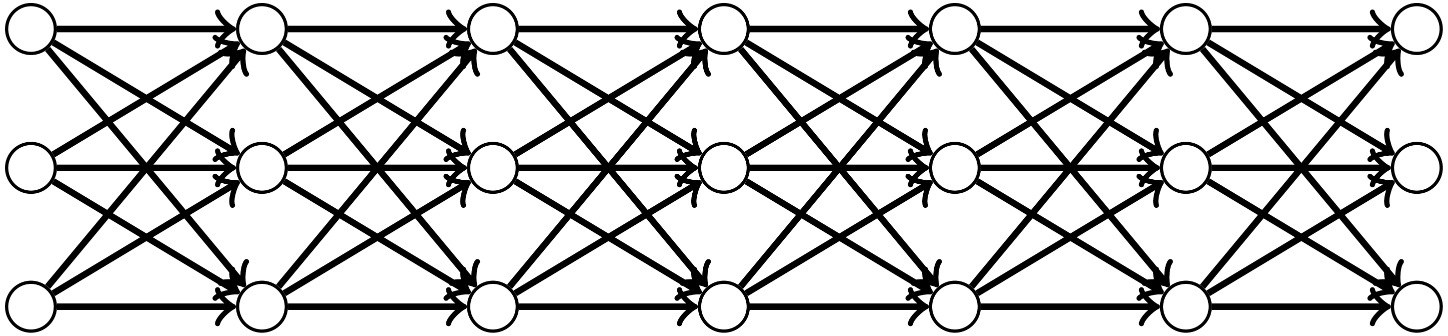


- At each node u, make a query based only on knowledge of *ancestors*' oracle interaction (plus shared randomness)
- Graph defines class of optimization algorithms $\mathcal{A}(\mathcal{G})$
- Come to our poster for details

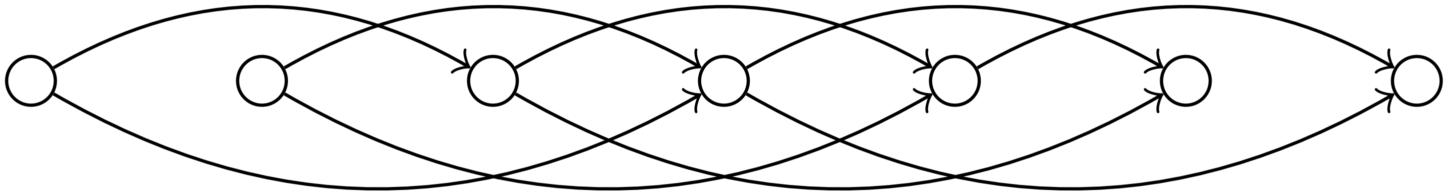
- Sequential:



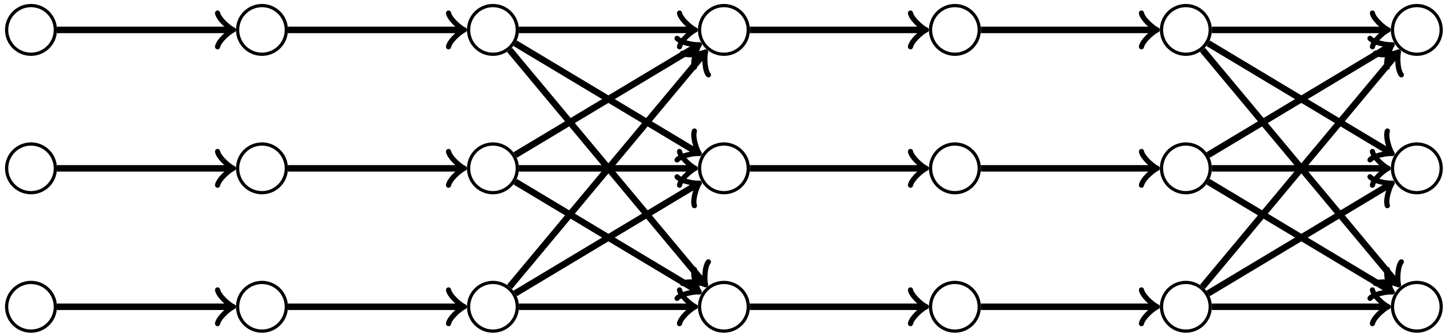
- Layer:



- Delays:



- Intermittent Communication:



Generic Lower Bounds

Theorem: For any dependency graph $\mathcal{G}$ with N nodes and depth D, no algorithm for optimizing convex, L-Lipschitz, H-smooth $f(x;z)$ on a bounded domain in high dimensions can guarantee error less than:

With stochastic gradient oracle:

$$\Omega\left(\min\left\{\frac{L}{\sqrt{D}}, \frac{H}{D^2}\right\} + \frac{L}{\sqrt{N}}\right)$$

Generic Lower Bounds

Theorem: For any dependency graph $\mathcal{G}$ with N nodes and depth D, no algorithm for optimizing convex, L-Lipschitz, H-smooth $f(x;z)$ on a bounded domain in high dimensions can guarantee error less than:

With stochastic gradient oracle:

$$\Omega\left(\min\left\{\frac{L}{\sqrt{D}}, \frac{H}{D^2}\right\} + \frac{L}{\sqrt{N}}\right)$$

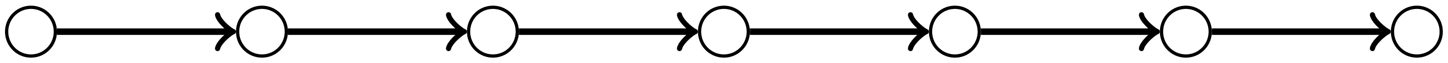
With stochastic prox oracle:

$$\Omega\left(\min\left\{\frac{L}{D}, \frac{H}{D^2}\right\} + \frac{L}{\sqrt{N}}\right)$$

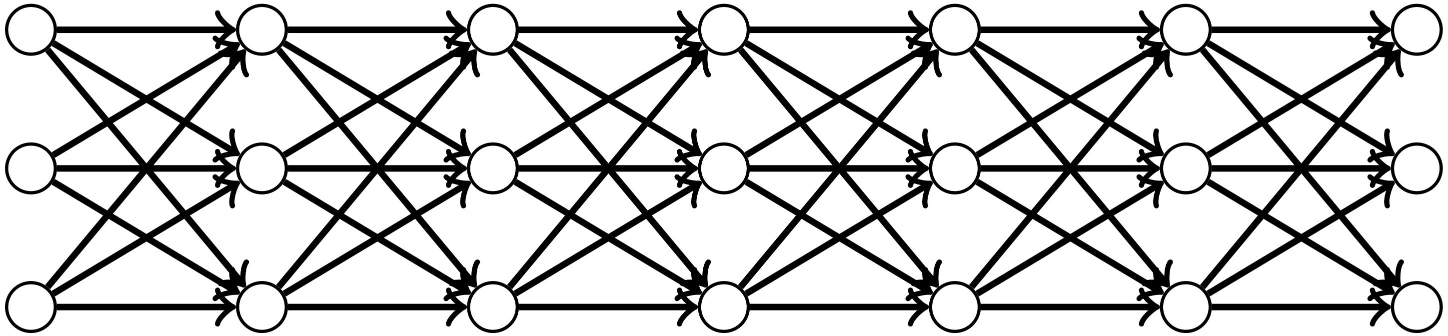
Prox oracle: $x, \beta, z \mapsto \arg\min_y f(y;z) + \frac{\beta}{2}\|y-x\|^2$

i.e. exactly optimize subproblem in each node (ADMM, DANE, etc.)

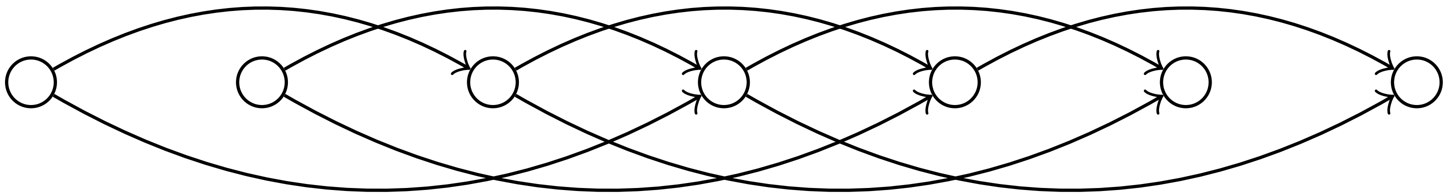
- Sequential:
  - SGD is optimal



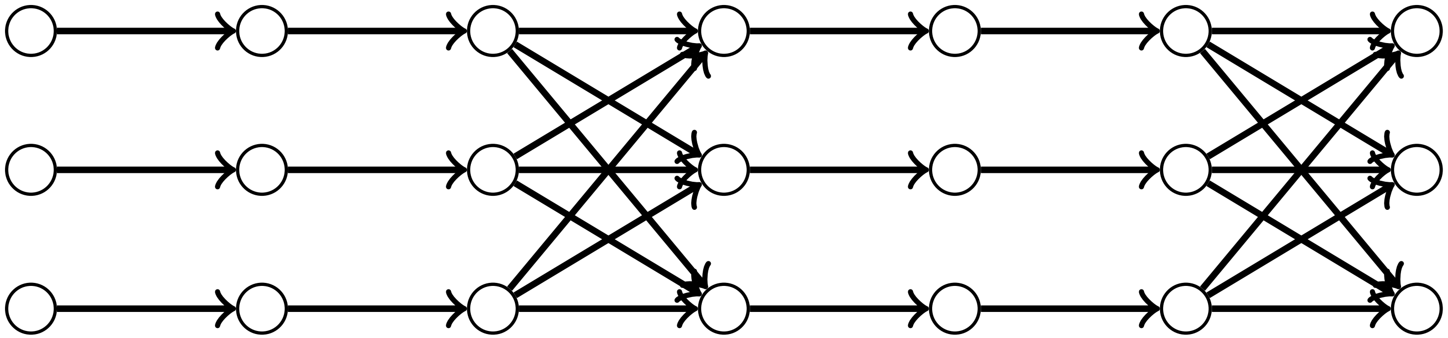
- Layers:
  - Accelerated minibatch SGD is optimal



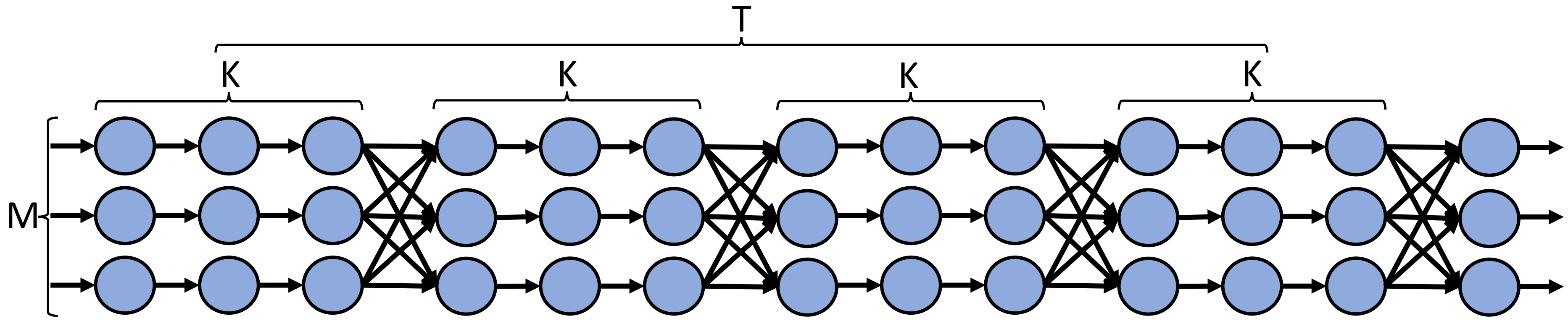
- Delays:
  - Delayed-update SGD is *not* optimal
  - "Wait-and-Collect" minibatch is optimal



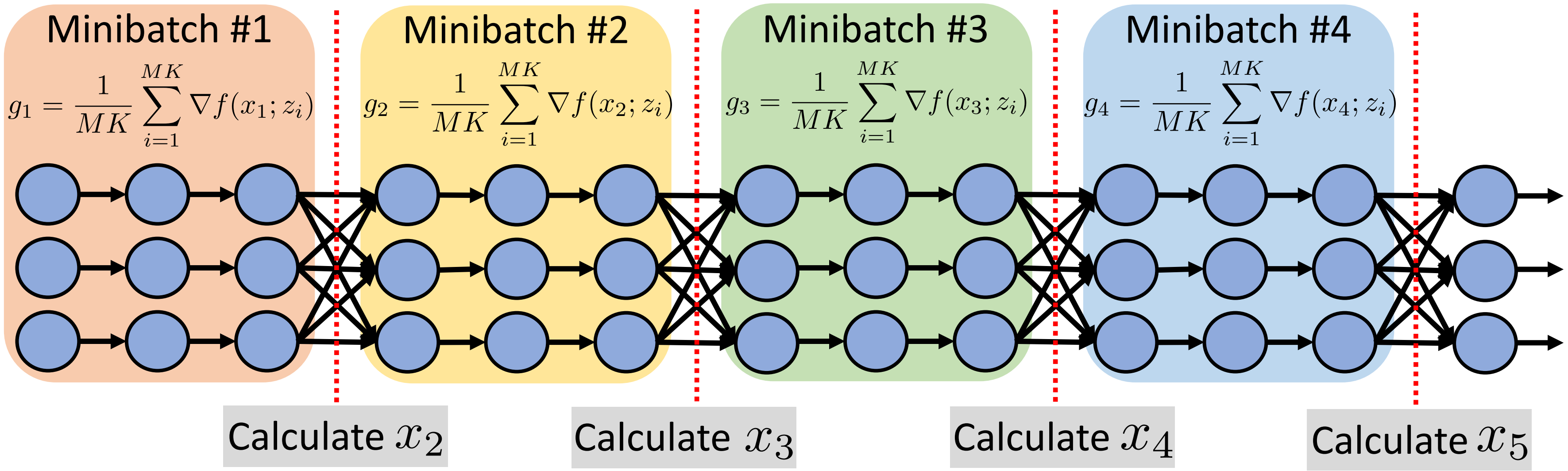
- Intermittent Communication:
  - Gaps between existing algorithms and lower bound



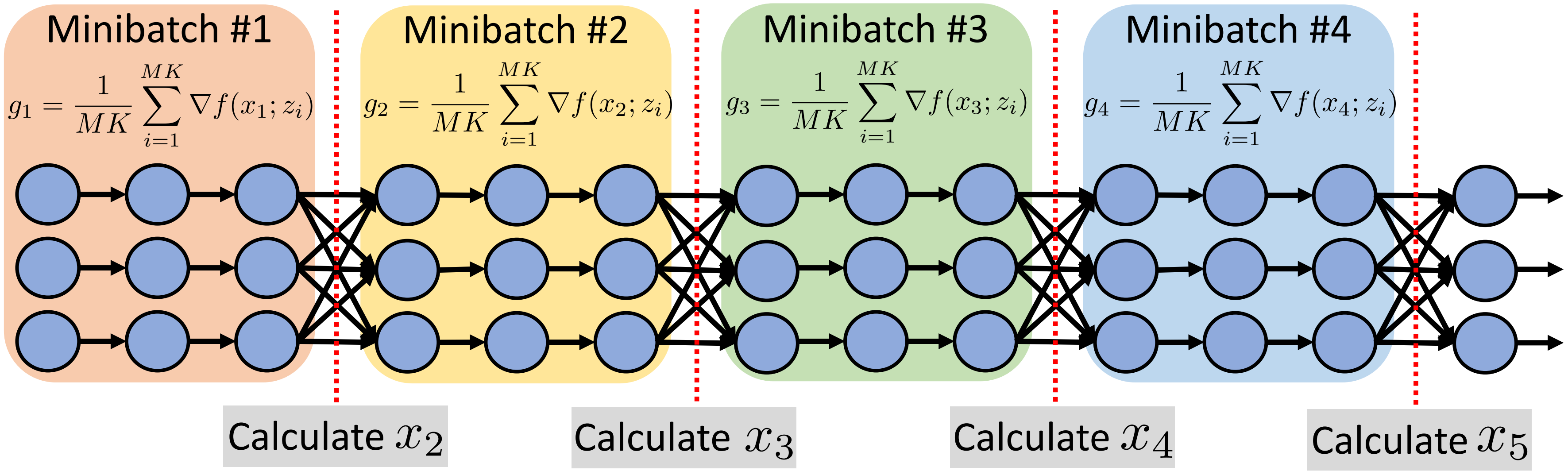
- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$



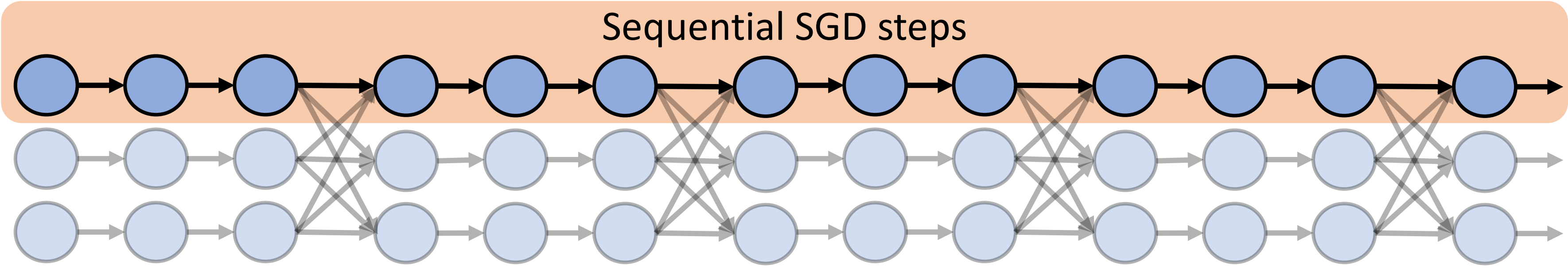
- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$



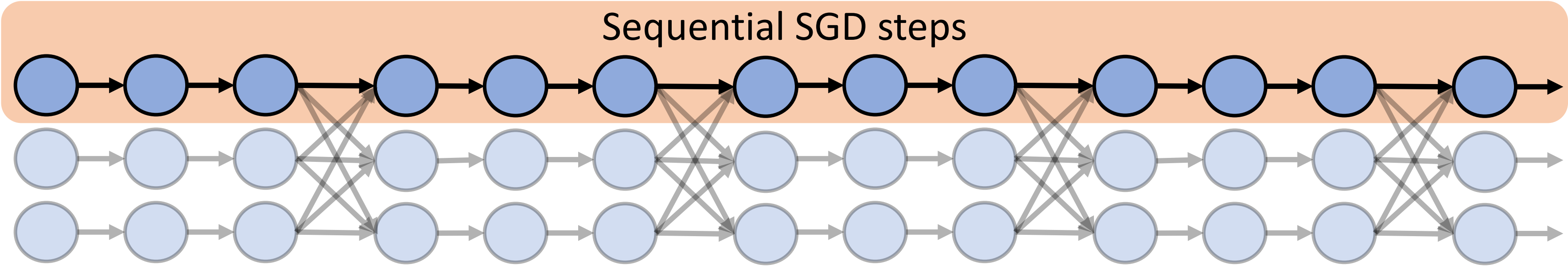
- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$



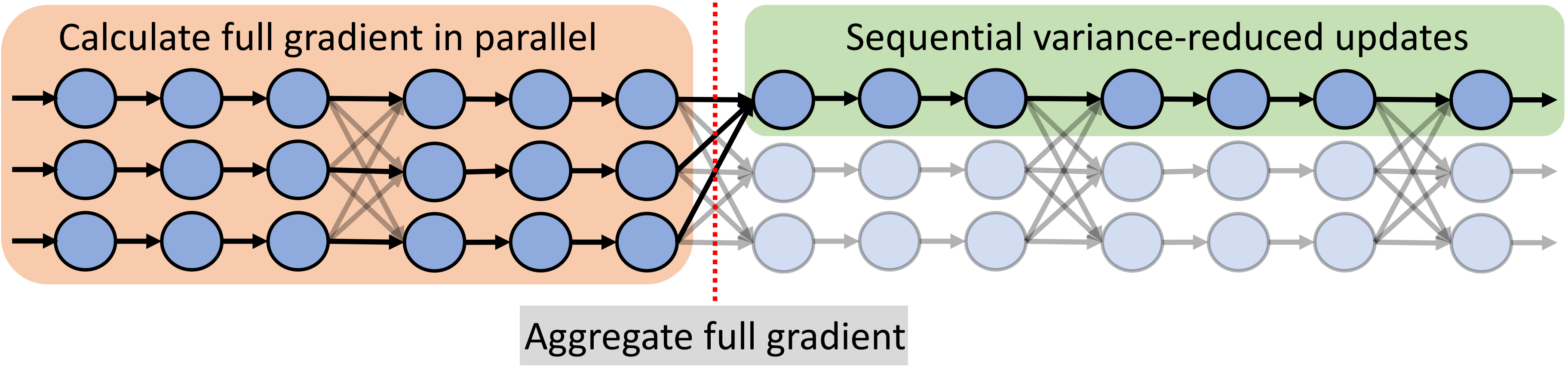
- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$



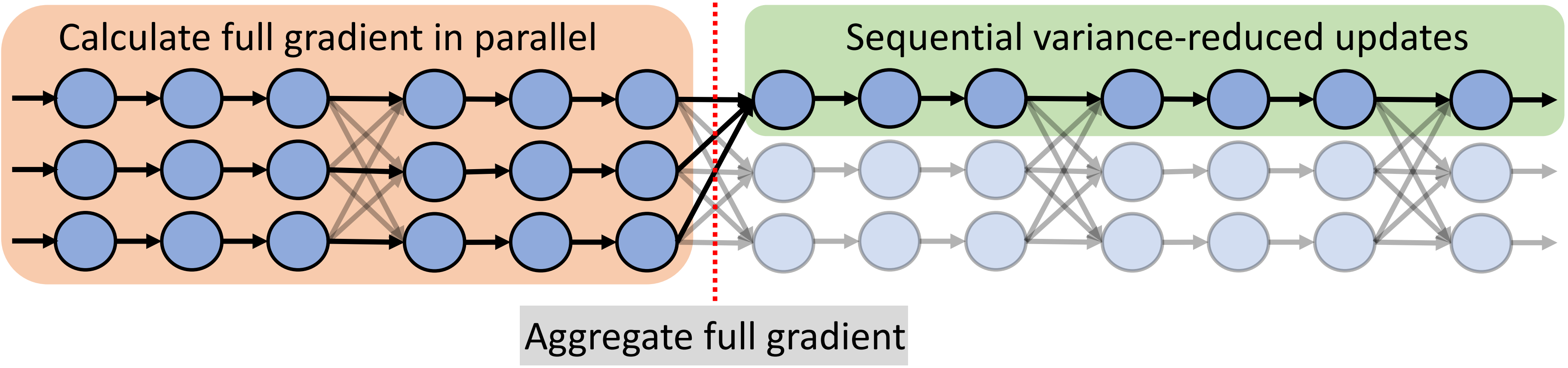
- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$



- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$
- Option 3: SVRG on empirical objective $\tilde{O}\left(\frac{H}{TK} + \frac{L}{\sqrt{TKM}}\right)$



- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}},\frac{H}{T^2K^2}\right\}+\frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2}+\frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$
- Option 3: SVRG on empirical objective $\tilde{O}\left(\frac{H}{TK}+\frac{L}{\sqrt{TKM}}\right)$



- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$
- Option 3: SVRG on empirical objective $\tilde{O}\left(\frac{H}{TK} + \frac{L}{\sqrt{TKM}}\right)$
- Combining 1-3:

$$\tilde{O}\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{TK}, \frac{H}{T^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$$

- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$
- Option 3: SVRG on empirical objective $\tilde{O}\left(\frac{H}{TK} + \frac{L}{\sqrt{TKM}}\right)$
- Combining 1-3:

$$\tilde{O}\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{TK}, \frac{H}{T^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$$

- Lower bound: $\Omega\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{T^2K^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$
- Option 1: Accelerated Minibatch SGD $O\left(\frac{H}{T^2} + \frac{L}{\sqrt{TKM}}\right)$
- Option 2: Sequential SGD $O\left(\frac{L}{\sqrt{TK}}\right)$
- Option 3: SVRG on empirical objective $\tilde{O}\left(\frac{H}{TK} + \frac{L}{\sqrt{TKM}}\right)$
- Combining 1-3:

$$\tilde{O}\left(\min\left\{\frac{L}{\sqrt{TK}}, \frac{H}{TK}, \frac{H}{T^2}\right\} + \frac{L}{\sqrt{TKM}}\right)$$

- Option 4: Parallel SGD ???

Come to our poster
tonight from 5-7pm