Expected smoothness is the key to understanding minibatching for stochastic gradient methods

Robert M. Gower

Joint work with Francis Bach, Nidham Gazagnadou, Nicolas Loizou, Xun Qian, Peter Richtarik, Alibek Sailanbayev, Othmane Sebbouh and Egor Shulgin.

July, 2019
Optimization in Machine Learning

(1) Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)
Optimization in Machine Learning

(1) Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)

(2) Choose a classifier: \(h_w(x) \rightarrow y\)

\[
\begin{align*}
h_w(\text{Cat}) & \rightarrow \text{Cat} \\
\end{align*}
\]
Optimization in Machine Learning

(1) Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)

(2) Choose a classifier: \(h_w(x) \mapsto y\)

(3) Choose a loss function: \(\ell(h_w(x), y) \geq 0\)
Optimization in Machine Learning

1. Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)

2. Choose a classifier: \(h_w(x) \rightarrow y\)

3. Choose a loss function: \(\ell(h_w(x), y) \geq 0\)

4. Solve the training problem:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell \left( h_w(x^i), y^i \right)
\]
Optimization in Machine Learning

(1) Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)

(2) Choose a classifier: \(h_w(x) \mapsto y\)

(3) Choose a loss function: \(\ell(h_w(x), y) \geq 0\)

(4) Solve the *training problem*:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x^i), y^i)
\]

(5) Test and cross-validate. If fail, go back a few steps
Optimization in Machine Learning

(1) Get data: \((x^1, y^1), \ldots, (x^n, y^n)\)

(2) Choose a classifier: \(h_w(x) \rightarrow y\)

(3) Choose a loss function: \(\ell(h_w(x), y) \geq 0\)

(4) Solve the training problem:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x^i), y^i)
\]

(5) Test and cross-validate. If fail, go back a few steps
Finite sum minimization

\[
(I) \quad \min_{w \in \mathbb{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Mission statement:
“Develop an \textit{informative} analysis for stochastic gradient algorithms for solving (I) that \textit{saves time} for practitioners and theorists.”
Finite sum minimization

\[(I) \quad \min_{w \in \mathbb{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \ell(h_w(x^i), y^i) \]

Mission statement:
“Develop an \textit{informative} analysis for stochastic gradient algorithms for solving (I) that \textit{saves time} for practitioners and theorists.”

\textit{informative}: tight with realistic assumptions \rightarrow inform \textit{parameter choices} and implementations
Finite sum minimization

\[ (I) \quad \min_{w \in \mathbb{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \]

Mission statement:
“Develop an \textit{informative} analysis for stochastic gradient algorithms for solving (I) that \textit{saves time} for practitioners and theorists.”

\textit{informative}: tight with realistic assumptions \quad inform parameter choices and implementations

\textit{saves time for practitioners}: Less hyper-parameter tuning \quad works out of the box

\textit{saves time for theorists}: Simplify and unifies existing theory.
Finite sum minimization

\[ (I) \quad \min_{w \in \mathbb{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \]

Mission statement:
“Develop an **informative** analysis for stochastic gradient algorithms for solving \((I)\) that **saves time** for practitioners and theorists.”

*informative*: tight with realistic assumptions ▶ inform parameter choices and implementations

*saves time for practitioners*: Less **hyper-parameter tuning** ▶ works out of the box

*saves time for theorists*: Simplify and unifies existing theory.

Case study today: **Learning rates/stepsizes** and **minibatch size** for SGD and stochastic variance reduced methods SAGA and SVRG
The Stochastic Gradient Method

Solving the training problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Baseline method: Stochastic Gradient Descent (SGD)

$$w^{t+1} = w^t - \gamma_t \nabla f_j(w^t)$$

Step size/Learning rate

Sampled i.i.d

\(j \in \{1, \ldots, n\}\)
The Stochastic Gradient Method

Solving the training problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Baseline method: Stochastic Gradient Descent (SGD)

$$w^{t+1} = w^t - \gamma_t \nabla f_j(w^t)$$

What about mini-batching

Step size/ Learning rate

Sampled i.i.d

$$j \in \{1, \ldots, n\}$$
The Stochastic Gradient Method

Solving the \textit{training problem}:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Baseline method: Stochastic Gradient Descent (SGD)

\[
\omega^{t+1} = \omega^t - \gamma_t \frac{1}{b} \sum_{j \in B} \nabla f_j(\omega^t)
\]

Minibatch where \( B \in \{1, \ldots, n\} \) with \( |B| = b \)
The Stochastic Gradient Method

Solving the training problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Baseline method: Stochastic Gradient Descent (SGD)

$$w^{t+1} = w^t - \gamma_t \frac{1}{b} \sum_{j \in B} \nabla f_j(w^t)$$

• What should $b$ be?

Minibatch where $B \in \{1, \ldots, n\}$ with $|B| = b$
The Stochastic Gradient Method

Solving the *training problem*:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Baseline method: Stochastic Gradient Descent (SGD)

$$w^{t+1} = w^t - \gamma_t \cdot \frac{1}{b} \sum_{j \in B} \nabla f_j(w^t)$$

- What should $b$ be?
- How does $b$ influence the stepsizes?

Minibatch where $B \in \{1, \ldots, n\}$ with $|B| = b$
The Stochastic Gradient Method

Solving the training problem:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Baseline method: Stochastic Gradient Descent (SGD)

\[
\omega^{t+1} = \omega^t - \gamma_t \frac{1}{b} \sum_{j \in B} \nabla f_j(\omega^t)
\]

- What should \( b \) be?
- How does \( b \) influence the stepsizes?
- How does the data influence the best mini-batch and stepsize?
How to choose the minibatch size?

Cross validation score

$\gamma$

step size

0

1 minibatch size $b$
How to choose the minibatch size?

Cross validation score

$\gamma$

step size

0

1 minibatch size $b$
How to choose the minibatch size?

Cross validation score

$\gamma$

step size

0

1 minibatch size

$b$
How to choose the minibatch size?

Cross validation score

γ

step size

0

1 minibatch size b
How to choose the minibatch size?

Cross validation score

γ
step size
0
1 minibatch size b
How to choose the minibatch size?

Cross validation score

\[ \gamma \]

step size

0

1 minibatch size

\[ b \]
How to choose the minibatch size?

Cross validation score

\[ \gamma \]

step size

0

1 minibatch size

\[ b \]
How to choose the minibatch size?

Cross validation score

$\gamma$

Step size

0

1 minibatch size

$b$
How to choose the minibatch size?
How to choose the minibatch size?

Cross validation score

- **Good**
- **Bad**

Step size vs. minibatch size
How to choose the minibatch size?

Best parameters

Cross validation score

Minibatch size

Step size

Bad

Good
Linear Scaling Rule: When the minibatch size is multiplied by $k$, multiply the learning rate by $k$.

**How to choose the minibatch size?**

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017

**Best parameters**

Cross validation score

good

bad

0 1 minibatch size $b$

step size

$\gamma$
Linear Scaling Rule: When the minibatch size is multiplied by $k$, multiply the learning rate by $k$.

How to choose the minibatch size?

Cross validation score

$\gamma(b) = \text{const} \times b$?

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017
How to choose the minibatch size?

Linear Scaling Rule: When the minibatch size is multiplied by \( k \), multiply the learning rate by \( k \).

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017
How to choose the minibatch size?

**Linear Scaling Rule:** When the minibatch size is multiplied by $k$, multiply the learning rate by $k$.

![Graph showing the relationship between minibatch size and cross validation score.](image-url)

**Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017**
How to choose the minibatch size?

Linear Scaling Rule: When the minibatch size is multiplied by $k$, multiply the learning rate by $k$. 

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017
How to choose the minibatch size?

Linear Scaling Rule: When the minibatch size is multiplied by $k$, multiply the learning rate by $k$.

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017

$\gamma(b) = \text{const} \times b$?
How to choose the minibatch size?

- Need to figure out functional relationship between minibatch size and step size.
- Missed the best one.
- \[ \gamma(b) = \text{const} \times b? \]

**Linear Scaling Rule:** When the minibatch size is multiplied by \( k \), multiply the learning rate by \( k \).

**Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., CoRR 2017**
Stochastic Reformulation of Finite sum problems
Simple Stochastic Reformulation

Random sampling vector \( v = (v_1, \ldots, v_n) \sim \mathcal{D} \) with

\[ \mathbb{E}[v_i] = 1, \quad \text{for } i = 1, \ldots, n \]
Simple Stochastic Reformulation

Random sampling vector $\mathbf{v} = (v_1, \ldots, v_n) \sim \mathcal{D}$ with
\[ \mathbb{E}[v_i] = 1, \quad \text{for } i = 1, \ldots, n \]

\[
 f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_i] f_i(w) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]
Random sampling vector \( v = (v_1, \ldots, v_n) \sim \mathcal{D} \) with
\[
\mathbb{E}[v_i] = 1, \quad \text{for } i = 1, \ldots, n
\]

\[
f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_i] f_i(w) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]

\[=: f_v(w)\]

Simple Stochastic Reformulation
Simple Stochastic Reformulation

Random sampling vector $v = (v_1, \ldots, v_n) \sim D$ with
\[
\mathbb{E}[v_i] = 1, \quad \text{for } i = 1, \ldots, n
\]

\[
f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_i] f_i(w) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right] =: f_v(w)
\]

Original finite sum problem
\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Stochastic Reformulation
\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) \right]
\]

Minimizing the expectation of random linear combinations of original function
SGD with arbitrary sampling

$$\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]$$
SGD with arbitrary sampling

$$\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]$$

Sample $v^t \sim \mathcal{D}$

$$w^{t+1} = w^t - \gamma_t \nabla f_v^t(w^t)$$

By design we have that

$$\mathbb{E} [\nabla f_v^t(w^t)] = \nabla f(w^t)$$
SGD with arbitrary sampling

\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_{\nu}(w) := \frac{1}{n} \sum_{i=1}^{n} \nu_i f_i(w) \right]
\]

By design we have that

Example: Gradient descent

\[\nu \equiv (1, \ldots, 1) \implies w^{t+1} = w^t - \gamma_t \nabla f(w^t)\]

The distribution \( \mathcal{D} \) encodes any form of mini-batching/ non-uniform sampling. Our analysis is done for any distribution \( \mathcal{D} \).

By design we have that

\[\mathbb{E}[\nabla f_{\nu^t}(w^t)] = \nabla f(w^t)\]
SGD with arbitrary sampling

$$\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]$$

**saves time** for theorists: One representation for all forms of sampling

Sample $\nu^t \sim \mathcal{D}$

$$w^{t+1} = w^t - \gamma_t \nabla f_{\nu^t}(w^t)$$

The distribution $\mathcal{D}$ encodes any form of mini-batching/ non-uniform sampling. Our analysis is done for any distribution $\mathcal{D}$.

Example: Gradient descent

$\nu \equiv (1, \ldots, 1)$

$$w^{t+1} = w^t - \gamma_t \nabla f(w^t)$$

By design we have that

$$\mathbb{E}[\nabla f_{\nu^t}(w^t)] = \nabla f(w^t)$$
Examples of arbitrary sampling: uniform single element

Random set $S \subset \{1, \ldots, n\}$, $|S| = 1$

$\text{Prob}[i \in S] = 1/n$, for $i = 1, \ldots, n$
Examples of arbitrary sampling:
uniform single element

Random set $S \subset \{1, \ldots, n\}$, $|S| = 1$

$\text{Prob}[i \in S] = 1/n$, for $i = 1, \ldots, n$

$v_i = \begin{cases} n & i \in S \\ 0 & i \notin S \end{cases}$

$\mathbb{E}[v_i] = 1$
Examples of arbitrary sampling: uniform single element

Random set $S \subset \{1, \ldots, n\}$, $|S| = 1$

$\text{Prob}[i \in S] = 1/n$, for $i = 1, \ldots, n$

$v_i = \begin{cases} n & i \in S \\ 0 & i \not\in S \end{cases}$

$\mathbb{E}[v_i] = 1$

$\nabla f_v(w) = \nabla f_i(w)$

$\mathbb{E}[\nabla f_v(w)] = \nabla f(w)$
Examples of arbitrary sampling: uniform single element

Random set $S \subseteq \{1, \ldots, n\}$, $|S| = 1$

$\text{Prob}[i \in S] = 1/n$, for $i = 1, \ldots, n$

$\mathbb{E}[v_i] = 1$

$\nabla f_\nu(w) = \nabla f_i(w)$

$\mathbb{E}[\nabla f_\nu(w)] = \nabla f(w)$

Single element SGD

Sample $v^t \sim \mathcal{D}$

$w^{t+1} = w^t - \gamma_t \nabla f_\nu^t(w^t)$
Examples of arbitrary sampling: uniform mini-batching

Random set $S \subset \{1, \ldots, n\}$, $|S| = b$

$\text{Prob}[i \in S] = b/n$, for $i = 1, \ldots, n$

Mini-batch SGD without replacement

Sample $v^t \sim \mathcal{D}$

$w^{t+1} = w^t - \gamma_t \nabla f_{v^t}(w^t)$

$v_i = \begin{cases} 
\frac{n}{b} & i \in S \\
0 & i \not\in S 
\end{cases}$

$\mathbb{E}[v_i] = 1$

$\nabla f_v(w) = \frac{1}{b} \sum_{i \in S} \nabla f_i(w)$

$\mathbb{E}[\nabla f_v(w)] = \nabla f(w)$
Examples of arbitrary sampling: non-uniform mini-batching

Random set $S \subset \{1, \ldots, n\}$, $\mathbb{E}|S| = b$

$\text{Prob}[i \in S] = p_i$, for $i = 1, \ldots, n$
Examples of arbitrary sampling: non-uniform mini-batching

Random set $S \subset \{1, \ldots, n\}$, $\mathbb{E}|S| = b$

$\text{Prob}[i \in S] = p_i$, for $i = 1, \ldots, n$

$v_i = \begin{cases} \frac{1}{p_i} & i \in S \\ 0 & i \notin S \end{cases}$

$\mathbb{E}[v_i] = 1$
Examples of arbitrary sampling: non-uniform mini-batching

Random set $S \subset \{1, \ldots, n\}$, $\mathbb{E}|S| = b$

$\text{Prob}[i \in S] = p_i$, for $i = 1, \ldots, n$

$v_i = \begin{cases} 
\frac{1}{p_i} & i \in S \\
0 & i \notin S 
\end{cases}$

$\mathbb{E}[v_i] = 1$

$\nabla f_v(w) = \frac{n}{p_i} \sum_{i \in S} \nabla f_i(w)$

$\mathbb{E}[\nabla f_v(w)] = \nabla f(w)$

Examples of arbitrary sampling: non-uniform mini-batching

Random set $S \subset \{1, \ldots, n\}$, $\mathbb{E}|S| = b$

$\text{Prob}[i \in S] = p_i$, for $i = 1, \ldots, n$

$$v_i = \begin{cases} 
\frac{1}{p_i} & i \in S \\
0 & i \not\in S 
\end{cases}$$

$\mathbb{E}[v_i] = 1$

Arbitrary sampling SGD

Sample $v_t \sim \mathcal{D}$

$$w^{t+1} = w^t - \gamma_t \nabla f_{v^t}(w^t)$$

$$\nabla f_v(w) = \frac{n}{p_i} \sum_{i \in S} \nabla f_i(w)$$

$\mathbb{E}[\nabla f_v(w)] = \nabla f(w)$

SGD with arbitrary sampling

$$\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]$$

Sample $v^t \sim \mathcal{D}$

$$w^{t+1} = w^t - \gamma_t \nabla f_{v^t}(w^t)$$

Includes all forms of SGD (and GD)
SGD with arbitrary sampling

\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]

Sample \( v^t \sim \mathcal{D} \)

\[
w^{t+1} = w^t - \gamma_t \nabla f_{v^t}(w^t)
\]

Includes all forms of SGD (and GD)

It’s a SGD general, but how to analyse this?
Assumption and convergence of SGD
Assumptions and Convergence of Gradient Descent

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \|w^* - y\|^2 \]

\[ \|\nabla f(w) - \nabla f(w^*)\|^2 \leq 2L (f(w) - f(w^*)) \]
Assumptions and Convergence of Gradient Descent

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \| w^* - y \|_2^2 \]

\[ \| \nabla f(w) - \nabla f(w^*) \|_2^2 \leq 2L \left( f(w) - f(w^*) \right) \]

\[ w^{t+1} = w^t - \frac{1}{L} \nabla f(w^t), \quad v \equiv (1, \ldots, 1) \]

\[ w^* = \arg \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \]

Iteration complexity of gradient descent

Given \( \epsilon > 0 \) and \( t \geq \frac{L}{\mu} \log \left( \frac{1}{\epsilon} \right) \),

\[ \frac{\| w^t - w^* \|}{\| w^0 - w^* \|} \leq \epsilon \]
Assumptions and Convergence of SGD

$$f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \|w^* - y\|^2_2$$

Bigger smoothness constant/ stronger assumption

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f_i(w^*)\|^2_2 \leq 2L_{\text{max}} (f(w) - f(w^*))$$
Assumptions and Convergence of SGD

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \| w^* - y \|_2^2 \]

Bigger smoothness constant/ stronger assumption

\[ \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w) - \nabla f_i(w^*) \|_2^2 \leq 2L_{\text{max}} (f(w) - f(w^*)) \]

Definition \[ \sigma_*^2 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|_2^2 \]
Assumptions and Convergence of SGD

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \| w^* - y \|^2 \]

Bigger smoothness constant/stronger assumption

\[ \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w) - \nabla f_i(w^*) \|^2 \leq 2L_{\text{max}} (f(w) - f(w^*)) \]

Definition

\[ \sigma_*^2 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 \]

Iteration complexity of SGD

\[ t \geq \left( \frac{L_{\text{max}}}{\mu} + \frac{\sigma_*^2}{\epsilon \mu^2} \right) \log \left( \frac{1}{\epsilon} \right) \]

\[ \mathbb{E}[\| w^t - w^* \|] \leq \epsilon \]

Informal comparison between GD and SGD iteration complexity
Informal comparison between GD and SGD iteration complexity

**GD**

\[ t \geq O \left( \frac{L}{\mu} \right) \]

**SGD**

\[ t \geq O \left( \frac{L_{\text{max}}}{\mu} + \frac{\sigma^2_*}{\epsilon \mu^2} \right) \]

\[
\frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon
\]
Informal comparison between GD and SGD iteration complexity

GD

\[ t \geq O \left( \frac{L}{\mu} \right) \]

\[
\frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon
\]

How do they compare?

In general: \( L \leq L_{\text{max}} \leq nL \)
Informal comparison between GD and SGD iteration complexity

**GD**

\[ t \geq O \left( \frac{L}{\mu} \right) \]

**SGD**

\[ t \geq O \left( \frac{L_{\text{max}}}{\mu} + \frac{\sigma^2}{\epsilon \mu^2} \right) \]

\[ \frac{E[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon \]

When \( n \) is big \( L \ll L_{\text{max}} \)

In general: \( L \leq L_{\text{max}} \leq nL \)

How do they compare?
Informal comparison between GD and SGD iteration complexity

GD
\[ t \geq O \left( \frac{L}{\mu} \right) \]

SGD
\[ t \geq O \left( \frac{L_{\text{max}}}{\mu} + \frac{\sigma^2}{\epsilon \mu^2} \right) \]

\[ \frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon \]

How do they compare?

In general: \( L \leq L_{\text{max}} \leq nL \)

Need new “interpolating” notion of smoothness

When \( n \) is big \( L \ll L_{\text{max}} \)

\( L \leq ? L(D) ? \leq L_{\text{max}} \)
Key constant: Expected smoothness

Ass: Expected Smoothness. We write $(f, \mathcal{D}) \sim ES(\mathcal{L})$ when

$$\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2] \leq 2\mathcal{L} (f(w) - f(w*))$$
Ass: Expected Smoothness. We write \((f, \mathcal{D}) \sim ES(\mathcal{L})\) when

\[
\mathbb{E} [\|\nabla f_{v}(w) - \nabla f_{v}(w^*)\|^2_2] \leq 2\mathcal{L} (f(w) - f(w^*))
\]

\[
\nabla f_{v}(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]
Ass: Expected Smoothness. We write \((f, \mathcal{D}) \sim ES(\mathcal{L})\) when

\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2_2] \leq 2\mathcal{L} (f(w) - f(w^*))
\]

\[
\nabla f_v(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]

Expected smoothness constant

Depends on \(v\) and \(f\)

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)
Key constant: Expected smoothness

**Ass: Expected Smoothness.** We write \( (f, \mathcal{D}) \sim ES(\mathcal{L}) \) when
\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2_2] \leq 2\mathcal{L} (f(w) - f(w^*))
\]

\[
\nabla f_v(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]

**Expected smoothness constant**
- Depends on \( v \) and \( f \)

**Lemma:**
- \( f_i \) convex and \( L_{\text{max}} \)-smooth
- \( (f, \mathcal{D}) \sim ES(\mathcal{L}) \)
- \( \mathcal{L} \leq L_{\text{max}} \lambda_{\text{max}} (\mathbb{E}[vv^T]) \)

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)
**Key constant: Expected smoothness**

**Ass: Expected Smoothness.** We write \((f, \mathcal{D}) \sim ES(\mathcal{L})\) when

\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2_2] \leq 2\mathcal{L} \left(f(w) - f(w^*)\right)
\]

\[
\nabla f_v(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]

**Lemma:**

\(f_i\) convex and \(L_{\text{max}}\)-smooth

\((f, \mathcal{D}) \sim ES(\mathcal{L})\)

\[
\mathcal{L} \leq L_{\text{max}} \lambda_{\text{max}} \left(\mathbb{E}[vv^\top]\right)
\]

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)
Key constant: Expected smoothness

Ass: Expected Smoothness. We write \((f, D) \sim ES(\mathcal{L})\) when
\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2] \leq 2\mathcal{L} (f(w) - f(w^*))
\]

\[
\nabla f_v(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]

Lemma:

\(f_i\) convex and \(L_{\text{max}}\)-smooth
\[(f, D) \sim ES(\mathcal{L})\]
\[
\mathcal{L} \leq L_{\text{max}} \lambda_{\text{max}} (\mathbb{E}[vv^\top])
\]

Expected smoothness constant
Depends on \(v\) and \(f\)

Definition: Gradient noise
\[
\sigma^2 := \mathbb{E}_{v \sim \mathcal{D}} [\|\nabla f_v(w^*)\|^2]
\]

Rough estimate (we can do better)

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)
Key constant: Expected smoothness

**Ass: Expected Smoothness.** We write \((f, \mathcal{D}) \sim ES(\mathcal{L})\) when

\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2_2] \leq 2\mathcal{L} (f(w) - f(w^*))
\]

\[
\nabla f_v(w) = \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(w)
\]

**Lemma:**

- \(f_i\) convex and \(L_{\text{max}}\)-smooth

- \((f, \mathcal{D}) \sim ES(\mathcal{L})\)

- \(\mathcal{L} \leq L_{\text{max}} \lambda_{\text{max}} (\mathbb{E}[v v^\top])\)

**Definition:** Gradient noise

\[
\sigma^2 := \mathbb{E}_{v \sim \mathcal{D}}[\|\nabla f_v(w^*)\|^2]
\]

**Expected smoothness constant**

Depends on \(v\) and \(f\)

Rough estimate (we can do better)

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)
Example of Expected Smoothness

$S$ is chosen uniformly at random from all subsets of size $b$

$$L(b) = \frac{n(b - 1)}{b(n - 1)}L + \frac{n - b}{b(n - 1)}L_{\max}$$

$$v_i = \begin{cases} \frac{n}{b} & i \in S \\ 0 & i \notin S \end{cases}$$
Example of Expected Smoothness

$S$ is chosen uniformly at random from all subsets of size $b$

$$\mathcal{L}(b) = \frac{n(b-1)}{b(n-1)} L + \frac{n-b}{b(n-1)} L_{\text{max}}$$

$$v_i = \begin{cases} \frac{n}{b} & i \in S \\ 0 & i \notin S \end{cases}$$

$$\mathcal{L}(1) = L_{\text{max}} \quad \text{and} \quad \mathcal{L}(n) = L_{\text{max}}$$
Example of Expected Smoothness

\[ S \text{ is chosen uniformly at random from all subsets of size } b \]

\[ \mathcal{L}(b) = \frac{n(b - 1)}{b(n - 1)}L + \frac{n - b}{b(n - 1)}L_{\max} \]

\[ v_i = \begin{cases} \frac{n}{b} & i \in S \\ 0 & i \notin S \end{cases} \]

\[ \mathcal{L}(1) = L_{\max} \text{ and } \mathcal{L}(n) = L_{\max} \]

What about \( \sigma^2 \)?

\[ \sigma^2 := \mathbb{E}[\|\nabla f_v(w^*)\|^2] \]
Example of Expected Smoothness

$S$ is chosen uniformly at random from all subsets of size $b$

$$v_i = \begin{cases} \frac{n}{b} & i \in S \\ 0 & i \notin S \end{cases}$$

$$\mathcal{L}(b) = \frac{n(b-1)}{b(n-1)} L + \frac{n-b}{b(n-1)} L_{\max}$$

$$\sigma^2(b) = \frac{n-b}{b(n-1)} \sigma^*_2$$

$\sigma^2$ measures how much model fits data

$$\sigma^2 :\ = \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2$$

$L_{\max} = \sigma^2 = 0$

What about $\sigma^2$?

$$\sigma^2 := \mathbb{E}[\| \nabla f_v(w^*) \|^2]$$

$\sigma^2 = \sigma^*_2$

$L = \mathcal{L}(1) = L_{\max}$ and $\mathcal{L}(n) = L_{\max}$
Expected smoothness gives awesome bound on gradient

Lemma \((f, \mathcal{D}) \sim ES(L)\)

\[\mathbb{E}[\|\nabla f_v(\omega)\|^2] \leq 4L(f(\omega) - f(\omega^*)) + 2\sigma^2\]
Expected smoothness gives awesome bound on gradient

Lemma \( (f, D) \sim ES(\mathcal{L}) \)

\[
\mathbb{E}[\|\nabla f_v(w)\|^2] \leq 4\mathcal{L}(f(w) - f(w^*)) + 2\sigma^2
\]

Normally bound on gradient is an **assumption**

**Assumption** There exists \( B > 0 \)

\[
\mathbb{E}[\|\nabla f_v(w^t)\|^2] \leq B^2
\]

References:
- Hazan & Kale, JMLR 2014.
- Rakhlin, Shamir, & Sridharan, ICML 2012
- Shamir & Zhang, ICML 2013.
Expected smoothness gives awesome bound on gradient

Lemma \((f, \mathcal{D}) \sim ES(\mathcal{L})\)

\[ \mathbb{E}[\|\nabla f_v(w)\|^2] \leq 4\mathcal{L}(f(w) - f(w^*)) + 2\sigma^2 \]

\(\sigma^2 := \mathbb{E}[\|\nabla f_v(w^*)\|^2]\)

Normally bound on gradient is an \textit{assumption}

Assumption There exists \(B > 0\)

\[ \mathbb{E}[\|\nabla f_v(w^t)\|^2] \leq B^2 \]

References:
- Hazan & Kale, JMLR 2014.
- Rakhlin, Shamir, & Sridharan, ICML 2012
- Shamir & Zhang, ICML 2013.
Expected smoothness gives awesome bound on gradient

\[ \mathbb{E}[\|\nabla f_v(w)\|^2] \leq 4\mathcal{L}(f(w) - f(w^*)) + 2\sigma^2 \]

Informative: with realistic assumptions

Lemma \((f,\mathcal{D}) \sim ES(\mathcal{L})\)

\[ \sigma^2 := \mathbb{E}[\|\nabla f_v(w^*)\|^2] \]

Normally bound on gradient is an \textit{assumption}

Assumption There exists \(B > 0\)

\[ \mathbb{E}[\|\nabla f_v(w^t)\|^2] \leq B^2 \]

References:
- Hazan & Kale, JMLR 2014.
- Rakhlin, Shamir, & Sridharan, ICML 2012
- Shamir & Zhang, ICML 2013.
Main Theorem (Linear convergence to a neighborhood)

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \| w^* - y \|^2 \]

**Theorem** \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[ \mathbb{E} [\| w^t - w^* \|^2] \leq (1 - \gamma \mu)^t \| w^0 - w^* \|^2 + \frac{2\gamma \sigma^2}{\mu} \]
**Main Theorem** (Linear convergence to a neighborhood)

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} ||w^* - y||^2 \]

**Theorem**  
\( (f, \mathcal{D}) \sim ES(\mathcal{L}) \) and \( \mu \)-quasi strongly convex

\[
\mathbb{E}[||w^t - w^*||^2] \leq (1 - \gamma \mu)^t ||w^0 - w^*||^2 + \frac{2\gamma \sigma^2}{\mu}
\]

Fixed stepsise \( \gamma_t \equiv \gamma \leq \frac{1}{2\mathcal{L}} \)
**Main Theorem** (Linear convergence to a neighborhood)

\[ f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \| w^* - y \|_2^2 \]

\[ \sigma^2 := \mathbb{E}[\| \nabla f_0(w^*) \|^2] \]

**Theorem**  \( (f, \mathcal{D}) \sim ES(\mathcal{L}) \) and \( \mu \)-quasi strongly convex

\[ \mathbb{E}[\| w^t - w^* \|^2] \leq (1 - \gamma \mu)^t \| w^0 - w^* \|^2 + \frac{2\gamma\sigma^2}{\mu} \]

Fixed stepsize \( \gamma_t \equiv \gamma \leq \frac{1}{2\mathcal{L}} \)

**Corollary**

\[ \gamma = \frac{1}{2} \max \left\{ \frac{1}{\mathcal{L}}, \frac{\epsilon\mu}{2\sigma^2} \right\} \]

\[ t \geq \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \quad \Rightarrow \quad \frac{\mathbb{E}[\| w^t - w^* \|]}{\| w^0 - w^* \|} \leq \epsilon \]
**Main Theorem** (Linear convergence to a neighborhood)

\[
f(w^*) \geq f(y) + \langle \nabla f(y), w^* - y \rangle + \frac{\mu}{2} \|w^* - y\|^2
\]

\[
\sigma^2 := \mathbb{E}[\|\nabla f_0(w^*)\|^2]
\]

**Theorem** \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[
\mathbb{E}[\|w^t - w^*\|^2] \leq (1 - \gamma \mu)^t \|w^0 - w^*\|^2 + \frac{2\gamma \sigma^2}{\mu}
\]

Fixed stepsize \(\gamma_t \equiv \gamma \leq \frac{1}{2\mathcal{L}}\)

**Corollary** \(\gamma = \frac{1}{2} \max \left\{ \frac{1}{\mathcal{L}}, \frac{\epsilon \mu}{2\sigma^2} \right\}\)

\[
t \geq \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right)
\]

\[
\frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon
\]

**saves time** for theorists: Includes GD and SGD as special cases. Also tighter!
Proof is SUPER EASY:

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \gamma \nabla f_v(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 - 2\gamma \langle \nabla f_v(w^t), w^t - w^* \rangle + \gamma^2 ||\nabla f_v(w^t)||_2^2.$$ 

Taking expectation with respect to $v \sim \mathcal{D}$

$$\mathbb{E}_v [||w^{t+1} - w^*||_2^2] = ||w^t - w^*||_2^2 - 2\gamma \langle \nabla f(w^t), w^t - w^* \rangle + \gamma^2 \mathbb{E}_v [||\nabla f_v(w^t)||_2^2]$$

$$\leq (1 - \gamma \mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_v [||\nabla f_v(w^t)||_2^2]$$

$$\leq (1 - \gamma \mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma \mathcal{L} - 1)(f(w) - f(w^*)) + 2\gamma^2 \sigma^2$$

$$\leq (1 - \gamma \mu)||w^t - w^*||_2^2 + 2\gamma^2 \sigma^2$$

Taking total expectation

$$\mathbb{E} [||w^{t+1} - w^*||_2^2] \leq (1 - \gamma \mu)\mathbb{E} [||w^t - w^*||_2^2] + 2\gamma^2 \sigma^2$$

$$= (1 - \gamma \mu)^{t+1}||w^0 - w^*||_2^2 + 2 \sum_{i=0}^{t} (1 - \gamma \mu)^i \gamma^2 \sigma^2$$

$$\leq (1 - \gamma \mu)^{t+1}||w^0 - w^*||_2^2 + \frac{2\gamma\sigma^2}{\mu}$$

$$\sum_{i=0}^{t} (1 - \gamma \mu)^i = \frac{1 - (1 - \gamma \mu)^{t+1}}{\gamma \mu} \leq \frac{1}{\gamma \mu}$$
Stochastic Gradient Descent
\( \gamma = 0.01 \)
Stochastic Gradient Descent

$\gamma = 0.2$
Stochastic Gradient Descent

\( \gamma = 0.5 \)
Total complexity for mini-batch SGD

**Corollary**

$$\gamma = \max \left\{ \frac{1}{L'}, \frac{\epsilon \mu}{4\sigma^2} \right\}$$

$$t \geq \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \quad \Rightarrow \quad \mathbb{E}[\|w^t - w^*\|] \leq \epsilon$$
Total complexity for mini-batch SGD

\[ C(b) := \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \times b \]
Total complexity for mini-batch SGD

\[ C(b) := \max \left\{ \frac{2\mathcal{L}}{\mu} , \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \times b \]

Corollary:
\[ t \geq \max \left\{ \frac{2\mathcal{L}}{\mu} , \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \Rightarrow \mathbb{E}[||w_t - w^*||] \leq \epsilon \]
Total complexity for mini-batch SGD

\[ C(b) := \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \times b \]

**Corollary**

\[ t \geq \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \Rightarrow \frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon \]  

#stochastic gradient evaluation in 1 iteration
Total complexity for mini-batch SGD

\[ C(b) := \max \left\{ \frac{2 \mathcal{L}}{\mu}, \frac{4 \sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \times b \]

Corollary

\[ t \geq \max \left\{ \frac{2 \mathcal{L}}{\mu}, \frac{4 \sigma^2}{\epsilon \mu^2} \right\} \gamma \log \left( \frac{2}{\epsilon} \right) \Rightarrow \frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon \]

\[ \mathcal{L} = \frac{n(b - 1)}{b(n - 1)} L + \frac{n - b}{b(n - 1)} L_{\max} \]

\[ \sigma^2 = \frac{n - b}{b(n - 1)} \sigma_*^2 \]

Total complexity is a simple function of mini-batch size \( b \)
Optimal mini-batch size

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\text{max}}, \frac{2(n-b)\sigma_*^2}{\epsilon \mu} \right\} \]

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(w^*)\|^2 \times \log \left( \frac{2}{\epsilon} \right) \]
Optimal mini-batch size

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\max}, \frac{2(n-b)\sigma_*^2}{\epsilon \mu} \right\} \]

Linearly increasing

\[ \sigma_* := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 \times \log \left( \frac{2}{\epsilon} \right) \]
Optimal mini-batch size

$$C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b - 1)L + (n - b)L_{\max}, \frac{2(n - b)\sigma^2_*}{\epsilon \mu} \right\}$$

Linearly increasing

Linearly decreasing

$$\sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 \times \log \left( \frac{2}{\epsilon} \right)$$
Optimal mini-batch size

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\text{max}}, \frac{2(n-b)^2\sigma_\ast^2}{\epsilon \mu} \right\} \]

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^\ast) \|^2 \times \log \left( \frac{2/\epsilon}{\epsilon} \right) \]

Linearly increasing

Linearly decreasing
Optimal mini-batch size

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ \frac{n(b-1)L + (n-b)L_{\text{max}}}{\gamma(b)}, \frac{2(n-b)\sigma^2_*}{\epsilon \mu \cdot \log \left( \frac{2}{\epsilon} \right)} \right\} \]

Linearity:
- Linearly increasing
- Linearly decreasing

Steps:
- \( b^* = n \left[ \frac{L - L_{\text{max}} + \frac{2}{\epsilon \mu} \cdot \sigma^2_*}{nL - L_{\text{max}} + \frac{2}{\epsilon \mu} \cdot \sigma^2_*} \right] \)

Diagram:
-轴：最小批大小
- 若干直线和曲线展示了不同的情况下最优批大小的变化。
Optimal mini-batch size for models that interpolate data

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 = 0 \times \log \left( \frac{2}{\epsilon} \right) \]

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\text{max}}, \frac{2(n-b)\sigma_*^2}{\epsilon \mu} \right\} \]
Optimal mini-batch size for models that interpolate data

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 = 0 \]

\[ C(b) := \frac{2}{\mu(n-1) \max \left\{ n(b-1)L + (n-b)L_{\max}, \frac{2(n-b)\sigma^2}{\varepsilon \mu} \right\} \times \log \left( \frac{2}{\varepsilon} \right)} \]
Optimal mini-batch size for models that interpolate data

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 = 0 \times \log \left( \frac{2}{\epsilon} \right) \]

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\text{max}}, \frac{2(n-b)\sigma_*^2}{\epsilon \mu} \right\} \]

\[ = \frac{2}{\mu(n-1)} (n(b-1)L + (n-b)L_{\text{max}}) \]
Optimal mini-batch size for models that interpolate data

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 = 0 \]

\[
C(b) := \frac{2}{\mu(n-1)} \max \left\{ \frac{n(b-1)L + (n-b)L_{\text{max}}}{\frac{2(n-b)\sigma_*^2}{\epsilon \mu} \times \log \left( \frac{2}{\epsilon} \right)} \right\} \\
= \frac{2}{\mu(n-1)} \left( n(b-1)L + (n-b)L_{\text{max}} \right)
\]

\[
\gamma(b) := \frac{n-1}{2} \frac{b}{n(b-1)L + (n-b)L_{\text{max}}}
\]
Optimal mini-batch size for models that interpolate data

$$\sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \left\| \nabla f_i(w^*) \right\|^2 = 0 \times \log \left( \frac{2}{\epsilon} \right)$$

$$C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{max}, \frac{2(n-b)\sigma^2_*}{\epsilon \mu} \right\}$$

$$= \frac{2}{\mu(n-1)} \left( n(b-1)L + (n-b)L_{max} \right)$$

Linearly increasing

$$\gamma(b) := \frac{n-1}{2} \frac{b}{n(b-1)L + (n-b)L_{max}}$$

increases with $b$

$$b^* = 1$$
Optimal mini-batch size for models that interpolate data

\[ \sigma_1 := \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(w^*) \|^2 = 0 \]

\[ C(b) := \frac{2}{\mu(n-1)} \max \left\{ n(b-1)L + (n-b)L_{\text{max}}, \frac{2(n-b)\sigma_*^2}{\epsilon \mu} \times \log \left( \frac{2}{\epsilon} \right) \right\} \]

\[ = \frac{2}{\mu(n-1)} \left( n(b-1)L + (n-b)L_{\text{max}} \right) \]

\[ \gamma(b) := \frac{n-1}{2} \frac{b}{n(b-1)L + (n-b)L_{\text{max}}} \]

All gains in mini-batching are due to multi-threading and cache memory?

\[ b^* = 1 \]
Stochastic Gradient Descent

$\gamma = 0.2$
Learning schedule: Constant & decreasing step sizes

\[ \gamma_t = \begin{cases} 
\frac{1}{2\mathcal{L}} & \text{for } t \leq 4\left[\frac{\mathcal{L}}{\mu}\right] \\
\frac{2t + 1}{(t + 1)^2 \mu} & \text{for } t > 4\left[\frac{\mathcal{L}}{\mu}\right]
\end{cases} \]
Learning schedule: Constant & decreasing step sizes

Theorem \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[
\gamma_t = \begin{cases} 
\frac{1}{2\mathcal{L}} & \text{for } t \leq 4\lfloor \mathcal{L}/\mu \rfloor \\
\frac{2t + 1}{(t + 1)^2\mu} & \text{for } t > 4\lfloor \mathcal{L}/\mu \rfloor
\end{cases}
\]
Learning schedule: Constant & decreasing step sizes

**Theorem** \((f, D) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[
\gamma_t = \begin{cases} 
\frac{1}{2\mathcal{L}} & \text{for } t \leq 4\left[\frac{\mathcal{L}}{\mu}\right] \\
\frac{2t + 1}{(t + 1)^2 \mu} & \text{for } t > 4\left[\frac{\mathcal{L}}{\mu}\right]
\end{cases}
\]

Learning rate with switch point

A stochastic condition number
Learning schedule: Constant & decreasing step sizes

**Theorem** \((f, D) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[
\gamma_t = \begin{cases} 
\frac{1}{2\mathcal{L}} & \text{for } t \leq 4\lceil \mathcal{L}/\mu \rceil \\
\frac{2t + 1}{(t + 1)^2\mu} & \text{for } t > 4\lceil \mathcal{L}/\mu \rceil
\end{cases}
\]

Learning rate with switch point

\[
\sigma^2 := \mathbb{E}[\|\nabla f_v(w^*)\|^2]
\]

\[
\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2 8}{\mu^2 t} + \frac{16\lceil \mathcal{L}/\mu \rceil^2}{e^2t^2} \|w^0 - w^*\|^2
\]

for \(t > 4\lceil \mathcal{L}/\mu \rceil\)

A stochastic condition number
Stochastic Gradient Descent with switch to decreasing stepsizes

Switch point $t = 4[K]$
Stochastic variance reduced methods
Simple Stochastic Reformulation

Random sampling vector \( v = (v_1, \ldots, v_n) \in \mathbb{R}^n \) with
\[
\mathbb{E}[v_i] = 1, \quad \text{for } i = 1, \ldots, n
\]

\[
f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_i] f_i(w) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]

What to do about the variance?

Original finite sum problem
\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Stochastic Reformulation
\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) \right]
\]
Minimizing the expectation of random linear combinations of original function
Controlled Stochastic Reformulation

\[ \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_v(w)] = \mathbb{E}[f_v(w)] - \mathbb{E}[z_v(w)] + \mathbb{E}[z_v(w)] \]
Controlled Stochastic Reformulation

\[
\frac{1}{n} \sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_v(w)] = \mathbb{E}[f_v(w)] - \mathbb{E}[z_v(w)] + \mathbb{E}[z_v(w)]
\]
Controlled Stochastic Reformulation

\[
\frac{1}{n} \sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_v(w)] = \mathbb{E}[f_v(w)] - \mathbb{E}[z_v(w)] + \mathbb{E}[z_v(w)]
\]

\[
= \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]]
\]
Controlled Stochastic Reformulation

\[
\frac{1}{n} \sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_v(w)] = \mathbb{E}[f_v(w)] - \mathbb{E}[z_v(w)] + \mathbb{E}[z_v(w)] = \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]]
\]

Original finite sum problem

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Controlled Stochastic Reformulation

\[
\min_{w \in \mathbb{R}^d} \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]]
\]

Use covariates to control the variance

covariate \( z_v(w) \in \mathbb{R} \)

Cancel out
Variance reduction with arbitrary sampling

$$\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) - z_v(w) + \mathbb{E}[z_v(w)] \right]$$
Variance reduction with arbitrary sampling

\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) - z_v(w) + \mathbb{E}[z_v(w)] \right]
\]

Sample \( v^t \sim D \)

\[
w^{t+1} = w^t - \gamma_t g_{v^t}(w^t)
\]
Variance reduction with arbitrary sampling

\[
\min_{w \in \mathbb{R}^d} \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]]
\]

Sample \( v^t \sim D \)

\[
w^{t+1} = w^t - \gamma_t g_{v^t}(w^t)
\]

\[
g_v(w) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)]
\]
Variance reduction with arbitrary sampling

\[ \min_{w \in \mathbb{R}^d} \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]] \]

By design we have that
\[ \mathbb{E}[g_{v^t}(w^t)] = \nabla f(w^t) \]

Sample \( v^t \sim D \)

\[ w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) \]

\[ g_v(w) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)] \]
Variance reduction with arbitrary sampling

By design we have that

$$\min_{w \in \mathbb{R}^d} \mathbb{E}[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]]$$

Sample $v^t \sim D$

$$w^{t+1} = w^t - \gamma_t g_{v^t}(w^t)$$

How to choose $z_v(w)$?

$$g_v(w) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)]$$

By design we have that

$$\mathbb{E}[g_{v^t}(w^t)] = \nabla f(w^t)$$
Choosing the covariate

Sample $v^t \sim D$

$$w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)]$$
Choosing the covariate

Sample $v^t \sim \mathcal{D}$

$$w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E} [\nabla z_v(w)]$$

We would like:

$$g_v(w) \approx \nabla f(w)$$
Choosing the covariate

Sample \( u^t \sim D \)

\[ w^{t+1} = w^t - \gamma_t g_{u^t}(w^t) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)] \]

We would like:

\[ g_v(w) \approx \nabla f(w) \quad \nabla z_v(w) \approx \nabla f_v(w) \]
Choosing the covariate

We would like:

Sample $v^t \sim \mathcal{D}$

\[ w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) := \nabla f_v(w) - \nabla z_v(w) + \mathbb{E}[\nabla z_v(w)] \]

We would like:

\[ g_v(w) \approx \nabla f(w) \quad \Rightarrow \quad \nabla z_v(w) \approx \nabla f_v(w) \]

Linear approximation

\[ z_v(w) = f_v(\tilde{w}) + \langle \nabla f_v(\tilde{w}), w - \tilde{w} \rangle \]

A reference point/ snap shot
SVRG: Stochastic Variance Reduced Gradients

\[ w^{t+1} = w^t - \gamma_t g_{\nu^t}(w^t) \]

- **Reference point**
  \[ \tilde{w} \in \mathbb{R}^d \]

- **Sample**
  \[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

- **Grad. estimate**
  \[ g_{\nu^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w}) \]

Johnson & Zhang, 2013 NIPS
SVRG: Stochastic Variance Reduced Gradients

Reference point:
\[ \tilde{w} \in \mathbb{R}^d \]

Sample:
\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate:
\[ g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w}) \]

Single element sampling:
\[ v_j = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases} \]
SVRG: Stochastic Variance Reduced Gradients

Reference point

\( \tilde{w} \in \mathbb{R}^d \)

Sample

\( \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \) uniformly

Grad. estimate

\[
g_{\nu^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})
\]

Single element sampling

\( \nu_j = \begin{cases} 
  n & j = i \\
  0 & j \neq i 
\end{cases} \)

\( \nabla z_{\nu^t}(w^t) = \nabla f_i(\tilde{w}) \)
SVRG: Stochastic Variance Reduced Gradients

Reference point

\[ \tilde{w} \in \mathbb{R}^d \]

Sample

\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate

\[ g_{vt}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w}) \]

Single element sampling

\[ v_j = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases} \]

\[ z_{vt}(w) = f_i(\tilde{w}) + \langle \nabla f_i(\tilde{w}), w - \tilde{w} \rangle \]

\[ \nabla z_{vt}(w^t) = \nabla f_i(\tilde{w}) \]
**SVRG: Stochastic Variance Reduced Gradients**

\[
\omega^{t+1} = \omega^t - \gamma_t g^{t}(\omega^t)
\]

Reference point

\[\tilde{\omega} \in \mathbb{R}^d\]

Sample

\[\nabla f_i(\omega^t), \quad i \in \{1, \ldots, n\} \text{ uniformly}\]

Grad. estimate

\[g^{t}(\omega^t) = \nabla f_i(\omega^t) - \nabla f_i(\tilde{\omega}) + \nabla f(\tilde{\omega})\]

Single element sampling

\[v_j = \begin{cases} 
  n & j = i \\
  0 & j \neq i 
\end{cases}\]

\[z^{t}(w) = f_i(\tilde{\omega}) + \langle \nabla f_i(\tilde{\omega}), w - \tilde{\omega} \rangle \]

\[\nabla z^{t}(w^t) = \nabla f_i(\tilde{\omega})\]

\[\mathbb{E}[\nabla z^{t}(w^t)] = \nabla f(\tilde{\omega})\]
SVRG: Stochastic Variance Reduced Gradients

Set $w^0 = 0$, choose $\gamma > 0$, $m \in \mathbb{N}$,
$\alpha_k > 0$ for $k = 0, \ldots, m - 1$
$\tilde{w}^0 = w^0$
for $t = 0, 1, 2, \ldots, T - 1$
calculate $\nabla f(\tilde{w}^t)$
for $k = 0, 1, 2, \ldots, m - 1$
sample $i \in \{1, \ldots, n\}$
$g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)$
$w^{k+1} = w^k - \gamma g^k$
$\tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k$
Output $\tilde{w}^T$
SVRG: Stochastic Variance Reduced Gradients

Set \( w^0 = 0 \), choose \( \gamma > 0 \), \( m \in \mathbb{N} \),
\[ \alpha_k > 0 \text{ for } k = 0, \ldots, m - 1 \]
\[ \tilde{w}^0 = w^0 \]
for \( t = 0, 1, 2, \ldots, T - 1 \)
calculate \( \nabla f(\tilde{w}^t) \)
for \( k = 0, 1, 2, \ldots, m - 1 \)
sample \( i \in \{1, \ldots, n\} \)
\[ g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t) \]
\[ w^{k+1} = w^k - \gamma g^k \]
\[ \tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k \]
Output \( \tilde{w}^T \)
SVRG: Stochastic Variance Reduced Gradients

Set \( w^0 = 0 \), choose \( \gamma > 0 \), \( m \in \mathbb{N} \),
\[ \alpha_k > 0 \text{ for } k = 0, \ldots, m - 1 \]
\[ \tilde{w}^0 = w^0 \]
for \( t = 0, 1, 2, \ldots, T - 1 \)
calculate \( \nabla f(\tilde{w}^t) \)
for \( k = 0, 1, 2, \ldots, m - 1 \)
sample \( i \in \{1, \ldots, n\} \)
\[ g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t) \]
\[ w^{k+1} = w^k - \gamma g^k \]
\[ \tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k \]
Output \( \tilde{w}^T \)

Freeze reference point for \( m \) iterations

Weighted average of inner iterates
SAGA: Stochastic Average Gradient

\[ w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) \]

Sample
\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate
\[ g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w^{t_j}) \]

Store grad.
\[ \nabla f_i(w^{t_i}) = \nabla f_i(w^t) \]

Single element sampling
\[ v_j = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases} \]
SAGA: Stochastic Average Gradient

\[ w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) \]

Sample

\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate

\[ g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w^{t_j}) \]

\[ \nabla z_{v^t}(w^t) = \nabla f_i(w^{t_i}) \]

Store grad.

\[ \nabla f_i(w^{t_i}) = \nabla f_i(w^t) \]
**SAGA: Stochastic Average Gradient**

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

Single element sampling

\[ v_j = \begin{cases} 
n & j = i \\
0 & j \neq i \end{cases} \]

Sample

\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate

\[ g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w^{t_j}) \]

Store grad.

\[ \nabla f_i(w^{t_i}) = \nabla f_i(w^t) \]

\[ z_{v^t}(w) = f_i(w^{t_i}) + \langle \nabla f_i(w^{t_i}), w - w^{t_i} \rangle \]

\[ \nabla z_{v^t}(w^t) = \nabla f_i(w^{t_i}) \]
SAGA: Stochastic Average Gradient

\[ w^{t+1} = w^t - \gamma_t g_{v^t}(w^t) \]

Sample

\[ \nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly} \]

Grad. estimate

\[ g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w^{t_j}) \]

\[ z_{v^t}(w) = f_i(w^{t_i}) + \langle \nabla f_i(w^{t_i}), w - w^{t_i} \rangle \]

Store grad.

\[ \nabla f_i(w^{t_i}) = \nabla f_i(w^t) \]

Single element sampling

\[ v_j = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases} \]

Expected gradient:

\[ \mathbb{E}[\nabla z_{v^t}(w^t)] \]

\[ \nabla z_{v^t}(w^t) = \nabla f_i(w^{t_i}) \]
SAGA: Stochastic Average Gradient

Set $w^0 = 0$, $g_i = \nabla f_i(w^0)$, for $i = 1 \ldots, n$
Choose $\gamma > 0$
for $t = 0, 1, 2, \ldots, T - 1$
    sample $i \in \{1, \ldots, n\}$
    $g^t = \nabla f_i(w^t) - g_i + \frac{1}{n} \sum_{j=1}^{n} g_j$
    $w^{t+1} = w^t - \gamma g^t$
    $g_i = \nabla f_i(w^t)$
Output $w^T$

No inner loop, rolling update
Stores a $d \times n$ matrix
Complexity of Variance Reduced
Iteration complexity for SVRG and SAGA for arbitrary sampling

Theorem for SVRG \((f, D) \sim ES(L)\) and \(\mu\)-strongly convex

\[
\text{stepsize } \gamma \leq \frac{1}{6L} \quad \implies \quad \text{Iteration complexity } \approx O\left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon}\right)\right)
\]
Iteration complexity for SVRG and SAGA for arbitrary sampling

**Theorem for SVRG** \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-strongly convex

Stepsize \(\gamma \leq \frac{1}{6\mathcal{L}}\)  

Iteration complexity \(\approx O\left(\frac{\mathcal{L}}{\mu} \log \left(\frac{1}{\epsilon}\right)\right)\)

Sebbouh, Gazagnadou, Jelassi, Bach, G., 2019

**Theorem for SAGA** (and the JacSketch family of methods) \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

Stepsize \(\gamma \leq \frac{1}{4\mathcal{L}}\)  

Iteration complexity \(\approx O\left(\frac{\mathcal{L}}{\mu} \log \left(\frac{1}{\epsilon}\right)\right)\)

G., Bach, Richtarik, 2018
Theorem for SVRG \((f, D) \sim E\!S(L)\) and \(\mu\)–strongly convex

\[
\text{stepsizes } \gamma \leq \frac{1}{6L} \quad \text{Iteration complexity} \quad \approx O\left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon}\right)\right)
\]

Sebbouh, Gazagnadou, Jelassi, Bach, G., 2019

Theorem for SAGA (and the JacSketch family of methods) \((f, D) \sim E\!S(L)\) and \(\mu\)–quasi strongly convex

\[
\text{stepsizes } \gamma \leq \frac{1}{4L} \quad \text{Iteration complexity} \quad \approx O\left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon}\right)\right)
\]

G., Bach, Richtarik, 2018

Missing details due to extra definitions
Total Complexity of mini-batch SVRG

\[ C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3n - b}{b(n - 1)} \frac{L_{\text{max}}}{\mu} + \frac{3n}{b} \frac{b - 1}{n - 1} \frac{L}{\mu}, m \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

\[ \gamma = \frac{1}{6} \frac{b(n - 1)}{(n - b)L_{\text{max}} + n(b - 1)L} \]
Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

\[
C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3}{b} \frac{n - b}{n - 1} \frac{L_{\text{max}}}{\mu} + \frac{3n}{b} \frac{b - 1}{n - 1} \frac{L}{\mu}, m \right\} \times \log \left( \frac{2}{\epsilon} \right)
\]

Non-linearly increasing

\[
\gamma = \frac{1}{6} \frac{b(n - 1)}{(n - b)L_{\text{max}} + n(b - 1)L}
\]
Total Complexity of mini-batch SVRG

\[ C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3}{b} \frac{n - b}{n - 1} \frac{L_{\text{max}}}{\mu} + \frac{3n}{b} \frac{b - 1}{n - 1} \frac{L}{\mu}, m \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

\[ \gamma = \frac{1}{6} \frac{b(n - 1)}{(n - b)L_{\text{max}} + n(b - 1)L} \]
Total Complexity of mini-batch SVRG

\[ C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3 n - b}{b n - 1} \frac{L_{\text{max}}}{\mu} + \frac{3n}{b} \frac{b - 1}{n - 1} \frac{L}{\mu}, m \right\} \times \log \left( \frac{2}{\varepsilon} \right) \]

- Linearly decreasing
- Non-linearly increasing

\[ \gamma = \frac{1}{6} \frac{b(n - 1)}{(n - b)L_{\text{max}} + n(b - 1)L} \]

Graph showing the total complexity as a function of mini-batch size.
Total Complexity of mini-batch SVRG

\[ C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3n - b}{b(n - 1)} \frac{L_{\max}}{\mu} + \frac{3n(b - 1)}{b(n - 1)} \frac{L}{\mu}, m \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

Non-linearly increasing

\[ \gamma = \frac{1}{6} \frac{b(n - 1)}{(n - b)L_{\max} + n(b - 1)L} \]
Total Complexity of mini-batch SVRG

\[
C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3n-b}{b(n-1)} \frac{L_{\text{max}}}{\mu} + \frac{3n}{b(n-1)} \frac{b-1}{\mu} \right\} \times \log \left( \frac{2}{\epsilon} \right)
\]

Non-linearly increasing

\[
\gamma = \frac{1}{6} \frac{b(n-1)}{(n-b)L_{\text{max}} + n(b-1)L}
\]

Stepsize increasing with \(b\)
Total Complexity of mini-batch SAGA

\[ C(b) = \max \left\{ n \frac{b - 1}{n - 1} \frac{4L}{\mu} + \frac{n - b}{n - 1} \frac{4L_{\max}}{\mu}, \quad n + \frac{n - b}{n - 1} \frac{4L_{\max}}{\mu} \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

\[ \gamma = \frac{1}{4} \frac{b(n - 1)}{\max \left\{ n(b - 1)L + (n - b)L_{\max}, (n - b)L_{\max} + \frac{n(n - 1)\mu}{4} \right\}} \]
Total Complexity of mini-batch SAGA

\[ C(b) = \max \left\{ \frac{b - 1}{n - 1} \frac{4L}{\mu} + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu}, \quad n + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu} \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

Linearly increasing

\[ \gamma = \frac{1}{4} \frac{b(n - 1)}{\max \left\{ n(b - 1)L + (n - b)L_{\text{max}}, (n - b)L_{\text{max}} + \frac{n(n - 1)\mu}{4} \right\}} \]
Total Complexity of mini-batch SAGA

\[ C(b) = \max \left\{ n \frac{b - 1}{n - 1} \frac{4L}{\mu} + n - b \frac{4L_{\text{max}}}{\mu}, \quad n + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu} \times \log \left( \frac{2}{\epsilon} \right) \right\} \]

**Linearly increasing**

\[ \gamma = \frac{1}{4} \frac{b(n - 1)}{\max \left\{ n(b - 1)L + (n - b)L_{\text{max}}, (n - b)L_{\text{max}} + \frac{n(n - 1)\mu}{4} \right\}} \]
Total Complexity of mini-batch SAGA

\[ C(b) = \max \left\{ n \frac{b - 1}{n - 1} \frac{4L}{\mu} + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu}, n + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu} \right\} \times \log \left( \frac{2}{\epsilon} \right) \]

Linearly increasing  \hspace{2cm} \text{Linearly decreasing}

\[ \gamma = \frac{1}{4} \max \left\{ n(b - 1)L + (n - b)L_{\text{max}}, (n - b)L_{\text{max}} + \frac{n(n - 1)\mu}{4} \right\} \]

\[ b^* = \left[ 1 + \frac{\mu(n - 1)}{4L} \right] \]
Total Complexity of mini-batch SAGA

$$C(b) = \max \left\{ \frac{b - 1}{n - 1} \frac{4L}{\mu} + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu}, \frac{n + \frac{n - b}{n - 1} \frac{4L_{\text{max}}}{\mu}}{\times \log \left(\frac{2}{\epsilon}\right)} \right\}$$

Linearly increasing

Linearly decreasing

$$C(b) = \max \left\{ \frac{b(n - 1)}{4} \frac{L}{\mu} \right\}$$

$$\gamma = \frac{1}{4} \frac{b(n - 1)}{\max \left\{ n(b - 1)L + (n - b)L_{\text{max}}, (n - b)L_{\text{max}} + \frac{n(n - 1)\mu}{4} \right\}}$$

$$b^* = \left[ 1 + \frac{\mu(n - 1)}{4L} \right]$$

Always smaller than 25% of data
Total Complexity of mini-batch SAGA

\[ b^{\text{empirical}} = 16384 \]

\[ b^* = 1 + \frac{\mu(n - 1)}{4L} \]

![Graph showing empirical total complexity vs. mini-batch size](image-url)
Total Complexity of mini-batch

SAGA

So accurate, close to empirical best mini-batch size

\[ b_{\text{empirical}} = 16384 \]

\[ b^* = 1 + \frac{\mu(n - 1)}{4L} \]
Take home message

Stochastic reformulations allow to view all variants as simple SGD

To analyse all forms of sampling used through expected smooth

How to calculate optimal mini-batch size of SGD, SAGA and SVRG

Stepsize increase by orders when mini-batch size increases

\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]

\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2] \leq \mathcal{L} (f(w) - f(w^*))
\]

\[
(f, \mathcal{D}) \sim ES(\mathcal{L})
\]
Take home message

Stochastic reformulations allow to view all variants as simple SGD

To analyse all forms of sampling used through expected smooth

How to calculate optimal mini-batch size of SGD, SAGA and SVRG

Stepsize increase by orders when mini-batch size increases

\[
\min_{w \in \mathbb{R}^d} \mathbb{E} \left[ f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]
\]

\[
\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2_{2}] \leq \mathcal{L} (f(w) - f(w^*))
\]

\[
(f, \mathcal{D}) \sim ES(\mathcal{L})
\]

![Graph showing stepsize increase with mini-batch size](image)
RMG, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin and Peter Richtárik (2019), ICML
SGD: general analysis and improved rates

RMG, P. Richtarik, F. Bach (2018), preprint online
Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching

Optimal mini-batch and step sizes for SAGA

Optimal mini-batch size

\[ n = 4912, \ d = 300, \ \lambda = \frac{100}{n}, \ \epsilon = 10^{-3}, \ \tau = \frac{n}{5} \]

![Graph showing the relationship between error and epoch number for different batch sizes and methods, including singletons, \( \tau \)-ind, \( \tau \)-nice, and \( 2633 = \tau^* \)-ind, \( 2633 = \tau^* \)-nice.](image-url)
Learning rate schedules
Main Theorem (Linear convergence to a neighborhood)

**Theorem** \((f, \mathcal{D}) \sim ES(\mathcal{L})\) and \(\mu\)-quasi strongly convex

\[
\mathbb{E}[\|w^t - w^*\|^2] \leq (1 - \gamma \mu)^t \|w^0 - w^*\|^2 + \frac{2\gamma \sigma^2}{\mu}
\]

Fixed stepsize \(\gamma_t \equiv \gamma \leq \frac{1}{2\mathcal{L}}\)

**Corollary** \(\gamma = \frac{1}{2} \max \left\{ \frac{1}{\mathcal{L}}, \frac{\epsilon \mu}{2\sigma^2} \right\}\)

\[t \geq \max \left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon \mu^2} \right\} \log \left( \frac{2}{\epsilon} \right) \quad \Rightarrow \quad \frac{\mathbb{E}[\|w^t - w^*\|]}{\|w^0 - w^*\|} \leq \epsilon\]

*saves time* for theorists: Includes GD and SGD as special cases. Also tighter!