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

Robert M. Gower

TELECOM ParisTech

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 : $h_w(x) \mapsto y$

$h_w\Bigg($  $\Bigg) \mapsto$ **Cat**

Optimization in Machine Learning

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

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

$h_w\Bigg($  $\Bigg) \quad \mapsto$ **Cat**

(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 : $h_w\,(x) \quad \mapsto \quad y$

$h_w \Big($  $\Big) \quad \mapsto$ **Cat**

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

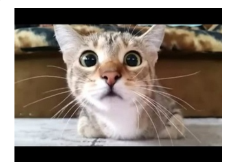
(4) Solve the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

Optimization in Machine Learning

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

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

$h_w\Big($  $\Big) \quad \mapsto$ **Cat**

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

(4) Solve the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

(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 : $h_w\ (x) \quad \mapsto \quad y$

$h_w\Bigg($  $\Bigg) \quad \mapsto$ **Cat**

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

(4) Solve the *training problem*: Optimization

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

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

Finite sum minimization

$= \ell\left(h_w(x^i), y^i\right)$
loss function of
ith data point

$$(I) \qquad \min_{w\in\mathbb{R}^d} f(w) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

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

Finite sum minimization

$= \ell\left(h_w(x^i), y^i\right)$
loss function of
ith data point

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

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

Finite sum minimization

$= \ell\left(h_w(x^i), y^i\right)$
loss function of
ith data point

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

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 <u>parameter choices</u> and implementations

saves time *for practitioners:* Less <u>hyper-parameter tuning</u> ➡ works out of the box

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

Finite sum minimization

$= \ell\left(h_w(x^i), y^i\right)$
loss function of
ith data point

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

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 <u>parameter choices</u> and implementations

saves time *for practitioners:* Less <u>hyper-parameter tuning</u> ➡ works out of the box

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

<u>Case study today</u>: **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)

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

Step size/ Learning rate

Sampled i.i.d $j \in \{1,\ldots,n\}$

What about mini-batching

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)

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

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)

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

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

- What should b be?

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?
- How does b influence the stepsizes?

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

The Stochastic Gradient Method

Solving the *training problem*:

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

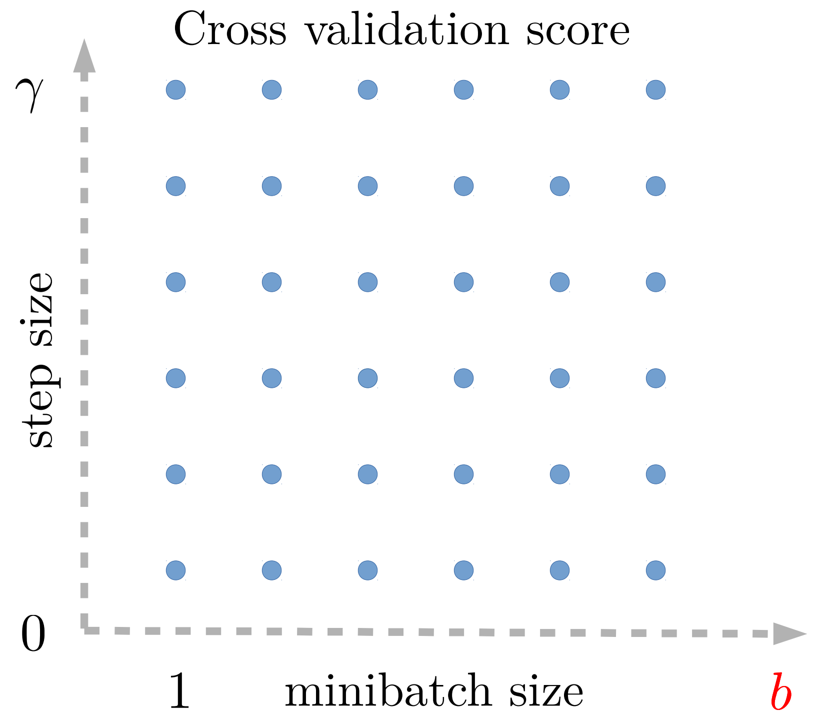
Baseline method: Stochastic Gradient Descent (SGD)

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

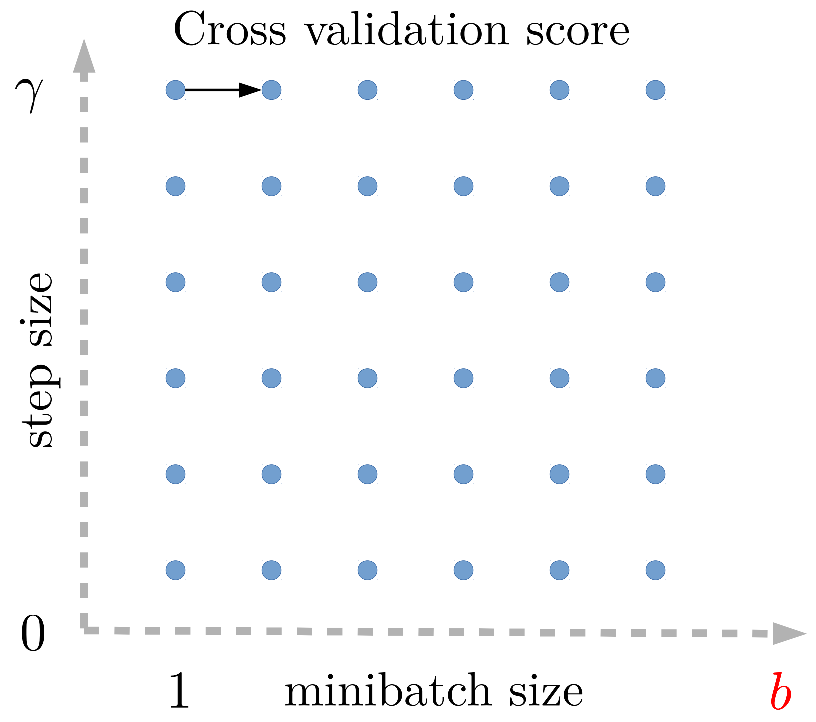
Minibatch where $B \in \{1,\ldots,n\}$ with $|B| = b$

- What should b be?
- How does b influence the stepsizes?
- How does the data influence the best mini-batch and stepsize?

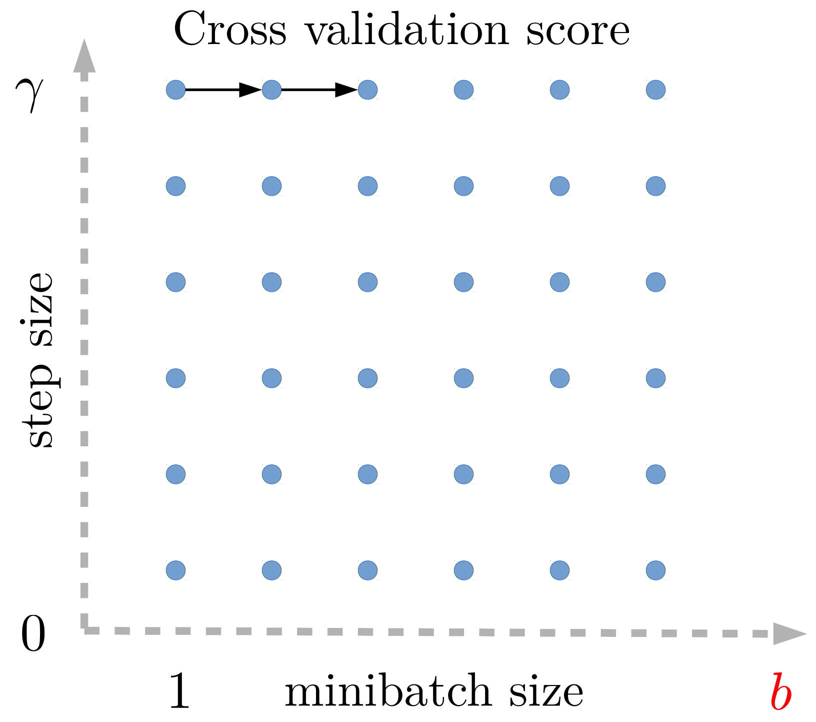
How to choose the minibatch size?



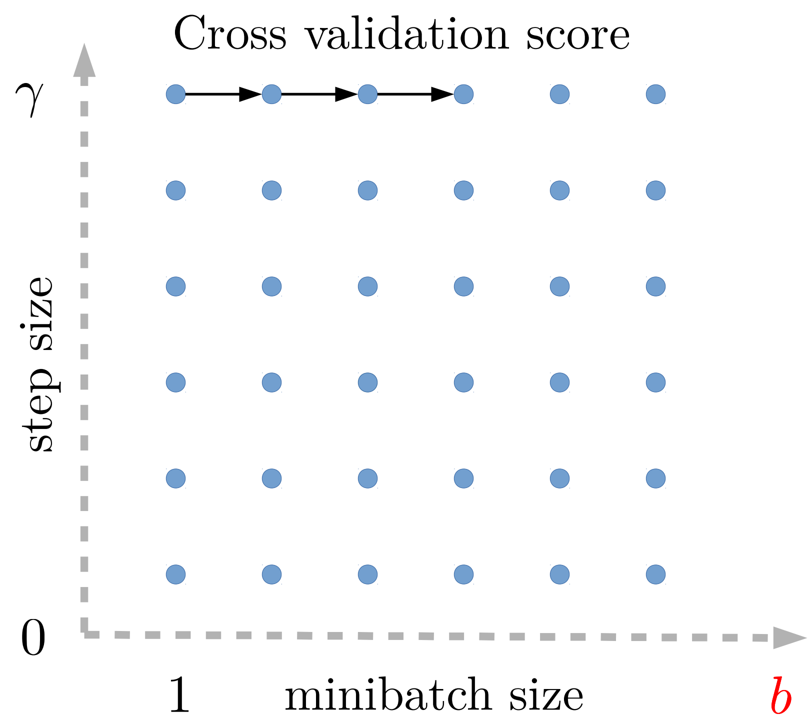
How to choose the minibatch size?



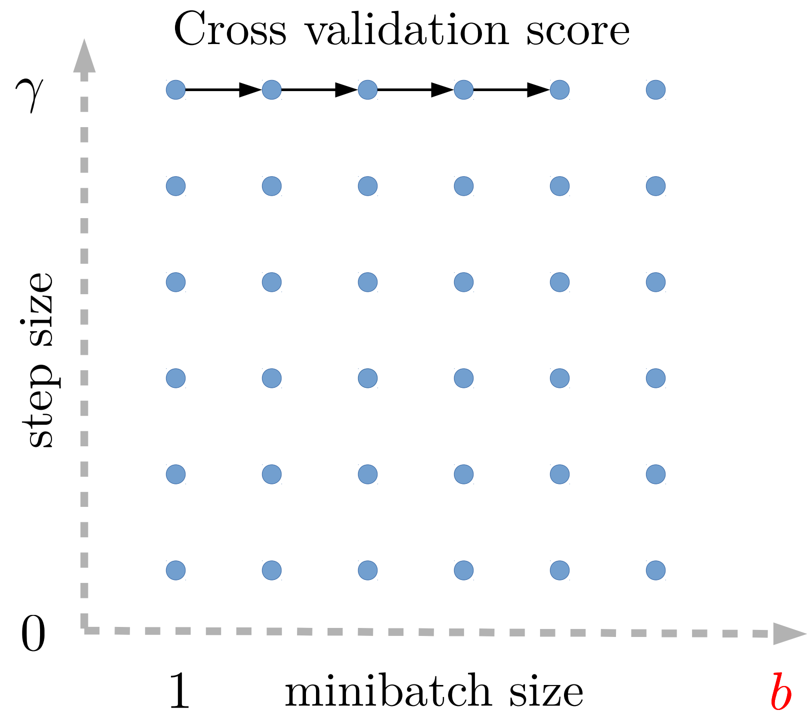
How to choose the minibatch size?



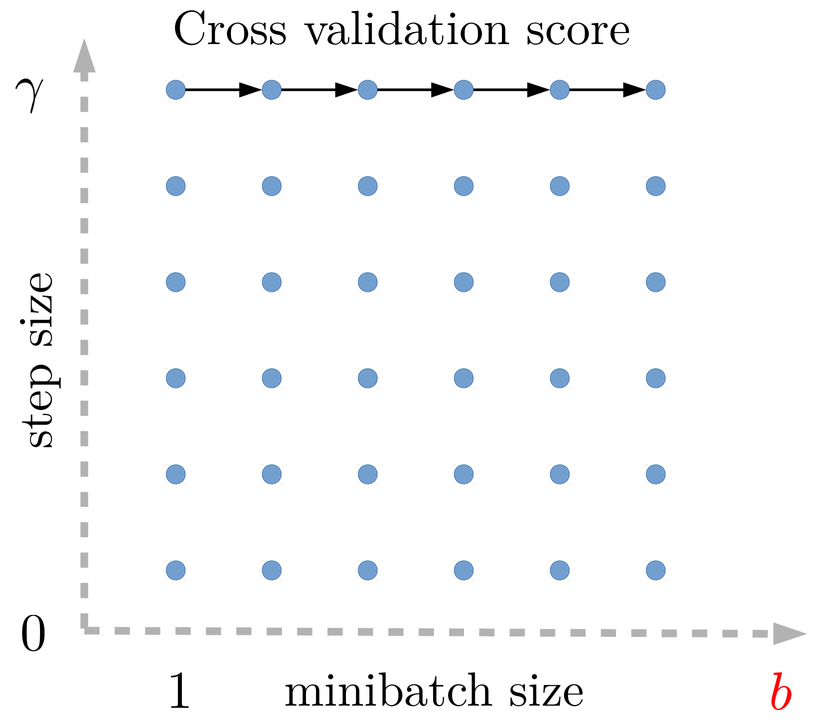
How to choose the minibatch size?



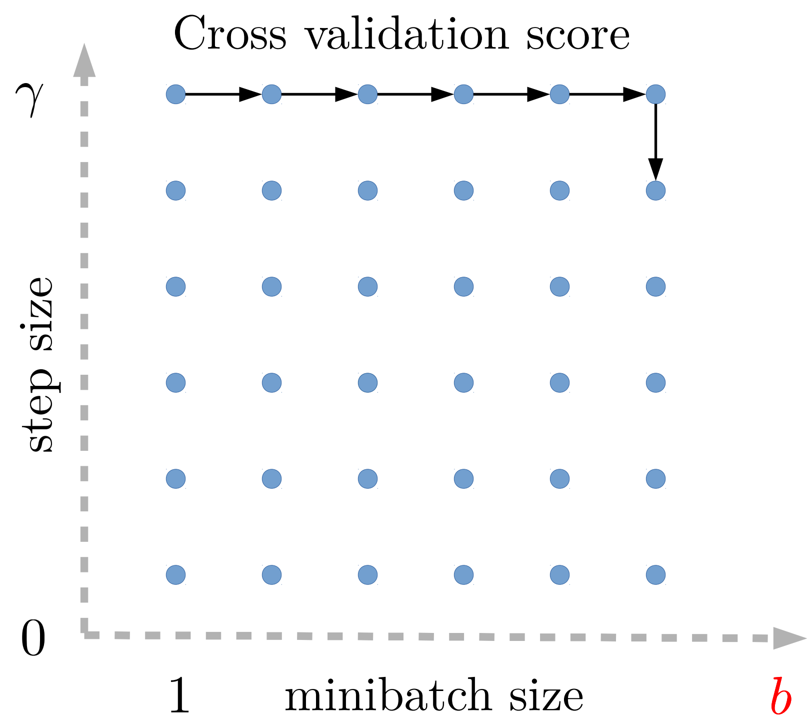
How to choose the minibatch size?



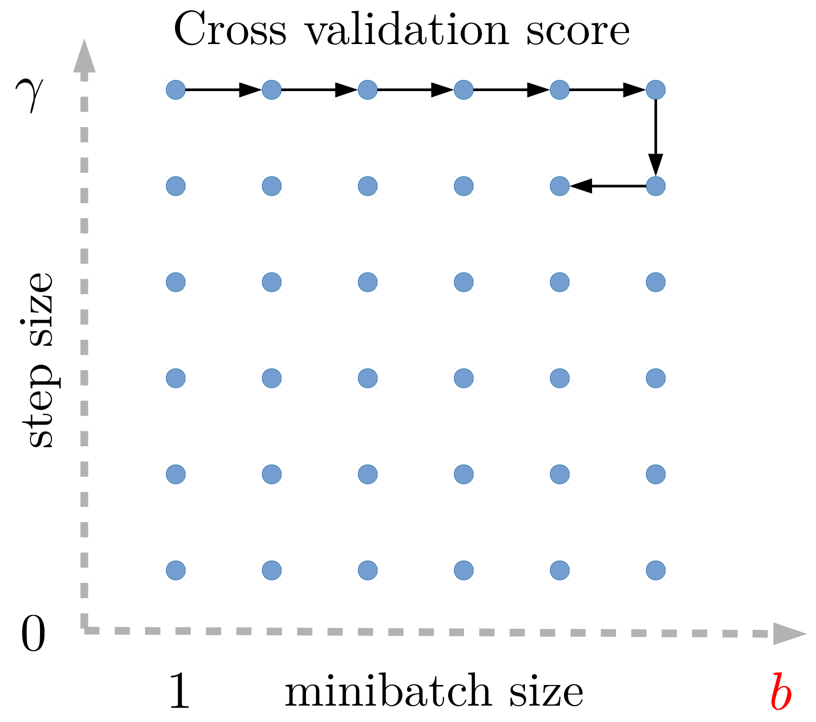
How to choose the minibatch size?



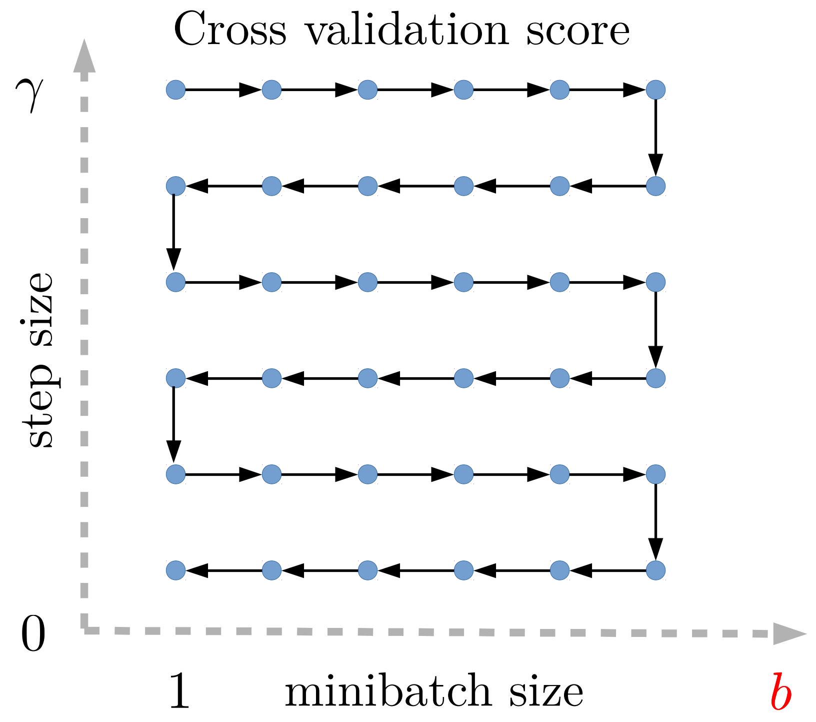
How to choose the minibatch size?



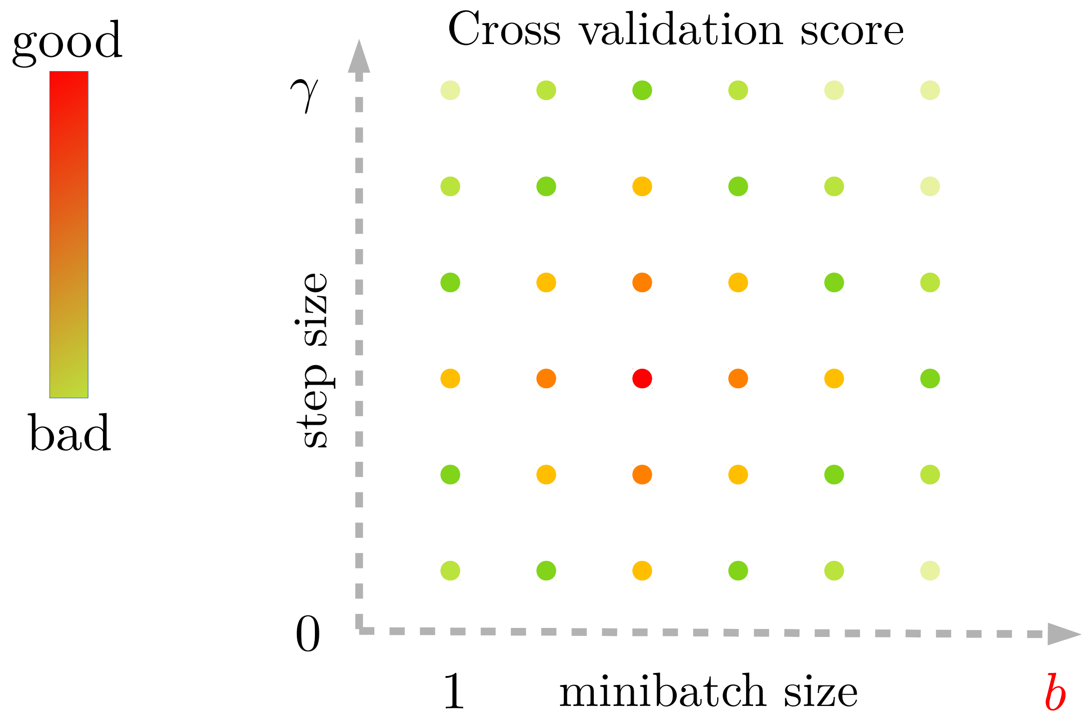
How to choose the minibatch size?



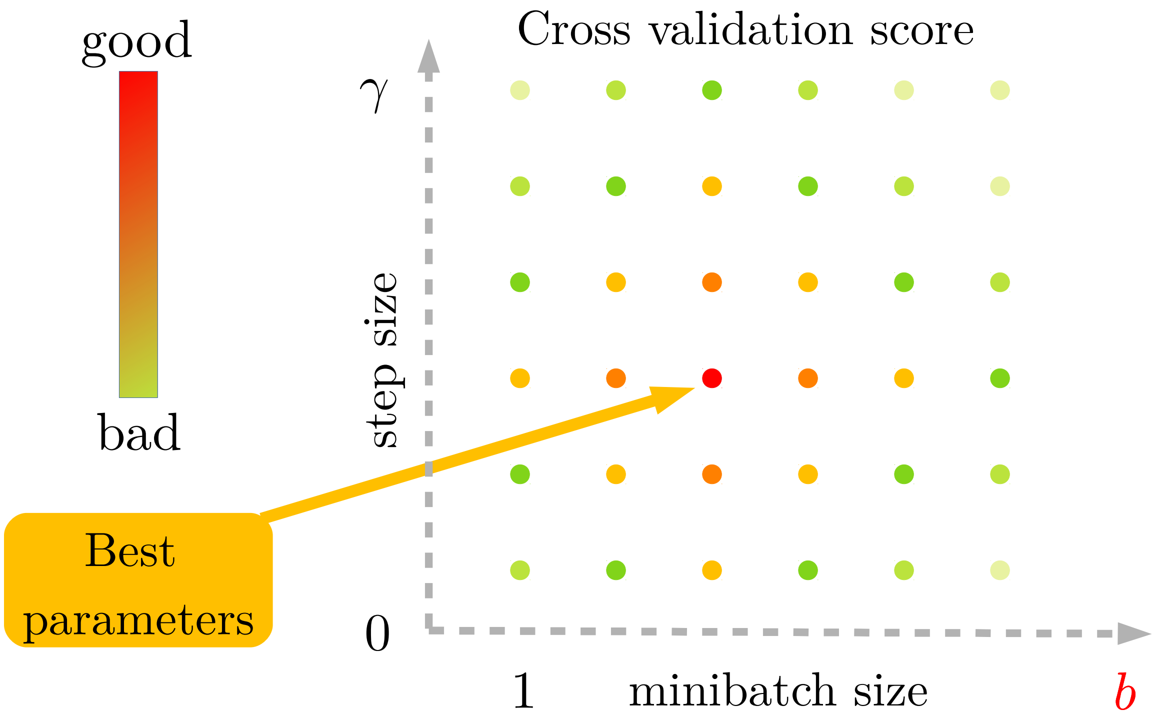
How to choose the minibatch size?



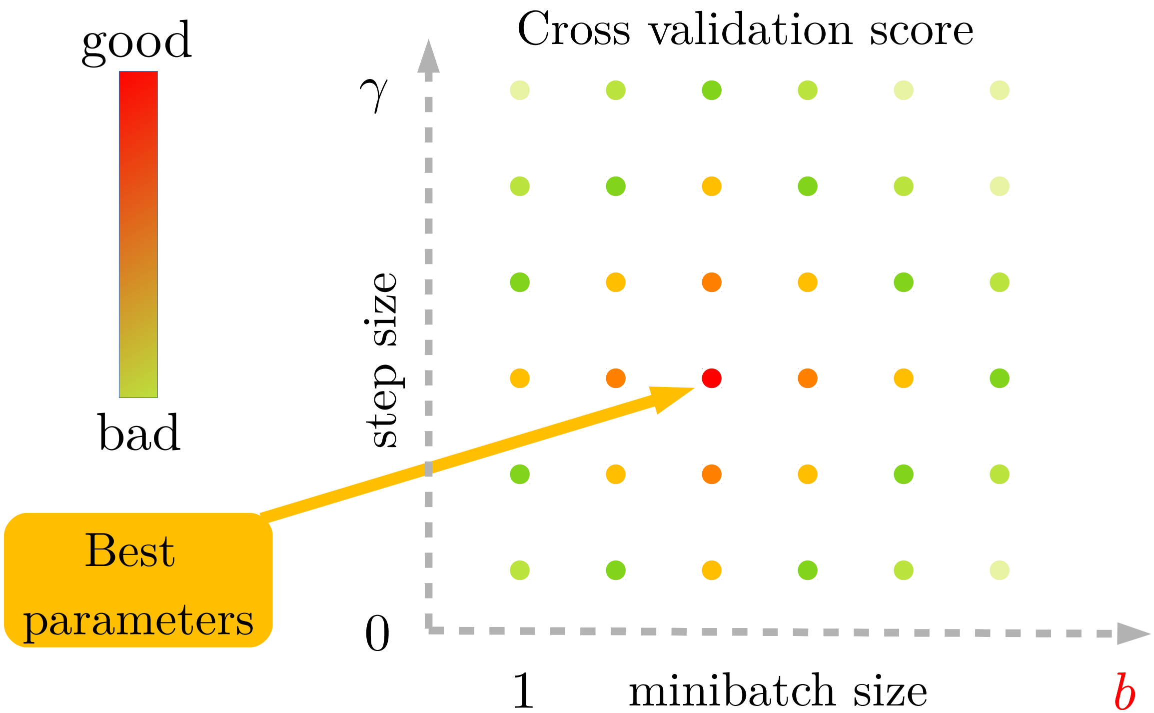
How to choose the minibatch size?



How to choose the minibatch size?



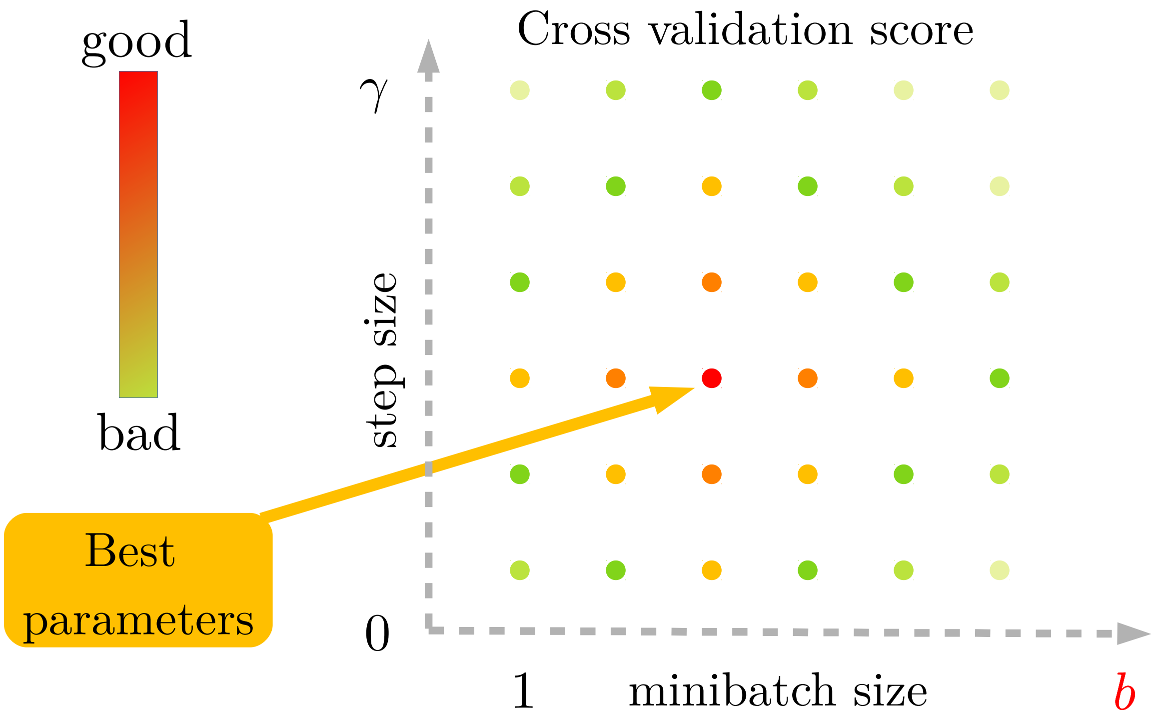
How to choose the minibatch size?



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

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

How to choose the minibatch size?

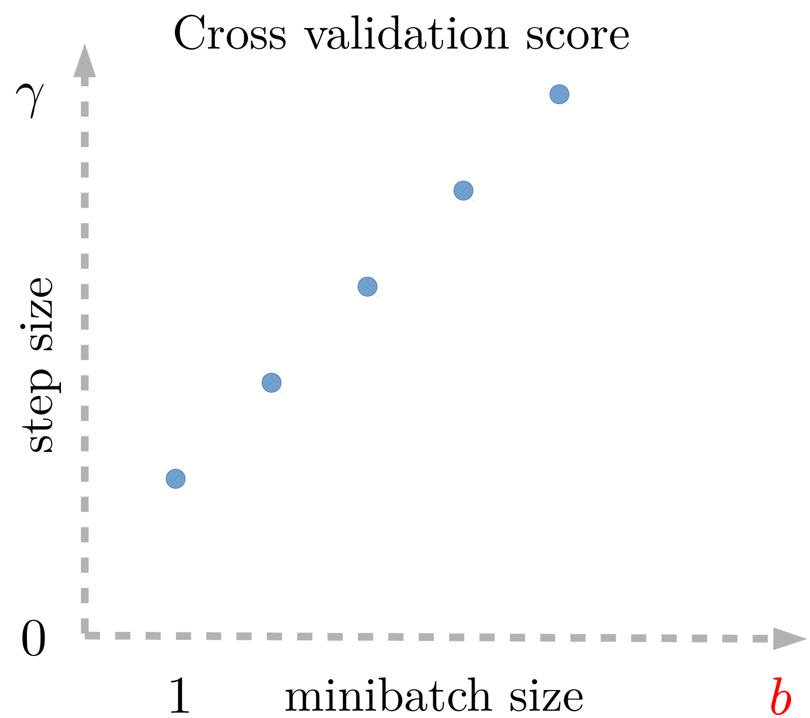


$$\gamma(b) = \text{const} \times b?$$

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

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

How to choose the minibatch size?

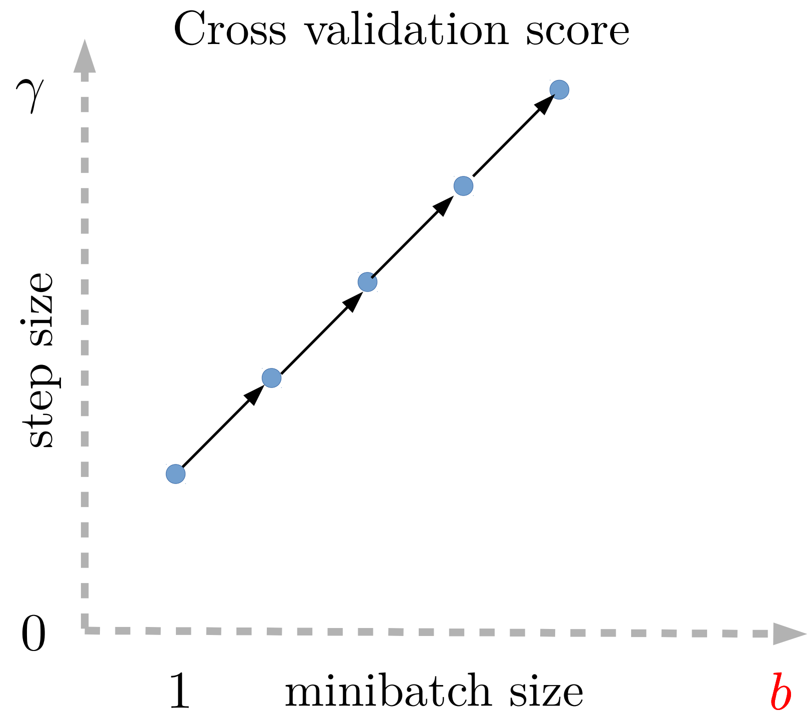


$\gamma(b) = \text{const} \times b$?

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

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

How to choose the minibatch size?

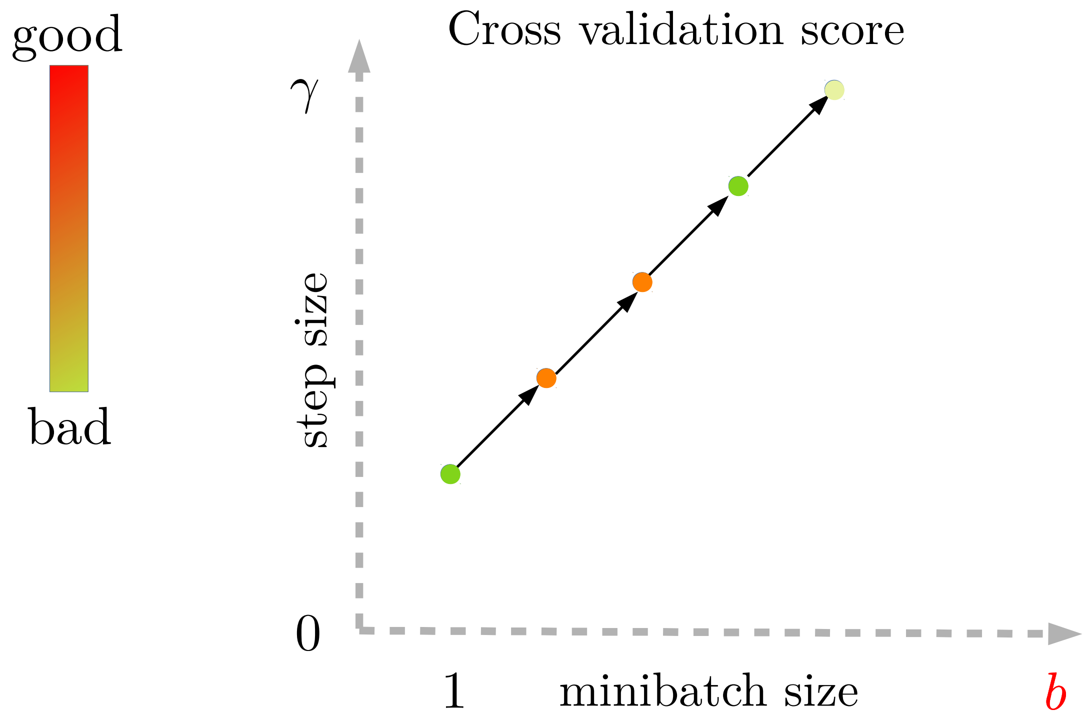


$\gamma(b) = \text{const} \times b$?

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

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

How to choose the minibatch size?

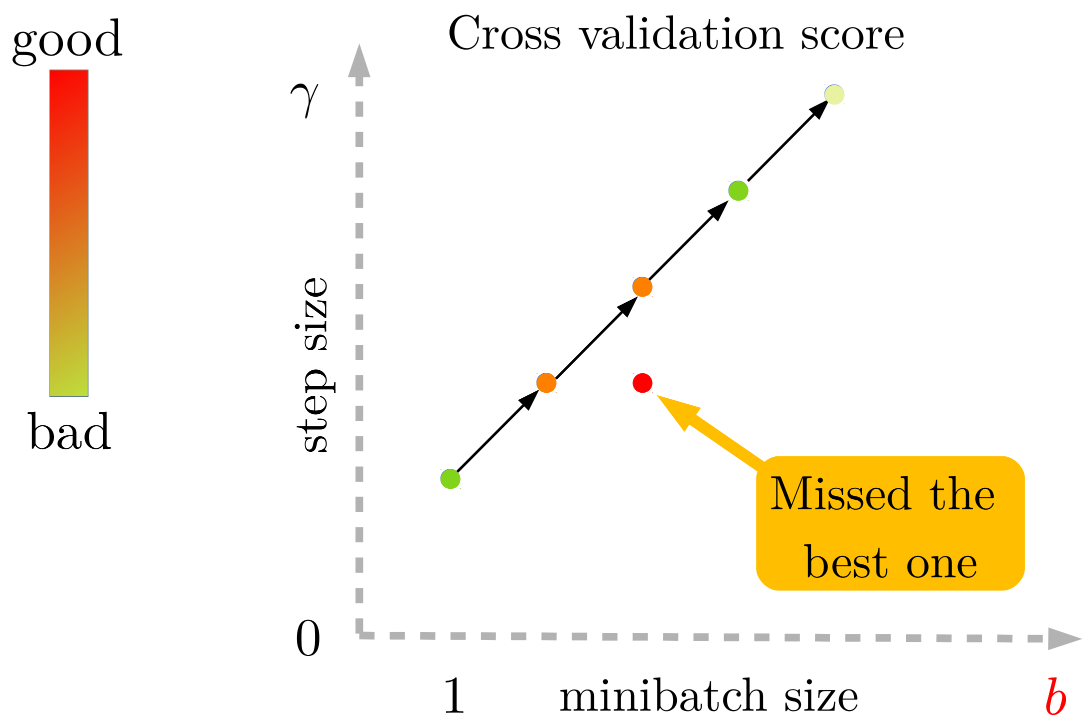


$$\gamma(b) = \text{const} \times b?$$

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

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

How to choose the minibatch size?

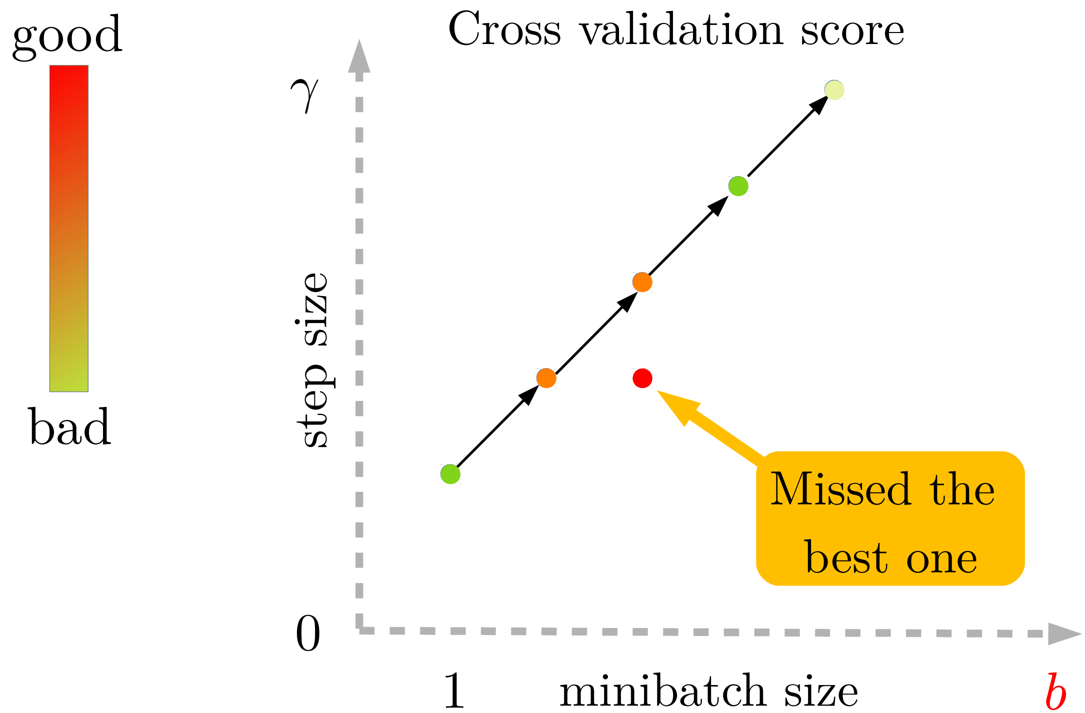


$$\gamma(b) = \text{const} \times b?$$

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

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

How to choose the minibatch size?



Need to figure out functional relationship between minibatch size and step size

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

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

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

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 $v = (v_1, \ldots, v_n) \sim \mathcal{D}$ with

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

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

Simple Stochastic Reformulation

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

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

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

Simple Stochastic Reformulation

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

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

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

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

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

SGD with arbitrary sampling

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

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

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

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

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

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
$$\mathbb{E}[\nabla f_{v^t}(w^t)] = \nabla f(w^t)$$

Example: Gradient descent

$$v \equiv (1, \ldots, 1) \quad \Rightarrow \quad w^{t+1} = w^t - \gamma_t \nabla f(w^t)$$

SGD with arbitrary sampling

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

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

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

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

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

$v \equiv (1, \ldots, 1)$ ➡ $w^{t+1} = w^t - \gamma_t \nabla f(w^t)$

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

Examples of arbitrary sampling: uniform single element

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

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

Examples of arbitrary sampling: uniform single element

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

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

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

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

Examples of arbitrary sampling: uniform single element

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

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

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

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

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

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

Examples of arbitrary sampling: uniform single element

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

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

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

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

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

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

Single element SGD

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

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

Examples of arbitrary sampling: uniform mini-batching

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

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

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

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

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

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

Mini-batch SGD without replacement

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

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

Examples of arbitrary sampling: non-uniform mini-batching

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

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

Richtárik and Takáč (arXiv:1310.3438; Opt Letters 2016)

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

Richtárik and Takáč (arXiv:1310.3438; Opt Letters 2016)

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

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

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

Richtárik and Takáč (arXiv:1310.3438; Opt Letters 2016)

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

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

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

Arbitrary sampling SGD

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

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

Richtárik and Takáč (arXiv:1310.3438; Opt Letters 2016)

SGD with arbitrary sampling

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

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

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

Includes all forms of SGD (and GD)

SGD with arbitrary sampling

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

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

quasi strong convexity constant

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

Smoothness constant

$$||\nabla f(w) - \nabla f(w^*)||_2^2 \;\leq\; 2L\,(f(w) - f(w^*))$$

Assumptions and Convergence of Gradient Descent

quasi strong convexity constant

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

Smoothness constant

$$||\nabla f(w) - \nabla f(w^*)||_2^2 \;\leq\; 2L\,(f(w) - f(w^*))$$

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

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

Iteration complexity of gradient descent

Given $\epsilon > 0$ and $t \geq \frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) \Rightarrow \frac{||w^t - w^*||}{||w^0 - w^*||} \leq \epsilon$

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

$$\frac{1}{n}\sum_{i=1}^{n} ||\nabla f_i(w) - \nabla f_i(w^*)||_2^2 \;\leq\; 2L_{\max}\left(f(w) - f(w^*)\right)$$

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

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

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

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

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

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

Iteration complexity of SGD

$$t \geq \left(\frac{L_{\max}}{\mu} + \frac{\sigma_*^2}{\epsilon\mu^2}\right)\log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \frac{\mathbb{E}[||w^t - w^*||]}{||w^0 - w^*||} \leq \epsilon$$

Needell, Srebro, Ward: Math. Prog, 2016.

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

$$t \geq O\left(\frac{L_{\max}}{\mu} + \frac{\sigma_*^2}{\epsilon \mu^2}\right)$$

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

Informal comparison between GD and SGD iteration complexity

GD

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

SGD

$$t \geq O\left(\frac{L_{\max}}{\mu} + \frac{\sigma_*^2}{\epsilon\mu^2}\right)$$

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

How do they compare?

In general: $L \leq L_{\max} \leq nL$

Informal comparison between GD and SGD iteration complexity

GD

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

SGD

$$t \geq O\left(\frac{L_{\max}}{\mu} + \frac{\sigma_*^2}{\epsilon\mu^2}\right)$$

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

When n is big
$L \ll L_{\max}$

How do they compare?

In general: $L \leq L_{\max} \leq nL$

Informal comparison between GD and SGD iteration complexity

GD

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

SGD

$$t \geq O\left(\frac{L_{\max}}{\mu} + \frac{\sigma_*^2}{\epsilon \mu^2}\right)$$

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

When n is big
$L \ll L_{\max}$

How do they compare?

In general: $L \leq L_{\max} \leq nL$

Need new "interpolating" notion of smoothness

$$L \leq ?\ L(\mathcal{D})\ ? \ \leq L_{\max}$$

Key constant: Expected smoothness

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

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

Key constant: Expected smoothness

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

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

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

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 smoothnes 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

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

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

Expected smoothnes constant

Depends on v and f

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)

Lemma:

f_i convex and $L_{\max}$–smooth

$$\Downarrow$$

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

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

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 smoothnes constant

Depends on v and f

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)

Lemma:

f_i convex and $L_{\max}$–smooth

$\Downarrow$

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

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

Rough estimate
(we can do better)

Key constant: Expected smoothness

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

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

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

Expected smoothnes constant

Depends on v and f

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)

Lemma:

f_i convex and $L_{\max}$–smooth

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

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

Rough estimate
(we can do better)

Definition: Gradient noise

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

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 smoothnes constant

Depends on v and f

RMG, Richtárik and Bach (arXiv:1805.02632, 2018)

Lemma:

f_i convex and $L_{\max}$–smooth

$\Downarrow$

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

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

Rough estimate
(we can do better)

Definition: Gradient noise

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

Generalization of

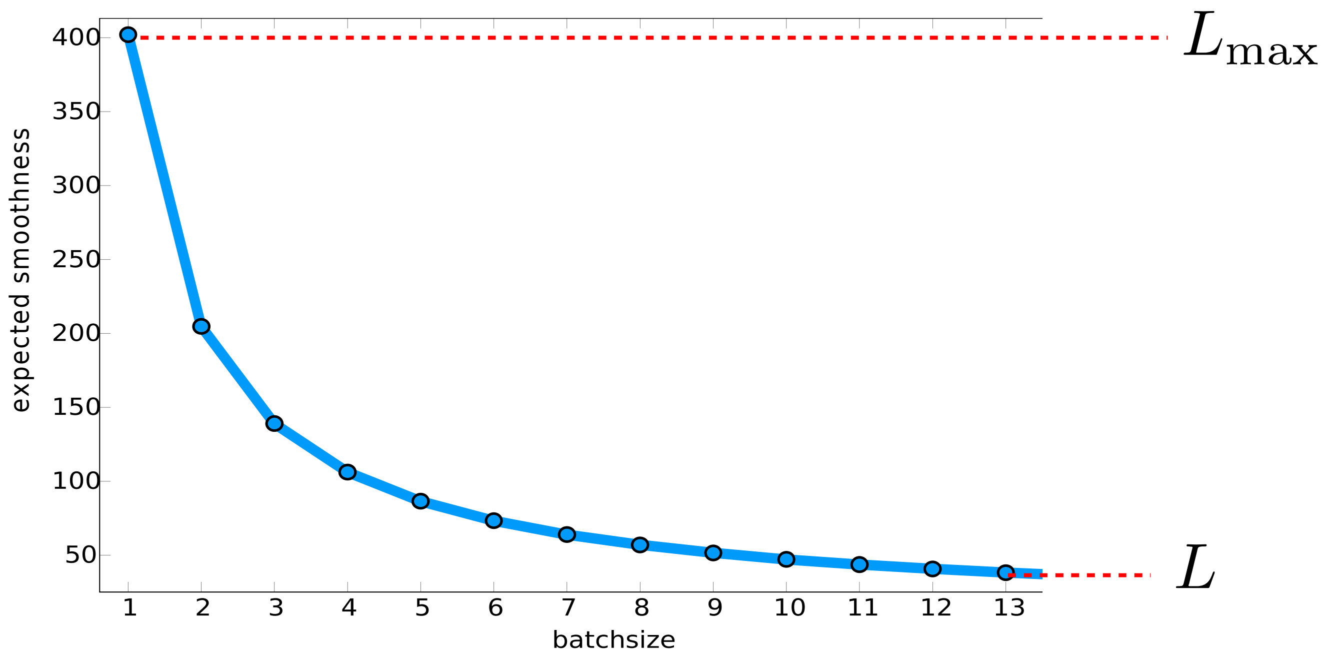
$$\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w^*)\|^2$$

Example of Expected Smoothness

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

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

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

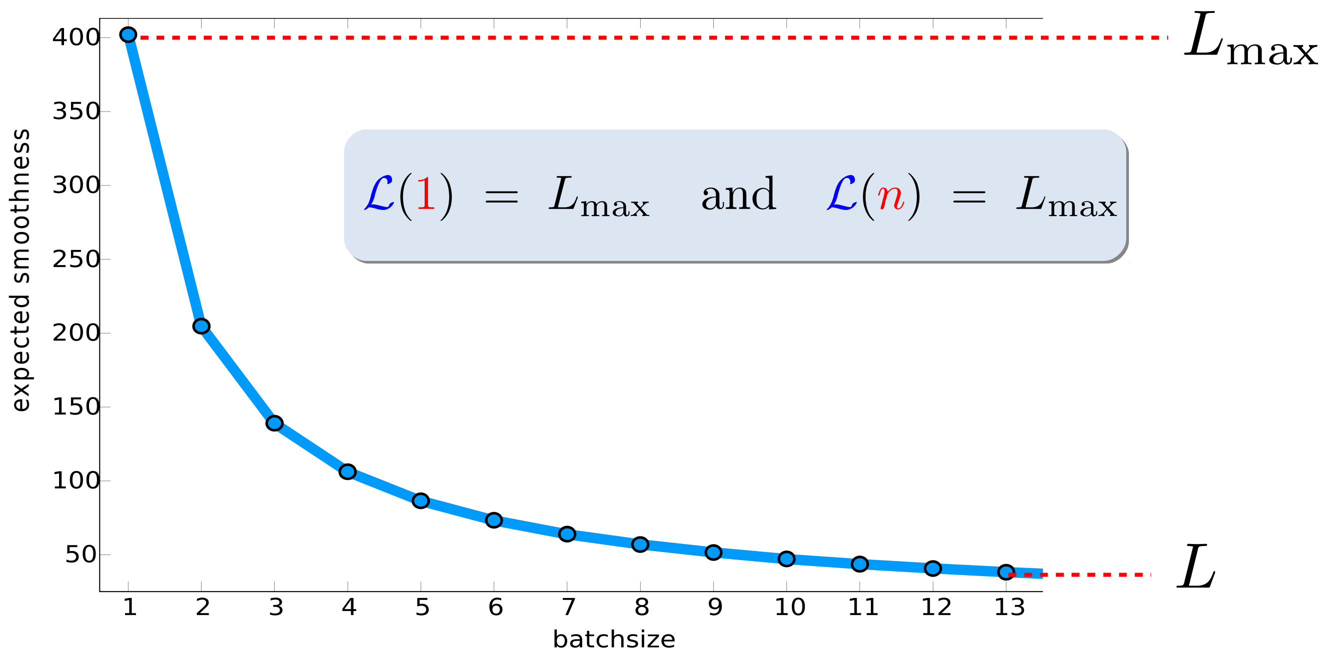


Example of Expected Smoothness

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

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

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



$$\mathcal{L}(1) \;=\; L_{\max} \quad \text{and} \quad \mathcal{L}(n) \;=\; L_{\max}$$

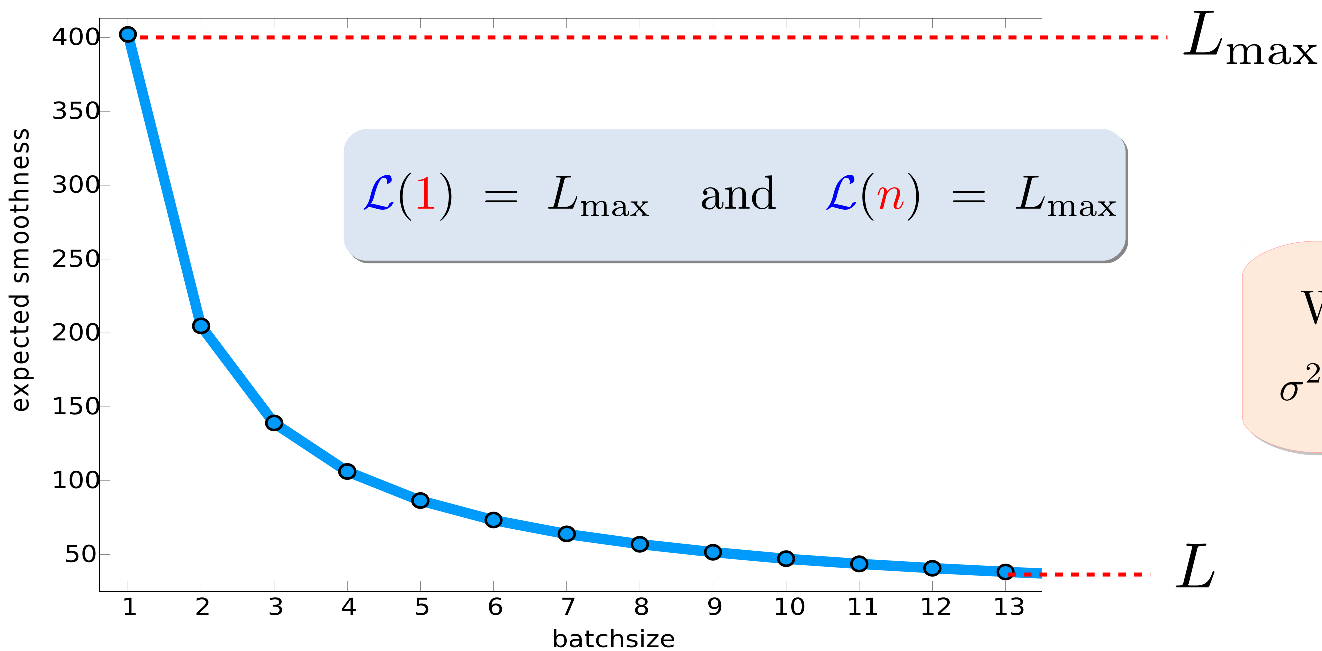
Example of Expected Smoothness

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

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

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

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



What about σ^2?

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

Example of Expected Smoothness

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

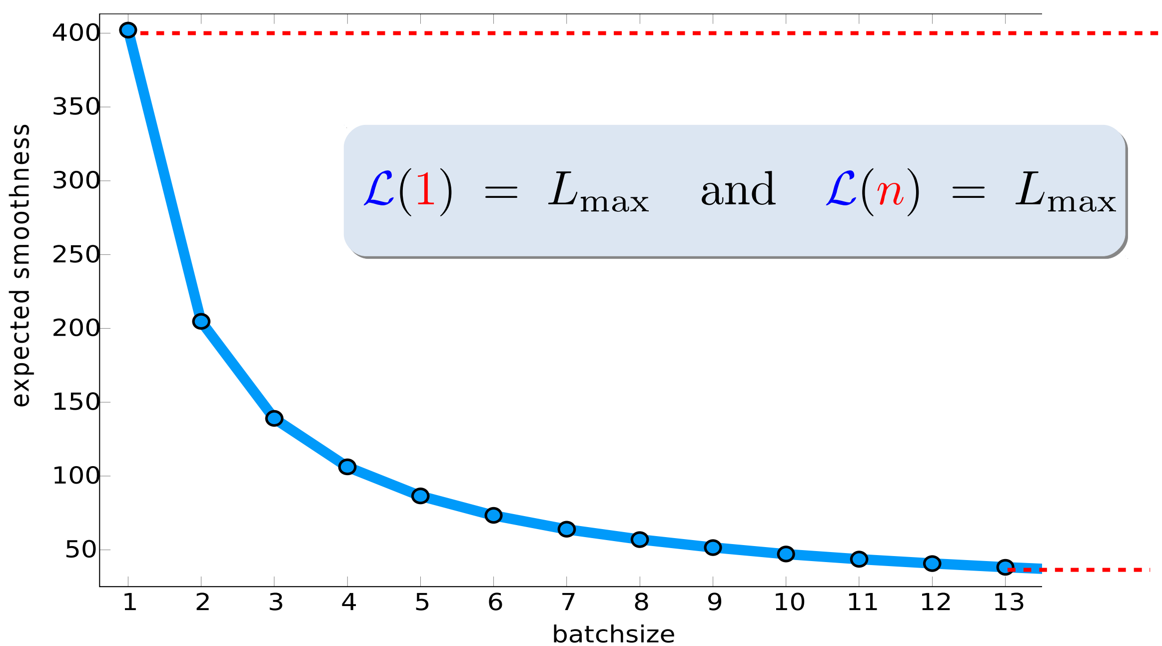
S is chosen uniformly at random from all subsets of size b

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

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

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

Measures how much model fits data



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

$L_{\max}$ $\sigma^2 = \sigma_*^2$

L $\sigma^2 = 0$

What about σ^2?

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

Expected smoothness gives awesome bound on gradient

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

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

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

Expected smoothness gives awesome bound on gradient

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

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

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

Normally bound on gradient is an *assumption*

Assumption There exists $B>0$

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

PDF Adobe Recht, Wright & Niu, F. Hogwild: Neurips, 2011.

PDF Adobe Hazan & Kale, JMLR 2014.

PDF Adobe Rakhlin, Shamir, & Sridharan, ICML 2012

PDF Adobe Shamir & Zhang, ICML 2013.

Expected smoothness gives awesome bound on gradient

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

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

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

Normally bound on gradient is an *assumption*

~~**Assumption** There exists $B>0$~~

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

Recht, Wright & Niu, F. Hogwild: Neurips, 2011.

Hazan & Kale, JMLR 2014.

Rakhlin, Shamir, & Sridharan, ICML 2012

Shamir & Zhang, ICML 2013.

Expected smoothness gives awesome bound on gradient

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

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

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

informative: with realistic assumptions

Normally bound on gradient is an *assumption*

~~**Assumption** There exists $B > 0$~~

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

Recht, Wright & Niu, F. Hogwild: Neurips, 2011.

Hazan & Kale, JMLR 2014.

Rakhlin, Shamir, & Sridharan, ICML 2012

Shamir & Zhang, ICML 2013.

Main Theorem (Linear convergence to a neighborhood)

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

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

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

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

Main Theorem (Linear convergence to a neighborhood)

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

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

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

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

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

Main Theorem (Linear convergence to a neighborhood)

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

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

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–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 $\gamma = \frac{1}{2}\max\left\{\frac{1}{\mathcal{L}}, \frac{\epsilon\mu}{2\sigma^2}\right\}$

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

Main Theorem (Linear convergence to a neighborhood)

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

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

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–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!

Proof is SUPER EASY:

$$
\begin{aligned}
||w^{t+1}-w^*||_2^2 &= ||w^t - w^* - \gamma \nabla f_v(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\gamma\langle \nabla f_v(w^t), w^t - w^*\rangle + \gamma^2 ||\nabla f_v(w^t)||_2^2.
\end{aligned}
$$

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

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

$$
\begin{aligned}
\mathbb{E}_v\left[||w^{t+1}-w^*||_2^2\right] &= ||w^t - w^*||_2^2 - 2\gamma\langle \nabla f(w^t), w^t - w^*\rangle + \gamma^2 \mathbb{E}_v\left[||\nabla f_v(w^t)||_2^2\right] \\
&\leq (1-\gamma\mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_v\left[||\nabla f_v(w^t)||_2^2\right] \\
&\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma\mathcal{L} - 1)(f(w) - f(w^*)) + 2\gamma^2\sigma^2 \\
&\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma^2\sigma^2
\end{aligned}
$$

quasi strong conv

$\gamma \leq \frac{1}{2\mathcal{L}}$

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

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

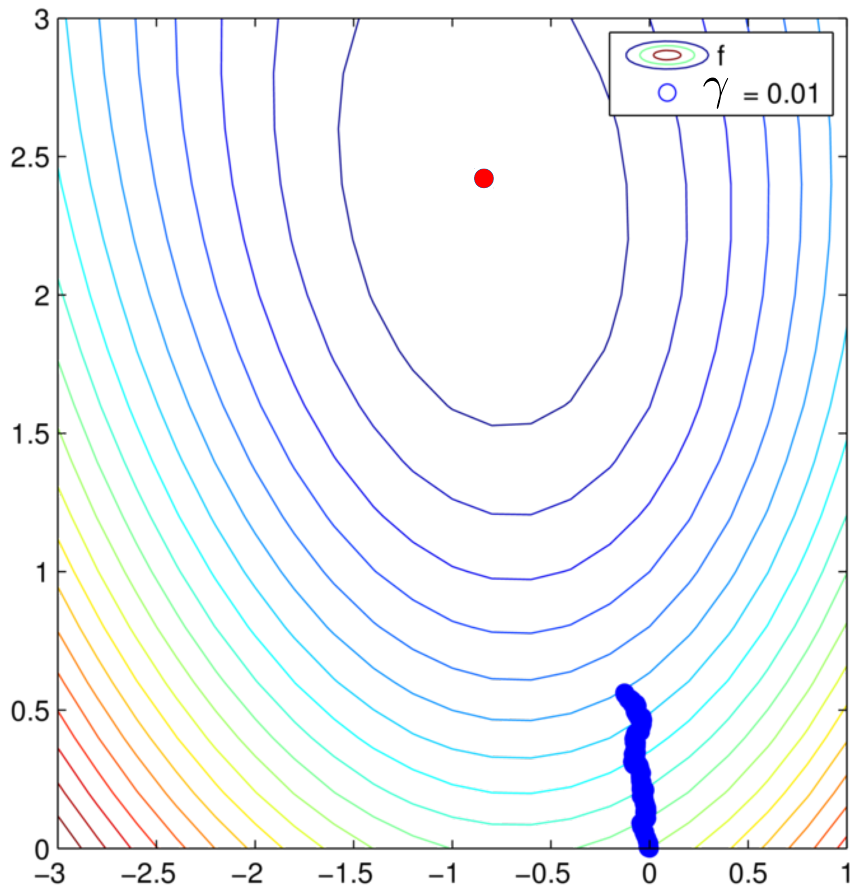
Taking total expectation

$$
\begin{aligned}
\mathbb{E}\left[||w^{t+1}-w^*||_2^2\right] &\leq (1-\gamma\mu)\mathbb{E}\left[||w^t - w^*||_2^2\right] + 2\gamma^2\sigma^2 \\
&= (1-\gamma\mu)^{t+1}||w^0 - w^*||_2^2 + 2\sum_{i=0}^{t}(1-\gamma\mu)^i\gamma^2\sigma^2 \\
&\leq (1-\gamma\mu)^{t+1}||w^0 - w^*||_2^2 + \frac{2\gamma\sigma^2}{\mu}
\end{aligned}
$$

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

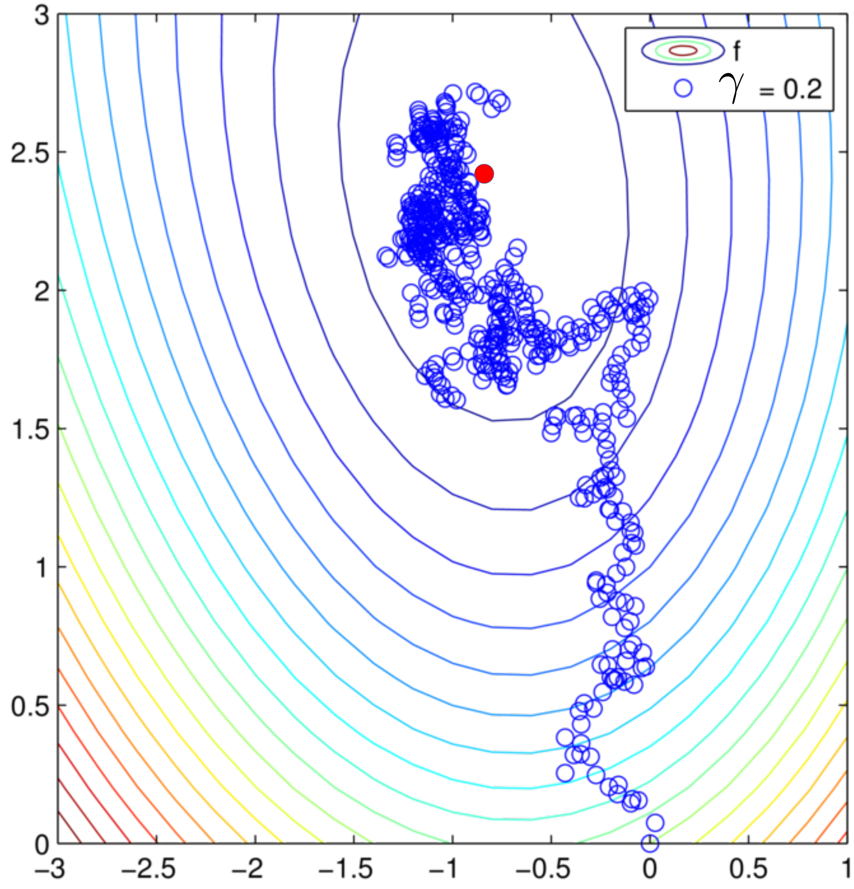
Stochastic Gradient Descent

$\gamma = 0.01$



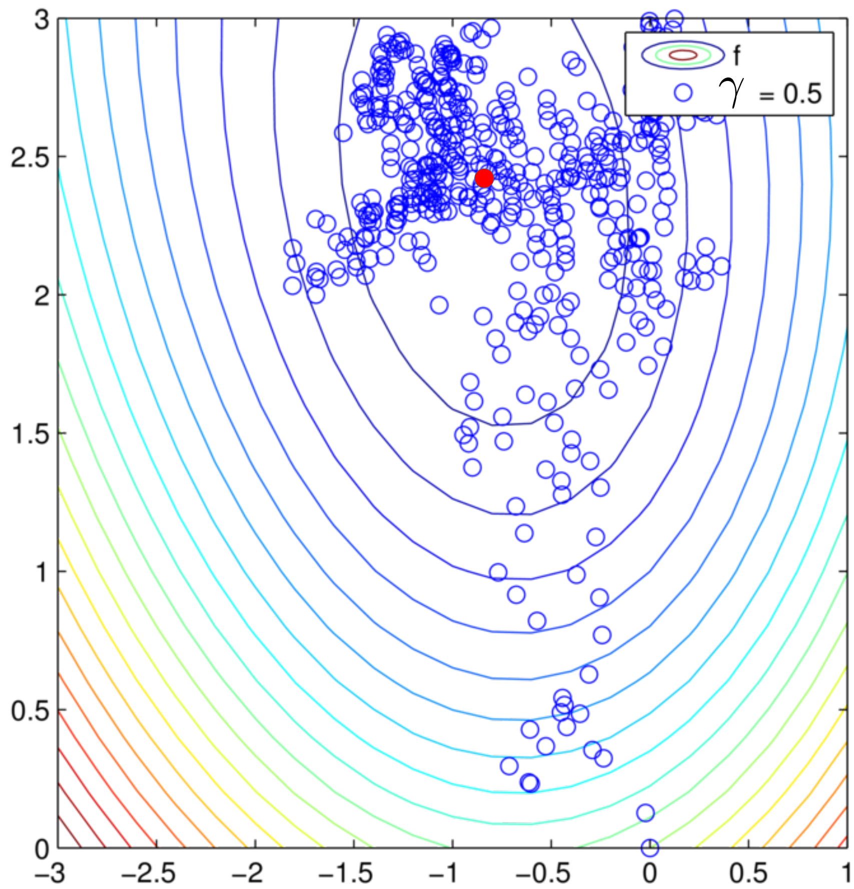
Stochastic Gradient Descent

$\gamma = 0.2$



Stochastic Gradient Descent

$\gamma = 0.5$



Total complexity for mini-batch SGD

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

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

Total complexity for mini-batch SGD

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

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

Corollary

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

Total complexity for mini-batch SGD

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

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

Corollary

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

Total complexity for mini-batch SGD

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

#stochastic gradient evaluation in 1 iteration

Corollary

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

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

Total complexity for mini-batch SGD

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

#stochastic gradient evaluation in 1 iteration

Corollary

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

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

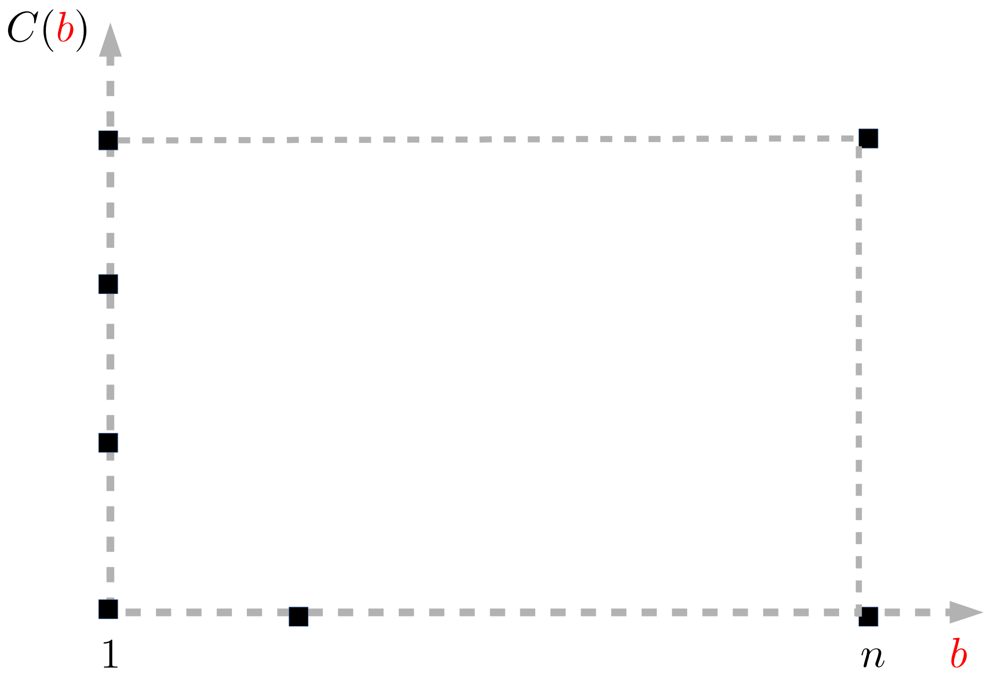
$$\left.\begin{aligned} \mathcal{L} &= \frac{n(b-1)}{b(n-1)}L + \frac{n-b}{b(n-1)}L_{\max} \\ \sigma^2 &= \frac{n-b}{b(n-1)}\sigma_*^2 \end{aligned}\right\} \Rightarrow$$

Total complexity is a simple function of mini-batch size b

Optimal mini-batch size

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

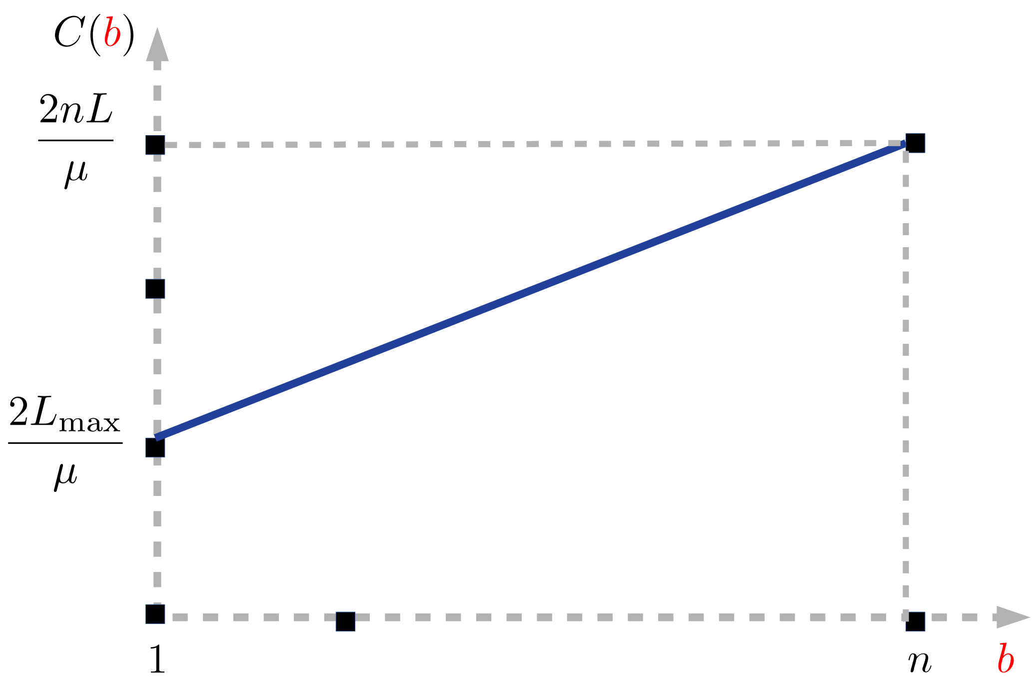
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ n(b-1)L + (n-b)L_{\max},\ \frac{2(n-b)\sigma_*^2}{\epsilon\mu} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$



Optimal mini-batch size

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

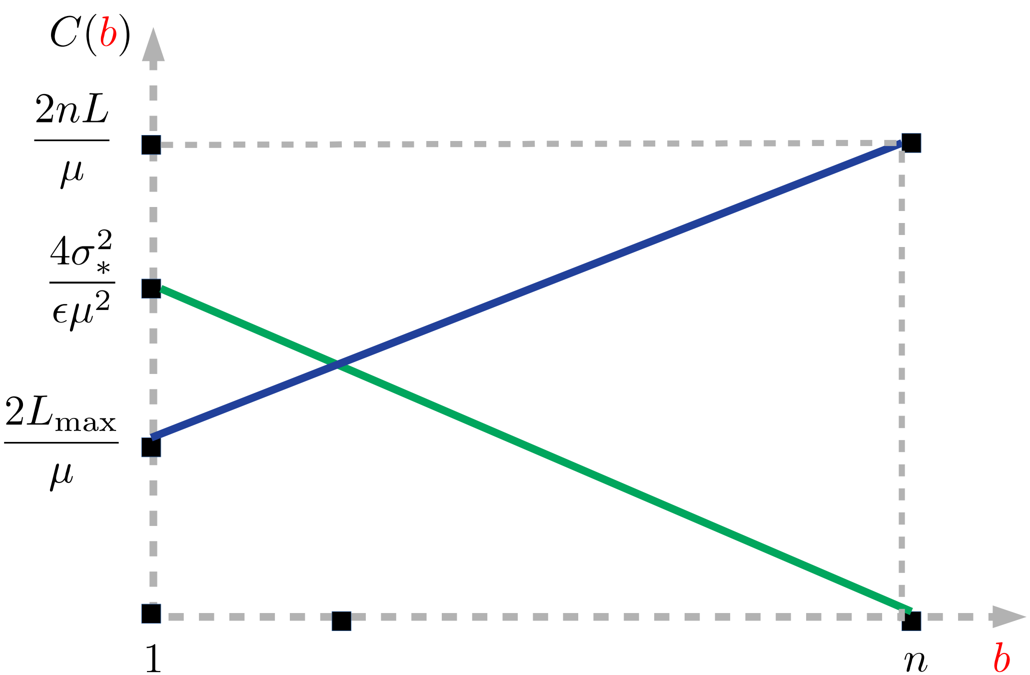
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \frac{2(n-b)\sigma_*^2}{\epsilon\mu} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$



Optimal mini-batch size

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

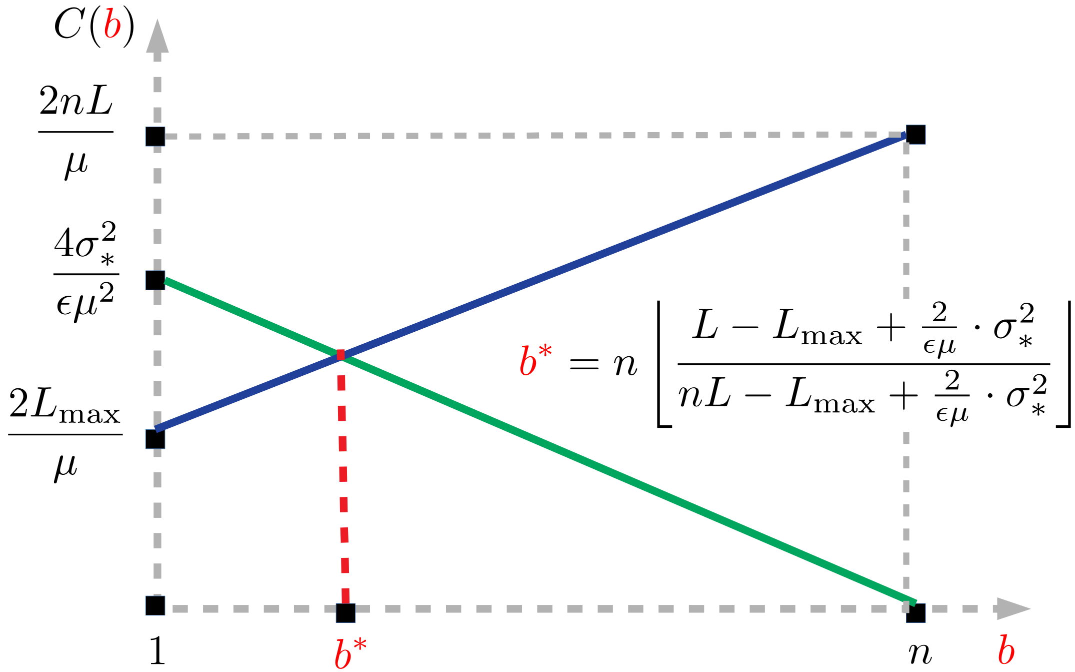
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{\underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}}\right\} \times \log\left(\frac{2}{\epsilon}\right)$$



Optimal mini-batch size

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

$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$



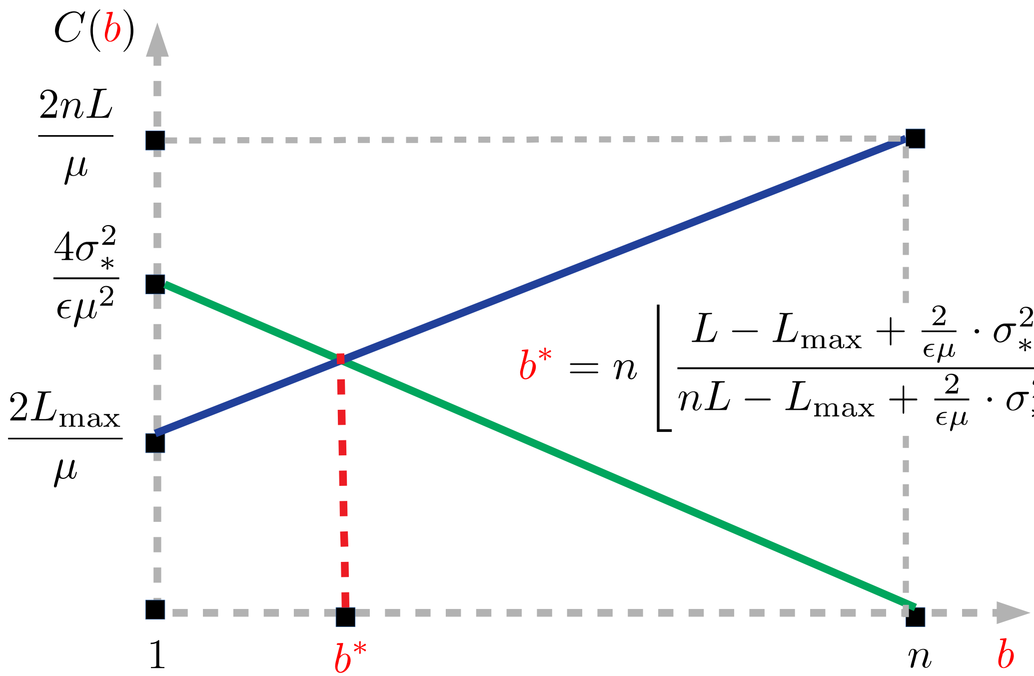
Optimal mini-batch size

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

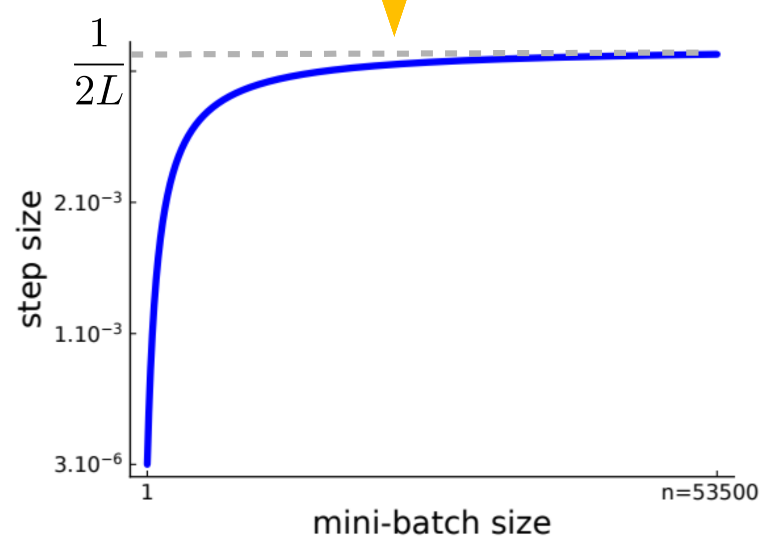
$$C(b) := \frac{2}{\mu(n-1)} \max\Big\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \Big\} \times \log\left(\frac{2}{\epsilon}\right)$$

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

Stepsize increases with b



$$b^* = n\left\lfloor \frac{L - L_{\max} + \frac{2}{\epsilon\mu}\cdot\sigma_*^2}{nL - L_{\max} + \frac{2}{\epsilon\mu}\cdot\sigma_*^2} \right\rfloor$$



Optimal mini-batch size for models that interpolate data

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

$$\times \log\left(\frac{2}{\epsilon}\right)$$

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

Optimal mini-batch size for models that interpolate data

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

$$0 \qquad \times \log\left(\frac{2}{\epsilon}\right)$$

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

Optimal mini-batch size for models that interpolate data

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

$$\times \log\left(\frac{2}{\epsilon}\right)$$

0

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

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

Optimal mini-batch size for models that interpolate data

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

$$0 \qquad\qquad \times \log\left(\frac{2}{\epsilon}\right)$$

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

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

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

Optimal mini-batch size for models that interpolate data

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

$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ n(b-1)L + (n-b)L_{\max},\ \cancel{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}^{\,0} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$

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

increases with b

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

$$\boxed{b^* = 1}$$

Optimal mini-batch size for models that interpolate data

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

$$\times \log\left(\frac{2}{\epsilon}\right)$$

$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ n(b-1)L + (n-b)L_{\max},\ \cancelto{0}{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}} \right\}$$

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

increases with b

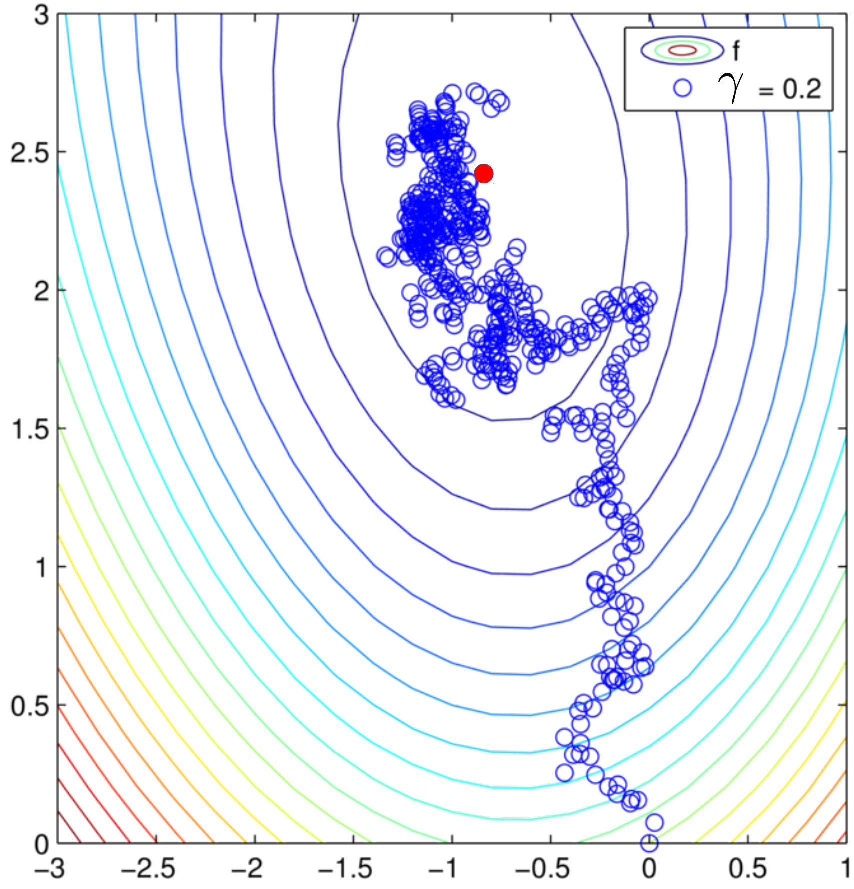
$$\gamma(b) := \frac{n-1}{2} \frac{b}{n(b-1)L + (n-b)L_{\max}}$$

$$b^* = 1$$

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

Stochastic Gradient Descent

$\gamma = 0.2$



Learning schedule: Constant & decreasing step sizes

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

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

Learning schedule: Constant & decreasing step sizes

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

Learning rate with switch point

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

Learning schedule: Constant & decreasing step sizes

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

Learning rate
with switch point

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

A stochastic
condition number

Learning schedule: Constant & decreasing step sizes

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

Learning rate with switch point

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

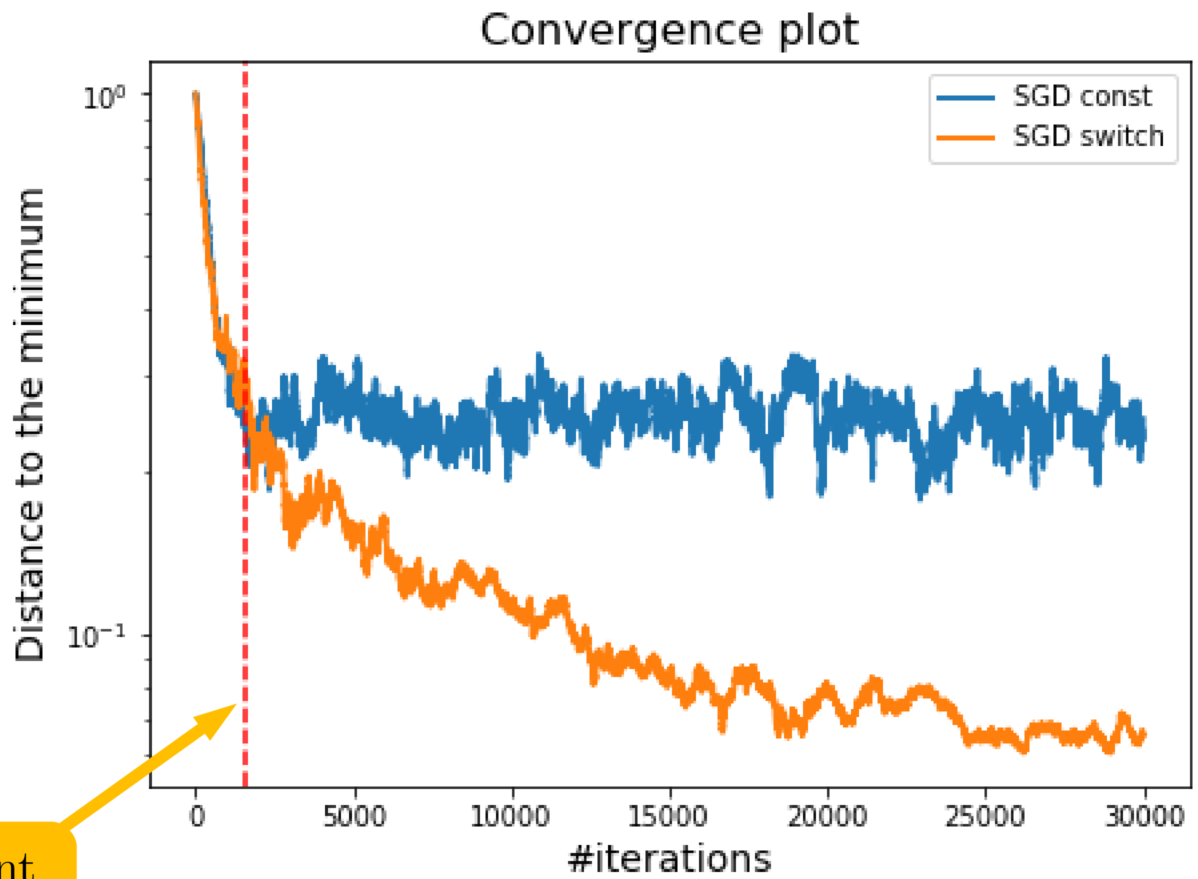
A stochastic condition number

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

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

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

Stochastic Gradient Descent with switch to decreasing stepsizes



Switch point
$t = 4\lceil \mathcal{K} \rceil$

Stochastic variance reduced methods

Simple Stochastic Reformulation

Random sampling vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ with

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

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

What to do about the variance?

Original finite sum problem

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

$\Longleftrightarrow$

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

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) \;=\; \mathbb{E}[f_{v}(w)] \;=\; \mathbb{E}[f_{v}(w)] - \mathbb{E}[z_{v}(w)] + \mathbb{E}[z_{v}(w)]$$

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

Cancel out

Controlled Stochastic Reformulation

covariate $z_{v}(w) \in \mathbb{R}$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_{v}(w)] = \mathbb{E}[f_{v}(w)] - \mathbb{E}[z_{v}(w)] + \mathbb{E}[z_{v}(w)]$$

$$= \mathbb{E}\big[f_{v}(w) - z_{v}(w) + \mathbb{E}[z_{v}(w)]\big]$$

Controlled Stochastic Reformulation

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

Cancel out

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

$$= \mathbb{E}\left[f_v(w) - z_v(w) + \mathbb{E}[z_v(w)]\right]$$

Original finite sum problem

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

Controlled Stochastic Reformulation

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

Use covariates to **control the variance**

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

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

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

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

Variance reduction with arbitrary sampling

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

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

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

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

Variance reduction with arbitrary sampling

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_{v}(w) - z_{v}(w) + \mathbb{E}[z_{v}(w)]\right]$$

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

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

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

$$g_{v}(w) := \nabla f_{v}(w) - \nabla z_{v}(w) + \mathbb{E}[\nabla z_{v}(w)]$$

Variance reduction with arbitrary sampling

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

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

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

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

How to choose $z_v(w)$?

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

Choosing the covariate

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

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

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

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

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

We would like:

$$g_v(w) \approx \nabla f(w)$$

$$\nabla z_v(w) \approx \nabla f_v(w)$$

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:

$$g_v(w) \approx \nabla f(w)$$

$$\nabla z_v(w) \approx \nabla f_v(w)$$

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

Johnson & Zhang, 2013 NIPS

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

Reference point $\tilde{w} \in \mathbb{R}^d$

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

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

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Single element sampling

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t), \quad i \in \{1, \ldots, n\}$ uniformly
Grad. estimate	$g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Single element sampling

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

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

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

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

$$\nabla z_{v^t}(w^t) = \nabla f_i(\tilde{w})$$

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Single element sampling

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t), \quad i \in \{1, \ldots, n\}$ uniformly
Grad. estimate	$g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

$$z_{v^t}(w) = f_i(\tilde{w}) + \langle \nabla f_i(\tilde{w}), w - \tilde{w} \rangle$$

$$\nabla z_{v^t}(w^t) = \nabla f_i(\tilde{w})$$

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Single element sampling

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t), \quad i \in \{1, \ldots, n\}$ uniformly
Grad. estimate	$g_{v^t}(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

$$z_{v^t}(w) = f_i(\tilde{w}) + \langle \nabla f_i(\tilde{w}), w - \tilde{w} \rangle \Longleftarrow \nabla z_{v^t}(w^t) = \nabla f_i(\tilde{w}) \Longrightarrow \mathbb{E}[\nabla z_{v^t}(w^t)] = \nabla f(\tilde{w})$$

SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang, NIPS 2013

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 $\tilde{w}^T$

Sebbouh, Gazagnadou, Jelassi, Bach, Gower, 2019

SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang, NIPS 2013

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

Freeze reference point for m iterations

Sebbouh, Gazagnadou, Jelassi, Bach, Gower, 2019

SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang, NIPS 2013

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

Freeze reference point for m iterations

Weighted average of inner iterates

Sebbouh, Gazagnadou, Jelassi, Bach, Gower, 2019

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

Single element sampling

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

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

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.

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

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

Single element sampling

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

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

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

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

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

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

Grad. estimate:

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

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

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

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

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

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

Sample

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

Grad. estimate

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

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

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

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

Store grad.

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

SAGA: Stochastic Average Gradient

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

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 $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–strongly convex

stepsize $\gamma \leq \dfrac{1}{6\mathcal{L}}$ $\Rightarrow$ Iteration complexity $\approx O\left(\dfrac{\mathcal{L}}{\mu} \log\left(\dfrac{1}{\epsilon}\right)\right)$

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

Iteration complexity for SVRG and SAGA for arbitrary sampling

Theorem for SVRG $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–strongly convex

stepsize $\gamma \leq \frac{1}{6\mathcal{L}}$ $\Rightarrow$ Iteration complexity $\approx O\left(\frac{\mathcal{L}}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$

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

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

stepsize $\gamma \leq \frac{1}{4\mathcal{L}}$ $\Rightarrow$ Iteration complexity $\approx O\left(\frac{\mathcal{L}}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$

PDF Adobe G., Bach, Richtarik, 2018

Iteration complexity for SVRG and SAGA for arbitrary sampling

Theorem for SVRG $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–strongly convex

stepsize $\gamma \leq \dfrac{1}{6\mathcal{L}}$ ⇒ Iteration complexity $\approx O\left(\dfrac{\mathcal{L}}{\mu}\log\left(\dfrac{1}{\epsilon}\right)\right)$

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

Missing details due to extra definitions

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

stepsize $\gamma \leq \dfrac{1}{4\mathcal{L}}$ ⇒ Iteration complexity $\approx O\left(\dfrac{\mathcal{L}}{\mu}\log\left(\dfrac{1}{\epsilon}\right)\right)$

G., Bach, Richtarik, 2018

Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

$$\times \log\left(\frac{2}{\epsilon}\right)$$

$$C(b) \;=\; 2\left(\frac{n}{m} + 2b\right) \max\left\{\frac{3}{b}\,\frac{n-b}{n-1}\,\frac{L_{\max}}{\mu} + \frac{3n}{b}\,\frac{b-1}{n-1}\,\frac{L}{\mu}, m\right\}$$

$$\gamma = \frac{1}{6}\,\frac{b(n-1)}{(n-b)L_{\max} + n(b-1)L}$$

Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

$$C(b) = \underbrace{2\left(\frac{n}{m} + 2b\right) \max\left\{\frac{3}{b}\frac{n-b}{n-1}\frac{L_{\max}}{\mu} + \frac{3n}{b}\frac{b-1}{n-1}\frac{L}{\mu}}_{\text{Non-linearly increasing}}, m\right\} \times \log\left(\frac{2}{\epsilon}\right)$$

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

Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Linearly decreasing

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

Non-linearly increasing

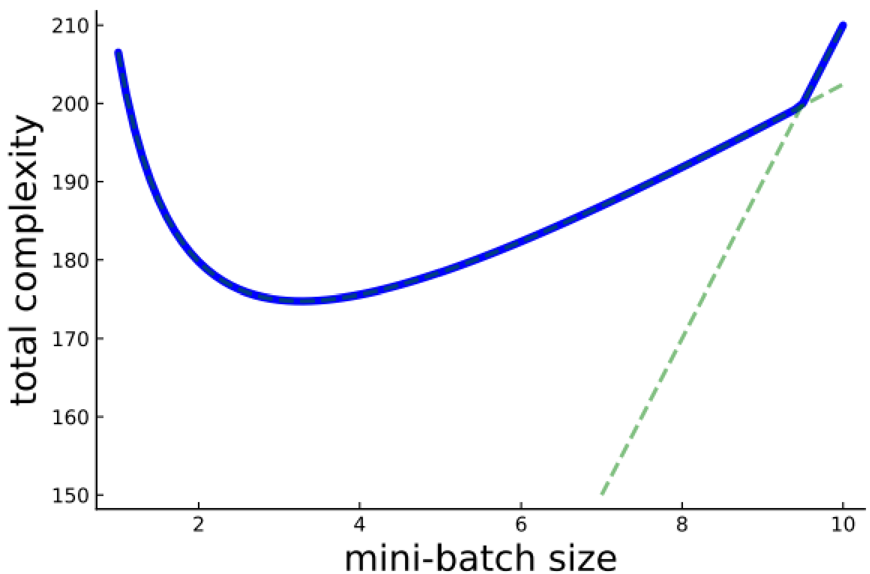
$$\gamma = \frac{1}{6}\frac{b(n-1)}{(n-b)L_{\max}+n(b-1)L}$$

Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

$$C(b) = \underbrace{\overbrace{2\left(\frac{n}{m}+2b\right)}^{\text{Linearly decreasing}} \max\left\{\frac{3}{b}\frac{n-b}{n-1}\frac{L_{\max}}{\mu} + \frac{3n}{b}\frac{b-1}{n-1}\frac{L}{\mu}\right.}_{\text{Non-linearly increasing}}, m\left.\vphantom{\frac{3}{b}}\right\} \times \log\left(\frac{2}{\epsilon}\right)$$

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



Total Complexity of mini-batch SVRG

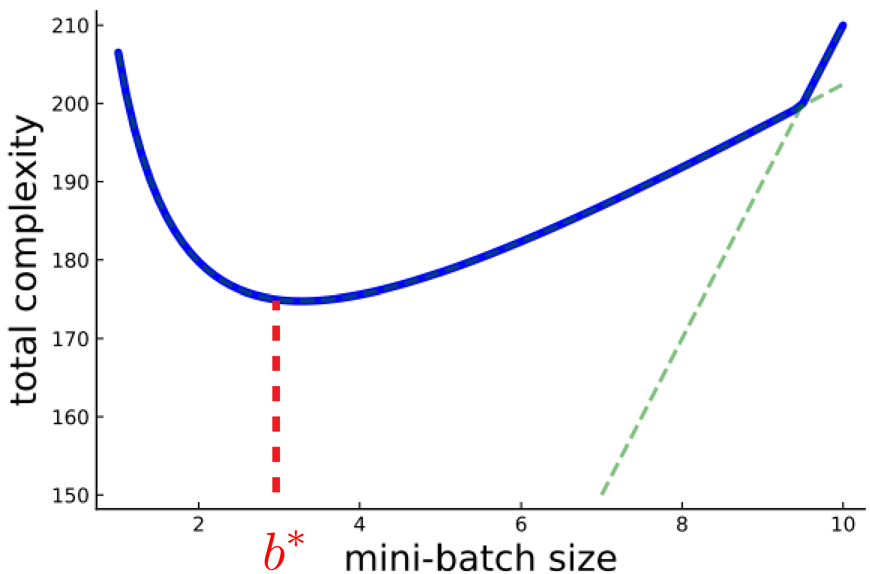
Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Linearly decreasing

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

Non-linearly increasing

$$\gamma = \frac{1}{6}\,\frac{b(n-1)}{(n-b)L_{\max} + n(b-1)L}$$



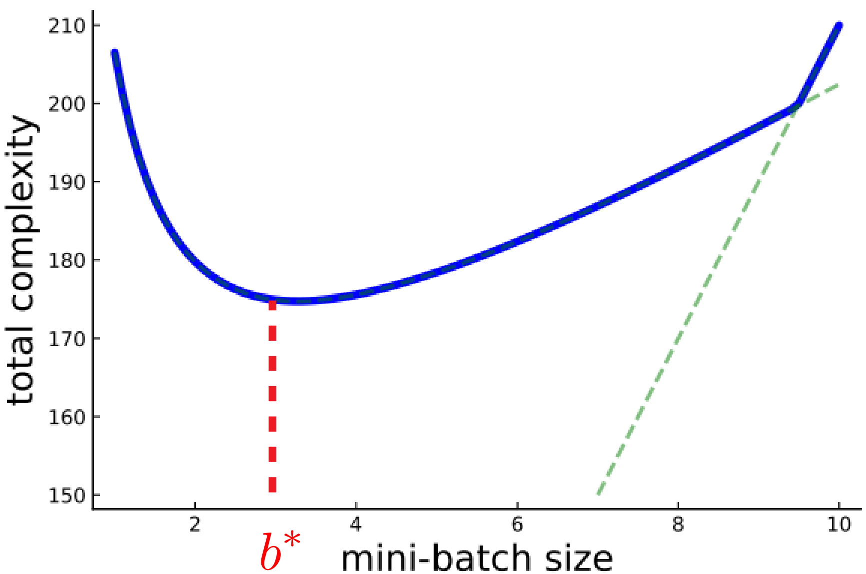
Total Complexity of mini-batch SVRG

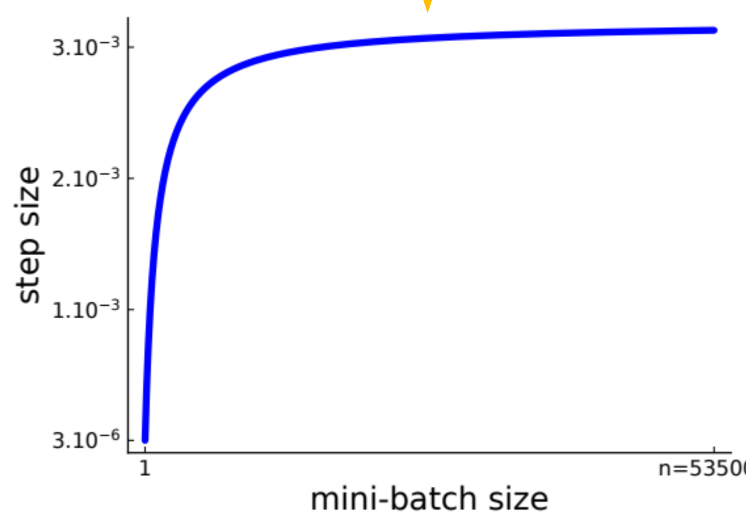
Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

$$C(b) \;=\; \underbrace{\overbrace{2\left(\frac{n}{m}+2b\right)}^{\text{Linearly decreasing}} \max\left\{\frac{3}{b}\frac{n-b}{n-1}\frac{L_{\max}}{\mu}+\frac{3n}{b}\frac{b-1}{n-1}\frac{L}{\mu}, m\right\}}_{\text{Non-linearly increasing}} \times \log\left(\frac{2}{\epsilon}\right)$$

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

Stepsize increasing with b



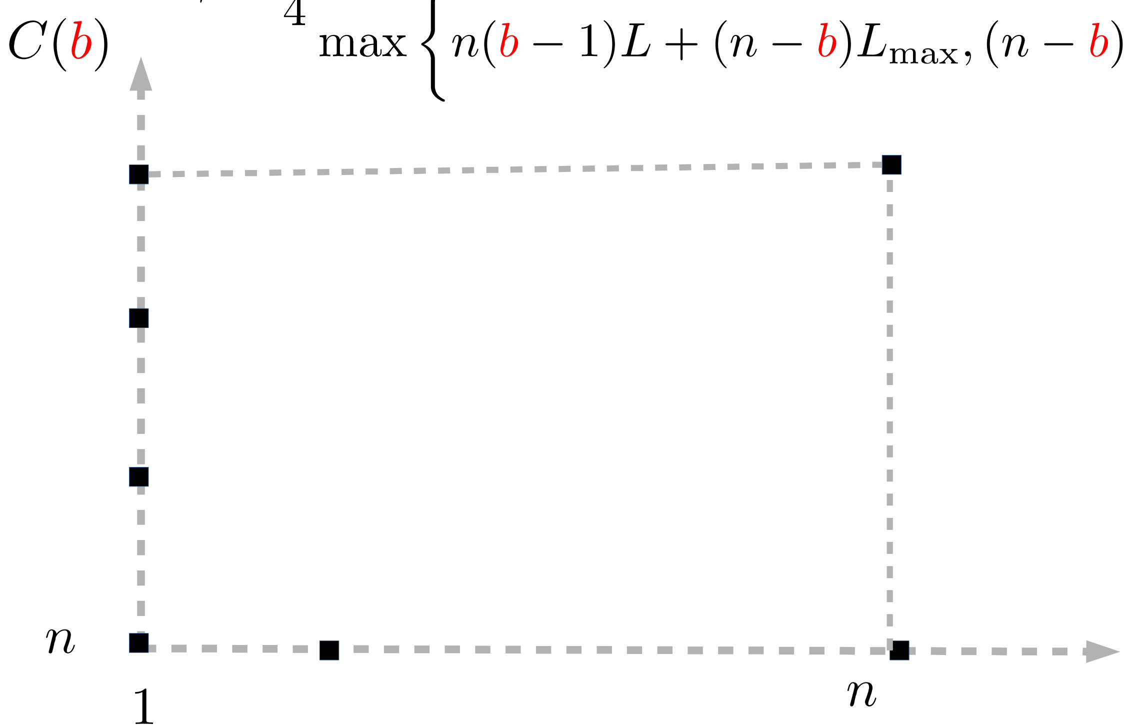


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

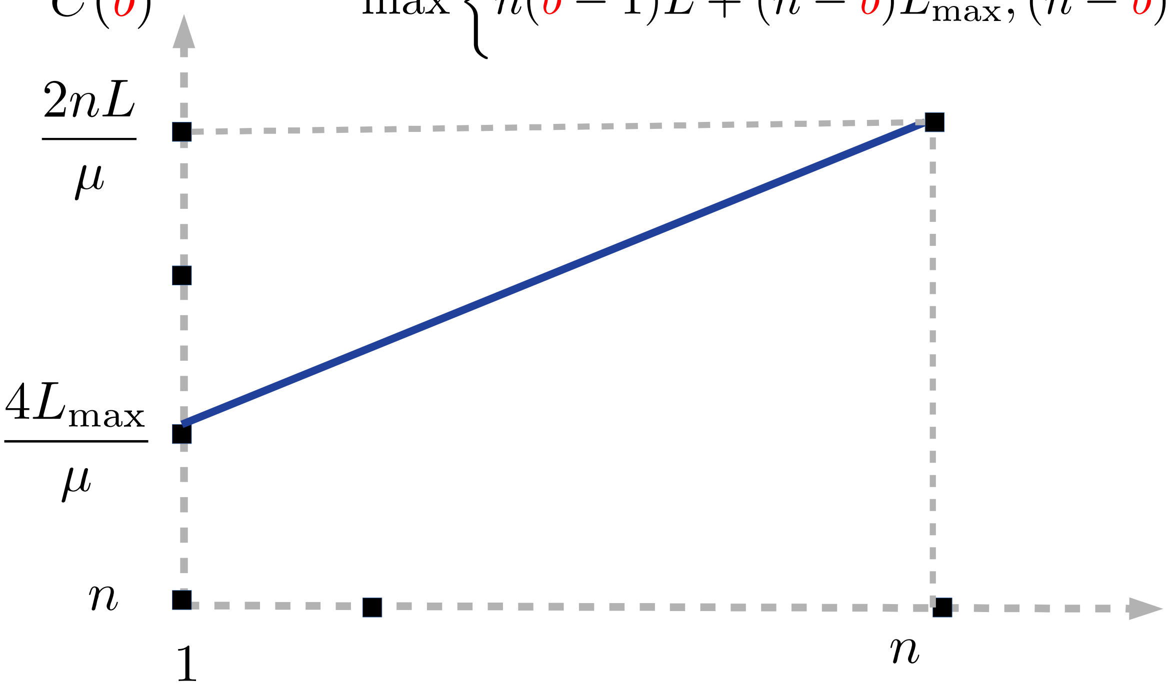


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

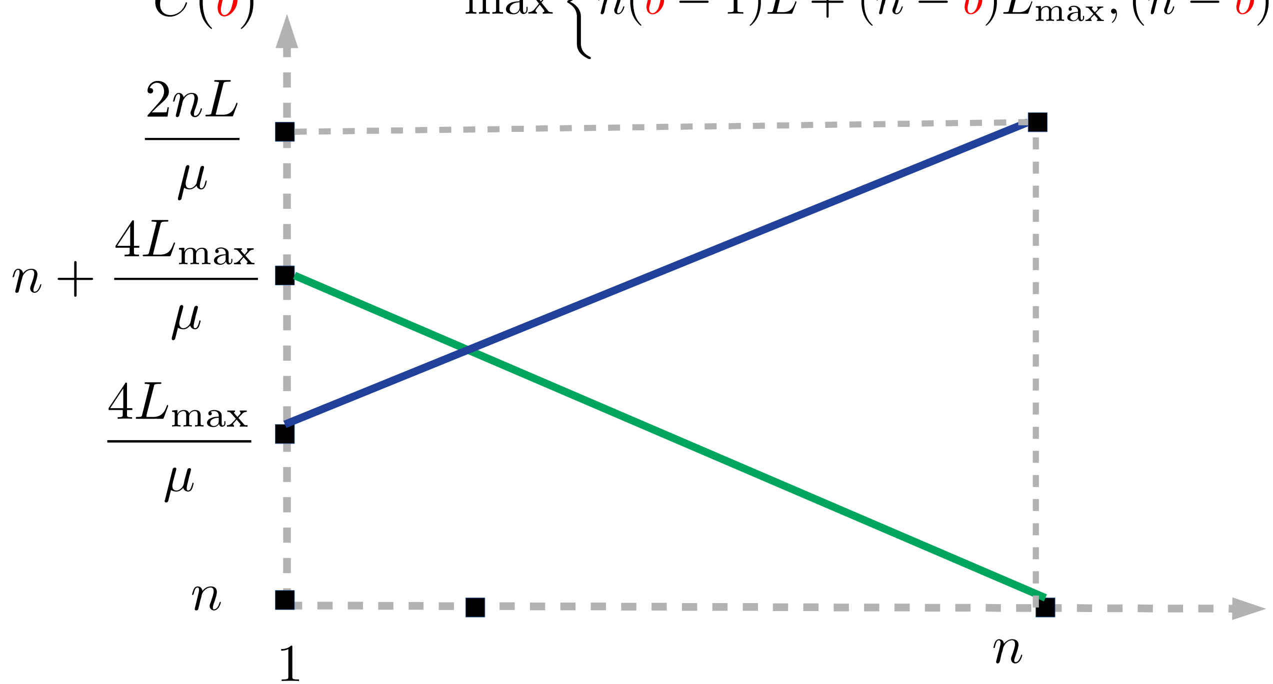


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

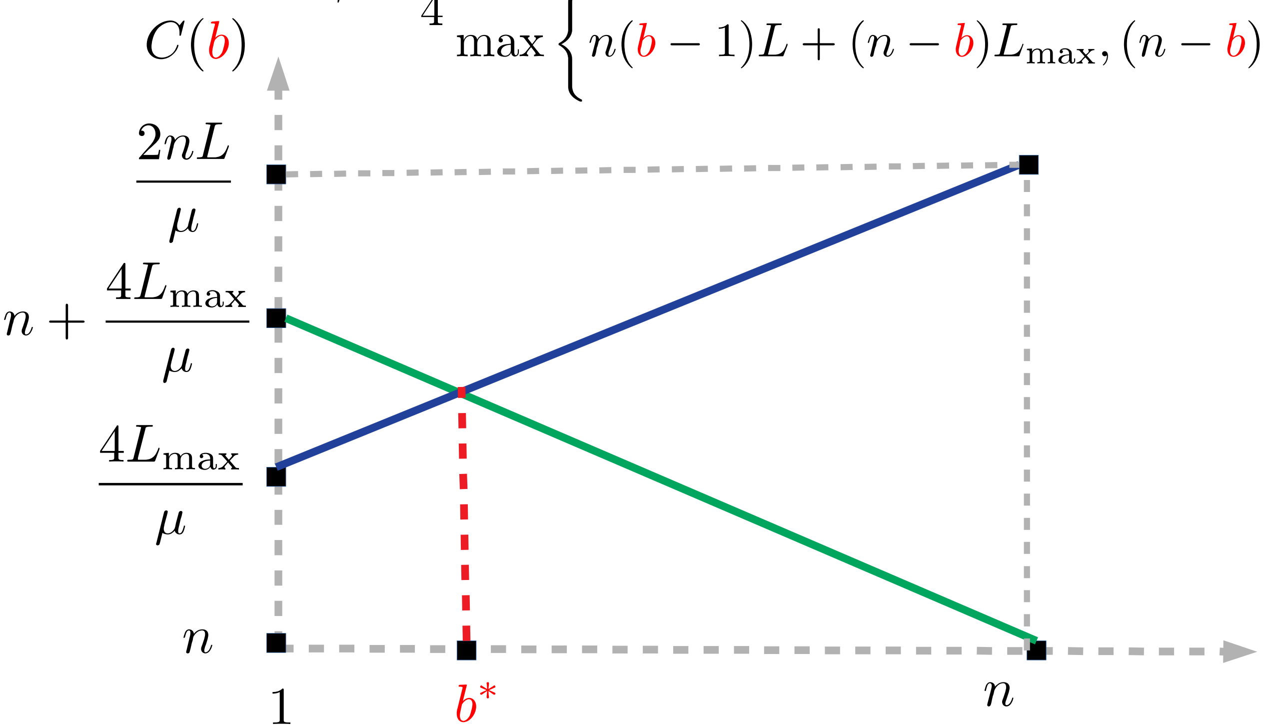


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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



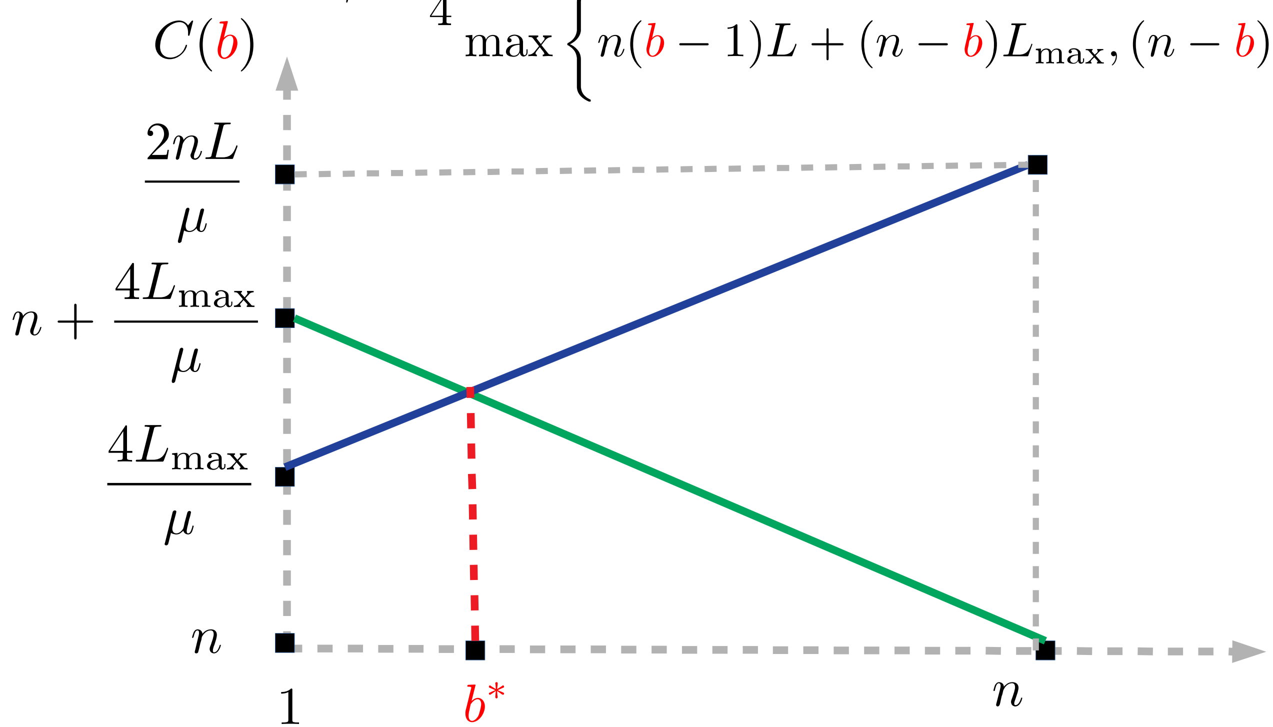
$$b^* = \left\lfloor 1 + \frac{\mu(n-1)}{4L} \right\rfloor$$

Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

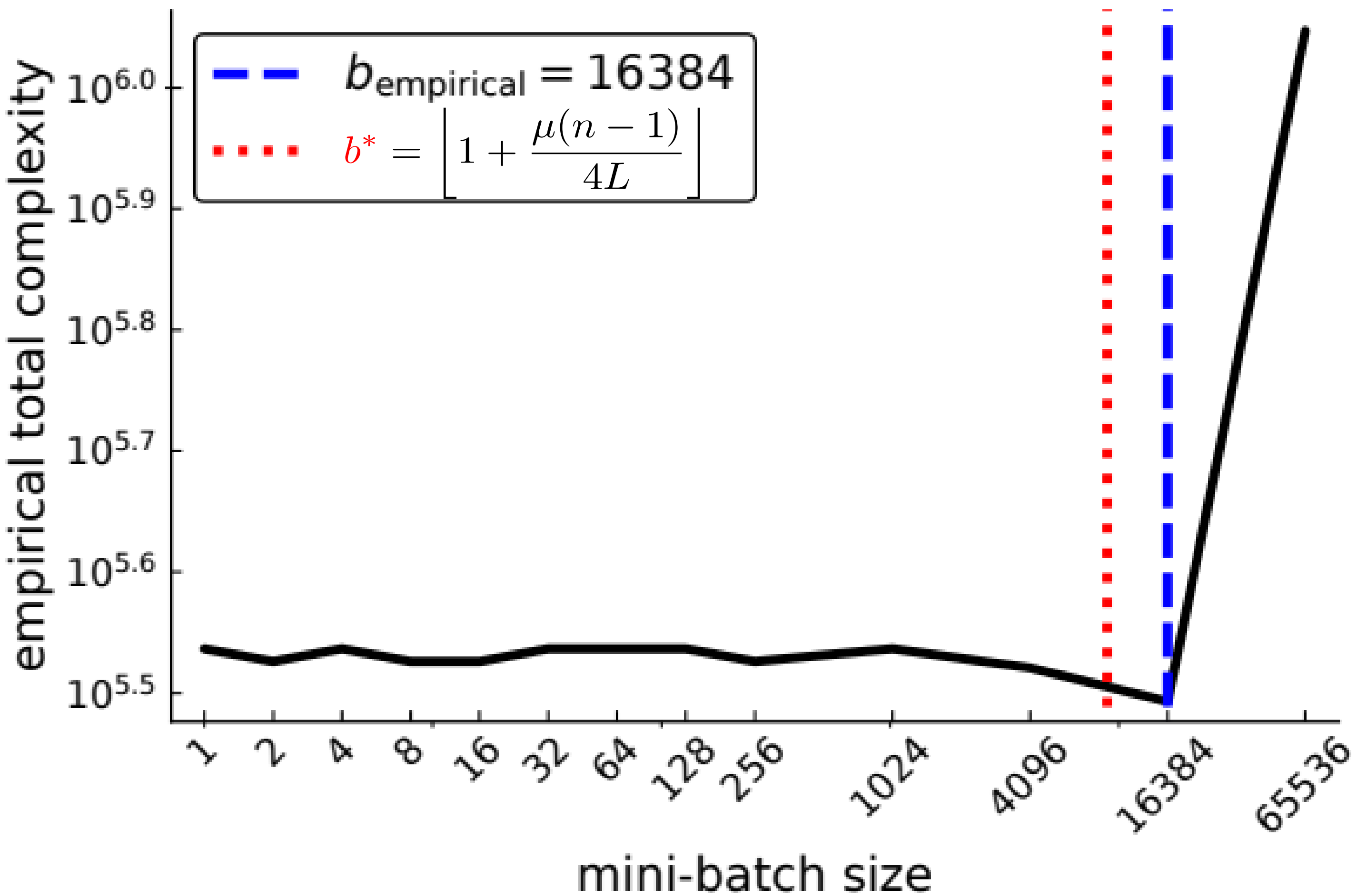
$$\gamma = \frac{1}{4}\frac{b(n-1)}{\max\left\{ n(b-1)L + (n-b)L_{\max},\ (n-b)L_{\max} + \frac{n(n-1)\mu}{4} \right\}}$$



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

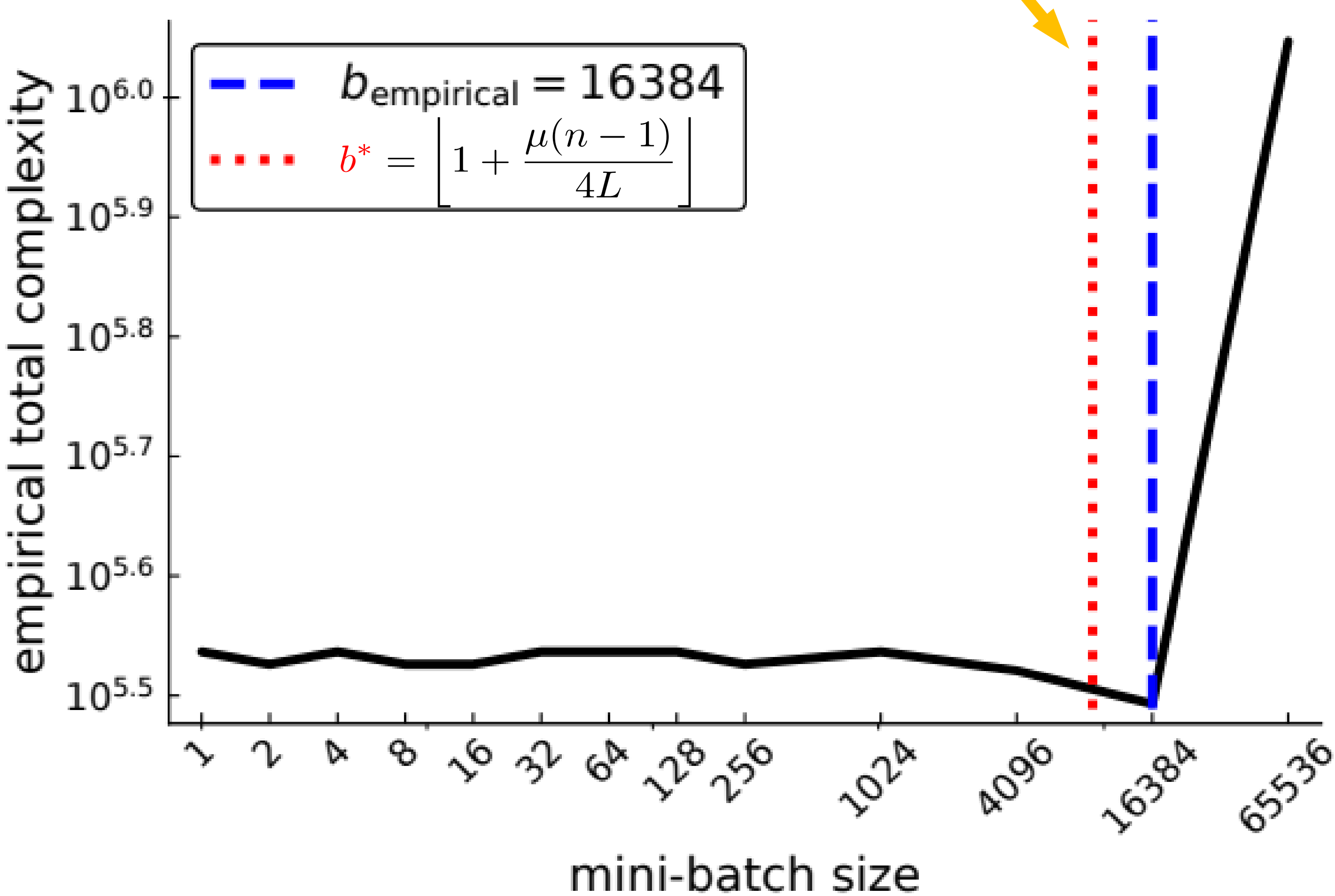
Always smaller than 25% of data

Total Complexity of mini-batch SAGA



Total Complexity of mini-batch SAGA

So accurate, close to empirical best mini-batch size



Take home message

Stochastic reformulations allow to view all variants as simple SGD

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

To analyse all forms of sampling used through expected smooth

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

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

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

Stepsize increase by orders when mini-batch size increases

Take home message

Stochastic reformulations allow to view all variants as simple SGD

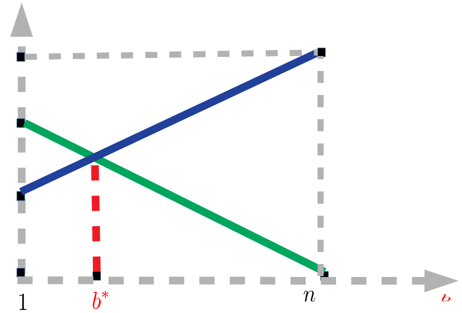
$$\min_{w\in\mathbf{R}^d} \mathbb{E}\left[f_v(w) := \frac{1}{n}\sum_{i=1}^n v_i f_i(w)\right]$$

To analyse all forms of sampling used through expected smooth

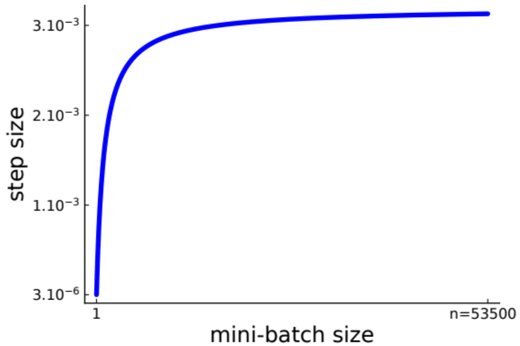
$$\mathbb{E}[||\nabla f_v(w) - \nabla f_v(w^*)||_2^2] \leq \mathcal{L}\left(f(w) - f(w^*)\right)$$

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

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



Stepsize increase by orders when mini-batch size increases



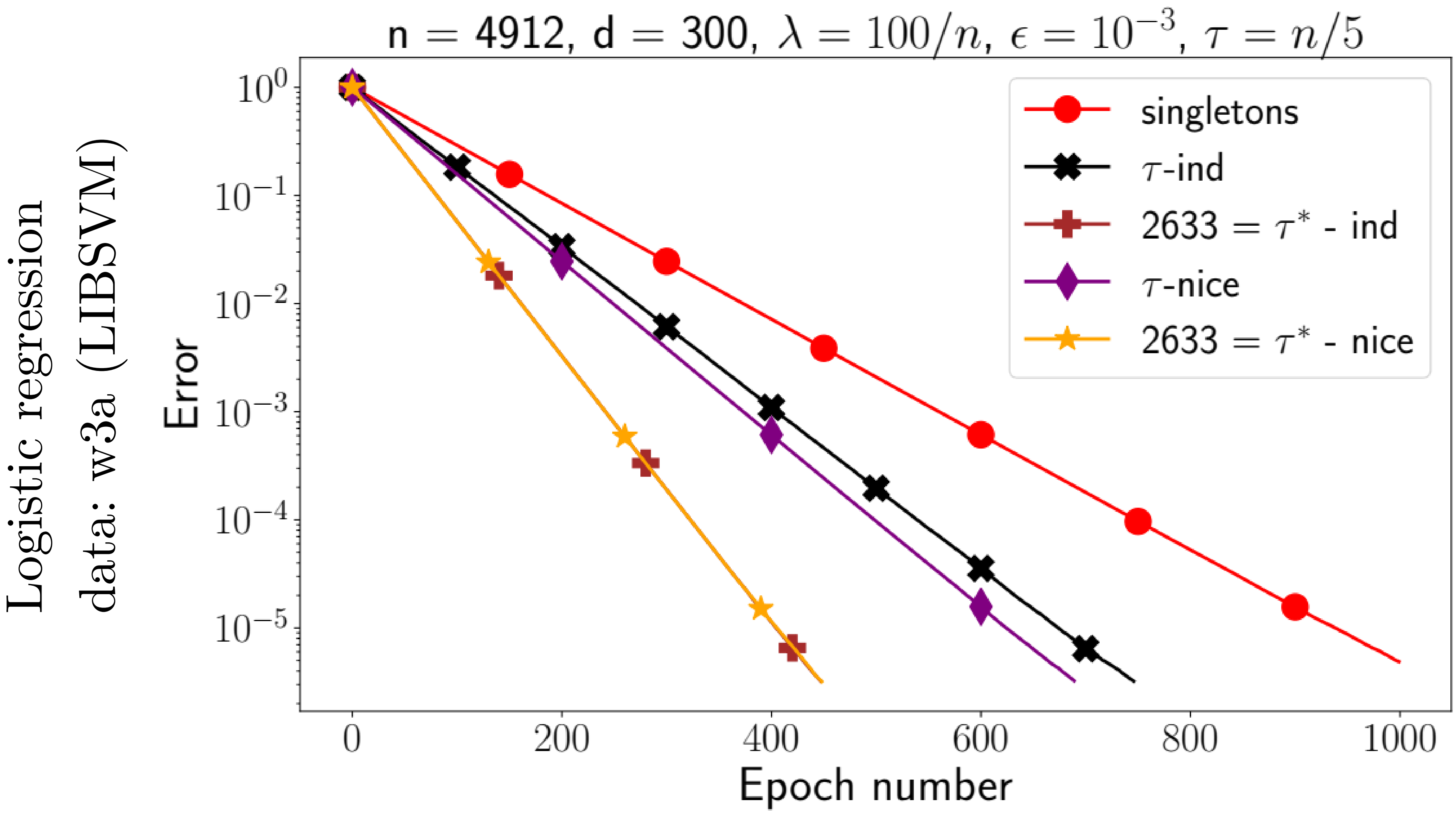
RMG, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin and Peter Richtárik (2019), ICML
SGD: general analysis and improved rates

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

N. Gazagnadou, RMG, J. Salmon (2019) , ICML 2019.
Optimal mini-batch and step sizes for SAGA

O. Sebbouh, N. Gazagnadou, S. Jelassi, F. Bach, RMG (2019), preprint online. **Towards closing the gap between the theory and practice of SVRG**

Optimal mini-batch size



Learning rate schedules

Main Theorem (Linear convergence to a neighborhood)

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and μ–quasi strongly convex

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

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

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!