

# 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**.

Google Research New York, 16<sup>th</sup> of July, 2019

# Optimization in Machine Learning

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

# Optimization in Machine Learning

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

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

$$h_w \left( \begin{array}{c} \text{Cat} \\ \text{Image} \end{array} \right) \mapsto \text{Cat}$$

# Optimization in Machine Learning

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

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

$$h_w \left( \begin{array}{c} \text{Cat} \\ \text{Image of a cat} \end{array} \right) \mapsto \text{Cat}$$

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

# Optimization in Machine Learning

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

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

$$h_w \left( \begin{array}{c} \text{Cat} \\ \text{Image of a cat} \end{array} \right) \mapsto \text{Cat}$$

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

(4) Solve the *training problem*:

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

# Optimization in Machine Learning

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

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

$$h_w \left( \begin{array}{c} \text{[Image of a cat]} \\ \hline \end{array} \right) \mapsto \text{Cat}$$

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

(4) Solve the *training problem*:

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

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

# Optimization in Machine Learning

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

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

$h_w\left(\begin{array}{c} \text{Cat} \\ \text{Image of a cat} \end{array}\right) \mapsto \text{Cat}$

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

(4) Solve the *training problem*: Optimization

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

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

# Finite sum minimization

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

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

## Mission statement:

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

# Finite sum minimization

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

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

## Mission statement:

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

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

# Finite sum minimization

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

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

## Mission statement:

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

*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.

# Finite sum minimization

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

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

## Mission statement:

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

*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  
 $j \in \{1, \dots, n\}$

# 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  
 $j \in \{1, \dots, n\}$

# 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)

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

Minibatch where  
 $B \in \{1, \dots, n\}$  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  
 $B \in \{1, \dots, n\}$  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?
- How does  $b$  influence the stepsizes?

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

# The Stochastic Gradient Method

Solving the *training problem*:

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

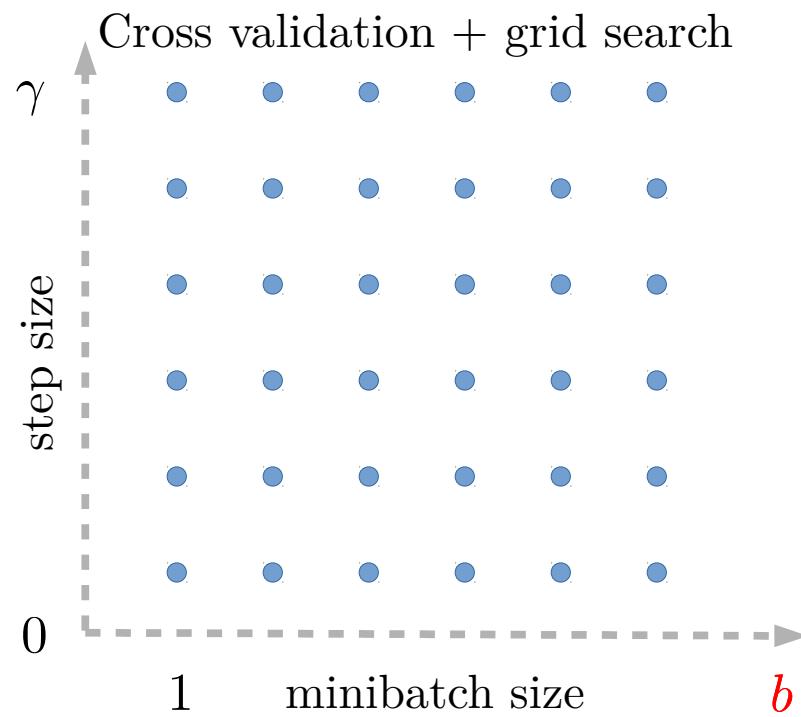
Baseline method: Stochastic Gradient Descent (SGD)

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

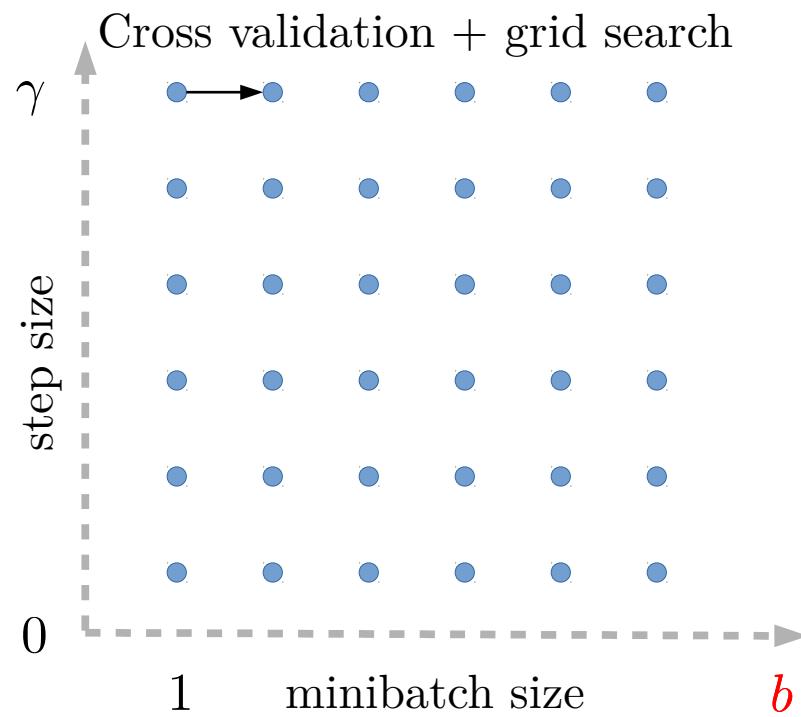
- What should  $b$  be?
- How does  $b$  influence the stepsizes?
- How does the data influence the best mini-batch and stepsize?

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

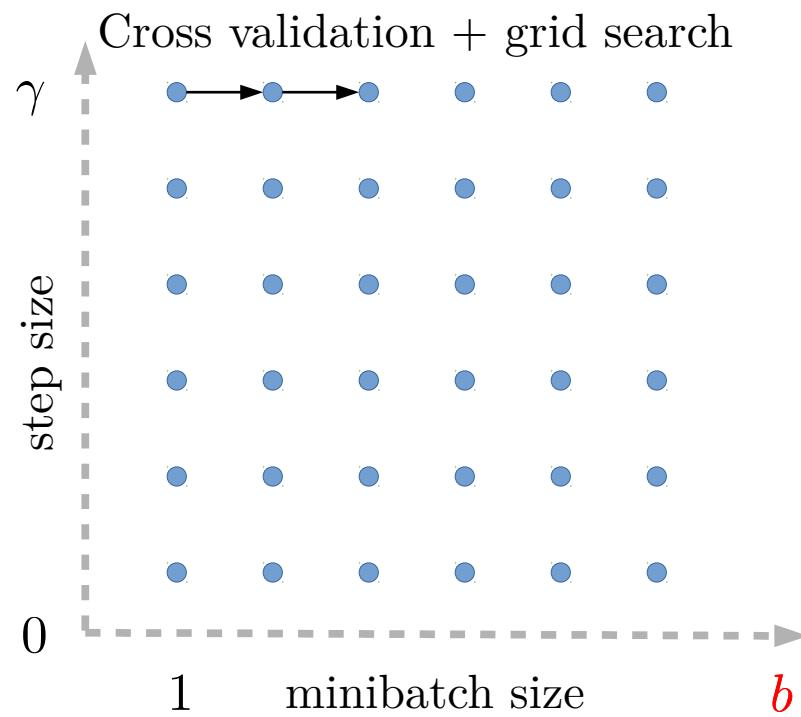
# How to choose the minibatch size?



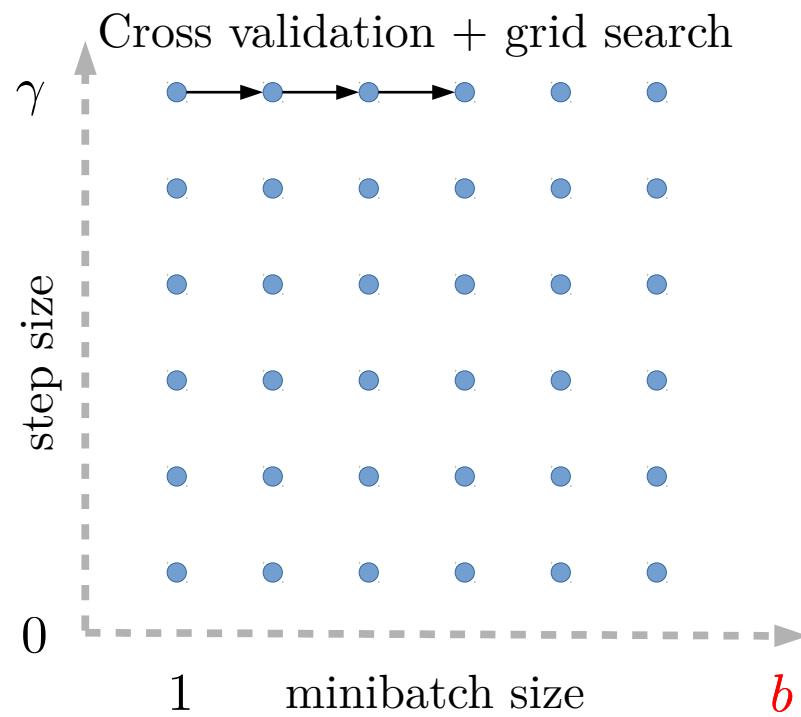
# How to choose the minibatch size?



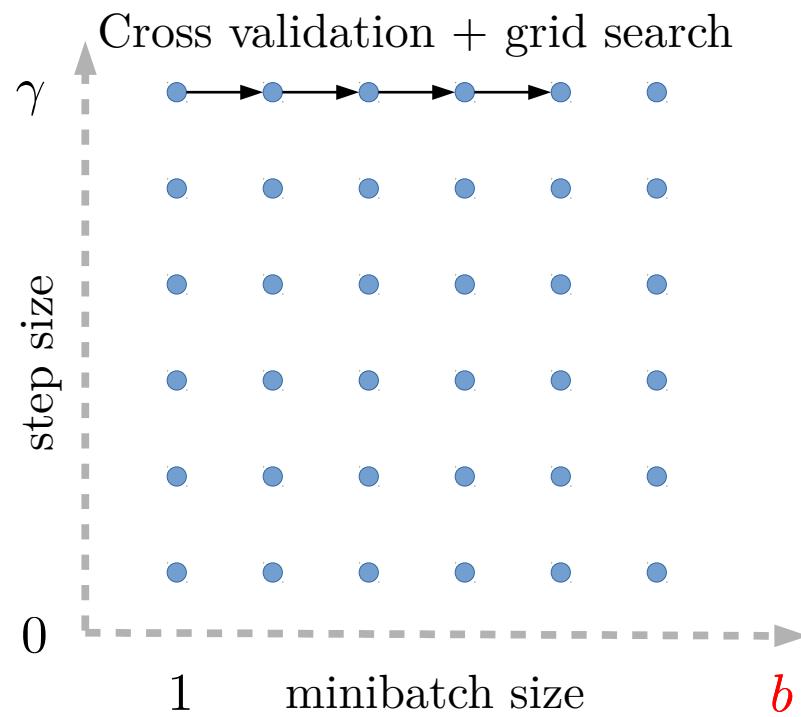
# How to choose the minibatch size?



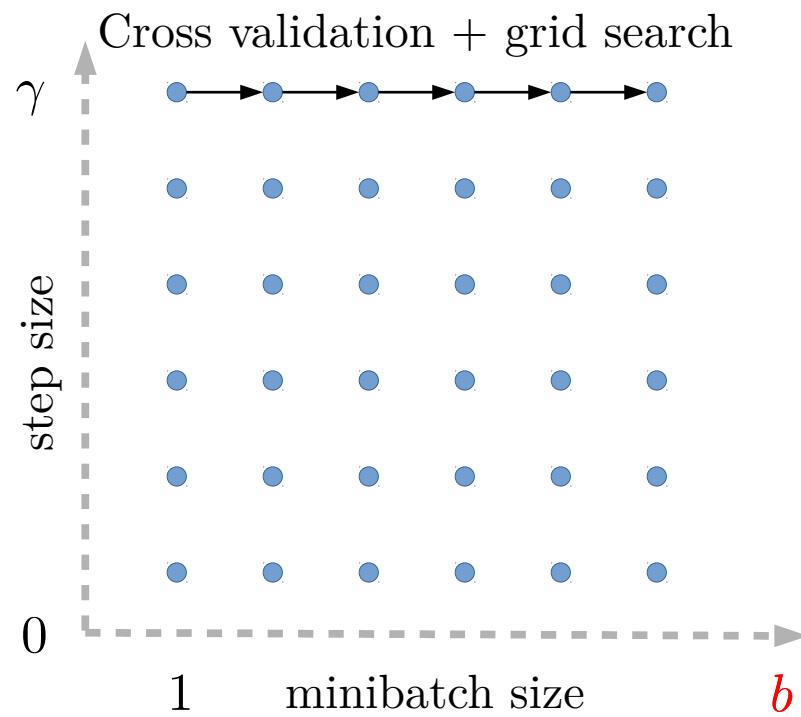
# How to choose the minibatch size?



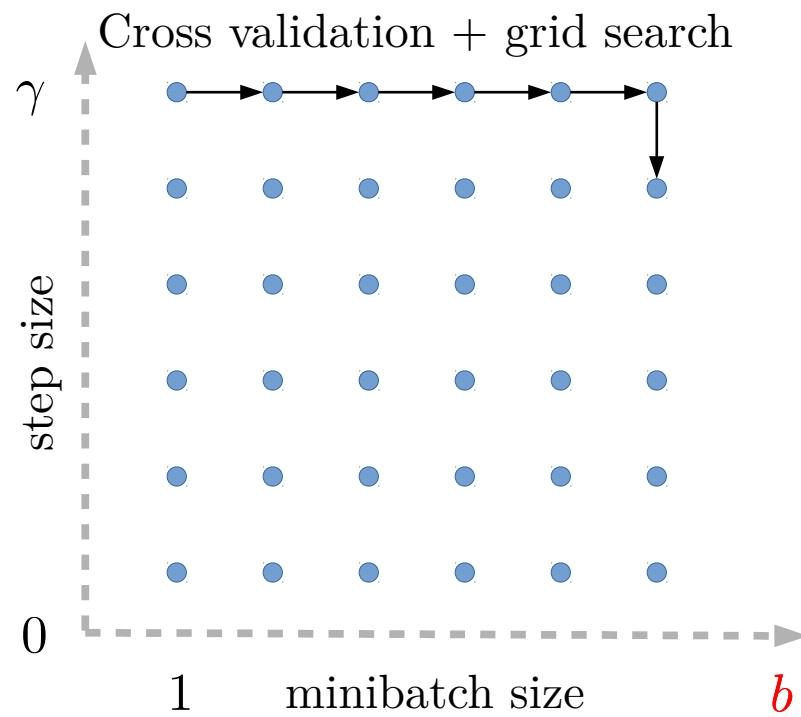
# How to choose the minibatch size?



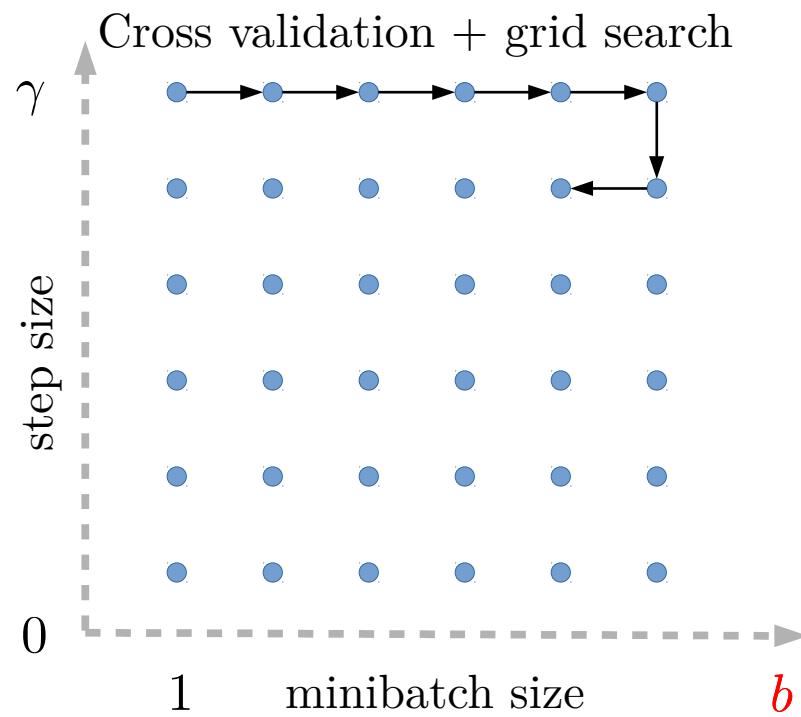
# How to choose the minibatch size?



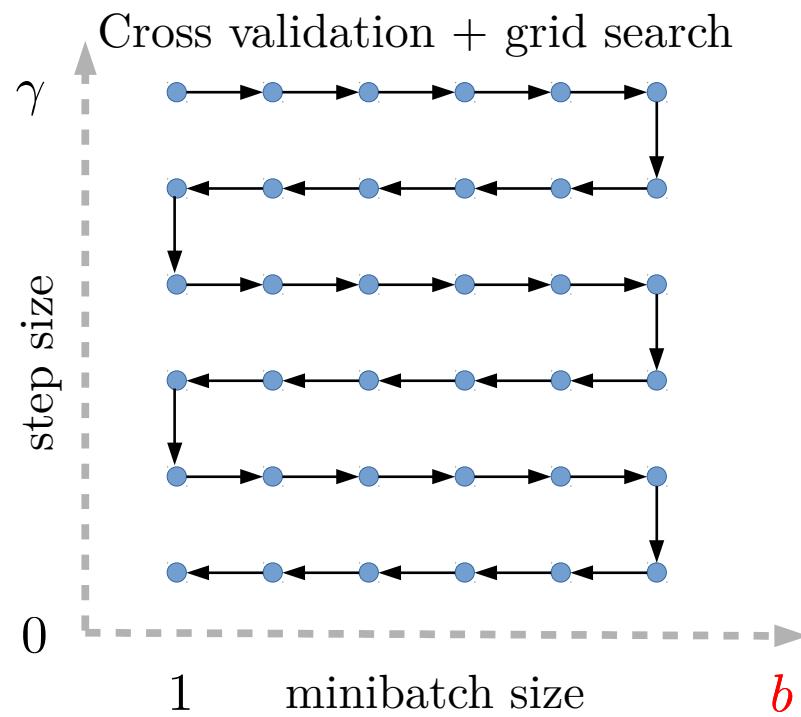
# How to choose the minibatch size?



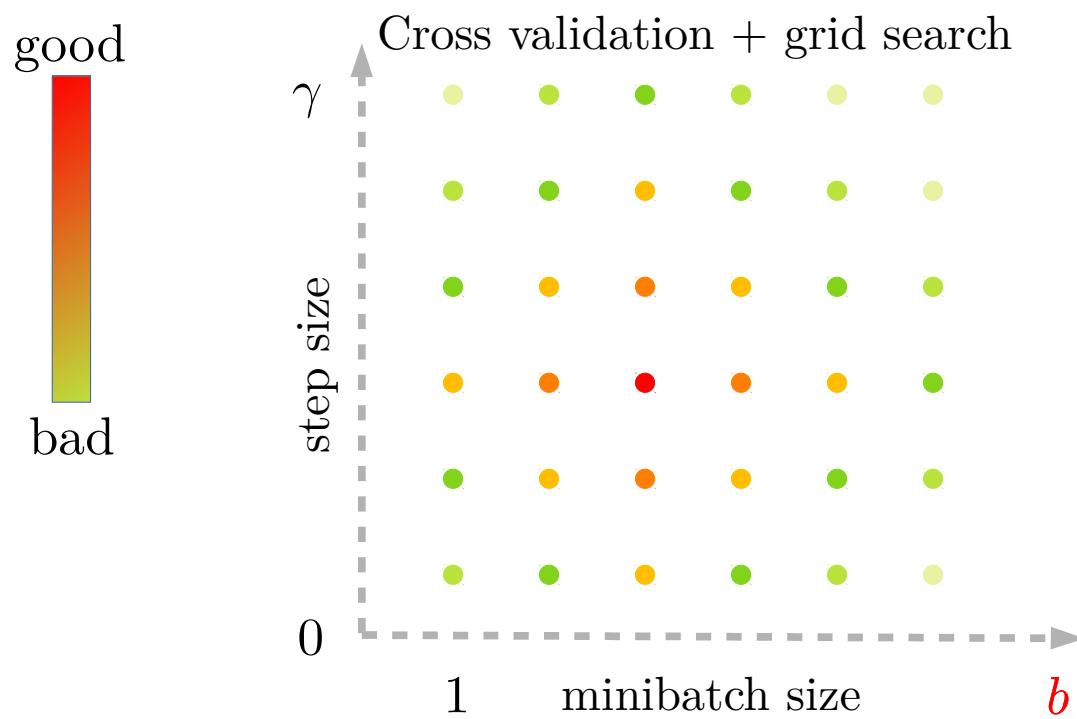
# How to choose the minibatch size?



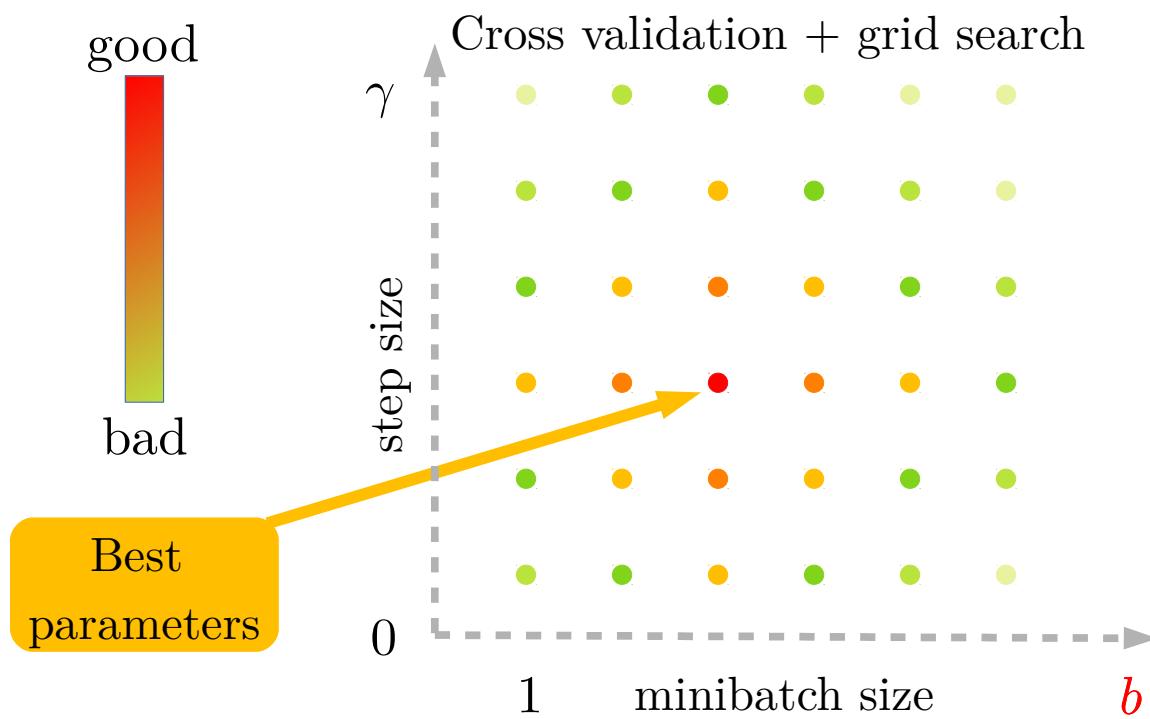
# How to choose the minibatch size?



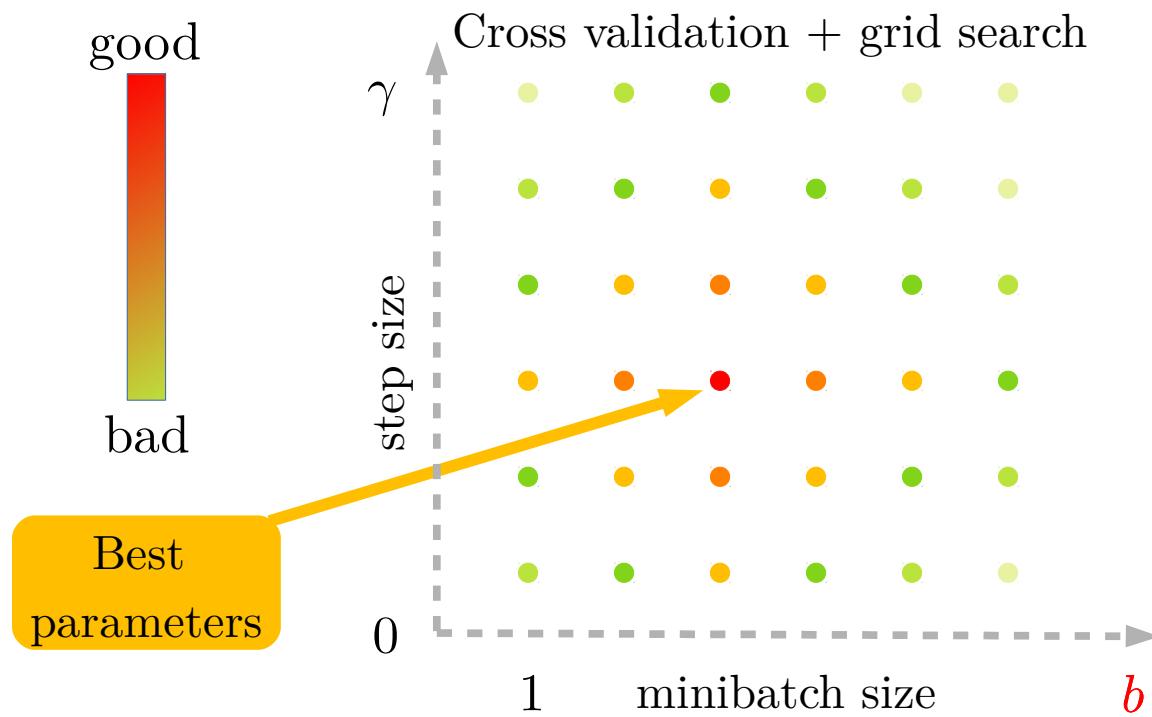
# How to choose the minibatch size?



# How to choose the minibatch size?



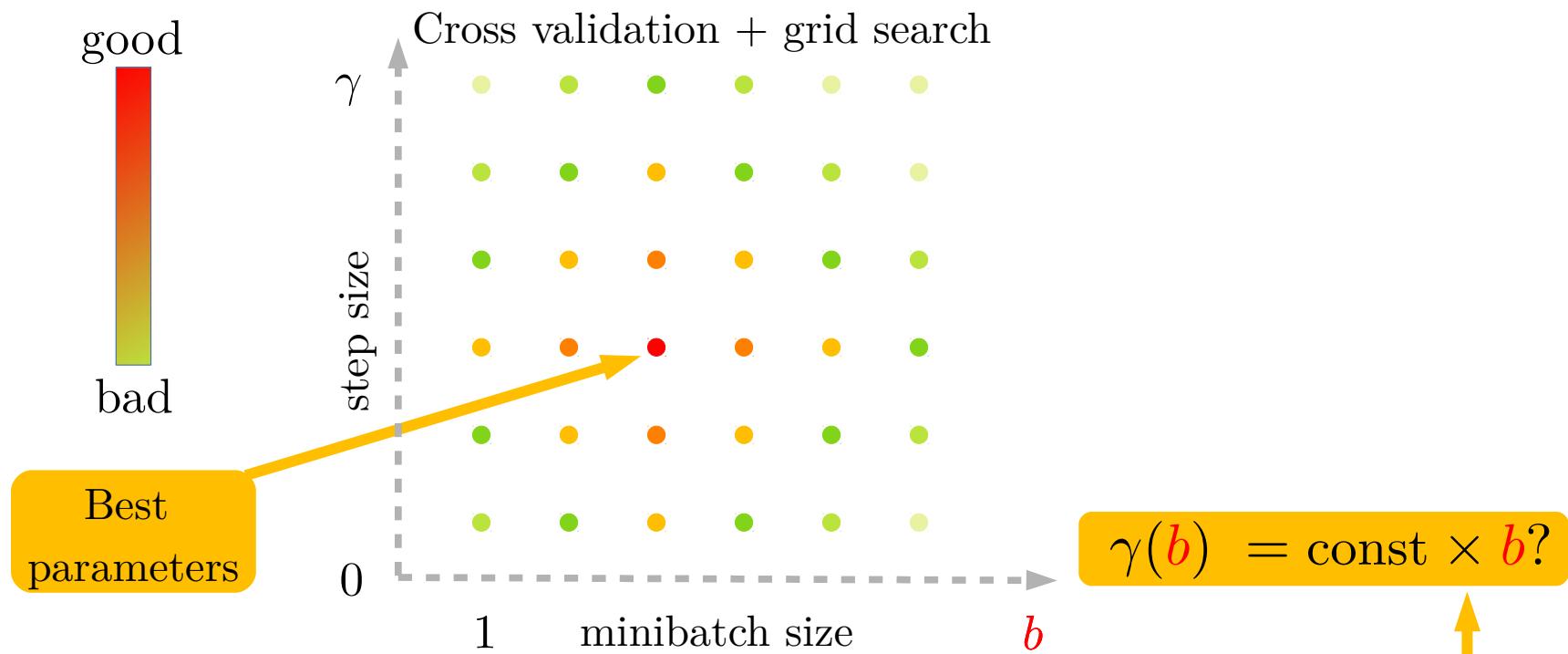
# How to choose the minibatch size?



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  $k$ .

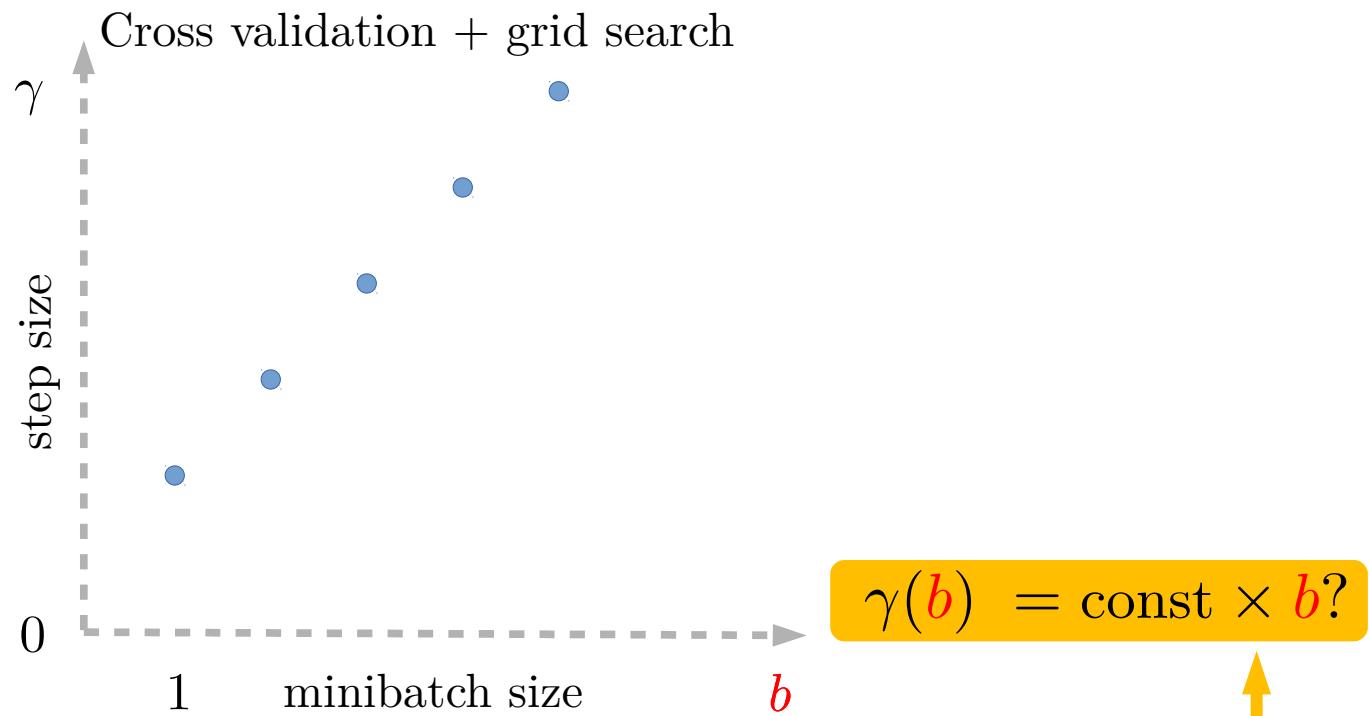
# How to choose the minibatch size?



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  $k$ .

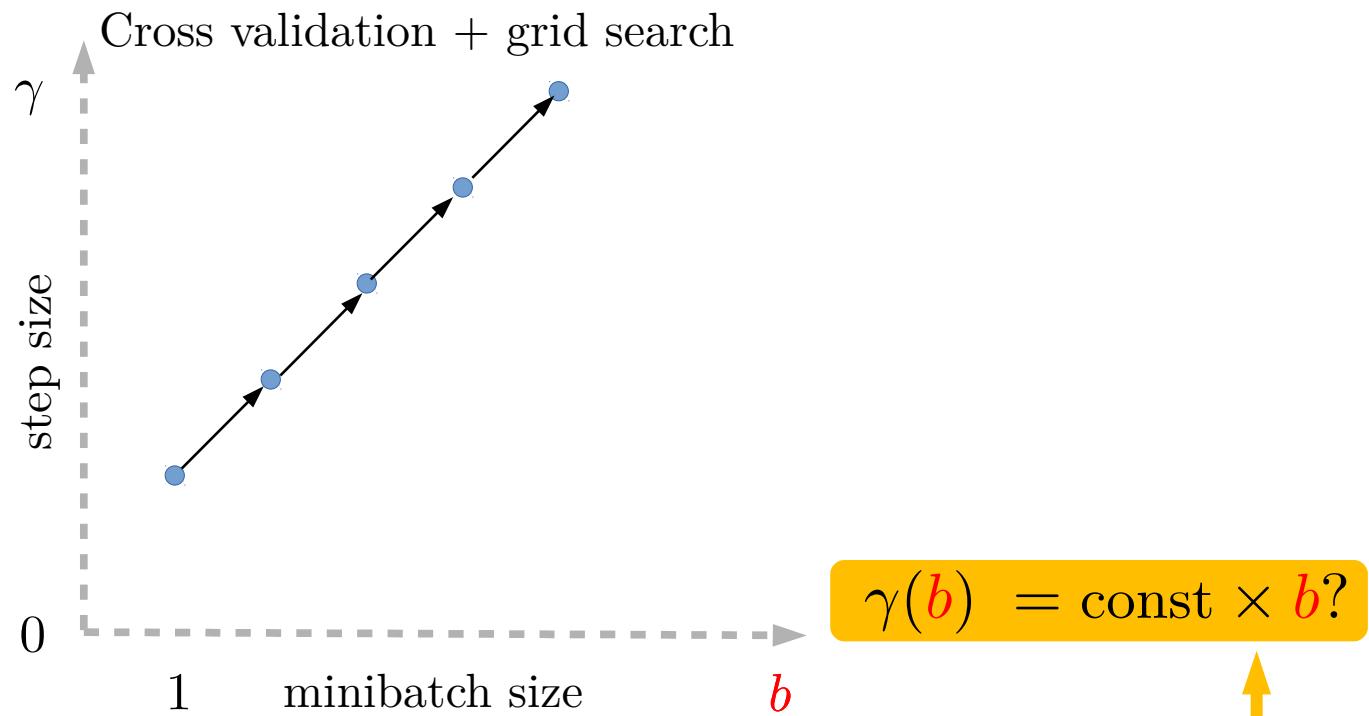
# How to choose the minibatch size?



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  $k$ .

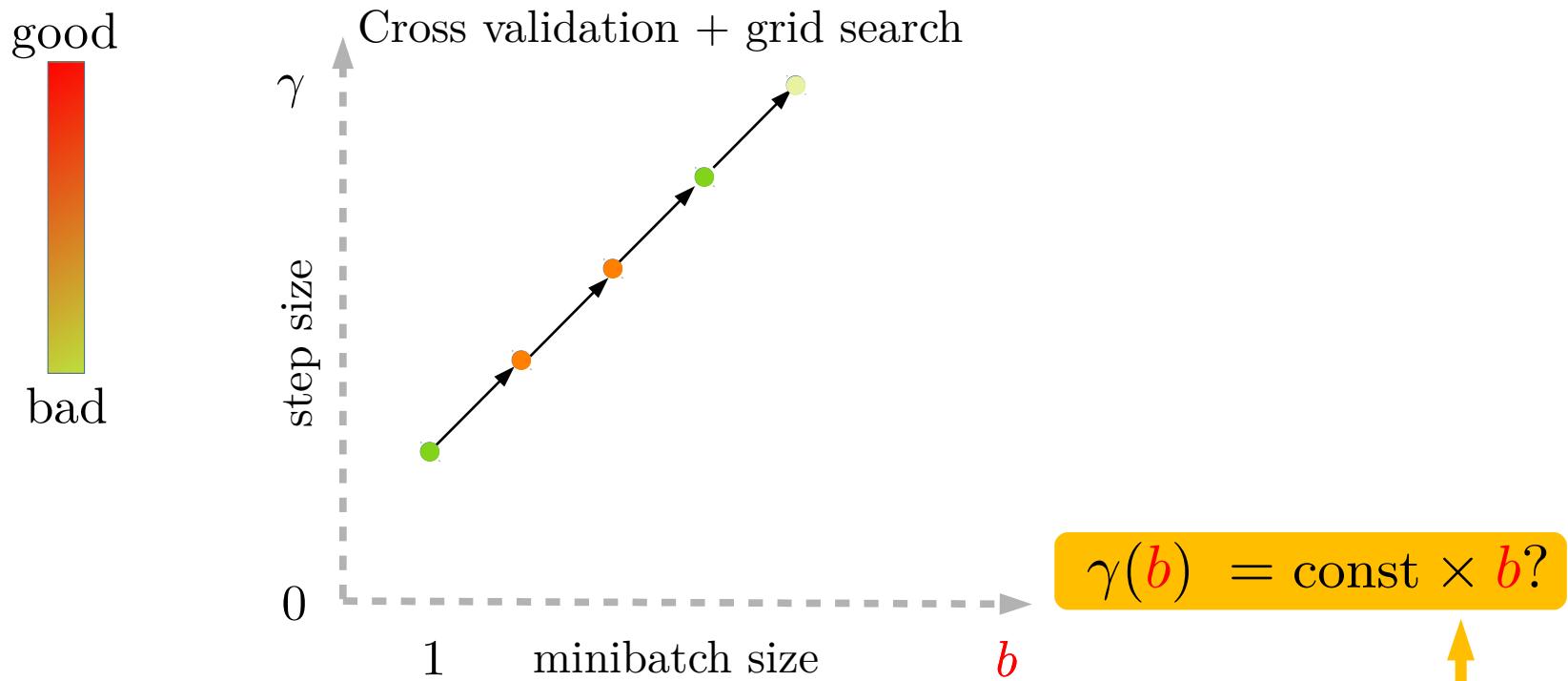
# How to choose the minibatch size?



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  $k$ .

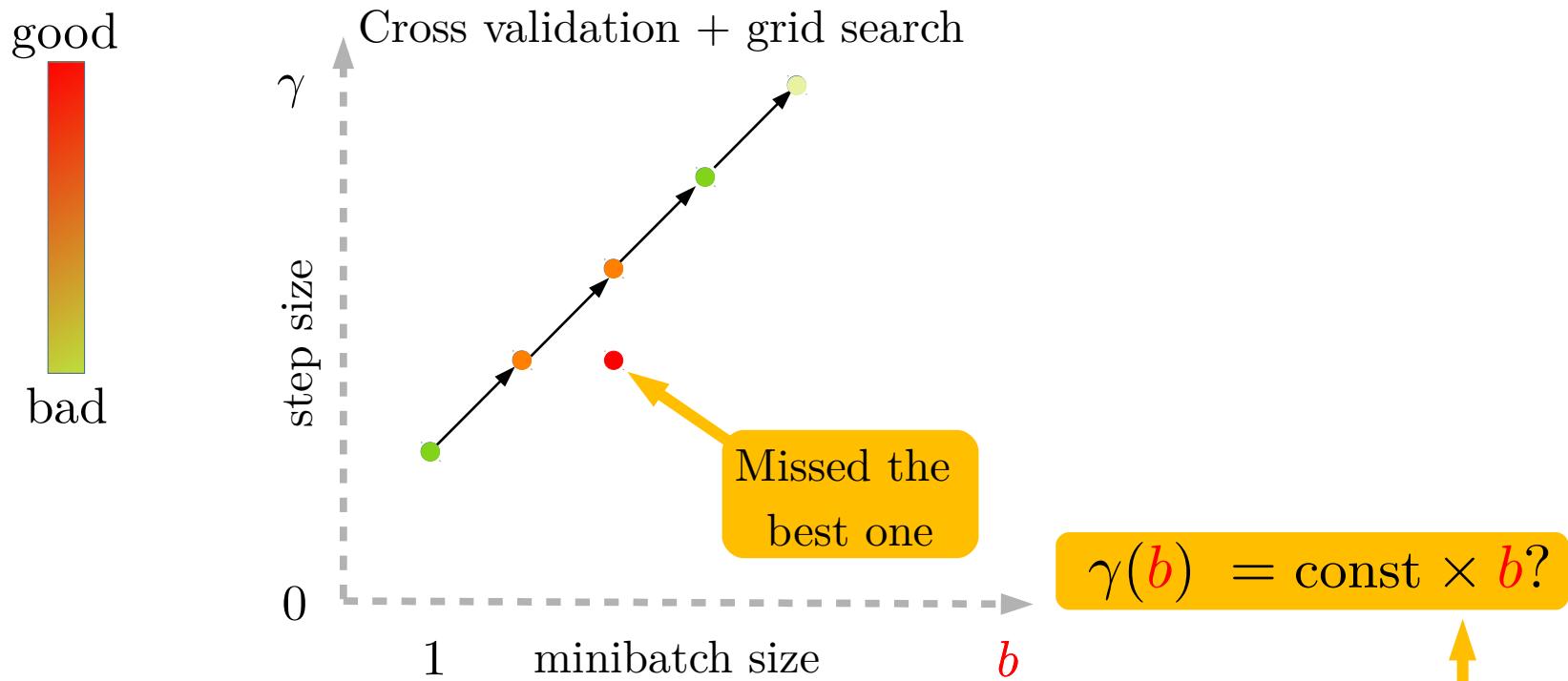
# How to choose the minibatch size?



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  $k$ .

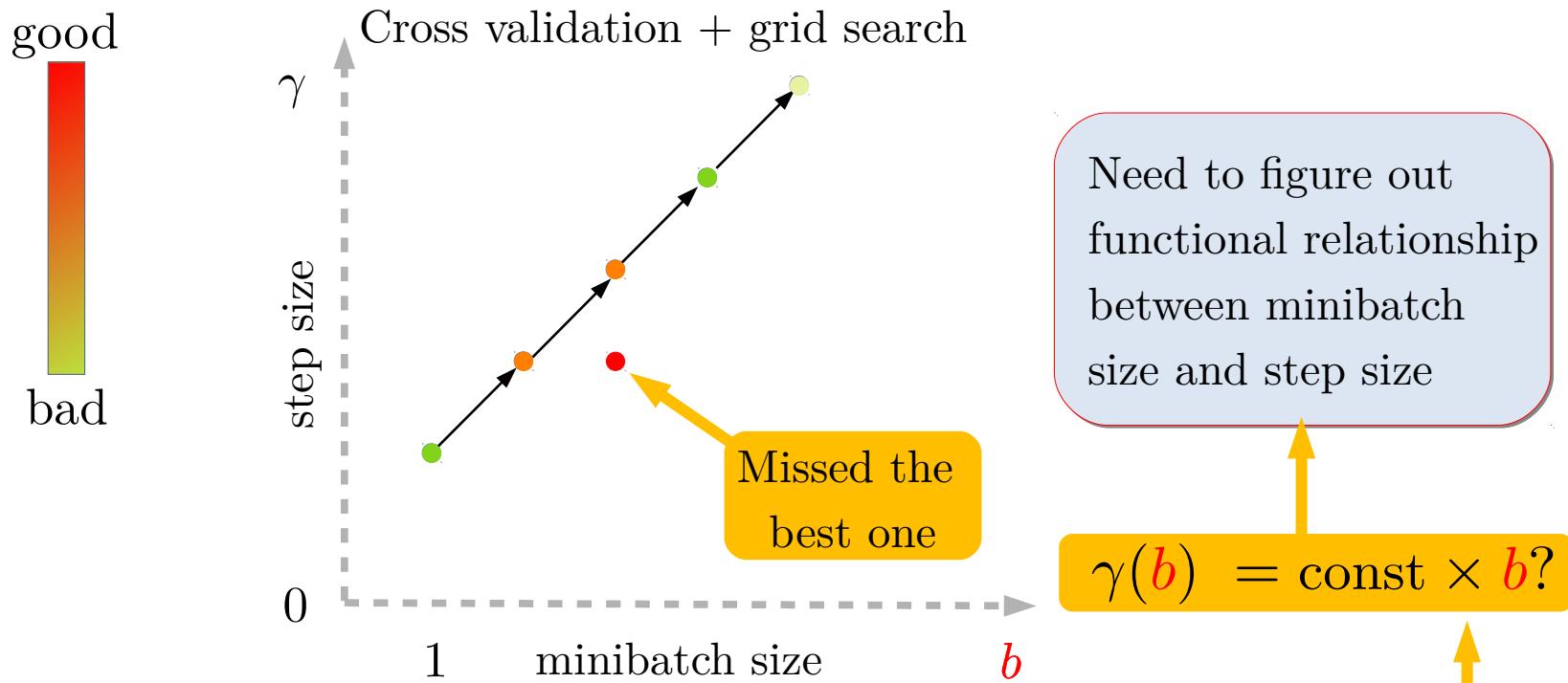
# How to choose the minibatch size?



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?



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  $\textcolor{red}{v} = (\textcolor{red}{v}_1, \dots, \textcolor{red}{v}_n) \in \mathbb{R}^n$  with

$$\mathbb{E}[\textcolor{red}{v}_i] = 1, \quad \text{for } i = 1, \dots, n$$

# Simple Stochastic Reformulation

Random sampling vector  $\textcolor{red}{v} = (\textcolor{red}{v}_1, \dots, \textcolor{red}{v}_n) \in \mathbb{R}^n$  with

$$\mathbb{E}[\textcolor{red}{v}_i] = 1, \quad \text{for } i = 1, \dots, n$$

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

# Simple Stochastic Reformulation

Random sampling vector  $\textcolor{red}{v} = (\textcolor{red}{v}_1, \dots, \textcolor{red}{v}_n) \in \mathbb{R}^n$  with

$$\mathbb{E}[\textcolor{red}{v}_i] = 1, \quad \text{for } i = 1, \dots, n$$

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

# Simple Stochastic Reformulation

Random sampling vector  $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_n) \in \mathbb{R}^n$  with

$$\mathbb{E}[\mathbf{v}_i] = 1, \quad \text{for } i = 1, \dots, n$$

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

Original finite sum problem

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



Stochastic Reformulation

$$\min_{w \in \mathbb{R}^d} \mathbb{E} [f_{\mathbf{v}}(w)]$$

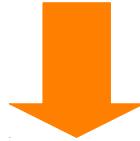
Minimizing the expectation of **random linear combinations** of original function

# SGD with arbitrary sampling

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

# SGD with arbitrary sampling

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



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

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

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

# SGD with arbitrary sampling

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



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

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

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

$$\mathbf{v} \equiv (1, \dots, 1) \rightarrow w^{t+1} = w^t - \gamma_t \nabla f(w^t)$$

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

# SGD with arbitrary sampling

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



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

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

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

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

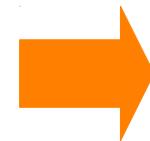
$$\mathbf{v} \equiv (1, \dots, 1) \rightarrow w^{t+1} = w^t - \gamma_t \nabla f(w^t)$$

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

# Examples of arbitrary sampling: uniform single element

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

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



# Examples of arbitrary sampling: uniform single element

Random set  $S \subset \{1, \dots, n\}$ ,  $|S| = 1$   
 $\text{Prob}[i \in S] = 1/n$ , for  $i = 1, \dots, n$



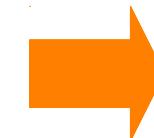
$$v_i = \begin{cases} n & i \in S \\ 0 & i \notin S \end{cases}$$

$\uparrow$

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

# Examples of arbitrary sampling: uniform single element

Random set  $S \subset \{1, \dots, n\}$ ,  $|S| = 1$   
 $\text{Prob}[i \in S] = 1/n$ , for  $i = 1, \dots, n$



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

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



$$\nabla f_{\textcolor{red}{v}}(w) = \nabla f_i(w)$$



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

# Examples of arbitrary sampling: uniform single element

Random set  $S \subset \{1, \dots, n\}$ ,  $|S| = 1$   
 $\text{Prob}[i \in S] = 1/n$ , for  $i = 1, \dots, n$



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

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



Single element SGD



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

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

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

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

# Examples of arbitrary sampling: uniform mini-batching

Random set  $S \subset \{1, \dots, n\}$ ,  $|S| = b$   
 $\text{Prob}[i \in S] = b/n$ , for  $i = 1, \dots, n$



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

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



Mini-batch SGD  
without replacement

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

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



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

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

# Examples of arbitrary sampling: non-uniform mini-batching

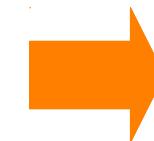
Random set  $S \subset \{1, \dots, n\}$ ,  $\mathbb{E}|S| = b$

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



# Examples of arbitrary sampling: non-uniform mini-batching

Random set  $S \subset \{1, \dots, n\}$ ,  $\mathbb{E}|S| = b$   
 $\text{Prob}[i \in S] = p_i, \quad \text{for } i = 1, \dots, n$



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

$\uparrow$

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



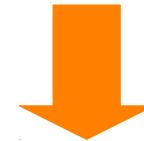
# Examples of arbitrary sampling: non-uniform mini-batching

Random set  $S \subset \{1, \dots, n\}$ ,  $\mathbb{E}|S| = b$   
 $\text{Prob}[i \in S] = p_i, \quad \text{for } i = 1, \dots, n$



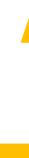
$$v_i = \begin{cases} \frac{1}{p_i} & i \in S \\ 0 & i \notin S \end{cases}$$

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



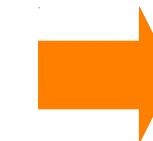
$$\nabla f_{\mathbf{v}}(w) = \frac{n}{p_i} \sum_{i \in S} \nabla f_i(w)$$

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



# Examples of arbitrary sampling: non-uniform mini-batching

Random set  $S \subset \{1, \dots, n\}$ ,  $\mathbb{E}|S| = b$   
 $\text{Prob}[i \in S] = p_i, \quad \text{for } i = 1, \dots, n$



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

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



## Arbitrary sampling SGD

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



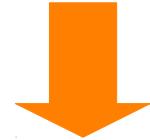
$$\nabla f_{v^t}(w) = \frac{n}{p_i} \sum_{i \in S} \nabla f_i(w)$$

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



# SGD with arbitrary sampling

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



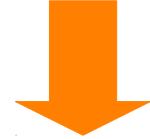
Sample  $\mathbf{v}^t \sim \mathcal{D}$

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

Includes all forms of  
SGD (and GD)

# SGD with arbitrary sampling

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



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

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

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

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

# 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

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

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

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

Iteration complexity of gradient descent

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

$$\frac{\|w^t - w^*\|}{\|w^0 - w^*\|} \leq \epsilon$$

# 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/ strong assumption

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_{\max}}{2} \|w - y\|_2^2$$

# Assumptions and Convergence of SGD

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

Bigger smoothness constant/ strong assumption

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_{\max}}{2} \|w - y\|_2^2$$

**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/ strong assumption

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_{\max}}{2} \|w - y\|_2^2$$

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

Iteration complexity of SGD

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



# 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/ strong assumption

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_{\max}}{2} \|w - y\|_2^2$$

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

$L \ll L_{\max}$

Iteration complexity of SGD

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



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



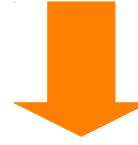
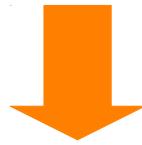
# Informal comparison between GD and SGD iteration complexity

GD

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

SGD

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



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

# Informal comparison between GD and SGD iteration complexity

GD

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

SGD

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



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

Ex:  $f(w) = \frac{1}{n} \|Aw - y\|^2 = \frac{1}{n} \sum_{i=1}^n (a_i w - y_i)^2$

$$L = \frac{1}{n} \lambda_{\max}(A^\top A) \leq \max_{i=1, \dots, n} \|a_i\|^2 = L_{\max}$$

# Informal comparison between GD and SGD iteration complexity

GD

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

SGD

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



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

Ex:  $f(w) = \frac{1}{n} \|Aw - y\|^2 = \frac{1}{n} \sum_{i=1}^n (a_i w - y_i)^2$

$$L = \frac{1}{n} \lambda_{\max}(A^\top A) \leq \max_{i=1, \dots, n} \|a_i\|^2 = L_{\max}$$

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

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



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

Ex:  $f(w) = \frac{1}{n} \|Aw - y\|^2 = \frac{1}{n} \sum_{i=1}^n (a_i w - y_i)^2$

$$L = \frac{1}{n} \lambda_{\max}(A^\top A) \leq \max_{i=1, \dots, n} \|a_i\|^2 = L_{\max}$$

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

Can one theorem include both SGD and GD analysis?

# Key constant: Expected smoothness

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

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

# Key constant: Expected smoothness

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

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

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

# Key constant: Expected smoothness

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

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

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

**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

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

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

Expected smoothnes constant  
Depends on  $\mathbf{v}$  and  $f$



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

**Lemma:**

$f_i$  convex and  $L_{\max}$ -smooth



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

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

# Key constant: Expected smoothness

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

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

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

**Expected smoothnes constant**  
Depends on  $\mathbf{v}$  and  $f$



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

**Lemma:**

$f_i$  convex and  $L_{\max}$ -smooth



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

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

Rough estimate  
(we can do better)

# Key constant: Expected smoothness

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

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

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

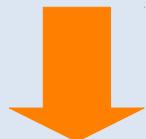
**Expected smoothnes constant**  
Depends on  $\mathbf{v}$  and  $f$



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

**Lemma:**

$f_i$  convex and  $L_{\max}$ -smooth



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

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

**Definition:** Gradient noise

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

Rough estimate  
(we can do better)

# Key constant: Expected smoothness

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

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

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

**Expected smoothnes constant**  
Depends on  $\mathbf{v}$  and  $f$



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

**Lemma:**

$f_i$  convex and  $L_{\max}$ -smooth



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

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

**Definition:** Gradient noise

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

Rough estimate  
(we can do better)

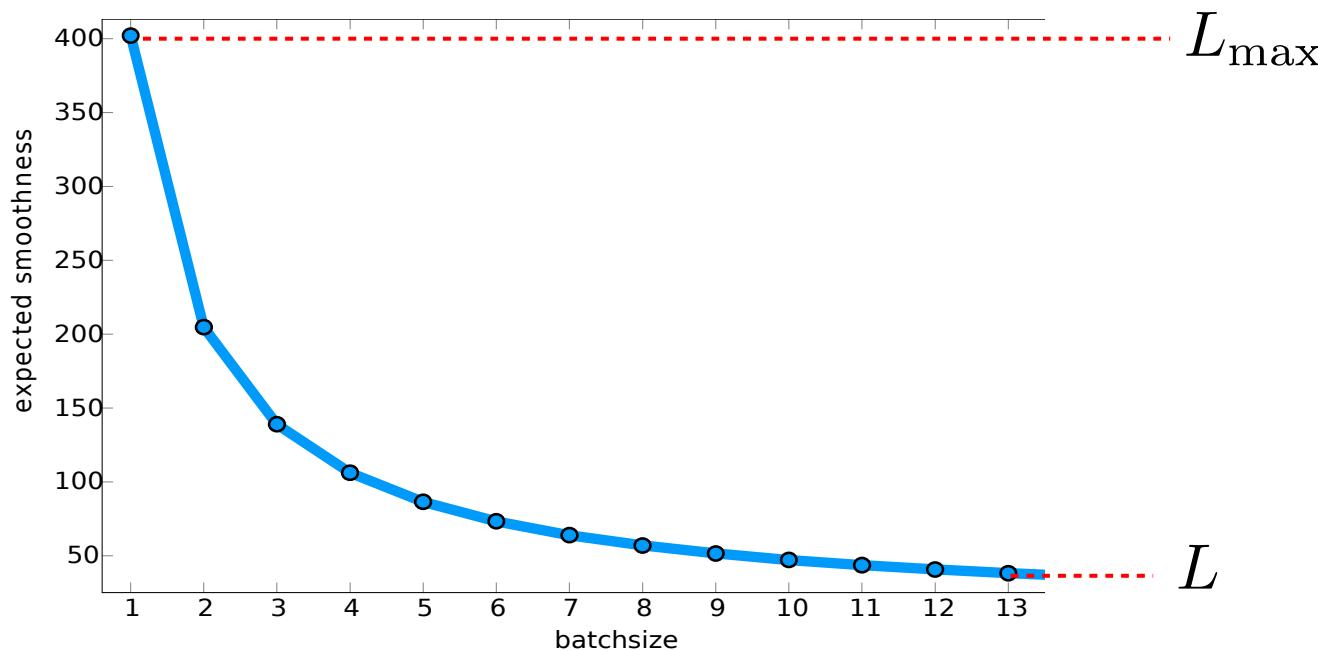
Generalization of  
 $\sigma_*^2 := \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w^*)\|^2$

# Example of Expected Smoothness

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

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

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

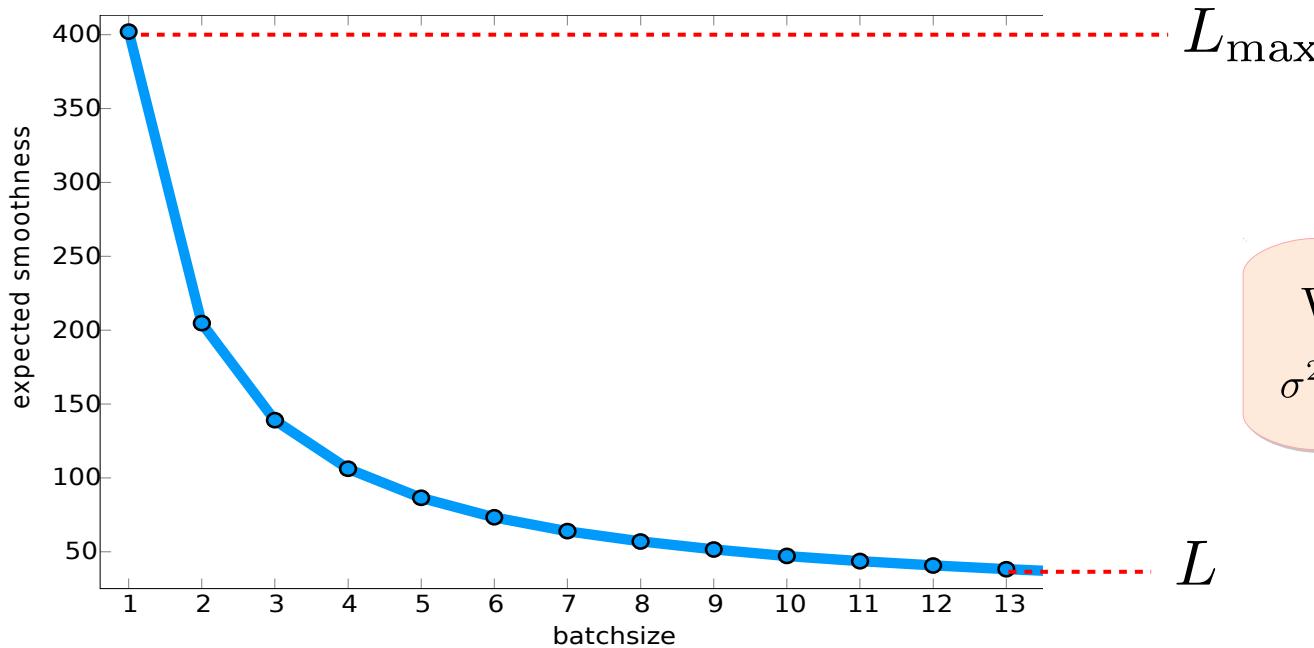


# Example of Expected Smoothness

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

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

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



What about  $\sigma^2$ ?  
 $\sigma^2 := \mathbb{E}[\|\nabla f_{\mathbf{v}}(w^*)\|^2]$

# Example of Expected Smoothness

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

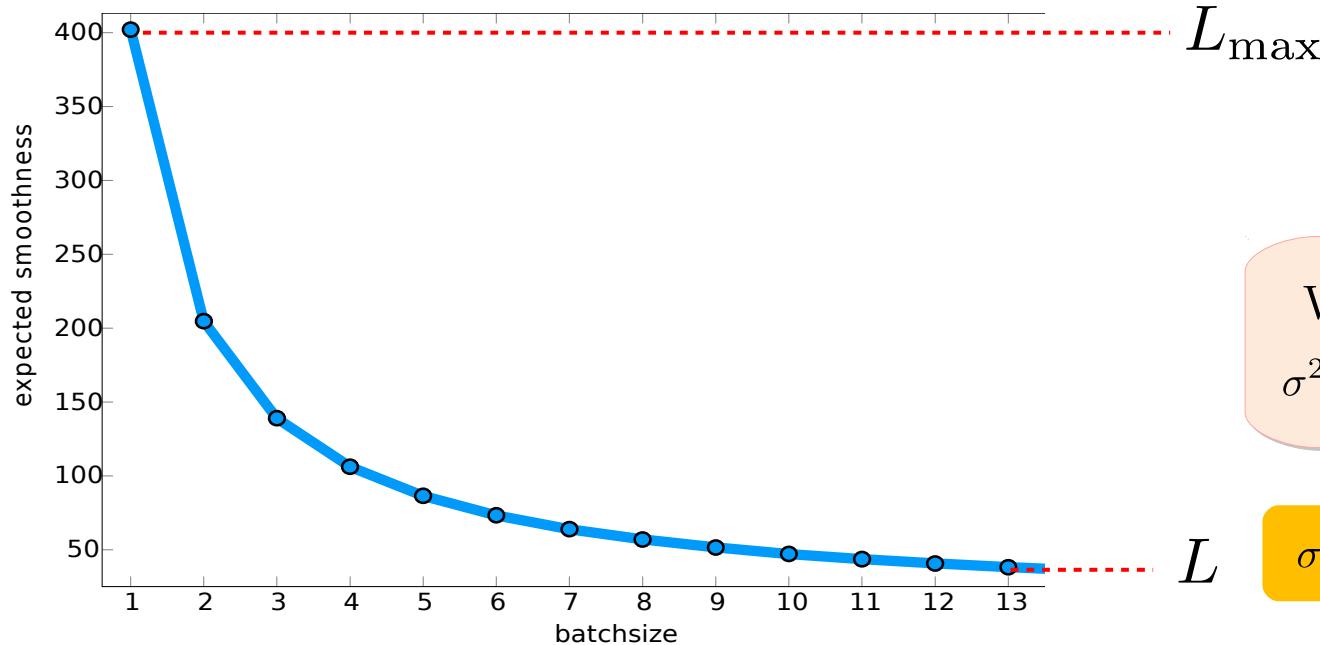
$$\mathcal{L} = \frac{n(b-1)}{b(n-1)}L + \frac{n-b}{b(n-1)}L_{\max}$$

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

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

Measure how much model fits data

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



What about  $\sigma^2$ ?

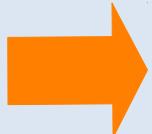
$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

$$\sigma^2 = 0$$

# Expected smoothness gives awesome bound on gradient

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$


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

# Expected smoothness gives awesome bound on gradient

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

Normally bound on  
gradient is an *assumption*

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

$$\mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^t)\|^2] \leq B^2$$



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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

Normally bound on  
gradient is an *assumption*

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

$$\mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^t)\|^2] \leq B^2$$



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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

Normally bound on gradient is an assumption

*informative:* with realistic assumptions

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

$$\mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^t)\|^2] \leq B^2$$



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)

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

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

# Main Theorem

(Linear convergence to a neighborhood)

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

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

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

# Main Theorem

(Linear convergence to a neighborhood)

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

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

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)

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

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

$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$

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

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

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

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

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

## Proof is SUPER EASY:

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

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

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

$$\mathbb{E}_{\textcolor{red}{v}} [\|w^{t+1} - w^*\|_2^2] = \|w^t - w^*\|_2^2 - 2\gamma \langle \nabla f(w^t), w^t - w^* \rangle + \gamma^2 \mathbb{E}_{\textcolor{red}{v}} [\|\nabla f_{\textcolor{red}{v}}(w^t)\|_2^2]$$

quasi strong conv  $\rightarrow \leq$   $(1 - \gamma\mu) \|w^t - w^*\|_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_{\textcolor{red}{v}} [\|\nabla f_{\textcolor{red}{v}}(w^t)\|_2^2]$

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

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

Taking total expectation

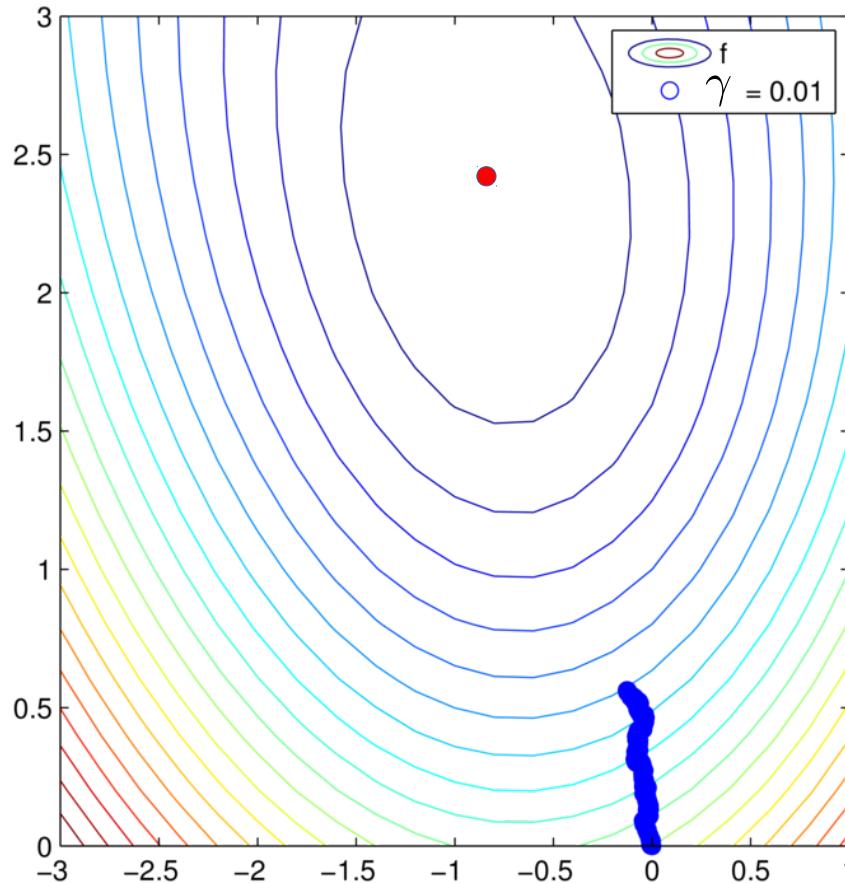
$$\begin{aligned}
 \mathbb{E} [\|w^{t+1} - w^*\|_2^2] &\leq (1 - \gamma\mu) \mathbb{E} [\|w^t - w^*\|_2^2] + 2\gamma^2\sigma^2 \\
 &= (1 - \gamma\mu)^{t+1} \|w^0 - w^*\|_2^2 + 2 \sum_{i=0}^t (1 - \gamma\mu)^i \gamma^2\sigma^2 \\
 &\leq (1 - \gamma\mu)^{t+1} \|w^0 - w^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}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

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

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

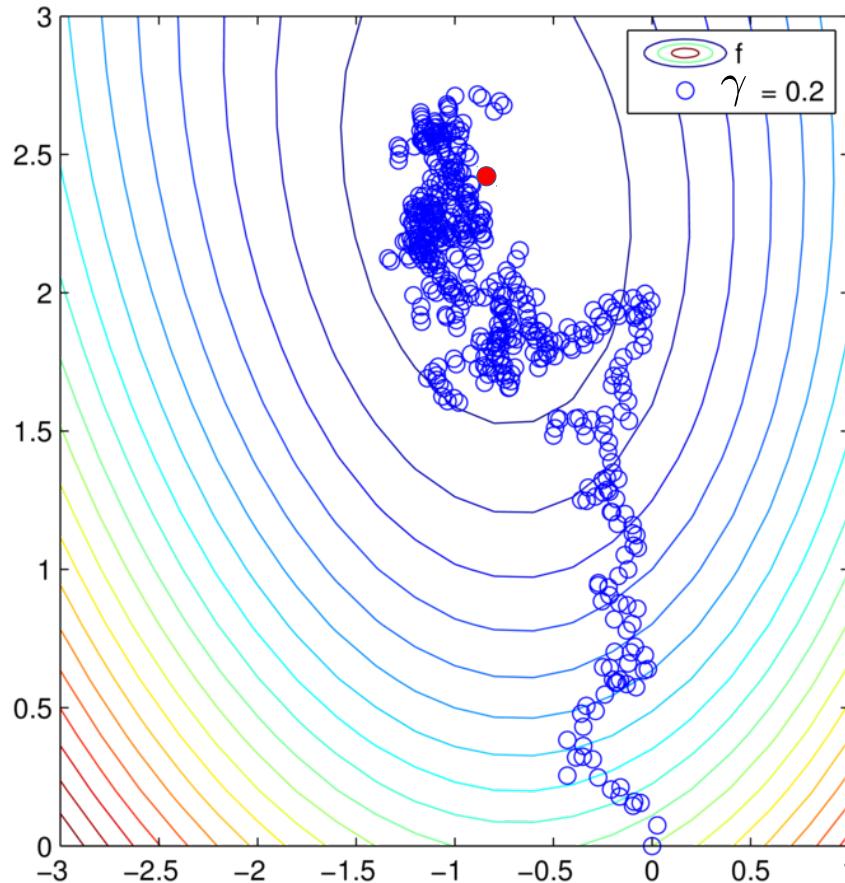
# Stochastic Gradient Descent

$\gamma = 0.01$



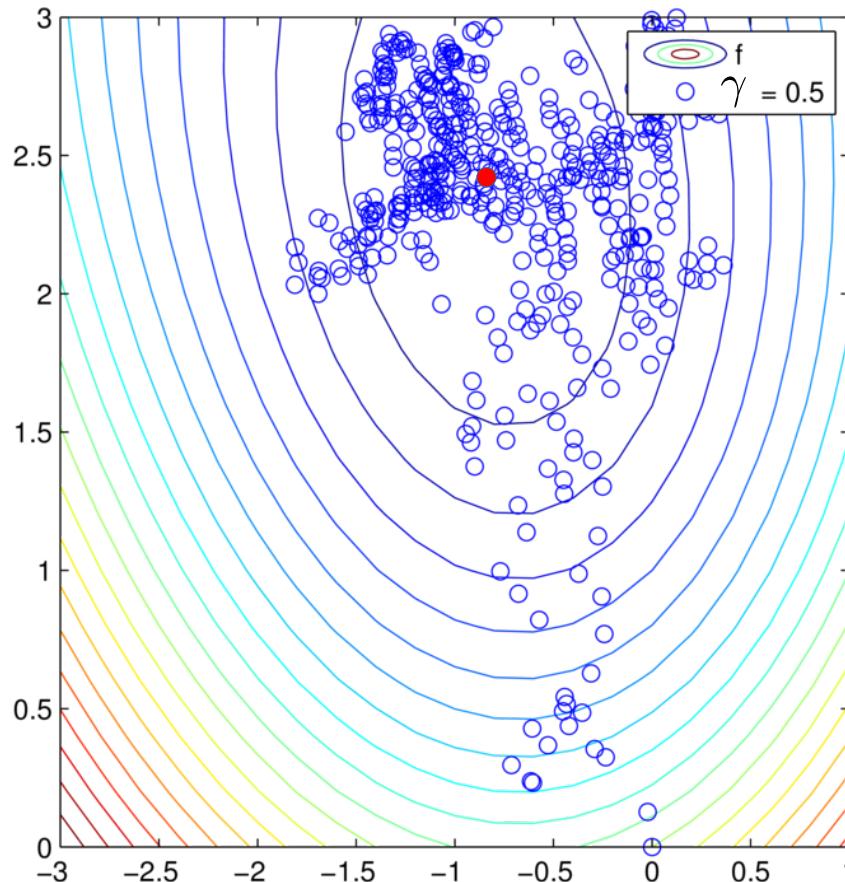
# Stochastic Gradient Descent

$$\gamma = 0.2$$



# 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

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

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

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

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

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

# 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

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

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

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

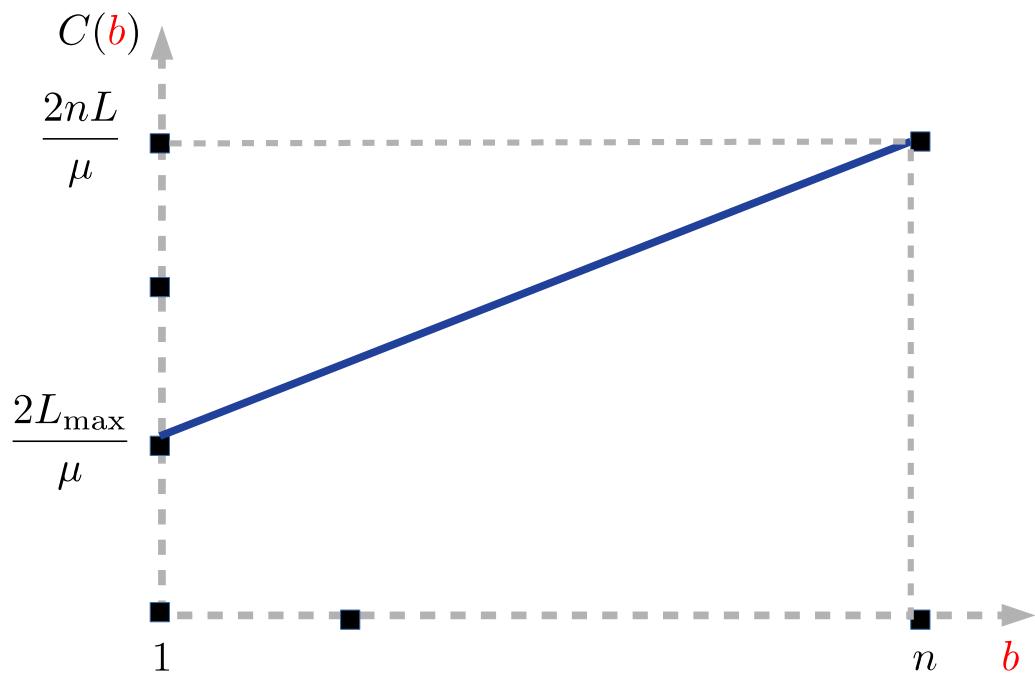


# Optimal mini-batch size

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

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

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



# Optimal mini-batch size

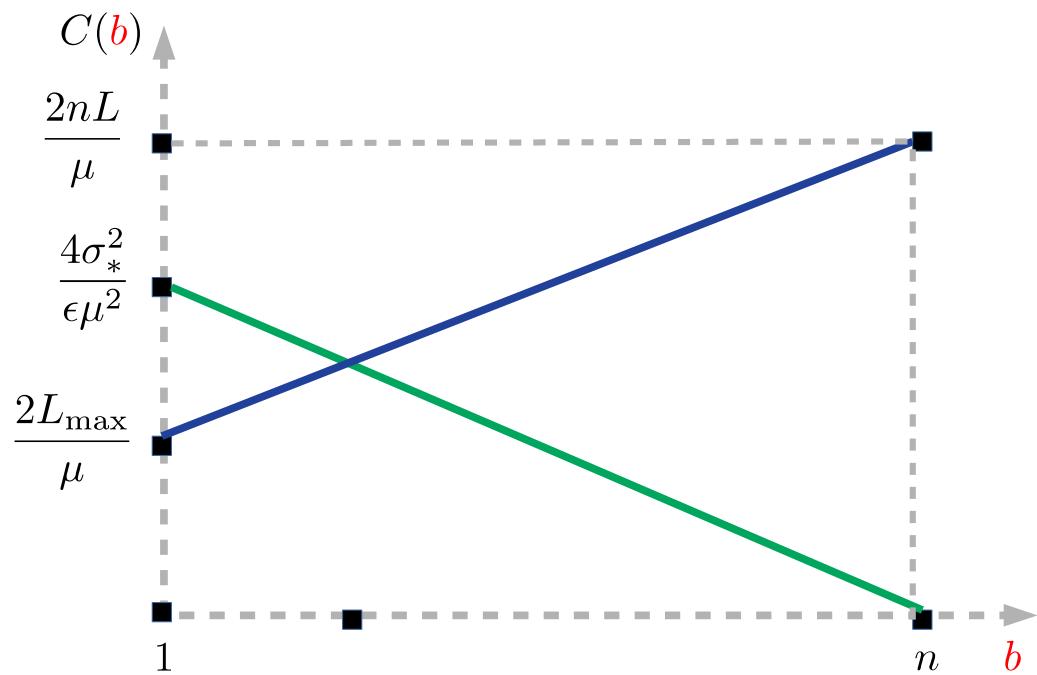
$$C(b) := \frac{2}{\mu(n-1)} \max \left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}}, \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \right\}$$

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

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

Linearly increasing

Linearly decreasing



# Optimal mini-batch size

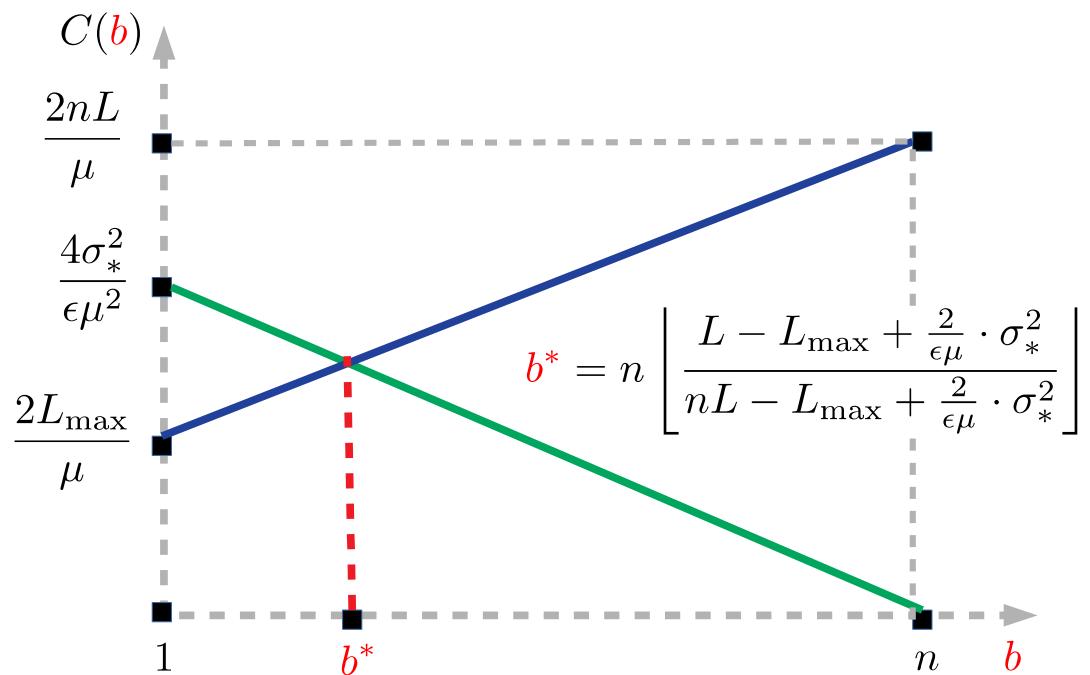
$$C(b) := \frac{2}{\mu(n-1)} \max \left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}}, \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \right\}$$

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

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

Linearly increasing

Linearly decreasing



# Optimal mini-batch size

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

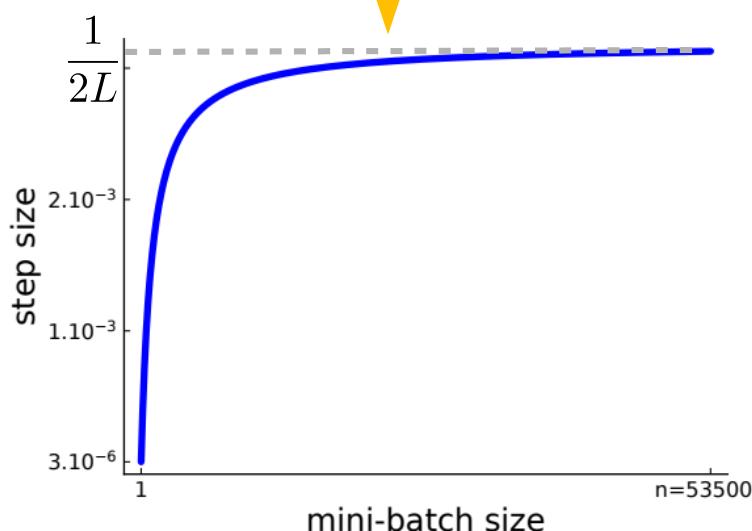
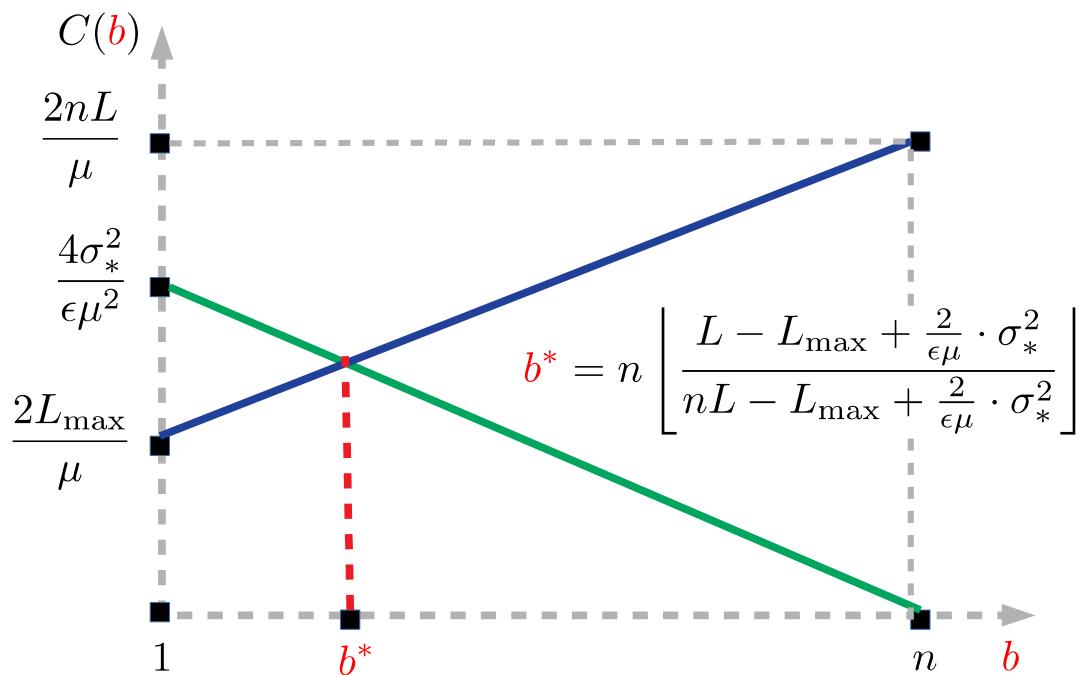
$$\sigma_1 := \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w^*)\|^2$$

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

Linearily increasing      Linearily decreasing

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

Stepsize increases with  $b$



# Optimal mini-batch size for models that interpolate data

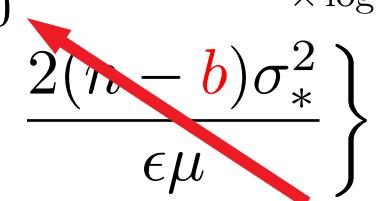
$$\sigma_1 := \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w^*)\|^2 = 0 \\ \times \log \left( \frac{2}{\epsilon} \right)$$

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

# Optimal mini-batch size for models that interpolate data

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

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

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


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

# Optimal mini-batch size for models that interpolate data

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

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

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

~~$\frac{2(n-\mathbf{b})\sigma_*^2}{\epsilon\mu}$~~

# Optimal mini-batch size for models that interpolate data

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

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

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

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

# Optimal mini-batch size for models that interpolate data

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

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

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

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

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

increases with  $b$



$$b^* = 1$$

# Optimal mini-batch size for models that interpolate data

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

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

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

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

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

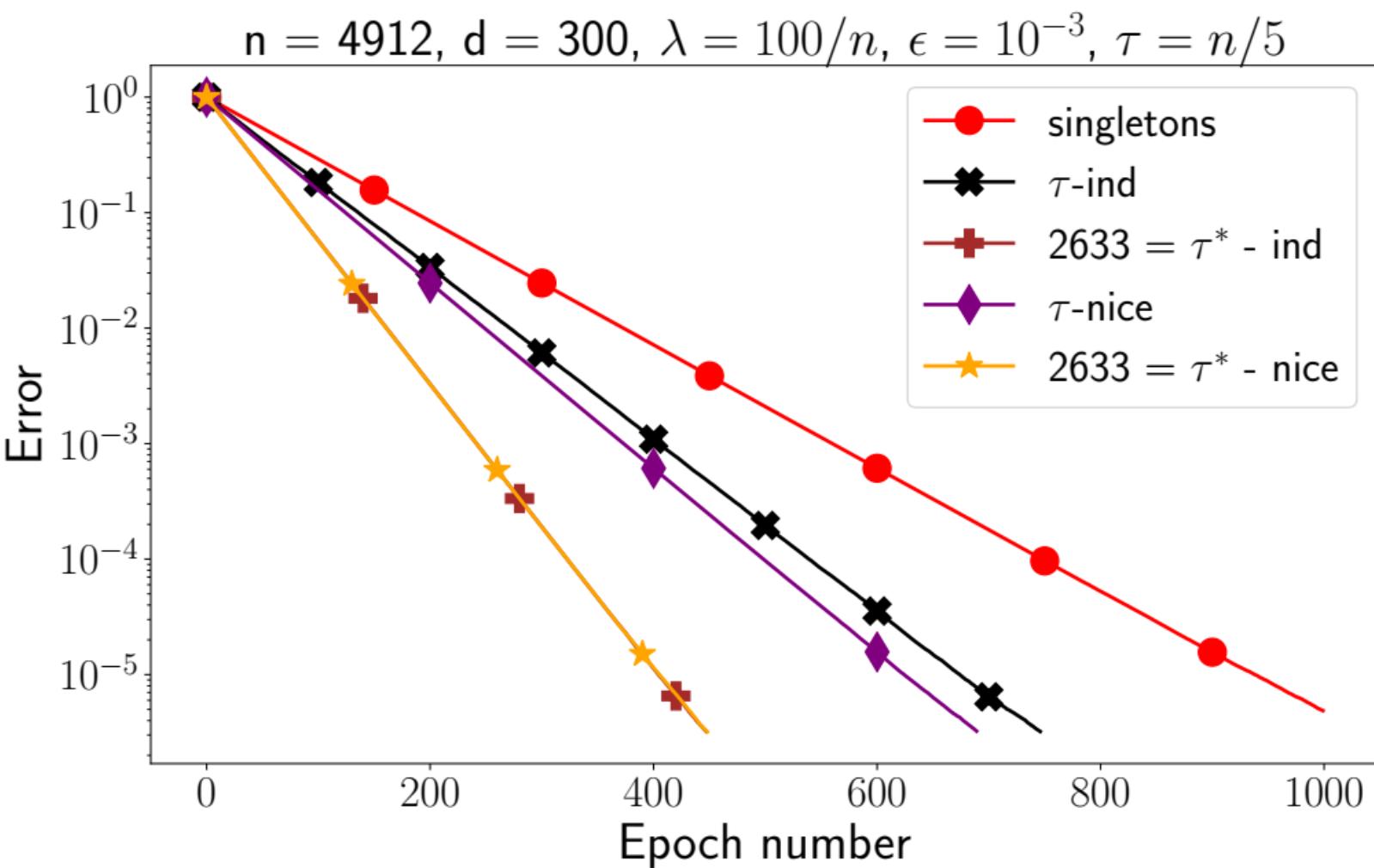
increases with  $b$



$b^* = 1$

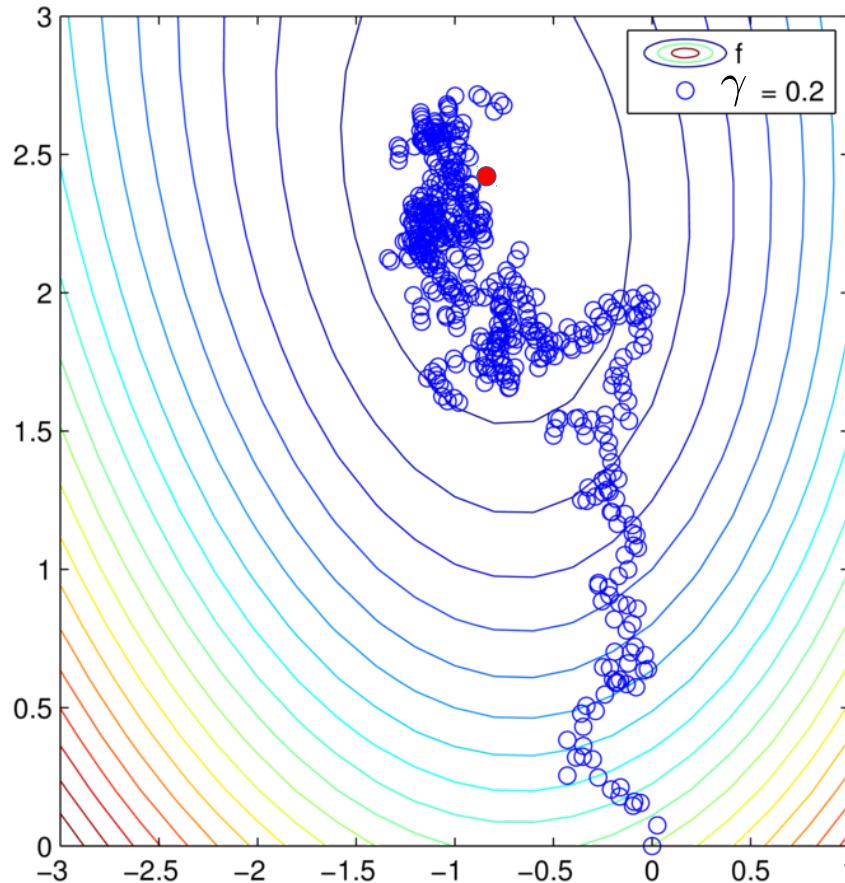
# Optimal mini-batch size

Logistic regression  
data: w3a (LIBSVM)



# Stochastic Gradient Descent

$$\gamma = 0.2$$



# Learning schedule: Constant & decreasing step sizes

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

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

A stochastic condition number

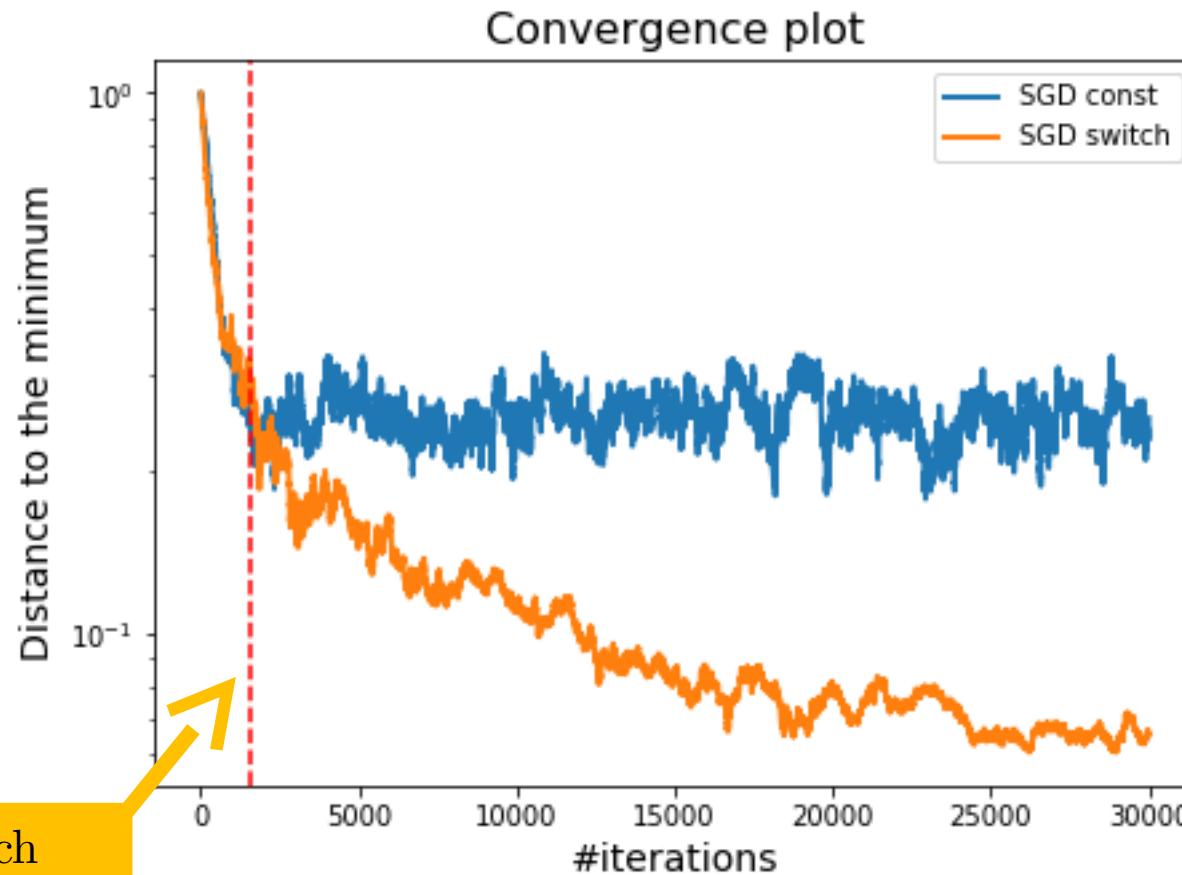
$$\sigma^2 := \mathbb{E}[\|\nabla f_{\textcolor{red}{v}}(w^*)\|^2]$$



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

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

# Stochastic Gradient Descent with switch to decreasing stepsizes



# Stochastic variance reduced methods

# Simple Stochastic Reformulation

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

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

$$f(w) := \frac{1}{n} \sum_{i=1}^n f_i(w) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\textcolor{red}{v}_i] f_i(w) = \mathbb{E} \left[ \underbrace{\frac{1}{n} \sum_{i=1}^n \textcolor{red}{v}_i}_{=: f_{\textcolor{red}{v}}(w)} f_i(w) \right]$$

What to do about the variance?

$=: f_{\textcolor{red}{v}}(w)$

**Original finite sum problem**

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



**Stochastic Reformulation**

$$\min_{w \in \mathbb{R}^d} \mathbb{E} [f_{\textcolor{red}{v}}(w)]$$

Minimizing the expectation of **random linear combinations** of original function

# Controlled Stochastic Reformulation

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

# Controlled Stochastic Reformulation

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

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

Cancel out

```
graph TD; A["covariate  $z_{\textcolor{red}{v}}(w) \in \mathbb{R}$ "] --> B[" $\mathbb{E}[f_{\textcolor{red}{v}}(w)] - \mathbb{E}[z_{\textcolor{red}{v}}(w)]$ "]; A --> C[" $\mathbb{E}[z_{\textcolor{red}{v}}(w)]$ "]; D["Cancel out"] --> C
```

# Controlled Stochastic Reformulation

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

The diagram illustrates the controlled stochastic reformulation. It shows the decomposition of the average function value into three components: the expected value of the function  $f_{\textcolor{red}{v}}(w)$ , the expected value of the covariate  $z_{\textcolor{red}{v}}(w)$ , and the difference between them. The term involving the covariate is further broken down into its expected value and the deviation from its expected value. Yellow boxes highlight the terms  $z_{\textcolor{red}{v}}(w)$  and  $\mathbb{E}[z_{\textcolor{red}{v}}(w)]$ , and a yellow box labeled "Cancel out" points to the cancellation of these terms.

# Controlled Stochastic Reformulation

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

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

Original finite sum problem

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



Controlled Stochastic Reformulation

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

Use covariates to **control the variance**

# Variance reduction with arbitrary sampling

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

# Variance reduction with arbitrary sampling

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



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

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

# Variance reduction with arbitrary sampling

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



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

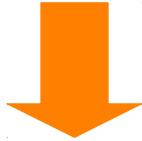
$$w^{t+1} = w^t - \gamma_t g_{\textcolor{red}{v}^t}(w^t)$$

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



# Variance reduction with arbitrary sampling

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



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

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

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

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

# Variance reduction with arbitrary sampling

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



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

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

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

How to choose  $z_{\mathbf{v}}(w)$  ?

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

# SVRG: Stochastic Variance Reduced Gradients



Johnson & Zhang, 2013 NIPS

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

Reference point

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

Sample

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

Grad. estimate

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

# SVRG: Stochastic Variance Reduced Gradients



Johnson & Zhang, 2013 NIPS

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

Reference point

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

Sample

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

Grad. estimate

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

Single element sampling

$$\mathbf{v}_j = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases}$$



# SVRG: Stochastic Variance Reduced Gradients



Johnson & Zhang, 2013 NIPS

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

Reference point

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

Single element sampling

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

Sample

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

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

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

Reference point

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

Single element sampling

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

Sample

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

Grad. estimate

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

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

# SVRG: Stochastic Variance Reduced Gradients



Johnson & Zhang, 2013 NIPS

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

Reference point

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

Single element sampling

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

Sample

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

Grad. estimate

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

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

# SVRG: Stochastic Variance Reduced Gradients



Jonhson & Zhang, NIPS 2013

Set  $w^0 = 0$ , choose  $\gamma > 0, m \in \mathbb{N}$ ,

$\alpha_k > 0$  for  $k = 0, \dots, m - 1$

$\tilde{w}^0 = w^0$

for  $t = 0, 1, 2, \dots, T - 1$

calculate  $\nabla f(\tilde{w}^t)$

for  $k = 0, 1, 2, \dots, m - 1$

sample  $i \in \{1, \dots, n\}$

$g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)$

$w^{k+1} = w^k - \gamma g^k$

$\tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k$

Output  $\tilde{w}^T$



Sebbouh, Gazagnadou & Gower, (In a near future)

# SVRG: Stochastic Variance Reduced Gradients



Jonhson & Zhang, NIPS 2013

Set  $w^0 = 0$ , choose  $\gamma > 0, m \in \mathbb{N}$ ,

$\alpha_k > 0$  for  $k = 0, \dots, m - 1$

$\tilde{w}^0 = w^0$

for  $t = 0, 1, 2, \dots, T - 1$

calculate  $\nabla f(\tilde{w}^t)$

Freeze reference point  
for  $m$  iterations

for  $k = 0, 1, 2, \dots, m - 1$

sample  $i \in \{1, \dots, n\}$

$g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)$

$w^{k+1} = w^k - \gamma g^k$

$\tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k$

Output  $\tilde{w}^T$



Sebbouh, Gazagnadou & Gower, (In a near future)

# SVRG: Stochastic Variance Reduced Gradients



Jonhson & Zhang, NIPS 2013

Set  $w^0 = 0$ , choose  $\gamma > 0, m \in \mathbb{N}$ ,

$\alpha_k > 0$  for  $k = 0, \dots, m - 1$

$\tilde{w}^0 = w^0$

for  $t = 0, 1, 2, \dots, T - 1$

calculate  $\nabla f(\tilde{w}^t)$

Freeze reference point  
for  $m$  iterations

for  $k = 0, 1, 2, \dots, m - 1$

sample  $i \in \{1, \dots, n\}$

$$g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)$$

$$w^{k+1} = w^k - \gamma g^k$$

$$\tilde{w}^{t+1} = \frac{1}{m} \sum_{k=0}^{m-1} \alpha_k w^k$$

Weighted average of  
inner iterates

Output  $\tilde{w}^T$



Sebbouh, Gazagnadou & Gower, (In a near future)

# SAGA: Stochastic Average Gradient



Defazio, Bach, & Lacoste-Julien, 2014 NIPS

Sample

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

Grad. estimate

$$g_{v^t}(w^t) = \nabla f_{\textcolor{red}{i}}(w^t) - \nabla f_{\textcolor{red}{i}}(w^{t_{\textcolor{red}{i}}}) + \frac{1}{n} \sum_{j=1}^n \nabla f_j(w^{t_j})$$

Store grad.

$$\nabla f_i(w^{t_{\textcolor{red}{i}}}) = \nabla f_i(w^t)$$

Single element sampling

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

# SAGA: Stochastic Average Gradient



Defazio, Bach, & Lacoste-Julien, 2014 NIPS

Sample

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

Single element sampling

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

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

Sample

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

Single element sampling

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

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

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

Single element sampling

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

Sample

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

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

Set  $w^0 = 0, g_i = \nabla f_i(w^0)$ , for  $i = 1 \dots, n$

Choose  $\gamma > 0$

for  $t = 0, 1, 2, \dots, T - 1$

sample  $i \in \{1, \dots, n\}$

$$g^t = \nabla f_i(w^t) - g_i + \frac{1}{n} \sum_{j=1}^n g_j$$

$$w^{t+1} = w^t - \gamma g^t$$

$$g_i = \nabla f_i(w^t)$$

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

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



Sebbouh, Gazagnadou & Gower, (In a near future)

# Iteration complexity for SVRG and SAGA for arbitrary sampling

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

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



Sebbouh, Gazagnadou & Gower, (In a near future)

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

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



Gower, Bach, Richtarik, 2018

# Iteration complexity for SVRG and SAGA for arbitrary sampling

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

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



Sebbouh, Gazagnadou & Gower, (In a near future)

Missing details due to extra definitions

**Theorem for SAGA (and the JacSketch family of methods)**

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

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



Gower, Bach, Richtarik, 2018

# Total Complexity of mini-batch

SVRG



Sebbouh, Gazagnadou & Gower, (In a near future)

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

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

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

# Total Complexity of mini-batch

SVRG



Sebbouh, Gazagnadou & Gower, (In a near future)

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

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

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

# Total Complexity of mini-batch

## SVRG



Sebbouh, Gazagnadou & Gower, (In a near future)

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

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

# Total Complexity of mini-batch

SVRG

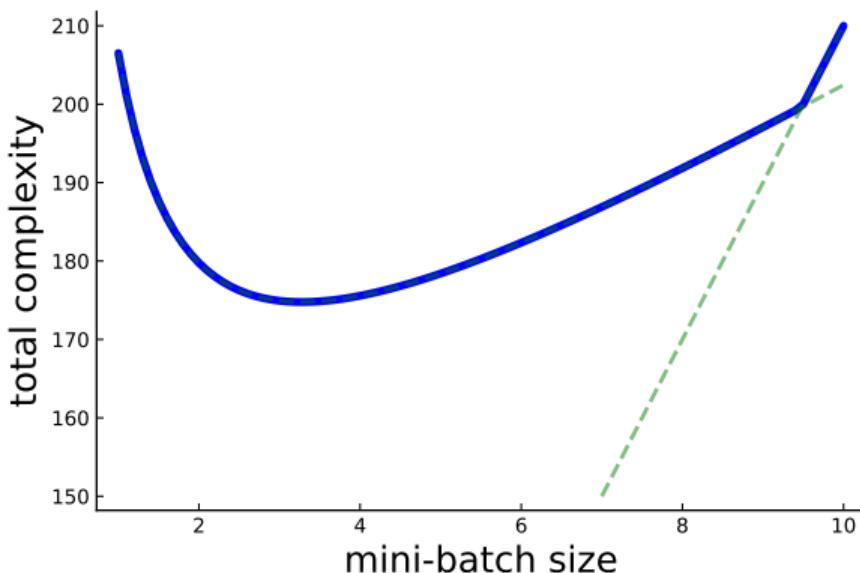


Sebbouh, Gazagnadou & Gower, (In a near future)

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

Linearly decreasing

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



# Total Complexity of mini-batch

SVRG

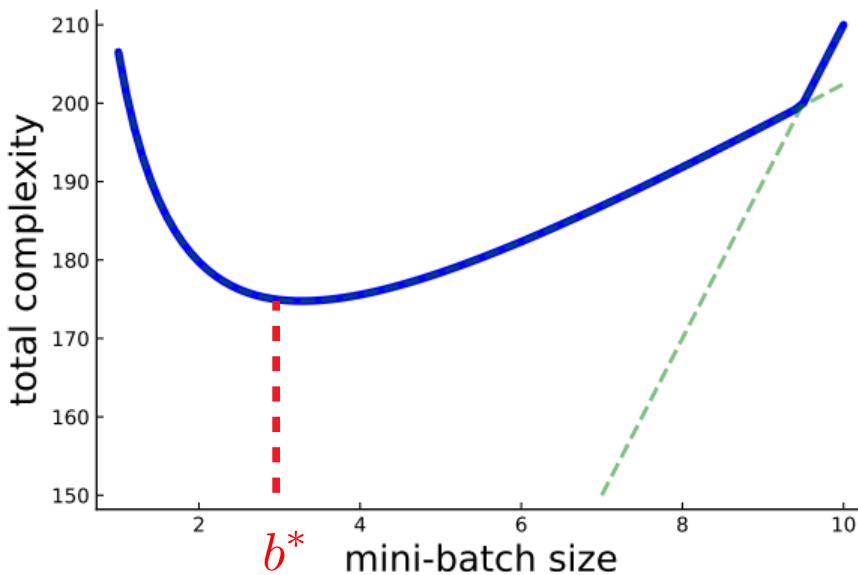


Sebbouh, Gazagnadou & Gower, (In a near future)

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

Linearly decreasing

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



# Total Complexity of mini-batch

SVRG



Sebbouh, Gazagnadou & Gower, (In a near future)

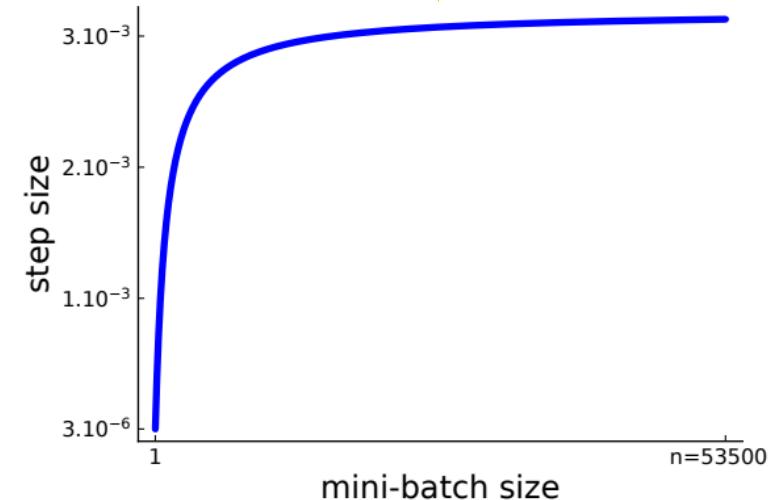
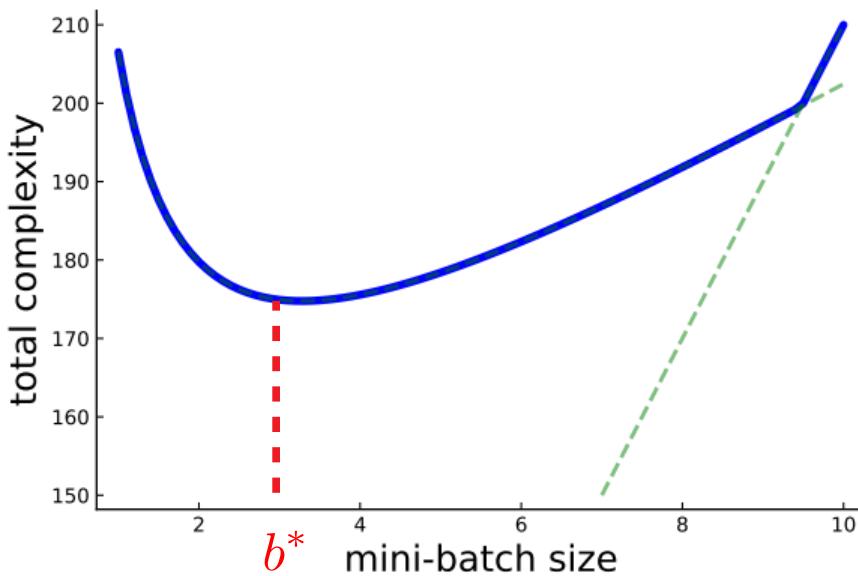
$$C(b) = 2 \left( \frac{n}{m} + 2b \right) \max \left\{ \frac{3}{b} \frac{n-b}{n-1} \frac{L_{\max}}{\mu} + \frac{3n}{b} \frac{b-1}{n-1} \frac{L}{\mu}, m \right\}$$

Non-linearly increasing

Linearly decreasing

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

Stepsize increasing with  $b$

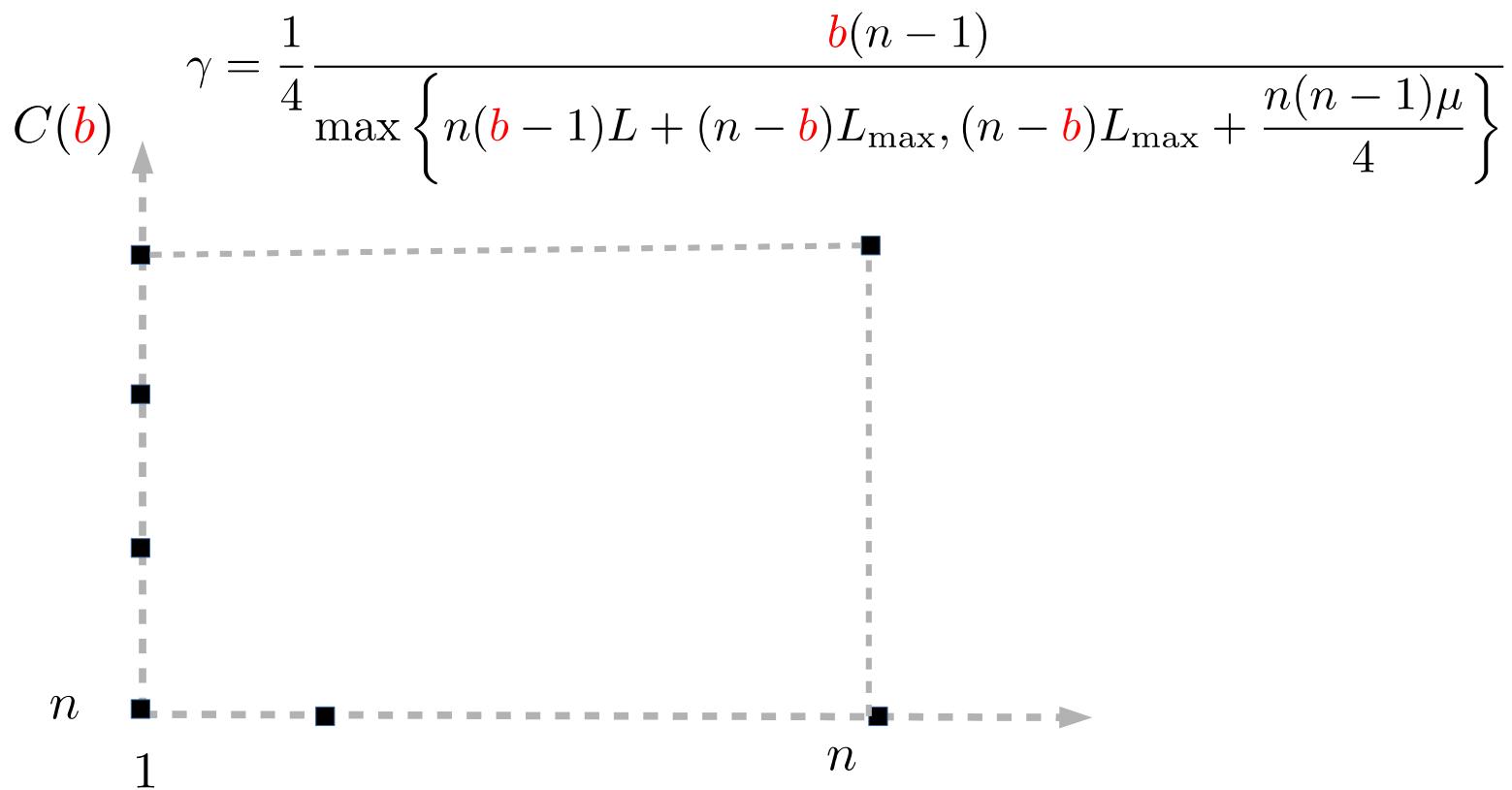


# Total Complexity of mini-batch SAGA



Gazagnadou, RMG & Salmon, ICML 2019

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



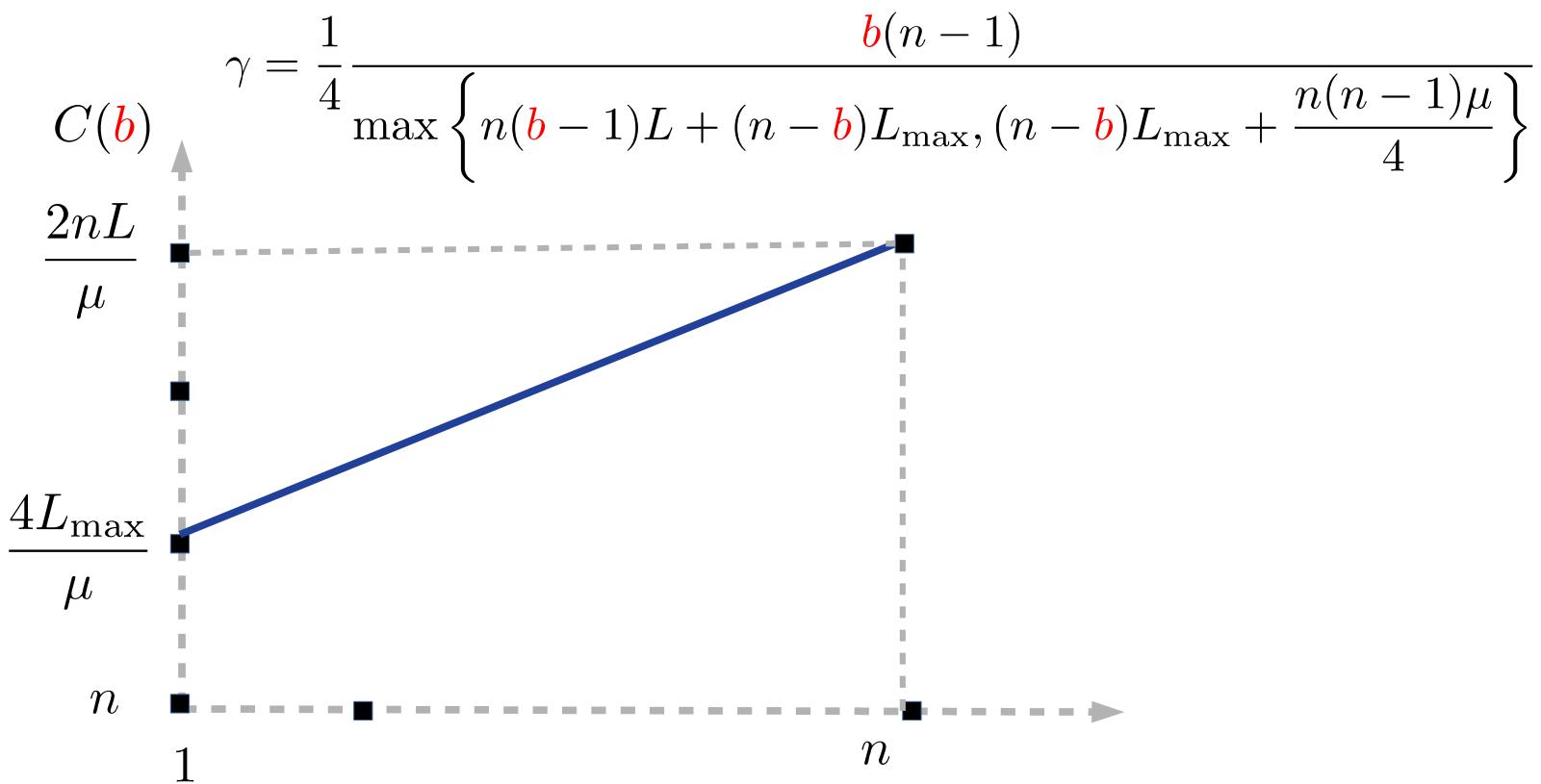
# Total Complexity of mini-batch SAGA



Gazagnadou, RMG & Salmon, ICML 2019

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

Linearly increasing



# Total Complexity of mini-batch SAGA

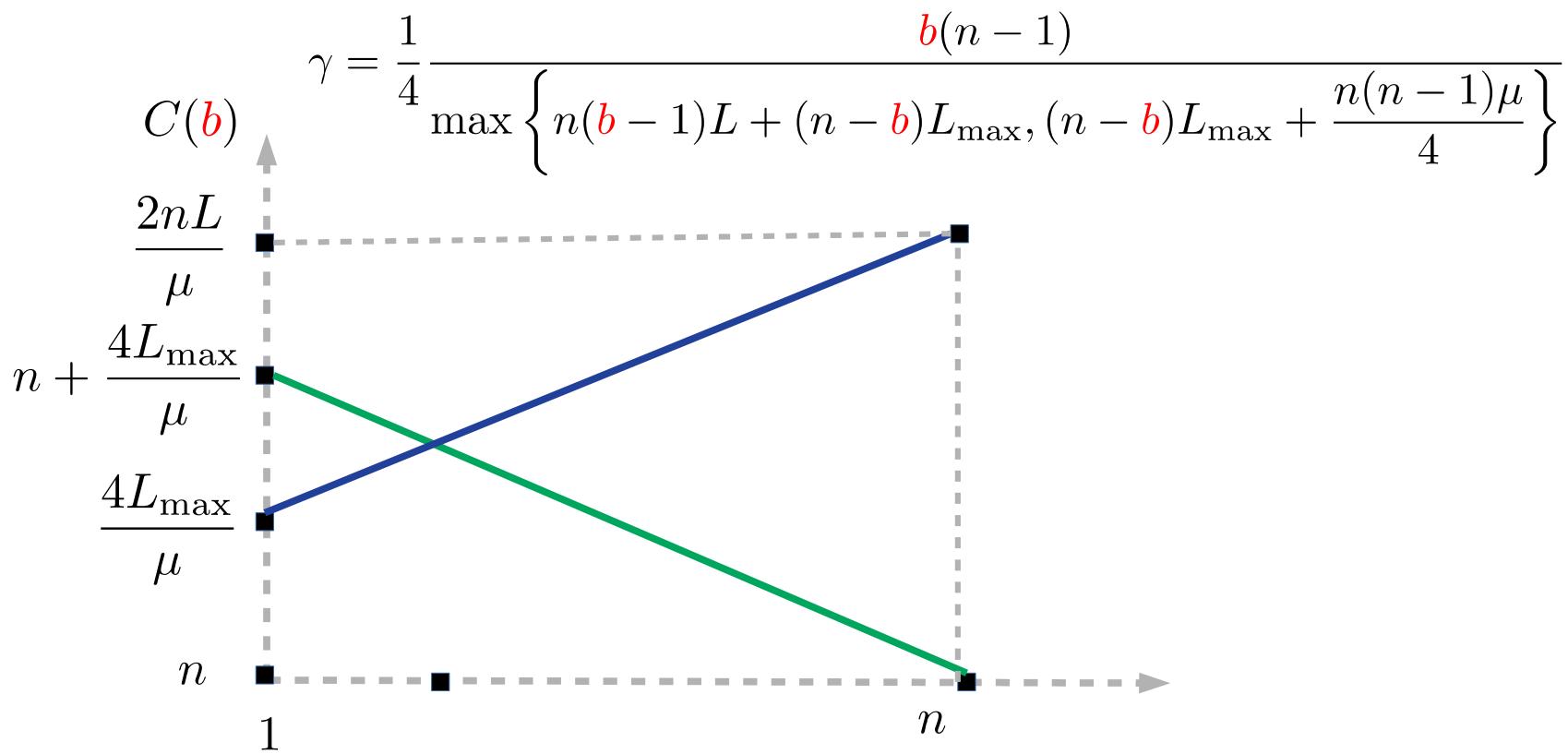


Gazagnadou, RMG & Salmon, ICML 2019

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

Linearly increasing

Linearly decreasing



# Total Complexity of mini-batch SAGA

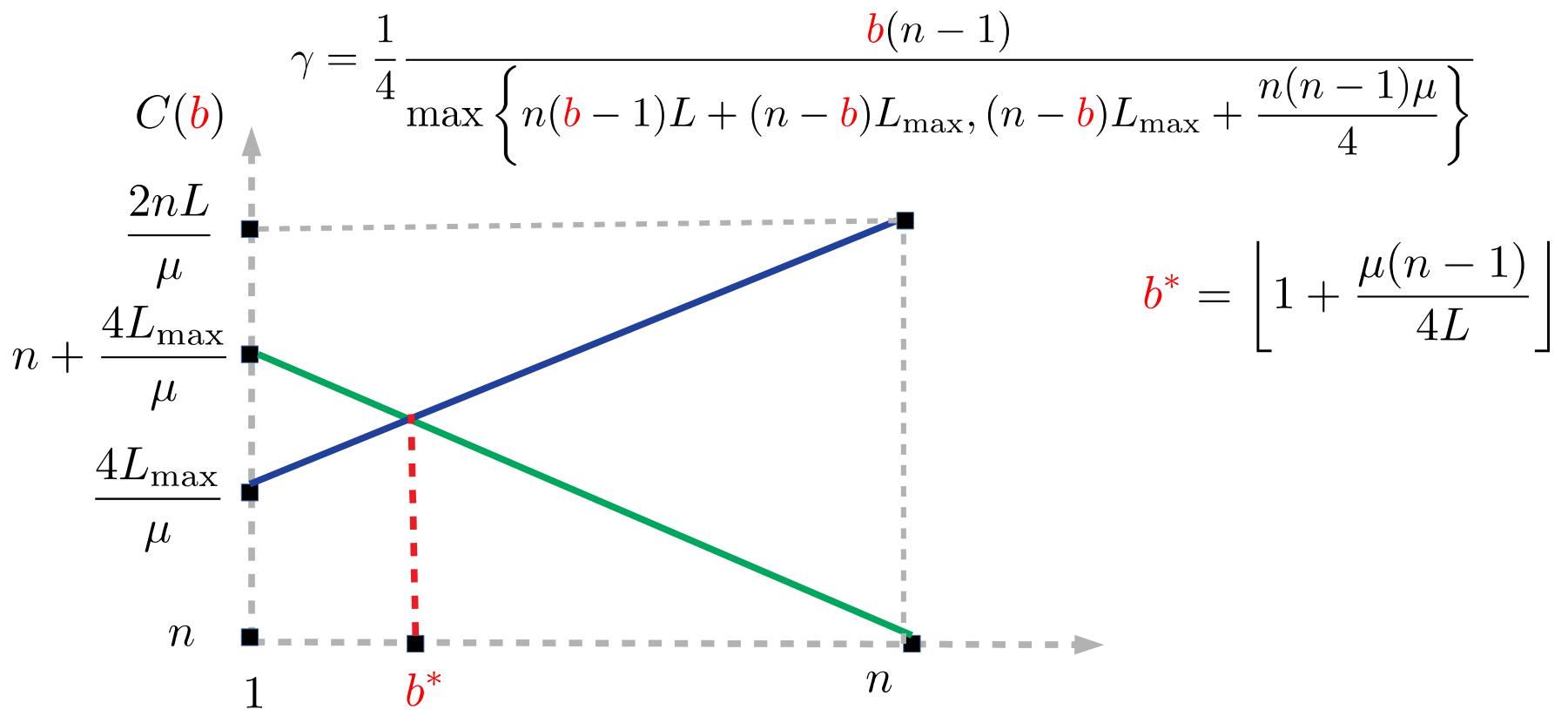


Gazagnadou, RMG & Salmon, ICML 2019

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

Linearly increasing

Linearly decreasing



# Total Complexity of mini-batch SAGA

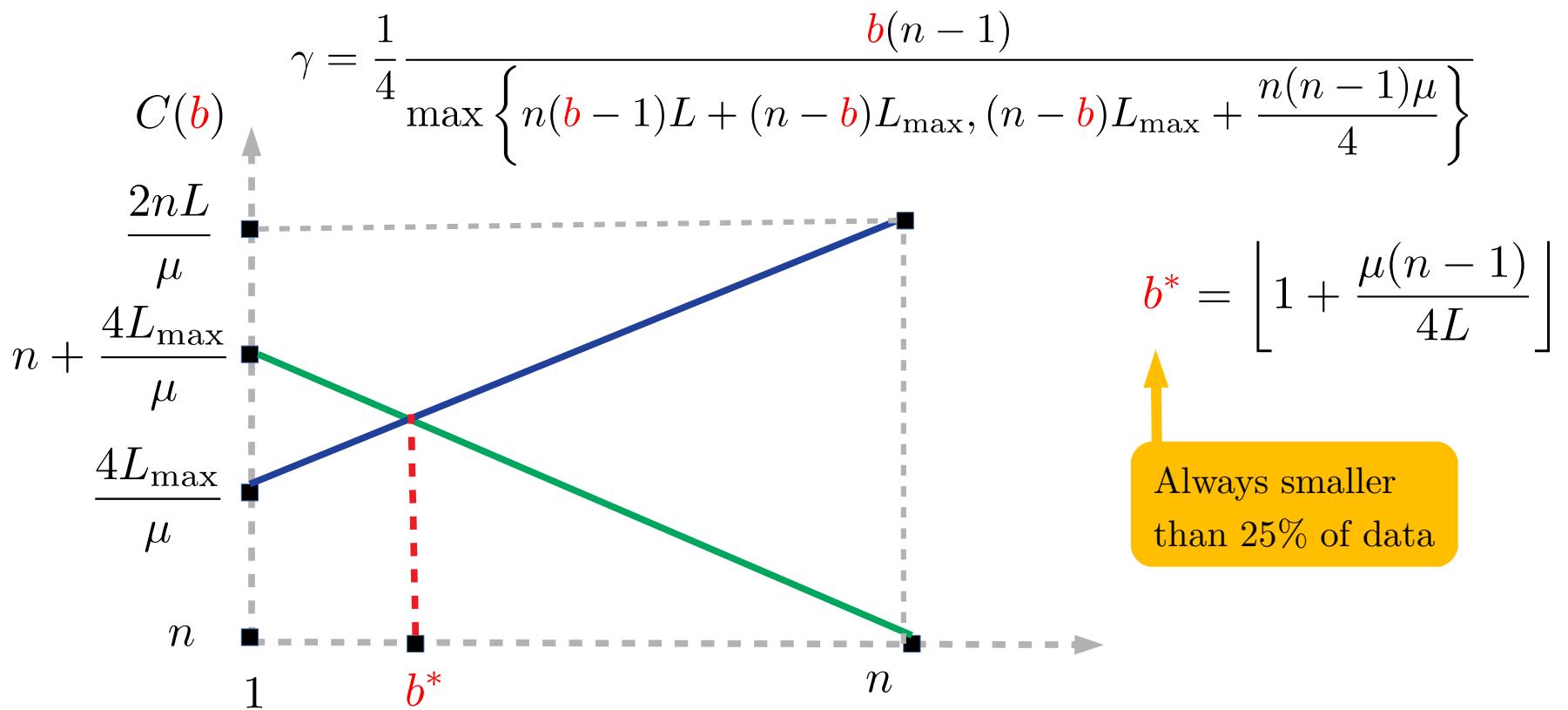


Gazagnadou, RMG & Salmon, ICML 2019

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

Linearly increasing

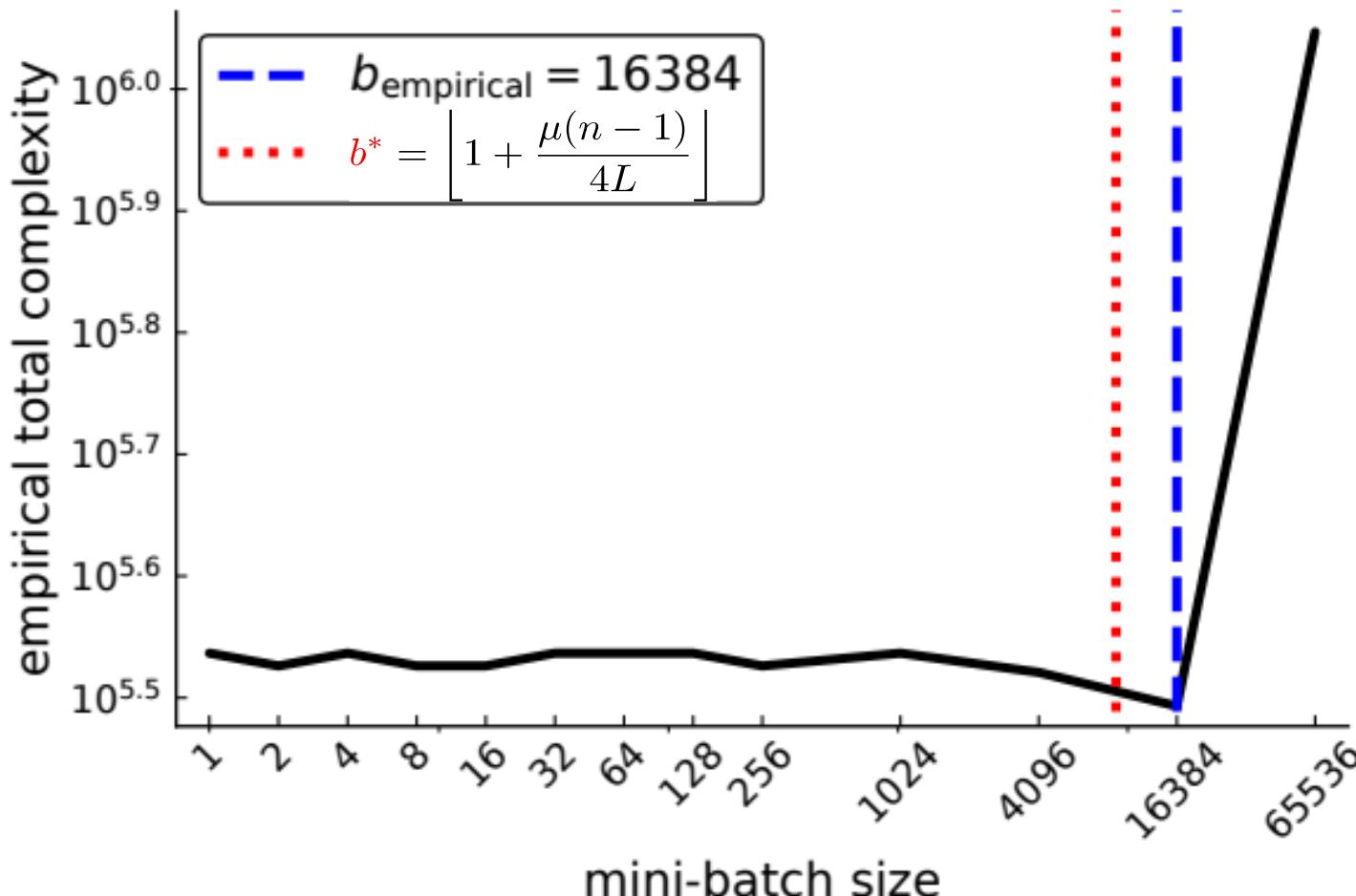
Linearly decreasing



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

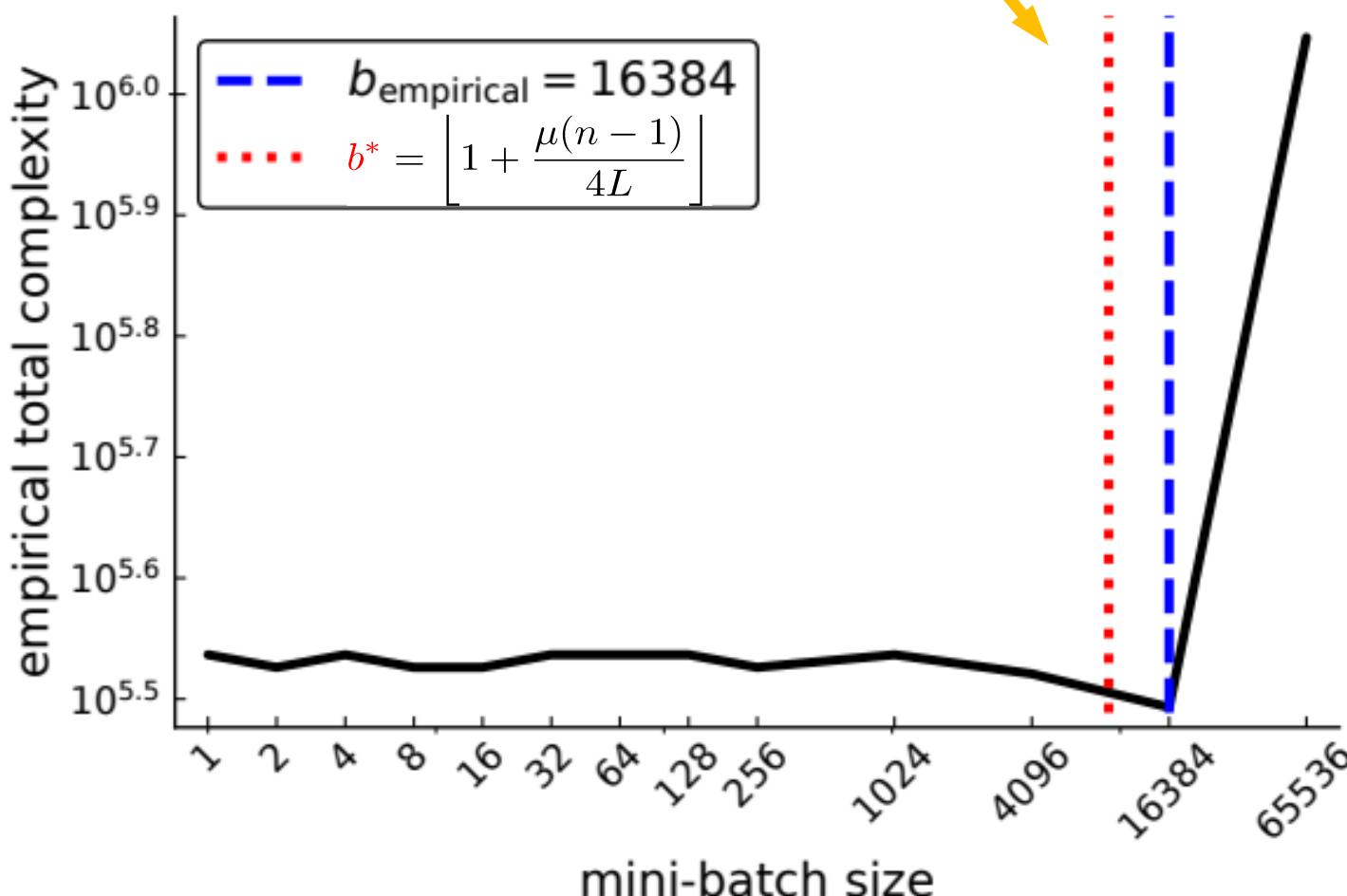
Always smaller  
than 25% of data

# 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

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

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

$$\mathbb{E}[||\nabla f_{\textcolor{red}{v}}(w) - \nabla f_{\textcolor{red}{v}}(w^*)||_2^2] \leq \textcolor{blue}{L} (f(w) - f(w^*)) \\ (f, \mathcal{D}) \sim ES(\textcolor{blue}{L})$$

# Take home message

Stochastic reformulations allow to view all variants as simple SGD

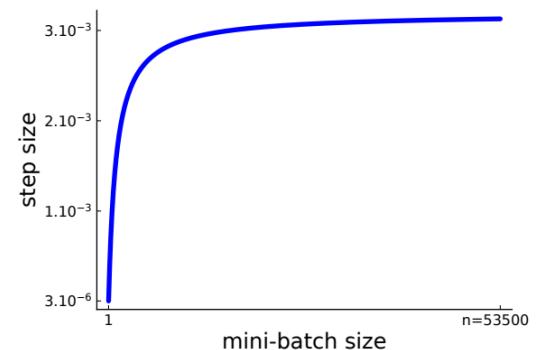
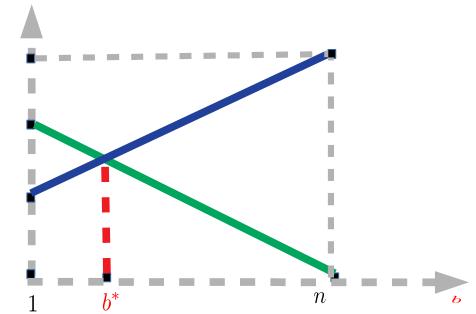
To analyse all forms of sampling used through expected smooth

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

Stepsize increase by orders when mini-batch size increases

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

$$\mathbb{E}[||\nabla f_{\textcolor{red}{v}}(w) - \nabla f_{\textcolor{red}{v}}(w^*)||_2^2] \leq \textcolor{blue}{L} (f(w) - f(w^*)) \\ (f, \mathcal{D}) \sim ES(\textcolor{blue}{L})$$





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**

# Learning rate schedules

# Main Theorem

(Linear convergence to a neighborhood)

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

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

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

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

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

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