Optimization for Data Science

Mini-batching, sampling, momentum and other tricks

Lecturer: Robert M. Gower & Alexandre Gramfort

Tutorials: Quentin Bertrand, Nidham Gazagnadou

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)

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

Step size/
Learning rate

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

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)

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

Step size/
Learning rate

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

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)

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

Sample mini-batch with $B \subset \{1, \ldots, n\}$ 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)

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

- What should b and be?

Sample mini-batch with
$B \subset \{1,\ldots,n\}$ 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)

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

Sample mini-batch with $B \subset \{1, \ldots, n\}$ with $|B| = b$

- What should b and   be?
- How does b influence the stepsize ?

The Stochastic Gradient Method

Solving the *training problem*:

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

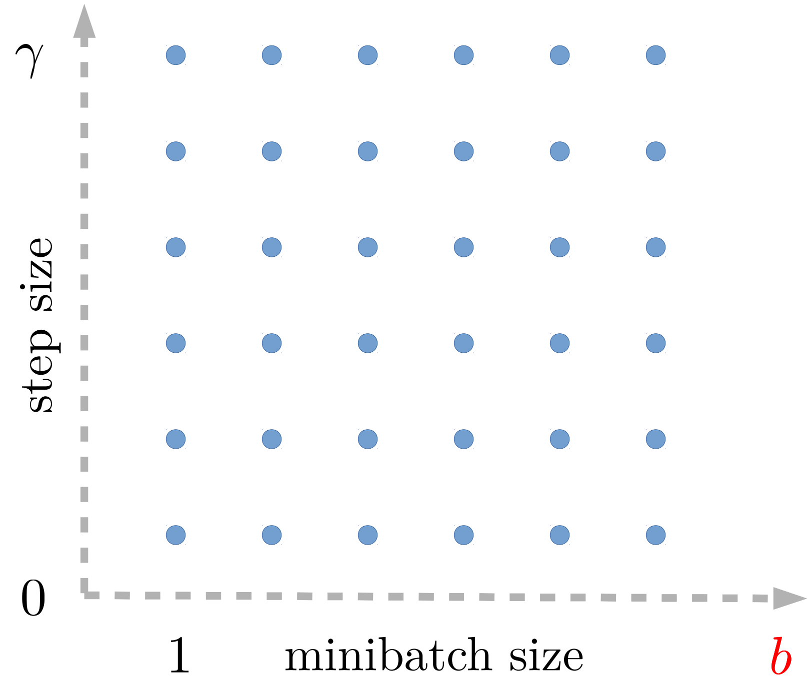
Baseline method: Stochastic Gradient Descent (SGD)

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

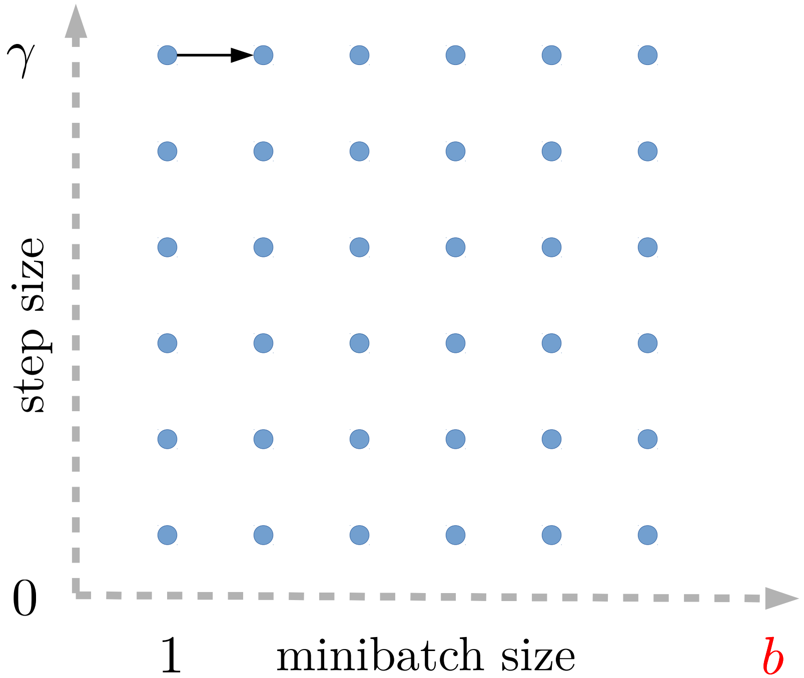
- What should b and γ be?
- How does b influence the stepsize γ?
- How does the data influence the best mini-batch and stepsize?

Sample mini-batch with $B \subset \{1,\dots,n\}$ with $|B| = b$

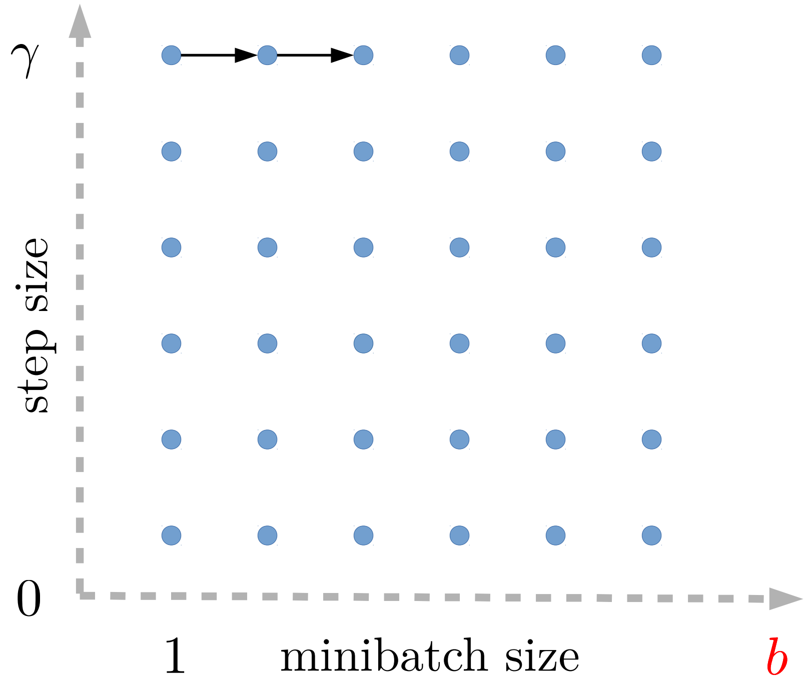
How to choose the minibatch size?



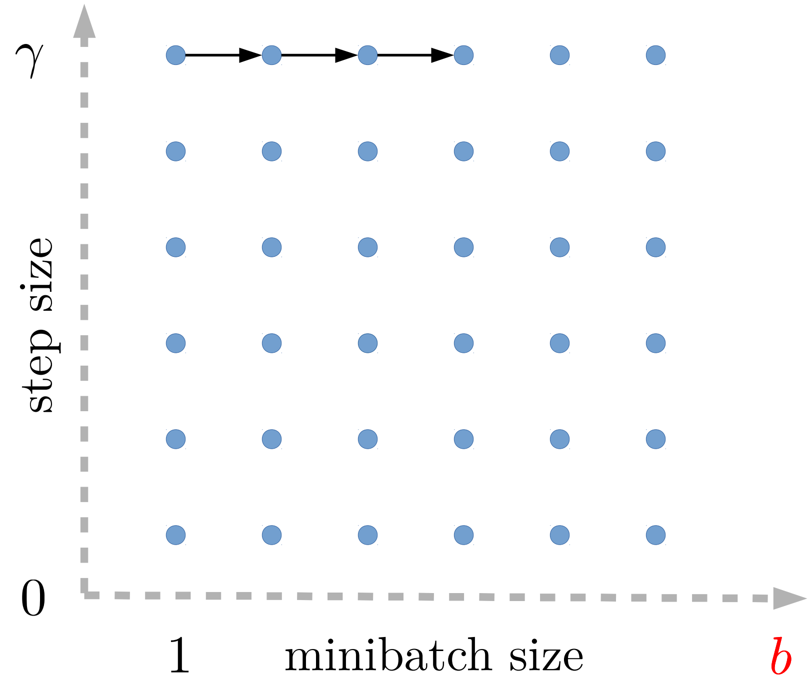
How to choose the minibatch size?



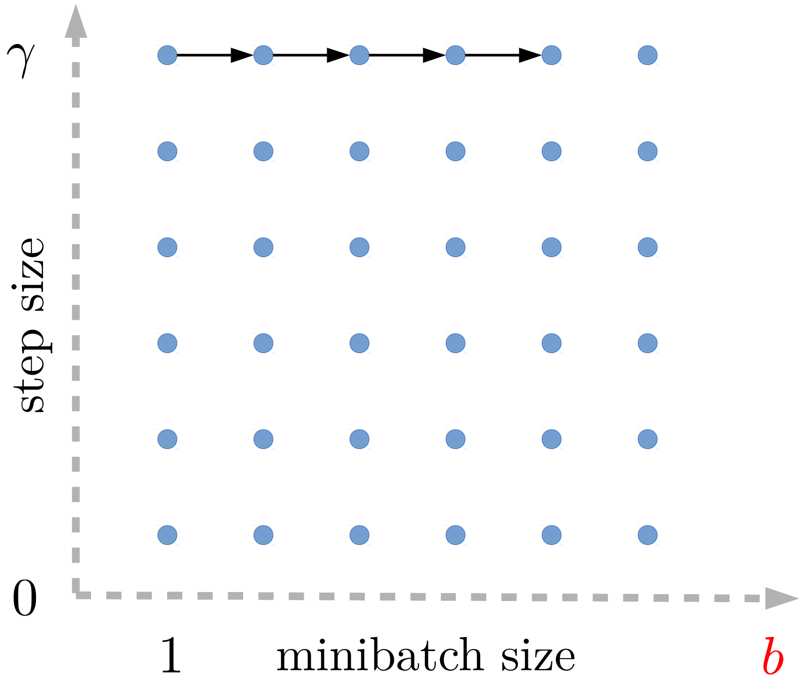
How to choose the minibatch size?



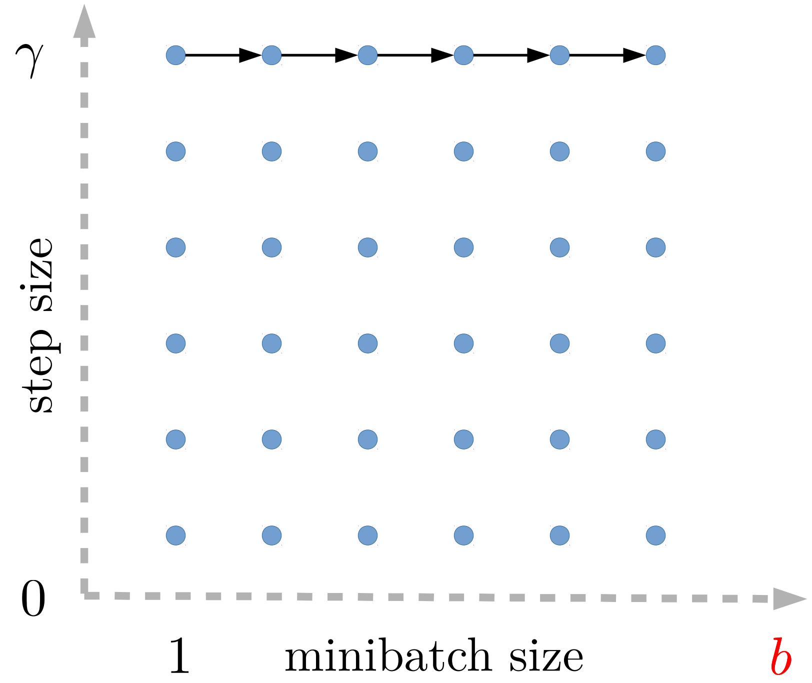
How to choose the minibatch size?



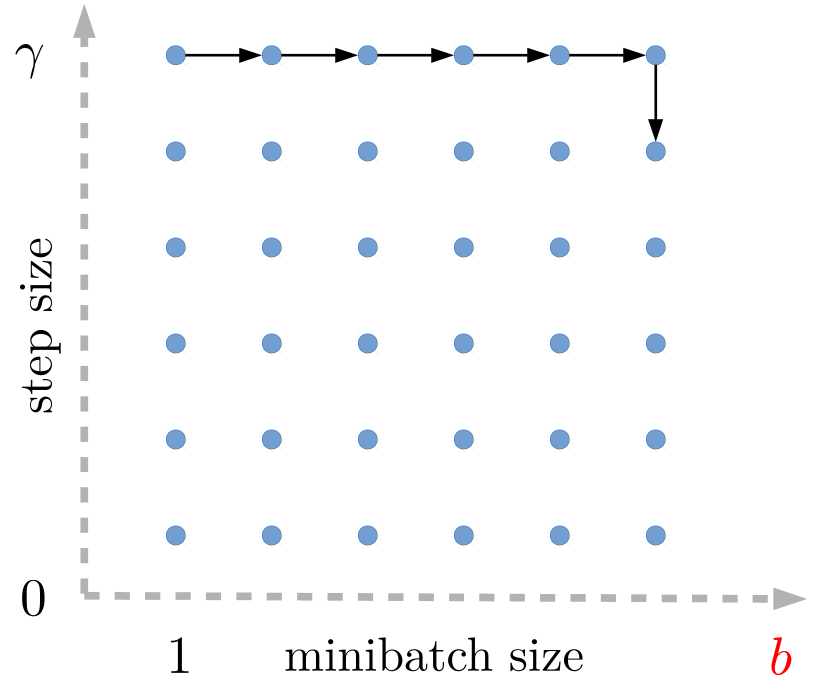
How to choose the minibatch size?



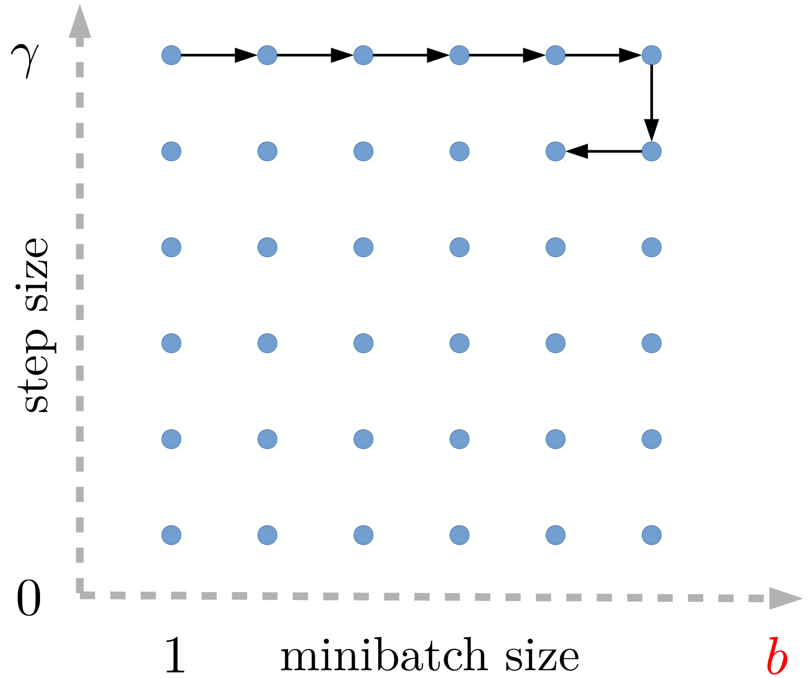
How to choose the minibatch size?



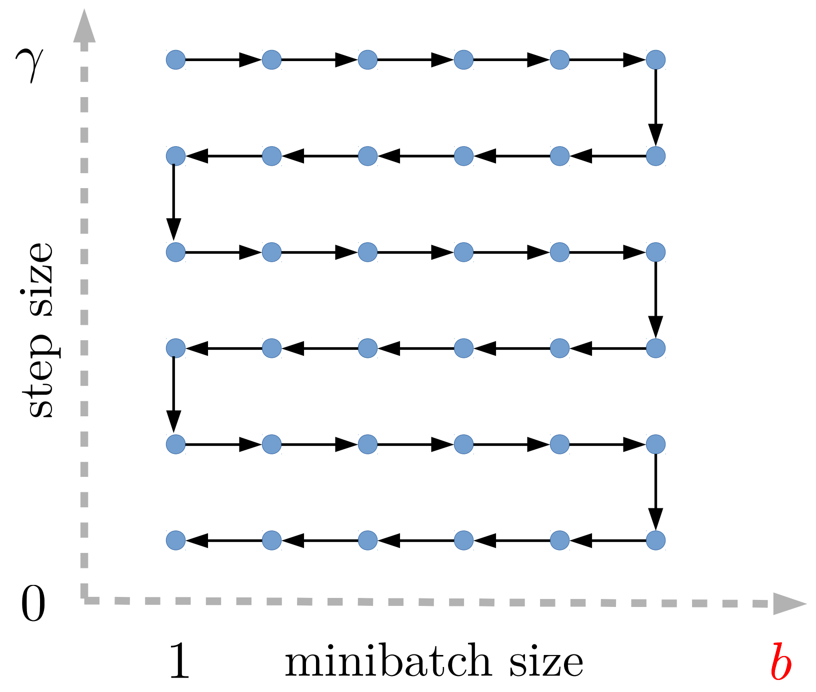
How to choose the minibatch size?



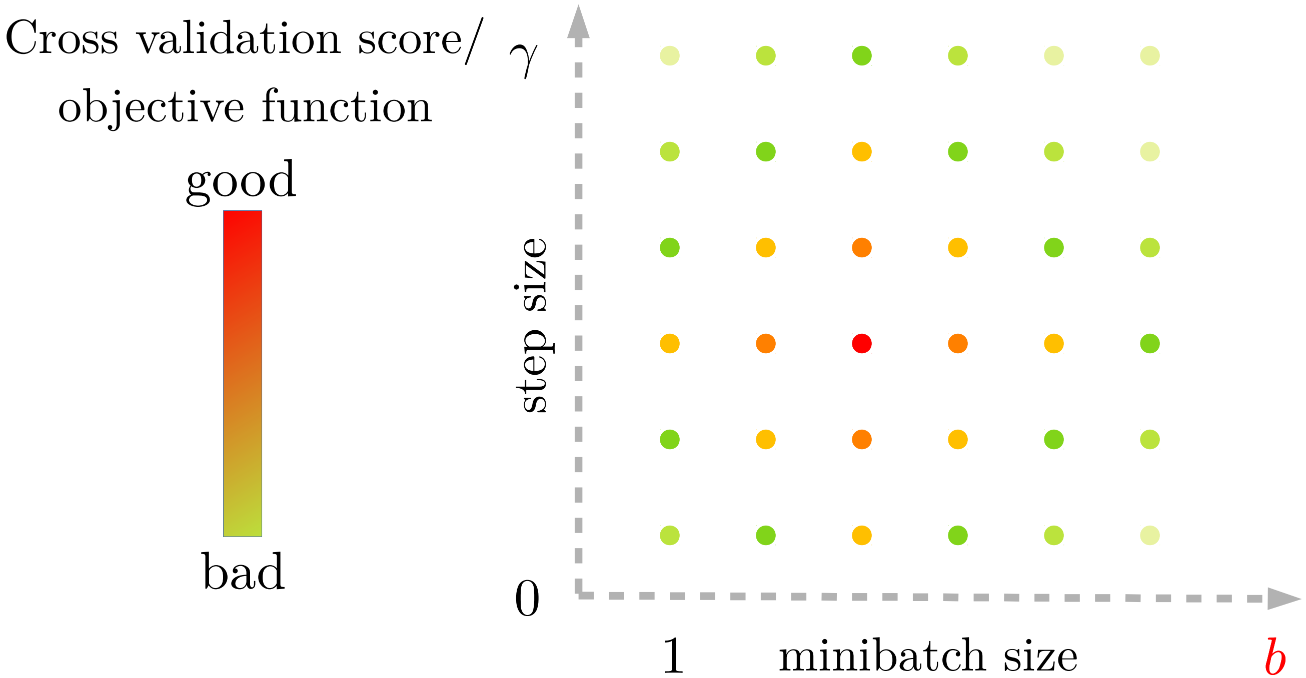
How to choose the minibatch size?



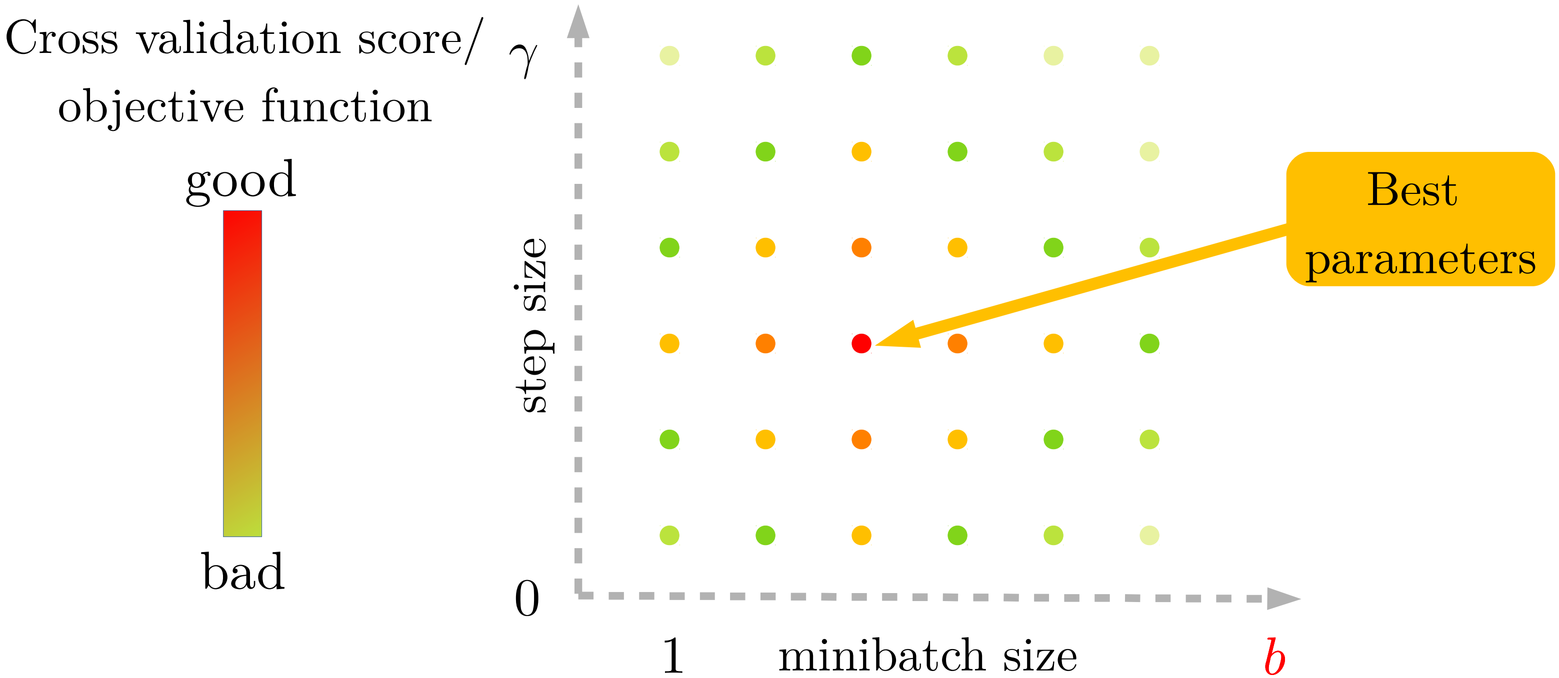
How to choose the minibatch size?



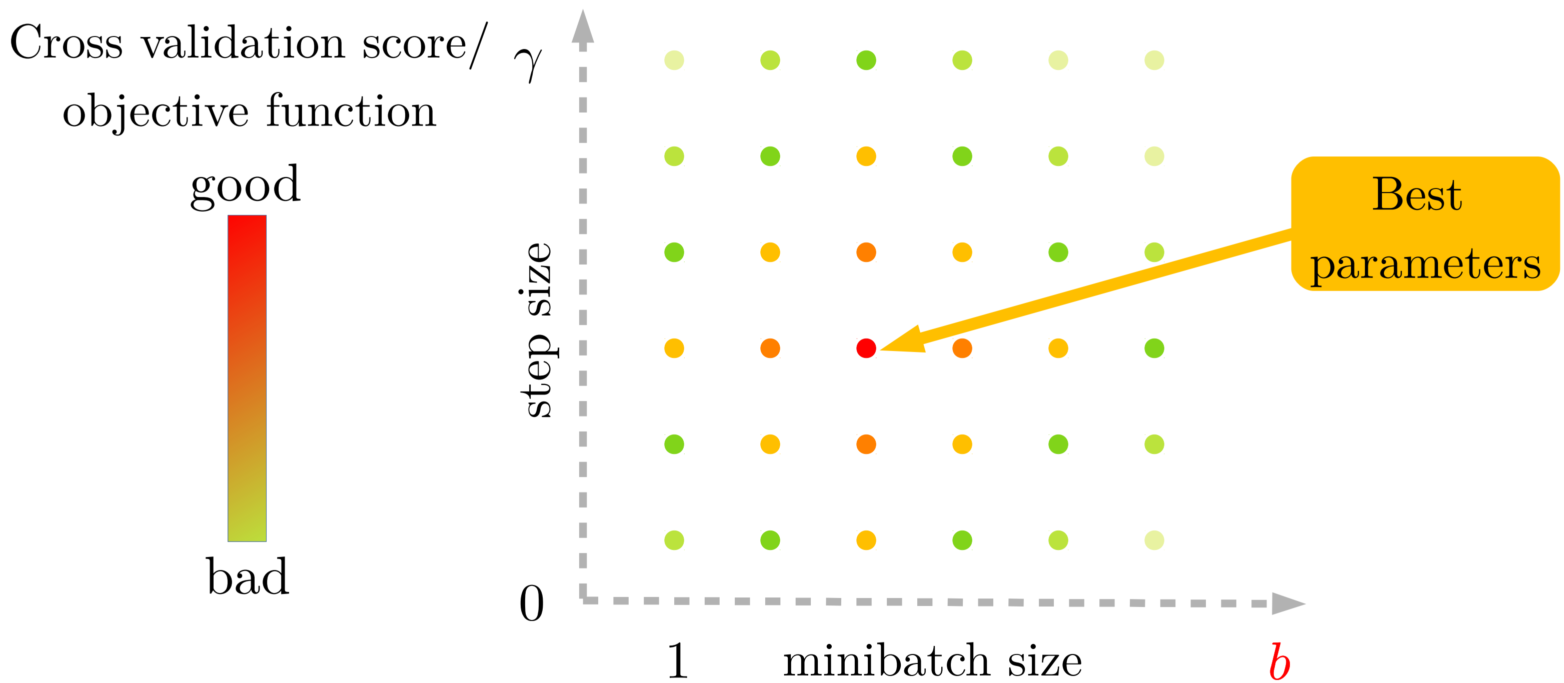
How to choose the minibatch size?



How to choose the minibatch size?



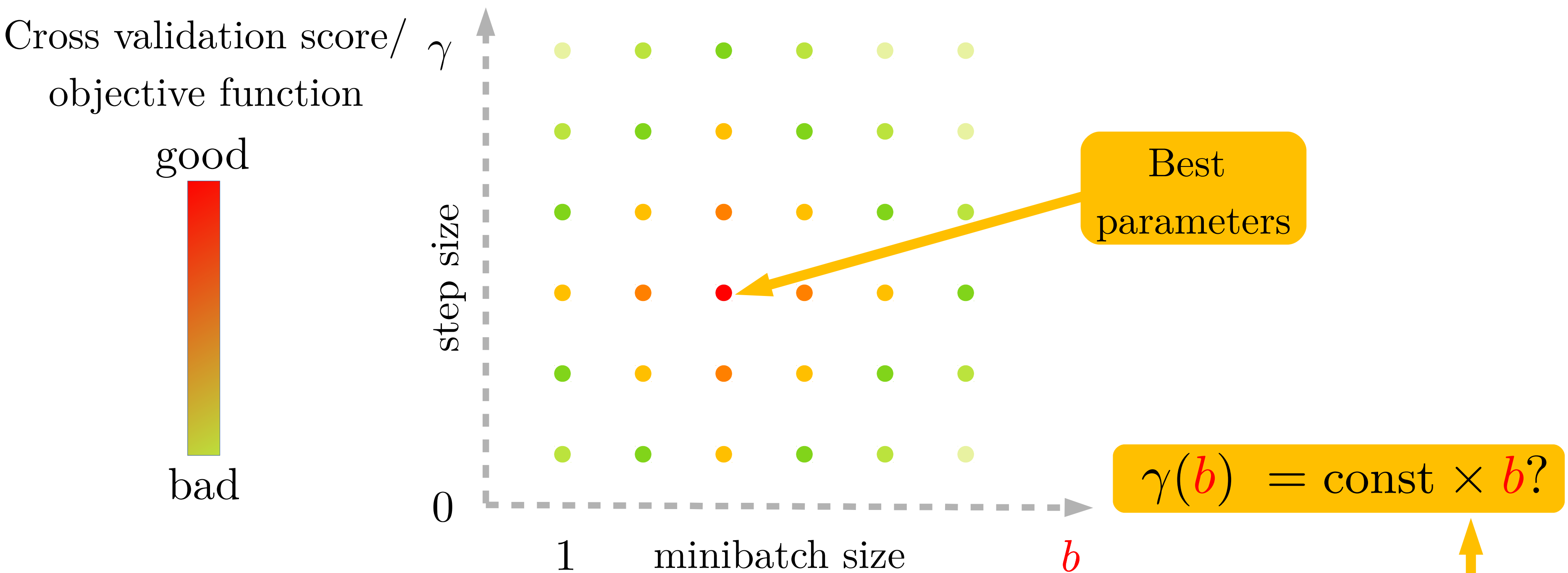
How to choose the minibatch size?



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.

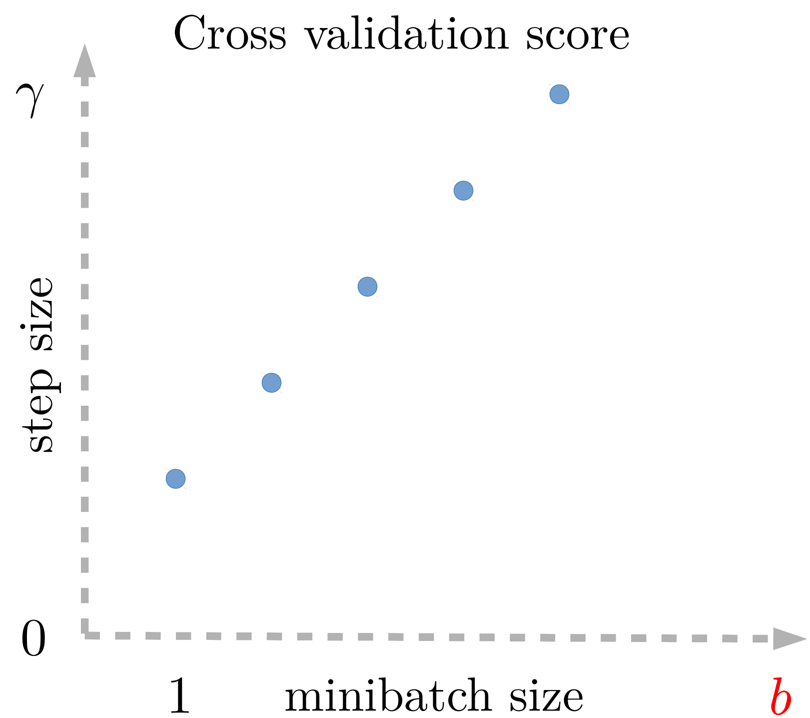
How to choose the minibatch size?



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?

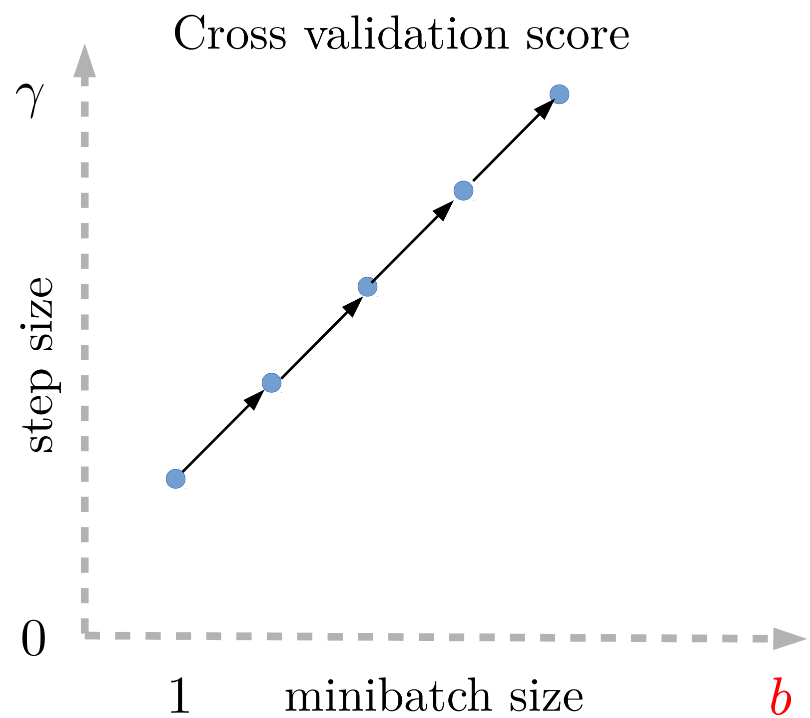


$\gamma(b) = \text{const} \times b$?

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

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

How to choose the minibatch size?

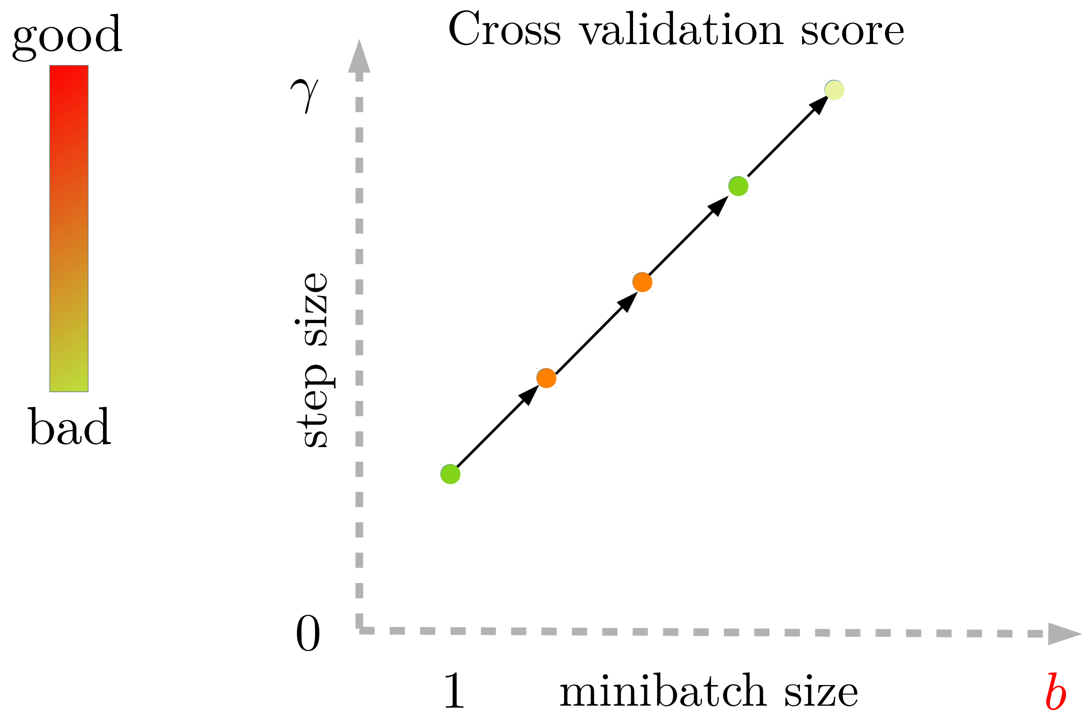


$\gamma(b) = \text{const} \times b?$

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

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

How to choose the minibatch size?

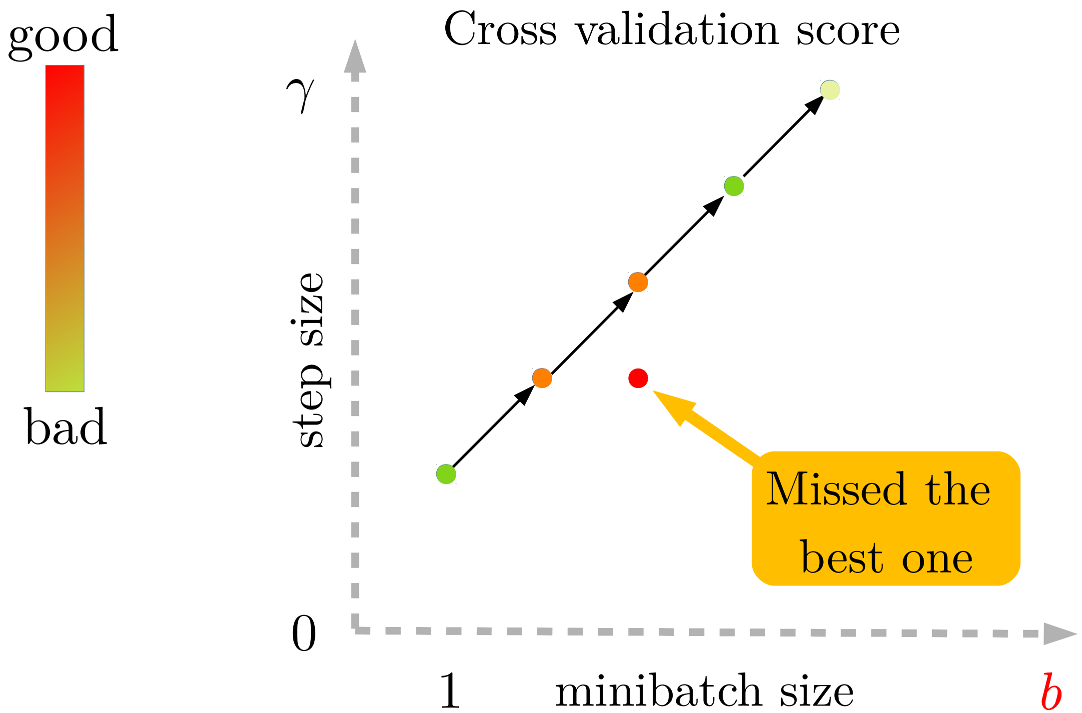


$\gamma(b) = \text{const} \times b$?

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

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

How to choose the minibatch size?

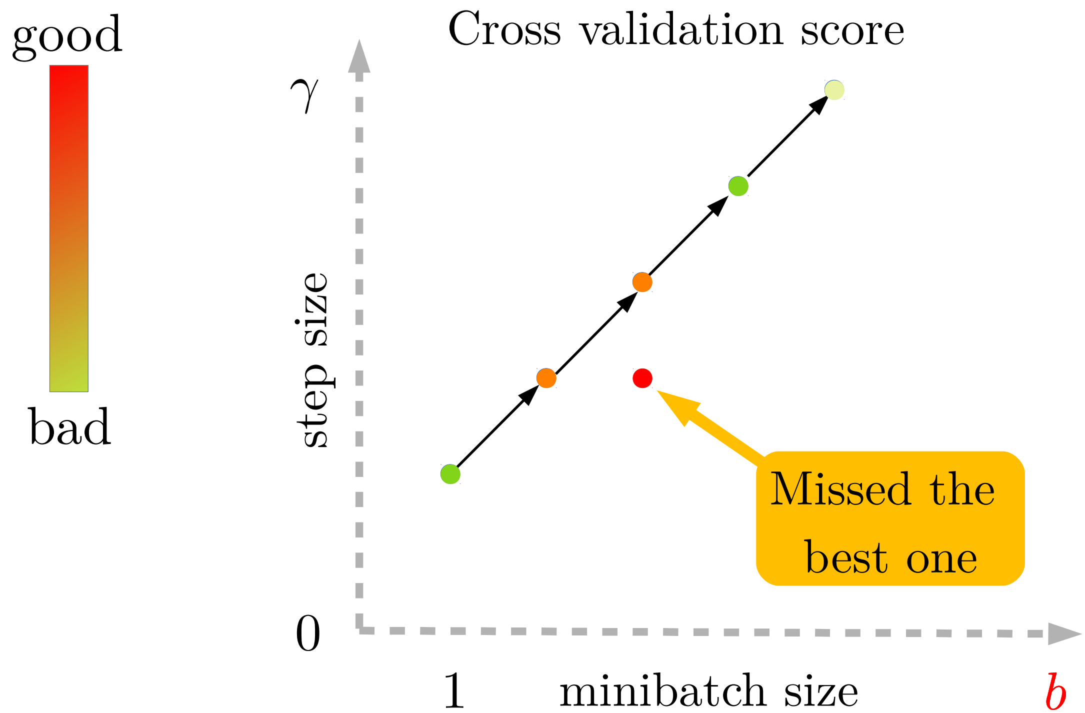


$$\gamma(b) = \text{const} \times b?$$

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

Linear Scaling Rule: When the mini-batch 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 mini-batch 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

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

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

Simple Stochastic Reformulation

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

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

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

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

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

$$\Downarrow$$

Sample $v^t \sim \mathcal{D}$ i.i.d

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

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

SGD with arbitrary sampling

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

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

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

The distribution $\mathcal{D}$ encodes any form of i.i.d mini-batching/ non-uniform sampling.

Example: Gradient descent

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

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

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

The distribution $\mathcal{D}$ encodes any form of i.i.d mini-batching/ non-uniform sampling.

Sample $v^t \sim \mathcal{D}$ i.i.d

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

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

Example: Gradient descent

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

Examples of arbitrary sampling: uniform single element

Random set

$$\mathbb{P}\left[S = \{i\}\right] = 1/\,n, \quad \text{for } i = 1, \ldots, n$$

Examples of arbitrary sampling: uniform single element

Random set

$$\mathbb{P}\left[S = \{i\}\right] = 1/n, \quad \text{for } i = 1, \ldots, n$$

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

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

Examples of arbitrary sampling: uniform single element

Random set

$$\mathbb{P}\left[S = \{i\}\right] = 1/\, n, \quad \text{for } i = 1, \ldots, n$$

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

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

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

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

Examples of arbitrary sampling: uniform single element

Random set

$$\mathbb{P}\left[S = \{i\}\right] = 1/\,n, \quad \text{for } i = 1, \ldots, n$$

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

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

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

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

Single element SGD

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

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

Examples of arbitrary sampling: uniform mini-batching

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

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

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

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

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

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

Mini-batch SGD without replacement

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

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

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

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

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

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

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

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

Examples of arbitrary sampling: non-uniform mini-batching

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

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

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

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

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

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

Arbitrary sampling SGD

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

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

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

SGD with arbitrary sampling

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

Includes all forms of SGD (including GD)

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

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

SGD with arbitrary sampling

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

Includes all forms of SGD (including GD)

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

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

How to analyse this general SGD?

SGD with arbitrary sampling

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

Includes all forms of SGD (including GD)

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

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

How to analyse this general SGD?

Look at the extremes: GD and single element SGD

Assumption and convergence of Gradient Descent and SGD

Reminder: Convergence GD strongly convex + smooth

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\langle \nabla f(w^t), w^* - w^t \rangle + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Now smoothness gives

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2$$

$$\Downarrow$$

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

Assumptions and Convergence of Gradient Descent

quasi strong convexity constant

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

Smoothness constant

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

Assumptions and Convergence of Gradient Descent

quasi strong convexity constant

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

Smoothness constant

$$||\nabla f(w) - \nabla f(w^*)||_2^2 \; \leq \; 2L\left(f(w) - f(w^*)\right) \quad \forall w$$

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

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

Iteration complexity of gradient descent

$$\text{Given } \epsilon > 0 \text{ and } t \geq \frac{L}{\mu} \log\left(\frac{1}{\epsilon}\right) \Rightarrow \frac{\|w^t - w^*\|^2}{\|w^0 - w^*\|^2} \leq \epsilon$$

Assumptions and Convergence of Stochastic Gradient Descent

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

Bigger smoothness constant/ stronger assumption

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

Assumptions and Convergence of Stochastic Gradient Descent

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

Bigger smoothness constant/ stronger assumption

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

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

Assumptions and Convergence of Stochastic Gradient Descent

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

Bigger smoothness constant/ stronger assumption

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

$$w^{t+1} = w^t - \frac{1}{2L_{\max}}\nabla f_j(w^t)$$

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

Iteration complexity of SGD

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

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

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

$$\frac{\mathbb{E}[\|w^t - w^*\|^2]}{\|w^0 - w^*\|^2} \leq \epsilon$$

Informal comparison between GD and SGD iteration complexity

GD

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

SGD

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

$$\frac{\mathbb{E}[\|w^t - w^*\|^2]}{\|w^0 - w^*\|^2} \leq \epsilon$$

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

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

$$\frac{\mathbb{E}[\|w^t - w^*\|^2]}{\|w^0 - w^*\|^2} \leq \epsilon$$

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

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

$$\frac{\mathbb{E}[\|w^t - w^*\|^2]}{\|w^0 - w^*\|^2} \leq \epsilon$$

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

$$L \leq \, ? \; L(v) \; ? \; \leq L_{\max}$$

Key constant: Expected smoothness

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

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

Key constant: Expected smoothness

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

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

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

Key constant: Expected smoothness

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

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

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

Expected smoothness constant

Depends on v and f

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

Key constant: Expected smoothness

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

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

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

Expected smoothness constant

Depends on v and f

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

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

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

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

Expected smoothness constant

Depends on v and f

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

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

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

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

Expected smoothness constant

Depends on v and f

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

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

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

Key constant: Expected smoothness

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

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

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

Expected smoothness constant

Depends on v and f

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

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

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

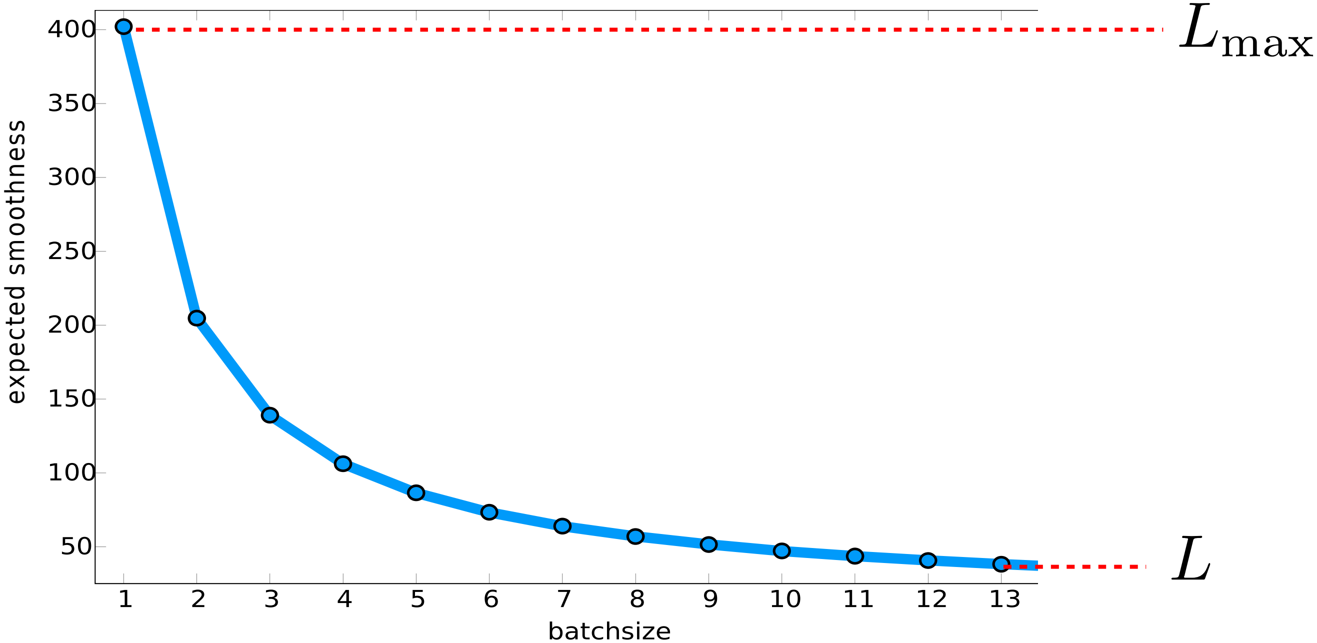
Example of Expected Smoothness

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

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

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

EXE: In your list!



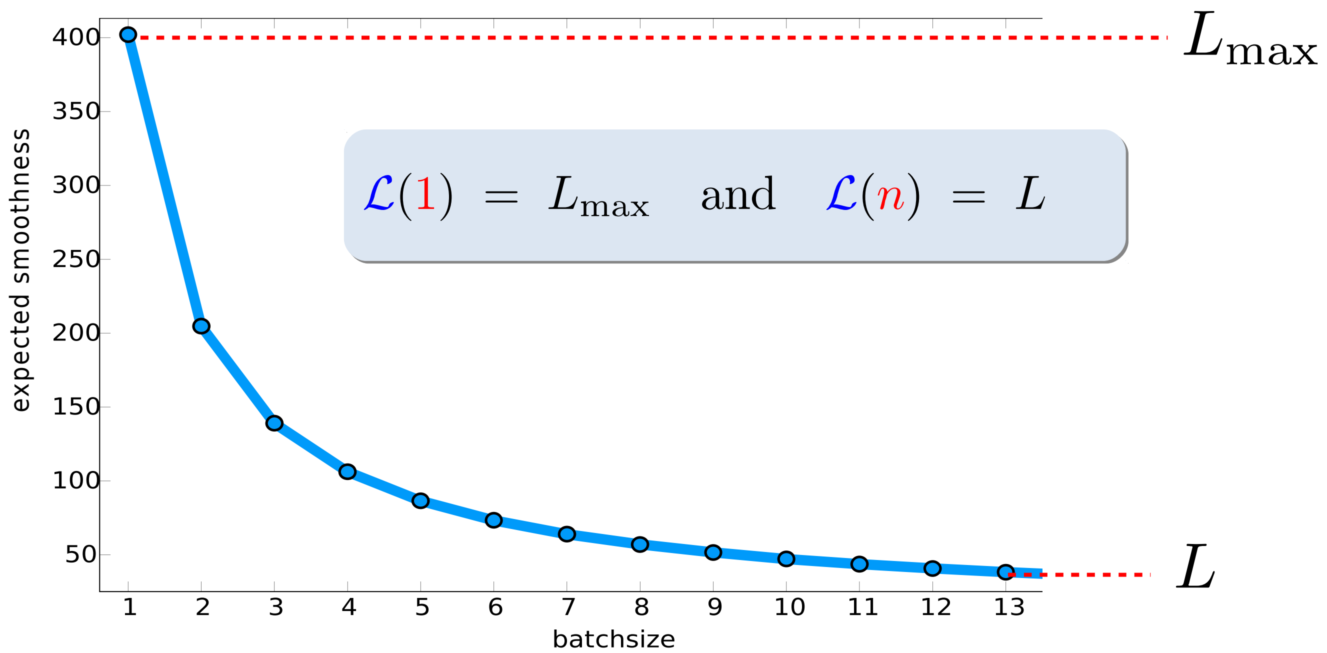
Example of Expected Smoothness

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

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

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

EXE: In your list!



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

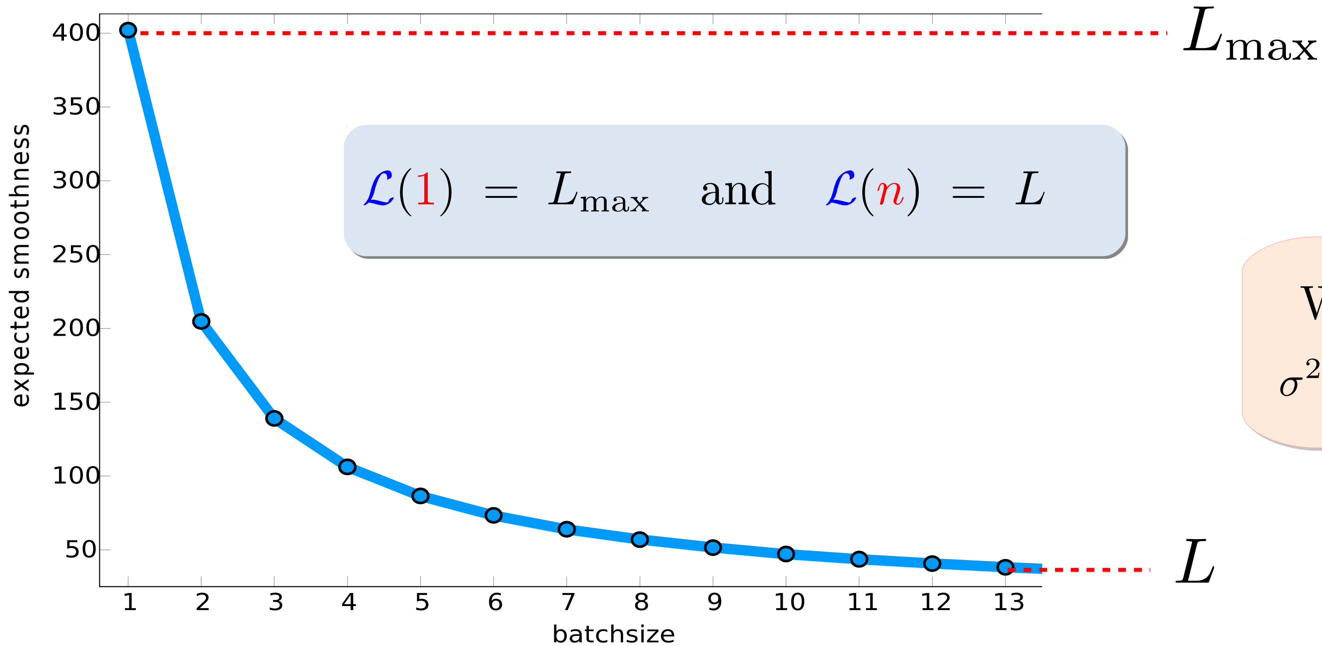
Example of Expected Smoothness

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

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

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

EXE: In your list!



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

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

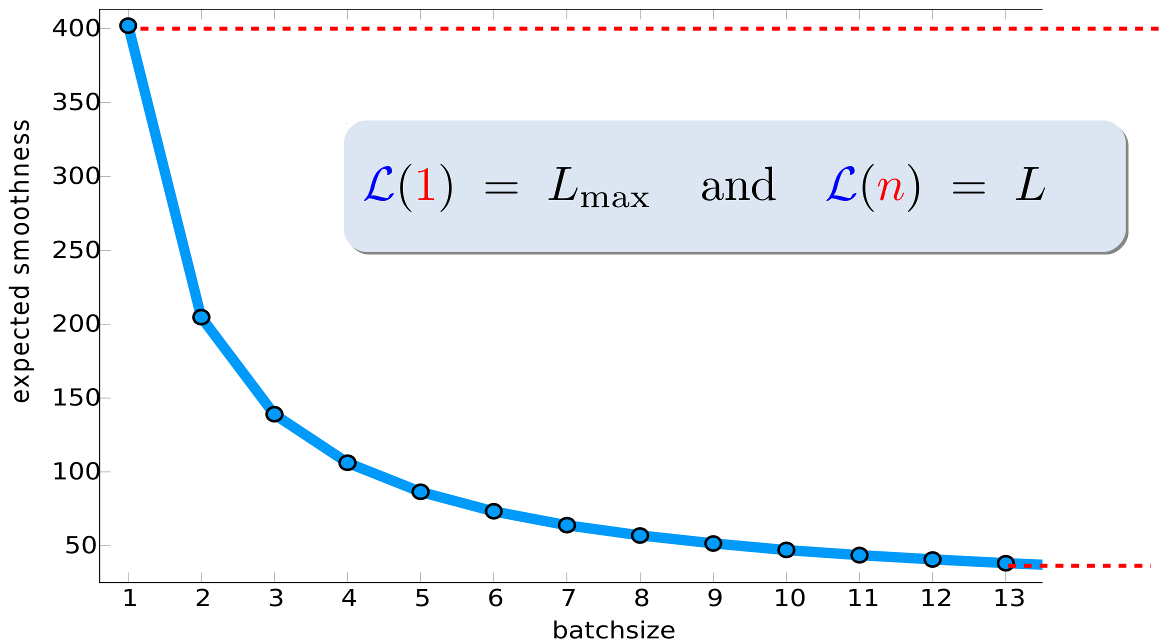
$$\mathcal{L}(b) = \frac{n(b-1)}{b(n-1)}L + \frac{n-b}{b(n-1)}L_{\max}$$

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

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

EXE: In your list!

Measures how much model fits data



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

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

What about σ^2?

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

L $\sigma^2 = 0$

Expected smoothness gives awesome bound on 2nd moment

Normally bound on gradient is an *assumption*

Assumption There exists $B>0$

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

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

Hazan & Kale, JMLR 2014.

Rakhlin, Shamir, & Sridharan, ICML 2012

Shamir & Zhang, ICML 2013.

$\forall w$

Expected smoothness gives awesome bound on 2nd moment

Normally bound on gradient is an *assumption*

Assumption There exists $B>0$

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

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

Hazan & Kale, JMLR 2014.

Rakhlin, Shamir, & Sridharan, ICML 2012

Shamir & Zhang, ICML 2013.

$\forall w$

Expected smoothness gives awesome bound on 2nd moment

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.

$\forall w$

Expected smoothness gives awesome bound on 2nd moment

Normally bound on gradient is an *assumption*

Assumption There exists $B > 0$

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

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

Hazan & Kale, JMLR 2014.

Rakhlin, Shamir, & Sridharan, ICML 2012

Shamir & Zhang, ICML 2013.

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

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

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

Expected smoothness gives awesome bound on 2nd moment

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.

informative: with realistic assumptions

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

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

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

Main Theorem (Linear convergence to a neighborhood)

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

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

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)

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

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

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

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

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

Main Theorem (Linear convergence to a neighborhood)

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

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

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

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

Main Theorem (Linear convergence to a neighborhood)

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

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

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

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

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

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 $v \sim \mathcal{D}$

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

quasi strong conv

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

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

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

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

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

Exercises on Sampling, Expected Smoothness + gradient noise

Optimal mini-batch sizes

Total complexity for mini-batch SGD

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

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

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

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

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

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

Total complexity for mini-batch SGD

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

Total Complexity =
#stochastic gradient calculated
in each iteration

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

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

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

Total Complexity = #stochastic gradient calculated in each iteration

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

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

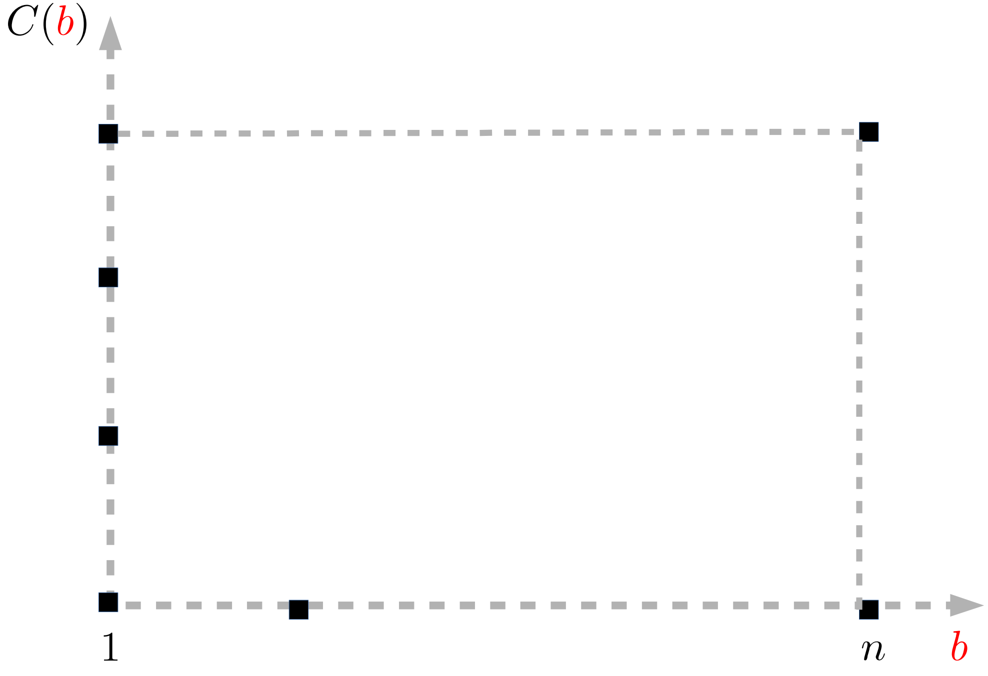
$$\left.\begin{aligned}\mathcal{L} &= \frac{n(b-1)}{b(n-1)}L + \frac{n-b}{b(n-1)}L_{\max} \\ \sigma^2 &= \frac{n-b}{b(n-1)}\sigma_*^2\end{aligned}\right\} \Rightarrow$$

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

Optimal mini-batch size

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

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

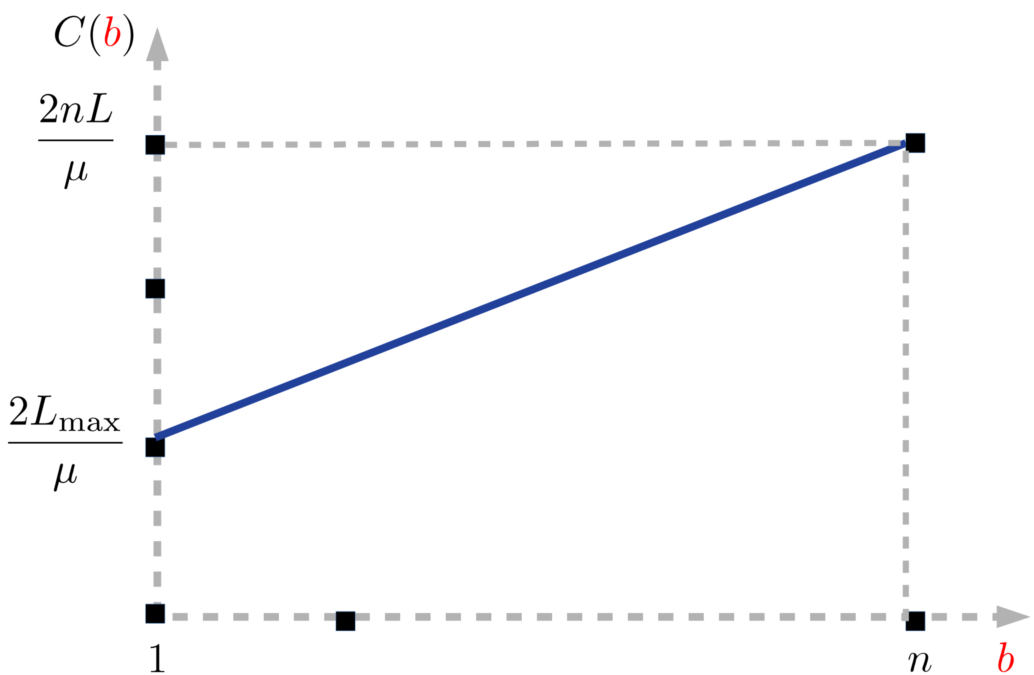


Optimal mini-batch size

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

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

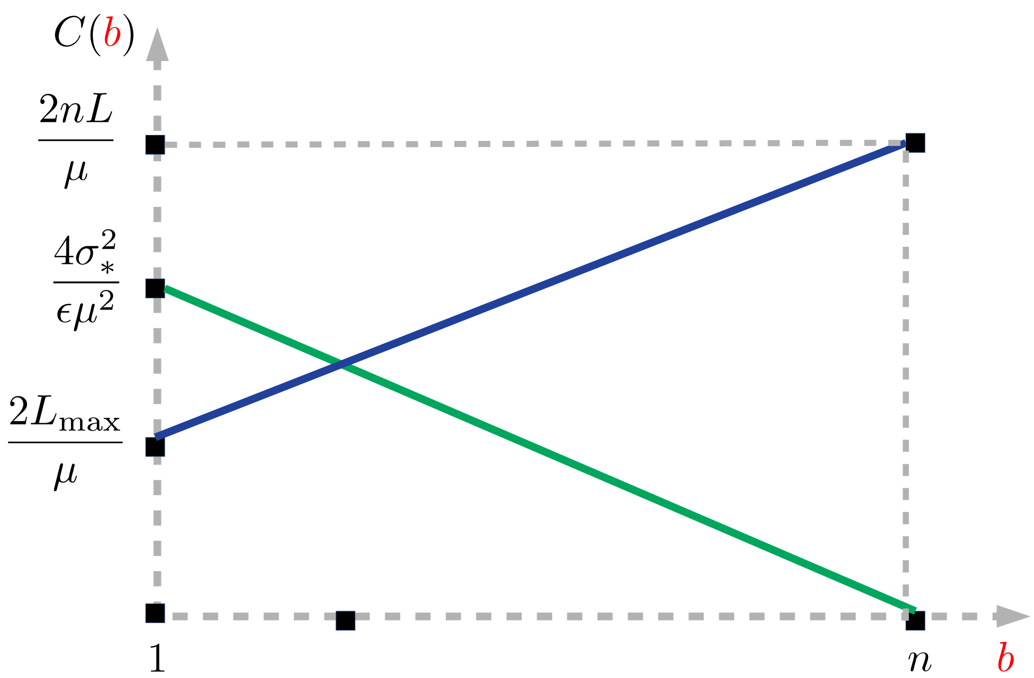
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{\underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \frac{2(n-b)\sigma_*^2}{\epsilon\mu}\right\}$$



Optimal mini-batch size

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

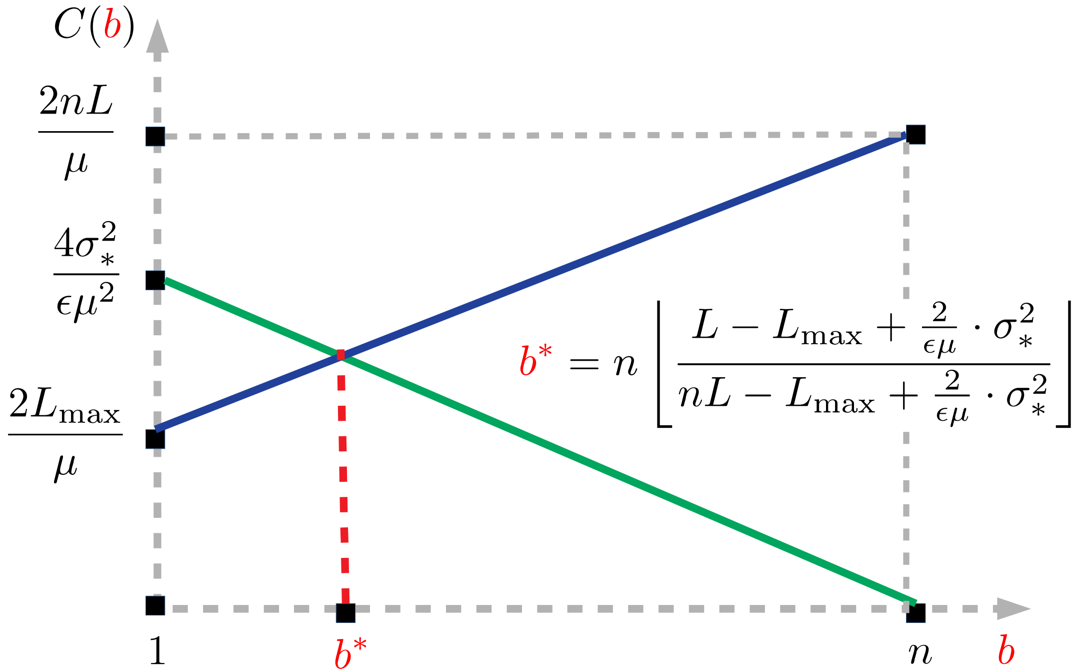
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}}, \; \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$



Optimal mini-batch size

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

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



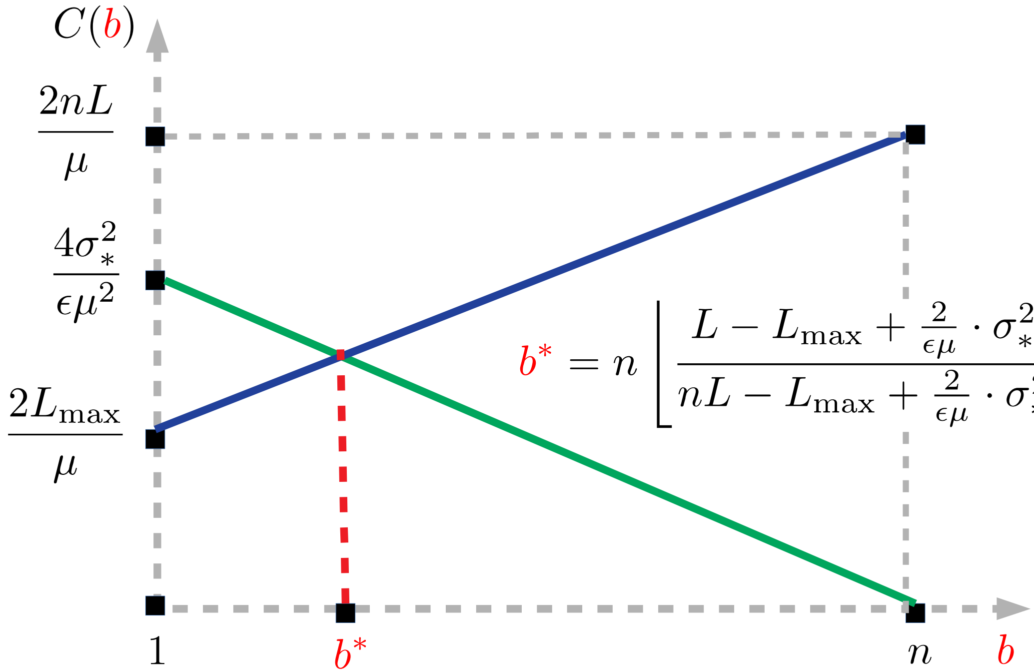
Optimal mini-batch size

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

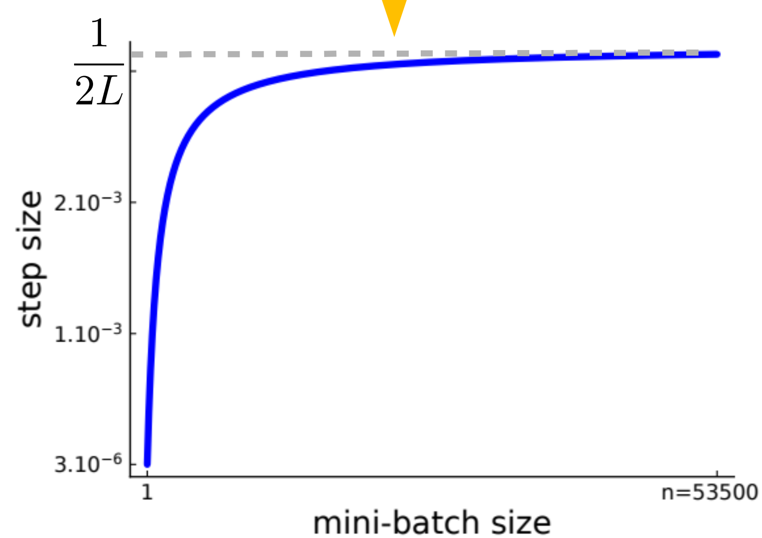
$$C(b) := \frac{2}{\mu(n-1)} \max\left\{ \underbrace{n(b-1)L + (n-b)L_{\max}}_{\text{Linearly increasing}},\ \underbrace{\frac{2(n-b)\sigma_*^2}{\epsilon\mu}}_{\text{Linearly decreasing}} \right\} \times \log\left(\frac{2}{\epsilon}\right)$$

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

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

$$\nabla f_i(w^*) = 0, \; \forall i$$

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

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

Optimal mini-batch size for models that interpolate data

$$\boxed{\nabla f_i(w^*) = 0, \ \forall i}$$

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

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

Optimal mini-batch size for models that interpolate data

$$\boxed{\nabla f_i(w^*) = 0, \ \forall i}$$

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

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

Optimal mini-batch size for models that interpolate data

$$\nabla f_i(w^*) = 0, \ \forall i$$

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

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

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

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

Optimal mini-batch size for models that interpolate data

$$\nabla f_i(w^*) = 0, \ \forall i$$

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

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

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

increases with b

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

$$b^* = 1$$

Optimal mini-batch size for models that interpolate data

$$\nabla f_i(w^*) = 0, \; \forall i$$

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

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

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

increases with b

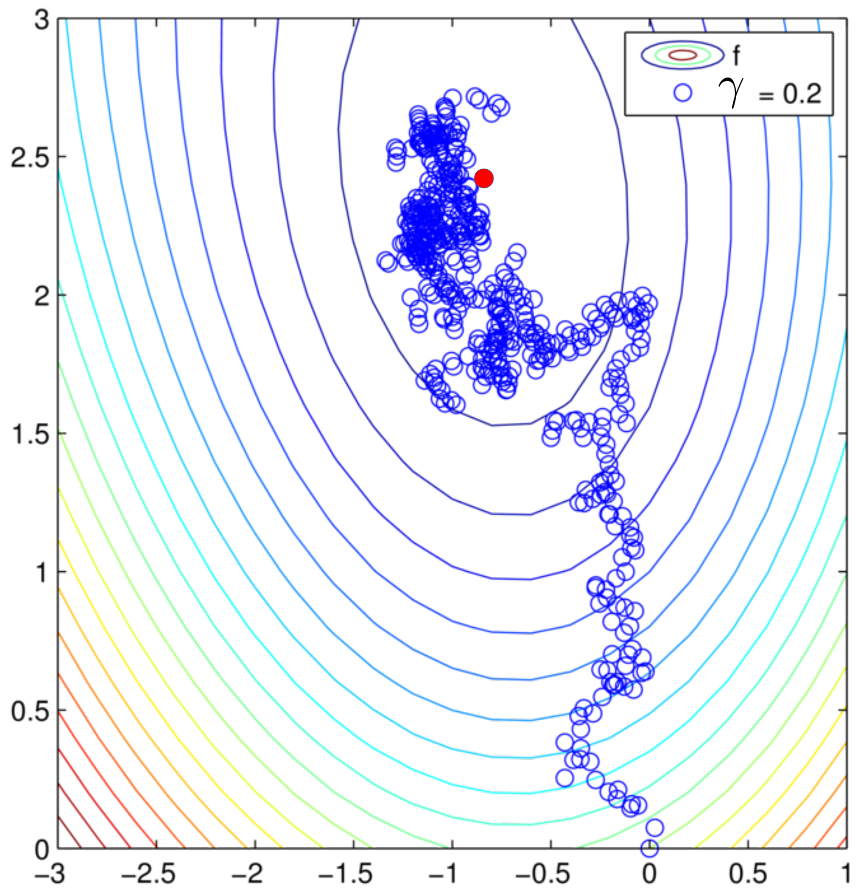
$$\gamma(b) := \frac{n-1}{2} \frac{b}{n(b-1)L + (n-b)L_{\max}}$$

$$b^* = 1$$

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

Learning rate with switch point

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

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

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

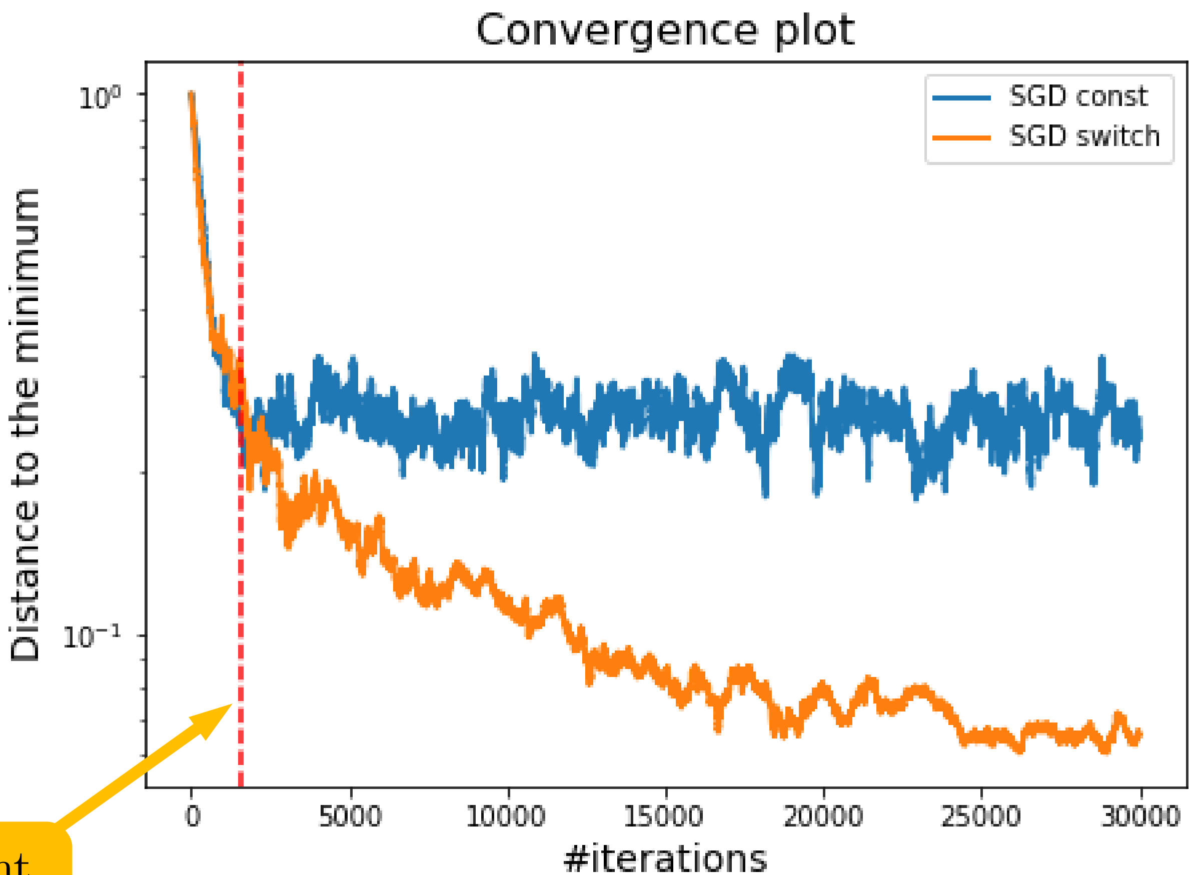
$\sigma^2 := \mathbb{E}[\|\nabla f_v(w^*)\|^2]$

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

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

$\nabla f_i(w^*) = 0, \ \forall i$ $\Rightarrow$ $\mathbb{E}\|w^t - w^*\|^2 \leq O\left(\frac{1}{t^2}\right)$

Stochastic Gradient Descent with switch to decreasing stepsizes



Switch point
$t = 4\lceil \mathcal{L}/\mu \rceil$

Stochastic variance reduced methods

Simple Stochastic Reformulation

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

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

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

What to do about the variance?

Original finite sum problem

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

Stochastic Reformulation

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

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

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

Cancel out

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

Controlled Stochastic Reformulation

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

Cancel out

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

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

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

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

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

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

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

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

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

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

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:

$$g_v(w) \approx \nabla f(w) \Rightarrow \nabla z_v(w) \approx \nabla f_v(w)$$

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:

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

Linear approximation

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

A reference point/ snap shot

SVRG: Stochastic Variance Reduced Gradients

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	$\nabla f_{v^t}(w^t), \quad v^t \sim \mathcal{D}$ Sampled i.i.d
Grad. estimate	$g_{v^t}(w^t) = \nabla f_{v^t}(w^t) - \nabla f_{v^t}(\tilde{w}) + \nabla f(\tilde{w})$

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_{v^t}(w^t), \quad v^t \sim \mathcal{D} \quad$ Sampled i.i.d
Grad. estimate	$g_{v^t}(w^t) = \nabla f_{v^t}(w^t) - \nabla f_{v^t}(\tilde{w}) + \nabla f(\tilde{w})$

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

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

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

Sample $\nabla f_{v^t}(w^t), \quad v^t \sim \mathcal{D} \quad$ Sampled i.i.d

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

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

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

SVRG: Stochastic Variance Reduced Gradients

Johnson & Zhang, 2013 NIPS

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_{v^t}(w^t), \quad v^t \sim \mathcal{D}$ Sampled i.i.d
Grad. estimate	$g_{v^t}(w^t) = \nabla f_{v^t}(w^t) - \nabla f_{v^t}(\tilde{w}) + \nabla f(\tilde{w})$

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

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

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

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

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

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

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

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

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

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

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

Total Complexity of mini-batch SVRG

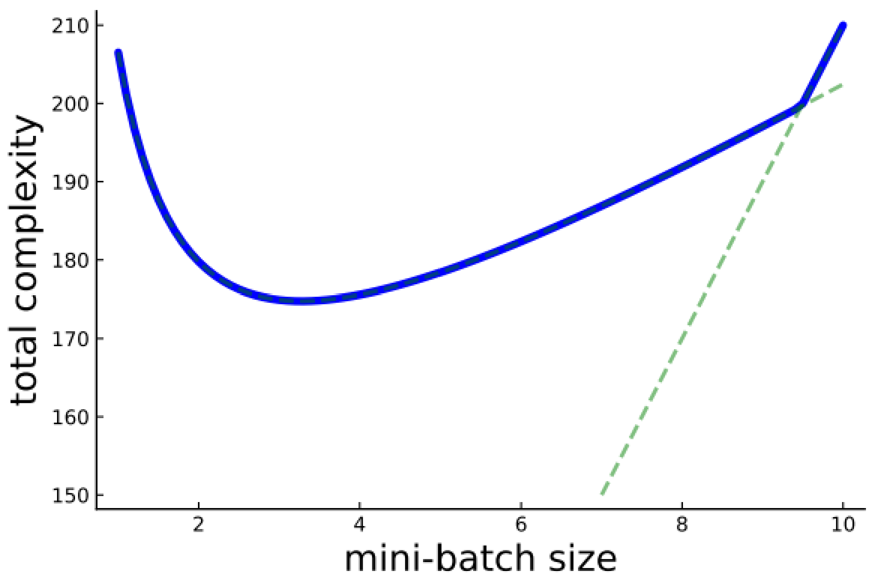
Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

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

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



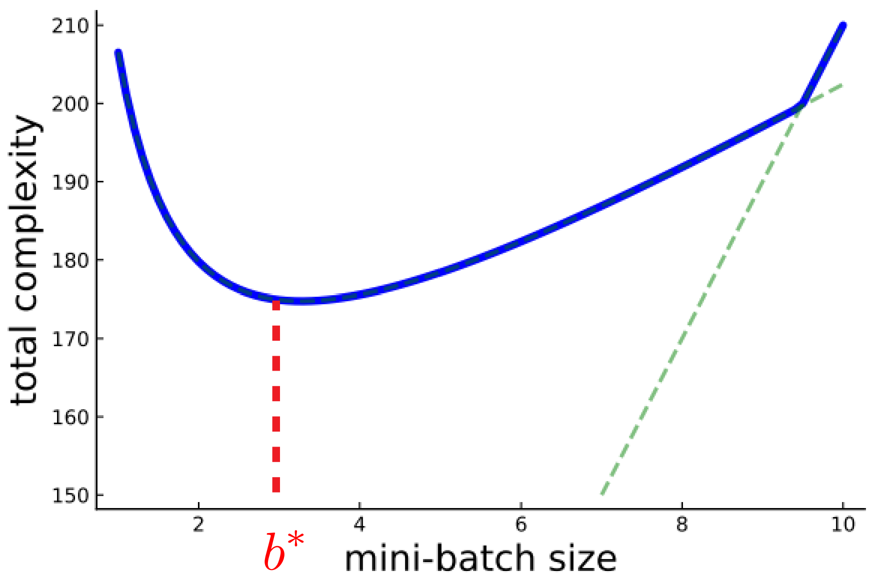
Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Linearly decreasing

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

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



Total Complexity of mini-batch SVRG

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

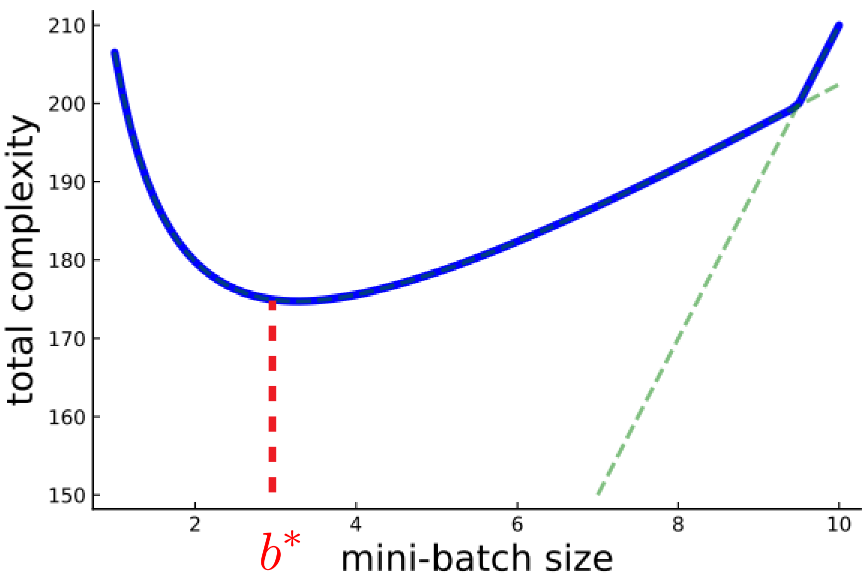
Linearly decreasing

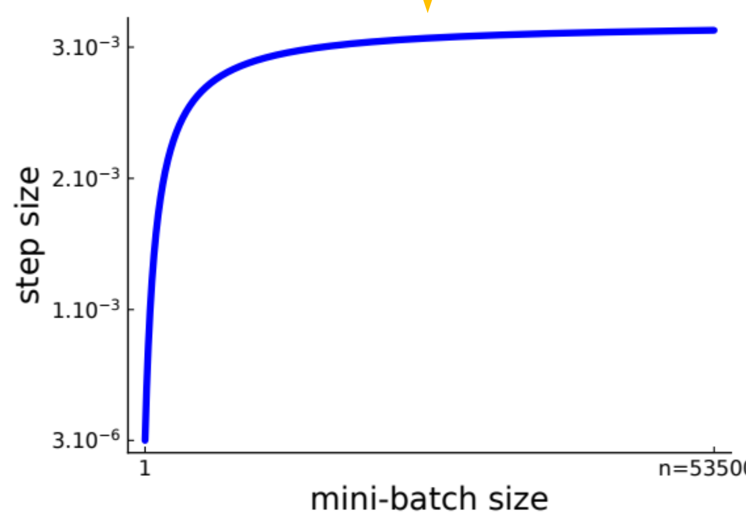
$$C(b) = 2\left(\frac{n}{m}+2b\right)\max\left\{\frac{3}{b}\frac{n-b}{n-1}\frac{L_{\max}}{\mu}+\frac{3n}{b}\frac{b-1}{n-1}\frac{L}{\mu}, m\right\} \times \log\left(\frac{2}{\epsilon}\right)$$

Non-linearly increasing

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

Stepsize increasing with b



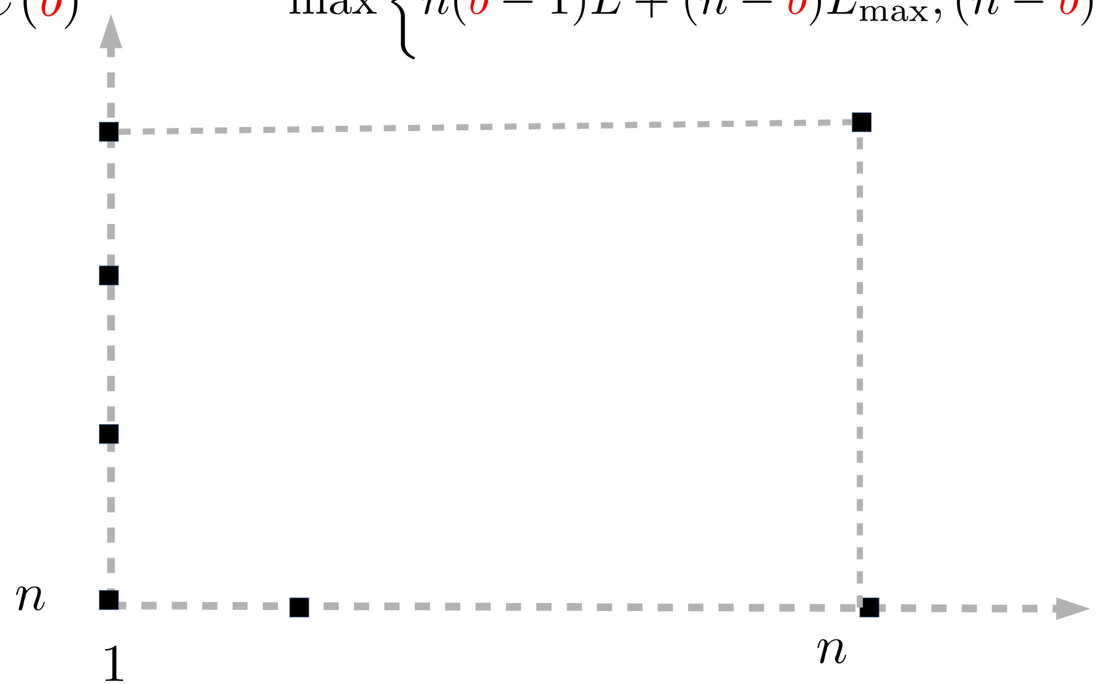


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

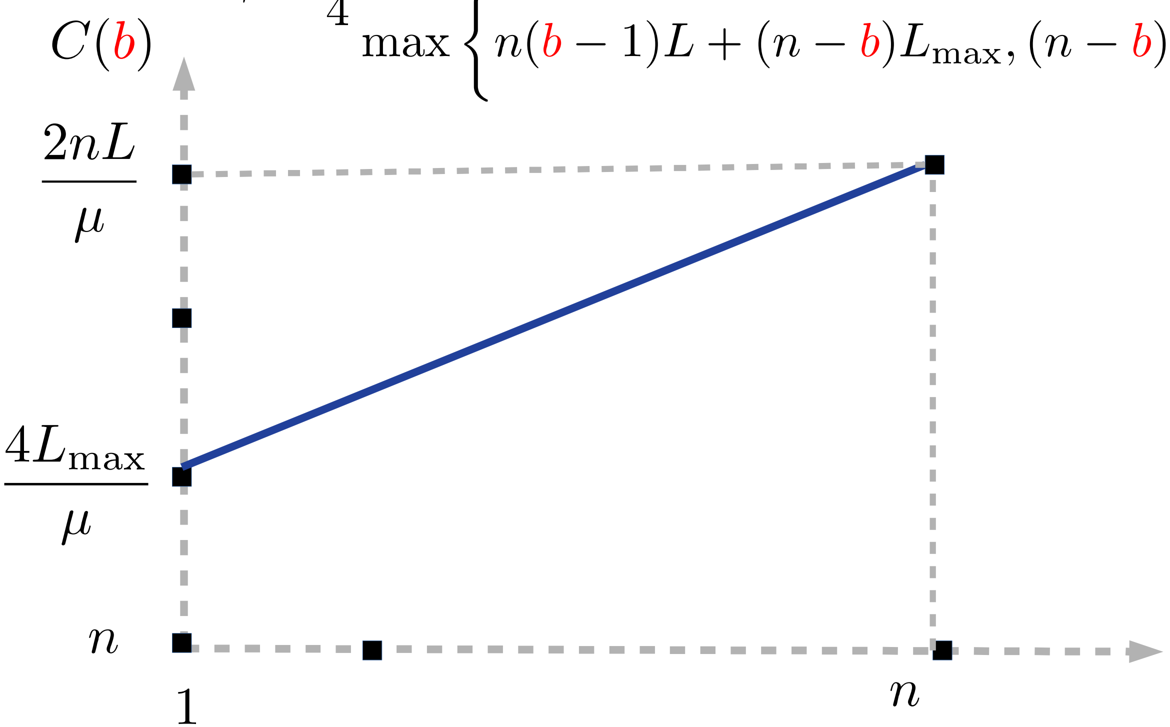


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

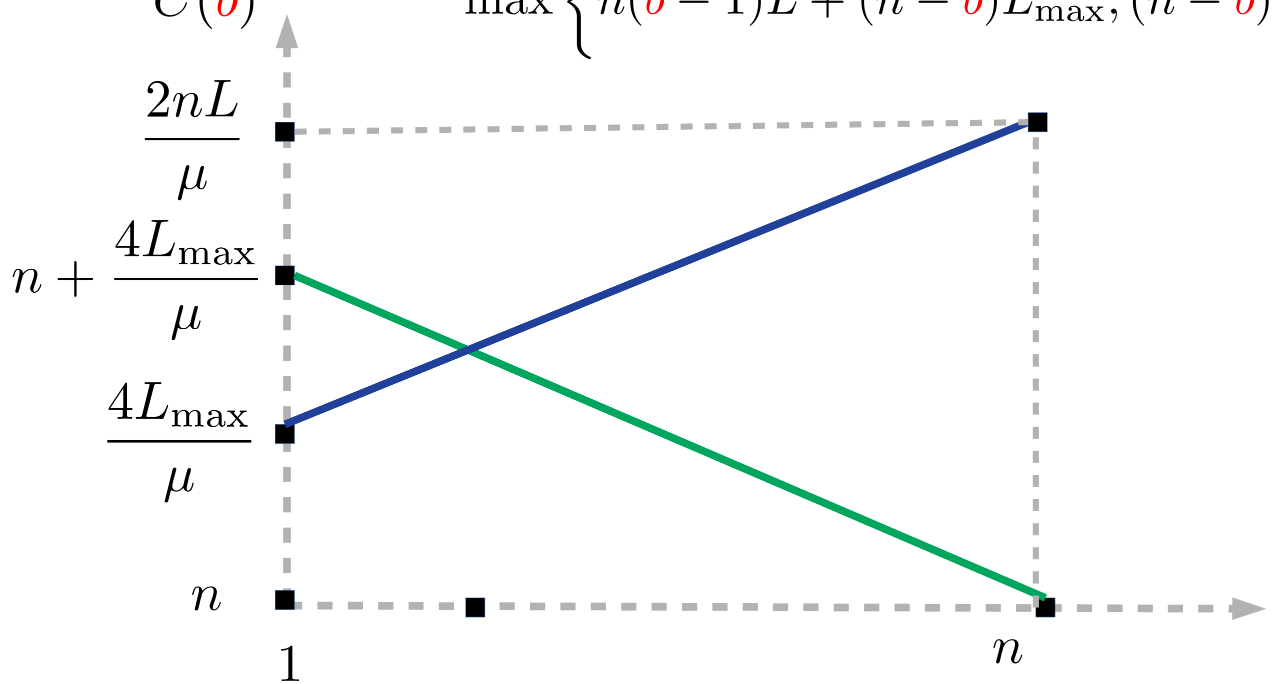


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

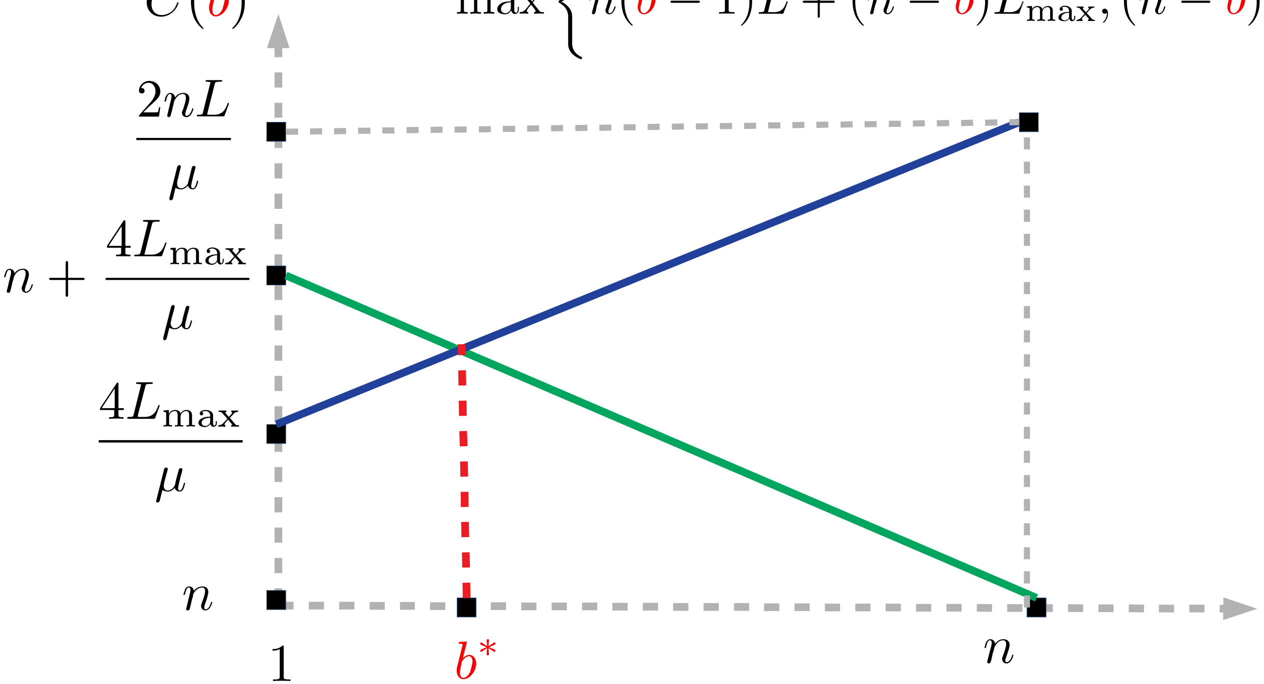


Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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



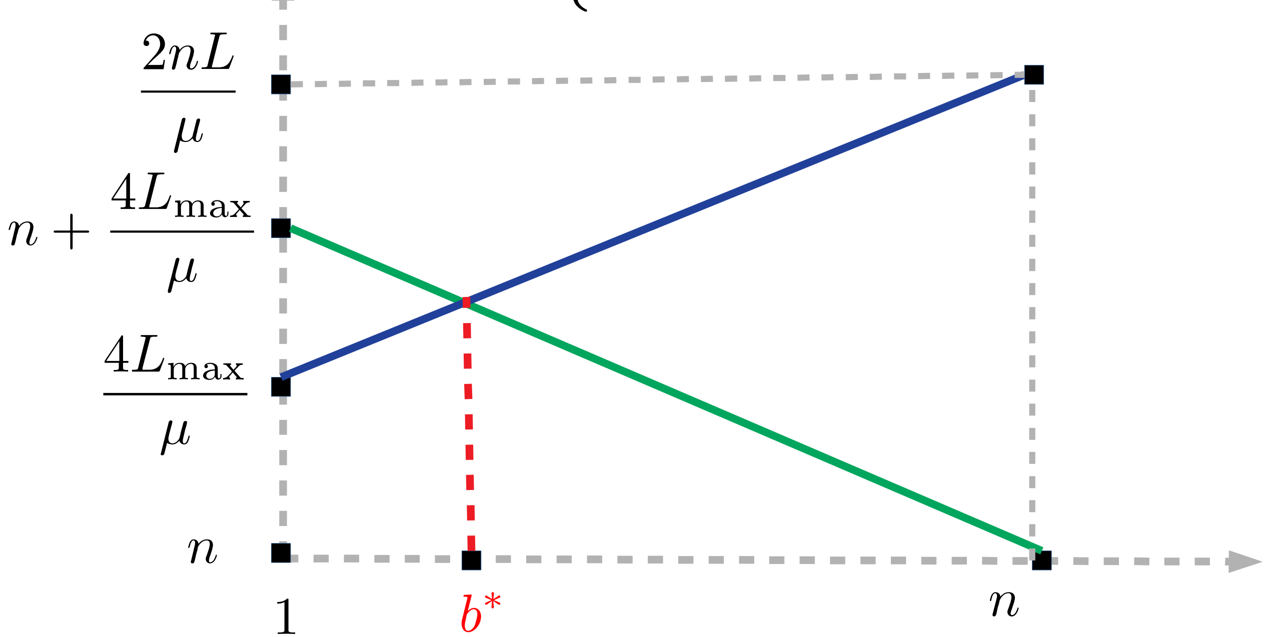
$$b^* = \left\lfloor 1 + \frac{\mu(n-1)}{4L} \right\rfloor$$

Total Complexity of mini-batch SAGA

Gazagnadou, G & Salmon, ICML 2019

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

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

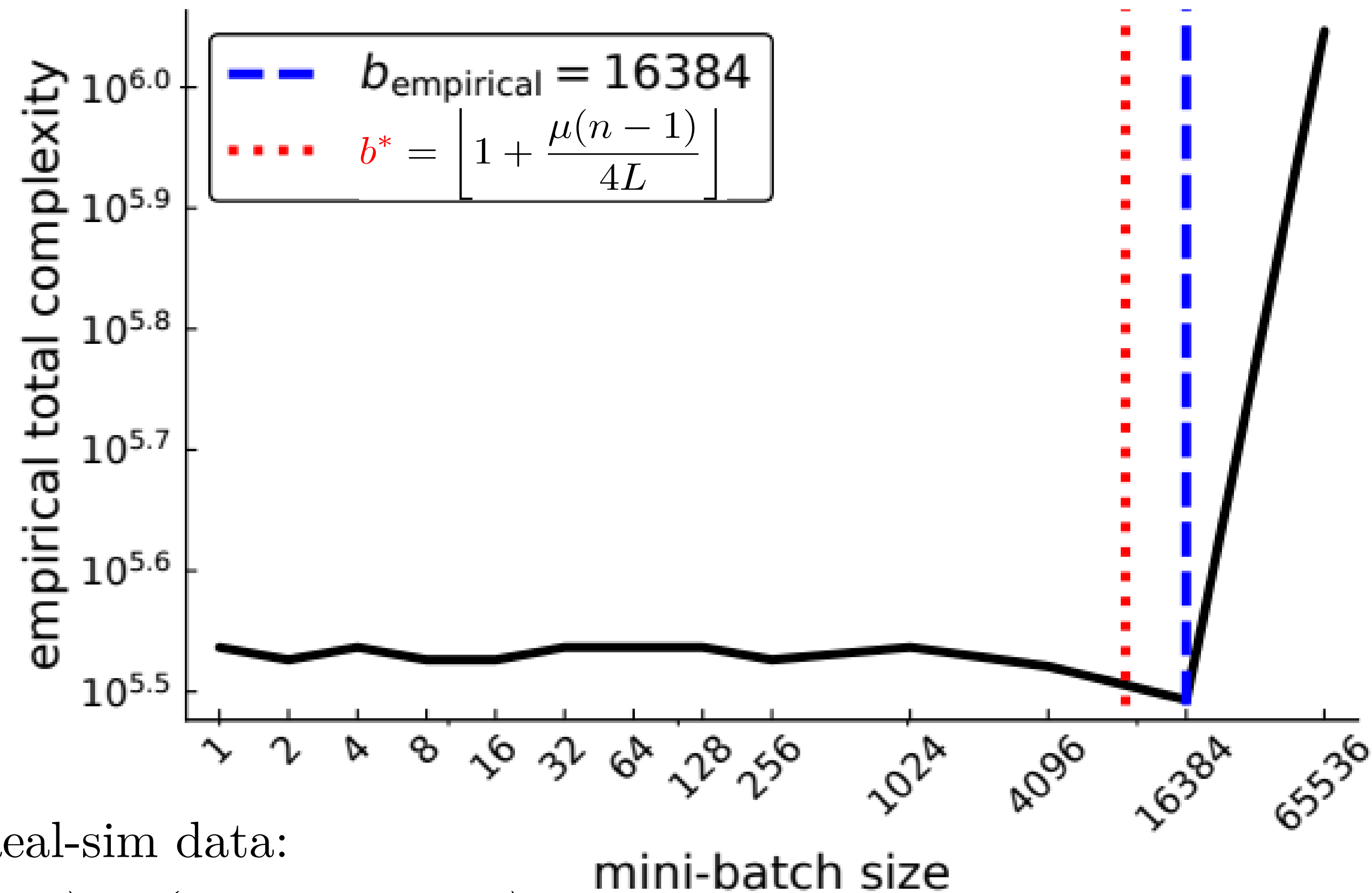


$$b^* = \left\lfloor 1 + \frac{\mu(n-1)}{4L} \right\rfloor$$

Always smaller than 25% of data

Total Complexity of mini-batch SAGA

Predicts good total complexity

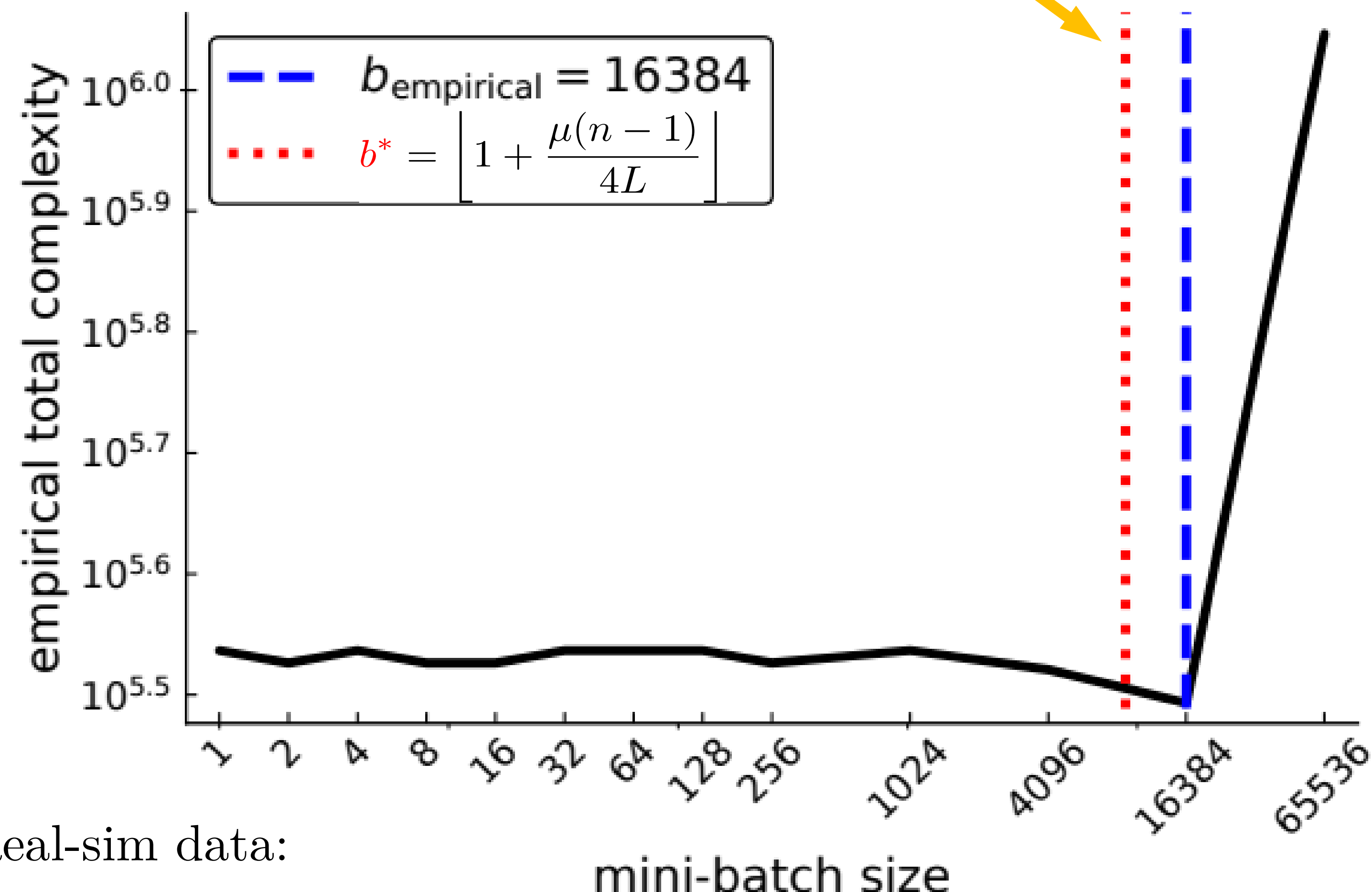


Real-sim data:

(n,d) = (72'309, 20'958)

Total Complexity of mini-batch SAGA

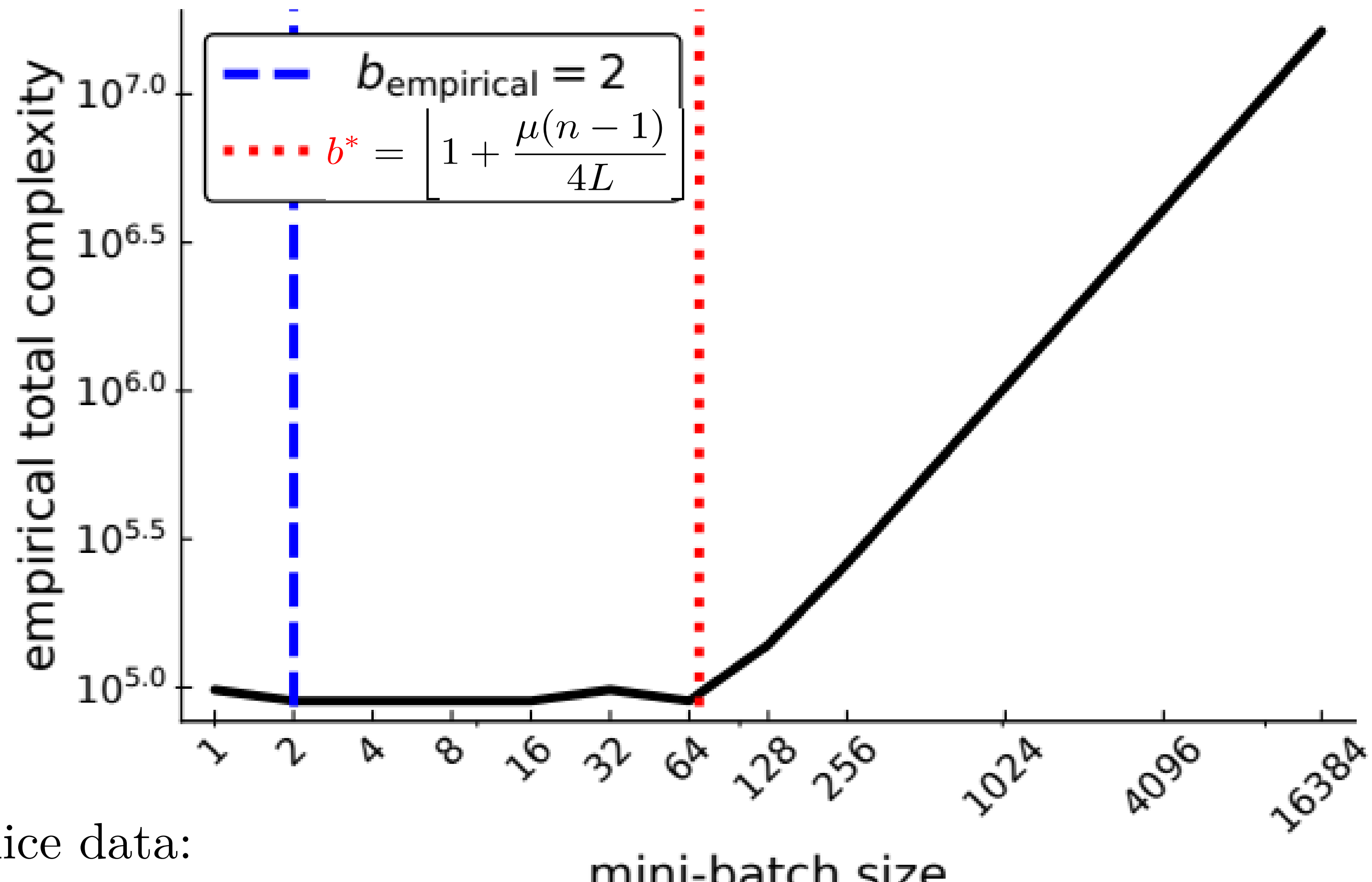
Predicts good total complexity



Real-sim data:

$(n,d) = (72'309,\ 20'958)$

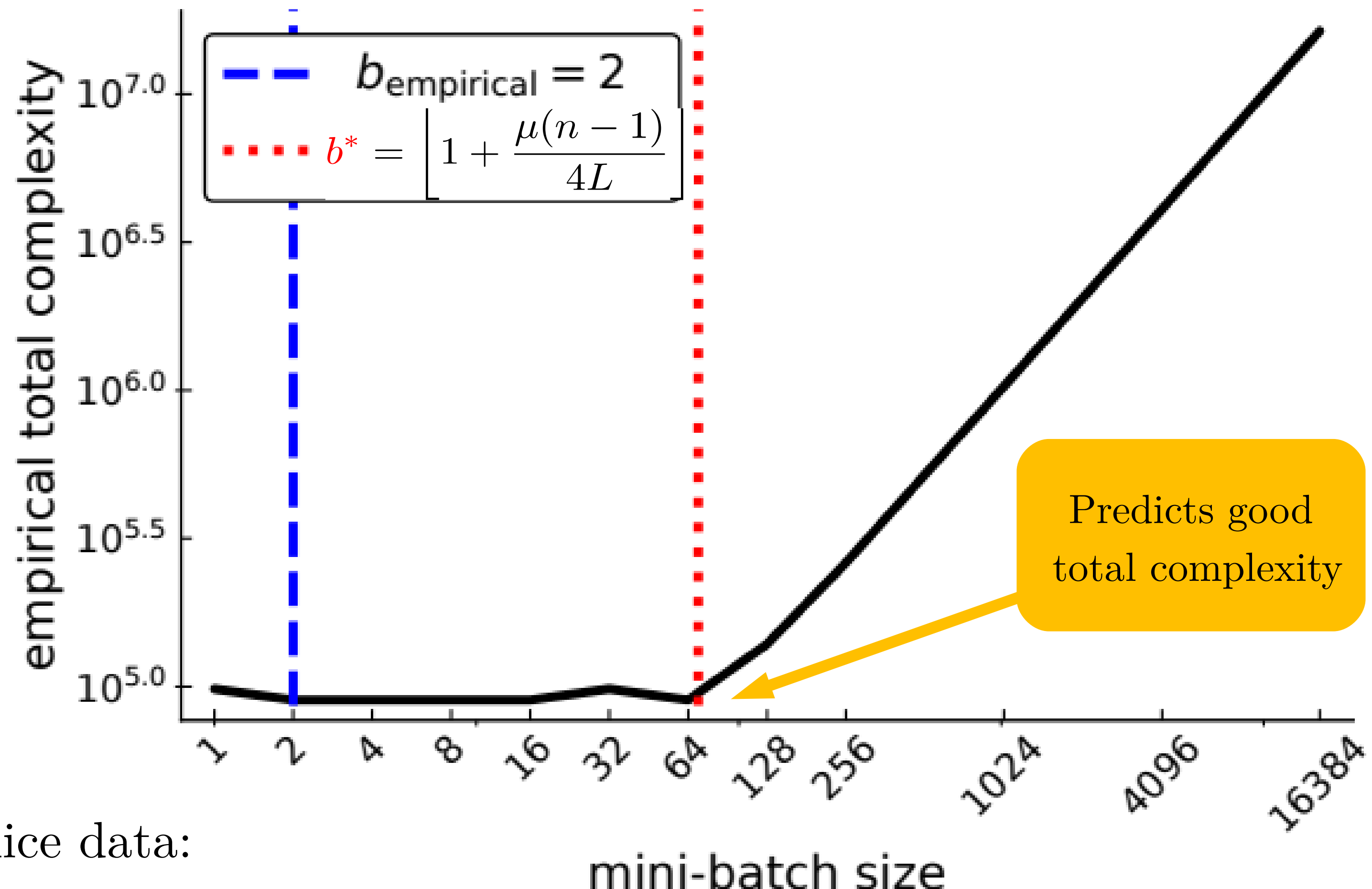
Total Complexity of mini-batch SAGA



Slice data:

$(n,d) = (53'500,\ 386)$

Total Complexity of mini-batch SAGA



Slice data:

(n,d) = (53'500, 386)

Take home message so far

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 so far

Stochastic reformulations allow to view all variants as simple SGD

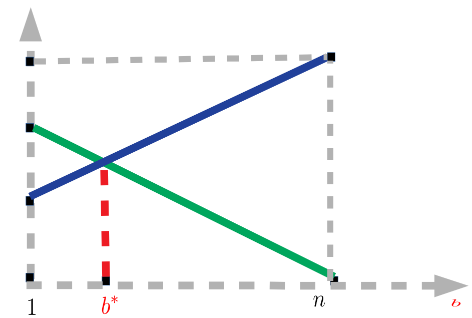
$$\min_{w \in \mathbf{R}^d} \mathbb{E}\left[f_v(w) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(w) \right]$$

To analyse all forms of sampling used through expected smooth

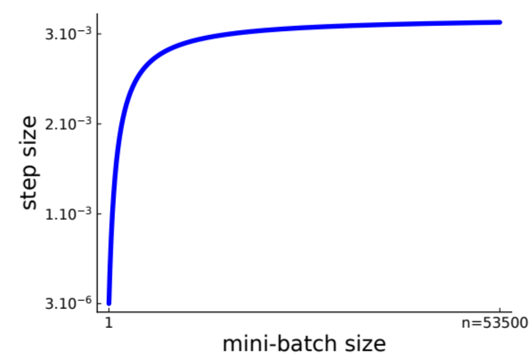
$$\mathbb{E}[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2] \leq \mathcal{L}\,(f(w) - f(w^*))$$

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

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



Stepsize increase by orders when mini-batch size increases



Momentum

Issue with Gradient Descent

Solving the *training problem*: $$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) =: f(w)$$

Baseline method: Gradient Descent (GD)

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

Step size/
Learning rate

Issue with Gradient Descent

Local rate of change

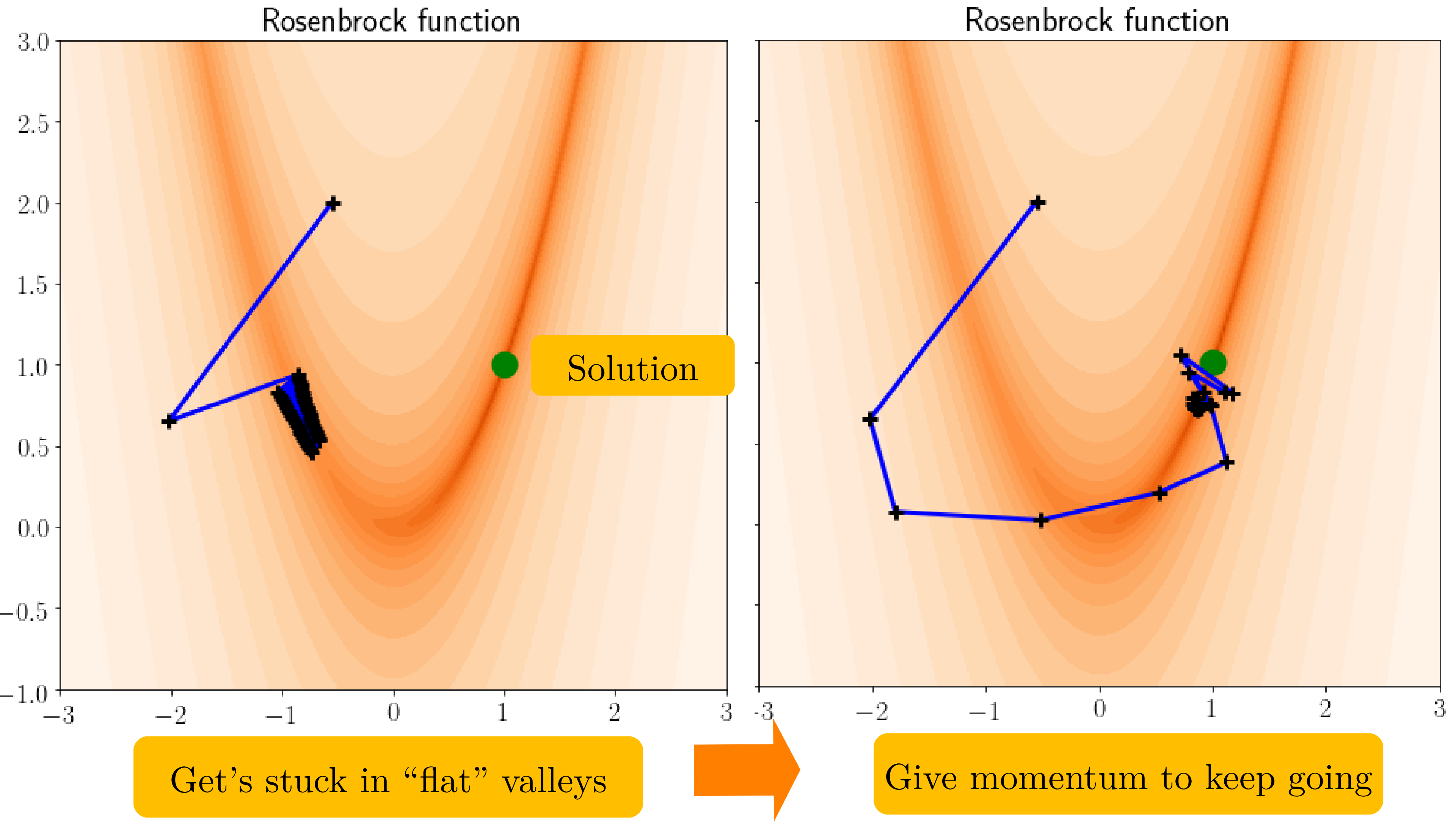
$$\Delta(d) := \lim_{s \to 0^+} \frac{f(x + ds) - f(x)}{s}$$

Max local rate

$$\frac{\nabla f(w^t)}{\|\nabla f(w^t)\|} := \max_{w \in \mathbb{R}^d} \Delta(d)$$
$$\text{subject to} \quad \|d\| = 1$$

GD is the "steepest descent"

Issue with Gradient Descent



Adding some Momentum to GD

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds ``Inertia'' to update}}$$

Adding some Momentum to GD

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds ``Inertia'' to update}}$$

GD with momentum (GDm):

Adds “Momentum” to update

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$

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

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$

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

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$\begin{aligned}
w^{t+1} &= w^t - \gamma\, m^t \\
&= w^t - \gamma\, (\beta m^{t-1} + \nabla f(w^t)) \\
&= w^t - \gamma\, \nabla f(w^t) - \gamma\beta\, m^{t-1} \\
&= w^t - \gamma\, \nabla f(w^t) + \frac{\gamma\beta}{\gamma}\, (w^t - w^{t-1})
\end{aligned}$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned}
w^{t+1} &= w^t - \gamma\, m^t \\
&= w^t - \gamma\,(\beta m^{t-1} + \nabla f(w^t)) \\
&= w^t - \gamma\,\nabla f(w^t) - \gamma\beta\, m^{t-1} \\
&= w^t - \gamma\,\nabla f(w^t) + \frac{\gamma\beta}{\gamma}\,\overbrace{(w^t - w^{t-1})}
\end{aligned}$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned}
w^{t+1} &= w^t - \gamma\, m^t \\
&= w^t - \gamma\,(\beta m^{t-1} + \nabla f(w^t)) \\
&= w^t - \gamma\,\nabla f(w^t) - \gamma\beta\, m^{t-1} \\
&= w^t - \gamma\,\nabla f(w^t) + \frac{\gamma\beta}{\gamma}\,\overbrace{(w^t - w^{t-1})}
\end{aligned}$$

$$w^{t+1} = w^t - \gamma\,\nabla f(w^t) + \beta(w^t - w^{t-1})$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned}
w^{t+1} &= w^t - \gamma\, m^t \\
&= w^t - \gamma\left(\beta m^{t-1} + \nabla f(w^t)\right) \\
&= w^t - \gamma\, \nabla f(w^t) - \gamma\beta\, m^{t-1} \\
&= w^t - \gamma\, \nabla f(w^t) + \frac{\gamma\beta}{\gamma}\overbrace{(w^t - w^{t-1})}
\end{aligned}$$

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma\, \nabla f(w^t) + \beta(w^t - w^{t-1})$$

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let f be μ–strongly convex and L–smooth, that is

stepsize

$$\mu I \quad \preceq \quad \nabla^2 f(w) \quad \preceq \quad LI, \quad \forall w \in \mathbb{R}^d$$

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then SGDm converges

momentum parameter

$$\|w^t - w^*\| \;\leq\; \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|w^0 - w^*\|$$

$\kappa := L/\mu$

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let f be μ–strongly convex and L–smooth, that is

stepsize

$$\mu I \preceq \nabla^2 f(w) \preceq LI, \quad \forall w \in \mathbb{R}^d$$

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then SGDm converges

momentum parameter

$$\Rightarrow \quad \|w^t - w^*\| \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|w^0 - w^*\|$$

$\kappa := L/\mu$

Corollary $t \geq \dfrac{1}{\sqrt{\kappa}+1}\log\left(\dfrac{1}{\epsilon}\right) \quad \Rightarrow \quad \dfrac{\|w^t - w^*\|}{\|w^0 - w^*\|} \leq \epsilon$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)}) ds (w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)})ds(w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$+w^* - w^*$

$$\begin{aligned} w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \\ &= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\ &= \left((1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) - \beta(w^{t-1} - w^*) \end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)})ds(w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \qquad +w^* - w^* \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \left(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=: A_s}\right)(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)}) ds (w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$+w^* - w^*$

$$\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \Big(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=: A_s}\Big)(w^t - w^*) - \beta(w^{t-1} - w^*) \\
&= A_s(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)}) ds (w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$
\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \qquad +w^* - w^* \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \left(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=:\, A_s}\right)(w^t - w^*) - \beta(w^{t-1} - w^*) \\
&= A_s(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}
$$

Depends on past. Difficult recurrence

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Simple recurrence!

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Simple recurrence!

$$\|z^{t+1}\| \leq \left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| \|z^t\|$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \quad \leq \quad \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|} \; \|z^t\|$$

$$\|A\| \quad := \quad \max_{i=1,\ldots,2n} |\lambda_i(A)|$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \quad \leq \quad \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|}_{\|A\| \; := \; \max_{i=1,\ldots,2n} |\lambda_i(A)|} \; \|z^t\|$$

EXE on Eigenvalues:

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then

$$\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| \quad = \quad \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \leq \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|}_{\|A\| := \max_{i=1,\ldots,2n} |\lambda_i(A)|} \|z^t\|$$

$$(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)$$

EXE on Eigenvalues:

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then

$$\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

Adding Momentum to SGD

Rumelhart, Hinton, Geoffrey, Ronald, 1986, Nature

Stochastic Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f_{j_t}(w^t) + \beta(w^t - w^{t-1})$$

Adds "Inertia" to update

SGD with momentum (SGDm):

$$m^t = \beta \, m^{t-1} + \nabla f_{j_t}(w^t)$$

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

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

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^t \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

SGD with momentum (SGDm):

$$w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^t \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

SGD with momentum (SGDm):

$$w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})$$

Acts like an approximate variance reduction since

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^t \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

SGD with momentum (SGDm):

$$w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})$$

Acts like an approximate variance reduction since

$$\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \approx \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(w^t)$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

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 Neurips 2019, preprint online. **Towards closing the gap between the theory and practice of SVRG**