Optimization Methods for Machine Learning
Part II – The theory of SG

Leon Bottou
Facebook AI Research

Frank E. Curtis
Lehigh University

Jorge Nocedal
Northwestern University

Summary

1. Setup
2. Fundamental Lemmas
3. SG for Strongly Convex Objectives
4. SG for General Objectives
5. Work complexity for Large-Scale Learning
6. Comments

1- Setup

The generic SG algorithm

The SG algorithm produces successive iterates $w_k \in \mathbb{R}^d$ with the goal to minimize a certain function $F : \mathbb{R}^d \to \mathbb{R}$.

We assume that we have access to three mechanisms

1. Given an iteration number k ,
   a mechanism to generate a realization of a random variable ξ_k.
   The $\{\xi_k\}$ form a sequence of jointly independent random variables
2. Given an iterate w_k and a realization ξ_k,
   a mechanism to compute a stochastic vector $g(w_k, \xi_k) \in \mathbb{R}^d$
3. Given an iteration number,
   a mechanism to compute a scalar stepsize $\alpha_k > 0$

The generic SG algorithm

Algorithm 4.1 (Stochastic Gradient (SG) Method)

Choose an initial iterate w_1.
for $k = 1, 2, \ldots$ **do**
    Generate a realization of the random variable ξ_k.
    Compute a stochastic vector $g(w_k, \xi_k)$.
    Choose a stepsize $\alpha_k > 0$.
    Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$.
end for

The generic SG algorithm

The function $F : \mathbb{R}^d \to \mathbb{R}$ could be

$$F(w) = \begin{cases} R(w) = \mathbb{E}[f(w;\xi)] & \text{the expected risk,} \\ R_n(w) = \frac{1}{n}\sum_{\xi=1}^{n} f(w;\xi) & \text{the empirical risk.} \end{cases}$$

The stochastic vector could be

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k; \xi_k) & \text{the gradient for one example,} \\ \dfrac{1}{n_k}\displaystyle\sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}) & \text{the gradient for a minibatch,} \\ H_k \dfrac{1}{n_k}\displaystyle\sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}), & \text{possibly rescaled} \end{cases}$$

The generic SG algorithm

Stochastic processes

- We assume that the $\{\xi_k\}$ are jointly independent to avoid the full machinery of stochastic processes. But everything still holds if the $\{\xi_k\}$ form an adapted stochastic process, where each ξ_k can depend on the previous ones.

Active learning

- We can handle more complex setups by view ξ_k as a "random seed". For instance, in active learning, $g(w_k, \xi_k)$ firsts construct a multinomial distribution on the training examples in a manner that depends on w_k, then uses the random seed ξ_k to pick one according to that distribution.

The same mathematics cover all these cases.

2- Fundamental lemmas

Smoothness

Smoothness

- Our analysis relies on a smoothness assumption.
  We chose this path because it also gives results for the nonconvex case.
  We'll discuss other paths in the commentary section.

Assumption 4.1 (Lipschitz-continuous gradients). *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and its gradient, $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$, is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leq L\|w - \overline{w}\|_2 \ \textit{ for all } \ \{w, \overline{w}\} \subset \mathbb{R}^d.$$

Well known consequence

$$F(w) \leq F(\overline{w}) + \nabla F(\overline{w})^T(w - \overline{w}) + \tfrac{1}{2}L\|w - \overline{w}\|_2^2 \ \text{ for all } \ \{w, \overline{w}\} \subset \mathbb{R}^d. \qquad (4.3)$$

Smoothness

- $\mathbb{E}_{\xi_k}[\quad]$ is the expectation with respect to the distribution of ξ_k only.
- $\mathbb{E}_{\xi_k}[F(w_{k+1})]$ is meaningful because w_{k+1} depends on ξ_k (step 6 of SG)

Lemma 4.2. *Under Assumption 4.1, the iterates of SG (Algorithm 4.1) satisfy the following inequality for all $k \in \mathbb{N}$:*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]. \quad (4.4)$$

Expected decrease

Noise

Smoothness

Lemma 4.2. *Under Assumption 4.1, the iterates of SG (Algorithm 4.1) satisfy the following inequality for all $k \in \mathbb{N}$:*

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \\ \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k,\xi_k)] + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2].\end{aligned} \quad (4.4)$$

Proof. By Assumption 4.1, the iterates generated by SG satisfy

$$\begin{aligned}F(w_{k+1}) - F(w_k) &\leq \nabla F(w_k)^T (w_{k+1} - w_k) + \tfrac{1}{2}L\|w_{k+1} - w_k\|_2^2 \\ &\leq -\alpha_k \nabla F(w_k)^T g(w_k,\xi_k) + \tfrac{1}{2}\alpha_k^2 L\|g(w_k,\xi_k)\|_2^2.\end{aligned}$$

Taking expectations in these inequalities with respect to the distribution of ξ_k, and noting that w_{k+1}—but not w_k—depends on ξ_k, we obtain the desired bound. $\square$

Moments

Assumption 4.3 (First and second moment limits). *The objective function and SG (Algorithm 4.1) satisfy the following:*

(a) *The sequence of iterates $\{w_k\}$ is contained in an open set over which F is bounded below by a scalar $F_{\inf}$.*

(b) *There exist scalars $\mu_G \geq \mu > 0$ such that, for all $k \in \mathbb{N}$,*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad \text{and} \tag{4.7a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2. \tag{4.7b}$$

(c) *There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$,*

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2. \tag{4.8}$$

Moments

(b) There exist scalars $\mu_G \geq \mu > 0$ such that, for all $k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad \text{and} \tag{4.7a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2. \tag{4.7b}$$

(c) There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + \ldots F(w_k)\|_2^2. \tag{4.8}$$

- In expectation $g(w_k, \xi_k)$ is a sufficient descent direction.
- True if $\mathbb{E}_{\xi_k}[g(w_k, \xi_k)] = \nabla F(w_k)$ with $\mu = \mu_G = 1$.
- True if $\mathbb{E}_{\xi_k}[g(w_k, \xi_k)] = H_k \nabla F(w_k)$ with bounded spectrum.

Moments

(b) There exist scalars $\mu_G \geq \mu > 0$ such that, for all $k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad \text{and} \tag{4.7a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2. \tag{4.7b}$$

(c) There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2. \tag{4.8}$$

- $\mathbb{V}_{\xi_k}[\quad]$ denotes the variance w.r.t. ξ_k
- Variance of the noise must be bounded in a mild manner.

Moments

(b) There exist scalars $\mu_G \geq \mu > 0$ such that, for all $k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k,\xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad \text{and} \tag{4.7a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k,\xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2. \tag{4.7b}$$

(c) There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\mathbb{V}_{\xi_k}[g(w_k,\xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2. \tag{4.8}$$

- Combining (4.7b) and (4.8) gives

$$\mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2] \leq M + M_G \|\nabla F(w_k)\|_2^2 \tag{4.9}$$

with $M_G := M_V + \mu_G^2 \geq \mu^2 > 0$.

Moments

Lemma 4.4. *Under Assumptions 4.1 and 4.3, the iterates of SG (Algorithm 4.1) satisfy the following inequalities for all $k \in \mathbb{N}$:*

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \\
&\qquad \le -\mu\alpha_k\|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L\mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2] \qquad (4.10a) \\
&\qquad \le -(\mu - \tfrac{1}{2}\alpha_k L M_G)\alpha_k\|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L M. \qquad (4.10b)
\end{aligned}
$$

Expected decrease

Noise

- The convergence of SG depends on the balance between these two terms.

Moments

Lemma 4.4. *Under Assumptions 4.1 and 4.3, the iterates of SG (Algorithm 4.1) satisfy the following inequalities for all $k \in \mathbb{N}$:*

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \\ &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] && \text{(4.10a)} \\ &\leq -(\mu - \tfrac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L M. && \text{(4.10b)}\end{aligned}$$

Proof. By Lemma 4.2 and (4.7a), it follows that

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \\ &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2],\end{aligned}$$

which is (4.10a). Assumption 4.3, giving (4.9), then yields (4.10b). □

3- SG for Strongly Convex Objectives

Strong convexity

Assumption 4.5 (Strong convexity). *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is strongly convex in that there exists a constant $c > 0$ such that for all $(\overline{w}, w) \in \mathbb{R}^d \times \mathbb{R}^d$*

$$F(\overline{w}) \geq F(w) + \nabla F(w)^T(\overline{w} - w) + \tfrac{1}{2}c\|\overline{w} - w\|_2^2 . \tag{4.11}$$

Hence, F has a unique minimizer, denoted as $w_* \in \mathbb{R}^d$ with $F_* := F(w_*)$.

Known consequence

$$2c(F(w) - F_*) \leq \|\nabla F(w)\|_2^2 \quad \text{for all} \quad w \in \mathbb{R}^d. \tag{4.12}$$

Why does strong convexity matter?

- It gives the strongest results.
- It often happens in practice (one regularizes to facilitate optimization!)
- It describes any smooth function near a strong local minimum.

Total expectation

Different expectations

- $\mathbb{E}_{\xi_k}[\quad]$ is the expectation with respect to the distribution of ξ_k only.
- $\mathbb{E}[\quad]$ is the total expectation w.r.t. the joint distribution of all ξ_k.

For instance, since w_k depends only on $\xi_1, \xi_2, \dots, \xi_{k-1}$,

$$\mathbb{E}[F(w_k)] = \mathbb{E}_{\xi_1}\mathbb{E}_{\xi_2} \dots \mathbb{E}_{\xi_{k-1}}[F(w_k)]$$

Results in expectation

- We focus on results that characterize the properties of SG in expectation.
- The stochastic approximation literature usually relies on rather complex martingale techniques to establish almost sure convergence results. We avoid them because they do not give much additional insight.

SG with fixed stepsize

Theorem 4.6 (Strongly Convex Objective, Fixed Stepsize). *Under Assumptions 4.1, 4.3, and 4.5 (with $F_{\inf} = F_*$), suppose that the SG method (Algorithm 4.1) is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{4.13}$$

Then, for all $k \in \mathbb{N}$ the expected optimality gap satisfies :

$$\begin{aligned} \mathbb{E}[F(w_k) - F_*] &\leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\xrightarrow{k\to\infty} \frac{\bar{\alpha}LM}{2c\mu}. \end{aligned} \tag{4.14}$$

- Only converges to a neighborhood of the optimal value.
- Both (4.13) and (4.14) describe well the actual behavior.

SG with fixed stepsize (proof)

Proof. Using Lemma 4.4 with (4.13) and (4.12), we have for all $k \in \mathbb{N}$ that

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)] &\leq -(\mu - \tfrac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\bar{\alpha}^2 LM \\
&\leq -\tfrac{1}{2}\bar{\alpha}\mu\|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\bar{\alpha}^2 LM \\
&\leq -\bar{\alpha}c\mu(F(w_k) - F_*) + \tfrac{1}{2}\bar{\alpha}^2 LM.
\end{aligned}$$

Subtracting F_* from both sides and taking total expectations,

$$\mathbb{E}[F(w_{k+1}) - F_*] \leq (1 - \bar{\alpha}c\mu)\mathbb{E}[F(w_k) - F_*] + \tfrac{1}{2}\bar{\alpha}^2 LM.$$

Subtracting the constant $\bar{\alpha}LM/(2c\mu)$ from both sides, one obtains

$$\mathbb{E}[F(w_{k+1}) - F_*] - \frac{\bar{\alpha}LM}{2c\mu} \leq (1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_k) - F_*] - \frac{\bar{\alpha}LM}{2c\mu}\right). \tag{4.15}$$

Observe that (4.15) is a contraction inequality since, by (4.13) and (4.9),

$$0 < \bar{\alpha}c\mu \leq \frac{c\mu^2}{LM_G} \leq \frac{c\mu^2}{L\mu^2} = \frac{c}{L} \leq 1. \tag{4.16}$$

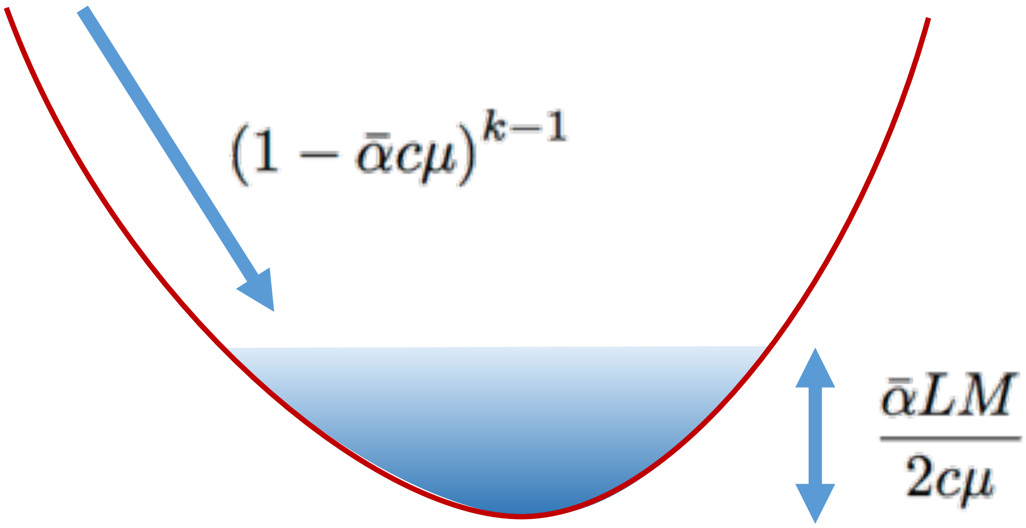
The result thus follows by applying (4.15) repeatedly. □

SG with fixed stepsize

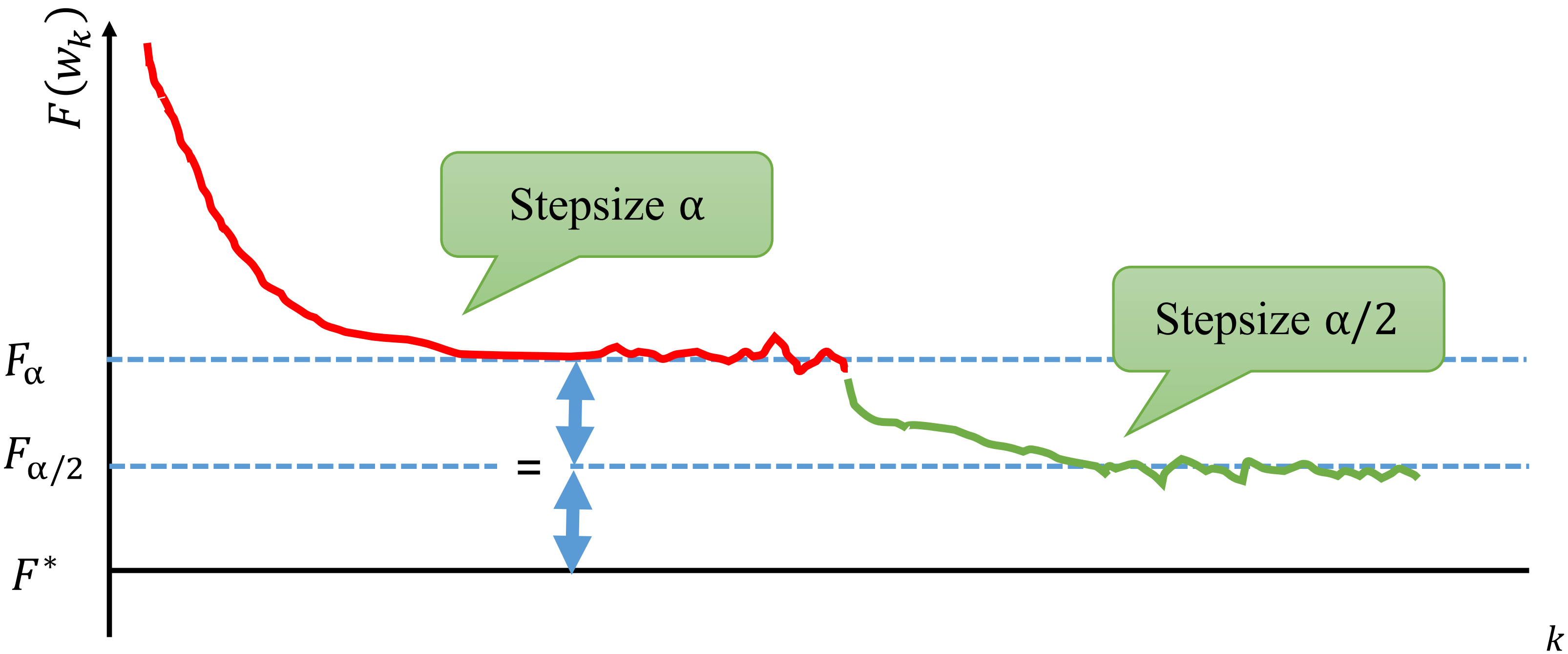
$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha} L M}{2c\mu} + (1 - \bar{\alpha} c\mu)^{k-1} \left(F(w_1) - F_* - \frac{\bar{\alpha} L M}{2c\mu} \right) \quad (4.14)$$

Note the interplay between the stepsize $\bar{\alpha}$ and the variance bound M.

- If $M = 0$, one recovers the linear convergence of batch gradient descent.
- If $M > 0$, one reaches a point where the noise prevents further progress.

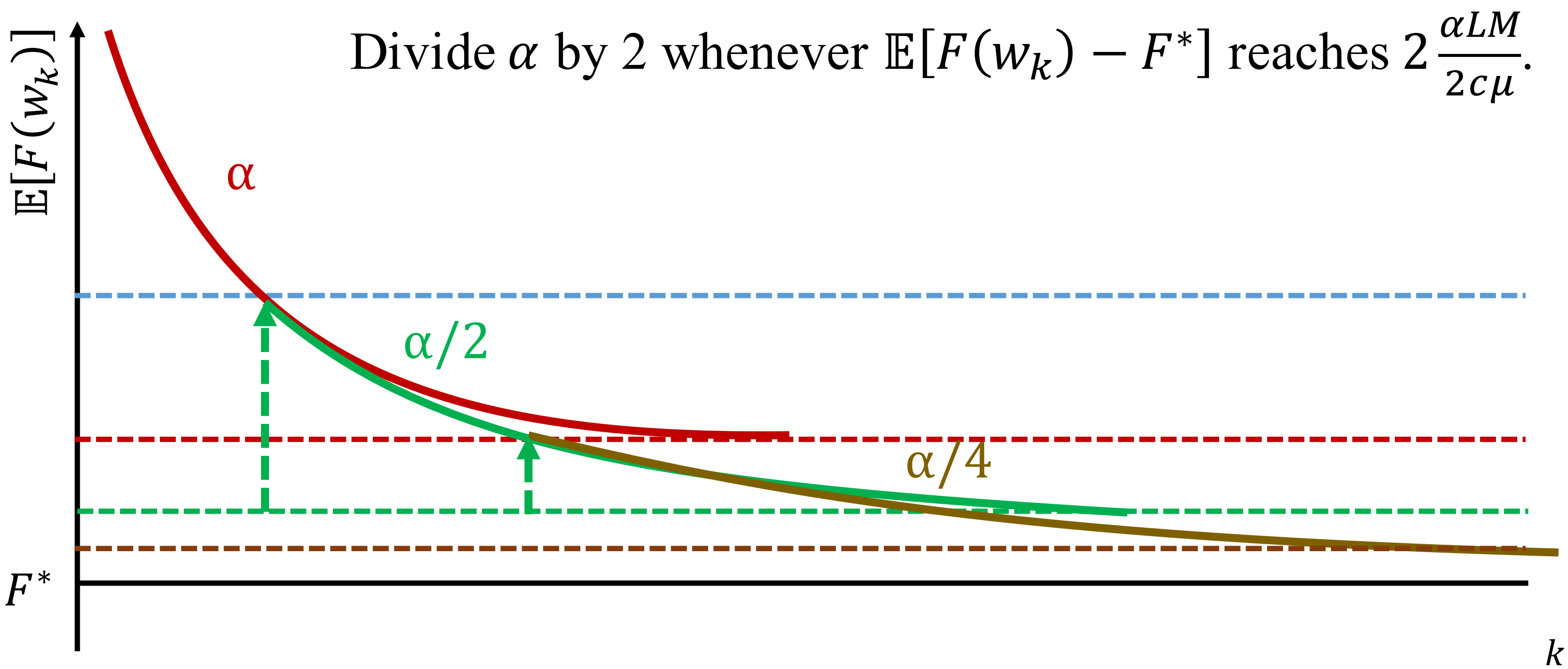


Diminishing the stepsizes



- If we wait long enough, halving the stepsize α eventually halves $F(w_k) - F^*$.
- We can even estimate $F^* \approx 2F_{\alpha/2} - F_\alpha$

Diminishing the stepsizes faster



- Divide α by 2 whenever $\mathbb{E}[F(w_k)]$ reaches $\alpha LM/c\mu$.
- Time τ_α between changes : $(1-\alpha c\mu)^{\tau_\alpha} = 1/3$ means $\tau_\alpha \propto 1/\alpha$.
- Whenever we halve α we must wait twice as long to halve $F(w) - F^*$.
- Overall convergence rate in $\mathcal{O}(1/k)$.

SG with diminishing stepsizes

Theorem 4.7 (Strongly Convex Objective, Diminishing Stepsizes). *Under Assumptions 4.1, 4.3, and 4.5 (with $F_{\inf} = F_*$), suppose that SG (Algorithm 4.1) is run with a stepsize sequence such that, for all $k \in \mathbb{N}$,*

$$\alpha_k = \frac{\beta}{\gamma + k} \text{ for some } \beta > \frac{1}{c\mu} \text{ and } \gamma > 0 \text{ s.t. } \alpha_1 \leq \frac{\mu}{LM_G}. \tag{4.18}$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma + k}, \tag{4.19}$$

where

$$\nu := \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - F_*)\right\}. \tag{4.20}$$

SG with diminishing stepsizes

Theorem 4.7 (Strongly Convex Objectiv… …**s).** … .1, 4.3, and 4.5 (with $F_{\text{inf}} = F_*$), su… …hat SG …n with a stepsize sequence such that, fo… …l $k \in \mathbb{N}$,

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \textit{for some} \quad \beta > \frac{1}{c\mu} \quad \textit{and} \quad \gamma > 0 \quad \textit{s.t.} \quad \alpha_1 \leq \frac{\mu}{L M_G}. \tag{4.18}$$

… $\mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma + k}, \tag{4.19}$$

where

$$\nu := \max\left\{\frac{\beta^2 L M}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - F_*)\right\}. \tag{4.20}$$

Same maximal stepsize

Stepsize decreases in 1/k

Not too slow…

…otherwise

gap ∝ stepsize

SG with diminishing stepsizes (proof)

Proof. Proceeding as in the proof of Theorem 4.6, one gets

$$\mathbb{E}[F(w_{k+1}) - F_*] \leq (1 - \alpha_k c\mu)\mathbb{E}[F(w_k) - F_*] + \tfrac{1}{2}\alpha_k^2 LM. \qquad (4.21)$$

We now prove (4.19) by induction. First, the definition of ν ensures that it holds for $k = 1$. Then, assuming (4.19) holds for some $k \geq 1$, it follows from (4.21) that

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - F_*] &\leq \left(1 - \frac{\beta c\mu}{\hat{k}}\right)\frac{\nu}{\hat{k}} + \frac{\beta^2 LM}{2\hat{k}^2} \qquad (\text{with } \hat{k} := \gamma + k) \\ &= \left(\frac{\hat{k}-1}{\hat{k}^2}\right)\nu \underbrace{- \left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2}}_{\text{nonpositive by the definition of } \nu} \leq \frac{\nu}{\hat{k}+1},\end{aligned}$$

where the last inequality follows because $\hat{k}^2 \geq (\hat{k}+1)(\hat{k}-1)$. $\square$

Mini batching

	Computation	Noise
$g(w_k,\xi_k) = \begin{cases} \nabla f(w_k;\xi_k) \\ \frac{1}{n_{\rm mb}}\sum_{i=1}^{n_{\rm mb}} \nabla f(w_k;\xi_{k,i}) \end{cases}$	1	M
	$n_{\rm mb}$	$M/n_{\rm mb}$

Using minibatches with stepsize $\bar{\alpha}$:

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu\, n_{\rm mb}} + [1-\bar{\alpha}c\mu]^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu\, n_{\rm mb}}\right).$$

Using single example with stepsize $\bar{\alpha}\,/\,n_{\rm mb}$:

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu\, n_{\rm mb}} + \left[1-\frac{\bar{\alpha}c\mu}{n_{\rm mb}}\right]^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu\, n_{\rm mb}}\right).$$

$n_{\rm mb}$ times more iterations that are $n_{\rm mb}$ times cheaper.

same

Minibatching

Ignoring implementation issues

- We can match minibatch SG with stepsize $\bar{\alpha}$
  using single example SG with stepsize $\bar{\alpha} \,/\, n_{\mathrm{mb}}$.
- We can match single example SG with stepsize $\bar{\alpha}$
  using minibatch SG with stepsize $\bar{\alpha} \times n_{\mathrm{mb}}$
  provided $\bar{\alpha} \times n_{\mathrm{mb}}$ is smaller than the max stepsize.

With implementation issues

- Minibatch implementations use the hardware better.
- Especially on GPU.

4- SG for General Objectives

Nonconvex objectives

Nonconvex training objectives are pervasive in deep learning.

Nonconvex landscape in high dimension can be very complex.
- Critical points can be local minima or saddle points.
- Critical points can be first order of high order.
- Critical points can be part of critical manifolds.
- A critical manifold can contain both local minima and saddle points.

We describe meaningful (but weak) guarantees
- Essentially, SG goes to critical points.

The SG noise plays an important role in practice
- It seems to help navigating local minima and saddle points.
- More noise has been found to sometimes help optimization.
- But the theoretical understanding of these facts is weak.

Nonconvex SG with fixed stepsize

Theorem 4.8 (Nonconvex Objective, Fixed Stepsize). *Under Assumptions 4.1 and 4.3, suppose that the SG method (Algorithm 4.1) is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \bar{\alpha} \leq \frac{\mu}{L M_G}. \tag{4.25}$$

Then, the expected sum-of-squares and average-squared gradients of F corresponding to the SG iterates satisfy the following inequalities for all $K \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{K\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\inf})}{\mu\bar{\alpha}} \tag{4.26a}$$

$$\text{and therefore} \quad \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\inf})}{K\mu\bar{\alpha}} \tag{4.26b}$$

Nonconvex SG with fixed stepsize

Theorem 4.8 (Nonconvex Objectiv[…] *[Und]er Assump-tions 4.1 and 4.3, suppose that the SG* […] *run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, sa*[…]

$$0 < \bar{\alpha} \le \frac{\mu}{LM_G}. \tag{4.25}$$

[…] *of-squares and average-* […] *e-* […] *ates satisfy the followin* […] *$\mathbb{N}$:*

$$\ldots \left[\sum_{k=1} \|\nabla F(w_k)\|_2 \ldots\right] \le \ldots \frac{\ldots F(w_1) - \ldots_{\inf})}{\mu\bar{\alpha}} \tag{4.26a}$$

$$\text{and therefore } \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \le \frac{\bar{\alpha} L M}{\mu} + \frac{2(F(w_1) - F_{\inf})}{K\mu\bar{\alpha}} \tag{4.26b}$$

Same max stepsize

If the average norm of the gradient is small, then the norm of the gradient cannot be often large…

This goes to zero like 1/K

This does not

Nonconvex SG with fixed stepsize (proof)

Proof. Taking the total expectation of (4.10b) and from (4.25),

$$\begin{aligned}\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] &\leq -(\mu - \tfrac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \tfrac{1}{2}\bar{\alpha}^2 LM \\ &\leq -\tfrac{1}{2}\mu\bar{\alpha}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \tfrac{1}{2}\bar{\alpha}^2 LM.\end{aligned}$$

Summing both sides of this inequality for $k \in \{1, \ldots, K\}$ and recalling Assumption 4.3(a) gives

$$F_{\text{inf}} - F(w_1) \leq \mathbb{E}[F(w_{K+1})] - F(w_1) \leq -\tfrac{1}{2}\mu\bar{\alpha}\sum_{k=1}^{K}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \tfrac{1}{2}K\bar{\alpha}^2 LM.$$

Rearranging yields (4.26a), and dividing further by K yields (4.26b). $\square$

Nonconvex SG with diminishing step sizes

Theorem 4.10 (Nonconvex Objective, Diminishing Stepsizes). *Under Assumptions 4.1 and 4.3, suppose that the SG method (Algorithm 4.1) is run with a stepsize sequence satisfying*

$$\sum_{k=1}^{\infty} \alpha_k = \infty \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty ,$$

then

$$\mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right] < \infty$$

Corollary 4.12. *Under the conditions of Theorem 4.10, if we further assume that the objective function F is twice differentiable, and that the mapping $w \mapsto \|\nabla F(w)\|_2^2$ has Lipschitz-continuous derivatives, then*

$$\lim_{k\to\infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0.$$

5- Work complexity for Large-Scale Learning

Large-Scale Learning

Assume that we are in the large data regime

- Training data is essentially unlimited.
- Computation time is limited.

The good

- More training data $\Rightarrow$ less overfitting
- Less overfitting $\Rightarrow$ richer models.

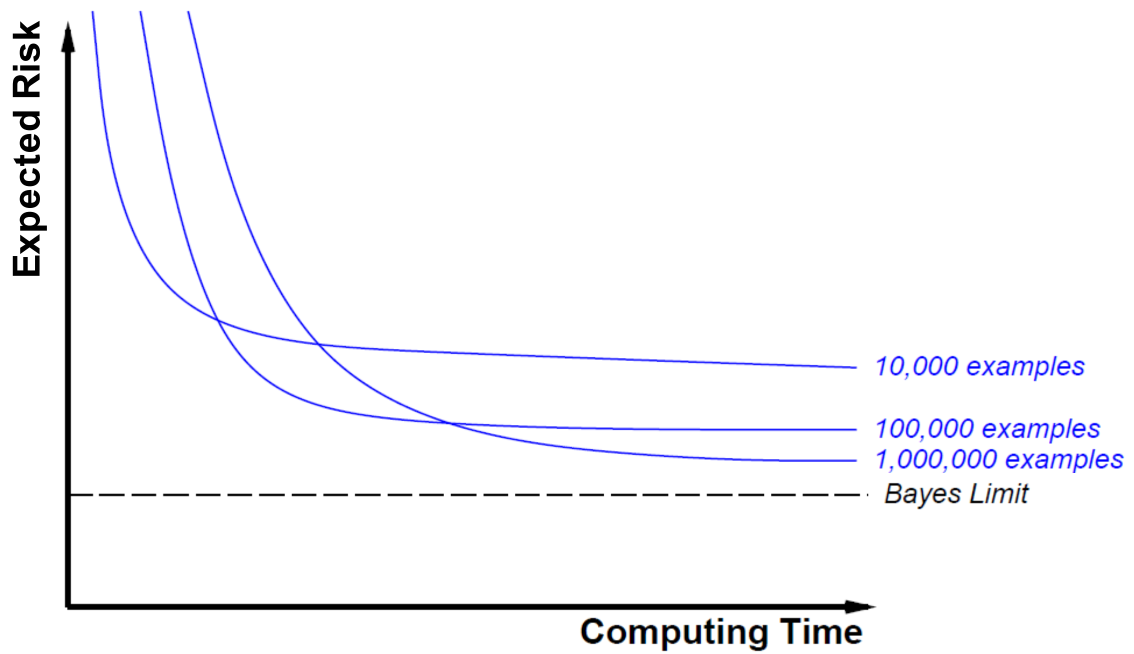
The bad

- Using more training data or rich models quickly exhausts the time budget.

The hope

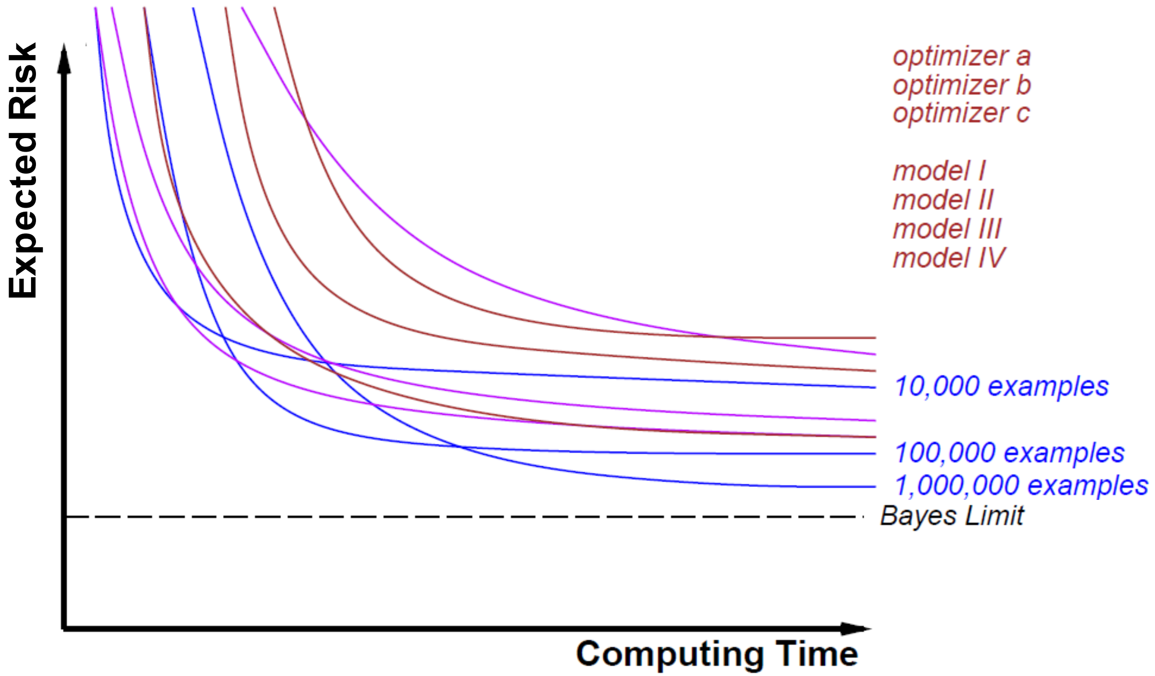
- How thoroughly do we need to optimize $R_n(w)$
  when we actually want another function $R(w)$ to be small?

Expected risk versus training time



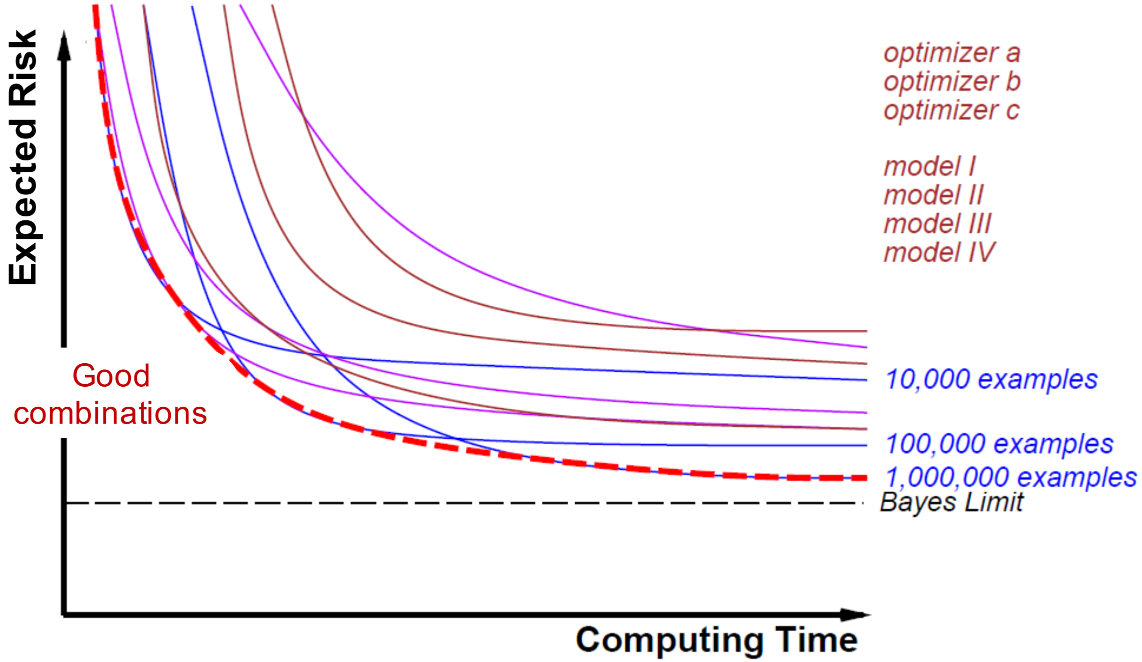
- When we vary the number of examples

Expected risk versus training time



- When we vary the number of examples, the model, and the optimizer…

Expected risk versus training time



- The optimal combination depends on the computing time budget

Formalization

The components of the expected risk

$$\mathbb{E}[R(\tilde{w}_n)] = \underbrace{R(w_*)}_{\mathcal{E}_{app}(\mathcal{H})} + \underbrace{\mathbb{E}[R(w_n) - R(w_*)]}_{\mathcal{E}_{est}(\mathcal{H},n)} + \underbrace{\mathbb{E}[R(\tilde{w}_n) - R(w_n)]}_{\mathcal{E}_{opt}(\mathcal{H},n,\epsilon)} \qquad (4.29)$$

Question

- Given a fixed model $\mathcal{H}$ and a time budget $\mathcal{T}_{\max}$, choose $n, \epsilon \ldots$

$$\min_{n,\epsilon} \ \mathcal{E}(n,\epsilon) = \mathbb{E}[R(\tilde{w}_n) - R(w_*)] \ \text{ s.t. } \ \mathcal{T}(n,\epsilon) \le \mathcal{T}_{\max}. \qquad (4.30)$$

Approach

- Statistics tell us $\mathcal{E}_{est}(n)$ decreases with a rate in range $1/\sqrt{n} \ldots 1/n$.
- For now, let's work with the fastest rate compatible with statistics

$$\mathcal{E}(n,\epsilon) \sim \frac{1}{n} + \epsilon \qquad (4.32)$$

Batch versus Stochastic

Typical convergence rates

- Batch algorithm: $\mathcal{T}(n,\epsilon) \sim n\log(1/\epsilon)$
- Stochastic algorithm: $\mathcal{T}(n,\epsilon) \sim 1/n$

Rate analysis

		Batch	Stochastic
$\mathcal{T}(n,\epsilon)$	$\sim$	$n \log\left(\frac{1}{\epsilon}\right)$	$\frac{1}{\epsilon}$
n^*	$\sim$	$\frac{\mathcal{T}_{\max}}{\log(\mathcal{T}_{\max})}$	$\mathcal{T}_{\max}$
$\mathcal{E}^*$	$\sim$	$\frac{\log(\mathcal{T}_{\max})}{\mathcal{T}_{\max}} + \frac{1}{\mathcal{T}_{\max}}$	$\frac{1}{\mathcal{T}_{\max}}$

Processing more training examples beats optimizing more thoroughly.

This effect only grows if $\mathcal{E}_{est}(n)$ decreases slower than $1/n$.

6- Comments

Asymptotic performance of SG is fragile

Diminishing stepsizes are tricky

- Theorem 4.7 (strongly convex function) suggests

$$\alpha_k = \frac{\beta}{\gamma + k}$$

SG converges very slowly if $\beta < \frac{1}{c\mu}$

SG usually diverges when α is above $\frac{2\mu}{LM_G}$

Constant stepsizes are often used in practice

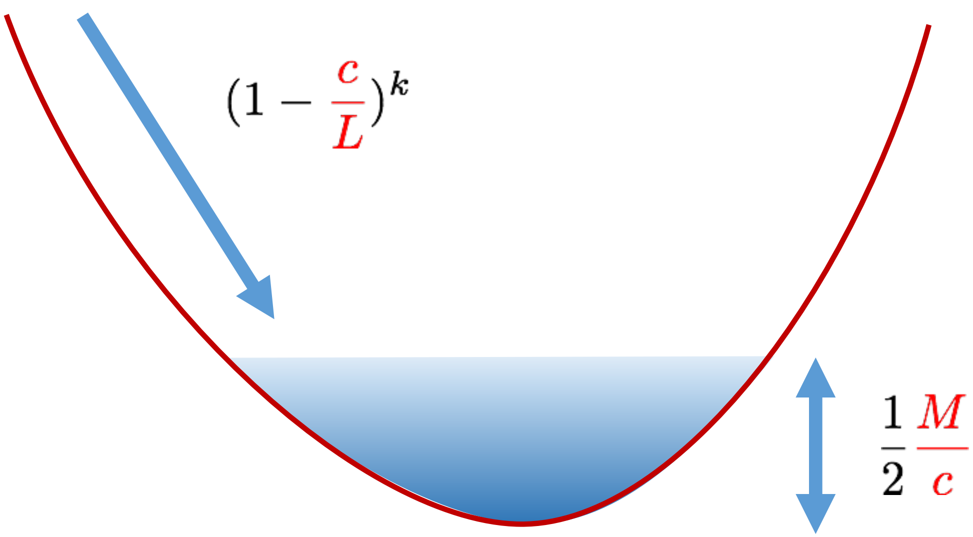
- Sometimes with a simple halving protocol.

Spoiler – Certain SG variants are more robust.

Condition numbers

The ratios $\frac{L}{c}$ and $\frac{M}{c}$ appear in critical places

- Theorem 4.6. With $\mu = 1, M_V = 0$, the optimal stepsize is $\bar{\alpha} = \frac{1}{L}$



Distributed computing

SG is notoriously hard to parallelize

- Because it updates the parameters w with high frequency
- Because it slows down with delayed updates.

SG still works with relaxed synchronization

- Because this is just a little bit more noise.

Communication overhead give room for new opportunities

- There is ample time to compute things while communication takes place.
- Opportunity for optimization algorithms with higher per-iteration costs

➔ SG may not be the best answer for distributed training.

Smoothness versus Convexity

Analyses of SG that only rely on convexity

- Bounding $\|w_k - w^*\|^2$ instead of $F(w_k) - F^*$ and assuming $\mathbb{E}_{\xi_k}[g(w_k, \xi_k)] = \hat{g}(w_k) \in \partial F(w_k)$ gives a result similar to Lemma 4.4.

$$\begin{aligned} &\mathbb{E}_{\xi_k}[\|w_{k+1} - w_*\|_2^2] - \|w_k - w_*\|_2^2 \\ &\qquad = \underbrace{-2\alpha_k \hat{g}(w_k)^T (w_k - w_*)}_{\text{Expected decrease}} + \underbrace{\alpha_k^2 \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]}_{\text{Noise}}, \end{aligned} \tag{A.2}$$

- Ways to bound the expected decrease

$$\text{General convexity}: \quad \hat{g}(w_k)^T (w_k - w_*) \geq F(w_k) - F(w_*) \geq 0$$
$$\text{Strong convexity}: \quad \hat{g}(w_k)^T (w_k - w_*) \geq c\|w_k - w_*\|^2 \geq 0$$

- Proof does not easily support second order methods.