CS 331: Stochastic Gradient Descent Methods

Peter Richtárik
https://richtarik.org

KAUST
King Abdullah University of Science and Technology

Fall 2022

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 1: Course Organization

Outline

- About Me
- Logistics
- Brief Course Description
- Goals and Objectives
- Knowledge Required
- Reference Text
- Method of Evaluation

About Me - I

- **Name:** Peter Richtárik
- **Position:** Professor of Computer Science at King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
  - also affiliated with Applied Mathematics and Computational Sciences (AMCS) and Statistics (STAT) programs
- **Website:** `https://richtarik.org`
- **Email:** `peter.richtarik@kaust.edu.sa`
- **Brief Academic CV:**
  - 2019–now, Professor, KAUST
  - 2017–2019, Associate Professor, KAUST
  - 2009–2017, Assistant and later Associate Professor, University of Edinburgh, United Kingdom
  - 2007–2009, Postdoc, UC Louvain, Belgium
  - 2002–2007, PhD student, Cornell, USA
  - 1996–2001, Bc and Mgr student, Comenius University, Slovakia
- **Research:**
  - optimization for machine learning
  - convex and nonconvex optimization

About Me - II

- stochastic/ramdomized zero, first and second order methods
- distributed centralized and decentralized optimization
- federated learning
- deep learning
- randomized linear algebra
- sampling algorithms
- saddle-point problems
- operator splitting algorithms

- **Publishing in:** ICML, NeurIPS, AISTATS, ICLR, JMLR, SIAM Journals, IEEE Journals, Optimization Methods and Software, Mathematical Programming ...

Logistics I

- **Communication:** All course-related communication will be done via Piazza
- **Lecture Time & Room:**
  - Tuesdays: 8am-11am
  - Building 9, Room 3222
- **Office Hours:** This is time when students can ask questions about the lectures. I will stay in class immediately after each lecture and will take questions then. Additional office hours can be scheduled by email appointment. The TAs will hold office hours as well.
- **Teaching Assistants (TAs):**
  1. Grigory Malinovsky

Figure: Grigory Malinovsky

Logistics II

- Web: `https://grigory-malinovsky.github.io`
- Email: `grigorii.malinovskii@kaust.edu.sa`

2. Igor Sokolov

Figure: Igor Sokolov

- Web: `https://cemse.kaust.edu.sa/people/person/igor-sokolov`
- Email: `igor.sokolov.1@kaust.edu.sa`

Brief Course Description I

Stochastic gradient descent (SGD) in one or another of its many variants is the workhorse method for training modern supervised machine learning models. However, the world of SGD methods is vast and expanding, which makes it hard for practitioners and even experts to understand its landscape and inhabitants.

- This course is a **mathematically rigorous and comprehensive introduction** to the field, and is based on the latest results and insights.
- The course develops a **convergence and complexity theory** for serial, parallel, and distributed variants of SGD, in the **strongly convex regime**[1] with randomness coming from sources such as **subsampling** and **compression.**
- A substantial part of the course offers a **unified analysis of a large family of variants of SGD** which have so far required different intuitions, convergence analyses, have different applications, and which have been developed separately in various communities. This framework includes methods with and without the following tricks, and their combinations:

Brief Course Description II

- variance reduction,
- data sampling,
- coordinate sampling,
- arbitrary sampling,
- importance sampling,
- mini-batching,
- sketching,
- sparsification,
- quantization.

[1] We do not have time to cover convex and nonconvex regimes in these 12 lectures.

Goals and Objectives

- Detailed understanding of the **key variants of** `SGD` for training **supervised machine learning models** and their differences
- Understanding the underlying **mathematical theory** and of the insights the theory offers for practice
- Preparation for **original theoretical and applied research** in the field of randomized methods for optimization and machine learning

Knowledge Required

- **Programming:** Strong experience with at least one high level computing language, such as MATLAB, Python, R, Julia, and C
- **Mathematical maturity:** Ability to comprehend and generate proofs
- **Linear algebra:** Abstract vector spaces, linear independence, basis, linear operators, quadratic forms, Euclidean spaces, inner product, norm
- **Matrix theory:** matrices, determinants, singular values, eigenvalues, matrix decompositions
- **Multivariate calculus:** gradient, Hessian, Taylor approximation, chain rule
- **Probability theory:** probability spaces, expectation, law of large numbers, tower property of expectation

Reference Text

- **Book:** There is no book that covers the material contained in this course. In fact, there is no course offered anywhere that covers this material. The contents are based on a fresh view of SGD methods based on latest research conducted in the last few years.
- **Slides:**
  - Slides will include all relevant material (e.g., some limited/terse explanation, theorems, algorithms, proofs).
  - However, course slides are *not* a book. Listening to the lecture is necessary for the slides to make sense; and asking questions is important and highly desirable for deeper understanding!
  - Slides will be made available in `pdf` and/or printed form before each lecture.
  - Recommendation: Upload the pdf version of the slides to an iPad, and annotate with an Apple pencil during the lecture. Alternatively, annotate the printed slides with a pen & highlighter.
- **Papers:** Relevant research papers will be referred to and recommended in lectures (optional reading).

Method of Evaluation

- **70% Assignments**
  - 14 weekly assignments
  - A mix of theoretical and coding problems
  - Each assignment is worth 5 points for a total of $14 \times 5 = 70$ points
  - No late submissions will be accepted (submit all you have done, even if you did not fully complete the assignment, by the deadline)
- **25% Final Exam**
  - Theoretical
- **5% Active Participation**
  - Attendance at lectures (this is compulsory)
  - Asking questions (lectures are designed in an interactive fashion – there is plenty of room for discussion)

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 2: Introduction to Optimization

Outline

- Optimization Problems and Solution Concepts
- Optimization Modelling and Data
- Optimization: Basic Terminology
- Optimization Algorithms: Basic Terminology
- Basic Algorithms
- Convergence

Optimization Problems and Solution Concepts

Optimization Problems

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x = (x_1, \ldots, x_d) \in Q \subseteq \mathbb{R}^d \end{aligned} \tag{1}$$

- a set Q **(feasible set / constraint)**
- a real-valued function $f : Q \to \mathbb{R}$ **(objective function / cost function / loss function)**

Condition	**Problem** (1) **is called**
$Q = \mathbb{R}^d$	**unconstrained**
$Q \neq \mathbb{R}^d$	**constrained**
$Q \neq \emptyset$	**feasible**
$Q = \emptyset$	**infeasible**
$\exists C \in \mathbb{R}$ such that $f(x) \geq C$ for all $x \in Q$	**bounded**
$\exists \{x^k\}_{k=0}^{\infty} \subseteq Q$ such that $f(x^k) \to -\infty$	**unbounded**

Basic Types of Optimization Problems

Q is often described as a solution of a (finite) system of equations and inequalities:

$$Q = \{x \ : \ g_i(x) = 0, \ i \in E; \ g_i(x) \leq 0, \ i \in I\}.$$

- E is some finite index set, indexing **E**qualities
- I is some finite index set, indexing **I**nequalities

Condition	**Problem** (1) **is called**
f, g_i are linear	**Linear Program** (LP)
f not linear	**Nonlinear Program** (NLP)
f quadratic, g_i linear	**Quadratic Program** (QP)
f, g_i quadratic	**Quadratically Constrained QP** (QCQP)
f, Q convex	**Convex Program** (CP)
f not convex	**Noncovex Program**
LP with $x_i \in \mathbb{Z}$ for some i	**Integer Program** (IP)

We shall deal with **convex problems** and **nonconvex problems** in this course, either without any constraints, or with a convex constraint set Q.

Solution Concepts I

Selection of some key solution concepts:

Condition	x **is called**
any x	**a solution/point**
$x \in Q$	**feasible solution/point**
$x \notin Q$	**infeasible solution/point**
$f(x) \leq f(y)$ for all $y \in Q \cap N(x, \delta)$	**local minimizer**
$f(x) \leq f(y)$ for all $y \in Q$	**global minimizer** ($x^\star$)
$f(x) \leq f(y) + \epsilon$ for all $y \in Q \cap N(x, \delta)$	**ϵ-solution** (local)
$f(x) \leq f(y) + \epsilon$ for all $y \in Q$	**ϵ-solution** (global)
$\|x - x^\star\| \leq \epsilon$	**ϵ-solution** (global)
$\nabla f(x) = 0$ and $Q = \mathbb{R}^d$	**stationary point**
$\|\nabla f(x)\| \leq \epsilon$ and $Q = \mathbb{R}^d$	**ϵ-stationary point**

Notation used above:

- For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, by $\|x\|$ we denote the **standard Euclidean norm** of x, defines as:

$$\|x\| = \sqrt{\sum_{i=1}^d x_i^2}.$$

Solution Concepts II

- The **δ-neighborhood** of $x \in \mathbb{R}^d$ is the set:

$$N(x, \delta) \stackrel{\text{def}}{=} \{y \ : \ \|y - x\| \leq \delta\}$$

For convex problems we will be interested in a **global minimizer**, and for nonconvex problems in a **stationary point.**

Optimization Modelling and Data

Optimization Modelling

1. One wants to solve some **problem**
   - Predict what advert a given user will click on
   - Deblurr an image
   - Rank a collection webpages based on their relevance to a given query
2. **Represent the solution** of the problem as a mathematical object; typically a vector $x \in \mathbb{R}^d$
   - x_i represents the probability of clicking on advert i
   - x_i represents the color of pixel i
   - x_i represents relevance of website i to the query
3. Choose a set $Q \subseteq \mathbb{R}^d$ of solutions which you are willing the consider. Q is the **feasibility/constraint set.**
   - $Q = \{x \in \mathbb{R}^d \ : \ 0 \leq x_i \leq 1, \ \sum_i x_i = 1\}$
   - $Q = \{x \in \mathbb{R}^d \ : \ 0 \leq x_i \leq 255\}$
   - $Q = \mathbb{R}^d$ or $Q = \{x \in \mathbb{R}^d \ : \ x \geq 0\}$
4. Choose a **cost/loss/objective function** $f : \mathbb{R}^d \to \mathbb{R}$ measuring the quality of each potential solution. In particular, $f(x) < f(y)$ should mean that x is a better solution than y.
5. Solve problem (OPT) using a suitable **optimization algorithm**

Optimization and Data

Sources of data:

1. **Description data:** Data forming the initial logical description of the problem
2. **Representation data:** Our choice of representation may depend on some synthetic / computed / collected data
3. **Feasibility data:** Data describing Q
4. **Objective data:** Data describing f

Optimization Algorithms: Basic Terminology

Direct vs Iterative Methods - I

Direct methods

- Perform a **finite number of steps.**
- Intermediary steps can't be used to produce an approximate solution.
- The solution found is (typically) **exact.**

Iterative methods

- Perform an **infinite number of steps** (in principle).
- Intermediary steps generate a sequence of iterates $x^0, x^1, x^2, \ldots$ which get progressively better (in some sense).
- Any solution in the sequence is **approximate** only

Direct vs Iterative Methods - II

Example 1

Solve problem (1) with $f(x) = \frac{1}{2}\|\boldsymbol{A}x - b\|^2$ and $Q = \mathbb{R}^d$.

- **Direct method:** To find the stationary point, which happens to be the global solution since f is convex, set the gradient to zero:

$$\nabla f(x) = \boldsymbol{A}^\top(\boldsymbol{A}x - b) = 0.$$

  This results to the linear system $\boldsymbol{A}^\top\boldsymbol{A}x = \boldsymbol{A}^\top b$. Solve the system using any direct solver for linear systems, such as Gaussian elimination. The result is obtained after a finite number of steps.

- **Iterative method:** Choose some initial guess $x^0 \in \mathbb{R}^d$, compute constant L satisfying the bound $L \geq \lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})$ and perform the following iterative process ("gradient descent"):

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k).$$

  It can be shown that $f(x^k) \leq \frac{C}{k}$ for all k and some constant C (which depends on x^0, $\boldsymbol{A}$ and b).

We will only talk about iterative methods in this course; these are by far the most prevalent and useful in optimization and machine learning.

Iterative Optimization Methods

All iterative methods can be written in the generic form

$$x^{k+1} = x^k + \alpha_k h^k,$$

where

- $h^k \in \mathbb{R}^d$ is the **search/descent direction**
- $\alpha_k \in \mathbb{R}$ is the **stepsize**
- $\alpha_k h^k \in \mathbb{R}^d$ is the **step**

Iterative optimization methods differ in how:

- the direction h^k is computed (a major distinguishing factor)
- the stepsize α_k is determined (a minor distinguishing factor)

Complexity

Two key questions when choosing an optimization method:

- How costly is it to get from x^k to x^{k+1}? Typically, we care about time. Some proxies for time:
  - **computation cost = # of arithmetic operations (additions, multiplications) performed,**
  - **communication cost = # bits transfered.**
- How many iterations k do we need to perform to get a solution of acceptable quality?

Definition 2 (Complexity of an Optimization Method)

Let $\mathcal{M}$ be an (iterative) optimization method for solving (1). Let $\epsilon > 0$ be a desired error tolerance with respect to some measure of success. Let W be the cost of performing one step of method $\mathcal{M}$. Let $k(\epsilon)$ be the number of iterations sufficient for the output of $\mathcal{M}$ to be within the desired tolerance ϵ. The **total complexity** of $\mathcal{M}$ is the function

$$\epsilon \mapsto k(\epsilon) \times W.$$

Remark: Typically, both W and $k(\epsilon)$ depend on the problem at hand (i.e., on f and Q).

Classification of Optimization Methods

Definition 3

We say that an (iterative) optimization algorithm is

- **feasible** if all iterates x^k are feasible, i.e., if $x^k \in Q$ for all k
- **infeasible** if it is not feasible
- **deterministic** if, given x^0, always the same sequence of iterates $\{x^k\}$ is produced
- **randomized** if the iterates $\{x^k\}$ form a random process (i.e., if they are random vectors)
- **zero order** if it is only allowed to compute zero-order information about f (i.e., function values)
- **first order** if it is only allowed to compute up to first-order information about f (i.e., function values and gradients)
- **second order** of it is allowed to compute up to second-order information about f (i.e., function values, gradients and Hessian)

Zero vs First vs Second Order Methods

Method	**Local information at x about f utilized**
zero order	function value $f(x) \in \mathbb{R}$
first order	gradient $\nabla f(x) \in \mathbb{R}^d$
second order	Hessian $\nabla^2 f(x) \in \mathbb{R}^{d \times d}$

Gradient and Hessian

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} \in \mathbb{R}^d \qquad \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ & \vdots & & \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{pmatrix} \in \mathbb{R}^{d \times d}$$

Key to Higher Order Methods: Taylor's Theorem

Landau's little "oh" notation

By $o(t)$ we mean any function $o : \mathbb{R} \to \mathbb{R}$ such that $\lim_{t\to 0} \frac{o(t)}{t} = 0$. Informally, these are functions which approach zero faster than "linearly"; i.e., faster than the $t \mapsto t$ function does.

Example 4

$t^2 = o(t)$ since $\lim_{t\to 0} \frac{t^2}{t} = t = 0$.

Theorem 5 (Multivariate Taylor's Theorem)

Let $U \subseteq \mathbb{R}^d$ be an open subset of $\mathbb{R}^d$. Fix any $x \in U$ and $h \in \mathbb{R}^d$ such that $x + th \in U$ for all $t \in [0, 1]$.

1. *If $f : U \to \mathbb{R}$ is* **differentiable,** *then*

$$f(x + h) = f(x) + \langle \nabla f(x), h\rangle + o(\|h\|).$$

2. *If $f : U \to \mathbb{R}$ is* **twice differentiable,** *then*

$$f(x + h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + o(\|h\|^2).$$

Standard Algorithms

A Quick Review of Standard Methods

Direct Search (Zero-order method)

- Fix a set $\mathcal{D} \stackrel{\text{def}}{=} \{d_1, \ldots, d_t\} \subset \mathbb{R}^d$, and constants $\alpha_0 > 0$, $c > 0$
- If $f(x^k + \alpha_k d_i) < f(x^k) - c\alpha_k^2$ for some i,
  - then set $x^{k+1} = x^k + \alpha_k d_i$
  - otherwise $x^{k+1} = x^k$, $\alpha_{k+1} \leftarrow \alpha_k/2$

Gradient Descent (First-order method)

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Accelerated Gradient Descent (Nesterov's Method) (First-order method)

$$y^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad x^{k+1} = (1 + \beta_k) y^{k+1} - \beta_k y^k$$

Newton's Method (Second-order method)

$$x^{k+1} = x^k - \alpha_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

Complexity of Standard Methods I

Consider the least-squares function

$$f(x) = \frac{1}{2}\|\boldsymbol{A}x - b\|^2,$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is a data matrix, $x \in \mathbb{R}^d$ is a vector of parameters to be tuned and $b \in \mathbb{R}^n$ is a data vector. Note that

$$\nabla f(x) = \boldsymbol{A}^\top(\boldsymbol{A}x - b), \qquad \nabla^2 f(x) = \boldsymbol{A}^\top \boldsymbol{A}.$$

Cost of evaluation of $f(x)$:

- Multiply $\boldsymbol{A}$ by x: $O(nd)$
- Subtract b from the result: $O(n)$
- Compute the squared norm of the resulting vector: $O(n)$
- Total cost: $O(nd)$

Cost of evaluation of $\nabla f(x)$:

- Multiply $\boldsymbol{A}$ by x: $O(nd)$

Complexity of Standard Methods II

- Subtract b from the result: $O(n)$
- Multiply the result by $\boldsymbol{A}^\top$: $O(nd)$
- Total cost: $O(nd)$

Cost of evaluation of $(\nabla^2 f(x))^{-1}\nabla f(x)$:

- Form the matrix $\nabla^2 f(x) = \boldsymbol{A}^\top \boldsymbol{A}$: $O(nd^2)$
- Form the gradient $\nabla f(x)$: $O(nd)$
- Solve the $n \times n$ linear system $\nabla^2 f(x)h = \nabla f(x)$ using a direct method (e.g., Gaussian elimination): $O(d^3)$
- Total cost: $O(d^3 + nd^2)$

Convergence

What Quantity do we Want to Converge?

Let $x^\star$ be a solution we are interested in (global, local, feasible, ...).

- If the method is **deterministic,** we may wish one of the following quantities to converge to zero:
  - Squared distance to a solution: $a_k \stackrel{\text{def}}{=} \|x^k - x^\star\|^2 \to 0$
  - Function values: $a_k \stackrel{\text{def}}{=} f(x^k) - f(x^\star) \to 0$
  - Squared norm of the gradient: $a_k \stackrel{\text{def}}{=} \|\nabla f(x^k)\|^2 \to 0$
- If the method is **randomized,** we may wish one of the following quantities to converge to zero:
  - Expected squared distance to a solution: $a_k \stackrel{\text{def}}{=} \mathbf{E}\left[\|x^k - x^\star\|^2\right] \to 0$
  - Expected function values: $a_k \stackrel{\text{def}}{=} \mathbf{E}\left[f(x^k) - f(x^\star)\right] \to 0$
  - Expected squared gradient norm: $a_k \stackrel{\text{def}}{=} \mathbf{E}\left[\|\nabla f(x^k)\|^2\right] \to 0$

The above are just some of the most common measures of success, several other quantities are studied as well.

The expectation above is with respect to the coin flips / randomness inherent in the randomized algorithm.

Typical Convergence Rates Encountered in Optimization

Bound on a_k	Bound on k implying $a_k \leq \epsilon$	Typical Recursion	Speed
$a_k \leq \frac{C}{\sqrt{k}}$	$k \geq \frac{C}{\epsilon^2}$		slow
$a_k \leq \frac{C}{k}$	$k \geq \frac{C}{\epsilon}$	$a_{k+1} \leq a_k - \frac{a_k^2}{C}$	ok
$a_k \leq \frac{C}{k^2}$	$k \geq \sqrt{\frac{C}{\epsilon}}$		fast
$a_k \leq a_0 e^{-k/C}$	$k \geq C \log\left(\frac{a_0}{\epsilon}\right)$	$a_{k+1} \leq \left(1 - \frac{1}{C}\right) a_k$	very fast

Recursions & Associated Convergence Rates I

Proposition 1

Let $\{a_k\}_{k=0}^{\infty}$ *be a sequence of positive real numbers.*

1. *Assume that the recursion* $a_{k+1} \leq \left(1 - \frac{1}{C}\right) a_k$ *holds for all* $k \geq 0$ *and some constant* $C > 1$. *Then*

$$a_k \leq \left(1 - \frac{1}{C}\right)^k a_0 \leq e^{-k/C} \cdot a_0.$$

2. *Assume that the recursion* $a_{k+1} \leq a_k - \frac{a_k^2}{C}$ *holds for all* $k \geq 0$ *and some constant* $C > 0$. *Then for all* $k \geq 0$, *we have*[2]

$$a_k \leq \frac{Ca_0}{C + a_0 k} \leq \frac{C}{k}.$$

[2] The second inequality only makes sense for $k \geq 0$

Proof of Proposition 1 - I

1. By applying the recurrence repeatedly, we get

$$a_k \leq \left(1 - \frac{1}{C}\right) a_{k-1} \leq \cdots \leq \left(1 - \frac{1}{C}\right)^k a_0.$$

Next, we can use the approximation

$$\left(1 - \frac{1}{C}\right)^k = \left[\left(1 - \frac{1}{C}\right)^C\right]^{k/C} \leq \left(e^{-1}\right)^{k/C} = e^{-k/C}.$$

Proof of Proposition 1 - II

2. Note that

$$\frac{1}{a_{k+1}} - \frac{1}{a_k} = \frac{a_k - a_{k+1}}{a_k a_{k+1}} \geq \frac{a_k - a_{k+1}}{a_k^2} \geq \frac{1}{C},$$

where the second inequality follows from the recursion. Therefore, the quantity $\frac{1}{a_k}$ grows at each step at least by $1/C$. By adding up the contributions, this implies that

$$\frac{1}{a_k} \geq \frac{1}{a_0} + \frac{k}{C}.$$

Therefore,

$$a_k \leq \left(\frac{1}{a_0} + \frac{k}{C}\right)^{-1} \leq \left(\frac{k}{C}\right)^{-1} = \frac{C}{k}.$$

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 3: Empirical Risk Minimization

Notation

- $\mathbb{R}^d$: Euclidean space of d-dimensional real vectors $x = (x_1, \ldots, x_d)$
- $\langle x, y \rangle \stackrel{\text{def}}{=} \sum_i x_i y_i$ is the **standard Euclidean inner product**
- $\|x\| \stackrel{\text{def}}{=} \langle x, x \rangle^{1/2}$ is the **standard Euclidean norm**
- $[k] \stackrel{\text{def}}{=} \{1, 2, \ldots, k\}$ for any positive integer k

Supervised Machine Learning

The Goal of Supervised Machine Learning

We wish to **learn** an approximation $h : \mathcal{A} \to \mathcal{B}$ of a function $h^\star : \mathcal{A} \to \mathcal{B}$ mapping inputs from a **domain space** $\mathcal{A}$ to outputs from a **label space** $\mathcal{B}$.
The function h is called a **predictor**, a **classifier**, or a **hypothesis**.
In other words, we wish to **learn a predictor** that will be able to predict the label $h^\star(a)$ of an unseen/new input $a \in \mathcal{A}$.

Set of "natural" objects $\mathcal{A}$	Set of labels $\mathcal{B}$		Prediction task
NYT articles	**Article category**	**(finite set)**	**Multi-class classification**
E-mails	Spam / not-spam	$\{-1, 1\}$	Binary classification
Images	Image category	(finite set)	Multi-class classification
Surveillance videos	Probability of a threat	$[0, 1]$	Regression
User clicks	Age	$(0, 150]$	Regression

Figure: Examples of prediction tasks.

Where Does the Data Come From?

We have access to **unbiased samples** of input-output pairs

$$(a, b) = (a, h^\star(a)) \in \mathcal{A} \times \mathcal{B}$$

following some distribution $\mathcal{D}$.

- We do not assume (even in theory) to know $\mathcal{D}$. Instead, we can learn about $\mathcal{D}$ by repeatedly sampling input-output pairs. All theory and methods should be distribution agnostic.
- Typically, we sample n (where n is large enough) input-output pairs, and call it the **training dataset**. We then use this dataset to learn a "good" approximation of $h^\star$.

What Function are we Learning?

We choose some **parametric class** of functions/hypotheses/models

$$h_x : \mathcal{A} \to \mathcal{B},$$

where the parameter x is described by d features; i.e., $x \in \mathcal{X} \subseteq \mathbb{R}^d$.

- The choice of the hypothesis class is crucial. If we choose a class which does not contain any function close to $h^\star$, we can't do well.
- Decisions about the hypothesis class are often "art"/"black magic" in practice. **Prior knowledge** about $h^\star$ is encoded in the selection of the hypothesis class.

Example 6 (Hypothesis classes)

- **linear model:** $h_x(a) = x^\top a$
- **linear model with feature map:** $h_x(a) = x^\top \phi(a)$, where $\phi : \mathcal{A} \to \mathbb{R}^d$ is a **feature map**
- **neural network:** $h_x(a) = x_l^\top \phi_{x_1,\ldots,x_{l-1}}(a)$, where $x = (x_1, \ldots, x_{l-1}, x_l)$ and $\phi_{x_1,\ldots,x_{l-1}} : \mathcal{A} \to \mathbb{R}^{\dim(x_l)}$ is a **learnable feature map** of the structure $\phi_{x_1,\ldots,x_{l-1}}(a) = \sigma(x_{l-1} \cdots \sigma(x_2 \sigma(x_1 a)))$, where σ is a **nonlinear activation function**

How do we Know a Prediction is Good?

We choose a **loss function**

$$\ell : \mathcal{B} \times \mathcal{B} \to \mathbb{R},$$

where $\ell(b', b)$ measures the loss of predicting b' when in fact the real output is b.

Example 7 (Loss functions)

Assume $\mathcal{B} \subseteq \mathbb{R}$.

- **hard loss:** $\ell(b', b) = 0$ if $b' = b$ and $\ell(b', b) = +\infty$ if $b' \neq b$ (used for consistent linear systems)
- **square loss:** $\ell(b', b) = \frac{1}{2}(b' - b)^2$, $b \in \mathbb{R}$ (used for regression)
- **absolute loss:** $\ell(b', b) = |b' - b|$, $b \in \mathbb{R}$ (used for regression)
- **hinge loss:** $\ell(b', b) = \max\{0, 1 - b'b\}$, $b \in \{-1, 1\}$ (used for classification)
- **logistic loss:** $\ell(b', b) = \log(1 + e^{-b'b})$, $b \in \{-1, 1\}$ (used for classification)

What Parameters are Good?

We wish to find $x \in \mathbb{R}^d$ which solves the following **stochastic optimization problem:**

$$\min_{x \in \mathbb{R}^d} \mathbf{E}_{(a,b)\sim\mathcal{D}} \left[f_{a,b}(x) \stackrel{\text{def}}{=} \ell(h_x(a), b) \right]. \tag{2}$$

- The function $\mathbf{E}_{(a,b)\sim\mathcal{D}} [f_{a,b}(x)]$ is also called **population risk**, **true risk**, or **generalization loss.**
- Problem (2) is also called the **(population) risk minimization** problem, and the value $\mathbf{E}_{(a,b)\sim\mathcal{D}} [\ell(h_x(a), b)]$ is the **generalization loss** (i.e., expected prediction loss) of model h_x.
- In problem (2), **we seek to find the parameters $x \in \mathbb{R}^d$ describing a model h_x which minimizes the population risk.**
- A key problem with (2) is that since we do not have access to $\mathcal{D}$, we may not be able to solve (2).

How to Solve (2)? Direct Attack

Attack the **risk minimization** problem directly via some stochastic optimization algorithm. Most popular method is **stochastic gradient descent (SGD):**

Algorithm 1 SGD for True Risk Minimization

1: **Parameters:** learning rates $\gamma_k > 0$, starting point $x^0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample an input-output pair $(a_k, b_k) \sim \mathcal{D}$
4: Set $g^k = \nabla f_{a_k, b_k}(x^k)$
5: $x^{k+1} = x^k - \gamma_k g^k$

Note that under some regularity conditions (conditions which allow swapping of expectation and differentiation), g^k is an unbiased estimator of the gradient of the true risk function:

$$\mathbf{E}_{(a_k,b_k)\sim\mathcal{D}}\left[g^k \mid x^k\right] = \nabla \mathbf{E}_{(a,b)\sim\mathcal{D}}\left[f_{a,b}(x^k) \mid x^k\right]$$

How to Solve (2)? Indirect Attack

- **Collect training data.** First collect a (finite) **training dataset** consisting of input-output pairs sampled from $\mathcal{D}$:

$$S_n \stackrel{\text{def}}{=} \{(a_1, b_1), (a_2, b_2), \cdots, (a_n, b_n)\}.$$

- **Work with empirical risk instead of the true risk.** If n is "large enough", the **empirical risk**, defined below, is a good approximation of the true risk

$$f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_{a_i, b_i}(x) \approx \mathbf{E}_{(a,b)\sim\mathcal{D}} \left[f_{a,b}(x) \right].$$

Note that f is a random function of the training data S_n, and that f is an unbiased estimator of the true risk. As n increases, its variance decreases.

Empirical Risk Minimization (ERM)

Instead of attempting to minimize the true risk (2), solve the **empirical risk minimization (ERM)** problem:

$$\min_{x \in \mathbb{R}^d} \left[f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \underbrace{f_{a_i,b_i}(x)}_{\stackrel{\text{def}}{=} f_i(x)} \right]. \tag{3}$$

- Any solution x^{ERM} of the ERM problem (3) is a function of S_n.
- The value $f(x^{\text{ERM}})$ is the **training loss** associated with the ERM solution x^{ERM}
- The value $\mathbf{E}_{(a,b)\sim\mathcal{D}} \left[f_{a,b}(x^{\text{ERM}}) \right]$ is the **generalization loss** associated with the ERM solution x^{ERM}.
- **Use `SGD` to solve ERM (3).**

Regularized ERM

Often in practice, we add a **regularizer** $R(x)$ to the empirical risk and solve the **regularized ERM** problem instead:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + R(x). \tag{4}$$

- There are **learning theoretic reasons to add a regularizer.** We will not talk about them. Refer to the book "Understanding Machine Learning: From Theory to Algorithms" by Shai Shalev-Shwartz and Shai Ben-David for explanation [76].
- **Regularizers can be used to encode "prior knowledge"** about the model. For instance, the L_1 regularizer,

$$R(x) = \|x\|_1 = \sum_{i=1}^{d} |x_i|,$$

  encourages sparsity in x.
- **Use Proximal SGD to solve Regularized ERM (4).**

Examples of Regularized ERM Problems I

regularized ERM problem	**loss function** $\ell(b', b)$	**loss function plot** ($b = 1$)	**regularizer** $R(x)$
Least Squares (Linear Regression)	$\frac{1}{2}(b' - b)^2$		0
L_1 Regression	$\lvert b' - b \rvert$		0
Ridge Regression (L_2 regularized Least Squares)	$\frac{1}{2}(b' - b)^2$		$\frac{\lambda}{2} \lVert x \rVert^2$
LASSO (L_1 regularized Least Squares)	$\frac{1}{2}(b' - b)^2$		$\lambda \lVert x \rVert_1$

Examples of Regularized ERM Problems II

regularized ERM problem	**loss function** $\ell(b', b)$	**loss function plot** ($b = 1$)	**regularizer** $R(x)$
Nonnegative Least Squares	$\frac{1}{2}(b' - b)^2$		$R(x) = \begin{cases} 0 & x \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$
Support Vector Machine (SVM)	$\max\{0, 1 - b'b\}$		$\frac{\lambda}{2}\ \|x\|^2$
Logistic Regression	$\log\left(1 + e^{-b'b}\right)$		$\frac{\lambda}{2}\ \|x\|^2$
Best Approximation	$\ell(b', b) = \begin{cases} 0 & b' = b \\ +\infty & b' \neq b \end{cases}$		$\|x - x^0\|$

Optimization Problems Arising in Training Supervised ML Models

Optimization Problems Arising in Machine Learning

In this course, we are interested in the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x) \tag{5}$$

- **Infinite sum:**

$$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[f_\xi(x) \right], \tag{6}$$

- **Finite sum:**

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{7}$$

  - **Finite Sum of Finite Sums:**

$$f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} f_{ij}(x) \tag{8}$$

  - **Finite Sum of Infinite Sums:**

$$f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i} \left[f_{i\xi_i}(x) \right] \tag{9}$$

These problems are of key importance in **supervised learning theory and practice.** **Common feature:** It is prohibitively expensive to compute the gradient of f, while an unbiased estimator of the gradient can be computed efficiently/cheaply.

Stochastic Optimization and Machine Learning

In the stochastic optimization problem

$$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[f_\xi(x) \right]$$

- x represents a **machine learning model** described by d parameters/features (e.g., logistic regression or a deep neural network),
- $\mathcal{D}$ is an unknown **distribution of labelled examples**,
- $f_\xi(x)$ represents the **loss** of model x on data point ξ, and
- f is the **generalization error.**

Problem (5) seeks to find the model x minimizing the generalization error.

- In statistical learning theory one assumes that while $\mathcal{D}$ is not known, samples $\xi \sim \mathcal{D}$ are available.
- In such a case, $\nabla f(x)$ is not computable, while $\nabla f_\xi(x)$, which is an unbiased estimator of the gradient of f at x, is easily computable.

Finite Sum Problems

In this course we will focus on functions f which arise as averages of a very large number of (smooth) functions:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

- This problem often arises by approximation of the stochastic optimization loss function (6) via **Monte Carlo integration.**
- Known as the **empirical risk minimization (ERM)** problem.
- ERM is currently the **dominant paradigm for solving supervised learning problems** [76].
- If index i is chosen uniformly at random from $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$, $\nabla f_i(x)$ is an **unbiased estimator of $\nabla f(x)$.**
- Typically, **$\nabla f_i(x)$ is about n times less expensive** to compute than $\nabla f(x)$.

Distributed & Federated Training

In distributed training of supervised models, one considers the finite sum problem (7), with n being the number of machines/devices, and each f_i

- also having a **finite sum structure,** i.e.,

$$f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} f_{ij}(x), \tag{10}$$

  where m_i **corresponds to the number of training examples stored on machine i.**

- or an **infinite-sum structure,** i.e.,

$$f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i} \left[f_{i\xi_i}(x) \right], \tag{11}$$

  where $\mathcal{D}_i$ **is the distribution of data stored on machine i.**

Proximal Stochastic Gradient Descent (SGD)

Proximal SGD I

Proximal stochastic gradient descent (SGD) [67, 50, 83] is a state-of-the-art algorithmic paradigm for solving optimization problems (5) in situations when f is either of structure (6) or (7).

In its generic form, (proximal) SGD defines the new iterate by subtracting a multiple of a stochastic gradient from the current iterate, and subsequently applying the proximal operator of R:

$$\boxed{x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)} \tag{12}$$

- g^k is an **unbiased estimator of the gradient** (i.e., a "stochastic gradient"):
$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k). \tag{13}$$
- $x \mapsto \mathrm{prox}_R(x)$ is the **proximal operator** (a.k.a. proximity operator):
$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \underset{u \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ R(u) + \frac{1}{2}\|u - x\|^2 \right\}. \tag{14}$$

Proximal SGD II

- Notice that
$$\mathrm{prox}_R(x) = \operatorname*{argmin}_u \phi_x(u),$$
where
$$\phi_x(u) \stackrel{\text{def}}{=} R(u) + \frac{1}{2}\|u - x\|^2.$$
- By $\operatorname{argmin}_u \phi_x(u)$ we refer to the **set of all minimizers** of ϕ_x. If ϕ_x does not have a minimizer, we set $\operatorname{argmin}_u \phi_x(u) = \emptyset$. Otherwise, $\operatorname{argmin}_u \phi_x(u)$ is defined as the **set of all minimizers** of ϕ_x.
- **Under some regularity assumptions** (we shall mention them a bit later) on R, $\mathrm{prox}_R(x)$ **is always a nonempty set consisting of a single element** (such a set is called **singleton**), which means that $\mathrm{prox}_R(x)$ is a **function.** We will be interested in this case in much of this course.

ZOO of Stochastic Gradients

There are **infinitely many** ways of obtaining a random vector g^k satisfying (13).

- **Pros: flexibility** to construct stochastic gradients in various ways **based on problem structure,** and in order to **target desirable properties** such as
  - convergence speed,
  - iteration cost,
  - overall complexity,
  - parallelizability,
  - suitability for given computing architecture,
  - communication cost,
  - generalization properties.
- **Cons: A crazy ZOO of methods.**
  - Hard to get into the field, hard to keep up with new results
  - Considerable **challenges in terms of convergence analysis.** Indeed, if one aims to, as one should, obtain the sharpest bounds possible, dedicated analyses are needed to handle each of the particular variants of SGD.

The Proximal Operator

The Proximal Operator I

The regularity property of R mentioned above is this:

Assumption 1

$R : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ *is a proper, closed and convex function.*

Next, we provide formal definitions of these three concepts:

Definition 8 (Proper functions)

An extended real-valued function $R : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is **proper** if there exists $x \in \mathbb{R}^d$ such that $R(x) < +\infty$.

Definition 9 (Closed functions)

An extended real-valued function $R : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is **closed** if its **epigraph**

$$\mathrm{epi}_R(x) \stackrel{\text{def}}{=} \{(x, t) \in \mathbb{R}^d \times \mathbb{R} \ : \ R(x) \leq t\}$$

is a closed set.

Closed functions are often also called **lower-semicontinuous.**

The Proximal Operator II

Definition 10 (Convex functions)

An extended real-valued function $R : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is **convex** if its **epigraph** is a convex set[3].

We will give a more in-depth treatment of convexity later in the course.

The proximity operator of a proper, closed and convex function has some nice properties. Some of these we will use repeatedly in the course, without proof.[4]

Proposition 2 (Single-valuedness of the prox)

If R is proper, closed and convex, then the proximal operator $x \mapsto \mathrm{prox}_R(x)$ is **single-valued.** *That is, $\mathrm{prox}_R(x)$ is a singleton for every $x \in \mathbb{R}^d$*

This is a nice property since it means that whenever we call $\mathrm{prox}_R(x)$ inside the proximal `GD` method, we get a unique output.

The Proximal Operator III

Proposition 3 (Non-expansiveness of the prox)

If R is proper, closed and convex, then the proximal operator $x \mapsto \mathrm{prox}_R(x)$ is **non-expansive:**

$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{15}$$

We will use this property often in the proofs of the SGD methods we will encounter in this class. In fact, the proximity operator of a proper, closed, and convex function R satisfies a stronger property, called **firm-nonexpansiveness:**

$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|^2 \leq \langle \mathrm{prox}_R(x) - \mathrm{prox}_R(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d.$$

It is easy to verify that this is a stronger property, i.e., that firm-nonexpansiveness implies non-expansiveness.

The Proximal Operator IV

Proposition 4 (Fixed points of prox and minimization)

Let $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed and convex function. Then

$$x^\star \in \arg\min_x \phi(x) \quad \Longleftrightarrow \quad x^\star = \mathrm{prox}_\phi(x^\star).$$

The above result can be shown to imply the following about the problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x)$$

we are interested in. If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, and

$$x^\star \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^d} f(x) + R(x),$$

then

$$x^\star = \mathrm{prox}_{\gamma R}\left(x^\star - \gamma \nabla f(x^\star)\right) \tag{16}$$

for all $\gamma \geq 0$. We will use identity (16) often in the proofs of the SGD methods in this course.

The Proximal Operator V

Proposition 5 (Moreau decomposition of prox)

Assume R is proper, closed and convex. Then

$$\mathrm{prox}_R(x) + \mathrm{prox}_{R^\star}(x) = x, \quad \forall x \in \mathbb{R}^d,$$

where $R^\star$ is the **Fenchel conjugate** *of R defined as follows:*

$$R^\star(x) \stackrel{def}{=} \sup_{u \in \mathbb{R}^d} \left\{ \langle x, u \rangle - R(u) \right\}.$$

While we will not use it in the proofs, this identity has the following use: if we know how to compute the prox of R, we know how to compute the prox of its Fenchel conjugate $R^\star$, and vice versa.

The Proximal Operator VI

Proposition 6 (Fenchel conjugate)

The Fenchel conjugate $R^\star$ of an extended real-valued function $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed and convex. If R is proper and convex, then $R^\star$ is proper. If R is proper, closed and convex, then so is $R^\star$, and moreover, $R = R^\star$.

[3] Recall that a subset S of a vector space is convex if $\alpha x + (1-\alpha)y \in S$ whenever $x, y \in S$ and $\alpha \in [0, 1]$.

[4] For more information about the proximity operator, we refer to the very nice book of Amir Beck [3].

Exercises

Exercises I

Exercise 1 (Proximal operator of univariate functions)

Compute the proximal operator $x \mapsto \mathrm{prox}_R(x)$ *(defined in* (14)*) of the following functions* $R : \mathbb{R} \to \mathbb{R}$*:*

(i) $R(x) = c$ *for some* $c \in \mathbb{R}$.

(ii) $R(x) = bx + c$, *where* $b, c \in \mathbb{R}$.

(iii) $R(x) = \frac{1}{2}ax^2$, *where* $a > 0$.

(iv) $R(x) = \begin{cases} 0 & x \neq 0 \\ c & x = 0 \end{cases}$.

(v) $R(x) = \begin{cases} 0 & x \in [a, b] \\ +\infty & \textit{otherwise} \end{cases}$, *where* $a, b \in \mathbb{R}$ *and* $a < b$.

(vi) $R(x) = c|x|$, *where* $c > 0$.

(vii) $R(x) = \begin{cases} cx & x \geq 0 \\ +\infty & \textit{otherwise} \end{cases}$, *where* $c \in \mathbb{R}$.

Exercises II

Exercise 2 (Proximal operator of convex quadratic functions)

Show that

(i) *The proximal operator of the constant function*

$$R(x) = c,$$

where $c \in \mathbb{R}$, *is*

$$\operatorname{prox}_R(x) = x.$$

(ii) *The proximal operator of the affine function*

$$R(x) = \langle b, x \rangle + c,$$

where $b \in \mathbb{R}^d$ *and* $c \in \mathbb{R}$, *is*

$$\operatorname{prox}_R(x) = x - b.$$

(iii) *The proximal operator of the convex quadratic function*

$$R(x) = \frac{1}{2}\langle \boldsymbol{A}x, x \rangle + \langle b, x \rangle + c,$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ *is symmetric and positive semidefinite,* $b \in \mathbb{R}^d$ *and* $c \in \mathbb{R}$, *is*

$$\operatorname{prox}_R(x) = (\boldsymbol{A} + \boldsymbol{I})^{-1}(x - b).$$

Exercises III

Exercise 3 (Proximal operator of separable functions)

Let $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *be of the form*

$$R(x) = R_1(x_1) + \cdots + R_d(x_d).$$

Then

$$\operatorname{prox}_R(x) = \begin{pmatrix} \operatorname{prox}_{R_1}(x_1) \\ \operatorname{prox}_{R_2}(x_2) \\ \vdots \\ \operatorname{prox}_{R_d}(x_d) \end{pmatrix}.$$

Exercise 4 (Firm nonexspansiveness implies non-expansiveness)

Prove that firm-nonexpansiveness implies non-expansiveness.

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 4: Gradient Descent

Gradient Descent

We first describe the (proximal) gradient descent (GD) method for solving the regularized convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x). \tag{17}$$

Arguably, this is the most basic of all SGD methods, and a starting point for the development of more elaborate variants.

Algorithm 2 GD

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Set $g^k = \nabla f(x^k)$
4: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

Convexity and Smoothness: Quick Introduction

Convexity and L-smoothness

Definition 11 (Convexity)

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$0 \leq \underbrace{D_f(x,y) \stackrel{\text{def}}{=} f(x) - f(y) - \langle \nabla f(y), x - y \rangle}_{\textbf{Bregman divergence} \text{ associated with } f}, \quad \forall x, y \in \mathbb{R}^d.$$

Definition 12 (L-smoothness)

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if $\exists$ constant $L \geq 0$ such that

$$2D_f(x,y) \leq L \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Remark: We will show later (see Proposition 17) that a convex and L-smooth function satisfies the inequality

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2D_f(x,y), \quad \forall x, y \in \mathbb{R}^d. \tag{18}$$

This inequality will play a crucial role in the analyses of the methods in this course. In the proofs we will use it for $x = x^k$ (iterate produced by SGD) and $y = x^\star$ (optimal solution).

Strong Convexity

Definition 13 (μ-strong convexity)

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-convex (or μ–strongly-convex for short) if there exists $\mu > 0$ such that

$$\mu \|x - y\|^2 \leq 2D_f(x, y), \quad \forall x, y \in \mathbb{R}^d. \tag{19}$$

It is easy to check that (19) can be equivalently written in the form

$$D_f(x, y) + \frac{\mu}{2} \|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d. \tag{20}$$

In fact, we will only use this inequality for $y = x^\star$, which can be equivalently written in the following forms

$$D_f(x, x^\star) + \frac{\mu}{2} \|x - x^\star\|^2 \leq \langle \nabla f(x) - \nabla f(x^\star), x - x^\star \rangle, \quad \forall x \in \mathbb{R}^d, \tag{21}$$

and

$$f(x^\star) \geq f(x) + \langle \nabla f(x), x^\star - x \rangle + \frac{\mu}{2} \|x^\star - x\|^2, \quad \forall x \in \mathbb{R}^d. \tag{22}$$

Convergence Analysis of Gradient Descent

A Descent Lemma

We first start with a lemma which goes almost all the way to establishing the convergence of GD.

Lemma 14

Assume f is L-smooth and μ-convex. Then the following inequality holds for all $\gamma \geq 0$ and all $k \geq 0$:

$$\left\|x^{k+1} - x^\star\right\|^2 \leq (1 - \gamma\mu) \left\|x^k - x^\star\right\|^2 - 2\gamma (1 - \gamma L) D_f(x^k, x^\star). \quad (23)$$

Proof of Lemma 14

Letting $r^k \stackrel{\text{def}}{=} x^k - x^\star$, we get

$$\begin{aligned}
\left\|r^{k+1}\right\|^2 &= \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma\nabla f(x^k)\right) - x^\star\right\|^2 \\
&\stackrel{(16)}{=} \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma\nabla f(x^k)\right) - \mathrm{prox}_{\gamma R}\left(x^\star - \gamma\nabla f(x^\star)\right)\right\|^2 \\
&\stackrel{(15)}{\leq} \left\|(x^k - \gamma\nabla f(x^k)) - (x^\star - \gamma\nabla f(x^\star))\right\|^2 \\
&= \left\|x^k - x^\star - \gamma\left(\nabla f(x^k) - \nabla f(x^\star)\right)\right\|^2 \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle + \gamma^2\left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2,
\end{aligned}$$

where the inequality follows from nonexpansiveness of the proximal operator. We now use strong convexity to bound

$$\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle \stackrel{(21)}{\geq} D_f(x^k, x^\star) + \frac{\mu}{2}\left\|x^k - x^\star\right\|^2 \tag{24}$$

and L-smoothness to bound

$$\left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2 \stackrel{(18)}{\leq} 2LD_f(x^k, x^\star). \tag{25}$$

Applying the bounds (24) and (25), we obtain

$$\left\|r^{k+1}\right\|^2 \leq (1 - \gamma\mu)\left\|r^k\right\|^2 - 2\gamma(1 - \gamma L)D_f(x^k, x^\star),$$

which finalizes the proof.

Convergence of Gradient Descent

We can now state the main result of this chapter.

Theorem 15 (Convergence of GD)

Assume f is L-smooth and μ-convex. Then the iterates of Gradient Descent (Algorithm 2) with stepsize $0 \leq \gamma \leq \frac{1}{L}$ satisfy

$$\left\|x^k - x^\star\right\|^2 \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2. \tag{26}$$

Proof.

Follows from Lemma 14 using the observation that the stepsize choice ensures that $2\gamma(1 - \gamma L) \geq 0$, and the fact that convexity of f implies the inequality $D_f(x^k, x^\star) \geq 0$. □

We may want to ask how many GD iterations it takes for the error to reach a certain target level ε. Standard arguments starting from the inequality (26) show that with $\gamma = \frac{1}{L}$, we have

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \left\|x^k - x^\star\right\|^2 \leq \varepsilon \left\|x^0 - x^\star\right\|^2. \tag{27}$$

Exercises

Exercises I

Exercise 5 (Strong convexity 1)

Show that (20) *follows from* (19).

Exercise 6 (Strong convexity 2)

Show that (20) *implies that* $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$.

Exercise 7 (Iteration complexity of GD)

Show that inequality (27) *follows from* (26).

Exercise 8 (An inequality satisfied by μ-convex functions)

Show that if a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-convex, then*

$$2D_f(x, x^\star) \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(x^\star)\|^2 .$$

Exercises II

Exercise 9 (Larger constant stepsize)

Let $R \equiv 0$, and assume that $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, with $\mu \boldsymbol{I} \preceq \nabla^2 f(x) \preceq L\boldsymbol{I}$ for all $x \in \mathbb{R}^d$, where $0 < \mu \leq L$ (in particular, this means that f is L-smooth and μ-convex). Let

$$T_\gamma(x) \stackrel{def}{=} x - \gamma \nabla f(x)$$

be the gradient mapping.

(i) *Show that as long as $0 < \gamma < \frac{2}{L}$, we have*

$$\| T_\gamma(x) - T_\gamma(y) \| \leq \max\{|1 - \gamma\mu|, |1 - \gamma L|\} \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{28}$$

(ii) *Show that the optimal stepsize (in terms of minimizing the upper bound in (28)) is*

$$\gamma = \frac{2}{\mu + L},$$

which gives the contraction

$$\| T_\gamma(x) - T_\gamma(y) \| \leq \frac{L - \mu}{L + \mu} \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

(iii) *Formulate a convergence theorem for* GD *from the inequality from part (ii). Compare this rate to that from Theorem 15. Which one is better? By how much?*

Exercises III

Exercise 10 (Polyak stepsizes)

Consider a variant of GD *with* **Polyak stepsize**[5]*, i.e., with the* **adaptive** *(i.e., iteration-dependent)* **stepsize** *rule*

$$\gamma_k = \frac{2D_f(x^k, x^\star)}{\left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2},$$

where we assume that $\nabla f(x^k) \neq \nabla f(x^\star)$ *for this to be well-defined. That is, we consider the method*

$$x^{k+1} = \mathrm{prox}_{\gamma_k R}\left(x^k - \gamma_k \nabla f(x^k)\right).$$

(i) *Show that* $\gamma_k \geq \frac{1}{L}$*. That is, the Polyak stepsize is always at least as large as the largest constant stepsize* $\gamma = \frac{1}{L}$ *allowed by Theorem 15.*

(ii)* *Show that* $\gamma_k \leq \frac{1}{\mu}$*.*

(iii) *Prove that after one step of* GD *with Polyak stepsize we have*

$$\left\|x^{k+1} - x^\star\right\|^2 \leq (1 - \gamma_k \mu)\left\|x^k - x^\star\right\|^2. \tag{29}$$

[5] Polyak stepsize makes most sense when we know that $\nabla f(x^\star) = 0$ (which happens, for example, when $R \equiv 0$), and when we know $f(x^\star)$. If we do not know these quantities, Polyak stepsize obviously cannot be implemented.

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 5: Convexity and Smoothness

Differentiable Functions

Fundamental Theorem of Calculus - I

We will use the following theorem repeatedly in various disguises.

Theorem 16

Let $\phi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable on an open interval containing points a and b. Then

$$\phi(b) - \phi(a) = \int_a^b \phi'(t)\, dt. \tag{30}$$

Corollary 17

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. Then for any $x, y \in \mathbb{R}^d$,

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt. \tag{31}$$

Proof.

Let $\phi(t) = f(x + t(y - x))$. Note that $\phi(0) = f(x)$, $\phi(1) = f(y)$ and that by chain rule, $\phi'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$. The rest follows by applying Theorem 16. $\square$

Fundamental Theorem of Calculus - II

Corollary 18

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Then for any $x, y, s \in \mathbb{R}^d$,

$$\langle \nabla f(y) - \nabla f(x), s \rangle = \int_0^1 \langle \nabla^2 f(x + t(y - x))(y - x), s \rangle \, dt. \tag{32}$$

Proof.

Let $\phi(t) \stackrel{\text{def}}{=} \langle \nabla f(x + t(y - x)), s \rangle$. Note that $\phi(0) = \langle \nabla f(x), s \rangle$, $\phi(1) = \langle \nabla f(y), s \rangle$, and that by chain rule,

$$\phi'(t) = \langle \nabla^2 f(x + t(y - x))(y - x), s \rangle.$$

The rest follows by applying Theorem 16. $\square$

Bregman Divergence

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.

Definition 19 (Bregman Divergence)

Bregman divergence of f is the mapping $D_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by:

$$D_f(x, y) \stackrel{\text{def}}{=} f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

- Note that D_f is not necessarily symmetric, i.e., in general $D_f(x, y) \neq D_f(y, x)$.
- This can be fixed by defining a symmetric version of Bregman divergence.

Definition 20 (Symmetrized Bregman Divergence)

Symmetrized Bregman divergence of f is defined by:

$$D_f(x, y) + D_f(y, x) = \langle \nabla f(x) - \nabla f(y), x - y \rangle = \langle \nabla f(y) - \nabla f(x), y - x \rangle$$

Convex Functions

Convex Functions: Definition

Definition 21 (Convex function)

Function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$0 \leq \lambda f(x) + (1 - \lambda) f(y) - f(\lambda x + (1 - \lambda) y) \stackrel{\text{def}}{=} \mathbf{C_f^\lambda(x, y)}$$

Some facts:

- If f_1, f_2 are convex and $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 f_1 + \alpha_2 f_2$ is convex.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, then $x \mapsto f(\boldsymbol{A}x + b)$ is convex.
- If f_1, f_2 are convex, then so is $\max\{f_1, f_2\}$.

Convex Functions: Characterization

Proposition 7 (Convexity and gradient)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. Then the following statements are equivalent:

(i) *f is convex (**convexity of the function**)*

(ii) $0 \leq D_f(x, y)$ *for all* $x, y \in \mathbb{R}^d$
*(**nonnegativity of the Bregman divergence**)*

(iii) $0 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ *for all* $x, y \in \mathbb{R}^d$
*(**monotonicity of the gradient**)*

Proposition 8 (Convexity and Hessian)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Then f is convex if and only if

$$0 \preceq \nabla^2 f(x) \quad \text{for all} \quad x \in \mathbb{R}^d.$$

*(**positive semidefiniteness of the Hessian**)*

Convex Functions: Examples

Example 22

- Linear function $f(x) = \langle b, x \rangle + c$ is convex.
- Quadratic function $f(x) = \frac{1}{2}\langle \boldsymbol{A}x, x \rangle + \langle b, x \rangle + c$ is convex if and only if $\boldsymbol{A} \succeq 0$.
- Exponential function $f(t) = e^t$ is convex.
- The Fenchel conjugate $f^*(x) \stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^d} \langle x, y \rangle - f(y)$ of any function $f : \mathbb{R}^d \to \mathbb{R}$ is convex.[6]

[6] The supremum, and hence f^*, can take the value $+\infty$. We only defined convexity of real-valued functions so far. We say that f^* is convex if $f^*(\lambda x + (1-\lambda)y) \leq \lambda f^*(x) + (1-\lambda) f^*(y)$ for all $x, y \in \mathbb{R}^d$ for which $f^*(x)$ and $f^*(y)$ are finite and for all $\lambda \in [0, 1]$. Equivalently, we say that a function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex if its domain $\operatorname{dom} \phi \stackrel{\text{def}}{=} \{x \mid \phi(x) < +\infty\}$ is a convex set.

Proof of Proposition 7 - I

- (i)⇒(ii) Choose any $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1)$. Define $z = \lambda x + (1 - \lambda)y$. Then

$$\begin{aligned} f(y) &\geq \frac{1}{1-\lambda}(f(z) - \lambda f(x)) \\ &= f(x) + \frac{1}{1-\lambda}(f(z) - f(x)) \\ &= f(x) + \frac{1}{1-\lambda}(f(x + (1-\lambda)(y-x)) - f(x)). \end{aligned}$$

  By taking limit $\lambda \to 1$, we get

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle,$$

  and hence (ii) holds.

- (ii)⇒(i) Choose any $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$ and define $z = \lambda x + (1 - \lambda)y$. Since $D_f(y, z) \geq 0$ and $D_f(x, z) \geq 0$, we have

$$f(z) \leq f(y) + \langle \nabla f(z), z - y \rangle = f(y) + \lambda \langle \nabla f(z), x - y \rangle$$

  and

$$f(z) \leq f(x) + \langle \nabla f(z), z - x \rangle = f(x) - (1 - \lambda) \langle \nabla f(z), x - y \rangle$$

  Multiplying the first inequality by $1 - \lambda$, the second by λ and adding the resulting inequalities, we get $f(z) \leq \lambda f(x) + (1 - \lambda) f(y)$.

Proof of Proposition 7 - II

- (ii) $\Rightarrow$ (iii) Choose any $x, y \in \mathbb{R}^d$. Adding the inequalities $D_f(x, y) \geq 0$ and $D_f(y, x) \geq 0$, we get (iii).
- (iii) $\Rightarrow$ (ii) Choose any $x, y \in \mathbb{R}^d$ and define $z = x + t(y - x)$. Then by the fundamental theorem of calculus,

$$
\begin{aligned}
f(y) &\overset{(31)}{=} f(x) + \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(z) - \nabla f(x), y - x \rangle dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \frac{1}{t} \langle \nabla f(z) - \nabla f(x), z - x \rangle dt \\
&\geq f(x) + \langle \nabla f(x), y - x \rangle,
\end{aligned}
$$

where the inequality follows from the bound $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0$, which holds due to (iii).

Proof of Proposition 8 - I

- Assume f is convex. Choose any $x, y' \in \mathbb{R}^d$ and $\theta > 0$. Then in view of Proposition 7(iii), we have

$$\begin{aligned} 0 &\leq \frac{1}{\theta^2}\langle \nabla f(x+\theta(y'-x)) - \nabla f(x), (x+\theta(y'-x)) - x\rangle \\ &= \underbrace{\frac{1}{\theta}\langle \nabla f(x+\theta(y'-x)) - \nabla f(x), y'-x\rangle}_{\stackrel{\text{def}}{=}\mathcal{A}_\theta}, \end{aligned}$$

and hence $\mathcal{A}_\theta \geq 0$ for all $\theta > 0$. Applying the fundamental theorem of calculus (Corollary 18) with $x \leftarrow x$, $y \leftarrow x + \theta(y'-x)$ and $s \leftarrow y' - x$ gives

$$\begin{aligned} \mathcal{A}_\theta &\stackrel{(32)}{=} \frac{1}{\theta}\int_0^1 \langle \nabla^2 f(x+t(y-x))(y-x), y'-x\rangle \, dt \\ &= \int_0^1 \langle \nabla^2 f(x+\theta t(y'-x))(y'-x), y'-x\rangle \, dt. \end{aligned}$$

Proof of Proposition 8 - II

Finally, taking limit $\theta \to 0$, we get

$$\begin{aligned} 0 &\leq \lim_{\theta \to 0} \mathcal{A}_\theta \\ &= \lim_{\theta \to 0} \int_0^1 \langle \nabla^2 f(x + \theta t(y' - x))(y' - x), y' - x \rangle \, dt \\ &= \int_0^1 \lim_{\theta \to 0} \langle \nabla^2 f(x + \theta t(y' - x))(y' - x), y' - x \rangle \, dt \\ &= \int_0^1 \langle \nabla^2 f(x)(y' - x), y' - x \rangle \, dt \\ &= \langle \nabla^2 f(x)(y' - x), y' - x \rangle. \end{aligned}$$

Since y' was arbitrary, this means that $\nabla^2 f(x) \succeq 0$.

- Proof of the reverse implication is left as an exercise.

Jensen's Inequality

Proposition 9 (Jensen's Inequality)

If f is convex and $X \in \mathbb{R}^d$ is a random vector, then

$$0 \leq \mathbf{E}\left[f(X)\right] - f(\mathbf{E}\left[X\right]).$$

Strongly Convex Functions

Strongly Convex Functions: Definition

Definition 23 (Strongly convex function)

Let $\mu \geq 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is **μ-strongly convex** (**μ-convex** for short) if the function

$$x \mapsto f(x) - \frac{\mu}{2} \|x\|^2$$

is convex. That is, if

$$\frac{\mu}{2}\lambda(1-\lambda) \|x-y\|^2 \leq C_f^\lambda(x,y)$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$.

Note that the $\mu = 0$ case reduces to convexity.

Strongly Convex Functions: Examples

Example 24

Let $\mu \geq 0$. $f(x) = \frac{\mu}{2} \|x\|^2$ is **μ-convex** since $f(x) - \frac{\mu}{2} \|x\|^2 = 0$ is convex.

Example 25

Let $f(x) = \frac{1}{2} x^\top \boldsymbol{A} x$, where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Let

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$$

be the eigenvalue decomposition of $\boldsymbol{A}$. We know that all eigenvalues $\{\lambda_i\}$ are real, and that $\sum_i u_i u_i^\top = \boldsymbol{I}$. Thus,

$$f(x) - \frac{\mu}{2} \|x\|^2 = \frac{1}{2} \sum_{i=1}^{d} (\lambda_i - \mu)(x^\top u_i)^2.$$

Clearly, f is **μ-convex** if and only if the last expression is convex in x. Since $x \mapsto (x^\top u_i)^2$ is convex for all i, and since a conic (i.e., linear with nonnegative coefficients) combination of convex functions is convex, f is **μ-convex** if $\min_i \lambda_i \geq \mu$; i.e., if the smallest eigenvalue of $\boldsymbol{A}$ is lower bounded by μ.

Strongly Convex Functions: Characterization

Proposition 10 (Strong convexity and gradient)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. Then the following statements are equivalent:

(i) *f is* **μ-convex**

(ii) $\mu \|x - y\|^2 \leq 2D_f(x, y)$ *for all* $x, y \in \mathbb{R}^d$

(iii) $\mu \|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ *for all* $x, y \in \mathbb{R}^d$

Proposition 11 (Strong convexity and gradient II)

If $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and **μ-convex**, *then*[7]

(ii') $2D_f(x, y) \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2$ *for all* $x, y \in \mathbb{R}^d$

(iii') $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2$ *for all* $x, y \in \mathbb{R}^d$

Proposition 12 (Strong convexity and Hessian)

If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then it is **μ-convex** *if and only if*

$$\mu \boldsymbol{I} \preceq \nabla^2 f(x) \quad \text{for all} \quad x \in \mathbb{R}^d.$$

[7] If $\mu = 0$, we interpret $1/\mu$ as $+\infty$, and the bounds hold trivially.

Strong Jensen's Inequality

Proposition 13 (Strong Jensen's Inequality)

*If f is μ-**convex** and $X \in \mathbb{R}^d$ is a random vector, then*

$$\frac{\mu}{2}\mathbf{Var}\left[X\right] \leq \mathbf{E}\left[f(X)\right] - f(\mathbf{E}\left[X\right]).$$

Proof.

Let us apply Jensen's inequality $F(\mathbf{E}\left[X\right]) \leq \mathbf{E}\left[F(X)\right]$ to $F(x) \stackrel{\text{def}}{=} f(x) - \frac{\mu}{2}\left\|x\right\|^2$, which is convex by definition. We get

$$f(\mathbf{E}\left[X\right]) - \frac{\mu}{2}\left\|\mathbf{E}\left[X\right]\right\|^2 \leq \mathbf{E}\left[f(X)\right] - \frac{\mu}{2}\mathbf{E}\left[\left\|X\right\|^2\right].$$

It remains to rearrange the inequality and use the identity

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[\left\|X\right\|^2\right] - \left\|\mathbf{E}\left[X\right]\right\|^2.$$

□

Smooth Functions

Smooth Functions: Definition

Definition 26 (Smooth function)

Function $f : \mathbb{R}^d \to \mathbb{R}$ is **L-smooth** if the function

$$x \mapsto \frac{L}{2} \|x\|^2 - f(x)$$

is convex. That is, if

$$C_f^\lambda(x, y) \leq \frac{L}{2} \lambda(1 - \lambda) \|x - y\|^2$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$.

Smooth Functions: Examples

Example 27

Let $L \geq 0$. $f(x) = \frac{L}{2}\|x\|^2$ is **L-smooth** since $\frac{L}{2}\|x\|^2 - f(x) = 0$ is convex.

Example 28

Let $f(x) = \frac{1}{2}x^\top \boldsymbol{A}x$, where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Let

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$$

be the eigenvalue decomposition of $\boldsymbol{A}$. We know that all eigenvalues $\{\lambda_i\}$ are real, and that $\sum_i u_i u_i^\top = \boldsymbol{I}$. Thus,

$$\frac{L}{2}\|x\|^2 - f(x) = \frac{1}{2}\sum_{i=1}^{d}(L - \lambda_i)(x^\top u_i)^2.$$

Clearly, f is **L-smooth** if and only if the last expression is convex in x. Since $x \mapsto (x^\top u_i)^2$ is convex for all i, and since a conic (i.e., linear with nonnegative coefficients) combination of convex functions is convex, f is **L-smooth** if $\max_i \lambda_i \leq L$; i.e., if the largest eigenvalue of $\boldsymbol{A}$ is upper bounded by L.

Smooth Functions: Characterization

Proposition 14 (L-smooth functions and gradient)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. Then the following statements are equivalent:

(i) *f is* **L-smooth**

(ii) $2D_f(x, y) \leq L \|x - y\|^2$ *for all* $x, y \in \mathbb{R}^d$

(iii) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \|x - y\|^2$ *for all* $x, y \in \mathbb{R}^d$

Proposition 15 (L-smooth functions and Hessian)

If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then it is **L-smooth** *if and only if*

$$\nabla^2 f(x) \preceq L\boldsymbol{I} \quad \text{for all} \quad x \in \mathbb{R}^d.$$

Reverse Jensen's inequality

Proposition 16 (Reverse Jensen's inequality)

If f is **L-smooth** *and $X \in \mathbb{R}^d$ is a random vector, then*

$$\mathbf{E}\left[f(X)\right] - f(\mathbf{E}\left[X\right]) \leq \frac{L}{2}\mathbf{Var}\left[X\right].$$

Proof.

Let us apply Jensen's inequality $F(\mathbf{E}\left[X\right]) \leq \mathbf{E}\left[F(X)\right]$ to $F(x) \stackrel{\text{def}}{=} \frac{L}{2}\left\|x\right\|^2 - f(x)$, which is convex by definition. We get

$$\frac{L}{2}\left\|\mathbf{E}\left[X\right]\right\|^2 - f(\mathbf{E}\left[X\right]) \leq \frac{L}{2}\mathbf{E}\left[\left\|X\right\|^2\right] - \mathbf{E}\left[f(X)\right].$$

It remains to rearrange the inequality and use the identity

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[\left\|X\right\|^2\right] - \left\|\mathbf{E}\left[X\right]\right\|^2.$$

□

Convex and Smooth Functions

Convex and Smooth Functions: Characterization

Proposition 17 (Convexity & smoothness versus gradient)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then the following statements are equivalent:

(i) *f is* **convex** *and* ***L*-smooth**

(ii) $\mathbf{0} \leq 2D_f(x, y) \leq L \|x - y\|^2$ *for all* $x, y \in \mathbb{R}^d$

(iii) $\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq 2D_f(x, y)$ *for all* $x, y \in \mathbb{R}^d$

(iv) $\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ *for all* $x, y \in \mathbb{R}^d$

(v) $\frac{\lambda(1-\lambda)}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq C_f^\lambda(x, y)$ *for all* $x, y \in \mathbb{R}^d$ *and* $\lambda \in [0, 1]$

Proposition 18 (Convexity & smoothness versus Hessian)

If f is twice differentiable, then it is **convex** *and* ***L*-smooth** *if and only if*

$$\mathbf{0}\boldsymbol{I} \preceq \nabla^2 f(x) \preceq L\boldsymbol{I} \quad \text{for all} \quad x \in \mathbb{R}^d.$$

Jensen's and Reverse Jensen's inequality

Proposition 19 (Jensen's and reverse Jensen's inequality)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be **convex** *and* ***L*-smooth** *and let $X \in \mathbb{R}^d$ be a random vector. Then*

$$\mathbf{0} \leq \mathbf{E}\left[f(X)\right] - f(\mathbf{E}\left[X\right]) \leq \frac{L}{2}\mathbf{Var}\left[X\right].$$

Strongly Convex and Smooth Functions

Strongly Convex and Smooth Functions I

Recall that if $f : \mathbb{R}^d \to \mathbb{R}$ is μ-convex and L-smooth, then

$$\frac{\mu}{2}\|x-y\|^2 \le D_f(x,y) \le \frac{L}{2}\|x-y\|^2,$$

and hence

$$\underbrace{C_f(x,y) + \frac{\mu}{2}\|x-y\|^2}_{L_f(x,y)} \le f(x) \le \underbrace{C_f(x,y) + \frac{L}{2}\|x-y\|^2}_{U_f(x,y)},$$

where

$$C_f(x,y) \stackrel{\text{def}}{=} f(y) + \langle \nabla f(y), x-y \rangle.$$

is the linear approximation of $f(x)$ computed at y.

On the next slide, we will illustrate this geometrically via an example.

Strongly Convex and Smooth Functions II

Example 29

The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \ln(1 + e^{-\theta x}) + \frac{\zeta}{2}x^2$ is L-smooth for any $L \geq \frac{\theta^2}{4} + \zeta$ and μ-convex for any $0 \leq \mu \leq \zeta$. In the below plots we choose $\theta = 2$ and $\zeta = 0.1$.

Figure: Functions $f(x)$, $C_f(x, y)$, $U_f(x, y)$ and $L_f(x, y)$ for $y = 10$, $L = 1.01$ and $\mu = 0.05$.

Strongly Convex and Smooth Functions III

Note that if f is L-smooth and μ-convex and the following inequalities hold:

$$\mu \|x - y\|^2 \leq 2D_f(x, y), \qquad \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2D_f(x, y).$$

By adding them up and applying the identity

$$D_f(x, y) + D_f(y, x) = \langle \nabla f(x) - \nabla f(y), x - y \rangle,$$

we get

$$\mu \|x - y\|^2 + \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2\langle \nabla f(x) - \nabla f(y), x - y \rangle. \quad (33)$$

Strongly Convex and Smooth Functions IV

However, inequality (33) can be strengthened via a more refined argument:

Proposition 20

*If $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, μ-**convex** and L-**smooth,** then*

$$\mu \|x - y\|^2 + \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \left(1 + \frac{\mu}{L}\right) \langle \nabla f(x) - \nabla f(y), x - y \rangle$$

for all $x, y \in \mathbb{R}^d$.

Remarks:

- Since $\frac{\mu}{L} \leq 1$, this is indeed a strengthening of inequality (33).
- This inequality can be used to obtain a sharper analysis of GD, and can also be used to obtain a sharper analysis of some variants of SGD.

Exercises

Exercises I

Exercise 11 (Three point identity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Show that the following identity holds for all $x, y, z \in \mathbb{R}^d$:

$$D_f(x, z) = D_f(x, y) + \langle x - y, \nabla f(y) - \nabla f(z) \rangle + D_f(y, z). \tag{34}$$

(i) *Show that in the special case with $x = z$, (34) reduces to*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = D_f(x, y) + D_f(y, x).$$

(ii) *Show that in the special case for $f(x) = \frac{1}{2} \|x\|^2$, (34) reduces to*

$$\frac{1}{2} \|x - z\|^2 = \frac{1}{2} \|x - y\|^2 + \langle x - y, y - z \rangle + \frac{1}{2} \|y - z\|^2 .$$

Exercises II

Exercise 12 (Completing the proof of Proposition 8)

Prove the other implication to complete the proof of Proposition 8.

Exercise 13 (Jensen's Inequality)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, $x_1, \ldots, x_m \in \mathbb{R}^d$ and $\lambda_1, \ldots, \lambda_m$ be nonnegative real numbers adding up to 1. Establish the following variant of Jensen's inequality:

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i). \tag{35}$$

Exercise 14 (Characterization of strong convexity)

Prove Proposition 10.

Exercise 15 (Strong convexity upper bounds)

Prove Proposition 11.

Exercises III

Exercise 16 (Strong convexity and Hessian)

Prove Proposition 12.

Exercise 17 (Characterization of L-smoothness)

Prove Proposition 14.

Exercise 18 (L-smoothness and Hessian)

Prove Proposition 15.

Exercise 19 (Bregman divergence)

Assume $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Show that then the Bregman divergence satisfies the following properties:

(i) *If f is convex, then $D_f(x, y) \geq 0$ for all $x, y \in \mathbb{R}^d$*

(ii) *$D_f(x, x) = 0$ for all $x \in \mathbb{R}^d$*

(iii) *If f is strongly convex, then $D_f(x, y) = 0$ if and only if $x = y$*

Exercises IV

Exercise 20 (Lipschitz gradient)

We say that f has L-Lipschitz gradient if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d. \tag{36}$$

Show that if f has L-Lipschitz gradient, then

$$-\frac{L}{2}\|x - y\|^2 \leq D_f(x, y) \leq \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

Exercise 21

Show that the univariate function $f(x) = \log(x + \sqrt{1 + x^2}) + x^2$

- *is strongly convex,*
- *has Lipschitz gradient.*

Exercises V

Exercise 22 (Gradient mapping)

Let $T_\gamma(x) \stackrel{def}{=} x - \gamma \nabla f(x)$ *be the* **gradient mapping** *operator. Show that if* f *has* *L-Lipschitz gradient,* *then for all* $x \in \mathbb{R}^d$ *and* $\gamma \geq 0$*:*

$$\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla f(x)\|^2 \leq f(x) - f(T_\gamma(x)) \leq \gamma \left(1 + \frac{3\gamma L}{2}\right) \|\nabla f(x)\|^2 .$$

Exercise 23

In the proof of Lemma 14, we have relied on inequality (20) *to capitalize on strong convexity of* f*. Strong convexity was used in a single step of the proof. That step did not use the fact that* f *is* also *L-smooth. It is possible to perform the proof differently and obtain a slightly stronger result. Do the analysis.*

Hint: Use Proposition 20.

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 6: Stochastic Gradient Descent

Based on:

[23] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P.R. **SGD: General Analysis and Improved Rates**, ICML 2019

[19] E. Gorbunov, F. Hanzely and P.R. **A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent**, AISTATS 2020

Basic Facts

Basic Facts - I

• **Young's inequality.** For all $a, b \in \mathbb{R}^d$ and $t > 0$ the following inequalities holds:

$$\langle a, b \rangle \leq \frac{\|a\|^2}{2t} + \frac{t\,\|b\|^2}{2}, \tag{37}$$

$$\|a + b\|^2 \leq 2\,\|a\|^2 + 2\,\|b\|^2\,, \tag{38}$$

and

$$\frac{1}{2}\,\|a\|^2 - \|b\|^2 \leq \|a + b\|^2\,. \tag{39}$$

• **Variance decomposition.** For a random vector $X \in \mathbb{R}^d$ (with finite second moment) and any $c \in \mathbb{R}^d$ the variance can be decomposed as

$$\mathbf{E}\left[\|X - \mathbf{E}[X]\|^2\right] = \mathbf{E}\left[\|X - c\|^2\right] - \|\mathbf{E}[X] - c\|^2\ . \tag{40}$$

Some consequences:

Basic Facts - II

- The solution of the optimization problem

$$\min_{c \in \mathbb{R}^d} \mathbf{E}\left[\|X - c\|^2\right]$$

  is $c = \mathbf{E}[X]$.
- For $c = 0$ we get the identity

$$\mathbf{Var}[X] \stackrel{\text{def}}{=} \mathbf{E}\left[\|X - \mathbf{E}[X]\|^2\right] = \mathbf{E}\left[\|X\|^2\right] - \|\mathbf{E}[X]\|^2 . \tag{41}$$

- Since $f(x) = \|x\|^2$ is 2-convex and 2-smooth, the strong and reverse Jensen's inequalities applied to f say that

$$\mathbf{Var}[X] = \frac{\mu}{2}\mathbf{Var}[X] \leq \mathbf{E}\left[\|X\|^2\right] - \|\mathbf{E}[X]\|^2 \leq \frac{L}{2}\mathbf{Var}[X] = \mathbf{Var}[X] .$$

  - This can be seen as an alternative proof of (41).
  - Perhaps more interestingly, this shows that the **strong and reverse Jensen's inequalities are tight for $f(x) = \|x\|^2$.**

Basic Facts - III

Identity (41) immediately leads to the well known inequality

$$\mathbf{E}\left[\|X\|^2\right] \geq \|\mathbf{E}[X]\|^2 . \tag{42}$$

- **Markov's inequality.** For any nonnegative random variable X and any $t > 0$,

$$\mathrm{Prob}(X \geq t) \leq \frac{\mathbf{E}[X]}{t}. \tag{43}$$

Tower Property

Lemma 30 (Tower Property / Iterated Expectation)

For any random variables X and Y, we have

$$\mathbf{E}\left[\mathbf{E}\left[X \mid Y\right]\right] = \mathbf{E}\left[X\right]. \tag{44}$$

Proof.

We shall only prove this for discrete random variables; the proof is more technical in the continuous case.

$$\begin{aligned}
\mathbf{E}\left[X\right] = \sum_x x\mathrm{Prob}(X = x) &= \sum_x x \sum_y \mathrm{Prob}(X = x \;\&\; Y = y) \\
&= \sum_y \sum_x x\mathrm{Prob}(X = x \;\&\; Y = y) \\
&= \sum_y \sum_x \mathrm{Prob}(Y = y) x \frac{\mathrm{Prob}(X = x \;\&\; Y = y)}{\mathrm{Prob}(Y = y)} \\
&= \sum_y \mathrm{Prob}(Y = y) \underbrace{\sum_x x\mathrm{Prob}(X = x \mid Y = y)}_{\mathbf{E}[X \mid Y=y]} \\
&= \mathbf{E}\left[\mathbf{E}\left[X \mid Y\right]\right].
\end{aligned}$$

Tower Property of Expectations: Intuition via Example

Example 31

Consider discrete random variables $\mathbf{X}$ and $\mathbf{Y}$:

- $\mathbf{X}$ has 2 outcomes: $\mathbf{x}_1$ and $\mathbf{x}_2$
- $\mathbf{Y}$ has 3 outcomes: $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$

Their joint probability mass function is given in this table:

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	
$\mathbf{x}_1$	0.05	0.20	0.03	**0.28**
$\mathbf{x}_2$	0.25	0.30	0.17	**0.72**
	0.30	**0.50**	**0.20**	1

Obviously, $\mathbf{E}[\mathbf{X}] = 0.28\mathbf{x}_1 + 0.72\mathbf{x}_2$. But we can also write:

$$
\begin{aligned}
\mathbf{E}[\mathbf{X}] &= (0.05\mathbf{x}_1 + 0.25\mathbf{x}_2) + (0.20\mathbf{x}_1 + 0.30\mathbf{x}_2) + (0.03\mathbf{x}_1 + 0.17\mathbf{x}_2) \\
&= \underbrace{0.30}_{\text{Prob}(\mathbf{Y}=\mathbf{y}_1)} \underbrace{\left(\frac{0.05}{0.30}\mathbf{x}_1 + \frac{0.25}{0.30}\mathbf{x}_2\right)}_{\mathbf{E}[\mathbf{X} \mid \mathbf{Y}=\mathbf{y}_1]} + 0.50\left(\frac{0.20}{0.50}\mathbf{x}_1 + \frac{0.30}{0.50}\mathbf{x}_2\right) \\
&\quad + 0.20\left(\frac{0.03}{0.20}\mathbf{x}_1 + \frac{0.17}{0.20}\mathbf{x}_2\right) \\
&= \mathbf{E}[\mathbf{E}[\mathbf{X} \mid \mathbf{Y}]].
\end{aligned}
$$

From GD to SGD: Theory Viewpoint

SGD in Generic Form

Consider the general regularized problem (6), i.e.,

$$\min_{x \in \mathbb{R}^d} f(x) + R(x).$$

In much of the course, we will talk about specific instances of the **generic (proximal) SGD method** (12)+(13):

Algorithm 3 SGD (generic)

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Construct an **unbiased estimator** $g^k \overset{\text{def}}{=} g(x^k)$ of the gradient
4: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

Unbiasedness

Assumption 2 (Unbiasedness)

For all $k \geq 0$, $g^k = g(x^k)$ is an **unbiased estimator** *of the gradient $\nabla f(x^k)$. That is,*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k). \tag{45}$$

Remarks:

- Note that in the above assumption we only need to construct an unbiased stochastic gradient along the path of the iterates only.
- In practice, we often construct a random mapping $x \mapsto g(x)$ satisfying the property

$$\mathbf{E}\left[g(x)\right] = \nabla f(x) \qquad \text{for all} \qquad x \in \mathbb{R}^d$$

  and simply apply it to the iterates.
  - This is a somewhat stronger assumption that what we need for analysis of SGD, but it has the advantage of being independent of the iterates of the algorithm.

Analysis of Generic SGD

- We would like to establish an analogue of Lemma 14, and similarly as in the GD case use it to establish a convergence result.
- We want to follow the proof of Lemma 14 as closely as possible, and only do adaptations when and where necessary.

This is what we get:

Lemma 32

Consider Algorithm 3. If f is μ-convex and Assumption 2 (unbiasedness) is satisfied, then for all $\gamma \geq 0$ and $k \geq 0$ we have

$$\begin{aligned} \mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] \quad \leq \quad & (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) \\ & +\gamma^2 \underbrace{\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right]}_{G^k}. \end{aligned} \tag{46}$$

Proof of Lemma 32 - I

- As in the GD case, let $r^k \stackrel{\text{def}}{=} x^k - x^\star$. Performing the exact same steps as in the case of GD, but replacing the gradient $\nabla f(x^k)$ by the stochastic gradient g^k, we get

$$\begin{aligned}
\left\|r^{k+1}\right\|^2 &= \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - x^\star\right\|^2 \\
&\stackrel{(16)}{=} \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - \mathrm{prox}_{\gamma R}\left(x^\star - \gamma \nabla f(x^\star)\right)\right\|^2 \\
&\stackrel{(15)}{\leq} \left\|(x^k - \gamma g^k) - (x^\star - \gamma \nabla f(x^\star))\right\|^2 \\
&= \left\|x^k - x^\star - \gamma\left(g^k - \nabla f(x^\star)\right)\right\|^2 \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle + \gamma^2\left\|g^k - \nabla f(x^\star)\right\|^2,
\end{aligned}$$

where, as before, the inequality follows from nonexpansiveness of the proximal operator.

Proof of Lemma 32 - II

- Now comes a slight change in the argument. We compute expectation of both sides of the above inequality, conditional on x^k. This gives

$$\begin{aligned}
\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\leq \left\|r^k\right\|^2 - 2\gamma\mathbf{E}\left[\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle \mid x^k\right] + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \mathbf{E}\left[g^k \mid x^k\right] - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \\
&\overset{(45)}{=} \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right],
\end{aligned}$$

  where in the second step we used linearity of expectation and the fact that r^k is a constant vector conditioned on x^k, and in the last step we used the unbiasedness assumption.

- As in the GD case, we again use μ-convexity to bound

$$\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle \overset{(21)}{\geq} D_f(x^k, x^\star) + \frac{\mu}{2}\left\|x^k - x^\star\right\|^2,$$

  which leads to

$$\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] \leq (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right],$$

  which is what we wanted to prove.

Bounding The Third Term - I

To analyze SGD using Lemma 32, we will need to come up with some appropriate upper bound on

$$G^k \stackrel{\text{def}}{=} \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right].$$

- Recall that in GD, we used the bound

$$G^k = \left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2 \leq 2LD_f(x^k, x^\star),$$

  which happens to be the consequence of L-smoothness and convexity.

  - Stepsize γ was then chosen so that the last two terms in (46) can be neglected (i.e., upper bounded by 0).
  - This led to the recursion $\left\|r^{k+1}\right\|^2 \leq (1 - \gamma\mu)\left\|r^k\right\|^2$.
- Motivated by this, it looks tempting to use an analogous inequality in the SGD case as well. Specifically, we require that there exists $A \geq 0$ such that

$$G^k \leq 2AD_f(x^k, x^\star)$$

  holds for all $k \geq 0$.

Bounding The Third Term - II

- However, this inequality does not hold for many typical choices of the stochastic gradient g^k unless the problem $\min f + R$ has some very special properties.
- Fortunately, there is a simple trick that fixes the issue: we require the inequality to hold up to some additive error $C \geq 0$:

$$G^k \leq 2AD_f(x^k, x^\star) + C.$$

Key Assumption

Assumption 3 (AC)

There exist constants $A \geq 0$ *and* $C \geq 0$ *such that for all* $k \geq 0$,

$$G^k \stackrel{def}{=} \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \leq 2AD_f(x^k, x^\star) + C. \qquad (47)$$

Remarks:

- In several subsequent lectures, we will give examples of practical stochastic gradient estimators g^k satisfying the above "AC" assumption/inequality.
- In a somewhat more distant future, we will further generalize this inequality in several ways, which will enable us to also consider
  - "variance reduced" SGD methods,
  - nonconvex f.

Convergence of SGD

Convergence Theory for SGD

Theorem 33

Assume that f is μ-convex, g^k is unbiased (Assumption 2) and that the AC assumption (Assumption 3) is satisfied. Choose a stepsize satisfying

$$0 < \gamma \leq \frac{1}{A}. \tag{48}$$

Then the iterates $\{x^k\}_{k\geq 0}$ of SGD *(Algorithm 3) satisfy*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{\gamma C}{\mu}. \tag{49}$$

As we shall see, the above theorem **covers many useful variants of** SGD **methods**, and describes a **linear convergence rate up to a certain oscillation radius** controlled by the constant C.

Convergence Theory: Commentary I

Recovering GD convergence result. In the case of GD, $g(x) = \nabla f(x)$ is trivially unbiased, and provided that f is convex and L-smooth, the AC assumption is satisfied with $A = L$ and $C = 0$. Moreover, the sequence of iterates $\{x^k\}_{k\geq 0}$ is deterministic.

- So, if f is μ-convex, then by Theorem 33 for all $0 < \gamma \leq \frac{1}{L}$ we get

$$\left\|x^k - x^\star\right\|^2 \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2.$$

- Note that we recover Theorem 15 exactly.

Convergence of the first term. The first term

$$T_1 \stackrel{\text{def}}{=} (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2$$

appearing on the right hand side of (185) is smaller (=better) if

- the number of iterations k is larger (we have this under control),
- the stepsize γ is larger (we have this under control),
- the modulus of strong convexity μ is larger (typically, we do not have this under control),

Convergence Theory: Commentary II

- the starting point x^0 is closer to the optimal point $x^\star$ (we can control this in principle, but this is hard to do as we do not know what $x^\star$ is).

Typically, k (knowing when to stop) and γ (knowing what stepsize to choose) are the hyperparameters we directly control. Note that

- $k \to \infty \quad \Rightarrow \quad T_1 \to 0$
- More precisely, for any $\varepsilon > 0$,

$$k \geq \frac{1}{\gamma\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad T_1 \leq \varepsilon \left\| x^0 - x^\star \right\|^2 . \tag{50}$$

- Focusing on T_1 only and ignoring the effect of the stepsize on the second term (which does not matter in the lucky/favorable case when $C = 0$), the optimal choice of the stepsize is $\gamma = \frac{1}{A}$.

Convergence of the second term. The second term

$$T_2 \stackrel{\text{def}}{=} \frac{\gamma C}{\mu}$$

appearing on the right hand side of (185) is smaller (=better) when

Convergence Theory: Commentary III

- the stepsize γ is smaller
- the strong convexity modulus μ is larger,
- C is smaller
  - The favorable case $C = 0$ is special, as then the second term vanishes, and so do the associated convergence complications.
  - In the general $C > 0$ case, we can decrease the size of T_1 by decreasing the stepsize γ.

We have

$$\gamma \leq \frac{\mu\varepsilon \left\|x^0 - x^\star\right\|^2}{C} \quad \Rightarrow \quad T_2 \leq \varepsilon \left\|x^0 - x^\star\right\|^2 . \tag{51}$$

Driving $T_1 + T_2$ to zero in the $C > 0$ case.
By combining (50) and (51), we get

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq T_1 + T_2 \leq 2\varepsilon \left\|x^0 - x^\star\right\|^2 \tag{52}$$

as long as $\gamma = \min\left\{\frac{1}{A}, \frac{\mu\varepsilon\left\|x^0 - x^\star\right\|^2}{C}\right\}$ and $k \geq \frac{1}{\gamma\mu}\log\frac{1}{\varepsilon}$.

Convergence Theory: Commentary IV

- Note that
$$k = \mathcal{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right) ,$$
which is **much worse** than $\mathcal{O}(\log\frac{1}{\varepsilon})$ (the latter was sufficient in the favorable $C = 0$ case).
- Indeed, the $C > 0$ case **forces us to forgo fast (linear) convergence rate and settle for slow (sublinear) convergence rate.**
- Later on in the course, we will talk about algorithmic strategies (commonly referred to as **variance reduction**) which enable us to maintain the fast (linear) rate even in the $C > 0$ case. This will require more sophisticated algorithmic and convergence analysis tools.

Proof of Theorem 33 - I

By combining Lemma 32 with Assumption 3, we get

$$
\begin{aligned}
\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\overset{(46)}{\leq} (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) + \gamma^2 \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \\
&\overset{(3)}{\leq} (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) + 2\gamma^2 A D_f(x^k, x^\star) + \gamma^2 C \\
&= (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma(1-\gamma A) D_f(x^k, x^\star) + \gamma^2 C \\
&\leq (1-\gamma\mu)\left\|r^k\right\|^2 + \gamma^2 C,
\end{aligned}
$$

where in the last step we dropped the Bregman divergence term because $\gamma \leq \frac{1}{A}$ and because Bregman divergence of a convex function is nonnegative.

Proof of Theorem 33 - II

We now take expectation again on both sides, and use the tower property

$$\mathbf{E}\left[\left\|r^{k+1}\right\|^2\right] \overset{(44)}{=} \mathbf{E}\left[\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right]\right],$$

arriving at

$$\mathbf{E}\left[\left\|r^{k+1}\right\|^2\right] \quad \leq \quad (1-\gamma\mu)\mathbf{E}\left[\left\|r^k\right\|^2\right] + \gamma^2 C.$$

Unrolling the recurrence and noting that $\mathbf{E}\left[\left\|r^0\right\|^2\right] = \left\|r^0\right\|^2$ gives us

$$\begin{aligned}\mathbf{E}\left[\left\|r^k\right\|^2\right] &\leq (1-\gamma\mu)^k \left\|r^0\right\|^2 + \gamma^2 C \sum_{i=0}^{k-1}(1-\gamma\mu)^i \\ &\leq (1-\gamma\mu)^k \left\|r^0\right\|^2 + \frac{\gamma C}{\mu}.\end{aligned}$$

More on the AC Assumption

When is the AC Assumption Satisfied?

Theorem 33 depends on the AC assumption (Assumption 33). When is it satisfied?

- It will be useful (and notationally cleaner) to "decouple" the AC assumption from the iterates of SGD and cast it in a more abstract setting.
- In particular, let $g(x)$ be an unbiased estimator of $\nabla f(x)$. That is, $\mathbf{E}\left[g(x)\right] = \nabla f(x)$. We define

$$G(x,y) \stackrel{\text{def}}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right]. \tag{53}$$

- Using the above notation, the AC assumption can be written in the form

$$G(x^k, x^\star) \leq 2AD_f(x^k, x^\star) + C.$$

- Typically, providing an upper bound $G(x,y)$ for *any* x and y is equally complicated as providing a bound for $x = x^k$ and $y = y^\star$. So, in order to establish the AC assumption, we will typically seek to establish an inequality of the form

$$G(x,y) \leq 2AD_f(x,y) + C(y), \qquad x, y \in \mathbb{R}^d, \tag{54}$$

where $C(y)$ is allowed to depend on y, after which we specialize it to $x = x^k$ and $y = y^\star$.

Bounding $G(x, y)$: First Approach

Theorem 34

Let $g(x)$ be an unbiased estimator of $\nabla f(x)$ and assume that $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth. Then

$$G(x, y) \leq 2LD_f(x, y) + \mathbf{Var}\left[g(x)\right]. \tag{55}$$

If, moreover, there exist constants A' and $C'(y)$ such that

$$\mathbf{Var}\left[g(x)\right] \leq 2A'D_f(x, y) + C'(y), \tag{56}$$

then

$$G(x, y) \leq 2(L + A')D_f(x, y) + C'(y). \tag{57}$$

Proof of Theorem 34

By applying the variance decomposition identity (40) (with $X = g(x)$ and $c = \nabla f(y)$) and because $\nabla f(x) = \mathbf{E}\left[g(x)\right]$, we can write

$$\begin{aligned} G(x,y) &\overset{(53)}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \\ &\overset{(40)}{=} \mathbf{E}\left[\|g(x) - \mathbf{E}\left[g(x)\right]\|^2\right] + \|\mathbf{E}\left[g(x)\right] - \nabla f(y)\|^2 \\ &= \mathbf{Var}\left[g(x)\right] + \|\nabla f(x) - \nabla f(y)\|^2 \ . \end{aligned}$$

It remains to apply the inequality $\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x,y)$, which holds since f is assumed to be convex and L-smooth (see Proposition 17(iii)).

Example: Stochastic Gradient with Bounded Variance

Example 35

The most classic assumption on the stochastic gradient beyond unbiasedness is **boundedness of its variance:**

$$\mathbf{Var}\left[g(x)\right] \leq \sigma^2 \quad \text{for all} \quad x \in \mathbb{R}^d.$$

This arises in practice in the situation when a zero-mean bounded-variance noise is added to the true gradient:

$$g(x) \stackrel{\text{def}}{=} \nabla f(x) + \xi, \qquad \mathbf{E}\left[\xi\right] = 0, \qquad \mathbf{Var}\left[\xi\right] \leq \sigma^2.$$

Since $\mathbf{Var}\left[g(x)\right] = \mathbf{Var}\left[\xi\right]$, inequality (55) then implies that

$$G(x,y) \leq 2LD_f(x,y) + \sigma^2.$$

Theorem 33 now says that as long as $\gamma \leq \frac{1}{L}$, we have

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{\gamma\sigma^2}{\mu}.$$

Bounding $G(x, y)$: Second Approach

Theorem 36

Let $g(x)$ be an unbiased estimator of $\nabla f(x)$ and assume that

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y) + C''(y). \tag{58}$$

Then

$$G(x, y) \leq 2(2A'')D_f(x, y) + 2(\mathbf{Var}\left[g(y)\right] + C''(y)) \ . \tag{59}$$

Proof.

$$\begin{aligned}
G(x, y) &\overset{(53)}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \\
&= \mathbf{E}\left[\|g(x) - g(y) + g(y) - \nabla f(y)\|^2\right] \\
&\overset{(38)}{\leq} \mathbf{E}\left[2\,\|g(x) - g(y)\|^2 + 2\,\|g(y) - \nabla f(y)\|^2\right] \\
&\overset{(58)}{\leq} 2(2A'')D_f(x, y) + 2(\mathbf{Var}\left[g(y)\right] + C''(y)).
\end{aligned}$$

□

Bounding $G(x, y)$: Commentary

Theorem 34

- The AC inequality holds with

$$A = L + A', \quad C = C'(x^\star).$$

- **Hard part:** Establishing (56)

Theorem 36

- The AC inequality holds with

$$A = 2A'', \quad C = 2(\mathbf{Var}\,[g(x^\star)] + C''(x^\star)).$$

- **Hard part:** Establishing (58)

Exercises

Exercises I

Exercise 24 (Young's inequality)

Prove inequalities (37), (38) *and* (39).

Exercise 25 (Variance decomposition)

Provide two different proofs of the variance decomposition identity (40).

Exercise 26 (Chebyshev's inequality)

Use Markov's inequality to prove that for a random vector $X \in \mathbb{R}^d$ *and any* $t > 0$, *we have*

$$\mathrm{Prob}\left(\|X - \mathbf{E}\left[X\right]\| \geq t\right) \leq \frac{\mathbf{Var}\left[X\right]}{t^2},$$

Exercises II

Exercise 27 (Convergence in mean square)

We say that a sequence $\{X^k\}_{k\geq 0}$ *of random vectors* **converges in mean square** *to random vector* $X \in \mathbb{R}^d$ *if the second moments* $\mathbf{E}\left[\|X^k\|^2\right]$ *and* $\mathbf{E}\left[\|X\|^2\right]$ *exist, and if*

$$\lim_{k\to\infty} \mathbf{E}\left[\left\|X^k - X\right\|^2\right] = 0.$$

Under the assumptions of Theorem 33, show that the sequence of random vectors $\{x^k\}_{k\geq 0}$ *produced by* `SGD` *converges in mean square to* $x^\star$ *if the AC assumption holds with* $C = 0$.

Exercise 28 (Convergence in probability)

We say that a sequence $\{X^k\}_{k\geq 0}$ *of random vectors* **converges in probability** *to random vector* $X \in \mathbb{R}^d$ *if for all* $\varepsilon > 0$

$$\lim_{k\to\infty} \mathrm{Prob}\left(\left\|X^k - X\right\| > \varepsilon\right) = 0.$$

Under the assumptions of Theorem 33, show that the sequence of random vectors $\{x^k\}_{k\geq 0}$ *produced by* `SGD` *converges in probability to* $x^\star$ *if the AC assumption holds with* $C = 0$.

Exercises III

Exercise 29

Let $f = \frac{1}{n}\sum_{i=1}^{n} f_i$*, and assume that each* f_i *is convex and* L_i*-smooth. Define a gradient estimator by setting* $g(x) = \nabla f_i(x)$*, where the index* i *is chosen uniformly at random. Establish a bound of the form*

$$G(x, y) \leq 2AD_f(x, y) + C,$$

and specify what A *and* C *are. Use this to derive a convergence result for* `SGD` *using this gradient estimator.*

Exercise 30 (Code)

Code up the `SGD` *method described in the previous example on an optimization problem of your choice. Perform a sequence of numerical experiments testing various parts of the established theory.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 7: SGD with Simple Sampling

Based on:

[23] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P.R. **SGD: General Analysis and Improved Rates**, ICML 2019

[87] P. Zhao and T. Zhang, **Stochastic optimization with importance sampling for regularized loss minimization**, ICML 2015

Introduction

Recall we consider the regularized **finite-sum** problem

$$\min_{x \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} f_i(x)}_{f(x)} + R(x). \tag{60}$$

- We now consider **a non-uniform sampling variant of** `SGD` which uses the gradient estimator

$$g(x) = \frac{\nabla f_i(x)}{n p_i}, \tag{61}$$

  where i **is chosen with probability** $p_i > 0$**.**

- Note that the gradient estimator is **unbiased:**

$$\mathbf{E}\left[g(x)\right] = \sum_{i=1}^{n} p_i \frac{\nabla f_i(x)}{n p_i} = \sum_{i=1}^{n} \frac{\nabla f_i(x)}{n} = \nabla f(x). \tag{62}$$

SGD-NS: the Algorithm

For the record, here is the formal algorithm:

Algorithm 4 SGD-NS

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, probabilities $p_1, \ldots, p_n$ summing up to one
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample $i^k = i \in \{1, 2, \ldots, n\}$ with probability $p_i > 0$
4: $g^k = g(x^k) = \frac{\nabla f_i(x^k)}{np_i}$
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Two notable special cases:

- SGD-US (SGD with uniform sampling): $p_i = \frac{1}{n}$
- SGD-IS (SGD with importance sampling): $p_i = \frac{L_i}{\sum_j L_j}$

Convergence Theory

Applying Theorem 33: What's Missing?

In order to analyze SGD-NS, we will use Theorem 33, which requires the following assumptions (written here in the "decoupled" form):

- μ-convexity of f,
- unbiaseness: $\mathbf{E}\left[g(x)\right] = \nabla f(x)$ for all $x \in \mathbb{R}^d$,
- AC assumption: $\mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \leq 2AD_f(x,y) + C(y)$ for all $x, y \in \mathbb{R}^d$.

We will continue to assume that f is μ-convex, and we already checked that the stochastic gradient estimator (61) used in SGD-NS is unbiased; see (62). **So, in order to be able to apply Theorem 33, we only need to establish the AC inequality.**

However, the AC inequality will not hold without us imposing further assumptions on the functions $f_1, f_2, \ldots, f_n$. Here is an assumption that works.

Assumption 4

Each f_i is convex and L_i-smooth.

Computing the AC Constants

In view of Proposition 17 characterizing convexity and smoothness, the following inequalities hold for all $i \in [n]$ if Assumption 4 is satisfied:

$$0 \leq 2D_{f_i}(x, y) \leq L_i \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d, \tag{63}$$

$$\frac{1}{L_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2D_{f_i}(x, y), \quad \forall x, y \in \mathbb{R}^d. \tag{64}$$

Lemma 37

Let Assumption 4 hold (convexity and L_i-smoothness of f_i for all i). Let the gradient estimator g be defined as in (61). Then

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y), \quad \text{for all} \quad x, y \in \mathbb{R}^d,$$

where $A'' = \max_i \frac{L_i}{np_i}$. Moreover,

$$G(x, y) \leq 2(2A'')D_f(x, y) + 2\mathbf{Var}\left[g(y)\right] \quad \text{for all} \quad x, y \in \mathbb{R}^d. \tag{65}$$

Proof of Lemma 37

Fix any $x, y \in \mathbb{R}^d$. Then

$$\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(61)}{=} \sum_{i=1}^n p_i \left\| \frac{\nabla f_i(x)}{np_i} - \frac{\nabla f_i(y)}{np_i} \right\|^2 && (66)\\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{np_i}\|\nabla f_i(x)-\nabla f_i(y)\|^2 && (67)\\
&\overset{(64)}{\leq} \frac{1}{n}\sum_{i=1}^n \frac{1}{np_i} 2L_i D_{f_i}(x,y)\\
&\leq 2\left(\max_i \frac{L_i}{np_i}\right)\frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y)\\
&= 2\left(\max_i \frac{L_i}{np_i}\right) D_f(x,y). && (68)
\end{aligned}$$

Finally, (65) follows by applying Theorem 36.

Complexity of SGD-NS

The following result is a direct consequence of Theorem 33 and Lemma 37.

Theorem 38

Assume that

- *f is μ-convex,*
- *Each f_i is convex and L_i-smooth (Assumption 4)*

Then SGD-NS *with stepsize $0 < \gamma \leq \frac{1}{2A''}$, where $A'' = \max_i \frac{L_i}{np_i}$, satisfies*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\sigma_\star^2}{\mu}, \tag{69}$$

where

$$\begin{aligned}\sigma_\star^2 \overset{def}{=} \mathbf{Var}\left[g(x^\star)\right] &\overset{(41)}{=} \mathbf{E}\left[\left\|g(x^\star)\right\|^2\right] - \left\|\mathbf{E}\left[g(x^\star)\right]\right\|^2 \\ &= \left(\frac{1}{n^2}\sum_{i=1}^n \frac{\left\|\nabla f_i(x^\star)\right\|^2}{p_i}\right) - \left\|\nabla f(x^\star)\right\|^2. \end{aligned} \tag{70}$$

Observation 1: Interpolation Regime

Zero Variance

The regime in which the variance of the stochastic gradient evaluated at $x^\star$ vanishes (i.e., $\sigma_\star^2 = 0$), is particularly favorable, and it makes sense to spend a bit of time to think about it. Main reason why it is special:

Theorem 39 (Factor of 2 Improvement)

Assume that $\sigma_\star^2 = 0$ *and let* $A'' = \max_i \frac{L_i}{np_i}$. *Then*

(i) *The upper bound in* (65) *can be improved by a factor of 2:*

$$G(x, x^\star) \leq 2A'' D_f(x, x^\star). \tag{71}$$

(ii) *Theorem 38 has the following stronger form:* `SGD-NS` *allows for a twice larger stepsize* $0 < \gamma \leq \frac{1}{A''}$, *and satisfies*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2. \tag{72}$$

In particular, `SGD-NS` *in this case converges to the solution* $x^\star$ *(in mean square) rather than to a neighborhood thereof only.*

Interpolation Regime: Definition

Definition 40 (Interpolation regime)

Assume that there is no regularizer (i.e., $R \equiv 0$). We say that problem (60) is in the **interpolation regime** if

$$\nabla f_i(x^\star) = 0 \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

Interpretation:

- If f_i is convex and $f_i(x) = \ell(h_x(a_i), b_i)$, where ℓ is a loss function and (a_i, b_i) is a data example (see problem (2)), then $\nabla f_i(x^\star) = 0$ means that **model $h_{x^\star}$ incurs minimum (typically zero) loss on example (a_i, b_i).**
- If this holds for all i, this means that model $h_{x^\star}$ perfectly fits (i.e., **interpolates**) all training data.

Interpolation Regime: Example

Example 41 (Linear system)

Assume the function $h^\star : \mathcal{A} \to \mathcal{B}$ we are learning is given by

$$h^\star(a) = \langle x^\star, a \rangle,$$

where $\mathcal{A} = \mathbb{R}^d$ and $\mathcal{B} = \mathbb{R}$. Given an input a_i, the true label is given by $b_i = h^\star(a_i) = \langle x^\star, a_i \rangle$. Assume we use model

$$h_x(a) = \langle x, a \rangle.$$

If we use quadratic loss

$$\ell(b, b') = \frac{1}{2}(b - b')^2,$$

then

$$f_i(x) = \ell(h_x(a_i), b_i) = \frac{1}{2}(\langle a_i, x \rangle - b_i)^2.$$

Notice that $f_i(x^\star) = 0$, and $\nabla f_i(x^\star) = 0$ for all i. Hence, in this case we are in the interpolation regime.

Observation 2: Uniform Sampling

Uniform Sampling: SGD-US

For the record, here is the formal algorithm:

Algorithm 5 SGD-US

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
  Sample $i^k = i \in \{1, 2, \ldots, n\}$ with probability $\frac{1}{n}$
  $g^k = \nabla f_i(x^k)$ obtain a stochastic gradient
  $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Comparing GD and SGD-US in the Interpolation Regime

Let us compare the convergence of GD and SGD-US:

- GD
  - By Theorem 15, using stepsize $\gamma = \frac{1}{L}$, GD achieves iteration complexity

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \left\|x^k - x^\star\right\|^2 \leq \varepsilon \left\|x^0 - x^\star\right\|^2$$

  - Cost of 1 iteration: n gradient evaluations
  - **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$
- SGD-US
  - By Theorem 39, using the uniform probabilities $p_i = \frac{1}{n}$, stepsize $\gamma = \frac{1}{A''} = \frac{1}{\max_i L_i}$, SGD-US achieves iteration complexity

$$k \geq \frac{\max_i L_i}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2$$

  - Cost of 1 iteration: 1 gradient evaluation
  - **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{\max_i L_i}{\mu}\right)$

Key Question: How does $\max_i L_i$ compare to nL?

Example: Large L_i

Example 42 (One can have $L_i = nL$ for all i)

Choose $\theta > 0$ and let

$$f_i(x) = \tfrac{\theta}{2} x_i^2$$

for $i = 1, 2, \ldots, n$. We have $f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i f_i(x) = \frac{\theta/n}{2} \|x\|^2$.

- Since

$$\tfrac{\theta}{2} \|x\|^2 - f_i(x) = \tfrac{\theta}{2} \sum_{j \neq i} x_j^2$$

  is convex, f_i is L_i-smooth with $L_i = \theta$. Moreover, the constant L_i can't be improved further.

- Since

$$\frac{\theta/n}{2} \|x\|^2 - f(x) = 0$$

  is convex, f is L-smooth with $L = \frac{\theta}{n}$. Moreover, the constant L can't be improved further.

Note that

$$L_i = nL \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

Example: Small L_i

Example 43 (One can have $L_i = L$ for all i)

Choose $\theta > 0$ and let

$$f_i(x) = \frac{\theta}{2} \|x\|^2$$

for $i = 1, 2, \ldots, n$. We have $f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i f_i(x) = \frac{\theta}{2} \|x\|^2$. Clearly, all functions $f_1, \ldots, f_n$ and f are θ-smooth, and the smoothness constant can't be further improved. Note that

$$L_i = L \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

Generalizing the Examples

It turns out that the two examples above are extreme cases. In particular, we will show in Lemma 44 that (in a certain precise sense)

$$L \leq \max_i L_i \leq nL.$$

- Because of the inequality $\max_i L_i \leq nL$, we concluded that in the interpolation regime, `SGD-US` **is not worse than** `GD` **in terms of total complexity!**
- In situations when $\max_i L_i \ll nL$ (recall that there exists a scenario in which $\max_i L_i = L$), in the interpolation regime, `SGD-US` **can be up to** n **times faster than** `GD`**!**

Smoothness Constants of Functions f_i and f

Lemma 44

Let $f_1, f_2, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be differentiable functions and define $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$.

(i) *Assume f_i is L_i-smooth for all i. Then f is L-smooth with*

$$L \leq \frac{1}{n}\sum_{i=1}^n L_i.$$

(ii) *Assume that f is L-smooth and f_i is convex for all i. Then each f_i is L_i-smooth with*

$$L_i \leq nL.$$

(iii) *Assume that f is L'-smooth, f_i is convex and L_i'-smooth for all i. Then there exist constants $L \leq L'$ and $L_i' \leq L_i$ for all i such that f is L-smooth, f_i is L_i-smooth for all i, and*

$$L \leq \max_i L_i \leq nL.$$

Proof of Lemma 44

(i) Since f_i is L_i-smooth, the function

$$\phi_i(x) \stackrel{\text{def}}{=} \frac{L_i}{2}\|x\|^2 - f_i(x)$$

is convex. Hence, the average of these functions,

$$\frac{1}{n}\sum_{i=1}^n \phi_i(x) = \frac{\frac{1}{n}\sum_{i=1}^n L_i}{2}\|x\|^2 - f(x),$$

is also convex. This means that f is $\left(\frac{1}{n}\sum_i L_i\right)$-smooth.

(ii) Since f is L-smooth, the function

$$\phi(x) \stackrel{\text{def}}{=} \frac{L}{2}\|x\|^2 - f(x)$$

is convex, and hence the function

$$n\phi(x) = \frac{nL}{2}\|x\|^2 - \sum_{i=1}^n f_i(x)$$

is also convex. Now, since all f_i's are convex, also $\sum_{i:i\neq j} f_i$ is convex, and hence

$$\frac{nL}{2}\|x\|^2 - \sum_{i=1}^n f_i(x) + \sum_{i:i\neq j} f_i = \frac{nL}{2}\|x\|^2 - f_j(x)$$

is convex. This means that f_j is L_j-smooth with $L_j \leq nL$.

(iii) Left as an exercise.

Observation 3: Importance Sampling

The Many Faces of Importance Sampling I

The (open) **standard simplex** in $\mathbb{R}^n$ is the set

$$\Delta_n \stackrel{\text{def}}{=} \left\{ p = (p_1, \ldots, p_n) \ : \ \sum_{i=1}^n p_i = 1 \quad \text{and} \quad p_i > 0 \text{ for all } i \right\}.$$

1. Optimizing the linear convergence rate without worrying about the neighborhood:

$$\begin{aligned} \arg \min_{p \in \Delta_n, \gamma} \left\{ (1 - \gamma\mu)^k \ : \ 0 < \gamma \leq \frac{1}{2A''} \right\} &= \arg \min_{p \in \Delta_n} \left(1 - \frac{\mu}{2 \max_i \frac{L_i}{np_i}} \right)^k \\ &= \arg \min_{p \in \Delta_n} \max_i \frac{L_i}{np_i}. \end{aligned}$$

The solution of this problem leads to the optimal stepsize $\gamma = \frac{1}{2\bar{L}}$ and importance sampling probabilities

$$p_i = \frac{L_i}{\sum_j L_j} \quad \Rightarrow \quad A'' = \frac{\sum_i L_i}{n} \stackrel{\text{def}}{=} \bar{L}.$$

The Many Faces of Importance Sampling II

We will refer to SGD with these probabilities as SGD-IS.

Algorithm 6 SGD-IS

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample $i^k = i \in \{1, 2, \ldots, n\}$ with probability $p_i = \frac{L_i}{\sum_j L_j}$
4: $g^k = \frac{\nabla f_i(x^k)}{np_i}$ obtain a stochastic gradient
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

On the next slide we compare the convergence of GD and SGD-IS in the **interpolation regime**.

The Many Faces of Importance Sampling III

- GD
  - By Theorem 15, using stepsize $\gamma = \frac{1}{L}$, GD achieves iteration complexity

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \left\|x^k - x^\star\right\|^2 \leq \varepsilon \left\|x^0 - x^\star\right\|^2$$

  - Cost of 1 iteration: n gradient evaluations
  - **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$
- SGD-IS
  - By Theorem 39, using the importance sampling probabilities $p_i = \frac{L_i}{\sum_j L_j}$, stepsize $\gamma = \frac{1}{A''} = \frac{1}{\frac{1}{n}\sum_i L_i}$, SGD-IS achieves iteration complexity

$$k \geq \frac{\frac{1}{n}\sum_i L_i}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2$$

  - Cost of 1 iteration: 1 gradient evaluation
  - **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{\frac{1}{n}\sum_i L_i}{\mu}\right)$

The Many Faces of Importance Sampling IV

Key Question: How does $\frac{1}{n}\sum_i L_i$ compare to nL?

By Lemma 44,

$$L \leq \frac{1}{n}\sum_{i=1}^{n} L_i \leq nL.$$

- Because of the inequality $\frac{1}{n}\sum_i L_i \leq \max_i L_i \leq nL$, we conclude that in the interpolation regime, `SGD-IS` **is better than** `SGD-US`**, which in turn is better than** `GD` **(in terms of total complexity!)**
- In situations when $\frac{1}{n}\sum_i L_i \ll \max_i L_i$, in the interpolation regime, `SGD-IS` **can be much faster than** `SGD-US`**!**

The Many Faces of Importance Sampling V

2. Optimizing the neighborhood size without worrying about convergence speed:

$$\arg\min_{p\in\Delta_n,\gamma}\left\{\frac{2\gamma\sigma_\star^2}{\mu}\ :\ 0<\gamma\leq\frac{1}{2A''}\right\}.$$

Notice we can push the above objective arbitrarily close to 0 for any fixed p by choosing γ sufficiently small. So, there is no useful notion of importance sampling in this case.

3. Optimizing the neighborhood size while making sure the convergence speed is good enough. We choose $\gamma_{\min}>0$ (can't be too big) and require that $\gamma\geq\gamma_{\min}$. This ensures that

$$(1-\gamma\mu)^k\leq(1-\gamma_{\min}\mu)^k.$$

The Many Faces of Importance Sampling VI

Using this as a constraint, we can now pose the problem

$$\begin{aligned}
\arg\min_{p\in\Delta_n,\gamma} & \left\{ \frac{2\gamma\sigma_\star^2}{\mu} \;:\; 0<\gamma\leq\frac{1}{2A''},\ \gamma\geq\gamma_{\min} \right\} \\
= & \arg\min_{p\in\Delta_n} \left\{ \frac{2\gamma_{\min}\sigma_\star^2}{\mu} \;:\; \gamma_{\min}\leq\frac{1}{2A''} \right\} \\
= & \arg\min_{p\in\Delta_n} \left\{ \sigma_\star^2 \;:\; \gamma_{\min}\leq\min_i \frac{np_i}{2L_i} \right\}.
\end{aligned}$$

We'll stop here and not worry about computing the optimal probabilities in this setting.

4. Optimizing the variance $\sigma_\star^2$. Since

$$\sigma_\star^2 \overset{(70)}{=} \left(\frac{1}{n^2}\sum_{i=1}^n \frac{\|\nabla f_i(x^\star)\|^2}{p_i} \right) - \|\nabla f(x^\star)\|^2,$$

The Many Faces of Importance Sampling VII

the optimal probabilities are the solution of

$$\min_{p \in \Delta_n} \sum_{i=1}^{n} \frac{\|\nabla f_i(x^\star)\|^2}{p_i}. \tag{73}$$

It turns out the optimal solution of (73) is given by the importance sampling probabilities

$$p_i = \frac{\|\nabla f_i(x^\star)\|}{\sum_{j=1}^{n} \|\nabla f_j(x^\star)\|}. \tag{74}$$

Using these optimal probabilities, the variance is equal to

$$\sigma_\star^2 = \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^\star)\| \right)^2 - \|\nabla f(x^\star)\|^2. \tag{75}$$

Exercises

Exercises I

Exercise 31 (Decomposition of Bregman divergence)

Let $f = \frac{1}{n}\sum_{i=1}^{n} f_i$. *Prove that* $D_f(x,y) = \frac{1}{n}\sum\limits_{i=1}^{n} D_{f_i}(x,y)$.

Exercise 32

Write down the complete proof of Theorem 38.

Exercise 33

Prove that $\sigma_\star^2 = 0$ *if and only if* $\nabla f_i(x^\star) = np_i \nabla f(x^\star)$ *for all* i.

Exercise 34

Prove part (iii) of Lemma 44.

Exercise 35

Give an example of L_i*-smooth functions* f_i *for which* $\frac{1}{n}\sum_i L_i \ll \max_i L_i$.

Exercises II

Exercise 36 (Importance sampling 1)

Let $a_1, \ldots, a_n > 0$. Prove that the solution to the optimization problem

$$\min_{p \in \Delta_n} \max_i \frac{a_i}{p_i}$$

is given by $p_i = \frac{a_i}{\sum_j a_j}$ for all $i = 1, 2, \ldots, n$.

Exercise 37 (Importance sampling 2 - simple version)

Let $a, b > 0$. Prove that the solution to the optimization problem

$$\min_{0<p<1} \frac{a}{p} + \frac{b}{1-p}$$

is given by $p = \frac{\sqrt{a}}{\sqrt{a}+\sqrt{b}}$.

Exercises III

Exercise 38 (Importance sampling 2)

Let $a_1, \ldots, a_n > 0$. Prove that the solution to the optimization problem

$$\min_{p \in \Delta_n} \sum_{i=1}^{n} \frac{a_i}{p_i}$$

is given by $p_i = \frac{\sqrt{a_i}}{\sum_j \sqrt{a_j}}$ for all $i = 1, 2, \ldots, n$.

Exercise 39

Let $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$ and let $a = \frac{1}{n} \sum_{i=1}^{n} a_i$. Show that

$$\left(\frac{1}{n} \sum_{i=1}^{n} \|a_i\| \right)^2 \geq \|a\|^2 . \tag{76}$$

(i) *Deduce that the expression in (75) is nonnegative.*[8]

(ii) *Is it true that (76) holds as an equality if and only if there exists $c \in \mathbb{R}^d$ such that for each i we have $a_i = t_i c$ for some $t_i \geq 0$?*

[8] Note that we knew this was nonnegative already through other means: it is equal to variance of $g(x^\star)$ wrt the probabilities given in (74).

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 8: SGD with Minibatch Sampling

Based on:

[23] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P.R.
SGD: General Analysis and Improved Rates, ICML 2019

Introduction

We will describe **three families of SGD methods** which allow us to use multiple functions f_i, chosen at random, in the formation of the stochastic gradient. If we assume that each function f_i corresponds to a single training data point only, then this means that **a random collection of all training data points (a "minibatch") is processed in each iteration.**

- SGD-NS in each iteration samples & processes a single training data point only
- GD in each iteration samples & processes all the training data points
- In some sense, these methods we will **interpolate between SGD-NS (or a variant thereof) and GD.** The methods are also known as **minibatch** SGD methods.

To do define and analyze each method, we need to:

- Describe the **unbiased** stochastic gradient estimator g
- Compute **expected smoothness** constant $A'' \geq 0$ such that

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y)$$

Sampling without Replacement (Nice Sampling)

Sampling without Replacement: Nice Sampling

Fix a minibatch size $\tau \in \{1, 2, \ldots, n\}$ and let S be a random subset of $\{1, 2, \ldots, n\}$ of size τ, chosen uniformly at random.[9] Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{i \in S} \nabla f_i(x). \tag{77}$$

This estimator leads to the following SGD algorithm:

Algorithm 7 SGD-NICE

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, minibatch size $\tau \in \{1, 2, \ldots, n\}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample $S^k \subseteq \{1, \ldots, n\}$ uniformly from all subsets of cardinality τ
4: $g^k = \frac{1}{\tau} \sum_{i \in S^k} \nabla f_i(x^k)$ obtain a stochastic gradient
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

[9] That is, we choose a single subset from the $\binom{n}{\tau}$ subsets of $\{1, 2, \ldots, n\}$ of cardinality τ, each with probability $1/\binom{n}{\tau}$. Such a random set S is also known in the literature under the name τ-nice sampling [66].

SGD-NICE: Unbiasedness and Expected Smoothness

Lemma 45

The gradient estimator g defined in (77) is unbiased. If we further assume that $n \geq 2$, f_i is convex and L_i-smooth for all i, and f is L-smooth, then

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{n-\tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L. \tag{78}$$

Finally,

$$\mathbf{Var}\left[g(y)\right] = \frac{n-\tau}{\tau(n-1)} \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2 \right). \tag{79}$$

Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{n-\tau}{\tau(n-1)}, \qquad b(\tau) \stackrel{\text{def}}{=} \frac{n(\tau-1)}{\tau(n-1)}.$$

Notice that

- $a(\tau) + b(\tau) = 1$ for all $\tau \in \{0, 1, \ldots, n\}$
- a is decreasing, with $a(1) = 1$, $a(n) = 0$
- b is increasing, with $b(1) = 0$, $b(n) = 1$

Figure: Functions a and b plotted for $1 \leq \tau \leq n$ and $n = 10$.

Commentary - II

Summary table:

τ	$a(\tau)$	$b(\tau)$	A''	Algorithm
1	1	0	$\max_i L_i$	SGD-US
τ	$\frac{n-\tau}{\tau(n-1)}$	$\frac{n(\tau-1)}{\tau(n-1)}$	$\frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)}L$	SGD-NICE
n	0	1	L	GD

Key insights:

- For $\tau = 1$, we recover SGD-US, its maximum stepsize, and hence also its rate
- For $\tau = n$, we recover GD, its maximum stepsize, and hence also its rate
- SGD-NICE is therefore a minibatch SGD method that interpolates between SGD-US and GD as τ moves from 1 to n.

Proof of Lemma 45 - I

Unbiasedness. Let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbf{E}[\chi_i] = \mathrm{Prob}(i \in S) = \frac{\tau}{n}. \tag{80}$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned}
\mathbf{E}[g(x)] &\overset{(77)}{=} \mathbf{E}\left[\frac{1}{\tau}\sum_{i\in S}\nabla f_i(x)\right] = \mathbf{E}\left[\frac{1}{\tau}\sum_{i=1}^{n}\chi_i\nabla f_i(x)\right] \\
&= \frac{1}{\tau}\sum_{i=1}^{n}\mathbf{E}[\chi_i]\nabla f_i(x) \\
&\overset{(80)}{=} \frac{1}{\tau}\sum_{i=1}^{n}\mathrm{Prob}(i\in S)\nabla f_i(x) \\
&\overset{(80)}{=} \frac{1}{\tau}\sum_{i=1}^{n}\frac{\tau}{n}\nabla f_i(x) \\
&= \nabla f(x).
\end{aligned}$$

Proof of Lemma 45 - II

Expected smoothness (i.e., computing constant A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{81}$$

Let χ_{ij} be the random variable defined by

$$\chi_{ij} = \begin{cases} 1 & i \in S \text{ and } j \in S \\ 0 & \text{otherwise} \end{cases}.$$

Note that

$$\chi_{ij} = \chi_i \chi_j. \tag{82}$$

Further, it is easy to show that

$$\mathbf{E}\left[\chi_{ij}\right] = \mathrm{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}. \tag{83}$$

It is easy to check that for any vectors $b_1, \ldots, b_n \in \mathbb{R}^d$ we have the identity

$$\left\| \sum_{i=1}^n b_i \right\|^2 - \sum_{i=1}^n \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle. \tag{84}$$

Proof of Lemma 45 - III

We will use this identity twice in what follows:

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(77)}{=} \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S}\nabla f_i(x)-\frac{1}{\tau}\sum_{i\in S}\nabla f_i(y)\right\|^2\right] \\
&\overset{(81)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{i\in S}a_i\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{i=1}^{n}\chi_i a_i\right\|^2\right] \\
&\overset{(84)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\sum_{i=1}^{n}\|\chi_i a_i\|^2+\sum_{i\neq j}\langle\chi_i a_i,\chi_j a_j\rangle\right] \\
&\overset{(82)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\sum_{i=1}^{n}\|\chi_i a_i\|^2+\sum_{i\neq j}\chi_{ij}\langle a_i,a_j\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{i=1}^{n}\mathbf{E}\left[\chi_i\right]\|a_i\|^2+\sum_{i\neq j}\mathbf{E}\left[\chi_{ij}\right]\langle a_i,a_j\rangle. \qquad (85)
\end{aligned}
$$

Proof of Lemma 45 - IV

Using the formulas (80) and (83) and the decomposition identity (84) again, we can continue:

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(85)}{=} \frac{1}{\tau^2}\left(\frac{\tau}{n}\sum_{i=1}^n \|a_i\|^2 + \frac{\tau(\tau-1)}{n(n-1)}\sum_{i\neq j}\langle a_i, a_j\rangle\right) \\
&= \frac{1}{\tau n}\sum_{i=1}^n \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\sum_{i\neq j}\langle a_i, a_j\rangle \\
&\overset{(84)}{=} \frac{1}{\tau n}\sum_{i=1}^n \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\left(\left\|\sum_{i=1}^n a_i\right\|^2 - \sum_{i=1}^n \|a_i\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^n \|a_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2. \qquad (86)
\end{aligned}
$$

Since f_i is convex and L_i-smooth, we know that

$$\|a_i\|^2 \overset{(81)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x,y).$$

Since f is convex and L-smooth, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2 \overset{(81)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x,y).$$

Proof of Lemma 45 - V

It only remains to plug these bounds to (86), apply the bound $L_i \leq \max_i L_i$ and use the identity $D_f(x,y) = \frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y)$:

$$\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(86)}{\leq} \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^n 2L_i D_{f_i}(x,y) + \frac{n(\tau-1)}{\tau(n-1)}2LD_f(x,y) \\
&\leq 2\frac{n-\tau}{\tau(n-1)}\max_i L_i \frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y) + 2\frac{n(\tau-1)}{\tau(n-1)}LD_f(x,y) \\
&= 2\frac{n-\tau}{\tau(n-1)}\max_i L_i D_f(x,y) + 2\frac{n(\tau-1)}{\tau(n-1)}LD_f(x,y) \\
&= 2\left(\frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)}L\right)D_f(x,y).
\end{aligned}$$

Variance. Using the variance decomposition and then rewriting the second moment using the exact same steps that led to (86), but with $a_i' = \nabla f_i(y)$ instead of a_i, we

Proof of Lemma 45 - VI

arrive at the identity

$$
\begin{aligned}
\mathbf{Var}\left[g(y)\right] &= \mathbf{E}\left[\|g(y)\|^2\right] - \|\mathbf{E}\left[g(y)\right]\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n}\|a_i'\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^{n}a_i'\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i'\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\|a_i'\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i'\right\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2\right).
\end{aligned}
$$

Convergence of SGD-NICE - I

By combining

- Theorem 33 (main convergence theorem for SGD under the AC assumption),
- Theorem 36 (result that reduced checking the AC assumption to checking expected smoothness: $A = 2A''$ and $C = 2\mathbf{Var}\left[g(x^\star)\right]$) and
- Lemma 45 (computation of expected smoothness constant A''),

we arrive at the complexity result:

Consider the SGD-NICE method with minibatch size $\tau \in \{1, 2, \ldots, n\}$. Assume f_i is convex and L_i-smooth for all i, and that f is μ-convex and L-smooth. Choose any relative error tolerance $0 < \delta < 1$, stepsize $\gamma = \min\left\{\frac{1}{2A''}, \frac{\mu\delta\left\|x^0 - x^\star\right\|^2}{2\mathbf{Var}[g(x^\star)]}\right\}$, where

$$A'' = \frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)}L$$

Convergence of SGD-NICE - II

and

$$\mathbf{Var}\left[g(x^\star)\right] = \frac{n-\tau}{\tau(n-1)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^\star)\|^2 - \|\nabla f(x^\star)\|^2\right)}_{=\sigma_\star^2}. \tag{87}$$

Then

$$k \geq \max\left\{\frac{2A''}{\mu}, \frac{2\mathbf{Var}\left[g(x^\star)\right]}{\delta\mu^2\left\|x^0-x^\star\right\|^2}\right\}\log\left(\frac{1}{\delta}\right) \Rightarrow \mathbf{E}\left[\left\|x^k-x^\star\right\|^2\right] \leq 2\delta\left\|x^0-x^\star\right\|^2.$$

Moving from relative error δ to absolute error $\varepsilon = 2\delta\left\|x^0-x^\star\right\|^2$, the above translates to

$$k \geq \max\left\{\frac{2A''}{\mu}, \frac{4\mathbf{Var}\left[g(x^\star)\right]}{\varepsilon\mu^2}\right\}\log\left(\frac{2\left\|x^0-x^\star\right\|^2}{\varepsilon}\right) \Rightarrow \mathbf{E}\left[\left\|x^k-x^\star\right\|^2\right] \leq \varepsilon.$$

Optimal Minibatch Size - I

Given the above convergence result, we may wish to ask which minibatch size is optimal with respect to the **total complexity** of SGD-NICE, defined as the **product of the number of iterations and the cost of one iteration.** Since

- the number of iterations is $\max\left\{\frac{2A''}{\mu}, \frac{4\mathbf{Var}[g(x^\star)]}{\varepsilon\mu^2}\right\}$ (we ignore the logarithmic factor which does not depend on τ), and
- cost of each iteration is τ,

we arrive at the following **total complexity minimization** problem:

$$\min_{1\leq\tau\leq n} \mathcal{C}(\tau),$$

where

$$\mathcal{C}(\tau) \stackrel{\text{def}}{=} \frac{2}{\mu(n-1)} \max\left\{\underbrace{(n-\tau)\max_i L_i + n(\tau-1)L}_{\text{increasing linear}}, \underbrace{(n-\tau)\frac{2\sigma_\star^2}{\varepsilon\mu}}_{\text{decreasing linear}}\right\}.$$

Optimal Minibatch Size - II

Observations:

- If $\sigma_\star^2 = 0$ (e.g., in the interpolation regime), then the "decreasing" part is equal to zero, and hence $\mathcal{C}(\tau)$ is an increasing function. So, the optimal minibatch size is

$$\tau^\star = 1.$$

- Further, if $\sigma_\star^2$ is not too large, then the increasing linear function dominates the decerasing linear function on the interval $[1, n]$, and hence the optimal minibatch size is again

$$\tau^\star = 1.$$

This happens for

$$\sigma_\star^2 \leq \frac{\varepsilon \mu \max_i L_i}{2}.$$

Optimal Minibatch Size - III

- Otherwise, the increasing linear and the decreasing linear lines intersect on $(1, n)$, and the optimal minibatch size can be found by computing the intersection:

$$\tau^\star = \frac{n(\theta + L - \max_i L_i)}{\theta + nL - \max_i L_i},$$

where $\theta = \frac{2\sigma_\star^2}{\varepsilon\mu}$.

- Notice that

$$\sigma_\star^2 = \frac{\varepsilon\mu \max_i L_i}{2} \;\Rightarrow\; \theta = \max_i L_i \;\Rightarrow\; \tau^\star = \frac{n(\max_i L_i + L - \max_i L_i)}{\max_i L_i + nL - \max_i L_i} = 1.$$

- Notice that

$$\sigma_\star^2 \to \infty \quad \Rightarrow \quad \tau^\star \to n.$$

Key takeaway: If we care about the total complexity of SGD-NICE, the larger the quantity $\sigma_\star^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 - \|\nabla f(x^\star)\|^2$ is, the larger minibatch size should be chosen. As this quantity grows, the optimal minibatch size approaches n. On the other hand, in the "small" $\sigma_\star^2$ regime (and in particular if $\sigma_\star^2 = 0$), the optimal minibatch size is $\tau^\star = 1$.

Sampling without Replacement (Independent Sampling)

Sampling without Replacement: Independent Sampling

- Let $p_1, p_2, \ldots, p_n$ be probabilities ($0 < p_i \leq 1$ for all i).
- We do not require these probabilities to add up to 1! So, $\sum_i p_i$ can be anywhere in the interval $(0, n]$.

For each i define a random set as follows:

$$S_i \stackrel{\text{def}}{=} \begin{cases} \{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1 - p_i \end{cases}.$$

We now define a random subset $S \subseteq \{1, 2, \ldots, n\}$ by taking the union of these simple sets

$$S \stackrel{\text{def}}{=} \bigcup_{i=1}^{n} S_i. \tag{88}$$

Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \sum_{i \in S} \frac{1}{np_i} \nabla f_i(x). \tag{89}$$

SGD-IND: The Algorithm

Gradient estimator (89) leads to the following new variant of SGD:

Algorithm 8 SGD-IND

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, probabilities $0 < p_i \leq 1$ for $i = 1, 2, \ldots, n$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample set $S^k = \cup_{i=1}^n S_i^k$, where $S_i^k = \{i\}$ with probability p_i
4: $g^k = \sum_{i \in S^k} \frac{1}{np_i} \nabla f_i(x^k)$ obtain a stochastic gradient
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Minibatch Size

Note that S has a random size/cardinality. In such cases, we will use the word minibatch size to refer to the expected cardinality:

$$\tau = \mathbf{E}\left[|S|\right].$$

Note that

$$\mathbf{E}\left[|S|\right] = \mathbf{E}\left[\sum_{i=1}^{n} |S_i|\right] = \sum_{i=1}^{n} \mathbf{E}\left[|S_i|\right] = \sum_{i=1}^{n} 1p_i + 0(1-p_i) = \sum_{i=1}^{n} p_i. \quad (90)$$

- If we choose $\tau = n$, we must necessarily have $p_i = 1$ for all i, $S \equiv \{1, 2, \ldots, n\}$ and hence we recover the gradient estimator used by GD: $g(x) = \nabla f(x)$.
- If we choose $\tau = 1$, we **do not** recover the gradient estimator used by SGD-NS:

Algorithm	Gradient estimator $g(x)$	Minibatch size τ
SGD-NS	$\frac{1}{np_i}\nabla f_i(x)$	1 (deterministically)
SGD-IND	$\sum_{i\in S} \frac{1}{np_i}\nabla f_i(x)$	1 (in expectation)

SGD-IND: Unbiasedness and Expected Smoothness

Lemma 46

The gradient estimator g defined in (89) is unbiased. If we further assume that f_i is convex and L_i-smooth for all i, and f is L-smooth, then

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{\max_i \left(\frac{1}{p_i} - 1\right) L_i}{n} + L. \tag{91}$$

Commentary - I

Minibatch size = 1:

Algorithm / probabilities	uniform ($p_i = 1/n$)	nonuniform
SGD-NS	$\max_i L_i$	$\max_i \frac{L_i}{np_i}$
SGD-IND	$L + \frac{n-1}{n} \max_i L_i$	$\frac{\max_i \left(\frac{1}{p_i} - 1\right) L_i}{n} + L$

Table: The value of A'' for two variants of SGD under uniform and nonuniform probabilities.

Commentary - II

Note that in the $\tau = 1$ case with uniform probabilities, we have

$$\begin{aligned} A''_{\text{SGD-IND}} &= L + \frac{n-1}{n} \max_i L_i \\ &= \frac{1}{n} nL + \left(1 - \frac{1}{n}\right) \max_i L_i \\ &\geq \frac{1}{n} \max_i L_i + \left(1 - \frac{1}{n}\right) \max_i L_i \\ &= \max_i L_i \\ &= A''_{\text{SGD-US}} \end{aligned}$$

where the inequality follows from Lemma 44(ii). So, `SGD-IND` has a worse A'' constant than `SGD-US`, which means its rate is worse.

Minibatch size = n: In the $\tau = n$ case, we recover `GD`, its maximum stepsize, and hence also its rate since $A'' = L$

Commentary - III

More insights:

- The estimator (89) thus leads to a minibatch SGD method that interpolates between something "similar" to SGD-NS and GD as τ moves from 1 to n.
- Unlike in the case of τ-nice sampling, here we can make use of nonuniform probabilities, which means we can think about constructing **importance sampling for minibatches.**

Proof of Lemma 46 - I

Unbiasedness. As before, let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbf{E}[\chi_i] = \mathrm{Prob}(i \in S) = \mathrm{Prob}(i \in S_i) = p_i. \tag{92}$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned}
\mathbf{E}[g(x)] &\overset{(89)}{=} \mathbf{E}\left[\sum_{i \in S} \frac{1}{np_i} \nabla f_i(x)\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{n} \chi_i \frac{1}{np_i} \nabla f_i(x)\right] \\
&= \sum_{i=1}^{n} \mathbf{E}[\chi_i] \frac{1}{np_i} \nabla f_i(x) \\
&\overset{(92)}{=} \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(x) \\
&= \nabla f(x).
\end{aligned}$$

Proof of Lemma 46 - II

Expected smoothness (i.e., computing A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{93}$$

We will use this identity twice in what follows:

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &\overset{(89)}{=} \mathbf{E}\left[\left\|\sum_{i\in S}\frac{1}{np_i}\nabla f_i(x) - \sum_{i\in S}\frac{1}{np_i}\nabla f_i(y)\right\|^2\right] \\
&\overset{(93)}{=} \mathbf{E}\left[\left\|\sum_{i\in S}\frac{a_i}{np_i}\right\|^2\right] \\
&= \mathbf{E}\left[\left\|\sum_{i=1}^{n}\chi_i\frac{a_i}{np_i}\right\|^2\right] \\
&\overset{(84)}{=} \mathbf{E}\left[\sum_{i=1}^{n}\left\|\chi_i\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\left\langle \chi_i\frac{a_i}{np_i}, \chi_j\frac{a_j}{np_j}\right\rangle\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{n}\chi_i\left\|\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\chi_i\chi_j\left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j}\right\rangle\right] \\
&= \sum_{i=1}^{n}\mathbf{E}\left[\chi_i\right]\left\|\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\mathbf{E}\left[\chi_i\chi_j\right]\left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j}\right\rangle. \qquad (94)
\end{aligned}$$

Proof of Lemma 46 - III

Since $\mathbf{E}[\chi_i] = p_i$, and since by independence we have $\mathbf{E}[\chi_i \chi_j] = \mathbf{E}[\chi_i]\,\mathbf{E}[\chi_j] = p_i p_j$, we can further write

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p_i}\|a_i\|^2 + \sum_{i\neq j}\left\langle \frac{a_i}{n}, \frac{a_j}{n}\right\rangle \\
&\overset{(84)}{=} \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p_i}\|a_i\|^2 + \left(\left\|\sum_{i=1}^n \frac{a_i}{n}\right\|^2 - \sum_{i=1}^n \left\|\frac{a_i}{n}\right\|^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{p_i} - 1\right)\|a_i\|^2 + \left\|\sum_{i=1}^n \frac{a_i}{n}\right\|^2. \qquad (95)
\end{aligned}$$

Since f_i is convex and L_i-smooth, we know that

$$\|a_i\|^2 \overset{(93)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x, y).$$

Since f is convex and L-smooth, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2 \overset{(93)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x, y).$$

Plugging these estimates into (95), and using the identity

$$D_f(x, y) = \frac{1}{n}\sum_{i=1}^n D_{f_i}(x, y),$$

Proof of Lemma 46 - IV

we finally get

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\leq \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)2L_iD_{f_i}(x,y)+2LD_f(x,y)\\
&\leq 2\frac{\max_i\left(\frac{1}{p_i}-1\right)L_i}{n}\frac{1}{n}\sum_{i=1}^{n}D_{f_i}(x,y)+2LD_f(x,y)\\
&= 2\frac{\max_i\left(\frac{1}{p_i}-1\right)L_i}{n}D_f(x,y)+2LD_f(x,y)\\
&= 2\left(\frac{\max_i\left(\frac{1}{p_i}-1\right)L_i}{n}+L\right)D_f(x,y).
\end{aligned}
$$

Sampling with Replacement

Sampling with Replacement: Multisampling

Let $q_1, q_2, \ldots, q_n$ be probabilities summing up to 1 and let s be the random variable equal to i with probability q_i. Fix a minibatch size $\tau \in \{1, 2, \ldots\}$ and let $s_1, s_2, \ldots, s_\tau$ be independent copies of s. Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{n q_{s_t}} \nabla f_{s_t}(x). \tag{96}$$

Gradient estimator (96) leads to the following new variant of SGD:

Algorithm 9 SGD-MULTI

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, positive probabilities $q_1, \ldots, q_n$ summing up to 1, minibatch size $\tau \in \{1, 2, \ldots\}$

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: Sample τ i.i.d. random variables $s_1^k, \ldots, s_\tau^k$, where each is equal to i with probability q_i

4: $g^k = \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{n q_{s_t^k}} \nabla f_{s_t^k}(x^k)$ obtain a stochastic gradient

5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

SGD-MULTI: Unbiasedness and Expected Smoothness

Lemma 47

The gradient estimator g defined in (96) is unbiased. If we further assume that f_i is convex and L_i-smooth for all i, and f is L-smooth, then

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right) + \left(1 - \frac{1}{\tau}\right) L. \tag{97}$$

Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{1}{\tau}, \qquad b(\tau) \stackrel{\text{def}}{=} 1 - \frac{1}{\tau}.$$

Notice that

- $a(\tau) + b(\tau) = 1$ for all $\tau \in \{0, 1, \dots\}$
- a is decreasing, with $a(1) = 1$, $a(+\infty) = 0$
- b is increasing, with $b(1) = 0$, $b(+\infty) = 1$

Figure: Functions a and b plotted for $1 \leq \tau \leq +\infty$.

Commentary - II

Summary table:

τ	$a(\tau)$	$b(\tau)$	A''	Algorithm
1	1	0	$\max_i \frac{L_i}{nq_i}$	SGD-NS
τ	$\frac{1}{\tau}$	$1-\frac{1}{\tau}$	$\frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right)+\left(1-\frac{1}{\tau}\right)L$	SGD-MULTI
$+\infty$	0	1	L	GD

Key insights:

- For $\tau = 1$, we recover SGD-NS, its maximum stepsize, and hence also its rate
- For $\tau = +\infty$, we recover GD, its maximum stepsize, and hence also its rate
- The estimator (77) thus leads to a minibatch SGD method that interpolates between SGD-NS and GD as τ moves from 1 to $+\infty$.

Proof of Lemma 47 - I

Unbiasedness. Unbiasedness of $g(x)$ follows via direct computation:

$$\begin{aligned}
\mathbf{E}\left[g(x)\right] &= \mathbf{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\mathbf{E}\left[\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{i=1}^{n}q_i\frac{1}{nq_i}\nabla f_i(x) \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\nabla f(x) \\
&= \nabla f(x).
\end{aligned}$$

Expected smoothness (i.e., computing A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{98}$$

Proof of Lemma 47 - II

Then

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &= \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)-\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(y)\right\|^2\right] \\
&\overset{(98)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{t=1}^{\tau}\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbf{E}\left[\sum_{t=1}^{\tau}\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2+\sum_{t\neq u}^{\tau}\left\langle\frac{a_{s_t}}{nq_{s_t}},\frac{a_{s_u}}{nq_{s_u}}\right\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{t=1}^{\tau}\mathbf{E}\left[\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right]+\frac{1}{\tau^2}\sum_{t\neq u}^{\tau}\mathbf{E}\left[\left\langle\frac{a_{s_t}}{nq_{s_t}},\frac{a_{s_u}}{nq_{s_u}}\right\rangle\right]. \quad (99)
\end{aligned}
$$

Proof of Lemma 47 - III

We now separately bound the two terms in (99) by a multiple of the Bregman divergence $D_f(x, y)$. First, we estimate

$$\begin{aligned}
\mathbf{E}\left[\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right] &= \sum_{i=1}^{n} q_i \left\|\frac{a_i}{nq_i}\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{nq_i} \|a_i\|^2 \\
&\overset{(98)}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{nq_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{2L_i}{nq_i} D_{f_i}(x, y) \\
&\leq 2\left(\max_i \frac{L_i}{nq_i}\right) \frac{1}{n}\sum_{i=1}^{n} D_{f_i}(x, y) \\
&= 2\left(\max_i \frac{L_i}{nq_i}\right) D_f(x, y). \qquad (100)
\end{aligned}$$

The first inequality follows from Proposition 17 (iii) since f_i is convex and L_i-smooth, the second inequality follows by bounding $\frac{L_i}{nq_i} \leq \max_i \frac{L_i}{nq_i}$, and the last identity was the subject of an exercise.

Proof of Lemma 47 - IV

Next, since s_t and s_u are independent for $t \neq u$, we have

$$\begin{aligned}
\mathbf{E}\left[\left\langle \frac{a_{s_t}}{nq_{s_t}}, \frac{a_{s_u}}{nq_{s_u}} \right\rangle\right] &= \left\langle \mathbf{E}\left[\frac{a_{s_t}}{nq_{s_t}}\right], \mathbf{E}\left[\frac{a_{s_u}}{nq_{s_u}}\right] \right\rangle \\
&= \left\langle \sum_{i=1}^{n} q_i \frac{a_i}{nq_i}, \sum_{i=1}^{n} q_i \frac{a_i}{nq_i} \right\rangle \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} a_i \right\|^2 \\
&\overset{(98)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq 2LD_f(x,y). \qquad (101)
\end{aligned}$$

Finally, plugging (100) and (101) into (99), we obtain

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &\leq \frac{1}{\tau^2} \sum_{t=1}^{\tau} 2\left(\max_i \frac{L_i}{nq_i}\right) D_f(x,y) + \frac{1}{\tau^2} \sum_{t \neq u} 2LD_f(x,y) \\
&= 2\left(\frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right) + \left(1 - \frac{1}{\tau}\right) L\right) D_f(x,y),
\end{aligned}$$

as desired.

Exercises

Exercises I

Exercise 40

Show that if S is a τ-nice sampling (i.e., a random subset of $\{1, 2, \ldots, n\}$ of cardinality τ chosen uniformly at random), then

$$\mathrm{Prob}(i \in S) = \frac{\tau}{n}.$$

Exercise 41

Show that if S is a τ-nice sampling (i.e., a random subset of $\{1, 2, \ldots, n\}$ of cardinality τ chosen uniformly at random), then

$$\mathrm{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}.$$

Exercises II

Exercise 42

Prove identity (84). *That is, prove that for any vectors* $b_1, \ldots, b_n \in \mathbb{R}^d$,

$$\left\| \sum_{i=1}^{n} b_i \right\|^2 - \sum_{i=1}^{n} \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle.$$

Exercise 43 (Second moment of SGD-IND)

Compute $\mathbf{E}\left[\|g(y)\|^2\right]$ *for the gradient estimator* $g(x)$ *defined in* (89).

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 9: CGD: Compressed Gradient Descent

Based on:

[38] S. Khirirat, H.R. Feyzmahdavian and M. Johansson, **Distributed Learning with Compressed Gradients**, arXiv:1806.06573v2, 2018

P.R., Stochastic Gradient Descent Methods, Manuscript, 2019

Distributed Training I

We are now considering the **finite-sum** problem

$$\min_{x \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} f_i(x)}_{f(x)} + R(x) \tag{102}$$

and move onto **distributed optimization/training:**

- There are n **machines (worker nodes)** which can work in parallel.
- Machine i contains information about:
  - f_i (this function can in turn be very complex)
  - regularizer R
- There is **one machine (master node / parameter server)** which serves as a hub for communication.

Distributed Training II

We employ a distributed compute system because of one or more of the following reasons:

- f is described by **too much data to be stored on a single computer,** and hence the data (i.e., functions) need to be stored on different computers of a distributed system (e.g., cluster, supercomputer)
- a **single computer is not powerful enough** for the task at hand and we have access to multiple computers
- the training **data is naturally distributed** among multiple workers (e.g., photos stored on mobile devices - federated learning)

Distributed Gradient Descent and Communication Bottleneck

Distributed Gradient Descent

- We **want all machines to be fully utilized,** so in each iteration we want to access all n functions $f_1, \ldots, f_n$
- Ideally, we would want to implement `GD`:
  1. Each machine computes $\nabla f_i(x^k)$
  2. These gradients are sent to the **master,** which aggregates (i.e., averages) them and broadcasts the average

     $$\nabla f(x^k) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^k) \tag{103}$$

     back to the workers
  3. Each worker performs the prox gradient step to obtain x^{k+1}:

     $$x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma \nabla f(x^k)\right)$$

  4. The process is repeated

Communication Bottleneck - I

- In distributed systems, **communication of messages is often the key bottleneck**, and the key quantity of interest is the **total complexity / communication complexity:**

$$\#\text{Communication Rounds} \times \#\text{Bits Communicated in Each Round}$$

sufficient to find a solution of suitable quality/accuracy. There are two ways in which we can improve communication complexity:

  - **Decrease the # of communication rounds** while making sure the # of bits communicated in each round does not increase.
  - **Decrease the # of communicated bits in each communication round** while making sure the # communications does not increase.
- In distributed GD, we have two types of communication:
  - **Workers to master:** Communication of the (often very high dimensional) gradients $\nabla f_i(x^k) \in \mathbb{R}^d$ from workers $i = 1, 2, \ldots, n$ to the master
  - **Master to workers:** Communication of the (often very high dimensional) aggregated gradient $\nabla f(x^k) = \frac{1}{n}\sum_{i=1}^n f_i(x^k)$ from the master to the n workers

Communication Bottleneck - II

- A popular solution to the communication bottleneck is **communication/gradient compression:** instead of communicating message $x \in \mathbb{R}^d$ (typically a gradient or a gradient estimator), we communicate message $\mathcal{C}(x)$ where

$$\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$$

is a **randomized mapping** (with certain properties) which we call a **compression operator** (to be defined later).

Gradient Descent with Compressed Communication

- **Worker to Master Compression:** Each worker i sends to the master the **compressed gradient**

$$g_i^k \stackrel{\text{def}}{=} \mathcal{C}_i^k(\nabla f_i(x^k))$$

instead of the (uncompressed) gradient $\nabla f_i(x^k) \in \mathbb{R}^d$, where

$$\mathcal{C}_i^k : \mathbb{R}^d \to \mathbb{R}^d$$

is a randomized mapping which outputs **easy to communicate vectors.** We call such mappings **compression operators.**

- **Master to Worker Compression:** The master computes the average of the compressed gradients, and applies a compression operator

$$\mathcal{C}^k : \mathbb{R}^d \to \mathbb{R}^d$$

to this, the result of which is broadcasted to the workers:

$$g^k \stackrel{\text{def}}{=} \mathcal{C}^k \left(\frac{1}{n} \sum_{i=1}^n \mathcal{C}_i^k \left(\nabla f_i(x^k) \right) \right). \tag{104}$$

Trade-off Between Compression and Convergence

We want the compressed messages to be "easy to communicate". However, the more we compress, the less accurate will g^k be as an estimator of the gradient, which will adversely affects the algorithm.

Intuition behind compression vs convergence:

- **No compression:** recovers gradient decent (`GD`)
- **A bit of compression:** may lead to improvement in the overall complexity
- **A lot of compression:** will lead to a worse performance than no compression.
- **Too much compression:** will break the method altogether (divergence).

Compression Operators

Definition 48 (Compression Operator)

We say that a randomized map $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a compression operator ($\omega > 0$) if for all $x \in \mathbb{R}^d$ it satisfies

$$\mathbf{E}\left[\mathcal{C}(x)\right] = x, \qquad \mathbf{E}\left[\left\|\mathcal{C}(x) - x\right\|^2\right] \leq \omega \left\|x\right\|^2. \tag{105}$$

We write $\mathcal{C} \in \mathbb{B}^d(\omega)$.

Lemma 49 (Second Moment)

Compression operator $\mathcal{C} \in \mathbb{B}^d(\omega)$ for all $x \in \mathbb{R}^d$ satisfies

$$\mathbf{E}\left[\left\|\mathcal{C}(x)\right\|^2\right] \leq (\omega + 1)\left\|x\right\|^2. \tag{106}$$

Proof.

Using the variance decomposition identity, we get

$$\mathbf{E}\left[\left\|\mathcal{C}(x)\right\|^2\right] \overset{(41)}{=} \mathbf{E}\left[\left\|\mathcal{C}(x) - x\right\|^2\right] + \left\|x\right\|^2 \overset{(106)}{\leq} \omega\left\|x\right\|^2 + \left\|x\right\|^2 = (\omega + 1)\left\|x\right\|^2.$$

Quick-and-Dirty Examples of Compression Operators

- **Sketch:**

$$\mathcal{C}(x) = \boldsymbol{S}x,$$

where $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ is a random matrix

  - **Identity compressor:** $\mathcal{C}(x) = x$ for all x
  - **Lazy/Bernoulli compressor:** Choose $0 < p \leq 1$ and define

$$\mathcal{C}(x) = \begin{cases} \frac{x}{p} & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}.$$

  - **Random/Nice sparsification** (Rand-k):

$$\mathcal{C}(x) = \frac{d}{k} \sum_{i \in S} x_i e_i,$$

where S is a random subset of $\{1, 2, \ldots, d\}$ of cardinality k chosen uniformly at random, and e_i is the ith standard unit basis vector in $\mathbb{R}^d$

- **Quantization**
  - **Ternary quantization**
  - **Random dithering**
  - **Natural dithering** [30] (exponentially better than random dithering)
    - Natural compression [30] (rounding all entries of x to positive or negative powers of two)

Distributed Gradient Descent with Compressed Gradients

Algorithm 10 Distributed Compressed Gradient Descent (DCGD)

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, compression operators $\mathcal{C}_1 \in \mathbb{B}^d(\omega_1), \ldots, \mathcal{C}_n \in \mathbb{B}^d(\omega_n), \mathcal{C} \in \mathbb{B}^d(\omega_{\mathrm{M}})$
for $k = 0, 1, 2, \ldots$ **do**
  for all workers $i \in \{1, 2, \ldots, n\}$ **in parallel do**
    Compute local gradient $\nabla f_i(x^k)$
    Compress local gradient $g_i^k = \mathcal{C}_i^k(\nabla f_i(x^k))$
    Send g_i^k to master
  Master computes the aggregate $\hat{g}^k = \frac{1}{n}\sum_{i=1}^n g_i^k$
  Master broadcasts the compressed aggregate $g^k = \mathcal{C}^k(\hat{g}^k)$ to all workers
  for all workers $i \in \{1, 2, \ldots, n\}$ **in parallel do**
    Compute the next iterate $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Remarks:

- $\mathcal{C}_i^k$ is an independent copy of $\mathcal{C}_i$ generated at iteration k
- $\mathcal{C}^k$ is an independent copy of $\mathcal{C}$ generated at iteration k

Unbiasedness

DCGD uses the gradient estimator

$$g(x) \stackrel{\text{def}}{=} \mathcal{C}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i\left(\nabla f_i(x)\right)\right), \tag{107}$$

where $\mathcal{C} \in \mathbb{B}^d(\omega_{\mathrm{M}})$, $\mathcal{C}_1 \in \mathbb{B}^d(\omega_1), \ldots, \mathcal{C}_n \in \mathbb{B}^d(\omega_n)$ are independent.

Lemma 50

The gradient estimator (107) *is unbiased.*

Remark: Since $g(x)$ is an unbiased estimator of the gradient $\nabla f(x)$, DCGD is a variant of SGD.

Proof of Lemma 50

In view of the tower property of conditional expectation (44), unbiasedness of the compression operators (105) and linearity of expectation, we have

$$
\begin{aligned}
\mathbf{E}\left[g(x)\right] &= \mathbf{E}\left[\mathcal{C}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i\left(\nabla f_i(x)\right)\right)\right] \\
&\overset{(44)}{=} \mathbf{E}\left[\mathbf{E}\left[\mathcal{C}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i\left(\nabla f_i(x)\right)\right) \mid \mathcal{C}_1,\ldots,\mathcal{C}_n\right]\right] \\
&\overset{(105)}{=} \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i\left(\nabla f_i(x)\right)\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\mathcal{C}_i\left(\nabla f_i(x)\right)\right] \\
&\overset{(105)}{=} \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) \\
&= \nabla f(x).
\end{aligned}
$$

Analysis in the $n = 1$ Case

CGD

If the non-distributed case (i.e., if $n = 1$ and hence $f = f_1$) and when there is no compression applied from the Master to the Worker, **Distributed Compressed Gradient Descent** (DCGD) reduces to **Compressed Gradient Descent** (CGD):

Algorithm 11 CGD

1: **Parameters:** learning rate $\gamma > 0$, compression operator $\mathcal{C} \in \mathbb{B}^d(\omega)$, starting point $x^0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute gradient $\nabla f(x^k)$
4: Compress gradient $g^k = \mathcal{C}^k\left(\nabla f(x^k)\right)$
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

The AC Inequality for CGD

Besides unbiasedness, we need to establish that the AC inequality holds:

Lemma 51

Assume f is convex and L-smooth. Define the gradient estimator

$$g(x) \stackrel{def}{=} \mathcal{C}\left(\nabla f(x)\right), \tag{108}$$

where $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a compression operator satisfying $\mathcal{C} \in \mathbb{B}^d(\omega)$. Then

$$G(x,y) \leq 2\underbrace{(2\omega+1)L}_{A} D_f(x,y) + \underbrace{2\omega \left\|\nabla f(y)\right\|^2}_{C(y)}. \tag{109}$$

If $\nabla f(y) = 0$, then we have the sharper bound

$$G(x,y) \leq 2\underbrace{(\omega+1)L}_{A} D_f(x,y). \tag{110}$$

Proof of Lemma 51 - I

Fix $x, y \in \mathbb{R}^d$. Note that since $\mathcal{C} \in \mathbb{B}^d(\omega)$, we have

$$\mathbf{E}\left[\|g(x)\|^2\right] \overset{(108)}{=} \mathbf{E}\left[\|\mathcal{C}(\nabla f(x))\|^2\right] \leq (\omega+1)\|\nabla f(x)\|^2. \tag{111}$$

In the case $\nabla f(y) = 0$, we can now proceed as follows:

$$\begin{aligned} G(x,y) &\overset{\text{def}}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \\ &= \mathbf{E}\left[\|g(x)\|^2\right] \\ &\overset{(111)}{\leq} (\omega+1)\|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq 2(\omega+1)LD_f(x,y), \end{aligned}$$

where the last inequality holds due to convexity and L-smoothness of f.

Proof of Lemma 51 - II

In the case $\nabla f(y) \neq 0$, we use the variance decomposition identity to write

$$
\begin{aligned}
G(x,y) &\overset{\text{def}}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \\
&\overset{(40)}{=} \mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(y)\|^2 \\
&\overset{(108)}{=} \mathbf{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(y)\|^2 \\
&\overset{(105)}{\leq} \omega \|\nabla f(x)\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 \\
&= \omega \|\nabla f(x) - \nabla f(y) + \nabla f(y)\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 \\
&\overset{(38)}{\leq} 2\omega \|\nabla f(x) - \nabla f(y)\|^2 + 2\omega \|\nabla f(y)\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 \\
&= (2\omega + 1) \|\nabla f(x) - \nabla f(y)\|^2 + 2\omega \|\nabla f(y)\|^2 \\
&\leq 2(2\omega + 1) L D_f(x,y) + 2\omega \|\nabla f(y)\|^2,
\end{aligned}
$$

where in the last step we have used convexity and L-smoothness of f.

Convergence of CGD

As a direct corollary of Theorem 33, we get

Corollary 52

- *Assume f is L-smooth and μ-convex*
- *Let $\mathcal{C} \in \mathbb{B}^d(\omega)$*
- *Let $A = (2\omega + 1)L$. If $\nabla f(x^\star) = 0$ (which happens, for instance, when $R \equiv 0$), we can instead set $A = (\omega + 1)L$*

If the stepsize satisfy $0 < \gamma \leq \frac{1}{A}$, then the iterates of CGD *(Algorithm 11) satisfy*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\omega \left\|\nabla f(x^\star)\right\|^2}{\mu}.$$

Remarks:

- Notice that compression **induces** an oscillation radius unless we are in the favorable situation with $\nabla f(x^\star) = 0$ (which happens, for instance, when $R \equiv 0$).
- If there is no compression, oscillation radius vanishes.

Iteration Complexity

In the $\nabla f(x^\star) = 0$ case, the above result says that if we choose

$$\gamma = \frac{1}{(\omega+1)L},$$

then CGD satisfies

$$k \geq \frac{(\omega+1)L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2.$$

Recall that GD satisfies:

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \left\|x^k - x^\star\right\|^2 \leq \varepsilon \left\|x^0 - x^\star\right\|^2.$$

Comparison:

- We recover GD and its rate as a special case for $\omega = 0$.
- CGD needs $(\omega+1)\times$ more iterations to solve the problem than GD.

Analysis in the general $n \geq 1$ Case

The AC Inequality for DCGD

Lemma 53

Assume f_i is convex and L_i-smooth for all i, and let f be L-smooth. Let the gradient estimator g be defined as in (107), *where*

$$\mathcal{C}_1 \in \mathbb{B}^d(\omega_1), \quad \mathcal{C}_2 \in \mathbb{B}^d(\omega_2), \quad \ldots \quad , \quad \mathcal{C}_n \in \mathbb{B}^d(\omega_n); \quad \mathcal{C} \in \mathbb{B}^d(\omega_{\mathrm{M}})$$

are independent compression operators. Then

$$G(x,y) \leq 2AD_f(x,y) + C(y), \tag{112}$$

where

$$A = L + 2(\omega_{\mathrm{M}} + 1)\max_i \left(L_i \frac{\omega_i}{n}\right) + L\omega_{\mathrm{M}}$$

and

$$C(y) = 2(\omega_{\mathrm{M}} + 1)\frac{1}{n}\sum_{i=1}^{n} \frac{\omega_i}{n} \|\nabla f_i(y)\|^2 + 2\omega_{\mathrm{M}} \|\nabla f(y)\|^2 .$$

Commentary - I

In order to get insight into the theoretical properties of DCGD implied by Lemma 53, for simplicity assume that

- There is **no compression done by the master,** i.e., $\omega_{\mathrm{M}} = 0$.
  - In some distributed systems, broadcasting from master to the workers is much faster than communicating from workers to the master [47]. Hence, the special case when there is no compression done by the master (i.e., $\omega_{\mathrm{M}} = 0$) is of interest.
  - Virtually all papers on the topic of communication compression consider compression done by the workers only. Notable exceptions: [30, 81] and these notes.
- We are in the **interpolation regime**, i.e., $\nabla f_i(x^\star) = 0$ for all i.

Under these assumptions, $C(x^\star) = 0$, and hence there is no oscillation radius. Moreover, $A = L + 2\max_i \left(L_i \frac{\omega_i}{n}\right)$, and hence if f is in addition assumed to be μ-convex, using Theorem 33 with the stepsize $\gamma = \frac{1}{A}$ yields

$$k \geq \frac{L + 2\max_i \left(L_i \frac{\omega_i}{n}\right)}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2.$$

Commentary - II

- **Compression for free.** For fixed n choose

$$\omega_i = \frac{nL}{L_i} \qquad \text{for all} \qquad i.$$

Then

$$L + 2\max_i \left(L_i \frac{\omega_i}{n} \right) = 3L.$$

- So, at the cost of performing $3\times$ as many communication rounds as GD, we are able to compress messages and hence reduce the time of one communication round.
- The communication time in each communication round depends on the slowest communicating worker i, which is the one whose ω_i is the smallest:

$$i_{\text{slowest}} = \arg\min_i \omega_i = \arg\max_i L_i.$$

Commentary - III

- If there exist a compressor $\mathcal{C} \in \mathbb{B}^d\left(\frac{nL}{\max_i L_i}\right)$ which can save more than $3\times$ in communicated bits, we win, and it is better to compress than not to compress.
- For each $k \in \{1, 2, \ldots, d\}$ there exists a compressor $\mathcal{C} \in \mathbb{B}^d\left(\frac{d}{k} - 1\right)$ which can save a factor of $\frac{d}{k}$ in communicated bits (Rand-k sparsifier). Choose k such that

$$\omega_{i_{\text{slowest}}} = \frac{nL}{\max_i L_i} = \frac{d}{k} - 1,$$

that is,

$$k^\star = \frac{d \max_i L_i}{nL + \max_i L_i}.$$

Total complexity of Rand-$k^\star$ (ignoring logarithmic factors) is

$$\underbrace{\frac{3L}{\mu} k^\star}_{\text{total complexity of DCGD}} = \frac{3L}{\mu} \frac{d \max_i L_i}{nL + \max_i L_i} = \underbrace{\left(\frac{L}{\mu} d\right)}_{\text{total complexity of GD}} \times \frac{3 \max_i L_i}{nL + \max_i L_i}.$$

Commentary - IV

So, DCGD with Rand-k is strictly better than GD if

$$s \stackrel{\text{def}}{=} \frac{nL + \max_i L_i}{3 \max_i L_i} > 1.$$

This quantity defines the total complexity speedup. Hence, the larger s is, the more speedup in total complexity we gain by using random sparsifier than not compressing at all.

- **Infinite workers regime.** Fix a μ-convex and L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ and for each integer $n \geq 1$ consider a representation of f as

$$f = \frac{1}{n} \sum_{i=1}^{n} f_{in},$$

where f_{in} is convex and L_{in}-smooth. This allows us to consider the idealized $n \to +\infty$ regime. If we choose compressors $\mathcal{C}_{in} \in \mathbb{B}^d(\omega_{in})$ such that $\omega_{in} \leq \omega$ for all i and n and assume that[10]

$$\lim_{n \to +\infty} \frac{1}{n} \max\{L_{1n}, L_{2n}, \ldots, L_{nn}\} = 0, \tag{113}$$

Commentary - V

then

$$\lim_{n\to\infty} L + 2 \max_{i\in\{1,2,\ldots,n\}} \left\{ L_{in} \frac{\omega_{in}}{n} \right\} = L.$$

- Hence, we get the iteration complexity of GD:

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2.$$

- It turns out there is a compression operator (e.g., a sparsification operator) for which $\omega_{in} = d$ and which compresses d-dimensional vector into (effectively) a 1-dimensional scalar. This means DCGD obtains a $d\times$ speedup in total complexity when compared to GD.

[10] We know that $L_{in} \leq nL$ for all i and n. Clearly, this is not enough to guarantee that (113) holds. So, indeed, this is an extra assumption.

Proof of Lemma 53 - I

Consider the $\nabla f(y) \neq 0$ case; the other (easier) case is left as an exercise.
First, notice that

$$\begin{aligned} G(x,y) &= \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \\ &\overset{(40)}{=} \mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right] + 2LD_f(x,y). \end{aligned}$$

So, it only remains to find an appropriate bound for $\mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right]$ and combine it with the above inequality. Recall that

$$g(x) = \mathcal{C}\left(\hat{g}(x)\right), \quad \hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} g_i(x), \quad g_i(x) = \mathcal{C}_i(\nabla f_i(x)) \tag{114}$$

Proof of Lemma 53 - II

and that $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ and $\mathcal{C} \in \mathbb{B}^d(\omega_{\mathrm{M}})$ are independent. We get the estimate

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right] &= \mathbf{E}\left[\|\mathcal{C}(\hat{g}(x)) - \nabla f(x)\|^2\right] \\
&= \mathbf{E}\left[\|\mathcal{C}(\hat{g}(x)) - \hat{g}(x) + \hat{g}(x) - \nabla f(x)\|^2\right] \qquad (115) \\
&= \mathbf{E}\left[\mathbf{E}\left[\|\mathcal{C}(\hat{g}(x)) - \hat{g}(x) + \hat{g}(x) - \nabla f(x)\|^2 \mid \mathcal{C}_1, \ldots, \mathcal{C}_n\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\|\mathcal{C}(\hat{g}(x)) - \hat{g}(x)\|^2 \mid \mathcal{C}_1, \ldots, \mathcal{C}_n\right] + \|\hat{g}(x) - \nabla f(x)\|^2\right] \\
&\leq \mathbf{E}\left[\omega_{\mathrm{M}} \|\hat{g}(x)\|^2 + \|\hat{g}(x) - \nabla f(x)\|^2\right] \\
&= \omega_{\mathrm{M}} \mathbf{E}\left[\|\hat{g}(x)\|^2\right] + \mathbf{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right]. \qquad (116)
\end{aligned}$$

Proof of Lemma 53 - III

We now bound each of the above terms individually. Let $a_i \overset{\text{def}}{=} g_i(x) - \nabla f_i(x)$ and note that $\mathbf{E}[a_i] = 0$. The second term can be bounded as

$$
\begin{aligned}
\mathbf{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right] &\overset{(114)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(g_i(x) - \nabla f_i(x))\right\|^2\right] \\
&= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} a_i\right\|^2\right] \\
&\overset{(84)}{=} \frac{1}{n^2}\mathbf{E}\left[\sum_{i=1}^{n}\|a_i\|^2 + \sum_{i\neq j}\langle a_i, a_j\rangle\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\left[\|a_i\|^2\right] + \sum_{i\neq j}\mathbf{E}\left[\langle a_i, a_j\rangle\right] \\
&\overset{(105)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\left[\|a_i\|^2\right] + \sum_{i\neq j}\langle \underbrace{\mathbf{E}[a_i]}_{=0}, \underbrace{\mathbf{E}[a_j]}_{0}\rangle \\
&\overset{(105)}{\leq} \frac{1}{n^2}\sum_{i=1}^{n}\omega_i\|\nabla f_i(x)\|^2, \qquad (117)
\end{aligned}
$$

Proof of Lemma 53 - IV

and the first term can be bounded as

$$
\begin{aligned}
\mathbf{E}\left[\|\hat{g}(x)\|^2\right] &\overset{(40)}{=} \mathbf{E}\left[\|\hat{g}(x) - \mathbf{E}\left[\hat{g}(x)\right]\|^2\right] + \|\mathbf{E}\left[\hat{g}(x)\right]\|^2 \\
&= \mathbf{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right] + \|\nabla f(x)\|^2 \\
&\overset{(117)}{\leq} \left(\frac{1}{n^2}\sum_{i=1}^{n} \omega_i \|\nabla f_i(x)\|^2\right) + \|\nabla f(x) - \nabla f(y) + \nabla f(y)\|^2 \\
&\overset{(38)}{\leq} \left(\frac{1}{n^2}\sum_{i=1}^{n} \omega_i \|\nabla f_i(x)\|^2\right) + 2\|\nabla f(x) - \nabla f(y)\|^2 + 2\|\nabla f(y)\|^2 \\
&\leq \left(\frac{1}{n^2}\sum_{i=1}^{n} \omega_i \|\nabla f_i(x)\|^2\right) + 2LD_f(x,y) + 2\|\nabla f(y)\|^2 . \qquad (118)
\end{aligned}
$$

Next, notice that

$$
\begin{aligned}
\|\nabla f_i(x)\|^2 &= \|\nabla f_i(x) - \nabla f_i(y) + \nabla f_i(y)\|^2 \\
&\overset{(38)}{\leq} 2\|\nabla f_i(x) - \nabla f_i(y)\|^2 + 2\|\nabla f_i(y)\|^2 \\
&\leq 4L_i D_{f_i}(x,y) + 2\|\nabla f_i(y)\|^2 , \qquad (119)
\end{aligned}
$$

where the last inequality follows from convexity and L_i-smoothness of f_i.

Proof of Lemma 53 - V

Combining (116), (117), (118) and (119), we get

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-\nabla f(x)\|^2\right] &\overset{(116)}{\leq} \omega_{\mathrm{M}}\mathbf{E}\left[\|\hat{g}(x)\|^2\right]+\mathbf{E}\left[\|\hat{g}(x)-\nabla f(x)\|^2\right] \\
&\overset{(117)+(118)}{=} (\omega_{\mathrm{M}}+1)\left(\frac{1}{n^2}\sum_{i=1}^n \omega_i\,\|\nabla f_i(x)\|^2\right) \\
&\qquad +2L\omega_{\mathrm{M}}D_f(x,y)+2\omega_{\mathrm{M}}\,\|\nabla f(y)\|^2 \\
&\overset{(119)}{\leq} (\omega_{\mathrm{M}}+1)\left(\frac{1}{n^2}\sum_{i=1}^n 4L_i\omega_i D_{f_i}(x,y)+2\omega_i\,\|\nabla f_i(y)\|^2\right) \\
&\qquad +2L\omega_{\mathrm{M}}D_f(x,y)+2\omega_{\mathrm{M}}\,\|\nabla f(y)\|^2 \\
&\leq \frac{4(\omega_{\mathrm{M}}+1)\max_i(L_i\omega_i)}{n}\,\frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y) \\
&\qquad +2(\omega_{\mathrm{M}}+1)\frac{1}{n}\sum_{i=1}^n\frac{\omega_i}{n}\,\|\nabla f_i(y)\|^2+2L\omega_{\mathrm{M}}D_f(x,y)+2\omega_{\mathrm{M}}\,\|\nabla f(y)\|^2 \\
&= 2\left(2(\omega_{\mathrm{M}}+1)\max_i\left(L_i\frac{\omega_i}{n}\right)+L\omega_{\mathrm{M}}\right)D_f(x,y) \\
&\qquad +2(\omega_{\mathrm{M}}+1)\frac{1}{n}\sum_{i=1}^n\frac{\omega_i}{n}\,\|\nabla f_i(y)\|^2+2\omega_{\mathrm{M}}\,\|\nabla f(y)\|^2\,. \qquad (120)
\end{aligned}
$$

The AC Inequality for DCGD without Master-Workers Compression

Lemma 54

Under the same assumption as Lemma 53 and in the special case when there is no compression done by the master (i.e., $\mathcal{C}(x) = x$ for all x), and when $\omega_i = \omega$ for all i, we have

$$G(x,y) \leq 2\underbrace{\left(L + 2L_{\max}\frac{\omega}{n}\right)}_{A} D_f(x,y) + \underbrace{2\frac{\omega}{n}\sigma^2(y)}_{C(y)}, \tag{121}$$

where

$$\sigma^2(y) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(y)\|^2 .$$

In the special case when $\sigma^2(y) = 0$, we have the sharper bound

$$G(x,y) \leq 2\underbrace{\left(L + L_{\max}\frac{\omega}{n}\right)}_{A} D_f(x,y). \tag{122}$$

Exercises

Exercises I

Exercise 44 (Generalized Compressor)

Assume $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ *is a randomized mapping satisfying the following properties:*

(i) $\mathbf{E}\left[\mathcal{C}(x)\right] = x$ *for all* $x \in \mathbb{R}^d$

(ii) $\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2 + \delta$ *for all* $x \in \mathbb{R}^d$ *and some constants* $\omega, \delta \geq 0$

Establish a convergence result for `CGD` *in the single worker case* ($n = 1$) *using this compressor.*

Exercise 45 (Compressing a Random Vector)

If $\mathcal{C} \in \mathbb{B}^d(\omega)$ *and* z *is a random vector independent of* $\mathcal{C}$*, then*

(i) $\mathbf{E}\left[\|\mathcal{C}(z) - z\|^2\right] \leq \omega \mathbf{E}\left[\|z\|^2\right]$*, and*

(ii) $\mathbf{E}\left[\|\mathcal{C}(z)\|^2\right] \leq (\omega + 1)\mathbf{E}\left[\|z\|^2\right]$.

Exercise 46

Prove Lemma 54.

Exercises II

Exercise 47 (A Random Process)

Fix $x^\star \in \mathbb{R}^d$ and consider the following simple algorithm. Choose $x^0 \in \mathbb{R}^d$ and $p \in (0,1)$ and for $k \geq 0$ iterate

$$x^{k+1} = \begin{cases} x^k & \text{with probability} \quad 1-p \\ x^\star & \text{with probability} \quad p. \end{cases} \tag{123}$$

(a) *Find function $f : \mathbb{R}^d \to \mathbb{R}$ so that Algorithm (123) is a variant of* `SGD` *applied to the problem*

$$\min_{x \in \mathbb{R}^d} f(x).$$

Identify the source of randomness and the stepsize used.

(b) *Use a theorem covered in class to formulate a convergence result for this method. Show that all assumptions of the theorem are satisfied.*

(c) *Deduce a convergence result for the random process directly from the definition of the process, without resorting to any convergence theorem from class. Can you get a better rate than that obtained in part (b)?*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 10: Randomized Coordinate Descent (RCD)

Based on:

P.R., Stochastic Gradient Descent Methods, Manuscript, 2019

[70] M. Safaryan, F. Hanzely and P.R., **Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization**, ICLR Workshop: Distributed and Private Machine Learning, 2021

Problem I

We now consider the problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x),$$

and assume that d **is very large.**

In this high dimensional regime, **randomized coordinate descent** (RCD) methods are the state of the art. As before, we will assume that f is smooth and strongly convex.

Plan:

- **Convergence Analysis I:** We first develop a simple "coarse" theory of RCD by viewing the method as CGD with a suitable compression operator. This theory will give some insight, but not the state-of-the-art convergence rates.
- **Convergence Analysis II and III:** We then improve upon this under an additional assumption on f ($\boldsymbol{L}$-smoothness). Here we only consider the $R \equiv 0$ setting.

Problem II

Suboptimal analysis: Convergence analysis I and II of RCD provided in this lecture is suboptimal. However, it **follows our general framework** and still gives some insight. **Convergence analysis III of RCD gives the best known rates.** However, it differs from the standard analysis provided by, for example, Nesterov [52], R. and Takáč [66, 65], and Qu and R. [60, 59]. Convergence analysis III is most similar to Qu and R. [60, 59], but uses an **unbiased** gradient estimator, whereas Qu and R. use a biased estimator. Still, the convergence result is the same.

Sketches

Sketches: Linear Compression Operators

Definition 55 (Sketch)

A **sketch** is a **linear** compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$. It is given by the formula

$$\mathcal{C}(x) = \boldsymbol{C}x, \tag{124}$$

where $\boldsymbol{C} \in \mathbb{R}^{d \times d}$ is a random matrix drawn from some ensemble of matrices. We say that $\mathcal{C}$ is a **diagonal sketch** if $\boldsymbol{C}$ is diagonal.

Lemma 56

Let $\mathcal{C}(x) \equiv \boldsymbol{C}x$ be a sketch. If $\mathbf{E}[\boldsymbol{C}] = \boldsymbol{I}_d$ and $\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]$ exists, then $\mathcal{C} \in \mathbb{B}^d(\omega)$ for

$$\omega = \lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]\right) - 1. \tag{125}$$

Proof.

Unbiasedness of $\mathcal{C}$ follows directly from the assumption $\mathbf{E}[\boldsymbol{C}] = \boldsymbol{I}_d$. The variance bound now follows by direct computation:

$$\begin{aligned}
\mathbf{E}\left[\|\boldsymbol{C}x - x\|^2\right] &= \mathbf{E}\left[\|\boldsymbol{C}x\|^2\right] - \|x\|^2 = x^\top \mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right] x - \|x\|^2 \\
&\leq x^\top \lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]\right) \boldsymbol{I}_d x - \|x\|^2 = \left(\lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]\right) - 1\right) \|x\|^2 .
\end{aligned}$$

Diagonal Sketches

We will now describe a concrete way of obtaining a sketching operator $\mathcal{C} \in \mathbb{B}^d(\omega)$.

Let $c^1, c^2, \ldots, c^d$ be a collection of random variables and define

$$\boldsymbol{C} \overset{\text{def}}{=} \mathrm{Diag}\left(c^1, c^2, \ldots, c^d\right). \tag{126}$$

- If all random variables c^i have mean 1, then $\mathbf{E}[\boldsymbol{C}] = \boldsymbol{I}_d$, and $\mathcal{C}$ is unbiased.
- Moreover, $\boldsymbol{C}^\top \boldsymbol{C} = \mathrm{Diag}\left((c^1)^2, (c^2)^2, \ldots, (c^d)^2\right)$, from which it follows that

$$\omega \overset{(125)}{=} \lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]\right) - 1 = \max_i \mathbf{E}\left[(c^i)^2\right] - 1. \tag{127}$$

Random Sparsification I

We now define a general **random sparsification** compression operator as a diagonal sketch defined via a **proper sampling** of a subset of coordinates.

Definition 57 (Sampling)

A random subset S of $[d] = \{1, 2, \ldots, d\}$ is called a sampling (of coordinates).

Definition 58 (Proper sampling)

We say that a sampling S is proper if

$$p_i \stackrel{\text{def}}{=} \mathrm{Prob}(i \in S) > 0$$

for all $i = 1, 2, \ldots, d$.

Random Sparsification II

Definition 59 (Random Sparsification)

Let S be a proper sampling of the set of coordinates and define

$$c_S^i = \begin{cases} \frac{1}{p_i} & i \in S \\ 0 & i \notin S \end{cases}. \tag{128}$$

Then the **random sparsification operator associated with S** is the diagonal sketch given by matrix $\boldsymbol{C} = \boldsymbol{C}_S$ given by (126) where $c^i = c_S^i$ for all $i = 1, 2, \ldots, d$. That is,

$$\boldsymbol{C}_S \overset{\text{def}}{=} \operatorname{Diag}\left(c_S^1, c_S^2, \ldots, c_S^d\right). \tag{129}$$

Note that

$$\mathcal{C}_S(x) \overset{(124)}{=} \boldsymbol{C}_S x \overset{(129)}{=} \operatorname{Diag}\left(c_S^1, \ldots, c_S^d\right) x \quad \Rightarrow \quad \left(\mathcal{C}_S(x)\right)_i = \begin{cases} \frac{x_i}{p_i} & i \in S \\ 0 & i \notin S \end{cases}.$$

In view of (127), we have $\mathcal{C}_S \in \mathbb{B}^d(\omega)$ with $\omega = \max_i \frac{1}{p_i} - 1$.

RCD and its Convergence Analysis

RCD: Randomized Coordinate Descent

We now consider the following RCD method:

$$x_i^{k+1} = \begin{cases} x_i^k - \gamma \frac{\left(\nabla f(x^k)\right)_i}{p_i} & i \in S^k \\ x_i^k & i \notin S^k \end{cases}.$$

Another way of writing the same method:

Algorithm 12 RCD with Arbitrary Sampling

1: **Parameters:** learning rate $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$; proper sampling S of the coordinates
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute the gradient: $\nabla f(x^k)$
4: Sample fresh $S^k \subseteq \{1, 2, \ldots, d\}$
5: Sparsify the gradient: $g^k = \boldsymbol{C}_{S^k} \nabla f(x^k) = \mathcal{C}_{S^k}\left(\nabla f(x^k)\right)$
6: Take SGD step: $x^{k+1} = x^k - \gamma g^k$

Convergence Analysis I

RCD as a Direct Special Case of CGD

- We have seen that $\mathcal{C}_S \in \mathbb{B}^d(\omega)$ with $\omega = \max_i \frac{1}{p_i} - 1$.
- Since RCD is a variant of CGD, we can apply Corollary 52 (convergence of CGD).

Corollary 60

- *Assume f is L-smooth and μ-convex*
- **Let S be a proper sampling** *and consider the* **sparsification operator** $\mathcal{C}_S \in \mathbb{B}^d(\omega)$ *with* $\omega = \max_i \frac{1}{p_i} - 1$.
- *Let $A^I = (2\omega + 1)L$. If $\nabla f(x^\star) = 0$ (which happens, for instance, when $R \equiv 0$), we can instead set $A^I = (\omega + 1)L$*

If the stepsize satisfies $0 < \gamma \leq \frac{1}{A^I}$, then the iterates of RCD *(Algorithm 12) satisfy*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\omega \left\|\nabla f(x^\star)\right\|^2}{\mu}.$$

RCD vs GD I

For simplicity, consider the special case $R \equiv 0$. Then Corollary 60 implies

$$k \geq \left(\max_i \frac{1}{p_i}\right) \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2 \qquad (130)$$

- **Using this analysis, it seems that RCD** requires $\left(\max_i \frac{1}{p_i}\right) \times$ more iterations than GD.
  - If $p_i = 1$ for all i, RCD reduces to GD, and (130) reduces to the

    $$k = \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\varepsilon}\right)$$

    rate of GD.
  - If $p_i = \frac{1}{d}$ for all i, then in view of (130), RCD needs $d\times$ more iterations than GD to converge:

    $$k = \mathcal{O}\left(d \frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$

    In many cases, this version of RCD can be implemented so that each iteration is $d\times$ less expensive. In such situations, **this analysis predicts identical performance.**

RCD vs GD II

- However, this conclusion is the **weakness of this analysis and not the weakness of RCD.**

Correct behaviour: RCD should take between $1\times$ to $d\times$ as many iterations as GD. Each iteration of RCD is typically $d\times$ cheaper than one iteration of GD. So, RCD is faster than GD.

Convergence Analysis II

No Regularizer

In this analysis we assume that $R \equiv 0$. That is, we consider the unconstrained/unregularized problem

$$\min_{x \in \mathbb{R}^d} f(x).$$

$\boldsymbol{L}$-smoothness

Assumption 5 ($\boldsymbol{L}$-smoothness)

There exists a positive semidefinite matrix $\boldsymbol{L} \in \mathbb{R}^{d\times d}$ such that f is $\boldsymbol{L}$-smooth. That is,

$$f(x+h) \leq f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\underbrace{\langle \boldsymbol{L}h, h\rangle}_{\stackrel{def}{=}\|h\|_{\boldsymbol{L}}^2} \qquad \forall x, h \in \mathbb{R}^d. \tag{131}$$

Remarks:

- The $\boldsymbol{L}$-smoothness assumption can be equivalently written as
$$2D_f(x,y) \leq \|x-y\|_{\boldsymbol{L}}^2, \qquad \forall x, y \in \mathbb{R}^d.$$
- Typically, one uses $\boldsymbol{L} = L\boldsymbol{I}_d$ for some $L > 0$. In this case, $\boldsymbol{L}$-smoothness reduces to the standard notion of L-smoothness.
- Functions appearing in practice are typically $\boldsymbol{L}$-smooth and $\boldsymbol{L}$ can be found relatively easily. Clearly, in that case, f is also L-smooth with
$$L = \lambda_{\max}(\boldsymbol{L}).$$
However, it is useful to keep the valuable information encoded in $\boldsymbol{L}$ and not replace it by $L\boldsymbol{I}_d$.

Key Lemma

Lemma 61

Assume f is $\boldsymbol{L}$-smooth, let $x^\star$ be a minimizer of f and let $\boldsymbol{W} = \mathrm{Diag}\,(w_1, \ldots, w_d)$, where w_i are positive entries. If we define

$$L_{\boldsymbol{W}} \overset{def}{=} \lambda_{\max}\left(\boldsymbol{W}^{1/2}\boldsymbol{L}\boldsymbol{W}^{1/2}\right), \tag{132}$$

then

$$\|\nabla f(x)\|_{\boldsymbol{W}}^2 \leq 2L_{\boldsymbol{W}}\left(f(x) - f(x^\star)\right), \qquad \forall x \in \mathbb{R}^d. \tag{133}$$

Special cases:

- Choose $\boldsymbol{W} = \boldsymbol{I}_d$. Then $L_{\boldsymbol{W}} = \lambda_{\max}(\boldsymbol{L}) = L$ and (133) reduces to the standard inequality (18) (for $y = x^\star$ and $R \equiv 0$).

Proof of Lemma 61

Fix $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$. Since $x^\star$ is a minimizer of f and since f is $\boldsymbol{L}$-smooth, we have

$$\begin{aligned} f(x^\star) &\leq f(x - \alpha \boldsymbol{W} \nabla f(x)) \\ &\overset{(131)}{\leq} f(x) - \alpha \langle \nabla f(x), \boldsymbol{W} \nabla f(x) \rangle + \frac{\alpha^2}{2} \|\boldsymbol{W} \nabla f(x)\|_{\boldsymbol{L}}^2 \\ &= f(x) - \alpha \langle \nabla f(x), \boldsymbol{W} \nabla f(x) \rangle + \frac{\alpha^2}{2} \langle \nabla f(x), \boldsymbol{W} \boldsymbol{L} \boldsymbol{W} \nabla f(x) \rangle . \end{aligned} \tag{134}$$

In view of (132), we have $\boldsymbol{W}^{1/2} \boldsymbol{L} \boldsymbol{W}^{1/2} \preceq L_{\boldsymbol{W}} \boldsymbol{I}_d$, which implies that $\boldsymbol{W} \boldsymbol{L} \boldsymbol{W} \preceq L_{\boldsymbol{W}} \boldsymbol{W}$. Plugging this estimate into (134), we get

$$\begin{aligned} f(x^\star) &\leq f(x) - \alpha \|\nabla f(x)\|_{\boldsymbol{W}}^2 + \frac{\alpha^2 L_{\boldsymbol{W}}}{2} \|\nabla f(x)\|_{\boldsymbol{W}}^2 \\ &= f(x) - \left(\alpha - \frac{\alpha^2 L_{\boldsymbol{W}}}{2} \right) \|\nabla f(x)\|_{\boldsymbol{W}}^2 . \end{aligned} \tag{135}$$

Since the above inequality holds for any $\alpha \in \mathbb{R}$, it also holds for $\alpha^\star = \frac{1}{L_{\boldsymbol{W}}}$ (this α maximizes the expression $\alpha - \frac{\alpha^2 L_{\boldsymbol{W}}}{2}$). It only remains to plug $\alpha^\star$ into (135) and rearrange the inequality.

AC Inequality

Since we consider the $R \equiv 0$ case, we have $\nabla f(x^\star) = 0$, and thus:

- $G(x, x^\star) = \mathbf{E}\left[\|g(x) - \nabla f(x^\star)\|^2\right] = \mathbf{E}\left[\|g(x)\|^2\right]$
- $D_f(x, x^\star) = f(x) - f(x^\star)$

Lemma 62

The gradient estimator of Algorithm 12 (RCD) *satisfies* $\mathbf{E}\left[\|g(x)\|^2\right] = \|\nabla f(x)\|^2_{\boldsymbol{P}^{-1}}$, *where*

$$\boldsymbol{P}^{-1} \stackrel{def}{=} \mathrm{Diag}\left(\frac{1}{p_1}, \ldots, \frac{1}{p_d}\right). \tag{136}$$

Furthermore, if f is $\boldsymbol{L}$-smooth, then $\|\nabla f(x)\|^2_{\boldsymbol{P}^{-1}} \leq 2L_{\boldsymbol{P}^{-1}} D_f(x, x^\star)$, *where*

$$L_{\boldsymbol{P}^{-1}} = \lambda_{\max}\left(\boldsymbol{P}^{-1/2}\boldsymbol{L}\boldsymbol{P}^{-1/2}\right). \tag{137}$$

So, the AC assumption holds:

$$G(x, x^\star) \stackrel{def}{=} \mathbf{E}\left[\|g(x) - \nabla f(x^\star)\|^2\right] \leq 2L_{\boldsymbol{P}^{-1}} D_f(x, x^\star).$$

Proof of Lemma 62

Recall that

$$g(x) = \boldsymbol{C}\nabla f(x),$$

where $\boldsymbol{C} = \mathrm{Diag}\left(c_S^1, \ldots, c_S^d\right)$. The first claim follows via:

$$\begin{aligned}
\mathbf{E}\left[\|g(x)\|^2\right] &= \mathbf{E}\left[\|\boldsymbol{C}\nabla f(x)\|^2\right] \\
&= \mathbf{E}\left[(\nabla f(x))^\top \boldsymbol{C}^2 \nabla f(x)\right] \\
&= (\nabla f(x))^\top \mathbf{E}\left[\boldsymbol{C}^2\right] \nabla f(x) \\
&= (\nabla f(x))^\top \mathrm{Diag}\left(\mathbf{E}\left[\left(c_S^1\right)^2\right], \ldots, \mathbf{E}\left[\left(c_S^d\right)^2\right]\right) \nabla f(x) \\
&\overset{(128)}{=} (\nabla f(x))^\top \mathrm{Diag}\left(\frac{1}{p_1}, \ldots, \frac{1}{p_d}\right) \nabla f(x) \\
&\overset{(136)}{=} \|\nabla f(x)\|^2_{\boldsymbol{P}^{-1}}.
\end{aligned}$$

The second claim follows by applying Lemma 61.

Convergence Theorem

Having established Lemma 62, as a direct corollary of Theorem 33, we get

Corollary 63

- *Assume f is $\boldsymbol{L}$-smooth and μ-convex*
- **Let S be a proper sampling** *and consider the* **sparsification operator** $\mathcal{C}_S$
- *Let the stepsize satisfy* $0 < \gamma \leq \frac{1}{A^{II}}$, *where*

$$A^{II} \overset{def}{=} L_{\boldsymbol{P}^{-1}} = \lambda_{\max}\left(\boldsymbol{P}^{-1/2}\boldsymbol{L}\boldsymbol{P}^{-1/2}\right).$$

Then the iterates of RCD *(Algorithm 12) satisfy*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2.$$

In particular, if we use the largest stepsize $\gamma = \frac{1}{A^{II}}$, then the above says:

$$k \geq \frac{L_{\boldsymbol{P}^{-1}}}{\mu} \log \frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon \left\|x^0 - x^\star\right\|^2.$$

Estimating $L_{\boldsymbol{P}^{-1}}$

We have the following estimates:

- If f is L-smooth (i.e., $\boldsymbol{L}$-smooth with $\boldsymbol{L} = L\boldsymbol{I}_d$), then

$$L_{\boldsymbol{P}^{-1}} \overset{(137)}{=} \lambda_{\max}\left(\boldsymbol{P}^{-1/2} L \boldsymbol{I}_d \boldsymbol{P}^{-1/2}\right) = L\lambda_{\max}\left(\boldsymbol{P}^{-1}\right) \overset{(136)}{=} \max_i \frac{L}{p_i}. \tag{138}$$

- If f is $\boldsymbol{L}$-smooth with $\boldsymbol{L} = \mathrm{Diag}\left(L_1, \ldots, L_d\right)$, then

$$L_{\boldsymbol{P}^{-1}} \overset{(137)}{=} \lambda_{\max}\left(\boldsymbol{P}^{-1/2} \mathrm{Diag}\left(L_1, \ldots, L_d\right) \boldsymbol{P}^{-1/2}\right) \overset{(136)}{=} \max_i \frac{L_i}{p_i}. \tag{139}$$

Case Study: Diagonal $\boldsymbol{L}$ and Single Element Sampling

Consider the following setup:

- $\boldsymbol{L}$ is **diagonal:** $\boldsymbol{L} = \mathrm{Diag}\left(L_1, \ldots, L_d\right)$.
- We use **single element sampling:** $|S| = 1$ with probability 1

Two bounds:

- Applying bound (138), we have $L = \max_i L_i$ and optimizing over the probabilities, we get the importance sampling $p_i = \frac{1}{d}$. This gives the rate
$$\mathcal{O}\left(\frac{d \max_i L_i}{\mu} \log \frac{1}{\varepsilon}\right).$$
- Applying bound (139) and optimizing over the probabilities, we get the importance sampling $p_i = \frac{L_i}{\sum_j L_j}$. This gives the rate
$$\mathcal{O}\left(\frac{\sum_i L_i}{\mu} \log \frac{1}{\varepsilon}\right).$$

Comparison: Utilizing the information contained in the matrix $\boldsymbol{L}$ can lead to substantial speedup. The second approach, where we take advantage of the heterogeneous diagonal structure of $\boldsymbol{L}$, leads to a speedup by the factor $1 \leq \frac{d \max_i L_i}{\sum_i L_i} \leq d$.

Convergence Analysis III

No Regularizer

As in Convergence Analysis II, in this analysis we also assume that $R \equiv 0$. That is, we consider the unconstrained/unregularized problem

$$\min_{x \in \mathbb{R}^d} f(x).$$

Main Theorem

Theorem 64

- *Assume f is $\boldsymbol{L}$-smooth and μ-convex*
- **Let S be a proper sampling** *and consider the* **sparsification operator** $\mathcal{C}_S$ *given by*
$$\mathcal{C}_S(x) = \boldsymbol{C}_S x.$$
- *Assume that the sparsification operator $\mathcal{C}_S$ is unbiased, i.e., assume that*
$$\mathbf{E}\left[\boldsymbol{C}_S\right] = \boldsymbol{I}_d.$$
- *Let the stepsize satisfy $0 < \gamma \leq \frac{1}{A^{III}}$, where*
$$A^{III} \overset{def}{=} \lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right]\right).$$

Then the iterates of RCD *(Algorithm 12) satisfy*

$$\mathbf{E}\left[f(x^k) - f(x^\star)\right] \quad \leq \quad (1-\gamma\mu)^k \left(f(x^0) - f(x^\star)\right).$$

Proof of Theorem 64 - I

Step 1: Use L-smoothness

Since f is L-smooth, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k - \gamma C_S \nabla f(x^k)) \\ &\overset{(131)}{\leq} f(x^k) - \gamma \left\langle \nabla f(x^k), C_S \nabla f(x^k) \right\rangle + \frac{\gamma^2}{2} \left\| C_S \nabla f(x^k) \right\|_L^2 \\ &= f(x^k) - \gamma \left\langle \nabla f(x^k), C_S \nabla f(x^k) \right\rangle + \frac{\gamma^2}{2} \left\langle \nabla f(x^k), C_S L C_S \nabla f(x^k) \right\rangle \end{aligned} \tag{140}$$

Step 2: Take conditional expectation

Taking expectation conditional on x^k on both sides, and using the fact that $\mathbf{E}[C_S] = I_d$, we get

$$\begin{aligned} \mathbf{E}\left[f(x^{k+1}) \mid x^k \right] &\overset{(140)}{\leq} \mathbf{E}\left[f(x^k) - \gamma \left\langle \nabla f(x^k), C_S \nabla f(x^k) \right\rangle + \frac{\gamma^2}{2} \left\langle \nabla f(x), C_S L C_S \nabla f(x^k) \right\rangle \mid x^k \right] \\ &\leq f(x^k) - \gamma \left\langle \nabla f(x^k), \underbrace{\mathbf{E}[C_S]}_{I_d} \nabla f(x^k) \right\rangle + \frac{\gamma^2}{2} \left\langle \nabla f(x), \mathbf{E}[C_S L C_S] \nabla f(x^k) \right\rangle . \end{aligned} \tag{141}$$

Proof of Theorem 64 - II

Step 3: Diagonalize the quadratic

Note that $\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S$ is symmetric and positive semidefinite, and so is $\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right]$. Therefore

$$\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right] \preceq \underbrace{\lambda_{\max}(\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right])}_{\stackrel{\text{def}}{=} A^{III}} \boldsymbol{I}_d,$$

which implies the inequality

$$\langle z, \mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right] z\rangle \leq \left\langle z, A^{III} \boldsymbol{I}_d z\right\rangle = A^{III} \|z\|^2, \qquad \forall z \in \mathbb{R}^d. \tag{142}$$

Plugging (142) into (141), and subtracting $f(x^\star)$ on both sides, we get

$$\begin{aligned}
\mathbf{E}\left[f(x^{k+1}) - f(x^\star) \mid x^k\right] &\stackrel{(142)}{=} f(x^k) - f(x^\star) - \gamma \left\|\nabla f(x^k)\right\|^2 + \frac{\gamma^2}{2} A^{III} \left\|\nabla f(x^k)\right\|^2 \\
&= f(x^k) - f(x^\star) - \gamma \left(1 - \frac{\gamma A^{III}}{2}\right) \left\|\nabla f(x^k)\right\|^2.
\end{aligned}$$

Step 4: Simplify by applying stepsize bound

Since $\gamma \leq \frac{1}{A^{III}}$, we have $\left(1 - \frac{\gamma A^{III}}{2}\right) \geq \frac{1}{2}$, and hence

$$\mathbf{E}\left[f(x^{k+1}) - f(x^\star) \mid x^k\right] \leq f(x^k) - f(x^\star) - \frac{\gamma}{2} \left\|\nabla f(x^k)\right\|^2. \tag{143}$$

Proof of Theorem 64 - III

Step 5: Apply strong convexity

Since f is μ-convex, we have

$$2(f(x^k) - f(x^\star)) \leq \frac{1}{\mu} \left\| \nabla f(x^k) - \underbrace{\nabla f(x^\star)}_{0} \right\|^2 . \tag{144}$$

Applying (144) in (143), we get

$$\mathbf{E}\left[f(x^{k+1}) - f(x^\star) \mid x^k \right] \quad \leq \quad (1 - \gamma\mu)(f(x^k) - f(x^\star)). \tag{145}$$

Step 6: Apply tower property to form a recurrence

Taking expectation again, and using the tower property, we get

$$\begin{aligned} \mathbf{E}\left[f(x^{k+1}) - f(x^\star) \right] &= \mathbf{E}\left[\mathbf{E}\left[f(x^{k+1}) - f(x^\star) \mid x^k \right] \right] \\ &\overset{(145)}{\leq} (1 - \gamma\mu)\mathbf{E}\left[f(x^k) - f(x^\star) \right]. \end{aligned}$$

Step 7: Solve the recurrence

It only remains to unroll the recursion.

Exercises

Exercises - I

Exercise 48

Let S be any sampling (i.e., a random subset of $\{1, 2, \ldots, d\}$) and let $p_i = \mathrm{Prob}(i \in S)$. Show that $\mathbf{E}[|S|] = \sum_{i=1}^{d} p_i$.

Exercise 49

Show that if $\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]$ is finite, then it must necessarily be a symmetric and positive semidefinite matrix.

Exercise 50

Show that if $\mathbf{E}[\boldsymbol{C}] = \boldsymbol{I}_d$ and $\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]$ is finite, then $\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right] \succeq \boldsymbol{I}_d$. In particular, $\lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]\right) \geq 1$.

Exercise 51

Let $f(x) = \frac{1}{2} x^\top \boldsymbol{L} x$, where $\boldsymbol{L}$ is a $d \times d$ positive definite matrix. Perform an analysis of RCD *applied to minimizing this function, aiming to obtain a bound for $\mathbf{E}\left[\left\|x^k - x^\star\right\|_{\boldsymbol{L}}^2\right]$. Compare the complexity result with that of the two results obtained in this lecture.*

Exercises - II

Exercise 52

Compute $\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right]$ *for* S *being a* τ*-nice sampling of* $[d]$*. What is the largest eigenvalue of this matrix? That is, calculate* $\lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right]\right)$*.*

Exercise 53

Show that for any proper sampling S *and any positive semidefinite matrix* $\boldsymbol{L}$*, we have*

$$\lambda_{\max}\left(\mathbf{E}\left[\boldsymbol{C}_S \boldsymbol{L} \boldsymbol{C}_S\right]\right) \leq \lambda_{\max}(\boldsymbol{P}^{-1/2} \boldsymbol{L} \boldsymbol{P}^{-1/2}) \leq \lambda_{\max}(\boldsymbol{L}) \lambda_{\max}(\boldsymbol{P}^{-1}).$$

What does this say about the convergence rate of `RCD` *as described by the three convergence analyses?*

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 11: Compression Operators

Based on:

P.R., Stochastic Gradient Descent Methods, Manuscript, 2019

[86] Wen, Wei and Xu, Cong and Yan, Feng and Wu, Chunpeng and Wang, Yandan and Chen, Yiran and Li, Hai, **Terngrad: Ternary gradients to reduce communication in distributed deep learning**, NeurIPS 2017

[47] K. Mishchenko, E. Gorbunov, M. Takáč and P. Richtárik, **Distributed learning with compressed gradient differences**, arXiv:1901.09269, 2019

[30] S. Horváth, C.-Y. Ho, Ľ. Horváth, A.N. Sahu, M. Canini and P. Richtárik, **Natural compression for distributed deep learning**, Mathematical and Scientific Machine Learning 2022 (arXiv:1905.10988, 2019)

Introduction

In today's lecture we'll

- see how to **combine compression operators** to produce new compressor operators
- give a couple of examples of compression operators which perform **quantization**

Manipulating Compression Operators

Manipulating Compression Operators

- **Composition**
- **Direct sum**
- **Perspective**
- **Probabilistic switching**
- **Convex combination**

Composition of Compression Operators

Theorem 65 (Composition)

Let $\mathcal{C}_1 \in \mathbb{B}^d(\omega_1)$ and $\mathcal{C}_2 \in \mathbb{B}^d(\omega_2)$ be independent compression operators, and let $\mathcal{C}$ be their composition:

$$\boxed{\mathcal{C}(x) \overset{def}{=} (\mathcal{C}_1 \circ \mathcal{C}_2)(x) = \mathcal{C}_1(\mathcal{C}_2(x))}$$

Then

$$\mathcal{C} \in \mathbb{B}^d\left((\omega_1+1)(\omega_2+1)-1\right). \tag{146}$$

Example 66

Let $\mathcal{C}_1$ be natural compression and $\mathcal{C}_2$ be a sparsifier generated by a proper sampling. Then $\mathcal{C}_1 \in \mathbb{B}^d(1/8)$ and $\mathcal{C}_2 \in \mathbb{B}^d\left(\max_i \frac{1}{p_i} - 1\right)$, and hence $\mathcal{C}_1 \circ \mathcal{C}_2 \in \mathbb{B}^d\left(\max_i \frac{9}{8p_i} - 1\right)$.

Proof of Theorem 65

Let us first show unbiasedness. By independence and unbiasedness of $\mathcal{C}_1$, we get

$$\mathbf{E}\left[\mathcal{C}_1(\mathcal{C}_2(x)) \mid \mathcal{C}_2\right] = \mathcal{C}_2(x). \tag{147}$$

Applying the tower property and using unbiasedness of $\mathcal{C}_2$, we get

$$\begin{aligned}
\mathbf{E}\left[\mathcal{C}_1(\mathcal{C}_2(x))\right] &= \mathbf{E}\left[\mathbf{E}\left[\mathcal{C}_1(\mathcal{C}_2(x)) \mid \mathcal{C}_2\right]\right] \\
&\overset{(147)}{=} \mathbf{E}\left[\mathcal{C}_2(x)\right] \\
&= x.
\end{aligned}$$

So, $\mathcal{C}$ is unbiased. Next, recall that $\mathbf{E}\left[\|\mathcal{C}_i(x)\|^2\right] \leq (\omega_i + 1)\|x\|^2$. Using this fact for $i = 1$, by independence we get

$$\mathbf{E}\left[\|\mathcal{C}_1(\mathcal{C}_2(x))\|^2 \mid \mathcal{C}_2\right] \leq (\omega_1 + 1)\|\mathcal{C}_2(x)\|^2. \tag{148}$$

By the tower property and the assumption that $\mathcal{C}_2 \in \mathbb{B}^d(\omega_2)$, we get

$$\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}_1(\mathcal{C}_2(x))\|^2\right] &= \mathbf{E}\left[\mathbf{E}\left[\|\mathcal{C}_1(\mathcal{C}_2(x))\|^2 \mid \mathcal{C}_2\right]\right] \\
&\overset{(148)}{\leq} \mathbf{E}\left[(\omega_1 + 1)\|\mathcal{C}_2(x)\|^2\right] \\
&= (\omega_1 + 1)\mathbf{E}\left[\|\mathcal{C}_2(x)\|^2\right] \\
&\leq (\omega_1 + 1)(\omega_2 + 1)\|x\|^2.
\end{aligned}$$

Hence, $\mathcal{C} \in \mathbb{B}^d(\omega)$ for $\omega = (\omega_1 + 1)(\omega_2 + 1) - 1$.

Direct Sum of Compression Operators

Theorem 67 (Direct Sum)

Let $d = \sum_{i=1}^{m} d_i$*, where* $\{d_i\}$ *are positive integers, and let* $x \in \mathbb{R}^d$ *be written as* $x = (x_1, \ldots, x_m)$*, where* $x_i \in \mathbb{R}^{d_i}$*. Consider compression operators*

$$\mathcal{C}_1 \in \mathbb{B}^{d_1}(\omega_1), \quad \mathcal{C}_2 \in \mathbb{B}^{d_2}(\omega_2), \quad \ldots, \quad \mathcal{C}_m \in \mathbb{B}^{d_m}(\omega_m).$$

Define $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ *as follows:*

$$\boxed{\mathcal{C}(x_1, \ldots, x_m) = (\mathcal{C}_1(x_1), \ldots, \mathcal{C}_m(x_m))}$$

Then

$$\mathcal{C} \in \mathbb{B}^d\left(\max_i \omega_i\right). \tag{149}$$

Example 68

Natural compression is constructed this way with $d_i = 1$ for all i. Indeed, we first define natural compression as a compressor from $\mathbb{R}$ to $\mathbb{R}$, and then apply it element-wise to $x \in \mathbb{R}^d$.

Proof of Theorem 67

Fix any $x = (x_1, \ldots, x_m) \in \mathbb{R}^d$. Let us first show unbiasedness. Applying linearity of expectation and then using unbiasedness of the constituent compressors $\mathcal{C}_1, \ldots, \mathcal{C}_m$, we get

$$\begin{aligned} \mathbf{E}\left[\mathcal{C}(x)\right] &= \mathbf{E}\left[(\mathcal{C}_1(x_1), \ldots, \mathcal{C}_m(x_m))\right] \\ &= \left(\mathbf{E}\left[\mathcal{C}_1(x_1)\right], \ldots, \mathbf{E}\left[\mathcal{C}_m(x_m)\right]\right) \\ &= (x_1, \ldots, x_m) \;=\; x. \end{aligned}$$

It now suffices to establish a bound on the second moment:

$$\begin{aligned} \mathbf{E}\left[\|\mathcal{C}(x)\|^2\right] &= \mathbf{E}\left[\|(\mathcal{C}_1(x_1), \ldots, \mathcal{C}_m(x_m))\|^2\right] \\ &= \mathbf{E}\left[\sum_{i=1}^m \|\mathcal{C}_i(x_i)\|^2\right] \\ &= \sum_{i=1}^m \mathbf{E}\left[\|\mathcal{C}_i(x_i)\|^2\right] \\ &\overset{(106)}{\leq} \sum_{i=1}^m (1+\omega_i)\,\|x_i\|^2 \\ &\leq (1+\max_i \omega_i)\sum_{i=1}^m \|x_i\|^2 \;=\; (1+\max_i \omega_i)\,\|x\|^2\,. \end{aligned}$$

So, $\mathcal{C} \in \mathbb{B}^d\left(\max_i \omega_i\right)$.

Perspective of a Compression Operator

Theorem 69 (Perspective)

Let $\mathcal{C} \in \mathbb{B}^d(\omega)$. *Let* $\phi : \mathbb{R}^d \to \mathbb{R}/\{0\}$ *and define*

$$\boxed{\mathcal{C}_\phi(x) \stackrel{def}{=} \phi(x)\mathcal{C}\left(\frac{x}{\phi(x)}\right)}$$

Then

$$\mathcal{C}_\phi \in \mathbb{B}^d(\omega). \tag{150}$$

Example 70

Dithering compressor arises in this way, with $\phi(x) = \|x\|$, and $\mathcal{C}$ being a direct sum of 1-dimensional randomized rounding operators.

Proof of Theorem

Choose any $x \in \mathbb{R}^d$ and let $t = \phi(x)$.

Unbiasedness follows trivially:

$$\mathbf{E}\left[\mathcal{C}_t(x)\right] = \mathbf{E}\left[t\mathcal{C}\left(\frac{x}{t}\right)\right] = t\mathbf{E}\left[\mathcal{C}\left(\frac{x}{t}\right)\right] = t\frac{x}{t} = x.$$

The second moment can be bounded as

$$\begin{aligned}
\mathbf{E}\left[\left\|\mathcal{C}_t(x)\right\|^2\right] &= \mathbf{E}\left[\left\|t\mathcal{C}\left(\frac{x}{t}\right)\right\|^2\right] \\
&= t^2\mathbf{E}\left[\left\|\mathcal{C}\left(\frac{x}{t}\right)\right\|^2\right] \\
&\leq t^2(\omega+1)\left\|\frac{x}{t}\right\|^2 \\
&= (\omega+1)\left\|x\right\|^2.
\end{aligned}$$

Therefore, $\mathcal{C}_\phi \in \mathbb{B}^d(\omega)$.

Probabilistic Switching Among Compression Operators

Theorem 71 (Probabilistic Switching)

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m : \mathbb{R}^d \to \mathbb{R}^d$ *be random compression operators, and assume that* $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ *for some constants* $\omega_i \geq 0$ *for all* i*. Let* $p_1, p_2, \ldots, p_m$ *be non-negative numbers summing up to one, and define a new randomized compression map* $\mathcal{C}$ *as follows:*

$$\boxed{\mathcal{C}(x) \stackrel{def}{=} \begin{cases} \mathcal{C}_1(x) & \text{with probability } p_1 \\ \mathcal{C}_2(x) & \text{with probability } p_2 \\ \vdots & \vdots \\ \mathcal{C}_m(x) & \text{with probability } p_m \end{cases}}$$

Then

$$\mathcal{C} \in \mathbb{B}^d\left(\sum_{i=1}^m p_i \omega_i\right). \tag{151}$$

Proof of Theorem 71

Let us first show unbiasedness. Using the tower property and linearity of expectation, and subsequently unbiasedness of the compression operators $\mathcal{C}_1, \ldots, \mathcal{C}_m$, we get

$$\begin{aligned}
\mathbf{E}\left[\mathcal{C}(x)\right] &= \mathbf{E}\left[\mathbf{E}\left[\mathcal{C}(x) \mid \mathcal{C}_1, \ldots, \mathcal{C}_m\right]\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m} p_i \mathcal{C}_i(x)\right] \\
&= \sum_{i=1}^{m} p_i \mathbf{E}\left[\mathcal{C}_i(x)\right] = \sum_{i=1}^{m} p_i x = x,
\end{aligned}$$

establishing that $\mathcal{C}$ is also unbiased. Using similar arguments, we get

$$\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] &= \mathbf{E}\left[\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2 \mid \mathcal{C}_1, \ldots, \mathcal{C}_m\right]\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m} p_i \|\mathcal{C}_i(x) - x\|^2\right] \\
&= \sum_{i=1}^{m} p_i \mathbf{E}\left[\|\mathcal{C}_i(x) - x\|^2\right] \\
&\leq \sum_{i=1}^{m} p_i \omega_i \|x\|^2 = \left(\sum_{i=1}^{m} p_i \omega_i\right) \|x\|^2 .
\end{aligned}$$

Therefore, $\mathcal{C} \in \mathbb{B}^d\left(\sum_{i=1}^{m} p_i \omega_i\right)$.

Convex Combination of Compression Operators

Theorem 72 (Convex Combination)

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m : \mathbb{R}^d \to \mathbb{R}^d$ *be random compression operators, and assume that* $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ *for some constants* $\omega_i \geq 0$ *for all* i*. Let* $p_1, p_2, \ldots, p_m$ *be non-negative numbers summing up to one, and define a new randomized compression map* $\mathcal{C}$ *as follows:*

$$\boxed{\mathcal{C}(x) \stackrel{def}{=} \sum_{i=1}^{m} p_i \mathcal{C}_i(x)}$$

Then

$$\mathcal{C} \in \mathbb{B}^d\left(\sum_{i=1}^{m} p_i \omega_i\right). \tag{152}$$

Proof of Theorem 72

Using linearity of expectation and unbiasedness of the compression operators $\mathcal{C}_1, \ldots, \mathcal{C}_m$, we get

$$\mathbf{E}\left[\mathcal{C}(x)\right] = \mathbf{E}\left[\sum_{i=1}^{m} p_i \mathcal{C}_i(x)\right] = \sum_{i=1}^{m} p_i \mathbf{E}\left[\mathcal{C}_i(x)\right] = \sum_{i=1}^{m} p_i x = x,$$

establishing that $\mathcal{C}$ is also unbiased.

$$\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] &= \mathbf{E}\left[\left\|\sum_{i=1}^{m} p_i \mathcal{C}_i(x) - x\right\|^2\right] = \mathbf{E}\left[\left\|\sum_{i=1}^{m} p_i \left(\mathcal{C}_i(x) - x\right)\right\|^2\right] \\
&\leq \mathbf{E}\left[\sum_{i=1}^{m} p_i \|\mathcal{C}_i(x) - x\|^2\right] \\
&= \sum_{i=1}^{m} p_i \mathbf{E}\left[\|\mathcal{C}_i(x) - x\|^2\right] \\
&\leq \sum_{i=1}^{m} p_i \omega_i \|x\|^2 = \left(\sum_{i=1}^{m} p_i \omega_i\right) \|x\|^2 .
\end{aligned}$$

The first inequality above follows from Jensen's inequality applied to the convex function $x \mapsto \|x\|^2$, the identity that follows is due to linearity of expectation, and the second inequality follows from the assumption that $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ for all i.

Quantization Operators

We will describe several quantization compressors:

- natural compression
- general dithering
  - ternary quantization
  - standard dithering
  - natural dithering

Quantization compressor	Notation	# levels	level structure	Definition
Natural compression	$\mathcal{C}_{\mathrm{nat}}$	$+\infty$	geometric	74
General dithering	$\mathcal{D}_{\mathrm{gen}}^{\mathcal{C},p,s}$	s	any	77
Ternary quantization	$\mathcal{D}_{\mathrm{gen}}^{\mathcal{C},p,1}$	1	–	79
Standard dithering	$\mathcal{D}_{\mathrm{stand}}^{\mathcal{C},p,s}$	s	linear	80
Natural dithering	$\mathcal{D}_{\mathrm{nat}}^{\mathcal{C},p,s}$	s	geometric	81

Natural Compression

Natural Compression: Informal Definition

Natural compression, denoted $\mathcal{C}_{\text{nat}} : \mathbb{R}^d \to \mathbb{R}^d$, is a compression operator introduced in 2019 by Horváth et al [30].

It is **applied element-wise** to $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$(\mathcal{C}_{\text{nat}}(x))_i = \mathcal{C}_{\text{nat}}(x_i).$$

So, we only need to define $\mathcal{C}_{\text{nat}} : \mathbb{R} \to \mathbb{R}$:

- When $t = 0$, we set $\mathcal{C}_{\text{nat}}(0) = 0$.
- When $t = 2^k$ for some integer k, we set $\mathcal{C}_{\text{nat}}(t) = t$.
- Otherwise we perform a **randomized rounding of the input $t \in \mathbb{R}$ to one of the two "closest" integer powers of 2:**
  - Let $\alpha \in \mathbb{R}$ be such that $|t| = 2^\alpha$ (i.e., $\alpha = \log_2 |t|$). Then

$$2^{\lfloor \alpha \rfloor} \leq |t| = 2^\alpha \leq 2^{\lceil \alpha \rceil} \tag{153}$$

    and we round t to either $\text{sign}(t) 2^{\lfloor \alpha \rfloor}$, or to $\text{sign}(t) 2^{\lceil \alpha \rceil}$.
  - The probabilities are chosen so that $\mathcal{C}_{\text{nat}}(t)$ is an **unbiased estimator** of t, i.e., $\mathbf{E}\left[\mathcal{C}_{\text{nat}}(t)\right] = t$ for all t.

Natural Compression: Examples

Figure: An illustration of natural compression applied to $t = 2.5$: $\mathcal{C}_{\rm nat}(2.5) = 2$ with probability $\frac{4-2.5}{2} = 0.75$, and $\mathcal{C}_{\rm nat}(2.5) = 4$ with probability $\frac{2.5-2}{2} = 0.25$. This choice of probabilities ensures that the compression operator is unbiased, i.e., $\mathbf{E}\left[\mathcal{C}_{\rm nat}(t)\right] = t$ for all t.

Example 73

- $t = 2.5$ will be rounded to either 2 or 4 (since $2^1 \leq 2.5 \leq 2^4$) - see the Figure
- $t = -2.75$ will be rounded to either -4 or -2 (since $-2^2 \leq -2.75 \leq -2^1$)
- $t = 0.75$ will be rounded to either $\frac{1}{2}$ or 1 (since $2^{-1} \leq 0.75 \leq 2^0$)

Natural Compression: Formal Definition I

The above informal definition can be formalized as follows:

Definition 74 (Natural Compression for $d = 1$)

Natural compression is a random function $\mathcal{C}_{\rm nat} : \mathbb{R} \mapsto \mathbb{R}$ defined as follows. We set

$$\mathcal{C}_{\rm nat}(0) = 0.$$

If $t \neq 0$, we let

$$\mathcal{C}_{\rm nat}(t) \stackrel{\rm def}{=} \begin{cases} \mathrm{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}, & \text{with probability } p(t), \\ \mathrm{sign}(t) \cdot 2^{\lceil \log_2 |t| \rceil}, & \text{with probability } 1 - p(t), \end{cases} \tag{154}$$

where the probability is given by

$$p(t) \stackrel{\rm def}{=} \frac{2^{\lceil \log_2 |t| \rceil} - |t|}{2^{\lfloor \log_2 |t| \rfloor}}.$$

Natural Compression: Formal Definition II

Alternatively, for t which is not 0 or an integer power of 2, (154) can be written as

$$\mathcal{C}_{\mathrm{nat}}(t) = \mathrm{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}(1 + \lambda(t)),$$

where

$$\lambda(t) \sim \mathrm{Bernoulli}(1 - p(t)).$$

That is,

$$\lambda(t) = \begin{cases} 1 & \text{with probability } 1 - p(t) \\ 0 & \text{with probability } p(t) \end{cases}$$

The key properties of any (unbiased) compression operator are **variance**, **ease of implementation**, and **compression level.** We characterize the remarkably low variance of $\mathcal{C}_{\mathrm{nat}}$ and describe an (almost) effortless and **natural** implementation, and the compression it offers.

Natural Compression: Variance

Theorem 75 (Horváth et al [30])

$\mathcal{C}_{\mathrm{nat}} \in \mathbb{B}^d\left(\frac{1}{8}\right)$.

Proof of Theorem 75 - I

Unbiasedness in the $d = 1$ case. Recall that $\mathcal{C}_{\text{nat}}(t)$ can be written in the form

$$\mathcal{C}_{\text{nat}}(t) = \text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}(1 + \lambda(t)), \tag{155}$$

where $\lambda(t) \sim \text{Bernoulli}(1 - p(t))$ and

$$p(t) = \frac{2^{\lceil \log_2 |t| \rceil} - |t|}{2^{\lfloor \log_2 |t| \rfloor}}. \tag{156}$$

Hence,

$$\begin{aligned}
\mathbf{E}\left[\mathcal{C}_{\text{nat}}(t)\right] &\overset{(155)}{=} \mathbf{E}\left[\text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}(1 + \lambda(t))\right] \\
&= \text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}\left(1 + \mathbf{E}\left[\lambda(t)\right]\right) \\
&= \text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}\left(1 + 1 - p(t)\right) \\
&\overset{(156)}{=} \text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}\left(\frac{2 \cdot 2^{\lfloor \log_2 |t| \rfloor}}{2^{\lfloor \log_2 |t| \rfloor}} - \frac{2^{\lceil \log_2 |t| \rceil} - |t|}{2^{\lfloor \log_2 |t| \rfloor}}\right) \\
&\overset{(156)}{=} \text{sign}(t) \cdot 2^{\lfloor \log_2 |t| \rfloor}\left(\frac{|t|}{2^{\lfloor \log_2 |t| \rfloor}}\right) \\
&= \text{sign}(t)|t| \\
&= t.
\end{aligned}$$

This establishes unbiasedness.

Proof of Theorem 75 - II

Variance in the $d = 1$ case. It suffices to show that

$$\mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(t))^2\right] \leq (\omega + 1)t^2, \qquad \forall t \in \mathbb{R} \tag{157}$$

for $\omega = 1/8$.

- If $t = 0$ or $t = \mathrm{sign}(t)2^\alpha$ with α being an integer, then $\mathcal{C}_{\mathrm{nat}}(t) = t$, and (157) holds as an identity with $\omega = 0$, and hence inequality (157) holds for $\omega = 1/8$.
- Otherwise $t = \mathrm{sign}(t)2^\alpha$ where $a \stackrel{\mathrm{def}}{=} \lfloor \alpha \rfloor < \alpha < \lceil \alpha \rceil = a + 1$. With this notation, we can write

$$\mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(t))^2\right] = 2^{2a}\frac{2^{a+1} - |t|}{2^a} + 2^{2(a+1)}\frac{|t| - 2^a}{2^a} = 2^a(3|t| - 2^{a+1}).$$

So,

$$\begin{aligned}\frac{\mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(t))^2\right]}{t^2} &= \frac{2^a(3|t| - 2^{a+1})}{t^2} \leq \sup_{2^a < t < 2^{a+1}} \frac{2^a(3|t| - 2^{a+1})}{t^2} \\ &= \sup_{1<\theta<2} \frac{2^a(3 \cdot 2^a\theta - 2^{a+1})}{(2^a\theta)^2} = \sup_{1<\theta<2} \frac{3\theta - 2}{\theta^2}.\end{aligned}$$

The optimal solution of the last maximization problem is $\theta = 4/3$, with optimal objective value $9/8$. This implies that (157) holds with $\omega = 1/8$.

Proof of Theorem 75 - III

Unbiasedness in the general d case. Due to linearity of expectation, the element-wise structure of natural compression, and unbiasedness in the $d = 1$ case established above, for each i we have

$$(\mathbf{E}\left[\mathcal{C}_{\mathrm{nat}}(x)\right])_i = \mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(x))_i\right] = \mathbf{E}\left[\mathcal{C}_{\mathrm{nat}}(x_i)\right] = x_i,$$

and hence $\mathcal{C}_{\mathrm{nat}} : \mathbb{R}^d \to \mathbb{R}^d$ is unbiased.

Variance in the general d case. For any $x \in \mathbb{R}^d$ we have

$$\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}_{\mathrm{nat}}(x)\|^2\right] &= \mathbf{E}\left[\sum_{i=1}^d (\mathcal{C}_{\mathrm{nat}}(x))_i^2\right] = \sum_{i=1}^d \mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(x))_i^2\right] = \sum_{i=1}^d \mathbf{E}\left[(\mathcal{C}_{\mathrm{nat}}(x_i))^2\right] \\
&\leq \sum_{i=1}^d (1/8 + 1) x_i^2 = (1/8 + 1) \sum_{i=1}^d x_i^2 = (1/8 + 1)\|x\|^2 .
\end{aligned}$$

The first equality follows from the definition of standard Euclidean norm, the second equality is due to linearity of expectation, and the third equality is due to the element-wise definition of natural compression. The first inequality is due to the variance bound in the $d = 1$ case, and the rest is straightforward simplification.

Implementation of Natural Compression I

Key message:

- Performing natural compression of a real number in a binary floating point format is computationally cheap.
- Excluding the randomization step, $\mathcal{C}_{\mathrm{nat}}$ amounts to simply **dispensing off the mantissa in the binary representation.**

We'll now justify these claims.

Implementation of Natural Compression II

Binary32

- The most common computer format for real numbers, **binary**32 (resp. *binary*64) **of the IEEE 754 standard, represents each number with** 32 (resp. 64) **bits,** where the first bit represents the sign, 8 (resp. 11) bits are used for the exponent, and the remaining 23 (resp. 52) bits are used for the mantissa.
- A scalar $t \in \mathbb{R}$ is represented in the form $(s, e_7, e_6, \ldots, e_0, m_1, m_2, \ldots, m_{23})$, where $s, e_i, m_j \in \{0, 1\}$ are bits, via the relationship

$$t = (-1)^s \times 2^{e-127} \times (1 + m), \quad e = \sum_{i=0}^{7} e_i 2^i, \quad m = \sum_{j=1}^{23} m_j 2^{-j}, \tag{158}$$

where s is the **sign**, e is the **exponent** and m is the **mantissa**.

Implementation of Natural Compression III

Example 76

A *binary*32 representation of $t = -2.75$ is visualized below. In this case, $s = 1$, $e_7 = 1$, $m_2 = m_3 = 1$ and hence

$$t = (-1)^s \times 2^{e-127} \times (1+m) = -1 \times 2 \times (1 + 2^{-2} + 2^{-3}) = -2.75.$$

Figure: IEEE 754 single-precision binary floating-point format: *binary*32.

- It is clear from (158) that $0 \leq m < 1$, and hence $2^{e-127} \leq |t| < 2^{e-126}$ (compare this with (153)). Moreover,

$$p(t) = \frac{2^{e-126} - |t|}{2^{e-127}} = 2 - |t|2^{127-e} = 1 - m.$$

Implementation of Natural Compression IV

- Hence, **natural compression of t represented as binary32 is given as follows:**

$$\mathcal{C}_{\mathrm{nat}}(t) = \begin{cases} (-1)^s \times 2^{e-127}, & \text{with probability} \quad 1-m, \\ (-1)^s \times 2^{e-126}, & \text{with probability} \quad m. \end{cases}$$

- Observe that $(-1)^s \times 2^{e-127}$ is **obtained from t by setting the mantissa m to zero, and keeping both the sign s and exponent e unchanged.**
- Similarly, $(-1)^s \times 2^{e-126}$ is **obtained from t by setting the mantissa m to zero, keeping the sign s, and increasing the exponent by one, which amounts to a simple shift of the bits forming the exponent to the left by one spot.**
- Hence, **both values can be computed from t essentially without any computation.**

Implementation of Natural Compression V

Communication savings

- In case of $\mathrm{binary}32$, the output $\mathcal{C}_{\mathrm{nat}}(t)$ of natural compression is encoded using the 8 bits in the exponent and an extra bit for the sign. **This is $3.56\times$ less communication.**
- In case of $\mathrm{binary}64$, we only need 11 bits for the exponent and 1 bit for the sign, and **this is $5.82\times$ less communication.**

General Dithering and Special Cases

p norm in $\mathbb{R}^d$

For $1 \leq p \leq +\infty$, the p-norm of vector $x \in \mathbb{R}^d$ is defined by

$$\|x\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} .$$

Monotonicity:

$$q \leq p \quad \Rightarrow \|x\|_p \leq \|x\|_q$$

For $q = 2$ we get

$$2 \leq p \quad \Rightarrow \quad \|x\|_p \leq \|x\|_2 \tag{159}$$

Dithering

Definition 77 (General dithering [30])

The **general dithering** operator associated with $\mathcal{C} \in \mathbb{B}^d(\omega)$, $1 \leq p \leq +\infty$, and s levels

$$0 = l_s < l_{s-1} < l_{s-2} < \cdots < l_1 < l_0 = 1,$$

denoted $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}$, is defined as follows. Let $x \in \mathbb{R}^d$.

- If $x = 0$, we let $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}(x) = 0$.
- If $x \neq 0$, we let

$$y_i \overset{\text{def}}{=} \frac{|x_i|}{\|x\|_p} \tag{160}$$

for all $i \in [d]$. For fixed i let $u \in \{0, 1, \ldots, s-1\}$ be such that $l_{u+1} \leq y_i \leq l_u$. Then we set

$$\left(\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}(x)\right)_i = \mathcal{C}(\|x\|_p) \times \text{sign}(x_i) \times \xi(y_i)\ , \tag{161}$$

where

$$\xi(y_i) \overset{\text{def}}{=} \begin{cases} l_u & \text{with probability } \frac{y_i - l_{u+1}}{l_u - l_{u+1}} \\ l_{u+1} & \text{with probability } \frac{l_u - y_i}{l_u - l_{u+1}} \end{cases}. \tag{162}$$

General Dithering: Commentary

When $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}$ is used to compress gradients, each worker communicates

- the norm compressed by $\mathcal{C}$ (at most 1 float),
- vector of signs (d bits), and
- efficient encoding of $\xi(y_i) \in \{l_0, \ldots, l_s\}$ for each entry $i = 1, 2, \ldots, d$.

Unbiasedness of General Dithering

Lemma 78

The general dithering operator $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}$ is unbiased.

Proof.

For $x = 0$ the statement holds trivially. Hence, fix $x \neq 0$. First, note that

$$\begin{aligned}
\mathbf{E}\left[\xi(y_i)\right] &\overset{(162)}{=} \frac{y_i - l_{u+1}}{l_u - l_{u+1}} l_u + \frac{l_u - y_i}{l_u - l_{u+1}} l_{u+1} \\
&= \frac{y_i l_u - l_{u+1} l_u + l_u l_{u+1} - y_i l_{u+1}}{l_u - l_{u+1}} = y_i \frac{l_u - l_{u+1}}{l_u - l_{u+1}} = y_i. \qquad (163)
\end{aligned}$$

Using the unbiasedness of $\mathcal{C}$ and independence of $\mathcal{C}$ and ξ, we get

$$\begin{aligned}
\mathbf{E}\left[(\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}(x))_i\right] &\overset{(161)}{=} \mathbf{E}\left[\mathcal{C}(\|x\|_p) \times \text{sign}(x_i) \times \xi(y_i)\right] \\
&= \mathbf{E}\left[\mathcal{C}(\|x\|_p)\right] \times \text{sign}(x_i) \times \mathbf{E}\left[\xi(y_i)\right] \\
&\overset{(163)}{=} \|x\|_p \times \text{sign}(x_i) \times y_i \\
&\overset{(160)}{=} \|x\|_p \times \text{sign}(x_i) \times \frac{|x_i|}{\|x\|_p} \\
&= x_i.
\end{aligned}$$

Ternary Quantization

Definition 79 (Ternary quantization)

In the special case when $s = 1$, general dithering $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}$ reduces to the **ternary quantization operator.**

Remarks:

- Note that the output of $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,1}$ is a vector whose all entries assume one of **three values** only:

$$\left(\mathcal{D}_{\text{gen}}^{\mathcal{C},p,1}(x)\right)_i \in \left\{-\mathcal{C}(\|x\|_p), 0, +\mathcal{C}(\|x\|_p)\right\}.$$

- For $\mathcal{C}$ being the identity compressor, $\mathcal{D}_{\text{gen}}^{\mathcal{C},\infty,1}$ was used in the `Terngrad` [86] mehod, without any analysis.
- `Terngrad` was first analyzed in [47]. However, [47] considered the more general **ternary quantization operator** $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,1}$ for $p \in [1, +\infty]$, and proposed a more efficient method to `Terngrad`: `DIANA`. We will learn about `DIANA` more later.

Standard Dithering

Definition 80 (Standard dithering [1])

Standard dithering $\mathcal{D}^{p,s}_{\mathrm{sta}}$ [17, 68, 1] is obtained as a special case of general dithering for $\mathcal{C}$ equal to the identity operator, and a **linear partition of the unit interval:**

$$l_s = \frac{0}{s} = 0, \quad l_{s-1} = \frac{1}{s}, \quad l_{s-2} = \frac{2}{s}, \quad \ldots, \quad l_1 = \frac{s-1}{s}, \quad l_0 = \frac{s}{s} = 1.$$

- $\mathcal{D}^{2,s}_{\mathrm{sta}}$ was used in QSGD [1].

Natural Dithering

Natural Dithering

Definition 81 (Natural dithering [30])

Natural dithering $\mathcal{D}^{p,s}_{\text{nat}}$ is obtained as a special case of general dithering for $\mathcal{C}$ equal to the identity operator, and a **binary geometric partition of the unit interval:**

$$l_s = 0, \quad l_{s-1} = \frac{1}{2^{s-1}}, \quad l_{s-2} = \frac{1}{2^{s-2}}, \quad \ldots, \quad l_1 = \frac{1}{2}, \quad l_0 = \frac{1}{2^0} = 1. \tag{164}$$

Example 82 (Standard vs natural dithering)

A comparison of the ξ operators for standard and natural dithering with $s = 3$ levels applied to $t = 3/8$.

Natural Dithering: Commentary

Natural dithering $\mathcal{D}^{p,s}_{\mathrm{nat}}$ is essentially an application of $\mathcal{C}_{\mathrm{nat}}$ to all normalized entries of $x \in \mathbb{R}^d$, with two differences:

- we also communicate the compressed norm $\|x\|_p$,
- in $\mathcal{C}_{\mathrm{nat}}$ the interval $[0, 1/2^{s-1}]$ is subdivided further, to machine precision, and in this sense $\mathcal{D}^{p,s}_{\mathrm{nat}}$ **can be seen as a limited precision variant of** $\mathcal{C}_{\mathrm{nat}}$**.**

Further comments:

- As is the case with $\mathcal{C}_{\mathrm{nat}}$, when implementing $\mathcal{D}^{p,s}_{\mathrm{nat}}$, the mantissa is ignored, and one communicates exponents only.
- The norm compression is particularly useful on the master side since multiplication by a naturally compressed norm is just summation of the exponents.

Natural Dithering: Variance

Theorem 83

$$\mathcal{D}_{\mathrm{nat}}^{p,s} \in \mathbb{B}^d(\omega),$$

where

$$\omega = \frac{1}{8} + \frac{d^{\frac{1}{r}}}{2^{s-1}} \min\left\{1, \frac{d^{\frac{1}{r}}}{2^{s-1}}\right\},$$

and $r = \min\{p, 2\}$.

Key insight: **Natural dithering is exponentially better than standard dithering.** In particular, for the same level of variance, $\mathcal{D}_{\mathrm{nat}}^{p,s}$ uses only s levels while $\mathcal{D}_{\mathrm{sta}}^{p,u}$ uses $u = 2^{s-1}$ levels.

Proof of Theorem 83 - I

Unbiasedness. Unbiasedness of $\mathcal{D}_{\text{nat}}^{p,s}$ is a direct consequence of unbiasedness of $\mathcal{D}_{\text{gen}}^{\mathcal{C},p,s}$.

Next, we establish a bound on the second moment of ξ. Fix x_i and let $y_i = \frac{|x_i|}{\|x\|_p}$. We consider two cases:

- If $y_i \geq \frac{1}{2^{s-1}}$, then the second moment of $\xi(y_i)$ can be bounded in the same way as in the case of natural compression:

$$\mathbf{E}\left[\xi^2\left(y_i\right)\right] \leq \frac{9}{8} y_i^2.$$

- If $0 < y_i < \frac{1}{2^{s-1}}$, then $0 = l_{u+1} < y_i < l_u = \frac{1}{2^{s-1}}$, and the second moment of $\xi(y_i)$ can be bounded as

$$\begin{aligned}
\mathbf{E}\left[\xi^2\left(y_i\right)\right] &= \tfrac{y_i - l_{u+1}}{l_u - l_{u+1}} l_u^2 + \tfrac{l_u - y_i}{l_u - l_{u+1}} l_{u+1}^2 \\
&= y_i 2^{s-1} 2^{2(1-s)} \\
&= 2^{1-s} y_i.
\end{aligned}$$

Proof of Theorem 83 - II

Let $\mathbb{1}(A) = 1$ if statement A is true, and $\mathbb{1}(A) = 0$ if it is false. Collecting the two cases into a single inequality, we get

$$\begin{aligned} \mathbf{E}\left[\xi^2\left(y_i\right)\right] &\leq \mathbb{1}\left(y_i \geq 2^{1-s}\right)\frac{9}{8}y_i^2 + \mathbb{1}\left(y_i < 2^{1-s}\right)2^{1-s}y_i \\ &\leq \frac{9}{8}y_i^2 + \mathbb{1}\left(y_i < 2^{1-s}\right)2^{1-s}y_i\ . \end{aligned} \tag{165}$$

$$\begin{aligned} \left\|\mathcal{D}_{\mathrm{nat}}^{p,s}(x)\right\|^2 &= \sum_{i=1}^{d}\left[\left(\mathcal{D}_{\mathrm{nat}}^{p,s}(x)\right)_i\right]^2 \\ &\overset{(161)}{=} \sum_{i=1}^{d}\left[\|x\|_p \times \mathrm{sign}(x_i) \times \xi(y_i)\right]^2 \\ &= \|x\|_p^2 \sum_{i=1}^{d}\xi^2(y_i). \end{aligned}$$

Proof of Theorem 83 - III

Taking expectation on both sides and applying inequality (165), we get

$$
\begin{aligned}
\mathbf{E}\left[\left\|\mathcal{D}_{\mathrm{nat}}^{p,s}(x)\right\|^2\right] &= \|x\|_p^2 \sum_{i=1}^{d} \mathbf{E}\left[\xi^2(y_i)\right] \\
&\overset{(165)}{\leq} \|x\|_p^2 \sum_{i=1}^{d} \left(\frac{9}{8} y_i^2 + \mathbb{1}\left(y_i < 2^{1-s}\right) 2^{1-s} y_i\right) \\
&= \|x\|_p^2 \left[\frac{9}{8}\|y\|^2 + 2^{1-s} \sum_{i=1}^{d} \mathbb{1}\left(y_i < 2^{1-s}\right) y_i\right] \\
&\overset{(160)}{=} \frac{9}{8}\|x\|^2 + \|x\|_p^2 2^{1-s} \sum_{i=1}^{d} \mathbb{1}\left(y_i < 2^{1-s}\right) y_i \\
&\leq \frac{9}{8}\|x\|^2 + \|x\|_p^2 2^{1-s} \sum_{i=1}^{d} \min\left\{y_i, 2^{1-s}\right\} \\
&= \frac{9}{8}\|x\|^2 + \|x\|_p^2 2^{1-s} \sum_{i=1}^{d} \min\left\{\frac{|x_i|}{\|x\|_p}, 2^{1-s}\right\} \\
&= \frac{9}{8}\|x\|^2 + 2^{1-s} \sum_{i=1}^{d} \min\left\{|x_i| \|x\|_p, 2^{1-s}\|x\|_p^2\right\}.
\end{aligned}
$$

Proof of Theorem 83 - IV

Applying the inequality

$$\sum_i \min\{a_i, b_i\} \leq \min\left\{\sum_i a_i, \sum_i b_i\right\} \tag{166}$$

leads to

$$\mathbf{E}\left[\left\|\mathcal{D}_{\mathrm{nat}}^{p,s}(x)\right\|^2\right] \leq \frac{9}{8}\|x\|^2 + 2^{1-s}\min\left\{\|x\|_1\|x\|_p, d2^{1-s}\|x\|_p^2\right\}. \tag{167}$$

It remains to apply the following bounds:

- $\|x\|_p \leq \|x\|_2$ for $p \geq 2$ (see (159))
- $\|x\|_1 \leq \sqrt{d}\|x\|_2$ (see (169))
- $\|x\|_p \leq d^{1/p-1/2}\|x\|_2$ for $1 < p < 2$ (see (170))

Hölder's Inequality

Theorem 84 (Hölder's inequality)

For any $1 < p < \infty$ *and* p' *such that* $\frac{1}{p} + \frac{1}{p'} = 1$ *and all* $a, b \in \mathbb{R}^d$*:*

$$\sum_{i=1}^{d} |a_i b_i| \leq \left(\sum_{i=1}^{d} |a_i|^p \right)^{1/p} \left(\sum_{i=1}^{d} |b_i|^{p'} \right)^{1/p'} .$$

This can be also written as

$$\|a \circ b\|_1 \leq \|a\|_p \|b\|_{p'} ,$$

where $\circ$ *denotes the Hadamard (i.e., element-wise) product of vectors.*

Consequences of Hölder's Inequality

Lemma 85

Let $1 < p < \infty$ *and* p' *such that* $\frac{1}{p} + \frac{1}{p'} = 1$. *Then for all* $x \in \mathbb{R}^d$ *we have*

$$\|x\|_1 \leq \|x\|_p \, d^{1/p'}, \tag{168}$$

$$\|x\|_1 \leq \|x\|_2 \, d^{1/2}. \tag{169}$$

If $1 < p < 2$, *then for all* $x \in \mathbb{R}^d$ *we have*

$$\|x\|_p \leq \|x\|_2 \, d^{1/p-1/2}. \tag{170}$$

Proof of Lemma 85

- Specializing Hölder's inequality to $a = x$ and $b_i = 1$ for all i gives the result.
- This is special case of the above for $p = 2$.
- Note that $p' = \frac{p}{p-1} > 2$ and choose $|a_i| = |x_i|^p$ and $b_i = 1$ for all i. Further, let $r = \frac{2}{p} > 1$. Then $r' = \frac{1}{1-\frac{1}{r}} = \frac{r}{r-1} = \frac{2}{2-p}$. By Hölder inequality,

$$
\begin{aligned}
\|x\|_p^p &= \sum_{i=1}^d |x_i|^p = \sum_{i=1}^d |x_i|^p \cdot 1 = \|a \circ b\|_1 \leq \|a\|_r \|b\|_{r'} \\
&= \left(\sum_{i=1}^d (|x_i|^p)^r\right)^{1/r} \left(\sum_{i=1}^d 1^{r'}\right)^{1/r'} \\
&= \left(\sum_{i=1}^d |x_i|^2\right)^{p/2} d^{(2-p)/2} \\
&= \|x\|_2^p \, d^{1-p/2}.
\end{aligned}
$$

It only remains to raise both sides to the power of $1/p$.

Exercises

Exercises I

Exercise 54 (Composition of Multiple Compression Operators)

Formulate and prove a generalization of Theorem 65 to arbitrary number of compression operators.

Exercise 55

Prove inequality (166). *That is, prove that for any* $a, b \in \mathbb{R}^d$ *we have*

$$\sum_i \min\{a_i, b_i\} \leq \min\left\{\sum_i a_i, \sum_i b_i\right\}.$$

Exercise 56

Prove that for all $x \in \mathbb{R}^d$ *and any* $1 < p < 2$ *we have*

$$\|x\|_p \leq \|x\|_{p'} \, d^{2/p-1},$$

where p' *is such that* $\frac{1}{p} + \frac{1}{p'} = 1$.

Exercise 57 (Ternary compressor)

Show that the ternary compressor belongs to $\mathbb{B}^d(\omega)$ *for some* $\omega \geq 0$, *and find* ω. *Easier option: do this for* $\mathcal{C}$ *being the identity compressor and* $p = 2$. *More difficult option: do this for any* $\mathcal{C} \in \mathbb{B}^d(\omega')$ *and* $p \in [1, +\infty]$.

Exercises II

Exercise 58 (Integer compression / randomized rounding)

Let $\mathcal{C}_{\mathrm{int}} : \mathbb{R} \to \mathbb{R}$ *be a randomized rounding operator similar to* $\mathcal{C}_{\mathrm{nat}}$*, but one that rounds (in an unbiased way) to integers instead of powers of two. That is,*

$$\mathcal{C}_{\mathrm{int}}(t) \stackrel{def}{=} \begin{cases} \lfloor t \rfloor + 1 & \text{with probability} \quad t - \lfloor t \rfloor \\ \lfloor t \rfloor & \text{with probability} \quad 1 - t + \lfloor t \rfloor \end{cases}.$$

We define $\mathcal{C}_{\mathrm{int}} : \mathbb{R}^d \to \mathbb{R}^d$ *via the* **direct sum** *construction:*

$$(\mathcal{C}_{\mathrm{int}}(x))_i = \mathcal{C}_{\mathrm{int}}(x_i), \quad i = 1, 2, \ldots, d.$$

(a) *Prove that* $\mathcal{C}_{\mathrm{int}}$ *is unbiased.*

(b) *Prove that there is no finite* ω *such that* $\mathcal{C}_{\mathrm{int}} \in \mathbb{B}^d(\omega)$.

(c) *Prove that* $\mathbf{E}\left[\|\mathcal{C}_{\mathrm{int}}(x) - x\|^2\right] \leq \frac{d}{4}$ *for all* $x \in \mathbb{R}^d$.

Exercises III

Exercise 59 (Top_k compressor)

Let $1 \leq k \leq d$. By $\mathrm{Top}_k : \mathbb{R}^d \to \mathbb{R}^d$ we denote the compressor that preserves the top k entries of the input x in absolute value (breaking ties arbitrarily), and zeroes out the remaining entries. That is, given $x \in \mathbb{R}^d$, let $\pi_1, \ldots, \pi_d$ be a permutation of the indices $1, \ldots, d$ such that

$$|x_{\pi_1}| \geq |x_{\pi_2}| \geq \cdots \geq |x_{\pi_d}|.$$

Then

$$(\mathrm{Top}_k(x))_i = \begin{cases} x_i & \text{if} \quad i \leq \pi_k, \\ 0 & \text{otherwise} \end{cases}.$$

Clearly, Top_k is biased for $k = 1, 2, \ldots, d-1$, and unbiased for $k = d$, in which case it becomes the identity compressor.
Prove that

$$\|\mathrm{Top}_k(x) - x\|^2 \leq \left(1 - \frac{k}{d}\right) \|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

Exercises IV

Exercise 60 (Top_k vs Rand_k compressor)

Fix $1 \leq k \leq d$ and let S be the k-nice sampling. That is, S is a random subset of $\{1, 2, \ldots, d\}$ of cardinality k, chosen uniformly at random. Let $\mathrm{Rand}_k : \mathbb{R}^d \to \mathbb{R}^d$ be defined as follows:

$$(\mathrm{Rand}_k(x))_i = \begin{cases} x_i & \text{if} \quad i \in S, \\ 0 & \text{otherwise} \end{cases}.$$

Prove that

$$\|\mathrm{Top}_k(x) - x\|^2 \leq \mathbf{E}\left[\|\mathrm{Rand}_k(x) - x\|^2\right] \quad \forall x \in \mathbb{R}^d.$$

Exercises V

Exercise 61 (Rank_k compressor)

Sometimes a vector $x \in \mathbb{R}^d$ is more naturally expressed in the form of a matrix. Let $d = a \times b$, and let $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ encode the d entries of x in a matrix form.

- *Let r be the rank of $\boldsymbol{X}$ (necessarily, $r \leq \min\{a, b\}$). This means that $\boldsymbol{X}$ has r nonzero* **singular values** $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. *Let*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

  be the (compact) **singular value decomposition** *of $\boldsymbol{X}$. That is, $\boldsymbol{U} \in \mathbb{R}^{a \times r}$ has orthonormal columns $\{u_i\}$, $\boldsymbol{S} \in \mathbb{R}^{r \times r}$ is diagonal with the singular values on the diagonal, and $\boldsymbol{V}^\top \in \mathbb{R}^{r \times b}$ has orthonormal rows $\{v_i^\top\}$.*

- *Choose $1 \leq k \leq \min\{a, b\}$ and define the Rank_k compressor as follows:*

$$\mathrm{Rank}_k(\boldsymbol{X}) \stackrel{def}{=} \sum_{i=1}^{\min\{k,r\}} \sigma_i u_i v_i^\top .$$

Prove that

$$\|\mathrm{Rank}_k(\boldsymbol{X}) - \boldsymbol{X}\|_F^2 \leq \left(1 - \frac{k}{\min\{a, b\}}\right) \|\boldsymbol{X}\|_F^2 , \quad \forall \boldsymbol{X} \in \mathbb{R}^{a \times b},$$

where $\|\cdot\|_F$ denotes the **Frobenius** *norm.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 12: Variance Reduction via Gradient Shift

Based on:

[19] E. Gorbunov, F. Hanzely and P.R., **A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent**, AISTATS, 2020

[28] F. Hanzely and P.R., **One method to rule them all: variance reduction for data, parameters and many new methods**, arXiv:1905.11266, 2019

Introduction

Introduction I

Problem: As before, we consider solving the general optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, closed and convex.

Algorithm: The method we study is SGD, performing iterations of the form

$$\boxed{x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g(x^k))},$$

where $g : \mathbb{R}^d \to \mathbb{R}^d$ is an unbiased estimator of the gradient, i.e.,

$$\mathbf{E}\left[g(x)\right] = \nabla f(x).$$

Assumptions: We performed a convergence analysis of SGD under the following assumptions:

1. f is μ-convex,

Introduction II

- Recall that we only needed *consequence* (21) of strong convexity, i.e., we only required the following inequality to hold:

$$D_f(x, x^\star) + \frac{\mu}{2}\|x - x^\star\|^2 \leq \langle \nabla f(x) - \nabla f(x^\star), x - x^\star \rangle, \quad \forall x \in \mathbb{R}^d$$

2. the AC assumption holds, i.e., there exists a constant $A \geq 0$ and function $C : \mathbb{R}^d \to [0, \infty)$ such that

$$G(x, y) \overset{\text{def}}{=} \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right] \leq 2AD_f(x, y) + C(y)$$

for all $x, y \in \mathbb{R}^d$.

- Recall that we only need this inequality for $y = x^\star$, i.e., we only need the inequality

$$G(x, x^\star) \leq 2AD_f(x, x^\star) + C(x^\star). \tag{171}$$

Convergence theory: Our main convergence theorem is Theorem 33, which says that the iterates $\{x^k\}_{k \geq 0}$ of SGD satisfy

$$\mathbf{E}\left[\|x^k - x^\star\|^2\right] \leq (1 - \gamma\mu)^k \|x^0 - x^\star\|^2 + \frac{\gamma C(x^\star)}{\mu},$$

provided that the stepsize satisfies $0 < \gamma \leq \frac{1}{A}$.

Nobody Likes the Neighborhood

In today's lecture we will ask the following question:

> In general, $C(x^\star)$ may be positive, which means that SGD converges linearly to a neighborhood of the optimal solution $x^\star$ only. Can something be done about this? **Can we design a gradient estimator for which $C(x^\star) = 0$?**

- We will show that any gradient estimator g can be converted into one for which $C(x^\star) = 0$
- However, this estimator is utterly **impractical** as it depends on $x^\star$, which we do not know.
- In further lectures we will show how this impracticality issue can be lifted.

Designing an Estimator for Which $C(x^\star) = 0$

We described two approaches for establishing the AC inequality:

	Approach 1	Approach 2
Theorem	Thm 34	Thm 36
Assumptions	f is convex and L-smooth	$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y)$
$G(x, y) \leq$	$2LD_f(x, y) + \mathbf{Var}\left[g(x)\right]$	$2(2A'')D_f(x, y) + 2\mathbf{Var}\left[g(y)\right]$
No Neighborhood if	$\mathbf{Var}\left[g(x)\right] \leq 2A' D_f(x, x^\star)$	$\mathbf{Var}\left[g(x^\star)\right] = 0$
Main Use	Compression	Subsampling

Shifted Gradient Estimators for Compression

Shifted Gradient Estimator for Compression

Definition 86 (Shifted Gradient Estimator for Compression)

Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a gradient estimator of the form

$$g(x) = \mathcal{C}(\nabla f(x)),$$

where $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a random mapping satisfying

$$\mathbf{E}\left[\mathcal{C}(s)\right] = s$$

for all s. The **shifted gradient estimator** with shift $h \in \mathbb{R}^d$ is the estimator $g_h : \mathbb{R}^d \to \mathbb{R}^d$ defined as follows:

$$g_h(x) \stackrel{\text{def}}{=} \mathcal{C}(\nabla f(x) - h) + h. \tag{172}$$

Bias and Variance

Lemma 87

The shifted estimator g_h is unbiased for all $h \in \mathbb{R}^d$. If $\mathcal{C} \in \mathbb{B}^d(\omega)$ and f is convex and L-smooth, then for every $y \in \mathbb{R}^d$ we have

$$\mathbf{Var}\left[g_{\nabla f(y)}(x)\right] \leq 2 \underbrace{L\omega}_{=A'} D_f(x, y).$$

Commentary:

- As a result, for the estimator $g_{\nabla f(x^\star)}$ we have

$$G(x, x^\star) \leq 2AD_f(x, x^\star) + C(x^\star)$$

  satisfied with $A = L + L\omega = (\omega + 1)L$ and $C(x^\star) = 0$ (see (171)).
- So, SGD which uses this estimator converges to $x^\star$ at a linear rate. **That is, there is no oscillation neighborhood!**
- This is a new version of CGD which we refer to by the name CGD-SHIFT.

CGD-SHIFT

We have just described the following method.

Algorithm 13 CGD-SHIFT

1: **Parameters:** learning rate $\gamma > 0$, compression operator $\mathcal{C} \in \mathbb{B}^d(\omega)$, starting point $x^0 \in \mathbb{R}^d$, shift $h = \nabla f(x^\star)$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute gradient $\nabla f(x^k)$
4: Compress gradient $g^k = \mathcal{C}\left(\nabla f(x^k) - h\right) + h$
5: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Proof of Lemma 87

The first claim is straightforward:

$$\begin{aligned}
\mathbf{E}\left[g_h(x)\right] &\overset{(172)}{=} \mathbf{E}\left[\mathcal{C}(\nabla f(x) - h) + h\right] \\
&= \mathbf{E}\left[\mathcal{C}(\nabla f(x) - h)\right] + h \\
&= \nabla f(x) - h + h \\
&= \nabla f(x).
\end{aligned}$$

Because of this, we can evaluate the variance for $h = \nabla f(y)$ via

$$\begin{aligned}
\mathbf{Var}\left[g_h(x)\right] &= \mathbf{E}\left[\|g_h(x) - \mathbf{E}\left[g_h(x)\right]\|^2\right] \\
&\overset{(172)}{=} \mathbf{E}\left[\|\mathcal{C}(\nabla f(x) - h) - (\nabla f(x) - h)\|^2\right] \\
&= \mathbf{Var}\left[\mathcal{C}(\nabla f(x) - h)\right] \\
&\leq \omega \|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq 2\omega L D_f(x, y).
\end{aligned}$$

Convergence of CGD-SHIFT

Corollary 88

- *Assume f is L-smooth and μ-convex*
- *Let $\mathcal{C} \in \mathbb{B}^d(\omega)$*
- *Let $A = (\omega + 1)L$*
- *Let the stepsize satisfy $0 < \gamma \leq \frac{1}{A}$*

Then the iterates of CGD-SHIFT *(Algorithm 13) satisfy*

$$\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2.$$

Shifted Compression Operators

Zero is Special

Recall that compression operators $\mathcal{C} \in \mathbb{B}^d(\omega)$ satisfy two properties:

(i) $\mathbf{E}[\mathcal{C}(x)] = x$ for all $x \in \mathbb{R}^d$

(ii) $\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2$ for all $x \in \mathbb{R}^d$

Observe that while property (i) is uniform/democratic across all vectors x, property (ii) is not. Indeed, **the vector $x = 0$ gets a special treatment** because

$$\mathbf{E}\left[\|\mathcal{C}(0) - 0\|^2\right] = 0.$$

This means that

- The zero vector enjoys zero variance when compressed!
- In other words, the zero vector is compressed to 0 with probability 1 (check why!)

Some questions:

- What is special about the zero vector?
- Could other vectors be made special? What would the consequences be?

Shifted Compression Operators

We now generalize the family $\mathbb{B}^d(\omega)$ to allow for operators where other vectors besides 0 can be "special" as well. That is, we allow for **shifts away from the origin.**

Definition 89 (Shifted compression operators)

Let $h \in \mathbb{R}^d$. We say that a randomized mapping $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ belongs to $\mathbb{B}^d(\omega; h)$ if

(i) $\mathbf{E}[\mathcal{C}(x)] = x$ for all $x \in \mathbb{R}^d$

(ii) $\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x - h\|^2$ for all $x \in \mathbb{R}^d$

We call the vector h a **shift,** and $\mathcal{C}$ is a **shifted compression operator.**

Note that

$$\mathbb{B}^d(\omega) \equiv \mathbb{B}^d(\omega; 0)$$

Shifting a Shifted Compression Operator

Lemma 90

Let $\mathcal{C}_h \in \mathbb{B}^d(\omega; h)$ and $v \in \mathbb{R}^d$. Then the randomized mapping $\mathcal{C}$ defined by

$$\mathcal{C}(x) \stackrel{def}{=} v + \mathcal{C}_h(x - v)$$

satisfies $\mathcal{C} \in \mathbb{B}^d(\omega; h + v)$.

Proof.

(i) Unbiasedness:

$$\mathbf{E}[\mathcal{C}(x)] = \mathbf{E}[v + \mathcal{C}_h(x - v)] = v + \mathbf{E}[\mathcal{C}_h(x - v)] = v + x - v = x.$$

(ii) Variance:

$$\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] &= \mathbf{E}\left[\|v + \mathcal{C}_h(x - v) - x\|^2\right] \\
&= \mathbf{E}\left[\|\mathcal{C}_h(x - v) - (x - v)\|^2\right] \\
&\leq \omega\|x - v - h\|^2 \quad = \quad \omega\|x - (h + v)\|^2.
\end{aligned}$$

□

Representation of Shifted Compression Operators

It follows from the previous lemma that all shifted compression operators $\mathcal{C}_h \in \mathbb{B}^d(\omega; h)$ arise by a shift of some operator $\mathcal{C}_0$ from $\mathbb{B}^d(\omega; 0)$. Let us make this more formal.

Theorem 91 (Representation)

Let $\mathcal{C}_h \in \mathbb{B}^d(\omega; h)$. Then there exists $\mathcal{C}_0 \in \mathbb{B}^d(\omega; 0)$ such that

$$\mathcal{C}_h(x) = h + \mathcal{C}_0(x - h). \tag{173}$$

Proof.

Let $\mathcal{C}_0$ be the shifted compression operator we get from $\mathcal{C}_h$ by applying the shift $-h$:

$$\mathcal{C}_0(x) \stackrel{\text{def}}{=} -h + \mathcal{C}_h(x + h). \tag{174}$$

By Lemma 90, we have $\mathcal{C}_0 \in \mathbb{B}^d(\omega; 0)$. Finally, we verify that shifting $\mathcal{C}_0$ by h will get us back to $\mathcal{C}_h$:

$$h + \mathcal{C}_0(x - h) \stackrel{(174)}{=} h + (-h + \mathcal{C}_h(x - h + h)) = \mathcal{C}_h(x).$$

□

AC Inequality in the $\nabla f(x^\star) \neq 0$ Case

Existing proof: We use a compression operator $\mathcal{C} \in \mathbb{B}^d(\omega; 0)$, and hence the extra term involving $\nabla f(x^\star)$ appears.

$$
\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x^\star)\|^2\right] &\overset{(40)}{=} \mathbf{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(*)}{\leq} \omega\,\|\nabla f(x) - 0\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= \omega\,\|\nabla f(x) - \nabla f(x^\star) + \nabla f(x^\star)\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(38)}{\leq} 2\omega\,\|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega\,\|\nabla f(x^\star)\|^2 \\
&\qquad + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= (2\omega + 1)\,\|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega\,\|\nabla f(x^\star)\|^2 \\
&\leq 2(2\omega + 1) L D_f(x, x^\star) + 2\omega\,\|\nabla f(x^\star)\|^2,
\end{aligned}
$$

where in (*) we used the fact that $\mathcal{C} \in \mathbb{B}^d(\omega; 0)$ and in the last step we have used convexity and L-smoothness of f.

AC Inequality in the $\nabla f(x^\star) \neq 0$ Case

New proof: Let us use a compression operator $\mathcal{C} \in \mathbb{B}^d(\omega; h)$ for some $h \neq 0$ instead.

$$
\begin{aligned}
\mathbf{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x^\star)\|^2\right] &\overset{(40)}{=} \mathbf{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(*)}{\leq} \omega \|\nabla f(x) - h\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= \omega \left\|\nabla f(x^k) - \nabla f(x^\star) + \nabla f(x^\star) - h\right\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(38)}{\leq} 2\omega \|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega \|\nabla f(x^\star) - h\|^2 \\
&\qquad + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= (2\omega + 1) \|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega \|\nabla f(x^\star) - h\|^2 \\
&\leq 2(2\omega + 1) L D_f(x, x^\star) + 2\omega \|\nabla f(x^\star) - h\|^2, \qquad (175)
\end{aligned}
$$

where in (*) we used the fact that $\mathcal{C} \in \mathbb{B}^d(\omega; h)$ and in the last step we have used convexity and L-smoothness of f.

Shifted Gradient Estimator for Subsampling

Shifted Gradient Estimator for Subsampling

Definition 92 (Shifted Gradient Estimator for Subsampling)

Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a gradient estimator. The **shifted gradient estimator** with shift $y \in \mathbb{R}^d$ is the estimator g_y defined as follows:

$$g_y(x) \stackrel{\text{def}}{=} g(x) - g(y) + \mathbf{E}\left[g(y)\right]. \tag{176}$$

Lemma 93

If g is unbiased, then the shifted estimator g_y is unbiased for all $y \in \mathbb{R}^d$. Moreover, $\mathbf{Var}\left[g_y(y)\right] = 0$.

Proof.

The first claim is straightforward:

$$\begin{aligned} \mathbf{E}\left[g_y(x)\right] &= \mathbf{E}\left[g(x) - g(y) + \mathbf{E}\left[g(y)\right]\right] = \mathbf{E}\left[g(x)\right] - \mathbf{E}\left[g(y)\right] + \mathbf{E}\left[g(y)\right] \\ &= \mathbf{E}\left[g(x)\right] = \nabla f(x). \end{aligned}$$

The second claim follows from

$$g_y(y) \stackrel{\text{def}}{=} g(y) - g(y) + \mathbf{E}\left[g(y)\right] = \mathbf{E}\left[g(y)\right], \tag{177}$$

which is deterministic, and hence its variance is zero. □

Two Special Shifts

Gradient Descent. The choice $y = x$ gives rise to the shifted estimator

$$g_y(x) = g_x(x) \overset{(177)}{=} g(x) - g(x) + \mathbf{E}\left[g(x)\right] = \nabla f(x),$$

i.e., the actual gradient of f at x.

- **Key issue: expensive to compute**

Shifted SGD**.** The choice $y = x^\star$ gives rise to the shifted estimator

$$g_y(x) = g_{x^\star}(x) \overset{(177)}{=} g(x) - g(x^\star) + \mathbf{E}\left[g(x^\star)\right] = g(x) - g(x^\star) + \nabla f(x^\star).$$

- **Key issue: we do not know the shift $x^\star$**

Shifted SGD

Algorithm 14 SGD-SHIFT

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, optimal point $x^\star$, gradient estimator g

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: Compute shifted stochastic gradient

$$g^k = g_{x^\star}(x^k) = g(x^k) - g(x^\star) + \nabla f(x^\star)$$

4: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Shifted Gradient Estimator: Expected Smoothness

Lemma 94

Assume we have an unbiased gradient estimator g satisfying the expected smoothness inequality

$$\mathbf{E}\left[\|g(x)-g(y)\|^2\right] \leq 2A'' D_f(x,y) \quad \forall x,y \in \mathbb{R}^d.$$

Then the shifted estimator g_y with shift $y \in \mathbb{R}^d$ satisfies the inequality

$$\mathbf{E}\left[\|g_y(x)-g_y(y)\|^2\right] \leq 2A'' D_f(x,y) \quad \forall x,y \in \mathbb{R}^d. \tag{178}$$

Proof.

$$\begin{aligned}
\mathbf{E}\left[\|g_y(x)-g_y(y)\|^2\right] &\overset{(176)+(177)}{=} \mathbf{E}\left[\|g(x)-g(y)+\mathbf{E}\left[g(y)\right]-\mathbf{E}\left[g(y)\right]\|^2\right] \\
&= \mathbf{E}\left[\|g(x)-g(y)\|^2\right] \\
&\leq 2A'' D_f(x,y),
\end{aligned}$$

where the last inequality follows from the expected smoothness assumption. □

Summary: Gradient Estimators Satisfying Expected Smoothness

Recall that expected smoothness inequality is satisfied by many gradient estimators we gave seen so far:

Method	Algorithm #	Estimator $g(x) =$	Expected Smoothness
SGD-NS	Alg 4	$\frac{\nabla f_i(x)}{np_i}$	Lemma 37
SGD-NICE	Alg 7	$\frac{1}{\tau}\sum_{i\in S}\nabla f_i(x)$	Lemma 45
SGD-IND	Alg 8	$\sum_{i\in\bigcup_{i=1}^n S_i}\frac{1}{np_i}\nabla f_i(x)$	Lemma 46
SGD-MULTI	Alg 9	$\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)$	Lemma 47

Exercises

Exercise 62

Design `DCGD-SHIFT` *and formulate and prove an associated convergence theorem.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 13: General Analysis of Variance Reduced SGD

Based on:

[19] E. Gorbunov, F. Hanzely and P.R., **A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent**, AISTATS, 2020

The Problem

We still consider the general regularized problem (6), i.e.,

$$\min_{x \in \mathbb{R}^d} f(x) + R(x).$$

Limited Expressive Power of the AC Inequality

As we have seen, Theorem 33 (which combines Lemma 32 and the **AC inequality**) covers many interesting instances of SGD, including:

- **Gradient Descent:** GD
- **Subsampling:** SGD-NS, SGD-US, SGD-IS, SGD-NICE, SGD-IND, SGD-MULTI, L2GD
- **Compression:** CGD, DCGD, RCD
- **Shift:** SGD-SHIFT, CGD-SHIFT, DCGD-SHIFT

Unless we are lucky and are in the interpolation regime, **only GD and the methods using an optimally shifted gradient operator converge linearly to $x^\star$.** Issues:

- GD has **expensive iterations**
- Shifted methods are **not practical at all** since they rely on the knowledge of $x^\star$ (through $\nabla f(x^\star)$ or $\nabla f_i(x^\star)$ for all i)

We will now switch our attention to so-called **variance-reduced methods,** i.e., methods able to **learn/construct the optimal shift iteratively. However, variance reduced methods can't be captured by the AC inequality: we need a more refined approach.**

Analysis of Generic SGD

Recall the statement of Lemma 32:

Lemma 32
Consider Algorithm 3 (SGD). If f is μ-convex and Assumption 2 (unbiasedness) is satisfied, then for all $\gamma \geq 0$ and $k \geq 0$ we have

$$\begin{aligned} \mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\leq (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) \\ &\quad + \gamma^2 \underbrace{\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right]}_{G^k}. \end{aligned}$$

We have used this lemma together with the AC inequality to prove our main convergence result for SGD: Theorem 33.

Recalling the Proof of Lemma 32 - I

- As in the GD case, let $r^k \stackrel{\text{def}}{=} x^k - x^\star$. Performing the exact same steps as in the case of GD, but replacing the gradient $\nabla f(x^k)$ by the stochastic gradient g^k, we get

$$\begin{aligned}
\left\|r^{k+1}\right\|^2 &= \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - x^\star\right\|^2 \\
&\stackrel{(16)}{=} \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - \mathrm{prox}_{\gamma R}\left(x^\star - \gamma \nabla f(x^\star)\right)\right\|^2 \\
&\stackrel{(15)}{\leq} \left\|(x^k - \gamma g^k) - (x^\star - \gamma \nabla f(x^\star))\right\|^2 \\
&= \left\|x^k - x^\star - \gamma\left(g^k - \nabla f(x^\star)\right)\right\|^2 \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle + \gamma^2\left\|g^k - \nabla f(x^\star)\right\|^2,
\end{aligned}$$

where, as before, the inequality follows from nonexpansiveness of the proximal operator.

Recalling the Proof of Lemma 32 - II

- Next, we compute expectation of both sides of the above inequality, conditional on x^k. This gives

$$\begin{aligned}
\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\leq \left\|r^k\right\|^2 - 2\gamma\mathbf{E}\left[\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle \mid x^k\right] + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \mathbf{E}\left[g^k \mid x^k\right] - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right] \\
&\overset{(45)}{=} \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right],
\end{aligned}$$

  where in the second step we used linearity of expectation and the fact that r^k is a constant vector conditioned on x^k, and in the last step we used the unbiasedness assumption.

- As in the GD case, we again use μ-convexity to bound

$$\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle \overset{(21)}{\geq} D_f(x^k, x^\star) + \frac{\mu}{2}\left\|x^k - x^\star\right\|^2,$$

  which leads to

$$\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] \leq (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k\right],$$

  which is what we wanted to prove.

Stochastic Gradient with Control

Stochastic Gradient with Control - I

- So far we have assumed that the stochastic gradient has the simple structure $g(x)$, where $g : \mathbb{R}^d \to \mathbb{R}^d$ is a random mapping such that

$$\mathbf{E}\left[g(x)\right] = \nabla f(x), \qquad \forall x \in \mathbb{R}^d.$$

- From now on, we will want to apply **adaptive control** to construct the stochastic gradient in an iterative manner in order to reduce its variance. In particular, we now consider stochastic gradient of the form

$$g(x, \xi),$$

where $g : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ (for some m, typically a multiple of d) is a random mapping, and $\xi \in \mathbb{R}^m$ **is a control variable.** We assume that

$$\mathbf{E}\left[g(x, \xi)\right] = \nabla f(x) \qquad \forall x \in \mathbb{R}^d, \forall \xi \in \mathbb{R}^m. \tag{179}$$

  - In practice, we let

$$g^k = g(x^k, \xi^k),$$

where ξ^k **is a suitably constructed random variable.**

Stochastic Gradient with Control - II

- Unbiasedness assumption (179) can be written in the form

$$\mathbf{E}\left[g(x^k, \xi^k) \mid x^k, \xi^k\right] = \nabla f(x^k).$$

Let us state the unbiasedness assumption formally:

Assumption 6 (Unbiasedness of Stochastic Gradient with Control)

$g(x, \xi)$ *is an unbiased estimator, i.e.,* (179) *holds.*

SGD-CTRL: SGD with Control

Algorithm 15 SGD-CTRL

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, starting control $\xi^0 \in \mathbb{R}^m$

for $k = 0, 1, 2, \ldots$ **do**

  Construct an **unbiased estimator** $g^k \overset{\text{def}}{=} g(x^k, \xi^k)$ of the gradient

  $x^{k+1} = \operatorname{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

  Construct new control ξ^{k+1}

SGD-CTRL: First Lemma

The following is a variant of Lemma 32 suitable for SGD-CTRL:

Lemma 95

Consider Algorithm 15 (SGD-CTRL). If f is μ-convex and Assumption 6 (unbiasedness) is satisfied, then for all $\gamma \geq 0$ and $k \geq 0$ we have

$$\begin{aligned} \mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k, \xi^k\right] &\leq (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) \\ &\quad + \gamma^2 \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right]. \end{aligned} \tag{180}$$

Proof of Lemma 95 - I

- Let $r^k \overset{\text{def}}{=} x^k - x^\star$. Performing the exact same steps as in the case of SGD, we get

$$\begin{aligned}
\left\|r^{k+1}\right\|^2 &= \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - x^\star\right\|^2 \\
&\overset{(16)}{=} \left\|\mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right) - \mathrm{prox}_{\gamma R}\left(x^\star - \gamma \nabla f(x^\star)\right)\right\|^2 \\
&\overset{(15)}{\leq} \left\|(x^k - \gamma g^k) - (x^\star - \gamma \nabla f(x^\star))\right\|^2 \\
&= \left\|x^k - x^\star - \gamma\left(g^k - \nabla f(x^\star)\right)\right\|^2 \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle + \gamma^2\left\|g^k - \nabla f(x^\star)\right\|^2,
\end{aligned}$$

where, as before, the inequality follows from nonexpansiveness of the proximal operator.

Proof of Lemma 95 - II

- Now comes a slight change in the argument. We compute expectation of both sides of the above inequality, conditional on x^k and on ξ^k. This gives

$$\begin{aligned}
\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k, \xi^k\right] &\leq \left\|r^k\right\|^2 - 2\gamma\mathbf{E}\left[\left\langle r^k, g^k - \nabla f(x^\star)\right\rangle \mid x^k, \xi^k\right] + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right] \\
&= \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \mathbf{E}\left[g^k \mid x^k, \xi^k\right] - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right] \\
&\overset{(179)}{=} \left\|r^k\right\|^2 - 2\gamma\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right],
\end{aligned}$$

  where in the second step we used linearity of expectation and the fact that r^k is a constant vector conditioned on x^k and ξ^k, and in the last step we used the unbiasedness assumption.

- We now use μ-convexity to bound

$$\left\langle r^k, \nabla f(x^k) - \nabla f(x^\star)\right\rangle \overset{(21)}{\geq} D_f(x^k, x^\star) + \frac{\mu}{2}\left\|x^k - x^\star\right\|^2,$$

  which leads to

$$\mathbf{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k, \xi^k\right] \leq (1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f(x^k, x^\star) + \gamma^2\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right],$$

  which is what we wanted to prove.

σ^k Assumption: A More General Model for the Evolution of the Stochastic Gradients

We now introduce an assumption on the stochastic gradients $\{g^k = g(x^k, \xi^k)\}_{k\geq 0}$ generated by SGD-CTRL which describes **the key properties of control process.**

Assumption 7 (Gorbunov, Hanzely & R 2019 [19])

Let $\{x^k, \xi^k\}_{k\geq 0}$ be the random iterates produced by SGD-CTRL *and assume that there exists a function $\sigma : \mathbb{R}^m \to [0, \infty)$ and non-negative constants $A, B, C, \tilde{A}, \tilde{B}, \tilde{C}$, with $\tilde{B} < 1$ such that the following two relations hold:*

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right] \leq 2AD_f(x^k, x^\star) + B\sigma^k + C, \tag{181}$$

and

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, \xi^k\right] \leq 2\tilde{A}D_f(x^k, x^\star) + \tilde{B}\sigma^k + \tilde{C}, \tag{182}$$

where $\sigma^k \stackrel{def}{=} \sigma(\xi^k)$.

Variance-reduced Methods

Definition 96 (Variance-reduced methods)

We say that SGD-CTRL with stochastic gradients satisfying Assumption 7 is variance reduced if $C = \tilde{C} = 0$.

- This is the **first quantitative definition of what it means for an SGD method to be variance-reduced.**
- We shall see (Theorem 97) that **variance-reduced methods converge to the solution, i.e., $x^k \to x^\star$,** whereas **non-variance reduced methods only converge to a neighbourhood of the solution.**
- Meaning of $\{\sigma^k\}_{k\geq 0}$: the sequence encodes the progress of the **variance-reduction process** employed.
  - If no variance reduction process is employed, then one has $\sigma^k \equiv 0$. In this case, one can set the parameters $B, \tilde{A}, \tilde{B}$ and $\tilde{C}$ to zero:

$$B = \tilde{A} = \tilde{B} = \tilde{C} = 0. \tag{183}$$

    This eliminates the second recursion, which is not necessary.
  - All variants of SGD we have considered so far are in this category.

Analysis of SGD-CTRL

Main Result

Theorem 97 (Gorbunov, Hanzely & R 2019 [19])

Assume that f is μ-convex, and that the stochastic gradient g^k is unbiased for all k and satisfies the σ^k assumption (Assumption 7). For any $M > \frac{B}{1-\tilde{B}}$ choose a stepsize γ satisfying

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + M\tilde{A}}\right\}. \tag{184}$$

Then the iterates $\{x^k, \sigma^k\}_{k\geq 0}$ of `SGD-CTRL` *satisfy*

$$\mathbf{E}\left[V^k\right] \leq \max\left\{(1-\gamma\mu)^k, \left(\frac{B + M\tilde{B}}{M}\right)^k\right\} V^0 + \frac{(C + M\tilde{C})\gamma^2}{\min\left\{\gamma\mu, 1 - \frac{B+M\tilde{B}}{M}\right\}}, \tag{185}$$

where the **Lyapunov function** *V^k is defined by*

$$V^k \stackrel{def}{=} \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k. \tag{186}$$

In the special case when $B = 0$, we can choose $M = 0$ and the expression $\frac{B+M\tilde{B}}{M}$ should be interpreted as zero.

Main Result: Commentary

- If the AC inequality holds, we have $B = \tilde{A} = \tilde{B} = \tilde{C} = 0$, and can choose $M = 0$, which means that we recover Theorem 33.
- Note that if $C = \tilde{C} = 0$ (this is the variance reduced setting), the Lyapunov function (and hence also $\left\|x^k - x^\star\right\|^2$) **converges linearly to zero.**
- Otherwise, `SGD-CTRL` converges linearly to a certain neighborhood of the solution whose size is determined by the size of the parameters C and $\tilde{C}$.

A Key Lemma I

For simplicity, let

$$d^k \stackrel{\text{def}}{=} \left\|x^k - x^\star\right\|^2.$$

The following lemma will be used in the proof of Theorem 97.

Lemma 98 (Key single iteration recurrence)

Consider the setting of Theorem 97. Then the following inequality holds for any $M \geq 0$ and all $k \geq 0$:

$$\begin{aligned}
&\mathbf{E}\left[d^{k+1}\right] + M\gamma^2\mathbf{E}\left[\sigma^{k+1}\right] + 2\gamma\left(1 - \gamma(A + M\tilde{A})\right)\mathbf{E}\left[D_f(x^k, x^\star)\right] \\
&\leq (1 - \gamma\mu)\mathbf{E}\left[d^k\right] + \left(B + M\tilde{B}\right)\gamma^2\mathbf{E}\left[\sigma^k\right] + (C + M\tilde{C})\gamma^2. \qquad (187)
\end{aligned}$$

Remark: In the regime when the σ^k inequality reduced to the AC inequality (i.e., if $B = \tilde{A} = \tilde{B} = \tilde{C} = \sigma^k = 0$), inequality (187) reduces to (we choose $M = 0$)

$$\mathbf{E}\left[d^{k+1}\right] \quad \leq \quad (1 - \gamma\mu)\mathbf{E}\left[d^k\right] - 2\gamma\left(1 - \gamma A\right)\mathbf{E}\left[D_f(x^k, x^\star)\right] + C\gamma^2,$$

which is an inequality that was establish during the proof of Theorem 33 .

Proof of Lemma 98 - I

Let $r^k = x^k - x^\star$, so that $d^k = \left\|r^k\right\|^2$. In view of Lemma 95, we have

$$\mathbf{E}\left[d^{k+1} \mid x^k, \xi^k\right] \leq (1-\gamma\mu)d^k - 2\gamma D_f(x^k, x^\star) + \gamma^2 \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right].$$

On the other hand, the first relation in the σ^k assumption (see (181)) states that

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, \xi^k\right] \leq 2AD_f(x^k, x^\star) + B\sigma^k + C.$$

Combining the above inequalities, we get

$$\mathbf{E}\left[d^{k+1} \mid x^k, \xi^k\right] \leq (1-\gamma\mu)d^k + 2\gamma\left(A\gamma - 1\right)D_f(x^k, x^\star) + B\gamma^2\sigma^k + \gamma^2 C. \tag{188}$$

Applying the tower property of conditional expectation, we get

$$\begin{aligned}\mathbf{E}\left[d^{k+1}\right] &= \mathbf{E}\left[\mathbf{E}\left[d^{k+1} \mid x^k, \xi^k\right]\right] \\ &\overset{(188)}{\leq} (1-\gamma\mu)\mathbf{E}\left[d^k\right] + 2\gamma\left(A\gamma - 1\right)\mathbf{E}\left[D_f(x^k, x^\star)\right] \\ &\quad + B\gamma^2\mathbf{E}\left[\sigma^k\right] + \gamma^2 C. \end{aligned} \tag{189}$$

Next, the second relation in the in the σ^k assumption (see (182)) states that

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, \xi^k\right] \leq 2\tilde{A}D_f(x^k, x^\star) + \tilde{B}\sigma^k + \tilde{C}. \tag{190}$$

Proof of Lemma 98 - II

Applying the tower property of conditional expectation, this leads to

$$\begin{aligned}\mathbf{E}\left[\sigma^{k+1}\right] &= \mathbf{E}\left[\mathbf{E}\left[\sigma^{k+1} \mid x^k, \xi^k\right]\right] \\ &\overset{(190)}{\leq} 2\tilde{A}\mathbf{E}\left[D_f(x^k, x^\star)\right] + \tilde{B}\mathbf{E}\left[\sigma^k\right] + \tilde{C}. \end{aligned} \tag{191}$$

In order to bound $\mathbf{E}\left[V^{k+1}\right] \overset{\text{def}}{=} \mathbf{E}\left[d^{k+1}\right] + M\gamma^2\mathbf{E}\left[\sigma^{k+1}\right]$, we combine inequalities (189) and (191):

$$\begin{aligned}\mathbf{E}\left[V^{k+1}\right] &\overset{(189)+(191)}{\leq} (1-\gamma\mu)\mathbf{E}\left[d^k\right] + 2\gamma\left(A\gamma - 1\right)\mathbf{E}\left[D_f(x^k, x^\star)\right] + B\gamma^2\mathbf{E}\left[\sigma^k\right] \\ &\quad + M\tilde{B}\gamma^2\mathbf{E}\left[\sigma^k\right] + 2M\tilde{A}\gamma^2\mathbf{E}\left[D_f(x^k, x^\star)\right] + (C + M\tilde{C})\gamma^2 \\ &= (1-\gamma\mu)\mathbf{E}\left[d^k\right] + \left(B + M\tilde{B}\right)\gamma^2\mathbf{E}\left[\sigma^k\right] \\ &\quad + 2\gamma\left(\gamma(A + M\tilde{A}) - 1\right)\mathbf{E}\left[D_f(x^k, x^\star)\right] + (C + M\tilde{C})\gamma^2 .\end{aligned}$$

It remains to rearrange the terms.

Proof of Theorem 97

Proof of Theorem 97

Note first that due to the stepsize restriction (184) we have

$$2\gamma\left(1-\gamma(A+M\tilde{A})\right)\mathbf{E}\left[D_f(x^k,x^\star)\right]>0,$$

and hence we can omit this term. Hence, Lemma 98 simplifies to

$$\begin{aligned}
\mathbf{E}\left[V^{k+1}\right] &\leq (1-\gamma\mu)\mathbf{E}\left[d^k\right]+\left(B+M\tilde{B}\right)\gamma^2\mathbf{E}\left[\sigma^k\right]+(C+M\tilde{C})\gamma^2\\
&= (1-\gamma\mu)\mathbf{E}\left[d^k\right]+\left(\frac{B}{M}+\tilde{B}\right)M\gamma^2\mathbf{E}\left[\sigma^k\right]+(C+M\tilde{C})\gamma^2\\
&\leq \underbrace{\max\left\{(1-\gamma\mu),\left(\frac{B}{M}+\tilde{B}\right)\right\}}_{\overset{\text{def}}{=}\theta}\underbrace{\left(\mathbf{E}\left[d^k\right]+M\gamma^2\mathbf{E}\left[\sigma^k\right]\right)}_{\mathbf{E}[V^k]}+(C+M\tilde{C})\gamma^2\\
&= \theta\mathbf{E}\left[V^k\right]+(C+M\tilde{C})\gamma^2.
\end{aligned}$$

Unrolling this recurrence, we get

$$\begin{aligned}
\mathbf{E}\left[V^k\right] \leq \theta^k V^0+(C+M\tilde{C})\gamma^2\sum_{s=0}^{k-1}\theta^s &\leq \theta^k V^0+(C+M\tilde{C})\gamma^2\sum_{s=0}^{\infty}\theta^s\\
&= \theta^k V^0+(C+M\tilde{C})\gamma^2\frac{1}{1-\theta}\\
&\leq \theta^k V^0+\frac{(C+M\tilde{C})\gamma^2}{\min\left\{\gamma\mu,1-\frac{B}{M}-\tilde{B}\right\}}.
\end{aligned}$$

Exercises

Exercise 63

Theorem 97 states that in the special case when $B = 0$*, we can choose* $M = 0$ *and the expression* $\frac{B+M\tilde{B}}{M}$ *should be interpreted as zero. Prove this.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 14: Loopless SVRG

Based on:

[29] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. **Variance reduced stochastic gradient descent with neighbors.** NIPS 2015

[40] D. Kovalev, S. Horváth and P. R. **Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop**, ALT 2020

Motivation

- So far we have seen a few variance-reduced SGD methods, **all with serious issues:**
  - GD: this method needs to perform n gradient computations in each iteration, which is **inefficient.**
  - SGD-US: in the special **lucky** case when $\mathbf{Var}\left[g(x^\star)\right] = 0$
  - SGD-SHIFT: this method is practically **not implementable** since we do not know $x^\star$
- **We will now describe and analyze the first variance-reduced method which is both efficient and implementable:**

Loopless SVRG (L-SVRG)

- L-SVRG was independently developed by Hofmann et al [29] (earlier) and Kovalev, Horváth & R [40].
  - The name L-SVRG was coined in [40]
  - Here we follow the general analysis in [19] (Theorem 97) to get a result that is more general than that in [40]
- L-SVRG is a **"loopless" variant of the famous SVRG method** of Johnson and Zhang [35]. See also S2GD [39] which appeared in the same year (2013).

The Glory and Plight of Shifted Gradient Estimators

Recall from (176) that given a gradient estimator g, the **shifted gradient estimator** with shift $y \in \mathbb{R}^d$ is defined by

$$g_y(x) = g(x) - g(y) + \mathbf{E}\left[g(y)\right].$$

- g_y is unbiased if g is unbiased
- $\mathbf{Var}\left[g_y(y)\right] = 0$ for all $y \in \mathbb{R}^d$
- shifted gradient estimator is a special case of the **controlled gradient estimator** $g(x, \xi)$ with control $\xi = y$

Special cases:

Method	Shift y	$g_y(x) = g(x) - g(y) + \nabla f(y)$	Issue
GD	x	$g(x) - g(x) + \nabla f(x)$	expensive estimator
SGD-SHIFT	$x^\star$	$g(x) - g(x^\star) + \nabla f(x^\star)$	we do not know $x^\star$

L-SVRG: SGD with Learned Shift

To resolve the issues with GD and SGD-SHIFT, we now aim to design an SGD method able to **learn the optimal shift, i.e., one able to construct a sequence $y^k \to x^\star$.**

Algorithm 16 Loopless SVRG (L-SVRG)

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, starting control variable (shift) $y^0 = x^0$, gradient estimator g
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad g^k = g(x^k) - g(y^k) + \nabla f(y^k)$
4: $\quad x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$
5: $\quad$ Update control variable (shift):

$$y^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ y^k & \text{with probability} \quad 1-p \end{cases}$$

Analysis of L-SVRG

L-SVRG and σ^k Assumption

We now show that L-SVRG satisfies Assumption 7.

Lemma 99

Assume the gradient estimator g is unbiased and satisfies the expected smoothness bound

$$\sigma(x) \overset{def}{=} \mathbf{E}\left[\|g(x) - g(x^\star)\|^2\right] \leq 2A'' D_f(x, x^\star).$$

Then the iterates $\{x^k, y^k\}_{k\geq 0}$ of L-SVRG *satisfy the relations*

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, y^k\right] \leq 2\underbrace{2A''}_{A} D_f(x^k, x^\star) + \underbrace{2}_{B}\sigma^k, \tag{192}$$

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, y^k\right] \leq 2\underbrace{pA''}_{\tilde{A}} D_f(x^k, x^\star) + \underbrace{(1-p)}_{\tilde{B}}\sigma^k, \tag{193}$$

where $\sigma^k \overset{def}{=} \sigma(y^k)$.

Remark: Since $C = \tilde{C} = 0$, L-SVRG is a **variance-reduced** variant of SGD.

Proof of Lemma 99

All expectations in the first part of the proof are conditioned on x^k and y^k. Since $g^k = g(x^k) - g(y^k) + \nabla f(y^k)$, we get:

$$
\begin{aligned}
\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, y^k\right] &= \mathbf{E}\left[\left\|g(x^k) - g(x^\star) + g(x^\star) - g(y^k) + \nabla f(y^k) - \nabla f(x^\star)\right\|^2\right] \\
&\overset{(38)}{\leq} 2\mathbf{E}\left[\left\|g(x^k) - g(x^\star)\right\|^2\right] \\
&\quad + 2\mathbf{E}\left[\left\|g(x^\star) - g(y^k) - \left(\nabla f(x^\star) - \nabla f(y^k)\right)\right\|\right]^2 \\
&\leq 2\mathbf{E}\left[\left\|g(x^k) - g(x^\star)\right\|^2\right] + 2\mathbf{E}\left[\left\|g(y^k) - g(x^\star)\right\|^2\right] \\
&\leq 4A'' D_f(x^k, x^\star) + 2\sigma(y^k).
\end{aligned}
$$

The second relation can be obtained directly from the definition of σ^{k+1} and the update rule for y^k. Indeed, since

$$\sigma^{k+1} = \sigma(y^{k+1})$$

and $y^{k+1} = x^k$ with probability p and $y^{k+1} = y^k$ with probability $1 - p$, we have

$$
\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, y^k\right] &= p\sigma(x^k) + (1-p)\sigma(y^k) \\
&\leq 2pA'' D_f(x^k, x^\star) + (1-p)\sigma(y^k).
\end{aligned}
$$

Convergence of L-SVRG

This follows from our main result for analyzing SGD-CTRL (Theorem 97):

Corollary 100 (Convergence of L-SVRG)

Assume f is μ-convex, and that the gradient estimator g is unbiased and satisfies the expected smoothness bound

$$\sigma(x) \overset{def}{=} \mathbf{E}\left[\|g(x) - g(x^\star)\|^2\right] \leq 2A'' D_f(x, x^\star).$$

Then L-SVRG *with stepsize $\gamma = \frac{1}{6A''}$ satisfies*

$$\mathbf{E}\left[V^k\right] \leq \left(1 - \min\left\{\frac{\mu}{6A''}, \frac{p}{2}\right\}\right)^k V^0, \tag{194}$$

where

$$V^k \overset{(186)}{=} \|x^k - x^\star\|^2 + M\gamma^2\sigma^k, \quad M = \frac{4}{p}, \quad \sigma^k = \sigma(y^k) \overset{def}{=} \mathbf{E}\left[\|g(y^k) - g(x^\star)\|^2\right].$$

So,

$$k \geq \max\left\{\frac{6A''}{\mu}, \frac{2}{p}\right\} \log\frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[V^k\right] \leq \varepsilon V^0.$$

Proof of Corollary 100

- In view of Lemma 99, the σ^k assumption is satisfied with
$$A = 2A'', \quad B = 2, \quad C = 0, \quad \tilde{A} = pA'', \quad \tilde{B} = 1 - p, \quad \tilde{C} = 0.$$
- We choose $M = \frac{4}{p} > \frac{2}{p} = \frac{B}{1-\tilde{B}}$, and hence
$$A + M\tilde{A} = 2A'' + \frac{4}{p}pA'' = 6A''.$$
Moreover,
$$\frac{B + M\tilde{B}}{M} = \frac{p}{2} + 1 - p = 1 - \frac{p}{2}.$$
- Theorem 97 says that the stepsize has to satisfy
$$\gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + M\tilde{A}}\right\} = \min\left\{\frac{1}{\mu}, \frac{1}{6A''}\right\} = \frac{1}{6A''},$$
where the last inequality follows from the observation that $\mu \leq A''$.
- Since $C = \tilde{C}$, and using the largest possible stepsize $\gamma = \frac{1}{6A''}$, Theorem 97 says that
$$\begin{aligned} \mathbf{E}\left[V^k\right] &\leq \max\left\{(1-\gamma\mu)^k, \left(\frac{B + M\tilde{B}}{M}\right)^k\right\} V^0 \\ &= \max\left\{\left(1 - \frac{\mu}{6A''}\right)^k, \left(1 - \frac{p}{2}\right)^k\right\} V^0 \\ &= \left(1 - \min\left\{\frac{\mu}{6A''}, \frac{p}{2}\right\}\right)^k V^0. \end{aligned}$$

L-SVRG-US

L-SVRG-US

- The name **Loopless SVRG (L-SVRG)** was coined by Kovalev, Horváth and R. [40]
- The only considered L-SVRG in the special case when g is the SGD-US estimator, i.e.,

$$g(x) = \nabla f_i(x) \quad \text{with probability} \quad \frac{1}{n}$$

- Recall that the SGD-US estimator satisfies the expected smoothness inequality

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y)$$

with $A'' = \max_i L_i = L_{\max}$.

L-SVRG-US: The Algorithm

Algorithm 17 Loopless SVRG with Uniform Sampling (L-SVRG-US) [40]

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, starting control variable (shift) $y^0 = x^0$, probability $0 < p \leq 1$

for $k = 0, 1, 2, \ldots$ **do**

  Sample $i^k = i \in \{1, 2, \ldots, n\}$ with probability $\frac{1}{n}$

  $g^k = \nabla f_i(x^k) - \nabla f_i(y^k) + \nabla f(y^k)$

  $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

  Update control variable (shift):

$$y^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ y^k & \text{with probability} \quad 1-p \end{cases}$$

L-SVRG-US: Cost per Iteration

In each iteration, one has to compute

- 1 gradient: $\nabla f_i(x^k)$ (always)
- n gradients: $\nabla f(y^k) = \frac{1}{n}\sum_i \nabla f_i(y^k)$ (with probability p)
- 1 gradient: $\nabla f_i(y^k)$ (with probability $1-p$)

So, the expected # of gradient evaluations per 1 iteration of L-SVRG is:

$$\begin{aligned} Cost &= 1 + pn + (1-p)1 \\ &= 2 + p(n-1). \end{aligned}$$

- A practical (and theoretically optimal) choice is $p = \frac{1}{n}$, for which

$$Cost \leq 3 = \mathcal{O}(1). \tag{195}$$

- Another good (and theoretically optimal) choice is $p = \frac{\mu}{L_{\max}}$, for which

$$Cost = \mathcal{O}\left(1 + \frac{n\mu}{L_{\max}}\right). \tag{196}$$

Total Complexity

Number of iterations: In view of Corollary 100, we have

$$k \geq \max\left\{\frac{6L_{\max}}{\mu}, \frac{2}{p}\right\}\log\frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[V^k\right] \leq \varepsilon V^0.$$

Cost of 1 iteration:

- If we choose $p = \frac{1}{n}$, in view of (195) the expected cost of 1 iteration is $Cost \leq 3 = \mathcal{O}(1)$. Hence, the total complexity of `L-SVRG-US` is

$$\text{Total complexity} = \mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}, n\right\}\log\frac{1}{\varepsilon}\right) \tag{197}$$

- If we choose $p = \frac{\mu}{L_{\max}}$, in view of (196) the expected cost of 1 iteration is $Cost = \mathcal{O}\left(1 + \frac{n\mu}{L_{\max}}\right)$. Hence, the total complexity of `L-SVRG-US` is

$$\text{Total complexity} = \mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}, n\right\}\log\frac{1}{\varepsilon}\right) \tag{198}$$

Comparison of Total Complexities

Let us compare the complexities of L-SVRG-US and SGD-US. In the case of SGD-US we will assume that $\mathbf{Var}\left[g(x^\star)\right] = 0$, so as to force the methods to converge to the solution $x^\star$.

Method	GD	SGD-US for $\mathbf{Var}\left[g(x^\star)\right] = 0$	L-SVRG-US $p = \mathcal{O}(\frac{1}{n})$ or $p = \mathcal{O}(\frac{\mu}{L_{\max}})$
Total complexity	$\frac{nL}{\mu} \log \frac{1}{\varepsilon}$	$\frac{L_{\max}}{\mu} \log \frac{1}{\varepsilon}$	$\max\left\{\frac{L_{\max}}{\mu}, n\right\} \log \frac{1}{\varepsilon}$

- **Big data regime:** $n \geq \frac{L_{\max}}{\mu}$
  - L-SVRG-US is slower than SGD-US
  - But one needs to be very lucky for $\mathbf{Var}\left[g(x^\star)\right] = 0$ to be the case!
  - If $\mathbf{Var}\left[g(x^\star)\right] > 0$, then SGD-US converges to a neighbourhood of the solution only!
- **Small data regime:** $n \leq \frac{L_{\max}}{\mu}$
  - L-SVRG-US and SGD-US are equally fast

Insight

Insight: When we are (very) lucky to be in the $\mathbf{Var}\left[g(x^\star)\right] = 0$ regime, variance reduction is not needed in the first place, and SGD-US is better than L-SVRG-US in the big data regime. They are the same in the small data regime. However, **when variance reduction is desirable ($\mathbf{Var}\left[g(x^\star)\right] > 0$), then L-SVRG-US shines: it delivers on its promise and is an efficient method which converges as fast (in the small data regime) or almost as fast (in the big data regime) to $x^\star$ as if variance reduction was not needed at all.**

Extensions

- **Arbitrary sampling:** `L-SVRG` with arbitrary sampling (i.e., a subset of $\{1, 2, \ldots, n\}$ can be sampled to form g^k in each step, following an arbitrary distribution) was developed and analyzed by Qu, Qian and R [58].
- **Acceleration:** An accelerated variant (`L-Katyusha`) was also developed in [58]. Katyusha was developed by Alen-Zhu [2].
- **Beyond strong convexity:** An analysis in the convex and smooth nonconvex regimes can be found in [58].

Exercises I

Exercise 64

Show that the L-SVRG *gradient estimator is unbiased, i.e.,*

$$\mathbf{E}\left[g^k \mid x^k, y^k\right] = \nabla f(x^k).$$

Exercise 65

Consider L-SVRG *using the gradient estimator* $g(x) = \nabla f(x) + \xi$*, where* ξ *is a zero-mean random vector.*

(i) *Show that this method reduces to* GD *for any choice of* p*.*

(ii) *Formulate a convergence theorem for this method by applying Corollary 100. How does the rate compare to* GD*'s rate* $O(\frac{L}{\mu}\log\frac{1}{\varepsilon})$*? What role does* p *play here?*

Exercise 66

Finish the proof of Corollary 100 by showing that $\mu \leq A''$*.*

Exercises II

Exercise 67

Formulate and prove a convergence theorem for `L-SVRG-NS`*: variant of* `L-SVRG` *using the* `SGD-NS` *estimator*

$$g(x) = \frac{\nabla f_i(x)}{np_i},$$

where p_i are positive probabilities summing up to 1.

Exercises III

Exercise 68

Let $x^\star \in \mathbb{R}^d$ *and let* $\{x^k\}_{k\geq 0}$ *be a random sequence for each* k *satisfying*

$$\mathbf{E}\left[\|x^{k+1} - x^\star\|^2 \mid x^k\right] \leq \alpha \|x^k - x^\star\|^2,$$

where $0 < \alpha < 1$. *Fix* $0 < p \leq 1$ *and define a derived sequence* $\{y^k\}_{k\geq 0}$ *as follows*

$$y^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ y^k & \text{with probability} \quad 1-p \end{cases}.$$

(i) *Prove that* $\mathbf{E}\left[\|y^k - x^\star\|^2\right] \to 0$ *at a linear rate, and find an expression for the rate.*

(ii) *What is the best rate as a function of* α *and* p*?*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 15: Compression with Learned Shift (DIANA)

Based on:

[47] K. Mishchenko, E. Gorbunov, M. Takáč and P. R. **Distributed Learning with Compressed Gradient Differences**, arXiv:1901.09269, 2019

[31] S. Horváth, D. Kovalev, K. Mishchenko, P. R. and S. Stich, **Stochastic Distributed Learning with Gradient Quantization and Variance Reduction**, arXiv:1904.05115, 2019

The Problem

We are solving the problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x),$$

where

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Goal of this Lecture: Develop a **practical** variance reduction scheme for

- compressed gradient descent (`CGD`)
- distributed compressed gradient descent (`DCGD`)

CGD-SHIFT

CGD-SHIFT is CGD employing the shifted compression operator

$$\mathcal{C}_h(x) \stackrel{\text{def}}{=} h + \mathcal{C}(x - h),$$

where

- $\mathcal{C} \in \mathbb{B}^d(\omega)$ for some $\omega \geq 0$, and
- h is the **optimal shift** $h = \nabla f(x^\star)$.

CGD-SHIFT:

$$\begin{aligned} g^k &= h + \mathcal{C}(\nabla f(x^k) - h) \\ x^{k+1} &= \operatorname{prox}_{\gamma R}\left(x^k - \gamma g^k\right) \end{aligned}$$

ISSUE: The optimal shift $h = \nabla f(x^\star)$ is **unknown,** and hence CGD-SHIFT is impractical.

DCGD-SHIFT

Recall DCGD-SHIFT: this is DCGD on each node i employing the shifted compression operator

$$\mathcal{C}_{h_i}(x) \stackrel{\text{def}}{=} h_i + \mathcal{C}_i(x - h_i),$$

where

- $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ for some $\omega_i \geq 0$, and
- h_i is the **optimal shift** $h_i = \nabla f_i(x^\star)$.

DCGD-SHIFT:

$$\begin{aligned} g^k &= \frac{1}{n}\sum_{i=1}^n \left[h_i + \mathcal{C}_i(\nabla f_i(x^k) - h_i)\right] \\ x^{k+1} &= \operatorname{prox}_{\gamma R}\left(x^k - \gamma g^k\right) \end{aligned}$$

ISSUE: The optimal shifts $h_i = \nabla f_i(x^\star)$ are **unknown,** and hence DCGD-SHIFT is impractical.

Challenges with Learning the Optimal Shifts

Challenges with Learning the Optimal Shifts - I

- For each node $i \in \{1, 2, \ldots, n\}$ we will produce a (random) sequence $\{h_i^k\}$ of shifts converging to the optimal shifts $\nabla f_i(x^\star)$:

$$h_i^k \to \nabla f_i(x^\star).$$

  That is, we will **iteratively learn the unknown optimal shifts**

$$\nabla f_1(x^\star), \nabla f_2(x^\star), \ldots, \nabla f_n(x^\star).$$

- In iteration k, node i applies the shifted compression operator with shift h_i^k to the local gradient $\nabla f_i(x^k)$, producing the local gradient estimator

$$g_i^k \stackrel{\text{def}}{=} h_i^k + \mathcal{C}_i(\nabla f_i(x^k) - h_i^k). \tag{199}$$

- How can this be done?
  - Each node i sends the compressed vector $\mathcal{C}_i(\nabla f_i(x^k) - h_i^k)$ to the master.
  - However, we **obviously DO NOT want node i to send the uncompressed vector h_i^k** to the master!

Challenges with Learning the Optimal Shifts - II

- Still, the master is **supposed to** compute the overall gradient estimator

$$g^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n g_i^k \overset{(199)}{=} \frac{1}{n}\sum_{i=1}^n \left[h_i^k + \mathcal{C}_i(\nabla f_i(x^k) - h_i^k)\right]$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^n h_i^k}_{g_1^k} + \underbrace{\frac{1}{n}\sum_{i=1}^n \mathcal{C}_i(\nabla f_i(x^k) - h_i^k)}_{g_2^k}$$

  and send g^k (or possibly a compressed version thereof) back to the nodes.

- Note that g^k consists of **two components:**
  - g_1^k: **it is not clear how the master can compute this** since the shifts h_i^k, computed by the nodes, **are uncompressed vectors,** which is why **we do not want to send them to the master!**
  - g_2^k: **can be computed by the master** from the compressed messages sent by the nodes to the master
- So, when designing the sequences $\{h_1^k\}_{k\geq 0}, \{h_2^k\}_{k\geq 0}, \ldots, \{h_n^k\}_{k\geq 0}$, we need to address **two challenging issues at the same time:**

Challenges with Learning the Optimal Shifts - III

- **CHALLENGE 1: Make sure the master node is somehow magically able to compute g_1^k!**
- **CHALLENGE 2:** Make sure we learn the optimal shifts, i.e., make sure that

$$h_i^k \to \nabla f_i(x^\star) \qquad \text{as} \qquad k \to \infty$$

for all $i \in \{1, 2, \ldots, n\}$.

The DIANA method of Mishchenko et al [47] resolves both of these issues! DIANA is the first CGD method able to converge linearly to the exact solution. DIANA can be seen as a variance reduction technique for reducing the variance introduced by gradient compression.

How Does DIANA Address these Challenges?

DIANA: Resolving Challenge 1

- Store the initial shifts $h_1^0, h_2^0, \ldots, h_n^0$ also on the master!
- Make sure master is able to compute the new shifts $h_1^{k+1}, h_2^{k+1}, \ldots, h_n^{k+1}$ from the current shifts $h_1^k, h_2^k, \ldots, h_n^k$ and the latest information obtained from the nodes, i.e.,

$$\mathcal{C}_1(\nabla f_1(x^k) - h_1^k), \mathcal{C}_2(\nabla f_2(x^k) - h_2^k), \ldots, \mathcal{C}_n(\nabla f_n(x^k) - h_n^k).$$

- In particular, **in DIANA we will update the shifts via the rule**

$$h_i^{k+1} = h_i^k + \alpha \mathcal{C}_i(\nabla f_i(x^k) - h_i^k), \quad i = 1, 2, \ldots, n, \tag{200}$$

where $\alpha > 0$ is a suitably chosen stepsize. **This can clearly be done by the master as well!**

DIANA: Resolving Challenge 2 - I

While, as we have seen, the update of the form (200) addresses **CHALLENGE 1**, why on earth should it also address **CHALLENGE 2?** In other words, why should the sequence $\{h_i^k\}_{k\geq 0}$ formed via (200) satisfy

$$h_i^k \to \nabla f_i(x^\star) \quad \text{as} \quad k \to \infty \ ?$$

It turns out that **the update rule (200) is equivalent to 1 step of CGD with stepsize α applied to a certain optimization problem whose solution is $\nabla f_i(x^k)$.** In particular, this is the optimization problem:

$$\max_{h_i \in \mathbb{R}^d} \phi_i^k(h_i) \overset{\text{def}}{=} -\frac{1}{2}\left\|h_i - \nabla f_i(x^k)\right\|^2. \tag{201}$$

- Note that this problem keeps changing from iteration to iteration because $\nabla f_i(x^k)$ keeps changing.
- ϕ_i^k is 1-smooth and 1-strongly concave (i.e., $-\phi_i^k$ is 1-strongly convex)

DIANA: Resolving Challenge 2 - II

- The only solution of ϕ_i^k is $\nabla f_i(x^k)$
- The gradient is given by

$$\nabla \phi_i^k(h_i) = \nabla f_i(x^k) - h_i$$

- The compressed gradient is given by

$$\mathcal{C}_i(\nabla \phi_i^k(h_i)) = \mathcal{C}_i(\nabla f_i(x^k) - h_i)$$

- One step of CGD with stepsize $\alpha > 0$ applied to (201) is given by

$$h_i^{k+1} = h_i^k + \alpha \mathcal{C}_i(\nabla f_i(x^k) - h_i^k) \tag{202}$$

- We know from Corollary 52 (convergence of CGD) applied to the auxiliary problem (201) that

$$\mathbf{E}\left[\left\|h_i^{k+1} - \nabla f_i(x^k)\right\|^2 \mid x^k, h_i^k\right] \leq (1-\alpha)\left\|h_i^k - \nabla f_i(x^k)\right\|^2, \tag{203}$$

where $\alpha \leq \frac{1}{\omega_i+1}$.

DIANA: Resolving Challenge 2 - III

- So, h_i^{k+1} is closer to the true gradient $\nabla f_i(x^k)$ than h_i^k. In this sense, the sequence $\{h_i^k\}_{k\geq 0}$ is "tracking" the current gradient vector $\nabla f_i(x^k)$. So, intuitively, if we believe that

$$\nabla f_i(x^k) \to \nabla f_i(x^\star),$$

we must have

$$h_i^k \to \nabla f_i(x^\star)$$

as well, albeit at a slower rate.

Non-distributed Case ($n = 1$)

DIANA for $n = 1$

Algorithm 18 DIANA for 1 node [47, 31]

Parameters: learning rates $\alpha > 0$ and $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$; initial shift $h^0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

  Compute the gradient: $\nabla f(x^k)$

  Compress shifted gradient: $m^k = \mathcal{C}\left(\nabla f(x^k) - h^k\right)$

  Compute gradient estimator: $g^k = h^k + m^k$

  Take proximal SGD step: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

  Update the shift: $h^{k+1} = h^k + \alpha m^k$

Establishing the σ^k Assumption

Lemma 101

Assume f is convex and L-smooth, and $\mathcal{C} \in \mathbb{B}^d(\omega)$. Let $0 \leq \alpha \leq \frac{1}{\omega+1}$ and let $\sigma : \mathbb{R}^d \to [0, \infty)$ and σ^k be defined as follows:

$$\sigma(h) \overset{def}{=} \|h - \nabla f(x^\star)\|^2, \qquad \sigma^k \overset{def}{=} \sigma(h^k) = \left\|h^k - \nabla f(x^\star)\right\|^2.$$

Then for all iterations of `DIANA` *(Algorithm 18), we have*

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, h^k\right] \leq 2\underbrace{(2\omega+1)L}_{A} D_f(x^k, x^\star) + \underbrace{2\omega}_{B}\sigma^k, \tag{204}$$

and

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] \leq 2\underbrace{\alpha L}_{\tilde{A}} D_f(x^k, x^\star) + \underbrace{(1-\alpha)}_{\tilde{B}}\sigma^k. \tag{205}$$

Proof of Lemma 101 - I

First inequality. Noting that $g^k = h^k + \mathcal{C}(\nabla f(x^k) - h^k)$, observe that we have shown inequality (204) already. Indeed, compare with (175):

AC Inequality in the $\nabla f(x^\star) \neq 0$ Case

New proof: Let us use a compression operator $\mathcal{C} \in \mathbb{B}^d(\omega; h)$ for some $h \neq 0$ instead.

$$
\begin{aligned}
\mathrm{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x^\star)\|^2\right] &\overset{(37)}{=} \mathrm{E}\left[\|\mathcal{C}(\nabla f(x)) - \nabla f(x)\|^2\right] + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(*)}{\leq} \omega \|\nabla f(x) - h\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= \omega \left\|\nabla f(x^k) - \nabla f(x^\star) + \nabla f(x^\star) - h\right\|^2 + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&\overset{(35)}{\leq} 2\omega \|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega \|\nabla f(x^\star) - h\|^2 \\
&\qquad + \|\nabla f(x) - \nabla f(x^\star)\|^2 \\
&= (2\omega + 1) \|\nabla f(x) - \nabla f(x^\star)\|^2 + 2\omega \|\nabla f(x^\star) - h\|^2 \\
&\leq 2(2\omega + 1) L D_f(x, x^\star) + 2\omega \|\nabla f(x^\star) - h\|^2, \qquad (172)
\end{aligned}
$$

where in (*) we used the fact that $\mathcal{C} \in \mathbb{B}^d(\omega; h)$ and in the last step we have used convexity and L-smoothness of f.

Proof of Lemma 101 - II

Second inequality. It only remains to establish Inequality (205). Recall that $m^k = \mathcal{C}(\nabla f(x^k) - h^k)$ and let $h^\star = \nabla f(x^\star)$. Due to unbiasedness of $\mathcal{C}$,

$$\mathbf{E}\left[m^k \mid x^k, h^k\right] = \nabla f(x^k) - h^k. \tag{206}$$

Further, since $\mathcal{C} \in \mathbb{B}^d(\omega)$, we have

$$\mathbf{E}\left[\left\|m^k\right\|^2 \mid x^k, h^k\right] \leq (\omega + 1)\left\|\nabla f(x^k) - h^k\right\|^2. \tag{207}$$

Since $h^{k+1} = h^k + \alpha m^k$, we can decompose

$$\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &= \mathbf{E}\left[\left\|h^{k+1} - h^\star\right\|^2 \mid x^k, h^k\right] \\
&= \mathbf{E}\left[\left\|\alpha m^k + (h^k - h^\star)\right\|^2 \mid x^k, h^k\right] \\
&= \left\|h^k - h^\star\right\|^2 + 2\mathbf{E}\left[\left\langle \alpha m^k, h^k - h^\star\right\rangle \mid x^k, h^k\right] + \mathbf{E}\left[\left\|\alpha m^k\right\|^2 \mid x^k, h^k\right] \\
&\overset{(206)}{=} \left\|h^k - h^\star\right\|^2 + 2\alpha\left\langle \nabla f(x^k) - h^k, h^k - h^\star\right\rangle + \alpha^2\mathbf{E}\left[\left\|m^k\right\|^2 \mid x^k, h^k\right] \\
&\overset{(207)}{\leq} \left\|h^k - h^\star\right\|^2 + 2\alpha\left\langle \nabla f(x^k) - h^k, h^k - h^\star\right\rangle \\
&\qquad + \alpha^2(\omega + 1)\left\|\nabla f(x^k) - h^k\right\|^2.
\end{aligned}$$

Proof of Lemma 101 - III

Due to the stepsize restriction $0 < \alpha \leq \frac{1}{\omega+1}$, and using the identity

$$2\langle a, b\rangle + \|a\|^2 = \|a+b\|^2 - \|b\|^2 \tag{208}$$

for $a = \nabla f(x^k) - h^k$ and $b = h^k - h^\star$, we can continue as follows

$$\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &\leq \left\|h^k - h^\star\right\|^2 + \alpha\left(2\left\langle \nabla f(x^k) - h^k, h^k - h^\star\right\rangle + \left\|\nabla f(x^k) - h^k\right\|^2\right) \\
&\overset{(208)}{=} \left\|h^k - h^\star\right\|^2 + \alpha\left(\left\|\nabla f(x^k) - h^\star\right\|^2 - \left\|h^k - h^\star\right\|^2\right) \\
&= (1-\alpha)\left\|h^k - h^\star\right\|^2 + \alpha\left\|\nabla f(x^k) - h^\star\right\|^2 \\
&\leq (1-\alpha)\sigma^k + 2\alpha L D_f(x^k, x^\star).
\end{aligned}$$

In the last step we have used convexity and L-smoothness of f.

Convergence of DIANA for $n = 1$

Now that we have established the σ^k assumption in Lemma 101, we can now apply Theorem 97 (convergence of SGD-CTRL):

Corollary 102

Assume that f is μ-convex and L-smooth. If the stepsizes satisfy

$$\alpha \leq \frac{1}{\omega + 1}, \qquad \gamma \leq \frac{1}{(2\omega + 1)\, L + M\alpha L},$$

where $M > \frac{2\omega}{\alpha}$, then the iterates of DIANA *satisfy*

$$\mathbf{E}\left[V^k\right] \leq \max\left\{(1 - \gamma\mu)^k, \left(\frac{2\omega + M(1-\alpha)}{M}\right)^k\right\} V^0, \tag{209}$$

where the Lyapunov function V^k is defined by

$$V^k \stackrel{def}{=} \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k.$$

Understanding the Rate

For the particular choice

$$\alpha = \frac{1}{\omega + 1}, \qquad M = 4\omega(\omega + 1), \qquad \gamma = \frac{1}{(6\omega + 1)\, L},$$

the iteration complexity of DIANA ($n = 1$ case) is

$$\max\left\{\frac{1}{\gamma\mu}, \frac{1}{\alpha - \frac{2\omega}{M}}\right\} \log\frac{1}{\varepsilon} = \boxed{\max\left\{(6\omega + 1)\,\frac{L}{\mu}, 2(\omega + 1)\right\} \log\frac{1}{\varepsilon}}$$

Key Insight: Same $\mathcal{O}(\frac{(\omega+1)L}{\mu} \log \frac{1}{\varepsilon})$ rate as CGD, but always converging to $x^\star$ rather than to a neighbourhood!

Distributed Case ($n \geq 1$)

Two Variance Reduction Techniques: Fixed and Learned Shift

Method	Alg #	Gradient Estimator $g^k =$
CGD-SHIFT	Alg 13	$\mathcal{C}(\nabla f(x^k) - \nabla f(x^\star)) + \nabla f(x^\star)$
DIANA (1 node)	Alg 18	$\mathcal{C}(\nabla f(x^k) - h^k) + h^k$
DCGD-SHIFT		$\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{C}_i(\nabla f_i(x^k) - \nabla f_i(x^\star)) + \nabla f_i(x^\star)\right]$
DIANA (n nodes)	Alg 19	$\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{C}_i(\nabla f_i(x^k) - h_i^k) + h_i^k\right]$

DIANA: the Algorithm

We can finally present an implementable distributed variance reduced CGD method: DIANA.

Algorithm 19 DIANA [47, 31]

Parameters: learning rates $\alpha > 0$ and $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$; initial local shifts $h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$ (stored on the n nodes)
Initialize: $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$ (stored on the master)
for $k = 0, 1, 2, \ldots$ **do**
    Broadcast x^k to all workers
    for $i = 1, \ldots, n$ in parallel **do**
        Compute local gradient: $\nabla f_i(x^k)$
        Compress shifted local gradient: $m_i^k = \mathcal{C}_i(\nabla f_i(x^k) - h_i^k)$
        Send message m_i^k to the master (define local gradient estimator: $g_i^k = h_i^k + m_i^k$)
        Update the local shift: $h_i^{k+1} = h_i^k + \alpha m_i^k$
    Aggregate received messages: $m^k = \frac{1}{n}\sum_{i=1}^n m_i^k$
    Compute global gradient estimator: $g^k = h^k + m^k$ (note: $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$)
    Take proximal SGD step: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
    Update aggregated shift: $h^{k+1} = h^k + \alpha m^k$ (note: $h^{k+1} = \frac{1}{n}\sum_{i=1}^n h_i^{k+1}$)

σ^k Assumption for DIANA for General n

The following result shows that the σ^k assumption (i.e., Assumption 7) holds.

Lemma 103 (Lemma 1 and consequence of Lemma 2 from [31])

Assume that functions f_i are convex and L_i-smooth for all i. Let $\mathcal{C}_i \in \mathbb{B}^d(\omega)$ for all i. Suppose that $\alpha \leq \frac{1}{\omega+1}$. Let $h = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^d \times \mathbb{R}^d \cdots \times \mathbb{R}^d = \mathbb{R}^{nd}$ and define $\sigma : \mathbb{R}^{nd} \to [0, \infty)$ and σ^k by

$$\sigma(h) = \frac{1}{n}\sum_{i=1}^n \|h_i - \nabla f_i(x^\star)\|^2 \qquad \sigma^k \overset{def}{=} \sigma(h^k) = \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^\star)\right\|^2.$$

Then for all iterations $k \geq 0$ of Algorithm 19 (DIANA) we have

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k, h^k\right] &= \nabla f(x^k), \\
\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, h^k\right] &\leq \underbrace{2\left(\frac{2\omega}{n}+1\right) L_{\max}}_{A} D_f(x^k, x^\star) + \underbrace{\frac{2\omega}{n}}_{B} \sigma^k, \\
\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &\leq 2\underbrace{\alpha L_{\max}}_{\tilde{A}} D_f(x^k, x^\star) + \underbrace{(1-\alpha)}_{\tilde{B}} \sigma^k.
\end{aligned}$$

Convergence of DIANA

Since we have established the σ^k assumption in Lemma 103, Theorem 97 applies (convergence of SGD-CTRL), and we get the following corollary:

Corollary 104

Assume that f_i is convex and L_i-smooth for all $i \in [n]$ and f is μ-convex. If the stepsizes satisfy

$$\alpha \leq \frac{1}{\omega + 1}, \qquad \gamma \leq \frac{1}{\left(1 + \frac{2\omega}{n}\right) L_{\max} + M L_{\max} \alpha},$$

where $M > \frac{2\omega}{n\alpha}$, then the iterates of DIANA *satisfy*

$$\mathbf{E}\left[V^k\right] \leq \max\left\{(1-\gamma\mu)^k, \left(\frac{2\frac{\omega}{n} + M(1-\alpha)}{M}\right)^k\right\} V^0, \tag{210}$$

where the Lyapunov function V^k is defined by

$$V^k \stackrel{def}{=} \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k.$$

Understanding the Rate - I

For the particular choice

$$\alpha = \frac{1}{\omega + 1}, \qquad M = \frac{4\omega(\omega+1)}{n}, \qquad \gamma = \frac{1}{\left(1 + \frac{6\omega}{n}\right) L_{\max}},$$

the iteration complexity of DIANA is

$$\max\left\{\frac{1}{\gamma\mu}, \frac{1}{\alpha - \frac{2\omega}{nM}}\right\} \log\frac{1}{\varepsilon} = \boxed{\max\left\{\kappa + \kappa\frac{6\omega}{n}, 2(\omega+1)\right\} \log\frac{1}{\varepsilon}} \quad (211)$$

where

$$\kappa \stackrel{\text{def}}{=} \frac{L_{\max}}{\mu}.$$

Understanding the Rate - II

Example 105 (DIANA with Rand-1)

Consider DIANA with random sparsification Rand-1 (henceforth: DIANA-Rand-1). This compressor has variance parameter $\omega = d - 1$, and hence (211) reduces to

$$\mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}\left(1+\frac{d}{n}\right), d\right\}\log\frac{1}{\varepsilon}\right) \tag{212}$$

Each node only sends 1 float to the master. If we declare this as 1 unit of communication (and consider master-worker communication to be negligible), then (212) is the total communication complexity.

Comparison with GD ($n = 1$): If $n = 1$, then $L_{\max} = L$ and communication complexity (212) of DIANA-Rand-1 is

$$\mathcal{O}\left(d\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

Understanding the Rate - III

This is the same as communication complexity of GD since GD needs $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ communication rounds and sends d floats in each.

Key Insight: DIANA-Rand-1 and GD have identical total communication complexity in the single node ($n = 1$) regime. While GD needs $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ communication rounds and sends d floats in each, DIANA-Rand-1 needs $\mathcal{O}\left(d\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ communication rounds and sends 1 float in each.

Comparison with GD (big n): If $d \leq \frac{L_{\max}}{\mu}$ and $d \leq n$, then the total communication complexity of (212) of DIANA-Rand-1 simplifies to

$$\mathcal{O}\left(\frac{L_{\max}}{\mu}\log\frac{1}{\varepsilon}\right).$$

Understanding the Rate - IV

On the other hand, since GD always communicates d floats, the total communication complexity of GD is

$$\mathcal{O}\left(d\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

Key Insight: DIANA-Rand-1 is better than GD if $d \leq n$ and $d\mu \leq L_{\max} \leq dL$. Recall that $L \leq L_{\max} \leq nL$, and that both extremes are possible. So, DIANA-Rand-1 can be up to d times better than GD (in the favorable case when $L_{\max} = L$).

Comparison with L-SVRG-US: Let us think of L-SVRG-US as a distributed method, even though it was not designed for that. As a distributed method, L-SVRG-US always communicates full dimensional vectors, but only a small number of workers do any work in any given communication round (1 worker with probability $1 - p = 1 - \frac{1}{n}$ and all n workers with probability $p = \frac{1}{n}$). This is in some sense the exact opposite approach to how DIANA-Rand-1 works. Indeed, DIANA-Rand-1 makes

Understanding the Rate - V

sure all workers are busy in each communication round, and subsequently makes sure only 1 float is sent by each worker to the master. This approach is chosen because the workers can work in parallel and hence involving more does not take any additional time. Instead, the bottleneck is communication, which is why gradients are sparsified.

So, it is no surprise that `DIANA-Rand-1` performs better than `L-SVRG` as a distributed method in terms of total (communication) complexity:

Method	`L-SVRG-US`	`DIANA` with Rand-1
# communications	$\mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}, n\right\}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}\left(1+\frac{d}{n}\right), d\right\}\log\frac{1}{\varepsilon}\right)$
cost per communication	$\mathcal{O}(d)$	$\mathcal{O}(1)$
# workers participating	$\mathcal{O}(1)$	n
total complexity	$\mathcal{O}\left(\max\left\{d\frac{L_{\max}}{\mu}, dn\right\}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}\left(1+\frac{d}{n}\right), d\right\}\log\frac{1}{\varepsilon}\right)$
total complexity ($n \leq d$)	$\mathcal{O}\left(\max\left\{d\frac{L_{\max}}{\mu}, dn\right\}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\max\left\{\frac{d}{n}\frac{L_{\max}}{\mu}, d\right\}\log\frac{1}{\varepsilon}\right)$
total complexity ($n \geq d$)	$\mathcal{O}\left(\max\left\{d\frac{L_{\max}}{\mu}, dn\right\}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\max\left\{\frac{L_{\max}}{\mu}, d\right\}\log\frac{1}{\varepsilon}\right)$

Extensions

Extensions I

- **Variance reduction for sketches and $n = 1$**
  - **Hanzely et al [27] (NeurIPS 2018)**
  - Algorithm name: SEGA
  - This is the first time a variance reduced CGD method was developed
  - The motivation of the paper was *not* to produced variance reduction for sketches, but instead to develop a generalized variant of RCD that can work with any (even non-separable) regularizer R. No prior RCD method could do that.
- **Any compressor from $\mathbb{B}^d(\omega)$**
  - **Horváth et al [31]**
  - The first DIANA paper [47] only worked with a very specific compressor (ternary quantizer). In [31] this was extended to any compressor from $\mathbb{B}^d(\omega)$.
- **General analysis of SGD**
  - **Gorbunov et al [19] (AISTATS 2020)**
  - The convergence proof for DIANA we present here is based on the general analysis technique based on the "σ^k assumption" developed in [19].
- **Alternative to DIANA but for sparsification and general n**

Extensions II

- **Mishchenko et al [48] (UAI 2020)**
- Based on the idea of "independent coordinate descent"

- **Acceleration**
  - **Li et al [44] (ICML 2020)**
  - `ADIANA`: the first accelerated `CGD` method (combination of Nesterov acceleration and `DIANA`)
  - Improved communication complexity
- **Finding fixed points**
  - **Chraibi et al [9]**
  - Problem: Find a fixed point of operator $T(x) = \frac{1}{n}\sum_{i=1}^{n} T_i(x)$, where $T_i : \mathbb{R}^d \to \mathbb{R}^d$ for $i = 1, 2, \ldots, n$ are operators (T is a strict contraction, and T_i is Lipschitz for all i).
  - Compression is applied to $T_i(x)$ and not to gradients (there are no gradients here).
- **Decentralized optimization**
  - **Kovalev et al [41]**
  - Workers are the nodes of a connected network. There is no master node, and communication can take place along the edges of the network only.

Extensions III

- **Biased compressors**
  - **Gorbunov et al [20] (NeurIPS 2020)**
  - First linearly converging `DCGD` using biased compressors.
  - Combination of Error Compensation and `DIANA` is needed to achieve this

Exercises

Exercises - I

Exercise 69

Give a detailed justification for why (203) *holds.*

Exercise 70

Consider the following alternative update rule (202)*:*

$$h^{k+1} = h^k - \alpha \mathcal{C}(h^k - \nabla f(x^k)). \tag{213}$$

(i) *Give an interpretation of this rule.*

(ii) *Show that* (203) *holds for this rule as well.*

(iii) *Show that a variant of* `DIANA` *using this update rule has the same convergence properties as* `DIANA`

Exercise 71

Provide a detailed proof of Corollary 102.

Exercises - II

Exercise 72

Provide a detailed proof of Lemma 103.

Exercise 73

Generalize `DIANA` *to the setting in which node i uses compressor $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$, and establish a convergence theorem for this new method.*

Exercise 74

Generalize `DIANA` *to bidirectional compression; that is, allow the master to compress the message it broadcasts to the workers. Establish a convergence theorem for this new method.*

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 16: Stochastic Average Gradient Algorithm (SAGA)

Based on:

[10] Aaron Defazio, Francis Bach and Simon Lacoste-Julien, **SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives**, NeurIPS, 2014

[25] Robert M. Gower, P.R. and Francis Bach, **Stochastic quasi-gradient methods: variance reduction via Jacobian sketching**, Mathematical Programming 188:135–192, 2021 [arXiv:1805.02632, 2018]

[57] Xun Qian, Zheng Qu and P.R., **SAGA with arbitrary sampling**, ICML, 2019

Problem

We consider the finite-sum problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + R(x), \tag{214}$$

but return to the non-distributed setting.

Plan for this lecture:

- Describe one more variance reduced SGD method for finite sum problems in addition to SGD-SHIFT and L-SVRG
- The method—SAGA **(Stochastic Average Gradient Algorithm)**—was developed by Defazio, Bach and Lacoste-Julien in 2014 [10]
- Offer alternative interpretations of the method:
  - **Jacobian sketching**
  - **Controlled stochastic reformulation**

SAGA - The True Origin of the Method's Name

Figure: Francis Bach deciding on the SAGA name for their method. In his own words: "the true reason behind SAGA (in my daily walk in Paris, I bumped into this shop)"

Variance Reduced Gradient Estimators

Randomness in g^k	Method	Alg #	Gradient Estimator $g^k =$	Useful?
Subsampling	GD	Alg 2	$\nabla f_i(x^k) - \nabla f_i(x^k) + \nabla f(x^k)$	Inefficient
Subsampling	SGD-SHIFT	Alg 14	$\nabla f_i(x^k) - \nabla f_i(x^\star) + \nabla f(x^\star)$	Theoretical
Subsampling	L-SVRG	Alg 16	$\nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$	Practical
Compression	GD	Alg 2	$\mathcal{C}(\nabla f(x^k) - \nabla f(x^k)) + \nabla f(x^k)$	Inefficient
Compression	CGD-SHIFT	Alg 13	$\mathcal{C}(\nabla f(x^k) - \nabla f(x^\star)) + \nabla f(x^\star)$	Theoretical
Compression	DIANA	Alg 18	$\mathcal{C}(\nabla f(x^k) - h^k) + h^k$	Practical

SAGA: Basic Description

L-SVRG vs SAGA: Comparison of Gradient Estimators I

1. L-SVRG

Sample index i uniformly at random

$$g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$$

$$w^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ w^k & \text{with probability} \quad 1-p \end{cases}$$

2. SAGA

Sample index i uniformly at random

$$g^k = \nabla f_i(x^k) - \nabla f_i(w_i^k) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w_j^k)$$

$$w_j^{k+1} = \begin{cases} x^k & \text{for} \quad j = i \\ w_j^k & \text{for} \quad j \neq i \end{cases}$$

L-SVRG vs SAGA: Comparison of Gradient Estimators II

1. L-SVRG
   - Samples a single index i, uniformly at random
   - All $j = 1, 2, \ldots, n$ share **one global auxiliary vector** $w^k \in \mathbb{R}^d$
     - **Low memory requirement:** $1 \times d$ ($w^k \in \mathbb{R}^d$)
   - The global vector w^{k+1} is updated to x^k with probability p in each iteration (independently of other randomness)
     - **Gradient learning frequency:** Update all gradients once in a while; i.e., update n gradients once every $1/p$ iterations (on average)

$$\nabla f_j(w^k) \to \nabla f_j(x^k) \quad \text{for all} \quad j = 1, 2, \ldots, n$$

2. SAGA
   - Samples a single index i, uniformly at random
   - Each $j = 1, 2, \ldots, n$ has its own **local auxiliary vector** $w_j^k \in \mathbb{R}^d$
     - **Large memory requirement:** $n \times d$ ($w_1^k, \ldots, w_n^k \in \mathbb{R}^d$)
   - One local vector (w_i^{k+1}) gets updated to x^k in each iteration
     - **Gradient learning frequency:** Update one gradient in each iteration

$$\nabla f_i(w_i^k) \to \nabla f_i(x^k)$$

The SAGA Algorithm

Algorithm 20 SAGA (Stochastic Average Gradient Algorithm)

1: **Parameters:** learning rate $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$, initial local iterates $w_1^0, w_2^0, \ldots, w_n^0 \in \mathbb{R}^d$

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: Sample $i \in \{1, 2, \ldots, n\}$ uniformly at random

4: $g^k = \nabla f_i\left(x^k\right) - \nabla f_i\left(w_i^k\right) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j\left(w_j^k\right)$

5: Take SGD step: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

6: Set $w_j^{k+1} = \begin{cases} x^k & j = i \\ w_j^k & j \neq i \end{cases}$

Unbiasedness of the Stochastic Gradient

For simplicity, define

$$w^k \stackrel{\text{def}}{=} (w_1^k, w_2^k, \ldots, w_n^k) \in \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d = \mathbb{R}^{nd}.$$

Let us check that the gradient estimator g^k employed by SAGA is unbiased. Taking expectation in the random index selection i at iteration k, we get:

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k, w^k\right] &= \mathbf{E}\left[\nabla f_i\left(x^k\right) - \nabla f_i\left(w_i^k\right) + \frac{1}{n}\sum_{j=1}^n \nabla f_j\left(w_j^k\right) \mid x^k, w^k\right] \\
&= \frac{1}{n}\sum_{j=1}^n \nabla f_j(x^k) - \frac{1}{n}\sum_{j=1}^n \nabla f_j\left(w_j^k\right) + \frac{1}{n}\sum_{j=1}^n \nabla f_j\left(w_j^k\right) \\
&= \nabla f(x^k).
\end{aligned}$$

The σ^k Assumption

We now show that the σ^k assumption (Assumption 7) is satisfied:

Lemma 106

Assume that for each $i = 1, 2, \ldots, n$, the function f_i is convex and L_i-smooth. Then for Algorithm 20 (SAGA) we have the following recursions:

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right] \leq 2\underbrace{(2L_{\max})}_{A} D_f(x^k, x^\star) + \underbrace{2}_{B}\sigma^k$$

and

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, w^k\right] \leq 2\underbrace{\frac{L_{\max}}{n}}_{\tilde{A}} D_f(x^k, x^\star) + \underbrace{\left(1 - \frac{1}{n}\right)}_{\tilde{B}}\sigma^k, \qquad (215)$$

where $\sigma(w) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_i) - \nabla f_i(x^\star)\|^2$ *and*

$$\sigma^k \overset{def}{=} \sigma(w^k) = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w_i^k) - \nabla f_i(x^\star)\right\|^2.$$

Proof of Lemma 106 - I

Using standard arguments, $\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right]$ is equal to

$$
\begin{aligned}
&= \mathbf{E}\left[\left\|\nabla f_i\left(w^k\right) - \nabla f_i\left(w_i^k\right) + \frac{1}{n}\sum_{j=1}^{n}\nabla f_j\left(w_j^k\right) - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&= \mathbf{E}\left[\left\|\nabla f_i\left(x^k\right) - \nabla f_i(x^\star) + \nabla f_i(x^\star) - \nabla f_i\left(w_i^k\right) + \frac{1}{n}\sum_{j=1}^{n}\nabla f_j\left(w_j^k\right) - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&\leq \mathbf{E}\left[2\left\|\nabla f_i\left(x^k\right) - \nabla f_i(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&\qquad + \mathbf{E}\left[2\left\|\nabla f_i(x^\star) - \nabla f_i\left(w_i^k\right) + \frac{1}{n}\sum_{j=1}^{n}\nabla f_j\left(w_j^k\right) - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&= 2\mathbf{E}\left[\left\|\nabla f_i\left(x^k\right) - \nabla f_i(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&\qquad + 2\mathbf{E}\left[\left\|\underbrace{\nabla f_i(x^\star) - \nabla f_i\left(w_i^k\right)}_{A_i} - \underbrace{\left(\nabla f(x^\star) - \frac{1}{n}\sum_{j=1}^{n}\nabla f_j\left(w_j^k\right)\right)}_{\mathbf{E}[A_i \mid x^k, w^k]}\right\|^2 \mid x^k, w^k\right].
\end{aligned}
$$

Proof of Lemma 106 - II

Applying the inequality

$$\mathbf{E}\left[\left\|A_i - \mathbf{E}\left[A_i \mid x^k, w^k\right]\right\|^2 \mid x^k, w^k\right] \leq \mathbf{E}\left[\|A_i\|^2 \mid x^k, w^k\right]$$

(i.e., variance is upper bounded by second moment), we can continue:

$$\begin{aligned}
\mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, w^k\right] &\leq 2\mathbf{E}\left[\left\|\nabla f_i\left(x^k\right) - \nabla f_i(x^\star)\right\|^2 \mid x^k, w^k\right] \\
&\quad + 2\mathbf{E}\left[\left\|\nabla f_i(x^\star) - \nabla f_i\left(w_i^k\right)\right\|^2 \mid x^k, w^k\right] \\
&= 2\frac{1}{n}\sum_{j=1}^{n}\underbrace{\left\|\nabla f_j\left(x^k\right) - \nabla f_j(x^\star)\right\|^2}_{\leq 2L_j D_{f_j}(x^k, x^\star)} \\
&\quad + 2\underbrace{\frac{1}{n}\sum_{j=1}^{n}\left\|\nabla f_j(x^\star) - \nabla f_j\left(w_j^k\right)\right\|^2}_{\sigma^k} \\
&\leq 4L_{\max}\frac{1}{n}\sum_{j=1}^{n} D_{f_j}(x^k, x^\star) + 2\sigma^k \\
&= 4L_{\max} D_f(x^k, x^\star) + 2\sigma^k.
\end{aligned}$$

Proof of Lemma 106 - III

We now proceed to the second recursion. First, note that for every i we have

$$w_i^{k+1} = \begin{cases} w_i^k & \text{with probability} \quad 1 - \frac{1}{n} \\ x^k & \text{with probability} \quad \frac{1}{n} \end{cases}. \tag{216}$$

Therefore,

$$\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, w^k\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[\left\|\nabla f_i\left(w_i^{k+1}\right) - \nabla f_i\left(x^\star\right)\right\|^2 \mid x^k, w^k\right] \\
&\overset{(216)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[\left(1-\frac{1}{n}\right)\left\|\nabla f_i\left(w_i^k\right) - \nabla f_i\left(x^\star\right)\right\|^2 + \frac{1}{n}\left\|\nabla f_i\left(x^k\right) - \nabla f_i\left(x^\star\right)\right\|^2\right] \\
&= \left(1-\frac{1}{n}\right)\sigma^k + \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{\left\|\nabla f_i\left(x^k\right) - \nabla f_i\left(x^\star\right)\right\|^2}_{\leq 2L_i D_{f_i}(x^k, x^\star)} \\
&\leq \left(1-\frac{1}{n}\right)\sigma^k + \frac{2L_{\max}}{n} D_f\left(x^k, x^\star\right).
\end{aligned}$$

Convergence of SAGA

In view of Lemma 106, our main result for analyzing SGD-CTRL (Theorem 97) implies the following convergence result:

Corollary 107

Assume that for each i, the function f_i is convex L_i-smooth. Further assume f is μ-convex. Choose $\gamma = \frac{1}{6L_{\max}}$. Then Algorithm 20 (SAGA) satisfies:

$$\mathbf{E}\left[V^k\right] \leq \left(1 - \min\left\{\frac{\mu}{6L_{\max}}, \frac{1}{2n}\right\}\right)^k V^0,$$

where the Lyapunov function is defined by

$$V^k \stackrel{def}{=} \left\|x^k - x^\star\right\|^2 + \frac{n}{9L_{\max}^2}\sigma^k,$$

and

$$\sigma^k \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w_i^k) - \nabla f_i(x^\star)\right\|^2.$$

Number of Iterations: L-SVRG vs SAGA

- L-SVRG: In view of Corollary 100, we have

$$k \geq \max\left\{\frac{6L_{\max}}{\mu}, \frac{2}{p}\right\}\log\frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[V^k\right] \leq \varepsilon V^0.$$

- SAGA: In view of Corollary 107, we have

$$k \geq \max\left\{\frac{6L_{\max}}{\mu}, 2n\right\}\log\frac{1}{\varepsilon} \quad \Rightarrow \quad \mathbf{E}\left[V^k\right] \leq \varepsilon V^0.$$

Conclusions:

- Same rate if we choose $p = \frac{1}{n}$ in L-SVRG (which is one of the optimal choices for p)
- L-SVRG has extra flexibility to choose p

Proof of Corollary 107

In view of Lemma 106, Assumption 7 holds with

$$A = 2L_{\max}, \quad B = 2, \quad \tilde{A} = \frac{L_{\max}}{n}, \quad \tilde{B} = 1 - \frac{1}{n}, \quad C = \tilde{C} = 0.$$

We can now apply Theorem 97 with $M = 4n > 2n = \frac{B}{1-\tilde{B}}$. Note that

$$A + M\tilde{A} = 2L_{\max} + 8n > \mu,$$

and hence the stepsize bound becomes

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + M\tilde{A}}\right\} = \frac{1}{A + M\tilde{A}} = \frac{1}{6L_{\max}}.$$

So, the choice $\gamma = \frac{1}{6L_{\max}}$ is justified. Since $\frac{B+M\tilde{B}}{M} = 1 + \frac{2}{4n} - \frac{1}{n} = 1 - \frac{1}{2n}$ and using $\gamma = \frac{1}{6L_{\max}}$, the rate in Theorem 97 becomes

$$\begin{aligned}\max\left\{(1-\gamma\mu)^k, \left(\frac{B+M\tilde{B}}{M}\right)^k\right\} &= \max\left\{\left(1 - \frac{\mu}{6L_{\max}}\right)^k, \left(1 - \frac{1}{2n}\right)^k\right\} \\ &= \left(1 - \min\left\{\frac{\mu}{6L_{\max}}, \frac{1}{2n}\right\}\right)^k.\end{aligned}$$

Finally, the Lyapunov function is

$$V^k = \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k = \left\|x^k - x^\star\right\|^2 \frac{4n}{36L_{\max}^2}\sigma^k.$$

Understanding SAGA

Recall that we were able to interpret the gradient learning mechanism of DIANA as one step of CGD applied to an auxiliary optimization problem whose solution is the gradient.

Can SAGA be interpreted in a similar way? Yes! [25].

In fact, we will give **2 more interpretations,** each offering a unique insight:

- **Jacobian learning**
- **Controlled stochastic reformulation**

SAGA: Jacobian Learning

Jacobian Learning

The **Jacobian learning** interpretation of SAGA [25] relies on the following steps:

1. Lifting
2. Sketching
3. Random linear system
4. Projection/SGD
5. Gradient estimate
6. SGD

We will now explain these steps.

Jacobian Learning: Lifting

We can't access individual functions f_i or their gradients by applying some transformation to f, where $f = \frac{1}{n}\sum_{i=1}^{n} f_i$.

- The information about individual functions $f_1, f_2, \ldots, f_n$ is not present in f.
- f only provides a statistic about the functions $f_1, \ldots, f_n$: their average.

It will be useful to define an object which does not have the above issue—**Lifted function**—defined as

$$F : \mathbb{R}^d \to \mathbb{R}^n \qquad F(x) = (f_1(x), f_2(x), \ldots, f_n(x)).$$

- We can access f_i by looking at the **coordinates of F:**

$$(F(x))_i = f_i(x).$$

- We can access ∇f_i be looking at the columns of the **Jacobian of F,** defined as

$$\boldsymbol{J}(x) \stackrel{\text{def}}{=} [\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_n(x)] \in \mathbb{R}^{d \times n}.$$

Jacobian Learning: Sketching

The Jacobian $\boldsymbol{J}(x)$ contains useful information about the gradient of f and also the gradients of the individual functions f_i:

- If e_i is the ith standard unit basis vector in $\mathbb{R}^n$, then

$$\boldsymbol{J}(x)e_i = \nabla f_i(x).$$

- If $e = \sum_i e_i \in \mathbb{R}^n$ is the vector of all ones, then

$$\boldsymbol{J}(x)\frac{e}{n} = \nabla f(x).$$

We **can** think of the process of obtaining/computing the stochastic gradient $\nabla f_i(x^k)$ as performing a **sketch (i.e., a random linear transformation) of the unknown Jacobian $\boldsymbol{J}(x^k)$:**

$$\boldsymbol{J}(x^k)e_i = \nabla f_i(x^k). \tag{217}$$

Jacobian Learning: Random Linear System

While we do not know $\boldsymbol{J}(x^k)$, **its sketch** (217) gives us some **(random) information about it.** Note that

$$\boldsymbol{J} e_i = \nabla f_i(x^k) \tag{218}$$

can be seen as a **random linear system** (since e_i is random), where

- e_i is the (random and known) system matrix
- $\nabla f_i(x^k)$ is the (known) right hand side
- $\boldsymbol{J} \in \mathbb{R}^{d \times n}$ is the unknown variable to be found

Solutions of the random linear system:

- $\boldsymbol{J} = \boldsymbol{J}(x^k)$ solves (218)
- However, besides $\boldsymbol{J} = \boldsymbol{J}(x^k)$, the system (218) has an infinity of solutions: any matrix $\boldsymbol{J} \in \mathbb{R}^{d \times n}$ whose ith column is $\nabla f_i(x^k)$ is a solution.

Jacobian Learning: Projection I

We will use the information contained in the Jacobian sketch $\nabla f_i(x^k)$ to **learn the Jacobian:**

- Given current estimate $\boldsymbol{J}^k \in \mathbb{R}^{d\times n}$ of the Jacobian, let us define the next estimate as

$$\begin{aligned} \boldsymbol{J}^{k+1} &= \arg\min \left\|\boldsymbol{J} - \boldsymbol{J}^k\right\|_F^2 \\ \text{subject to} \quad & \boldsymbol{J}e_i = \nabla f_i(x^k) \end{aligned} \tag{219}$$

  where $\|\cdot\|_F$ is the **Frobenius matrix norm** defined by

$$\|\boldsymbol{J}\|_F^2 \stackrel{\text{def}}{=} \mathrm{Trace}\left(\boldsymbol{J}^\top \boldsymbol{J}\right) = \mathrm{Trace}\left(\boldsymbol{J}\boldsymbol{J}^\top\right) = \sum_{i=1}^d \|\boldsymbol{J}_{i:}\|^2 = \sum_{j=1}^n \left\|\boldsymbol{J}_{:j}\right\|^2 = \sum_{i=1}^d \sum_{j=1}^n \boldsymbol{J}_{ij}^2.$$

- It is clear that the solution of (219) is given by

$$\boldsymbol{J}_{:j}^{k+1} = \begin{cases} \boldsymbol{J}_{:j}^k & j \neq i \\ \nabla f_i(x^k) & j = i \end{cases}. \tag{220}$$

Jacobian Learning: Projection II

Why does it make sense to perform the **sketch-and-project** iteration (219)?

- **Intuitive justification:**
  - We **believe** that $\boldsymbol{J}^k$ is good, and do not want to change it much, unless necessary.
  - At the same time, we **observe the Jacobian sketch** $\boldsymbol{J}(x^k)e_i = \nabla f_i(x^k)$, which provides us with **new information about the Jacobian.**
  - In particular, we know that the **true Jacobian $J(x^k)$ satisfies the random linear system** $\boldsymbol{J}e_i = \nabla f_i(x^k)$.
  - So, we define $\boldsymbol{J}^{k+1}$ as the **closest matrix to J^k consistent with this (random) linear system.**
- **Formal justification: J^{k+1} can be seen as one step of `SGD-US`, with unit stepsize, applied to a strongly convex quadratic optimization problem whose unique solution is $J(x^k)$:**

Jacobian Learning: Projection III

- Define $\phi : \mathbb{R}^{d\times n} \to \mathbb{R}$ via

$$\phi(\boldsymbol{J}) \stackrel{\text{def}}{=} \frac{1}{2n}\left\|\boldsymbol{J} - \boldsymbol{J}(x^k)\right\|_F^2 = \frac{1}{n}\sum_{j=1}^n \underbrace{\frac{1}{2}\left\|\boldsymbol{J}_{:j} - \nabla f_j(x^k)\right\|^2}_{\phi_j(\boldsymbol{J})}$$

  and consider the optimization problem

$$\min_{\boldsymbol{J}\in\mathbb{R}^{d\times n}} \left[\phi(x) = \frac{1}{n}\sum_{j=1}^n \phi_j(\boldsymbol{J})\right].$$

- By construction, ϕ is a $\frac{1}{n}$-strongly convex ($\mu = \frac{1}{n}$) quadratic function with a unique minimizer: $\boldsymbol{J}^\star \stackrel{\text{def}}{=} \boldsymbol{J}(x^k)$. It can be checked that each ϕ_j is convex and 1-smooth. Thus, the average smoothness constant is $L_{\max} = 1$. Also, we know that the gradient of ϕ at the optimum is zero:

$$\nabla\phi(\boldsymbol{J}^\star) = 0 \in \mathbb{R}^{d\times n}.$$

- One step of `SGD-US` with stepsize γ applied to this problem has the form

$$\boldsymbol{J}^{k+1} = \boldsymbol{J}^k - \gamma\nabla\phi_i(\boldsymbol{J}^k). \tag{221}$$

Jacobian Learning: Projection IV

- Since ϕ_i does not depend on $\boldsymbol{J}_{:j}$ for $j \neq i$, we have

$$(\nabla \phi_i(\boldsymbol{J}))_{:j} = \begin{cases} 0 \in \mathbb{R}^d & j \neq i \\ \boldsymbol{J}_{:j} - \nabla f_j(x^k) & j = i \end{cases}. \tag{222}$$

- By plugging the gradient (222) into `SGD-US` (221) and using the unit stepsize $\gamma = 1$, we get

$$\boldsymbol{J}_{:j}^{k+1} = \boldsymbol{J}_{:j}^k - \left(\nabla \phi_i(\boldsymbol{J}^k)\right)_{:j} = \begin{cases} \boldsymbol{J}_{:j}^k & j \neq i \\ \nabla f_i(x^k) & j = i \end{cases},$$

recovering (220).

Jacobian Learning: Projection V

- We can now apply the convergence result for SGD-US. The result allows stepsizes
$$0 < \gamma \leq \frac{1}{L_{\max}} = 1.$$
Using the largest stepsize $\gamma = 1$, and since the strong convexity constant of ϕ is $\mu = \frac{1}{n}$, we get the rate
$$(1 - \gamma\mu)^k = \left(1 - \frac{1}{n}\right)^k.$$
The one-step decrease can be captured as
$$\mathbf{E}\left[\left\|\boldsymbol{J}^{k+1} - \boldsymbol{J}^\star\right\|_F^2 \mid \boldsymbol{J}^k, x^k\right] \leq \left(1 - \frac{1}{n}\right)\left\|\boldsymbol{J}^k - \boldsymbol{J}^\star\right\|_F^2. \tag{223}$$
- The Jacobian learning factor $1 - \frac{1}{n}$ in (223) manifests itself in the decrease rate for σ^k in the second recursion (215).

Jacobian Learning: Gradient Estimate I

Recall that the only reason why we are trying to learn the Jacobian is because it may be helpful in generating a **good gradient estimator.**

- Since for the true Jacobian we have $\boldsymbol{J}(x^k)\frac{e}{n} = \nabla f(x^k)$, it seems reasonable to define the gradient estimator from $\boldsymbol{J}^{k+1}$ in the same way:

$$\nabla f(x^k) \approx \boldsymbol{J}^{k+1}\frac{e}{n}.$$

- However, such an estimator would be **biased,** and very hard to analyze. In fact, using this estimator leads to the famous SAG method of Le Roux, Schmidt and Bach [69] which won the Lagrange Prize in Continuous Optimization in 2018.

Jacobian Learning: Gradient Estimate II

- It turns out that we get an **unbiased** estimator by a linear combination of $\boldsymbol{J}^{k+1}\frac{e}{n}$ and $\boldsymbol{J}^{k}\frac{e}{n}$:

$$
\begin{aligned}
g^k &\stackrel{\text{def}}{=} n\boldsymbol{J}^{k+1}\frac{e}{n} + (1-n)\boldsymbol{J}^{k}\frac{e}{n} \\
&= \left(\boldsymbol{J}^{k+1} - \boldsymbol{J}^{k}\right) e + \boldsymbol{J}^{k}\frac{e}{n} \\
&= \sum_{j=1}^{n}\left(\boldsymbol{J}_{:j}^{k+1} - \boldsymbol{J}_{:j}^{k}\right) + \boldsymbol{J}^{k}\frac{e}{n} \\
&\stackrel{(220)}{=} \nabla f_i(x^k) - \boldsymbol{J}_{:i}^{k} + \boldsymbol{J}^{k}\frac{e}{n} \\
&= \nabla f_i(x^k) - \boldsymbol{J}_{:i}^{k} + \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{J}_{:j}^{k}. \qquad (224)
\end{aligned}
$$

Jacobian Learning: Gradient Estimate III

- If we now let w_j^k to be such vectors for which

$$\nabla f_j(w_j^k) = \boldsymbol{J}_j^k,$$

then the estimator g^k in (224) can be written in the form

$$g^k = \nabla f_i(x^k) - \nabla f_i(w_i^k) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w_j^k),$$

which is the SAGA **estimator!**

Jacobian Learning: SGD

Finally, SAGA is SGD with the gradient estimator g^k built in the way described above:

$$x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$$

The SAGA Algorithm: Jacobian Learning View

Algorithm 21 SAGA (Stochastic Average Gradient Algorithm)

1: **Parameters:** learning rate $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$, initial Jacobian estimate $\boldsymbol{J}^0 \in \mathbb{R}^{d \times n}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample $i \in \{1, 2, \ldots, n\}$ uniformly at random
4: $g^k = \nabla f_i(x^k) - \boldsymbol{J}^k_{:i} + \frac{1}{n} \sum_{j=1}^n \boldsymbol{J}^k_{:j}$
5: Take SGD step: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
6: Update the Jacobian estimate via sketch-and-project

$$\boldsymbol{J}^{k+1} = \arg \min_{\boldsymbol{J} \in \mathbb{R}^{d \times n}} \left\{ \left\| \boldsymbol{J} - \boldsymbol{J}^k \right\|_F^2 \mid \boldsymbol{J} e_i = \nabla f_i(x^k) \right\}$$

SAGA: Controlled Stochastic Reformulation

Controlled Stochastic Reformulation

The **controlled stochastic reformulation** interpretation [25] of SAGA relies on the following steps:

1. Simple stochastic reformulation
2. Zero mean random function
3. Controlled stochastic reformulation

We will now explain these steps.

CSR: Simple Stochastic Reformulation

- Let i be an index from $\{1, 2, \ldots, n\}$ chosen uniformly at random and for each $j \in \{1, 2, \ldots, n\}$ define the random variable

$$v_j \stackrel{\text{def}}{=} \begin{cases} n & j = i \\ 0 & j \neq i \end{cases}. \tag{225}$$

  Clearly, $\mathbf{E}[v_j] = 1$ for all j.
- Now define the **random function**

$$f_v(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} v_j f_j(x) = \frac{1}{n} \langle F(x), v \rangle. \tag{226}$$

  Clearly,

$$\mathbf{E}[f_v(x)] = \mathbf{E}\left[\frac{1}{n} \sum_{j=1}^{n} v_j f_j(x)\right] = \frac{1}{n} \sum_{j=1}^{n} \mathbf{E}[v_j] f_j(x) = \frac{1}{n} \sum_{j=1}^{n} f_j(x) = f(x).$$

This leads to the **simple stochastic reformulation** of the finite-sum problem (214):

$$\min_{x \in \mathbb{R}^d} \mathbf{E}[f_v(x)] + R(x).$$

CSR: Zero Mean Random Function

Given $\boldsymbol{J} \in \mathbb{R}^{d \times n}$, define the **random control function**

$$z_{\boldsymbol{J},v}(x) \stackrel{\text{def}}{=} \frac{1}{n} \left\langle \boldsymbol{J}^\top x, v \right\rangle = \frac{1}{n} \left\langle \boldsymbol{J} v, x \right\rangle . \tag{227}$$

Clearly, if we subtract from $z_{\boldsymbol{J},v}(x)$ its expectation, we get a **zero mean random function.** Indeed, if we define

$$Z_{\boldsymbol{J},v}(x) \stackrel{\text{def}}{=} z_{\boldsymbol{J},v}(x) - \mathbf{E}\left[z_{\boldsymbol{J},v}(x)\right], \tag{228}$$

then

$$\mathbf{E}\left[Z_{\boldsymbol{J},v}(x)\right] \equiv 0.$$

Controlled Stochastic Reformulation I

Define

$$f_{\boldsymbol{J},v}(x) \stackrel{\text{def}}{=} f_v(x) - Z_{\boldsymbol{J},v}(x). \tag{229}$$

Since $\mathbf{E}\left[Z_{\boldsymbol{J},v}(x)\right] \equiv 0$, we have

$$f(x) = \mathbf{E}\left[f_v(x)\right] = \mathbf{E}\left[f_{\boldsymbol{J},v}(x)\right].$$

This leads to the **controlled stochastic reformulation** of the finite-sum problem (214):

$$\min_{x \in \mathbb{R}^d} \mathbf{E}\left[f_{\boldsymbol{J},v}(x)\right] + R(x). \tag{230}$$

We now wish to solve this via SGD:

$$x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma \nabla f_{\boldsymbol{J},v^k}(x^k)\right).$$

- The matrix $\boldsymbol{J}$ acts as the **control**
- If $\boldsymbol{J} = 0$ (no control), we have $z_{\boldsymbol{J},v}(x) \equiv Z_{\boldsymbol{J},v}(x) \equiv 0$, and hence the controlled reformulation (230) is identical to the simple reformulation.
- Is it useful to consider $\boldsymbol{J} \neq 0$?

Controlled Stochastic Reformulation II

Observe that

$$\begin{aligned}
\nabla f_{\boldsymbol{J},v}(x) &\overset{(229)}{=} \nabla f_v(x) - \nabla Z_{\boldsymbol{J},v}(x) \\
&\overset{(228)}{=} \nabla f_v(x) - \nabla z_{\boldsymbol{J},v}(x) + \nabla \mathbf{E}\left[z_{\boldsymbol{J},v}(x)\right] \\
&\overset{(226)+(227)}{=} \frac{1}{n}\left\langle \boldsymbol{J}(x), v\right\rangle - \frac{1}{n}\boldsymbol{J}v + \frac{1}{n}\boldsymbol{J}e \\
&= \frac{1}{n}\left\langle \boldsymbol{J}(x) - \boldsymbol{J}, v\right\rangle + \frac{1}{n}\boldsymbol{J}e. \qquad (231)
\end{aligned}$$

Notice that

$$\nabla f_{\boldsymbol{J},v}(x) \overset{(231)}{=} \frac{1}{n}\left\langle \boldsymbol{J}(x) - \boldsymbol{J}, v\right\rangle + \frac{1}{n}\boldsymbol{J}e \overset{(225)}{=} \left(\boldsymbol{J}(x)\right)_{:i} - \boldsymbol{J}_{:i} + \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{J}_{:j}.$$

For an appropriate choice of $\boldsymbol{J}$, **this is the gradient estimator of SAGA !**
Consider now the **Jacobian-at-optimum** matrix

$$\boldsymbol{J}^\star \overset{\text{def}}{=} \left[\nabla f_1(x^\star), \nabla f_2(x^\star), \ldots, \nabla f_n(x^\star)\right] \in \mathbb{R}^{d\times n}$$

Controlled Stochastic Reformulation III

and observe that

$$\nabla f_{\boldsymbol{J}^\star,v}(x) \overset{(231)}{=} \frac{1}{n}\left\langle \boldsymbol{J}(x) - \boldsymbol{J}^\star, v\right\rangle + \frac{1}{n}\boldsymbol{J}^\star e, \tag{232}$$

which implies

$$\nabla f_{\boldsymbol{J}^\star,v}(x^\star) \overset{(232)}{=} \frac{1}{n}\left\langle \boldsymbol{J}(x^\star) - \boldsymbol{J}^\star, v\right\rangle + \frac{1}{n}\boldsymbol{J}^\star e = \frac{1}{n}\boldsymbol{J}^\star e = \nabla f(x^\star).$$

Hence, the variance of the stochastic gradient at the optimum (i.e., of $\nabla f_{\boldsymbol{J}^\star,v}(x^\star)$) is zero:

$$\mathbf{E}\left[\left\|\nabla f_{\boldsymbol{J}^\star,v}(x^\star) - \nabla f(x^\star)\right\|^2\right] = 0.$$

This means (after an expected smoothness condition is verified) that SGD **applied to the controlled stochastic reformulation with control $J = J^\star$ is a variance-reduced method!**

- We do not know $\boldsymbol{J}^\star$

Controlled Stochastic Reformulation IV

- However, we can try to learn $\boldsymbol{J}^\star$ by producing a sequence of approximations $\{\boldsymbol{J}^k\}$ with $\boldsymbol{J}^k \to \boldsymbol{J}^\star$
- At iteration k we apply one step of SGD for solving the controlled stochastic reformulation with control $\boldsymbol{J}^k$:

$$\min_{x \in \mathbb{R}^d} \mathbf{E}\left[f_{\boldsymbol{J}^k, v}(x)\right] + R(x).$$

- We then update $\boldsymbol{J}^k$ to $\boldsymbol{J}^{k+1}$ via Jacobian learning: (220)
- The process is repeated

The method described above is SAGA!

The SAGA Algorithm: CSR View

Here is a description of SAGA as SGD **applied to a sequence of controlled stochastic reformulations of the original problem**, with the control $\boldsymbol{J}^k$ evolving via the Jacobian sketching mechanism (which is, as we described, one step of SGD-US applied to a sequence of auxiliary convex optimization problem).

Algorithm 22 SAGA (Stochastic Average Gradient Algorithm)

Parameters: learning rate $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$, initial Jacobian estimate $\boldsymbol{J}^0 \in \mathbb{R}^{d\times n}$

for $k = 0, 1, 2, \ldots$ **do**

  Sample $i = i^k \in \{1, 2, \ldots, n\}$ uniformly at random

  For all j define $v_j^k = \begin{cases} n & j = i \\ 0 & j \neq i \end{cases}$

  $g^k = \nabla f_{\boldsymbol{J}^k, v^k}(x^k)$

  Take SGD step: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$

  Update the Jacobian estimate $\boldsymbol{J}_{:j}^{k+1} = \begin{cases} \boldsymbol{J}_{:j}^k & j \neq i \\ \nabla f_i(x^k) & j = i \end{cases}$

Extensions

- SAGA can be seen as a **special case of a more general variance-reduced method called JacSketch:** Gower, Richtárik, Bach 2018 [25].
  - The first importance sampling for SAGA with the "right rate".
  - Best minibatch variants of SAGA.
- Designing an arbitrary sampling variant of SAGA was an open problem until recently: Qian, Qu and Richtárik, ICML 2019 [57]
- It is possible to **interpolate** between L-SVRG and SAGA: Hoffman et al, NeurIPS 2015 [29].
- SAGA can be analyzed also for smooth convex functions, and smooth nonconvex functions.

Exercises

Exercises

Exercise 75

Argue formally why (220) *is the solution to the sketch-and-project iteration* (219).

Exercise 76

Let $\boldsymbol{S} \in \mathbb{R}^{n\times b}$ *be a (random) matrix and assume that the following generalized Jacobian sketch is known:* $\boldsymbol{J}(x^k)\boldsymbol{S} \in \mathbb{R}^{d\times b}$. *This gives rise to the generalized sketch-and-project iteration*

$$\begin{aligned} \boldsymbol{J}^{k+1} &= \arg\min \left\|\boldsymbol{J} - \boldsymbol{J}^k\right\|_F^2 \\ \textit{subject to} \quad & \boldsymbol{J}\boldsymbol{S} = \boldsymbol{J}(x^k)\boldsymbol{S} \end{aligned} \tag{233}$$

Prove that

$$\boldsymbol{J}^{k+1} = \boldsymbol{J}^k + \left(\boldsymbol{J}(x^k) - \boldsymbol{J}^k\right)\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{S})^\dagger\boldsymbol{S}^\top, \tag{234}$$

where $\dagger$ *refers to the Moore-Penrose pseudoinverse.*

Exercise 77

Verify that in the case of $\boldsymbol{S} = e_i$, *the general formula* (234) *reduces to* (220).

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 17: Nonconvex Setting: the ABC Assumption

Based on:

[37] Ahmed Khaled and P.R., **Better Theory for SGD in the Nonconvex World**, arXiv:2002.03329, 2020

Problem and Goal

The Problem

We study the complexity of SGD,

$$x^{k+1} = x^k - \gamma g(x^k), \tag{235}$$

for solving the **non-regularized** optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{236}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is **possibly nonconvex** and satisfies the following conditions:

Assumption 8

Function f is

- *bounded from below, i.e., $f(x) \geq f^{\inf} \in \mathbb{R}$ for all $x \in \mathbb{R}^d$,*
- *L-smooth.*

Goal: Finding a Stationary Point

In this lecture we will give up on the goal of finding a global optimum, and will instead be satisfied with finding a stationary point, defined next:

Definition 108 (Approximately stationary point)

We say that a (possibly random) vector $x \in \mathbb{R}^d$ is an **ε-stationary point** if

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon^2, \tag{237}$$

where $\mathbf{E}[\cdot]$ is the expectation over the randomness inherent in x.

Remark: Since $(\mathbf{E}[\|\nabla f(x)\|])^2 \leq \mathbf{E}\left[\|\nabla f(x)\|^2\right]$, (237) implies the bound

$$\mathbf{E}[\|\nabla f(x)\|] \leq \varepsilon.$$

Some Initial Thinking I

Recall that in the **convex** and L-smooth case, the following inequality played a key role in the analysis if GD:

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x, y). \tag{238}$$

- We only needed to use (238) for $y = x^\star$: the (global and local) solution.
- The AC assumption—key assumption in our convergence analysis of SGD—was a generalization of (238) to the case of stochastic gradient estimators.

Some Initial Thinking II

- In the nonconvex case we now consider the non-regularized problem

$$\min f$$

only, and we do not aim to find a global solution $x^\star$. We merely aim to find a stationary point y, i.e., a point for which

$$\nabla f(y) = 0.$$

In such a case we have

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle = f(x) - f(y)$$

and (238) simplifies to

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f(y)). \tag{239}$$

Some Initial Thinking III

- For a stationary point y, the value $f(y)$ is a-priori unknown, and could be arbitrarily close to $\inf f$. However, Assumption 8 does not guarantee that a minimizer exists. So, in order to make sure an inequality of the type (239) can hold in the nonconvex setting, we could replace the right hand side by the supremum over y:

$$\begin{aligned} \|\nabla f(x)\|^2 &\leq \sup_y 2L(f(x) - f(y)) \\ &= 2L(f(x) - \inf_y f(y)) \\ &\leq 2L(f(x) - f^{\inf}). \end{aligned}$$

- In this heuristic way we have arrived at (not proved!) an inequality which might be relevant in the nonconvex L-smooth setting. We shall next **prove** that this inequality indeed holds under Assumption 8.

A Key Lemma

Lemma 109

Let f be a function for which Assumption 8 is satisfied. Then

$$\|\nabla f(x)\|^2 \leq 2L\left(f(x) - f^{\inf}\right), \qquad \forall x \in \mathbb{R}^d. \tag{240}$$

Remarks:

- **The good:** This lemma holds **without the need for convexity.**
- **The bad:** We have **lost the ability to consider arbitrary y** and have to consider y for which $\nabla f(y) = 0$ only. But this is OK as we will not need to use it for any other y.
- **The really bad:** It turns out that (240) **is not strong enough to lead to a good analysis of GD in the nonconvex setting.** Indeed, even if x is a stationary point, the right hand side of (240) is not necessarily equal to zero! Therefore, it is not possible to control the level of stationarity of x using the right hand side of (240) for any other point other than the global minimum.

Proof of Lemma 109

If we let $x_+ = x - \frac{1}{L}\nabla f(x)$, then using the L-smoothness of f we get

$$f(x_+) \leq f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{L}{2}\|x_+ - x\|^2 .$$

Using that $f^{\inf} \leq f(x_+)$ and the definition of x_+ we have,

$$f^{\inf} \leq f(x_+) \leq f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

Rearranging, we get the claim.

The ABC Assumption

The ABC Assumption - I

The following assumption introduced by Khaled and R. (2020) [37] plays a key role in our analysis.

Assumption 9 (ABC)

The second moment of the stochastic gradient satisfies

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq 2A\left(f(x) - f^{\inf}\right) + B\|\nabla f(x)\|^2 + C, \qquad \text{(ABC)}$$

for some $A, B, C \geq 0$ *and all* $x \in \mathbb{R}^d$.

Remarks:

- **Terminology.** In [37], the (ABC) inequality is called **expected smoothness.** However, in an attempt to come up with a more meaningful terminology, in this course we reserved the notion "expected smoothness "to something else.
  - Indeed, in this course we reserved the term "expected smoothness" to a bound of the form

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y).$$

The ABC Assumption - II

- Note that for any stationary point y we have $\nabla f(y) = 0$, and hence

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle = f(x) - f(y).$$

  So, if y is a stationary point, the expected smoothness inequality simplifies to

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A''(f(x) - f(y)).$$

- **ABC vs AC Assumption.** (ABC) Assumption is a generalization of the AC assumption which we used to analyze non-variance reduced variants of SGD in the strongly convex and convex cases.
  - **Left hand side.** Since $\nabla f(y) = 0$ for any stationary point y (e.g., the global solution $y = x^\star$, if one exists), we have

$$\mathbf{E}\left[\|g(x)\|^2\right] = \mathbf{E}\left[\|g(x) - \nabla f(y)\|^2\right],$$

    which is what we were bounding in the AC assumption.
  - **A term.** Instead of $D_f(x, y) = f(x) - f(y)$, we have the term $f(x) - f^{\inf}$ on the right hand side of (ABC).

The ABC Assumption - III

- B **term.** We have an extra B term which we did not have in the AC assumption.

- **Purpose of the B term.** The extra B term is necessary if we want to get a tight convergence result in all special cases of SGD.
  - For instance, the GD estimator $g(x) = \nabla f(x)$ satisfies (ABC) with $A = 0$, $B = 1$ and $C = 0$.
  - In the strongly convex and convex cases we did not need the B term because we were able to use the bound $\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^\star))$ instead. Indeed, convexity and L-smoothness imply this bound. So, the A term was enough.

Existing Models of Stochastic Gradient

Existing Models of Stochastic Gradient I

1. Bounded variance. Ghadimi and Lan [16] analyze SGD under the assumption that f is lower bounded, that the stochastic gradients $g(x)$ are unbiased and have **bounded variance**

$$\mathbf{E}\left[\|g(x) - \nabla f(x)\|^2\right] \leq \sigma^2. \tag{BV}$$

Note that due to unbiasedness, this is equivalent to

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \|\nabla f(x)\|^2 + \sigma^2. \tag{241}$$

With an appropriately chosen constant stepsize γ, their results imply a $\mathcal{O}(\varepsilon^{-4})$ rate of convergence.

2. Maximal strong growth. In the context of finite-sum problems

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{242}$$

Existing Models of Stochastic Gradient II

and SGD-US gradient estimator $g(x) = \nabla f_i(x)$, the **maximal strong growth** condition requires the inequality

$$\|g(x)\|^2 \leq \alpha \|\nabla f(x)\|^2 \tag{M-SG}$$

to hold almost surely for some $\alpha \geq 0$.

- Tseng (1998) [82] used the maximal strong growth condition early to establish the convergence of the incremental gradient method, a closely related variant of SGD.
- Schmidt and Le Roux (2013) [73] prove linear convergence of SGD for strongly convex objectives under (M-SG).

3. Expected strong growth. We may also assume that (M-SG) holds in expectation rather than uniformly, leading to **expected strong growth**:

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \alpha \|\nabla f(x)\|^2 \tag{E-SG}$$

for some $\alpha > 0$.

Existing Models of Stochastic Gradient III

- Like (M-SG), variants of (E-SG) have also been used in convergence results for incremental gradient methods [78].
- Vaswani et al (2019) [84] prove that under (E-SG), SGD converges to an ε-stationary point in $\mathcal{O}(\varepsilon^{-2})$ steps.
- (E-SG) is quite strong and necessitates interpolation: if $\nabla f(x) = 0$, then $g(x) = 0$ almost surely. This is typically not true in the finite-sum case where the functions f_i can be very different [46].

4. Relaxed growth condition. Bottou et al (2018) [7] consider the **relaxed growth condition** which is a version of (E-SG) featuring an additive constant:

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \alpha \|\nabla f(x)\|^2 + \beta. \tag{RG}$$

In view of (241), (RG) can be also seen as a generalization of the (BV) assumption.

- Bertsekas and Tsitsiklis (2000) [4] establish the almost-sure convergence of SGD under (RG).

Existing Models of Stochastic Gradient IV

- Bottou et al (2018) [7] give a $\mathcal{O}(\varepsilon^{-2})$ convergence rate for nonconvex objectives to a neighborhood of a stationary point of radius linearly proportional to β.
- Unfortunately, (RG) is quite difficult to verify in practice and **can be shown not to hold for some simple problems**, as the following proposition shows.

Proposition 21

There is a simple finite-sum minimization problem with two functions for which (RG) *is not satisfied.*

5. Gradient confusion. Sankararaman et al (2019) [72] postulate a **gradient confusion** bound for finite-sum problems and SGD with uniform single-element sampling. This bound requires the existence of $\eta > 0$ such that

$$\langle \nabla f_i(x), \nabla f_j(x) \rangle \geq -\eta \tag{GC}$$

holds for all $i \neq j$ and all $x \in \mathbb{R}^d$.

Existing Models of Stochastic Gradient V

- For general nonconvex objectives, they prove convergence to a neighborhood only.

6. Sure smoothness. Lei et al (2019) [43] analyze SGD for f of the form

$$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[f_\xi(x) \right]. \tag{243}$$

They assume that f_ξ is nonnegative and has L-Lipschitz gradient almost surely, i.e.,

$$\|g(x) - g(y)\| \leq L \|x - y\| \quad \text{and} \quad f_\xi(x) \geq 0 \tag{SS}$$

almost surely for all $x, y \in \mathbb{R}^d$. They then use $g(x) = \nabla f_\xi(x)$ for ξ sampled i.i.d. from $\mathcal{D}$. We call this condition **sure-smoothness** (SS).

- They establish the sample complexity $\mathcal{O}(\varepsilon^{-4} \log(\varepsilon^{-1}))$.
- Unfortunately, their results do not recover the rate of GD and their analysis is not easily extendable to compression and subsampling.

Proof of Proposition 21 - I

We define the function $f : \mathbb{R} \to \mathbb{R}$ as,

$$f(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| < 1, \\ |x| - \frac{1}{2} & \text{otherwise.} \end{cases}$$

Then f is 1-smooth and lower bounded by 0. We consider using SGD with the stochastic gradient estimator

$$g(x) = \begin{cases} \nabla f(x) + \sqrt{|x|} & \text{with probability } \frac{1}{2}, \\ \nabla f(x) - \sqrt{|x|} & \text{with probability } \frac{1}{2}. \end{cases}$$

We now show that (RG) does not hold: Assuming that (RG) holds, there exist constants α and β such that,

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \alpha \|\nabla f(x)\|^2 + \beta.$$

Consider $x = 2\max\{1, 2(\alpha + \beta)\}$. Then $|x| > 1$ and hence $\|\nabla f(x)\| = 1$ by the definition of f. Specializing (RG) we get,

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \alpha + \beta. \tag{244}$$

Proof of Proposition 21 - II

On the other hand,

$$\begin{aligned}\mathbf{E}\left[\|g(x)\|^2\right] = \frac{1}{2}\left(\left(1+\sqrt{x}\right)^2 + \left(1-\sqrt{x}\right)^2\right) &\geq \frac{1}{2}\left(1+\sqrt{x}\right)^2 \\ &\geq \frac{1}{2}|x| \\ &= \max\left\{1, 2\left(\alpha+\beta\right)\right\}. \quad (245)\end{aligned}$$

We see that (244) and (245) are in clear contradiction. It follows that (RG) does not hold.

We now show that (ABC) holds:

- First, suppose that $|x| \geq 1$. Then

$$\begin{aligned}\mathbf{E}\left[\|g(x)\|^2\right] &= \frac{1}{2}\left(\left(1+\sqrt{|x|}\right)^2 + \left(1-\sqrt{|x|}\right)^2\right) \\ &= \frac{1}{2}\left(2+2|x|\right) \\ &= 1+|x| \\ &= \frac{3}{2} + \left(f(x) - f^{\inf}\right), \quad (246)\end{aligned}$$

where in the last line we used that when $|x| \geq 1$ we have $f(x) - f^{\inf} = |x| - \frac{1}{2}$.

Proof of Proposition 21 - III

- Now suppose that $|x| \leq 1$. Then

$$\begin{aligned} \mathbf{E}\left[\|g(x)\|^2\right] &= \frac{1}{2}\left(\left(x+\sqrt{|x|}\right)^2 + \left(x-\sqrt{|x|}\right)^2\right) \\ &= x^2 + |x| \\ &\leq 1 + 1 = 2. \end{aligned} \tag{247}$$

Combining (246) and (247), we get the following inequality for all $x \in \mathbb{R}$:

$$\mathbf{E}\left[\|g(x)\|^2\right] \leq f(x) - f^{\inf} + 2.$$

It follows that (ABC) is satisfied with $A = \frac{1}{2}$, $B = 0$, and $C = 2$.

(ABC) is the Weakest Condition

Relationship Between the Assumptions

Figure: Relationships between the 6 existing conditions and the ABC condition established in Theorem 110. An arrow indicates an implication. (ABC) is the weakest condition of them all.

Relationship Between the Assumptions

Theorem 110 ((ABC) is the Weakest Condition)

If Assumption 8 holds (f is L-smooth and lower bounded), then (ABC) *is a weaker condition than the 6 described above:* (BV), (M-SG), (E-SG), (RG), (GC), (SS). *In particular, the following relations hold:*

1. (M-SG) $\Rightarrow$ (E-SG)
2. (E-SG) $\Rightarrow$ (RG)
3. (BV) $\Rightarrow$ (RG)
4. *If f has finite sum structure* (242), *then* (GC) $\Rightarrow$ (RG)
5. (RG) $\Rightarrow$ (ABC)
6. *If f has the expectation structure* (243), *then* (SS) $\Rightarrow$ (ABC)

Remark: We only need Assumption 8 for the last statement: (6).

Proof of Theorem 110 - I

1. (M-SG) $\Rightarrow$ (E-SG). Suppose that (M-SG) holds. Then
$$\mathbf{E}\left[\|g(x)\|^2\right] \leq \mathbf{E}\left[\alpha\|\nabla f(x)\|^2\right] = \alpha\,\|\nabla f(x)\|^2\,.$$
Hence (E-SG) holds.
2. (E-SG) $\Rightarrow$ (RG). If (E-SG) holds, then plugging $\beta = 0$ shows that (RG) holds.
3. (BV) $\Rightarrow$ (RG). If (BV) holds, then plugging $\alpha = 0$ and $\beta = \sigma^2$ shows that (RG) holds.
4. (GC) $\Rightarrow$ (RG) . We use the definition of $\|\nabla f(x)\|^2$,
$$\begin{aligned}
\|\nabla f(x)\|^2 &= \left\langle \sum_{i=1}^n \frac{\nabla f_i(x)}{n}, \sum_{j=1}^n \frac{\nabla f_j(x)}{n} \right\rangle \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \langle \nabla f_i(x), \nabla f_j(x)\rangle \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \|\nabla f_i(x)\|^2 + \sum_{i=1}^n\sum_{j=1,j\neq i}^n \langle \nabla f_i(x), \nabla f_j(x)\rangle\right) \\
&\overset{(GC)}{\geq} \frac{1}{n}\left(\mathbf{E}_i\left[\|\nabla f_i(x)\|^2\right] - \frac{\eta(n^2-n)}{n}\right) \\
&= \frac{1}{n}\mathbf{E}_i\left[\|\nabla f_i(x)\|^2\right] - \eta\cdot\left(1-\frac{1}{n}\right).
\end{aligned}$$

Proof of Theorem 110 - II

Rearranging we get,

$$\mathbf{E}\left[\|\nabla f_i(x)\|^2\right] \leq n \cdot \|\nabla f(x)\|^2 + \eta\,(n-1)\,.$$

Hence, (RG) holds with $\alpha = n$ and $\beta = \eta\,(n-1)$.

5. (RG) $\Rightarrow$ (ABC). Putting $A = 0$, $B = \alpha$ and $C = \beta$ shows that (ABC) is satisfied.
6. (SS) $\Rightarrow$ (ABC). Note that $f_\xi(x)$ is almost surely bounded from below by 0, and that by (SS), $f_\xi(x)$ is almost-surely L-smooth. Then using Lemma 109 we have almost surely,

$$\left\|\nabla f_\xi(x)\right\|^2 \leq 2L f_\xi(x).$$

Let $f^{\inf}$ be the infimum of f (necessarily exists, since $f(x) = \mathbf{E}\left[f_\xi(x)\right] \geq 0$). Then,

$$\left\|\nabla f_\xi(x)\right\|^2 \leq 2L\left(f_\xi(x) - f^{\inf}\right) + 2Lf^{\inf}.$$

Taking expectations and using that $f(x) = \mathbf{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right]$ we see that (ABC) is satisfied with $A = 2L$ and $C = 2Lf^{\inf}$.

General Convergence Theory

Lemma Providing a Key Recursion I

Our main convergence result relies on the following key lemma.

Lemma 111

Let Assumption 8 (L-smoothness and $f \geq f^{\inf}$) and Assumption 9 (ABC) be satisfied. Choose constant stepsize γ satisfying $0 < \gamma \leq \frac{1}{LB}$. Then for any $K \geq 1$, the iterates $\{x^k\}$ of SGD *satisfy*

$$\frac{1}{2}\sum_{k=0}^{K-1} w_k r^k + \frac{w_{K-1}}{\gamma}\delta^K \leq \frac{w_{-1}}{\gamma}\delta^0 + \frac{LC}{2}\sum_{k=0}^{K-1} w_k \gamma. \tag{248}$$

where $r^k \overset{def}{=} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right]$, $w_k \overset{def}{=} \frac{w_{-1}}{(1+LA\gamma^2)^{k+1}}$ for $w_{-1} > 0$ arbitrary, and $\delta^k \overset{def}{=} \mathbf{E}\left[f(x^k)\right] - f^{\inf}$.

Remark: Lemma 111 bounds a **weighted sum** of stochastic gradients over the entire run of the algorithm. This idea of weighting different iterates has been used in the analysis of SGD in the convex case [62, 77, 79] typically with the goal of returning a weighted average of the iterates $\bar{x}_K$ at the end. In contrast, we only use the weighting to facilitate the proof.

The Main Convergence Result

The main convergence result follows:

Theorem 112 (Main)

Let Assumption 8 (L-smoothness and $f \geq f^{\inf}$) and Assumption 9 (ABC) be satisfied. Choose constant stepsize γ satisfying $0 < \gamma \leq \frac{1}{LB}$. Then for any $K \geq 1$, the iterates $\{x^k\}$ of SGD *satisfy*

$$\min_{0 \leq k \leq K-1} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq LC\gamma + \frac{2\left(1 + LA\gamma^2\right)^K}{\gamma K}\delta^0,$$

where $\delta^0 \stackrel{def}{=} f(x^0) - f^{\inf}$.

Remarks:

- While the bound of Theorem 112 shows a possible **exponential blow-up**, we can show that by carefully controlling the stepsize we can nevertheless attain an **ε-stationary point given $\mathcal{O}(\varepsilon^{-4})$ stochastic gradient evaluations.**
- This dependence is in fact **optimal** for SGD without extra assumptions on second-order smoothness or disruptiveness of the stochastic gradient noise [13]. A similar stepsize was used in [16].

Proof of Theorem 112

We start with Lemma 111, which says that

$$\frac{1}{2}\sum_{k=0}^{K-1} w_k r^k \quad \leq \quad \frac{1}{2}\sum_{k=0}^{K-1} w_k r^k + \frac{w_{K-1}}{\gamma}\delta^K \quad \overset{(248)}{\leq} \quad \frac{w_{-1}}{\gamma}\delta^0 + \frac{LC}{2}\sum_{k=0}^{K-1} w_k \gamma.$$

Let $W_K \overset{\text{def}}{=} \sum_{k=0}^{K-1} w_k$. Dividing both sides by W_K, we obtain

$$\frac{1}{2}\min_{0\leq k\leq K-1} r^k \quad \leq \quad \frac{1}{2W_K}\sum_{k=0}^{K-1} w_k r^k \quad \leq \quad \frac{w_{-1}}{W_K}\frac{\delta^0}{\gamma} + \frac{LC\gamma}{2}. \tag{249}$$

Note that

$$W_K = \sum_{k=0}^{K-1} w_k \geq \sum_{k=0}^{K-1} \min_{0\leq i\leq K-1} w_i = K w_{K-1} = \frac{K w_{-1}}{(1+LA\gamma^2)^K}. \tag{250}$$

Using this in (249) yields

$$\frac{1}{2}\min_{0\leq k\leq K-1} r^k \leq \frac{(1+LA\gamma^2)^K}{\gamma K}\delta^0 + \frac{LC\gamma}{2}.$$

Multiplying both sides by 2 yields the theorem's claim.

Corollary: Iteration Complexity - I

Corollary 113

Fix $\varepsilon > 0$. *Choose the stepsize* $\gamma > 0$ *as*

$$\gamma = \min\left\{\frac{1}{\sqrt{LAK}}, \frac{1}{LB}, \frac{\varepsilon}{2LC}\right\}.$$

Then provided that

$$K \geq \frac{12\delta^0 L}{\varepsilon^2} \max\left\{\frac{12\delta^0 A}{\varepsilon^2}, B, \frac{2C}{\varepsilon^2}\right\}, \tag{251}$$

we have

$$\min_{0 \leq k \leq K-1} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \varepsilon^2.$$

Corollary: Iteration Complexity - II

Remarks:

- **Rate of GD:** The iteration complexity given by (251) **recovers the rate of GD:** plugging in $B = 1$ and $A = C = 0$ shows that we require a total of
$$\frac{12\delta^0 L}{\varepsilon^2}$$
iterations in required to a reach an ε-stationary point. This is the standard rate of convergence for GD on nonconvex objectives [3], up to absolute (non-problem-specific) constants.
- **Faster rate under** (E-SG)**:** Plugging in $A = C = 0$ and B to be any nonnegative constant recovers the fast convergence of SGD under strong growth (E-SG). Our bounds are similar to [43] but improve upon them by recovering full gradient descent, assuming the weaker (ABC) condition, and attaining the optimal
$$\mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$$
rate without logarithmic terms.

Proof of Corollary 113 - I

Theorem 112 says that if the stepsize satisfies

$$0 < \gamma \leq \frac{1}{LB}, \tag{252}$$

then

$$\min_{0 \leq k \leq K-1} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq LC\gamma + \frac{2\left(1 + LA\gamma^2\right)^K}{\gamma K}\delta^0. \tag{253}$$

Using the fact that $1 + x \leq \exp(x)$, we can bound

$$\left(1 + LA\gamma^2\right)^K \leq \left(\exp(LA\gamma^2)\right)^K = \exp(LA\gamma^2 K) \leq \exp(1) \leq 3,$$

where the second inequality holds because

$$\gamma \leq \frac{1}{\sqrt{LAK}} \tag{254}$$

by assumption. Substituting in (253), we get

$$\min_{0 \leq k \leq K-1} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq LC\gamma + \frac{6}{\gamma K}\delta^0. \tag{255}$$

To make the right hand side of (255) smaller than ε^2, we require that the first term satisfies $LC\gamma \leq \frac{\varepsilon^2}{2}$, which holds as long as

$$\gamma \leq \frac{\varepsilon^2}{2LC}. \tag{256}$$

Proof of Corollary 113 - II

Similarly, for the second term we require that $\frac{6\delta^0}{\gamma K} \leq \frac{\varepsilon^2}{2}$, and this will hold as long as the number of iterations satisfies

$$K \geq \frac{12\delta^0}{\gamma\varepsilon^2}. \tag{257}$$

Note that we have three requirements on the stepsize γ so far, (252), (254) and (256):

$$\gamma \leq \frac{1}{LB} \qquad \gamma \leq \frac{1}{\sqrt{LAK}} \qquad \gamma \leq \frac{\varepsilon^2}{2LC}.$$

Plugging all these three bounds on the stepsize into (257), we get

$$K \geq \frac{12\delta^0 LB}{\varepsilon^2} \qquad K \geq \frac{12\delta^0\sqrt{LAK}}{\varepsilon^2} \qquad K \geq \frac{24\delta^0 LC}{\varepsilon^4}, \tag{258}$$

respectively. Since K appears on both sides of the second requirement in (258), by cancellation and squaring we simplify (258) to

$$K \geq \frac{12\delta^0 LB}{\varepsilon^2} \qquad K \geq \frac{12^2(\delta^0)^2 LA}{\varepsilon^4} \qquad K \geq \frac{24\delta^0 LC}{\varepsilon^4}.$$

Finally, we collect the terms into a single bound:

$$K \geq \frac{12\delta^0 L}{\varepsilon^2} \max\left\{\frac{12\delta^0 A}{\varepsilon^2}, B, \frac{2C}{\varepsilon^2}\right\}.$$

Proof of Lemma 111 - I

Step 1: L-smoothness.

We start with the L-smoothness of f, which implies

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &\overset{(235)}{=} f(x^k) - \gamma \left\langle \nabla f(x^k), g(x^k) \right\rangle + \frac{L\gamma^2}{2}\left\|g(x^k)\right\|^2. \end{aligned} \tag{259}$$

Step 2: Applying the (ABC) Assumption.

Taking expectation in (259) conditional on x^k, and using Assumption 9 (ABC), we get

$$\begin{aligned} \mathbf{E}\left[f(x^{k+1}) \mid x^k\right] &= f(x^k) - \gamma\left\|\nabla f(x^k)\right\|^2 + \frac{L\gamma^2}{2}\mathbf{E}\left[\left\|g(x^k)\right\|^2\right] \\ &\overset{(ABC)}{\leq} f(x^k) - \gamma\left\|\nabla f(x^k)\right\|^2 + \frac{L\gamma^2}{2}\left(2A\left(f(x^k) - f^{\inf}\right) + B\left\|\nabla f(x^k)\right\|^2 + C\right) \\ &= f(x^k) - \gamma\left(1 - \frac{LB\gamma}{2}\right)\left\|\nabla f(x^k)\right\|^2 + LA\gamma^2\left(f(x^k) - f^{\inf}\right) + \frac{LC\gamma^2}{2}. \end{aligned}$$

Proof of Lemma 111 - II

Step 3: Full expectation and reshuffling terms.

Subtracting $f^{\inf}$ from both sides gives

$$\mathbf{E}\left[f(x^{k+1}) \mid x^k\right] - f^{\inf} \leq \left(1 + LA\gamma^2\right)\left(f(x_k) - f^{\inf}\right) - \gamma\left(1 - \frac{LB\gamma}{2}\right)\left\|\nabla f(x^k)\right\|^2 + \frac{LC\gamma^2}{2},$$

taking expectation again, using the tower property and rearranging, we get

$$\mathbf{E}\left[f(x^{k+1}) - f^{\inf}\right] + \gamma\left(1 - \frac{LB\gamma}{2}\right)\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \left(1 + LA\gamma^2\right)\mathbf{E}\left[f(x^k) - f^{\inf}\right] + \frac{LC\gamma^2}{2}.$$

Letting $\delta^k \overset{\text{def}}{=} \mathbf{E}\left[f(x^k) - f^{\inf}\right]$ and $r^k \overset{\text{def}}{=} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right]$, we can rewrite the last inequality as

$$\gamma\left(1 - \frac{LB\gamma}{2}\right) r^k \leq \left(1 + LA\gamma^2\right)\delta^k - \delta^{k+1} + \frac{LC\gamma^2}{2}.$$

Our choice of stepsize guarantees that $1 - \frac{LB\gamma}{2} \geq \frac{1}{2}$. As such,

$$\frac{\gamma}{2} r^k \leq \left(1 + LA\gamma^2\right)\delta^k - \delta^{k+1} + \frac{LC\gamma^2}{2} \tag{260}$$

for $k \geq 0$.

Proof of Lemma 111 - III

Step 4: Weighting sequence trick.

We now follow Stich (2019) [79] and define an exponentially decaying weighting sequence $w_0, w_1, w_2, \ldots, w_K$ that is used to weight the terms in (260). However, unlike [79], we are interested in the weighting sequence solely as a proof technique, and it does not show up in the final bounds.

Fix $w_{-1} > 0$ and define

$$w_k = \frac{w_{k-1}}{1 + LA\gamma^2} \qquad \text{for all} \qquad k \geq 0.$$

Multiplying recursion (260) by $\frac{w_k}{\gamma}$, we get

$$\begin{aligned} \frac{1}{2} w_k r^k &\leq \frac{w_k \left(1 + LA\gamma^2\right)}{\gamma} \delta^k - \frac{w_k}{\gamma} \delta^{k+1} + \frac{LC\gamma w_k}{2} \\ &= \frac{w_{k-1}}{\gamma} \delta^k - \frac{w_k}{\gamma} \delta^{k+1} + \frac{LC\gamma w_k}{2}. \end{aligned}$$

Summing up both sides for $k = 0, 1, \ldots, K-1$, and noticing that many terms telescope, we get

$$\frac{1}{2} \sum_{k=0}^{K-1} w_k r^k \leq \frac{w_{-1}}{\gamma} \delta^0 - \frac{w_{K-1}}{\gamma} \delta^K + \frac{LC\gamma}{2} \sum_{k=0}^{K-1} w_k.$$

Rearranging we get the lemma's statement.

Exercises

Exercises I

Exercise 78

Formulate and prove a theorem stating that under some assumptions, the `SGD-NICE` *estimator satisfies the (ABC) assumption. Find values of A, B and C. You can't rely on convexity.*

Exercise 79

Construct a stochastic gradient estimator $g(x)$ for some L-smooth function f which does not *satisfy the (ABC) assumption.*

Exercise 80

Formulate and prove a variant of Theorem 112 in which we do not output the **best** *iterate, but a* **random** *iterate.*

Exercise 81

Formulate and prove a variant of Theorem 112 and Corollary 113 in which we assume that the (ABC) assumption holds with $B = 0$.

Exercise 82

Formulate and prove a variant of Theorem 112 and Corollary 113 in which we assume that the (ABC) assumption holds with $A = C = 0$.

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 18: Nonconvex Setting: the ABC Assumption and the Polyak-Łojasiewicz Inequality

Based on:

[37] Ahmed Khaled and P.R., **Better Theory for SGD in the Nonconvex World**, arXiv:2002.03329, 2020

Plan

Problem: We consider the problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

and assume that the set

$$X^\star \stackrel{\text{def}}{=} \{x^\star \in \mathbb{R}^d \mid f(x^\star) \leq f(x) \; \forall x \in \mathbb{R}^d\}$$

is nonempty. Let $f^\star \stackrel{\text{def}}{=} f(x^\star)$ where $x^\star \in X^\star$.

Plan: Establish convergence of SGD for the following class of functions f:

- f is L-smooth
- f satisfies the **Polyak-Łojasiewicz inequality**

Remark: This is a smaller class of functions than L-smooth functions, and we will be able to prove stronger guarantees.

Two Relaxations of Strong Convexity - I

Recall that if $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and μ-convex, then

$$\mu \|x - y\|^2 \overset{\text{Prop 10}}{\leq} 2D_f(x, y) \overset{\text{Prop 11}}{\leq} \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2 \qquad (261)$$

for all $x, y \in \mathbb{R}^d$. We will consider **relaxed versions of both inequalities.** If we choose $y = x^\star \in X^\star$, then

$$\nabla f(x^\star) = 0, \quad \text{and} \quad D_f(x, x^\star) = f(x) - f^\star.$$

- **Quadratic Growth inequality.** With the choice $y = x^\star$, the first inequality in (261) reduces to

$$\frac{\mu}{2} \|x - x^\star\|^2 \leq f(x) - f^\star.$$

Notice that the left hand side depends on the choice $x^\star \in X^\star$. To make the inequality as weak as possible, we can get rid of the dependence by picking the $x^\star$ which is as close as possible to x:

$$\frac{\mu}{2} \min_{x^\star \in X^\star} \|x - x^\star\|^2 \leq f(x) - f^\star \qquad \forall x \in \mathbb{R}^d$$

Two Relaxations of Strong Convexity - II

If we write

$$\pi(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in X^\star} \|x - y\|^2$$

for the projection onto the set of minima $X^\star$, this can be written compactly as

$$\boxed{\frac{\mu}{2} \|x - \pi(x)\|^2 \leq f(x) - f^\star \qquad \forall x \in \mathbb{R}^d}$$

- **Polyak-Łojasiewicz inequality.** With the choice $y = x^\star$, the second inequality in (261) reduces to

$$\boxed{f(x) - f^\star \leq \frac{1}{2\mu} \|\nabla f(x)\|^2 \qquad \forall x \in \mathbb{R}^d}$$

Example 114

- $f(x) = x^2 + 3\sin^2(x)$ is PL.
- Let $\phi : \mathbb{R}^m \to \mathbb{R}$ be strongly convex and $\boldsymbol{A} \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$. Then $f(x) = \phi(\boldsymbol{A}x + b)$ is PL.

More Relaxations of Strong Convexity - I

Assume $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable.

1. **Strong convexity (SC):**

$$\mu \|x - y\|^2 \leq 2D_f(x, y) \qquad \forall x, y \in \mathbb{R}^d$$

2. **Essential strong convexity (ESC):**

$$\mu \|x - y\|^2 \leq 2D_f(x, y) \qquad \forall x, y : \pi(x) = \pi(y)$$

3. **Weak strong convexity (WSC) / Quasi strong convexity (QSC):**

$$\mu \|\pi(y) - y\|^2 \leq 2D_f(\pi(y), y) \qquad \forall y \in \mathbb{R}^d$$

4. **Restricted secant inequality (RSI) / Quadratic gradient growth (QGG):**

$$\mu \|\pi(y) - y\|^2 \leq \left\langle \underbrace{\nabla f(\pi(y))}_{0} - \nabla f(y), \pi(y) - y \right\rangle \qquad \forall y \in \mathbb{R}^d$$

More Relaxations of Strong Convexity - II

5. **Error bound (EB):**

$$\mu \left\| \pi(y) - y \right\| \leq \left\| \nabla f(y) \right\| \qquad \forall y \in \mathbb{R}^d$$

6. **Polyak-Łojasiewicz (PL):**

$$\underbrace{2D_f(x, \pi(x))}_{2(f(x)-f^\star)} \leq \underbrace{\frac{1}{\mu} \left\| \nabla f(x) - \nabla f(\pi(x)) \right\|^2}_{\frac{1}{\mu}\|\nabla f(x)\|^2} \qquad \forall x \in \mathbb{R}^d$$

7. **Quadratic growth (QG) / Quadratic functional growth (QFG):**

$$\mu \left\| x - \pi(x) \right\|^2 \leq \underbrace{2D_f(x, \pi(x))}_{2(f(x)-f^\star)} \qquad \forall x \in \mathbb{R}^d$$

More Relaxations of Strong Convexity - III

Theorem 115 (Karimi, Nutini and Schmidt 2016 [36])

If f has Lipschitz gradient, then

$$(SC) \Rightarrow (ESC) \Rightarrow (WSC) \Rightarrow (RSI) \Rightarrow (EB) \equiv (PL) \Rightarrow (QG)$$

If we further assume that f is convex, then

$$(RSI) \equiv (EB) \equiv (PL) \equiv (QG)$$

More Relaxations of Strong Convexity - IV

Figure: Chain of inclusions between several function classes. Figure from [49].

Remark: Refer to Karimi, Nutini and Schmidt [36] and Necoara, Nesterov and Glineur [49] for more details.

Polyak-Łojasiewicz Condition

Assumption 10

We say that $f : \mathbb{R}^d \to \mathbb{R}$ *satisfies the* **Polyak-Łojasiewicz (PL) condition/inequality** *if there exists* $\mu > 0$ *such that*

$$f(x) - f^\star \leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \qquad \forall x \in \mathbb{R}^d.$$

Smart Stepsizes for a Key Recursion

We rely on the following lemma where we use the stepsize sequence recently introduced by Stich (2019) [79] but without iterate averaging, as averaging in general may not make sense for nonconvex models.

Lemma 116

Let $a, b, c \geq 0$ with $0 < a \leq b$. Consider a sequence $\{d_t\}_{t \geq 0}$ satisfying

$$d_{t+1} \leq (1 - \gamma_t a)\, d_t + \gamma_t^2 c, \tag{262}$$

where $\gamma_t \leq \frac{1}{b}$ for all $t \geq 0$. Fix $K > 0$ and let $\theta = \lceil \frac{K}{2} \rceil$ and $s = \frac{2b}{a}$. Then choosing the stepsize as

$$\gamma_t = \begin{cases} \frac{1}{b}, & \text{if } K \leq 2(s-1), \\ \frac{1}{b}, & \text{if } K > 2(s-1) \text{ and } t < \theta, \\ \frac{2}{a(s+t-\theta)} & \text{if } K > 2(s-1) \text{ and } t \geq \theta, \end{cases}$$

gives

$$d_K \leq \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{12c}{a^2 K}.$$

Proof of Lemma 116 - I

Step 1. If $K \leq 2(s-1)$, then starting from (262) we have for $\gamma_t \equiv \gamma = \frac{1}{b}$,

$$d_K \quad \leq \quad (1-\gamma a)\, d_{K-1} + \gamma^2 c.$$

Recursing the above inequality, we get

$$\begin{aligned} d_K &\leq (1-\gamma a)^K d_0 + \gamma^2 c \sum_{t=0}^{K-1} (1-\gamma a)^t \\ &\leq (1-\gamma a)^K d_0 + \frac{\gamma c}{a} \\ &\leq \exp(-\gamma a K)\, d_0 + \frac{\gamma c}{a} \\ &= \exp\left(-\frac{aK}{b}\right) d_0 + \frac{c}{ab} \qquad (263) \\ &\leq \exp\left(-\frac{aK}{b}\right) d_0 + \frac{4c}{a^2 K}, \qquad (264) \end{aligned}$$

where in the second inequality we used the fact that $1 - x \leq e^{-x}$, and in last inequality we used that $K \leq 2(s-1) = 2\left(2\frac{b}{a} - 1\right) \leq \frac{4b}{a}$.

Step 2. If $K > 2(s-1)$, then for $t < \theta = \lceil \frac{K}{2} \rceil$ we have

$$\gamma_t \equiv \gamma = \frac{1}{b},$$

Proof of Lemma 116 - II

and hence using the same argument as in **Step 1**, Equation (263), we get

$$\begin{aligned} d_\theta &\leq \exp\left(-\frac{a\theta}{b}\right) d_0 + \frac{c}{ab} \\ &\leq \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{c}{ab}, \end{aligned} \tag{265}$$

where the last inequality follows due to the estimate $\theta \geq \frac{K}{2}$.

Step 3. If $K > 2(s-1)$, then for $t \geq \theta$ we have

$$\gamma_t = \frac{2}{a(s+t-\theta)},$$

and hence

$$d_{t+1} \leq (1-\gamma_t a)\, d_t + \gamma_t^2 c = \frac{s+t-\theta-2}{s+t-\theta} d_t + \frac{4c}{a^2 (s+t-\theta)^2}.$$

Proof of Lemma 116 - III

1. Multiplying both sides by $(s+t-\theta)^2$, we get

$$\begin{aligned}(s+t-\theta)^2\, d_{t+1} &\leq (s+t-\theta-2)\,(s+t-\theta)\, d_t + \frac{4c}{a^2} \\ &= \left((s+t-1-\theta)^2 - 1\right) d_t + \frac{4c}{a^2} \\ &\leq (s+t-\theta-1)^2\, d_t + \frac{4c}{a^2}.\end{aligned}$$

2. Letting $w_t = (s+t-\theta-1)^2$, we get

$$w_{t+1} d_{t+1} \leq w_t d_t + \frac{4c}{a^2}.$$

3. Summing up for $t = \theta, \theta+1, \theta+2, \ldots, K-1$ and telescoping we get,

$$w_K d_K \leq w_\theta d_\theta + \frac{4c\,(K-\theta)}{a^2} = (s-1)^2 d_\theta + \frac{4c\,(K-\theta)}{a^2}.$$

4. Dividing both sides by w_K and using that since $s \geq 2$, then $w_K = (s+K-\theta-1)^2 \geq (2+K-\theta-1)^2 = (K-\theta+1)^2$, we get

$$\begin{aligned}d_K &\leq \frac{(s-1)^2 d_\theta}{w_K} + \frac{4c(K-\theta)}{a^2 w_K} \\ &\leq \frac{(s-1)^2 d_\theta}{(K-\theta+1)^2} + \frac{4c}{a^2(K-\theta+1)}.\end{aligned} \tag{266}$$

Proof of Lemma 116 - IV

5. By the definition of θ we have

$$K - \theta + 1 = K - \left\lceil \frac{K}{2} \right\rceil + 1 \geq K - \left(\frac{K}{2} + 1 \right) + 1 = \frac{K}{2}.$$

6. Plugging this estimate into (266), we get

$$d_K \leq \frac{4(s-1)^2 d_\theta}{K^2} + \frac{8c}{a^2 K}. \tag{267}$$

Step 4: Combining Step 2 and Step 3.

1. Using (265) and the fact that $K > 2(s-1)$, we have

$$\begin{aligned} \frac{4(s-1)^2 d_\theta}{K^2} &\overset{(265)}{\leq} \frac{4(s-1)^2}{K^2} \left(\exp\left(-\frac{aK}{2b}\right) d_0 + \frac{c}{ab} \right) \\ &= \frac{4(s-1)^2}{K^2} \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{4c(s-1)^2}{abK^2} \\ &\leq \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{4c(s-1)^2}{abK^2} \\ &= \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{c}{a^2 K} \cdot \frac{4a(s-1)^2}{bK}. \end{aligned} \tag{268}$$

Proof of Lemma 116 - V

2. The last term in (268) can be bounded as follows. Since $K > 2(s-1)$, we have

$$\frac{4a(s-1)^2}{bK} \leq \frac{2a(s-1)}{b} = \frac{2a\left(2\frac{b}{a}-1\right)}{b} \leq 4.$$

Therefore, by substituting this back into (268), we get

$$\frac{4(s-1)^2 d_\theta}{K^2} \quad \leq \quad \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{4c}{a^2 K}. \tag{269}$$

3. Substituting (269) in (267), we get

$$d_K \quad \leq \quad \exp\left(-\frac{aK}{2b}\right) d_0 + \frac{12c}{a^2 K}. \tag{270}$$

Step 5: Combining Step 4 and Step 1.
It remains to take the maximum of the two bounds (270) and (264).

Convergence of SGD under PL Condition

Using the stepsize scheme of Lemma 116, we can show that SGD finds an optimal global solution at a $\mathcal{O}(1/K)$ rate if $C > 0$ and $\mathcal{O}(\exp(-K))$ rate if $C = 0$, where K is the total number of iterations.

Theorem 117

Assume f is L-smooth and that $X^\star \neq \emptyset$. Suppose that Assumptions 9 (ABC) and 10 (PL) hold. Suppose that SGD is run for $K > 0$ iterations with the stepsize sequence $\{\gamma_k\}_{k\geq 0}$ of Lemma 116 with

$$\gamma_k \leq \min\left\{\frac{\mu}{2LA}, \frac{1}{2LB}\right\}$$

for all $k \geq 0$. Then

$$\mathbf{E}\left[f(x^K) - f^\star\right] \leq \frac{12\kappa_f C}{2\mu K} + \exp\left(-\frac{K}{4\kappa_f \max\{\kappa_S, B\}}\right)\left(f(x^0) - f^\star\right),$$

where $\kappa_f \overset{def}{=} \frac{L}{\mu}$ is the condition number of f and $\kappa_S \overset{def}{=} \frac{A}{\mu}$ is the stochastic condition number.

Proof of Theorem 117 - I

Using L-smoothness of f,

$$\begin{aligned} f(x^{k+1}) - f^\star &\leq f(x^k) - f^\star + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &= f(x^k) - f^\star - \gamma_k \left\langle \nabla f(x^k), g(x^k) \right\rangle + \frac{L\gamma_k^2}{2}\left\|g(x^k)\right\|^2. \end{aligned}$$

Taking expectation conditional on x^k and using the ABC Assumption (Assumption 9), we get

$$\begin{aligned} \mathbf{E}\left[f(x^{k+1}) - f^\star \mid x^k\right] &\leq f(x^k) - f^\star - \gamma_k\left\|\nabla f(x^k)\right\|^2 + \frac{L\gamma_k^2}{2}\mathbf{E}\left[\left\|g(x^k)\right\|^2 \mid x^k\right] \\ &\overset{(\mathrm{ABC})}{\leq} f(x^k) - f^\star - \gamma_k\left(1 - \frac{LB\gamma_k}{2}\right)\left\|\nabla f(x^k)\right\|^2 \\ &\quad + LA\gamma_k^2\left(f(x^k) - f^\star\right) + \frac{LC\gamma_k^2}{2}. \end{aligned}$$

Using $1 - \frac{LB\gamma_k}{2} \geq \frac{3}{4}$ and the PL inequality (Assumption 10), we get

$$\mathbf{E}\left[f(x^{k+1}) - f^\star \mid x^k\right] \leq \left(1 - \frac{3\gamma_k\mu}{2} + LA\gamma_k^2\right)\left(f(x^k) - f^\star\right) + \frac{LC\gamma_k^2}{2}.$$

Proof of Theorem 117 - II

Our choice of stepsize implies $LA\gamma_k \leq \frac{\mu}{2}$, hence

$$\mathbf{E}\left[f(x^{k+1}) - f^\star \mid x^k\right] \quad \leq \quad (1-\gamma_k\mu)\left(f(x^k) - f^\star\right) + \frac{LC\gamma_k^2}{2}.$$

Taking expectation again, using tower property and letting $r^k \stackrel{\text{def}}{=} \mathbf{E}\left[f(x^k) - f^\star\right]$, we have arrived at the recursion

$$r^{k+1} \quad \leq \quad (1-\gamma_k\mu)\, r^k + \frac{LC\gamma_k^2}{2}.$$

Applying Lemma 116 with $a = \mu$, $b = \max\{2\kappa_f A, 2LB\}$, and $c = \frac{LC}{2}$ yields

$$\mathbf{E}\left[f(x^K) - f^\star\right] \quad \leq \quad \exp\left(-\frac{\mu K}{4\max\{\kappa_f A, LB\}}\right)\left(f(x^0) - f^\star\right) + \frac{12LC}{2\mu^2 K}.$$

Using the definition of κ_S yields the theorem's statement.

Corollary: Iteration Complexity

We now obtain $\mathcal{O}(\varepsilon^{-1})$ convergence rate for PL functions in the $C > 0$ regime (which is the optimal dependence on the accuracy ε for non variance-reduced SGD even in the strongly convex case; see Nguyen et al (2019) [55]).

Corollary 118

In the same setting of Theorem 117, fix $\varepsilon > 0$. Let $r^0 \stackrel{def}{=} f(x^0) - f^\star$. Then $\mathbf{E}\left[f(x^K) - f^\star\right] \leq \varepsilon$ as long as

$$K \geq \kappa_f \max\left\{2\kappa_S \log\left(\frac{2r^0}{\varepsilon}\right), 2B\log\left(\frac{2r^0}{\varepsilon}\right), \frac{12C}{2\mu\varepsilon}\right\}.$$

Remarks:

- While the dependence on ε is optimal, the situation is different when we consider the dependence on problem constants, and in particular the dependence on κ_S: Corollary 118 shows a possibly **multiplicative dependence $\kappa_f\kappa_S$.**
- This is different for objectives where we assume convexity, and we show this next.

Exercises

Exercises

Exercise 83

Let $\phi : \mathbb{R}^m \to \mathbb{R}$ *be strongly convex and* $\boldsymbol{A} \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$. *Prove that* $f(x) = \phi(\boldsymbol{A}x + b)$ *satisfies the PL inequality.*

Exercise 84

Prove that PL implies QFG.

Exercise 85

Prove Corollary 118.

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 19: Variance Reduction for Nonconvex SGD

Based on:

[63] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnábas Póczos and Alex Smola, **Stochastic Variance Reduction for Nonconvex Optimization**, ICML 2016

[45] Zhize Li and P.R., **A Unified Analysis of Stochastic Gradient Methods for Nonconvex Federated Optimization**, arXiv:2006.07013, 2020

Introduction

Plan for the Lecture

- We study the non-regularized finite sum problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n f_i(x)\right\},$$

  where the functions f are smooth but not necessarily convex.
- We want to find a (random) point $x\in\mathbb{R}^d$ for which we can guarantee that

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right]\leq\varepsilon^2. \tag{271}$$

- We study the convergence of SGD-CTRL

$$x^{k+1}=x^k-\gamma g^k, \tag{272}$$

  under a **new variant of the σ^k assumption; one that works in the nonconvex setting.**
- We show that **variance reduction helps, and that a variant of L-SVRG-US improves on GD (always) and on the non-variance-reduced method SGD-US (in the not very big n setting).**

Nonconvex σ^k Assumption

Assumption 11

f is lower bounded below, i.e., $f \geq f^{\inf}$. The gradient estimator g^k in `SGD-CTRL` *is unbiased, i.e.,*

$$\mathbf{E}\left[g^k \mid x^k, \xi^k\right] = \nabla f(x^k), \tag{273}$$

and there exist non-negative constants $A, B_1, B_2, C, \tilde{A}, \tilde{B}_1, \tilde{B}_2, \tilde{C}$, with $0 \leq \tilde{B}_1 < 1$, and a random sequence $\{\sigma^k\}_{k\geq 0}$ such that the following two inequalities hold for all $k \geq 0$:

$$\mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, \xi^k\right] \leq 2A(f(x^k) - f^{\inf}) + B_1\sigma^k + B_2\left\|\nabla f(x^k)\right\|^2 + C, \tag{274}$$

$$\mathbf{E}\left[\sigma^{k+1} \mid x^k, \xi^k\right] \leq 2\tilde{A}(f(x^k) - f^{\inf}) + \tilde{B}_1\sigma^k + \tilde{B}_2\left\|\nabla f(x^k)\right\|^2 + \tilde{C}. \tag{275}$$

A New Recursion

Lemma: A New Recursion

In the proof of our main theorem, we will establish a recursion whose solution is described by the next lemma.

Lemma 119

Let $\{X_k\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$ be sequences of nonnegative real numbers which for each $k \geq 0$ satisfy the recursion

$$X_{k+1} \leq aX_k - Y_k + c,$$

where $a > 1$ and $c \geq 0$ are constants. Fix $K \geq 1$ and for each $k = 0, 1, \ldots, K-1$ define probabilities

$$p_k \stackrel{def}{=} \frac{a^{K-(k+1)}}{S_K}, \qquad \text{where} \qquad S_K \stackrel{def}{=} \sum_{k=0}^{K-1} a^{K-(k+1)}.$$

Define random variable Y as follows: $Y = Y_k$ with probability p_k. Then

$$\mathbf{E}[Y] \leq \frac{a^K X_0 - X_K}{S_K} + c \leq \frac{a^K}{S_K} X_0 + c. \tag{276}$$

Lemma 119: Commentary

- Using the inequality $1 + x \leq e^x$, which holds for all $x \geq 0$, and the fact that $S_K \geq K$, we can further bound

$$\frac{a^K}{S_K} \leq \frac{(1 + (a-1))^K}{K} \leq \frac{e^{(a-1)K}}{K}. \tag{277}$$

- When applying Lemma 119, we will have

$$a = 1 + \gamma^2 h \tag{278}$$

for some $h > 0$, where $\gamma > 0$ is a stepsize. This allows us to control the exponential blowup in (277). Indeed, if we limit the stepsize to

$$\gamma \leq \sqrt{\frac{\log 2}{hK}}, \tag{279}$$

then $\gamma^2 hK \leq \log 2$, whence

$$\frac{a^K}{S_K} \overset{(277)}{\leq} \frac{e^{(a-1)K}}{K} \overset{(278)}{\leq} \frac{e^{\gamma^2 hK}}{K} \leq \frac{2}{K}. \tag{280}$$

Proof of Lemma 119

Multiplying the inequality $Y_k \leq aX_k - X_{k+1} + c$ by $a^{K-(k+1)}$, we obtain

$$a^{K-(k+1)} Y_k \leq a^{K-k} X_k - a^{K-(k+1)} X_{k+1} + a^{K-(k+1)} c.$$

Summing up these inequalities for $k = 0, 1, \ldots, K-1$, many terms telescope, and we obtain

$$\begin{aligned} \sum_{k=0}^{K-1} a^{K-(k+1)} Y_k &\leq a^K X_0 - X_K + \sum_{k=0}^{K-1} a^{K-(k+1)} c \\ &= a^K X_0 - X_K + S_K c. \end{aligned}$$

Dividing both sides by S_K, we get

$$\sum_{k=0}^{K-1} p_k Y_k \leq \frac{a^K X_0 - X_K}{S_K} + c.$$

It remains to observe that the left hand side can be interpreted as $\mathbf{E}[Y]$.

Main Theorem

Main Theorem

Theorem 120 (Main Theorem)

Let Assumption 8 (L-smoothness and $f \geq f^{\inf}$) and Assumption 11 (σ^k assumption for nonconvex functions) be satisfied. Choose constant stepsize γ satisfying

$$0 < \gamma \leq \frac{1}{L\left(B_2 + \theta \tilde{B}_2\right)}, \tag{281}$$

where $\theta \overset{def}{=} \frac{B_1}{1-\tilde{B}_1}$. Then for any $K \geq 1$, `SGD-CTRL` *can output a random point x (chosen as one of the points $x^0, x^1, \ldots, x^{K-1}$ at random with certain probabilities) satisfying*

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq L\left(C + \theta \tilde{C}\right)\gamma + \frac{2\left(1 + L\left(A + \theta \tilde{A}\right)\gamma^2\right)^K}{\gamma K}\Delta^0,$$

where $\Delta^0 \overset{def}{=} f(x^0) - f^{\inf} + \frac{1}{2}L\theta\sigma^0\gamma^2$.

Proof of Theorem 120 - I

Step 1: Starting off by following one's nose.

Since f is L-smooth, we have

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &\overset{(272)}{=} f(x^k) - f^{\inf} - \gamma \langle \nabla f(x^k), g^k \rangle + \frac{L\gamma^2}{2}\left\|g^k\right\|^2. \end{aligned} \tag{282}$$

By applying expectation to both sides of (282) and subsequently using unbiasedness of g^k and the assumed bound (274) on the second moment of the stochastic gradient, we get

$$\begin{aligned} \mathbf{E}\left[f(x^{k+1}) - f^{\inf} \mid x^k, \xi^k\right] &\overset{(282)}{\leq} f(x^k) - f^{\inf} - \gamma\left\|\nabla f(x^k)\right\|^2 + \frac{L\gamma^2}{2}\mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, \xi^k\right] \\ &\overset{(274)}{\leq} f(x^k) - f^{\inf} - \gamma\left\|\nabla f(x^k)\right\|^2 \\ &\quad + \frac{L\gamma^2}{2}\left[2A(f(x^k) - f^{\inf}) + B_1\sigma^k + B_2\left\|\nabla f(x^k)\right\|^2 + C\right] \\ &= (1 + LA\gamma^2)(f(x^k) - f^{\inf}) + \frac{LB_1\gamma^2}{2}\sigma^k \\ &\quad - \left(\gamma - \frac{LB_2\gamma^2}{2}\right)\left\|\nabla f(x^k)\right\|^2 + \frac{LC\gamma^2}{2}. \end{aligned} \tag{283}$$

Proof of Theorem 120 - II

Step 2: Trying to set up a recursion.

Choose any $M > 0$ and define

$$\Delta^{k+1} \stackrel{\text{def}}{=} f(x^{k+1}) - f^{\inf} + M\gamma^2\sigma^{k+1}, \tag{284}$$

By combining inequality (283) with inequality (275), we get

$$\begin{aligned}
\mathbf{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right] &\stackrel{(283)+(275)}{\leq} \underbrace{(1 + LA\gamma^2 + 2M\tilde{A}\gamma^2)}_{=a}(f(x^k) - f^{\inf}) + \left(\frac{LB_1}{2} + M\tilde{B}_1\right)\gamma^2\sigma^k \\
&\quad - \underbrace{\left(\gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2\right)}_{=b}\left\|\nabla f(x^k)\right\|^2 + \underbrace{\frac{LC\gamma^2}{2} + M\tilde{C}\gamma^2}_{=c} \\
&= a\left[f(x^k) - f^{\inf} + \frac{\frac{LB_1}{2} + M\tilde{B}_1}{a}\gamma^2\sigma^k\right] - b\left\|\nabla f(x^k)\right\|^2 + c,
\end{aligned} \tag{285}$$

where

$$a \stackrel{\text{def}}{=} 1 + LA\gamma^2 + 2M\tilde{A}\gamma^2, \tag{286}$$

$$b \stackrel{\text{def}}{=} \gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2, \tag{287}$$

$$c \stackrel{\text{def}}{=} \frac{LC\gamma^2}{2} + M\tilde{C}\gamma^2. \tag{288}$$

Proof of Theorem 120 - III

Step 3: Lyapunov function: setting up a recursion.

In order to turn (291) into a recursion which has Δ^k on the right hand side, we need to make sure that

$$\frac{\frac{LB_1}{2} + M\tilde{B}_1}{a} \leq M. \tag{289}$$

Fortunately, it is easy to see (prove this!) that we can make sure this holds by an appropriate choice of M. In particular, inequality (289) holds if we choose

$$M \stackrel{\text{def}}{=} \frac{LB_1}{2(1-\tilde{B}_1)} = \frac{L\theta}{2}. \tag{290}$$

With this choice of M, we can continue from (291) and obtain the recursion

$$\mathbf{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right] \stackrel{(284)+(291)+(289)}{\leq} a\Delta^k - b\left\|\nabla f(x^k)\right\|^2 + c. \tag{291}$$

By applying expectation to both sides of this, and using the tower property of expectation, we get the recursion

$$\begin{aligned} \mathbf{E}\left[\Delta^{k+1}\right] &= \mathbf{E}\left[\mathbf{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right]\right] \\ &\stackrel{(291)}{\leq} a\mathbf{E}\left[\Delta^k\right] - b\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + c. \end{aligned} \tag{292}$$

Proof of Theorem 120 - IV

Step 4: Solving the recursion.

We now apply Lemma 119 to recursion (292) with $X_k = \mathbf{E}\left[\Delta^k\right]$ and $Y_k = b\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right]$. If we set $x = x^k$ with probability p_k (where p_k is as in Lemma 119), which means that $Y = Y_k$ with probability p_k, we conclude that

$$\begin{aligned} b\mathbf{E}\left[\|\nabla f(x)\|^2\right] &= \mathbf{E}\left[Y\right] \\ &\overset{(276)}{\leq} \frac{a^K}{S_K}\Delta^0 + c \qquad (293) \\ &\leq \frac{a^K}{K}\Delta^0 + c, \qquad (294) \end{aligned}$$

where the last inequality follows since $a \geq 1$, which implies that $S_K \geq K$. We now evaluate the expressions for b and c in (294). First,

$$\begin{aligned} b &\overset{(287)}{=} \gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2 \\ &\overset{(290)}{=} \gamma - \frac{\gamma}{2}\left(LB_2\gamma + L\theta\tilde{B}_2\gamma\right) \\ &\geq \frac{\gamma}{2}, \qquad (295) \end{aligned}$$

Proof of Theorem 120 - V

where the last inequality holds by setting

$$\boxed{\gamma \leq \frac{1}{L\left(B_2 + \theta \tilde{B}_2\right)}} \tag{296}$$

Moreover,

$$c \overset{(288)}{=} \frac{LC}{2}\gamma^2 + M\tilde{C}\gamma^2 \overset{(290)}{=} \frac{L}{2}\left(C + \theta\tilde{C}\right)\gamma^2. \tag{297}$$

By plugging the bound (295) on b and expression (297) for c into (294), we obtain the result.

Special Cases: GD and SGD

GD

GD satisfies Assumption 11 with

$$A = 0, B_1 = 0, B_2 = 1, C = 0, \tilde{A} = 0, \tilde{B}_1 = 0, \tilde{B}_2 = 0, \tilde{C} = 0, \sigma^k \equiv 0.$$

In this case, the stepsize bound (281) simplifies to

$$0 < \gamma \leq \frac{1}{L},$$

and Theorem 120 states that

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \frac{2(f(x^0) - f^{\inf})}{K\gamma}. \tag{298}$$

By choosing the maximal stepsize $\gamma = \frac{1}{L}$, we get

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \frac{2L(f(x^0) - f^{\inf})}{K} = \mathcal{O}\left(\frac{1}{K}\right).$$

So,

$$K \geq \frac{2L(f(x^0) - f^{\inf})}{\varepsilon^2} \quad \Rightarrow \quad \mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon^2.$$

SGD Under the ABC Assumption

Assume SGD satisfies the ABC assumption. This is a special case of Assumption 11 with

$$\tilde{A} = 0, \tilde{B}_1 = 0, \tilde{B}_2 = 0, \tilde{C} = 0, \sigma^k \equiv 0.$$

In this case, the stepsize bound (281) simplifies to

$$0 < \gamma \leq \frac{1}{LB_2}, \tag{299}$$

and Theorem 120 states that

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq LC\gamma + \frac{2\left(1 + LA\gamma^2\right)^K}{K\gamma}(f(x^0) - f^{\inf}),$$

Notice that this is the same result (same right hand side) as that of Theorem 112 describing the convergence of SGD under the ABC assumption.

Special Case: L-SVRG

L-SVRG-NICE I

Algorithm 23 Minibatch Loopless SVRG (L-SVRG-NICE)

1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, starting control variable (shift) $y^0 = x^0$, SGD-NICE gradient estimator $g(x)$ with minibatch size τ

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: $\quad g^k = g(x^k, y^k) \overset{\text{def}}{=} g(x^k) - g(y^k) + \nabla f(y^k)$

4: $\quad x^{k+1} = x^k - \gamma g^k$

5: $\quad$ Update control variable (shift):

$$y^{k+1} = \begin{cases} x^{k+1} & \text{with probability} \quad p \\ y^k & \text{with probability} \quad 1-p \end{cases} \tag{300}$$

Remark: This is (almost) a special case of L-SVRG *with* $R \equiv 0$ *and* g *chosen as the* SGD-NICE *estimator. I say "almost" because here we update the control vector* y^{k+1} *to* x^{k+1} *and not* x^k*.*

L-SVRG-NICE II

- The SGD-NICE estimator $g(x)$ has the form

$$g(x) = \frac{1}{\tau} \sum_{i \in S} \nabla f_i(x), \tag{301}$$

where S is a random subset of $\{1, 2, \ldots, n\}$ of cardinality τ chosen uniformly at random.

- The L-SVRG-NICE estimator has the form

$$\begin{aligned} g(x, y) &\overset{\text{def}}{=} g(x) - g(y) + \nabla f(y) \\ &\overset{(301)}{=} \frac{1}{\tau} \sum_{i \in S} [\nabla f_i(x) - \nabla f_i(y)] + \nabla f(y). \end{aligned} \tag{302}$$

A Formula for Sample Variance

Lemma 121

Let $a_1, \ldots, a_n \in \mathbb{R}^d$ *and*

$$\bar{a} \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} a_i. \tag{303}$$

Let $S \subseteq \{1, 2, \ldots, n\}$ *be a random set of cardinality* $\tau \in \{1, 2, \ldots, n\}$ *chosen uniformly at random. Then*

$$\mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S}(a_i - \bar{a})\right\|^2\right] = \frac{n-\tau}{(n-1)\tau}\frac{1}{n}\sum_{i=1}^{n}\|a_i - \bar{a}\|^2 \tag{304}$$

Proof.

Follow similar steps as in the proof of the last claim of Lemma 45. □

Interpretation of Lemma 121 - I

- Let $b_i = a_i - \bar{a}$ for all i. Let b be a **random variable** defined by

$$b = b_i \qquad \text{with probability} \qquad 1/n. \tag{305}$$

- The **mean and variance of b** are given by

$$\mathbf{E}[b] \overset{(305)}{=} \frac{1}{n}\sum_{i=1}^{n} b_i = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a}) = \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) - \bar{a} \overset{(303)}{=} 0 \tag{306}$$

and

$$\mathbf{Var}[b] = \frac{1}{n}\sum_{i=1}^{n} \|b_i - \mathbf{E}[b]\|^2 \overset{(306)}{=} \frac{1}{n}\sum_{i=1}^{n} \|b_i\|^2, \tag{307}$$

respectively.

Interpretation of Lemma 121 - II

- Given a random set $S \subseteq \{1, 2, \ldots, n\}$, the mean of b can be approximated by the **sample mean**

$$b_S \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{i \in S} b_i. \tag{308}$$

Note that

$$\mathbf{E}\left[b_S\right] = 0. \tag{309}$$

- The **sample variance** is given by

$$\mathbf{Var}\left[b_S\right] = \mathbf{E}\left[\left\|b_S - \mathbf{E}\left[b_S\right]\right\|^2\right] \stackrel{(309)}{=} \mathbf{E}\left[\left\|b_S\right\|^2\right] \stackrel{(308)}{=} \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i \in S} b_i\right\|^2\right]. \tag{310}$$

In view of (307) and (310), Lemma 121 says that

$$\mathbf{Var}\left[b_S\right] = \frac{n - \tau}{(n-1)\tau}\mathbf{Var}\left[b\right],$$

i.e., it says that the **sample variance is equal to a certain fraction of the variance.**

A New Assumption

Let

$$a_i(x, y) \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y)$$

and

$$\bar{a}(x, y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} a_i(x, y) = \nabla f(x) - \nabla f(y).$$

Assumption 12

There exists $L_{\text{avg}}^2 \geq 0$ *such that*

$$\frac{1}{n} \sum_{i=1}^{n} \|a_i(x, y) - \bar{a}(x, y)\|^2 \leq L_{\text{avg}}^2 \|x - y\|^2 \tag{311}$$

for all $x, y \in \mathbb{R}^d$.

A Sufficient Condition for Assumption 12

Lemma 122

If f_i has L_i-Lipschitz continuous gradient for all i, then Assumption 12 holds with $L_{\text{avg}}^2 = \frac{1}{n}\sum_{i=1}^n L_i^2$. If additionally f is μ-convex, then Assumption 12 holds with $L_{\text{avg}}^2 = \frac{1}{n}\sum_{i=1}^n L_i^2 - \mu^2$.

Proof.

Note that the left hand side in (311) is the variance of the random variable $a(x,y)$ defined as follows: $a(x,y) = a_i(x,y)$ with probability $1/n$. Since for any random vector X we have the identity

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[\|X\|^2\right] - \|\mathbf{E}\left[X\right]\|^2,$$

we can write

$$\frac{1}{n}\sum_{i=1}^n \|a_i(x,y) - \bar{a}(x,y)\|^2 \quad = \quad \frac{1}{n}\sum_{i=1}^n \|a_i(x,y)\|^2 - \|\bar{a}(x,y)\|^2.$$

It remains to note that if the gradient of f_i is L_i-Lipschitz continuous, then $\|a_i(x,y)\|^2 = \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_i^2 \|x-y\|^2$, and if f is μ-convex, then $\|\bar{a}(x,y)\|^2 = \|\nabla f(x) - \nabla f(y)\|^2 \geq \mu^2 \|x-y\|^2$.

A Tower Property with Three Random Variables

Lemma 123

For three random variables X, Y, Z we have

$$\mathbf{E}\left[\mathbf{E}\left[\mathbf{E}\left[X \mid Y, Z\right] \mid Y\right]\right] = \mathbf{E}\left[X\right]. \tag{312}$$

Proof.

Let $W = \mathbf{E}\left[X \mid Y, Z\right]$. Then by the classical tower property, the left hand side in (312) is equal to

$$\mathbf{E}\left[\mathbf{E}\left[W \mid Y\right]\right] \overset{(44)}{=} \mathbf{E}\left[W\right].$$

By applying the classical tower property again, we get

$$\mathbf{E}\left[W\right] = \mathbf{E}\left[\mathbf{E}\left[X \mid Y, Z\right]\right] \overset{(44)}{=} \mathbf{E}\left[X\right].$$

□

L-SVRG-NICE Satisfies a Variant of the Nonconvex σ^k Assumption

Lemma 124

Suppose that Assumption 12 holds. Let

$$\sigma(x,y) \stackrel{def}{=} \|x-y\|^2 .$$

The L-SVRG-NICE *gradient estimator* (302) *is unbiased, and for each* $\beta > 0$ *satisfies the recursions*

$$\mathbf{E}\left[\|g^k\|^2\right] \leq \underbrace{\alpha}_{B_1} \mathbf{E}\left[\sigma^k\right] + \underbrace{1}_{B_2} \mathbf{E}\left[\|\nabla f(x^k)\|^2\right], \tag{313}$$

$$\mathbf{E}\left[\sigma^{k+1}\right] \leq \underbrace{(1-p)\left(1+\gamma\beta+\gamma^2\alpha\right)}_{\tilde{B}_1} \mathbf{E}\left[\sigma^k\right] + \underbrace{(1-p)\left(\gamma\beta^{-1}+\gamma^2\right)}_{\tilde{B}_2} \mathbf{E}\left[\|\nabla f(x^k)\|^2\right], \tag{314}$$

where $\alpha \stackrel{def}{=} \frac{(n-\tau)L_{\text{avg}}^2}{(n-1)\tau}$ *and* $\sigma^k \stackrel{def}{=} \sigma(x^k,y^k) = \|x^k-y^k\|^2$.

Commentary on Lemma 124

- The recursions (313) and (314) are slightly different from the recursions required by Assumption 11. Indeed, we use full expectations here and not conditional expectations, and as a result, there is no randomness in these inequalities.
- If Assumption 11 holds, then by taking expectation and applying the tower property, we get recursions of this type.
- **It turns out that it is sufficient to assume inequalities of the type (313) and (314) in order to prove Theorem 120 (see Exercise 87).**

Proof of Lemma 124 - I

Step 1: Unbiasedness. We already know that the estimator is unbiased.

Step 2: First recursion. Let is now establish the first recursion. If we let

$$a_i(x,y) \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y)$$

and $\bar{a} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n a_i(x,y) = \nabla f(x) - \nabla f(y)$, then by variance decomposition we can write

$$\begin{aligned}
\mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] &= \mathbf{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2 \mid x^k, y^k\right] + \left\|\nabla f(x^k)\right\|^2 \\
&\stackrel{(302)}{=} \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S^k}\left[\nabla f_i(x^k) - \nabla f_i(y^k)\right] + \nabla f(y^k) - \nabla f(x^k)\right\|^2 \mid x^k, y^k\right] + \left\|\nabla f(x^k)\right\|^2 \\
&= \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S^k}\left[a_i(x^k,y^k) - \bar{a}(x^k,y^k)\right]\right\|^2 \mid x^k, y^k\right] + \left\|\nabla f(x^k)\right\|^2 \\
&\stackrel{(304)}{\leq} \frac{n-\tau}{(n-1)\tau}\frac{1}{n}\sum_{i=1}^n \left\|a_i(x^k,y^k) - \bar{a}(x^k,y^k)\right\|^2 + \left\|\nabla f(x^k)\right\|^2 \\
&\stackrel{(311)}{\leq} \frac{(n-\tau)L_{\text{avg}}^2}{(n-1)\tau}\left\|x^k - y^k\right\|^2 + \left\|\nabla f(x^k)\right\|^2 \\
&= \frac{(n-\tau)L_{\text{avg}}^2}{(n-1)\tau}\sigma^k + \left\|\nabla f(x^k)\right\|^2 . \qquad (315)
\end{aligned}$$

Recursion (313) is obtained by applying expectation again and using the tower property.

Proof of Lemma 124 - II

Step 3: Second recursion: the beginning.

$$
\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] &= \mathbf{E}\left[\left\|x^{k+1} - y^{k+1}\right\|^2 \mid x^{k+1}, x^k, y^k\right] \\
&\overset{(300)}{=} p\left\|x^{k+1} - x^{k+1}\right\|^2 + (1-p)\left\|x^{k+1} - y^k\right\|^2 \\
&= (1-p)\left\|x^{k+1} - y^k\right\|^2 \\
&= (1-p)\left\|x^{k+1} - x^k + x^k - y^k\right\|^2 \\
&= (1-p)\left\|x^{k+1} - x^k\right\|^2 + 2(1-p)\left\langle x^{k+1} - x^k, x^k - y^k\right\rangle + (1-p)\left\|x^k - y^k\right\|^2 \\
&= (1-p)\gamma^2\left\|g^k\right\|^2 + 2(1-p)\left\langle x^{k+1} - x^k, x^k - y^k\right\rangle + (1-p)\sigma^k \\
&= (1-p)\gamma^2\left\|g^k\right\|^2 + 2(1-p)\gamma\left\langle -g^k, x^k - y^k\right\rangle + (1-p)\sigma^k \qquad (316)
\end{aligned}
$$

Proof of Lemma 124 - III

Step 4: Second recursion: the middle game.

Now taking conditional expectation again, but this time conditioning on x^k and y^k only, and applying the inequality

$$\langle u, v\rangle \leq \frac{1}{2\beta}\|u\|^2 + \frac{\beta}{2}\|v\|^2 \tag{317}$$

which holds for any $\beta > 0$ and all $u, v \in \mathbb{R}^d$, we get

$$\begin{aligned}
\mathbf{E}\left[\mathbf{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] \mid x^k, y^k\right] &\overset{(316)}{\leq} (1-p)\gamma^2 \mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + 2(1-p)\gamma\left\langle \underbrace{\mathbf{E}\left[g^k \mid x^k, y^k\right]}_{\nabla f(x^k)}, y^k - x^k\right\rangle \\
&\quad + (1-p)\sigma^k \\
&\overset{(317)}{\leq} (1-p)\gamma^2 \mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + 2(1-p)\gamma\left(\frac{1}{2\beta}\left\|\nabla f(x^k)\right\|^2 + \frac{\beta}{2}\left\|x^k - y^k\right\|^2\right) \\
&\quad + (1-p)\sigma^k \\
&= (1-p)\gamma^2 \mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + \frac{(1-p)\gamma}{\beta}\left\|\nabla f(x^k)\right\|^2 \\
&\quad + (1-p)\left(1+\gamma\beta\right)\sigma^k. \qquad (318)
\end{aligned}$$

Proof of Lemma 124 - IV

Step 5: Second recursion: the end game.

For simplicity, in what follows we denote $\alpha \overset{\text{def}}{=} \frac{(n-\tau)L_{\text{avg}}^2}{(n-1)\tau}$. Applying expectation one more time, and using the more elaborate tower property of expectation (312), we get

$$
\begin{aligned}
\mathbf{E}\left[\sigma^{k+1}\right] &\overset{(312)}{=} \mathbf{E}\left[\mathbf{E}\left[\mathbf{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] \mid x^k, y^k\right]\right] \\
&\overset{(318)}{=} \mathbf{E}\left[(1-p)\gamma^2 \mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + \frac{(1-p)\gamma}{\beta}\left\|\nabla f(x^k)\right\|^2 + (1-p)(1+\gamma\beta)\sigma^k\right] \\
&= (1-p)\gamma^2 \mathbf{E}\left[\mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right]\right] + \frac{(1-p)\gamma}{\beta}\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathbf{E}\left[\sigma^k\right] \\
&\overset{(315)}{=} (1-p)\gamma^2 \mathbf{E}\left[\alpha\sigma^k + \left\|\nabla f(x^k)\right\|^2\right] + \frac{(1-p)\gamma}{\beta}\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathbf{E}\left[\sigma^k\right] \\
&= (1-p)\gamma^2\alpha\mathbf{E}\left[\sigma^k\right] + (1-p)\gamma^2\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + \frac{(1-p)\gamma}{\beta}\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathbf{E}\left[\sigma^k\right] \\
&= (1-p)\left(1+\gamma\beta+\gamma^2\alpha\right)\mathbf{E}\left[\sigma^k\right] + (1-p)\left(\gamma\beta^{-1}+\gamma^2\right)\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right].
\end{aligned}
$$

Rate of L-SVRG - I

1. We will use Theorem 120 to analyze the method. While this theorem requires Assumption 11 to hold, we know (see Exercise 87) that the theorem holds even under an alternative assumption, which is provided by Lemma 124.
2. In view of Lemma 124, we have

$$A = \tilde{A} = C = \tilde{C} = 0, \tag{319}$$

$$B_1 = \alpha, \quad B_2 = 1, \tag{320}$$

$$\tilde{B}_1 = (1-p)(1+\gamma\beta+\gamma^2\alpha), \quad \tilde{B}_2 = (1-p)(\gamma\beta^{-1}+\gamma^2). \tag{321}$$

3. Recall that we have freedom to choose $\beta > 0$. We will choose

$$\beta = \gamma\alpha c, \tag{322}$$

where $c > 0$ is to be determined.
(*Remark:* This is *not* the c from Lemma 119.)

Rate of L-SVRG - II

4. In order for Theorem 120 to be applicable, we need to make sure that $\tilde{B}_1 < 1$. We will achieve this by limiting the stepsize choice to be small enough. In particular, we claim that if

$$\frac{1}{\gamma} \geq \sqrt{\frac{4}{3}\frac{1-p}{p}\alpha(c+1)}, \tag{323}$$

then

$$\tilde{B}_1 \leq 1 - \frac{p}{4}. \tag{324}$$

Indeed,

$$\begin{aligned}
\tilde{B}_1 &= (1-p)(1+\gamma\beta+\gamma^2\alpha) \\
&\overset{(322)}{=} (1-p)(1+\gamma^2\alpha(c+1)) \\
&= 1-p+\gamma^2\alpha(c+1)(1-p) \\
&\overset{(323)}{\leq} 1-p+\frac{3p}{4} \\
&= 1-\frac{p}{4}.
\end{aligned}$$

Rate of L-SVRG - III

5. We can now apply Theorem 120. We know that as long as

$$0 < \gamma \leq \frac{1}{L\left(B_2 + \frac{B_1 \tilde{B}_2}{1-\tilde{B}_1}\right)}, \tag{325}$$

then for the random output x of the algorithm, after K iterations we have the guarantee

$$\begin{aligned} \mathbf{E}\left[\|\nabla f(x)\|^2\right] &\leq L\left(C + \theta\tilde{C}\right)\gamma + \frac{2\left(1 + L\left(A + \theta\tilde{A}\right)\gamma^2\right)^K}{\gamma K}\Delta^0 \\ &\overset{(319)}{=} \frac{2(f(x^0) - f^{\inf})}{\gamma K}. \end{aligned} \tag{326}$$

In the last step we have used the fact that

$$\Delta^0 \overset{\text{def}}{=} f(x^0) - f^{\inf} + \frac{1}{2}L\theta\sigma^0\gamma^2 = f(x^0) - f^{\inf},$$

which follows from $\sigma^0 = 0$, which is true because of the assumption $x^0 = y^0$.

Rate of L-SVRG - IV

6. In view of (326), the rate is optimized if we choose the stepsize as large as possible. Recall that we have two bounds on γ: (323) and (325). So, L-SVRG-NICE has the rate

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \quad \leq \quad \frac{2(f(x^0)-f^{\inf})}{K} \times \underbrace{\max\left\{\underbrace{\sqrt{\frac{4}{3}\frac{1-p}{p}\alpha(c+1)}}_{M_1}, \underbrace{L\left(B_2+\frac{B_1\tilde{B}_2}{1-\tilde{B}_1}\right)}_{M_2}\right\}}_{MAX}.$$

Rate of L-SVRG - V

7. We have

$$
\begin{aligned}
M_2 &= L\left(B_2 + \frac{B_1\tilde{B}_2}{1-\tilde{B}_1}\right) \\
&\overset{(320)}{=} L\left(1 + \frac{\alpha\tilde{B}_2}{1-\tilde{B}_1}\right) \\
&\overset{(324)}{\leq} L\left(1 + \frac{4\alpha\tilde{B}_2}{p}\right) \\
&\overset{(321)}{=} L\left(1 + \frac{4\alpha(1-p)(\gamma\beta^{-1}+\gamma^2)}{p}\right) \\
&\overset{(322)}{=} L\left(1 + \frac{4(1-p)}{p}\left(c^{-1}+\alpha\gamma^2\right)\right) \\
&\overset{(323)}{\leq} L\left(1 + \frac{4(1-p)}{p}\left(c^{-1} + \frac{3p}{4(1-p)(c+1)}\right)\right) \\
&= L\left(1 + \frac{3}{c+1} + \frac{4(1-p)}{pc}\right). \qquad (327)
\end{aligned}
$$

Rate of L-SVRG - VI

8. For simplicity, assume that $\tau = 1$, in which case $\alpha = L_{\text{avg}}^2$. Also, choose $p = \frac{1}{n}$. We have:

$$\begin{aligned} M &= \max\{M_1, M_2\} \\ &= \max\left\{\left(\frac{4(1-p)}{3p}(c+1)\right)^{1/2} L_{\text{avg}}, \left(1 + \frac{3}{c+1} + \frac{4(1-p)}{pc}\right) L\right\} \\ &= \max\left\{\left(\frac{4(n-1)}{3}(c+1)\right)^{1/2} L_{\text{avg}}, \left(1 + \frac{3}{c+1} + \frac{4(n-1)}{c}\right) L\right\}. \end{aligned}$$

If we choose $c = n^{1/3}$, then we get

$$M = \mathcal{O}\left(n^{2/3} \max\{L_{\text{avg}}, L\}\right).$$

Rate of L-SVRG - VII

9. Plugging this expression for M into the bound from Step 6, we get

$$\boxed{\mathbf{E}\left[\|\nabla f(x)\|^2\right] = \mathcal{O}\left(\frac{n^{2/3}\max\{L_{\text{avg}}, L\}(f(x^0) - f^{\inf})}{K}\right)}$$

This should be contrasted with the rate of SGD-US and GD for finding a point x with $\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon^2$:

Method	Rate	# Grad Evals / Iter	Total Complexity
SGD-US	$\mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$	1	$\mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$
GD	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	n	$\mathcal{O}\left(\frac{n}{\varepsilon^2}\right)$
L-SVRG-US	$\mathcal{O}\left(\frac{n^{2/3}}{\varepsilon^2}\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left(\frac{n^{2/3}}{\varepsilon^2}\right)$

Rate of L-SVRG - VIII

Conclusions:

- L-SVRG-US is better than GD.
- L-SVRG-US is better than SGD-US if $n^{2/3} \leq \frac{1}{\varepsilon^2}$, i.e., if n is not too large, or equivalently, if ε is small enough.

10. Variance reduced SGD (in particular, SVRG) in the nonconvex setting was first analyzed by Reddi et al (2016) [63]. This was the first paper that showed the $\mathcal{O}(n^{2/3}/\varepsilon^2)$ rate we proved here as well. We provided a different theorem for a (somewhat) different method, and a different proof.

Exercises

Exercises

Exercise 86

Show that (289) *holds if* M *is chosen as in* (290).

Exercise 87

Consider a variant of Assumption 11 in which relations (274) *and* (275) *are replaced by*

$$\mathbf{E}\left[\left\|g^k\right\|^2\right] \leq 2A\mathbf{E}\left[f(x^k) - f^{\inf}\right] + B_1\mathbf{E}\left[\sigma^k\right] + B_2\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + C,$$

$$\mathbf{E}\left[\sigma^{k+1}\right] \leq 2\tilde{A}\mathbf{E}\left[f(x^k) - f^{\inf}\right] + \tilde{B}_1\mathbf{E}\left[\sigma^k\right] + \tilde{B}_2\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + \tilde{C}.$$

Show that Theorem 120 still holds under this new assumption. That is, provide an alternative proof using this new assumption.

Exercise 88

Show that there exists a value of parameter p *such that* `L-SVRG-NICE` *with minibatch size* $\tau = n^{2/3}$ *achieves total complexity* $\mathcal{O}(n^{2/3}/\varepsilon^2)$.

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 20: PAGE: A Simple Optimal SGD for Nonconvex Problems

Based on:

[54] Lam M Nguyen, Jie Liu, Katya Scheinberg and Martin Takáč, **SARAH: A novel method for machine learning problems using stochastic recursive gradient**, ICML 2017

[32] Samuel Horváth, Lihua Lei, P.R. and Michael I. Jordan, **Adaptivity of stochastic gradient methods for nonconvex optimization**, SIAM J. on Mathematics of Data Science, 2022

[88] Zhize Li, Hongyan Bao, Xiangliang Zhang and P.R., **PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization**, ICML 2021

The Problem

The problem: We consider the finite-sum problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}, \tag{328}$$

where the functions f_i are smooth but **not necessarily convex (i.e., allowed to be nonconvex).**

Goal: Find an ε-stationary solution, i.e., a (possibly random) point $x \in \mathbb{R}^d$ such that

$$\mathbf{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon^2.$$

Main Takeaways I

- We will describe and analyze PAGE: **an optimal stochastic first order algorithm** for solving (328)
- The total complexity of PAGE (i.e., number of iterations multiplied by the number of evaluations of ∇f_i) is

$$\mathcal{O}\left(n^{1/2}/\varepsilon^2\right).$$

This is better than the total complexity of (non variance reduced) SGD, GD and L-SVRG:

Method	Rate	# Grad Evals / Iter	Total Complexity
SGD	$\mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$	1	$\mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$
GD	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	n	$\mathcal{O}\left(\frac{n}{\varepsilon^2}\right)$
L-SVRG-US	$\mathcal{O}\left(\frac{n^{2/3}}{\varepsilon^2}\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left(\frac{n^{2/3}}{\varepsilon^2}\right)$
PAGE ($\tau = 1$)	$\mathcal{O}\left(\frac{n^{1/2}}{\varepsilon^2}\right)$	2	$\mathcal{O}\left(\frac{n^{1/2}}{\varepsilon^2}\right)$
PAGE ($\tau = n^{1/2}$)	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$2n^{1/2}$	$\mathcal{O}\left(\frac{n^{1/2}}{\varepsilon^2}\right)$

Main Takeaways II

- **Original works showing this complexity:** SARAH by Pham et al [56] (2019), SPIDER by Fang et al (2018) [14], SPIDERBoost by Wang et al (2018) [85]
- SC-SARAH of Horváth et al (2020) [32] was the first method using (an implicitly) loopless update rule, as does PAGE
- PAGE of Li et al (2020) [88] is **not** the first optimal SGD method for nonconvex problems. However, PAGE **is very simple, and has a simple analysis.** Plus, [88] resolves some open problems related to existing lower bounds (we will not deal with this in this lecture).

The Algorithm

PAGE: The Algorithm

Algorithm 24 PAGE: ProbAbilistic Gradient Estimator

Parameters: Initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, minibatch size $\tau \in \{1, 2, \ldots, n\}$, probability $0 < p \leq 1$

Initialization: $g^0 = \nabla f(x^0)$

for $k = 0, 1, 2, \ldots$ **do**

  Take SGD step: $x^{k+1} = x^k - \gamma g^k$

  Pick a random set $S^k \subseteq \{1, 2, \ldots, n\}$ of size τ uniformly at random

  Update the gradient estimator:

$$g^{k+1} = \begin{cases} \nabla f(x^{k+1}) & \text{with probability } p \\ g^k + \frac{1}{\tau} \sum\limits_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(x^k)) & \text{with probability } 1 - p \end{cases}$$

Remarks:

- PAGE uses a **recursive update rule for the estimator** (the first recursive estimator is SARAH of Nguyen et al (2017))
- Note that if either $p = 1$ or $\tau = n$, PAGE reduces to GD.

PAGE Estimator vs L-SVRG-NICE Estimator

L-SVRG-NICE:

$$
\begin{aligned}
g^0 &= \nabla f(x^0) = \nabla f(y^0) \\
g^{k+1} &= \begin{cases} \nabla f(x^k) + \frac{1}{\tau} \sum\limits_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(x^k)) & \text{with prob. } p \\ \nabla f(y^k) + \frac{1}{\tau} \sum\limits_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(y^k)) & \text{with prob. } 1-p \end{cases}
\end{aligned}
$$

L-SVRG-NICE (new variant):

$$
\begin{aligned}
g^0 &= \nabla f(x^0) = \nabla f(y^0) \\
g^{k+1} &= \begin{cases} \nabla f(x^{k+1}) & \text{with prob. } p \\ \nabla f(y^k) + \frac{1}{\tau} \sum\limits_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(y^k)) & \text{with prob. } 1-p \end{cases}
\end{aligned}
$$

PAGE:

$$
\begin{aligned}
g^0 &= \nabla f(x^0) \\
g^{k+1} &= \begin{cases} \nabla f(x^{k+1}) & \text{with prob. } p \\ g^k + \frac{1}{\tau} \sum\limits_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(x^k)) & \text{with prob. } 1-p \end{cases}
\end{aligned}
$$

Main Theorem

PAGE: Convergence Theorem

Theorem 125

Assume f is L-smooth, lower bounded by $f^{\inf}$ and suppose that Assumption 12 holds[11]*. Assume $n > 1$, choose minibatch size $\tau \in \{1, 2, \ldots, n\}$, probability $p \in (0, 1]$ and stepsize*

$$0 < \gamma \leq \frac{1}{L + L_{\text{avg}} \sqrt{\frac{(1-p)(n-\tau)}{p(n-1)\tau}}} \overset{def}{=} \gamma_{p,\tau}. \tag{329}$$

Fix $K \geq 1$ and let $\hat{x}^K$ be chosen from the iterates $x^0, x^1, \ldots, x^{K-1}$ of PAGE *uniformly at random. Then*

$$\mathbf{E}\left[\left\|\nabla f(\hat{x}^K)\right\|^2\right] \leq \frac{2(f(x^0) - f^{\inf})}{\gamma K}. \tag{330}$$

[11] There exists $L_{\text{avg}}^2 \geq 0$ such that $\frac{1}{n}\sum_{i=1}^n \|a_i(x,y) - \bar{a}(x,y)\|^2 \leq L_{\text{avg}}^2 \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$, where $a_i(x,y) \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y)$ and $\bar{a}(x,y) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n a_i(x,y) = \nabla f(x) - \nabla f(y)$.

Theorem 125: Commentary - I

1. This is a slightly improved version of a theorem from [88]: the theorem in [88] uses the (typically) larger quantity $L_{\rm avg}$ instead of L, and does not include the beneficial $\frac{n-\tau}{\tau}$ factor.
2. **Iteration complexity.** All other things equal, larger stepsize leads to a better bound (330). Let $\gamma = \gamma_{p,\tau}$. Then $\mathbf{E}\left[\left\|\nabla f(\hat{x}^K)\right\|^2\right] \leq \varepsilon^2$ as long as

$$K \geq K_{p,\tau} \stackrel{\text{def}}{=} \frac{2\left(L + L_{\rm avg}\sqrt{\frac{(1-p)(n-\tau)}{p(n-1)\tau}}\right)(f(x^0) - f^{\rm inf})}{\varepsilon^2} = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right). \qquad (331)$$

- If $p = 1$ or $\tau = n$, the PAGE reduces to GD and $\gamma_{p,\tau} = \frac{1}{L}$.
- We recover iteration complexity of GD in the nonconvex setting:

$$K \geq \frac{2L\delta^0}{\varepsilon^2},$$

where $\delta^0 \stackrel{\text{def}}{=} f(x^0) - f^{\rm inf}$.

Theorem 125: Commentary - II

3. **Number of gradient computations in each iteration.** In each iteration, PAGE computes

$$G_{p,\tau} \stackrel{\text{def}}{=} pn + (1-p)\tau \tag{332}$$

new gradients on average.

- If $p = 1$ or $\tau = n$, then PAGE reduces to GD and $G_{p,\tau} = n$.

4. **Total gradient complexity.** Note that there is a pre-processing step where n gradients $\nabla f_1(x^0), \ldots, \nabla f_n(x^0)$ are computed. Using this and combining (331) and (332), the total (expected) number of gradients computed by PAGE in order to output point $\hat{x}^K$ for $K = K_{p,\tau}$ is

$$\begin{aligned} T_{p,\tau} &\stackrel{\text{def}}{=} n + \lceil K_{p,\tau} - 1 \rceil \, G_{p,\tau} \\ &\stackrel{(331)+(332)}{=} n + \left\lceil \frac{2\left(L + L_{\text{avg}}\sqrt{\frac{(1-p)(n-\tau)}{p(n-1)\tau}}\right)\delta^0}{\varepsilon^2} - 1 \right\rceil (pn + (1-p)\tau) \\ &\leq n + \frac{2\left(L + L_{\text{avg}}\sqrt{\frac{(1-p)(n-\tau)}{p(n-1)\tau}}\right)\delta^0}{\varepsilon^2} (pn + (1-p)\tau), \end{aligned} \tag{333}$$

where $\delta^0 \stackrel{\text{def}}{=} f(x^0) - f^{\text{inf}}$.

Theorem 125: Commentary - III

- If $p = 1$ or $\tau = n$, the PAGE reduces to GD and we recover the total gradient complexity of GD in the nonconvex setting:

$$T_{p,\tau} \overset{(333)}{\leq} n + \frac{2nL\delta^0}{\varepsilon^2}.$$

5. **An "optimal" choice of parameters.** Choose

$$\tau = \sqrt{n} \qquad \text{and} \qquad p = \frac{\tau}{n+\tau}.$$

(For simplicity, ignore the fact that n may not be a square; this is not important.) Then

$$\frac{(1-p)(n-\tau)}{p(n-1)\tau} = \frac{n(n-\tau)}{(n-1)\tau^2} = \frac{n-\tau}{n-1} = \frac{\sqrt{n}}{\sqrt{n}+1} \leq 1$$

and

$$pn + (1-p)\tau = \frac{2\tau n}{\tau n} = \frac{2n}{\sqrt{n}+1} \leq 2\sqrt{n}.$$

Theorem 125: Commentary - IV

Hence, the total complexity can be bounded as follows:

$$T_{p,\tau} \overset{(333)}{\leq} n + \frac{4\sqrt{n}\left(L + L_{\text{avg}}\right)\delta^0}{\varepsilon^2} = \boxed{\mathcal{O}\left(n + \frac{\sqrt{n}}{\varepsilon^2}\right)} \quad (334)$$

6. **Another "optimal" choice of parameters.** Choose

$$\tau = 1 \qquad \text{and} \qquad p = \frac{\tau}{n+\tau}.$$

Then

$$\frac{(1-p)(n-\tau)}{p(n-1)\tau} = n$$

and

$$pn + (1-p)\tau = \frac{n}{n+1} + \frac{n}{n+1} \leq 2.$$

Hence, the total complexity can be bounded as follows:

$$T_{p,\tau} \overset{(333)}{\leq} n + \frac{4\left(L + \sqrt{n}L_{\text{avg}}\right)\delta^0}{\varepsilon^2} = \boxed{\mathcal{O}\left(n + \frac{\sqrt{n}}{\varepsilon^2}\right)} \quad (335)$$

Proofs

Descent Identity I

Lemma 126

Let $x, g \in \mathbb{R}^d$ *and*

$$x^+ \stackrel{def}{=} x - \gamma g, \tag{336}$$

where $\gamma > 0$. *Then for any* $M > 0$ *we have the identity*

$$\langle \nabla f(x), x^+ - x \rangle + \frac{M}{2} \left\| x^+ - x \right\|^2 = -\frac{\gamma}{2} \left\| \nabla f(x) \right\|^2 - \left(\frac{1}{2\gamma} - \frac{M}{2} \right) \left\| x^+ - x \right\|^2 + \frac{\gamma}{2} \left\| g - \nabla f(x) \right\|^2 . \tag{337}$$

Remarks:

- The lemma holds if we replace $\nabla f(x)$ in all three places by any other vector $h \in \mathbb{R}^d$. We wrote it in this form because we will typically use this identity for $h = \nabla f(x)$.

Descent Identity II

- The intended use of this inequality is to give an upper bound on the function value $f(x^+)$ obtained after one step of a gradient-type method of the form (336). In particular, if f is L-smooth, we can apply this identity with $M = L$ and get

$$\begin{aligned} f(x^+) &\leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L}{2} \left\| x^+ - x \right\|^2 \\ &\overset{(337)}{=} f(x) - \frac{\gamma}{2} \left\| \nabla f(x) \right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \left\| x^+ - x \right\|^2 \\ &\quad + \frac{\gamma}{2} \left\| g - \nabla f(x) \right\|^2 . \end{aligned} \tag{338}$$

- **Application of (338) to GD:** For $g = \nabla f(x)$ we get

$$\begin{aligned} f(x^+) &\overset{(338)}{\leq} f(x) - \frac{\gamma}{2} \left\| \nabla f(x) \right\|^2 - \left(\frac{\gamma}{2} - \frac{\gamma^2 L}{2} \right) \left\| \nabla f(x) \right\|^2 \\ &= f(x) - \gamma \left(1 - \frac{\gamma L}{2} \right) \left\| \nabla f(x) \right\|^2 . \end{aligned}$$

Descent Identity III

The right hand side is optimized for the stepsize choice $\gamma = \frac{1}{L}$, leading to the well-known bound

$$f(x^+) \quad \leq \quad f(x) - \frac{1}{2L} \left\| \nabla f(x) \right\|^2 .$$

- **Application of (338) to SGD:** For $x = x^k$, and $g = g^k$ we have $x^+ = x^{k+1}$, and hence

$$\begin{aligned} f(x^{k+1}) \quad \overset{(338)}{\leq} \quad & f(x^k) - \frac{\gamma}{2} \left\| \nabla f(x^k) \right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \left\| x^{k+1} - x^k \right\|^2 \\ & + \frac{\gamma}{2} \left\| g^k - \nabla f(x^k) \right\|^2 . \end{aligned} \tag{339}$$

Proof of Lemma 126

Recall from (336) that $x^+ \stackrel{\text{def}}{=} x - \gamma g$. Letting

$$z \stackrel{\text{def}}{=} x - \gamma \nabla f(x), \tag{340}$$

we can write

$$\begin{aligned}
\langle \nabla f(x), x^+ - x\rangle + \frac{M}{2}\left\|x^+ - x\right\|^2 &= \langle \nabla f(x) - g, x^+ - x\rangle + \langle g, x^+ - x\rangle + \frac{M}{2}\left\|x^+ - x\right\|^2 \\
&\stackrel{(336)}{=} \gamma\langle \nabla f(x) - g, -g\rangle - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&= \gamma\langle \nabla f(x) - g, \nabla f(x) - g - \nabla f(x)\rangle - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&= \gamma\left\|\nabla f(x) - g\right\|^2 - \gamma\langle \nabla f(x) - g, \nabla f(x)\rangle - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&\stackrel{(340)}{=} \gamma\left\|\nabla f(x) - g\right\|^2 - \frac{1}{\gamma}\langle x^+ - z, x - z\rangle - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&= \gamma\left\|\nabla f(x) - g\right\|^2 - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&\qquad - \frac{1}{2\gamma}\left(\left\|x^+ - z\right\|^2 + \left\|x - z\right\|^2 - \left\|x^+ - x\right\|^2\right) \\
&\stackrel{(340)}{=} \gamma\left\|\nabla f(x) - g\right\|^2 - \left(\frac{1}{\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 \\
&\qquad - \frac{1}{2\gamma}\left(\gamma^2\left\|\nabla f(x) - g\right\|^2 + \gamma^2\left\|\nabla f(x)\right\|^2 - \left\|x^+ - x\right\|^2\right) \\
&= \frac{\gamma}{2}\left\|\nabla f(x)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{M}{2}\right)\left\|x^+ - x\right\|^2 + \frac{\gamma}{2}\left\|g - \nabla f(x)\right\|^2.
\end{aligned}$$

Linear Inequality Implying a Quadratic Inequality

In the main proof, we will need to require that the stepsize γ satisfies a quadratic inequality of the form

$$a\gamma^2 + b\gamma \leq 1, \tag{341}$$

where $a > 0$ and $b > 0$ are some constants.

- Note that if (341) holds, then necessarily

$$\gamma \leq \min\left\{\frac{1}{\sqrt{a}}, \frac{1}{b}\right\}.$$

- Solving for the largest gamma which satisfies (341) leads to an inconvenient expression. It will be useful to find a **simple and tight** bound on the stepsize which implies (341). The following lemma provides this:

Lemma 127

If $0 \leq \gamma \leq \frac{1}{\sqrt{a}+b}$, *then* (341) *holds. Moreover, the bound is tight up to the factor of 2 since* $\frac{1}{\sqrt{a}+b} \leq \min\left\{\frac{1}{\sqrt{a}}, \frac{1}{b}\right\} \leq \frac{2}{\sqrt{a}+b}$.

Proof of Theorem 125 - I

STEP 1: Bounding $\mathbf{E}\left[\left\|g^{k+1} - \nabla f(x^{k+1})\right\|^2\right]$.

Recall the definition of the PAGE gradient estimator (see Line 6 in Algorithm 24):

$$g^{k+1} = \begin{cases} \nabla f(x^{k+1}) & \text{with probability } p, \\ g^k + \frac{1}{\tau} \sum_{i \in S^k} (\nabla f_i(x^{k+1}) - \nabla f_i(x^k)) & \text{with probability } 1-p. \end{cases} \tag{342}$$

A direct calculation now reveals that

$$\begin{aligned} G &\overset{\text{def}}{=} \mathbf{E}\left[\left\|g^{k+1} - \nabla f(x^{k+1})\right\|^2 \mid x^{k+1}, x^k, g^k, S^k\right] \\ &\overset{(342)}{=} p \underbrace{\left\|\nabla f(x^{k+1}) - \nabla f(x^{k+1})\right\|^2}_{=0} + (1-p) \left\| g^k + \frac{1}{\tau} \sum_{i \in S^k} \left(\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\right) - \nabla f(x^{k+1}) \right\|^2 \\ &= (1-p) \left\| \underbrace{g^k - \nabla f(x^k)}_{X} + \frac{1}{\tau} \sum_{i \in S^k} \underbrace{\left(\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\right)}_{a_i} - \underbrace{\left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)}_{\bar{a} = \frac{1}{n}\sum_i a_i} \right\|^2 \\ &= (1-p)\|X\|^2 + 2(1-p)\left\langle X, \frac{1}{\tau} \sum_{i \in S^k} a_i - \bar{a} \right\rangle + (1-p) \left\| \frac{1}{\tau} \sum_{i \in S^k} a_i - \bar{a} \right\|^2. \end{aligned} \tag{343}$$

Proof of Theorem 125 - II

Since $\mathbf{E}\left[\frac{1}{\tau}\sum_{i\in S^k} a_i - \bar{a} \mid x^{k+1}, x^k, g^k\right] = 0$ (make sure you see why) and because X is constant conditioned on x^{k+1}, x^k, g^k, we have

$$\begin{aligned}
\mathbf{E}\left[G \mid x^{k+1}, x^k, g^k\right] &\overset{(343)}{=} (1-p)\mathbf{E}\left[\|X\|^2 \mid x^{k+1}, x^k, g^k\right] + (1-p)\mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S^k} a_i - \bar{a}\right\|^2 \mid x^{k+1}, x^k, g^k\right] \\
&\overset{(304)}{=} (1-p)\mathbf{E}\left[\|X\|^2 \mid x^{k+1}, x^k, g^k\right] + (1-p)\frac{n-\tau}{(n-1)\tau}\frac{1}{n}\sum_{i=1}^n \|a_i - \bar{a}\|^2 \\
&\overset{(311)}{\leq} (1-p)\mathbf{E}\left[\|X\|^2 \mid x^{k+1}, x^k, g^k\right] + (1-p)\frac{n-\tau}{(n-1)\tau}L_{\text{avg}}^2\left\|x^{k+1} - x^k\right\|^2, \qquad (344)
\end{aligned}$$

where in the second step we have applied the identity from Lemma 121), and the last step is due to the smoothness assumption (Assumption 12).

Proof of Theorem 125 - III

By first applying the three term tower property (312) and subsequently the standard two-term tower property, we get

$$
\begin{aligned}
\mathbf{E}\left[\left\|g^{k+1}-\nabla f(x^{k+1})\right\|^2\right] &\overset{(312)}{=} \mathbf{E}\left[\mathbf{E}\left[\underbrace{\mathbf{E}\left[\left\|g^{k+1}-\nabla f(x^{k+1})\right\|^2 \mid x^{k+1},x^k,g^k,S^k\right]}_{G} \mid x^{k+1},x^k,g^k\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[G \mid x^{k+1},x^k,g^k\right]\right] \\
&\overset{(344)}{\leq} \mathbf{E}\left[(1-p)\mathbf{E}\left[\|X\|^2 \mid \mid x^{k+1},x^k,g^k\right]+(1-p)\frac{n-\tau}{(n-1)\tau}L_{\mathrm{avg}}^2\left\|x^{k+1}-x^k\right\|^2\right] \\
&= (1-p)\mathbf{E}\left[\mathbf{E}\left[\|X\|^2 \mid \mid x^{k+1},x^k,g^k\right]\right]+(1-p)\frac{n-\tau}{(n-1)\tau}L_{\mathrm{avg}}^2\mathbf{E}\left[\left\|x^{k+1}-x^k\right\|^2\right] \\
&= (1-p)\mathbf{E}\left[\|X\|^2\right]+(1-p)\frac{n-\tau}{(n-1)\tau}L_{\mathrm{avg}}^2\mathbf{E}\left[\left\|x^{k+1}-x^k\right\|^2\right] \\
&= (1-p)\mathbf{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]+(1-p)\frac{n-\tau}{(n-1)\tau}L_{\mathrm{avg}}^2\mathbf{E}\left[\left\|x^{k+1}-x^k\right\|^2\right]. \qquad (345)
\end{aligned}
$$

STEP 2: Bounding $\mathbf{E}\left[f(x^{k+1})-f^{\inf}\right]$.

Subtracting $f^{\inf}$ from both sides of (339) and taking expectation, we get

$$
\begin{aligned}
\mathbf{E}\left[f(x^{k+1})-f^{\inf}\right] \leq\ & \mathbf{E}\left[f(x^k)-f^{\inf}\right]-\frac{\gamma}{2}\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbf{E}\left[\left\|x^{k+1}-x^k\right\|^2\right] \\
&+\frac{\gamma}{2}\mathbf{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]. \qquad (346)
\end{aligned}
$$

Proof of Theorem 125 - IV

STEP 3: Combining the two bounds.

Let $\delta^k \overset{\text{def}}{=} \mathbf{E}\left[f(x^k) - f^{\inf}\right]$, $s^k \overset{\text{def}}{=} \mathbf{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$ and $r^k \overset{\text{def}}{=} \mathbf{E}\left[\left\|x^{k+1} - x^k\right\|^2\right]$.

Then by adding (346) with a $\frac{\gamma}{2p}$ multiple of (345), we obtain

$$\begin{aligned}
\delta^{k+1} + \frac{\gamma}{2p}s^{k+1} &\leq \delta^k - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)r^k + \frac{\gamma}{2}s^k + \frac{\gamma}{2p}\left(\frac{(1-p)(n-\tau)L_{\text{avg}}^2}{(n-1)\tau}r^k + (1-p)s^k\right) \\
&= \delta^k + \frac{\gamma}{2p}s^k - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{(1-p)(n-\tau)\gamma L_{\text{avg}}^2}{2p(n-1)\tau}\right)r^k \\
&\leq \delta^k + \frac{\gamma}{2p}s^k - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 . \qquad (347)
\end{aligned}$$

The last inequality follows from the bound $\frac{(1-p)(n-\tau)L_{\text{avg}}^2}{p(n-1)\tau}\gamma^2 + L\gamma \leq 1$, which holds because of Lemma 127 and our assumption (329) on the stepsize.

STEP 4: Solving the recursion.

By summing up inequalities (347) for $k = 0, \ldots, K-1$, we get

$$0 \leq \delta^K + \frac{\gamma}{2p}s^K \leq \delta^0 + \frac{\gamma}{2p}s^0 - \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right]. \qquad (348)$$

Proof of Theorem 125 - V

Multiplying both sides by $\frac{2}{\gamma K}$ and using the fact that $s^0 = 0$ (which follows from our choice $g^0 = \nabla f(x^0)$), after rearranging we get

$$\sum_{k=0}^{K-1} \frac{1}{K} \mathbf{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{2\delta^0}{\gamma K}. \tag{349}$$

It remains to notice that the left hand side can be interpreted as $\mathbf{E}\left[\left\|\nabla f(\hat{x}^K)\right\|^2\right]$, where $\hat{x}^K$ is chosen from $x^0, x^1, \ldots, x^{K-1}$ uniformly at random.

Exercises

Exercises I

Exercise 89

Prove that if either $p = 1$ or $\tau = n$, PAGE reduces to GD.

Exercise 90

Prove Lemma 127.

Exercise 91

Show that if there exists $L_{\mathrm{avg}} > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_{\mathrm{avg}}^2\|x-y\|^2, \quad \forall x, y \in \mathbb{R}^d,$$

then f has L-Lipschitz gradient with $L \leq L_{\mathrm{avg}}$, and hence it is also L-smooth.

Exercise 92

Formulate a convergence theorem for a variant of PAGE in which the initial gradient estimator g^0 is allowed to be arbitrary. Can we benefit theoretically from the choice $g^0 = 0$?

Exercises II

Exercise 93

Show that that if f satisfies the PL inequality with constant $\mu > 0$, then the rate of `PAGE` *is* $\mathcal{O}\left(\left(n + \sqrt{n}\kappa\right)\log\frac{1}{\epsilon}\right)$, *where* $\kappa \stackrel{def}{=} \frac{L'}{\mu}$ *is a condition number and* L' *is some smoothness constant (e.g., a combination of* L *and* L_{avg}*).*

Exercise 94

Establish a convergence theorem for `PAGE` *in the strongly convex setting.*

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 21: EF21: Dealing with Contractive Compressors via Error Feedback

Based on:

[75] Frank Seide, Hao Fu, Jasha Droppo, Gang Li and Dong, Yu, **1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs**, ISCA 2014

[64] P.R., Igor Sokolov and Ilyas Fatkhullin, **EF21: A new, simpler, theoretically better, and practically faster error feedback**, NeurIPS 2021

[15] Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li and P.R., **EF21 with bells & whistles: practical algorithmic extensions of modern error feedback**, arXiv preprint arXiv:2110.03294, 2021

Introduction - I

We are still interested in **distributed nonconvex optimization:**

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

Our goal is still to develop an efficient method relying **communication compression.**

The new:

- We now want to be able to work with a **new class of communication compression operators** we did not consider before!
- In particular, we will work with the new class of **contractive compressors,** which include operators such as
  - Top-K (**greedy sparsification**), and
  - Rank-K (**low rank approximation**).

Key advantage:

- **Contractive compressors are very powerful in practice,** and often (but not always) beat unbiased compressors.

Introduction - II

Key difficulties:

- **Contractive compressors are necessarily biased,** and this causes **difficulties in theoretical analysis.**
- When contractive compressors are used, `DCGD` **can diverge (exponentially!) in the $n > 1$ regime, even on strongly convex problems!** A simple example with $n = d = 3$ was provided by Beznosikov, Horváth, R. and Safaryan [5] (see Example 131).
- To fix this problem, one therefore needs to modify `DCGD`.
  - The first mechanism proposed to fix this issue is due to Seide et al (2014) [75], and is known under the name **Error Feedback (`EF`)**, a.k.a. **Error Compensation (`EC`).** `EF` was first analyzed in about 2018. However, **all known analyses of `EF` have major issues (see [64]):**
    - strong (and sometimes unreasonable) assumptions,
    - bad rates.

Introduction - III

- An alternative mechanism based on the idea of **transforming any contractive compressor $\mathcal{C}$ into an unbiased "induced" compressor** $\mathcal{Q}_{\mathcal{C}}$ was recently proposed by Horváth and R. (ICLR 2020) [33]. The induced compressor is

$$\mathcal{Q}_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \mathcal{C}(x) + \mathcal{Q}(x - \mathcal{C}(x)),$$

  where $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is a carefully chosen **unbiased** compressor.
  - Note that $\mathcal{Q}_{\mathcal{C}}$ is unbiased.
  - This allows to tap into analyses techniques we have seen before.

  We will not talk about this in this lecture.
- There is no theory (in the $n > 1$ regime) that would satisfactorily explain why and when contractive compressors can be better than unbiased compressors.

Plan

- We will cover a **new Error Feedback mechanism from 2021 called** EF21 due to R., Sokolov and Fatkhullin (NeurIPS 2021) [64].
- EF21 currently enjoys the **best rates** among all (nonaccelerated) methods that use **contractive compressors.**
  - However, EF21 **does not have better theory than** MARINA.
  - Convergence proof for EF21 is very very similar to that of MARINA, but in a certain key step one "pays" for the lack of unbiasedness.
  - On the other hand, EF21 **can often be better than** MARINA **in practice.**
- There is a large gap between theory and practice here!

Contractive Compressors

Definition 128 (Contractive compressor)

A (possibly) randomized mapping $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a **contractive compressor** if there exists $0 < \alpha \leq 1$ such that

$$\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\alpha)\|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

Examples:

- Any unbiased compressor with variance parameter ω (i.e., any $\mathcal{C} \in \mathbb{B}^d(\omega)$) is contractive **after appropriate scaling.**
- Top-K (a **greedy** version of the Rand-K sparsifier)
- Rank-K (best **rank-K approximation** of x)

Compressors in $\mathbb{B}^d(\omega)$ are Contractive after Scaling

Lemma 129

Let $\mathcal{C} \in \mathbb{B}^d(\omega)$ for some $\omega \geq 0$. Then

$$\alpha\mathcal{C} \in \mathbb{C}^d(\alpha),$$

where $\alpha \overset{def}{=} \frac{1}{\omega+1}$.

Proof.

First recall that for all $x \in \mathbb{R}^d$, we have

$$\mathbf{E}\left[\mathcal{C}(x)\right] = x, \quad \text{and} \quad \mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (\omega+1)\|x\|^2. \tag{350}$$

Now,

$$\begin{aligned}
\mathbf{E}\left[\|\alpha\mathcal{C}(x) - x\|^2\right] &= \mathbf{E}\left[\alpha^2\|\mathcal{C}(x)\|^2 - 2\langle\alpha\mathcal{C}(x), x\rangle + \|x\|^2\right] \\
&= \alpha^2 \underbrace{\mathbf{E}\left[\|\mathcal{C}(x)\|^2\right]}_{\overset{(350)}{\leq}(\omega+1)\|x\|^2} - 2\langle\alpha \underbrace{\mathbf{E}\left[\mathcal{C}(x)\right]}_{\overset{(350)}{=}x}, x\rangle + \|x\|^2 \\
&= (\alpha - 2\alpha + 1)\|x\|^2 = (1-\alpha)\|x\|^2.
\end{aligned}$$

□

Top-K: A Greedy Sparsification Strategy

Top-K: A Greedy Sparsifier

Top-K is an operator mapping vectors from $\mathbb{R}^d$ to vectors from $\mathbb{R}^d$, keeping the top K (in absolute value) entries[12] of the input vector, and zeroing out the rest.

Definition 130 (Top-K)

Choose $1 \leq K \leq d$. Then Top-K is the operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ defined as follows. Let $x \in \mathbb{R}^d$ and let $\pi = (\pi_1, \ldots, \pi_d)$ be a permutation of $(1, \ldots, d)$ such that

$$|x_{\pi_1}| \geq |x_{\pi_2}| \geq \cdots \geq |x_{\pi_d}|.$$

Then $\mathcal{C}(x)$ is defined coordinate-wise as follows:

$$(\mathcal{C}(x))_{\pi_i} \overset{\text{def}}{=} \begin{cases} x_{\pi_i} & 1 \leq i \leq K \\ 0 & \text{otherwise} \end{cases}.$$

Remark: **Top-K is not an unbiased compressor!**

[12] In case of ties, ties are broken arbitrarily.

Failure of DCGD with Biased Compressors

Failure of DCGD with Biased Compressors

Example 131 (Beznosikov, Horváth, R. and Safaryan [5])

Consider $n = d = 3$ and define

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|^2, \quad f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|^2, \quad f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|^2,$$

where $a = (-3, 2, 2)$, $b = (2, -3, 2)$, and $c = (2, 2, -3)$. Let the starting iterate be $x^0 = (t, t, t)$, where $t > 0$. Then

$$\nabla f_1(x^0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x^0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x^0) = \frac{t}{2}(9, 9, -11).$$

Using the Top-1 compressor, we get

$$\mathcal{C}(\nabla f_1(x^0)) = \frac{t}{2}(-11, 0, 0), \quad \mathcal{C}(\nabla f_2(x^0)) = \frac{t}{2}(0, -11, 0), \quad \mathcal{C}(\nabla f_3(x^0)) = \frac{t}{2}(0, 0, -11).$$

The next iterate of DCGD is

$$x^1 = x^0 - \gamma\frac{1}{3}\sum_{i=1}^{3}\mathcal{C}(\nabla f_i(x^0)) = \left(1 + \frac{11\eta}{6}\right)x^0.$$

Repeated application gives

$$x^k = \left(1 + \frac{11\eta}{6}\right)^k x^0.$$

Since $\eta > 0$, the entries of x^k diverge exponentially fast to $+\infty$.

Rank-K: Compression via Low Rank Approximation

Reshaping of Vectors into Matrices

- To define the Rank-K compressor as a mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$, we will need to **reshape the vector $x \in \mathbb{R}^d$ into a matrix $\boldsymbol{X} \in \mathbb{R}^{a \times b}$**, where $d = a \times b$.
- Clearly, given such a representation of x, we have

$$r \stackrel{\text{def}}{=} \mathrm{Rank}(\boldsymbol{X}) \leq \min\{a, b\}.$$

Example 132

Let

$$x = \begin{pmatrix} 1 \\ 3 \\ -5 \\ 7 \\ 0 \\ 2 \end{pmatrix} \in \mathbb{R}^6$$

and choose $a = 2$ and $b = 3$. Then we can represent x as the matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & -5 \\ 7 & 0 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 3}.$$

Singular Value Decomposition (SVD)

Theorem 133 (Singular Value Decomposition)

Let $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ *and* $r = \mathrm{Rank}(\boldsymbol{X})$. *Then there exist* **orthonormal matrices**[13] $\boldsymbol{U} \in \mathbb{R}^{a \times a}$, $\boldsymbol{V} \in \mathbb{R}^{b \times b}$ *and a* **diagonal** *matrix*

$$\boldsymbol{S} \in \mathbb{R}^{a \times b} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r, \underbrace{0, \ldots, 0}_{\min\{a,b\} - r \text{ zeros}})$$

such that

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ *are the* **positive singular values** *of* $\boldsymbol{X}$.

Remark: The remaining $\min\{a, b\} - r$ diagonal entries of $\boldsymbol{S}$ are all zeroes. These are the **zero singular values** of $\boldsymbol{X}$.

[13] A square matrix $\boldsymbol{M} \in \mathbb{R}^{v \times v}$ is orthonormal if its rows and columns are orthonormal vectors, i.e., if $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{M}^\top \boldsymbol{M} = \boldsymbol{I}$.

Best Low Rank Approximation of Matrices

Theorem 134 (Eckart-Young-Mirsky)

Choose $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ *and* $1 \leq K < r$, *where* $r = \mathrm{Rank}(\boldsymbol{X})$.

(i) *Then* **the best rank-K approximation** *of* $\boldsymbol{X}$ *is the matrix*

$$\boldsymbol{X}_K \stackrel{def}{=} \sum_{i=1}^{K} \sigma_i u_i v_i^\top . \tag{351}$$

That is, $\boldsymbol{X}_K$ *is a solution to the optimization problem*[14]

$$\min_{\boldsymbol{Y} \in \mathbb{R}^{a \times b}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\|_{\mathrm{F}} \ : \ \mathrm{Rank}(\boldsymbol{Y}) \leq K \right\}, \tag{352}$$

where $\|\cdot\|_{\mathrm{F}}$ *denotes the Frobenius norm.*

(ii) *Moreover,* **the Frobenius distance between $\boldsymbol{X}$ and its best rank-K approximation $\boldsymbol{X}_K$** *is given by*

$$\|\boldsymbol{X}_K - \boldsymbol{X}\|_{\mathrm{F}} = \left\| \sum_{i=K+1}^{r} \sigma_i u_i v_i^\top \right\|_{\mathrm{F}} = \sqrt{\sum_{i=K+1}^{r} \sigma_i^2}. \tag{353}$$

[14] One can replace the Frobenius norm by the **spectral norm,** and the statement (i) would still be true.

SVD and Frobenius Norm

The second equality in (353) follows from the following lemma.

Lemma 135

Let

- $\boldsymbol{U} \in \mathbb{R}^{a \times a}$ *and* $\boldsymbol{V} \in \mathbb{R}^{b \times b}$ *be orthonormal matrices,*
- I *be a nonempty subset of* $\{1, 2, \ldots, \min\{a, b\}\}$*, and*
- $\{\lambda_i \ : \ i \in I\}$ *be nonnegative real numbers.*

Then

$$\left\| \sum_{i \in I} \lambda_i u_i v_i^\top \right\|_{\mathrm{F}} = \sqrt{\sum_{i \in I} \lambda_i^2}. \tag{354}$$

Proof.

Solve this as an exercise. □

Rank-K Compressor

Remark: Let $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ and $r = \mathrm{Rank}(\boldsymbol{X})$.

- If $K < r$, then the best rank-K approximation of $\boldsymbol{X}$ is $\boldsymbol{X}_K$ given by (351).
- If $r \leq K$, then trivially the best rank-K approximation of $\boldsymbol{X}$, i.e., the solution of (352), is $\boldsymbol{X}$.

Definition 136 (Rank-K compressor)

Let $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ and $r = \mathrm{Rank}(\boldsymbol{X})$. Choose $1 \leq K \leq \min\{a, b\}$. Then the Rank-$K$ compressor is the mapping $\mathcal{C} : \mathbb{R}^{a \times b} \to \mathbb{R}^{a \times b}$ defined by

$$\mathcal{C}(\boldsymbol{X}) \stackrel{\text{def}}{=} \begin{cases} \boldsymbol{X}_K \stackrel{(352)}{=} \sum_{i=1}^{K} \sigma_i u_i v_i^\top, & K < r, \\ \boldsymbol{X}, & K \geq r, \end{cases} \tag{355}$$

where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{X}$.

Rank-K Compressor: Communication Savings

- Instead of sending $d = ab$ floats (entries of $\boldsymbol{X}$), we send the $2K$ vectors $\{(\tilde{u}_i, \tilde{v}_i) = (\sqrt{\sigma_i} u_i, \sqrt{\sigma_i} v_i) \in \mathbb{R}^a \times \mathbb{R}^b \ : \ i = 1, \ldots, K\}$. So, we send

$$K(a+b)$$

  floats.
- Note that $K \leq r = \mathrm{Rank}(\boldsymbol{X}) \leq \min\{a, b\}$, and hence

$$K(a+b) \leq \min\{a, b\}(a+b) = \min\{a^2+ab, b^2+ab\} = d+\min\{a^2, b^2\},$$

  which is larger than $d = ab$.
- So, one should choose K small enough:

$$K \ll \frac{ab}{a+b}.$$

  Say we choose $a = b = \sqrt{d}$. Then we want to choose

$$K \ll \frac{ab}{a+b} = \frac{d}{2\sqrt{d}} = \frac{\sqrt{d}}{2}.$$

Rank-K is Contractive

We now show that the Rank-K compressor is contractive, i.e., that it belongs to $\mathbb{C}^d(\alpha)$ for $\alpha = \frac{K}{\min\{a,b\}}$.

Theorem 137 (Rank-K compressor is contractive)

Let $\mathcal{C}(\cdot)$ be the Rank-K compressor, where $1 \leq K \leq \min\{a, b\}$. Then

$$\|\mathcal{C}(\boldsymbol{X}) - \boldsymbol{X}\|_{\mathrm{F}}^2 \leq \left(1 - \frac{K}{\min\{a, b\}}\right) \|\boldsymbol{X}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{X} \in \mathbb{R}^{a \times b}.$$

Remark: The above lemma is a straightforward consequence of the properties of SVD. It was explicitly stated/used/proved in the context of communication compression in [71].

Proof of Theorem 137 - I

Fix $\boldsymbol{X} \in \mathbb{R}^{a \times b}$ and let $r = \mathrm{Rank}(\boldsymbol{X})$.

Step 0 (Handling the easy case). If $K \geq r$, then since (in view of (355)) $\mathcal{C}(\boldsymbol{X}) = \boldsymbol{X}$, we have $\|\mathcal{C}(\boldsymbol{X}) - \boldsymbol{X}\|_{\mathrm{F}}^2 = 0$, and hence the statement of the theorem holds.

Step 1. If $K < r$, then in view of (355) and (353), we have

$$\|\mathcal{C}(\boldsymbol{X}) - \boldsymbol{X}\|_{\mathrm{F}}^2 \overset{(355)}{=} \left\| \sum_{i=K+1}^{r} \sigma_i u_i v_i^\top \right\|_{\mathrm{F}}^2 \overset{(353)}{=} \sum_{i=K+1}^{r} \sigma_i^2. \tag{356}$$

Moreover, in view of Lemma 135,

$$\|\boldsymbol{X}\|_{\mathrm{F}}^2 = \left\| \sum_{i=1}^{r} \sigma_i u_i v_i^\top \right\|_{\mathrm{F}}^2 = \sum_{i=1}^{r} \sigma_i^2. \tag{357}$$

Step 2. Next, since $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, the average of the tail values $\sigma_{K+1}^2, \sigma_{K+2}^2, \ldots, \sigma_r^2$ can't be larger than the average of all the values $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$, and thus

$$\frac{1}{r-K} \sum_{i=K+1}^{r} \sigma_i^2 \leq \frac{1}{r} \sum_{i=1}^{r} \sigma_i^2. \tag{358}$$

Proof of Theorem 137 - II

Step 3. This means that

$$\begin{aligned}
\|\mathcal{C}(\boldsymbol{X}) - \boldsymbol{X}\|_{\mathrm{F}}^2 &\overset{(356)}{=} \sum_{i=K+1}^{r} \sigma_i^2 \\
&\overset{(358)}{\leq} \frac{r-K}{r} \sum_{i=1}^{r} \sigma_i^2 \\
&\overset{(357)}{=} \left(1 - \frac{K}{r}\right) \|\boldsymbol{X}\|_{\mathrm{F}}^2 \\
&\leq \left(1 - \frac{K}{\min\{a,b\}}\right) \|\boldsymbol{X}\|_{\mathrm{F}}^2 ,
\end{aligned}$$

where in the last step we have used the fact that $r \leq \min\{a, b\}$.

EF21

EF21: The Algorithm (Single Node)

Algorithm 25 EF21 ($n = 1$): Error Feedback, 2021 version

1: **Parameters:** Initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, contractive compressor $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$
2: **Initialization:** $g^0 = \mathcal{C}(\nabla f(x^0))$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: Take SGD step: $x^{k+1} = x^k - \gamma g^k$
5: Update the gradient estimator:

$$g^{k+1} = g^k + \mathcal{C}(\nabla f(x^{k+1}) - g^k)$$

EF21: The Algorithm (Multiple Nodes)

Algorithm 26 EF21 (general n): Error Feedback, 2021 version

Parameters: Initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, contractive compressor $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$

Initialization on workers: Compute $g_i^0 = \mathcal{C}(\nabla f_i(x^0))$ for all $i \in [n]$, and communicate these vectors to the server

Initialization on server: Compute $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$

for $k = 0, 1, 2, \ldots$ **do**

  Take SGD step: $x^{k+1} = x^k - \gamma g^k$

  for nodes $i = 1, 2, \ldots, n$ in parallel **do**

    Compute the local gradient $\nabla f_i(x^{k+1})$

    Communicate the compressed message to the server:

$$m_i^k = \mathcal{C}(\nabla f_i(x^{k+1}) - g_i^k)$$

    Update the local gradient estimator: $g_i^{k+1} = g_i^k + m_i^k$

  Compute $g^{k+1} = \frac{1}{n}\sum_{i=1}^n g_i^{k+1}$ as follows: $g^{k+1} = g^k + \frac{1}{n}\sum_{i=1}^n m_i^k$

Theory for EF21

Assumptions

Assumption 13 (Global loss is lower bounded)

$f^{\inf} \stackrel{def}{=} \inf_{x \in \mathbb{R}^d} f(x) > -\infty.$

Assumption 14 (Local gradients are Lipschitz)

Every f_i has L_i-Lipschitz gradient, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

- If each f_i has L_i-Lipschitz gradient, then it is straightforward to check by Jensen's inequality that f is L-Lipschitz, with L satisfying the inequality
$$L \leq \frac{1}{n} \sum_i L_i.$$
- It will be also useful to define $\tilde{L} \stackrel{\text{def}}{=} (\frac{1}{n} \sum_{i=1}^n L_i^2)^{1/2}$. By the arithmetic-quadratic mean inequality, we have $\frac{1}{n} \sum_i L_i \leq \tilde{L}$.

Notation

Let

$$G^k \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \left\| g_i^k - \nabla f_i(x^k) \right\|^2 . \tag{359}$$

In EF21 we use a **contractive compressor $\mathcal{C} \in \mathbb{C}^d(\alpha)$,** where $0 < \alpha \leq 1$, and define

$$\theta \stackrel{\text{def}}{=} 1 - \sqrt{1-\alpha}, \qquad \beta \stackrel{\text{def}}{=} \frac{1-\alpha}{1-\sqrt{1-\alpha}}.$$

Theory in General Nonconvex Case

We now formulate our first complexity result.

Theorem 138

Let Assumptions 13 and 14 hold, and let the stepsize in Algorithm 26 (`EF21`*) be set as*

$$0 < \gamma \leq \left(L + \tilde{L}\sqrt{\frac{\beta}{\theta}} \right)^{-1}. \tag{360}$$

Fix $T \geq 1$ and let $\hat{x}^T$ be chosen from the iterates $x^0, x^1, \ldots, x^{T-1}$ uniformly at random. Then

$$\mathbf{E}\left[\left\| \nabla f(\hat{x}^T) \right\|^2 \right] \leq \frac{2\left(f(x^0) - f^{\inf}\right)}{\gamma T} + \frac{\mathbf{E}\left[G^0\right]}{\theta T}. \tag{361}$$

Remarks:

- Proof very similar to the proofs of `PAGE` and `MARINA`.
- Reuses two lemmas: Lemma 126 and Lemma 127.
- Read the proof in [64].

Theory in General Nonconvex Case: Commentary I

- Previous best rate for EF in the case of general smooth nonconvex functions was $\mathcal{O}(1/T^{2/3})$, which is worse than the $\mathcal{O}(1/T)$ rate of E21. Moreover, this worse rate was obtained using stronger assumptions (e.g., assuming that $\|\nabla f_i(x)\|$ is bounded for all nodes $i \in [n]$ and all $x \in \mathbb{R}^d$, which does not hold for quadratics).
- **No compression = Gradient Descent.** If we do not compress, i.e., if we use the identity compressor $\mathcal{C}(x) \equiv x$, then
  - EF21 reduces to GD,
  - $\alpha = 1$, and hence $\theta = 1$ and $\beta = 0$. Also, $G^0 = 0$. Choosing the maximal stepsize $\gamma = 1/L$, Theorem 138 reduces to

$$\mathbf{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \frac{2L\left(f(x^0) - f^{\inf}\right)}{T},$$

which is the rate of **gradient descent.**

Theory in General Nonconvex Case: Commentary II

- **Dependence on α.** Note that in general,

$$\alpha \mapsto \sqrt{\frac{\beta}{\theta}} = \frac{(1+\sqrt{1-\alpha})}{\alpha} - 1 \in \left[\frac{1}{\alpha} - 1, \frac{2}{\alpha} - 1\right)$$

  is decreasing in α. This makes sense since larger α means less dramatic compression, which leads to smaller $\sqrt{\frac{\beta}{\theta}}$, and this through (360) allows for larger stepsize, and hence fewer communication rounds.
- **Bad dependence on n.** Unfortunately, and unlike in all cases when we considered unbiased compressors, **the theoretical rate of EF21 does not improve with the number of nodes n.**
  - However, this is not an issue with the analysis!
  - In fact, EF21 can behave this way, and this is due to the fact that **contractive compressors can be too greedy.**

Theory in General Nonconvex Case: Commentary III

- Indeed, consider the **homogeneous data regime** with $f_1 = \cdots = f_n$, and consider the Top-K compressor. **Note that EF21 produces the same sequence of iterates independently of n! So, one can't prove speedup in n.** This does not happen for methods such as DCGD, DIANA, MARINA with unbiased and independent compressors, and not even for MARINA with the unbiased (and correlated) Perm-K compressor system. We will talk about MARINA later.
- For a comparison of the communication complexities of EF21 and MARINA (which shows that EF21 is worse), see the work of Szlendak, Tyurin and R. [80].

Extensions

- Linear rate under PL inequality (see [64]).
- A more aggressive variant called `EF21+` that works better in practice, but also suffers from larger oscillations. Has exactly the same theory as `EF21` (see [64]).
- Multiple extensions of vanilla `EF21` were proposed by Fatkhullin et al in [15]:
  1. stochastic approximation (i.e., replacing gradients by stochastic gradients),
  2. variance reduction (for stochastic approximation),
  3. partial participation (i.e., subsampling of nodes/clients),
  4. compression on the server (also via the EF21 trick),
  5. Polyak momentum,
  6. regularization.

Exercises

Exercises I

Exercise 95

Let $\mathcal{C} \in \mathbb{B}^d(\omega)$.

(i) *Find the set of all scalars* $t \geq 0$ *such that*

$$t\mathcal{C} \in \mathbb{C}^d(\alpha)$$

for some $0 \leq \alpha < 1$.

(ii) *Find the scalar* $t \geq 0$ *which maximizes* α.

Exercise 96

Prove Lemma 135.

Exercises II

Exercise 97

Consider a contractive compressor $\mathcal{C} \in \mathbb{C}^d(\alpha)$, and let $\{x^k\}_{k\geq 0}$ be a sequence of vectors in $\mathbb{R}^d$ converging to some $x^\star \in \mathbb{R}^d$. Define

$$y^k = x^\star + \mathcal{C}(x^k - x^\star)$$

for $k \geq 0$. We think of y^k as an estimator of x^k. Prove that this estimator is gradually improving in the following sense:

$$\mathbf{E}\left[\left\|y^k - x^k\right\|^2\right] \to 0.$$

Exercise 98

Establish a convergence theorem for `EF21` *under the PL inequality.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 22: MARINA: Distributed Compressed Gradient Descent with a Biased Estimator

Based on:

[18] Eduard Gorbunov, Konstantin Burlachenko, Zhize Li and P.R., **MARINA: Faster non-convex distributed learning with compression**, ICML 2021

[80] Rafał Szlendak, Alexander Tyurin and P.R., **Permutation compressors for provably faster distributed nonconvex optimization**, ICLR 2022

Outline of the Lecture

We now turn attention to **distributed nonconvex optimization:**

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

Our goal is to develop an efficient method relying **communication compression.**

The plan:

- Motivated by PAGE [88], we will describe structurally a very similar method, called MARINA [18]. main difference: MARINA uses communication compression instead of subsampling.
- As PAGE, MARINA uses a **biased** gradient estimator. This estimator is formed using **unbiased** compressors.
- We will not use the analysis provided in [18]. Instead, we will use the **refined analysis** due to Szlendak, Tyurin and R [80]. Improvements:
  - Analysis for **potentially correlated** compressors.
  - Analysis relies on the novel notion of **Hessian variance.**
  - Improved complexity is obtained for specially designed correlated compressors called **input variance compressors** in the regime when Hessian variance is small. This happens, for example, when the data is nearly homogeneous.

Content of the Lecture

We will not use slides. Instead, we will go through selected parts of [80].

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 23: Stochastic Newton Method

Based on:

[42] Dmitry Kovalev, Konstantin Mishchenko and P.R., **Stochastic Newton and Cubic Newton Methods with Simple Local Linear-Quadratic Rates**, NeurIPS Beyond First Order Methods Workshop, 2019.

Introduction

We consider to the problem

$$\min_{x \in \mathbb{R}^d} \left[f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right]$$

in the **big data scenario:**

$$n \gg d.$$

We considered this scenario in many of the SGD lectures.

Novelty: **In contrast with all previous lectures, we will now consider the use of second-order information.**

First-order Methods I

As we know very well by now, the **prototypical first-order method is gradient decent (GD).**

- GD is based on **local linear approximation** of f around the current iterate:

$$\begin{aligned} f(x^k + h) \approx G^k(h) &\stackrel{\text{def}}{=} f(x^k) + \langle \nabla f(x^k), h \rangle + \frac{L^k}{2} \|h\|^2 \\ &= f(x^k) + \langle \nabla f(x^k), h \rangle + \frac{1}{2} \langle L^k \boldsymbol{I}_d h, h \rangle . \end{aligned}$$

- GD is then constructed by setting x^{k+1} to be $x^k + h^k$, where h^k is the **minimizer of the model $G^k(\cdot)$:**

$$h^k \stackrel{\text{def}}{=} \arg\min_{h \in \mathbb{R}^d} G^k(h) \qquad \Rightarrow \qquad \boxed{x^{k+1} = x^k - \frac{1}{L^k} \nabla f(x^k)}$$

First-order Methods II

- L^k can be chosen so that

$$f(x^k + h) \leq G^k(h), \qquad \forall h \in \mathbb{R}^d,$$

which is a crucial property for establishing **global convergence (i.e., convergence from any starting point $x^0 \in \mathbb{R}^d$).** For instance, these choices of L^k will do the job:

  - $L^k \geq \sup_{x \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 f(x))$ if f is convex and twice differentiable, or
  - $L^k \geq L$, where L is the Lipschitz constant of the gradient of f.

Second-order Methods

The **prototypical second-order method is Newton's method (`Newton`).**

- `Newton` is based on **local quadratic approximation** of f around the current iterate:

$$f(x^k + h) \approx N^k(h) \quad \stackrel{\text{def}}{=} \quad f(x^k) + \langle \nabla f(x^k), h \rangle + \frac{1}{2} \langle \nabla^2 f(x^k) h, h \rangle .$$

- `Newton` is then constructed by setting x^{k+1} to be $x^k + h^k$, where h^k is the **minimizer of the model $N^k(\cdot)$:**

$$h^k \stackrel{\text{def}}{=} \arg\min_{h \in \mathbb{R}^d} N^k(h) \qquad \Rightarrow \qquad \boxed{x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k)}$$

- `Newton` **does not in general converge globally.** This is because **it is not in general true that**

$$f(x^k + h) \leq N^k(h), \qquad \forall h \in \mathbb{R}^d.$$

What to do with Non-Convergence of `Newton` I

There are two types of solutions:

1. **Keep `Newton` unchanged,** and instead study **local convergence (i.e., convergence for x^0 which is close enough to the solution)**
   - This is the classical approach. We adopt it in this lecture.
   - **Newton method converges extremely fast locally** [51]. It enjoys **local quadratic convergence.** Under appropriate assumptions, one can show that there exists a constant $c > 0$ (typically much larger than 1) such that

     $$\left\|x^{k+1} - x^\star\right\| \leq c \left\|x^k - x^\star\right\|^2.$$

     As we shall see (see Exercise 99), if $\left\|x^0 - x^\star\right\| < \frac{1}{2c}$, then for all $k \geq 0$ we have

     $$\left\|x^k - x^\star\right\| \leq (1/2)^{2^k} / c.$$

2. **Modify `Newton` to make it globally convergent.**
   - Use any globally convergent method (e.g., a first order method) to get from x^0 to a neighbourhood of the solution, and then switch to `Newton`.

What to do with Non-Convergence of Newton II

- Use **Damped Newton method:**

$$x^{k+1} = x^k - \gamma^k \nabla^2 f(x^k)^{-1} \nabla f(x^k),$$

  with damping (stepsize) γ^k.
- Use **cubically regularized Newton method.** This was developed in 1981 by Griewank [26] and in 2006 rediscovered and analyzed by Polyak & Nesterov [53].

Issues with Newton - I

1. **May not converge globally.**
2. **Not all functions** of interest in practice **are twice differentiable.** Examples:
   - $f(x) = \int_0^x \sum_{i=1}^n |t - a_i| dt$ for positive scalars $a_1, \ldots, a_n \in \mathbb{R}$.
   - LASSO: $f(x) = \frac{1}{2} \|\boldsymbol{A}x - b\|^2 + \lambda \|x\|_1$. Not differentiable at 0 and hence not twice differentiable.
3. The Hessian $\nabla^2 f(x^k)$ is **expensive to compute and/or invert if d is large.** For a general non-structured $d \times d$ matrix, computation of the inverse costs $\mathcal{O}(d^3)$ arithmetic operations.

Possible remedies:

(a) Do not invert. Instead, (approximately) **solve the Newton linear system**

$$\nabla^2 f(x^k) h = \nabla f(x^k)$$

to obtain h^k, and perform the step $x^{k+1} = x^k - h^k$.

Issues with Newton - II

(b) **Quasi-Newton (QN) methods** try to approximate Hessian from zero-order and first-order information, using the approximate identity (a.k.a. **secant equation**)

$$\nabla^2 f(x^{k+1})\delta_k \approx \gamma_k, \tag{362}$$

where

$$\gamma_k \stackrel{\text{def}}{=} \nabla f(x^{k+1}) - \nabla f(x^k), \qquad \delta_k \stackrel{\text{def}}{=} x^{k+1} - x^k.$$

Equation (362) is satisfied as an identity for convex quadratics (see Exercise 100). In particular, QN methods form a sequence of Hessian approximations $\{\boldsymbol{H}_k\}$ with the property

$$\boldsymbol{H}_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k) \tag{363}$$

and perform iterations of the form

$$x^{k+1} = x^k - \boldsymbol{H}_k^{-1}\nabla f(x^k).$$

There are many QN variants, differing in which solution to (363) they choose. Some notable examples:

Issues with Newton - III

(i) SR1 (Symmetric Rank 1):

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{(\delta_k - \boldsymbol{H}_k\gamma_k)(\delta_k - \boldsymbol{H}_k\gamma_k)^\top}{\langle \delta_k - \boldsymbol{H}_k\gamma_k, \gamma_k\rangle}$$

(ii) DFP (Davidon-Fletcher-Powell):

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\delta_k\delta_k^\top}{\langle \gamma_k, \delta_k\rangle} - \frac{\boldsymbol{H}_k\gamma_k\gamma_k^\top\boldsymbol{H}_k}{\langle \boldsymbol{H}_k\gamma_k, \gamma_k\rangle}$$

(iii) BFGS (Broyden-Fletcher-Goldfarb-Shanno):

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{H}_k\gamma_k\delta_k^\top + \delta_k\gamma_k^\top\boldsymbol{H}_k}{\langle \boldsymbol{H}_k\gamma_k, \gamma_k\rangle} - \beta_k\frac{\boldsymbol{H}_k\gamma_k\gamma_k^\top\boldsymbol{H}_k}{\langle \boldsymbol{H}_k\gamma_k, \gamma_k\rangle}$$

The above QN methods (and some others) can be written in the **least-change** form

$$\begin{aligned} \boldsymbol{H}_{k+1} &= \arg\min_{\boldsymbol{H}\in\mathbb{R}^{d\times d}} \|\boldsymbol{H} - \boldsymbol{H}_k\|_{\boldsymbol{B}} \\ \text{subject to} \quad & \boldsymbol{H}\delta_k = \gamma_k \\ & \boldsymbol{H} \in \mathcal{L}_k, \end{aligned}$$

Issues with Newton - IV

where

$$\|\boldsymbol{X}\|_{\boldsymbol{B}} \stackrel{\text{def}}{=} \left\|\boldsymbol{B}^{-1/2}\boldsymbol{X}\boldsymbol{B}^{-1/2}\right\|_F$$

is a weighted Frobenius norm with weight matrix $\boldsymbol{B} \succ 0$ (e.g., $\boldsymbol{B} = \boldsymbol{I}_d$ or $\boldsymbol{B} = \boldsymbol{H}_k^{-1}$), and $\mathcal{L}_k = \mathbb{R}^d$ or $\mathcal{L}_k = \{\boldsymbol{H} \ : \ \boldsymbol{H} = \boldsymbol{H}^\top\}$.

For a connection between QN methods and SGD, refer to [24].

(c) RSN: **Randomized Subspace Newton** [22] - in each iteration take a Newton step in a random subspace passing through the current point x^k.

- Global linear convergence rate under suitable assumptions.
- Scales to large d.
- Good performance on generalized linear models.

(d) RBCN: **Randomized Block Cubic Newton** [12]

Issues with Newton - V

4. The Hessian $\nabla^2 f(x^k)$ is **expensive to compute if n is large:**

$$\nabla^2 f(x^k) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f_i(x^k).$$

Indeed, even if $\nabla^2 f_i(x^k)$ is easy to compute and invert for any single i, $\nabla^2 f(x^k)$ is the average of a very large number of such matrices.

Possible remedies:

(i) **Approximate Hessian via subsampling** [74, 6, 8]:

$$\nabla^2 f(x^k) \approx \frac{1}{|S_k|} \sum_{i \in S_k} \nabla^2 f_i(x^k).$$

- Bad: Need large (or increasing to n) minibatch size $|S_k|$ to approximate the Hessian well.
- Bad: We also need to approximate the gradient by subsampling (if we do not, we will perform a full pass over data in each iteration).
- Bad: No theory or weak theory ("sensitivity analysis").
- Bad: Not faster than variance-reduced first-order methods in practice.

Issues with Newton - VI

(ii) **Take Newton-type steps in the dual: Stochastic Dual Newton Ascent (SDNA) [61]**.

- Good: Do not need twice differentiable f_i; it is enough for each f_i to be L_i -smooth.
- Bad: Need explicit L_2 regularization.
- Bad: One has to solve a $|S_k|$-dimensional **regularized** convex quadratic optimization sub-problem (in the dual) in each iteration. The regularizer is separable and strongly convex (will be nonsmooth if primal loss functions are not strongly convex).
- Good: There is no limitation on the minibatch size $|S_k|$; even $|S_k| = 1$ works (in which case the method reduces to SDCA)
- Good: Global linear convergence guarantee. Superlinear speedup in the minibatch size. Better than SDCA in runtime for an appropriate choice of the minibatch size.

(iii) **Stochastic Newton (SN) [42]**: this lecture.

- Bad: local convergence analysis
- Good: **works with any minibatch size $\tau = |S_k|$, even $\tau = 1$.**
- Good: **Rate independent of condition number; depends on τ and n only.**
- Good: **recovers Newton for $|S_k| = n$, and its local quadratic rate**. **Local linear rate** for minibatch sizes $|S_k| < n$, better for larger minibatches.

Issues with Newton - VII

- Good: **First stochastic variant of Newton for solving $\min_x \frac{1}{n} \sum_i f_i(x)$ which works even with small minibatch sizes.**
- Bad: A very new method; its practical benefits are not fully understood yet. Need to develop practical variants needed.

Stochastic Newton Method [42]

Stochastic Newton Method

Algorithm 27 Stochastic Newton (SN)

1: **Initialize:** Choose starting iterates $w_1^0, w_2^0, \ldots, w_n^0 \in \mathbb{R}^d$ and minibatch size $\tau \in \{1, 2, \ldots, n\}$

2: **for** $k = 0, 1, 2, \ldots$ **do**

3:
$$x^k = \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k)\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right]$$

4: Choose a subset $S^k \subseteq \{1, \ldots, n\}$ of size τ uniformly at random

5: $$w_i^{k+1} = \begin{cases} x^k, & i \in S^k \\ w_i^k, & i \notin S^k \end{cases}$$

Initialization: We may decide to set $w_i^0 = w_j^0$ for all i, j.

Generic Efficient Implementation

Unless the functions $f_1, \ldots, f_n$ have a special structure that can be exploited, the following generic efficient implementation can be used:

- Form the matrix $\boldsymbol{H}^0 = \frac{1}{n}\sum_i \nabla^2 f_i(w_i^0)$ (cost: $\mathcal{O}(nd^2)$)
- Form the vector $g^0 = \frac{1}{n}\sum_i \nabla^2 f_i(w_i^0) w_i^0 - \nabla f_i(w_i^0)$ (cost: $\mathcal{O}(nd^2)$)
- Given $\boldsymbol{H}^{k-1}$ and g^{k-1}, compute $\boldsymbol{H}^k$ and g^k as follows:
  - Compute
    - 2τ Hessians $\nabla^2 f_i(w_i^k)$ and $\nabla^2 f_i(w_i^{k+1})$ for $i \in S^k$ (cost: $\mathcal{O}(\tau d^2)$)
    - 2τ gradients $\nabla f_i(w_i^k)$ and $\nabla f_i(w_i^{k+1})$ for $i \in S^k$ (cost: $\mathcal{O}(\tau d)$)
  - Let

$$\boldsymbol{H}^k = \boldsymbol{H}^{k-1} + \frac{1}{n}\sum_{i \in S^k}\left(\nabla^2 f_i(w_i^{k+1}) - \nabla^2 f_i(w_i^k)\right) \qquad \text{(cost: } \mathcal{O}(\tau d^2)\text{)}$$

$$g^k = g^{k-1} + \frac{1}{n}\sum_{i \in S^k}\left(\nabla^2 f_i(w_i^{k+1}) w_i^{k+1} - \nabla f_i(w_i^{k+1}) - \left(\nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right)\right)$$

(cost: $\mathcal{O}(\tau d^2)$)

- Invert $\boldsymbol{H}^k$ (cost: $\mathcal{O}(d^3)$)
- Compute $x^k = \left(\boldsymbol{H}^k\right)^{-1} g^k$ (cost: $\mathcal{O}(d^2)$)
- Update w_i^{k+1} for $i \in S^k$ (cost: $\mathcal{O}(\tau d)$)

Total Cost

By adding the costs above, we get the following formula for the **cost of initialization and subsequent iterations:**

- Initialization: $\mathcal{O}(nd^2)$
- Iteration: $\mathcal{O}(\tau d^2 + d^3)$

Total cost after k iterations:

$$\mathcal{O}(nd^2 + k(\tau + d)d^2)$$

Local Convergence Analysis

Lipschitz Hessian

Definition 139 (Lipschitz Hessian)

Function $f : \mathbb{R}^d \to \mathbb{R}$ has H-Lipschitz Hessian if

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq H \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{364}$$

Lemma 140

If $f : \mathbb{R}^d \to \mathbb{R}$ has H-Lipschitz Hessian, then

$$\left\|\nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x - y)\right\| \leq \frac{H}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{365}$$

Proof of Lemma 140 I

STEP 1: Fundamental theorem of calculus.

Recall that by Corollary 18 (fundamental theorem of calculus),

$$\langle \nabla f(x) - \nabla f(y), s\rangle = \int_0^1 \langle \nabla^2 f(y + t(x-y))(x-y), s\rangle \, dt. \tag{366}$$

for any $x, y, s \in \mathbb{R}^d$. Applying this identity for the standard unit basis vectors $s = e_1, \ldots, e_d$ in $\mathbb{R}^d$, and writing the d resulting identities compactly in a vector form, we get the identity

$$\nabla f(x) = \nabla f(y) + \int_0^1 \nabla^2 f(y + t(x-y))(x-y) \, dt. \tag{367}$$

STEP 2: Getting the bound.

Fix any $x, y \in \mathbb{R}^d$. Using the above identity, we get

$$\begin{aligned} \nabla f(x) &\overset{(367)}{=} \nabla f(y) + \int_0^1 \nabla^2 f(y + t(x-y))(x-y) \, dt \\ &= \nabla f(y) + \nabla^2 f(y)(x-y) \\ &\qquad + \int_0^1 \left[\nabla^2 f(y + t(x-y)) - \nabla^2 f(y)\right](x-y) \, dt. \end{aligned} \tag{368}$$

Proof of Lemma 140 II

Letting $A \stackrel{\text{def}}{=} \nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x-y)$, we can use this identity to produce the desired upper bound on $\|A\|$ as follows:

$$
\begin{aligned}
\|A\| &\stackrel{(368)}{=} \left\| \int_0^1 \left[\nabla^2 f(y + t(x-y)) - \nabla^2 f(y)\right](x-y)\, dt \right\| \\
&\leq \int_0^1 \left\| \left[\nabla^2 f(y + t(x-y)) - \nabla^2 f(y)\right](x-y) \right\| dt \\
&\leq \int_0^1 \left\| \nabla^2 f(y + t(x-y)) - \nabla^2 f(y) \right\| \|x-y\|\, dt \\
&\stackrel{(364)}{\leq} \int_0^1 H \|t(x-y)\| \|x-y\|\, dt \\
&= H \|x-y\|^2 \int_0^1 t\, dt \\
&= \frac{H}{2} \|x-y\|^2 .
\end{aligned}
$$

Main Result

Let $\boldsymbol{W}^k \overset{\text{def}}{=} [w_1^k, \ldots, w_n^k] \in \mathbb{R}^{d \times n}$.

Theorem 141

Assume that every f_i is μ-convex and has H-Lipschitz Hessian and consider the following Lyapunov function:

$$\mathcal{A}^k \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} \left\| w_i^k - x^\star \right\|^2. \tag{369}$$

Then for the random iterates of Algorithm 27 we have the recursion

$$\mathbf{E}\left[\mathcal{A}^{k+1} \mid \boldsymbol{W}^k\right] \leq \left(1 - \frac{\tau}{n} + \frac{\tau}{n}\left(\frac{H}{2\mu}\right)^2 \mathcal{A}^k\right) \mathcal{A}^k. \tag{370}$$

Furthermore, if $\|w_i^0 - x^\star\| \leq \frac{\mu}{H}$ for all i, then

$$\mathbf{E}\left[\mathcal{A}^{k+1} \mid \boldsymbol{W}^k\right] \leq \left(1 - \frac{3\tau}{4n}\right) \mathcal{A}^k. \tag{371}$$

Commentary

- When $\tau = n$, we obtain `Newton` as a special case[15]. Recursion (370) simplifies to

$$\mathcal{A}^{k+1} \leq \left(\frac{H}{2\mu}\right)^2 \left(\mathcal{A}^k\right)^2. \tag{372}$$

  In view of Exercise 99, we recover **local quadratic rate of `Newton`.**

- When $1 \leq \tau < n$, recursion (371) implies

$$\mathbf{E}\left[\mathcal{A}^k\right] \leq \left(1 - \frac{3\tau}{4n}\right)^k \mathcal{A}^0.$$

  We obtain a **local linear convergence rate that is independent of the condition number, and improves as τ grows**. Thus, the method provably adapts to the curvature, and enjoys linear speedup as the minibatch size τ grows.

[15] To fully recover `Newton`, we also need to choose $w_i^0 = w_j^0$ for all i, j. Otherwise we get a method whose first step is different, and all subsequent steps are Newton steps.

Two Lemmas

Lemma 142

Assume that each f_i is μ-convex and has H-Lipschitz Hessian. Then for all $k \geq 0$ we have

$$\|x^k - x^\star\| \leq \frac{H}{2\mu} \cdot \mathcal{A}^k. \tag{373}$$

Lemma 143 (Local boundedness)

Assume that each f_i is μ-convex and has H-Lipschitz Hessian. If

$$\|w_i^0 - x^\star\| \leq \frac{\mu}{H}$$

for all i, then for all $k \geq 0$ the Lyapunov function is bounded above as

$$\mathcal{A}^k \leq \frac{\mu^2}{H^2}. \tag{374}$$

Proof of Lemma 142 I

Letting $\boldsymbol{H}_k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(w_i^k)$, one step of the method can be written in the form

$$x^k = \boldsymbol{H}_k^{-1}\left[\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right].$$

Subtracting $x^\star = \boldsymbol{H}_k^{-1}\boldsymbol{H}_k x^\star$ from both sides, and using the fact that $\sum_{i=1}^n \nabla f_i(x^\star) = 0$, this further leads to

$$x^k - x^\star = \boldsymbol{H}_k^{-1}\frac{1}{n}\sum_{i=1}^n \left[\nabla^2 f_i(w_i^k)\left(w_i^k - x^\star\right) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^\star)\right)\right]. \tag{375}$$

Next, note that since f_i is μ-convex, we have $\nabla^2 f_i(w_i^k) \succeq \mu \boldsymbol{I}_d$ for all i. This implies that $\boldsymbol{H}_k \succeq \mu \boldsymbol{I}_d$, which in turns gives

$$\left\|\boldsymbol{H}_k^{-1}\right\| \leq \left\|\frac{1}{\mu}\boldsymbol{I}_d\right\| = \frac{1}{\mu}. \tag{376}$$

Proof of Lemma 142 II

Finally,

$$
\begin{aligned}
\left\|x^k - x^\star\right\| &\overset{(375)}{\leq} \left\|\boldsymbol{H}_k^{-1}\right\| \left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla^2 f_i(w_i^k)\left(w_i^k - x^\star\right) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^\star)\right)\right]\right\| \\
&\overset{(376)}{\leq} \frac{1}{\mu}\left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla f_i(x^\star) - \nabla f_i(w_i^k) - \nabla^2 f_i(w_i^k)\left(x^\star - w_i^k\right)\right]\right\| \\
&\leq \frac{1}{\mu n}\sum_{i=1}^n \left\|\nabla f_i(x^\star) - \nabla f_i(w_i^k) - \nabla^2 f_i(w_i^k)\left(x^\star - w_i^k\right)\right\| \\
&\overset{(365)}{\leq} \frac{1}{\mu n}\sum_{i=1}^n \frac{H}{2}\left\|w_i^k - x^\star\right\|^2 \\
&\overset{(369)}{=} \frac{H}{2\mu}\mathcal{A}^k.
\end{aligned}
$$

Proof of Lemma 143

STEP 1: CLAIM.

We claim that

$$\left\| w_i^k - x^\star \right\|^2 \leq \frac{\mu^2}{H^2}, \qquad \forall i \in \{1, 2, \ldots, n\}. \tag{377}$$

The upper bound on $\mathcal{A}^k$ follows as a consequence.

STEP 2: PROOF OF THE CLAIM.

We will now prove the claim by induction in k.

- The statement (377) holds for $k = 0$ by our assumption.
- Assume it holds for k and let us show that it holds for $k+1$. If $i \notin S^k$, w_i^k is not updated and inequality (377) holds for $w_i^{k+1} = w_i^k$ by induction assumption. On the other hand, for every $i \in S^k$, we have

$$\left\| w_i^{k+1} - x^\star \right\| = \left\| x^k - x^\star \right\| \overset{(373)}{\leq} \frac{H}{2\mu} \cdot \frac{1}{n} \sum_{j=1}^n \left\| w_j^k - x^\star \right\|^2 \leq \frac{H}{2\mu} \frac{1}{n} \sum_{j=1}^n \frac{\mu^2}{H^2} < \frac{\mu}{H}.$$

  Thus, again, (377) holds for w_i^{k+1}.

Third Lemma

Lemma 144

The random iterates of Algorithm 27 satisfy the identity

$$\mathbf{E}\left[\mathcal{A}^{k+1}\right] = \frac{\tau}{n}\mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] + \left(1 - \frac{\tau}{n}\right)\mathcal{A}^k. \tag{378}$$

Proof.

For each $i \in \{1, 2, \ldots, n\}$, w_i^{k+1} is equal to x^k with probability $\frac{\tau}{n}$ and to w_i^k with probability $1 - \frac{\tau}{n}$, hence the result. □

Proof of Theorem 141

Using Lemma 142 and Lemma 144, we obtain

$$
\begin{aligned}
\mathbf{E}\left[\mathcal{A}^{k+1}\right] &\overset{(378)}{=} \frac{\tau}{n}\left\|x^k - x^\star\right\|^2 + \left(1 - \frac{\tau}{n}\right)\mathcal{A}^k \\
&\overset{(373)}{\leq} \frac{\tau}{n}\left(\frac{H}{2\mu}\right)^2\left(\mathcal{A}^k\right)^2 + \left(1 - \frac{\tau}{n}\right)\mathcal{A}^k \\
&= \left(1 - \frac{\tau}{n} + \frac{\tau}{n}\left(\frac{H}{2\mu}\right)^2\mathcal{A}^k\right)\mathcal{A}^k \\
&\overset{(374)}{\leq} \left(1 - \frac{3\tau}{4n}\right)\mathcal{A}^k,
\end{aligned}
$$

where in the last step we also used the assumption $\|w_i^0 - x^\star\| \leq \frac{\mu}{H}$ for $i = 1, \ldots, n$.

Exercises - I

Exercise 99 (Quadratic convergence rate)

Let $\{d_k\}_{k\geq 0}$ *be a sequence of nonnegative real numbers, with* $d_0 > 0$. *Assume there exists constant* $c > 0$ *such that for all* k *the sequence satisfy the recursion*

$$d_{k+1} \leq c\,(d_k)^2\,.$$

Assume that $d_0 \leq \frac{1}{2c}$. *Show that then*

(i) $d_{k+1} \leq \frac{1}{2} d_k$ *for all* k *(and hence the sequence is decreasing)*

(ii) $d_k \leq c^{2^k-1} d_0^{2^k}$ *for all* k

(iii) $d_k \leq \frac{1}{c}\left(\frac{1}{2}\right)^{2^k}$ *for all* k

(iv) *Chose any* $0 < \varepsilon < \frac{1}{2c}$. *Then*

$$k \geq \log_2 \log_2 \frac{1}{c\varepsilon} \qquad \Rightarrow \qquad d_k \leq \varepsilon.$$

Exercises - II

Exercise 100 (Secant Equation and Convex Quadratics)

Let $f(x) = \frac{1}{2}x^\top \boldsymbol{A}x + b^\top x + c$, *where* $\boldsymbol{A} \succ 0$, $b \in \mathbb{R}^d$ *and* $c \in \mathbb{R}$. *Show that*

$$\boldsymbol{A}(x - y) = \nabla f(x) - \nabla f(y), \qquad \forall x, y \in \mathbb{R}^d,$$

which means that the secant equation (362) *holds as an identity.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 24: Randomized BFGS

Based on:

[24] Robert M. Gower and P.R., **Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms**, SIAM Journal on Matrix Analysis and Applications 38(4):1380-1409, 2017.

[11] Dmitry Kovalev, Robert M. Gower, P.R. and Alexander Rogozin, **Fast linear convergence of randomized BFGS**, arXiv:2002.11337, 2020.

Introduction

We consider solving the problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

with a **stochastic quasi-Newton method.** Notation:

- $x_\star = \arg\min f(x)$
- $\mathbb{S}^d$: space of $d \times d$ symmetric matrices
- Hessian matrix:

$$\boldsymbol{H}_x \stackrel{\text{def}}{=} \nabla^2 f(x), \qquad \boldsymbol{H}_k \stackrel{\text{def}}{=} \boldsymbol{H}_{x_k} = \nabla^2 f(x_k), \qquad \boldsymbol{H}_\star \stackrel{\text{def}}{=} \boldsymbol{H}_{x_\star} = \nabla^2 f(x_\star)$$

Weighted Frobenius Norm

Definition 145

For every $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and a positive definite "weight" matrix $\boldsymbol{H} \in \mathbb{S}^d$, we write

$$\|\boldsymbol{A}\|_{F(\boldsymbol{H})}^2 \stackrel{\text{def}}{=} \operatorname{Trace}\left(\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{A}^\top\right) = \left\|\boldsymbol{H}^{1/2}\boldsymbol{A}\boldsymbol{H}^{1/2}\right\|_F^2 \tag{379}$$

to denote a weighted Frobenius norm.

Remark: In this lecture, the weight matrix $\boldsymbol{H}$ will be chosen as the Hessian of f evaluated at some point, for example

- $\boldsymbol{H} = \boldsymbol{H}_k = \nabla^2 f(x_k)$,
- $\boldsymbol{H} = \boldsymbol{H}_\star = \nabla^2 f(x_\star)$.

RBFGS: Three Equivalent Formulations

BFGS Update: Introduction

We consider the **randomized BFGS update** first introduced by Gower and R in 2016 [24], and later in 2016 used in the context of machine learning by Gower, Goldfarb and R. [21].

Ingredients:

- $\boldsymbol{H} \in \mathbb{S}^d$ is a Hessian matrix we are trying to estimate/approximate,
- $\boldsymbol{B}_k \in \mathbb{S}^d$ is the current estimate of $\boldsymbol{H}^{-1}$,
- $\boldsymbol{S}_k \in \mathbb{R}^{d \times \tau}$ is a (typically thin) random matrix (drawn independently at step k) with $\tau \ll d$ columns and full column rank, and
- $\boldsymbol{B}_{k+1} \in \mathbb{S}^d$ is the new estimate of $\boldsymbol{H}^{-1}$.

We will give three equivalent ways of defining the RBFGS update, i.e., of defining $\boldsymbol{B}_{k+1}$ as a function of $\boldsymbol{H}$, $\boldsymbol{B}_k$, and $\boldsymbol{S}_k$:

- primal formulation,
- dual formulation,
- closed form expression.

BFGS Update: Primal and Dual Formulations

RBFGS: Primal formulation

$$
\begin{aligned}
\boldsymbol{B}_{k+1} &\overset{\text{def}}{=} \arg\min_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{B}_k\|^2_{F(\boldsymbol{H})} \\
\text{subject to} \quad & \boldsymbol{S}_k^\top \boldsymbol{H}\boldsymbol{X} = \boldsymbol{S}_k^\top \boldsymbol{I} \\
& \boldsymbol{X} = \boldsymbol{X}^\top
\end{aligned}
$$

RBFGS: Dual formulation

$$
\begin{aligned}
\boldsymbol{B}_{k+1} &\overset{\text{def}}{=} \arg\min_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{H}^{-1}\right\|^2_{F(\boldsymbol{H})} \\
\text{subject to} \quad & \boldsymbol{X} = \boldsymbol{B}_k + \boldsymbol{S}_k\boldsymbol{Y}^\top + \boldsymbol{Y}\boldsymbol{S}_k^\top \\
& \boldsymbol{Y} \in \mathbb{R}^{d\times\tau}
\end{aligned}
$$

Ingredients: $\boldsymbol{H} \in \mathbb{S}^d$ is a Hessian matrix we are trying to estimate/approximate, $\boldsymbol{B}_k \in \mathbb{S}^d$ is the current estimate of $\boldsymbol{H}^{-1}$, $\boldsymbol{S}_k \in \mathbb{R}^{d\times\tau}$ is a (typically thin) random matrix with $\tau \ll d$ columns and full column rank, and $\boldsymbol{B}_{k+1} \in \mathbb{S}^d$ is the new estimate of $\boldsymbol{H}^{-1}$.

BFGS Update: Closed Form

RBFGS: Closed form

$$\boldsymbol{B}_{k+1} = \mathrm{BFGS}(\boldsymbol{B}_k, \boldsymbol{H}, \boldsymbol{S}_k) \stackrel{\text{def}}{=} \boldsymbol{G}_k + (\boldsymbol{I} - \boldsymbol{G}_k \boldsymbol{H})\, \boldsymbol{B}_k\, (\boldsymbol{I} - \boldsymbol{H}\boldsymbol{G}_k)\,, \qquad (380)$$

where

$$\boldsymbol{G}_k \stackrel{\text{def}}{=} \boldsymbol{S}_k (\boldsymbol{S}_k^\top \boldsymbol{H} \boldsymbol{S}_k)^{-1} \boldsymbol{S}_k^\top .$$

Remarks:

- Note that $\boldsymbol{S}_k^\top \boldsymbol{H} \boldsymbol{S}_k$ is invertible since we assume that $\boldsymbol{S}_k$ has full column rank and $\boldsymbol{H}$ is invertible.
- We use this BFGS update in the randomized BFGS method stated as Algorithm 28.

Theorem 146 (Equivalence)

The primal, dual and closed form expression formulations of RBFGS *are equivalent.*

RBFGS: The Algorithm

RBFGS: The Algorithm

Algorithm 28 Randomized BFGS (RBFGS)

1: **Initialize:** $x_0 \in \mathbb{R}^d$, $\boldsymbol{B}_0 \in \mathbb{S}^d$, distribution $\mathcal{D}$ over matrices from $\mathbb{R}^{d\times\tau}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Take a step:

$$x_{k+1} = x_k - \boldsymbol{B}_k \nabla f(x_k)$$

4: **Monotonic option**:

$$x_{k+1} = \arg\min \{f(x_{k+1}), f(x_k)\}$$

5: Sample random $\boldsymbol{S}_k \sim \mathcal{D}$ and compute

$$\boldsymbol{B}_{k+1} = \mathrm{BFGS}(\boldsymbol{B}_k, \boldsymbol{H}_k, \boldsymbol{S}_k)$$

Assumptions

Assumptions I

We consider the Monotonic option on line 8 of Algorithm 28. This enforces the function values $f(x_k)$ to be monotonically decreasing. That is, if we let

$$\mathcal{Q} \stackrel{\text{def}}{=} \{x \ : \ f(x) \leq f(x_0)\},$$

then by (4) we have that $x_k \in \mathcal{Q}$ for all k.

We now rely on strong convexity, and Lipschitz gradients and Hessians:

Assumption 15

Function f is μ-convex. That is,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{381}$$

Remark: We specifically write $\|\cdot\|_2$ for the standard Euclidean norm in this lecture since we will work with several other norms as well. Since we assume strong convexity, $x_\star = \arg\min f(x)$ is well defined and unique.

Assumptions II

Assumption 16

Function f has L_1-Lipschitz gradient. That is,

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L_1 \|y - x\|_2 , \quad \forall x, y \in \mathbb{R}^d. \tag{382}$$

Assumption 17

Function f has L_2-Lipschitz Hessian. That is,

$$\|\boldsymbol{H}_y - \boldsymbol{H}_x\|_2 \leq L_2 \|y - x\|_2 , \quad \forall x, y \in \mathbb{R}^d. \tag{383}$$

Main Theorem

Main Theorem

Theorem 147

Assume the monotonic option in Algorithm 28 is used. Let Assumptions 15, 16 and 17 hold and let

$$\Psi_k \stackrel{def}{=} \sqrt{f(x_k) - f(x_\star)} + \beta \left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_\star)}^2, \tag{384}$$

where $\beta = \frac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho}$ *and where* ρ *is defined as follows:*

$$\rho \stackrel{def}{=} \inf_{x \in \mathcal{Q}} \lambda_{\min} \left(\mathbf{E} \left[\boldsymbol{H}_x^{1/2} \boldsymbol{S} (\boldsymbol{S}^\top \boldsymbol{H}_x \boldsymbol{S})^{-1} \boldsymbol{S}^\top \boldsymbol{H}_x^{1/2} \right] \right). \tag{385}$$

If

$$f(x_0) - f(x_\star) \leq \mathcal{F} \stackrel{def}{=} \frac{1}{4} \left[\frac{\sqrt{2L_1} L_2}{\mu^2} + \frac{32\sqrt{2} d L_1^{5/2} L_2}{\rho \mu^4} \right]^{-2}, \tag{386}$$

then

$$\mathbf{E}\left[\Psi_{k+1}\right] \leq \left(1 - \frac{\rho}{2}\right) \mathbf{E}\left[\Psi_k\right]. \tag{387}$$

Five Lemmas

First Lemma: Approximation of One Weighted Frobenius Norm by Another

Lemma 148

Let $\boldsymbol{W} \in \mathbb{S}^d$ and let Assumption 17 hold. Then for all $x, y \in \mathbb{R}^d$ the following inequality holds:

$$\|\boldsymbol{W}\|_{F(\boldsymbol{H}_x)} \leq \left(1 + \frac{L_2}{\mu} \|x - y\|_2\right) \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} . \tag{388}$$

Comments:

- If $x \approx y$, then $\|\boldsymbol{W}\|_{F(\boldsymbol{H}_x)} \approx \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)}$

Basic Facts about Matrix Norms - I

Any matrix norm. All matrix norms need to satisfy:

- absolute homogeneity: $\|t\boldsymbol{A}\| = |t|\,\|\boldsymbol{A}\|$
- subadditivity: $\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$
- nonnegativity: $\|\boldsymbol{A}\| \geq 0$
- definiteness: $\|\boldsymbol{A}\| = 0 \Leftrightarrow \boldsymbol{A} = 0$

Spectral norm: matrix norm induced by the vector L_2-norm.

- definition:

$$\|\boldsymbol{A}\|_2 \stackrel{\text{def}}{=} \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|_2}{\|x\|_2} \tag{389}$$

- submultiplicativity:

$$\|\boldsymbol{A}\boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\|_2\,\|\boldsymbol{B}\|_2 \tag{390}$$

- $$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_2 = \left\|\boldsymbol{A}^\top\boldsymbol{A}\right\|_2 = \|\boldsymbol{A}\|_2^2 = \left\|\boldsymbol{A}^\top\right\|_2^2 \tag{391}$$

Basic Facts about Matrix Norms - II

- Using subadditivity of the spectral norm with $\boldsymbol{A} = \boldsymbol{C} - \boldsymbol{I}$ and $\boldsymbol{B} = \boldsymbol{I}$, we get

$$\|\boldsymbol{C}\|_2 \leq \|\boldsymbol{C} - \boldsymbol{I}\|_2 + \|\boldsymbol{I}\|_2 = \|\boldsymbol{C} - \boldsymbol{I}\|_2 + 1. \tag{392}$$

Connection between Frobenius and spectral norms.

- Mixed submultiplicativity:

$$\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_2 \tag{393}$$

$$\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F \tag{394}$$

- From these, we get the inequality

$$\begin{aligned} \left\|\boldsymbol{ABA}^\top\right\|_F &\overset{(393)}{\leq} \|\boldsymbol{AB}\|_F \left\|\boldsymbol{A}^\top\right\|_2 \\ &\overset{(394)}{\leq} \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F \left\|\boldsymbol{A}^\top\right\|_2 \\ &\overset{(391)}{=} \|\boldsymbol{B}\|_F \left\|\boldsymbol{A}^\top\boldsymbol{A}\right\|_2 . \end{aligned} \tag{395}$$

Proof of Lemma 148

$$
\begin{aligned}
\|\boldsymbol{W}\|_{F(\boldsymbol{H}_x)} &\overset{(379)}{=} \left\|\boldsymbol{H}_x^{1/2}\boldsymbol{W}\boldsymbol{H}_x^{1/2}\right\|_F \\
&= \left\|\underbrace{\boldsymbol{H}_x^{1/2}\boldsymbol{H}_y^{-1/2}}_{A}\underbrace{\boldsymbol{H}_y^{1/2}\boldsymbol{W}\boldsymbol{H}_y^{1/2}}_{B}\underbrace{\boldsymbol{H}_y^{-1/2}\boldsymbol{H}_x^{1/2}}_{A^\top}\right\|_F \\
&\overset{(395)}{\leq} \left\|\underbrace{\boldsymbol{H}_y^{1/2}\boldsymbol{W}\boldsymbol{H}_y^{1/2}}_{B}\right\|_F \left\|\underbrace{\boldsymbol{H}_y^{-1/2}\boldsymbol{H}_x^{1/2}}_{A^\top}\underbrace{\boldsymbol{H}_x^{1/2}\boldsymbol{H}_y^{-1/2}}_{A}\right\|_2 \\
&\overset{(379)}{=} \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} \left\|\boldsymbol{H}_y^{-1/2}\boldsymbol{H}_x\boldsymbol{H}_y^{-1/2}\right\|_2 \\
&\overset{(392)}{\leq} \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} \left(1 + \left\|\boldsymbol{H}_y^{-1/2}[\boldsymbol{H}_x - \boldsymbol{H}_y]\boldsymbol{H}_y^{-1/2}\right\|_2\right) \\
&\overset{(390)}{\leq} \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} \left(1 + \left\|\boldsymbol{H}_y^{-1/2}\right\|_2 \|\boldsymbol{H}_x - \boldsymbol{H}_y\|_2 \left\|\boldsymbol{H}_y^{-1/2}\right\|_2\right) \\
&\overset{(391)}{=} \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} \left(1 + \left\|\boldsymbol{H}_y^{-1}\right\|_2 \|\boldsymbol{H}_x - \boldsymbol{H}_y\|_2\right) \\
&\overset{(383)}{\leq} \|\boldsymbol{W}\|_{F(\boldsymbol{H}_y)} \left(1 + \frac{L_2}{\mu}\|x - y\|_2\right).
\end{aligned}
$$

Second Lemma: Improvement of the Iterate x_k After One Step of the Newton-Star Method

Lemma 149

Let $\boldsymbol{H}_\star \stackrel{def}{=} \nabla^2 f(x_\star)$. *It follows that*

$$\left\| \left(x_k - \boldsymbol{H}_\star^{-1} \nabla f(x_k)\right) - x_\star \right\|_2 \leq \frac{L_2}{2\mu} \left\| x_k - x_\star \right\|_2^2 . \tag{396}$$

Comments:

- One can think of
$$\hat{x}_{k+1} \stackrel{\text{def}}{=} x_k - \boldsymbol{H}_\star^{-1} \nabla f(x_k)$$
as one step of an **approximate Newton** method where we use $\boldsymbol{H}_\star$ instead of the true Hessian $\boldsymbol{H}_{x_k}$.
- It turns out that like Newton, Newton-Star also enjoys a **local quadratic convergence rate** [71].

Proof of Lemma 149

By the fundamental theorem of calculus,

$$\nabla f(x_k) = \nabla f(x_k) - \nabla f(x_\star) = \int_0^1 \nabla^2 f(x_\star + t(x_k - x_\star))(x_k - x_\star)dt. \tag{397}$$

The result now follows from

$$\begin{aligned}
\left\| x_k - x_\star - \boldsymbol{H}_\star^{-1} \nabla f(x_k) \right\|_2 &\overset{(397)}{=} \left\| \boldsymbol{H}_\star^{-1} \left(\boldsymbol{H}_\star - \int_0^1 \nabla^2 f(x_\star + t(x_k - x_\star))dt \right) (x_k - x_\star) \right\|_2 \\
&\overset{(389)+(390)}{\leq} \left\| \boldsymbol{H}_\star^{-1} \right\|_2 \left\| \boldsymbol{H}_\star - \int_0^1 \nabla^2 f(x_\star + t(x_k - x_\star))dt \right\|_2 \left\| x_k - x_\star \right\|_2 \\
&= \left\| \boldsymbol{H}_\star^{-1} \right\|_2 \left\| \int_0^1 \left[\boldsymbol{H}_\star - \nabla^2 f(x_\star + t(x_k - x_\star)) \right] dt \right\|_2 \left\| x_k - x_\star \right\|_2 \\
&\leq \left\| \boldsymbol{H}_\star^{-1} \right\|_2 \left[\int_0^1 \left\| \boldsymbol{H}_\star - \nabla^2 f(x_\star + t(x_k - x_\star)) \right\|_2 dt \right] \left\| x_k - x_\star \right\|_2 \\
&\overset{(383)+(381)}{\leq} \frac{L_2}{\mu} \left(\int_0^1 \left\| t(x_k - x_\star) \right\|_2 dt \right) \left\| x_k - x_\star \right\|_2 \\
&= \frac{L_2}{2\mu} \left\| x_k - x_\star \right\|_2^2 .
\end{aligned}$$

Third Lemma: Function Value Decrease After One Step of RBFGS (Form 1)

Lemma 150

After one step of RBFGS*, i.e., for the point*

$$x_+ \stackrel{def}{=} x_k - \boldsymbol{B}_k \nabla f(x_k),$$

we have the inequality

$$\sqrt{f(x_+) - f(x_\star)} \leq \frac{\sqrt{2L_1}L_2}{\mu^2}\left(f(x_k) - f(x_\star)\right) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2}\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2. \tag{398}$$

Remark: If $\boldsymbol{B}_k = \boldsymbol{H}_\star^{-1}$, i.e., if our inverse Hessian estimate was "perfect" (from the viewpoint of the Newton-Star method), the above inequality would say that

$$f(x_+) - f(x_\star) \leq \left(\frac{\sqrt{2L_1}L_2}{\mu^2}\right)^2 \left(f(x_k) - f(x_\star)\right)^2,$$

which is a recursion similar to the one that holds for Newton's method, which leads to its local quadratic convergence rate. However, here we work with function suboptimality rather than distance to the solution, and we pay a price for not knowing $\boldsymbol{H}_\star^{-1}$ precisely.

Proof of Lemma 150 - I

We start with an upper-bound for $\|x_+ - x_\star\|_2$:

$$\begin{aligned}
\|x_+ - x_\star\|_2 &= \|x_k - x_\star - \boldsymbol{B}_k \nabla f(x_k)\|_2 \\
&= \left\|x_k - x_\star - \boldsymbol{H}_\star^{-1}\nabla f(x_k) + (\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k)\nabla f(x_k)\right\|_2 \\
&\leq \left\|x_k - x_\star - \boldsymbol{H}_\star^{-1}\nabla f(x_k)\right\|_2 + \left\|(\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k)\nabla f(x_k)\right\|_2 \\
&\overset{(396)+(382)}{\leq} \frac{L_2}{2\mu}\|x_k - x_\star\|_2^2 + L_1 \left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_2 \|x_k - x_\star\|_2 .
\end{aligned}$$

Now using Young's inequality $ab \leq \frac{c}{2}a^2 + \frac{1}{2c}b^2$ for every $a, b, c > 0$ with $a = \|x_k - x_\star\|_2$, $b = \left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_2$ and $c = \frac{L_2}{\mu L_1}$ in the above gives

$$\begin{aligned}
\|x_+ - x_\star\|_2 &\leq \frac{L_2}{\mu}\|x_k - x_\star\|_2^2 + \frac{\mu L_1^2}{2L_2}\left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_2^2 \\
&\leq \frac{L_2}{\mu}\|x_k - x_\star\|_2^2 + \frac{\mu L_1^2}{2L_2}\left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_{F(\boldsymbol{H}_k)}^2 \left\|\boldsymbol{H}_k^{-1}\right\|_2^2 \\
&\leq \frac{L_2}{\mu}\|x_k - x_\star\|_2^2 + \frac{L_1^2}{2\mu L_2}\left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_{F(\boldsymbol{H}_k)}^2 .
\end{aligned}$$

Proof of Lemma 150 - II

Using smoothness (382), the above inequality and then strong convexity (381), in this order, we get

$$\begin{aligned}
\sqrt{f(x_+) - f(x_\star)} &\overset{(382)}{\leq} \sqrt{\frac{L_1}{2}} \|x_+ - x_\star\|_2 \\
&\leq \sqrt{\frac{L_1}{2}} \frac{L_2}{\mu} \|x_k - x_\star\|_2^2 + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_{F(\boldsymbol{H}_k)}^2 \\
&\overset{(381)}{\leq} \frac{\sqrt{2L_1}L_2}{\mu^2} (f(x_k) - f(x_\star)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \left\|\boldsymbol{H}_\star^{-1} - \boldsymbol{B}_k\right\|_{F(\boldsymbol{H}_k)}^2 .
\end{aligned}$$

Fourth Lemma: Improvement of the Matrix $\boldsymbol{B}_k$ After One Step of RBFGS (Form 1)

Lemma 151

$$\mathbf{E}\left[\left\|\boldsymbol{B}_{k+1}-\boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_k)} \mid x_k, \boldsymbol{B}_k\right] \leq (1-\rho)\left\|\boldsymbol{B}_k-\boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_k)} + \left\|\boldsymbol{H}_k-\boldsymbol{H}_\star\right\|^2_{F(\boldsymbol{H}_\star^{-1})}, \quad (399)$$

where

$$\rho \stackrel{def}{=} \inf_{x\in\mathbb{R}^d} \lambda_{\min}\left(\mathbf{E}\left[\boldsymbol{H}_x^{1/2}\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{H}_x\boldsymbol{S})^{-1}\boldsymbol{S}^\top\boldsymbol{H}_x^{1/2}\right]\right). \quad (400)$$

Proof of Lemma 151 - I

Define the random projection matrix:

$$\boldsymbol{Z}_k \stackrel{\text{def}}{=} \boldsymbol{H}_k^{1/2}\boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{H}_k \boldsymbol{S})^{-1}\boldsymbol{S}^\top \boldsymbol{H}_k^{1/2}. \tag{401}$$

Using the just established object $\boldsymbol{Z}_k$, the BFGS update can be written in the form

$$\boldsymbol{B}_{k+1} \stackrel{(380)}{=} \boldsymbol{H}_k^{-1/2}\left(\boldsymbol{Z}_k + (\boldsymbol{I}-\boldsymbol{Z}_k)\boldsymbol{H}_k^{1/2}\boldsymbol{B}_k\boldsymbol{H}_k^{1/2}(\boldsymbol{I}-\boldsymbol{Z}_k)\right)\boldsymbol{H}_k^{-1/2}. \tag{402}$$

To abbreviate the calculations, we define the linear operator $P : \mathbb{S}^d \to \mathbb{S}^d$ via:

$$P(\boldsymbol{A}) \stackrel{\text{def}}{=} (\boldsymbol{I}-\boldsymbol{Z}_k)\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{Z}_k). \tag{403}$$

Using this notation, we can write:

$$\begin{aligned}
\left\|\boldsymbol{B}_{k+1}-\boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_k)} &\stackrel{(379)}{=} \left\|\boldsymbol{H}_k^{1/2}(\boldsymbol{B}_{k+1}-\boldsymbol{H}_\star^{-1})\boldsymbol{H}_k^{1/2}\right\|_F^2 \\
&\stackrel{(402)+(403)}{=} \left\|\boldsymbol{Z}_k + P(\boldsymbol{H}_k^{1/2}\boldsymbol{B}_k\boldsymbol{H}_k^{1/2}) - \boldsymbol{H}_k^{1/2}\boldsymbol{H}_\star^{-1}\boldsymbol{H}_k^{1/2}\right\|_F^2 \\
&= \left\|P(\boldsymbol{H}_k^{1/2}(\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1})\boldsymbol{H}_k^{1/2}) + P(\boldsymbol{H}_k^{1/2}\boldsymbol{H}_\star^{-1}\boldsymbol{H}_k^{1/2}) + \boldsymbol{Z}_k - \boldsymbol{H}_k^{1/2}\boldsymbol{H}_\star^{-1}\boldsymbol{H}_k^{1/2}\right\|_F^2.
\end{aligned}$$

Now note that the following equality holds

$$\boldsymbol{Z}_k = \boldsymbol{I} - P(\boldsymbol{I}) = \boldsymbol{H}_k^{1/2}\boldsymbol{H}_k^{-1}\boldsymbol{H}_k^{1/2} - P(\boldsymbol{H}_k^{1/2}\boldsymbol{H}_k^{-1}\boldsymbol{H}_k^{1/2}). \tag{404}$$

Using this equality and the shorthand

$$\boldsymbol{R}_k \stackrel{\text{def}}{=} \boldsymbol{H}_k^{1/2}(\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1})\boldsymbol{H}_k^{1/2} \in \mathbb{S}^d \qquad \text{and} \qquad \boldsymbol{D}_k \stackrel{\text{def}}{=} \boldsymbol{H}_k^{1/2}(\boldsymbol{H}_\star^{-1} - \boldsymbol{H}_k^{-1})\boldsymbol{H}_k^{1/2} \in \mathbb{S}^d,$$

Proof of Lemma 151 - II

we have that

$$\left\|\boldsymbol{B}_{k+1} - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 = \left\|P(\boldsymbol{H}_k^{1/2}(\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1})\boldsymbol{H}_k^{1/2}) + P(\boldsymbol{H}_k^{1/2}(\boldsymbol{H}_\star^{-1} - \boldsymbol{H}_k^{-1})\boldsymbol{H}_k^{1/2}) - \boldsymbol{H}_k^{1/2}(\boldsymbol{H}_\star^{-1} - \boldsymbol{H}_k^{-1})\boldsymbol{H}_k^{1/2}\right\|_F^2$$
$$= \underbrace{\|P(\boldsymbol{R}_k)\|_F^2}_{(I)} + \underbrace{\|P(\boldsymbol{D}_k) - \boldsymbol{D}_k\|_F^2}_{(II)} + \underbrace{2\left\langle P(\boldsymbol{R}_k), P(\boldsymbol{D}_k) - \boldsymbol{D}_k\right\rangle}_{(III)}. \tag{405}$$

One can show that $(III) = 0$. Note that (I) has the following upper-bound:

$$\begin{aligned}\|P(\boldsymbol{R}_k)\|_F^2 &= \mathrm{Trace}\left(\boldsymbol{R}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{R}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\right)\\ &= \mathrm{Trace}\left(\boldsymbol{R}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{R}_k\right) - \mathrm{Trace}\left(\boldsymbol{Z}_k\boldsymbol{R}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{R}_k\boldsymbol{Z}_k\right)\\ &\leq \mathrm{Trace}\left(\boldsymbol{R}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{R}_k\right).\end{aligned} \tag{406}$$

Taking expectation conditioned on x_k now gives

$$\begin{aligned}\mathbf{E}\left[\|P(\boldsymbol{R}_k)\|_F^2 \mid x_k, \boldsymbol{B}_k\right] &\leq \mathrm{Trace}\left(\boldsymbol{R}_k\left(\boldsymbol{I} - \mathbf{E}\left[\boldsymbol{Z}_k\right]\right)\boldsymbol{R}_k\right)\\ &\leq \left(1 - \lambda_{\min}(\mathbf{E}\left[\boldsymbol{Z}_k\right])\right)\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 \qquad (407)\\ &\overset{(400)}{\leq} (1-\rho)\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2, \qquad (408)\end{aligned}$$

where we used that $\mathrm{Trace}\left(\boldsymbol{AB}\right) \geq \lambda_{\min}(\boldsymbol{A})\mathrm{Trace}\left(\boldsymbol{B}\right)$ for any symmetric positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

Proof of Lemma 151 - III

Furthermore, (II) has the following upper-bound:

$$\begin{aligned} \|P(\boldsymbol{D}_k) - \boldsymbol{D}_k\|_F^2 &= \mathrm{Trace}\left([\boldsymbol{D}_k - (\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{Z}_k)]^2\right) \\ &= \|\boldsymbol{D}_k\|_F^2 - \mathrm{Trace}\left((\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\right) \\ &\leq \|\boldsymbol{H}_k - \boldsymbol{H}_\star\|_{F(\boldsymbol{H}_\star^{-1})}^2, \end{aligned} \tag{409}$$

where we used that

$$\mathrm{Trace}\left((\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{Z}_k)\right) \geq 0$$

since the matrix within this trace is symmetric positive semi-definite. Using the above together with (408) and (405) gives

$$\mathbf{E}\left[\left\|\boldsymbol{B}_{k+1} - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 \mid x_k, \boldsymbol{B}_k\right] \leq (1-\rho)\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 + \|\boldsymbol{H}_k - \boldsymbol{H}_\star\|_{F(\boldsymbol{H}_\star^{-1})}^2. \tag{410}$$

Fifth Lemma: Improvement of the Iterate After One Step of RBFGS (Form 2)

Lemma 152

$$\mathbf{E}\left[\left\|\boldsymbol{B}_{k+1}-\boldsymbol{H}_{\star}^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 \mid x_k, \boldsymbol{B}_k\right] \leq (1-\rho)\left\|\boldsymbol{B}_k-\boldsymbol{H}_{\star}^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 + \frac{2dL_2^2}{\mu^3}(f(x_k)-f(x_{\star})). \quad (411)$$

Proof of Lemma 152

Recalling (410) we have that

$$\mathbf{E}\left[\left\|\boldsymbol{B}_{k+1}-\boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 \mid x_k, \boldsymbol{B}_k\right] \leq (1-\rho)\left\|\boldsymbol{B}_k-\boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2+\left\|\boldsymbol{H}_k-\boldsymbol{H}_\star\right\|_{F(\boldsymbol{H}_\star^{-1})}^2 . \tag{412}$$

Now note that

$$\begin{aligned}\left\|\boldsymbol{H}_\star^{-1/2}(\boldsymbol{H}_k-\boldsymbol{H}_\star)\boldsymbol{H}_\star^{-1/2}\right\|_F^2 &\leq d\left\|\boldsymbol{H}_\star^{-1}\right\|_2^2\left\|\boldsymbol{H}_k-\boldsymbol{H}_\star\right\|_2^2 \\ &\overset{(381)+(383)}{\leq} \frac{dL_2^2}{\mu^2}\left\|x_k-x_\star\right\|_2^2 . \end{aligned} \tag{413}$$

Finally, taking expectation over (412) and using the above gives

$$\begin{aligned}\mathbf{E}\left[\left\|\boldsymbol{B}_{k+1}-\boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2 \mid x_k, \boldsymbol{B}_k\right] &\leq (1-\rho)\left\|\boldsymbol{B}_k-\boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2+\frac{dL_2^2}{\mu^2}\left\|x_k-x_\star\right\|_2^2 \\ &\leq (1-\rho)\left\|\boldsymbol{B}_k-\boldsymbol{H}_\star^{-1}\right\|_{F(\boldsymbol{H}_k)}^2+\frac{2dL_2^2}{\mu^3}(f(x_k)-f(x_\star)).\end{aligned}$$

Proof of Theorem 147

Proof of Theorem 147 - I

Having established several key lemmas, we are now ready to prove Theorem 147. We consider the Lyapunov function

$$\Psi_k = \sqrt{f(x_k) - f(x_\star)} + \beta \left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_\star)}^2, \tag{414}$$

where $\beta = \dfrac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho}$.

$$\begin{aligned}
\mathbf{E}\left[\Psi_{k+1}\right] &= \sqrt{f(x_{k+1}) - f(x_\star)} + \beta \mathbf{E}\left[\left\| \boldsymbol{B}_{k+1} - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_\star)}^2\right] \\
&\leq \sqrt{f(x_+) - f(x_\star)} + \beta \mathbf{E}\left[\left\| \boldsymbol{B}_{k+1} - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_\star)}^2\right] \\
&\overset{(398)+(388)}{\leq} \frac{\sqrt{2L_1}L_2}{\mu^2}(f(x_k) - f(x_\star)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2}\left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_k)}^2 \\
&\quad + \beta\left[1 + \frac{L_2}{\mu}\left\|x_k - x_\star\right\|_2\right]^2 \mathbf{E}\left[\left\| \boldsymbol{B}_{k+1} - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_k)}^2\right] \\
&\overset{(411)}{\leq} \frac{\sqrt{2L_1}L_2}{\mu^2}(f(x_k) - f(x_\star)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2}\left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_k)}^2 \\
&\quad + \beta\left[1 + \frac{L_2}{\mu}\left\|x_k - x_\star\right\|_2\right]^2 \left[(1-\rho)\left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_k)}^2 + \frac{2dL_2^2}{\mu^3}(f(x_k) - f(x_\star))\right] \\
&\leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{2\beta d L_2^2}{\mu^3}\left[1 + \frac{L_2}{\mu}\left\|x_k - x_\star\right\|_2\right]^2\right](f(x_k) - f(x_\star)) \\
&\quad + \left[(1-\rho)\left[1 + \frac{L_2}{\mu}\left\|x_k - x_\star\right\|_2\right]^4 + \frac{L_1^{5/2}}{2\beta\sqrt{2}\mu L_2}\left[1 + \frac{L_2}{\mu}\left\|x_k - x_\star\right\|_2\right]^2\right]\beta\left\| \boldsymbol{B}_k - \boldsymbol{H}_\star^{-1} \right\|_{F(\boldsymbol{H}_\star)}^2.
\end{aligned}$$

Proof of Theorem 147 - II

Now, since $f(x_k) - f(x_\star)$ is non-increasing by line 4 in Algorithm 28, we have that

$$\|x_k - x_\star\|_2 \leq \sqrt{\frac{2}{\mu}(f(x_k) - f(x_\star))} \leq \sqrt{\frac{2}{\mu}(f(x_0) - f(x_\star))} \leq \sqrt{\frac{2\mathcal{F}}{\mu}},$$

and hence

$$\begin{aligned}
\mathbf{E}\left[\Psi_{k+1}\right] \quad \leq \quad & \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{2\beta d L_2^2}{\mu^3}\left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}}\right]^2\right](f(x_k) - f(x_\star)) \\
& + \left[(1-\rho)\left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}}\right]^4 + \frac{L_1^{5/2}}{2\beta\sqrt{2}\mu L_2}\left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}}\right]^2\right]\beta\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_\star)}.
\end{aligned}$$

Equation (386) implies $\frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \leq 1$ and we get

$$\begin{aligned}
\mathbf{E}\left[\Psi_{k+1}\right] \quad \leq \quad & \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{8\beta d L_2^2}{\mu^3}\right](f(x_k) - f(x_\star)) \\
& + \left[(1-\rho)\left[1 + \frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}}\right] + \frac{\sqrt{2}L_1^{5/2}}{\beta\mu L_2}\right]\beta\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_\star)} \\
\leq \quad & \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{8\beta d L_2^2}{\mu^3}\right](f(x_k) - f(x_\star)) \\
& + \left[1 - \rho + \frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} + \frac{\sqrt{2}L_1^{5/2}}{\beta\mu L_2}\right]\beta\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_\star)}.
\end{aligned}$$

Proof of Theorem 147 - III

Using $\beta = \frac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho}$ and using $\frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \leq \frac{\rho}{4}$, which follows from (386), we get

$$\mathbf{E}\left[\Psi_{k+1}\right] \leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{32\sqrt{2}dL_1^{5/2}L_2}{\rho\mu^4}\right]\sqrt{\mathcal{F}\cdot(f(x_k)-f(x_\star))} + \left(1-\frac{\rho}{2}\right)\beta\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_\star)}.$$

Finally, using (386) we get

$$\mathbf{E}\left[\Psi_{k+1}\right] \leq \frac{1}{2}\sqrt{(f(x_k)-f(x_\star))} + \left(1-\frac{\rho}{2}\right)\beta\left\|\boldsymbol{B}_k - \boldsymbol{H}_\star^{-1}\right\|^2_{F(\boldsymbol{H}_\star)} \leq \left(1-\frac{\rho}{2}\right)\Psi_k. \tag{415}$$

Exercises

Exercise 101

Prove that $\|AB\|_F \leq \|A\|_F \|B\|_2$.

Exercise 102

Prove the equivalence theorem.

Exercise 103

Study the expression

$$\rho \stackrel{def}{=} \inf_{x \in \mathcal{Q}} \lambda_{\min} \left(\mathbf{E} \left[\boldsymbol{H}_x^{1/2} \boldsymbol{S} (\boldsymbol{S}^\top \boldsymbol{H}_x \boldsymbol{S})^{-1} \boldsymbol{S}^\top \boldsymbol{H}_x^{1/2} \right] \right).$$

- *Evaluate* ρ *in the case when* $\boldsymbol{S} = \boldsymbol{I}$ *with probability 1.*
- *Assume* $\boldsymbol{H}_x = \boldsymbol{H}$ *for all* x *(i.e.,* f *is a quadratic). Let* $\boldsymbol{S} = e_i$ *with probability* $1/d$*, where* $\{e_1, \ldots, e_d\}$ *are the standard unit basis vectors in* $\mathbb{R}^d$*. Evaluate* ρ*.*
- *Prove that whenever* ρ *is well defined, then* $\rho \leq 1$*.*

CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 25: Federated Newton Learn (FedNL)

Based on:

[34] Rustem Islamov, Xun Qian, and P.R., **Distributed second order methods with fast rates and compressed communication**, ICML 2021.
[71] Mher Safaryan, Rustem Islamov, Xun Qian, and P.R., **FedNL: Making Newton-type methods applicable to federated learning**, arXiv:2106.02969, 2021.

Introduction

We consider the problem of distributed optimization

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right],$$

where n is the number of nodes, using **Newton-type methods.**

- We describe why Newton's method is inefficient in this regime (necessity to send $d \times d$ Hessian matrices from nodes to the server).
- We describe `FedNL` **(Federated Newton Learn) a compressed variance reduced Newton-type method able to work with unbiased and biased compressors.**
  - In brief, `FedNL` can be thought of as Newton-type method where the Hessian matrices are replaced by their `EF21` estimates!
  - The above description is almost but not entirely correct. Still, it conveys the basic idea.

Remark: We will go through selected parts of the paper in class. There are no slides.

Table of Contents

Contents I

1. Course Organization
   - About Me
   - Logistics
   - Brief Course Description
   - Goals and Objectives
   - Knowledge Required
   - Reference Text
   - Method of Evaluation
2. Introduction to Optimization
   - Optimization Problems and Solution Concepts
   - Optimization Problems
   - Types of Optimization Problems
   - Solution Concepts
   - Optimization Modelling and Data
   - Optimization Modelling
   - Optimization and Data
   - Optimization Algorithms
   - Basic Algorithms
   - Convergence
   - What Quantity do we Want to Converge?
   - Typical Convergence Rates Encountered in Optimization
   - Recursions & Associated Convergence Rates
3. Empirical Risk Minimization
   - Notation

Contents II

Supervised Machine Learning
Optimization Problems Arising in Machine Learning
Stochastic Optimization and Machine Learning
Finite Sum Problems
Distributed & Federated Training
Proximal SGD
ZOO of Stochastic Gradients
The Proximal Operator
Exercises

4. Gradient Descent
Gradient Descent
Convexity and L-smoothness
Strong Convexity
A Descent Lemma
Convergence of Gradient Descent
Exercises

5. Convexity and Smoothness
Differentiable Functions
Fundamental Theorem of Calculus
Bregman Divergence
Convex Functions
Convex Functions: Definition
Convex Functions: Characterization

Contents III

Convex Functions: Examples
Proof of a Proposition
Jensen's Inequality
Strongly Convex Functions
Strongly Convex Functions: Definition
Strongly Convex Functions: Examples
Strongly Convex Functions: Characterization
Smooth Functions
Smooth Functions: Definition
Smooth Functions: Examples
Smooth Functions: Characterization
Reverse Jensen's inequality
Convex and Smooth Functions
Convex and Smooth Functions: Characterization
Jensen's and Reverse Jensen's inequality
Strongly Convex and Smooth Functions
Exercises

6. Stochastic Gradient Descent
Basic Facts
Tower Property of Conditional Expectation
From GD to SGD: Theory Viewpoint
SGD in Generic Form
Analysis of Generic SGD

Contents IV

Bounding The Third Term
Key Assumption
Convergence of SGD
Convergence Theory: Commentary
Proof of the Main Result
More on the AC Assumption
Bounding $G(x, y)$: First Approach
Bounding $G(x, y)$: Second Approach
Exercises

7. SGD with Simple Sampling
Introduction
SGD-NS: the Algorithm
Computing the AC Constants
Complexity of SGD-NS
Observation 1: Interpolation Regime
Zero Variance
Interpolation Regime: Definition
Interpolation Regime: Example
Observation 2: Uniform Sampling
Uniform Sampling: SGD-US
Comparing GD and SGD-US in the Interpolation Regime
Example: Large L_i
Example: Small L_i

Contents V

Generalizing the Examples
Observation 3: Importance Sampling
The Many Faces of Importance Sampling
Exercises

8. SGD with Minibatch Sampling
Introduction
Sampling without Replacement
Sampling without Replacement: Nice Sampling
SGD-NICE: Unbiasedness and Expected Smoothness
Commentary
Proof of a Lemma
Convergence of SGD-NICE
Optimal Minibatch Size
Sampling without Replacement
Sampling without Replacement: Independent Sampling
SGD-IND: The Algorithm
Minibatch Size
SGD-IND: Unbiasedness and Expected Smoothness
Commentary
Proof of a Lemma
Sampling with Replacement
Sampling with Replacement: Multisampling
SGD-MULTI: Unbiasedness and Expected Smoothness

Contents VI

Commentary
Proof of a Lemma
Exercises

9. CGD: Compressed Gradient Descent
Distributed Training
Distributed Gradient Descent
Communication Bottleneck
Gradient Descent with Compressed Communication
Trade-off Between Compression and Convergence
Compression Operators
Quick-and-Dirty Examples of Compression Operators
Distributed Gradient Descent with Compressed Gradients
Unbiasedness
Proof of a Lemma
CGD
Compressed Gradient Descent
The AC Inequality for CGD
Convergence of CGD
Iteration Complexity
Analysis in the General Case
The AC Inequality for DCGD
Commentary
Proof of a Lemma

Contents VII

The AC Inequality for DCGD without Master-Workers Compression
Exercises

10. RCD: Randomized Coordinate Descent
Sketches: Linear Compression Operators
Diagonal Sketches
General Random Sparsification -
RCD: Randomized Coordinate Descent
Convergence Analysis I
RCD as a Direct Special Case of CGD
RCD vs GD
Convergence Analysis II
No Regularizer
Matrix Smoothness
Key Lemma
AC Inequality
Convergence Theorem
Estimation of Smoothness Constant
Case Study: Diagonal Smoothness Matrix and Single Element Sampling
Convergence Analysis III
No Regularizer
Main Theorem
Exercises

11. Compression Operators

Contents VIII

Manipulating Compression Operators
Composition of Compression Operators
Direct Sum of Compression Operators
Perspective of a Compression Operator
Probabilistic Switching Among Compression Operators
Convex Combination of Compression Operators
Quantization Operators
Natural Compression: Informal Definition
Natural Compression: Examples
Natural Compression: Formal Definition
Natural Compression: Variance
Implementation of Natural Compression
General Dithering and Special Cases
Ternary Quantization
Standard Dithering
Natural Dithering
Natural Dithering
Natural Dithering: Variance
Natural Dithering: Proof of the Key Result
Hölder's Inequality
Consequences of Hölder's Inequality
Exercises

Contents IX

Introduction
Nobody Likes the Neighborhood
Designing an Estimator for Which $C(x^\star) = 0$
Shifted Gradient Estimators for Compression
Bias and Variance
Compressed Gradient Descent
Proof of a Lemma
Convergence of CGD-SHIFT
Shifted Compression Operators
Zero is Special
Shifted Compression Operators
Shifting a Shifted Compression Operator
Representation of Shifted Compression Operators
Re-examining the Proof
Shifted Gradient Estimator for Subsampling
Two Special Shifts
Shifted SGD
Shifted Gradient Estimator: Expected Smoothness
Summary: Gradient Estimators Satisfying Expected Smoothness
Exercises

13. General Analysis of Variance Reduced SGD
Recall: Key Lemma in the Analysis of Generic SGD
Stochastic Gradient with Control

Contents X

SGD-CTRL: SGD with Control
SGD-CTRL: First Lemma
σ^k Assumption: A More General Model for the Evolution of the Stochastic Gradients
Variance-reduced Methods
Main Result
Main Result: Commentary
A Key Lemma
Proof of a Lemma
Proof of the Main Result
Exercises

14. Loopless SVRG (L-SVRG)
Motivation
The Glory and Plight of Shifted Gradient Estimators
L-SVRG: SGD with Learned Shift
L-SVRG and σ^k Assumption
Proof of a Lemma
Convergence of L-SVRG
L-SVRG-US
L-SVRG-US: The Algorithm
L-SVRG-US: Cost per Iteration
Total Complexity
Comparison of Total Complexities
Insight

Contents XI

Exercises

15. Compression with Learned Shift (DIANA)
The Problem
CGD-SHIFT
DCGD-SHIFT
Challenges with Learning the Optimal Shifts
DIANA: Resolving Challenge 1
DIANA: Resolving Challenge 2
DIANA for $n = 1$
Establishing the σ^k Assumption
Proof of a Lemma
Convergence of DIANA for $n = 1$
Understanding the Rate
Distributed Case
Two Variance Reduction Techniques: Fixed and Learned Shift
DIANA: Algorithm
σ^k Assumption for DIANA for General n
Convergence of DIANA
Understanding the Rate
Extensions
Exercises

16. SAGA
The Problem

Contents XII

Variance Reduced Gradient Estimators
L-SVRG vs SAGA: Comparison of Gradient Estimators I
L-SVRG vs SAGA: Comparison of Gradient Estimators II
The SAGA Algorithm
Unbiasedness of the Stochastic Gradient
The σ^k Assumption
Convergence of SAGA
Number of Iterations: L-SVRG vs SAGA
Proofs
Understanding SAGA
Jacobian Learning
Jacobian Learning: Lifting
Jacobian Learning: Sketching
Jacobian Learning: Random Linear System
Jacobian Learning: Projection
Jacobian Learning: Gradient Estimate
Jacobian Learning: SGD
The SAGA Algorithm: Jacobian Learning View
Controlled Stochastic Reformulation
CSR: Simple Stochastic Reformulation
CSR: Zero Mean Random Function
CSR
The SAGA Algorithm: CSR View

Contents XIII

Extensions
Exercises

17. Nonconvex Setting: the ABC Assumption
The Problem
Goal: Finding a Stationary Point
Some Initial Thinking
A Key Lemma
Proof of Key Lemma
The ABC Assumption
Existing Models of Stochastic Gradient
Proof of a Proposition: (RG) Does not Hold
(ABC) is the Weakest Condition
Proof of Theorem: (ABC) is the Weakest Condition
General Convergence Theory
Lemma Providing a Key Recursion
The Main Convergence Result
Proof of The Main Convergence Result
Corollary: Iteration Complexity
Corollary: Proof
Proof of the Key Recursion Lemma
Exercises

18. Nonconvex Setting: the ABC Assumption and the Polyak-Łojasiewicz Inequality
Plan

Contents XIV

Two Relaxations of Strong Convexity
More Relaxations of Strong Convexity
Smart Stepsizes for a Key Recursion
Proof of a Lemma
Convergence of SGD under PL Condition
Proof of a Theorem
Corollary: Iteration Complexity
Exercises

19. Variance Reduction for Nonconvex SGD
Introduction
Nonconvex σ^k Assumption
Lemma: A New Recursion
Lemma: Commentary
Lemma: Proof
Main Theorem
Main Theorem: Proof
Special Cases: GD
Special Cases: SGD
Special Cases: L-SVRG
A Formula for Sample Variance
A New Assumption
A Tower Property with Three Random Variables
L-SVRG-NICE Satisfies a Variant of the Nonconvex σ^k Assumption

Contents XV

Lemma: Proof
Rate of L-SVRG
Exercises

20. PAGE: A Simple Optimal SGD for Nonconvex Problems
The Problem
Main Takeaways
PAGE: The Algorithm
PAGE Estimator vs L-SVRG-NICE Estimator
Convergence Theorem
Convergence Theorem: Commentary
Proofs
Descent Identity
Proof of Descent Identity
Linear Inequality Implying a Quadratic Inequality
Proof of Convergence Theorem
Exercises

21. EF21: Dealing with Contractive Compressors via Error Feedback
Introduction
Plan
Contractive Compressors
Compressors in $\mathbb{B}^d$ are Contractive after Scaling
Top-K: A Greedy Sparsification Strategy
Failure of DCGD with Biased Compressors

Contents XVI

Rank-K: Compression via Low Rank Approximation
Reshaping of Vectors into Matrices
Singular Value Decomposition (SVD)
Best Low Rank Approximation of Matrices
Rank-K is Contractive
Proof
The EF21 Algorithm
Theory for EF21
Assumptions
Theory in General Nonconvex Case
Extensions
Exercises

22. MARINA: Distributed Compressed Gradient Descent with a Biased Estimator
Outline
Content

23. Stochastic Newton Method
First-order Methods
Second-order Methods
The Algorithm
Generic Efficient Implementation
Total Cost
Local Convergence Analysis
Lipschitz Hessian

Contents XVII

Main Theorem
Two Lemmas
Third Lemma
Proof of the Main Theorem

24. Randomized BFGS
RBFGS: The Algorithm
Assumptions
Main Theorem
Five Lemmas
Proof of the Main Theorem

25. Federated Newton Learn (FedNL)

Bibliography

Bibliography I

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic.
QSGD: Communication-efficient SGD via gradient quantization and encoding.
In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[2] Zeyuan Allen-Zhu.
Katyusha: The first direct acceleration of stochastic gradient methods.
In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

[3] Amir Beck.
First-Order Methods in Optimization.
Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[4] Dimitri P. Bertsekas and John N. Tsitsiklis.
Gradient Convergence in Gradient methods with Errors.
SIAM Journal on Optimization, 10(3):627–642, January 2000.

[5] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan.
On biased compression for distributed learning.
arXiv:2002.12410, 2020.

Bibliography II

[6] Antoine Bordes, Léon Bottou, and Patrick Gallinari.
SGD-QN: careful quasi-Newton stochastic gradient descent.
Journal of Machine Learning Research, 10:1737–1754, 2009.

[7] Léon. Bottou, Frank E. Curtis, and Jorge. Nocedal.
Optimization Methods for Large-Scale Machine Learning.
SIAM Review, 60(2):223–311, 2018.

[8] Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal.
On the use of stochastic Hessian information in optimization methods for machine learning.
SIAM Journal on Optimization, 21(3):977–995, 2011.

[9] Sélim Chraibi, Ahmed Khaled, Dmitry Kovalev, Adil Salim, Peter Richtárik, and Martin Takáč.
Distributed fixed point methods with compressed iterates.
arXiv preprint arXiv:1912.09925, 2019.

[10] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien.
SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives.
In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

Bibliography III

[11] Peter Richtárik Alexander Rogozin Dmitry Kovalev, Robert M. Gower.
Fast linear convergence of randomized bfgs.
arXiv:2002.11337, 2020.

[12] Nikita Doikov and Peter Richtárik.
Randomized block cubic Newton method.
In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[13] Yoel Drori and Ohad Shamir.
The Complexity of Finding Stationary Points with Stochastic Gradient Descent.
arXiv preprint arXiv:1910.01845, 2019.

[14] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang.
SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator.
In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

[15] Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik.
Ef21 with bells & whistles: practical algorithmic extensions of modern error feedback.
arXiv preprint arXiv:2110.03294, 2021.

Bibliography IV

[16] Saeed Ghadimi and Guanghui Lan.
Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming.
SIAM Journal on Optimization, 23(4):2341–2368, 2013.

[17] W. M. Goodall.
Television by pulse code modulation.
The Bell System Technical Journal, 30(1):33–49, Jan 1951.

[18] Eduard Gorbunov, Konstantin Burlachenko, Zhize Li, and Peter Richtárik.
MARINA: Faster non-convex distributed learning with compression.
In *38th International Conference on Machine Learning*, 2021.

[19] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik.
A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent.
In *AISTATS*, 2020.

[20] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik.
Linearly converging error compensated SGD.
In *Neural Information Processing Systems*, 2020.

Bibliography V

[21] Robert M. Gower, Donald Goldfarb, and Peter Richtárik.
Stochastic block BFGS: Squeezing more curvature out of data.
Proceedings of the 33rd International Conference on Machine Learning, 2016.

[22] Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik.
RSN: Randomized Subspace Newton.
In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[23] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik.
SGD: General analysis and improved rates.
In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[24] Robert Mansel Gower and Peter Richtárik.
Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms.
SIAM Journal on Matrix Analysis and Applications, 38(4):1380–1409, 2017.

Bibliography VI

[25] Robert Mansel Gower, Peter Richtárik, and Francis Bach.
Stochastic quasi-gradient methods: variance reduction via Jacobian sketching.
Mathematical Programming, 188:135–192, 2021.

[26] Andreas Griewank.
The modification of Newton's method for unconstrained optimization by bounding cubic terms.
Technical report, Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1981.

[27] Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik.
SEGA: Variance reduction via gradient sketching.
In *Advances in Neural Information Processing Systems 31*, pages 2082–2093, 2018.

[28] Filip Hanzely and Peter Richtárik.
One method to rule them all: variance reduction for data, parameters and many new methods.
arXiv preprint arXiv:1905.11266, 2019.

[29] Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams.
Variance reduced stochastic gradient descent with neighbors.
In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

Bibliography VII

[30] Samuel Horváth, Chen-Yu Ho, Ľudovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik.
Natural compression for distributed deep learning.
arXiv preprint arXiv:1905.10988, 2019.

[31] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Peter Richtárik, and Sebastian Stich.
Stochastic distributed learning with gradient quantization and variance reduction.
arXiv preprint arXiv:1904.05115, 2019.

[32] Samuel Horváth, Lihua Lei, Peter Richtárik, and Michael I.Jordan.
Adaptivity of stochastic gradient methods for nonconvex optimization.
SIAM Journal on Mathematics of Data Science, 4(2):634–648, 2022.

[33] Samuel Horváth and Peter Richtárik.
A better alternative to error feedback for communication-efficient distributed learning.
In *9th International Conference on Learning Representations (ICLR)*, 2021.

[34] Rustem Islamov, Xun Qian, and Peter Richtárik.
Distributed second order methods with fast rates and compressed communication.
In *International Conference on Machine Learning*, 2021.

Bibliography VIII

[35] Rie Johnson and Tong Zhang.
Accelerating stochastic gradient descent using predictive variance reduction.
In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

[36] Hamed Karimi, Julie Nutini, and Mark Schmidt.
Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition.
In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, page 795–811, Berlin, Heidelberg, 2016. Springer-Verlag.

[37] Ahmed Khaled and Peter Richtárik.
Better theory for SGD in the nonconvex world.
arXiv Preprint arXiv:2002.03329, 2020.

[38] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson.
Distributed learning with compressed gradients.
arXiv preprint arXiv:1806.06573, 2018.

[39] Jakub Konečný and Peter Richtárik.
Semi-stochastic gradient descent methods.
Frontiers in Applied Mathematics and Statistics, pages 1–14, 2017.

Bibliography IX

[40] Dmitry Kovalev, Samuel Horváth, and Peter Richtárik.
Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop.
In *31st International Conference on Algorithmic Learning Theory*, 2020.

[41] Dmitry Kovalev, Anastasia Koloskova, Martin Jaggi, Peter Richtárik, and Sebastian Stich.
A linearly convergent algorithm for decentralized optimization: Sending less bits for free!
Technical report, KAUST, 2020.

[42] Dmitry Kovalev, Konstanting Mishchenko, and Peter Richtárik.
Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates.
In *NeurIPS Beyond First Order Methods Workshop*, 2019.

[43] Yunwei Lei, Ting Hu, Guiying Li, and Ke Tang.
Stochastic Gradient Descent for Nonconvex Learning Without Bounded Gradient Assumptions.
IEEE Transactions on Neural Networks and Learning Systems, pages 1–7, 2019.

Bibliography X

[44] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik.
Acceleration for compressed gradient descent in distributed and federated optimization.
In *International Conference on Machine Learing*, 2020.

[45] Zhize Li and Peter Richtárik.
A unified analysis of stochastic gradient methods for nonconvex federated optimization.
arXiv preprint arXiv:2006.07013, 2020.

[46] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas.
Communication-Efficient Learning of Deep Networks from Decentralized Data.
In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[47] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik.
Distributed learning with compressed gradient differences.
arXiv preprint arXiv:1901.09269, 2019.

Bibliography XI

[48] Konstantin Mishchenko, Filip Hanzely, and Peter Richtárik.
99needed.
In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 979–988, 2020.

[49] I Necoara, Y. Nesterov, and F. Glineur.
Linear Convergence of First Order Methods for Non-Strongly Convex Optimization.
Mathematical Programming, 175(1–2):69–107, May 2019.

[50] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro.
Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609, 2009.

[51] Yurii Nesterov.
Introductory lectures on convex optimization: a basic course (Applied Optimization).
Kluwer Academic Publishers, 2004.

[52] Yurii Nesterov.
Efficiency of coordinate descent methods on huge-scale optimization problems.
SIAM Journal on Optimization, 22(2):341–362, 2012.

Bibliography XII

[53] Yurii Nesterov and Boris T. Polyak.
Cubic regularization of Newton method and its global performance.
Mathematical Programming, 108(1):177–205, 2006.

[54] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč.
SARAH: A novel method for machine learning problems using stochastic recursive gradient.
In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621. PMLR, 2017.

[55] Phuong Ha Nguyen, Lam Nguyen, and Marten van Dijk.
Tight Dimension Independent Lower Bound on the Expected Convergence Rate for Diminishing Step Sizes in SGD.
In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3660–3669. Curran Associates, Inc., 2019.

[56] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh.
ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization.
arXiv preprint arXiv:1902.05679, 2019.

Bibliography XIII

[57] Xun Qian, Zheng Qu, and Peter Richtárik.
SAGA with arbitrary sampling.
In *The 36th International Conference on Machine Learning*, 2019.

[58] Xun Qian, Zheng Qu, and Peter Richtárik.
L-SVRG and L-Katyusha with arbitrary sampling.
Journal of Machine Learning Research, 2020.

[59] Zheng Qu and Peter Richtárik.
Coordinate descent with arbitrary sampling I: Algorithms and complexity.
Optimization Methods and Software, 31(5):829–857, 2016.

[60] Zheng Qu and Peter Richtárik.
Coordinate descent with arbitrary sampling II: Expected separable overapproximation.
Optimization Methods and Software, 31(5):858–884, 2016.

[61] Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq.
SDNA: Stochastic dual Newton ascent for empirical risk minimization.
In *The 33rd International Conference on Machine Learning*, pages 1823–1832, 2016.

Bibliography XIV

[62] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan.
Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization.
In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[63] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola.
Stochastic variance reduction for nonconvex optimization.
In *The 33th International Conference on Machine Learning*, pages 314–323, 2016.

[64] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin.
EF21: A new, simpler, theoretically better, and practically faster error feedback.
In *Advances in Neural Information Processing Systems*, 2021.

[65] Peter Richtárik and Martin Takáč.
On optimal probabilities in stochastic coordinate descent methods.
Optimization Letters, 10(6):1233–1243, 2016.

[66] Peter Richtárik and Martin Takáč.
Parallel coordinate descent methods for big data optimization.
Mathematical Programming, 156(1-2):433–484, 2016.

Bibliography XV

[67] H. Robbins and S. Monro.
A stochastic approximation method.
Annals of Mathematical Statistics, 22:400–407, 1951.

[68] L. Roberts.
Picture coding using pseudo-random noise.
IRE Transactions on Information Theory, 8(2):145–154, February 1962.

[69] Nicolas Le Roux, Mark Schmidt, and Francis Bach.
A stochastic gradient method with an exponential convergence rate for finite training sets.
In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[70] Mher Safaryan, Filip Hanzely, and Peter Richtárik.
Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization.
In *ICLR Workshop: Distributed and Private Machine Learning*, 2021.

[71] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik.
FedNL: Making Newton-type methods applicable to federated learning.
arXiv preprint arXiv:2106.02969, 2021.

Bibliography XVI

[72] Karthik A. Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein.

The Impact of Neural Network Overparameterization on Gradient Confusion and Stochastic Gradient Descent.

arXiv preprint arXiv:1904.06963, 2019.

[73] Mark Schmidt and Nicolas Le Roux.

Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition.

arXiv preprint arXiv:1308.6370, 2013.

[74] Nicol N Schraudolph and G Simon.

A Stochastic Quasi-Newton Method for Online Convex Optimization.

In Proceedings of 11th International Conference on Artificial Intelligence and Statistics, 2007.

[75] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu.

1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs.

pages 1058–1062. ISCA, 2014.

Bibliography XVII

[76] Shai Shalev-Shwartz and Shai Ben-David.
Understanding machine learning: from theory to algorithms.
Cambridge University Press, 2014.

[77] Ohad Shamir and Tong Zhang.
Stochastic Gradient Descent for Non-Smooth Optimization: Convergence Results and Optimal Averaging Schemes.
In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–71–I–79. JMLR.org, 2013.

[78] M.V. Solodov.
Incremental Gradient Algorithms with Stepsizes Bounded Away from Zero.
Computational Optimization and Applications, 11(1):23–35, 1998.

[79] Sebastian U. Stich.
Unified Optimal Analysis of the (Stochastic) Gradient Method.
arXiv preprint arXiv:1907.04232, 2019.

[80] Rafał Szlendak, Alexander Tyurin, and Peter Richtárik.
Permutation compressors for provably faster distributed nonconvex optimization.
In *10th International Conference on Learning Representations (ICLR)*, 2022.

Bibliography XVIII

[81] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu.

`DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression.

In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6155–6165, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[82] Paul Tseng.

An Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule.

SIAM Journal on Optimization, 8(2):506–531, May 1998.

[83] Sharan Vaswani, Francis Bach, and Mark Schmidt.

Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron.

In *22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pages 1195–1204, 2019.

Bibliography XIX

[84] Sharan Vaswani, Francis Bach, and Mark Schmidt.

Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron.

In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 16–18 Apr 2019.

[85] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh.

Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization.

arXiv preprint arXiv:1810.10690, 2018.

[86] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li.

Terngrad: Ternary gradients to reduce communication in distributed deep learning.

In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

Bibliography XX

[87] Peilin Zhao and Tong Zhang.
Stochastic optimization with importance sampling for regularized loss minimization.
In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, volume 37, pages 1–9, 2015.

[88] Xiangliang Zhang Zhize Li, Hongyan Bao and Peter Richtárik.
PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization.
In *ICML 2021*, 2021.