Randomized Coordinate Descent for Big Data Optimization (Theory)

©2014 Peter Richtárik

University of Edinburgh

Grenoble, June 11-12, 2014



Contents I

1. NSync

Samplings Assumptions Complexity Proof

2. Blocks

Decomposition Projection Norms

3. Samplings

Definition Sampling Zoo Basic Identity Consequences of the Basic Identity Identities for Uniform Samplings Identities for Doubly Uniform Samplings Elementary Samplings Probability Matrices



2/116

臣

・ロト ・回ト ・ヨト ・ヨト

Contents II

Sampling Identity for a Quadratic Distributed Sampling

- 4. Functions
 - $\mathsf{Model}\ 1$
 - Model 2
 - Model 3
- 5. ESO
 - Model 1 General ESO Bounds Eigenvalues of Probability Matrices ESO 2 ESO2 Bounds Product Sampling τ -Nice Sampling Distributed τ -Nice Sampling Distributed NSync Model 3 ESO DSO ESO and Lipschitz Continuity



Contents III 6. APPROX Algorithm Complexity 4 Lemmas Lemmas

Proof of the Main Theorem



Lecture 1 NSync



The Problem

In order to quickly illustrate the important topics and notions that we will study in more depth later, we first consider the following simple problem:

minimize
$$f(x)$$
 (1)
subject to $x \in \mathbb{R}^n$

We will assume that f is:

- "smooth" (will be made precise later)
- strongly convex



Introduction to Parallel Coordinate Descent

This **NSync algorithm** was introduced in a brief 5p paper by R. and Takáč [11] and was meant to be an entry point to the field of parallel coordinate descent.

Algorithm (NSync)

Input: initial point $x_0 \in \mathbb{R}^n$

subset probabilities $\{p_S\}$ for each $S \subseteq [n] \stackrel{\text{def}}{=} \{1, 2, ..., n\}$ stepsize parameters $v_1, ..., v_n > 0$

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

a) Select a random set of coordinates $S_k \subseteq [n]$ following the law

$$\mathbf{P}(S_k=S)=p_S,\qquad S\subseteq [n]$$

b) Update (possibly in parallel) selected coordinates:

$$x_{k+1} = x_k - \sum_{i \in S_k} \frac{1}{v_i} (e_i^T \nabla f(x_k)) e_i$$

イロト イポト イヨト イヨト 二日

end for

Two More Ways of Writing the Update Step

1. Coordinate-by-coordinate:

$$x_{k+1}^{(i)} = \begin{cases} x_k^{(i)}, & i \notin S_k, \\ x_k^{(i)} - \frac{1}{v_i} (\nabla f(x_k))^{(i)}, & i \in S_k. \end{cases}$$

2. Via projection to a subset of blocks: If for $h \in \mathbb{R}^n$ and $S \subseteq [n]$ we write

$$h_{[S]} \stackrel{\text{def}}{=} \sum_{i \in S} h^{(i)} e_i,$$

then

 $x_{k+1} = x_k + h_{[S_k]}$ for $h = -(\operatorname{Diag}(v))^{-1} \nabla f(x_k)$.

We shall interchangeably write:

$$\nabla_i f(x) = e_i^T \nabla f(x) = (\nabla f(x))^{(i)}$$

8/116

Samplings

Definition 1 (Sampling)

By the name **sampling** we will refer to a set valued random mapping with values being subsets of $[n] = \{1, 2, ..., n\}$. For sampling \hat{S} we define $p = (p_1, ..., p_n)^T$, where

$$p_i = \mathbf{P}(i \in \hat{S}) \tag{2}$$

We say that \hat{S} is proper, if $p_i > 0$ for all *i*. Lemma 2 ([5])

$$\sum_{i=1}^{n} p_i = \mathbf{E}[|\hat{S}|].$$
 (3)

Proof.

$$\sum_{i=1}^{n} p_i \stackrel{(2)}{=} \sum_{i=1}^{n} \sum_{S \subseteq [n]: i \in S} p_S = \sum_{S \subseteq [n]} \sum_{i: i \in S} p_S = \sum_{S \subseteq [n]} p_S |S| = \mathbf{E}[|\hat{S}|].$$

Assumption: Strong convexity

Assumption 1 (Strong convexity)

f is differentiable and γ -strongly convex with respect to the norm $\|\cdot\|_s$ (weighted Euclidean norm with weights $s = (s_1, \ldots, s_n)^T > 0$). That is, for all $x, h \in \mathbb{R}^n$,

$$f(x+h) \ge f(x) + \langle \nabla f(x), h \rangle + \frac{\gamma}{2} \|h\|_s^2.$$
(4)

Notation used above:

$$\|h\|_s \stackrel{\text{def}}{=} \left(\sum_{i=1}^n s_i(h^{(i)})^2\right)^{1/2}$$

(weighted Euclidean norm)

Assumption: Expected Separable Overapproximation Assumption 2 (ESO)

Assume \hat{S} is proper and that for some vector of positive weights $v = (v_1, \dots, v_n)$ and all $x, h \in \mathbb{R}^n$,

$$\mathbf{E}[f(x+h_{[\hat{S}]})] \le f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|_{p \bullet v}^2.$$
(5)

For simplicity, we will often write

$$(f, \hat{S}) \sim ESO(v).$$

Note that the ESO parameters v, p depend on both f and \hat{S} .

Notation used above:

 $h_{[S]} \stackrel{\text{def}}{=} \sum_{i \in S} h^{(i)} e_i \quad (\text{projection of } h \in \mathbb{R}^n \text{ onto coordinates } i \in S)$ $\langle g, h \rangle_p \stackrel{\text{def}}{=} \sum_{i=1}^n p_i g^{(i)} h^{(i)} \quad (\text{weighted inner product})$ $p \bullet v \stackrel{\text{def}}{=} (p^{(1)} v^{(1)}, \dots, p^{(n)} v^{(n)}) \quad (\text{Hadamard product})$

Complexity of NSync

Theorem 3 ([11])

Let x_* be a minimizer of f. Let Assumptions 1 and 2 be satisfied for a proper sampling \hat{S} (that is, $(f, \hat{S}) \sim ESO(v)$). Choose

- starting point $x_0 \in \mathbb{R}^n$,
- error tolerance $0 < \epsilon < f(x_0) f(x_*)$ and
- confidence level $0 < \rho < 1$.

If $\{x_k\}$ are the random iterates generated by NSync where the random sets S_k are iid following the distribution of \hat{S} , then

$$\mathbf{K} \geq \frac{\mathbf{\Lambda}}{\gamma} \log \left(\frac{f(x_0) - f(x_*)}{\epsilon \rho} \right) \implies \mathbf{P}(f(\mathbf{x}_{\mathbf{K}}) - f(x_*) \leq \epsilon) \geq 1 - \rho, \quad (6)$$

where

$$\Lambda \stackrel{def}{=} \max_{i=1,\ldots,n} \frac{v_i}{p_i s_i} \geq \frac{\sum_{i=1}^n \frac{v_i}{s_i}}{\mathbf{E}[|\hat{S}|]}.$$

(7)

(日) (部) (注) (注) (言)

What does this mean?

- Linear convergence. NSync converges linearly (i.e., logarithmic dependence on ε)
- High confidence is not a problem. ρ appears inside the logarithm, so it easy to achieve high confidence (by running the method longer; there is no need to restart)
- Focus on the leading term. The leading term is Λ; and we have closed from expression for it in terms of
 - parameters v_1, \ldots, v_n (which depend on f and \hat{S})
 - parameters p_1, \ldots, p_n (which depend on \hat{S})
- ▶ **Parallelization speedup.** The lower bound suggests that *if it was* the case that the parameters v_i did not grow with increasing $\tau \stackrel{\text{def}}{=} \mathbf{E}[|\hat{S}|]$, then we could potentially be getting linear speedup in τ (average number of updates per iteration).
 - So we shall study the dependence of v_i on τ (this will depend on f and Ŝ)
 - As we shall see, speedup does is often guaranteed for sparse problems.



Question: How to design sampling \hat{S} so that Λ is minimized?



Proof of Theorem 3 - Part I

• If we let $\mu \stackrel{\text{def}}{=} \gamma / \Lambda$, then

$$f(x+h) \stackrel{(4)}{\geq} f(x) + \langle \nabla f(x), h \rangle + \frac{\gamma}{2} \|h\|_{s}^{2}$$

$$\geq f(x) + \langle \nabla f(x), h \rangle + \frac{\mu}{2} \|h\|_{v \bullet p^{-1}}^{2}.$$
(8)

Indeed, μ is defined to be the largest number for which $\gamma \|h\|_s^2 \ge \mu \|h\|_{v \bullet p^{-1}}^2$ holds for all *h*. Hence, *f* is μ -strongly convex with respect to the norm $\|\cdot\|_{v \bullet p^{-1}}$.

Let x_{*} be a minimizer of f, i.e., an optimal solution of (1).
 Minimizing both sides of (8) in h, we get

$$f(x_{*}) - f(x) \stackrel{(8)}{\geq} \min_{h \in \mathbb{R}^{n}} \langle \nabla f(x), h \rangle + \frac{\mu}{2} \|h\|_{v \bullet p^{-1}}^{2} \\ = -\frac{1}{2\mu} \|\nabla f(x)\|_{p \bullet v^{-1}}^{2}.$$
(9)

イロト イポト イヨト イヨト 二日

Proof of Theorem 3 - Part II

▶ Let $h_k \stackrel{\text{def}}{=} -v^{-1} \bullet \nabla f(x_k)$. Then $x_{k+1} = x_k + (h_k)_{[\hat{S}]}$, and utilizing Assumption 2, we get

$$\begin{aligned} \mathsf{E}[f(x_{k+1}) \mid x_k] &= \mathsf{E}\left[f(x_k + (h_k)_{[\hat{S}]}) \mid x_k\right] \\ &\stackrel{(5)}{\leq} f(x_k) + \langle \nabla f(x_k), h_k \rangle_p + \frac{1}{2} \|h_k\|_{p \bullet v}^2 \\ &= f(x_k) - \frac{1}{2} \|\nabla f(x_k)\|_{p \bullet v^{-1}}^2 \\ &\stackrel{(9)}{\leq} f(x_k) - \mu(f(x_k) - f(x_*)). \end{aligned}$$

Taking expectations in the last inequality,

$$\mathbf{E}[f(x_k) - f(x_*)] \le (1 - \mu)^k (f(x_0) - f(x_*)). \tag{10}$$

• Using Markov inequality, (10) and the definition of K, we finally get

$$\mathbf{P}(f(x_{\mathcal{K}}) - f(x_{*}) \ge \epsilon) \leq \mathbf{E}[f(x_{\mathcal{K}}) - f(x_{*})]/\epsilon$$

$$\stackrel{(10)}{\leq} (1 - \mu)^{\mathcal{K}}(f(x_{0}) - f(x_{*}))/\epsilon \leq \rho.$$

15 / 116

Proof of Theorem 3 - Part III

Finally, let us now establish the lower bound on Λ . Letting $\Delta \stackrel{\text{def}}{=} \{ p' \in \mathbb{R}^n : p' \ge 0, \sum_i p'_i = \mathbf{E}[|\hat{S}|] \}, \text{ we have}$ $\Lambda \stackrel{(7)}{=} \max_i \frac{v_i}{p_i s_i} \stackrel{(3)}{\ge} \min_{p' \in \Delta} \max_i \frac{v_i}{p'_i s_i} = \frac{1}{|\mathbf{E}[|\hat{S}|]} \sum_{i=1}^n \frac{v_i}{s_i},$

where the last equality follows since optimal p'_i is proportional to v_i/s_i .



Lecture 2 BLOCKS



The idea

We now assume the decision vector x has N coordinates

$$x \in \mathbb{R}^N$$

which we partition into *n* "blocks".

Idea: We let the algorithm operate on "block level" instead \Rightarrow **block** coordinate descent. That is, at iteration *k*,

イロト イポト イヨト イヨト 二日

- ▶ a random subset S_k of blocks $[n] = \{1, 2, ..., n\}$ is chosen
- and updated.

What do we gain by introducing blocks?

- Flexibility: We can partition the coordinates any way we like for any reason we might have.
 - Sometimes block structure is implied by the problem at hand. In L1 optimization, one often chooses N_i = 1 for all *i*. In group LASSO problems, groups correspond to blocks.
- Generality: By allowing for general block structure, we simultaneously analyze several classes of algorithms:
 - ► coordinate descent (if we choose N_i = 1 for all i)
 - **block coordinate descent** (if we choose $N_i > 1$ and n > 1)
 - ▶ gradient descent (if we choose *n* = 1)
 - fast $(O(1/k^2))$ versions of the above...

• Efficiency: It is sometimes more efficient to have blocks because:

- this leads to a more "chunky" workload for each processor if we think that each processor handles one block
- one can design block-norms based on data, which leads to better approximation and hence faster convergence
- one can try to optimize the partitioning of coordinates to blocks (say, by trying to optimize complexity bounds, which depend on block structure)



Block Decomposition of \mathbb{R}^N

- Partition. Let H₁,..., H_n be a partition of the set of coordinates/variables {1, 2, ..., N} into n nonempty subsets. Let N_i = |H_i|.
- ► Projection/lifting matrices. Let U_i ∈ ℝ^{N×N_i} be the column submatrix of the N × N identity matrix corresponding to coordinates in H_i.
- ▶ **Projection of** \mathbb{R}^N **to** \mathbb{R}^{N_i} For $x \in \mathbb{R}^N$, define

$$x^{(i)} \stackrel{\text{def}}{=} U_i^T x \in \mathbb{R}^{N_i}, \quad i = 1, 2, \dots, n.$$

Notice that $x^{(i)}$ is the block of coordinates of x belonging to H_i .

▶ Lifting \mathbb{R}^{N_i} to \mathbb{R}^N . Given $x^{(i)} \in \mathbb{R}^{N_i}$, notice that the vector $s = U_i x^{(i)} \in \mathbb{R}^N$ has all blocks equal to 0 except for block *i*, which is equal to $x^{(i)}$. That is,

$$s^{(j)} = \begin{cases} x^{(j)} & j = i \\ 0 & \text{otherwise.} \end{cases}$$



Examples - Part I

- Example 4
- 1. Single block.

$$n = 1;$$
 $H_1 = \{1, 2, \dots, N\};$ $U_1 = I$

2. Blocks of size 1. This is the setting already introduced in NSync:

$$N = n; \qquad H_i = \{i\}; \qquad U_i = e_i$$

3. Two blocks of different sizes. Let N = 5 (5 coordinates), n = 2 (2 blocks) and let the partitioning be given by

$$H_1 = \{1,3\}, \quad H_2 = \{2,4,5\}.$$

Then

$$U_{1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad U_{2} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



Examples - Part II

For $x \in \mathbb{R}^N = \mathbb{R}^5$ we have

$$x^{(1)} = U_1^T x = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} \in \mathbb{R}^{N_1} = \mathbb{R}^2$$
$$x^{(2)} = U_2^T x = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_4 \\ x_5 \end{pmatrix} \in \mathbb{R}^{N_2} = \mathbb{R}^3$$

On the other hand, for any $x \in \mathbb{R}^5$:

$$U_{1}x^{(1)} = U_{1}(U_{1}^{T}x) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{1} \\ x_{3} \\ x_{3} \end{pmatrix} = \begin{pmatrix} x_{1} \\ 0 \\ x_{3} \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{5}$$



22/116

★ロト ★御 と ★ 注 と ★ 注 と … 注

Examples - Part III

and

$$U_2 x^{(2)} = U_2 (U_2^T x) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ x_2 \\ 0 \\ x_4 \\ x_5 \end{pmatrix} \in \mathbb{R}^5$$

So, we have the unique decomposition:

$$x = U_1 x^{(1)} + U_2 x^{(2)}$$

The next simple result will formalize this.



Block Decomposition: Formal Statement Proposition 1 (Block Decomposition) Any vector $x \in \mathbb{R}^N$ can be written uniquely as

$$x = \sum_{i=1}^{n} U_i x^{(i)},$$
 (11)

where $x^{(i)} \in \mathbb{R}^{N_i}$. Moreover,

$$x^{(i)} = U_i^T x. (12)$$

Proof. Fix any $x \in \mathbb{R}^N$. Noting that $\sum_i U_i U_i^T$ is the $N \times N$ identity matrix, we have $x = \sum_i U_i U_i^T x$, where $U_i^T x \in \mathbb{R}^{N_i}$. Let us now show uniqueness. Assume that $x = \sum_i U_i x_1^{(i)} = \sum_i U_i x_2^{(i)}$, where $x_1^{(i)}, x_2^{(i)} \in \mathbb{R}^{N_i}$. Since $U_j^T U_i = \begin{cases} N_j \times N_j & \text{identity matrix,} & \text{if } i = j, \\ N_j \times N_i & \text{zero matrix,} & \text{otherwise,} \end{cases}$ (13) we get $0 = U_j^T (x - x) = U_j^T \sum_i U_i (x_1^{(i)} - x_2^{(i)}) = x_1^{(j)} - x_2^{(j)}$, for all j.

24 / 116

Projection onto (a subspace spanned by) a set of blocks

For $h \in \mathbb{R}^N$ and $\emptyset \neq S \subseteq [n] \stackrel{\text{def}}{=} \{1, 2, \dots, n\}$, we write

$$h_{[S]} = \sum_{i \in S} U_i h^{(i)}.$$
 (14)

In words, $h_{[S]}$ is a vector in \mathbb{R}^N obtained from $h \in \mathbb{R}^N$ by zeroing out the blocks that do not belong to S. Hence:

$$(h_{[S]})^{(i)} = \begin{cases} h^{(i)}, & i \in S, \\ 0, & i \notin S. \end{cases}$$



Norms in \mathbb{R}^{N_i} and \mathbb{R}^N

With each block $i \in [n]$ we associate a positive definite matrix $B_i \in \mathbb{R}^{N_i \times N_i}$ and a scalar $v_i > 0$, and equip \mathbb{R}^{N_i} and \mathbb{R}^N with the norms

$$\|x^{(i)}\|_{(i)} \stackrel{\text{def}}{=} \langle B_i x^{(i)}, x^{(i)} \rangle^{1/2}, \quad \|x\|_v \stackrel{\text{def}}{=} \left(\sum_{i=1}^n v_i \|x^{(i)}\|_{(i)}^2\right)^{1/2}.$$
(15)

The corresponding conjugate norms, defined by

$$\|s\|^* = \max\{\langle s, x \rangle : \|x\| \le 1\}$$

are given by

$$\|x^{(i)}\|_{(i)}^* \stackrel{\text{def}}{=} \langle B_i^{-1} x^{(i)}, x^{(i)} \rangle^{1/2}, \quad \|x\|_v^* = \left(\sum_{i=1}^n \frac{1}{v_i} \left(\|x^{(i)}\|_{(i)}^*\right)^2\right)^{1/2}.$$
(16)



Norms: Examples

Example 5

Consider the following extreme special cases:

1. Single block. Let n = 1, v = 1 and B be a positive definite matrix. Then

$$\|x\|_{(1)} = \|x\|_{\nu} = \langle Bx, x \rangle^{1/2}, \quad x \in \mathbb{R}^{N}.$$

For instance, if $f(x) = \frac{1}{2} ||Ax - b||^2$ we may choose:

- $B = A^T A$ (assuming $A^T A$ is positive definite)
- $B = \text{Diag}(A^T A)$ (assuming no column in A is zero, $A^T A$ is positive definite)
- 2. Blocks of size one. Let $N_i = 1$ for all *i* and set $B_i = 1$. Then

$$\|t\|_{(i)} = \|t\|_{(i)}^* = |t|, \qquad t \in \mathbb{R}$$

and

$$\|x\|_{v} = \left(\sum_{i=1}^{n} v_{i}(x^{(i)})^{2}\right)^{1/2}, \qquad x \in \mathbb{R}^{N}.$$



イロト イヨト イヨト イヨト 二日

Exercises

Exercise 1 Show that $\|\cdot\|_{\nu}^{*}$, as defined above, is indeed the conjugate norm of $\|\cdot\|_{\nu}$.

Exercise 2

Generalize NSync to the block setting and provide a complexity analysis.



Lecture 3 SAMPLINGS



Samplings: Definition

Definition 6 (Sampling)

Sampling is a *random set-valued mapping* \hat{S} with values in $2^{[n]}$, the collection of subsets of $[n] = \{1, 2, ..., n\}$.

• A sampling \hat{S} is uniquely characterized by the probability mass function

$$\mathbf{P}(S) \stackrel{\text{def}}{=} \mathbf{P}(\hat{S} = S), \quad S \subseteq [n]; \tag{17}$$

that is, by assigning probabilities to all subsets of [n].

Let

$$p_i \stackrel{\text{def}}{=} \mathbf{P}(i \in \hat{S}). \tag{18}$$

Let

$$p_{ij} \stackrel{\text{def}}{=} \mathbf{P}(i \in \hat{S}, j \in \hat{S}) = \sum_{S:\{i,j\} \subset S} \mathbf{P}(S).$$
(19)

Sampling Zoo - Part I

Why consider different samplings?

- 1. **Basic Considerations.** It is important that each block has a positive probability of being chosen, otherwise an algorithm will not be able to update some blocks and hence will not converge to optimum. For technical/sanity reasons, we define:
 - ▶ Proper sampling. $p_i = \mathbf{P}(i \in \hat{S}) > 0$ for all blocks $i \in [n]$
 - Nil sampling: $P(\hat{S} = \emptyset) = 1$
 - Vacuous sampling: $P(\hat{S} = \emptyset) > 0$
- 2. Parallelism. Choice of sampling affects the level of parallelism:
 - ► E[|Ŝ|] is the average number of updates performed in parallel in one iteration; and is hence closely related to the number of iterations.
 - serial sampling: picks one block:

$$\mathsf{P}(|\hat{S}|=1)=1$$

We call this sampling serial although nothing prevents us from computing the actual update to the block, and/or to apply he update in parallel.



Sampling Zoo - Part II

fully parallel sampling: always picks all blocks:

$$P(\hat{S} = \{1, 2, ..., n\}) = 1$$

- 3. **Processor reliability.** Sampling may be induced/informed by the computing environment:
 - ▶ Reliable/dedicated processors. If one has reliable processors, it is sensible to choose sampling \hat{S} such that $P(|\hat{S}| = \tau)$ 1 for some τ related to the number of processors.
 - ► Unreliable processors. If processors given a computing task are busy or unreliable, they return answer later or not at all - it is then sensible to ignore such updates and move on. This then means that *Ŝ* varies from iteration to iteration.
- 4. **Distributed computing.** In a distributed computing environment it is sensible:
 - to allow each node as much autonomy as possible so as to minimize communication cost,
 - to make sure all nodes are busy at all times



イロト イポト イヨト イヨト 二日

Sampling Zoo - Part III

This suggests a strategy where the set of blocks is partitioned, with each node owning a partition, and independently picking a "chunky" subset of blocks at each iteration it will update, ideally from local information.

- 5. **Uniformity.** It may or not may make sense to update some blocks more often than others:
 - uniform samplings:

$$\mathbf{P}(i \in \hat{S}) = \mathbf{P}(j \in \hat{S})$$
 for all $i, j \in [n]$

doubly uniform (DU): These are samplings characterized by:

$$|S'| = |S''| \ \Rightarrow \ \mathsf{P}(\hat{S} = S') = \mathsf{P}(\hat{S} = S'') \quad \text{for all} \quad S', S'' \subseteq [n]$$

• τ -nice: DU sampling with the additional property that

$$\mathsf{P}(|\hat{S}|= au)=1$$

- distributed \(\tau\)-nice: will define later
- independent sampling: union of independent uniform serial samplings
- nonuniform samplings



33 / 116

イロト イポト イヨト イヨト 二日

Sampling Zoo - Part IV

- 6. **Complexity of generating a sampling.** Some samplings are computationally more efficient to generate than others: the potential benefits of a sampling may be completely ruined by the difficulty to generate sets according to the sampling's distribution.
 - a τ-nice sampling can be well approximated by an independent sampling, which is easy to generate...
 - ▶ a general sampling, as considered in NSync, will be hard to generate

Basic Identity

Theorem 7 (Sum over a random index set) Let $\emptyset \neq J, J_1, J_2 \subset [n]$ and \hat{S} be any sampling. If θ_i , $i \in [n]$, and θ_{ij} , for $(i,j) \in [n] \times [n]$ are real constants, then¹

$$\mathsf{E}\left[\sum_{i\in J\cap \hat{\mathsf{S}}}\theta_i\right] = \sum_{i\in J}p_i\theta_i,$$

$$\mathbf{E}\left[\sum_{i\in J\cap\hat{S}}\theta_{i}\mid|J\cap\hat{S}|=k\right]=\sum_{i\in J}\mathbf{P}(i\in\hat{S}\mid|J\cap\hat{S}|=k)\theta_{i},\qquad(20)$$
$$\mathbf{E}\left[\sum_{i\in J_{1}\cap\hat{S}}\sum_{j\in J_{2}\cap\hat{S}}\theta_{ij}\right]=\sum_{i\in J_{1}}\sum_{j\in J_{2}}p_{ij}\theta_{ij}.\qquad(21)$$



Proof of Theorem 7

We prove the first statement, proof of the remaining statements is essentially identical:

$$\mathbf{E}\left[\sum_{i\in J\cap\hat{S}}\theta_i\right] \stackrel{(17)}{=} \sum_{S\subset[n]} \left(\sum_{i\in J\cap S}\theta_i\right) \mathbf{P}(\hat{S}=S)$$
$$= \sum_{i\in J}\sum_{S:i\in S}\theta_i \mathbf{P}(\hat{S}=S)$$
$$= \sum_{i\in J}\theta_i\sum_{S:i\in S}\mathbf{P}(\hat{S}=S)$$
$$= \sum_{i\in J}p_i\theta_i.$$


Consequences of Theorem 7

Corollary 8 ([5])

Let $\emptyset \neq J \subset [n]$ and \hat{S} be an arbitrary sampling. Further, let $a, h \in \mathbb{R}^N$, $w \in \mathbb{R}^n_+$ and let g be a block separable function, i.e., $g(x) = \sum_i g_i(x^{(i)})$. Then

$$\mathbf{E}\left[|J \cap \hat{S}|\right] = \sum_{i \in J} p_i, \qquad (22)$$

$$\mathsf{E}\left[|J \cap \hat{S}|^2\right] = \sum_{i \in J} \sum_{j \in J} p_{ij}, \qquad (23)$$

$$\mathsf{E}\left[\langle a, h_{[\hat{S}]} \rangle_{w}\right] = \langle a, h \rangle_{\rho \bullet w}, \qquad (24)$$

$$\mathbf{E}\left[\|h_{[\hat{S}]}\|_{w}^{2}\right] = \|h\|_{p \bullet w}^{2}, \qquad (25)$$

$$\mathbf{E}\left[g(x+h_{[\hat{S}]})\right] = \sum_{i=1}^{n} \left[p_i g_i(x^{(i)}+h^{(i)})+(1-p_i)g_i(x^{(i)})\right].$$
(26)

Moreover, the matrix $P \stackrel{\text{def}}{=} (p_{ij})$ is positive semidefinite.



37 / 110

Proof of Corollary 8

All 5 identities follow by applying Lemma 7 and observing that:

$$|J \cap \hat{S}| = \sum_{i \in J \cap \hat{S}} 1 |J \cap \hat{S}|^2 = (\sum_{i \in J \cap \hat{S}} 1)^2 = \sum_{i \in J \cap \hat{S}} \sum_{j \in J \cap \hat{S}} 1 \langle a, h_{[\hat{S}]} \rangle_w = \sum_{i \in \hat{S}} w_i \langle a^{(i)}, h^{(i)} \rangle \|h_{[\hat{S}]}\|_w^2 = \sum_{i \in \hat{S}} w_i \|h^{(i)}\|_{(i)}^2 \text{ and} g(x + h_{[\hat{S}]}) = \sum_{i \in \hat{S}} g_i(x^{(i)} + h^{(i)}) + \sum_{i \notin \hat{S}} g_i(x^{(i)}) = \sum_{i \in \hat{S}} g_i(x^{(i)} + h^{(i)}) + \sum_{i=1}^n g_i(x^{(i)}) - \sum_{i \in \hat{S}} g_i(x^{(i)}),$$

Finally, for any $\theta = (\theta_1, \dots, \theta_n)^T \in \mathbb{R}^n$,

$$\theta^{\mathsf{T}} P \theta = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \theta_i \theta_j \stackrel{(21)}{=} \mathsf{E} \left[\left(\sum_{i \in \hat{S}} \theta_i \right)^2 \right] \ge 0.$$

Remark: The above results hold for arbitrary samplings. Let us specialize them, in order of decreasing generality, to uniform, doubly uniform and nice samplings.



38 / 116

Identities: uniform samplings

If \hat{S} is uniform, then from (22) using J = [n] we get

$$p_i = \frac{\mathbf{E}[|\hat{S}|]}{n}, \qquad i \in [n].$$
(27)

Plugging (27) into (22), (24), (25) and (26) yields

$$\mathbf{E}\left[|J \cap \hat{S}|\right] = \frac{|J|}{n} \mathbf{E}[|\hat{S}|],\tag{28}$$

$$\mathbf{E}\left[\langle a, h_{[\hat{S}]} \rangle_{w}\right] = \frac{\mathbf{E}\left[|\hat{S}|\right]}{n} \langle a, h \rangle_{w},$$
(29)
$$\mathbf{E}\left[\|h_{[\hat{S}]}\|_{w}^{2}\right] = \frac{\mathbf{E}\left[|\hat{S}|\right]}{n} \|h\|_{w}^{2},$$
(30)

$$\mathbf{E}\left[g(x+h_{[\hat{S}]})\right] = \frac{\mathbf{E}[|\hat{S}|]}{n}g(x+h) + \left(1 - \frac{\mathbf{E}[|\hat{S}|]}{n}\right)g(x).$$
(31)

<ロ ▶ < 部 ▶ < 差 ▶ < 差 ▶ 差 の Q (~ 39/116

Identities: doubly uniform samplings

Consider the case n > 1; the case n = 1 is trivial. For doubly uniform \hat{S} , p_{ij} is constant for $i \neq j$:

$$p_{ij} = \frac{\mathsf{E}[|\hat{S}|^2 - |\hat{S}|]}{n(n-1)}.$$
(32)

Indeed, this follows from

$$p_{ij} = \sum_{k=1}^{n} \mathbf{P}(\{i,j\} \subseteq \hat{S} \mid |\hat{S}| = k) \mathbf{P}(|\hat{S}| = k) = \sum_{k=1}^{n} \frac{k(k-1)}{n(n-1)} \mathbf{P}(|\hat{S}| = k).$$

Substituting (32) and (27) into (23) then gives

$$\mathbf{E}[|J \cap \hat{S}|^2] = (|J|^2 - |J|) \frac{\mathbf{E}[|\hat{S}|^2 - |\hat{S}|]}{n \max\{1, n-1\}} + |J| \frac{|\hat{S}|}{n}.$$
 (33)



イロト イポト イヨト イヨト 二日

Identities: τ -nice sampling

Finally, if \hat{S} is τ -nice (and $\tau \neq 0$), then $\mathbf{E}[|\hat{S}|] = \tau$ and $\mathbf{E}[|\hat{S}|^2] = \tau^2$, which used in (33) gives

$$\mathbf{E}[|J \cap \hat{S}|^2] = \frac{|J|\tau}{n} \left(1 + \frac{(|J|-1)(\tau-1)}{\max\{1, n-1\}} \right).$$
(34)

Moreover, assume that $\mathbf{P}(|J \cap \hat{S}| = k) \neq 0$ (this happens precisely when $0 \leq k \leq |J|$ and $k \leq \tau \leq n - |J| + k$). Then for all $i \in J$,

$$\mathbf{P}(i \in \hat{S} \mid |J \cap \hat{S}| = k) = \frac{\binom{|J|-1}{k-1}\binom{n-|J|}{\tau-k}}{\binom{|J|}{k}\binom{n-|J|}{\tau-k}} = \frac{k}{|J|}$$

Substituting this into (20) yields

$$\mathbf{E}\left[\sum_{i\in J\cap\hat{S}}\theta_i\mid |J\cap\hat{S}|=k\right]=\frac{k}{|J|}\sum_{i\in J}\theta_i.$$
(35)

Elementary Samplings, Intersection and Restriction

Definition 9 (Elementary samplings)

Elementary sampling associated with $J \subseteq [n]$ is sampling \hat{E}_J for which

$$\mathbf{P}(\hat{E}_J=J)=1.$$

Definition 10 (Intersection of samplings)

For two samplings \hat{S}_1 and \hat{S}_2 we define the intersection $\hat{S} \stackrel{\text{def}}{=} \hat{S}_1 \cap \hat{S}_2$ as the sampling for which:

$$\mathbf{P}(\hat{S}=S)=\mathbf{P}(\hat{S}_1\cap\hat{S}_2=S),\quad S\subseteq [n].$$

Definition 11 (Restriction of a sampling to a subset) Let \hat{S} be a sampling and $J \subseteq [n]$. By restriction of \hat{S} to J we mean the sampling

$$\hat{E}_J \cap \hat{S}.$$



Probability matrices associated with samplings - Part I

Definition 12 (Probability matrix)

With arbitrary sampling \hat{S} we associate an *n*-by-*n* matrix $P = P(\hat{S})$ with entries

$$[P(\hat{S})]_{ij} = \mathbf{P}(i \in \hat{S}, j \in \hat{S}).$$

Lemma 13 (Intersection of independent samplings; [14]) Let \hat{S}_1 , \hat{S}_2 be independent samplings. Then

$$P(\hat{S}_1 \cap \hat{S}_2) = P(\hat{S}_1) \bullet P(\hat{S}_2).$$

That is, the probability matrix of an intersection of independent samplings is the Hadamard product of their probability matrices.

Proof. $[P(\hat{S}_1 \cap \hat{S}_2)]_{ij} = \mathbf{P}(\{i, j\} \in \hat{S}_1 \cap \hat{S}_2) = \mathbf{P}(\{i, j\} \in \hat{S}_1)\mathbf{P}(\{i, j\} \in \hat{S}_2) = [P(\hat{S}_1)]_{ij}[P(\hat{S}_2)]_{ij}.$



Probability matrices associated with samplings - Part II Example 14 (Probability Matrix of an Elementary Sampling) Note that the probability matrix of the elementary sampling \hat{E}_J is the matrix

$$P(\hat{E}_J) \stackrel{\text{def}}{=} e_J e_J^T, \tag{36}$$

where e_J we denote the binary vector in \mathbb{R}^n with ones in places corresponding to set J. That is,

$$[P(\hat{E}_J)]_{ij} = egin{cases} 1 & i,j\in J, \ 0 & ext{otherwise}. \end{cases}$$

Hence, for arbitrary sampling \hat{S} , the probability matrix of $J \cap \hat{S}$ is the submatrix of $P(\hat{S})$ corresponding to the rows and columns indexed by J:

$$[P(J \cap \hat{S})]_{ij} = [P(\hat{E}_J) \bullet P(\hat{S})]_{ij} = \begin{cases} [P(\hat{S})]_{ij}, & i, j \in J, \\ 0, & \text{otherwise.} \end{cases}$$
(37)



(日) (四) (注) (注) (三)

Probability matrices associated with samplings - Part III Lemma 15 (Decomposition of a Probability Matrix; [14]) Let Ŝ be any sampling. Then

$$P(\hat{S}) = \sum_{S \subseteq [n]} P(\hat{S} = S) P(\hat{E}_S).$$
(38)

That is, the probability matrix of arbitrary sampling is a convex combination of elementary probability matrices.

Proof.

Fix any $i, j \in [n]$. Since $(P(\hat{E}_S))_{ij} = 1$ iff $\{i, j\} \subseteq S$, from definition we have

$$(P(\hat{S}))_{ij} = \sum_{S:\{i,j\}\subseteq S} \mathbf{P}(\hat{S} = S)$$

= $\sum_{S:\{i,j\}\subseteq S} \mathbf{P}(\hat{S} = S)(P(\hat{E}_S))_{ij}$
= $\left(\sum_{S:\{i,j\}\subseteq S} \mathbf{P}(\hat{S} = S)P(\hat{E}_S)\right)_{ij}$



・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・

Sampling Identity for a Quadratic

Lemma 16 Let G be any real $n \times n$ matrix and \hat{S} an arbitrary sampling. Then for any $h \in \mathbb{R}^n$ we have

$$\mathbf{E}\left[h_{[\hat{S}]}^{T}Gh_{[\hat{S}]}\right] = h^{T}\left(P(\hat{S})\bullet G\right)h,\tag{39}$$

where • denotes the Hadamard (elementwise) product of matrices. Proof.

$$\mathbf{E} \begin{bmatrix} h_{[\hat{S}]}^T G h_{[\hat{S}]} \end{bmatrix} \stackrel{(14)}{=} \mathbf{E} \begin{bmatrix} \sum_{i \in \hat{S}} \sum_{j \in \hat{S}} G_{ij} h^{(i)} h^{(j)} \end{bmatrix}$$
$$\stackrel{(21)}{=} \sum_{i=1}^n \sum_{j=1}^n p_{ij} G_{ij} h^{(i)} h^{(j)} = h^T \left(P(\hat{S}) \bullet G \right) h.$$

46 / 116

・ロト ・四ト ・ヨト ・ヨト

Distributed sampling

The following sampling is useful in the design of a **distributed coordinate descent method**.

Definition 17 (Distributed τ -nice sampling; [10, 13]) Let $\mathcal{P}_1, \ldots, \mathcal{P}_c$ be a partition of $\{1, 2, \ldots, n\}$ such that $|\mathcal{P}_l| = s$ for all *l*. That is, sc = n. Now let $\hat{S}_1, \ldots, \hat{S}_c$ be independent τ -nice samplings from $\mathcal{P}_1, \ldots, \mathcal{P}_c$, respectively. Then the sampling

$$\hat{S} \stackrel{\text{def}}{=} \cup_{l=1}^{c} \hat{S}_{l},\tag{40}$$

is called distributed τ -nice sampling.

Idea: Blocks in \mathcal{P}_{l} , and all associated data, will be handled/stored by computer/node l only. Node l picks blocks in \hat{S}_{l} , computes the updates fro local information, and applies the updates to locally stored $x^{(i)}$ for $i \in \mathcal{P}_{l}$.



Probability Matrix of Distributed τ -nice Sampling

Consider the distributed $\tau\text{-nice}$ sampling and define:

- $E = P(\hat{E}_{[n]})$: the $n \times n$ matrix of all ones
- *I* be the $n \times n$ identity matrix
- $B = \sum_{l=1}^{c} P(\hat{E}_{\mathcal{P}_l})$: the 0-1 matrix with $B_{ij} = 1$ iff i, j belong to the same partition

Lemma 18 ([10]; presented in a different form)

Consider the distributed τ -nice sampling \hat{S} . Its probability matrix can be written as

$$P(\hat{S}) = \frac{\tau}{s} \left[\alpha_1 I + \alpha_2 E + \alpha_3 (E - B) \right], \tag{41}$$

where

$$\alpha_1 = 1 - \frac{\tau - 1}{ss_1}, \qquad \alpha_2 = \frac{\tau - 1}{s_1}, \qquad \alpha_3 = \frac{\tau}{s} - \frac{\tau - 1}{s_1},$$

and $s_1 = \max\{1, s - 1\}.$



Proof of Lemma 18

Let
$$P = P(\hat{S})$$
. It is easy to see that
 $P_{ij} = \frac{\tau}{s} \stackrel{\text{def}}{=} \beta_3$ if $i = j$,
 $P_{ij} = \frac{\tau(\tau-1)}{ss_1} \stackrel{\text{def}}{=} \beta_2$ if $i \neq j$ and i, j belong to the same partition,
 $P_{ij} = \frac{\tau^2}{s^2} \stackrel{\text{def}}{=} \beta_3$ if $i \neq j$ belong to different partitions.

So, we can write

$$P = \beta_1 I + \beta_2 (B - I) + \beta_3 (E - B) = (\beta_1 - \beta_2) I + \beta_2 E + (\beta_3 - \beta_2) (E - B).$$

Exercises

Exercise 3

Find an expression for the probability matrix of

- ► the \(\tau\)-nice sampling,
- arbitrary doubly uniform sampling.

Exercise 4

Let \hat{S} be any sampling. Show that

▶ $\lambda_{max}(P) \leq \mathbf{E}[|\hat{S}|]$ and that the bound is tight, ▶ $P \succ pp^{T}$.



Lecture 4 FUNCTIONS



Introduction

- ► In this part we describe three models for *f*.
- These models can be thought of as function classes described by a list of properties.
- However, a single function may belong to more function classes.

In big data setting, some information is computationally difficult to extract from data.

Consider $f(x) = \frac{1}{2} ||Ax - b||^2$.

- It is difficult to compute the largest eigenvalue of A^TA if A is large (this is the Lipschitz constant of ∇f with respect to the standard Euclidean norm)
- It is easier to compute the squared norm of each column (these correspond to coordinate-wise Lipschitz constants).

Important point: The models differ in the amount of information they reveal about f.



Model: Quadratic

Model 1 ([10, 13])

We assume that

1. Structure and Smoothness: $f : \mathbb{R}^N \to \mathbb{R}$ is differentiable and for all $x, h \in \mathbb{R}^N$ satisfies

$$f(x+h) \leq f(x) + (\nabla f(x))^T h + \frac{1}{2} h^T A^T A h, \qquad (42)$$

where $A \in \mathbb{R}^{m \times N}$.

2. **Sparsity:** Row *j* of *A* depends on blocks $i \in C_j$ only. Formally,

$$C_j \stackrel{def}{=} \{i : A_{ji} \neq 0\},\$$

where $A_{ji} \stackrel{\text{def}}{=} e_j^T A U_i \in \mathbb{R}^{1 \times N_i}$. Let $\omega_j \stackrel{\text{def}}{=} |C_j|$. 3. Convexity: f is convex.

Remark: Information about f is contained in the matrix A.



Examples

Example 19

In machine learning (ML), functions f of the following form are common:

$$f(x) = \sum_{j=1}^{m} f_j(x) = \sum_{j=1}^{m} \ell(x; a_j, y^j),$$

where N is the number of features, m number of examples, $a_j \in \mathbb{R}^N$ corresponds to *j*th example and y^j is a label associated with *j*th example.

Here are some convex loss functions ℓ often used in ML for which the total loss f satisfies (42):

Loss function ℓ	$f_j(x)$	(42) satisfied for A given by
square loss (SL)	$\frac{1}{2}(y^j - \boldsymbol{a}_j^T \boldsymbol{x})^2$	$A_{j:} = a_j^T$
logistic loss (LL)	$\log(1 + \exp(-y^j a_j^T x))$	$A_{j:}=rac{1}{2}oldsymbol{a}_{j}^{T}$
square hinge loss (HL)	$rac{1}{2}\max\{0,1-y^ja_j^Tx\}^2$	$\textit{A}_{j:} = \textit{a}_{j}^{T}$

Interpretation of ω_j (point 2 in Model 1) : # features in example j



54 / 116

Definition 20 (Block Gradients)

The *i*th **block gradient** of $f : \mathbb{R}^N \to \mathbb{R}$ at x is defined to be the *i*th block of the gradient of f at x:

$$\nabla_i f(x) \stackrel{\text{def}}{=} (\nabla f(x))^{(i)} = U_i^T \nabla f(x) \in \mathbb{R}^{N_i}.$$
(43)

In other words, $\nabla_i f(x)$ is the vector of partial derivatives with respect to coordinates belonging to block *i*.



Model: Classical

Model 2 ([2, 5, 9]) We assume that

1. Structure: Function $f : \mathbb{R}^N \to \mathbb{R}$ is of the form

$$f(x) = \sum_{j=1}^m f_j(x)$$

- 2. Sparsity: f_j depends on x via blocks $i \in C_j$ only.
- 3. Convexity: Functions $\{f_j\}$ are convex.
- 4. Smoothness: Function f has block-Lipschitz gradient with constants $L_1, \ldots, L_n > 0$. That is, for all $i = 1, 2, \ldots, n$,

$$\|\nabla_{i}f(x+U_{i}t)-\nabla_{i}f(x)\|_{(i)}^{*} \leq L_{i}\|t\|_{(i)}, \quad x \in \mathbb{R}^{N}, \ t \in \mathbb{R}^{N_{i}}.$$
 (44)

Remark: Information about f is contained in the constants L_1, \ldots, L_n .



Examples

Example 21 (Least squares)

Consider the quadratic function $f(x) = \frac{1}{2} ||Ax - b||^2$.

(i) Consider the block setup with $N_i = 1$ (all blocks are of size 1) and $B_i = 1$ for all $i \in [n]$ (standard Eucl. norms for each block: $||t||_{(i)} = |t|$). Then $U_i = e_i$ and

$$\begin{aligned} \|\nabla_i f(x+U_it) - \nabla_i f(x)\|_{(i)}^* &= |e_i^T A^T (A(x+te_i) - b) - e_i^T A^T (Ax - b)| \\ &= |e_i^T A^T A e_i||t| = \|A_{ij}\|^2 |t|, \end{aligned}$$

whence $L_i = ||A_{:i}||^2$.

(ii) Choose nontrivial block sizes $(N_i > 1)$ and define data-driven block norms with $B_i = A_i^T A_i$, where $A_i = AU_i$, assuming that $B_i \succ 0$. Then

$$\begin{aligned} \|\nabla_{i}f(x+U_{i}t)-\nabla_{i}f(x)\|_{(i)}^{*} &= \|U_{i}^{T}A^{T}(A(x+U_{i}t)-b)-U_{i}^{T}A^{T}(Ax-b)\|_{(i)}^{*} \\ &= \|U_{i}^{T}A^{T}AU_{i}t\|_{(i)}^{*} \\ &\stackrel{(16)}{=} \langle (A_{i}A_{i}^{T})^{-1}U_{i}^{T}A^{T}AU_{i}t, U_{i}^{T}A^{T}AU_{i}t\rangle^{1/2} \\ &= \langle B_{i}t,t\rangle^{1/2} \stackrel{(15)}{=} \|t\|_{(i)}, \end{aligned}$$

whence $L_i = 1$.



イロト イポト イヨト イヨト 二日

Model: Newest

Model 3 ([12]) We assume that

1. **Structure:** $f : \mathbb{R}^N \to \mathbb{R}$ is of the form

$$f(x) = \sum_{j=1}^{m} f_j(x).$$
 (45)

- Sparsity: f_j depends on x via blocks i ∈ C_j only. Let ω_j = |C_j|. (Note that i ∉ C_j ⇒ L_{ji} = 0)
- 3. Convexity: Functions $\{f_j\}$ are convex.
- 4. Smoothness: Functions $\{f_j\}$ have block-Lipschitz gradient with constants $L_{ji} \ge 0$. That is, for all j = 1, 2, ..., m and i = 1, 2, ..., n,

$$\|\nabla_i f_j(x+U_i t) - \nabla_i f_j(x)\|_{(i)}^* \le L_{ji} \|t\|_{(i)}, \quad x \in \mathbb{R}^N, \ t \in \mathbb{R}^{N_i}.$$
(46)

Remark: Information about f is contained in the constants $\{L_{ji}\}$



Computation of L_{ji}

We now give a formula for the constants L_{ji} in the case when f_j arises as a composition of a scalar function ϕ_j whose derivative has a known Lipschitz constant (this is often easy to compute), and a linear functional.

Proposition 2 ([12])

Let $f_j(x) = \phi_j(e_j^T A x)$, where $\phi_j : \mathbb{R} \to \mathbb{R}$ is a function with L_{ϕ_j} -Lipschitz derivative:

$$|\phi_j(\mathbf{s}) - \phi_j(\mathbf{s}')| \le L_{\phi_j}|\mathbf{s} - \mathbf{s}'|, \qquad \mathbf{s}, \mathbf{s}' \in \mathbb{R}.$$
 (47)

Then f_j has a block Lipshitz gradient (i.e., satisfies (46)) with constants

$$L_{ji} = L_{\phi_j} \left(\|A_{ji}^{\mathsf{T}}\|_{(i)}^* \right)^2, \qquad i = 1, 2, \dots, n,$$
(48)

where

$$A_{ji} = e_j^T A U_i \tag{49}$$

(i.e., A_{ji} is the *i*th block of *j*-th row of A).



Proof of Proposition 2

For any $x \in \mathbb{R}^N$, $t \in \mathbb{R}^{N_i}$ and i we have

$$\begin{split} \|\nabla_{i}f_{j}(x+U_{i}t)-\nabla_{i}f_{j}(x)\|_{(i)}^{*} \\ \stackrel{(43)}{=} & \|U_{i}^{T}(e_{j}^{T}A)^{T}\phi_{j}'(e_{j}^{T}A(x+U_{i}t))-U_{i}^{T}(e_{j}^{T}A)^{T}\phi_{j}'(e_{j}^{T}Ax)\|_{(i)}^{*} \\ & = & \|A_{ji}^{T}\phi_{j}'(e_{j}^{T}A(x+U_{i}t))-A_{ji}^{T}\phi_{j}'(e_{j}^{T}Ax)\|_{(i)}^{*} \\ & \leq & \|A_{ji}^{T}\|_{(i)}^{*}|\phi_{j}'(e_{j}^{T}A(x+U_{i}t))-\phi_{j}'(e_{j}^{T}Ax)| \\ \stackrel{(47)}{\leq} & \|A_{ji}^{T}\|_{(i)}^{*}L_{\phi_{j}}|A_{ji}t| \leq & \|A_{ji}^{T}\|_{(i)}^{*}L_{\phi_{j}}\|A_{ji}^{T}\|_{(i)}^{*}\|t\|_{(i)}^{*}, \end{split}$$

where the last step follows by applying the Cauchy-Schwartz inequality.

Examples

Example 22 (Least squares)

Consider the quadratic function

$$f(x) = \frac{1}{2} ||Ax - b||^2 = \frac{1}{2} \sum_{j=1}^{m} (e_j^T A x - b_j)^2.$$

Then $f_j(x) = \phi_j(e_j^T A x)$, where $\phi_j(s) = \frac{1}{2}(s - b_j)^2$ and $L_{\phi_j} = 1$.

(i) Consider the block setup with $N_i = 1$ (all blocks are of size 1) and $B_i = 1$ for all $i \in [n]$ (standard Euclidean norms for each block). Then by Proposition 2,

$$L_{ji} \stackrel{(48)}{=} L_{\phi_j} (\|A_{ji}^T\|_{(i)}^*)^2 = A_{ji}^2.$$

(ii) Choose nontrivial block sizes $(N_i > 1)$ and define data-driven block norms with $B_i = A_i^T A_i$, where $A_i = AU_i$, assuming that the matrices $A_i^T A_i$ are positive definite. Then by Proposition 2,

$$L_{ji} \stackrel{(48)}{=} L_{\phi_j} (\|A_{ji}^T\|_{(i)}^*)^2 \stackrel{(16)}{=} \langle (A_i^T A_i)^{-1} A_{ji}^T, A_{ji}^T \rangle \stackrel{(49)}{=} e_j^T A_i (A_i^T A_i)^{-1} A_i^T e_j.$$



イロト イヨト イヨト イヨト 二日

Lecture 5 Expected Separable Overapproximation



Introduction

In this part we shall look at the three models of f (Lecture 3) and various types of samplings \hat{S} (Lecture 4) and compute parameters $v = (v_1, \ldots, v_n)$ such $(f, \hat{S}) \sim ESO(v).$

These parameters are important since:

- They are stepsize parameters needed in the algorithm (in NSync, but also in other randomized block coordinate descent methods).
- ► Their size as a function of \(\tau = E[|S|]\) describes achievable parallelization speedup.
- By computing v we get one step closer to ultimate goal of designing sampling Ŝ optimizing the complexity bound.



 $\mathsf{ESO}(f \sim \mathsf{Model} \ 1, \ \hat{S} \sim \mathsf{arbitrary})$

Theorem 23 ([14])

Let f satisfy assumptions in Model 1, assume all blocks are of size 1 $(N_i = 1)$ and \hat{S} be any sampling. Then for all $x, h \in \mathbb{R}^N$,

$$\mathbf{E}\left[f(x+h_{[\hat{S}]})\right] \le f(x) + \langle \nabla f(x), h \rangle_{p} + \frac{1}{2} \|h\|_{p \bullet v}^{2}, \tag{50}$$

where v is any vector such that

$$P \bullet A^T A \preceq \text{Diag}(p \bullet v), \tag{51}$$

where $P = P(\hat{S})$ is the probability matrix associated with \hat{S} .

Remark: The Hadamard product of two PSD matrices is PSD (P is PSD by Corollary 8).



Proof of Theorem 23

We have

$$\mathbf{E}\left[f(x+h_{[\hat{S}]})\right] \stackrel{(42)}{\leq} \mathbf{E}\left[f(x)+\langle\nabla f(x),h_{[\hat{S}]}\rangle+\frac{1}{2}\langle A^{T}Ah_{[\hat{S}]},h_{[\hat{S}]}\rangle\right] \\ \stackrel{(24)}{=} f(x)+\langle\nabla f(x),h\rangle_{p}+\frac{1}{2}\mathbf{E}\left[h_{[\hat{S}]}^{T}A^{T}Ah_{[\hat{S}]}\right] \\ \stackrel{(*)}{=} f(x)+\langle\nabla f(x),h\rangle_{p}+\frac{1}{2}h^{T}\left(P\bullet A^{T}A\right)h \\ \leq f(x)+\langle\nabla f(x),h\rangle_{p}+\frac{1}{2}\underbrace{h^{T}\operatorname{Diag}(p\bullet v)h}_{=\|\|h\|_{p\bullet v}^{2}}$$

where (*) comes from Lemma 16.



Ways of satisfying (51)

Let us fix a sampling \hat{S} (and hence P) and data A. We can find v for which $P \bullet A^T A \preceq \text{Diag}(p \bullet v)$ in several ways:

1.
$$v_i = \lambda_1 \|A_{:i}\|^2$$
 and

$$\lambda_1 = \max_{\theta \in \mathbb{R}^n} \{ \theta^{\mathsf{T}} (P \bullet A^{\mathsf{T}} A) \theta \ : \ \theta^{\mathsf{T}} \operatorname{Diag} (P \bullet A^{\mathsf{T}} A) \theta \leq 1 \}.$$

2.
$$v_i = \frac{\lambda_{max}(P \bullet A^T A)}{p_i}$$
.
3. $v_i = \lambda_{max}(A^T A) \frac{(\max_i p_i)}{p_i}$ (using Lemma 24 with $X = P$)
4. $v_i = \frac{\lambda_{max}(P)}{p_i} \max_i ||A_{:i}||^2$ (using Lemma 24 with $X = A^T A$)

Lemma 24

For any two PSD matrices X, Y with nonnegative elements,

$$\lambda_{max}(X \bullet Y) \leq \lambda_{max}(X) \max_{i} Y_{jj}$$



66 / 116

Eigenvalues of Probability Matrices

Definition 25 (Eigenvalues) For arbitrary sampling \hat{S} we define

$$\lambda(\hat{S}) \stackrel{\text{def}}{=} \max_{\theta \in \mathbb{R}^n} \{\theta^T P(\hat{S})\theta : \theta^T \operatorname{Diag}(P(\hat{S}))\theta \le 1\}.$$
(52)

and

$$\lambda'(\hat{S}) \stackrel{\text{def}}{=} \max_{\theta \in \mathbb{R}^n} \{ \theta^T P(\hat{S}) \theta : \theta^T \theta \le 1 \}.$$
(53)

Example 26 (Elementary Sampling)

Fix $S \subseteq [n]$ and consider the elementary sampling \hat{E}_S . Note that

$$\lambda(\hat{E}_{\mathcal{S}}) = \lambda_{max}(\mathcal{P}(\hat{E}_{\mathcal{S}})) = \lambda_{max}(e_{\mathcal{S}}e_{\mathcal{S}}^{\mathsf{T}}) = \|e_{\mathcal{S}}\|^2 = |\mathcal{S}|.$$
(54)

Since $J \cap \hat{E}_S = \hat{E}_{J \cap S}$, we get

$$\lambda(J \cap \hat{E}_S) = \lambda(\hat{E}_{J \cap S}) \stackrel{(54)}{=} |J \cap S|.$$
(55)

57 / 116

Insightful and Easily Computable Bound

Issues with Theorem 23:

- It does not provide insightful nor easily computable expressions for v_i (which are needed to run the algorithm).
- ► It does not answer the following inverse problem: given data matrix A and/or its sparsity pattern {C_i}, design a "good" sampling.

The following two results go a good way to overcoming these issues.

Theorem 27 (Useful ESO; [14])

Let the assumptions of Theorem 23 be satisfied. Then (51) holds (i.e., $(f, \hat{S}) \sim ESO(v)$) with v given by:

$$v_i = \sum_{j=1}^m \lambda(C_j \cap \hat{S}) A_{ji}^2, \quad i = 1, 2, \dots, n.$$
 (56)



イロト イポト イヨト イヨト 二日

Proof of Theorem 27

Note that it follows from (21) that for any vector $\theta \in \mathbb{R}^n$ and any j the following identity holds:

$$\mathbf{E}\left[\left(\sum_{i\in C_j\cap\hat{S}}\theta_i\right)^2\right] = \sum_{i=1}^n [P(C_j\cap\hat{S})]_{ij}\theta_i\theta_j = \theta^T P(C_j\cap\hat{S})\theta.$$
(57)

Fix $h \in \mathbb{R}^n$. Let $z_j = (z_j^{(1)}, \dots, z_i^{(n)})^T \in \mathbb{R}^n$ be defined as follows: $z_i^{(i)} = h^{(i)}A_{ii}$. We then have

$$\mathbf{E} \left[h_{[\hat{S}]}^{T} A^{T} A h_{[\hat{S}]} \right] = \sum_{j=1}^{m} \mathbf{E} \left[h_{[\hat{S}]}^{T} A_{j:}^{T} A_{j:} h_{[\hat{S}]} \right] = \sum_{j=1}^{m} \mathbf{E} \left[\left(\sum_{i \in C_{j} \cap \hat{S}} h^{(i)} A_{ji} \right)^{2} \right]$$

$$\stackrel{(57)}{=} \sum_{j=1}^{m} z_{j}^{T} P(C_{j} \cap \hat{S}) z_{j} \stackrel{(52)}{\leq} \sum_{j=1}^{m} \lambda(C_{j} \cap \hat{S}) \left(z_{j}^{T} \operatorname{Diag}(P(C_{j} \cap \hat{S})) z_{j} \right)$$

$$\stackrel{(37)}{=} \sum_{j=1}^{m} \lambda(C_{j} \cap \hat{S}) \sum_{i \in C_{j}} p_{i} (h^{(i)} A_{ji})^{2} = \sum_{j=1}^{m} \lambda(C_{j} \cap \hat{S}) \sum_{i=1}^{n} p_{i} (h^{(i)} A_{ji})^{2}$$

$$= \sum_{i=1}^{n} p_{i} (h^{(i)})^{2} \sum_{j=1}^{m} \lambda(C_{j} \cap \hat{S}) A_{ji}^{2} = \sum_{i=1}^{n} p_{i} (h^{(i)})^{2} v_{i}.$$

★ロト ★御 と ★ 注 と ★ 注 と … 注 69 / 116

п.

Useful bounds on $\lambda(\hat{S})$

Theorem 28 ([14])

Let \hat{S} be an arbitrary sampling.

- 1. Lower bound. If \hat{S} is not nill, then $\frac{\mathsf{E}[[\hat{S}]^2]}{\mathsf{E}[[\hat{S}]]} \leq \lambda(\hat{S})$.
- 2. Upper bound. If $|\hat{S}| \leq \tau$ with probability 1, then $\lambda(\hat{S}) \leq \tau$.
- 3. Identity. If $|\hat{S}| = \tau$ with probability 1, then $\lambda(\hat{S}) = \tau$.

Let us apply the 2nd part of the above theorem to the sampling $J \cap \hat{S}$: Corollary 29

Let \hat{S} be an arbitrary sampling, $J \subseteq [n]$ and c a constant such that $|J \cap \hat{S}| \leq c$ with probability 1. Then

$$\lambda(J \cap \hat{S}) \leq c.$$

In particular, if $|\hat{S}| \leq \tau$ with probability 1, then $|J \cap \hat{S}| \leq \min\{|J|, \tau\}$ with probability 1, and hence $\lambda(J \cap \hat{S}) \leq \min\{|J|, \tau\}$.

Remark: The above corollary is useful as we can apply it in connection with Theorem 27 with $J = C_j$ for j = 1, 2, ..., m.



Computing $\lambda(J \cap \hat{S})$: Product Sampling

Example 30 (Product Sampling)

Assume that the sets $\{C_j\}$ in Model 1 form a partition of [n]. The consider the sampling \hat{S} defined as follows:

$$\mathbf{P}(\hat{S} = S) = \begin{cases} (\prod_{j=1}^{m} |C_j|)^{-1}, & S \in C_1 \times C_2 \times \cdots \times C_m, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $|C_j \cap \hat{S}| = 1$ with probability 1, and hence by Corollary 29,

 $\lambda(C_j \cap \hat{S}) \leq 1.$

On the other hand, by the first part of Theorem 28, $\lambda(C_j \cap \hat{S}) \ge 1$, and hence this sampling achieves the smallest possible value of the " λ parameters" in (56) (which is "good" as other things equal, ESO with small $\{v_i\}$ is better). Let us remark that $\mathbf{E}[|\hat{S}|] = m$.



Computing $\lambda(J \cap \hat{S})$: τ -Nice Sampling

Exercise 5 (τ -Nice Sampling)

Show by direct computation that if \hat{S} is a τ -nice sampling, then the lower bound in part 1 of Theorem 28 is attained for $C_j \cap \hat{S}$ for all j:

$$\lambda(C_{j} \cap \hat{S}) = \frac{\mathbf{E}[|C_{j} \cap \hat{S}|^{2}]}{\mathbf{E}[|C_{j} \cap \hat{S}|]} \stackrel{(34)+(28)}{=} 1 + \frac{(\omega_{j} - 1)(\tau - 1)}{\max\{n - 1, 1\}}, \quad (58)$$

where $\omega_{j} = |C_{j}|.$
Computing $\lambda(J \cap \hat{S})$: Distributed τ -Nice Sampling - Part I

Exercise 6 (Distributed τ -Nice Sampling; [14]) Show that if \hat{S} is the distributed τ -nice sampling, then

$$\lambda(C_j \cap \hat{S}) \leq \underbrace{1 + \frac{(\tau - 1)(\omega_j - 1)}{s_1}}_{\lambda_{1,j}} + \underbrace{\left(\frac{\tau}{s} - \frac{\tau - 1}{s_1}\right) \frac{\omega'_j - 1}{\omega'_j}}_{\lambda_{2,j}}\omega_j, \quad (59)$$

where $s_1 = \max\{1, s - 1\}$, $\omega_j = |C_j|$, and ω'_j is the number of partitions "active" at row j of A:

$$\omega'_{j} \stackrel{def}{=} |\{I : A_{ji} \neq 0 \text{ for some } i \in \mathcal{P}_{I}\}|.$$

Exercise 7

Show that if the number of partitions is 1 (c = 1), bound (59) for the distributed τ -nice sampling specializes to the bound (58) for the τ -nice sampling.



Computing $\lambda(J \cap \hat{S})$: Distributed τ -Nice Sampling - Part II

Lemma 31 ([14])

Consider the distributed τ -nice sampling. Suppose $\tau \ge 2$. For any $1 \le \eta \le s$ the following holds:

$$\left(\frac{\tau}{s}-\frac{\tau-1}{s-1}\right)\eta\leq \frac{1}{\tau-1}\left(1+\frac{(\tau-1)(\eta-1)}{s-1}\right).$$

Note that Lemma 31 implies that

$$\lambda_{1,j} + \lambda_{2,j} \le \left(1 + \frac{1}{\tau - 1}\right) \lambda_{1,j}.$$
(60)

Distributed NSync: Cost of Distribution

Assume *f* is 1-strongly convex, and consider running NSync with the distributed τ -nice sampling. Then $p_i = \frac{\mathbf{E}[\hat{S}]}{n} = \frac{\tau c}{sc} = \frac{\tau}{s}$ and hence the leading term in the complexity bound is

$$\Lambda = \max_{i} \frac{v_{i}}{p_{i}} \stackrel{(56)}{=} \max_{i} \frac{s \sum_{j=1}^{m} \lambda(C_{j} \cap \hat{S})}{\tau} \stackrel{(60)}{\leq} \max_{i} \frac{s \sum_{j=1}^{m} (\lambda_{1,j} + \lambda_{2,j}) A_{ji}^{2}}{\tau} \stackrel{\text{def}}{=} \Lambda'.$$

- Notice that the effect of partitioning on complexity comes only through λ_{2,j}.
- Define a new quantity that does not depend on partitioning:

$$\Lambda'' = \max_{i} \frac{s \sum_{j=1}^{m} \lambda_{1,j} A_{ji}^2}{\tau}$$

and notice that (60) implies that

$$\Lambda'' \leq \Lambda' \leq (1 + \frac{1}{\tau - 1})\Lambda''$$

This means that:

Theorem 32 (Cost of Distribution: compare with [10, 14]) If $\tau \ge 2$, the worst-case partitioning is at most $(1 + \frac{1}{\tau})$ times worse than the optimal partitioning, in terms of the number of iterations of NSync.



Proof of Theorem 28 - Part I

Point 1. For simplicity of notation, put $P = P(\hat{S})$. If we choose $\theta \in \mathbb{R}^n$ with $\theta_i = (\text{Tr}(P))^{-1/2}$ for all *i*, we get $\theta^T D^P \theta = \sum_i P_{ii} \theta_i^2 = 1$ and hence

$$\lambda(\hat{S}) \stackrel{(52)}{\geq} \theta^{T} P \theta \stackrel{(57)}{=} \mathbf{E} \Big[\Big(\sum_{i \in \hat{S}} \theta_i \Big)^2 \Big] = \frac{\mathbf{E} \Big[\Big(\sum_{i \in \hat{S}} 1 \Big)^2 \Big]}{\mathsf{Tr}(P)} \stackrel{(22)}{=} \frac{\mathbf{E} [|\hat{S}|^2]}{\mathbf{E} [|\hat{S}|]}.$$

Point 2. Let us represent \hat{S} as a convex combination of elementary samplings: $\hat{S} = \sum_{S \subseteq [n]} q_S \hat{E}_S$, where $q_S = \mathbf{P}(\hat{S} = S)$. Note that then we also have

$$P(\hat{S}) = \sum_{S \subseteq [n]} q_S P(\hat{E}_S) \stackrel{(52)}{=} \sum_{S \subseteq [n]} q_S e_S e_S^T.$$
(61)



Proof of Theorem 28 - Part II Since $|\hat{S}| \le \tau$ with probability 1, we have $|S| \le \tau$ whenever $q_S > 0$. For any $\theta \in \mathbb{R}^n$ we can now estimate:

$$\begin{aligned} \theta^T P(\hat{S}) \theta \stackrel{\text{(61)}}{=} \sum_{S:q_S > 0} q_S (e_S^T \theta)^2 &\leq \sum_{S:q_S > 0} q_S \|e_S\|^2 \sum_{i \in S} \theta_i^2 \\ \stackrel{\text{(54)}}{=} \sum_{S:q_S > 0} q_S |S| \sum_{i \in S} \theta_i^2 \\ &\leq \tau \sum_{S:q_S > 0} q_S \theta^T \operatorname{Diag}(e_S e_S^T) \theta \\ &= \tau \theta^T \left(\sum_{S:q_S > 0} q_S \operatorname{Diag}(e_S e_S^T) \right) \theta \\ \stackrel{\text{(61)}}{=} \tau \left(\theta^T \operatorname{Diag}(P(\hat{S})) \theta \right). \end{aligned}$$

We thus see that $\lambda(\hat{S}) \leq \tau$.

Point 3. The result follows by combining the upper and lower bounds. Alternatively, we can see this by inspecting the derivation in part 2. Indeed, if $|\hat{S}| = \tau$ with probability 1, then $|S| = \tau$ whenever $q_S > 0$, and hence the second inequality in point 2 above is an equality. By choosing $\theta_i = \alpha$ for any constant α , the first inequality turns into an equality (this is because we then have equality in the Cauchy-Schwartz inequality $e_S^T \theta \le ||e_S||^2 \sum_{i \in S} \theta_i^2$ for all S).



ESO($f \sim \text{Model 3}, \hat{S} \sim \tau$ -nice)

Theorem 33

Let f satisfy assumptions in Model 3 and \hat{S} be a τ -nice sampling. Then for all $x, h \in \mathbb{R}^N$.

$$\mathbf{E}\left[f(x+h_{[\hat{S}]})\right] \le f(x) + \frac{\tau}{n}\left(\langle \nabla f(x), h \rangle + \frac{1}{2} \|h\|_{\nu}^{2}\right), \qquad (62)$$

where

$$v_{i} \stackrel{def}{=} \sum_{j=1}^{m} \beta_{j} L_{ji} = \sum_{j:i \in C_{j}} \beta_{j} L_{ji}, \qquad i = 1, 2, \dots, n,$$
(63)
$$\beta_{j} \stackrel{def}{=} 1 + \frac{(\omega_{j} - 1)(\tau - 1)}{\max\{1, n - 1\}}, \qquad j = 1, 2, \dots, m.$$

That is, $(f, S) \sim ES$



Proof of Theorem 33 - Part I

▶ We first claim that for all *j*,

$$\mathbf{E}\left[f_{j}(x+h_{[\hat{S}]})\right] \leq f_{j}(x) + \frac{\tau}{n}\left(\langle \nabla f_{j}(x), h \rangle + \frac{\beta_{j}}{2} \|h\|_{L_{j}}^{2}\right), \quad (64)$$

where $L_{j:} = (L_{j1}, \ldots, L_{jn}) \in \mathbb{R}^n$. That is, $(f_j, \hat{S}) \sim ESO(\beta_j L_{j:})$. Equation (62) then follows by adding up the inequalities (64) for all j. In the rest we prove the claim.

▶ A well known consequence of (46) is that for all $x \in \mathbb{R}^N$, $t \in \mathbb{R}^{N_i}$,

$$f_j(x + U_i t) \le f_j(x) + \langle \nabla_i f_j(x), t \rangle + \frac{L_{ji}}{2} ||t||_{(i)}^2.$$
 (65)



Proof of Theorem 33 - Part II

We fix x and define

$$\hat{f}_j(h) \stackrel{\text{def}}{=} f_j(x+h) - f_j(x) - \langle \nabla f_j(x), h \rangle.$$
(66)

Since

it now only remains to show that

$$\mathbf{E}\left[\hat{f}_{j}(h_{[\hat{S}]})\right] \leq \frac{\tau_{\beta_{j}}}{2n} \|h\|_{L_{j}}^{2}.$$
(67)

► We now adopt the convention that expectation conditional on an event which happens with probability 0 is equal to 0. Let $\eta_j \stackrel{\text{def}}{=} |C_j \cap \hat{S}|$, and using this convention, we can write

$$\mathbf{E}\left[\hat{f}_{j}(h_{[\hat{S}]})\right] = \sum_{k=0}^{n} \mathbf{P}(\eta_{j}=k) \mathbf{E}\left[\hat{f}_{j}(h_{[\hat{S}]}) \mid \eta_{j}=k\right].$$
(68)

Proof of Theorem 33 - Part III

For any $k \ge 1$ for which $\mathbf{P}(\eta_j = k) > 0$, we now use use convexity of \hat{f}_j to write

$$\mathbf{E}\left[\hat{f}_{j}(h_{[\hat{S}]}) \mid \eta_{j} = k\right] = \mathbf{E}\left[\hat{f}_{j}\left(\frac{1}{k}\sum_{i \in C_{j} \cap \hat{S}} kU_{i}h^{(i)}\right) \mid \eta_{j} = k\right]$$

$$\leq \mathbf{E}\left[\frac{1}{k}\sum_{i \in C_{j} \cap \hat{S}} \hat{f}_{j}\left(kU_{i}h^{(i)}\right) \mid \eta_{j} = k\right]$$

$$\stackrel{(35)}{=} \frac{1}{\omega_{j}}\sum_{i \in C_{j}} \hat{f}_{j}\left(kU_{i}h^{(i)}\right)$$

$$\stackrel{(65)+(66)}{\leq} \frac{1}{\omega_{j}}\sum_{i \in C_{j}} \frac{L_{ji}}{2} ||kh^{(i)}||_{(i)}^{2} = \frac{k^{2}}{2\omega_{j}} ||h||_{L_{j:}}^{2}. (69)$$

82 / 116

<ロ> (四) (四) (三) (三) (三) (三)

Proof of Theorem 33 - Part IV

► Finally,

$$\mathbf{E}\left[\hat{f}_{j}(h_{[\hat{S}]})\right] \stackrel{(68)+(69)}{\leq} \sum_{k} \mathbf{P}(\eta_{j}=k) \frac{k^{2}}{2\omega_{j}} \|h\|_{L_{j:}}^{2}$$
$$= \frac{1}{2\omega_{j}} \|h\|_{L_{j:}}^{2} \mathbf{E}[|C_{j} \cap \hat{S}|^{2}]$$
$$\stackrel{(34)}{=} \frac{\tau\beta_{j}}{2n} \|h\|_{L_{j:}}^{2},$$

and hence (67) is proved.



$\mathsf{DSO}(f \sim \mathsf{Model 3})$

Corollary 34

Let f satisfy assumptions in Model 3 and \hat{S} be a τ -nice sampling. Then for all $x, h \in \mathbb{R}^N$ we have

$$f(x+h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{\bar{\omega}\bar{L}}{2} \|h\|_{w}^{2},$$
(70)

where

$$\bar{\omega} \stackrel{\text{def}}{=} \sum_{j} \omega_{j} \frac{\sum_{i} L_{ji}}{\sum_{k,i} L_{ki}}, \quad \bar{L} \stackrel{\text{def}}{=} \frac{\sum_{ji} L_{ji}}{n}, \quad w_{i} \stackrel{\text{def}}{=} \frac{n}{\sum_{j,i} \omega_{j} L_{ji}} \sum_{j} \omega_{j} L_{ji}.$$
(71)

Note that $\bar{\omega}$ is a data-weighted average of the values $\{\omega_j\}$ and that $\sum w_i = n$.

Proof.

This follows from Theorem 33 used with $\tau = n$ (notice that $\bar{\omega}\bar{L}w = v$).

・ロッ ・雪 ・ ・ ヨ ・ ・ ヨ ・

ESO and Lipschitz Continuity I

We will now study the collection of functions $\hat{\phi}_x : \mathbb{R}^N \to \mathbb{R}$ for $x \in \mathbb{R}^N$ defined by

$$\hat{\phi}_{x}(h) \stackrel{\text{def}}{=} \mathbf{E}\left[\phi(x+h_{[\hat{S}]})\right].$$
(72)

Let us first establish some basic connections between ϕ and $\hat{\phi}_{x}$.

Lemma 35 ([9])
Let Ŝ be any sampling and φ : ℝ^N → ℝ any function and x ∈ ℝ^N. Then
(i) if φ is convex, so is φ̂_x,
(ii) φ̂_x(0) = φ(x),
(iii) If Ŝ is proper and uniform, and φ : ℝ^N → ℝ is continuously differentiable, then

$$abla \hat{\phi}_x(0) = rac{\mathsf{E}[|\hat{S}|]}{n}
abla \phi(x).$$



イロト イヨト イヨト イヨト 二日

Proof of Lemma 35

Fix $x \in \mathbb{R}^N$. Notice that

$$\hat{\phi}_x(h) = \mathbf{E}[\phi(x+h_{[\hat{S}]})] = \sum_{S \subseteq [n]} \mathbf{P}(\hat{S}=S)\phi(x+U_Sh),$$

where

$$U_S \stackrel{\mathsf{def}}{=} \sum_{i \in S} U_i U_i^{\mathsf{T}}.$$

As $\hat{\phi}_x$ is a convex combination of convex functions, it is convex, establishing (i). Property (ii) is trivial. Finally,

$$\nabla \hat{\phi}_{x}(\mathbf{0}) = \mathbf{E} \left[\nabla \phi(x + h_{[\hat{S}]}) \Big|_{h=0} \right] = \mathbf{E} \left[U_{\hat{S}} \nabla \phi(x) \right] = \mathbf{E} \left[U_{\hat{S}} \right] \nabla \phi(x) = \frac{\mathbf{E}[|\hat{S}|]}{n} \nabla \phi(x).$$

The last equality follows from the observation that $U_{\hat{S}}$ is an $N \times N$ binary diagonal matrix with ones in positions (v, v) for coordinates $v \in \{1, 2, ..., N\}$ belonging to blocks $i \in \hat{S}$ only, coupled with the fact that for uniform samplings, $p_i = \mathbf{E}[|\hat{S}|]/n$.



ESO and Lipschitz Continuity II

We now establish a connection between ESO and a uniform bound in x on the Lipschitz constants of the gradient "at the origin" of the functions $\{\hat{\phi}_x, x \in \mathbb{R}^N\}$.

Theorem 36

Let \hat{S} be proper and uniform, and $\phi : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable. Then the following statements are equivalent:

(i)
$$(\phi, \hat{S}) \sim ESO(v)$$
,

(ii)
$$\hat{\phi}_x(h) \leq \hat{\phi}_x(0) + \langle \nabla \hat{\phi}_x(0), h \rangle + \frac{1}{2} \frac{\mathbb{E}[|\hat{S}|]}{n} ||h||_v^2, \qquad x, h \in \mathbb{R}^N.$$

Proof.

We only need to substitute (72) and Lemma 35(ii-iii) into inequality (ii) and compare the result with the definition of ESO (5).



Lecture 6 APPROX



The Problem

We are interested in solving the following optimization problem:

$$\min_{x \in \mathbb{R}^N} f(x) + \psi(x), \tag{73}$$

where

f is a "smooth" convex function (to be made precise later),

• ψ is block separable:

$$\psi(x) = \sum_{i=1}^{n} \psi_i(x^{(i)}), \tag{74}$$

where $\psi_i : \mathbb{R}^{N_i} \to \mathbb{R} \cup \{+\infty\}$ are convex and closed.



Examples of Regularizers

Smooth optimization:

$$\psi(x)\equiv 0$$

Box constraints: Let $X_i \subseteq \mathbb{R}^{N_i}$ be closed convex sets and

$$\psi(x) = \begin{cases} 0, & x^{(i)} \in X_i & \text{for all} & i \in [n] \\ +\infty, & \text{otherwise.} \end{cases}$$

L2/Ridge:

$$\psi(\mathbf{x}) = \lambda \|\mathbf{x}\|_2^2$$

► L1/LASSO:

$$\psi(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$$

Group LASSO:

$$\psi(x) = \sum_{i=1}^{n} \|x^{(i)}\|_2$$

All are block separable and convex.



90 / 116

APPROX algorithm – Version 1

1: Choose
$$x_0 \in \text{dom } \psi$$
 and set $z_0 = x_0$ and $\theta_0 > 0$
2: for $k \ge 0$ do
3: $y_k = (1 - \theta_k)x_k + \theta_k z_k$
4: Generate a random set of blocks $S_k \sim \hat{S}$
5: $z_{k+1} = z_k$
6: for $i \in S_k$ do
7: $z_{k+1}^{(i)} = \arg\min_{z \in \mathbb{R}^{N_i}} \left\{ \langle \nabla_i f(y_k), z \rangle + \frac{\theta_k v_i}{2p_i} ||z - z_k^{(i)}||_{(i)}^2 + \psi_i(z) \right\}$
8: end for
9: $x_{k+1} = y_k + \theta_k (z_{k+1} - z_k) \bullet p^{-1}$
10: $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2 - \theta_k^2}}{2}$ (fast) or $\theta_{k+1} = \theta_k$ (normal)
11: end for

Remark 1: Our analysis will follow this version.

Remark 2: The • product is to be applied block-wise, i.e., for $a \in \mathbb{R}^N$:

$$a \bullet p^{-1} = \sum_{i=1}^{n} \frac{1}{p_i} U_i a^{(i)}.$$



91/116

E

Reformulation: Change of Variables - Part I

Focusing on the iterates x_k , y_k , z_k only, the algorithm can schematically be written as follows:

APPROX Schema: Version 1

$$y_k \leftarrow (1-\theta_k)x_k + \theta_k z_k$$
 (75)

$$z_{k+1} \leftarrow Procedure(y_k; z_k; S_k)$$
 (76)

$$x_{k+1} \leftarrow y_k + \theta_k (z_{k+1} - z_k) \bullet p^{-1}$$
(77)

Consider the change of variables from $\{x_k, y_k, z_k,\}$ to $\{z_k, g_k\}$ where

$$g_k = y_k - z_k \tag{78}$$

Inverse change of variables: From $\{z_k, g_k\}$ we can recover $\{x_k, y_k, z_k\}$ as follows:

$$x_{k+1} \stackrel{(77)+(78)}{=} (z_k + g_k) + \theta_k (z_{k+1} - z_k) \bullet p^{-1}, \quad y_k \stackrel{(78)}{=} z_k + g_k \quad (79)$$

92/116

Reformulation: Change of Variables - Part II

It remains to show that g_{k+1} can be computed (from g and z):

$$g_{k+1} \stackrel{(78)}{=} y_{k+1} - z_{k+1} \stackrel{(75)}{=} (1 - \theta_{k+1})(x_{k+1} - z_{k+1})$$

$$\stackrel{(79)}{=} (1 - \theta_{k+1})(g_k - (e - \theta_k p^{-1}) \bullet (z_{k+1} - z_k)),$$

where $e \in \mathbb{R}^n$ is the vector of all ones.

Method (75)–(77) can thus be written in the form: APPROX Schema: Version 2

$$z_{k+1} \leftarrow Procedure(z_k + g_k; z_k; S_k)$$
(80)

$$g_{k+1} \leftarrow (1-\theta_{k+1}) \left(g_k - (e - \theta_k p^{-1}) \bullet (z_{k+1} - z_k) \right)$$
 (81)

Historical Notes

1. "Normal" & uniform. Choose $\theta_0 = \frac{\mathbb{E}[|\hat{S}|]}{n}$ and $\theta_k = \theta_0$ for all k and let \hat{S} be uniform, i.e., $p_i = \frac{\mathbb{E}[|\hat{S}|]}{n}$. Then $g_k = 0$ for all k and the method simplifies to:

$$z_{k+1} \leftarrow Procedure(z_k; z_k; S_k)$$
 (82)

This is the PCDM method of R. and Takáč [5].

- Fast & uniform. For uniform Ŝ, "fast" option in Step 10 and θ₀ = E[[Ŝ]]/n, this method reduces to the original APPROX method of Fercoq & R. [12].
- Fast & non-uniform. For non-uniform Ŝ presented here, θ₀ ≤ min_i p_i (and θ₀ ≤ 1 if ψ ≡ 0) and for the "fast" option in Step 10, it was analyzed by Qu & R. [14].



APPROX algorithm – Version 2 (variables g_k, z_k)

In detail, version 2 has the following form:

- 1: Choose $x_0 \in \text{dom } \psi$ and $\theta_0 > 0$, $g_0 = 0$ and $z_0 = x_0$
- 2: for k > 0 do
- Generate a random set of blocks $S_k \sim \hat{S}$ 3:
- 4: $Z_{k+1} \leftarrow Z_k$ for $i \in S_k$ do 5:
- $t_{i}^{(i)} =$ 6: $\arg\min_{t\in\mathbb{R}^{N_i}}\left\{\langle\nabla_i f(g_k+z_k),t\rangle+\tfrac{\theta_k v_i}{2p_i}\|t\|_{(i)}^2+\psi_i(z_k^{(i)}+t)\right\}$ $z_{k+1}^{(i)} \leftarrow z_{k}^{(i)} + t_{k}^{(i)}$ 7:
- end for 8.
- $g_{k+1} \leftarrow (1-\theta_{k+1})(g_k (e-\theta_k p^{-1}) \bullet t_k)$ 9: $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$ (fast) or $\theta_{k+1} = \theta_k$ (normal)
- 10.
- 11: end for
- 12: OUTPUT: $x_{k+1} = (z_k + g_k) + \theta_k (z_{k+1} z_k) \bullet p^{-1}$



Complexity

Theorem 37 ([12, 14])

Assume:

- $\{S_k\}_{k\geq 1}$ are iid following the distribution of a proper sampling \hat{S} ,
- f is convex and $(f, \hat{S}) \sim \text{ESO}(v)$,
- ψ is block separable, where ψ_i are convex and closed.

Let $x_0 \in \text{dom } F$ and choose $\theta_0 \in (0, \min_i p_i]$ (if $\psi = 0$, choose $\theta_0 \in (0, 1]$). Then for any point y such that $F(y) \leq F(x_0)$ (and hence also for the optimal point x_* if such a point exists), the iterates $\{x_k\}$ of APPROX satisfy

$$\mathbf{E}[F(x_k) - F(y)] \leq \frac{4}{((k-1)\theta_0 + 2)^2}C, \quad k \ge 1$$
 (83)

where

$$C \stackrel{\text{def}}{=} (1 - \theta_0) \left(F(x_0) - F(y) \right) + \frac{\theta_0^2}{2} \|x_0 - y\|_{\rho^{-2} \bullet v}^2. \tag{84}$$

Comments: Smooth Case ($\psi \equiv 0$)

▶ In the smooth case ($\psi \equiv 0$) we may choose $\theta_0 = 1$ and get

$$\mathbf{E}[F(x_k)-F(x_*)] \leq \frac{2\|x_0-x_*\|_{p^{-2}\bullet v}^2}{(k+1)^2} = \frac{2}{(k+1)^2} \sum_{i=1}^n \frac{v_i}{p_i^2} \|x_0^{(i)}-x_*^{(i)}\|_{(i)}^2.$$

If, moreover, we choose uniform sampling Ŝ and let τ = E[|Ŝ|], then since p_i = ^T/_n for all i, we get

$$\mathbf{E}[F(x_k) - F(x_*)] \le \frac{2n^2 \|x_0 - x_*\|_{\nu}^2}{\tau^2 (k+1)^2}$$

In other words, the number of iterations for obtaining an ϵ -solution (in expectation) does not exceed

$$k = \left[\frac{\sqrt{2}n\|x_0 - x_*\|_{\nu}}{\tau\sqrt{\epsilon}} - 1\right].$$
 (85)

Note that the bound gets better as the average number of processors (τ) increases (with the caveat that v will generally also grow in τ, but less so for sparse problems; as ESO predicts).





We shall now prove the Theorem. We first need to establish 4 lemmas.



Lemma: Properties of the sequence θ_k

In the first lemma we summarize well-known properties of the sequence θ_k used in APPROX.

Lemma 38

The sequence $\{\theta_k\}_{k\geq 0}$ defined APPROX, under the FAST option, is decreasing and satisfies

$$0 < \theta_k \le \frac{2}{k+2/\theta_0} \le 1 \tag{86}$$

and

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}.$$
(87)

Lemma: x_k is in the convex hull of z_0, \ldots, z_k

Lemma 39 Let $\{x_k, z_k\}_{k \ge 0}$ be the iterates of APPROX; and assume $0 < \theta_0 \le \min_i p_i$. Then for all $k \ge 0$ we have

$$\mathbf{x}_{k}^{(i)} = \sum_{l=0}^{k} \gamma_{kl}^{(i)} \mathbf{z}_{l}^{(i)}, \quad i = 1, 2, \dots, n$$
 (88)

where for each *i*, the coefficients $\gamma_{k0}^{(i)}, \ldots, \gamma_{kk}^{(i)}$ are non-negative and sum to 1. Moreover, the coefficients are defined recursively by setting $\gamma_{00}^{(i)} = 1, \ \gamma_{10}^{(i)} = 1 - \frac{\theta_0}{p_i}, \ \gamma_{11}^{(i)} = \frac{\theta_0}{p_i} \text{ and for } k \ge 1,$

$$\gamma_{k+1,l}^{(i)} = \begin{cases} (1-\theta_k)\gamma_{kl}^{(i)}, & l = 0, \dots, k-1, \\ (1-\theta_k)\gamma_{kk}^{(i)} + \theta_k - \frac{\theta_k}{p_i}, & l = k, \\ \frac{\theta_k}{p_i}, & l = k+1. \end{cases}$$
(89)

Moreover, for all $k \ge 0$ and $i \in [n]$, the following identity holds

$$\gamma_{k+1,k}^{(i)} + \gamma_{k+1,k+1}^{(i)} = (1 - \theta_k)\gamma_{kk}^{(i)} + \theta_k.$$
(90)

Remarks about Lemma 39

Note that if p_i = p_j for all i, j ∈ [n] (i.e., if Ŝ is a uniform sampling), then γ⁽ⁱ⁾_{kl} = γ^(j)_{kl} for all i, j, and hence the lemma says that x_k is a convex combination of the vectors z₀, z₁, ..., z_k.

- The lemma is only needed in the nonsmooth case ($\psi \neq 0$).
- ► The proof is straightforward of a "follow-your-nose" style.

Proof of Lemma 39 - Part I

We proceed by induction in k. Fix any $i \in [n]$.

Step 1 (Base case).

- Since $x_0 = z_0$, we have $\gamma_{00}^{(i)} = 1$.
- ► Since $x_1 = y_0 + \theta_0(z_1 z_0) \bullet p^{-1}$ and $y_0 = x_0$, we get $x_1^{(i)} = (1 \frac{\theta_0}{p_i})z_0^{(i)} + \frac{\theta_0}{p_i}z_1^{(i)}$, whence $\gamma_{10}^{(i)} = 1 \frac{\theta_0}{p_i}$, $\gamma_{11}^{(i)} = \frac{\theta_0}{p_i}$.

Note that for each k, the coefficients are nonnegative and sum to one.

Step 2 (Recursive relation). If the recursive relation (89) holds for some $k \ge 1$, then it holds for k + 1:

$$\begin{aligned} x_{k+1}^{(i)} &\stackrel{(Step \; 9)}{=} \quad y_k^{(i)} + \frac{\theta_k}{p_i} (z_{k+1}^{(i)} - z_k^{(i)}) \\ &\stackrel{(Step \; 3)}{=} \quad (1 - \theta_k) x_k^{(i)} + \theta_k z_k^{(i)} + \frac{\theta_k}{p_i} (z_{k+1}^{(i)} - z_k^{(i)}) \\ &\stackrel{(88)}{=} \quad (1 - \theta_k) \sum_{l=0}^k \gamma_{kl}^{(i)} z_l^{(i)} + \theta_k z_k^{(i)} + \frac{\theta_k}{p_i} (z_{k+1}^{(i)} - z_k^{(i)}) \\ &= \sum_{l=0}^{k-1} \underbrace{(1 - \theta_k) \gamma_{kl}^{(i)}}_{\gamma_{k+1,l}^{(i)}} z_l^{(i)} + \underbrace{((1 - \theta_k) \gamma_{kk}^{(i)} + \theta_k - \frac{\theta_k}{p_i})}_{\gamma_{k+1,k}^{(i)}} z_k^{(i)} + \underbrace{(y_{k+1}^{(i)} - y_{k+1}^{(i)})}_{\gamma_{k+1,k}^{(i)}} z_k^{(i)} + \underbrace{(y_{k+1}^{(i)} - y_{k+1}$$

Proof of Lemma 39 - Part II

Step 3 (Nonnegativity).

Since 0 < θ_k ≤ 1 (because θ₀ ≤ min_i p_i ≤ 1 and {θ_k} is a decreasing sequence of positive numbers), we deduce from (89) and, using the inductive non-negativity assumption, that γ⁽ⁱ⁾_{k+1,l} ≥ 0 for l = 0,..., k − 1.

Moreover,

$$\begin{split} \gamma_{k+1,k}^{(i)} &\stackrel{(89)}{=} (1-\theta_k)\gamma_{kk}^{(i)} + \theta_k - \frac{\theta_k}{\rho_i} \\ &= \theta_k(1-\gamma_{kk}^{(i)}) + \gamma_{kk}^{(i)} - \frac{\theta_k}{\rho_i} \\ &\stackrel{(89)}{=} \theta_k(1-\gamma_{kk}^{(i)}) + \frac{\theta_{k-1}-\theta_k}{\rho_i} > \theta_k(1-\gamma_{kk}^{(i)}) \ge 0. \end{split}$$

where the first inequality follows since $\{\theta_k\}$ is a decreasing sequence, and the last inequality by the inductive hypothesis that $\gamma_{kl}^{(i)}$, $l = 0, 1, \ldots, k$ are nonnegative and sum to 1.

• Finally,
$$\gamma_{k+1,k+1}^{(i)} = \frac{\theta_k}{p_i} > 0.$$



Proof of Lemma 39 - Part III

Step 4 (Unit sum). Finally, we can write

$$\sum_{l=0}^{k+1} \gamma_{k+1,l}^{(i)} = \sum_{l=0}^{k-1} \gamma_{k+1,l}^{(i)} + \gamma_{k+1,k}^{(i)} + \gamma_{k+1,k+1}^{(i)}$$

$$\stackrel{(89)}{=} (1-\theta_k) \sum_{l=0}^{k-1} \gamma_{kl}^{(i)} + \left((1-\theta_k)\gamma_{kk}^{(i)} + \theta_k - \frac{\theta_k}{p_i}\right) + \frac{\theta_k}{p_i}$$

$$= (1-\theta_k) \sum_{l=0}^{k} \gamma_{kl}^{(i)} + \theta_k$$

$$= 1,$$

where the last step follows from the inductive hypothesis that $\{\gamma_{kl}^{(i)}\}\$ for $l = 0, 1, \ldots, k$ sum to one.



Lemma: Tseng Define

$$\tilde{z}_{k+1} \stackrel{\text{def}}{=} \arg\min_{z \in \mathbb{R}^N} \left\{ \psi(z) + \langle \nabla f(y_k), z - y_k \rangle + \frac{n\theta_k}{2\tau} \|z - z_k\|_v^2 \right\}$$

$$(15)+(74) = \arg\min_{\substack{z^{(i)} \in \mathbb{R}^{N_i} \\ i \in [n]}} \sum_{i=1}^n \left\{ \psi_i(z^{(i)}) + \langle \nabla_i f(y_k), z^{(i)} - y_k^{(i)} \rangle + \frac{n\theta_k v_i}{2\tau} \|z^{(i)} - z_k^{(i)}\|_{(i)}^2 \right\}.$$

From this and the definition of z_{k+1} in APPROX, we see that

$$z_{k+1}^{(i)} = \begin{cases} \tilde{z}_{k+1}^{(i)}, & i \in S_k \\ z_k^{(i)}, & i \notin S_k. \end{cases}$$
(91)

イロト イポト イヨト イヨト 二日

105 / 116

Lemma 40 (Property 1 in [1]) Let $\xi(u) \stackrel{\text{def}}{=} f(y_k) + \langle \nabla f(y_k), u - y_k \rangle + \frac{\theta_k}{2} \|u - z_k\|_{p^{-1} \bullet v}^2$. Then for any $y \in \operatorname{dom} \psi$,

$$\psi(\tilde{z}_{k+1}) + \xi(\tilde{z}_{k+1}) \le \psi(y) + \xi(y) - \frac{\theta_k}{2} \|y - \tilde{z}_{k+1}\|_{p^{-1} \bullet v}^2.$$
(92)

Lemma: Gradient vs Stochastic Gradient Mapping

We now connect the gradient mapping (producing \tilde{z}_{k+1}) and the stochastic block gradient mapping (producing the random vector z_{k+1}).

From now on, by \mathbf{E}_k we denote the expectation with respect to S_k , conditioned on all history.

Lemma 41 ([12]) For any $y \in \mathbb{R}^N$ and $k \ge 0$, $\mathbf{E}_k \left[\|z_{k+1} - y\|_v^2 - \|z_k - y\|_v^2 \right] = \|\tilde{z}_{k+1} - y\|_{\rho \bullet v}^2 - \|z_k - y\|_{\rho \bullet v}^2$. (93)



Proof of Lemma 41

Let \hat{S} be any proper sampling and $a,h\in\mathbb{R}^N.$ Recall the following sampling identities:

$$\mathbf{E}[\|h_{[\hat{S}]}\|_{v}^{2}] \stackrel{(25)}{=} \|h\|_{p \bullet v}^{2}, \qquad \mathbf{E}[\langle a, h_{[\hat{S}]} \rangle_{v}] \stackrel{(24)}{=} \langle a, h \rangle_{p \bullet v}.$$
(94)
Let $h = \tilde{z}_{k+1} - z_{k}$. In view of (14) and (91), we can write
 $h_{[S_{k}]} = z_{k+1} - z_{k}$. Now,

$$\mathbf{E}_{k} \left[\|z_{k+1} - y\|_{v}^{2} - \|z_{k} - y\|_{v}^{2}\right] = \mathbf{E}_{k} \left[\|h_{[S_{k}]}\|_{v}^{2} + 2\langle z_{k} - y, h_{[S_{k}]} \rangle_{v}\right]$$

$$\stackrel{(94)}{=} \|h\|_{p \bullet v}^{2} + 2\langle z_{k} - y, h\rangle_{p \bullet v}$$

$$= \left(\|\tilde{z}_{k+1} - y\|_{p \bullet v}^{2} - \|z_{k} - y\|_{p \bullet v}^{2}\right).$$



107 / 116

イロト イポト イヨト イヨト 二日

Proof of the Main Result (Theorem 37) - Part I

Step 1 (Bounding f). From the definition of y_k in the algorithm:

$$\theta_k(y_k-z_k)=(1-\theta_k)(x_k-y_k). \tag{95}$$

Since $x_{k+1} = y_k + h_{[S_k]}$ with $h = \theta_k(\tilde{z}_{k+1} - z_k) \bullet \sigma$, we use ESO and obtain the following bound:

$$\begin{aligned} \mathbf{E}_{k}[f(x_{k+1})] &= \mathbf{E}_{k}[f(y_{k}+h_{[S_{k}]})] \\ &\leq f(y_{k}) + \langle \nabla f(y_{k}), h \rangle_{p} + \frac{1}{2} \|h\|_{p \bullet w}^{2} \\ &= f(y_{k}) + \theta_{k} \langle \nabla f(y_{k}), \tilde{z}_{k+1} - z_{k} \rangle + \frac{\theta_{k}^{2}}{2} \|\tilde{z}_{k+1} - z_{k}\|_{\sigma \bullet v}^{2} \\ &= (1-\theta_{k})f(y_{k}) - \theta_{k} \langle \nabla f(y_{k}), z_{k} - y_{k} \rangle \\ &+ \theta_{k}(f(y_{k}) + \langle \nabla f(y_{k}), \tilde{z}_{k+1} - y_{k} \rangle + \frac{\theta_{k}}{2} \|\tilde{z}_{k+1} - z_{k}\|_{\sigma \bullet v}^{2}) \end{aligned}$$

$$\begin{aligned} &\stackrel{(95)}{=} (1-\theta_{k})(f(y_{k}) + \langle \nabla f(y_{k}), x_{k} - y_{k} \rangle) \\ &+ \theta_{k}(f(y_{k}) + \langle \nabla f(y_{k}), \tilde{z}_{k+1} - y_{k} \rangle + \frac{\theta_{k}}{2} \|\tilde{z}_{k+1} - z_{k}\|_{\sigma \bullet v}^{2}). \end{aligned}$$



108 / 116

イロト イポト イヨト イヨト 二日
Proof of the Main Result (Theorem 37) - Part II

Step 2 (Bounding ψ for "fast θ_k "). By Lemma 39, each block of the vector x_k is a convex combination of the corresponding blocks of the vectors z_0, \ldots, z_k . By the convexity of each function ψ_i , for all $k \ge 0$ we have

$$\psi_{i}(x_{k}^{(i)}) \stackrel{(88)}{=} \psi_{i}\left(\sum_{l=0}^{k} \gamma_{kl}^{(i)} z_{l}^{(i)}\right) \leq \sum_{l=0}^{k} \gamma_{kl}^{(i)} \psi_{i}(z_{l}^{(i)}) \stackrel{\text{def}}{=} \alpha_{k}^{i}.$$
(97)

Moreover,

$$\psi(x_k) = \sum_{i=1}^n \psi_i(x_k^{(i)}) \stackrel{(97)}{\leq} \sum_{i=1}^n \alpha_k^i \stackrel{\text{def}}{=} \hat{\psi}_k.$$
 (98)



Proof of the Main Result (Theorem 37) - Part III

Then, for all $k \ge 0$ and $i \in \{1, \ldots, n\}$, we have:

$$\begin{aligned} \mathbf{E}_{k}[\alpha_{k+1}^{i}] & \stackrel{(97)+(89)}{=} \mathbf{E}_{k} \left[\sum_{l=0}^{k} \gamma_{k+1,l}^{(i)} \psi_{i}(z_{l}^{(i)}) + \frac{\theta_{k}}{\rho_{i}} \psi_{i}(z_{k+1}^{(i)}) \right] \\ &= \sum_{l=0}^{k} \gamma_{k+1,l}^{(i)} \psi_{i}(z_{l}^{(i)}) + \frac{\theta_{k}}{\rho_{i}} \mathbf{E}_{k}[\psi_{i}(z_{k+1}^{(i)})] \\ \stackrel{(91)}{=} \sum_{l=0}^{k} \gamma_{k+1,l}^{(i)} \psi_{i}(z_{l}^{(i)}) + \frac{\theta_{k}}{\rho_{i}} \left(\rho_{i}\psi_{i}(\tilde{z}_{k+1}^{(i)}) + (1-\rho_{i})\psi_{i}(z_{k}^{(i)})\right) \\ &= \sum_{l=0}^{k} \gamma_{k+1,l}^{(i)} \psi_{i}(z_{l}^{(i)}) + (\frac{1}{\rho_{i}} - 1)\theta_{k}\psi_{i}(z_{k}^{(i)}) + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)}) \\ \stackrel{(89)}{=} (1-\theta_{k}) \sum_{l=0}^{k-1} \gamma_{kl}^{(i)}\psi_{i}(z_{l}^{(i)}) + (\gamma_{k+1,k}^{(i)} + (\frac{1}{\rho_{i}} - 1)\theta_{k})\psi_{i}(z_{k}^{(i)}) + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)}) \\ \stackrel{(89)}{=} (1-\theta_{k}) \sum_{l=0}^{k-1} \gamma_{kl}^{(i)}\psi_{i}(z_{l}^{(i)}) + (\gamma_{k+1,k}^{(i)} + \gamma_{k+1,k+1}^{(i)} - \theta_{k})\psi_{i}(z_{k}^{(i)}) + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)}) \\ \stackrel{(90)}{=} (1-\theta_{k}) \sum_{l=0}^{k} \gamma_{kl}^{(i)}\psi_{i}(z_{l}^{(i)}) + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)}) \\ \stackrel{(97)}{=} (1-\theta_{k})\alpha_{k}^{i} + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)}). \end{aligned}$$

110 / 116

イロト イポト イヨト イヨト 二日

Proof of the Main Result (Theorem 37) - Part IV

Finally,

$$\mathbf{E}_{k}[\hat{\psi}_{k+1}] \stackrel{(98)}{=} \mathbf{E}_{k}\left[\sum_{i=1}^{n} \alpha_{k+1}^{i}\right]$$

$$= \sum_{i=1}^{n} \mathbf{E}_{k}[\alpha_{k+1}^{i}]$$

$$\stackrel{(99)}{=} \sum_{i=1}^{n} (1-\theta_{k})\alpha_{k}^{i} + \theta_{k}\psi_{i}(\tilde{z}_{k+1}^{(i)})$$

$$\stackrel{(98)}{=} (1-\theta_{k})\hat{\psi}_{k} + \theta_{k}\psi(\tilde{z}_{k+1}). \quad (100)$$



Proof of the Main Result (Theorem 37) - Part V Step 3 (Recursion). For all $k \ge 0$ define:

$$\hat{F}_k \stackrel{\text{def}}{=} \hat{\psi}_k + f(x_k), \tag{101}$$

and bound the expectation of \hat{F}_{k+1} as follows:

$$\mathbf{E}_{k}[\hat{F}_{k+1}] \stackrel{(101)}{=} \mathbf{E}_{k}[\hat{\psi}_{k+1} + f(x_{k+1})] \\
\stackrel{(100)}{=} (1 - \theta_{k})\hat{\psi}_{k} + \theta_{k}\psi(\tilde{z}_{k+1}) + \mathbf{E}_{k}[f(x_{k+1})] \\
\stackrel{(96)}{\leq} (1 - \theta_{k})\hat{\psi}_{k} + (1 - \theta_{k})(f(y_{k}) + \langle \nabla f(y_{k}), x_{k} - y_{k} \rangle) \\
+ \theta_{k}(\psi(\tilde{z}_{k+1}) + f(y_{k}) + \langle \nabla f(y_{k}), \tilde{z}_{k+1} - y_{k} \rangle + \frac{\theta_{k}}{2} \|\tilde{z}_{k+1} - z_{k}\|_{p^{-1} \bullet v}^{2}) \\
\stackrel{(92)}{\leq} (1 - \theta_{k})\hat{\psi}_{k} + (1 - \theta_{k})(f(y_{k}) + \langle \nabla f(y_{k}), x_{k} - y_{k} \rangle) \\
+ \theta_{k}(\psi(y) + f(y_{k}) + \langle \nabla f(y_{k}), y - y_{k} \rangle + \frac{\theta_{k}}{2} \|y - z_{k}\|_{p^{-1} \bullet v}^{2} \\
- \frac{\theta_{k}}{2} \|y - \tilde{z}_{k+1}\|_{p^{-1} \bullet v}^{2}) \\
\leq (1 - \theta_{k})\hat{\psi}_{k} + (1 - \theta_{k})f(x_{k}) \\
+ \theta_{k}(\psi(y) + f(y) + \frac{\theta_{k}}{2} \|y - z_{k}\|_{p^{-1} \bullet v}^{2} - \frac{\theta_{k}}{2} \|y - \tilde{z}_{k+1}\|_{p^{-1} \bullet v}^{2}) \\
= (1 - \theta_{k})\hat{F}_{k} + \theta_{k}F(y) + \frac{\theta_{k}^{2}}{2} (\|y - z_{k}\|_{p^{-2} \bullet v}^{2} - \|y - \tilde{z}_{k+1}\|_{p^{-1} \bullet v}^{2}) \\
\stackrel{(??)}{=} (1 - \theta_{k})\hat{F}_{k} + \theta_{k}F(y) + \frac{\theta_{k}^{2}}{2} \mathbf{E}_{k}[\|y - z_{k}\|_{p^{-2} \bullet v}^{2} - \|y - z_{k+1}\|_{p^{-2} \bullet v}^{2}]. \end{aligned}$$
(102)

112 / 116

Proof of the Main Result (Theorem 37) - Part VI

After rearranging (102), using (87), we obtain the recursion:

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \mathbf{E}_k[\hat{F}_{k+1} - F(y)] + \frac{1}{2} \mathbf{E}_k[\|z_{k+1} - y\|_{\rho^{-2} \bullet v}^2] \le \frac{1-\theta_k}{\theta_k^2}(\hat{F}_k - F(y)) + \frac{1}{2} \|z_k - y\|_{\rho^{-2} \bullet v}^2.$$

Step 4 (Analyzing the recursion). We now take total expectation in the above inequality and unroll the recurrence:

$$\frac{1-\theta_k}{\theta_k^2} \mathbf{E}[\hat{F}_k - F(y)] + \frac{1}{2} \mathbf{E}[\|z_k - y\|_{p^{-2} \bullet v}^2] \le \frac{1-\theta_0}{\theta_0^2} (\hat{F}_0 - F(y)) + \frac{1}{2} \|z_0 - y\|_{p^{-2} \bullet v}^2.$$

Hence, for all $k \ge 1$,

$$\mathbf{E}[\hat{F}_{k} - F(y)] \leq \frac{\theta_{k-1}^{2}(1-\theta_{0})}{\theta_{0}^{2}}(\hat{F}_{0} - F(y)) + \frac{\theta_{k-1}^{2}}{2} \|x_{0} - y\|_{p^{-2} \bullet v}^{2} \\ \leq \frac{4}{((k-1)\theta_{0}+2)^{2}} ((1-\theta_{0})(F(x_{0}) - F(y)) + \frac{\theta_{0}^{2}}{2} \|x_{0} - y\|_{p^{-2} \bullet v}^{2}).$$



References I

- [1] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical Report, 2008.
- Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 22(2):341-362, 2012
- [3] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(2):1–38, 2014
- [4] Peter Richtárik and Martin Takáč. Efficient serial and parallel coordinate descent methods for huge-scale truss topology design. *Operations Research Proceedings 2011*, pp. 27-32, 2012
- [5] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. arXiv:1212.0873, 12/2012
- [6] Martin Takáč, Avleen Bijral, Peter Richtárik and Nathan Srebro. Mini-batch primal and dual methods for SVMs. *ICML*, 2013



References II

- [7] Rachael Tappenden, Peter Richtárik and Jacek Gondzio. Inexact coordinate descent: complexity and preconditioning. arXiv:1304.5530, 04/2013
- [8] Rachael Tappenden, Peter Richtárik, Burak Büke. Separable approximations and decomposition methods for the augmented Lagrangian. to appear in *Optimization Methods and Software*, arXiv:1308.6774
- [9] Olivier Fercoq and Peter Richtárik. Smooth minimization of nonsmooth functions with parallel coordinate descent methods. arXiv:1309.5885, 09/2013
- [10] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. arXiv:1310.2059, 10/2013
- [11] Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. arXiv:1310.3438, 10/2013
- [12] Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. arXiv:1312.5799, 12/2013



- [13] Olivier Fercoq, Zheng Qu, Peter Richtárik, Martin Takáč. Fast distributed coordinate descent for minimizing non-strongly convex losses. arXiv:1405.5300, 05/2014
- [14] Zheng Qu and Peter Richtárik. Accelerated parallel coordinate descent with importance sampling. Manuscript, 2014

