Coordinate Descent with Arbitrary Sampling II: Expected Separable Overapproximation

Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: http://www.tandfonline.com/loi/goms20

Coordinate descent with arbitrary sampling II: expected separable overapproximation

Zheng Qu & Peter Richtárik

To cite this article: Zheng Qu & Peter Richtárik (2016) Coordinate descent with arbitrary sampling II: expected separable overapproximation, Optimization Methods and Software, 31:5, 858-884, DOI: 10.1080/10556788.2016.1190361

To link to this article: http://dx.doi.org/10.1080/10556788.2016.1190361

Optimization Methods & Software, 2016
Vol. 31, No. 5, 858–884, http://dx.doi.org/10.1080/10556788.2016.1190361

Taylor & Francis
Taylor & Francis Group

Coordinate descent with arbitrary sampling II: expected separable overapproximation

Zheng Qu[a*] and Peter Richtárik[b]

[a]*Department of Mathematics, The University of Hong Kong, Hong Kong, Hong Kong;* [b]*School of Mathematics, The University of Edinburgh, Edinburgh, UK*

(*Received 9 January 2016; accepted 12 May 2016*)

The design and complexity analysis of randomized coordinate descent methods, and in particular of variants which update a random subset (sampling) of coordinates in each iteration, depend on the notion of expected separable overapproximation (ESO). This refers to an inequality involving the objective function and the sampling, capturing in a compact way certain smoothness properties of the function in a random subspace spanned by the sampled coordinates. ESO inequalities were previously established for special classes of samplings only, almost invariably for uniform samplings. In this paper we develop a systematic technique for deriving these inequalities for a large class of functions and for *arbitrary samplings*. We demonstrate that one can recover existing ESO results using our general approach, which is based on the study of eigenvalues associated with samplings and the data describing the function.

Keywords: coordinate descent; randomized methods; arbitrary sampling; expected separable overapproximation; parallel and distributed coordinate descent

1. Introduction

Coordinate descent methods have been popular with practitioners for many decades due to their inherent conceptual simplicity and the ease with which one can produce a working code. However, up to a few exceptions [15,33], they have been largely ignored in the optimization community until recently when a renewed interest in coordinate descent was sparked by several reports of their remarkable success in certain applications [3,23,35]. Additional and perhaps more significant reason behind the recent flurry of research activity in the area of coordinate descent comes from breakthroughs in our theoretical understanding of these methods through the introduction of *randomization* in the iterative process [5–8,10–14,16,17,19,20,22,24,25,27,28,30–32]. Traditional variants of coordinate descent rely on cyclic or greedy rules for the selection of the next coordinate to be updated. The existing worst–case complexity bounds of the cyclic or greedy rules are in general weaker than that obtained for randomized methods [1,34], except in certain special regimes when the data matrix is very sparse [4,18].

*Corresponding author. Email: zhengqu@maths.hku.hk

1.1 *Expected separable overapproximation*

It has recently become increasingly clear that the design and complexity analysis of randomized coordinate descent methods is intimately linked with and can be better understood through the notion of *expected separable overapproximation (ESO)* [5,7,8,19,20,24,26,27,30,31]. This refers to an inequality involving the objective function and the sampling (a random set-valued mapping describing the law with which subsets of coordinates are selected at each iteration), capturing in a compact way certain smoothness properties of the function in a random subspace spanned by the sampled coordinates.

A (coordinate) sampling $\hat{S}$ is a random set-valued mapping with values being subsets of $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$. It will be useful to write

$$p_i \stackrel{\text{def}}{=} \mathbb{P}(i \in \hat{S}), \quad i \in [n]. \tag{1}$$

DEFINITION 1.1 (ESO) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\hat{S}$ a sampling. We say that f admits an ESO with respect to sampling $\hat{S}$ with parameters $v = (v_1, \ldots, v_n) > 0$ if the following inequality holds*[1] *for all $x, h \in \mathbb{R}^n$:*

$$\mathbb{E}\left[f\left(x + \sum_{i \in \hat{S}} h_i e_i\right)\right] \leq f(x) + \sum_{i=1}^{n} p_i (\nabla_i f(x))^\top h_i + \frac{1}{2} \sum_{i=1}^{n} p_i v_i h_i^2. \tag{2}$$

We will compactly write $(f, \hat{S}) \sim \text{ESO}(v)$.

In this definition, e_i is the ith unit coordinate vector in $\mathbb{R}^n$ and $\nabla_i f(x) = (\nabla f(x))^\top e_i$ is the ith partial derivative of f at x. In the context of *block* coordinate descent, the above definition refers to the case when all blocks correspond to coordinates. For the simplicity of the exposition, we focus on this case. However, all our results can be extended to the more general block set-up.

Instead of the above general definition, it will be useful for the reader to instead think about the form of this inequality in the simple case when $f(x) = \|Ax\|^2$, where $\|\cdot\|$ is the L^2-norm, and $x = 0$. Letting $A = [A_1, \ldots, A_n]$, in this case inequality (2) takes the form

$$\mathbb{E}\left[\left\|\sum_{i \in \hat{S}} A_i h_i\right\|^2\right] \leq h^\top \text{Diag}(p \circ v) h,$$

where $p \circ v$ denotes the Hadamard product of vectors $p = (p_1, \ldots, p_n)$ and $v = (v_1, \ldots, v_n)$; that is, $p \circ v = (p_1 v_1, \ldots, p_n v_n) \in \mathbb{R}^n$, and $\text{Diag}(p \circ v)$ is the n-by-n diagonal matrix with vector $p \circ v$ on the diagonal. The term on the left-hand side is a convex quadratic function of h, and so is the term on the right-hand side—however, the latter function has a diagonal Hessian. Hence, for quadratics, finding the smallest ESO parameter v reduces to an eigenvalue problem.

The ESO inequality is of key importance for randomized coordinate descent methods for several reasons:

- The parameters $v = (v_1, \ldots, v_n)$ for which ESO holds are needed[2] to run coordinate descent. Indeed, they are used to set the stepsizes to a suitable value.
- The size of these parameters directly influences the complexity of the method (see Table 1).
- There are problems for which updating more coordinates in each iteration, as opposed to updating just one, may *not* lead to fewer iterations [27] (which suggests that perhaps the resources should be instead utilized in some other way). Whether this happens or not can

Table 1. Complexity of randomized coordinate descent methods which were analysed for an arbitrary sampling (λ is a strong convexity constant, x_0 is the starting point and x_* is the optimal point.

Set-up	Complexity	Method / Paper / Year
Strongly convexsmooth	$\max_i(\frac{v_i}{p_i\lambda}) \times \log(\frac{1}{\epsilon})$	NSync [26], 10/2013
Strongly convex nonsmooth (primal–dual)	$\max_i(\frac{1}{p_i} + \frac{v_i}{p_i\lambda n}) \times \log(\frac{1}{\epsilon})$	QUARTZ [20], 11/2014
Convex smooth	$\sqrt{2\sum_{i=1}^n \frac{v_i(x_i^0 - x_i^*)^2}{p_i^2}} \times \frac{1}{\sqrt{\epsilon}}$	ALPHA [19], 12/2014

be understood through a careful study of the complexity result and its dependence, through the vectors p and v, on the number of coordinates updated in each iteration [5,8,20,24,27,30].

- The ESO assumption is *generic* in the sense that as soon as function f and sampling $\hat{S}$ satisfy it, the complexity result follows. This leads to a natural dichotomy in the study of coordinate descent: (i) the search for new variants of coordinate descent (e.g. parallel, accelerated, distributed) and study of their complexity under the ESO assumption, and (ii) the search for pairs $(f, \hat{S})$ for which one can compute v such that $(f, \hat{S}) \sim \text{ESO}(v)$. Our current study follows this dichotomy: in [19] we deal with the algorithmic and complexity aspects, and in this paper we deal with the ESO aspect.

1.2 *Complexity of coordinate descent*

As mentioned above, complexity of coordinate descent methods depends in a crucial way on the optimization problem, sampling employed, and on the ESO parameters $v = (v_1, \dots, v_n)$. In Table 1 we summarize all known complexity results[3] which hold for an *arbitrary sampling*. Note that in all cases, vectors p and v appear in the complexity bound. The bounds are not directly comparable as they apply to different optimization problems.

For instance, the NSync bound[4] in Table 1 applies to the problem of unconstrained minimization of a smooth strongly convex function. It was in [26] where the general form of the ESO inequality used in this paper was first mentioned and used to derive a complexity result for a coordinate descent method with arbitrary sampling.

The Quartz algorithm [20], on the other hand, applies to a much more serious problem—a problem of key importance in machine learning. In particular, it applies to the regularized *empirical risk minimization* problem, where the loss functions are convex and have Lipschitz gradients and the regularizer is strongly convex and possibly nonsmooth. Coordinate ascent is applied to the dual of this problem, and the bound appearing in Table 1 applies to the duality gap.[5]

The APPROX method was first proposed in [7] and then generalized to an arbitrary sampling (among other things) in [19]. In its accelerated variant, it enjoys a $O(1/\sqrt{\epsilon})$ rate, whereas its non-accelerated variant has a slower $O(1/\epsilon)$ rate. Again, the complexity of the method explicitly depends on the vector of probabilities p and the ESO parameter v.

1.3 *Historical remarks*

The ESO relation (2) was first introduced by Richtárik and Takáč [27] in the special case of *uniform samplings*, i.e. samplings for which $\mathbb{P}(i \in \hat{S}) = \mathbb{P}(j \in \hat{S})$ for all coordinates $i, j \in \{1, 2, \dots, n\}$. The uniformity condition is satisfied for a large variety of samplings, we refer the reader to [27] for a basic classification of uniform samplings (including overlapping,

non-overlapping, doubly uniform, binomial, nice, serial and parallel samplings) and to [8,20,24] for further examples (e.g. 'distributed sampling'). The study of non-uniform samplings has until recently been confined to *serial* sampling only, i.e. to samplings which only pick a single coordinate at a time. In [26] the authors propose a particular example of a *parallel non-uniform* sampling, where 'parallel' refers to samplings for which $\mathbb{P}(|\hat{S}| > 1) > 0$, and 'non-uniform' simply means not uniform. Furthermore, they derive an ESO inequality for their sampling and a partially separable function. The proposed sampling is easy to generate (note that in general a sampling is described by assigning distinct probabilities to all 2^n subsets of $[n]$, and hence most samplings will necessarily be hard to generate), and leads to strong ESO bounds which predict nearly linear speedup for NSync for sparse problems. A further example of a non-uniform sampling was given in [20]—the so-called 'product sampling'—and an associated ESO inequality derived. Intuitively speaking, this sampling samples sets of 'independent' coordinates, which leads to complexity scaling linearly with the size of the sampled sets. To the best of our knowledge, this is the state of the art—no further non-uniform samplings were proposed nor associated ESO inequalities derived.

1.4 *Contributions*

We now briefly list the contributions of this work.

(1) ESO inequalities were previously established for special classes of samplings only, almost invariably for *uniform samplings* [5,7,8,24,27], and often using seemingly disparate approaches. We give the first *systematic study of ESO inequalities* for *arbitrary* samplings.
(2) We recover existing ESO results by applying our general technique.
(3) Our approach to deriving ESO inequalities is via the study of random principal submatrices of a positive semidefinite matrix. In particular, we give bounds on the largest eigenvalue of the mean of the random submatrix. This may be of independent interest.

1.5 *Outline of the paper*

Our paper is organized as follows. In Section 2 we describe the class of functions (f), we consider in this paper and briefly establish some basic terminology related to samplings ($\hat{S}$). In Section 3 we study probability matrices associated with samplings ($\mathbf{P}(\hat{S})$), in Section 4 we study eigenvalues of these probability matrices ($\lambda(\mathbf{P}(\hat{S}))$ and $\lambda'(\mathbf{P}(\hat{S}))$) and in Section 5 we design a general technique for computing parameter $v = (v_1, \dots, v_n)$ for which the ESO inequality holds (i.e. for which $(f, \hat{S}) \sim \mathrm{ESO}(v)$). We illustrate the use of these techniques in Section 5.4 and conclude with Section 7. We provide in (Appendix) a list of frequently used notations.

1.6 *Notations*

For two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size, we denote by $\mathbf{M}_1 \circ \mathbf{M}_2$ their Hadamard (i.e. elementwise) product. We use the same notation for Hadamard product of vectors. For arbitrary matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $S \subseteq [n]$, we will use the notation $\mathbf{M}_{[S]}$ for the n-by-n matrix obtained from $\mathbf{M}$ by retaining elements $\mathbf{M}_{ij}$ for which both $i \in S$ and $j \in S$ and zeroing out all other elements. For a vector $v \in \mathbb{R}^n$, we denote by $\mathrm{Diag}(v)$ the diagonal matrix with v on the diagonal. For an n-by-n matrix $\mathbf{M}$, $\mathrm{Diag}(\mathbf{M})$ denotes the diagonal matrix containing the diagonal of $\mathbf{M}$. By $\mathbf{I}$ we denote the n-by-n identity matrix and by $\mathbf{E}$, we denote the n-by-n matrix of all ones. For any

$h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ and $S \subseteq [n]$, we will write

$$h_{[S]} \stackrel{\text{def}}{=} \sum_{i \in S} h_i e_i = \mathbf{I}_{[S]} h, \tag{3}$$

where $e_1, \ldots, e_n$ are the standard basis vectors in $\mathbb{R}^n$. Also note that

$$\mathbf{M}_{[S]} = \mathbf{E}_{[S]} \circ \mathbf{M} = \mathbf{I}_{[S]} \mathbf{M} \mathbf{I}_{[S]}. \tag{4}$$

2. Functions and samplings

Recall that in the paper we are concerned with establishing inequality (2) which we succinctly write as $(f, \hat{S}) \sim ESO(v)$. In Section 2.1 we describe the class of functions f, we consider in this paper and in Section 2.2, we briefly review several elementary facts related to samplings.

2.1 *Functions*

We assume in this paper that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and that it satisfies the following assumption.

ASSUMPTION 2.1 *There is an m-by-n matrix* $\mathbf{A}$ *such that for all* $x, h \in \mathbb{R}^n$,

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \tfrac{1}{2} h^\top \mathbf{A}^\top \mathbf{A} h. \tag{5}$$

In the subsequent text, we shall often refer to the set of columns of $\mathbf{A}$ for which the entry in the jth row of $\mathbf{A}$ is nonzero:

$$J_j \stackrel{\text{def}}{=} \{i \in [n] : \mathbf{A}_{ji} \neq 0\}. \tag{6}$$

Assumption 2.1 holds for many functions of interest in optimization and machine learning. Coordinate descent methods for functions f explicitly required to satisfy Assumption 2.1 were studied in [2,8,24].

The following simple observation will help us relate the above assumption with standing assumptions considered in various papers on randomized coordinate descent methods.

PROPOSITION 2.1 *Assume f is of the form*

$$f(x) = \sum_{j=1}^{s} \phi_j(\mathbf{M}_j x), \tag{7}$$

where for each j, $\mathbf{M}_j \in \mathbb{R}^{d \times n}$ *and function* $\phi_j : \mathbb{R}^d \to \mathbb{R}$ *has* γ_j*-Lipschitz continuous gradient* (*with respect to the L2 norm*). *Then f satisfies Assumption* 2.1 *for matrix* $\mathbf{A}$ *given by*

$$\mathbf{A}^\top \mathbf{A} = \sum_{j=1}^{s} \gamma_j \mathbf{M}_j^\top \mathbf{M}_j.$$

Proof Pick $x, h \in \mathbb{R}^n$ and let $f_j(x) \stackrel{\text{def}}{=} \phi_j(\mathbf{M}_j x)$. Then since ϕ_j is γ_j-smooth, we have

$$\begin{aligned} f_j(x+h) = \phi_j(\mathbf{M}_j x + \mathbf{M}_j h) &\le \phi_j(\mathbf{M}_j x) + \langle \nabla \phi_j(\mathbf{M}_j x), \mathbf{M}_j h\rangle + \frac{\gamma_j}{2}\|\mathbf{M}_j h\|^2 \\ &= f_j(x) + \langle \nabla f_j(x), h\rangle + \frac{\gamma_j}{2} h^{\mathrm{T}} \mathbf{M}_j^{\top} \mathbf{M}_j h. \end{aligned}$$

It remains to add these inequalities for $j = 1, \ldots, s$. ■

We now apply Proposition 2.1 to several special cases:

(1) *Partial separability.* Let $d = n$ and $\mathbf{M}_j = \mathbf{I}_{[C_j]}$, where for each j, $C_j \subseteq [n]$. Then f is of the form

$$f(x) = \sum_{j=1}^{s} \phi_j(\mathbf{I}_{[C_j]} x). \tag{8}$$

That is, ϕ_j depends on coordinates of x belonging to set C_j only. By Proposition 2.1, f satisfies (5), where $\mathbf{A}$ is the n-by-n diagonal matrix given by

$$\mathbf{A}_{ii} = \sqrt{\sum_{j: i \in C_j} \gamma_j}, \quad i \in [n].$$

Functions of the form (8) (i.e. partially separable functions) were considered in the context of *parallel coordinate descent methods* in [27]. However, in [27] the authors only assume the *sum* f to have a Lipschitz gradient (which is more general, but somewhat complicates the analysis), whereas we assume that all component functions $\{\phi_j\}_j$ have Lipschitz gradient.

(2) *Linear transformation of variables.* Let $s = 1$. Then f is of the form

$$f(x) = \phi_1(\mathbf{M}_1 x). \tag{9}$$

By Proposition 2.1, f satisfies (5), where $\mathbf{A}$ is given by

$$\mathbf{A} = \sqrt{\gamma_1}\mathbf{M}_1.$$

A function of the form (9) appears in the *dual problem* of the standard primal–dual formulation to which *stochastic dual coordinate ascent* methods are applied [11,20,28,29,36].

(3) *Sum of scalar functions depending on x through an inner product.* Let $d = 1$ and $\mathbf{M}_j = e_j^{\mathrm{T}}\mathbf{M}$, where $\mathbf{M} \in \mathbb{R}^{m \times n}$ and e_j is the jth unit coordinate vector in $\mathbb{R}^m$. Then f is of the form

$$f(x) = \sum_{j=1}^{m} \phi_j(e_j^{\top}\mathbf{M}x). \tag{10}$$

By Proposition 2.1, f satisfies (5), with $\mathbf{A}$ given by

$$\mathbf{A} = \mathrm{Diag}(\sqrt{\gamma_1}, \ldots, \sqrt{\gamma_m})\mathbf{M}.$$

Functions of the form (10) play an important role in the design of *efficiently implementable accelerated coordinate descent methods* [7,19]. These functions also appear in the *primal problem* of the standard primal–dual formulation to which *stochastic dual coordinate ascent* methods are applied.

2.2 *Samplings*

As defined in the introduction, by sampling we mean a random set-valued mapping with values in $2^{[n]}$ (the set of subsets of $[n]$).

Classification of samplings. Following the terminology established in [27], we say that sampling $\hat{S}$ is *proper* if $p_i = \mathbb{P}(i \in \hat{S}) > 0$ for all $i \in [n]$. We shall focus our attention on proper samplings as otherwise there is a coordinate which is never chosen (and hence never updated by the coordinate descent method). We say that $\hat{S}$ is *nil* if $\mathbb{P}(\hat{S} = \emptyset) = 1$.

Of key importance in this paper are *elementary samplings*, defined next.

DEFINITION 2.1 (Elementary samplings) *Elementary sampling associated with $S \subseteq [n]$ is a sampling which selects set S with probability one. We will denote it by $\hat{E}_S$: $\mathbb{P}(\hat{E}_S = S) = 1$.*

We say that the sampling is *uniform* if $\mathbb{P}(i \in \hat{S}) = \mathbb{P}(j \in \hat{S})$ for all $i, j \in [n]$. The class of uniform samplings is large, for examples (and properties) of notable subclasses, we refer the reader to [24,27].

We say that sampling $\hat{S}$ is *doubly uniform* if it satisfies the following condition: if $|S_1| = |S_2|$, then $\mathbb{P}(\hat{S} = S_1) = \mathbb{P}(\hat{S} = S_2)$. Necessarily, every doubly uniform sampling is uniform [27]. The definition postulates an additional 'uniformity' property ('equal cardinality implies equal probability'), whence the name. As described in [27], doubly uniform samplings are special in the sense that 'good' ESO results can be proved for them. A notable subclass of the class of doubly uniform samplings are the τ-nice samplings for $1 \leq \tau \leq n$. The τ-nice sampling is obtained by picking (all) subsets of cardinality τ, uniformly at random (we give a precise definition below). This sampling is by far the most common in stochastic optimization, and refers to standard mini-batching. The τ-nice sampling arises as a special case of the (c, τ)-distributed sampling (which, as its name suggests, can be used to design distributed variants of coordinate descent [8,24]), which we define next:

DEFINITION 2.2 ((c, τ)-distributed sampling; [8,20,24]) *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*

$$\hat{S} \overset{\text{def}}{=} \bigcup_{l=1}^{c} \hat{S}_l, \tag{11}$$

is called (c, τ)-distributed sampling.

The τ-nice sampling arises as a special case of the (c, τ)-distributed sampling (for $c = 1$) which we define next.

DEFINITION 2.3 (τ-nice sampling; [5,7,27,30,31]) *Sampling $\hat{S}$ from $[n]$ is called τ-nice if it picks only subsets of $[n]$ of cardinality τ, uniformly at random. More formally, it is defined by*

$$\mathbb{P}(\hat{S} = S) = \begin{cases} 1/\binom{n}{\tau}, & |S| = \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Operations with samplings. We now define several basic operations with samplings (convex combination, intersection and restriction).

DEFINITION 2.4 (Convex combination of samplings; [27]) *Let $\hat{S}_1, \dots, \hat{S}_k$ be samplings and let $q_1, \dots, q_k$ be non-negative scalars summing to 1. By $\sum_{t=1}^k q_t \hat{S}_t$ we denote the sampling obtained as follows: we first pick $t \in \{1, \dots, k\}$, with probability q_t, and then sample according to $\hat{S}_t$. More formally, $\hat{S}$ is defined as follows:*

$$\mathbb{P}(\hat{S} = S) = \sum_{t=1}^{k} q_t \mathbb{P}(\hat{S}_t = S), \quad S \subseteq [n]. \tag{13}$$

Note that (13) indeed defines a sampling, since

$$\sum_{S \subseteq [n]} \mathbb{P}(\hat{S} = S) = \sum_{S \subseteq [n]} \sum_{t=1}^{k} q_t \mathbb{P}(\hat{S}_t = S) = \sum_{t=1}^{k} q_t \sum_{S \subseteq [n]} \mathbb{P}(\hat{S}_t = S) = \sum_{t=1}^{k} q_t = 1.$$

Each sampling is a convex combination of elementary samplings. Indeed, for each $\hat{S}$, we have

$$\hat{S} = \sum_{S \subseteq [n]} \mathbb{P}(\hat{S} = S) \hat{E}_S. \tag{14}$$

We now show that each doubly uniform sampling arises as a convex combination of τ-nice samplings.

PROPOSITION 2.2 *Let $\hat{S}$ be a doubly uniform sampling and let $\hat{S}_\tau$ be the τ-nice sampling, for $\tau = 0, 1, \dots, n$. Then*

$$\hat{S} = \sum_{\tau=0}^{n} \mathbb{P}(|\hat{S}| = \tau) \hat{S}_\tau.$$

Proof Fix any $S \subseteq [n]$ and let $q_\tau = \mathbb{P}(|\hat{S}| = \tau)$. Note that

$$\mathbb{P}(\hat{S} = S) = \sum_{\tau=0}^{n} \mathbb{P}(\hat{S} = S \ \& \ |\hat{S}| = \tau) = \sum_{\tau=0}^{n} q_\tau \mathbb{P}(\hat{S} = S \,|\, |\hat{S}| = \tau) = \sum_{\tau=0}^{n} q_\tau \mathbb{P}(\hat{S}_\tau = S),$$

where the last equality follows from the definition of doubly uniform and τ-nice samplings. The statement then follows from (13) (i.e. by definition of convex combination of samplings). ■

It will be useful to define two more operations with samplings; intersection and restriction.

DEFINITION 2.5 (Intersection of samplings) *For two samplings $\hat{S}_1$ and $\hat{S}_2$, we define the intersection $\hat{S} \overset{\text{def}}{=} \hat{S}_1 \cap \hat{S}_2$ as the sampling for which:*

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

DEFINITION 2.6 (Restriction of a sampling) *Let $\hat{S}$ be a sampling and $J \subseteq [n]$. By restriction of $\hat{S}$ to J, we mean the sampling $\hat{E}_J \cap \hat{S}$. By abuse of notation, we will also write this sampling as $J \cap \hat{S}$.*

Graph sampling. Let $G = (V, E)$ be an undirected graph with $|V| = n$ vertex and (i, i') be an edge in E if and only if there is $j \in [m]$ such that $\{i, i'\} \subseteq J_j$. If S is an independent set of graph G, then necessarily

$$\max_{j \in [m]} |J_j \cap S| = 1.$$

Denote by $\mathcal{T}$ the collection of all independent sets of the graph G. We now define the graph sampling as follows:

DEFINITION 2.7 (Graph sampling) *Graph sampling associated with graph G is any sampling $\hat{S}$ for which $\mathbb{P}(\hat{S} = S) = 0$ if $S \notin \mathcal{T}$. In other words, a graph sampling can only assign positive weights to independent sets of G.*

Let $\hat{S}$ be a graph sampling. In view of (14), for some non-negative constants q_S adding up to 1:

$$\hat{S} = \sum_{S \in \mathcal{T}} q_S \hat{E}_S.$$

Note that, necessarily, $q_S = \mathbb{P}(\hat{S} = S)$ for all $S \in \mathcal{T}$.

DEFINITION 2.8 (Product sampling) *Let $X_1, \ldots, X_\tau$ be a partition of $[n]$, i.e.*

$$X_1 \cup \cdots X_\tau = [n]; \quad X_i \cap X_j = \emptyset, \quad \forall 1 \leq i < j \leq n.$$

Define:

$$\mathcal{S} \overset{\text{def}}{=} X_1 \times \cdots \times X_\tau.$$

The product sampling $\hat{S}$ is obtained by choosing $S \in \mathcal{S}$, uniformly at random; that is, via:

$$\mathbb{P}(\hat{S} = S) = \frac{1}{|\mathcal{S}|} = \frac{1}{\prod_{l=1}^{\tau} |X_l|}, \quad S \in \mathcal{S}. \tag{15}$$

A similar sampling was first considered in [20, Section 3.3] with an additional *group separability assumption* on the partition $X_1, \ldots, X_\tau$, which can be equivalently stated as

$$\max_{j \in m} |J_j \cap S| = 1, \quad \forall S \in \mathcal{S}.$$

In other words, it is both a product sampling and graph sampling. Note that in Definition 2.8 we do not make any assumption on the partition. Also, the product sampling is a non-uniform sampling as long as all the sets X_l do not have the same cardinality, which occurs necessarily if τ, representing the number of processors, is not divisible by n.

3. Probability matrix associated with a sampling

In this section we define the notion of *probability matrix* associated with a sampling. As we shall see in later sections, this matrix encodes all information about $\hat{S}$ which is relevant for development of ESO inequality.

DEFINITION 3.1 (Probability matrix) *With each sampling $\hat{S}$, we associate an n-by-n 'probability matrix' $\mathbf{P} = \mathbf{P}(\hat{S})$ defined by*

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

We shall write $\mathbf{P}(\hat{S})$ when it is important to indicate which sampling is behind the probability matrix, otherwise we simply write $\mathbf{P}$. Probability matrices of elementary samplings are given by

$$\mathbf{P}(\hat{E}_S) = \mathbf{E}_{[S]} = e_{[S]} e_{[S]}^\top, \tag{16}$$

where $e \in \mathbb{R}^n$ is the vector of all ones. In particular, the matrix is rank-one and positive semidefinite.

3.1 *Representation of probability matrices*

We now establish a simple but particularly insightful result, leading to many useful identities.

THEOREM 3.1 *For each sampling $\hat{S}$, we have*

$$\mathbf{P}(\hat{S}) = \mathbb{E}[\mathbf{E}_{[\hat{S}]}] = \sum_{S \subseteq [n]} \mathbb{P}(\hat{S} = S)\mathbf{E}_{[S]}. \tag{17}$$

In particular:

(i) *The set of probability matrices is the convex hull of the probability matrices corresponding to elementary samplings.*
(ii) $\mathbf{P}(\hat{S}) \succeq 0$ *for each* $\hat{S}$.

Proof The (i,j) element of the matrix on the right-hand side is $\mathbb{E}[(\mathbf{E}_{[\hat{S}]})_{ij}]$. Since $(\mathbf{E}_{[\hat{S}]})_{ij} = 1$ if $\{i,j\} \subseteq \hat{S}$ and $(\mathbf{E}_{[\hat{S}]})_{ij} = 0$ otherwise, we have $\mathbb{E}[(\mathbf{E}_{[\hat{S}]})_{ij}] = \mathbb{P}(\{i,j\} \subseteq \hat{S}) = (\mathbf{P}(\hat{S}))_{ij}$. Claim (i) follows from (17) since $\mathbf{E}_{[S]} = \mathbf{P}(\hat{E}_S)$. Claim (ii) follows from (17) since $\mathbf{E}_{[S]} \succeq 0$ for all $S \subseteq [n]$. ■

We have the following useful corollary:[6]

COROLLARY 3.1 *Let $\hat{S}$ be any sampling, $\mathbf{P} = \mathbf{P}(\hat{S})$, $\mathbf{M} \in \mathbb{R}^{n \times n}$ be an arbitrary matrix and $h \in \mathbb{R}^n$. Then the following identities hold:*

$$\mathbf{P} \circ \mathbf{M} = \mathbb{E}[\mathbf{M}_{[\hat{S}]}], \tag{18}$$

$$h^\top(\mathbf{P} \circ \mathbf{M})h = \mathbb{E}[h^\top \mathbf{M}_{[\hat{S}]}h] = \mathbb{E}[h_{[\hat{S}]}^\top \mathbf{M} h_{[\hat{S}]}], \tag{19}$$

$$h^\top \mathbf{P} h = \mathbb{E}\left[\left(\sum_{i \in \hat{S}} h_i\right)^2\right], \tag{20}$$

$$\sum_{i=1}^n \mathbf{P}_{ii} h_i = \mathbb{E}\left[\sum_{i \in \hat{S}} h_i\right], \tag{21}$$

$$\mathrm{e}^\mathrm{T} \mathbf{P} e = \mathbb{E}[|\hat{S}|^2], \tag{22}$$

$$\mathrm{Tr}(\mathbf{P}) = \mathbb{E}[|\hat{S}|]. \tag{23}$$

Proof Since multiplying a matrix in the Hadamard sense by a fixed matrix is a linear operation,

$$\mathbf{P} \circ \mathbf{M} \overset{(17)}{=} \mathbb{E}[\mathbf{E}_{[\hat{S}]}] \circ \mathbf{M} = \mathbb{E}[\mathbf{E}_{[\hat{S}]} \circ \mathbf{M}] \overset{(4)}{=} \mathbb{E}[\mathbf{M}_{[\hat{S}]}].$$

Next, identity (19) follows from (18):

$$h^\top(\mathbf{P} \circ \mathbf{M})h = h^\top \mathbb{E}[\mathbf{M}_{[\hat{S}]}]h = \mathbb{E}[h^\top \mathbf{M}_{[\hat{S}]}h] \overset{(4)}{=} \mathbb{E}[h^\top \mathbf{I}_{[\hat{S}]} \mathbf{M} \mathbf{I}_{[\hat{S}]} h] \overset{(3)}{=} \mathbb{E}[h_{[\hat{S}]}^\top \mathbf{M} h_{[\hat{S}]}].$$

Identity (20) follows from (19) by setting $\mathbf{M} = \mathbf{E}$:

$$\mathbb{E}[h_{[\hat{S}]}^\top \mathbf{E} h_{[\hat{S}]}] = \mathbb{E}\left[\sum_{i,j \in \hat{S}} h_i h_j\right] = \mathbb{E}\left[\left(\sum_{i \in \hat{S}} h_i\right)^2\right].$$

Identity (21) holds since

$$\sum_i \mathbf{P}_{ii}h_i = \sum_i \sum_{S:i\in S} \mathbb{P}(\hat{S}=S)h_i = \sum_{S\subseteq[n]} \mathbb{P}(\hat{S}=S)\sum_{i\in S} h_i = \mathbb{E}\left[\sum_{i\in\hat{S}} h_i\right].$$

Finally, (22) (resp. (23)) follows from (20) (resp. (21)) by setting $h=e$. ■

If $\hat{S}$ is a uniform sampling (i.e. if $\mathbb{P}(i\in\hat{S}) = \mathbb{P}(j\in\hat{S})$ for all $i,j\in[n]$), then from (23) we deduce that for all $i\in[n]$:

$$p_i \equiv \mathbb{P}(i\in\hat{S}) \equiv \mathbf{P}_{ii} = \frac{\mathbb{E}[|\hat{S}|]}{n}. \tag{24}$$

3.2 *Operations with samplings*

We now give formulae for the probability matrix of the sampling arising as a convex combination, intersection or a restriction, in terms of the probability matrices of the constituent samplings.

Convex combination of samplings. We have seen in (14) that each sampling is a convex combination of elementary samplings. In view of Theorem 3.1, the probability matrices of the samplings are related the same way:

$$\mathbf{P}(\hat{S}) = \sum_{S\subseteq[n]} \mathbb{P}(\hat{S}=S)\mathbf{P}(\hat{E}_S). \tag{25}$$

More generally, as formalized in the following lemma, the probability matrix of a convex combination of samplings is equal to the convex combination of the probability matrices of these samplings.

LEMMA 3.1 *Let $\hat{S}_1,\dots,\hat{S}_k$ be samplings and $q_1,\dots,q_k$ be non-negative scalars summing up to 1. Then*

$$\mathbf{P}\left(\sum_{t=1}^k q_t\hat{S}_t\right) = \sum_{t=1}^k q_t\mathbf{P}(\hat{S}_t). \tag{26}$$

Proof Let $\hat{S}$ be the convex combination of samplings $\hat{S}_1,\dots,\hat{S}_k$ and fix any $i,j\in[n]$. By definition,

$$\begin{aligned}(\mathbf{P}(\hat{S}))_{ij} = \mathbb{P}(\{i,j\}\subseteq\hat{S}) &= \sum_{S\subseteq[n]:\{i,j\}\subseteq S} \mathbb{P}(\hat{S}=S)\\ &\overset{(13)}{=} \sum_{S\subseteq[n]:\{i,j\}\subseteq S}\sum_{t=1}^k q_t\mathbb{P}(\hat{S}_t=S) = \sum_{t=1}^k q_t \sum_{S\subseteq[n]:\{i,j\}\subseteq S} \mathbb{P}(\hat{S}_t=S)\\ &= \sum_{t=1}^k q_t\mathbb{P}(\{i,j\}\subseteq\hat{S}_t) = \sum_{t=1}^k q_t(\mathbf{P}(\hat{S}_t))_{ij} = \left(\sum_{t=1}^k q_t\mathbf{P}(\hat{S}_t)\right)_{ij}.\end{aligned}$$

■

Intersection of samplings. The probability matrix of the intersection of two independent samplings is equal to the Hadamard product of the probability matrices of these samplings. This is formalized in the following lemma.

LEMMA 3.2 *Let $\hat{S}_1, \hat{S}_2$ be independent samplings. Then*

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

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

Restriction. By Lemma 3.2, the probability matrix of the restriction of arbitrary sampling $\hat{S}$ to $J \subseteq [n]$ is given by (we give several alternative ways of writing the result):

$$\mathbf{P}(J \cap \hat{S}) = \mathbf{P}(\hat{E}_J) \circ \mathbf{P}(\hat{S}) \overset{(16)}{=} \mathbf{E}_{[J]} \circ \mathbf{P}(\hat{S}) = \mathbf{I}_{[J]}\mathbf{P}(\hat{S})\mathbf{I}_{[J]}. \tag{27}$$

Note that $\mathbf{P}(J \cap \hat{S})$ is the matrix obtained from $\mathbf{P}(\hat{S})$ by keeping only elements $i, j \in J$ and zeroing out all the rest. Furthermore, by combining the formulae derived above, we get

$$\mathbf{P}\left(J \cap \sum_{t=1}^{k} q_t \hat{S}_t\right) \overset{(27)+(26)}{=} \mathbf{E}_{[J]} \circ \left(\sum_{t=1}^{k} q_t \mathbf{P}(\hat{S}_t)\right) = \sum_{t=1}^{k} q_t (\mathbf{E}_{[J]} \circ \mathbf{P}(\hat{S}_t)) \overset{(27)}{=} \sum_{t=1}^{k} q_t \mathbf{P}(J \cap \hat{S}_t). \tag{28}$$

3.3 *Probability matrix of special samplings*

The probability matrix of the (c, τ)-distributed samplings is computed in the following lemma.

LEMMA 3.3 *Let $\hat{S}$ be the (c, τ)-distributed sampling associated with the partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_c\}$ of $[n]$ such that $s = |\mathcal{P}_l|$ for $l \in [c]$ (see Definition 2.2). Then*

$$\mathbf{P}(\hat{S}) = \frac{\tau}{s}[\alpha_1 \mathbf{I} + \alpha_2 \mathbf{E} + \alpha_3(\mathbf{E} - \mathbf{B})], \tag{29}$$

where

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

$s_1 = \max(s - 1, 1)$ *and*

$$\mathbf{B} = \sum_{l=1}^{c} \mathbf{P}(\hat{E}_{\mathcal{P}_l}). \tag{30}$$

Note that $\mathbf{B}$ *is the 0–1 matrix with* $\mathbf{B}_{ij} = 1$ *if and only if i,j belong to the same partition.*

Proof Let $\mathbf{P} = \mathbf{P}(\hat{S})$. It is easy to see that

$$\mathbf{P}_{ij} = \begin{cases} \dfrac{\tau}{s} & \text{if } i = j, \\ \dfrac{\tau(\tau - 1)}{s s_1} & \text{if } i \neq j \quad \text{and} \quad i, j \in \mathcal{P}_l \text{ for some } l \in [c], \\ \dfrac{\tau^2}{s^2} & \text{otherwise.} \end{cases}$$

Hence,

$$\mathbf{P} = \frac{\tau}{s}\mathbf{I} + \frac{\tau(\tau - 1)}{s s_1}(\mathbf{B} - \mathbf{I}) + \frac{\tau^2}{s^2}(\mathbf{E} - \mathbf{B}) = \frac{\tau}{s}[\alpha_1 \mathbf{I} + \alpha_2 \mathbf{E} + \alpha_3(\mathbf{E} - \mathbf{B})].$$

■

As a corollary of the above in the $c=1$ case, we obtain the probability matrix of the τ-nice sampling:

LEMMA 3.4 *Fix $1 \le \tau \le n$ and let $\hat{S}$ be the τ-nice sampling. Then*

$$\mathbf{P}(\hat{S}) = \frac{\tau}{n}((1-\beta)\mathbf{I} + \beta\mathbf{E}), \tag{31}$$

where $\beta = (\tau-1)/\max(n-1,1)$. If $\tau = 0$, then $\mathbf{P}(\hat{S})$ is the zero matrix.

Proof For $\tau \ge 1$, this follows from Lemma 3.3 in the special case when $c=1$ (note that $\mathcal{P}_1 = [n], s = n$ and $\mathbf{B} = \mathbf{E}$). ■

Finally, we compute the probability matrix of a doubly uniform sampling.

LEMMA 3.5 *Let $\hat{S}$ be a doubly uniform sampling and assume it is not nil* (*i.e. assume that* $\mathbb{P}(\hat{S} = \emptyset) \ne 1$). *Then*

$$\mathbf{P}(\hat{S}) = \frac{\mathbb{E}[|\hat{S}|]}{n}((1-\beta)\mathbf{I} + \beta\mathbf{E}), \tag{32}$$

where

$$\beta = \left(\frac{\mathbb{E}[|\hat{S}|^2]}{\mathbb{E}[|\hat{S}|]} - 1\right) / \max(n-1,1). \tag{33}$$

Proof Letting $q_\tau = \mathbb{P}(|\hat{S}| = \tau)$, by Proposition 2.2 we can write $\hat{S} = \sum_{\tau=0}^{n} q_\tau \hat{S}_\tau$, where $\hat{S}_\tau$ is the τ-nice sampling. It only remains to combine Lemmas 3.1 and 3.4 and rearrange the result. ■

Note that Lemma 3.4 is a special case of Lemma 3.5 (covering the case when $\mathbb{P}(|\hat{S}| = \tau) = 1$ for some τ).

4. Largest eigenvalues of the probability matrix

For an $n \times n$ positive semidefinite matrix $\mathbf{M}$, we denote by $\lambda(\mathbf{M})$ the largest eigenvalue of $\mathbf{M}$:

$$\lambda(\mathbf{M}) \stackrel{\text{def}}{=} \max_{h \in \mathbb{R}^n} \{h^\top \mathbf{M} h : h^\top h \le 1\}, \tag{34}$$

and by $\lambda'(\mathbf{M})$ the 'normalized' largest eigenvalue of $\mathbf{M}$:

$$\lambda'(\mathbf{M}) \stackrel{\text{def}}{=} \max_{h \in \mathbb{R}^n} \{h^\top \mathbf{M} h : h^\top \mathrm{Diag}(\mathbf{M}) h \le 1\}. \tag{35}$$

Note that $1 \le \lambda'(\mathbf{M}) \le n$.

In this section we study (standard and normalized) largest eigenvalue of the probability matrix associated with a sampling:

$$\lambda(\hat{S}) \stackrel{\text{def}}{=} \lambda(\mathbf{P}(\hat{S})) \stackrel{(34)}{=} \max_{h \in \mathbb{R}^n} \{h^\top \mathbf{P}(\hat{S}) h : h^\top h \le 1\} \tag{36}$$

and

$$\lambda'(\hat{S}) \stackrel{\text{def}}{=} \lambda'(\mathbf{P}(\hat{S})) \stackrel{(35)}{=} \max_{h \in \mathbb{R}^n} \{h^\top \mathbf{P}(\hat{S}) h : h^\top \mathrm{Diag}(\mathbf{P}(\hat{S})) h \le 1\}. \tag{37}$$

Recall that by Theorem 3.1, $\mathbf{P}(\hat{S})$ is positive semidefinite for each sampling $\hat{S}$. For convenience, we write $\lambda(\hat{S})$ (resp. $\lambda'(\hat{S})$) instead of $\lambda(\mathbf{P}(\hat{S}))$ (resp. $\lambda'(\mathbf{P}(\hat{S}))$). We study these quantities since,

as we will show in later sections, they are useful in computing parameter $v = (v_1, \ldots, v_n)$ for which ESO holds.

If $\hat{S}$ is a uniform sampling (i.e. if $\mathbb{P}(i \in \hat{S}) = \mathbb{P}(j \in \hat{S})$ for all $i, j \in [n]$), then since $\mathrm{Tr}(\mathbf{P}(\hat{S})) = \mathbb{E}[|\hat{S}|]$ (see (24)), we have $\mathrm{Diag}(\mathbf{P}(\hat{S})) = (\mathbb{E}[|\hat{S}|]/n)\mathbf{I}$, from which we obtain (assuming that $\hat{S}$ is not nil):

$$\lambda'(\hat{S}) = \frac{n}{\mathbb{E}[|\hat{S}|]}\lambda(\hat{S}). \tag{38}$$

4.1 *Elementary samplings*

In the case of elementary samplings, the situation is simple. Indeed, for any $J \subseteq [n]$, we have

$$\lambda'(\hat{E}_J) = \lambda(\hat{E}_J) \overset{(16)}{=} \lambda(e_{[J]}e_{[J]}^\top) = e_{[J]}^\top e_{[J]} = |J|. \tag{39}$$

This can, in fact, be seen as a consequence of a more general identity[7] for arbitrary symmetric rank-one matrices: for any $x \in \mathbb{R}^n$, we have

$$\lambda'(xx^\top) = \|x\|_0 \overset{\text{def}}{=} |\{i : x_i \neq 0\}|. \tag{40}$$

Since $\mathbf{P}(\hat{E}_J) = \mathbf{E}_{[J]}$ and $\mathrm{Diag}(\mathbf{E}_{[J]}) = \mathbf{I}_{[J]}$, (39) can equivalently be written as

$$\mathbf{E}_{[J]} \preceq |J|\mathbf{I}_{[J]}, \tag{41}$$

and adding that the bound is tight.

4.2 *Bounds for arbitrary samplings*

In the first result of this section, we give sharp bounds for $\lambda'(\hat{S})$ for arbitrary sampling $\hat{S}$.

THEOREM 4.1 *Let $\hat{S}$ be an arbitrary sampling.*

(i) *Lower bound. If $\hat{S}$ is not nil (i.e. if $\mathbb{P}(\hat{S} \neq \emptyset) > 0$), then*

$$1 \leq \frac{\mathbb{E}[|\hat{S}|^2]}{\mathbb{E}[|\hat{S}|]} \leq \lambda'(\hat{S}).$$

(ii) *Upper bound. If τ is a constant such that $|\hat{S}| \leq \tau$ with probability 1, then $\lambda'(\hat{S}) \leq \tau$.*
(iii) *Identity. If $|\hat{S}| = \tau$ with probability 1, then $\lambda'(\hat{S}) = \tau$.*

Proof (i) For simplicity, let $\mathbf{P} = \mathbf{P}(\hat{S})$. If $e \in \mathbb{R}^n$ is the vector of all ones, then we get

$$\lambda'(\hat{S}) \overset{(37)}{\geq} \frac{e^\top \mathbf{P} e}{e^\top \mathrm{Diag}(\mathbf{P})e} = \frac{e^\top \mathbf{P} e}{\mathrm{Tr}(\mathbf{P})} \geq 1,$$

where the last inequality holds since $\mathrm{Tr}(\mathbf{P})$ is upper bounded by the sum of all elements of $\mathbf{P}$. It remains to apply identities (22) and (23).

(ii) In view of (14), we can represent $\hat{S}$ as a convex combination of elementary samplings:

$$\hat{S} = \sum_{S\subseteq[n]} q_S \hat{E}_S,$$

where $q_S = \mathbb{P}(\hat{S} = S)$. Since $|\hat{S}| \leq \tau$ with probability 1, we have $|S| \leq \tau$ whenever $q_S > 0$. Thus, we have

$$\mathbf{P}(\hat{S}) = \sum_{S\subseteq[n]} q_S \mathbf{P}(\hat{E}_S) \overset{(39)}{\preceq} \sum_{S\subseteq[n]} q_S |S| \mathrm{Diag}(\mathbf{P}(\hat{E}_S)) \preceq \tau \sum_{S\subseteq[n]} q_S \mathrm{Diag}(\mathbf{P}(\hat{E}_S)) \overset{(25)}{=} \tau \mathrm{Diag}(\mathbf{P}(\hat{S})).$$

(iii) The result follows by combining the upper and lower bounds.

■

In the next result, we study the quantity $\lambda(\hat{S})$.

THEOREM 4.2 *The following statements hold:*

(i) *Lower and upper bounds. For any sampling $\hat{S}$, we have*

$$\frac{\mathbb{E}[|\hat{S}|^2]}{n} \leq \lambda(\hat{S}) \leq \mathbb{E}[|\hat{S}|]. \tag{42}$$

(ii) *Sharper upper bound. If $\hat{S}$ is uniform and $|\hat{S}| \leq \tau$ with probability one, then the upper bound can be improved to*

$$\lambda(\hat{S}) \leq \frac{\mathbb{E}[|\hat{S}|]\tau}{n}.$$

(iii) *Identity. If $\hat{S}$ is uniform and $|\hat{S}| = \tau$ with probability one, then*

$$\lambda(\hat{S}) = \frac{\tau^2}{n}.$$

Proof (i) The upper bound holds since $\lambda(\hat{S})$ is the maximal eigenvalue of $\mathbf{P}(\hat{S})$ and by (24), $\mathbb{E}[|\hat{S}|] = \mathrm{Tr}(\mathbf{P}(\hat{S}))$. The lower bound follows from:

$$\lambda(\hat{S}) = \lambda(\mathbf{P}(\hat{S})) \geq \frac{e^\top \mathbf{P}(\hat{S}) e}{e^\top e} \overset{(22)}{=} \frac{\mathbb{E}[|\hat{S}|^2]}{n}.$$

(ii) By combining (38) and Theorem 4.1 (ii) we obtain

$$\lambda(\hat{S}) \overset{38}{=} \frac{\mathbb{E}[|\hat{S}|]}{n} \lambda'(\hat{S}) \overset{\text{Thm 4.1}}{\leq} \frac{\mathbb{E}[|\hat{S}|]\tau}{n}.$$

(iii) The result follows by combining the lower bound from (i) with the upper bound in (ii).

■

A natural lower bound for $\lambda(\hat{S})$ (largest eigenvalue of $\mathbf{P}(\hat{S})$) is $\mathbb{E}[|\hat{S}|]/n$ (the average of the eigenvalues of $\mathbf{P}(\hat{S})$). Notice that the lower bound in (42) is better than this. Moreover, observe that both bounds in (42) are tight. Indeed, in view of (39), the upper bound is achieved for any elementary sampling. The lower bound is also tight—in view of part (iii) of the theorem.

4.3 *Bounds for restrictions of selected samplings*

In this part we study the normalized eigenvalue associated with the restriction of a few selected samplings (or families of samplings). In particular, we first give a (necessarily rough) bound that holds for arbitrary samplings, followed by a bound for the (c,τ)-distributed sampling and the τ-nice sampling (both are specific uniform samplings). Finally, we give a bound for the family of doubly uniform samplings.

PROPOSITION 4.1 *Let $\hat{S}$ be an arbitrary sampling and let τ be such that $|\hat{S}| \leq \tau$ with probability 1. Then for all $\emptyset \neq J \subseteq [n]$, we have*

$$\lambda'(J \cap \hat{S}) \leq \min\{|J|, \tau\}. \tag{43}$$

Proof [8]Note that $|J \cap \hat{S}| \leq \min\{|J|, \tau\}$ with probability 1. We only need to apply the upper bound in Theorem 4.1 to the restriction sampling $J \cap \hat{S}$. ■

We now proceed to the (c,τ)-distributed sampling (recall Definition 2.2).

PROPOSITION 4.2 *Let $\hat{S}$ be the (c,τ)-distributed sampling associated with a partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_c\}$ of $[n]$ such that $s = |\mathcal{P}_l|$ for $l \in [c]$. Fix arbitrary $\emptyset \neq J \subseteq [n]$ and let ω' be the number of sets $\mathcal{P}_l$ which have a nonempty intersection with J; that is, let $\omega' \overset{\text{def}}{=} |\{l : J \cap \mathcal{P}_l \neq \emptyset\}|$. Then*

$$\lambda'(J \cap \hat{S}) \leq 1 + \frac{(|J|-1)(\tau-1)}{s_1} + |J|\left(\frac{\tau}{s} - \frac{\tau-1}{s_1}\right)\frac{\omega'-1}{\omega'}, \tag{44}$$

where $s_1 = \max(s-1, 1)$.

Proof By applying Lemmas 3.2 and 3.3, we get

$$\begin{aligned}\mathbf{P}(J \cap \hat{S}) &\overset{(27)}{=} \mathbf{P}(\hat{E}_J) \circ \mathbf{P}(\hat{S}) \overset{(29)}{=} \frac{\tau}{s}[\alpha_1 \mathbf{P}(\hat{E}_J) \circ \mathbf{I} + \alpha_2 \mathbf{P}(\hat{E}_J) \circ \mathbf{E} + \alpha_3 \mathbf{P}(\hat{E}_J) \circ (\mathbf{E} - \mathbf{B})] \\ &= \frac{\tau}{s}[\alpha_1 \mathrm{Diag}(\mathbf{P}(\hat{E}_J)) + \alpha_2 \mathbf{P}(\hat{E}_J) + \alpha_3 \mathbf{P}(\hat{E}_J) - \alpha_3 \mathbf{P}(\hat{E}_J) \circ \mathbf{B}].\end{aligned} \tag{45}$$

For any $h \in \mathbb{R}^n$,

$$h^\top \mathbf{P}(\hat{E}_J) h = \left(\sum_{i \in J} h_i\right)^2 = \left(\sum_{l=1}^{c} \sum_{i \in \mathcal{P}_l \cap J} h_i\right)^2 \leq \omega' \sum_{l=1}^{c} \left(\sum_{i \in \mathcal{P}_l \cap J} h_i\right)^2 = \omega' \sum_{l=1}^{c} h^\top \mathbf{P}(\hat{E}_{J \cap \mathcal{P}_l}) h, \tag{46}$$

where the inequality is an application of the Cauchy–Schwartz inequality. It follows that

$$\mathbf{P}(\hat{E}_J) \circ \mathbf{B} \overset{(30)}{=} \sum_{l=1}^{c} \mathbf{P}(\hat{E}_J) \circ \mathbf{P}(\hat{E}_{\mathcal{P}_l}) = \sum_{l=1}^{c} \mathbf{P}(\hat{E}_{J \cap \mathcal{P}_l}) \overset{(46)}{\succeq} \frac{1}{\omega'} \mathbf{P}(\hat{E}_J). \tag{47}$$

Plugging (47) into (45), we get

$$\begin{aligned}\mathbf{P}(J\cap\hat{S}) &\preceq \frac{\tau}{s}\left[\alpha_1\mathrm{Diag}(\mathbf{P}(\hat{E}_J))+\left(\alpha_2+\alpha_3\left(1-\frac{1}{\omega'}\right)\right)\mathbf{P}(\hat{E}_J)\right]\\ &\overset{(39)}{\preceq} \frac{\tau}{s}\left[\alpha_1+\left(\alpha_2+\alpha_3\left(1-\frac{1}{\omega'}\right)\right)|J|\right]\mathrm{Diag}(\mathbf{P}(\hat{E}_J))\\ &= \left[1+\frac{(|J|-1)(\tau-1)}{s_1}+|J|\left(\frac{\tau}{s}-\frac{\tau-1}{s_1}\right)\frac{\omega'-1}{\omega'}\right]\mathrm{Diag}(\mathbf{P}(\hat{E}_J))\circ\mathrm{Diag}(\mathbf{P}(\hat{S})).\end{aligned}$$

Finally, note that $\mathrm{Diag}(\mathbf{P}(\hat{E}_J))\circ\mathrm{Diag}(\mathbf{P}(\hat{S}))=\mathrm{Diag}(\mathbf{P}(\hat{E}_J)\circ\mathbf{P}(\hat{S}))\overset{(27)}{=}\mathrm{Diag}(\mathbf{P}(J\cap\hat{S}))$. ■

We now specialize the above result to the $c=1$ case, obtaining a formula for $\lambda'(J\cap\hat{S})$ in the case when $\hat{S}$ is the τ-nice sampling (recall Definition 2.3).

PROPOSITION 4.3 *Let $\hat{S}$ be the τ-nice sampling. Then for all $\emptyset\neq J\subseteq[n]$,*

$$\lambda'(J\cap\hat{S})=1+\frac{(|J|-1)(\tau-1)}{\max(n-1,1)}. \tag{48}$$

Proof Let $\emptyset\neq J\subseteq[n]$. Since τ-nice sampling is the $(1,\tau)$-distributed sampling, by applying Proposition 4.2, we get

$$\lambda'(J\cap\hat{S})\leq 1+\frac{(|J|-1)(\tau-1)}{\max(n-1,1)}.$$

Next, by direct calculation, we can verify that

$$\mathbb{E}[|J\cap\hat{S}|^2]=\frac{|J|\tau}{n}\left(1+\frac{(|J|-1)(\tau-1)}{\max(n-1,1)}\right)\quad\text{and}\quad\mathbb{E}[|J\cap\hat{S}|]=\frac{|J|\tau}{n},$$

which together with the lower bound established in Theorem 4.1 yields:

$$\lambda'(J\cap\hat{S})\geq\frac{\mathbb{E}[|J\cap\hat{S}|^2]}{\mathbb{E}[|J\cap\hat{S}|]}=1+\frac{(|J|-1)(\tau-1)}{\max(n-1,1)}.$$

■

Note that (48) is much better (i.e. smaller) than the right-hand side in (43). This is to be expected as the bound (43) applies to *all* samplings (which have size at most τ with probability 1).

Finally, we give a bound on the normalized largest eigenvalue of the restriction of a doubly uniform sampling.

PROPOSITION 4.4 *Let $\hat{S}$ be a doubly uniform sampling which is not nil (i.e. $\mathbb{P}(\hat{S}=\emptyset)\neq 1$). Then for all $\emptyset\neq J\subseteq[n]$,*

$$\lambda'(J\cap\hat{S})\leq 1+\frac{(|J|-1)\left(\frac{\mathbb{E}[|\hat{S}|^2]}{\mathbb{E}[|\hat{S}|]}-1\right)}{\max(n-1,1)}. \tag{49}$$

Proof Combining (27) and (32), we get

$$\mathbf{P}(J \cap \hat{S}) \overset{(27)}{=} \mathbf{P}(\hat{E}_J) \circ \mathbf{P}(\hat{S}) \overset{(32)}{=} \mathbf{E}_{[J]} \circ \left(\frac{\mathbb{E}[|\hat{S}|]}{n}((1-\beta)\mathbf{I} + \beta \mathbf{E}) \right)$$

$$= \frac{\mathbb{E}[|\hat{S}|]}{n}((1-\beta)\mathbf{I}_{[J]} + \beta \mathbf{E}_{[J]})$$

$$\overset{(41)}{\preceq} \frac{\mathbb{E}[|\hat{S}|]}{n}(1-\beta+\beta|J|)\mathbf{I}_{[J]} = (1+(|J|-1)\beta)\mathrm{Diag}(\mathbf{P}(J \cap \hat{S})),$$

where β is as in (33). ■

5. Expected separable overapproximation

In this section we develop a general technique for computing parameters $v = (v_1, \ldots, v_n)$ for which the ESO inequality (2) holds.

5.1 *General technique*

We will write $\mathbf{M}_1 \succeq \mathbf{M}_2$ to indicate that $\mathbf{M}_1 - \mathbf{M}_2$ is positive semidefinite. It is a well-known fact [9, Theorem 5.2.1] that the Hadamard product of two positive semidefinite matrices is positive semidefinite:

$$\mathbf{M}_1 \succeq 0 \quad \& \quad \mathbf{M}_2 \succeq 0 \Rightarrow \mathbf{M}_1 \circ \mathbf{M}_2 \succeq 0. \tag{50}$$

The reason for defining and studying probability matrices $\mathbf{P}(\hat{S})$ is motivated by the following result, which for functions satisfying Assumption 2.1 reduces the ESO Assumption $(f, \hat{S}) \sim \mathrm{ESO}(v)$ to the problem of bounding the Hadamard product of the probability matrix $\mathbf{P}(\hat{S})$ and the data matrix $\mathbf{A}^\top\mathbf{A}$ from above by a diagonal matrix. Note that because $\mathbf{P}(\hat{S}) \succeq 0$, in view of (50), the Hadamard product $\mathbf{P}(\hat{S}) \circ \mathbf{A}^\top\mathbf{A}$ is positive semidefinite.

LEMMA 5.1 *If f satisfies Assumption* 2.1 *and*

$$\mathbf{P}(\hat{S}) \circ (\mathbf{A}^\top\mathbf{A}) \preceq \mathrm{Diag}(v \circ p), \tag{51}$$

for some vector $v \in \mathbb{R}^n_{++}$, where p is the vector of probabilities defined in (1), *then*

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

Proof Let us substitute $h \leftarrow h_{[\hat{S}]}$ into (5) and take expectation in $\hat{S}$ of both sides. Applying (19), we obtain

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \mathrm{Diag}(\mathbf{P}(\hat{S}))\nabla f(x), h\rangle + \tfrac{1}{2}h^\top(\mathbf{P}(\hat{S}) \circ (\mathbf{A}^\top\mathbf{A}))h, \quad \forall x, h \in \mathbb{R}^n. \tag{52}$$

It remains to apply assumption (51). ■

We next focus on the problem of finding vector v for which (51) holds. The following direct consequence of (50) will be helpful in this regard:

$$(0 \preceq \mathbf{M}_1 \ \& \ \mathbf{M}_2 \preceq \mathbf{M}_3) \Rightarrow \mathbf{M}_1 \circ \mathbf{M}_2 \preceq \mathbf{M}_1 \circ \mathbf{M}_3. \tag{53}$$

In particular, (53) can be used to establish the first part of the following useful lemma.

LEMMA 5.2 *If* $\mathbf{M}_1 \succeq 0$ *and* $\mathbf{M}_2 \succeq 0$, *then*

$$\lambda'(\mathbf{M}_1 \circ \mathbf{M}_2) \leq \min\{\lambda'(\mathbf{M}_1), \lambda'(\mathbf{M}_2)\}, \tag{54}$$

$$\lambda'(\mathbf{M}_1 + \mathbf{M}_2) \leq \max\{\lambda'(\mathbf{M}_1), \lambda'(\mathbf{M}_2)\}. \tag{55}$$

Proof By definition, $\mathbf{M}_2 \preceq \lambda'(\mathbf{M}_2)\mathrm{Diag}(\mathbf{M}_2)$, which together with (53) implies:

$$\mathbf{M}_1 \circ \mathbf{M}_2 \preceq \lambda'(\mathbf{M}_2)(\mathbf{M}_1 \circ \mathrm{Diag}(\mathbf{M}_2)) = \lambda'(\mathbf{M}_2)\mathrm{Diag}(\mathbf{M}_1 \circ \mathbf{M}_2).$$

Applying the same reasoning to the matrix $\mathbf{M}_1$, we obtain $\mathbf{M}_1 \circ \mathbf{M}_2 \preceq \lambda'(\mathbf{M}_1)\mathrm{Diag}(\mathbf{M}_1 \circ \mathbf{M}_2)$. Combining the two results, we obtain (54). Inequality (55) follows from:

$$\begin{aligned}\mathbf{M}_1 + \mathbf{M}_2 &\preceq \lambda'(\mathbf{M}_1)\mathrm{Diag}(\mathbf{M}_1) + \lambda'(\mathbf{M}_2)\mathrm{Diag}(\mathbf{M}_2)\\ &\leq \max\{\lambda'(\mathbf{M}_1), \lambda'(\mathbf{M}_2)\}(\mathrm{Diag}(\mathbf{M}_1) + \mathrm{Diag}(\mathbf{M}_2))\\ &= \max\{\lambda'(\mathbf{M}_1), \lambda'(\mathbf{M}_2)\}\mathrm{Diag}(\mathbf{M}_1 + \mathbf{M}_2).\end{aligned}$$

■

5.2 *ESO I: no coupling between the sampling and data*

By applying Lemma 5.2, Equation (54), to $\mathbf{M}_1 = \mathbf{P}(\hat{S})$ and $\mathbf{M}_2 = \mathbf{A}^\top\mathbf{A}$, we obtain a formula for v satisfying (51).

THEOREM 5.1 (ESO without coupling between sampling and data) *Let f satisfy Assumption* 2.1 *and let $\hat{S}$ be an arbitrary sampling. Then $(f, \hat{S}) \sim \mathrm{ESO}(v)$ for $v = (v_1, \ldots, v_n)$ defined by*

$$v_i = \min\{\lambda'(\mathbf{P}(\hat{S})), \lambda'(\mathbf{A}^\top\mathbf{A})\}\sum_{j=1}^{m}\mathbf{A}_{ji}^2, \quad i \in [n]. \tag{56}$$

Proof Let $\mathbf{P} = \mathbf{P}(\hat{S})$. To establish the main statement, it is sufficient to apply Lemmas 5.1 and 5.2 and note that for v defined by (56), $\mathrm{Diag}(v \circ p) = \min(\lambda'(\mathbf{P}), \lambda'(\mathbf{A}^\top\mathbf{A}))\mathrm{Diag}(\mathbf{P} \circ \mathbf{A}^\top\mathbf{A})$. ■

If for some τ, $|\hat{S}| \leq \tau$ with probability 1, then in view of Theorem 4.1, we have $\lambda'(\mathbf{P}(\hat{S})) \leq \tau$. Furthermore,

$$\lambda'(\mathbf{A}^\top\mathbf{A}) = \lambda'\left(\sum_{j=1}^{m}\mathbf{A}_{j:}^\top\mathbf{A}_{j:}\right) \overset{(55)}{\leq} \max_j \lambda'(\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) \overset{(40)}{=} \max_j \|\mathbf{A}_{j:}\|_0 = \max_j |J_j|.$$

Hence, in view of Lemma 5.1, we can pick the ESO parameter conservatively as follows:

$$v_i = \min\{\tau, \max_j |J_j|\}\sum_{j=1}^{m}\mathbf{A}_{ji}^2, \quad i \in [n]. \tag{57}$$

An ESO inequality with v_i similar to (57) was established in [27], but for a different class of functions (ω-partially separable functions: functions expressed as a sum of functions each of which depends on at most ω coordinates) and uniform samplings only. Indeed, the bound established therein for arbitrary uniform samplings uses $v_i = \min\{\tau, \omega\}L_i$, where ω is the degree of

separability of f and L_i is the Lipschitz constant of ∇f associated with coordinate i. In our setting, $\omega = \max_j |J_j|$ and L_i corresponds to $\sum_j \mathbf{A}_{ji}^2$. Hence, (57) could be seen as a generalization of the ESO bound in [27] to arbitrary samplings.

Note that computation of the normalized eigenvalue $\lambda'(\mathbf{A}^\top\mathbf{A})$ could be time-consuming, and would require a number of passes through the data prior to running a coordinate descent method, which may be prohibitive. In the next section we follow a different approach, one in which this issue is avoided. The main idea is to decompose $\mathbf{A}^\top\mathbf{A}$ as a sum of the rank-one matrices $\mathbf{A}_{j:}^\top\mathbf{A}_{j:}$ and then bound each term $\mathbf{P}(\hat{S}) \circ \mathbf{A}_{j:}^\top\mathbf{A}_{j:}$ separately.

5.3 *ESO II: coupling the sampling with data*

In this section we use a different strategy for satisfying (51). We first write

$$\mathbf{P}(\hat{S}) \circ \mathbf{A}^\top\mathbf{A} = \mathbf{P}(\hat{S}) \circ \sum_{j=1}^m \mathbf{A}_{j:}^\top\mathbf{A}_{j:} = \sum_{j=1}^m \mathbf{P}(\hat{S}) \circ \mathbf{A}_{j:}^\top\mathbf{A}_{j:},$$

where $\mathbf{A}_{j:}$ denote the jth row vector of matrix $\mathbf{A}$ and then bound each term in the last sum individually. Recall the definition of set J_j from (6): $J_j = \{i \in [n] : \mathbf{A}_{ji} \neq 0\}$.

THEOREM 5.2 (ESO with coupling between sampling and data) *Let $\hat{S}$ be an arbitrary sampling and $v = (v_1, \ldots, v_n)$ be defined by*

$$v_i = \sum_{j=1}^m \lambda'(J_j \cap \hat{S})\mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n. \tag{58}$$

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

Proof Let $j \in [m]$ and $\mathbf{A}_{j:}$ denote the jth row vector of matrix $\mathbf{A}$. By the definition of J_j,

$$\mathbf{A}_{j:}^\top\mathbf{A}_{j:} = (e_{[J_j]}e_{[J_j]}^\top) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) = \mathbf{P}(\hat{E}_{J_j}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}).$$

Thus, $\mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) = \mathbf{P}(\hat{S}) \circ \mathbf{P}(\hat{E}_{J_j}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) = \mathbf{P}(J_j \cap \hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:})$. We now apply Lemma 5.2 to the sampling $J_j \cap \hat{S}$ and the matrix $\mathbf{A}_{j:}$ and obtain

$$\begin{aligned}\mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) &\preceq \min\{\lambda'(J_j \cap \hat{S}), \lambda'(\mathbf{A}_{j:}^\top\mathbf{A}_{j:})\}\mathrm{Diag}(\mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:})) \\ &\preceq \lambda'(J_j \cap \hat{S})\mathrm{Diag}(\mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:})).\end{aligned} \tag{59}$$

Therefore,

$$\begin{aligned}\mathbf{P}(\hat{S}) \circ \mathbf{A}^\top\mathbf{A} = \sum_{j=1}^m \mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) &\preceq \sum_{j=1}^m \lambda'(J_j \cap \hat{S})\mathrm{Diag}(\mathbf{P}(\hat{S}) \circ (\mathbf{A}_{j:}^\top\mathbf{A}_{j:})) \\ &= \mathrm{Diag}(\mathbf{P}(\hat{S})) \circ \sum_{j=1}^m \lambda'(J_j \cap \hat{S})\mathrm{Diag}(\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) = \mathrm{Diag}(p) \circ \mathrm{Diag}(v),\end{aligned}$$

where $p = (p_1, \ldots, p_n)$ is the vector of probability defined in (1) and $v = (v_1, \ldots, v_n)$ is defined in (58). For completeness, let us show that the second inequality in (59) can be replaced by equality. Indeed, from (40) and the fact that $|J_j| = \|A_{j:}\|_0$, we obtain $\lambda'(\mathbf{A}_{j:}^\top\mathbf{A}_{j:}) = |J_j|$. Finally,

using the upper bound in Theorem 4.1, we know that $\lambda'(J_j \cap \hat{S}) \leq |J_j|$. Hence, $\min\{\lambda'(J_j \cap \hat{S}), \lambda'(\mathbf{A}_{j:}^\top \mathbf{A}_{j:})\} = \lambda'(J_j \cap \hat{S})$. ■

The benefit of this approach is twofold: First, if the data matrix $\mathbf{A}$ is sparse, the sets J_j have small cardinality, and from Proposition 4.1 (or other results in Section 4.3, depending on the sampling $\hat{S}$ used), we conclude that $\lambda'(J_j \cap \hat{S})$ is small. Hence, the parameters v_i obtained through (58) get better (i.e. smaller) with sparser data. Second, the formula for v_i does not involve the need to compute an eigenvalue associated with the data matrix. On the other hand, instead of having to compute $\lambda'(\hat{S})$ (which, as we have seen, is equal to τ if $|\hat{S}| = \tau$ with probability 1), we now need to compute the normalized largest eigenvalue of m restrictions of $\hat{S}$, $\lambda'(J_j \cap \hat{S})$ for all $j = 1, 2, \ldots, m$. However, for this there is a good upper bound available through Proposition 4.1 for an arbitrary sampling, and refined bounds can be derived for specific samplings (for examples, see Section 4.3).

5.4 *ESO without eigenvalues*

In this section we illustrate the use of the techniques developed in the preceding sections to derive ESO inequalities, for selected samplings, which do not depend on any eigenvalues, and lead to easily computable ESO parameters $v = (v_1, \ldots, v_n)$. The techniques can be used to derive similar ESO inequalities for other samplings as well.

PROPOSITION 5.1 *Let f satisfy Assumption 2.1 and let sets $J_1, \ldots, J_m$ be defined as in (6). Then $(f, \hat{S}) \sim \text{ESO}(v)$ provided that the sampling $\hat{S}$ and vector v are chosen in any of the following ways:*

(i) *$\hat{S}$ is an arbitrary sampling such that $|\hat{S}| \leq \tau$ with probability 1, and*

$$v_i = \sum_{j=1}^{m} \min\{|J_j|, \tau\} \mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n. \tag{60}$$

(ii) *$\hat{S}$ is the (c, τ)-distributed sampling and*

$$v_i = \sum_{j=1}^{m} \left[1 + \frac{(|J_j| - 1)(\tau - 1)}{s_1} + |J_j| \left(\frac{\tau}{s} - \frac{\tau - 1}{s_1} \right) \frac{\omega_j' - 1}{\omega_j'} \right] \mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n, \tag{61}$$

where $\omega_j' \overset{\text{def}}{=} |\{l : \mathcal{P}_l \cap J_j \neq 0\}|$ for $j \in [m]$.

(iii) *$\hat{S}$ is the τ-nice sampling (for $\tau \geq 1$) and*

$$v_i = \sum_{j=1}^{m} \left[1 + \frac{(|J_j| - 1)(\tau - 1)}{\max(n - 1, 1)} \right] \mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n, \tag{62}$$

(iv) *$\hat{S}$ is a doubly uniform sampling (which is not nil) and*

$$v_i = \sum_{j=1}^{m} \left[1 + \frac{(|J_j| - 1) \left(\frac{\mathbb{E}[|\hat{S}|^2]}{\mathbb{E}[|\hat{S}|]} - 1 \right)}{\max(n - 1, 1)} \right] \mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n, \tag{63}$$

(v) $\hat{S}$ *is a graph sampling and*

$$v_i = \sum_{j=1}^{m} \mathbf{A}_{ji}^2, \quad i = 1, 2, \ldots, n. \tag{64}$$

(vi) $\hat{S}$ *is a serial sampling* (*i.e. a sampling for which* $|\hat{S}| = 1$ *with probability* 1) *and* $v = (v_1, \ldots, v_n)$ *is defined as in* (64).

Proof (i) A direct consequence of Theorem 5.2 and Proposition 4.1.
(ii) A direct consequence of Theorem 5.2 and Proposition 4.2.
(iii) This is a special case of part (ii) for $c = 1$.
(iv) A direct consequence of Theorem 5.2 and Proposition 4.4.
(v) For a graph sampling it is clear that $|J_j \cap \hat{S}| \leq 1$ with probability 1 for all $j \in [m]$. The result then follows from Theorem 5.2.
(vi) A special case of (v). Indeed, a single vertex is an independent set of a graph.

■

Remark Note that part (i) of Proposition 5.1 is a strict improvement on (57). Also, this is strict improvement, both in the quality of the bound and in generality of the sampling, on the result in [27], which was proved for uniform samplings only and where the bound involved $\max_j |J_j|$ instead of $|J_j|$. Part (ii) should be compared with the results obtained in [8] and part (iii) with those in [7,27].

6. Discussion

6.1 *Trade-off between preprocessing time and iteration complexity*

As stressed before, smaller parameter $v = (v_1, \ldots, v_n)$ leads to better convergence result (see Table 1) but computing the smallest admissible v would require too large computational effort. Nevertheless, using a cheaply computed parameter $v = (v_1, \ldots, v_n)$ would lead to large iteration complexity and slow convergence. The trade-off between the preprocessing time for computing the parameter $v = (v_1, \ldots, v_n)$ and the iteration complexity of the algorithm shall be discussed next.

For specific samplings such as τ-nice sampling and (c, τ)-distributed sampling, admissible v can be computed using dedicated formulae (61) and (62), which appeared, respectively in [7,8]. For arbitrary sampling $\hat{S}$, admissible parameter v can be computed according to (57), (58) or (60), which are given for the first time. While (58) requires computing the largest eigenvalue for m matrices of sizes $\{J_1, \ldots, J_m\}$, both (57) and (60) can be computed in at most two passes over the data. In return, (58) provides a smaller parameter v which improves the iteration complexity.

For approximating $\lambda'(J_j \cap \hat{S})$, one can apply power method on the positive semidefinite matrix $\mathbf{P}(J_j \cap \hat{S})$. The number of operations needed in one iteration of the power method is $|J_j|^2$ and if we apply T iterations of power method[9], then the total number of operations needed for computing v using (58) is

$$O\left(T \sum_{j=1}^{m} |J_j|^2\right) \leq O\left(T \max_j |J_j| \mathrm{nnz}(\mathbf{A})\right),$$

where the big O notation hides constants independent of the data matrix $\mathbf{A}$.

Recall from Table 1 how the iteration complexity of different methods depends on the parameter $v = (v_1, \dots, v_n)$. Let us consider the strongly convex smooth objective function set-up and assume that the random sampling $\hat{S}$ has cardinality τ with probability 1. Then the computational time of one epoch (n iterations) is of the same order as τ passes over the data. Therefore, given a parameter $v = (v_1, \dots, v_n)$, the number of passes over the data is bounded by

$$O\left(\frac{1}{\lambda}\max_i \frac{v_i\tau}{p_i n}\log\left(\frac{1}{\epsilon}\right)\right),$$

where ϵ is the target accuracy and λ is the strong convexity parameter of the problem.

The comparison of the three formulae in terms of overall complexity is reported in Table 2. It is clear from the table that the trade-off between the preprocessing and the iteration complexity mainly depends on the proportion between $T\sum_{j=1}^m |J_j|^2/\mathrm{nnz}(\mathbf{A})$ and $(1/\lambda)\max_i(v_i\tau/p_i n)$. In Table 3 we report the actual computing time of v using different formulae and the corresponding value of $\max_i(v_i\tau/p_i n)$, for two real data matrices w8a and dorothea. To facilitate the comparison we normalized the two data sets so that the diagonal elements of $\mathbf{A}^\top\mathbf{A}$ are all one. The samplings $\hat{S}$ that we used in the experiments are all product sampling (Definition 2.8) with respect to some random partition of the set $[n]$. The number of iterations T for the power method is fixed to 10 and we multiply the obtained value by 1.01. Because of the comparable processing time, Formula (60) is clearly better than Formula (57). From Table 3 we also see that Formula (58) requires significant computational effort for computing v comparing to the other two but also reduces the value of $\max_i(v_i\tau/p_i n)$ by order of magnitude in most of

Table 2. Total number of passes over data for three different admissible parameters v.

$v = (v_1, \dots, v_n)$	Number of passes over the data
(57)	$O\left(1 + \frac{1}{\lambda}\max_i \frac{v_i\tau}{p_i n}\log(\frac{1}{\epsilon})\right)$
(58)	$O\left(\frac{T\sum_{j=1}^m \lvert J_j\rvert^2}{\mathrm{nnz}(\mathbf{A})} + \frac{1}{\lambda}\max_i \frac{v_i\tau}{p_i n}\log(\frac{1}{\epsilon})\right)$
(60)	$O\left(1 + \frac{1}{\lambda}\max_i \frac{v_i\tau}{p_i n}\log(\frac{1}{\epsilon})\right)$

Table 3. Comparison of Formula (57), Formula (58) and Formula (60).

		$\frac{\text{Time of computing } v}{\text{Time of one pass over data}}$			$\max_i \frac{v_i\tau}{p_i n}$		
Data	τ	Equation (57)	Equation (60)	Equation (58)	Equation (57)	Equation (60)	Equation (58)
	1	1	1	1	1	1	1
w8a	8	1	1.9	382.5	8.11	8.11	1.92
$n = 300$	16	1	1.8	380.3	17.07	17.10	3.03
$m = 49{,}749$	24	1	2.0	377.2	24.96	24.97	3.98
$\frac{\mathrm{nnz}}{nm} \simeq 3.8\%$	128	1	1.8	331.4	145.92	50.45	20.80
	256	1	1.8	313.2	194.56	67.13	50.54
	1	1	1	1	1	1	1
dorothea	8	1	3.2	8057.1	8.01	8.01	1.44
$n = 100{,}000$	16	1	1.52	8442.4	16.01	16.01	1.93
$m = 800$	24	1	1.52	8546.5	24.01	24.01	2.42
$\frac{\mathrm{nnz}}{nm} \simeq 0.91\%$	128	1	1.52	8686.6	128.13	128.13	8.79
	256	1	1.52	8715.8	256.91	256.91	16.68
	1024	1	1.53	8724.1	1038.1	1038.1	64.5

the regimes. Let us take the example of dorothea with $\tau = 256$, then the overall number of passes over data is $O(1.52 + (256.91/\lambda)\log(1/\epsilon))$ if v is computed using Formula (60) and $O(8715.8 + (16.68/\lambda)\log(1/\epsilon))$ if v is computed using Formula (58). Hence for small enough strong convexity parameter λ, it is worth to spend more time in computing a good parameter v using Formula (58), which will then be compensated by a smaller iteration complexity.

6.2 *Optimal sampling*

Proposition 5.1 should be understood in the context of complexity results for randomized coordinate descent, such as those in Table 1. For instance, in view of (60) for an arbitrary sampling $\hat{S}$ such that $|\hat{S}| \leq \tau$ with probability 1, the accelerated coordinate descent method developed in [19] has complexity

$$\sqrt{2\sum_{i=1}^{n}\frac{v_i(x_i^0 - x_i^*)^2}{p_i^2}} \times \frac{1}{\sqrt{\epsilon}} = \sqrt{2\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\min\{|J_j|,\tau\}\mathbf{A}_{ji}^2(x_i^0 - x_i^*)^2}{p_i^2}} \times \frac{1}{\sqrt{\epsilon}}. \tag{65}$$

Naturally, the bound improves if we use a specialized sampling, such as the τ-nice sampling (since the constants v_i become smaller).

Sometimes, one can find a sampling which minimizes the complexity bound. For instance, if we restrict our attention to serial samplings only (samplings picking a single coordinate at a time), then one can find probabilities $p_1, \ldots, p_n$, which uniquely define a sampling, minimizing the complexity bound:

$$p_i = \frac{(w_i(x_i^0 - x_i^*)^2)^{1/3}}{\sum_{i=1}^{n}(w_i(x_i^0 - x_i^*)^2)^{1/3}}, \quad i \in [n], \tag{66}$$

where $w_i = \sum_j \mathbf{A}_{ji}^2$. Note that if the ith coordinate is optimal at the starting point (i.e. if $x_i^0 = x_i^*$), then the prediction is to choose $p_i = 0$ (i.e. to never update coordinate i)—this is what one would expect. Using the serial sampling defined by (66), the complexity (65) takes the form

$$C_{\text{opt}} = \sqrt{2}\left(\sum_{i=1}^{n} w_i^{1/3}(x_i^0 - x_i^*)^{2/3}\right)^{3/2} \times \frac{1}{\sqrt{\epsilon}} = \frac{\sqrt{2}\|d\|_2^3}{\sqrt{\epsilon}},$$

where $d \in \mathbb{R}^n$ with $d_i = w_i^{1/6}(x_i^0 - x_i^*)^{1/3}$ and $\|d\|_q = (\sum_{i=1}^{n} d_i^q)^{1/q}$. However, if the uniform serial sampling is used instead (each coordinate is chosen with probability $p_i = 1/n$), then the complexity (65) has the form

$$C_{\text{unif}} = \sqrt{2}n\left(\sum_{i=1}^{n} w_i(x_i^0 - x_i^*)^2\right)^{1/2} \times \frac{1}{\sqrt{\epsilon}} = \frac{\sqrt{2}n\|d\|_6^3}{\sqrt{\epsilon}}.$$

While $\|d\|_6 \leq \|d\|_2$ for all d, these quantities can be equal, in which case C_{opt} is n times better than C_{unif}.

7. Conclusion

We have conducted a systematic study of ESO inequalities for a large class of functions (those satisfying Assumption 2.1) and *arbitrary* samplings. These inequalities are crucial in the design

and complexity analysis of randomized coordinate descent methods. This led us to study standard and normalized largest eigenvalue of the Hadamard product of the probability matrix associated with a sampling and a certain positive semidefinite matrix containing the data defining the function. Using our approach we have established new ESO results and also re-derived ESO results already established in the literature (in the case of uniform samplings) via different techniques. Our approach can be used to derive further bounds for specific samplings and can potentially be of interest outside the domain of randomized coordinate descent.

Acknowledgments

Accelerated Coordinate Descent Methods for Big Data Optimization. Most of the material of this paper was obtained by the authors in Spring 2014, and was presented by PR in June 2014 at the 'Khronos Days Summer School' focused on 'High-Dimensional Learning and Optimization' in Grenoble, France [21]; http://www.maths.ed.ac.uk/%7Eprichtar/docs/cdm-talk.pdf.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors acknowledge support from the EPSRC Grant EP/K02325X/1.

Notes

1. This definition can in a straightforward way be extended the case when coordinates are replaced by *blocks* of coordinates [27]. In such a case, h_i would be a allowed to be a vector of size larger than one, e_i would be replaced by a column submatrix of the identity matrix (usually denoted U_i in the literature) and h_i^2 would be replaced by the squared norm of h_i (it is often useful to design this norm based on properties of f).
2. All existing *parallel* coordinate descent methods for which a complexity analysis has been performed are designed with fixed stepsizes. Designing a line-search procedure in such a set-up is a non-trivial task, and to the best of our knowledge, only a single paper in the literature deals with this issue [6]. Certainly, properly designed line search has the potential to improve the practical performance of these methods.
3. We exclude from the table some earlier results [25], where an arbitrary *serial* sampling was analysed; i.e. sampling $\hat{S}$ for which $\mathbb{P}(|\hat{S}| = 1) = 1$. The situation is much simpler for serial samplings.
4. Complexity of NSync depends on the initial (x_0) and optimal (x_*) points, but have hidden this dependence. The full bound is obtained by replacing $\log(1/\epsilon)$ by $\log((f(x_0) - f(x_*))/\epsilon)$, where f is the objective function.
5. Complexity of Quartz depends on an initial pair of primal and dual vectors; we have omitted this dependence from the table. The full complexity result is obtained by replacing $\log(1/\epsilon)$ by $\log(\Delta_0/\epsilon)$, where Δ_0 is the difference between the primal and dual function values for a pair of (primal and dual) starting points.
6. Identities (20)–(23) were already established in [27], in a different way without relying on Theorem 3.1, which is new. However, in this paper a key role is played by identities (18)–(19), which are also new. It was while proving these identities that we realized the fundamental nature of Theorem 3.1, as a vehicle for obtaining all identities in Corollary 3.1 as a consequence. The identities will be needed in further development. For illustration of a different proof technique, here is an alternative proof of (19):
   $\mathbb{E}[h_{[\hat{S}]}^{\mathrm{T}}\mathbf{M}h_{[\hat{S}]}] = \mathbb{E}\left[\sum_{(i,i')\in\hat{S}\times\hat{S}}\mathbf{M}_{ii'}h_{i'}h_i\right] = \sum_{(i,i')\in[n]\times[n]}\mathbb{P}(i\in\hat{S}, i'\in\hat{S})\mathbf{M}_{ii'}h_{i'}h_i = h^{\mathrm{T}}(\mathbf{P}\circ\mathbf{M})h.$
7. The proof is immediate: fixing x, for any $h\in\mathbb{R}^n$ we have $(h^\top x)^2 = ((h\circ x)^\top e)^2 = ((h\circ x)^\top e_{[S]})^2$, where e is the vector of all ones, $S = \{i : x_i \neq 0\}$ and the entries of $e_{[S]}$ are 1 for $i\in S$ and 0 otherwise. It only remains to apply the Cauchy–Schwartz inequality, which is attained, whence the identity.
8. This simple result can alternatively be proved by applying (54) (which we mention in a later section) together with (27), (39) and the upper bound in Theorem 4.1.
9. Note that as the matrix $\mathbf{P}(J_j\cap\hat{S})$ is positive semidefinite, the power method always converges to the largest eigenvalue even if it is not a dominant eigenvalue. We defer the study on the convergence rate of power method for different matrices $\mathbf{P}(J_j\cap\hat{S})$ to a future work.

References

[1] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim. 23(4) (2013), pp. 2037–2060.

[2] J.K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, *Parallel coordinate descent for l1-regularized loss minimization*, in *ICML*, Washington, 2011, pp. 321–328.

[3] A.A. Canutescu and R.L. Dunbrack, *Cyclic coordinate descent: A robotics algorithm for protein loop closure*, Protein Sci. 12 (2003), pp. 963–972.

[4] I.S. Dhillon, P.K. Ravikumar, and A. Tewari, *Nearest neighbor based greedy coordinate descent*, in *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., Montreal, 2011, pp. 2160–2168.

[5] O. Fercoq and P. Richtárik, *Smooth minimization of nonsmooth functions by parallel coordinate descent*, preprint (2013). Available at arXiv:1309.5885.

[6] O. Fercoq and P. Richtárik, *Universal coordinate descent methods*, Tech. rep., University of Edinburgh, Edinburgh, May 2014.

[7] O. Fercoq and P. Richtárik, *Accelerated, parallel, and proximal coordinate descent*, SIAM J. Optim. 25(4) (2015), pp. 1997–2023.

[8] O. Fercoq, Z. Qu, P. Richtárik, and M. Takáč, Fast distributed coordinate descent for minimizing non-strongly convex losses, in *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[9] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[10] Y.T. Lee and A. Sidford, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems*, preprint (2013). Available at arXiv:1305.1922.

[11] Q. Lin, Z. Lu, and L. Xiao, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*, SIAM J. Optim. 25(4) (2015), pp. 2244–2273.

[12] J. Liu and S.J. Wright, *Asynchronous stochastic coordinate descent: Parallelism and convergence properties*, SIAM J. Optim. 25(1) (2015), pp. 351–376.

[13] J. Liu, S.J. Wright, C. Ré, V. Bittorf, and S. Sridhar, *An asynchronous parallel stochastic coordinate descent algorithm*, J. Mach. Learn. Res. 16 (2015), pp. 285–322.

[14] Z. Lu and L. Xiao, *On the complexity analysis of randomized block-coordinate descent methods*, Math. Program. 152(1) (2014), pp. 615–642.

[15] Z.Q. Luo and P. Tseng, *A coordinate gradient descent method for nonsmooth separable minimization*, J. Optim. Theory Appl. 72(1) (2002).

[16] I. Necoara and A. Patrascu, *A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints*, Comput. Optim. Appl. 57 (2014), pp. 307–337.

[17] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim. 22(2) (2012), pp. 341–362.

[18] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander and H. Koepke, *Coordinate descent converges faster with the gauss-southwell rule than random selection*, in *ICML*, Lille, France, 2015, pp. 1632–1641.

[19] Z. Qu and P. Richtárik, *Coordinate descent methods with arbitrary sampling I: Algorithms and complexity*, preprint (2014). Available at arXiv:1412.8060.

[20] Z. Qu, P. Richtárik and T. Zhang, *Quartz: Randomized dual coordinate ascent with arbitrary sampling*, in *Advances in Neural Information Processing Systems 28*, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds., Curran Associates, Inc., 2015, pp. 865–873.

[21] P. Richtárik, *Randomized coordinate descent for big data optimization (theory)*, Tech. rep., School of Mathematics, University of Edinburgh, Edinburgh, June 2014.

[22] P. Richtárik and M. Takáč, *Efficiency of randomized coordinate descent methods on minimization problems with a composite objective function*, in *4th Workshop on Signal Processing with Adaptive Sparse Structured Representations*, June 2011.

[23] P. Richtárik and M. Takáč, *Efficient serial and parallel coordinate descent method for huge-scale truss topology design*, in *Operations Research Proceedings 2011: Selected Papers of the International Conference on Operations Research (OR 2011), August 30 – September 2, 2011, Zurich, Switzerland*, D. Klatte, H.J. Lüthi and K. Schmedders, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 27–32.

[24] P. Richtárik and M. Takáč, *Distributed coordinate descent method for learning with big data*, preprint (2013). Available at arXiv:1310.2059.

[25] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program. 144(2) (2014), pp. 1–38.

[26] P. Richtárik and M. Takáč, *On optimal probabilities in stochastic coordinate descent methods*, Optim. Lett. (2015), pp. 1–11.

[27] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program. 156 (2016), pp. 433–484.

[28] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss*, J. Mach. Learn. Res. 14(1) (2013), pp. 567–599.

[29] S. Shalev-Shwartz and T. Zhang, *Accelerated mini-batch stochastic dual coordinate ascent*, in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 378–385.

[30] M. Takáč, A. Bijral, P. Richtárik and N. Srebro, *Mini-batch primal and dual methods for SVMs*, in *ICML*, Atlanta, USA, 2013, pp. 1022–1030.
[31] R. Tappenden, P. Richtárik and B. Büke, *Separable approximations and decomposition methods for the augmented lagrangian*, Optim. Methods Softw. 30(3) (2015), pp. 643–668.
[32] R. Tappenden, P. Richtárik and J. Gondzio, *Inexact coordinate descent: complexity and preconditioning*, J. Optim. Theory Appl. (2016), pp. 1–33.
[33] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl. 109 (2001), pp. 475–494.
[34] S.J. Wright, *Coordinate descent algorithms*, Math. Program. 151(1) (2015), pp. 3–34.
[35] T.T. Wu and K. Lange, *Coordinate descent algorithms for lasso penalized regression*, Ann. Appl. Stat. 2(1) (2008), pp. 224–244.
[36] P. Zhao and T. Zhang, *Stochastic optimization with importance sampling*, in *ICML*, Lille, France, 2015.

Appendix. Frequently used notation

Table A1. Notation appearing frequently in the paper.

Samplings		
$\mathbb{P}$	Probability	
$\mathbb{E}$	Expectation	
S,J	Subsets of $[n] \stackrel{\text{def}}{=} \{1,2,\dots,n\}$	
$\hat{S}$	Sampling, i.e. a random subset of $[n]$	Sections 1.1,2.2
$\hat{E}_S$	Elementary sampling associated with set $S \subseteq [n]$	Definition 2.1
$\hat{S}_1 \cap \hat{S}_2$	Intersection of samplings $\hat{S}_1$ and $\hat{S}_2$	Definition 2.5
$J \cap \hat{S}$	Restriction of sampling $\hat{S}$ to set J $(= \hat{E}_S \cap \hat{S})$	Definition 2.6
$\mathbf{P} = \mathbf{P}(\hat{S})$	n-by-n probability matrix: $\mathbf{P}_{ij} = \mathbb{P}(\{i,j\} \subseteq \hat{S})$	Section 3
p_i	$p_i = \mathbf{P}_{ii} = \mathbb{P}(i \in \hat{S})$	(1)
p	$p = (p_1,\dots,p_n)^\top \in \mathbb{R}^n$	(1)
Matrices and vectors		
e	The n-by-1 vector of all ones	
e_i	The ith unit coordinate vector in $\mathbb{R}^n$	
$h_{[S]}$	For $h \in \mathbb{R}^n$ and $S \subseteq [n]$, this is defined by $h_{[S]} = \sum_{i \in S} h_i e_i$	
$\mathbf{A}$	m-by-n data matrix defining f	(5)
J_j	The set of $i \in [n]$ for which $\mathbf{A}_{ij} \neq 0$	(6)
$\mathbf{I}$	n-by-n identity matrix	
$\mathbf{E}$	n-by-n matrix of all ones	
Diag	Outputs a diagonal matrix based on its argument (matrix or vector)	
$\circ$	Hadamard (elementwise) product of two matrices or vectors	
$\mathbf{M}_{[S]}$	Restriction of matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ to rows and columns indexed by S	(4)
$\lambda(\mathbf{M})$	Maximal eigenvalue of n-by-n matrix $\mathbf{M}$	Section 4; (34)
$\lambda'(\mathbf{M})$	Normalized maximal eigenvalue of n-by-n matrix $\mathbf{M}$	Section 4; (35)
$\lambda'(\hat{S})$	Shorthand notation for $\lambda'(\mathbf{P}(\hat{S}))$	