Coordinate Descent Face-Off: Primal or Dual?

Dominik Csiba Peter Richtárik[*]

School of Mathematics
University of Edinburgh
United Kingdom

May 29, 2016

Abstract

Randomized coordinate descent (RCD) methods are state-of-the-art algorithms for training linear predictors via minimizing regularized empirical risk. When the number of examples (n) is much larger than the number of features (d), a common strategy is to apply RCD to the dual problem. On the other hand, when the number of features is much larger than the number of examples, it makes sense to apply RCD directly to the primal problem. In this paper we provide the first joint study of these two approaches when applied to L2-regularized ERM. First, we show through a rigorous analysis that for dense data, the above intuition is precisely correct. However, we find that for sparse and structured data, primal RCD can significantly outperform dual RCD even if $d \ll n$, and vice versa, dual RCD can be much faster than primal RCD even if $n \ll d$. Moreover, we show that, surprisingly, a single sampling strategy minimizes both the (bound on the) number of iterations and the overall expected complexity of RCD. Note that the latter complexity measure also takes into account the average cost of the iterations, which depends on the structure and sparsity of the data, and on the sampling strategy employed. We confirm our theoretical predictions using extensive experiments with both synthetic and real data sets.

1 Introduction

In the last 5 years or so, randomized coordinate descent (RCD) methods [22, 12, 18, 19] have become immensely popular in a variety of machine learning tasks, with supervised learning being a prime example. The main reasons behind the rise of RCD-type methods is that they can be easily implemented, have intuitive appeal, and enjoy superior theoretical and practical behaviour when compared to classical methods such as SGD [20], especially in high dimensions, and in situations when solutions of medium to high accuracy are needed. One of the most important success stories of RCD is in the domain of training linear predictors via regularized empirical risk minimization (ERM).

The highly popular SDCA algorithm [24] arises as the application of RCD [18] to the *dual problem* associated with the (primal) ERM problem[1]. In practice, SDCA is most effective in situations where the number of examples (n) exceeds the number of features (d). Since the dual of ERM is an n dimensional problem, it makes intuitive sense to apply RCD to the dual. Indeed, RCD can be seen as a randomized decomposition strategy, reducing the n dimensional problem to a sequence of (randomly generated) one-dimensional problems.

However, if the number of features exceeds the number of examples, and especially when the difference is very large, RCD methods [19] have been found very attractive for solving the *primal problem* (i.e., the ERM

[*] This author would like to acknowledge support from the EPSRC Fellowship EP/N005538/1, “Randomized Algorithms for Extreme Convex Optimization”.

[1] Indeed, the analysis of SDCA in [24] proceeds by applying the complexity result from [18] to the *dual problem*, and then arguing that the same rate applies to the primal suboptimality as well.

problem) directly. For instance, distributed variants of RCD, such as Hydra [17] and its accelerated cousin Hydra^2 [3] have been successfully applied to solving problems with billions of features.

Recently, a variety of novel primal methods for ERM have been designed, including SAG [21], SVRG [5], S2GD [8], proxSVRG [25], mS2GD [6], SAGA [2], MISO [10] and S2CD [7]. As SDCA, all these methods improve dramatically on SGD [20] as a benchmark, which they achieve by employing one of a number of variance-reduction strategies. However, these methods have essentially identical identical theoretical behavior to SDCA, including the property that these methods thrive in the data-laden domain (i.e., $n \gg d$). In this sense, in our comparison of primal vs dual RCD, these methods should be viewed as "dual methods".

1.1 Contributions

In this paper we provide the first joint study of these two approaches—applying RCD to the primal vs dual problems—and we do so in the context of L2-regularized ERM. First, we show through a rigorous theoretical analysis that for dense data, the intuition that the primal approach is better than the dual approach when $n \geq d$, and vice versa, is precisely correct. However, we show that for sparse data, this does not need to be the case: primal RCD can significantly outperform dual RCD even if $d \ll n$, and vice versa, dual RCD can be much faster than primal RCD even if $n \ll d$. In particular, we identify that the face-off between primal and dual RCD boils down to the comparison of as single quantity associated with the data matrix and its transpose. Moreover, we show that, surprisingly, a single sampling strategy minimizes both the (bound on the) number of iterations and the overall expected complexity of RCD. Note that the latter complexity measure takes into account also the average cost of the iterations, which depends on the structure and sparsity of the data, and on the sampling strategy employed. We confirm our theoretical findings using extensive experiments with both synthetic and real data sets.

2 Primal and Dual Formulations of ERM

Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be a data matrix, with n referring to the number of examples and d to the number of features. With each example $\mathbf{X}_{:j} \in \mathbb{R}^d$ we associate a loss function $\phi_j : \mathbb{R} \to \mathbb{R}$, and pick a regularization constant $\lambda > 0$. The key problem of this paper is the L2-regularized ERM problem

$$\min_{w\in\mathbb{R}^d} \left[P(w) := \frac{1}{n}\sum_{j=1}^n \phi_j(\langle \mathbf{X}_{:j}, w\rangle) + \frac{\lambda}{2}\|w\|_2^2 \right], \tag{1}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean inner product and $\|w\|_2 := \sqrt{\langle w, w\rangle}$. We refer to (1) as the *primal problem*. We assume throughout that the functions $\{\phi_j\}$ are convex and β-smooth:

$$\phi_j(s) + \phi_j'(s)t \leq \phi_j(s+t) \leq \phi_j(s) + \phi_j'(s)t + \frac{\beta}{2}t^2, \qquad \text{for all} \qquad s,t \in \mathbb{R}. \tag{2}$$

The *dual problem* of (1) is

$$\max_{\alpha\in\mathbb{R}^n} \left[D(\alpha) := \; -\frac{1}{2\lambda n^2}\|\mathbf{X}\alpha\|_2^2 - \frac{1}{n}\sum_{j=1}^n \phi_j^*(-\alpha_j) \right], \tag{3}$$

where $\phi_j^* : \mathbb{R} \to \mathbb{R}$ is the convex conjugate of ϕ_j, defined by $\phi_j^*(s) := \sup\{st - \phi_j(t) \; : \; t \in \mathbb{R}\}$. It is well known that [24, 15] that $P(w) \geq D(\alpha)$ for every pair $(w,\alpha) \in \mathbb{R}^d \times \mathbb{R}^n$ and $P(w^*) = D(\alpha^*)$. Moreover, the primal and dual optimal solutions, w^* and α^*, respectively, are unique, and satisfy the relations $w^* = \frac{1}{\lambda n}\mathbf{X}\alpha^*$ and $\alpha_j^* = \phi_j'(\langle \mathbf{X}_{:j}, w^*\rangle)$ for all $j \in [n] := \{1,\ldots,n\}$, which also uniquely characterize them.

3 Primal and Dual RCD

In its general "arbitrary sampling" form [16], RCD applied to the primal problem (1) has the form

$$w_i^{k+1} \leftarrow w_i^k - \frac{1}{u_i'}\nabla_i P(w^k) \quad \text{for} \quad i \in S_k, \qquad\qquad w_i^{k+1} \leftarrow w_k^k \quad \text{for} \quad i \notin S_k, \tag{4}$$

where $u_1', \dots, u_d' > 0$ are parameters of the method and $\nabla_i P(w) = \frac{1}{n}\sum_{j=1}^n \phi_j'(\langle \mathbf{X}_{:j}, w\rangle)\mathbf{X}_{ij} + \lambda w_i$ is the ith partial derivative of P at w. This update is performed for a random subset of the coordinates $i \in S_k \subseteq [d]$ chosen in an i.i.d. fashion according to some sampling $\hat{S}_P$. The parameters u_i' are computed ahead of the iterative process and need to be selected carefully in order for the method to work [16, 14]. Specifically, one can set $u_i' := \frac{\beta}{n}u_i + \lambda$, where $u = (u_1, \dots, u_d)$ is chosen so as to satisfy the ESO (expected separable overapproximation) inequality

$$\mathbf{P} \circ \mathbf{X}\mathbf{X}^\top \preceq \mathrm{Diag}(p \circ u), \tag{5}$$

where $\mathbf{P}$ is the $d \times d$ matrix with entries $\mathbf{P}_{ij} = \mathbb{P}(i \in \hat{S}, j \in \hat{S})$, $p = \mathrm{Diag}(\mathbf{P}) \in \mathbb{R}^d$ and $\circ$ denotes the Hadamard (element-wise) product of matrices. The method is formally described as Algorithm 1.

Algorithm 1 Primal RCD: NSync [16]

```
Input: initial iterate $w^0 \in \mathbb{R}^d$; sampling $\hat{S}_P$; ESO parameters $u_1, \dots, u_d > 0$
Initialize: $z^0 = \mathbf{X}^\top w^0$
for $k = 0, 1, \dots$ do
  Sample $S_k \subseteq [d]$ according to $\hat{S}_P$
  for $i \in S_k$ do
    Compute $\Delta_i^k = -\frac{n}{\beta u_i + \lambda n}\left(\frac{1}{n}\sum_{j=1}^n \phi_j'(z_j^k)\mathbf{X}_{ij} + \lambda w_i^k\right)$
    Update $w_i^{k+1} = w_i^k + \Delta_i^k$
  end for
  for $i \notin S_k$ do
    $w_i^{k+1} = w_i^k$
  end for
  Update $z^{k+1} = z^k + \sum_{i \in S_k} \Delta_i^k \mathbf{X}_{i:}^\top$
end for
```

When applying RCD to the dual problem 3, we can't proceed as above since the functions ϕ_j^* are not necessarily smooth, and hence we can't compute the partial derivatives of the dual objective. The standard approach here is to use a proximal variant of RCD [19]. In particular, Algorithm 2 has been analyzed in [15]. Like Algorithm 1, Algorithm 2 is also capable to work with an arbitrary sampling, which in this case is a random subset of $[n]$. The ESO parameters $v = (v_1, \dots, v_j)$ must in this case satisfy the ESO inequality

$$\mathbf{Q} \circ \mathbf{X}^\top\mathbf{X} \preceq \mathrm{Diag}(q \circ v), \tag{6}$$

where $\mathbf{Q}$ is an $n \times n$ matrix with entries $\mathbf{Q}_{ij} = \mathbb{P}(i \in \hat{S}_D, j \in \hat{S}_D)$ and $q = \mathrm{Diag}(\mathbf{Q}) \in \mathbb{R}^n$.

Algorithm 2 Dual RCD: Quartz [15]

```
Input: initial dual variables $\alpha^0 \in \mathbb{R}^n$, sampling $\hat{S}_D$; ESO parameters $v_1, \ldots, v_n > 0$
Initialize: set $w^0 = \frac{1}{\lambda n}\mathbf{X}\alpha^0$
for $k = 0, 1, \ldots$ do
    Sample $S_k \subseteq [n]$ according to $\hat{S}_D$
    for $j \in S_k$ do
        Compute $\Delta_j^k = \arg\max_{h \in \mathbb{R}} \left\{ -\phi_j^*(-(\alpha_j^k + h)) - h\langle \mathbf{X}_{:j}, w^k \rangle - \frac{v_j h^2}{2\lambda n} \right\}$
        Update $\alpha_j^{k+1} = \alpha_j^k + \Delta_j^k$
    end for
    for $j \notin S_k$ do
        $\alpha_j^{k+1} = \alpha_j^k$
    end for
    Update $w^{k+1} = w^k + \frac{1}{\lambda n}\sum_{j \in S_k} \Delta_j^k \mathbf{X}_{:j}$
end for
```

If we assume that $|\hat{S}_P| = 1$ (resp. $|\hat{S}_D| = 1$) with probability 1 (i.e., of the samplings are "serial"), then it is trivial to observe that (5) (resp. (6)) holds with

$$u = \mathrm{Diag}(\mathbf{X}\mathbf{X}^\top) \qquad (\text{resp. } v = \mathrm{Diag}(\mathbf{X}^\top\mathbf{X})). \tag{7}$$

Easily computable expressions for u (resp. v) for more complicated samplings can be found in [14].

4 Iteration Complexity and Total Arithmetic Complexity

In this section we give expressions for the total expected arithmetic complexity of the two algorithms.

4.1 Number of iterations

Iteration complexity of Algorithms 1 and 2 is described in the following theorem. We include a proof sketch in the appendix.

Theorem 1. ***(Complexity: Primal vs Dual RCD)*** *Let $\{\phi_j\}$ be convex and β-smooth.*
(i) If $\hat{S}_P$ is proper (i.e., $p_i > 0$ for all i), and u satisfies (5), then iterates of primal RCD satisfy

$$k \geq K_P = K_P(\hat{S}_P, \epsilon) := \max_{i \in [d]} \left(\frac{\beta u_i + \lambda n}{p_i \lambda n} \right) \log \left(\frac{C_P}{\epsilon} \right) \quad \Rightarrow \quad \mathbb{E}[P(w^k) - P(w^*)] \leq \epsilon, \tag{8}$$

where C_P is a constant depending on w^0 and w^*.
(ii) If $\hat{S}_D$ is proper (i.e., $q_i > 0$ for all i), and v satisfies (6), then iterates of dual RCD satisfy

$$k \geq K_D = K_D(\hat{S}_D, \epsilon) := \max_{j \in [n]} \left(\frac{\beta v_j + \lambda n}{q_j \lambda n} \right) \log \left(\frac{C_D}{\epsilon} \right) \quad \Rightarrow \quad \mathbb{E}[P(w^k) - P(w^*)] \leq \epsilon, \tag{9}$$

where C_D is a constant depending on w^0 and w^*.

For the dual method a stronger guarantee can be established (see [15]): as soon as $k \geq L_D(\hat{S}_D, \epsilon)$, we have $\mathbb{E}[P(w^k) - D(\alpha^k)] \leq \epsilon$. Clearly, this stronger result implies the claim in part ii) of the above theorem.

4.2 Average cost of a single iteration

Let $\|\cdot\|_0$ be the number of nonzeros in a matrix/vector. It is easy to observe that the average cost of a single iteration of Algorithm 1 is

$$W_P(\mathbf{X}, \hat{S}_P) := \mathcal{O}\left(\mathbb{E}\left[\sum_{i\in\hat{S}_P} \|\mathbf{X}_{i:}\|_0\right]\right) = \mathcal{O}\left(\sum_{i=1}^{d} p_i \|\mathbf{X}_{i:}\|_0\right), \tag{10}$$

and for Algorithm 2 it is

$$W_D(\mathbf{X}, \hat{S}_D) := \mathcal{O}\left(\mathbb{E}\left[\sum_{j\in\hat{S}_D} \|\mathbf{X}_{:j}\|_0\right]\right) = \mathcal{O}\left(\sum_{j=1}^{n} q_j \|\mathbf{X}_{:j}\|_0\right). \tag{11}$$

We remark that the constant hidden in $\mathcal{O}$ may be larger for Algorithm 1 than for Algorithm 2. The reason for this is that for Algorithm 1 we compute the one-dimensional derivative ϕ_j' for every nonzero term in the sum, while for Algorithm 2 we do this only once. Depending on the loss ϕ_j, this may lead to slower iterations. For example, if ϕ_j isthe logistic loss, experimentation shows that the constant is around 50. On the other hand, if ϕ_j is the squared loss, the constant is 1.

4.3 Total complexity

By combining the bounds on the number of iterations provided by Theorem 1 with the formulas (10) and (11) for the cost of a single iteration we obtain the following expressions for the *total complexity* of the two algorithms, where we ignore the logarithmic terms and drop the $\tilde{O}$ symbol:

$$T_P = T_P(\mathbf{X}, \hat{S}_P) := K_P W_P \overset{(8)+(10)}{=} \left(\max_{i\in[d]} \frac{\beta u_i + \lambda n}{p_i \lambda n}\right)\left(\sum_{i=1}^{d} p_i \|\mathbf{X}_{i:}\|_0\right), \tag{12}$$

$$T_D = T_D(\mathbf{X}, \hat{S}_D) := K_D W_D \overset{(9)+(11)}{=} \left(\max_{j\in[n]} \frac{\beta v_j + \lambda n}{q_j \lambda n}\right)\left(\sum_{j=1}^{n} q_j \|\mathbf{X}_{:j}\|_0\right). \tag{13}$$

5 Choosing a Sampling that Minimizes the Total Complexity

In this section we identify the *optimal sampling* in terms of the *total complexity*. This is different from previous results on *importance sampling*, which neglect to take into account the cost of the iterations [16, 15, 27, 11]. For simplicity, we shall only consider *serial* samplings, i.e., samplings which only pick a single coordinate at a time. The situation is much more complicated with non-serial samplings where first importance sampling results have only been derived recently [1].

5.1 Uniform Sampling

The simplest serial sampling is the *uniform sampling*: it selects every coordinate with the same probability, i.e. $p_i = 1/d$, $\forall i \in [d]$ and $q_j = 1/n$, $\forall j \in [n]$. In view of (12), (13) and (7), we get

$$T_P = \|\mathbf{X}\|_0 \left(1 + \frac{\beta}{\lambda n} \max_{i\in[d]} \|\mathbf{X}_{i:}\|_2^2\right)$$

and

$$T_D = \|\mathbf{X}\|_0 \left(1 + \frac{\beta}{\lambda n} \max_{j\in[n]} \|\mathbf{X}_{:j}\|_2^2\right).$$

We can now clearly see that whether $T_P \leq T_D$ or $T_P \geq T_D$ depends does not simply depend on d vs n, but instead depends on the relative value of the quantities $\max_{i\in[d]} \|\mathbf{X}_{i:}\|_2^2$ and $\max_{j\in[n]} \|\mathbf{X}_{:j}\|_2^2$. Having said that, we shall not study these quantities in this paper. The reason for this is that for the cake of brevity, we shall instead focus on comparing the primal and dual RCD methods for optimal sampling which minimizes the total complexity, in which case we will obtain different quantities.

5.2 Importance Sampling

By *importance sampling* we mean the serial sampling $\hat{S}_P$ (resp. $\hat{S}_D$) which minimizes the bounds K_P in 8 (resp. K_D in (9)). It can easily be seen (see also [16], [15], [27]), that importance sampling probabilities are given by

$$p_i^* = \frac{\beta u_i + \lambda n}{\sum_l (\beta u_l + \lambda n)} \qquad \text{and} \qquad q_j^* = \frac{\beta v_j + \lambda n}{\sum_l (\beta v_l + \lambda n)}. \tag{14}$$

On the other hand, one can observe that the average iteration cost of importance sampling may be larger than the average iteration cost of uniform serial sampling. Therefore, it is a natural question to ask, whether it is necessarily better. In view of (12), (13) and (14), the total complexities for importance sampling are

$$T_P = \|\mathbf{X}\|_0 + \frac{\beta}{\lambda n}\sum_{i=1}^{d} \|\mathbf{X}_{i:}\|_0 \|\mathbf{X}_{i:}\|_2^2, \qquad T_D = \|\mathbf{X}\|_0 + \frac{\beta}{\lambda n}\sum_{j=1}^{n} \|\mathbf{X}_{:j}\|_0 \|\mathbf{X}_{:j}\|_2^2. \tag{15}$$

Since a weighted average is smaller than the maximum, the total complexity of both methods with importance sampling is always better than with uniform sampling. However, this does not mean that importance sampling is the sampling that minimizes total complexity.

5.3 Optimal Sampling

The next theorem states that, in fact, importance sampling *does* minimize the total complexity.

Theorem 2. *The optimal serial sampling (i.e., the serial sampling minimizing the total expected complexity T_P (resp, T_D)) is the importance sampling* (14).

6 The Face-Off

In this section we investigate the two quantities in (15), T_P and T_D, measuring the total complexity of the two methods as functions of the data $\mathbf{X}$. Clearly, it is enough to focus on the quantities

$$C_P(\mathbf{X}) := \sum_{i=1}^{d} \|\mathbf{X}_{i:}\|_0 \|\mathbf{X}_{i:}\|^2 \qquad \text{and} \qquad C_D(\mathbf{X}) := \sum_{j=1}^{n} \|\mathbf{X}_{:j}\|_0 \|\mathbf{X}_{:j}\|^2. \tag{16}$$

We shall ask questions such as: when is $C_P(\mathbf{X})$ larger/smaller than $C_D(\mathbf{X})$, and by how much. In this regard, it is useful to note that $C_P(\mathbf{X}) = C_D(\mathbf{X}^\top)$. Our first result gives tight lower and upper bounds on their ratio.

Theorem 3. *For any $\mathbf{X} \in \mathbb{R}^{d\times n}$ with no zero rows or columns, we have the bounds $\|\mathbf{X}\|_F^2 \leq C_P(\mathbf{X}) \leq n\|\mathbf{X}\|_F^2$ and $\|\mathbf{X}\|_F^2 \leq C_D(\mathbf{X}) \leq d\|\mathbf{X}\|_F^2$. It follows that $1/d \leq C_P(\mathbf{X})/C_D(\mathbf{X}) \leq n$. Moreover, all these bounds are tight.*

Since $C_P(\mathbf{X})$ (resp. $C_D(\mathbf{X})$) can dominate the expression (12) (resp. (13)) for total complexity, it follows that, depending on the data matrix $\mathbf{X}$, *the primal method can be up to d times faster than the dual method, and up to n times slower than the dual method.*

6.1 Random Data and Dense Data

Assume now that the entries of $\mathbf{X}$ are chosen in an i.i.d. manner from some distribution with mean μ and variance σ^2. While this is not a realistic scenario, it will help us build intuition about what we can expect the quantities $C_P(\mathbf{X})$ and $C_D(\mathbf{X})$ to look like. A simple calculation reveals that $\mathbb{E}[C_P(\mathbf{X})] = dn\sigma^2 + dn^2\mu^2$, and $\mathbb{E}[C_D(\mathbf{X})] = dn\sigma^2 + nd^2\mu^2$. Hence,

$$\mathbb{E}[C_P(\mathbf{X})] \leq \mathbb{E}[C_D(\mathbf{X})]$$

precisely when $n \leq d$, which means that the primal method is better when $n < d$ and the dual method is better when $n > d$.

If $\mathbf{X}$ is a dense deterministic matrix ($\mathbf{X}_{ij} \neq 0$ for all i, j), then $C_P(\mathbf{X}) = n\|\mathbf{X}\|_F^2$ and $C_D(\mathbf{X}) = d\|\mathbf{X}\|_F^2$, and we reach the same conclusion as for random data: everything boils down to d vs n.

6.2 Binary Data

In this part we identify a class of data matrices for which one can have $C_P \leq C_D$ even if $d \ll n$. This class is by no means exhaustive, and serves as an example which we use to illustrate the phenomenon.

Let $\mathbb{B}^{d\times n}$ denote the set of $d \times n$ matrices $\mathbf{X}$ with (signed) binary elements, i.e., with $\mathbf{X}_{ij} \in \{-1, 0, 1\}$ for all i, j. For $\mathbf{X} \in \mathbb{B}^{d\times n}$, the expressions in (16) can be also written in the form $C_P(\mathbf{X}) = \sum_{i=1}^{d} \|\mathbf{X}_{i:}\|_0^2$ and $C_D(\mathbf{X}) = \sum_{j=1}^{n} \|\mathbf{X}_{:j}\|_0^2$. By $\mathbb{B}_{\neq 0}^{d\times n}$ we denote the set of all matrices in $\mathbb{B}^{d\times n}$ with nonzero columns and rows.

For positive integers a, b we write $\bar{a}_b := b\left\lfloor \frac{a}{b} \right\rfloor$ (i.e., a rounded down to the closest multiple of b). Further, we write

$$R(\alpha, d, n) := U(\alpha, d, n)/L(\alpha, n),$$

where

$$L(\alpha, n) := \frac{1}{n}(\bar{\alpha}_n^2 + (\alpha - \bar{\alpha}_n)(2\bar{\alpha}_n + n))$$

and

$$U(\alpha, d, n) := (d+1)\overline{(\alpha - n)}_{d-1} + n - 1 + (\alpha - n + 1 - \overline{(\alpha - n)}_{d-1})^2.$$

The following is a refinement of Theorem 3 for binary matrices of fixed cardinality α.

Theorem 4. *For all $\mathbf{X} \in \mathbb{B}_{\neq 0}^{d\times n}$ with $\alpha = \|\mathbf{X}\|_0$ we have the bounds $1/R(\alpha, n, d) \leq C_P(\mathbf{X})/C_D(\mathbf{X}) \leq R(\alpha, d, n)$. Moreover, these bounds are tight.*

The above theorem follows from Lemma 9, which we formulate and prove in the Appendix. This lemma establishes formulas for the minimum and maximum of C_D and C_P, subject to the constraint $\|\mathbf{X}\|_0 = \alpha$, in terms of the functions L and U. Further, as we show in Lemma 10 in the Appendix, if $d \geq n$ and $\alpha \geq n^2 + 3n$, then $R(\alpha, d, n) \leq 1$. Likewise, if $n \geq d$ and $\alpha \geq d^2 + 3d$, then $R(\alpha, n, d) \leq 1$. Combined with Theorem 4, this has an interesting consequence, spelled out in the next theorem and its corollary.

Theorem 5. *Let $\mathbf{X} \in \mathbb{B}_{\neq 0}^{d\times n}$. If $d \geq n$ and $\|\mathbf{X}\|_0 \geq n^2 + 3n$, then $C_P(\mathbf{X}) \leq C_D(\mathbf{X})$. By symmetry, if $n \geq d$ and $\|\mathbf{X}\|_0 \geq d^2 + 3d$, then $C_D(\mathbf{X}) \leq C_P(\mathbf{X})$.*

This result says that for binary data, and $d \geq n$, the primal method is better than the dual method even for non-dense data, as long as the the data is "dense enough". Observe that as long as $d \geq n^2 + 3n$, all matrices $\mathbf{X} \in \mathbb{B}_{\neq 0}^{d\times n}$ satisfy $\|\mathbf{X}\|_0 \geq d \geq n^2 + 3n \geq n$. This leads to the following corollary.

Corollary 6. *If $d \geq n^2 + 3n$, then for all $\mathbf{X} \in \mathbb{B}_{\neq 0}^{d\times n}$ we have $C_P(\mathbf{X}) \leq C_D(\mathbf{X})$. By symmetry, if $n \geq d^2 + 3d$, then for all $\mathbf{X} \in \mathbb{B}_{\neq 0}^{d\times n}$ we have $C_D(\mathbf{X}) \leq C_P(\mathbf{X})$.*

In words, the corollary states that for binary data where the number of features (d) is large enough in comparison with the number of examples (n), the primal method will be always better. On the other hand, if

n is large enough, the dual method will be always better. This behavior can be observed in Figure 1. For large enough d, all the values $R(\alpha, d, n)$ are below 1.

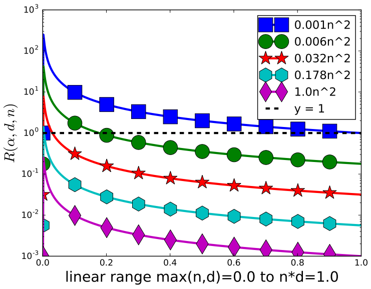


Figure 1: The value $R(\alpha, d, n)$ plotted for $n = 10^3$, $n \leq d \leq n^2$ and $\max\{d, n\} \leq \alpha \leq nd$.

7 Experiments

We conducted experiments on both real and synthetic data. The problem we were interested in is a standard logistic regression with an L2-regularizer, i.e.,

$$P(w) = \frac{1}{n} \sum_{j=1}^{n} \log(1 + \exp(-y_j \langle \mathbf{X}_{:j}, w \rangle)) + \frac{\lambda}{2} \|w\|_2^2.$$

In all our experiments we used $\lambda = 1/n$ and we normalized all the entries of $\mathbf{X}$ by the average column norm. Note that for logistic loss there is no closed form solution for Δ_j^k in Algorithm 2. Therefore we use a variant of Algorithm 2 where $\Delta_j^k = \eta(\phi_j'(\langle \mathbf{X}_{:j}, w \rangle) + \alpha_j^k)$ with the step size η defined as $\eta = \min_{j \in [n]} (q_j \lambda n)/(\beta v_j + \lambda n)$. This variant has the same convergence rate guarantees as Algorithm 2 and does not require exact minimization. Details can be found in [15].

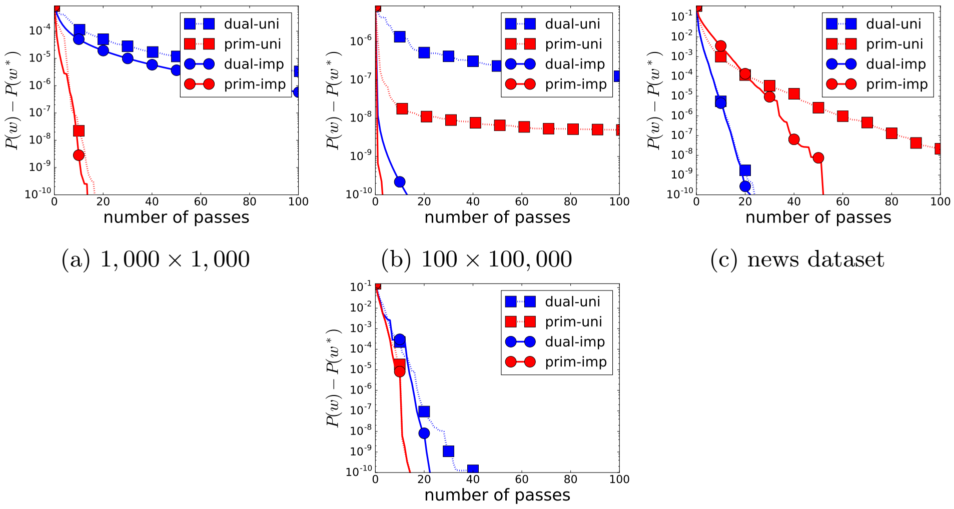
We plot the training error against the number of passes through the data. The number of passes is calculated according to the number of visited nonzero entries in the matrix $\mathbf{X}$. One pass means that we look at $\|\mathbf{X}\|_0$ nonzero entries of $\mathbf{X}$, but not necessarily all of them. We look at the problems from the perspective of the primal approach. The same could be done symmetrically for the dual approach.

7.1 General Data

We look at the matrices which give worst-case bounds for general matrices (Theorem 3) and their empirical properties for different choices of d and n. The corresponding figures are Figure 2a and 2b. For a square dataset, we can observe a large speed-up. For large n we can observe, that the theory holds and the primal method is still faster, but because of numerical issues (we need very small and very large numbers in matrix) and the fact that the optimal value is very close to an "initial guess" of the algorithm, the difference in speed is more difficult to observe.

7.2 Synthetic Binary Data

We looked at matrices with all entries in $\{a, -a, 0\}$ for some $a \neq 0$. We fixed the number of features to be $d = 100$ and we varied the number of examples n and the sparsity level $\alpha = \|\mathbf{X}\|_0$. For each triplet $[d, n, \alpha]$ we produced the worst-case matrix for dual RCD according to the developed theory. The results are in Figure 3.



(a) $1,000 \times 1,000$

(b) $100 \times 100,000$

(c) news dataset

(d) leukemia dataset

Figure 2: Testing the worst case for general matrices and real datasets

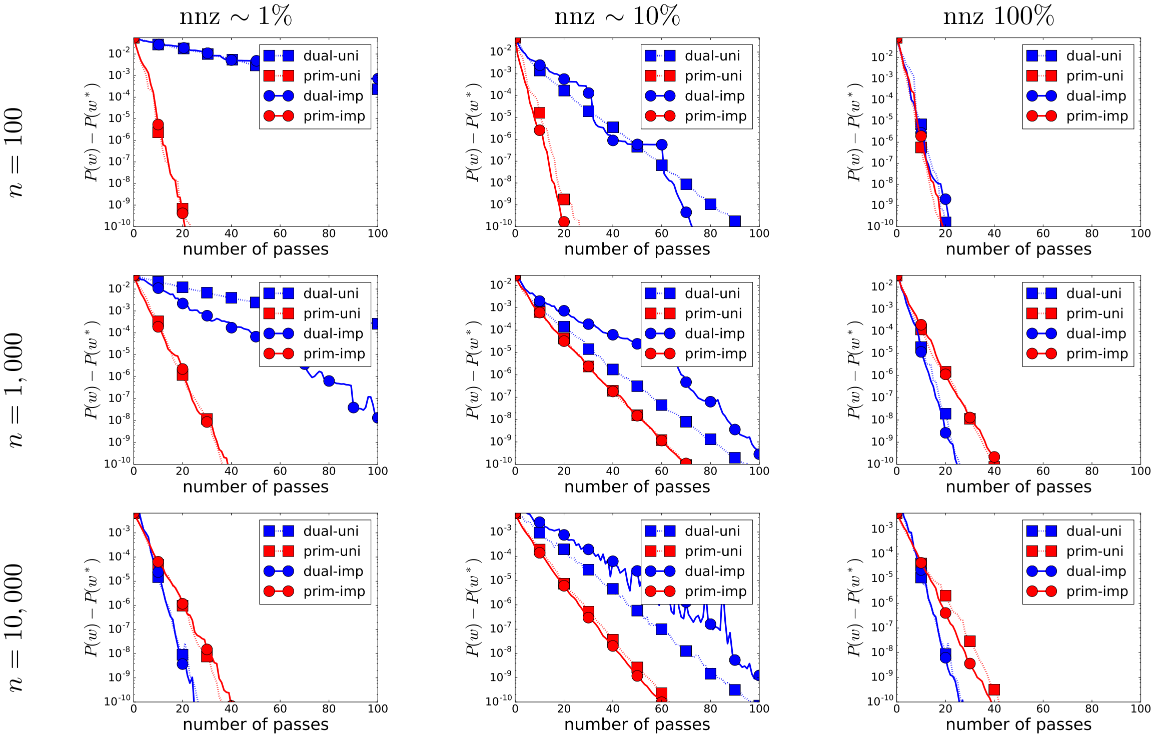


Figure 3: Worst-case experiments with various dimensions and sparsity levels for $d = 100$

dataset	d	n	density	$\|\mathbf{X}\|_0$	C_P	C_D	T_P/T_D
news	1,355,191	19,996	0.03%	9,097,916	3×10^7	9×10^6	2.0
leukemia	7,129	38	100.00%	270,902	1×10^7	2×10^9	0.5

Table 1: Details on the datasets used in the experiments

7.3 Real Data

We used two real datasets to showcase our theory: news and leukemia[2]. The news dataset in Figure 2c is a nice example of our theory in practice. As shown in Table 1 we have $d \gg n$, but the dual method is empirically faster than the primal one. The reason is simple: the news dataset uses a bag of words representation of news articles. If we look at the distribution of features (words), there are many words which appear just very rarely and there are words commonly used in many articles. The features have therefore a very skewed distribution of their nonzero entries. On the other hand, the examples have quite uniform distribution, as the number of distinct words in an article acts nicely. This distribution of nonzero entries highly favors the dual approach, as shown in the theory. The leukemia dataset in Figure 2d is a fully dense dataset and $d \gg n$. Therefore, as our theoretical analysis shows, the primal approach should be better. The ratio between the runtimes is not very large, as the constant $\|\mathbf{X}\|_0$ is of similar order as the additional term in the computation of the true runtime. The empirical speedup in Figures 2c and Figures 2d matches the theoretical predictions from Table 1.

8 Conclusions and Extensions

We have shown that the question whether RCD should be applied to the primal or the dual problem depends on the structure of the training dataset. For dense data, this simply boils down to whether we have more data or parameters, which is intuitively appealing. We have shown, both theoretically, and through experiments with synthetic and real datasets, that contrary to what seems to be a popular belief, primal RCD can outperform dual RCD even if $n \gg d$.

In order to focus on the main message, we have chosen to present our results for simple (as opposed to "accelerated") variants of RCD. However, our results can be naturally extended to accelerated variants of RCD, such as APPROX [4], ASDCA [23], APCG [9], ALPHA [13] and SPDC [26].

Likewise, for simplicity, we focused on serial sampling (i.e., sampling a single coordinate). However, it is possible to use our approach to gain insights into the performance of primal vs dual RCD for arbitrary sampling strategies [16, 15, 13, 14].

References

[1] D. Csiba and P. Richtárik. Importance sampling for minibatches. *arXiv:1602.02283*, 2016.

[2] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS 27*, pages 1646–1654, 2014.

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

[4] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[5] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS 26*, 2013.

[2] both datasets are available from https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[6] J. Konečný, J. Lu, P. Richtárik, and M. Takáč. mS2GD: Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE J. of Selected Topics in Sig. Proc.*, 10(2):242–255, 2016.

[7] J. Konečný, Z. Qu, and P. Richtárik. Semi-stochastic coordinate descent. *arXiv:1412.6293*, 2014.

[8] J. Konečný and P. Richtárik. S2GD: Semi-stochastic gradient descent methods. *arXiv:1312.1666*, 2013.

[9] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. In *NIPS 27*, pages 3059–3067, 2014.

[10] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[11] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *NIPS 27*, pages 1017–1025, 2014.

[12] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[13] Z. Qu and P. Richtárik. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 2016.

[14] Z. Qu and P. Richtárik. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 2016.

[15] Z. Qu, P. Richtárik, and T. Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *NIPS 28*, pages 865–873. 2015.

[16] P. Richtárik and M. Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, pages 1–11, 2015.

[17] P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *JMLR*, pages 1–25, 2016.

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

[19] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484, 2015.

[20] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 1951.

[21] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

[22] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *JMLR*, 12:1865–1892, 2011.

[23] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *NIPS 26*, pages 378–385. 2013.

[24] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *JMLR*, 14(1):567–599, 2013.

[25] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[26] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. ICML, 2015.

[27] P. Zhao and T. Zhang. Stochastic optimization with importance sampling. ICML, 2015.

APPENDIX

Proof of Theorem 1

We say that $P \in \mathcal{C}^1(\mathbf{M})$, if

$$P(w+h) \leq P(w) + \langle \nabla P(w), h\rangle + \frac{1}{2} h^\top \mathbf{M} h, \quad \forall w, h \in \mathbb{R}^d.$$

For three vectors $a, b, c \in \mathbb{R}^n$ we define $\langle a, b\rangle_c := \sum_{i=1}^d a_i b_i c_i$ and $\|a\|_c^2 := \langle a, a\rangle_c = \sum_{i=1}^d c_i a_i^2$. Also, let for $\emptyset \neq S \subseteq [d]$ and $h \in \mathbb{R}^d$, we write $h_S := \sum_{i \in S} h_i e_i$, where e_i is the i-th coordinate vector (i.e., standard basis vector) in $\mathbb{R}^d$.

We will need the following two lemmas.

Lemma 7. *The primal objective P satisfies $P \in \mathcal{C}^1(\mathbf{M})$, where $\mathbf{M} = \lambda \mathbf{I} + \frac{\beta}{n}\mathbf{X}\mathbf{X}^\top$.*

Proof.

$$\begin{aligned}
P(w+h) &\overset{(1)}{=} \frac{1}{n}\sum_{i=1}^n \phi_i(\langle \mathbf{X}_{:i}, w\rangle + \langle \mathbf{X}_{:i}, h\rangle) + \frac{\lambda}{2}\|w+h\|^2 \\
&\overset{(2)}{\leq} \frac{1}{n}\sum_{i=1}^n \left[\phi_i(\langle \mathbf{X}_{:i}, w\rangle) + \phi_i'(\langle \mathbf{X}_{:i}, w\rangle)\cdot\langle \mathbf{X}_{:i}, h\rangle + \frac{\beta}{2}\langle \mathbf{X}_{:i}, h\rangle^2\right] + \frac{\lambda}{2}\|w\|^2 + \lambda\langle w, h\rangle + \frac{\lambda}{2}\|h\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \phi_i(\langle \mathbf{X}_{:i}, w\rangle) + \frac{\lambda}{2}\|w\|^2 + \left\langle \frac{1}{n}\sum_{i=1}^n \phi_i'(\langle \mathbf{X}_{:i}, w\rangle)\mathbf{X}_{:i} + \lambda w\ ,\ h\right\rangle \\
&\quad + \frac{1}{2}h^\top\left(\frac{\beta}{n}\sum_{i=1}^n \mathbf{X}_{:i}(\mathbf{X}_{:i})^\top + \lambda \mathbf{I}\right) h \\
&= P(w) + \langle \nabla P(w), h\rangle + \frac{1}{2}h^\top \mathbf{M} h.
\end{aligned}$$

□

Lemma 8. *If $P \in \mathcal{C}^1(\mathbf{M})$ and $u' \in \mathbb{R}^d$ is such that $\mathbf{P} \circ \mathbf{M} \preceq \mathrm{Diag}(p \circ u')$, then*

$$\mathbb{E}[P(w + h_{[\hat{S}_P]})] \leq P(w) + \langle \nabla P(w), h\rangle_p + \frac{1}{2}\|h\|^2_{p\circ u'}.$$

Proof. See [14], Section 3. □

We can now proceed to the proof of Theorem 1.

First, note that

$$\mathbf{P}\circ\mathbf{M} = \lambda \mathrm{Diag}(p) + \frac{\beta}{n}(\mathbf{P}\circ\mathbf{X}\mathbf{X}^\top) \preceq \lambda\mathrm{Diag}(p) + \frac{\beta}{n}\mathrm{Diag}(p\circ u)$$

with u defined as in (5). We now separately establish the two complexity results; (i) for primal RCD and (ii) for dual RCD.

(i) The proof is a consequence of the proof of the main theorem of [16]. Assumption 1 from [16] holds with $w_i := \lambda + \frac{\beta}{n}u_i$ (Lemma 7 & Lemma 8) and Assumption 2 from [16] holds with standard Euclidean norm and $\gamma := \lambda$. We follow the proof all the way to the bound

$$\mathbb{E}[P(w^k) - P(w^*)] \leq (1-\mu)^k (P(w^0) - P(w^*))$$

which holds for μ defined by

$$\mu := \frac{\lambda}{\max_i \frac{n\lambda+\beta u_i}{np_i}}$$

by direct substitution of the quantities. The result follows by standard arguments. Note that $C_P = P(w^0) - P(w^*)$.

(ii) The proof is a direct consequence of the proof of the main theorem of [15], using the fact that $P(w^k) - P(w^*) \leq P(w^k) - D(\alpha^k)$, as the weak duality holds. Note that $C_D = P(w^0) - D(\alpha^0)$.

Proof of Theorem 2

The proofs for Algorithm 1 and Algorithm 2 are analogous, and hence we will establish the result for Algorithm 1 only. For brevity, denote $s_i = \beta u_i + \lambda n$. We aim to solve the optimization problem:

$$p^* \quad \leftarrow \quad \mathop{\arg\min}_{p \in \mathbb{R}^d_+ \;:\; \sum_i p_i = 1} \quad T_P \overset{(12)}{=} \left(\max_{i \in [d]} \frac{s_i}{p_i \lambda n} \right) \cdot \sum_{i=1}^d p_i \|\mathbf{X}_{i:}\|_0. \tag{17}$$

First observe, that the problem is homogeneous in p, i.e., if p is optimal, also cp will be optimal for $c > 0$, as the solution will be the same. Using this argument, we can remove the constraint $\sum_i p_i = 1$. Also, we can remove the multiplicative factor $1/(\lambda n)$ from the denominator as it does not change the arg min. Hence we get the simpler problem

$$p^* \quad \leftarrow \quad \mathop{\arg\min}_{p \in \mathbb{R}^d_+} \quad \left[\left(\max_{i \in [d]} \frac{s_i}{p_i} \right) \cdot \sum_{i=1}^d p_i \|\mathbf{X}_{i:}\|_0 \right]. \tag{18}$$

Now choose optimal p and assume that there exist j, k such that $s_j/p_j < s_k/p_k$. By a small decrease in p_j, we will still have $s_j/p_j \leq s_k/p_k$, and hence the term $\max_i s_i/p_i$ stays unchanged. However, the term $\sum_i p_i \|\mathbf{X}_{i:}\|_0$ decreased. This means that the optimal sampling must satisfy $s_i/p_i = \mathit{const}$ for all i. However, this is precisely the importance sampling.

Proof of Theorem 3

By assumption, all rows and columns of $\mathbf{X}$ are nonzero. Therefore, $1 \leq \|\mathbf{X}_{i:}\|_0 \leq n$ and $1 \leq \|\mathbf{X}_{:j}\|_0 \leq d$, and the bounds on C_P and C_D follow by applying this to (16). The bounds for the ratio follow immediately by combining the previous bounds. It remains to establish tightness. For $a, b, c \in \mathbb{R}$, let $\mathbf{X}(a, b, c) \in \mathbb{R}^{d \times n}$ be the matrix defined as follows:

$$\mathbf{X}_{ij}(a, b, c) = \begin{cases} a & i \neq 1 \wedge j = 1 \\ b & i = 1 \wedge j \neq 1 \\ c & i = 1 \wedge j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $\mathbf{X}(a, b, c)$ does not have any zero rows nor columns as long as a, b, c are nonzero. Since $C_P(\mathbf{X}(a, b, c)) = (d-1)a^2 + n(n-1)b^2 + nc^2$ and $C_D(\mathbf{X}(a, b, c)) = d(d-1)a^2 + (n-1)b^2 + dc^2$, one readily sees that

$$\lim_{\substack{b \to 0 \\ c \to 0}} \frac{C_P(\mathbf{X}(a, b, c))}{C_D(\mathbf{X}(a, b, c))} = \frac{1}{d} \qquad \text{and} \qquad \lim_{\substack{a \to 0 \\ c \to 0}} \frac{C_P(\mathbf{X}(a, b, c))}{C_D(\mathbf{X}(a, b, c))} = n.$$

Proof of Theorem 4

We first need a lemma.

Lemma 9. *Let α be an integer satisfying $\max\{d,n\} \leq \alpha \leq dn$ and let L and U be the functions defined in Section 6.2. We have the following identities:*

$$\begin{aligned}
L(\alpha, n) &= \min_{\mathbf{X}\in\mathbb{B}^{d\times n}_{\neq 0}} \{C_D(\mathbf{X}) \;:\; \|\mathbf{X}\|_0 = \alpha\} && (19)\\
L(\alpha, d) &= \min_{\mathbf{X}\in\mathbb{B}^{d\times n}_{\neq 0}} \{C_P(\mathbf{X}) \;:\; \|\mathbf{X}\|_0 = \alpha\} && (20)\\
U(\alpha, d, n) &= \max_{\mathbf{X}\in\mathbb{B}^{d\times n}_{\neq 0}} \{C_D(\mathbf{X}) \;:\; \|\mathbf{X}\|_0 = \alpha\} && (21)\\
U(\alpha, n, d) &= \max_{\mathbf{X}\in\mathbb{B}^{d\times n}_{\neq 0}} \{C_P(\mathbf{X}) \;:\; \|\mathbf{X}\|_0 = \alpha\}. && (22)
\end{aligned}$$

Proof. Let $\mathbf{X} \in \mathbb{B}^{d\times n}_{\neq 0}$ be an arbitrary matrix and let $\omega = (\omega_1, \dots, \omega_n)$, where $\omega_j := \|\mathbf{X}_{:j}\|_0$. Let $\alpha = \|\mathbf{X}\|_0 = \sum_j \omega_j$. Observe that $C_D(\mathbf{X}) = \sum_{j=1}^n \|\mathbf{X}_{:j}\|_0^2 = \|\omega\|_2^2$.

(i) We shall first establish (19). Assume that the exist two columns j, k of $\mathbf{X}$, such that $\omega_j + 2 \leq \omega_k$, i.e., their difference in the number of nonzeros is at least 2. Because $\omega_k > \omega_j$, there has to exist a row which has a nonzero entry in the k-th column and a zero entry in the j-th column. Let $\mathbf{X}'$ be the matrix obtained from $\mathbf{X}$ by switching these two entries. Note that $C_P(\mathbf{X}) = C_P(\mathbf{X}')$. However, we have

$$C_D(\mathbf{X}) - C_D(\mathbf{X}') = \omega_j^2 + \omega_k^2 - (\omega_j + 1)^2 - (\omega_k - 1)^2 = 2\omega_k - 2\omega_j - 2 > 0.$$

It follows that while there exist two such columns, the minimum is not achieved. So, we only need to consider matrices $\mathbf{X}$ for which there exists integer a such that $\omega_j = a$ or $\omega_j = a + 1$ for every j. Let $b = |\{j \;:\; \omega_j = a\}|$.

We can now without loss of generality assume that $0 \leq b \leq n - 1$. Indeed, we can do this is because the choices $b = 0$ and $b = n$ lead to the same matrices, and hence by focusing on $b = 0$ we have not removed any matrices from consideration. With simple calculations we get

$$\alpha = ba + (n - b)(a + 1) = n(a + 1) - b.$$

Note that $\alpha + b$ is a multiple of n. It follows that $b = n - \alpha + \bar{\alpha}_n$ and $a = \bar{\alpha}_n/n$. Up to the ordering of the columns (which does not affect $C_D(\mathbf{X})$) we have just one candidate $\mathbf{X}$, therefore it has to be the minimizer of C_D. Finally, we can easily calculate the minimum as

$$\begin{aligned}
\sum_{j=1}^n \omega_j^2 &= ba^2 + (n - b)(a + 1)^2 = (n - \alpha + \bar{\alpha}_n)\left(\frac{\bar{\alpha}_n}{n}\right)^2 + (\alpha - \bar{\alpha}_n)\left(\frac{\bar{\alpha}_n}{n} + 1\right)^2\\
&= \frac{1}{n}\left(\bar{\alpha}_n^2 + (\alpha - \bar{\alpha}_n)(2\bar{\alpha}_n + n)\right) = L(\alpha, n).
\end{aligned}$$

(ii) Claim (20) follows from part (19) via symmetry: $C_P(\mathbf{X}) = C_D(\mathbf{X}^\top)$ and $\|\mathbf{X}\|_0 = \|\mathbf{X}^\top\|_0$.

(iii) We now establish claim (21). Assume that there exist a pair of columns j, k such that $1 < \omega_j \leq \omega_k < d$. Let $\mathbf{X}'$ be the matrix obtained from $\mathbf{X}$ by zeroing out an entry in the j-th column and putting a nonzero inside the k-th column. Then

$$C_D(\mathbf{X}') - C_D(\mathbf{X}) = (\omega_j - 1)^2 + (\omega_k + 1)^2 - \omega_j^2 - \omega_k^2 = 2\omega_k - 2\omega_j + 2 > 0.$$

It follows that while there exist such a pair of columns, the maximum is not achieved. This condition leaves us with matrices $\mathbf{X}$ where at most one column j has ω_j *not* equal to 1 or d.

Formally, let $a = |\{j \,:\, \omega_j = d\}|$. Then we have $n - a - 1$ columns with 1 nonzero and 1 column with b nonzeros, where $1 \leq b < d$. This is correct, as $b = d$ is the same as $b = 1$ and a being one more. We can compute a and b from the equation

$$\begin{aligned}(n - a - 1) \cdot 1 + 1 \cdot b + a \cdot d &= \alpha \\ b + a(d-1) &= \alpha - n + 1\end{aligned}$$

as the only solution to the division with remainder of $\alpha - n + 1$ by $d - 1$, with the difference that $b \in \{1, \dots, d-1\}$ instead of the standard $\{0, \dots, d-2\}$. We get

$$a = \left\lfloor \frac{a - n}{d - 1} \right\rfloor \qquad \text{and} \qquad b = \alpha - n + 1 - \overline{(\alpha - n)}_{d-1}.$$

The maximum can now be easily computed as follows:

$$\begin{aligned}\sum_{j=1}^{n} \omega_j^2 &= (n - a - 1) + b^2 + ad^2 \\ &= n - \left\lfloor \frac{a-n}{d-1} \right\rfloor - 1 + \left(\alpha - n + 1 - \overline{(\alpha - n)}_{d-1}\right)^2 + \left\lfloor \frac{a-n}{d-1} \right\rfloor d^2 \\ &= U(\alpha, d, n).\end{aligned}$$

(iv) Again, claim (22) follows from (21) via symmetry.

□

We can now proceed to the proof of the theorem.

The quantity is the ratio between the maximal value of C_P and the minimal value of C_D, we have to show that there exists a matrix $\mathbf{X}$ such that this is achieved. Assume we have a matrix $\mathbf{X}$ which has the maximal C_P. In the proof of Lemma 9 we showed, that by switching entries in $\mathbf{X}$ we can get the minimal value of C_D without changing C_P. Therefore we can achieve maximal C_P and minimal C_D at the same time. Analogically for every other case.

Proof of Theorem 5

As shown in the main text, the theorem follows from the following lemma. Hence, we only need to prove the lemma.

Lemma 10. *If $d \geq n$ and $\alpha \geq n^2 + 3n$, then $R(\alpha, d, n) \leq 1$. If $n \geq d$ and $\alpha \geq d^2 + 3d$, then $R(\alpha, n, d) \leq 1$.*

Proof. We focus on the first part, the second follows in an analogous way. Using the two assumptions, we have $\alpha(n^2 + 3n) + n^3 \leq \alpha^2 + dn^2$. By adding $n^2 + n$ to the right hand side and after reshuffling, we obtain the inequality

$$n\left[(n+1)(\alpha - d) + d - 1 + n^2\right] \leq (\alpha - n)^2.$$

For positive scalars $a, b > 0$, we have the trivial estimates $a - b \leq \bar{a}_b := b\lfloor \frac{a}{b} \rfloor \leq a$. We use them to bound four expressions:

$$\begin{aligned}(\alpha - d) &\geq \overline{(\alpha - d)}_{n-1} \\ n^2 &\geq (\alpha - d + 1 - \overline{(\alpha - d)}_{n-1})^2 \\ \bar{\alpha}_n^2 &\geq (\alpha - n)^2 \\ (\alpha - \bar{\alpha}_n)(2\bar{\alpha}_n + n) &\geq 0\end{aligned}$$

Using these bounds one-by-one we get the result

$$\begin{aligned}
n\left[(n+1)(\alpha-d)+d-1+n^2\right] &\leq (\alpha-n)^2 \\
n\left[(n+1)\overline{(\alpha-d)}_{n-1}+d-1+n^2\right] &\leq (\alpha-n)^2 \\
n\left[(n+1)\overline{(\alpha-d)}_{n-1}+d-1+(\alpha-d+1-\overline{(\alpha-d)}_{n-1})^2\right] &\leq (\alpha-n)^2 \\
n\left[(n+1)\overline{(\alpha-d)}_{n-1}+d-1+(\alpha-d+1-\overline{(\alpha-d)}_{n-1})^2\right] &\leq \bar{\alpha}_n^2 \\
n\left[(n+1)\overline{(\alpha-d)}_{n-1}+d-1+(\alpha-d+1-\overline{(\alpha-d)}_{n-1})^2\right] &\leq \bar{\alpha}_n^2+(\alpha-\bar{\alpha}_n)(2\bar{\alpha}_n+n) \\
R(\alpha,d,n) &\leq 1
\end{aligned}$$

□