Accelerated Coordinate Descent with Arbitrary Sampling and Best Rates for Minibatches

Filip Hanzely
KAUST

Peter Richtárik
KAUST, University of Edinburgh, MIPT

Abstract

Accelerated coordinate descent is a widely popular optimization algorithm due to its efficiency on large-dimensional problems. It achieves state-of-the-art complexity on an important class of empirical risk minimization problems. In this paper we design and analyze an accelerated coordinate descent (`ACD`) method which in each iteration updates a random subset of coordinates according to an arbitrary but fixed probability law, which is a parameter of the method. While minibatch variants of `ACD` are more popular and relevant in practice, there is no importance sampling for `ACD` that outperforms the standard uniform minibatch sampling. Through insights enabled by our general analysis, we design new importance sampling for minibatch `ACD` which significantly outperforms previous state-of-the-art minibatch `ACD` in practice. We prove a rate that is at most $\mathcal{O}(\sqrt{\tau})$ times worse than the rate of minibatch `ACD` with uniform sampling, but can be $\mathcal{O}(n/\tau)$ times better, where τ is the minibatch size. Since in modern supervised learning training systems it is standard practice to choose $\tau \ll n$, and often $\tau = \mathcal{O}(1)$, our method can lead to dramatic speedups. Lastly, we obtain similar results for minibatch nonaccelerated `CD` as well, achieving improvements on previous best rates.

1 Introduction

Many key problems in machine learning and data science are routinely modeled as optimization problems and solved via optimization algorithms. With the increase of the volume of data used to formulate optimization models, there is a need for new efficient algorithms able to cope with the challenge. Through intensive research and algorithmic innovation during the last 10-15 years, gradient methods have become the methods of choice for large-scale optimization problems.

In this paper we consider the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where f a smooth and strongly convex function, and the main difficulty comes from the dimension n being very large (e.g., millions or billions). In this regime, *coordinate descent (CD)* variants of gradient methods are the state of the art.

The simplest variant of `CD` in each iterations updates a single variable of x by taking a one dimensional gradient step along the direction of ith unit basis vector $e_i \in \mathbb{R}^n$, which leads to the update rule

$$\boxed{x^{k+1} = x^k - \alpha_i \nabla_i f(x^k) e_i} \tag{2}$$

where $\nabla_i f(x^k) := e_i^\top \nabla f(x^k)$ is the ith partial derivative and α_i is a suitably chosen stepsize. The classical smoothness assumption used in the analysis of `CD` methods (Nesterov, 2012a) is to require the existence of constants $L_i > 0$ such that

$$f(x + t e_i) \leq f(x) + t \nabla_i f(x) + \frac{L_i}{2} t^2 \tag{3}$$

holds for all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$ and $i \in [n] := \{1, 2, \ldots, n\}$. In this setting, one can choose the stepsizes to be $\alpha_i = 1/L_i$.

There are several rules studied in the literature for choosing the coordinate i in iteration k, including cyclic rule (Luo and Tseng, 1992; Tseng, 2001; Saha and Tewari, 2013; Wright, 2015; Gurbuzbalaban et al., 2017), Gauss-Southwell or other greedy rules (Nutini et al., 2015; You et al., 2016; Stich et al., 2017a), random (stationary) rule (Nesterov, 2012a; Richtárik and Takáč, 2014, 2016; Shalev-Shwartz and Zhang, 2014;

Lin et al., 2014; Fercoq and Richtárik, 2015) and adaptive random rules (Csiba et al., 2015; Stich et al., 2017b). In this work we focus on stationary random rules, which are popular by practitioners and well understood in theory.

Updating one coordinate at a time. The simplest randomized CD method of the form (2) chooses coordinate i in each iteration uniformly at random. If f is σ–convex[1], then this method converges in $(n \max_i L_i/\sigma) \log(1/\epsilon)$ iterations in expectation. If index i is chosen with probability $p_i \propto L_i$, then the iteration complexity improves to $(\sum_i L_i/\sigma) \log(1/\epsilon)$. The latter result is always better than the former, and can be up to n times better. These results were established in a seminal paper by Nesterov (2012a). The analysis was later generalized to arbitrary probabilities $p_i > 0$ by Richtárik and Takáč (2014), who obtained the complexity

$$\left(\max_i \frac{L_i}{p_i \sigma}\right) \log \frac{1}{\epsilon}. \qquad (4)$$

Clearly, (4) includes the previous two results as special cases. Note that the importance sampling $p_i \propto L_i$ minimizes the complexity bound (4) and is therefore in this sense optimal.

Minibatching: updating more coordinates at a time. In many situations it is advantageous to update a small *subset (minibatch)* of coordinates in each iteration, which leads to the *minibatch CD method* which has the form

$$x_i^{k+1} = \begin{cases} x_i^k - \alpha_i \nabla_i f(x^k) & i \in S^k, \\ x_i^k & i \notin S^k. \end{cases} \qquad (5)$$

For instance, it is often equally easy to fetch information about a small batch of coordinates S^k from memory at the same or comparable time as it is to fetch information about a single coordinate. If this memory access time is the bottleneck as opposed to computing the actual updates to coordinates $i \in S^k$, then it is more efficient to update all coordinates belonging to the minibatch S^k. Alternatively, in situations where parallel processing is available, one is able to compute the updates to a small batch of coordinates simultaneously, leading to speedups in wall clock time. With this application in mind, minibatch CD methods are also often called *parallel* CD methods (Richtárik and Takáč, 2016).

---

[1]We say that f is σ–convex if it is strongly convex with strong convexity modulus $\sigma > 0$. That is, if $f(x + h) \geq f(x) + (\nabla f(x))^\top h + \frac{\sigma}{2}\|h\|^2$ for all $x, h \in \mathbb{R}^n$, where $\|h\| := (\sum_i h_i^2)^{1/2}$ is the standard Euclidean norm.

2   Arbitrary sampling and minibatching

Arbitrary sampling. Richtárik and Takáč (2016) analyzed method (5) for *uniform samplings* S^k, i.e., assuming that $\mathbb{P}(i \in S^k) = \mathbb{P}(j \in S^k)$ for all i, j. However, the ultimate generalization is captured by the notion of *arbitrary sampling* pioneered by Richtárik and Takáč (2016b). A *sampling* refers to a set-valued random mapping S with values being the subsets of $[n]$. The word *arbitrary* refers to the fact that no additional assumptions on the sampling, such as uniformity, are made. This result generalizes the results mentioned above.

M–smoothness. For minibatch CD methods it is useful to assume a more general notion of smoothness parameterized by a positive semidefinite matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$. We say that f is $\mathbf{M}$–smooth if

$$f(x + h) \leq f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top \mathbf{M} h \qquad (6)$$

for all $x, h \in \mathbb{R}^n$. The standard L–smoothness condition is obtained in the special case when $\mathbf{M} = L\mathbf{I}$, where $\mathbf{I}$ is the identity matrix in $\mathbb{R}^n$. Note that if f is $\mathbf{M}$–smooth, then (3) holds for $L_i = \mathbf{M}_{ii}$. Conversely, it is known that if (3) holds, then (6) holds for $\mathbf{M} = n\mathbf{Diag}(L_1, L_2, \ldots, L_n)$ (Nesterov, 2012a). If h has at most ω nonzero entries, then this result can be strengthened and (6) holds with $\mathbf{M} = \omega\mathbf{Diag}(L_1, L_2, \ldots, L_n)$ (Richtárik and Takáč, 2016, Theorem 8). In many situations, $\mathbf{M}$–smoothness is a very natural assumption. For instance, in the context of empirical risk minimization (ERM), which is a key problem in supervised machine learning, f is of the form $f(x) = \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{A}_i x) + \frac{\sigma}{2}\|x\|^2$, where $\mathbf{A}_i \in \mathbb{R}^{q \times n}$ are data matrices, $f_i : \mathbb{R}^q \to \mathbb{R}$ are loss functions and $\sigma \geq 0$ is a regularization constant. If f_i is convex and γ_i–smooth, then f is σ–convex and $\mathbf{M}$–smooth with $\mathbf{M} = (\frac{1}{m} \sum_i \gamma_i \mathbf{A}_i^\top \mathbf{A}_i) + \sigma\mathbf{I}$ (Qu and Richtárik, 2016). In these situations it is useful to design CD algorithms making full use of the information contained in the data as captured in the smoothness matrix $\mathbf{M}$.

Given a sampling S and $\mathbf{M}$–smooth function f, let $v = (v_1, \ldots, v_n)$ be positive constants satisfying the ESO (expected separable overapproximation) inequality

$$\mathbf{P} \circ \mathbf{M} \preceq \mathbf{Diag}(p_1 v_1, \ldots, p_n v_n), \qquad (7)$$

where $\mathbf{P}$ is the *probability matrix* associated with sampling S, defined by $\mathbf{P}_{ij} := \mathbb{P}(i \in S \ \& \ j \in S)$, $p_i := \mathbf{P}_{ii} = \mathbb{P}(i \in S)$ and $\circ$ denotes the Hadamard (i.e., elementwise) product of matrices. From now on we define the *probability vector* as $p := (p_1, \ldots, p_n) \in \mathbb{R}^n$ and let $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ be the vector of ESO parameters. With this notation, (7) can be equivalently

written as $\mathbf{P} \circ \mathbf{M} \preceq \mathbf{Diag}\,(p \circ v)$. We say that S is *proper* if $p_i > 0$ for all i.

It can be show by combining the results of Richtárik and Takáč (2016b) and Qu and Richtárik (2016) that under the above assumptions, the minibatch `CD` method (5) with stepsizes $\alpha_i = 1/v_i$ enjoys the iteration complexity

$$\left(\max_i \frac{v_i}{p_i \sigma}\right) \log \frac{1}{\epsilon}. \tag{8}$$

Since in situations when $|S^k| = 1$ with probability 1 once can choose $v_i = L_i$, the complexity result (8) generalizes (4). Inequality (7) is standard in minibatch coordinate descent literature. It was studied extensively by Qu and Richtárik (2016), and has been used to analyze parallel `CD` methods (Richtárik and Takáč, 2016; Richtárik and Takáč, 2016b; Fercoq and Richtárik, 2015), distributed `CD` methods (Richtárik and Takáč, 2016a; Fercoq et al., 2014), accelerated `CD` methods (Fercoq and Richtárik, 2015; Fercoq et al., 2014; Qu and Richtárik, 2016; Chambolle et al., 2017), and dual methods (Qu et al., 2015; Chambolle et al., 2017).

Importance sampling for minibatches. It is easy to see, for instance, that if we do not restrict the class of samplings over which we optimize, then the trivial *full sampling* $S^k = [n]$ with probability 1 is optimal. For this sampling, $\mathbf{P}$ is the matrix of all ones, $p_i = 1$ for all i, and (7) holds for $v_i = L := \lambda_{\max}(\mathbf{M})$ for all i. The minibatch `CD` method (5) reduces to gradient descent, and the complexity estimate (8) becomes $(L/\sigma)\log(1/\epsilon)$, which is the standard rate of gradient descent. However, typically we are interested in finding the best sampling from the class of samplings which use a minibatch of size τ, where $\tau \ll n$. While we have seen that the importance sampling $p_i = L_i / \sum_j L_j$ is optimal for $\tau = 1$, in the minibatch case $\tau > 1$ the problem of determining a sampling which minimizes the bound (8) is much more difficult. For instance, Richtárik and Takáč (2016b) consider a certain parametric family of samplings where the problem of finding the best sampling from this family reduces to a linear program.

Surprisingly, and in contrast to the situation in the $\tau = 1$ case where an optimal sampling is known and is in general non-uniform, there is no minibatch sampling that is guaranteed to outperform τ–nice sampling. We say that S is τ–nice if it samples uniformly from among all subsets of $[n]$ of cardinality τ. The probability matrix of this sampling is given by $\mathbf{P} = \frac{\tau}{n}\left((1-\beta)\mathbf{I} + \beta\mathbf{E}\right)$, where $\beta = \frac{\tau-1}{n-1}$ (assume $n > 1$) and $\mathbf{E}$ is the matrix of all ones, and $p_i = \frac{\tau}{n}$ (Qu and Richtárik, 2016). It follows that the ESO inequality (7) holds for $v_i = (1-\beta)\mathbf{M}_{ii} + \beta L$. By plugging

into (8), we get the iteration complexity

$$\frac{n}{\tau}\left(\frac{(1-\beta)\max_i \mathbf{M}_{ii} + \beta L}{\sigma}\right)\log\frac{1}{\epsilon}. \tag{9}$$

This rate interpolates between the rate of `CD` with uniform probabilities (for $\tau = 1$) and the rate of gradient descent (for $\tau = n$).

3 Contributions

For *accelerated coordinate descent (`ACD`)* without minibatching (i.e., when $\tau = 1$), the currently best known iteration complexity result, due to Allen-Zhu et al. (2016), is

$$\mathcal{O}\left(\frac{\sum_i \sqrt{L_i}}{\sqrt{\sigma}}\log\frac{1}{\epsilon}\right). \tag{10}$$

The probabilities used in the algorithm are proportional to the square roots of the coordinate-wise Lipschitz constants: $p_i \propto \sqrt{L_i}$. This is the first `CD` method with a complexity guarantee which does not explicitly depend on the dimension n, and is an improvement on the now-classical result of Nesterov (2012a) giving the complexity

$$\mathcal{O}\left(\sqrt{\frac{n\sum_i L_i}{\sigma}}\log\frac{1}{\epsilon}\right).$$

The rate (10) is always better than this, and can be up to $\sqrt{n}$ times better if the distribution of L_i is extremely non-uniform. Unlike in the non-accelerated case described in the previous section, there is no complexity result for `ACD` with general probabilities such as (4), or with an arbitrary sampling such as (8). In fact, an `ACD` method was not even designed in such settings, despite a significant recent development in accelerated coordinate descent methods (Nesterov, 2012b; Lee and Sidford, 2013; Lin et al., 2014; Qu and Richtárik, 2016; Allen-Zhu et al., 2016).

Our key contributions are:

$\diamond$ `ACD` **with arbitrary sampling.** We design an `ACD` method which is able to operate with an *arbitrary sampling* of subsets of coordinates. We describe our method in Section 4.

$\diamond$ **Iteration complexity.** We prove (see Theorem 4.2) that the iteration complexity of `ACD` is

$$\mathcal{O}\left(\sqrt{\max_i \frac{v_i}{p_i^2 \sigma}}\log\frac{1}{\epsilon}\right), \tag{11}$$

where v_i are ESO parameters given by (7) and $p_i > 0$ is the probability that coordinate i belongs to the sampled set S^k: $p_i := \mathbb{P}(i \in S^k)$. The result of Allen-Zhu et al. (10) (`NUACDM`) can be recovered as a special case of (11) by focusing on samplings defined by $S^k = \{i\}$ with

Table 1: Complexity results for non-accelerated (CD) and accelerated (ACD) coordinate descent methods for σ–convex functions and arbitrary sampling S. The last row corresponds to the setup with arbitrary proper sampling S (i.e., a random subset of $[n]$ with the property that $p_i := \mathbb{P}(i \in S) > 0$). We let $\tau := \mathbf{E}[|S|]$ be the expected mini-batch size. We assume that f is $\mathbf{M}$–smooth (see (6)). The positive constants $v_1, v_2, \ldots, v_n$ are the ESO parameters (depending on f and S), defined in (7). The first row arises as a special of the third row in the non-minibatch (i.e., $\tau = 1$) case. Here we have $v_i = L_i := \mathbf{M}_{ii}$. The second row is a special case of the first row for the optimal choice of the probabilities $p_1, p_2, \ldots, p_n$.

	CD	ACD
$\tau = 1, p_i > 0$	$\left(\max_i \frac{L_i}{p_i \sigma}\right) \log \frac{1}{\epsilon}$ (Richtárik and Takáč, 2014)	$\sqrt{\max_i \frac{L_i}{p_i^2 \sigma}} \log \frac{1}{\epsilon}$ **(this paper)**
$\tau = 1$, best p_i	$\frac{\sum_i L_i}{\sigma} \log \frac{1}{\epsilon}; \quad p_i \propto L_i$ (Nesterov, 2012a)	$\frac{\sum_i \sqrt{L_i}}{\sqrt{\sigma}} \log \frac{1}{\epsilon}; p_i \propto \sqrt{L_i}$ (Allen-Zhu et al., 2016)
arbitrary sampling S	$\left(\max_i \frac{v_i}{p_i \sigma}\right) \log \frac{1}{\epsilon}$ (Richtárik and Takáč, 2016b)	$\sqrt{\max_i \frac{v_i}{p_i^2 \sigma}} \log \frac{1}{\epsilon}$ **(this paper)**

probability $p_i \propto \sqrt{L_i}$ (recall that in this case $v_i = L_i$). When $S^k = [n]$ with probability 1, then our method reduces to accelerated gradient descent (AGD) (Nesterov, 1983, 2004), and since $p_i = 1$ and $v_i = L$ (the Lipschitz constant of ∇f) for all i, (11) reduces to the standard complexity of AGD: $\mathcal{O}(\sqrt{L/\sigma} \log(1/\epsilon))$.

⋄ **Weighted strong convexity.** We prove a slightly more general result than (11) in which we allow the strong convexity of f to be measured in a weighted Euclidean norm with weights v_i/p_i^2. In situations when f is naturally strongly convex with respect to a weighted norm, this more general result will typically lead to a better complexity result than (11), which is fine-tuned for standard strong convexity. There are applications when f is naturally a strongly convex with respect to some weighted norm (Allen-Zhu et al., 2016).

⋄ **Minibatch methods.** We design several *new* importance samplings for minibatches, calculate the associated complexity results, and show through experiments that they significantly outperform the standard uniform samplings used in practice and constitute the state of the art. Our importance sampling leads to rates which are provably within a small factor from the best known rates, but can lead to an improvement by a factor of $\mathcal{O}(n)$. We are the first to establish such a result, both for CD (Appendix B) and ACD (Section 5).

The key complexity results obtained in this paper are summarized and compared to prior results in Table 1.

4 The algorithm

The accelerated coordinate descent method (ACD) we propose is formalized as Algorithm 1. If we removed (13) and (16) from the method, and replaced y^{k+1} in (15) by x^{k+1}, we would recover the CD method. Acceleration is obtained by the inclusion of the extrapolation

steps (13) and (16). As mentioned before, we will analyze our method under a more general strong convexity assumption.

Assumption 4.1. *f is σ_w–convex with respect to the $\|\cdot\|_w$ norm. That is,*

$$f(x + h) \geq f(x) + \langle \nabla f(x), h \rangle + \frac{\sigma_w}{2} \|h\|_w^2, \quad (12)$$

for all $x, h \in \mathbb{R}^n$, where $\sigma_w > 0$.

Note that if f is σ–convex in the standard sense (i.e., for $w = (1, \ldots, 1)$), then f is σ_w–convex for any $w > 0$ with $\sigma_w = \min_i \frac{\sigma}{w_i}$. Considering a general σ_w–convexity allows us to get a tighter convergence rate in some cases (Allen-Zhu et al., 2016).

Algorithm 1 ACD (Accelerated coordinate descent with arbitrary sampling)

Input: i.i.d. proper samplings $S^k \sim \mathcal{D}$; ESO parameters $v \in \mathbb{R}_{++}^n$; $p_i = \mathbb{P}(i \in S^k)$ and $w_i = v_i/p_i^2$ for all $i \in [n]$; strong convexity constant $\sigma_w > 0$; stepsize parameters $\theta \approx 0.618\sqrt{\sigma_\omega}$ (see (20)) and $\eta = 1/\theta$
Initialize: Initial iterate $y^0 = z^0 \in \mathbb{R}^n$
for $k = 0, 1 \ldots$ **do**

$$x^{k+1} = (1 - \theta)y^k + \theta z^k \qquad (13)$$

$$\text{Get } S^k \sim \mathcal{D} \qquad (14)$$

$$y^{k+1} = x^{k+1} - \sum_{i \in S^k} \frac{1}{v_i} \nabla_i f(x^{k+1}) e_i \qquad (15)$$

$$z^{k+1} = \frac{1}{1 + \eta\sigma_w}\Big(z^k + \eta\sigma_w x^{k+1}$$
$$\qquad\qquad - \sum_{i \in S^k} \frac{\eta}{p_i w_i} \nabla_i f(x^{k+1}) e_i\Big) \quad (16)$$

end

Using the tricks developed by Lee and Sidford (2013); Fercoq and Richtárik (2015); Lin et al. (2014), Algo-

rithm 1 can be implemented so that only $|S^k|$ coordinates are updated in each iteration. We are now ready derive a convergence rate of ACD.

Theorem 4.2 (Convergence of ACD)**.** *Let S^k be i.i.d. proper (but otherwise arbitrary) samplings. Let $\mathbf{P}$ be the associated probability matrix and $p_i := \mathbb{P}(i \in S^k)$. Assume f is $\mathbf{M}$–smooth (see (6)) and let v be ESO parameters satisfying (7). Further, assume that f is σ_w–convex (with $\sigma_w > 0$) for*

$$w_i := \frac{v_i}{p_i^2}, \qquad i = 1, 2, \dots, n, \tag{17}$$

with respect to the weighted Euclidean norm $\|\cdot\|_w$ (i.e., we enforce Assumption 4.1). Then

$$\sigma_w \le \frac{\mathbf{M}_{ii} p_i^2}{v_i} \le p_i^2 \le 1, \qquad i = 1, 2, \dots n. \tag{18}$$

In particular, if f is σ–convex with respect to the standard Euclidean norm, then we can choose

$$\sigma_w = \min_i \frac{p_i^2 \sigma}{v_i}. \tag{19}$$

Finally, if we choose

$$\begin{aligned} \theta \quad &:= \quad \frac{\sqrt{\sigma_w^2 + 4\sigma_w} - \sigma_w}{2} = \frac{2\sigma_w}{\sqrt{\sigma_w^2 + 4\sigma_w} + \sigma_w} \\ &\ge \quad 0.618\sqrt{\sigma_\omega} \end{aligned} \tag{20}$$

and $\eta := \frac{1}{\theta}$, then the random iterates of ACD satisfy

$$\mathbb{E}\left[P^k\right] \le (1 - \theta)^k P^0, \tag{21}$$

where $P^k := \frac{1}{\theta^2}\left(f(y^k) - f(x^*)\right) + \frac{1}{2(1-\theta)}\|z^k - x^*\|_w^2$ and x^* is the optimal solution of (1).

Noting that $1/0.618 \le 1.619$, as an immediate consequence of (21) and (20) we get bound

$$k \ge \frac{1.619}{\sqrt{\sigma_w}} \log \frac{1}{\epsilon} \quad \Rightarrow \quad \mathbb{E}\left[P^k\right] \le \epsilon P^0. \tag{22}$$

If f is σ–convex, then by plugging (19) into (22) we obtain the iteration complexity bound

$$1.619 \cdot \sqrt{\max_i \frac{v_i}{p_i^2 \sigma}} \log \frac{1}{\epsilon}. \tag{23}$$

Complexity (23) is our key result (also mentioned in (11) and Table 1).

5 Importance sampling for minibatches

Let $\tau := \mathbb{E}\left[|S^k|\right]$ be the expected minibatch size. The next theorem provides an insightful lower bound for the complexity of ACD we established, one independent of p and v.

Theorem 5.1 (Limits of minibatch performance)**.** *Let the assumptions of Theorem 4.2 be satisfied and let f be σ–convex. Then the dominant term in the rate (23) of ACD admits the lower bound*

$$\sqrt{\max_i \frac{v_i}{p_i^2 \sigma}} \ge \frac{\sum_i \sqrt{\mathbf{M}_{ii}}}{\tau \sqrt{\sigma}}. \tag{24}$$

Note that for $\tau = 1$ we have $\mathbf{M}_{ii} = v_i = L_i$, and the lower bound is achieved by using the importance sampling $p_i \propto \sqrt{L_i}$. Hence, this bound gives a limit on how much speedup, compared to the best known complexity in the $\tau = 1$ case, we can hope for as we increase τ. The bound says we can not hope for better than linear speedup in the minibatch size. An analogous result (obtained by removing all the squares and square roots in (24)) was established by Richtárik and Takáč (2016b) for CD.

In what follows, it will be useful to write the complexity result (23) in a new form by considering a specific choice of the ESO vector v.

Lemma 5.2. *Choose any proper sampling S and let $\mathbf{P}$ be its probability matrix and p its probability vector. Let $c(S, \mathbf{M}) := \lambda_{\max}(\mathbf{P}' \circ \mathbf{M}')$, where $\mathbf{P}' := \mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{-1/2}$, $\mathbf{M}' := \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}$ and $\mathbf{D} := \mathbf{Diag}(p)$. Then the vector v defined by $v_i := c(S, \mathbf{M})p_i^2$ satisfies the ESO inequality (7) and the total complexity (23) becomes*

$$1.619 \cdot \frac{\sqrt{c(S, \mathbf{M})}}{\sqrt{\sigma}} \log \frac{1}{\epsilon}. \tag{25}$$

Since $\frac{1}{n}\text{Trace}(\mathbf{P}' \circ \mathbf{M}') \le c(S, \mathbf{M}) \le \text{Trace}(\mathbf{P}' \circ \mathbf{M}')$ and $\text{Trace}(\mathbf{P}' \circ \mathbf{M}') = \sum_i \mathbf{P}'_{ii} \mathbf{M}'_{ii} = \sum_i \mathbf{M}'_{ii} = \sum_i \mathbf{M}_{ii}/p_i^2$, we get the bounds:

$$\begin{aligned} \sqrt{\frac{c(S, \mathbf{M})}{\sigma}} \log \frac{1}{\epsilon} \quad &\ge \quad \sqrt{\frac{1}{n}\sum_i \frac{\mathbf{M}_{ii}}{p_i^2 \sigma}} \log \frac{1}{\epsilon} \\ \sqrt{\frac{c(S, \mathbf{M})}{\sigma}} \log \frac{1}{\epsilon} \quad &\le \quad \sqrt{\sum_i \frac{\mathbf{M}_{ii}}{p_i^2 \sigma}} \log \frac{1}{\epsilon}. \end{aligned} \tag{26}$$

5.1 Sampling 1: standard uniform minibatch samlpling (τ–nice sampling)

Let S_1 be the τ-nice sampling. It can be shown (see Lemma C.3) that $c(S_1, \mathbf{M}) \le \frac{n^2}{\tau^2}((1 - \beta)\max_i \mathbf{M}_{ii} + \beta L)$, and hence the iteration complexity (23) becomes

$$\mathcal{O}\left(\frac{n}{\tau}\sqrt{\frac{(1 - \beta)\max_i \mathbf{M}_{ii} + \beta L}{\sigma}} \log \frac{1}{\epsilon}\right). \tag{27}$$

This result interpolates between ACD with uniform probabilities (for $\tau = 1$) and accelerated gradient descent (for $\tau = n$). Note that the rate (27) is a strict improvement on the CD rate (9).

Table 2: New complexity results for `ACD` with minibatch size $\tau = \mathbb{E}\left[|S^k|\right]$ and various samplings (we suppress $\log(1/\epsilon)$ factors in all expressions). Constants: $\sigma =$ strong convexity constant of f, $L = \lambda_{\max}(\mathbf{M})$, $\beta = (\tau - 1)/(n - 1)$, $1 \le \gamma \le \sqrt{n}$, and $\omega \le \mathcal{O}(\sqrt{\tau})$ (ω can be as small as $\mathcal{O}(\tau/n)$).

Lower bound	$S_1 : p_i = \dfrac{\tau}{n}$	$S_2 : \dfrac{p_i^2}{\mathbf{M}_{ii}} \propto 1$	$S_3 : \dfrac{p_i^2}{\mathbf{M}_{ii}} \propto 1 - p_i$
$\dfrac{\sum_i \sqrt{\mathbf{M}_{ii}}}{\tau\sqrt{\sigma}}$	$\dfrac{n\sqrt{(1-\beta)\max_i \mathbf{M}_{ii} + \beta L}}{\tau\sqrt{\sigma}}$	$\dfrac{\gamma \sum_i \sqrt{\mathbf{M}_{ii}}}{\tau\sqrt{\sigma}}$	$\omega\dfrac{n\sqrt{(1-\beta)\max_i \mathbf{M}_{ii} + \beta L}}{\tau\sqrt{\sigma}}$
(24)	= uniform `ACD` for $\tau = 1$ = `AGD` for $\tau = n$	$\le \sqrt{n}\times$ lower bound • $\tau \le \dfrac{\sum_j \sqrt{\mathbf{M}_{jj}}}{\max_i \mathbf{M}_{ii}}$	• fastest in practice • any τ allowed

5.2 Sampling 2: importance sampling for minibatches

Consider now the sampling S_2 which includes every $i \in [n]$ in S_2, independently, with probability $p_i = \tau \frac{\sqrt{\mathbf{M}_{ii}}}{\sum_j \sqrt{\mathbf{M}_{jj}}}$. This sampling was not considered in the literature before. Note that $\mathbb{E}\left[|S_2|\right] = \sum_i p_i = \tau$. For this sampling, bounds (26) become:

$$\sqrt{\frac{c(S, \mathbf{M})}{\sigma}} \log \frac{1}{\epsilon} \ge \frac{\sum_i \sqrt{\mathbf{M}_{ii}}}{\tau\sqrt{\sigma}} \log \frac{1}{\epsilon}$$
$$\sqrt{\frac{c(S, \mathbf{M})}{\sigma}} \log \frac{1}{\epsilon} \le \frac{\sqrt{n} \sum_i \sqrt{\mathbf{M}_{ii}}}{\tau\sqrt{\sigma}} \log \frac{1}{\epsilon}. \quad (28)$$

Clearly, with this sampling we obtain an `ACD` method with complexity within a $\sqrt{n}$ factor from the lower bound established in Theorem 5.1. For $\tau = 1$ we have $\mathbf{P}' = \mathbf{I}$ and hence

$$\begin{aligned} c(S, \mathbf{M}) &= \lambda_{\max}(\mathbf{I} \circ \mathbf{M}') = \lambda_{\max}(\mathbf{Diag}(\mathbf{M}')) \\ &= \max_i \mathbf{M}_{ii}/p_i^2 = (\sum_j \sqrt{\mathbf{M}_{jj}})^2. \end{aligned}$$

Thus, the rate of `ACD` achieves the lower bound in (28) (see also (10)) and we recover the best current rate of `ACD` in the $\tau = 1$ case, established by Allen-Zhu et al. (2016). However, the sampling has an important limitation: it can be used for $\tau \le \sum_j \sqrt{\mathbf{M}_{jj}}/\max_i \mathbf{M}_{ii}$ only as otherwise the probabilities p_i exceed 1.

5.3 Sampling 3: another importance sampling for minibatches

Now consider sampling S_3 which includes each coordinate i within S_3 independently, with probability p_i satisfying the relation $p_i^2/\mathbf{M}_{ii} \propto 1 - p_i$. This is equivalent to setting

$$p_i := \frac{2\mathbf{M}_{ii}}{\sqrt{\mathbf{M}_{ii}^2 + 2\mathbf{M}_{ii}\delta} + \mathbf{M}_{ii}}, \quad (29)$$

where δ is a scalar for which $\sum_i p_i = \tau$. This sampling was not considered in the literature before. Probability

vector p was chosen as (29) for two reasons: i) $p_i \le 1$ for all i, and therefore the sampling can be used for all τ in contrast to S_1, and ii) we can prove Theorem 5.3.

Let $c_1 := c(S_1, \mathbf{M})$ and $c_3 := c(S_3, \mathbf{M})$. In light of (25), Theorem 5.3 compares S_1 and S_3 and says that `ACD` with S_3 has at most $\mathcal{O}(\sqrt{\tau})$ times worse rate compared to `ACD` with S_1, but has the capacity to be $\mathcal{O}(n/\tau)$ times better. We prove in Appendix B a similar theorem for `CD`. We stress that, despite some advances in the development of importance samplings for minibatch methods (Richtárik and Takáč, 2016b; Csiba and Richtárik, 2018), S_1 was until now the state-of-the-art in theory for `CD`. We are the first to give a provably better rate in the sense of Theorem B.3. The numerical experiments show that S_3 consistently outperforms S_1, and often dramatically so.

Theorem 5.3. *The leading complexity terms c_1 and c_3 of `ACD` (Algorithm (5)) with samplings S_1, and S_3, respectively, defined in Lemma 5.2, compare as follows:*

$$c_3 \le 2\frac{(2n - \tau)(n\tau + n - \tau)}{(n - \tau)^2} c_1 = \mathcal{O}(\tau)c_1. \quad (30)$$

Moreover, there exists $\mathbf{M}$ where $c_3 \le \mathcal{O}(\frac{\tau^2}{n^2})c_1$.

In real world applications, minibatch size τ is limited by hardware and in typical situations, one has $\tau \ll n$, oftentimes $\tau = \mathcal{O}(1)$. The importance of Theorem 5.3 is best understood from this perspective.

6 Experiments

We perform extensive numerical experiments to justify that minibatch `ACD` with importance sampling works well in practice. Here we present a few selected experiment only; more can be found in Appendix D.

In most of plots we compare of both accelerated and non-accelerated `CD` with all samplings S_1, S_2, S_3 introduced in Sections 5.1, 5.2 and 5.3 respectively. We refer to `ACD` with sampling S_3 as `AN` (Accelerated Nonuniform), `ACD` with sampling S_1 as `AU`, `ACD` with sampling S_2 as
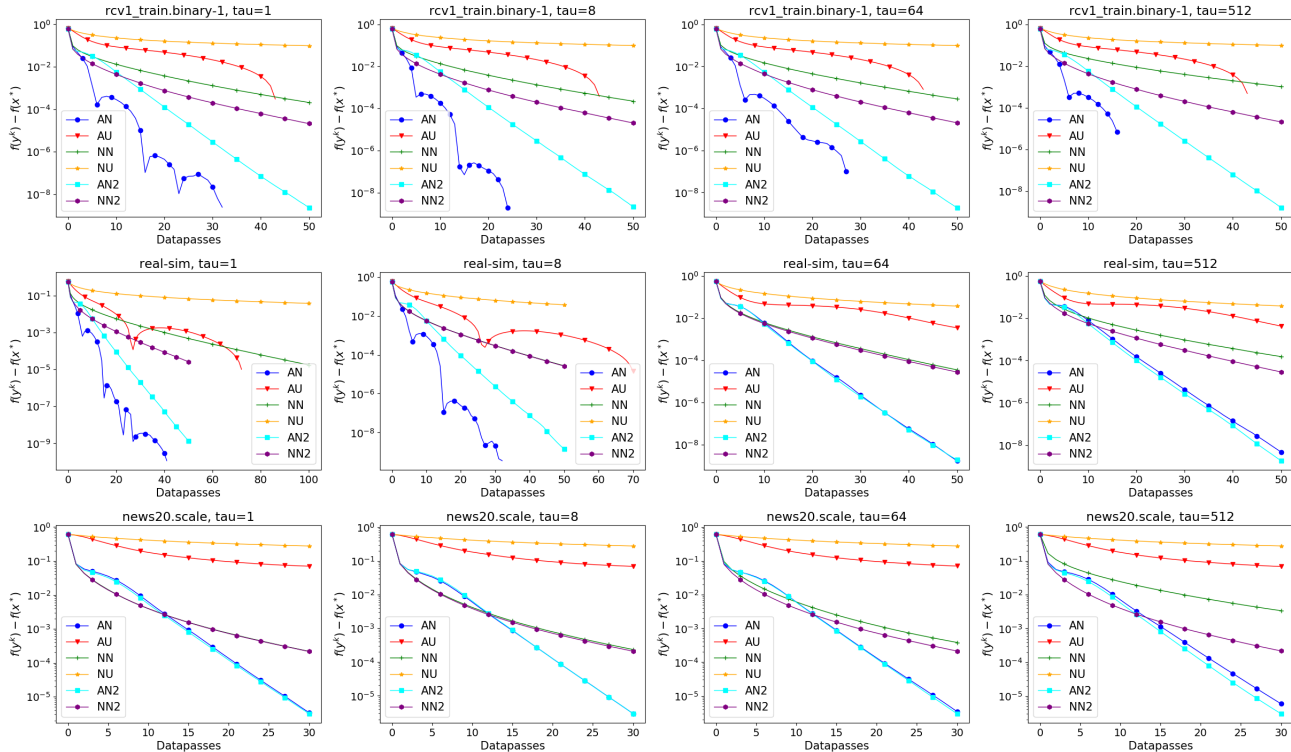
Figure 1: Six variants of coordinate descent (`AN`, `AU`, `NN`, `NU`, `AN2` and `AU2`) applied to a logistic regression problem, with minibatch sizes $\tau = 1, 8, 64$ and 512.

`AN2`, `CD` with sampling S_3 as `NN`, `CD` with sampling S_1 as `NU` and `CD` with sampling S_2 as `NN2`. We compare the methods for various choices of the expected minibatch sizes τ and on several problems.

In Figure 1, we report on a logistic regression problem with a few selected LibSVM Chang and Lin (2011) datasets. For larger datasets, pre-computing both strong convexity parameter σ and v may be expensive (however, recall that for v we need to tune only one scalar). Therefore, we choose ESO parameters v from Lemma 5.2, while estimating the smoothness matrix as $10\times$ its diagonal. An estimate of the strong convexity σ for acceleration was chosen to be the minimal diagonal element of the smoothness matrix. We provide a formal formulation of the logistic regression problem, along with more experiments applied to further datasets in Appendix D.2, where we choose v and σ in full accord with the theory.

Coordinate descent methods which allow for separable proximal operator were proven to be efficient to solve ERM problem, when applied on dual (Shalev-Shwartz and Tewari, 2011; Shalev-Shwartz and Zhang, 2013, 2014; Zhao and Zhang, 2015). Although we do not develop proximal methods in this paper, we empirically demonstrate that `ACD` allows for this extension as well. As a specific problem to solve, we choose dual of

SVM with hinge loss. Figure 2 presents the results. A detailed description of the experiment is presented in Appendix D.3. The results are indeed in favour of `ACD` with importance sampling. Therefore, `ACD` is not only suitable for *big dimensional* problems, it can handle the *big data* setting as well.

Finally, in Appendix D.1 we present several synthetic examples in order to shed more light on acceleration and importance sampling, and to see how its performance depends on the data. We also study how minibatch size influences the convergence rate. All the experimental results clearly show that acceleration, importance sampling and minibatching have a significant impact on practical performance of `CD` methods. Moreover, the difference in the performance of samplings S_2 and S_3 is negligible, and therefore we recommend using S_3, as it is not limited by the bound on expected minibatch size τ.

References

Allen-Zhu, Z. and Orecchia, L. (2017). Linear coupling: An ultimate unification of gradient and mirror descent. In *Innovations in Theoretical Computer Science*.

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. (2016). Even faster accelerated coordinate descent
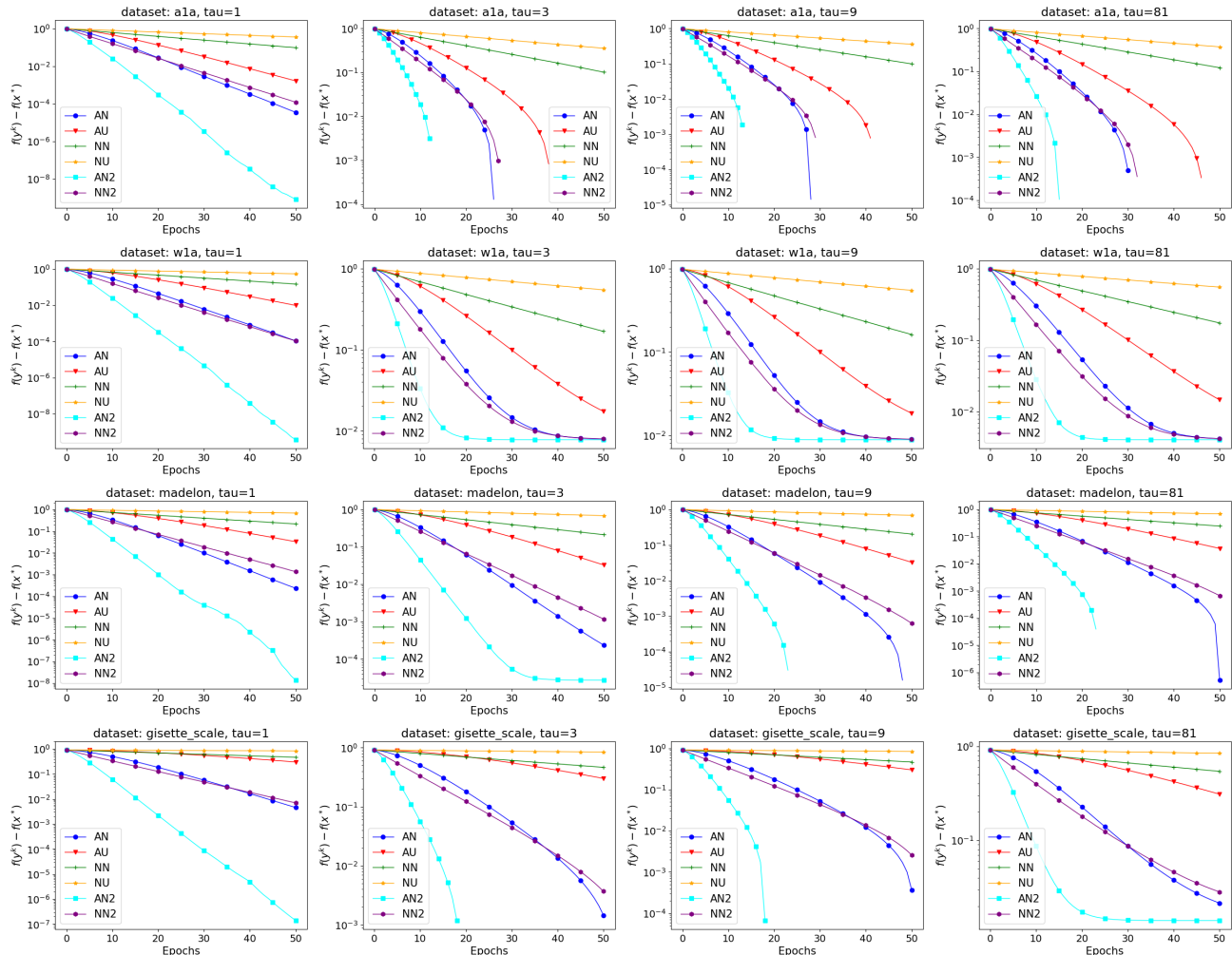
Figure 2: Six variants of coordinate descent (`AN`, `AU`, `NN`, `NU`, `AN2` and `AU2`) applied to a dual SVM problem, with minibatch sizes $\tau = 1, 2, 8$ and 81.

using non-uniform sampling. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1110–1119, New York, New York, USA.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schöenlieb, C.-B. (2017). Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *arXiv:1706.04957*.

Chang, C.-C. and Lin, C.-J. (2011). LibSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Csiba, D., Qu, Z., and Richtárik, P. (2015). Stochastic dual coordinate ascent with adaptive probabilities. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 674–683, Lille, France.

Csiba, D. and Richtárik, P. (2018). Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27).

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

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

Gurbuzbalaban, M., Ozdaglar, A., Parrilo, P. A., and Vanli, N. (2017). When cyclic coordinate descent outperforms randomized coordinate descent. In *Advances in Neural Information Processing Systems*, pages 7002–7010.

Lee, Y. T. and Sidford, A. (2013). Efficient accelerated coordinate descent methods and faster algorithms for

solving linear systems. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 147–156.

Lin, Q., Lu, Z., and Xiao, L. (2014). An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pages 3059–3067.

Luo, Z.-Q. and Tseng, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Kluwer Academic Publishers.

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

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

Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., and Koepke, H. (2015). Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1632–1641, Lille, France.

Qu, Z. and Richtárik, P. (2016). Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857.

Qu, Z. and Richtárik, P. (2016). Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884.

Qu, Z., Richtárik, P., and Zhang, T. (2015). Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems 28*.

Richtárik, P. and Takáč, M. (2016a). Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25.

Richtárik, P. and Takáč, M. (2016b). On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243.

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

Richtárik, P. and Takáč, M. (2016). Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484.

Saha, A. and Tewari, A. (2013). On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599.

Shalev-Shwartz, S. and Zhang, T. (2014). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 64–72, Bejing, China.

Stich, S. U., Raj, A., and Jaggi, M. (2017a). Approximate steepest coordinate descent. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3251–3259, International Convention Centre, Sydney, Australia.

Stich, S. U., Raj, A., and Jaggi, M. (2017b). Safe adaptive importance sampling. In *Advances in Neural Information Processing Systems*, pages 4384–4394.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

You, Y., Lian, X., Liu, J., Yu, H.-F., Dhillon, I. S., Demmel, J., and Hsieh, C.-J. (2016). Asynchronous parallel greedy coordinate descent. In *Advances in Neural Information Processing Systems*, pages 4682–4690.

Zhang, F. (1999). *Matrix Theory: Basic Results and Techniques*. Springer-Verlag New York.

Zhao, P. and Zhang, T. (2015). Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1–9, Lille, France.

Appendix

A   Proof of Theorem 4.2

Before starting the proof, we mention that the proof technique we use is inspired by Allen-Zhu and Orecchia (2017); Allen-Zhu et al. (2016), which takes the advantage of the coupling of gradient descent with mirror descent, resulting in a relatively simple proof.

A.1   Proof of inequality (18)

By comparing (12) and (6) for $h = e_i$, we get $\sigma_w w_i \leq \mathbf{M}_{ii}$, and the first inequality in (18) follows. Using (7) it follows that $e_i^\top (\mathbf{P} \circ \mathbf{M}) e_i \preceq e_i^\top \mathbf{Diag}\,(p \circ v)\, e_i$, which in turn implies $\mathbf{M}_{ii} \leq v_i$ and the second inequality in (18) follows.

A.2   Descent lemma

The following lemma is a consequence of $\mathbf{M}$–smoothness of f, and ESO inequality (7).

Lemma A.1. *Under the assumptions of Theorem 4.2, for all $k \geq 0$ we have the bound*

$$f(x^{k+1}) - \mathbb{E}\left[f(y^{k+1}) \,|\, x^{k+1}\right] \geq \frac{1}{2}\|\nabla f(x^{k+1})\|^2_{v^{-1}\circ p}. \tag{31}$$

Proof. We have

$$
\mathbb{E}\left[f(y^{k+1}) \mid x^{k+1}\right] \overset{(15)}{=} \mathbb{E}\left[f\left(x^{k+1} - \sum_{i \in S^k}\frac{1}{v_i}\nabla_i f(x^{k+1})e_i\right) \mid x^{k+1}\right]
$$

$$
\overset{(6)}{\leq} f(x^{k+1}) - \|\nabla f(x^{k+1})\|^2_{v^{-1}\circ p} + \frac{1}{2}\mathbb{E}\left[\left\|\sum_{i \in S^k}\frac{1}{v_i}\nabla_i f(x^{k+1})e_i\right\|^2_{\mathbf{M}} \mid x^{k+1}\right]
$$

$$
\overset{(7)}{\leq} f(x^{k+1}) - \|\nabla f(x^{k+1})\|^2_{v^{-1}\circ p} + \frac{1}{2}\left\|\nabla f(x^{k+1})\right\|^2_{v^{-1}\circ p}.
$$

$\square$

A.3   Key technical inequality

We first establish a lemma which will play a key part in the analysis.

Lemma A.2. *For every u we have*

$$
\eta \sum_{i \in S^k}\left\langle \frac{1}{p_i}\nabla_i f(x^{k+1})e_i, z^{k+1} - u\right\rangle - \frac{\eta\sigma_w}{2}\|x^{k+1} - u\|^2_w
$$

$$
\leq -\frac{1}{2}\|z^k - z^{k+1}\|^2_w + \frac{1}{2}\|z^k - u\|^2_w - \frac{1+\eta\sigma_w}{2}\|z^{k+1} - u\|^2_w.
$$

Proof. The proof is a direct generalization of the proof of analogous lemma of Allen-Zhu et al. (2016). We include it for completeness. Notice that (16) is equivalent to

$$
z^{k+1} = \operatorname{argmin}_z h^k(z) := \operatorname{argmin}_z \frac{1}{2}\|z - z^k\|^2_w + \eta\sum_{i \in S^k}\langle\frac{1}{p_i}\nabla_i f(x^{k+1})e_i, z\rangle + \frac{\eta\sigma_w}{2}\|z - x^{k+1}\|^2_w.
$$

Therefore, we have for every u

$$
\begin{aligned}
0 &= \langle \nabla h^k(z^{k+1}), z^{k+1} - u\rangle_w \\
&= \langle z^{k+1} - z^k, z^{k+1} - u\rangle_w + \eta\sum_{i \in S^k}\langle\frac{1}{p_i}\nabla_i f(x^{k+1})e_i, z^{k+1} - u\rangle + \eta\sigma_w\langle z^{k+1} - x^{k+1}, z^{k+1} - u\rangle_w. \quad (32)
\end{aligned}
$$

Next, by generalized Pythagorean theorem we have

$$\langle z^{k+1} - z^k, z^{k+1} - u\rangle_w = \frac{1}{2}\|z^k - z^{k+1}\|_w^2 - \frac{1}{2}\|z^k - u\|_w^2 + \frac{1}{2}\|u - z^{k+1}\|_w^2 \tag{33}$$

and

$$\langle z^{k+1} - x^{k+1}, z^{k+1} - u\rangle_w = \frac{1}{2}\|x^{k+1} - z^{k+1}\|_w^2 - \frac{1}{2}\|x^{k+1} - u\|_w^2 + \frac{1}{2}\|u - z^{k+1}\|_w^2. \tag{34}$$

It remains to put (34) and (35) into (33). $\qquad\square$

A.4 Proof of the theorem

To mitigate notational burden, consider all expectations in this proof to be taken with respect to the choice of the random subset of coordinates S^k. Using Lemma A.2 we have

$$\eta \sum_{i \in S^k} \langle \frac{1}{p_i}\nabla_i f(x^{k+1})e_i, z^k - u\rangle - \frac{\eta\sigma_w}{2}\|x^{k+1} - u\|_w^2$$

$$\leq \quad \eta \sum_{i \in S^k} \langle \frac{1}{p_i}\nabla_i f(x^{k+1})e_i, z^k - z^{k+1}\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_w^2 + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\|z^{k+1} - u\|_w^2$$

$$\leq \quad \frac{\eta^2}{2}\|\sum_{i \in S^k}\frac{1}{p_i}\nabla_i f(x^{k+1})e_i\|_{w^{-1}}^2 + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\|z^{k+1} - u\|_w^2$$

$$= \quad \frac{\eta^2}{2}\|\sum_{i \in S^k}\nabla_i f(x^{k+1})e_i\|_{w^{-1}\circ p^{-2}}^2 + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\|z^{k+1} - u\|_w^2.$$

Taking the expectation over the choice of S^k, we get

$$\eta\langle \nabla f(x^{k+1}), z^k - u\rangle - \frac{\eta\sigma_w}{2}\|x^{k+1} - u\|_w^2$$

$$\leq \quad \frac{\eta^2}{2}\|\nabla f(x^{k+1})\|_{w^{-1}\circ p^{-1}}^2 + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\mathbb{E}\left[\|z^{k+1} - u\|_w^2\right]$$

$$\overset{(17)}{=} \quad \frac{\eta^2}{2}\|\nabla f(x^{k+1})\|_{v^{-1}\circ p}^2 + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\mathbb{E}\left[\|z^{k+1} - u\|_w^2\right]$$

$$\overset{(31)}{\leq} \quad \eta^2\left(f(x^{k+1}) - \mathbb{E}\left[f(y^{k+1})\right]\right) + \frac{1}{2}\|z^k - u\|_w^2 - \frac{1+\eta\sigma_w}{2}\mathbb{E}\left[\|z^{k+1} - u\|_w^2\right].$$

Next, we have the following bounds

$$\eta\left(f(x^{k+1}) - f(x^*)\right) \overset{(12)}{\leq} \eta\langle \nabla f(x^{k+1}), x^{k+1} - x^*\rangle - \frac{\eta\sigma_w}{2}\|x^* - x^{k+1}\|_w^2$$

$$= \quad \eta\langle \nabla f(x^{k+1}), x^{k+1} - z^k\rangle + \eta\langle \nabla f(x^{k+1}), z^k - x^*\rangle - \frac{\eta\sigma_w}{2}\|x^* - x^{k+1}\|_w^2$$

$$\overset{(13)}{=} \quad \frac{(1-\theta)\eta}{\theta}\langle \nabla f(x^{k+1}), y^k - x^{k+1}\rangle + \eta\langle \nabla f(x^{k+1}), z^k - x^*\rangle - \frac{\eta\sigma_w}{2}\|x^* - x^{k+1}\|_w^2$$

$$\overset{(36)}{\leq} \quad \frac{(1-\theta)\eta}{\theta}\left(f(y^k) - f(x^{k+1})\right) + \eta^2\left(f(x^{k+1}) - \mathbb{E}\left[f(y^{k+1})\right]\right)$$

$$\quad + \frac{1}{2}\|z^k - x^*\|_w^2 - \frac{1+\eta\sigma_w}{2}\mathbb{E}\left[\|z^{k+1} - x^*\|_w^2\right].$$

Choosing $\eta = \frac{1}{\theta}$ and rearranging the above we obtain

$$\frac{1}{\theta^2}\left(\mathbb{E}\left[f(y^{k+1})\right] - f(x^*)\right) + \frac{1 + \frac{\sigma_w}{\theta}}{2}\mathbb{E}\left[\|z^{k+1} - x^*\|_w^2\right] \leq \frac{(1-\theta)}{\theta^2}\left(f(y^k) - f(x^*)\right) + \frac{1}{2}\|z^k - x^*\|_w^2.$$

Finally, setting θ such that $1 + \frac{\sigma_w}{\theta} = \frac{1}{1-\theta}$, which coincides with (20), we get

$$\mathbb{E}\left[P^{k+1}\right] \leq (1-\theta)P^k,$$

as desired.

B Better rates for minibatch `CD` (without acceleration)

In this section we establish better rates for minibatch `CD` method than the current state of the art. Our starting point is the following complexity theorem.

Theorem B.1. *Choose any proper sampling and let $\mathbf{P}$ be its probability matrix and p its probability vector. Let*

$$c(S, \mathbf{M}) := \lambda_{\max}(\mathbf{P}'' \circ \mathbf{M}),$$

where $\mathbf{P}'' := \mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}$ and $\mathbf{D} := \mathbf{Diag}\,(p)$. Then the vector v defined by $v_i = c(S, \mathbf{M})p_i$ satisfies the ESO inequality (7). Moreover, if we run the non-accelerated `CD` method (5) with this sampling and stepsizes $\alpha_i = \frac{1}{c(S,\mathbf{M})p_i}$, then the iteration complexity of the method is

$$\frac{c(S, \mathbf{M})}{\sigma} \log \frac{1}{\epsilon}. \tag{35}$$

Proof. Let $v_i = cp_i$ for all i. The ESO inequality holds for this choice of v if $\mathbf{P} \circ \mathbf{M} \preceq c\mathbf{D}^2$. This is equivalent to Since $\mathbf{D}^{-1}(\mathbf{P} \circ \mathbf{M})\mathbf{D}^{-1} = \mathbf{P}'' \circ \mathbf{M}$, the above inequality is equivalent to $\mathbf{P}'' \circ \mathbf{M} \preceq c\mathbf{I}$, which is equivalent to $c \geq \lambda_{\max}(\mathbf{P}'' \circ \mathbf{M})$. So, choosing $c = c(S, \mathbf{M})$ works. Plugging this choice of v into the complexity result (8) gives (36). $\square$

B.1 Two uniform samplings and one new importance sampling

In the next theorem we compute now consider several special samplings. All of them choose in expectation a minibatch of size τ and are hence directly comparable.

Theorem B.2. *The following statements hold:*

(i) *Let S_1 be the τ–nice sampling. Then*

$$c_1 := c(S_1, \mathbf{M}) = \frac{n}{\tau}\lambda_{\max}\left(\frac{\tau - 1}{n - 1}\mathbf{M} + \frac{n - \tau}{n - 1}\mathbf{Diag}\,(\mathbf{M})\right). \tag{36}$$

(ii) *Let S_2 be the independent uniform sampling with minibatch size τ. That is, for all i we independently decide whether $i \in S$, and do so by picking i with probability $p_i = \frac{\tau}{n}$. Then*

$$c_2 := c(S_2, \mathbf{M}) = \lambda_{\max}\left(\mathbf{M} + \frac{n - \tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right). \tag{37}$$

(iii) *Let S_3 be an independent sampling where we choose $p_i \propto \frac{\mathbf{M}_{ii}}{\delta + \mathbf{M}_{ii}}$ where $\delta > 0$ is chosen so that $\sum_i p_i = \tau$. Then*

$$c_3 := c(S_3, \mathbf{M}) = \lambda_{\max}(\mathbf{M}) + \delta. \tag{38}$$

Moreover,

$$\delta \leq \frac{\mathrm{Trace}\,(\mathbf{M})}{\tau}. \tag{39}$$

Proof. We will deal with each case separately:

(i) The probability matrix of S_1 is $\mathbf{P} = \frac{\tau}{n}\left(\beta\mathbf{E} + (1 - \beta)\mathbf{I}\right)$, where $\beta = \frac{\tau - 1}{n - 1}$, and $\mathbf{D} = \frac{\tau}{n}\mathbf{I}$. Hence,

$$
\begin{aligned}
\mathbf{P}'' \circ \mathbf{M} &= (\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}) \circ \mathbf{M} \\
&= \frac{\tau}{n}\left(\beta\mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1} + (1 - \beta)\mathbf{D}^{-2}\right) \circ \mathbf{M} \\
&= \frac{\tau}{n}\left(\frac{\tau - 1}{n - 1}\mathbf{E} + \frac{n - \tau}{n - 1}\mathbf{I}\right) \circ \mathbf{M} \\
&= \frac{\tau}{n}\left(\frac{\tau - 1}{n - 1}\mathbf{M} + \frac{n - \tau}{n - 1}\mathbf{Diag}\,(\mathbf{M})\right).
\end{aligned}
$$

(ii) The probability matrix of S_2 is $\mathbf{P} = \frac{\tau}{n}\left(\frac{\tau}{n}\mathbf{E} + (1 - \frac{\tau}{n})\mathbf{I}\right)$, and $\mathbf{D} = \frac{\tau}{n}\mathbf{I}$. Hence,

$$
\begin{aligned}
\mathbf{P}'' \circ \mathbf{M} &= (\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}) \circ \mathbf{M} \\
&= \frac{\tau}{n}\left(\frac{\tau}{n}\mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1} + \left(1 - \frac{\tau}{n}\right)\mathbf{D}^{-2}\right) \circ \mathbf{M} \\
&= \left(\mathbf{E} + \frac{n-\tau}{\tau}\mathbf{I}\right) \circ \mathbf{M} \\
&= \mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\,.
\end{aligned}
$$

(iii) The probability matrix of S_3 is $\mathbf{P} = pp^\top + \mathbf{D} - \mathbf{D}^2$. Therefore,

$$
\begin{aligned}
\mathbf{P}'' \circ \mathbf{M} &= (\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}) \circ \mathbf{M} \\
&= \left(\mathbf{D}^{-1}pp^\top\mathbf{D}^{-1} + \mathbf{D}^{-1} - \mathbf{I}\right) \circ \mathbf{M} \\
&= \left(\mathbf{E} + \mathbf{D}^{-1} - \mathbf{I}\right) \circ \mathbf{M} \\
&= \left(\mathbf{E} + \delta(\mathbf{Diag}\,(\mathbf{M}))^{-1}\right) \circ \mathbf{M} \\
&= \mathbf{M} + \delta\mathbf{I}.
\end{aligned}
$$

To establish the bound on δ, it suffices to note that

$$
\tau = \sum_i p_i = \sum_i \frac{\mathbf{M}_{ii}}{\delta + \mathbf{M}_{ii}} \leq \sum_i \frac{\mathbf{M}_{ii}}{\delta} = \frac{\text{Trace}\,(\mathbf{M})}{\delta}.
$$

$\square$

B.2 Comparing the samplings

In the next result we show that sampling S_3 is at most twice worse than S_2, which is at most twice worse than S_1. Note that S_1 is uniform; and it is the standard minibatch sampling used in the literature and applications. Our novel sampling S_3 is *non-uniform*, and is at most four times worse than S_1 in the worst case. However, *it can be substantially better*, as we shall show later by giving an example.

Theorem B.3. *The leading complexity terms c_1, c_2, and c_3 of CD (Algorithm (5)) with samplings S_1, S_2, and S_3, respectively, defined in Theorem B.2, compare as follows:*

(i) $c_3 \leq \frac{2n-\tau}{n-\tau}c_2$

(ii) $c_2 \leq \frac{(n-1)\tau}{n(\tau-1)}c_1 \leq 2c_1$

Proof. We have:

(i)

$$
\begin{aligned}
c_3 &\overset{(39)}{=} \lambda_{\max}(\mathbf{M}) + \delta \\
&\leq \lambda_{\max}\left(\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) + \delta \\
&\overset{(38)}{=} c_2 + \delta \\
&\overset{(40)}{\leq} c_2 + \frac{\mathrm{Trace}\,(\mathbf{M})}{\tau} \\
&\leq c_2 + \frac{n\max_i \mathbf{M}_{ii}}{\tau} \\
&= c_2 + \frac{n}{n-\tau}\frac{n-\tau}{\tau}\max_i \mathbf{M}_{ii} \\
&= c_2 + \frac{n}{n-\tau}\lambda_{\max}\left(\frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) \\
&\leq c_2 + \frac{n}{n-\tau}\lambda_{\max}\left(\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) \\
&\overset{(38)}{=} \frac{2n-\tau}{n-\tau}c_2.
\end{aligned}
$$

(ii)

$$
\begin{aligned}
c_2 &\overset{(38)}{=} \lambda_{\max}\left(\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) \\
&= \lambda_{\max}\left(\frac{n(\tau-1)}{\tau(n-1)}\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M}) + \left(1 - \frac{n(\tau-1)}{\tau(n-1)}\right)\mathbf{M}\right) \\
&\overset{(\dagger)}{\leq} \lambda_{\max}\left(\frac{n(\tau-1)}{\tau(n-1)}\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) + \lambda_{\max}\left(\left(1 - \frac{n(\tau-1)}{\tau(n-1)}\right)\mathbf{M}\right) \\
&\leq \lambda_{\max}\left(\frac{n(\tau-1)}{\tau(n-1)}\mathbf{M} + \frac{n(n-\tau)}{\tau(n-1)}\mathbf{Diag}\,(\mathbf{M})\right) + \frac{n-\tau}{(n-1)\tau}\lambda_{\max}\,(\mathbf{M}) \\
&\overset{(37)}{=} c_1 + \frac{n-\tau}{(n-1)\tau}\lambda_{\max}\,(\mathbf{M}) \\
&\overset{(38)}{\leq} c_1 + \frac{n-\tau}{(n-1)\tau}c_2.
\end{aligned}
$$

The statement follows by reshuffling the final inequality. In step ($\dagger$) we have used subadditivity of the function $\mathbf{A} \mapsto \lambda_{\max}(\mathbf{A})$.

$\square$

The next simple example shows that sampling S_3 can be arbitrarily better than sampling S_1.

Example 1. Consider $n \gg 1$, and choose any τ and

$$
\mathbf{M} := \begin{pmatrix} n & 0^\top \\ 0 & \mathbf{I} \end{pmatrix}
$$

for $\mathbf{I} \in \mathbb{R}^{(n-1)\times(n-1)}$. Then, it is easy to verify that $c_1 \overset{(37)}{=} \frac{n^2}{\tau}$ and $c_3 \overset{(39)+(40)}{\leq} n + \frac{2n-1}{\tau} = \mathcal{O}(\frac{n}{\tau})$. Thus, convergence rate of CD with S_3 sampling can be up to $\mathcal{O}(n)$ times better than convergence rate of CD with τ–nice sampling.

Remark 1. Looking only at diagonal elments of $\mathbf{M}$, an intuition tells us that one should sample a coordinate corresponding to larger diagonal entry of $\mathbf{M}$ with higher probability. However, this might lead to worse convergence, comparing to τ–nice sampling. Therefore the results we provide in this section cannot be qualitatively better, i.e.

there are examples of smoothness matrix, for which assigning bigger probability to bigger diagonal elements leads to worse rate. It is an easy exercise to verify that for $\mathbf{M} \in \mathbb{R}^{10 \times 10}$ such that

$$\mathbf{M} := \begin{pmatrix} 2 & 0^\top \\ 0 & 11^\top \end{pmatrix},$$

and $\tau \geq 2$ we have $c(S_{\text{nice}}, \mathbf{M}) \leq c(S', \mathbf{M})$ for any S' satisfying $p(S')_i \geq p(S')_j$ if and only if $\mathbf{M}_{ii} \geq \mathbf{M}_{jj}$.

C Proofs for Section 5

C.1 Proof of Theorem 5.1

We start with a lemma which allows us to focus on ESO parameters v_i which are proportional to the squares of the probabilities p_i.

Lemma C.1. *Assume that the ESO inequality* (7) *holds. Let* $j = \arg\max_i \frac{v_i}{p_i^2}$, $c = \frac{v_j}{p_j^2}$ *and* $v' = cp^2$ *(i.e.,* $v'_i = cp_i^2$ *for all* i*). Then the following statements hold:*

(i) $v' \geq v$.

(ii) *ESO inequality* (7) *holds for* v' *also.*

(iii) *Assuming* f *is* σ*–convex, Theorem 4.2 holds if we replace* v *by* v'*, and the rate* (24) *is unchanged if we replace* v *by* v'*.*

Proof. (i) $v'_i = cp_i^2 = \frac{v_j}{p_j^2}p_i^2 = \left(\frac{v_j}{p_j^2}\frac{p_i^2}{v_i}\right)v_i \geq v_i$.

(ii) This follows directly from (i).

(iii) Theorem 4.2 holds with v replaced by v' because ESO holds. To show that the rates are unchanged first note that $\max_i \frac{v_i}{p_i^2} = \frac{v_j}{p_j^2} = c$. On the other hand, by construction, we have $c = \frac{v'_i}{p_i^2}$ for all i. So, in particular, $c = \max_i \frac{v'_i}{p_i^2}$.

□

In view of the above lemma, we can assume without loss of generality that $v = cp^2$. Hence, the rate in (24) can be written in the form

$$\sqrt{\max_i \frac{v_i}{p_i^2 \sigma}} = \sqrt{\frac{c}{\sigma}}. \tag{40}$$

In what follows, we will establish a lower bound on c, which will lead to the lower bound on the rate expressed as inequality (24). As a starting point, note that directly from (7) we get the bound

$$\mathbf{P} \circ \mathbf{M} \preceq \mathbf{Diag}\,(p \circ v) = c\mathbf{Diag}\,\left(p^3\right). \tag{41}$$

Let $\mathbf{D}_1 = \mathbf{Diag}\,(p)^{-1/2}$ and $\mathbf{D}_2 = \mathbf{Diag}\,(p)^{-1}$. From (42) we get $\mathbf{D}_1\mathbf{D}_2(\mathbf{P} \circ \mathbf{M})\mathbf{D}_2\mathbf{D}_1 \preceq c\mathbf{I}$ and hence

$$c \geq c(S, \mathbf{M}) := \lambda_{\max}(\mathbf{D}_1\mathbf{D}_2(\mathbf{P} \circ \mathbf{M})\mathbf{D}_2\mathbf{D}_1). \tag{42}$$

At this point, the following identity will be useful.

Lemma C.2. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{n \times n}$*, with* $\mathbf{D}_1, \mathbf{D}_2$ *being diagonal. Then*

$$\mathbf{D}_1(\mathbf{A} \circ \mathbf{B})\mathbf{D}_2 = (\mathbf{D}_1\mathbf{A}\mathbf{D}_2) \circ \mathbf{B} = \mathbf{A} \circ (\mathbf{D}_1\mathbf{B}\mathbf{D}_2). \tag{43}$$

Proof. The proof is straightforward, and hence we do not include it. The identity is formulated as an exercise in (Zhang, 1999). □

Repeatedly applying Lemma C.2, we get

$$\mathbf{D}_1\mathbf{D}_2(\mathbf{P} \circ \mathbf{M})\mathbf{D}_2\mathbf{D}_1 = \underbrace{(\mathbf{D}_1\mathbf{P}\mathbf{D}_1)}_{\mathbf{P}'} \circ \underbrace{(\mathbf{D}_2\mathbf{M}\mathbf{D}_2)}_{\mathbf{M}'}.$$

Plugging this back into (43), and since $\mathbf{P}'_{ii} = 1$ for all i, we get the bound

$$
\begin{aligned}
c &\geq& c(S, \mathbf{M}) = \lambda_{\max}(\mathbf{P}' \circ \mathbf{M}') \geq \max_i (\mathbf{P}' \circ \mathbf{M}')_{ii} = \max_i \mathbf{P}'_{ii}\mathbf{M}'_{ii} = \max_i \mathbf{M}'_{ii} \\
&=& \max_i \frac{\mathbf{M}_{ii}}{p_i^2} \geq \frac{\left(\sum_{i=1}^n \mathbf{M}_{ii}^{1/2}\right)^2}{\tau^2}.
\end{aligned}
\tag{44}
$$

The last inequality follows by observing that the optimal solution of the optimization problem

$$\min_p \left\{ \max_i \frac{\mathbf{M}_{ii}}{p_i^2} \mid p_1, \ldots, p_n > 0, \ \sum_i p_i = \tau \right\}$$

is $p_i = \tau \frac{\mathbf{M}_{ii}^{1/2}}{\sum_j \mathbf{M}_{jj}^{1/2}}$. Inequality (24) now follows by substituting the lower bound on c obtained in (45) into (41).

C.2 Proof of Lemma 5.2

$$
\begin{aligned}
\mathbf{Diag}\left(p_1 v_1, \ldots, p_n v_n\right) &=& c(S, \mathbf{M})\mathbf{Diag}\left(p_1^3, \ldots, p_n^3\right) \\
&=& c(S, \mathbf{M})\mathbf{D}^3 \\
&=& \lambda_{\max}\left(\left(\mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{-1/2}\right) \circ \left(\mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\right)\right)\mathbf{D}^3 \\
&\succeq& \mathbf{D}^{\frac{3}{2}}\left(\left(\mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{-1/2}\right) \circ \left(\mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\right)\right)\mathbf{D}^{\frac{3}{2}} \\
&\overset{(44)}{=}& \mathbf{P} \circ \mathbf{M}.
\end{aligned}
$$

The last inequality came from the fact that $\mathbf{D}$ is diagonal.

C.3 Bound on $c(S_1, \mathbf{S})$

Lemma C.3. $c(S_1, \mathbf{M}) \leq \frac{n^2}{\tau^2}((1 - \beta)\max_i \mathbf{M}_{ii} + \beta L)$.

Proof. Recall that the probability matrix of S_1 is $\mathbf{P} = \frac{\tau}{n}((1 - \beta)\mathbf{I} + \beta\mathbf{E})$. Since $p_i = \frac{\tau}{n}$ and $\mathbf{M} \preceq L\mathbf{I}$, we have

$$
\begin{aligned}
c(S_1, \mathbf{M}) &=& \lambda_{\max}\left(\mathbf{P}' \circ \mathbf{M}'\right) \\
&=& \lambda_{\max}\left(\left(\mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{-1/2}\right) \circ \left(\mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\right)\right) \\
&=& \lambda_{\max}\left(\frac{\tau}{n}\left((1 - \beta)\mathbf{D}^{-1} + \beta\mathbf{D}^{-1/2}\mathbf{E}\mathbf{D}^{-1/2}\right) \circ \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\right) \\
&=& \frac{\tau}{n}\lambda_{\max}\left(\left((1 - \beta)\mathbf{D}^{-1} + \beta\mathbf{D}^{-1/2}\mathbf{E}\mathbf{D}^{-1/2}\right) \circ \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\right) \\
&=& \frac{\tau}{n}\lambda_{\max}\left((1 - \beta)\mathbf{Diag}\left(\mathbf{M}_{ii}/p_i^3\right) + \beta\mathbf{D}^{-3/2}\mathbf{M}\mathbf{D}^{-3/2}\right) \\
&\preceq& \frac{\tau}{n}\lambda_{\max}\left((1 - \beta)\mathbf{Diag}\left(\mathbf{M}_{ii}/p_i^3\right) + \beta L\mathbf{D}^{-3}\right) \\
&=& \frac{\tau}{n}\lambda_{\max}\left((1 - \beta)\frac{n^3}{\tau^3}\max_i \mathbf{M}_{ii} + \beta L\frac{n^3}{\tau^3}\right) \\
&=& \frac{n^2}{\tau^2}\left((1 - \beta)\max_i \mathbf{M}_{ii} + \beta L\right).
\end{aligned}
$$

$\square$

C.4 Proof of Theorem 5.3

For the purpose of this proof, let S_2 be the independent uniform sampling with minibatch size τ. That is, for all i we independently decide whether $i \in S$, and do so by picking i with probability $p_i = \frac{\tau}{n}$. Recall that S_3 is the independent importance sampling.

For simplicity, let $\mathbf{P}_i$ be the probability matrix of sampling S_i, $\mathbf{D}_i := \mathbf{Diag}\,(\mathbf{P}_i)$, and $\mathbf{M}'_i := \mathbf{D}_i^{-1/2}\mathbf{M}\mathbf{D}_i^{-1/2}$, for $i = 1, 3$. Next, we have

$$
\begin{aligned}
c(S_i, \mathbf{M}) &= \lambda_{\max}\left(\left(\mathbf{D}_i^{-1/2}\mathbf{P}_i\mathbf{D}_i^{-1/2}\right) \circ \left(\mathbf{D}_i^{-1}\mathbf{M}\mathbf{D}_i^{-1}\right)\right) \\
&\overset{(44)}{=} \lambda_{\max}\left(\left(\mathbf{D}_i^{-1}\mathbf{P}_i\mathbf{D}_i^{-1}\right) \circ \left(\mathbf{D}_i^{-1/2}\mathbf{M}\mathbf{D}_i^{-1/2}\right)\right) \\
&= \lambda_{\max}\left((\mathbf{E} + \mathbf{D}_i^{-1} - \mathbf{I}) \circ \mathbf{M}'_i\right) \\
&= \lambda_{\max}\left(\mathbf{M}'_i + \mathbf{Diag}\,(\mathbf{M}'_i) \circ (\mathbf{D}_i^{-1} - \mathbf{I})\right),
\end{aligned}
\tag{45}
$$

where the third identity holds since both S_i is an independent sampling, which means that $\left(\mathbf{D}_i^{-1}\mathbf{P}_i\mathbf{D}_i^{-1}\right)_{kl} = \frac{p_{kl}}{p_k p_l}$, where $p = \mathbf{Diag}\,(\mathbf{D}_i)$.

Denote $c_i := c(S_i, \mathbf{M})$. Thus for S_2 we have

$$
c_2 = \frac{\tau}{n}\lambda_{\max}\left(\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right).
\tag{46}
$$

Let us now establish a technical lemma.

Lemma C.4.

$$
\lambda_{\max}\left(\mathbf{M}'_3 + \mathrm{diag}(\mathbf{M}'_3) \circ (\mathbf{D}_3^{-1} - \mathbf{I})\right) \leq \frac{2n-\tau}{n-\tau}\lambda_{\max}\left(\mathbf{M}'_3 + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M}'_3)\right)
\tag{47}
$$

Proof. The statement follows immediately repeating the steps of the proof of (i) from Theorem B.3 using the fact that for sampling S_3 we have $p_i/\mathbf{M}_{ii} \propto p_i^{-1} - 1$. $\qquad\square$

We can now proceed with comparing c_2 to c_3.

$$
\begin{aligned}
c_3 &= \lambda_{\max}\left(\mathbf{M}'_3 + \mathbf{Diag}\,(\mathbf{M}'_3) \circ (\mathbf{D}_3^{-1} - \mathbf{I})\right) \\
&\overset{(48)}{\leq} \frac{2n-\tau}{n-\tau}\lambda_{\max}\left(\mathbf{M}'_3 + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M}'_3)\right) \\
&\overset{(*)}{\leq} \frac{2n-\tau}{n-\tau}\lambda_{\max}\left(n\mathbf{Diag}\,(\mathbf{M}'_3) + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M}'_3)\right) \\
&= \frac{2n-\tau}{n-\tau}\frac{n\tau+n-\tau}{\tau}\lambda_{\max}\left(\mathbf{Diag}\,(\mathbf{M}'_3)\right) \\
&\overset{(**)}{\leq} \frac{2n-\tau}{n-\tau}\frac{n\tau+n-\tau}{\tau}\frac{\tau}{n}\lambda_{\max}\left(\mathbf{Diag}\,(\mathbf{M})\right) \\
&\leq \frac{2n-\tau}{n-\tau}\frac{n\tau+n-\tau}{\tau}\frac{\tau}{n}\frac{\tau}{n-\tau}\lambda_{\max}\left(\mathbf{M} + \frac{n-\tau}{\tau}\mathbf{Diag}\,(\mathbf{M})\right) \\
&\overset{(46)}{=} \frac{2n-\tau}{n-\tau}\frac{n\tau+n-\tau}{\tau}\frac{\tau}{n-\tau}c_2
\end{aligned}
\tag{48}
$$

Above, inequality $(*)$ holds since for any $n \times n$ matrix $\mathbf{Q} \succ 0$ we have $\mathbf{Q} \preceq n\mathbf{Diag}\,(\mathbf{Q})$ and inequality $(**)$ holds since $(\mathbf{D}_3)_{ii} \geq (\mathbf{D}_3)_{jj}$ if and only if $\mathbf{M}_{ii} \geq \mathbf{M}_{jj}$ due to choice of p.

Let us now compare to c_2 and c_1. We have

$$
\begin{aligned}
c_1 &= \lambda_{\max}\left(\left(\mathbf{D}_1^{-1/2}\mathbf{P}_1\mathbf{D}_1^{-1/2}\right) \circ \left(\mathbf{D}_1^{-1}\mathbf{M}\mathbf{D}_1^{-1}\right)\right) \\
&\overset{(44)}{=} \lambda_{\max}\left(\left(\mathbf{D}_1^{-1}\mathbf{P}_1\mathbf{D}_1^{-1}\right) \circ \left(\mathbf{D}_1^{-1/2}\mathbf{M}\mathbf{D}_1^{-1/2}\right)\right) \\
&= \lambda_{\max}\left(\left(\frac{\tau-1}{n-1}\frac{n}{\tau}\mathbf{E} + \frac{n}{\tau}\mathbf{I} - \frac{\tau-1}{n-1}\frac{n}{\tau}\mathbf{I}\right) \circ \mathbf{M}_1''\right) \\
&= \frac{n}{\tau}\lambda_{\max}\left(\frac{\tau-1}{n-1}\mathbf{M}_1'' + \frac{n-\tau}{n-1}\mathbf{Diag}\left(\mathbf{M}_1''\right)\right) \\
&= \left(\frac{n}{\tau}\right)^2\lambda_{\max}\left(\frac{\tau-1}{n-1}\mathbf{M} + \frac{n-\tau}{n-1}\mathbf{Diag}\left(\mathbf{M}\right)\right).
\end{aligned}
\tag{49}
$$

As (47) and (50) are established, following the proof of (ii) from Theorem B.3, we arrive at

$$
c_2 \le \frac{(n-1)\tau}{n(\tau-1)}c_1 \le 2c_1.
\tag{50}
$$

It remains to combine (49) and (51) to establish (30).

As an example where $c_3 \approx \left(\frac{\tau}{n}\right)^2 c_2$, we propose Example 2.

Example 2. Consider $n \ge 1$, choose any $n \ge \tau \ge 1$ and

$$
\mathbf{M} := \begin{pmatrix} N & 0^\top \\ 0 & \mathbf{I} \end{pmatrix}
$$

for $\mathbf{I} \in \mathbb{R}^{(n-1)\times(n-1)}$. Then, it is easy to verify that $c_1 \overset{(50)}{=} \left(\frac{n}{\tau}\right)^2 N$. Moreover, for large enough N we have

$$
p \approx \left(1, \frac{\tau-1}{n-1}, \ldots, \frac{\tau-1}{n-1}\right)^\top \qquad \Rightarrow \qquad \mathbf{M}_3' \approx \mathbf{Diag}\left(N, \frac{n-1}{\tau-1}, \ldots, \frac{n-1}{\tau-1}\right).
$$

Therefore, using (46) and again for large enough N, we get $c_3 \approx N$. Thus, $c_3 \approx \left(\frac{\tau}{n}\right)^2 c_2$.

D   Extra Experiments

In this section we present additional numerical experiments. We first present some synthetic examples in Section D.1 in order to have better understanding of both acceleration and importance sampling, and to see how it performs on what type of data. We also study how minibatch size influences the convergence rate.

Then, in Section D.2, we work with logistic regression problem on LibSVM (Chang and Lin, 2011) data. For small datasets, we choose the parameters of `ACD` as theory suggests and for large ones, we estimate them, as we describe in the main body of the paper. Lastly, we tackle dual of SVM problem with squared hinge loss, which we present in Section D.3.

In most of plots we compare of both accelerated and non-accelerated `CD` with all samplings S_1, S_2, S_3 introduced in Sections 5.1, 5.2 and 5.3 respectively. We refer to `ACD` with sampling S_3 as `AN` (Accelerated Nonuniform), `ACD` with sampling S_1 as `AU`, `ACD` with sampling S_2 as `AN2`, `CD` with sampling S_3 as `NN`, `CD` with sampling S_1 as `NU` and `CD` with sampling S_2 as `NN2`. As for Sampling 2, it might happen that probabilities become larger than one if τ is large (see Section 5.2), we set those probabilities to 1 while keeping the rest as it is.

We compare the mentioned methods for various choices of the expected minibatch sizes τ and several problems.

D.1   Synthetic quadratics

As we mentioned, the goal of this section is to provide a better understanding of both acceleration and importance sampling. For this purpose we consider as simple setting as possible – minimizing quadratic

$$f(x) = \frac{1}{2}x^\top \mathbf{M}x - b^\top x, \tag{51}$$

where $b \sim N(0, I)$ and $\mathbf{M}$ is chosen as one of the 5 types, as the following table suggests.

Problem type	M
1	$\mathbf{A}^\top\mathbf{A} + \mathbf{I}$ for $\mathbf{A}^{\frac{n}{2}\times n}$; have independent entries from $N(0, 1)$
2	$\mathbf{A}^\top\mathbf{A} + \mathbf{I}$ for $\mathbf{A}^{2n\times n}$; have independent entries from $N(0, 1)$
3	$\mathrm{diag}(1, 2, \ldots, n)$
4	$\mathbf{A} + \mathbf{I}$, $\mathbf{A}_{n,n} = n$, $\mathbf{A}_{1:(n-1),1:(n-1)} = 1$, $\mathbf{A}_{1:(n-1),n} = \mathbf{A}_{n,1:(n-1)} = 0$
5	$\mathbf{A}^\top\mathbf{D}\mathbf{A} + \mathbf{I}$ for $\mathbf{A}^{\frac{n}{2}\times n}$; have independent entries from $N(0, 1)$, $\mathbf{D} = \frac{1}{\sqrt{n}}\mathbf{Diag}\,(1, 2, \ldots, n)$

Table 3: Problem types

In the first example we perform (Figure 3), we compare the performance of both accelerated and non-accelerated algorithm with both nonuniform and τ nice sampling on problems as per Table 3. In all experiments, we set $n = 1000$ and we plot a various choices of τ.

D.1.1 Comparison of methods on synthetic data

Figure 3 presents the numerical performance of `ACD` for various types of synthetic problems given by (52) and Table 3. It suggests what our theory shows – that accelerated algorithm is always faster than its non-accelerated counterpart, and on top of that, performance of τ–nice sampling (S_1) can be negligibly faster than importance sampling (S_2, S_3), but is usually significantly slower. A significance of the importance sampling is mainly demonstrated on problem type 4, which roughly coincides with Examples 1 and 2. Figure 3 presents Sampling 2 only for the cases when the bound on τ form Section 5.2 is satisfied.
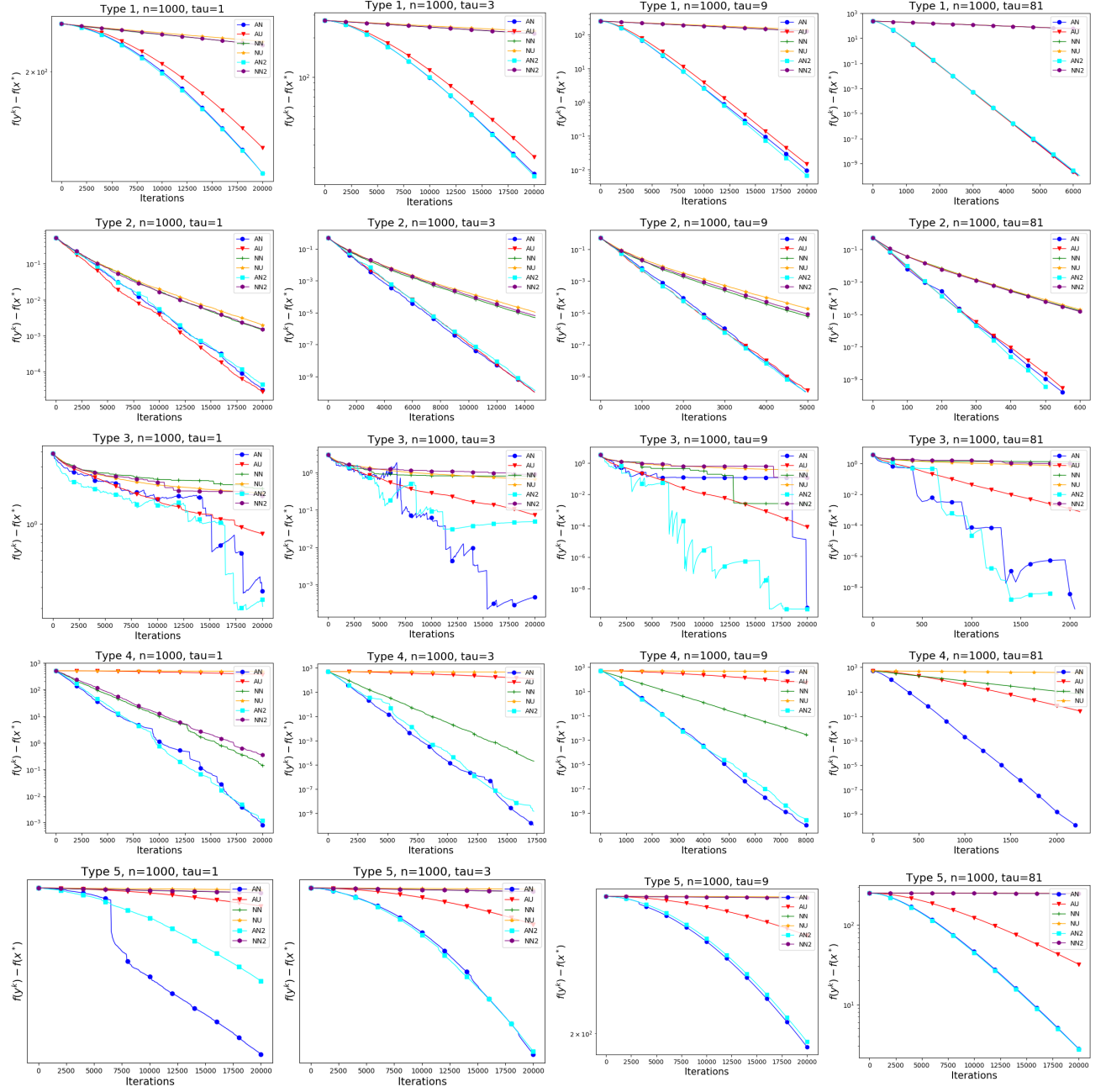
Figure 3: Comparison of accelerated, nonaccelerated algorithm with both importance and τ nice sampling for a various quadratic problems.

D.1.2 Speedup in τ

The next experiment shows an empirical speedup for the coordinate descent algorithms for a various types of problems. For simplicity, we do not include Sampling 2. Figure 4 provides the results. Oftentimes, the empirical speedup (in terms of the number of iteration) in τ is close to linear, which demonstrates the power and significance of minibatching.
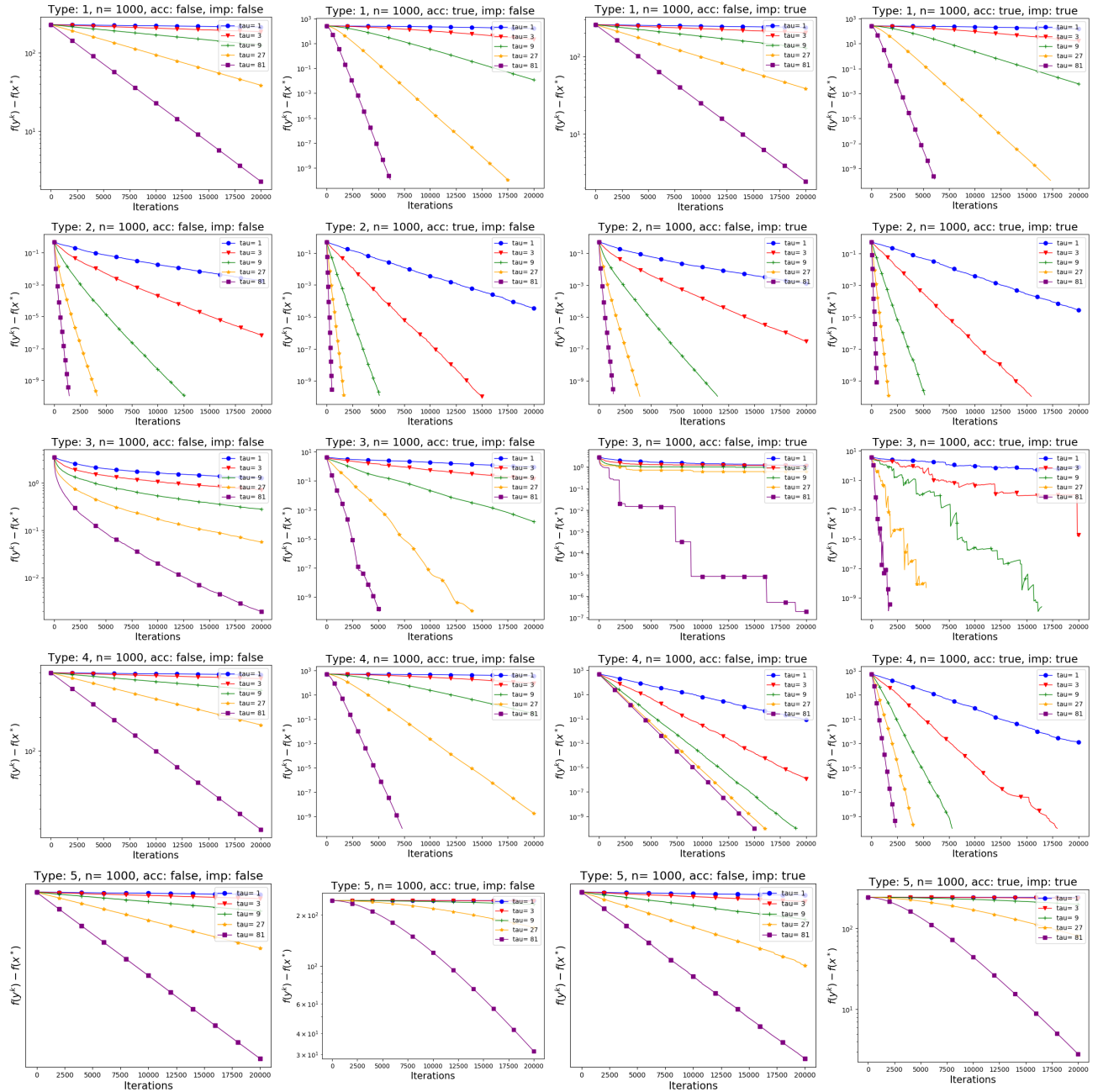

Figure 4: Comparison of speedup gained by both τ-nice sampling and importance sampling with and without acceleration on various quadratic problems.

D.2 Logistic Regression

In this section we apply `ACD` on the regularized logistic regression problem, i.e.

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(\mathbf{A}_{i,:} x \cdot b\right)\right) + \frac{\lambda}{2} \|x\|^2,$$

for $b \in \{-1, 1\}$ and data matrix $\mathbf{A}$ comes from LibSVM. In each experiment in this section, we have chosen regularization parameter λ to be the average diagonal element of the smoothness matrix. We first apply the methods with the optimal parameters as our theory suggests on smaller datasets. On larger ones (Section D.2.1), we set them in a cheaper way, which is not guaranteed to work by theory we provide.

In our first experiment, we apply `ACD` on LibSVM data directly for various minibatch sizes τ. Figure 5 shows the results. As expected, `ACD` is always better to `CD`, and importance sampling is always better to uniform one.
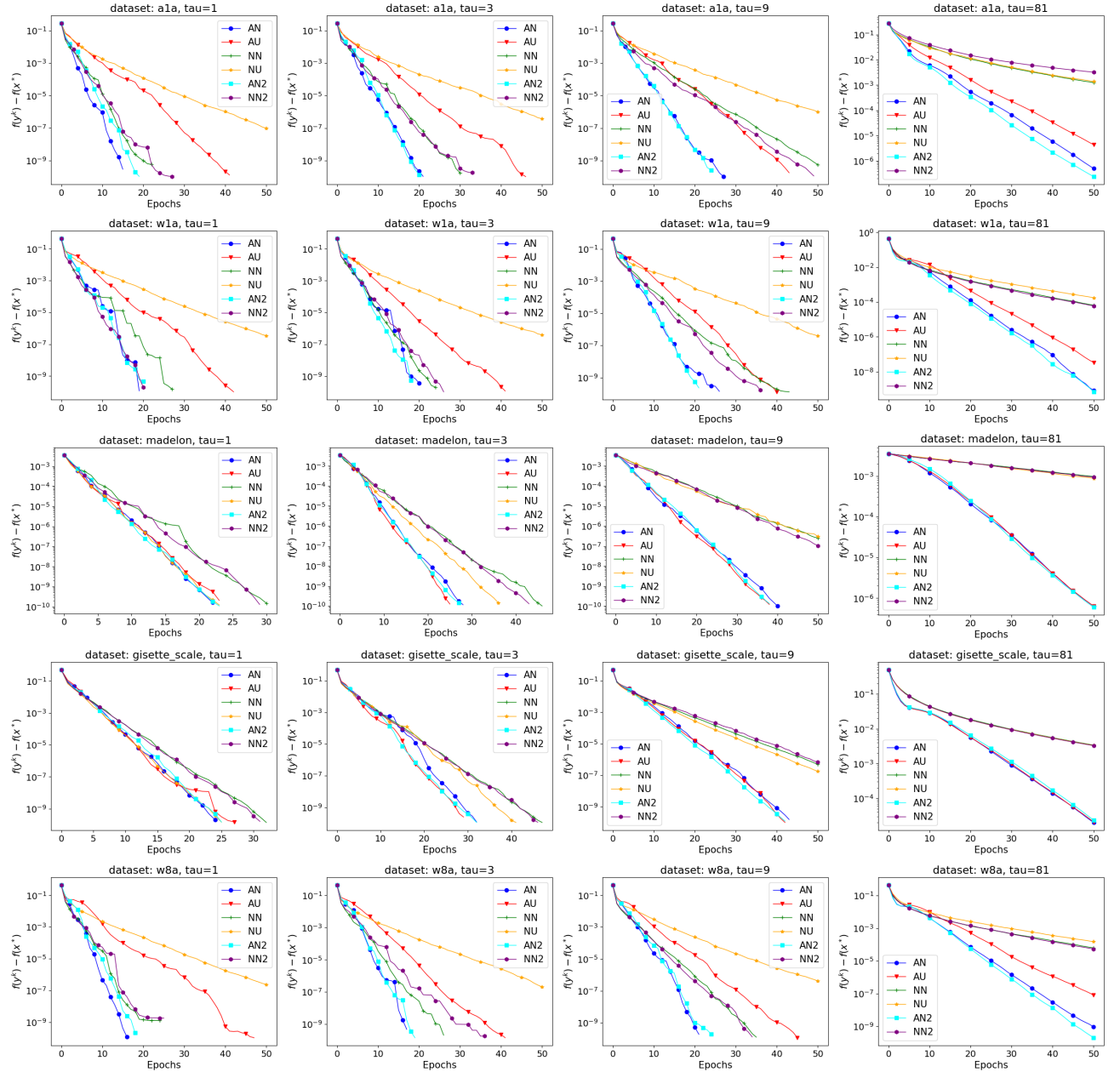

Figure 5: Accelerated coordinate desent applied on the logistic regression problem, for various LibSVM datasets and minibatch sizes τ

Note that, for some datasets and especially bigger minibatch sizes, the effect of importance sampling is sometimes negligible. To demonstrate the power of importance sampling, in the next experiment, we first corrupt the data – we multiply each row and column of the data matrix $\mathbf{A}$ by random number from uniform distribution over $[0, 1]$. The results can be seen in Figure 6. As expected, the effect of importance sampling becomes more significant.
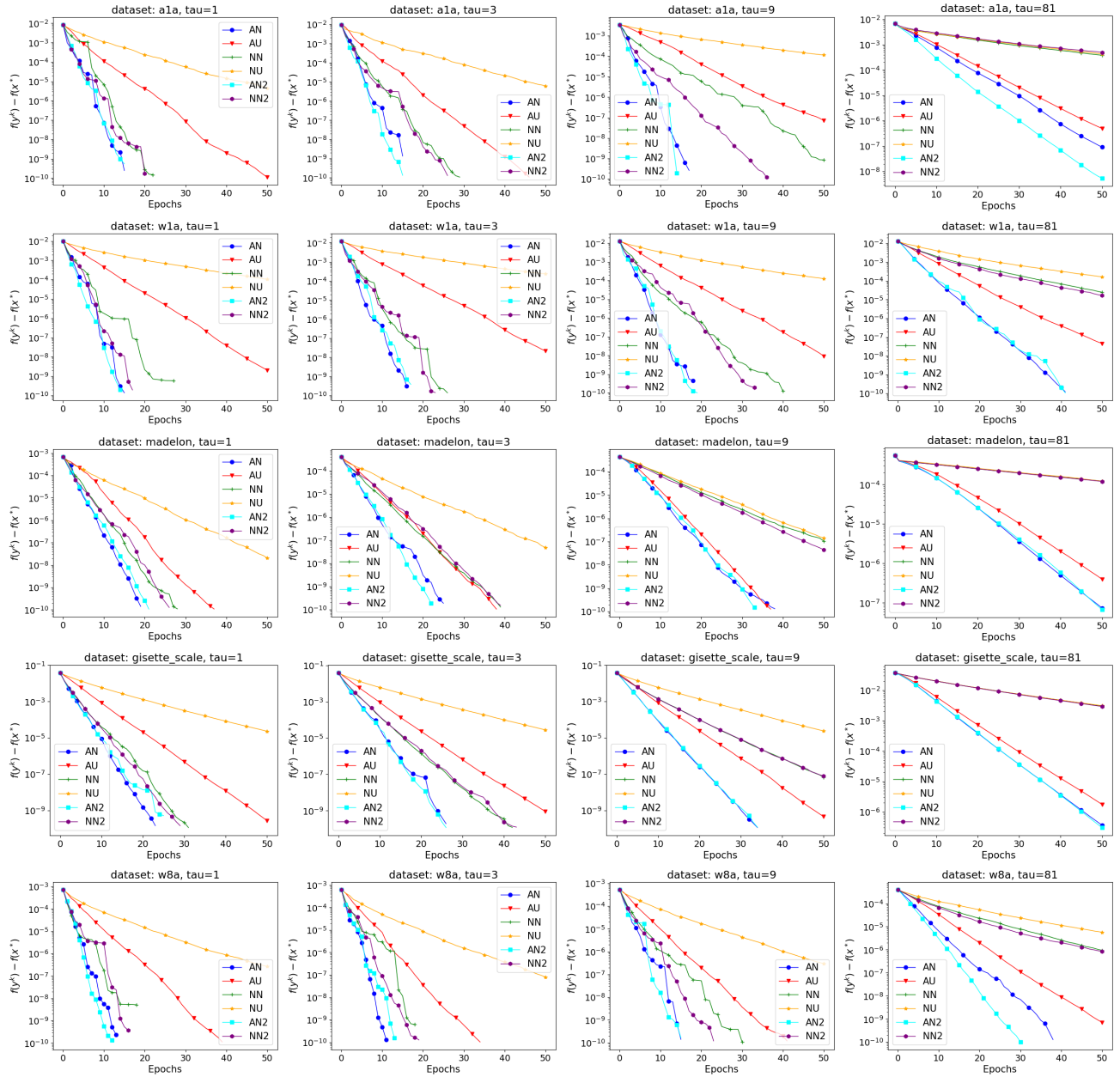

Figure 6: Accelerated coordinate desent applied on the logistic regression problem, for various rescaled LibSVM datasets and minibatch sizes τ

D.2.1 Practical method on larger dataset

For completeness, we restate here experiments from Figure 1. We have chosen regularization parameter λ to be the average diagonal element of the smoothness matrix and estimated v, σ as described in Section 6.
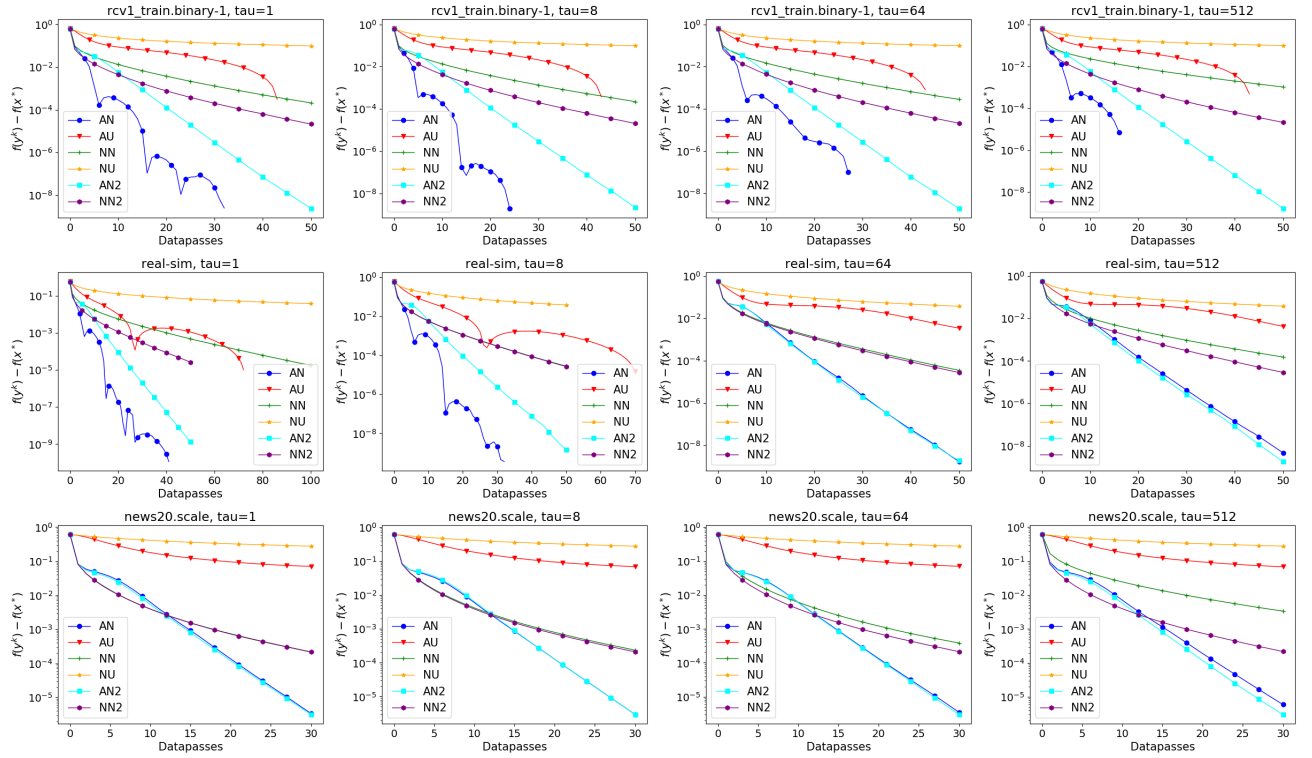

Figure 7: Six variants of coordinate descent (AN, AU, NN, NU, AN2 and AU2) applied to a logistic regression problem, with minibatch sizes $\tau = 1, 8, 64$ and 512.

D.3 Support Vector Machines

In this section we apply `ACD` on the dual of SVM problem with squared hinge loss, i.e.,

$$f(x) = \frac{1}{\lambda n^2} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} b_i \mathbf{A}_{ji} x_i \right)^2 - \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{4n} \sum_{i=1}^{n} x_i^2 + \mathcal{I}_{[0,\infty]}(x),$$

where $\mathcal{I}_{[0,\infty]}$ stands for indicator function of set $[0,\infty]$, i.e. $\mathcal{I}_{[0,\infty]}(x) = 0$ if $x \in \mathbb{R}^n_+$, otherwise $\mathcal{I}_{[0,\infty]}(x) = \infty$. As for the data, we have rescaled each row and each column of the data matrix coming frol LibSVM by random scalar generated from uniform distribution over $[0,1]$. We have chosen regularization parameter λ to be maximal diagonal element of the smoothness matrix divided by 10 in each experiment below. We deal with nonsmooth indicator function using proximal operator, which happens to be a projection in this case. We choose ESO parameters v from Lemma 5.2, while estimating the smoothness matrix as $\sqrt{n}$–times multiple of its diagonal. An estimate of the strong convexity σ for acceleration was chosen to be minimal diagonal element of the smoothness matrix, therefore we adapt a similar approach as in Section D.2.1.

Recall that we did not provide a theory for the proximal steps. However, we make the experiment to demonstrate that `ACD` can solve big data problems on top of large dimensional problems. Although the results are presented in the main body, we restate them here again (Figure 8) for the sake of readibility.
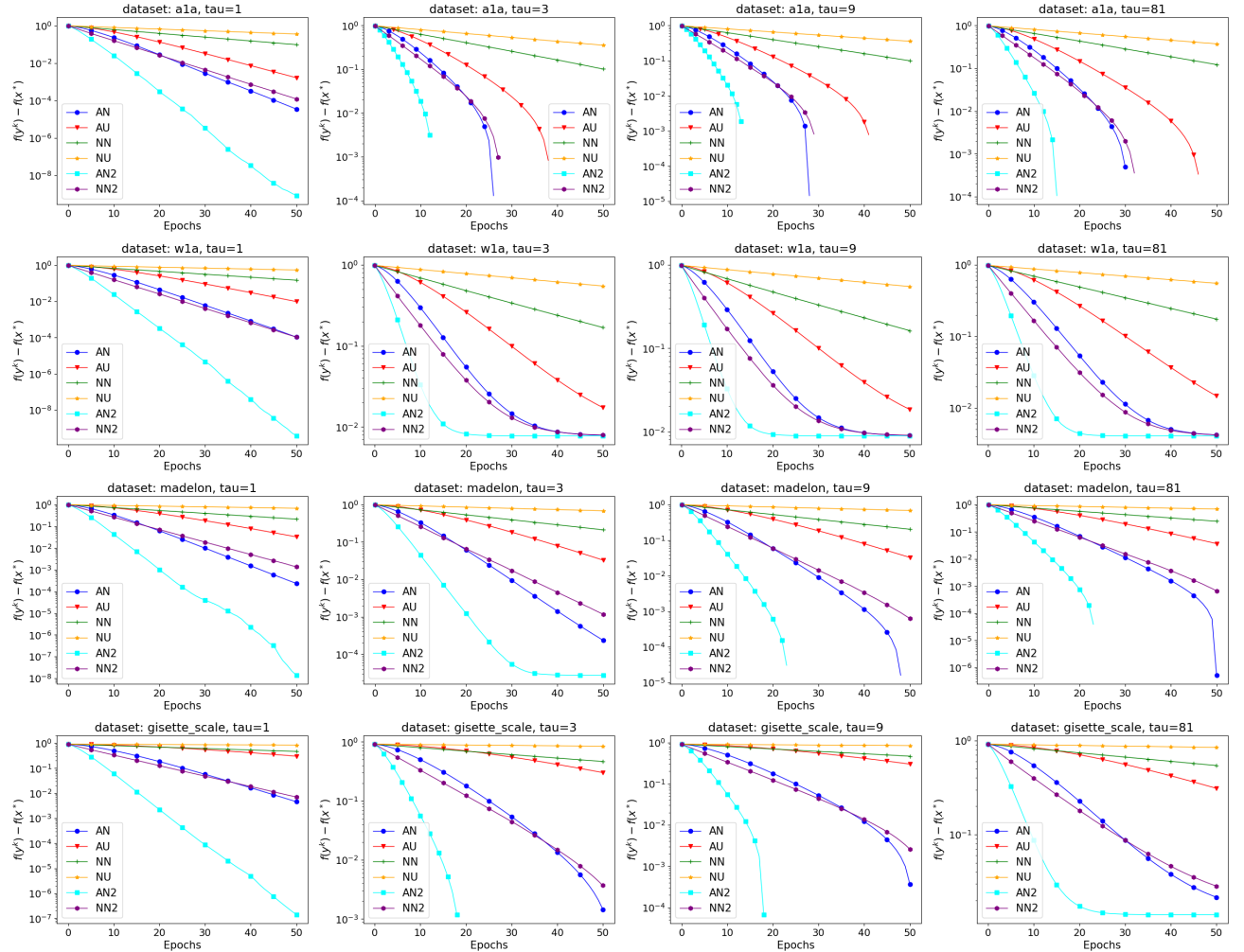

Figure 8: Accelerated coordinate desent applied on the dual of of SVM with squared hinge loss, for various LibSVM datasets