Faster Coordinate Descent via Adaptive Importance Sampling

Dmytro Perekrestenko
ETH Zurich

Volkan Cevher
EPFL

Martin Jaggi
EPFL

Abstract

Coordinate descent methods employ random partial updates of decision variables in order to solve huge-scale convex optimization problems. In this work, we introduce new *adaptive* rules for the random selection of their updates. By adaptive, we mean that our selection rules are based on the dual residual or the primal-dual gap estimates and can change at each iteration. We theoretically characterize the performance of our selection rules and demonstrate improvements over the state-of-the-art, and extend our theory and algorithms to general convex objectives. Numerical evidence with hinge-loss support vector machines and Lasso confirm that the practice follows the theory.

1 Introduction

Coordinate descent methods rely on random partial updates of decision variables for scalability. Indeed, due to their space and computational efficiency as well as their ease of implementation, these methods are the state-of-the-art for a wide selection of standard machine learning and signal processing applications [Fu, 1998, Hsieh et al., 2008, Wright, 2015].

Basic coordinate descent methods sample an active coordinate set for optimization uniformly at random, *cf.*, Stochastic Dual Coordinate Ascent (SDCA) [Shalev-Shwartz and Zhang, 2013] and other variants [Friedman et al., 2007, 2010, Shalev-Shwartz and Tewari, 2011]. However, recent results suggest that by employing an appropriately defined *non-uniform* fixed sampling strategy, the convergence can be improved both in the theory as well as in practice [Zhao and Zhang, 2014, Necorara et al., 2012, Nesterov, 2012].

In this work, we show that we can surpass the existing convergence rates by exploiting adaptive sampling strategies that change the sampling probability distribution during each iteration. For this purpose, we adopt the primal-dual framework of Dünner et al. [2016]. In contast, however, we also handle convex optimization problems with general convex regularizers without assuming strong convexity of the regularizer.

In particular, we consider an adaptive coordinate-wise duality gap based sampling. Hence, our work can be viewed as a natural continuation of the work of Csiba et al. [2015], where the authors introduce an adaptive version of SDCA for the smoothed hinge-loss support vector machine (SVM). However, our work generalizes the gap-based adaptive criterion of [Osokin et al., 2016] in a nontrivial way to a broader convex optimization template of the following form:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \; f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i), \tag{1}$$

where A is the data matrix, f is a smooth convex function, and each g_i is a general convex function.

The template problem class in (1) includes not only smoothed hinge-loss SVM, but also Lasso, Ridge Regression, (the dual formulation of the) original hinge-loss SVM, Logistic Regression, etc. As a result, our theoretical results for adaptive sampling can also recover the existing results for fixed non-uniform [Zhao and Zhang, 2014] and uniform [Dünner et al., 2016] sampling as special cases.

Contributions. Our contributions are as follows:

- We introduce new adaptive and fixed non-uniform sampling schemes for random coordinate descent for problems for the template (1).
- To our knowledge, we provide the first convergence rate analysis of coordinate descent methods with adaptive sampling for problems with general convex regularizer (i.e., the class in (1)).
- We derive convergence guarantees with arbitrary sampling distributions for both strongly convex and the general convex cases, and identify new convergence improvements.
- We support the theory with numerical evidence (i.e., Lasso and hinge-loss SVM) and illustrate significant performance improvements in practice.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

Outline: Section 2 provides basic theoretical preliminaries. Section 3 describes our theoretical results and introduces new sampling schemes. Section 4 discusses the application of our theory to machine learning, and compares the computational complexity of proposed sampling methods. Section 5 provides numerical evidence for the new methods. Section 6 discusses our contributions in the light of existing work.

2 Preliminaries

We recall some concepts from convex optimization used in the sequel. The *convex conjugate* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f^*(\boldsymbol{v}) := \sup_{\boldsymbol{u}\in\mathbb{R}^n} \boldsymbol{v}^\top \boldsymbol{u} - f(\boldsymbol{u})$.

Definition 2.1. *A function $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ has B-bounded support if its domain is bounded by B:*

$$f(\boldsymbol{u}) < +\infty \to \|\boldsymbol{u}\| \leq B.$$

Lemma 2.2 (Duality between Lipschitzness and L-Bounded Support, [Dünner et al., 2016])**.** *Given a proper convex function g, it holds that g has L-bounded support if and only if g^* is L-Lipschitz.*

2.1 Our primal-dual setup

In this paper, we develop coordinate descent methods for the following primal-dual optimization pair:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left[\mathcal{O}_A(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i)\right], \tag{A}$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\mathcal{O}_B(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w})\right], \tag{B}$$

where we have $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$.

Our primal-dual template generalizes the standard primal-dual SDCA setup in [Shalev-Shwartz and Zhang, 2013]. As a result, we can provide gap certificates for the quality of the numerical solutions while being applicable to broader set of problems.

Optimality conditions. The first-order optimality conditions for the problems (A) and (B) are given by

$$\begin{aligned} &\boldsymbol{w} \in \partial f(A\boldsymbol{\alpha}), && -\boldsymbol{a}_i^\top \boldsymbol{w} \in \partial g_i(\alpha_i) && \forall i \in [n] \\ &A\boldsymbol{\alpha} \in \partial f^*(\boldsymbol{w}), && \alpha_i \in \partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) && \forall i \in [n] \end{aligned} \tag{2}$$

For a proof, see [Bauschke and Combettes, 2011].

Duality gap. The duality gap is the difference between primal and dual solutions:

$$G(\boldsymbol{\alpha}, \boldsymbol{w}) := \mathcal{O}_A(\boldsymbol{\alpha}) - (-\mathcal{O}_B(\boldsymbol{w})), \tag{3}$$

which provides a certificate on the approximation accuracy both the primal and dual objective values. While always non-negative, under strong duality the gap reaches zero only in an optimal pair $(\boldsymbol{\alpha}^\star, \boldsymbol{w}^\star)$. When f is differentiable the optimality conditions (2) write as $\boldsymbol{w}^\star = \boldsymbol{w}(\boldsymbol{\alpha}^\star) = \nabla f(A\boldsymbol{\alpha}^\star)$. We will rely on the following relationship in our applications:

$$\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\alpha}) := \nabla f(A\boldsymbol{\alpha}) \ .$$

Choosing this mapping allows for running existing algorithms based on $\boldsymbol{\alpha}$ alone, without the algorithm needing to take care of $\boldsymbol{w}$. Nevertheless, the mapping allows us to express the gap purely in terms of the original variable $\boldsymbol{\alpha}$, at any time: We define $G(\boldsymbol{\alpha}) := G(\boldsymbol{\alpha}, \boldsymbol{w}(\boldsymbol{\alpha})) = \mathcal{O}_A(\boldsymbol{\alpha}) - (-\mathcal{O}_B(\boldsymbol{w}(\boldsymbol{\alpha})))$.

Coordinate-wise duality gaps. For our problem structure of partially separable problems (A) and (B), it is not hard to show that the duality gap can be written as a sum of coordinate-wise gaps:

$$G(\boldsymbol{\alpha}) = \sum_i G_i(\alpha_i) := \sum_i \left(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w}\right) \tag{4}$$

The relation holds since our mapping $\boldsymbol{w} = \nabla f(A\boldsymbol{\alpha})$ invokes the Fenchel-Young inequality for f with an equality. Moreover, the Fenchel-Young inequality for g_i implies that all $G_i(\alpha_i)$'s are non-negative.

2.2 Dual residuals

We base our fixed non-uniform and adaptive schemes on the concept of "dual residual," i.e., a measure of progress to the optimum of the dual variables $\boldsymbol{\alpha}$. Here we assume that $\boldsymbol{w} = \nabla f(A\boldsymbol{\alpha})$.

Definition 2.3 (Dual Residual. A generalization of [Csiba et al., 2015])**.** *Consider the primal-dual setting (A)-(B). Let each g_i be μ_i-strongly convex with convexity parameter $\mu_i \geq 0$ $\forall i \in [n]$. For the case $\mu_i = 0$ we require g_i to have a bounded support. Then, given $\boldsymbol{\alpha}$, the i-th dual residue on iteration t is given by:*

$$\kappa_i^{(t)} := \min_{u \in \partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)})} |u - \alpha_i^{(t)}|.$$

Remark 2.4. *Note that for u to be well defined, i.e., the subgradient in (6) not to be empty, we need the domain of g^* to be the whole space. For $\mu > 0$ this is given by strong convexity of g_i, while for $\mu_i = 0$ this follows from the bounded support assumption on g_i.*

Definition 2.5 (Coherent probability vector, [Csiba et al., 2015])**.** *We say that probability vector $\boldsymbol{p}^{(t)} \in \mathbb{R}^n$ is coherent with the dual residue vector $\boldsymbol{\kappa}^{(t)}$ if for all $i \in [n]$, we have $\kappa_i^{(t)} \neq 0 \ \Rightarrow \ p_i^{(t)} > 0$.*

Definition 2.6 (t-support set)**.** *We call the set*

$$I_t := \{i \in [n] : \kappa_i^{(t)} \neq 0\} \subseteq [n]$$

a t-support set.

Lemma 2.7. *Suppose that for each i, g_i^* is L_i-Lipschitz. Then, $\forall i : |\kappa_i| \le 2L_i$.*

Proof. By Lemma 2.2, the L_i-Lipschitzness g_i^* implies L_i-bounded support of $g_i(\alpha_i)$ and therefore $|\alpha_i| \le L_i$. By writing Lipschitzness as bounded subgradient, $|u_i| \le L_i$, and $|\kappa_i| = |\alpha_i - u_i| \le |\alpha_i| + |u_i| \le 2L_i$. □

2.3 Coordinate Descent

Algorithm 1 describes the Coordinate Descent (CD) method in the primal-dual setting (A) and (B). The method has 3 major steps: Coordinate selection, line-search along the chosen coordinate, and primal-dual parameter updates. While the standard CD methods chooses the coordinates at random with fixed distibutions, we develop adaptive strategies in the sequel that change the sampling distribution per iteration. The other steps remain essentially the same.

Algorithm 1 Coordinate Descent

1: Let $\boldsymbol{\alpha}^{(0)} := \mathbf{0} \in \mathbb{R}^n, \boldsymbol{w}^{(0)} := \boldsymbol{w}(\boldsymbol{\alpha}^{(0)})$
2: **for** $t = 0, 1, ..., T$ **do**
3: Sample $i \in [n]$ randomly according to $\boldsymbol{p}^{(t)}$
4: Find $\Delta\alpha_i$ minimizing $\mathcal{O}_A(\boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i\Delta\alpha_i)$
5: $\boldsymbol{\alpha}^{(t+1)} := \boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i\Delta\alpha_i$
6: $\boldsymbol{w}^{(t+1)} := \boldsymbol{w}(\boldsymbol{\alpha}^{(t+1)})$
7: **end for**

3 Adaptive Sampling-based CD

Our goal is to find a ε_B-suboptimal parameter $\boldsymbol{w}$ or ε_A-suboptimal parameter $\boldsymbol{\alpha}$, i.e., $\mathcal{O}_A(\boldsymbol{\alpha}) - \mathcal{O}_A(\boldsymbol{\alpha}^\star) \le \varepsilon_A$ or $\mathcal{O}_B(\boldsymbol{w}) - \mathcal{O}_B(\boldsymbol{w}^\star) \le \varepsilon_B$, for the following pair of dual optimization problems (A) and (B).

3.1 Key lemma

This subsection introduces a lemma that characterizes the relationship between any sampling distribution for the coordinates, denoted as $\boldsymbol{p}$, and the convergence rate of the CD method. For this purpose, we build upon the [Csiba et al., 2015, Lemma 3] to relax the strong-convexity restrictions on g_i's. That is, we derive a convergence result for the general convex g_i with coordinate-dependent strong convexity constants μ_i. In contrast to [Csiba et al., 2015, Lemma 3], we can have $\mu_i = 0$ when g_i has bounded support.

Lemma 3.1. *Let f be $1/\beta$-smooth and each g_i be μ_i-strongly convex with convexity parameter $\mu_i \ge 0$ $\forall i \in [n]$. For the case $\mu_i = 0$, we require g_i to have a bounded support. Then for any iteration t, any sampling distribution $\boldsymbol{p}^{(t)}$ and any arbitrary $s_i \in [0, 1]$ $\forall i \in [n]$, the iterates of the CD method satisfy*

$$\mathbb{E}[\mathcal{O}_A(\boldsymbol{\alpha}^{(t+1)}) \,|\, \boldsymbol{\alpha}^{(t)}] \le \mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) - \sum_i p_i^{(t)} \Big(\frac{\mu_i(s_i - s_i^2)}{2} - \frac{s_i^2 \|\boldsymbol{a}_i\|^2}{2\beta} \Big) |\kappa_i^{(t)}|^2, \tag{5}$$

here $\kappa_i^{(t)}$ is i-th dual residual (see Def. 2.3).

The proof is provided in Appendix A.

Remark 3.2. *If in addition to the conditions of Lemma 3.1 we require $\boldsymbol{p}^{(t)}$ to be coherent with $\boldsymbol{\kappa}^{(t)}$, then for any $\theta \in [0, \min_{i \in I_t} p_i^{(t)}]$ it holds that*

$$\mathbb{E}[\mathcal{O}_A(\boldsymbol{\alpha}^{(t+1)}) \,|\, \boldsymbol{\alpha}^{(t)}] \le \mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)}, \tag{6}$$

$$F^{(t)} := \frac{1}{n^2 \beta \theta} \sum_{i \in I_t} \Big(\frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^{(t)}} - \mu_i \beta \Big) |\kappa_i^{(t)}|^2.$$

Proof. Since s_i in Lemma 3.1 is an arbitrary number $\in [0, 1]$, we take $s_i = \frac{\theta}{p_i^{(t)}}$ for points with $i \in I_t$ and $s_i = 0$ for all other points, here $\theta \in [0, \min_i p_i^{(t)}]$. Then, (5) becomes the following, finalizing the proof:

$$\begin{aligned}
\mathbb{E}[\mathcal{O}_A(\boldsymbol{\alpha}^{(t+1)}) \,|\, \boldsymbol{\alpha}^{(t)}] &\le \mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \theta \sum_{i \in I_t} G_i(\boldsymbol{\alpha}^{(t)}) \\
&\quad - \sum_{i \in I_t} \Big(\frac{\mu_i \theta}{2} - \frac{\theta^2}{p_i^{(t)}} \frac{\mu_i \beta + \|\boldsymbol{a}_i\|^2}{2\beta} \Big) |\kappa_i^{(t)}|^2 \\
&= \mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) \\
&\quad - \frac{\theta}{2\beta} \sum_{i \in I_t} \Big(\mu_i \beta - \frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^{(t)}} \Big) |\kappa_i^{(t)}|^2
\end{aligned}$$

□

3.2 Why is the generalization important?

Consider the key lemma with $\mu_i = 0$. Then, Remark 3.2 implies the following:

$$\mathbb{E}[\mathcal{O}_A(\boldsymbol{\alpha}^{(t+1)}) \,|\, \boldsymbol{\alpha}^{(t)}] \le \mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)}, \tag{7}$$

where

$$F^{(t)} := \frac{1}{n^2 \beta} \sum_{i \in I_t} \Big(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \Big). \tag{8}$$

Contrary to the strongly convex case, $F^{(t)}$ is positive. Hence, the sampling distributions derived in [Csiba et al., 2015] with strongly convex g_i are not optimal.

The following theorem generalizes [Zhao and Zhang, 2014, Theorem 5], and [Dünner et al., 2016, Theorem 9] to allow adaptive sampling:

Theorem 3.3. *Assume f is a $\frac{1}{\beta}$-smooth function. Then, if g_i^* is L_i-Lipschitz for each i and $\boldsymbol{p}^{(t)}$ is coherent with $\boldsymbol{\kappa}^{(t)}$, then the CD iterates satisfy*

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{2F^\circ n^2 + \frac{2\varepsilon_A^{(0)}}{p_{\min}}}{\frac{2}{p_{\min}} + t}. \tag{9}$$

Moreover, we obtain a duality gap $G(\bar{\boldsymbol{\alpha}}) \leq \varepsilon$ *after an overall number of iterations T whenever*

$$T \geq \max\left\{0, \frac{1}{p_{\min}} \log\left(\frac{2\varepsilon_A^{(0)}}{n^2 p_{\min} F^\circ}\right)\right\} + \frac{5F^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}}. \tag{10}$$

Moreover, when $t \geq T_0$ with

$$T_0 := \max\left\{0, \frac{1}{p_{\min}} \log\left(\frac{2\varepsilon_A^{(0)}}{n^2 p_{\min} F^\circ}\right)\right\} + \frac{4F^\circ n^2}{\varepsilon} - \frac{2}{p_{\min}} \tag{11}$$

we have the suboptimality *bound of $\mathbb{E}[\mathcal{O}_A(\boldsymbol{\alpha}^{(t)}) - \mathcal{O}_A(\boldsymbol{\alpha}^\star)] \leq \varepsilon/2$. Here $\varepsilon_A^{(0)}$ is the initial dual suboptimality and F° is an upper bound on $\mathbb{E}[F^{(t)}]$ taken over the random choice of the sampled coordinate at $1, \ldots, T_0$ algorithm iterations.*

The proof is provided in Appendix A.

Remark 3.4. *We recover [Dünner et al., 2016, Theorem 9] as a special case of Theorem 3.3 by setting $p_i^{(t)} = \frac{1}{n}$. We recover [Zhao and Zhang, 2014, Theorem 5] by setting $p_i^{(t)} = \frac{L_i}{\sum_j L_j}$.*

3.2.1 Strategy I: Gap-wise sampling

Based on the results above, we first develop sampling strategies for the CD method based on the decomposibly of the duality gap, i.e., sampling each coordinate according to its duality gap.

Definition 3.5 (Nonuniformity measure, [Osokin et al., 2016])**.** *The nonuniformity measure $\chi(\boldsymbol{x})$ of a vector $\boldsymbol{x} \in \mathbb{R}^n$, is defined as:*

$$\chi(\boldsymbol{x}) := \sqrt{1 + n^2\, Var[\boldsymbol{p}]},$$

where $\boldsymbol{p} := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}$ is the normalized probability vector.

Lemma 3.6. *Let $\boldsymbol{x} \in \mathbb{R}^n_+$. Then, it holds that*

$$\|\boldsymbol{x}\|_2 = \frac{\chi(\boldsymbol{x})}{\sqrt{n}} \|\boldsymbol{x}\|_1.$$

Proof. The proof follows from Def. 3.5 and

$$\mathrm{Var}[\boldsymbol{p}] = \mathbb{E}[\boldsymbol{p}^2] - \mathbb{E}[\boldsymbol{p}]^2 = \tfrac{1}{n}\|\boldsymbol{p}\|_2^2 - \tfrac{1}{n^2}.$$ □

Theorem 3.7. *Let f be a $\frac{1}{\beta}$-smooth function. Then, if g_i^* is L_i-Lipschitz for each i and $p_i^{(t)} := \frac{G_i(\boldsymbol{\alpha}^{(t)})}{G(\boldsymbol{\alpha}^{(t)})}$, then the iterations of the CD method satisfies*

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{2F_g^\circ n^2 + 2n\varepsilon_A^{(0)}}{t + 2n}, \tag{12}$$

where F_g° is an upper bound on $\mathbb{E}\left[F_g^{(t)}\right]$, where the expectation is taken over the random choice of the sampled coordinate at iterations $1, \ldots, t$ of the algorithm. Here $\overrightarrow{\boldsymbol{G}}$ and $\overrightarrow{\boldsymbol{F}}$ are defined as follows:

$$\overrightarrow{\boldsymbol{G}} := (G_i(\boldsymbol{\alpha}^{(t)}))_{i=1}^n, \qquad \overrightarrow{\boldsymbol{F}} := (\|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2)_{i=1}^n,$$

and $F_g^{(t)}$ is defined analogously to (8):

$$F_g^{(t)} := \frac{\chi(\overrightarrow{\boldsymbol{F}})}{n\beta(\chi(\overrightarrow{\boldsymbol{G}}))^3} \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2. \tag{13}$$

The proof is provided in Appendix A.

Gap-wise vs. Uniform Sampling: Here we compare the rates obtained by Theorem 3.7 for gap-wise sampling and Theorem 3.3 for uniform sampling. According to the Theorem 3.3, the rate for any distribution can be written as follows

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{2F^\circ n^2 + \frac{2\varepsilon_A^{(0)}}{p_{\min}}}{\frac{2}{p_{\min}} + t} = \frac{\frac{2}{\beta}\mathbb{E}\left[\sum_i \frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^t}\right] + \frac{2\varepsilon_A^{(0)}}{p_{\min}}}{\frac{2}{p_{\min}} + t}.$$

For the uniform distribution ($p_i = 1/n$), this yields

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{\frac{2n}{\beta}\mathbb{E}\left[\sum_i |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2\right] + 2n\varepsilon_A^{(0)}}{2n + t}. \tag{14}$$

The rate of gap-wise sampling depends on non-uniformity measures $\chi(\overrightarrow{\boldsymbol{G}})$ and $\chi(\overrightarrow{\boldsymbol{F}})$:

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{\frac{2n}{\beta}\mathbb{E}\left[\frac{\chi(\overrightarrow{\boldsymbol{F}})}{(\chi(\overrightarrow{\boldsymbol{G}}))^3} \sum_i |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2\right] + 2n\varepsilon_A^{(0)}}{2n + t}.$$

In the best case for gap-wise sampling the variance in $(|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2)_{i=1}^n$ is 0, $\chi(\overrightarrow{\boldsymbol{F}}) \approx 1$, and variance of gaps is maximal $\chi(\overrightarrow{\boldsymbol{G}}) \approx \sqrt{n}$. When this condition holds, the convergence rate becomes the following:

$$\mathbb{E}[\varepsilon_A^{(t)}] \leq \frac{\frac{2}{\beta\sqrt{n}}\mathbb{E}\left[\sum_i |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2\right] + 2n\varepsilon_A^{(0)}}{2n + t}.$$

In the worst case scenario, when variance is maximal in $(|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2)_{i=1}^n$, $\chi(\overrightarrow{\boldsymbol{F}}) \approx \sqrt{n}$, the rate of gap-wise sampling is better than of uniform only when the gaps are non-uniform enough i.e., $\chi(\overrightarrow{\boldsymbol{G}}) \geq n^{\frac{1}{6}}$.

3.3 Strategy II: Adaptive & Uniform

Instead of minimizing (10), we here find an optimal sampling distribution as to minimize our bound:

$$T \geq \frac{5F^\circ n^2}{\varepsilon} + \frac{5\varepsilon_A^{(0)}}{\varepsilon p_{\min}}. \tag{15}$$

The number of iterations T is directly proportional to F° and $1/p_{\min}$. Therefore, the optimal distribution $\boldsymbol{p}$ should minimize F° and $1/p_{\min}$ at the same time.

We denote the distribution minimizing $1/p_{\min}$ as *supportSet uniform*, which is the following rule:

$$p_i^{(t)} := \begin{cases} \frac{1}{m_t}, & \text{if } \kappa_i^{(t)} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Above, m_t is a cardinality of the support set on iteration t. The distribution minimizing F°, called *adaptive*:

$$p_i^{(t)} := \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}. \tag{17}$$

The mix of the two aforementioned distributions balances two terms and gives a good suboptimal T in (15). We define mixed distribution as:

$$p_i^{(t)} := \begin{cases} \frac{\sigma}{m_t} + (1-\sigma)\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where $\sigma \in [0,1]$. This distribution gives us the following bounds on F° and $1/p_{\min}$:

$$F^\circ_{\text{mix}} \leq \frac{F^\circ_{\text{ada}}}{1-\sigma} \qquad \frac{1}{p_{\min}} \leq \frac{m}{\sigma}.$$

and bound on the number of iterations:

$$T \geq \frac{5F^\circ_{\text{ada}} n^2}{\varepsilon(1-\sigma)} + \frac{5\varepsilon_A^{(0)} m}{\varepsilon\sigma}. \tag{19}$$

Above $m := \max_t m_t$. Since the process of finding F°_{ada} is rather problematic, a good σ can be found by replacing F°_{ada} with its upper bound and minimizing (18) w.r.t. σ. Another option is to use a "safe" choice of $\sigma = 0.5$, and provide a balance between two distributions. This strategy benefits from convergence guarantees in case of unknown F°_{ada}. In the applications section we use the latter option and call this sampling variant *ada-uniform sampling*.

3.4 Variations along the theme

Based on the discussion above, we summarize our new variants of sampling schemes for Algorithm 1:

- *uniform* - sample uniformly at random.
- *supportSet uniform* - sample uniformly at random inside the support set, defined in (16). The distribution is recomputed every iteration.
- *adaptive* - sample adaptively based on dual residual, defined in (17). The distribution is recomputed every iteration.
- *ada-uniform* - sample based on a mixture between *supportSet uniform* and *adaptive*, defined in (18). The distribution is recomputed every iteration.
- *importance* - sample with a fixed non-uniform variant of *adaptive* obtained by bounding $\kappa_i^{(t)}$ with $2L_i$ (Lemma 2.7): $p_i := \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|}$. The distribution is computed only once. When the data is normalized, this sampling variant coincides with "importance sampling" of [Zhao and Zhang, 2014].
- *ada-gap* - sample randomly based on coordinate-wise duality gaps, defined in Section 3.2.1. The distribution is recomputed every iteration.
- *gap-per-epoch* - Use *ada-gap* but with updates per-epoch. The gap-based distribution is only recomputed at the beginning of each epoch and stays fixed during each epoch.

Full descriptions of the variants are in Appendix B.

4 Applications

4.1 Lasso and Sparse Logistic Regression

The Lasso and L1-regularized Logistic Regression are quintessential problems with a general convex regularizer. Given a data matrix $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]$ and a vector $\boldsymbol{y} \in \mathbb{R}^d$, the Lasso is stated as:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1. \tag{20}$$

Both problems can easily be reformulated in our primal-dual setting (A)-(B), by choosing $g_i(\alpha_i) := \lambda|\alpha_i|$. We have $f(A\boldsymbol{\alpha}) := \|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2$ for Lasso, and $f(A\boldsymbol{\alpha})$ is the logistic loss for classification respectively. To our knowledge, there is no importance sampling or adaptive sampling techniques for CD in this setting.

Lipschitzing trick. In order to have duality gap convergence guarantees (Theorem 3.3) we need g_i^* to be Lipschitz continuous, which however is not the case for $g_i = |.|$ being the absolute value function. We modify the function g_i without affecting the iterate sequence of CD using the "Lipschitzing trick" from [Dünner et al., 2016], as follows.

According to Lemma 2.2, a proper convex function g_i has bounded support if and only if g_i^* is Lipschitz continuous. We modify $g_i(\alpha_i) = \lambda|\alpha_i|$ by restricting its support to the interval with radius $B := \frac{1}{\lambda}(f(A\boldsymbol{\alpha}^{(0)}) + \lambda\|\boldsymbol{\alpha}^{(0)}\|_1)$. Since Algorithm 1 is monotone, we can choose B big enough to guarantee that $\boldsymbol{\alpha}^{(t)}$ will stay inside the ball during optimization, i.e. that the algorithm's iterate sequence will not be affected by B. By

modifying g_i to bounded support of size B, we guarantee g_i^* to be B-Lipschitz continuous.

$$\bar{g}_i(\alpha_i) := \begin{cases} \lambda|\alpha_i|, & \text{if } |\alpha_i| \le B \\ +\infty, & \text{otherwise} \end{cases}$$

The conjugate of $\bar{g}_i$ is:

$$\bar{g}_i^*(u_i) = \max_{\alpha_i : |\alpha_i| \le B} u_i\alpha_i - \lambda|\alpha_i| = B\big[|u_i| - \lambda\big]_+.$$

Duality gap. Using the gap decomposition (4) we obtain coordinate-wise duality gaps for modified Lasso and sparse logistic regression, which now depends on the chosen parameter B:

$$G(\boldsymbol{\alpha}) = \sum_i \Big(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w} \Big) \qquad (21)$$
$$= \sum_i \Big(B\big[|\boldsymbol{a}_i^\top \boldsymbol{w}| - \lambda\big]_+ + \lambda|\alpha_i| + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w} \Big).$$

4.2 Hinge-Loss SVM

Our framework directly covers the original hinge-loss SVM formulation. The importance sampling technique [Zhao and Zhang, 2014] are not applicable to the original hinge-loss, but relies on a smoothed version of the hinge-loss, changing the problem.

When $\varphi_i(.)$ is the hinge-loss, defined as $\varphi_i(s) := [1 - sy_i]_+$, our framework is directly applicable by mapping the SVM dual problem to our template (A), that is

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \mathcal{O}_A(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^n \varphi_i^*(-\alpha_i) + \frac{\lambda}{2}\Big\| \frac{1}{\lambda n}\sum_{i=1}^n \alpha_i \boldsymbol{a}_i \Big\|_2^2. \qquad (22)$$

The conjugate of the hinge-loss is $\varphi_i^*(\alpha_i) = \alpha_i y_i$, with $\alpha_i y_i \in [0,1]$. In other words, $g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) = \frac{1}{n}\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w})$ in our notation, and $f^*(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$.

Duality gap. Section 2.1 shows that the duality gap decomposes into a sum of coordinate-wise gaps.

4.3 Computational costs

We discuss the computational costs of the proposed variants under the different sampling schemes. Table 1 states the costs in detail, where *nnz* is the number of non-zero entries in the data matrix A. In the table, one *epoch* means n consecutive coordinate updates, where n is the number features in the Lasso, and is the number of datapoints in the SVM.

Sampling and probability update In each iteration, we sample a coordinate from a non-uniform probability distribution. While the straightforward approach requires $\Theta(n)$ per sample, it is not hard to see that this can be improved to $\Theta(\log(n))$ when using a tree data structure to maintain the probability vector [Nesterov, 2013, Shalev-Shwartz and Wexler, 2016]. The tree structure can be built in $\mathcal{O}(n\log(n))$.

Table 1: A summary of computational costs

Algorithm	Cost per Epoch
uniform	$\mathcal{O}(\text{nnz})$
importance	$\mathcal{O}(\text{nnz} + n\log(n))$
gap-per-epoch	$\mathcal{O}(\text{nnz} + n\log(n))$
supportSet-uniform	$\mathcal{O}(n \cdot \text{nnz})$
adaptive	$\mathcal{O}(n \cdot \text{nnz})$
ada-uniform	$\mathcal{O}(n \cdot \text{nnz})$
ada-gap	$\mathcal{O}(n \cdot \text{nnz})$

Variable update and distribution generation Computing all dual residuals κ_i or all coordinate-wise duality gaps G_i is as expensive as an epoch of the classic CD method, i.e., we need to do $\Theta(\text{nnz})$ operations (one matrix-vector multiplication). In contrast, updating one coordinate α_i is cheap, being $\Theta(\text{nnz}/n)$.

Total cost per epoch In a naive implementation, the most expensive sampling schemes are *adaptive, supportSet-uniform, ada-uniform* and *ada-gap*. Those completely recompute the sampling distribution after each iteration, giving a total per-epoch complexity of $\mathcal{O}(n \cdot \text{nnz})$. In contrast, the fixed non-uniform sampling scheme *importance* requires to build the sampling distribution only once, or once per epoch for *gap-per-epoch* (both giving $\mathcal{O}(\text{nnz})$ operations). The complexity of n samplings using the tree structure is $\mathcal{O}(n\log(n))$, the complexity of a variable update is $\mathcal{O}(\text{nnz})$. Overall, the asymptotic complexity therefore is $\mathcal{O}(n\log(n) + \text{nnz})$ per epoch, compared to $\mathcal{O}(\text{nnz})$ for simple uniform sampling.

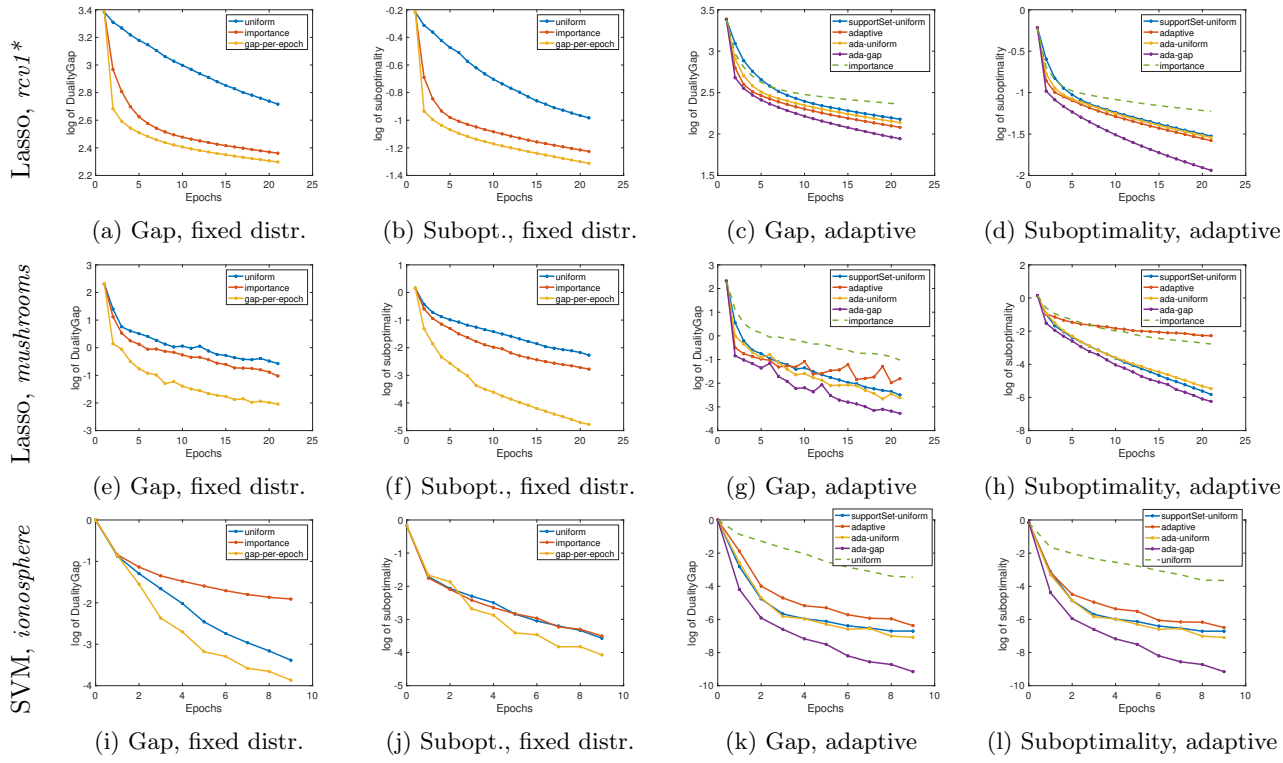
5 Experimental results

We provide numerical evidence for our CD sampling strategies on two key machine learning problems: The Lasso and the hinge-loss SVM. All our algorithms and theory are also directly applicable to sparse logistic regression and others, but we omit experiments due to space limitations.

Datasets. The experiments are performed on three standard datasets listed in Table 2, available[1] from the UCI repository [Asuncion and Newman, 2007]. Note that *rcv1** is a randomly subsampled[2] version the *rcv1*

[1] www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[2] We randomly picked 10000 datapoints and 1000 features, and then removed zero rows and columns.

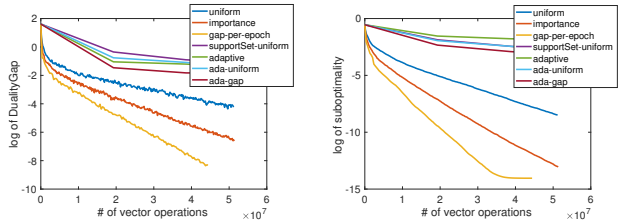


(a) Gap, fixed distr. (b) Subopt., fixed distr. (c) Gap, adaptive (d) Suboptimality, adaptive

(e) Gap, fixed distr. (f) Subopt., fixed distr. (g) Gap, adaptive (h) Suboptimality, adaptive

(i) Gap, fixed distr. (j) Subopt., fixed distr. (k) Gap, adaptive (l) Suboptimality, adaptive

Figure 1: Lasso (first two rows) and SVM (bottom row). Comparison of different fixed and adaptive variants of CD, reporting duality gap and suboptimality measures vs. epochs - rcv1*, mushrooms and ionosphere datasets.



(a) Gap (b) Suboptimality

Figure 2: Lasso on the *mushrooms* dataset. Performance in terms of duality gap and suboptimality, plotted against the total number of vector operations.

dataset. Experiments on the full *rcv1* dataset are provided in Appendix C.

Table 2: Datasets

Dataset	d	n	nnz/(nd)	$c_v = \frac{\mu(\|\boldsymbol{a}_i\|)}{\sigma(\|\boldsymbol{a}_i\|)}$
mushrooms	112	8124	18.8%	1.34
rcv1*	809	7438	0.3%	0.62
ionosphere	351	33	88%	3.07

Setup. For Lasso, the regularization parameter λ in (20) is set such that the cardinality of the true support set is between 10% and 15 % of the total number of features n. We use $\lambda = 0.05$ for *mushrooms*, and $\lambda = 7 \cdot 10^{-4}$ for *rcv1**. For hinge-loss SVM, the regularization parameter λ is chosen such that the classification error on the test set was comparable to training error. We use $\lambda = 0.1$ for *ionosphere*.

Performance. Figures 1 and 2 show the performance of all our studied variants of CD. We record *suboptimality* and *duality gap* (see (3)) as the main measures of algorithm performance. All reported results are averaged over 5 runs of each algorithm.

Methods with fixed sampling distributions. For the three efficient sampling schemes, our results show that CD importance converges faster than CD uniform on both datasets for Lasso, however it is worse than the uniform on SVM. The "mildly" adaptive strategy gap-per-epoch, based on our coordinate-wise duality gap theory but computed only per-epoch, significantly outperforms both of them. This is observed both in number of epochs (Figure 1) as well as number of total vector operations (Figure 2).

Methods with adaptive sampling distributions For the adaptive methods updating the probabilities after each coordinate step in Figure 1, we show importance sampling as a baseline method (dashed line). We see that measured per epoch, all adaptive meth-

ods outperform the fixed sampling methods.Among all adaptive methods, the ada-gap algorithm shows better convergence speed with both suboptimality and duality gap measures.

Highlights.

- The experiments for Lasso show a clear advantage of non-uniform sampling over the uniform, and superiority of the adaptive sampling over the fixed non-uniform, which is supported by our theory.
- Among the adaptive methods per iteration, the best performance for both Lasso and SVM in terms of epochs is by ada-gap, which has proven convergence bounds (Theorem 3.7), but also has high computational cost ($\Theta(d \cdot \text{nnz})$).
- The best sampling scheme in terms of total computational cost is gap-per-epoch, which is the epoch-wise variant of the ada-gap algorithm (based on recomputing duality gaps once per epoch), as shown in Figure 2.

6 A discussion on the results

Coordinate descent methods have a rich history in the discipline of optimization as well as many machine learning applications, *cf.*, [Wright, 2015] for a review.

For SVMs, CD related methods have been studied since their introduction, e.g., by [Friess et al., 1998]. Hsieh et al. [2008] is the first to propose CD in the partially separable primal-dual setting for hinge-loss. Theoretical convergence rates beyond the application of the hinge-loss can be found in the SDCA line of work [Shalev-Shwartz and Zhang, 2013], which is the primal-dual analog of the primal-only SGD algorithms. However, the main limitation of SDCA is that it is only applicable to strongly convex regularizers, or requires smoothing techniques [Nesterov, 2005] to be applicable to general regularizers such as L1. The technique of Dünner et al. [2016] can extend to the CD algorithms as well as the primal-dual analysis to the problem class of interest here, using a bounded set of interest for the iterates instead of relying on smoothing.

The convergence rate of stochastic methods (such as CD and SGD) naturally depends on a sampling probability distribution over the coordinates or datapoints respectively. While virtually all existing methods use sampling uniformly at random [Hsieh et al., 2008, Shalev-Shwartz and Tewari, 2011, Lacoste-Julien et al., 2013, Shalev-Shwartz and Zhang, 2012], recently [Nesterov, 2012, Qu and Richtárik, 2016, Zhao and Zhang, 2014, Allen-Zhu et al., 2016] showed that an appropriately defined fixed non-uniform sampling distribution, dubbed as *importance sampling*, can significantly improve the convergence.

The work of [Csiba et al., 2015] has taken the non-uniform sampling a step further towards adaptive sampling. While restricted to strongly convex regularizers, the rates provided for the AdaSDCA algorithm of [Csiba et al., 2015] - when updating all probabilities after each step - can beat the ones for uniform and *importance sampling*. A different approach for adapting the sampling distribution is proposed in [Osokin et al., 2016], where the block coordinate Frank-Wolfe algorithm is enhanced with sampling proportional to values of block-wise duality gaps. An adaptive variant of SGD is studied by [Papa et al., 2015], where they proposed an adaptive sampling scheme dependent on the past iterations in a Markovian manner, without giving explicit convergence rates. Other adaptive heuristics without proven convergence guarantees include ACF [Glasmachers and Dogan, 2014] and ACiD [Loshchilov et al., 2011].

For general convex regularizers such as L1, the development of CD algorithms includes [Fu, 1998, Friedman et al., 2007, 2010] and more recent extensions also improving the theoretical convergence rates [Shalev-Shwartz and Tewari, 2011, Johnson and Guestrin, 2015, Zhao et al., 2014]. All are restricted to uniform sampling, and we are not aware of proven convergence rates showing improvements of non-uniform or even adaptive sampling for unmodified L1 problems. [Zhao and Zhang, 2014, Allen-Zhu et al., 2016] show improved rates for non-uniform sampling for L1 but require a smoothing modification of the original problem, and are not covering adaptive sampling.

Conclusion. In this work, we have investigated *adaptive* rules for adjusting the sampling probabilities in coordinate descent. Our theoretical results provide improved convergence rates for a more general class of algorithm schemes on one hand, and optimization problems on the other hand, where we are able to directly analyze CD on general convex objectives (as opposed to strongly convex regularizers in previous works). This is particularly useful for L1 problems and (original) hinge-loss objectives, which were not covered by previous schemes. Our practical experiments confirm the strong performance of the adaptive algorithms, and confirm that the behavior predicted by our theory. Finally, we advocate the use of the computationally efficient gap-per-epoch sampling scheme in practice. While the scheme is close to the ones supported by our theory, an explicit primal-dual convergence analysis remains a future research question.

Acknowledgements. We thank Dominik Csiba and Peter Richtárik for helpful discussions. VC was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750, and SNF CRSII2-147633.

References

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even Faster Accelerated Coordinate Descent Using Non-Uniform Sampling. In *ICML 2016 - Proceedings of the 33th International Conference on Machine Learning*, 2016.

A Asuncion and DJ Newman. UCI Machine Learning Repository. *Miscellaneous*, 2007.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. 2011.

Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic Dual Coordinate Ascent with Adaptive Probabilities. In *ICML 2015 - Proceedings of the 32th International Conference on Machine Learning*, 2015.

Celestine Dünner, Simone Forte, Martin Takáč, and Martin Jaggi. Primal-Dual Rates and Certificates. In *ICML 2016 - Proceedings of the 33th International Conference on Machine Learning*, 2016.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33 (1):1–22, 2010.

Thilo-Thomas Friess, Nello Cristianini, and Colin Campbell. The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines. *ICML 1998 - Proceedings of the 15th International Conference on Machine Learning*, 1998.

Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

Tobias Glasmachers and Ürün Dogan. Coordinate Descent with Online Adaptation of Coordinate Frequencies. *arXiv*, 2014.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and S Sundararajan. A Dual Coordinate Descent Method for Large-scale Linear SVM. ICML - *the 25th International Conference on Machine Learning*, 2008.

Tyler Johnson and Carlos Guestrin. Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization. In *ICML 2015 - Proceedings of the 32th International Conference on Machine Learning*, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.

Ilya Loshchilov, Marc Schoenauer, and Michèle Sebag. Adaptive Coordinate Descent. In Natalio Krasnogor and Pier Luca Lanzi, editors, *Genetic and Evolutionary Computation Conference (GECCO 2011)*, pages 885–992, Dublin, Ireland, 2011. ACM-SIGEVO.

I Necorara, Yurii Nesterov, and François Glineur. Efficiency of Randomized Coordinate Descent Methods on Optimization Problems With Linearly Coupled Constraints. *Technical Report*, 2012.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Yurii Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov. Subgradient methods for huge-scale optimization problems. *Mathematical Programming*, 146 (1-2):275–297, 2013.

Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, and Simon Lacoste-Julien. Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs. In *ICML 2016 - Proceedings of the 33th International Conference on Machine Learning*, 2016.

Guillaume Papa, Pascal Bianchi, and Stéphan Clémençon. Adaptive Sampling for Incremental Optimization Using Stochastic Gradient Descent. In *ALT - 26th International Conference on Algorithmic Learning Theory*, pages 317–331. 2015.

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

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic Methods for l_1-regularized Loss Minimization. *JMLR*, 12: 1865–1892, 2011.

Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the Maximal Loss: How and Why? *arXiv*, 2016.

Shai Shalev-Shwartz and Tong Zhang. Proximal Stochastic Dual Coordinate Ascent. *arXiv*, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *JMLR*, 14:567–599, 2013.

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

Peilin Zhao and Tong Zhang. Stochastic Optimization with Importance Sampling. *arXiv*, 2014.

Tuo Zhao, Han Liu, and Tong Zhang. Pathwise Coordinate Optimization for Sparse Learning: Algorithm and Theory. *arXiv*, 2014.