Faster optimization through adaptive importance sampling

Master Thesis

Author(s):
Perekrestenko, Dmytro

Publication date:
2016

Permanent link:
https://doi.org/10.3929/ethz-a-010702800

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Master Thesis
Spring 2016

Faster Optimization through Adaptive Importance Sampling

Section:
Electrical and Electronics Engineering

Student:
Dmytro Perekrestenko

Supervisors:
Prof. Martin Jaggi
Prof. Volkan Cevher

Submission date:
August 12, 2016

Abstract

The current state of the art stochastic optimization algorithms (SGD, SVRG, SCD, SDCA, etc.) are based on sampling one active datapoint uniformly at random in each iteration. Changing these probabilities to better reflect the importance of each datapoint is a natural and powerful idea. In this thesis we analyze Stochastic Coordinate Descent methods with fixed non-uniform and adaptive sampling. We consider problems with strongly convex (e.g. SVM) and general convex (e.g. Lasso) regularizers and obtain new upper bounds on the number of iterations needed to reach given duality gap or suboptimality. Experiments on smoothed hinge loss SVM and Lasso are provided to confirm the theoretical analysis.

Keywords. *adaptive sampling, non-uniform sampling, gap-wise sampling, stochastic coordinate descent, primal-dual framework, lasso, svm*

Contents

Chapter 1

Introduction

Coordinate descent methods [6] solve optimization problems by successively performing line-search along coordinate directions. Their advantages are simplicity, ease of implementation and scalability. A lot of their implementations have become state-of-the-art for many machine learning problems [12, 7].

The traditional coordinate descent algorithms such as Stochastic Dual Coordinate Ascent (SDCA) [28] and Stochastic Coordinate Descent (SCD) [24] sample one active datapoint uniformly at random in each iteration. While the uniform sampling guarantees the unbiasedness of the estimate, observably high variance aquired by uniform sampling negatively affects the convergence rate. In the recent work [32] it is shown that by employing an appropriately defined non-uniform fixed sampling strategy the convergence rate can be significantly improved. In this thesis we go further and achieve even better convergence rate by exploiting the notion of adaptive sampling, which is based on changing the sampling probability distribution each iteration according to the data and values of the dual variables.

The well-known duality setting of SDCA [28, 27] is restricted to strongly convex regularizers and finite sum optimization problems. Therefore the majority of recent research on Coordinate Descent methods is focused only on minimization with strongly convex regularizers, e.g. [20], [4], etc.. In our work we adopt the new primal-dual framework from Dünner et al. (2016) [5], which contrary to the existing methods of adding a strongly-convex (L_2) term [17, 21], allows us to optimize problems with general convex regularizers leaving the algorithms and optima unaffected.

This thesis is a natural continuation of the work of Csiba et al. (2015) [4],

where they developed an adaptive version of SDCA for the one particular problem of smoothed hinge loss SVM. Using the same approach, we equip each dual variable with a measure of progress called "dual residue" (we use a generalized version of "dual residue" in [4]) and base our adaptive sampling scheme on these measures. We also generalize their theory from smoothed hinge loss SVM to the general problems of minimizing convex partially separable functions of type:

$$f(\boldsymbol{\alpha}) = h(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i), \tag{1.1}$$

where A is data matrix, h - smooth function and each g_i - general convex function. This problem class includes not only smoothed hinge loss SVM, but also Lasso, Ridge Regression, (the dual formulation of the) hinge loss SVM, Logistic Regression, etc.. Our theoretical results for adaptive sampling include the existing results for fixed non-uniform [32] and uniform [5] sampling as special cases.

Contributions. The main contributions of this work can be summarized as follows:

- We provide a novel theoretical analysis of convergence rate of Stochastic Coordinate Descent for partially separable problems with strongly convex and general convex regularizers.
- To the best of our knowledge, we are the first to provide a convergence rate analysis of Stochastic Coordinate Descent with adaptive sampling schemes for problems with general convex regularizer, e.g. Lasso.
- We introduce new adaptive and fixed non-uniform sampling schemes for Stochastic Coordinate Descent for problems with general convex regularizers.
- We derive the duality gap and suboptimality convergence guarantees for arbitrary sampling distributions for both strongly convex and general convex cases and theoretically prove that the new schemes have a faster convergence than the conventional ones.
- We support the developed theory with numerical experiments on Lasso and smoothed hinge loss SVM and show in practice that the new adaptive schemes significantly outperform the non-adaptive ones.

The thesis is structured as follows. In Chapter 2, we do the general overview of existing literature on Stochastic Coordinate Descent. Chapter 3 presents basic theoretical preliminaries on which the thesis is based. In Chapter 4 we propose a new convergence lemma and derive optimal adaptive and non-adaptive schemes for strongly convex and general convex problems.

Chapter 5 discusses the application of the developed theory on two particular problems and compares the computational complexity of proposed sampling methods. In Chapter 6, we evaluate the performance of SCD with the derived in preceding chapters sampling schemes. Chapter 7 summarizes the main findings of the thesis and concludes it.

Chapter 2

Related work

In this chapter, we review the existing works on coordinate descent for two typical machine learning problems - SVM and Lasso.

Coordinate descent/ascent methods for optimization were studied from the foundation of discipline. They first appeared in the work of Hildreth (1957) [11] as a method to solve unconstrained quadratic programming problems.

The first paper to discuss coordinate descent method for Support Vector Machine (SVM) problem (Friess et al. (1998) [8]) was focused on solving SVM dual formulation. The drawback of their algorithm was its inability take advantage of the data sparsity, which made it not suitable for large sparse data problems. This problem was solved in Chang et al. (2008) [3], where they were the first to propose coordinate descent methods for solving a primal SVM. While their approach was experimentally quicker to find an optimal solution than the other methods at that time, the non-differentiability of primal hinge loss SVM restricted their approach to only squared hinge loss SVM problems.

The major breakthrough was made in 2008 by Hsieh et al. [12], who introduced Stochastic Dual Coordinate Ascent (SDCA) method for solving linear SVM with hinge and squared hinge loss functions. Experiments showed that in some regimes SDCA is significantly faster than the other solvers, such as the primal method in [3], Stochastic Gradient Descent (SGD / pegasos) [23], and SVM$^{\text{perf}}$ [13]. The theoretical analysis of SDCA was made in Shalev-Shwartz et al. (2013) [28]. In that work the primal-dual theoretical convergence guarantees for SDCA were derived, and it was shown that its convergence is comparable or better than the one of Stochastic Gradient De-

scent.

The convergence rate of stochastic coordinate descent methods naturally depends on a sampling probability distribution over the datapoints. While the majority of existing stochastic coordinate methods sample one active coordinate uniformly at random [12, 24, 15, 26], it was shown in Zhao & Zhang (2014) [32] that an appropriately defined non-uniform fixed sampling distribution, which they called *importance sampling*, can significantly improve the convergence rate.

In the work of Csiba et al. (2015) [4] the non-uniform sampling strategy was further developed into an SDCA variant with adaptive sampling, called the AdaSDCA algorithm. AdaSDCA updates the sampling distribution at the beginning of each iteration based on measures of progress of dual variables to the optimum. It was theoretically proven that in some cases, AdaSDCA achieves a faster convergence than uniform and *importance sampling*. An another approach for construction of adaptive sampling distribution was proposed in [18], in this work the block coordinate Frank-Wolfe algorithm was enhanced with sampling proportional to values of block-wise duality gaps. An adaptive variant of SGD was studied by Papa et al., (2015) [19], where they proposed an adaptive sampling scheme dependent on the past iterations in a Markovian manner. Other adaptive methods are heuristics without proven convergence guarantees, they include ACF [10] and ACiD [16].

The first use of coordinate descent for solving Lasso problem[1] was made in Fu (1998) [9], where he introduced a "shooting" algorithm. However the work of Fu was largely ignored by machine learning community and the approach gained interest only after paper of Friedman et al. (2007) [6], where they developed a path-wise coordinate descent method, which showed competitive performance with current at that time state-of-the-art LARS and homotopy methods. The latest state-of-the-art methods for Lasso are coordinate methods on the primal formulation such as GLMNET [7] and its extensions [24, 30].

In this thesis we derive primal-dual guarantees for coordinate descent on problems of minimizing convex partially separable functions of type (1.1) with general convex g_i. While the primal-dual guarantees for coordinate descent on dual SVM with $L2$ regularization (or strongly convex g_i) were extensively studied in [29, 31, 32, 4], the $L1$ regularization case (or general convex g_i) was an open problem until the paper of Dünner et al., (2016) [5],

[1] also known as L_1-regularized least squares linear regression problem

where the new primal-dual framework was introduced and the new primal-dual convergence guarantees for coordinate descent with uniform sampling were obtained for a wider class of problems with general convex regularizer (general convex g_i). This class of problems includes Lasso, Ridge Regression, hinge loss SVM, etc..

In this work we adopt the primal-dual framework of Dünner et al., (2016) [5] to derive primal-dual guarantees for fixed non-uniform and adaptive sampling of coordinates for both strongly convex and general convex g_i. While in strongly convex scenario our work reproves the results of [4], in general convex case we introduce a new type of adaptive algorithms which show state-of-the-art practical and theoretical performance.

Chapter 3

Preliminaries

In this chapter, we give basic mathematical concepts, which are the building blocks of the theory we introduce in the following chapters.

3.1 Basic definitions of convex optimization

Definition 3.1 *A set $X \subseteq \mathbb{R}^n$ is called convex if and only if:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in X \quad \forall 0 \leq s \leq 1 : \quad s\boldsymbol{x} + (1-s)\boldsymbol{y} \in X$$

Definition 3.2 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its domain D is a convex set and:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in D \quad \forall 0 \leq s \leq 1 : \quad f(s\boldsymbol{x} + (1-s)\boldsymbol{y}) \leq sf(\boldsymbol{x}) + (1-s)f(\boldsymbol{y})$$

Definition 3.3 *A function f: $\mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz continuous if $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we have*

$$|f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq L\|\boldsymbol{a} - \boldsymbol{b}\|.$$

Definition 3.4 *A function f: $\mathbb{R}^n \to \mathbb{R}$ is called L-smooth, for some $L > 0$, if it is differentiable and its derivative is L-Lipschitz continuous, or equivalently*

$$f(\boldsymbol{u}) \leq f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle + \frac{L}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2 \quad \forall \boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^n.$$

Definition 3.5 *A function f: $\mathbb{R}^n \to \mathbb{R}$ is called μ-strongly convex, for $\mu \geq 0$, if*

$$f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2 \quad \forall \boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^n.$$

Definition 3.6 *A function f: $\mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ has B-bounded support if its effective domain is bounded by B:*

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

3.2 Convex conjugates

Definition 3.7 *The convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as*

$$f^*(v) := \sup_{u \in \mathbb{R}^n} v^\top u - f(u).$$

Lemma 3.8 (Duality between Lipschitzness and L-Bounded Support, [5]) *Given a proper convex function g, it holds that g has L-bounded support if and only if g^* is L-Lipschitz.*

Lemma 3.9 (Duality between Smoothness and Strong Convexity, [14]) *Given a closed convex function f, it holds that f is μ-strongly convex if and only if f^* is $(1/\mu)$-smooth.*

Theorem 3.10 (Fenchel-Young Inequality) *Let $f^*(w) = \sup_{x \in \mathbb{R}^n}\{\langle w, x\rangle - f(x)\}$. Then the following inequality holds:*

$$f(x) + f^*(w) \geq \langle w, x\rangle$$

Equality holds if w is a subgradient of f at x, $w \in \partial f(x)$.

3.3 Primal-dual setting

In this thesis we adopt the primal-dual structure of Dünner et al. (2016) [5]. We consider the following pair of optimization problems which are dual to each other:

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^n} \Big[D(\alpha) := f(A\alpha) + \sum_i g_i(\alpha_i) \Big], \\ &\min_{w \in \mathbb{R}^d} \Big[P(w) := f^*(w) + \sum_i g_i^*(-a_i^\top w) \Big], \end{aligned} \tag{3.1}$$

here $A = [a_1, \ldots, a_n]$. The new structure has two advantages over the standard primal-dual SDCA setting. Firstly, as will be shown in the next subsections, the new framework generalizes the duality gap in SDCA and makes it act as a certificate for achieved solution accuracy. Secondly, the two problems in (3.1) are symmetrical to each other relatively to functions f and g, which enables the new structure to include much more machine learning methods than the traditional primal-dual setting of SDCA.

3.3.1 Optimality conditions

The first-order optimality conditions for problems (3.1) are given by

$$\begin{aligned} &w \in \partial f(A\alpha), && -a_i^\top w \in \partial g_i(\alpha_i) \text{ for all } i \in [n] \\ &A\alpha \in \partial f^*(w), && \alpha_i \in \partial g_i^*(-a_i^\top w) \text{ for all } i \in [n] \end{aligned} \tag{3.2}$$

For the proof see [2].

3.3.2 Duality gap

The duality gap is a difference between primal and dual solutions. From the definition of convex conjugate and the primal-dual setting (3.1), we have:

$$P(\boldsymbol{w}) \geq P(\boldsymbol{w}^*) \geq -D(\boldsymbol{\alpha}^*) \geq -D(\boldsymbol{\alpha}), \tag{3.3}$$

where $(\boldsymbol{w}^*, \boldsymbol{\alpha}^*)$ are optimal solutions of (3.1). The property (3.3) leads to the definition of the duality gap:

$$G(\boldsymbol{\alpha}, \boldsymbol{w}) := P(\boldsymbol{w}) - (-D(\boldsymbol{\alpha})). \tag{3.4}$$

The duality gap is a non-negative function and under strong duality it reaches zero value only in an optimal pair $(\boldsymbol{w}^*, \boldsymbol{\alpha}^*)$. When f is differentiable the conditions (3.2) imply:

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

The duality gap is a certificate on approximation accuracy of variable $\boldsymbol{\alpha}$:

$$G(\boldsymbol{\alpha}, \boldsymbol{w}) := P(\boldsymbol{w}(\boldsymbol{\alpha})) - (-D(\boldsymbol{\alpha})) \geq P(\boldsymbol{w}(\boldsymbol{\alpha})) - P(\boldsymbol{w}^*).$$

Coordinate-wise duality gaps

Below we will show that for our problem structure, the duality gap can be decomposed to the sum of non-negative coordinate-wise gaps. Here we assume that $\boldsymbol{w}$ is chosen according to the first order optimality, $\boldsymbol{w} = \nabla f(A\boldsymbol{\alpha})$. In the primal-dual setting (3.1), the duality gap is defined as

$$\begin{aligned} G(\boldsymbol{\alpha}) :&= P(\boldsymbol{w}(\boldsymbol{\alpha})) - (-D(\boldsymbol{\alpha})) = P(\boldsymbol{w}(\boldsymbol{\alpha})) + D(\boldsymbol{\alpha}) \\ &= f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i) + f^*(\boldsymbol{w}(\boldsymbol{\alpha})) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}(\boldsymbol{\alpha})). \end{aligned} \tag{3.5}$$

Since $\boldsymbol{w} = \nabla f(A\boldsymbol{\alpha})$, the Fenchel-Young inequality (Theorem 3.10) holds with equality:

$$f^*(\boldsymbol{w}(\boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) = (A\boldsymbol{\alpha})^\top \boldsymbol{w}(\boldsymbol{\alpha}) = \sum_i \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w}.$$

As a result the duality gap can be decomposed as a sum of the following coordinate-wise gaps:

$$G(\boldsymbol{\alpha}) = \sum_i G_i(\alpha_i) = \sum_i \left(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w} \right) \tag{3.6}$$

Each coordinate-wise gap $G_i(\alpha_i)$ is non-negative due to Fenchel-Young inequality (Theorem 3.10).

3.3.3 Dual residuals

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

Definition 3.11 (Dual Residual. A generalization of [4]) *Consider the primal-dual setting (3.1). 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:*

$$\kappa_i^{(t)} = u_i^{(t)} - \alpha_i^{(t)},$$

where $u_i^{(t)} \in \partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)})$.

Remark 3.12 *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 3.13 (Coherent probability vector, [4]) *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 $\kappa_i^{(t)} \neq 0 \quad \rightarrow \quad p_i^{(t)} > 0$.*

Definition 3.14 (t-support set) *We call set I_t:*

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

a t-support set.

Lemma 3.15 *Suppose that for all i, g_i^* is L-Lipschitz. Then, $\forall i : |\kappa_i| \leq 2L$.*

Proof According to Lemma 3.8, the L-Lipschitzness g_i^* implies L-bounded support of $g_i(\alpha_i)$ and therefore $|\alpha_i| \leq L$. Moreover, by the equivalence of Lipschitzness and bounded subgradient ([22], Lemma 2.6) we have $|u_i| \leq L$ and thus $|\kappa_i| = |\alpha_i - u_i| \leq |\alpha_i| + |u_i| \leq 2L$. □

3.4 Stochastic Coordinate Descent algorithm

A formal definition of Stochastic Coordinate Descent (SCD) in the primal-dual setting (3.1) is given in Algorithm 1 below. As can be seen, the algorithm has 3 major steps. Firstly it randomly chooses the coordinate, secondly it is doing line-search on $D(\boldsymbol{\alpha})$ along the chosen coordinate, thirdly it updates primal and dual variables. Note that all the steps of the algorithm are fixed, except for the way of choosing the coordinate on step one. The standard SCD chooses coordinates uniformly at random, in this thesis we propose a different approach.

Algorithm 1 Stochastic Coordinate Descent

1: let $\boldsymbol{\alpha}^{(0)} = 0 \in \mathbb{R}^n, \boldsymbol{w}^{(0)} = \boldsymbol{w}(\boldsymbol{\alpha}^{(0)})$
2: **for** $t = 0, 1, ...T$ **do**
3: Sample $i \in [n]$ randomly
4: Find $\Delta\alpha_i$ minimizing $D(\boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i \Delta\alpha_i)$
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**

Chapter 4

Theoretical analysis of Stochastic Coordinate Descent

In this chapter we introduce the necessary tools for the development of good sampling distributions for the Stochastic Coordinate Descent (SCD) algorithm. We start from generalizing ([4], Lemma 3) to apply to the setting of general Lipschitz functions and allowing them to have non-uniform strong-convexity constants. Further we use the generalized lemma to derive the convergence guarantees for strongly convex and general convex scenarios.

We consider the pair of dual to each other empirical loss minimization problems. Our goal is to find a ε_P-suboptimal parameter $\boldsymbol{w}$ or ε_D-suboptimal parameter $\boldsymbol{\alpha}$, i.e., $P(\boldsymbol{w}) - P(\boldsymbol{w}^*) \leq \varepsilon_P$ or $D(\boldsymbol{\alpha}) - D(\boldsymbol{\alpha}^*) \leq \varepsilon_D$, for the following pair of dual optimization problems:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \Big[D(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i) \Big],$$

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \Big[P(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) \Big].$$

For example, if we are given examples $(A, \boldsymbol{y})$ or equivalently $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{a}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, we obtain the Support Vector Machine (SVM) problem by setting $g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) = \max(0, 1 - y_i \boldsymbol{a}_i^\top \boldsymbol{w})$, and $f^*(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$. We obtain the Lasso problem by setting $f(A\boldsymbol{\alpha}) = \|\boldsymbol{y} - A\boldsymbol{\alpha}\|_2^2$ and $g_i(\alpha_i) = \lambda|\alpha_i|$.

4.1 Core lemma

In this section we develop a new lemma which states the relationship between any sampling distribution $\boldsymbol{p}$ and the convergence rate of SCD. This lemma is a natural extension of ([4], Lemma 3) with relaxation of constraints

on g_i's. While ([4], Lemma 3) is proven only for strongly convex g_i with one for all strong convexity constant μ, we derive the same convergence result for the general convex g_i with with coordinate-dependent strong convexity constants μ_i. The new lemma allows us to derive the convergence rates of SCD with adaptive sampling for common machine learning problems with non-strongly convex g_i, e.g. SVM and Lasso.

Lemma 4.1 *Consider Stochastic Coordinate Descent. Let f be $1/\beta$-smooth and 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 for any iteration t, any sampling distribution $\boldsymbol{p}^t$ and any arbitrary $s_i \in [0,1]$ $\forall i \in [n]$ it holds that*

$$\begin{aligned}\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) \\ - \sum_i p_i^{(t)} \Big(\frac{\mu_i(s_i - s_i^2)}{2} - \frac{s_i^2 \|\boldsymbol{a}_i\|^2}{2\beta} \Big) |\kappa_i^{(t)}|^2,\end{aligned} \tag{4.1}$$

here $\kappa_i^{(t)}$ is i-th dual residual (see Def. 3.11).

Proof Since in SCD update ($\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i \Delta\alpha_i$, see Algorithm 1) only one coordinate per iteration is changed, the one iteration improvement in dual objective can be written as:

$$D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) = \underbrace{\Big[g_i(\alpha_i^{(t)}) + f(A\boldsymbol{\alpha}^{(t)}) \Big]}_{(A)} - \underbrace{\Big[g_i(\alpha_i^{(t+1)}) + f(A\boldsymbol{\alpha}^{(t+1)}) \Big]}_{(B)}.$$

To bound part (B) we use a suboptimal update $\Delta\alpha_i = s_i \kappa_i^{(t)}$, for all $s_i \in [0,1]$:

$$\begin{aligned}(B) &= g_i(\alpha_i^{(t+1)}) + f(A\boldsymbol{\alpha}^{(t+1)}) \\ &\leq \min_{\Delta\alpha_i} \Big[g_i(\alpha_i^{(t)} + \Delta\alpha_i) + f(A\boldsymbol{\alpha}^{(t)} + \boldsymbol{a}_i \Delta\alpha_i) \Big] \\ &\leq g_i(\alpha_i^{(t)} + s_i\kappa_i^{(t)}) + f(A\boldsymbol{\alpha}^{(t)} + \boldsymbol{a}_i s_i \kappa_i^{(t)}).\end{aligned}$$

Each of g_i is μ_i-strongly convex, therefore:

$$\begin{aligned}g_i(\alpha_i^{(t)} + s_i\kappa_i^{(t)}) &= g_i(\alpha_i^{(t)} + s_i(u_i^{(t)} - \alpha_i^{(t)})) \\ &= g_i(s_i(u_i^{(t)}) + (1 - s_i)(\alpha_i^{(t)})) \\ &\leq s_i g_i(u_i^{(t)}) + (1 - s_i) g_i(\alpha_i^{(t)}) - \frac{\mu_i}{2} s_i (1 - s_i)(\kappa_i^{(t)})^2.\end{aligned}$$

The function f is $\frac{1}{\beta}$-smooth:

$$f(A\boldsymbol{\alpha}^{(t)} + \boldsymbol{a}_i s_i \kappa_i^{(t)}) \leq f(A\boldsymbol{\alpha}^{(t)}) + \nabla f(A\boldsymbol{\alpha}^{(t)})^\top (s_i \kappa_i^{(t)} \boldsymbol{a}_i) + \frac{1}{2\beta} \|s_i \kappa_i^{(t)} \boldsymbol{a}_i\|^2.$$

As a result:

$$(B) \leq s_i g_i(u_i^{(t)}) - s_i g_i(\alpha_i^{(t)}) - \frac{\mu_i}{2} s_i (1 - s_i)(\kappa_i^{(t)})^2 + \underbrace{g_i(\alpha_i^{(t)}) + f(A\boldsymbol{\alpha}^{(t)})}_{(A)} + \nabla f(A\boldsymbol{\alpha}^{(t)})^\top (s_i \kappa_i^{(t)} \boldsymbol{a}_i) + \frac{1}{2\beta} \|s_i \kappa_i^{(t)} \boldsymbol{a}_i\|^2.$$

With obtained results above and optimality condition (3.2) $\boldsymbol{w}(\boldsymbol{\alpha}) = \nabla f(A\boldsymbol{\alpha})$, the improvement in dual objective can be written as:

$$\begin{aligned} D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) &= (A) - (B) \\ &\geq -s_i g_i(u_i^{(t)}) + s_i g_i(\alpha_i^{(t)}) + \frac{\mu_i}{2} s_i (1 - s_i)(\kappa_i^{(t)})^2 \\ &\quad - \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(s_i u_i^{(t)} \boldsymbol{a}_i) + \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(s_i \alpha_i^{(t)} \boldsymbol{a}_i) - \frac{1}{2\beta} \|s_i \kappa_i^{(t)} \boldsymbol{a}_i\|^2 \\ &= s_i \Big(- g_i(u_i^{(t)}) + g_i(\alpha_i^{(t)}) - \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(u_i^{t-1} \boldsymbol{a}_i) \\ &\quad + \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(\alpha_i^{t-1} \boldsymbol{a}_i) + \frac{\mu_i}{2} (1 - s_i)(\kappa_i^{(t)})^2 - \frac{s_i}{2\beta} \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2 \Big). \end{aligned}$$

Since $u_i^{(t)} \in \partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}(\boldsymbol{\alpha}^{(t)}))$, the Fenchel-Young inequality (3.10) becomes equality for $g_i(u_i^{(t)})$:

$$g_i(u_i^{(t)}) + g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) = -\boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(u_i^{(t)} \boldsymbol{a}_i)$$

Using this fact, the bound on the improvement in dual objective becomes:

$$\begin{aligned} D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) \geq s_i \Big(& g_i(\alpha_i^{(t)}) + g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) + \boldsymbol{w}(\boldsymbol{\alpha}^{(t)})(\alpha_i^{(t)} \boldsymbol{a}_i) \\ & + \frac{\mu_i}{2} (1 - s_i)(\kappa_i^{(t)})^2 - \frac{s_i}{2\beta} \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2 \Big) \end{aligned}$$

Therefore for any $s_i \in [0, 1]$ it holds that:

$$D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) \geq s_i \Big[G_i(\boldsymbol{\alpha}^{(t)}) + \frac{\mu_i}{2} (1 - s_i) |\kappa_i^{(t)}|^2 - \frac{s_i}{2\beta} \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2 \Big], \quad (4.2)$$

where G_i is i-th coordinate-wise duality gap:

$$G(\boldsymbol{\alpha}^{(t)}) = \sum_i G_i(\boldsymbol{\alpha}^{(t)}), \qquad G_i(\boldsymbol{\alpha}^{(t)}) = g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i^{(t)}) + \alpha_i^{(t)} \boldsymbol{a}_i^\top \boldsymbol{w}.$$

By taking an expectation of the both sides with respect to i, conditioned on $\boldsymbol{\alpha}^{(t)}$, we obtain:

$$\begin{aligned} \mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)}) | \boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) & - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) \\ & - \sum_i p_i^{(t)} \Big(\frac{\mu_i (s_i - s_i^2)}{2} - \frac{s_i^2 \|\boldsymbol{a}_i\|^2}{2\beta} \Big) |\kappa_i^{(t)}|^2 \end{aligned}$$

and thus finalize the proof. □

Remark 4.2 *If in addition to the conditions of Lemma 4.1 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:*

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)}, \tag{4.3}$$

where

$$F^{(t)} = \frac{1}{n^2 \beta \theta} \sum_{i \in I_t} \Big(\frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^{(t)}} - \mu_i \beta \Big) |\kappa_i^{(t)}|^2. \tag{4.4}$$

Proof Since s_i in Lemma 4.1 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 $\theta \in [0, \min_i p_i^{(t)}]$. The inequality (4.1) becomes:

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

and thus finalizes the proof. □

4.2 Strongly convex objectives

In this section we focus on finding optimal sampling distributions for problems of the form (3.1) with strongly convex functions g_i. For instance, the dual of the smoothed hinge loss SVM fits into this assumptions. For simplicity reasons we drop the conditioning on $\boldsymbol{\alpha}^{(t)}$ in each of the expectations in this section, e.g. we use $\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)})]$ when we mean $\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) \mid \boldsymbol{\alpha}^{(t)}]$.
To achieve a direct bound on the duality gap:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)})] \geq \theta G(\boldsymbol{\alpha}^{(t)}) \tag{4.5}$$

in (4.3) when $\mu_i > 0$ we need $\left[\frac{\theta^2 n^2}{2} F^{(t)} := \frac{\theta}{2\beta} \sum_{i \in I_t} \Big(\mu_i \beta - \frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^{(t)}} \Big) |\kappa_i^{(t)}|^2 \right]$ to be non-positive. This can be achieved in two different ways:

- the first is to find θ and $\boldsymbol{p}$ which for all i satisfy:

$$\Big(\mu_i \beta - \frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^t} \Big) \geq 0, \tag{4.6}$$

  in this case $\boldsymbol{p}$ and θ are constant during the whole optimization process. We call *importance sampling scheme* the sampling scheme based on $\boldsymbol{p}$ that maximizes θ in this case.

- the second is to find $\theta^{(t)}$ and $\boldsymbol{p}^{(t)}$ which satisfy:

$$\left(\frac{\theta}{2\beta}\sum_i\left(\mu_i\beta-\frac{\theta(\mu_i\beta+\|\boldsymbol{a}_i\|^2)}{p_i^{(t)}}\right)|\kappa_i^{(t)}|^2\right)\geq 0, \tag{4.7}$$

  in this case $\boldsymbol{p}^{(t)}$ and $\theta^{(t)}$ are changing every iteration. We call *adaptive importance sampling scheme* the sampling scheme based on $\boldsymbol{p}^{(t)}$ that maximizes $\theta^{(t)}$ in this case.

Remark 4.3 *Note that the parameter θ in (4.5) is directly related to the convergence rate of SCD. Since it multiplicatively measures the improvement per step, our goal is to find distributions which maximize θ.*

Remark 4.4 (first way - importance sampling case) *The condition (4.6) holds when:*

$$\forall i\in[n]:\ \ \frac{\theta}{p_i}\in\left[0,\frac{\mu_i\beta}{\mu_i\beta+\|\boldsymbol{a}_i\|^2}\right].$$

To find a probability vector $\boldsymbol{p}$ which maximizes the convergence rate we need to solve:

$$\begin{gathered}\max\theta\\ s.t.\ \forall i\ \ \frac{\theta}{p_i}\in\left[0,\frac{\mu_i\beta}{\mu_i\beta+\|\boldsymbol{a}_i\|^2}\right],\\ \sum_i p_i=1.\end{gathered}$$

The solution is:

$$\theta_{imp}:=\frac{1}{\sum_j\frac{\|\boldsymbol{a}_j\|^2+\mu_j\beta}{\mu_j\beta}},\qquad p_i:=\frac{\frac{\|\boldsymbol{a}_i\|^2+\mu_i\beta}{\mu_i}}{\sum_j\frac{\|\boldsymbol{a}_j\|^2+\mu_j\beta}{\mu_j}}. \tag{4.8}$$

Remark 4.5 (second way - adaptive importance sampling case) *This case was the central result in [4], here we cite its derivation for completeness and comparison reasons. The condition (4.7) holds for:*

$$\theta^{(t)}\leq\frac{\sum_i\mu_i\beta|\kappa_i^{(t)}|^2}{\sum_i(\mu_i\beta+\|\boldsymbol{a}_i\|^2)|\kappa_i^{(t)}|^2(p_i^{(t)})^{-1}}.$$

The probability vector $\boldsymbol{p}^{(t)}$ should maximize $\theta^{(t)}$. Using the same approach as in [4] we compute the optimal parameter $\theta_{ada}^{(t)}$ and the distribution:

$$p_j^{(t)}:=\frac{|\kappa_j^{(t)}|\sqrt{\mu_j\beta+\|\boldsymbol{a}_j\|^2}}{\sum_i|\kappa_i^{(t)}|\sqrt{\mu_i\beta+\|\boldsymbol{a}_i\|^2}}, \tag{4.9}$$

$$\theta_{ada}^{(t)}:=\frac{\sum_i\mu_i\beta|\kappa_i^{(t)}|^2}{\left(\sum_i|\kappa_i^{(t)}|\sqrt{\mu_i\beta+\|\boldsymbol{a}_i\|^2}\right)^2}. \tag{4.10}$$

Remark 4.6 (default case - uniform sampling) *The case of uniform sampling is a special case of Remark 4.4 with fixed probability distribution $\boldsymbol{p}$, $p_i = 1/n$. The only unknown parameter left to find is θ. We need to find maximal θ which satisfies (4.6) $\forall i \in [1, n]$:*

$$\Big(\mu_i\beta - \theta n(\mu_i\beta + \|\boldsymbol{a}_i\|^2)\Big) \geq 0,$$

according to the inequality above θ should satisfy:

$$\forall i \;\; \theta \leq \frac{\mu_i\beta}{n(\mu_i\beta + \|\boldsymbol{a}_i\|^2)}.$$

As a result, the maximal possible θ for uniform sampling case is:

$$\theta_{unif} = \min_{i\in[1,n]} \left(\frac{\mu_i\beta}{n(\mu_i\beta + \|\boldsymbol{a}_i\|^2)}\right). \tag{4.11}$$

The following corollary will show that the adaptive importance sampling is never worse then the importance sampling and that uniform is never better than the non-uniform sampling schemes.

Corollary 4.7 *Consider Stochastic Coordinate Descent. For any iteration $t \geq 0$ it holds that:*

$$\theta_{unif} \leq \theta_{imp} \leq \theta_{ada}^{(t)}.$$

Proof This fact directly follows from the settings of optimization problems we solve in Remarks 4.4, 4.5 and 4.6. □

Using these probability distributions and inequality (4.3) we can find convergence rates for SCD with importance and adaptive importance sampling schemes:

$$\begin{aligned}\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)})] &\geq \mathbb{E}\left[\theta^{(t)}\left(P(\boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) + D(\boldsymbol{\alpha}^{(t)})\right)\right] \\ &= \hat{\theta}^{(t)}\mathbb{E}[P(\boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) + D(\boldsymbol{\alpha}^{(t)})] \\ &\geq \hat{\theta}^{(t)}\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{*})],\end{aligned} \tag{4.12}$$

where $\hat{\theta}^{(t)} := \frac{\mathbb{E}\left[\theta^{(t)}\left(P(\boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) + D(\boldsymbol{\alpha}^{(t)})\right)\right]}{\mathbb{E}[P(\boldsymbol{w}(\boldsymbol{\alpha}^{(t)})) + D(\boldsymbol{\alpha}^{(t)})]}$.
By plugging in into (4.12) suboptimality $\varepsilon_D^{(t)} := D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{*})$ and $D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) = \varepsilon_D^{(t)} - \varepsilon_D^{(t+1)}$ we obtain:

$$\mathbb{E}[\varepsilon_D^{(t)} - \varepsilon_D^{(t+1)}] \geq \hat{\theta}^{(t)}\mathbb{E}[\varepsilon_D^{(t)}]$$

and finally

$$\mathbb{E}[\varepsilon_D^{(t+1)}] \leq (1 - \hat{\theta}^{(t)})\mathbb{E}[\varepsilon_D^{(t)}] \leq \varepsilon_D^{(0)}\prod_{k=0}^{t}(1 - \hat{\theta}^{(k)}). \tag{4.13}$$

In SCD with uniform and importance sampling schemes $\hat{\theta}^{(t)} = \theta$ is constant during the whole optimization process (see (4.8), (4.11)). Therefore for them we can easily find convergence rates and number of iterations required to reach any given suboptimality:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq (1-\theta)^t \varepsilon_D^{(0)} \leq \exp\big(-t\theta\big)\varepsilon_D^{(0)}. \tag{4.14}$$

The number of iterations T required to get a suboptimality $\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^*)] \leq \varepsilon_D$:

$$T \geq \frac{1}{\theta} \log\Big(\frac{\varepsilon_D^{(0)}}{\varepsilon_D}\Big). \tag{4.15}$$

By plugging in the corresponding θ in (4.14) and (4.15), we obtain:

- **Uniform sampling**

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \exp\left(-t \min_{i\in[1,n]}\left(\frac{\mu_i\beta}{n(\mu_i\beta + \|\boldsymbol{a}_i\|^2)}\right)\right)\varepsilon_D^{(0)}, \tag{4.16}$$

$$T_{\text{unif}} \geq \frac{\log\left(\frac{\varepsilon_D^{(0)}}{\varepsilon_D}\right)}{\min_{i\in[1,n]}\left(\frac{\mu_i\beta}{n(\mu_i\beta+\|\boldsymbol{a}_i\|^2)}\right)}. \tag{4.17}$$

- **Importance sampling**

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \exp\left(\frac{-t}{\sum_j \frac{\|\boldsymbol{a}_j\|^2+\mu_j\beta}{\mu_j\beta}}\right)\varepsilon_D^{(0)}, \tag{4.18}$$

$$T_{\text{imp}} \geq \sum_j \frac{\|\boldsymbol{a}_j\|^2 + \mu_j\beta}{\mu_j\beta} \log\Big(\frac{\varepsilon_D^{(0)}}{\varepsilon_D}\Big). \tag{4.19}$$

As can be easily seen $T_{\text{imp}} \leq T_{\text{unif}}$.

Adaptive importance sampling. Since in adaptive importance sampling scenario parameter $\hat{\theta}^{(t)}$ (4.10) changes with iteration, we cannot derive its convergence rate in the same way as we did for schemes with constant θ. However due to $\theta_{\text{imp}} \leq \theta_{\text{ada}}^{(t)}$ (see Corollary 4.7) for SCD with adaptive importance sampling it holds that:

$$\begin{aligned}
\mathbb{E}[\varepsilon_D^{(t)}] &\leq \varepsilon_D^{(0)} \prod_{k=0}^{t-1}(1-\theta_{\text{ada}}^{(k)}) \\
&\leq \varepsilon_D^{(0)} \prod_{k=0}^{t-1}(1-\theta_{\text{ada}}^{(k)}) \\
&\leq \varepsilon_D^{(0)}(1-\theta_{\text{imp}})^t.
\end{aligned}$$

Above we proved that SCD with adaptive importance sampling has a faster convergence rate than SCD with the importance sampling, therefore $T_{\text{ada}} \leq T_{\text{imp}} \leq T_{\text{unif}}$.
To conclude, in this section we analyzed two sampling schemes for SCD - importance and adaptive importance. New schemes were compared with each other and with standard uniform sampling. The importance sampling provides a better rate than uniform and adaptive importance outperforms the importance sampling.

4.3 General convex objectives

In this section we propose several ways to improve the sampling scheme of coordinate descent algorithm in the general convex case, i.e. when all strong convexity parameters are zero ($\mu_i = 0$). In this case the statement of Remark 4.2 becomes:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)}, \tag{4.20}$$

where

$$F^{(t)} := \frac{1}{n^2 \beta} \sum_{i \in I_t} \Big(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \Big). \tag{4.21}$$

As could be seen, contrary to the strongly convex case, in the general convex case $F^{(t)}$ is always positive and therefore the sampling distributions derived in Section 4.2 are not optimal for general convex problems with general convex g_i, e.g. SVM and Lasso. In the following theorem we generalize ([32], Theorem 5) and ([5], Theorem 9) where only cases of sampling with particular fixed distributions are considered and find the convergence rate for an arbitrary chosen adaptive sampling distribution.

Theorem 4.8 *Consider Stochastic Coordinate Descent. Assume f is $\frac{1}{\beta}$-smooth function. Then, if g_i^* is L_i-Lipschitz for each i and $\boldsymbol{p}^{(t)}$ is coherent with $\boldsymbol{\kappa}^{(t)}$, it suffices to have a total number of iterations of*

$$T \geq \max \left\{ 0, \frac{1}{p_{\min}} \log \Big(\frac{2\varepsilon_D^0}{n^2 p_{\min} F^\circ} \Big) \right\} + \frac{5F^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}} \tag{4.22}$$

or alternatively

$$T \geq \frac{5F'^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}} \tag{4.23}$$

to obtain a duality gap $G(\bar{\boldsymbol{\alpha}}) \leq \varepsilon$. Moreover, when $t \geq T_0$ with

$$T_0 = \max \left\{ 0, \frac{1}{p_{\min}} \log \Big(\frac{2\varepsilon_D^0}{n^2 p_{\min} F^\circ} \Big) \right\} + \frac{4F^\circ n^2}{\varepsilon} - \frac{2}{p_{\min}} \tag{4.24}$$

or alternatively

$$T_0 = \frac{4F'^{\circ}n^2}{\varepsilon} - \frac{2}{p_{\min}} \tag{4.25}$$

we have the suboptimality bound of $\mathbb{E}[D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{\star})] \leq \varepsilon/2$*, where* ε_D^0 *is the initial dual suboptimality and* F° *is an upper bound on* $\mathbb{E}[F^{(t)}]$ *taken over all coordinates at* $1, \ldots, T$ *algorithm iterations and* $F'^{\circ} := F^{\circ} + \frac{\varepsilon_D^{(0)}}{n^2 p_{\min}}$.

Proof From Remark 4.2 we know:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)} \tag{4.26}$$

With $D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t+1)}) = \varepsilon_D^{(t)} - \varepsilon_D^{(t+1)}$ and $\varepsilon_D^{(t)} = D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{*}) \leq G(\boldsymbol{\alpha}^{(t)})$, this implies:

$$\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] \geq \varepsilon_D^{(t)} - \theta\varepsilon_D^{(t)} + \frac{\theta^2 n^2}{2} F^{(t)}$$

by taking unconditional expectation over all iterations and using definition of F° we obtain:

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t+1)}] &\leq (1-\theta)\mathbb{E}[\varepsilon_D^{(t)}] + \frac{\theta^2 n^2}{2}\mathbb{E}[F^{(t)}] \\ &\leq (1-\theta)\mathbb{E}[\varepsilon_D^{(t)}] + \frac{\theta^2 n^2}{2} F^{\circ}\end{aligned}$$

Now we will show using induction that we can bound the dual suboptimality as:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \frac{2F^{\circ}n^2}{\frac{2}{p_{\min}} + t - t_0}, \tag{4.27}$$

where $t \geq t_0 = \max\left\{0, \frac{1}{p_{\min}} \log\left(\frac{2\varepsilon_D^0}{n^2 p_{\min} F^{\circ}}\right)\right\}$. Indeed, let's choose $\theta = p_{\min}$, then the basis of induction at $t = t_0$ is verified as:

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t)}] &\leq (1-p_{\min})^t \varepsilon_D^{(0)} + \sum_{i=0}^{t-1} (1-p_{\min})^i p_{\min}^2 n^2 \frac{F^{\circ}}{2} \\ &\leq e^{-t p_{\min}} \varepsilon_D^{(0)} + n^2 p_{\min} \frac{F^{\circ}}{2} \\ &\leq n^2 p_{\min} F^{\circ}.\end{aligned}$$

Note that if in (4.27) instead of F° we take $F'^{\circ} := F^{\circ} + \frac{\varepsilon_D^{(0)}}{n^2 p_{\min}}$, the condition holds with $t_0 = 0$:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \frac{2F'^{\circ}n^2}{\frac{2}{p_{\min}} + t}.$$

Now let's prove the inductive step, for $t > t_0$. Suppose claim holds for $t-1$, then

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t)}] &\leq (1-\theta)\mathbb{E}[\varepsilon_D^{(t-1)}] + \theta^2 n^2 \frac{F^\circ}{2} \\ &\leq (1-\theta)\frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0} + \theta^2 n^2 \frac{F^\circ}{2},\end{aligned}$$

choosing $\theta = \frac{2}{\frac{2}{p_{\min}}+t-1-t_0} \leq p_{\min}$ yeilds:

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t)}] &\leq \left(1 - \frac{2}{\frac{2}{p_{\min}} + t - 1 - t_0}\right)\frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0} \\ &+ \left(\frac{2}{\frac{2}{p_{\min}} + t - 1 - t_0}\right)^2 \frac{F^\circ n^2}{2} \\ &= \left(1 - \frac{2}{\frac{2}{p_{\min}} + t - 1 - t_0}\right)\frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0} \\ &+ \left(\frac{1}{\frac{2}{p_{\min}} + t - 1 - t_0}\right)\frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0} \\ &= \left(1 - \frac{1}{\frac{2}{p_{\min}} + t - 1 - t_0}\right)\frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0} \\ &= \frac{2F^\circ n^2}{\frac{2}{p_{\min}} + (t-1) - t_0}\left(\frac{\frac{2}{p_{\min}} + t - 2 - t_0}{\frac{2}{p_{\min}} + t - 1 - t_0}\right) \\ &\leq \frac{2F^\circ n^2}{\frac{2}{p_{\min}} + t - t_0}.\end{aligned}$$

This proves the bound (4.27) on suboptimality. To bound the duality gap we sum the inequality (4.26) over the interval $t = T_0 + 1, ..., T$ and obtain

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(T_0)}) - D(\boldsymbol{\alpha}^{(T)})] \geq \theta\mathbb{E}\Big[\sum_{t=T_0+1}^{T} D(\boldsymbol{\alpha}^{(t-1)}) + P(\boldsymbol{w}^{(t-1)})\Big] - (T-T_0)\frac{\theta^2 n^2}{2}F^\circ,$$

by rearranging terms and choosing $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$ to be the average vectors over $t \in \{T_0, T-1\}$ we get:

$$\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] = \mathbb{E}[D(\bar{\boldsymbol{\alpha}}) + P(\bar{\boldsymbol{w}})] \leq \frac{\mathbb{E}[D(\boldsymbol{\alpha}^{(T_0)}) - D(\boldsymbol{\alpha}^{(T)})]}{\theta(T-T_0)} + \theta n^2 \frac{F^\circ}{2}.$$

If $T \geq \frac{1}{p_{\min}} + T_0$ and $T_0 \geq t_0$, we can set $\theta = 1/(T-T_0)$ and combining this

with (4.27) we get:

$$\begin{aligned}\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] &\leq \mathbb{E}[D(\boldsymbol{\alpha}^{(T_0)}) - D(\boldsymbol{\alpha}^{(T)})] + \frac{F^\circ n^2}{2(T-T_0)} \\ &\leq \mathbb{E}[D(\boldsymbol{\alpha}^{(T_0)}) - D(\boldsymbol{\alpha}^*)] + \frac{F^\circ n^2}{2(T-T_0)} \\ &\leq \frac{2F_T n^2}{\frac{2}{p_{\min}} + t - t_0} + \frac{F^\circ n^2}{2(T-T_0)}.\end{aligned}$$

A sufficient condition to bound the duality gap by ε is that $T_0 \geq t_0 - \frac{2}{p_{\min}} + \frac{4F^\circ n^2}{\varepsilon}$ and $T \geq T_0 + \frac{F^\circ n^2}{\varepsilon}$ which also implies $\mathbb{E}[D(\boldsymbol{\alpha}^{(T_0)}) - D(\boldsymbol{\alpha}^*)] \leq \varepsilon/2$. Since we also need $T_0 \geq t_0$ and $T - T_0 \geq \frac{1}{p_{\min}}$, the overall number of iterations should satisfy:

$$T_0 \geq \max\left\{t_0, \frac{4F_T n^2}{\varepsilon} - \frac{2}{p_{\min}} + t_0\right\} \quad \text{and} \quad T - T_0 \geq \max\left\{\frac{1}{p_{\min}}, \frac{F^\circ n^2}{\varepsilon}\right\}.$$

Using $a + b \geq \max(a, b)$ we finally can bound the total number of required iterations to reach a duality gap of ε by:

$$\begin{aligned}T &\geq T_0 + \frac{1}{p_{\min}} + \frac{F^\circ n^2}{\varepsilon} \\ &\geq t_0 + \frac{4F^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}} + \frac{F^\circ n^2}{\varepsilon}. \\ &= t_0 + \frac{5F^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}}\end{aligned}$$

This concludes the proof. □

Remark 4.9 *We recover ([5], Theorem 9) as a special case of Theorem 4.8 by setting* $p_i^{(t)} = \frac{1}{n}$. *We recover ([32], Theorem 5) by setting* $p_i^{(t)} = \frac{L_i}{\sum_j L_j}$ [1].

4.3.1 Maximizing convergence rate in case of infinitesimal ε

The inequalities (4.22) and (4.24) claim that the number of iterations is directly related to F°. Indeed, if we take a limit $\varepsilon \to 0$, then the only significant term in (4.22) is the one dependent on ε, i.e. $\frac{5F^\circ n^2}{\varepsilon}$. To achieve a better convergence speed we have to find a distribution $\boldsymbol{p}^{(t)}$ which minimizes the F°. Below we derive optimal adaptive and non-adaptive distributions to minimize $F^{(t)}$ and consequently F°.
By definition:

$$F^{(t)} := \frac{1}{n^2\beta}\sum_i \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^t}\right).$$

[1] this distribution is a Lipschitz-variant of our "importance sampling" (4.29) distribution

From Lemma 3.15 we know that if each g_i^* is L_i-Lipschitz, then $|\kappa_i| = |\alpha_i - u_i| \leq 2L$. Let's first find what $F^{(t)}$ value we achieve with uniform sampling $p_i^t = 1/n$:

$$F^{(t)} = \frac{1}{n^2\beta}\sum_i \Big(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2}{1/n}\Big) \leq \frac{4}{\beta}\sum_i \frac{L_i^2\|\boldsymbol{a}_i\|^2}{n}.$$

Now let us find an adaptive probability distribution which minimizes $F^{(t)}$, where $|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2 = c_i^2$:

$$\min_p \sum_i \frac{c_i^2}{p_i}$$
$$s.t. \quad \sum_i p_i = 1$$

We define the Lagrangian:

$$L(p,\eta) = \sum_i \frac{c_i^2}{p_i} - \eta(\sum_i p_i - 1) \to \min$$

The optimality conditions are:

$$\forall i, \forall j \in [n] \to \frac{c_i^2}{p_i^2} = \frac{c_j^2}{p_j^2},$$

$$\sum_i p_i = 1.$$

The solution is

$$p_i^{(t)} := \frac{c_i}{\sum_j c_j} = \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}. \tag{4.28}$$

Now let us put derived distribution into the initial expression for $F^{(t)}$ and find the target upper bound:

$$F^{(t)} = \frac{1}{n^2\beta}\sum_i \Big(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big) = \frac{1}{\beta}\sum_i \Big(\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{n}\Big)\sum_j \Big(\frac{|\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}{n}\Big)$$

$$= \frac{1}{\beta}\Big(\sum_i \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{n}\Big)^2 \leq \frac{4}{\beta}\Big(\sum_i \frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)^2.$$

To find an optimal probability distribution which does not depend on dual residuals, we solve the same opimization problem as in adaptive case, but instead of "true" $F^{(t)}$ we minimize its independent of dual-residuals upper bound $\bar{F} := \frac{4}{n^2\beta}\sum_i \Big(\frac{L_i^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big)$.

$$F^{(t)} = \frac{1}{n^2\beta}\sum_i \Big(\frac{|\kappa_i|^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big) \leq \frac{4}{n^2\beta}\sum_i \Big(\frac{L_i^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big) = \bar{F}.$$

The solution is to choose the sampling probabilities proportional to $L_i\|\boldsymbol{a}_i\|$:

$$p_i := \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|}, \tag{4.29}$$

we call this sampling distribution "importance sampling" analogously to the strongly convex case. We achieve the same upper bound on $F^{(t)}$ as with the adaptive sampling:

$$F^{(t)} = \frac{1}{n^2\beta}\sum_i\Big(\frac{|\kappa_i|^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big) = \frac{1}{\beta}\sum_i\Big(\frac{|\kappa_i|^2\|\boldsymbol{a}_i\|}{L_i n}\Big)\sum_j\Big(\frac{L_j\|\boldsymbol{a}_j\|}{n}\Big)$$
$$\leq \frac{4}{\beta}\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big) \leq \frac{4}{\beta}\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)^2.$$

As a result, analogously to the strongly convex case we found adaptive and fixed sampling distributions, which minimize the upper bound on the number of iterations required to get an infinitesimal suboptimality.
Summary:

- Uniform sampling case:

$$p_i := \frac{1}{n};\quad F^{\circ}_{\text{unif}} \geq \mathbb{E}\left[\frac{1}{\beta}\sum_i\Big(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2}{n}\Big)\right];$$
$$F^{(t)} \leq F_{\text{unif}} := \frac{4}{\beta}\sum_i\frac{L_i^2\|\boldsymbol{a}_i\|^2}{n}$$

- Importance sampling case:

$$p_i := \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|};\quad F^{\circ}_{\text{imp}} \geq \mathbb{E}\left[\frac{1}{\beta}\sum_i\Big(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|}{L_i n}\Big)\sum_j\Big(\frac{L_j\|\boldsymbol{a}_j\|}{n}\Big)\right];$$
$$F^{(t)} \leq F_{\text{imp}} := \frac{4}{\beta}\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)^2$$

- Adaptive importance sampling case:

$$p_i^{(t)} := \frac{c_i}{\sum_j c_j} = \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j|\kappa_j^{(t)}|\|\boldsymbol{a}_j\|};\quad F^{\circ}_{\text{ada}} \geq \mathbb{E}\left[\frac{1}{\beta}\Big(\sum_i\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{n}\Big)^2\right];$$
$$F^{(t)} \leq F_{\text{ada}} := \frac{4}{\beta}\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)^2$$

Remark 4.10 *The inequality* $\Big(\sum_i\frac{L_i\|\boldsymbol{a}_i\|}{n}\Big)^2 \leq \sum_i\frac{L_i^2\|\boldsymbol{a}_i\|^2}{n}$ *is a direct consequence of the Cauchy-Schwarz inequality, therefore the adaptive and importance sampling schemes always give a better lower bound on* $F^{(t)}$ *than uniform sampling.*

4.3.2 Comparison with uniform sampling

Note that Theorem 4.8 states that the number of iterations sufficient to get an ε-optimal duality gap (4.22) does not depend on sampling probabilities of points out of the set I_t at step t. Therefore sampling schemes which sample only from the t-support set have an advantage over uniform sampling in the convergence speed due to the increased $p_{\min}$ and lower bound on $F^{(t)}$.
Let's assume that the size of the t-support set never exceeds some $m \in [1, n]$ and compare two sampling methods:

- Uniform sampling: $p_i^{(t)} = \frac{1}{n}$.
  $p_{\min} = \frac{1}{n}$, $F^{(t)} = \frac{1}{n^2\beta} \sum_{i \in I_t} \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \right) = \frac{1}{n\beta} \sum_{i \in I_t} |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2 \leq F_{\text{unif}}$
  Number of iterations according to (4.22): $T \geq \max \left\{ 0, n \log \left(\frac{2\varepsilon_D^0}{nF_{\text{unif}}} \right) \right\} + \frac{5n^2 F_{\text{unif}}}{\varepsilon}$
- Support set uniform:

$$\begin{cases} p_i^{(t)} = \frac{1}{m}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases} \tag{4.30}$$

  $p_{\min} = \frac{1}{m}$, $F^{(t)} = \frac{1}{n^2\beta} \sum_{i \in I_t} \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \right) = \frac{m}{n^2\beta} \sum_{i \in I_t} |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2 \leq \frac{m}{n} F_{\text{unif}}$
  Number of iterations according to (4.22): $T \geq \max \left\{ 0, m \log \left(\frac{2\varepsilon_D^{(0)}}{nF_{\text{unif}}} \right) \right\} + \frac{5nmF_{\text{unif}}}{\varepsilon}$

If we define $T_p(\varepsilon) = \max \left\{ 0, \log \left(\frac{2\varepsilon_D^{(0)}}{nF_{\text{unif}}} \right) \right\} + \frac{5nF_{\text{unif}}}{\varepsilon}$, then with uniform sampling we reach desired duality gap in $nT_p(\varepsilon)$ iterations and with the support set sampling in $mT_p(\varepsilon)$ iterations. In practice we have observed $m \sim 0.1n$ for several realistic Lasso experiments, in that case, support set uniform sampling can provide a 10-fold improvement over uniform sampling.

4.3.3 Maximizing convergence speed in case of constant ε

While we already derived distributions for the theoretical case of infinitesimal ε in Section 4.3.1, in practice we run the method until we achieve some constant accuracy (e.g. $\varepsilon \sim 10^{-6}$). In this case in (4.22) the term containing $\frac{1}{p_{\min}} \log \left(\frac{2\varepsilon_D^0}{n^2 p_{\min} F^\circ} \right)$ is not negligible compared to the term $\frac{5F^\circ n^2}{\varepsilon}$ and we need to consider distributions which minimize the whole expression in (4.22). To see the importance of this, consider the case when the values of the dual

residuals κ_i have very high variance, in this case some points in adaptive distribution (4.28) have extreme values, e.g. $\frac{1}{p_{\min}} \sim F^\circ n^3$. In this case the inequality (4.22) becomes:

$$T \geq F^\circ n^3 \log(2n\varepsilon_D^0) + \frac{5F^\circ n^2}{\varepsilon},$$

which shows that for $n \sim \varepsilon^{-1}$ the independent of ε term has a non-negligible impact on T.
To find sampling distributions which minimize the whole bound on T we will use the bound (4.23):

$$T \geq \frac{5F^\circ n^2}{\varepsilon} + \frac{\varepsilon_D^{(0)}}{\varepsilon p_{\min}}. \tag{4.31}$$

Below we consider an approach to find an optimal distribution based on mixing adaptive distribution (4.28) with uniform (4.30).

Mixture of uniform and adaptive

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. As was shown above, the optimal distribution minimizing F° is given by (4.28) and the distribution minimizing $1/p_{\min}$ is given by (4.30). Below we are finding a mix of the two aforementioned distributions which minimize the T in (4.31). We define mixed distribution as:

$$\begin{cases} p_i^{(t)} = \frac{\sigma}{m} + (1-\sigma)\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases} \tag{4.32}$$

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

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

and bound on the number of iterations:

$$T \geq \frac{5F^\circ_{\text{ada}} n^2}{\varepsilon(1-\sigma)} + \frac{\varepsilon_D^{(0)} m}{\varepsilon\sigma}. \tag{4.33}$$

Since the process of finding F°_{ada} is rather problematic, a good σ can be found by replacing F°_{ada} with its upper bound F_{ada} and minimizing (4.33). Another option is to use a "safe" choice of $\sigma = 0.5$, which provides a balance between two distributions and guarantees decent convergence in case of unknown F°_{ada}. In the applications section below (Chapter 5) we use the latter option and call this sampling variant *ada-uniform sampling*.

4.3.4 Gap-wise sampling

In this subsection we develop the theory and convergence rates of SCD with gap-wise sampling, i.e. sampling each coordinate according to its duality gap. This chapter is motivated by the paper [18] where the gap-wise sampling distribution proved to be the optimal in the Structured Support Vector Machine (SSVM) problem. We start by deriving the convergence theorem and then we compare the obtained rate with ones developed in previous subsections.

Definition 4.11 (Nonuniformity measure, [18]) *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 probability vector obtained by normalizing $\boldsymbol{x}$.

Lemma 4.12 *Let $\boldsymbol{x} \in \mathbb{R}^n_+$. The following relation holds:*

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

Proof. It directly follows from:

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

and Def. 4.11.

Theorem 4.13 *Consider Stochastic Coordinate Descent. Assume f is $\frac{1}{\beta}$-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 on each iteration it holds that*

$$\mathbb{E}[\varepsilon_D^{(t+1)}] \leq \frac{C + 2n\varepsilon_D^{(0)}}{t + 2n + 1}, \tag{4.34}$$

where C is an upper bound on $\mathbb{E}\left[\frac{2n\chi(\overrightarrow{F})\sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{(\chi(\overrightarrow{G}))^3\beta}\right]$, where the expectation is taken over the random choice of the sampled coordinate at iterations $1, \ldots, t$ of the algorithm. Here $\overrightarrow{G}$ and $\overrightarrow{F}$ are defined as:

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

Proof We start from the result (4.1) of Lemma 4.1 when $\mu_i = 0$:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) + \sum_i p_i^{(t)} \frac{s_i^2 \|\boldsymbol{a}_i\|^2}{2\beta} |\kappa_i^{(t)}|^2,$$

by regrouping the elements and subtracting the optimal function value $D(\boldsymbol{\alpha}^*)$ from both sides we obtain:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)}) - D(\boldsymbol{\alpha}^*)|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^*) - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) + \sum_i \frac{p_i^t s_i^2}{2\beta} \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2.$$

With $\varepsilon_D^{(t)} := D(\boldsymbol{\alpha}^t) - D(\boldsymbol{\alpha}^*)$:

$$\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] \leq \varepsilon_D^{(t)} - \sum_i s_i p_i^{(t)} G_i(\boldsymbol{\alpha}^{(t)}) + \sum_i \frac{p_i^{(t)} s_i^2}{2\beta} \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2.$$

We take $p_i^{(t)} := \frac{G_i(\boldsymbol{\alpha}^{(t)})}{G(\boldsymbol{\alpha}^{(t)})}$ and $s_i := s$, then:

$$\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] \leq \varepsilon_D^{(t)} - \frac{s}{G(\boldsymbol{\alpha}^{(t)})} \sum_i (G_i(\boldsymbol{\alpha}^{(t)}))^2 + \frac{s^2}{2\beta G(\boldsymbol{\alpha}^{(t)})} \sum_i G_i(\boldsymbol{\alpha}^{(t)}) \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2.$$

To simplify the following derivation we define a duality gap vector $\overrightarrow{\boldsymbol{G}} := (G_i(\boldsymbol{\alpha}^{(t)}))_{i=1}^n$ and residual vector $\overrightarrow{\boldsymbol{F}} := (\|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2)_{i=1}^n$, the inequality becomes:

$$\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] \leq \varepsilon_D^{(t)} - \frac{s}{G(\boldsymbol{\alpha}^{(t)})} \|\overrightarrow{\boldsymbol{G}}\|_2^2 + \frac{s^2}{2\beta G(\boldsymbol{\alpha}^{(t)})} \langle \overrightarrow{\boldsymbol{G}}, \overrightarrow{\boldsymbol{F}} \rangle.$$

By bounding the last term using the Cauchy-Schwarz inequality $\langle \overrightarrow{\boldsymbol{G}}, \overrightarrow{\boldsymbol{F}} \rangle \leq \|\overrightarrow{\boldsymbol{G}}\|_2 \|\overrightarrow{\boldsymbol{F}}\|_2$ and using Lemma 4.12 we obtain:

$$\begin{aligned}
\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] &\leq \varepsilon_D^{(t)} - \frac{s}{G(\boldsymbol{\alpha}^{(t)})} \|\overrightarrow{\boldsymbol{G}}\|_2^2 + \frac{s^2}{2\beta G(\boldsymbol{\alpha}^{(t)})} \|\overrightarrow{\boldsymbol{G}}\|_2 \|\overrightarrow{\boldsymbol{F}}\|_2 \\
&= \varepsilon_D^{(t)} - \frac{s G(\boldsymbol{\alpha}^{(t)}) (\chi(\overrightarrow{\boldsymbol{G}}))^2}{n} + \frac{s^2 \chi(\overrightarrow{\boldsymbol{G}}) \chi(\overrightarrow{\boldsymbol{F}}) \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{2n\beta} \\
&\leq \varepsilon_D^{(t)} - \varepsilon_D^{(t)} \frac{(\chi(\overrightarrow{\boldsymbol{G}}))^2 s}{n} + \frac{s^2 \chi(\overrightarrow{\boldsymbol{G}}) \chi(\overrightarrow{\boldsymbol{F}}) \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{2n\beta} \\
&= \Big(1 - \frac{(\chi(\overrightarrow{\boldsymbol{G}}))^2 s}{n}\Big) \varepsilon_D^{(t)} + \frac{s^2 \chi(\overrightarrow{\boldsymbol{G}}) \chi(\overrightarrow{\boldsymbol{F}}) \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{2n\beta}.
\end{aligned}$$

In the third line we have used weak duality, that is $G(\boldsymbol{\alpha}^{(t)}) \geq \varepsilon_D^{(t)}$. Analogously to the proof of Theorem 4.8 we now prove that the suboptimality is bounded by:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \frac{2n(C + \varepsilon_D^{(0)})}{t + 2n + 1}, \tag{4.35}$$

where

$$C \geq \mathbb{E}\Big[\frac{2n\chi(\overrightarrow{\boldsymbol{F}}) \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{(\chi(\overrightarrow{\boldsymbol{G}}))^3 \beta}\Big].$$

The basis of induction at $t = 0$ obviously follows from the non-negativity of C.
Now let us prove the induction step, assume that condition (4.35) holds at step t, then by taking $s := \frac{2n}{(t+2n)(\chi(\overrightarrow{\boldsymbol{G}}))^2}$ we get:

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t+1)}|\boldsymbol{\alpha}^{(t)}] &\leq \Big(1 - \frac{(\chi(\overrightarrow{\boldsymbol{G}}))^2 s}{n}\Big)\varepsilon_D^{(t)} + \frac{s^2\chi(\overrightarrow{\boldsymbol{G}})\chi(\overrightarrow{\boldsymbol{F}})\sum_i \|\boldsymbol{a}_i\|^2|\kappa_i^{(t)}|^2}{2n\beta} \\ &\leq \Big(1 - \frac{2}{(t+2n)}\Big)\frac{2n(C+\varepsilon_D^{(0)})}{t+2n} + \frac{2n\chi(\overrightarrow{\boldsymbol{F}})\sum_i \|\boldsymbol{a}_i\|^2|\kappa_i^{(t)}|^2}{\beta(t+2n)^2(\chi(\overrightarrow{\boldsymbol{G}}))^3}.\end{aligned} \tag{4.36}$$

By taking an unconditional expectation of (4.36) and bounding by $\hat{C} := C + \varepsilon_D^{(0)}$ we obtain:

$$\begin{aligned}\mathbb{E}[\varepsilon_D^{(t+1)}] &\leq \Big(1 - \frac{2}{(t+2n)}\Big)\frac{2n\hat{C}}{t+2n} + \frac{2n}{(t+2n)^2}\mathbb{E}\left[\frac{\chi(\overrightarrow{\boldsymbol{F}})\sum_i \|\boldsymbol{a}_i\|^2|\kappa_i^{(t)}|^2}{\beta(\chi(\overrightarrow{\boldsymbol{G}}))^3}\right] \\ &\leq \Big(1 - \frac{2}{(t+2n)}\Big)\frac{2n\hat{C}}{t+2n} + \frac{2n\hat{C}}{(t+2n)^2} \\ &= \frac{2n\hat{C}}{t+2n}\Big(1 - \frac{2}{(t+2n)} + \frac{1}{(t+2n)}\Big) \\ &= \frac{2n\hat{C}}{t+2n}\frac{t+2n-1}{t+2n} \\ &\leq \frac{2n\hat{C}}{t+2n}\frac{t+2n}{t+2n+1} \\ &= \frac{2n\hat{C}}{t+2n+1}.\end{aligned}$$

And this concludes the proof. □

4.3.5 Theoretical comparison of different sampling schemes

Here we compare the rates obtained by Theorem 4.13 for gap-wise sampling and Theorem 4.8 for uniform sampling. Recall that according to the Theorem 4.8 the rate for any distribution is:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \frac{2F'^{\circ}n^2}{\frac{2}{p_{\min}} + t} = \frac{\frac{2}{\beta}\mathbb{E}\Big[\sum_i \frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2}{p_i^t}\Big] + \frac{2\varepsilon_D^{(0)}}{p_{\min}}}{\frac{2}{p_{\min}} + t}.$$

For the uniform distribution ($p_i = 1/n$) this gives:

$$\mathbb{E}[\varepsilon_D^{(t)}] \leq \frac{\frac{2n}{\beta}\mathbb{E}\Big[\sum_i |\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2\Big] + 2n\varepsilon_D^{(0)}}{2n + t}. \tag{4.37}$$

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

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

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

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

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

Chapter 5

Applications

In this chapter, we adopt the sampling distributions derived in the previous chapter and propose algorithms for two particular machine learning problems - smoothed hinge loss SVM and Lasso (L_1-regularized least squares linear regression).

5.1 Smoothed hinge loss SVM

The smoothed hinge loss SVM is a typical problem with strongly convex regularizer. This problem will be used as an application of theory developed in Section 4.2.

Primal SVM problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} P(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n}\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2. \tag{5.1}$$

Dual problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} D(\boldsymbol{\alpha}) := -\frac{1}{n}\sum_{i=1}^{n}\varphi_i^*(-\alpha_i) + \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i\boldsymbol{a}_i\right\|_2^2. \tag{5.2}$$

Here $\varphi_i(.)$ is a smoothed version of hinge loss, defined as:

$$\varphi_i(x) = \max_{v\in[-1,0]}\left[vxy_i - v - \frac{1}{2}v^2\right].$$

The conjugate of the smooth hinge loss is $\varphi_i^*(\alpha_i) = \alpha_i y_i + \frac{1}{2}\alpha_i^2$. As can be seen, the problems (5.1, 5.2) are stated in our primal-dual setting (3.1) with $g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) = \frac{1}{n}\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w})$, and $f^*(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$.

5.1.1 Duality gap

According to the first-order optimality conditions (3.2), in optimum $\boldsymbol{w}$ is:

$$\boldsymbol{w}(\boldsymbol{\alpha}) = \frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i \boldsymbol{a}_i,$$

this gives the following expression for the duality gap, in terms of a given iterate $\boldsymbol{\alpha}$:

$$G(\boldsymbol{\alpha}) = P(\boldsymbol{w}(\boldsymbol{\alpha})) - (-D(\boldsymbol{\alpha})) = \frac{1}{n}\sum_{i=1}^{n}\Big(\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) - \varphi_i^*(-\alpha_i)\Big) + \lambda\|\boldsymbol{w}\|_2^2.$$

We can rewrite it as a sum of non-negative coordinate-wise gaps, as we discussed above in Section 3.3.2:

$$\begin{aligned} G(\boldsymbol{\alpha}) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) - \varphi_i^*(-\alpha_i)\Big) + \lambda\|\boldsymbol{w}\|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) - \varphi_i^*(-\alpha_i)\Big) + \lambda \boldsymbol{w}^\top \frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i \boldsymbol{a}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) - \varphi_i^*(-\alpha_i) + \boldsymbol{a}_i^\top \alpha_i \boldsymbol{w}\Big). \end{aligned}$$

Using the definition of $\varphi_i^*(\alpha_i)$, we finally obtain:

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{n} G_i(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}) - \alpha_i y_i + \frac{1}{2}\alpha_i^2 + \boldsymbol{w}^\top \alpha_i \boldsymbol{a}_i\Big).$$

5.1.2 Algorithms

In this subsection we propose 4 variants of sampling schemes for Stochastic Coordinate Descent (SCD) for solving the dual smoothed hinge loss SVM. They are:

- **uniform** - basic SDCA algorithm from [28] adapted to our primal-dual setting. Discussed in Remark 4.6.
- **importance** - non-uniform fixed sampling scheme. Discussed and defined in Remark 4.4. In our case $\beta := n$, $\mu_i := \lambda$.
- **adaptive** - the algorithm of Csiba et al. (2015) [4] adapted to our setting. Discussed in Remark 4.5.
- (*heuristic*) - **adaptive+** the practical version of **adaptive** with lower per-iteration computational complexity. It was also introduced in [4].

Algorithm 2 Stochastic Dual Coordinate Descent (uniform & importance)

Choose mode $\in$ [uniform, importance]
let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = 0$
switch mode **do**
    case uniform
        $p_i = \frac{1}{n}$
    case importance
        $p_i = \frac{\|\boldsymbol{a}_i\|^2 + n\lambda}{\sum_j \|\boldsymbol{a}_j\|^2 + n\lambda}$
for t = 0,1,... **do**
    sample j from $[n]$ according to distribution $\boldsymbol{p}$
    $\Delta\alpha_j = y_j \max\left(0, \min\left(1, \frac{1 - y_j \boldsymbol{a}_j^\top \boldsymbol{w}^{(t)} - y_i \alpha_j^{(t)}}{\|\boldsymbol{a}_j\|^2/(\lambda n) + 1} + y_j \alpha_j^{(t)}\right)\right) - \alpha_j^{(t)}$
    $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \Delta\alpha_j \boldsymbol{e}_j$
    $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + (\lambda n)^{-1} \Delta\alpha_j \boldsymbol{a}_j$
end for

Algorithm 3 Stochastic Dual Coordinate Descent (adaptive & adaptive+)

Choose mode $\in$ [adaptive, adaptive+]
choose parameter $m > 1$ (empirically)
let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = 0$
for t = 0,1,... **do**
  switch mode **do**
    case adaptive+
      if $\mathrm{mod}(t, n) == 0$ **then**
        goto 13
      else
        $\boldsymbol{p}^{(t)} = \boldsymbol{p}^{(t-1)}$
      end if
    case adaptive
      calculate $\kappa_j^{(t)} = \alpha_j^{(t)} + \nabla\varphi_i(\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)})$, $\forall j \in [n]$
      generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$p_j^{(t)} = \frac{|\kappa_j^{(t)}|\sqrt{\lambda n + \|\boldsymbol{a}_j\|^2}}{\sum_i |\kappa_i^{(t)}|\sqrt{\lambda n + \|\boldsymbol{a}_i\|^2}}$$

  sample j from $[n]$ according to distribution $\boldsymbol{p}$
  $\Delta\alpha_j = y_j \max\left(0, \min\left(1, \frac{1 - y_j \boldsymbol{a}_j^\top \boldsymbol{w}^{(t)} - y_i \alpha_j^{(t)}}{\|\boldsymbol{a}_j\|^2/(\lambda n)+1} + y_j \alpha_j^{(t)}\right)\right) - \alpha_j^{(t)}$
  $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \Delta\alpha_j \boldsymbol{e}_j$
  $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + (\lambda n)^{-1} \Delta\alpha_j \boldsymbol{a}_j$
  $p_j^{(t)} = p_j^{(t)}/m$
end for

5.2 Lasso

The Lasso is a typical problem with general convex regularizer. This problem will be used as an application of theory developed in Section 4.3. Given a data matrix $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ and right hand side vector $\boldsymbol{y}$, the Lasso problem is stated as:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (\langle \boldsymbol{\alpha}, \boldsymbol{a}_i \rangle - y_i)^2 + \lambda \|\boldsymbol{\alpha}\|_1,$$

or alternatively

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \frac{1}{2n} \|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{5.3}$$

The problem can easily be reformulated in our primal-dual setting (3.1):

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\Big[D(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i)\Big],$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^d}\Big[P(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w})\Big].$$

Here $f(A\boldsymbol{\alpha}) = \frac{1}{2n}\|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2$ and $g_i(\alpha_i) = \lambda|\alpha_i|$.

5.2.1 Lipschitzing trick

In order to have duality gap convergence guarantees (Theorem 4.8) we need g_i^* to be Lipschitz continuous, however it is not the case for the conjugate of the absolute value function $g_i = |.|$. We modify the function g_i without affecting the iterate sequence of SCD using "Lipschitzing trick" from [5]. This method is described below.

According to Lemma 3.8, for a proper function g_i having bounded support is equivalent with g_i^* being 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 $\boldsymbol{\alpha}$ will stay inside the ball during optimization, and that the algorithm's iterate sequence will not be affected. By modifying g_i to be bounded by 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 $\bar{g}_i$-conjugate will be:

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

5.2.2 Duality gap

Using gap decomposition equation (3.6) we obtain coordinate-wise duality gaps for modified Lasso, which now depends on the chosen parameter B:

$$\begin{aligned} G(\boldsymbol{\alpha}) = \sum_i G_i(\alpha_i) &= \sum_i \Big(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w}\Big) \\ &= \sum_i \Big(B\big[|\boldsymbol{a}_i^\top \boldsymbol{w}| - \lambda\big]_+ + \lambda|\alpha_i| + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w}\Big). \end{aligned} \tag{5.4}$$

5.2.3 Algorithms with fixed sampling

In this subsection we give the coordinate descent algorithms with sampling schemes with fixed probabilities which we derived the in previous chapter. The basic Stochastic Coordinate Descent (Algorithm 1) with uniform sampling for Lasso was presented in ([24], Algorithm 1). Here we give this algorithm along with its enhanced fixed non-uniform sampling versions of *importance sampling* (see (4.29)) and heuristic *gap-init* sampling, which is based on initial coordinate-wise duality gaps.

To describe the algorithm the "soft-threshold" function $s_\tau(w)$ is defined:

$$s_\tau(w) := sign(w)(|w| - \tau)_+ = sign(w)\max\Big\{|w| - \tau, 0\Big\}$$

Algorithm 4 Stochastic Coordinate Descent

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: sample j from $[d]$ according to distribution $\boldsymbol{p}$
4: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
5: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
6: $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
7: **end for**

The name of the algorithm depends on type of sampling distribution $\boldsymbol{p}$:

- **uniform** $p_i = \frac{1}{d}$
- **importance** $p_i = \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|}$ (defined in (4.29))
- (*heuristic*) **gap-init** (defined according to (5.4))

$$p_i = \frac{G_i^{(0)}}{\sum_j G_j^{(0)}} = \frac{B\left[|\boldsymbol{a}_i^\top \boldsymbol{w}^{(0)}| - \lambda\right]_+ + \lambda|\alpha_i^{(0)}| + \alpha_i^{(0)}\boldsymbol{a}_i^\top \boldsymbol{w}^{(0)}}{\sum_j \left(B\left[|\boldsymbol{a}_j^\top \boldsymbol{w}^{(0)}| - \lambda\right]_+ + \lambda|\alpha_j^{(0)}| + \alpha_j^{(0)}\boldsymbol{a}_j^\top \boldsymbol{w}^{(0)}\right)}$$

5.2.4 Algorithms with adaptive sampling

In this subsection we consider Stochastic Coordinate Descent with adaptive sampling schemes. Here we present 5 different schemes, 4 of them are supported by the theory in the previous chapter and *ada-division* is a heuristic proposed in [4] as AdaSDCA+ and adapted by us to Lasso problem:

- **supportSet-uniform** discussed in Section 4.3.2, defined in (4.30).
- **ada-uniform** discussed in Section 4.3.3, defined in (4.32).
- **ada-gap** discussed in Section 4.3.4, defined in Theorem 4.13.
- **adaptive** discussed in Section 4.3.1, defined in (4.28).
- (*heuristic*) **ada-division** is a heuristic, which tries to resemble **adaptive** with less computational complexity.

The algorithms with aforementioned sampling schemes are given below:

Algorithm 5 Stochastic Coordinate Descent (supportSet-uniform)

let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
for t = 0,1,... **do**
  calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$

$$|\kappa_j^{(t)}| = \left|\alpha_j - B \cdot \operatorname{sign}(\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}) \cdot \left[|\boldsymbol{a}_j^\top \boldsymbol{w}^{(t)}| - \lambda\right]_+\right|$$

  find t-support set $I_t = \{i \in [d] : \kappa_i^{(t)} \neq 0\} \subseteq [d]$
  generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$\begin{cases} p_i^{(t)} = \frac{1}{|I_t|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$$

  sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
  let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
  $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
  $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
end for

Algorithm 6 Stochastic Coordinate Descent (ada-gap)

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: calculate feature-wise duality gaps $G_j^{(t)}$ for all $j \in [d]$

$$G_j^{(t)} = B\left[|\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}| - \lambda\right]_+ + \lambda|\alpha_i^{(t)}| + \alpha_i^{(t)}\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}$$

4: generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$p_i^{(t)} = \frac{G_i^{(t)}}{\sum_j G_j^{(t)}}$$

5: sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
6: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
7: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
8: $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
9: **end for**

Algorithm 7 Stochastic Coordinate Descent (adaptive)

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$

$$|\kappa_j^{(t)}| = \left|\alpha_j - B \cdot \operatorname{sign}(\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}) \cdot \left[|\boldsymbol{a}_j^\top \boldsymbol{w}^{(t)}| - \lambda\right]_+\right|$$

4: generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

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

5: sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
6: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
7: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
8: $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
9: **end for**

Algorithm 8 Stochastic Coordinate Descent (ada-uniform)

let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
for t = 0,1,... **do**
    calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$

$$|\kappa_j^{(t)}| = \left|\alpha_j - B \cdot \operatorname{sign}(\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}) \cdot \left[|\boldsymbol{a}_j^\top \boldsymbol{w}^{(t)}| - \lambda\right]_+\right|$$

    find t-support set $I_t = \{i \in [d] : \kappa_i^{(t)} \neq 0\} \subseteq [d]$
    generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$\begin{cases} p_i^{(t)} = \frac{1}{2|I_t|} + \frac{|\kappa_i^{(t)}| \|\boldsymbol{a}_i\|}{2\sum_j |\kappa_j^{(t)}| \|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$$

    sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
    let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
    $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
    $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
end for

Algorithm 9 Stochastic Coordinate Descent (ada-division)

choose parameter $m > 1$ (empirically)
let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
for t = 0,1,... **do**
  if $\mathrm{mod}(t, d) == 0$ **then**
    calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$

$$|\kappa_j^{(t)}| = \left|\alpha_j - B \cdot \mathrm{sign}(\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)}) \cdot \left[|\boldsymbol{a}_j^\top \boldsymbol{w}^{(t)}| - \lambda\right]_+\right|$$

    generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

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

  end if
  sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
  let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
  $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
  $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
  generate $\boldsymbol{p}^{(t+1)}$:

$$\begin{cases} p_i^{(t+1)} = p_i^{(t)}/m, & \text{if } i == j \\ p_i^{(t+1)} = p_i^{(t)}, & \text{otherwise} \end{cases}$$

  Normalize $\boldsymbol{p}^{(t+1)}$
end for

5.3 Comparison of computational complexity

In this section we discuss the computational cost of the Stochastic Coordinate Descent Algorithm with various sampling schemes. We first discuss the complexity of sampling and updating the probability vector, then we talk about complexity of distribution generation and variable update. The comparison is summarized in Table 5.1.

Remark 5.1 ***nnz*** *is the number of non-zero elements in the data matrix, i.e. matrix A in (3.1).*

Remark 5.2 *An* ***epoch*** *consists of n consecutive coordinate update iterations, where n is the number of datapoints in the SVM case or the number features in the Lasso case.*

Sampling and probability update In each iteration, the algorithm samples a coordinate (datapoint in SVM case or feature in Lasso case) from some non-uniform probability distribution. The straightforward approach requires $O(n)$ operations to sample a point from non-uniform distribution. To do it more efficiently we use tree data structure from [25] to maintain sampling probability vector. Tree structure requires $O(n\log(n))$ operations to build itself and $O(\log(n))$ operations to sample a point from the distribution or update one of the probabilities.

Variable update and distribution generation To compute all dual residuals κ_i or all coordinate-wise duality gaps G_i we need to do $O(\text{nnz})$ operations (due to the matrix-vector multiplication). The cost of updating the variable $\boldsymbol{\alpha}$ is $O(\text{nnz}/n)$.

Total cost of an epoch The most expensive sampling schemes are *SVM adaptive, Lasso adaptive, Lasso supportSet-uniform, Lasso ada-uniform* and *Lasso ada-gap*. For all of them we completely recompute the sampling distribution at the beginning of each iteration and this has a per-epoch complexity of $O(n \cdot \text{nnz})$. During each epoch we n times update the variable $\boldsymbol{\alpha}$, this has complexity $O(n \cdot \text{nnz}/n) = O(\text{nnz})$. Since the sampling distribution is completely recomputed each iteration, the tree structure does not give an advantage and the complexity of n samplings is $O(n^2)$. The total complexity is $O(\text{nnz} + n^2 + n \cdot \text{nnz}) = O(n \cdot \text{nnz})$.
Two heuristic adaptive schemes *SVM adaptive+* and *Lasso ada-division* recompute sampling distribution only once per epoch ($O(\text{nnz})$ operations) and update only one of the probabilities on each iteration. Here we use the tree structure with which the per-epoch cost of sampling is $O(n\log(n))$. The cost of variable update stays $O(\text{nnz})$. Overall complexity is $O(n\log(n) + 2\text{nnz}) = O(n\log(n) + \text{nnz})$.
Fixed non-uniform sampling schemes *SVM importance*, *Lasso importance* and *Lasso gap-init* require only one computation of the sampling distribution (that is $O(\text{nnz})$ operations). The complexity of n sampling operations using tree structure is $O(n\log(n))$, the complexity of a variable update is $O(\text{nnz})$. Overall asymptotic complexity is the same as in the heuristic adaptive schemes: $O(n\log(n) + \text{nnz})$.
The uniform sampling schemes have complexity per epoch of $O(\text{nnz})$.

Table 5.1: Computational cost of one epoch of SCD with various sampling schemes

Algorithm	**Cost of an Epoch**	**Mode**
SVM uniform	$O(\text{nnz})$	uniform
SVM importance	$O(\text{nnz} + n\log(n))$	fixed non-uniform
SVM adaptive	$O(n \cdot \text{nnz})$	adaptive
SVM adaptive+	$O(\text{nnz} + n\log(n))$	adaptive
Lasso uniform	$O(\text{nnz})$	uniform
Lasso importance	$O(\text{nnz} + n\log(n))$	fixed non-uniform
Lasso gap-init	$O(\text{nnz} + n\log(n))$	fixed non-uniform
Lasso supportSet-uniform	$O(n \cdot \text{nnz})$	adaptive
Lasso adaptive	$O(n \cdot \text{nnz})$	adaptive
Lasso ada-uniform	$O(n \cdot \text{nnz})$	adaptive
Lasso ada-division	$O(\text{nnz} + n\log(n))$	adaptive
Lasso ada-gap	$O(n \cdot \text{nnz})$	adaptive

Chapter 6

Experimental results

In this chapter, we evaluate the discussed algorithms on two typical convex optimization problems - smoothed hinge loss SVM and Lasso.

Performance metric We use **suboptimality** and **duality gap** (see (3.4)) as measures of algorithm performance. All figures in this chapter are given in the log scale and all the reported results are averaged over 5 runs.

Datasets The experiments were performed on two datasets from LIBSVM website[1], which are listed in Tables 6.1, 6.2. *mushrooms* is dataset from UCI repository [1], *rcv1(subsampled)* is a subsampled version of Reuters news stories corpus (*rcv1* dataset). Below we describe how the subsampling was done.
The initial *rcv1* dataset has 20242 features and 47236 datapoints. We randomly picked 10000 datapoints and 1000 features and then removed datapoints which had zero values on all picked features and features which had zero value for all picked datapoints. The resulting dataset *rcv1(subsampled)* has 7438 datapoints and 809 features.

6.1 Smoothed hinge loss SVM problem

In this section we perform experiments on smoothed hinge loss SVM. The value of parameter λ in (5.1) was chosen in a way to minimize the test error. We use $\lambda = 0.05$ for both *rcv1(subsampled)* and *mushrooms* datasets. SCD adaptive+ showed the best practical convergence with value of parameter m in range $(5, 10)$. In our experiments m was chosen to be equal to 10. The results of the experiments are shown in Figures 6.1 and 6.2. From figures we see that the adaptive method clearly outperforms all other methods in terms

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

of number of iterations. The heuristic SCD adaptive+ is slower than adaptive, but significantly faster than the method with the same asymptotic complexity - SCD importance. SCD importance shows no advantage over SCD uniform. This can be explained by relatively small variance of datapoint norms (see table 6.1), which leads to proximity of importance (4.8) and uniform sampling distributions.

Table 6.1: Datasets SVM

Dataset	Features	Datapoints	nnz/(nd)	mean of $\|\boldsymbol{a}_i\|$	Var($\|\boldsymbol{a}_i\|$)
mushrooms	112	8124	18.8%	4.58	0
rcv1(subsampled)	809	7438	0.3%	1.53	0.27

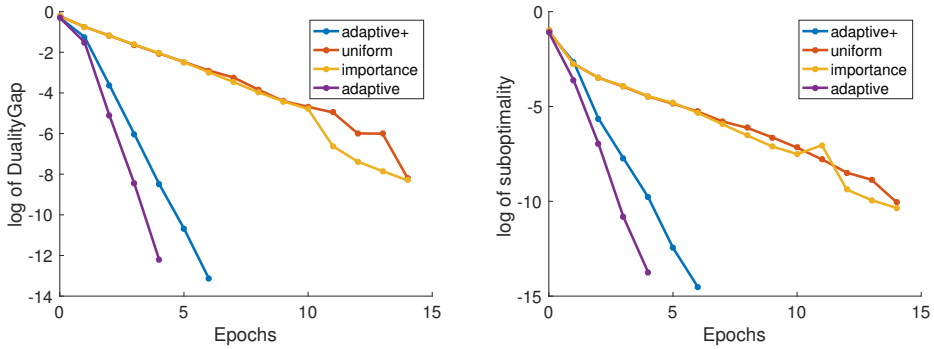

Figure 6.1: SVM. Comparison of different versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - rcv1 dataset

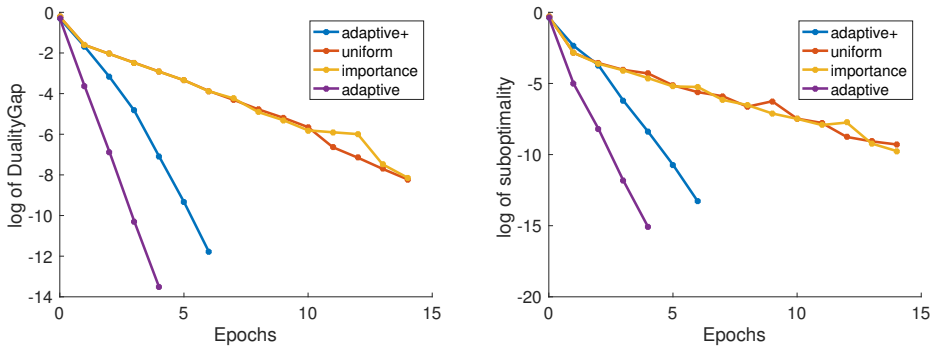

Figure 6.2: SVM. Comparison of different versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - mushrooms dataset

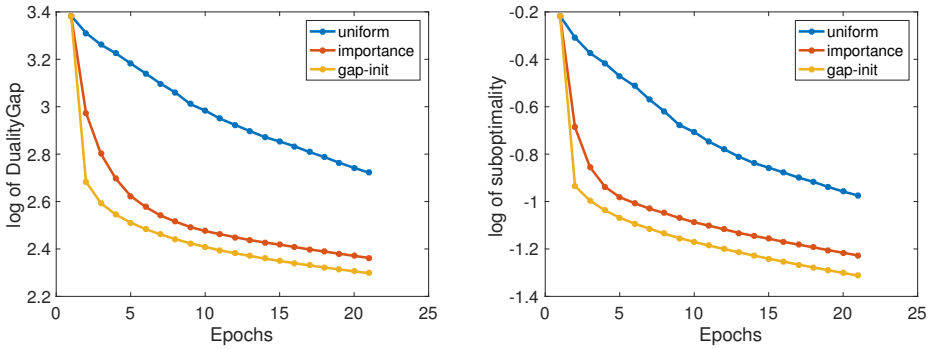

Figure 6.3: Lasso. Comparison of different fixed distribution versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and sub-optimality(right) measure - rcv1 dataset

6.2 Lasso

In this section we evaluate SCD with various sampling schemes on the Lasso problem (5.3). The parameter λ in (5.3) was chosen in a way such that the cardinality of the true support set was between 10% and 15 % of the total number of features. For *mushrooms* we use $\lambda = 0.05$, for *rcv1(subsampled)* we use $\lambda = 7 \cdot 10^{-4}$.

Table 6.2: Datasets Lasso

Dataset	Features	Datapoints	nnz/(nd)	mean of $\|\boldsymbol{a}_i\|$	Var($\|\boldsymbol{a}_i\|$)
mushrooms	112	8124	18.8%	31.35	545
rcv1(subsampled)	809	7438	0.3%	2.58	17.3

6.2.1 Methods with fixed sampling distributions

We first evaluate Lasso SCD with fixed non-uniform distributions. The results on *rcv1(subsampled)* and *mushrooms* are given in Figure 6.3 and Figure 6.4 correspondingly. As we see, SCD importance converges faster than SCD uniform on both datasets, and naturally the performance difference grows with the value of variance in the dataset (see Table 6.2). The heuristic sampling scheme based on coordinate-wise duality gap initialization - SCD gap-init significantly outperforms both of them.

6.2.2 Methods with adaptive sampling distributions

In this section we evaluate the adaptive methods. The results on *rcv1(subsampled)* and *mushrooms* are given in Figure 6.5 and Figure 6.6 correspondingly. The importance sampling is shown as a baseline method. From the figures we

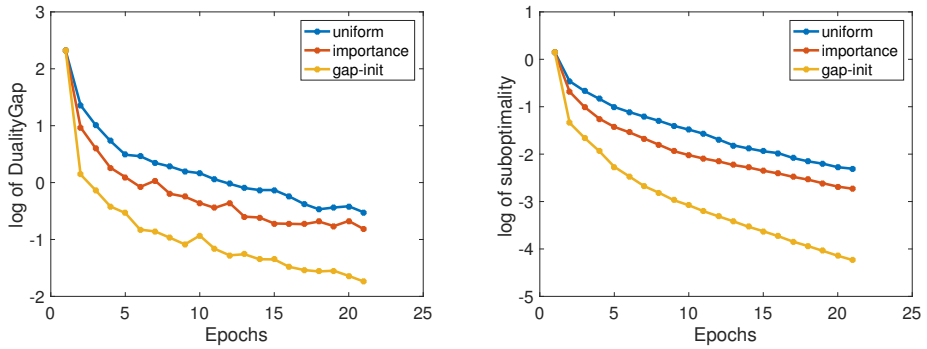

Figure 6.4: Lasso. Comparison of different fixed distribution versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - mushrooms dataset

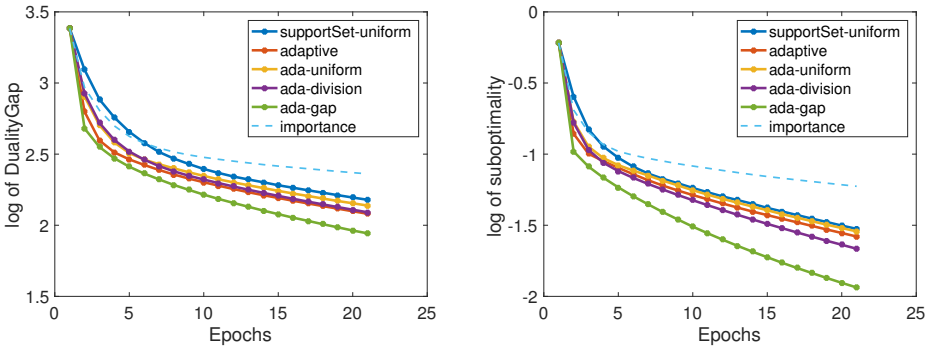

Figure 6.5: Lasso. Comparison of different adaptive versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - rcv1 dataset

see that in general all adaptive methods outperform the fixed sampling methods and have approximately same convergence speed with exception of suboptimality convergence of the SCD adaptive in Figure 6.6. The convergence process of SCD adaptive on *mushrooms* dataset shows an unusual dynamic - it has a fast (but not stable) convergence in terms of duality gap, but very poor convergence in terms of primal suboptimality.
Out of all adaptive methods SCD ada-gap and SCD ada-division algorithms showed slightly better convergence speed with both suboptimality and duality gap measures.

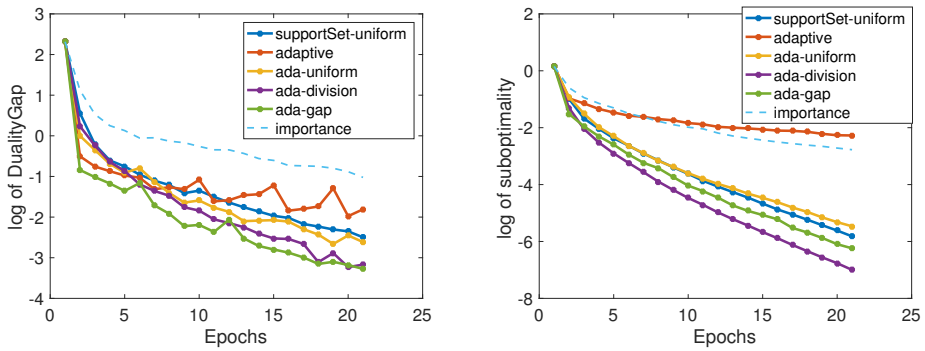

Figure 6.6: Lasso. Comparison of different adaptive versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - mushrooms dataset

6.3 Summary

In this section we summarize the results of the conducted experiments.

- the best convergence on smoothed hinge loss SVM problem showed the SCD adaptive method. This supports the developed theory. However high asymptotic computational complexity of SCD adaptive makes it impractical, instead of it we advise to use adaptive+ heuristic, which has the same moderate asymptotic computational complexity as SCD importance, but significantly faster than it in terms of convergence speed.
- the experiments on Lasso problem showed an advantage of fixed non-uniform sampling over the uniform and superiority of the adaptive sampling over the fixed. This also supports out theory.
- the experiments showed that the advantage of importance sampling over the uniform depends on the variance of $\|\boldsymbol{a}_i\|$ in the dataset.
- the best fixed non-uniform sampling proved to be gap-init, which is sampling based on initial coordinate-wise duality gaps.
- the proposed adaptive methods have approximately same performance with exception of SCD adaptive, which we do not recommend to use due to its instability.
- the best performance on Lasso was shown by SCD ada-division and SCD ada-gap. The advantage of SCD ada-division is low asymptotic computational complexity ($O(\text{nnz} + d\log(d))$), however the drawbacks are lack of theoretical basis and dependence on empirically chosen parameter m. On the other hand SCD ada-gap has proven convergence bounds

(Theorem 4.13), but the per-epoch computational cost ($O(d \cdot \text{nnz})$) is higher than the one of SCD ada-division.

Chapter 7

Conclusion

In this thesis, we discussed the Stochastic Coordinate Descent algorithm with fixed non-uniform and adaptive sampling schemes for convex partially separable optimization problems with strongly convex and general convex regularizers. We proposed a novel analysis of convergence rate for SCD with arbitrary adaptive sampling distributions. To the best of our knowledge we were first to develop such analysis for problems with general convex regularizer, e.g. Lasso. Based on our analysis we introduced new adaptive and fixed non-uniform sampling schemes and theoretically proved their superiority over conventional uniform sampling approach. We supported the developed theory with numerical experiments on smoothed hinge loss SVM and Lasso and showed in practice that the new adaptive schemes significantly outperform the non-adaptive ones.

Bibliography

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

[2] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, NY, 2011.

[3] Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines. *JMLR*, 9:1369–1398, 2008.

[4] 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*, February 2015.

[5] 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.

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

[7] 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.

[8] 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 Fifteenth International Conference on Machine Learning*, pages 188–196, 1998.

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

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

[11] Clifford Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.

[12] 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. In *the 25th International Conference on Machine Learning*, pages 408–415, New York, USA, 2008. ACM Press.

[13] Thorsten Joachims. Training linear SVMs in linear time. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2006.

[14] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Technical report, Toyota Technological Institute - Chicago, USA, 2009.

[15] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. *arXiv*, 2012.

[16] 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, July 2011. ACM-SIGEVO, ACM Press.

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

[18] 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*, pages 593–602, 2016.

[19] 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. Springer International Publishing, Cham, 2015.

[20] Zheng Qu, Peter Richtárik, and Tong Zhang. Randomized Dual Coordinate Ascent with Arbitrary Sampling.

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

[22] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[23] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. *ICML '07: Proceedings of the 24th international conference on Machine learning*, June 2007.

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

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

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

[27] Shai Shalev-Shwartz and Tong Zhang. Accelerated Mini-Batch Stochastic Dual Coordinate Ascent. *arXiv*, May 2013.

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

[29] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, Series A:1–41, November 2014.

[30] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An Improved GLMNET for L1-regularized Logistic Regression. *JMLR*, 13:1999–2030, 2012.

[31] Yuchen Zhang and Xiao Lin. Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization. In *ICML 2015 - Proceedings of the 32th International Conference on Machine Learning*, pages 353–361, 2015.

[32] Peilin Zhao and Tong Zhang. Stochastic Optimization with Importance Sampling. *arXiv*, January 2014.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

FASTER OPTIMIZATION THROUGH ADAPTIVE IMPORTANCE SAMPLING

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
PEREKRESTENKO	DMYTRO

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
12th of August Zürich	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.