Primal-Dual Rates and Certificates

Celestine Dünner   CDU@ZURICH.IBM.COM
IBM Research, Zürich, Switzerland

Simone Forte   FORTESIMONE90@GMAIL.COM
ETH Zürich, Switzerland

Martin Takáč   TAKAC.MT@GMAIL.COM
Lehigh University, USA

Martin Jaggi   JAGGIM@INF.ETHZ.CH
ETH Zürich, Switzerland

Abstract

We propose an algorithm-independent framework to equip existing optimization methods with primal-dual certificates. Such certificates and corresponding rate of convergence guarantees are important for practitioners to diagnose progress, in particular in machine learning applications.

We obtain new primal-dual convergence rates, e.g., for the Lasso as well as many L_1, Elastic Net, group Lasso and TV-regularized problems. The theory applies to any norm-regularized generalized linear model. Our approach provides efficiently computable duality gaps which are globally defined, without modifying the original problems in the region of interest.

1. Introduction

The massive growth of available data has moved data analysis and machine learning to center stage in many industrial as well as scientific fields, ranging from web and sensor data to astronomy, health science, and countless other applications. With the increasing size of datasets, machine learning methods are limited by the scalability of the underlying optimization algorithms to train these models, which has spurred significant research interest in recent years.

However, practitioners face a significant problem arising with the larger model complexity in large-scale machine learning and in particular deep-learning methods - it is increasingly hard to diagnose if the optimization algorithm

used for training works well or not. With the optimization algorithms also becoming more complex (e.g., in a distributed setting), it can often be very hard to pin down if *bad performance of a predictive model either comes from slow optimization, or from poor modeling choices.* In this light, easily verifiable guarantees for the quality of an optimization algorithm are very useful — note that the optimum solution of the problem is unknown in most cases. For convex optimization problems, a primal-dual gap can serve as such a certificate. If available, the gap also serves as a useful stopping criterion for the optimizer.

So far, the majority of popular optimization algorithms for learning applications comes without a notion of primal-dual gap. In this paper, we aim to change this for a relevant class of machine learning problems. We propose a primal-dual framework which is algorithm-independent, and allows to equip existing algorithms with additional primal-dual certificates as an add-on.

Our approach is motivated by the recent analysis of SDCA (Shalev-Shwartz & Zhang, 2013). We extend their setting to the significantly larger class of convex optimization problems of the form

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \ f(A\boldsymbol{\alpha}) + g(\boldsymbol{\alpha})$$

for a given matrix $A \in \mathbb{R}^{d \times n}$, f being smooth, and g being a general convex function. This problem class includes the most prominent regression and classification methods as well as generalized linear models. We will formalize the setting in more details in Section 3, and highlight the associated dual problem, which has the same structure. An overview over some popular examples that can be formulated in this setting either as primal or dual problem is given in Table 1.

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

- Our new primal-dual framework is algorithm-independent, that is it allows users to equip existing algorithms with primal-dual certificates and convergence rates.

- We introduce a new Lipschitzing trick allowing duality gaps (and thus accuracy certificates) which are globally defined, for a large class of convex problems which did not have this property so far. Our approach does not modify the original problems in the region of interest. In contrast to existing methods adding a small strongly-convex (L_2) term as, e.g., in (Shalev-Shwartz & Zhang, 2014; Zhang & Lin, 2015), our approach leaves both the algorithms and the optima unaffected.

- Compared with the well-known duality setting of SDCA (Shalev-Shwartz & Zhang, 2013; 2014) which is restricted to strongly convex regularizers and finite sum optimization problems, our framework encompasses a significantly larger class of problems. We obtain new primal-dual convergence rates, e.g., for the Lasso as well as many L_1, Elastic Net, group Lasso and TV-regularized problems. The theory applies to any norm-regularized generalized linear model.

- Existing primal-dual guarantees for the class of ERM problems with Lipschitz loss from (Shalev-Shwartz & Zhang, 2013) (e.g., SVM) are valid only for an "average" iteration. We show that the same rate of convergence can be achieved, e.g., for accelerated SDCA, but without the necessity of computing an average iterate.

- Our primal-dual theory captures a more precise notion of data-dependency compared with existing results (which relied on per-coordinate information only). To be precise, our shown convergence rate for the general algorithms is dependent on the *spectral norm of the data*, see also (Takáč et al., 2013; 2015).

2. Related Work

Linearized ADMM solvers. For the problem structure of our interest here, one of the most natural algorithms is the splitting method known as the Chambolle-Pock algorithm (also known as Primal-Dual Hybrid Gradient, Arrow-Hurwicz method, or linearized ADMM) (Chambolle & Pock, 2010). While this algorithm can give rise to a duality gap, it is significantly less general compared to our framework. In each iteration, it requires a complete solution of the proximal operators of both f and g, which can be computationally expensive. Its convergence rate is sensitive to the used step-size (Goldstein et al., 2015). Our framework is not algorithm-specific, and holds for arbitrary iterate sequences. More recently, the SPDC method (Zhang & Lin, 2015) was proposed as a coordinate-wise variant of (Chambolle & Pock, 2010), but only for strongly convex g.

Stochastic coordinate solvers. Coordinate descent/ascent methods have become state-of-the-art for many machine learning problems (Hsieh et al., 2008; Friedman et al., 2010). In recent years, theoretical convergence rate guarantees have been developed for the primal-only setting, e.g., by (Nesterov, 2012; Richtárik & Takáč, 2014), as well as more recently also for primal-dual guarantees, see, e.g., (Lacoste-Julien et al., 2013; Shalev-Shwartz & Zhang, 2013; 2014). The influential Stochastic[1] Dual Coordinate Ascent (SDCA) framework (Shalev-Shwartz & Zhang, 2013) was motivated by the L_2-regularized SVM, where coordinate descent is very efficient on the dual SVM formulation, with every iteration only requiring access to a single datapoint (i.e. a column of the matrix A in our setup). In contrast to primal stochastic gradient descent (SGD) methods, the SDCA algorithm family is often preferred as it is free of learning-rate parameters, and has a fast linear (geometric) convergence rate. SDCA and recent extensions (Takáč et al., 2013; Shalev-Shwartz & Zhang, 2014; Qu et al., 2014; Shalev-Shwartz, 2015; Zhang & Lin, 2015) require g to be strongly convex.

Under the weaker assumption of weak strong convexity (Necoara, 2015), a linear rate for the primal-dual convergence of SDCA was recently shown by (Ma et al., 2015b).

In the Lasso literature, a similar trend in terms of solvers has been observed recently, but with the roles of primal and dual problems reversed. For those problems, coordinate descent algorithms on the primal formulation have become the state-of-the-art, as in GLMNET (Friedman et al., 2010) and extensions (Shalev-Shwartz & Tewari, 2011; Yuan et al., 2012; 2010). Despite the prototypes of both problem types—SVM for the L_2-regularized case and Lasso for the L_1-regularized case—being closely related (Jaggi, 2014), we are not aware of existing primal-dual guarantees for coordinate descent on unmodified L_1-regularized problems.

Comparison to smoothing techniques. Existing techniques for bringing L_1-regularized and related problems into a primal-dual setting compatible with SDCA do rely on the classical Nesterov-smoothing approach - By adding a small amount of L_2 to the part g, the objective becomes strongly convex; see, e.g., (Nesterov, 2005; 2012; Tran-Dinh & Cevher, 2014; Richtárik & Takáč, 2014; Shalev-Shwartz & Zhang, 2014). However, the appropriate strength of smoothing is difficult to tune. It will depend on the accuracy level, and will influence the chosen algorithm, as it will change the iterates, the resulting convergence rate as well as the tightness of the resulting duality gaps. The line of work of Tran-Dinh & Cevher (2014; 2015) provides duality gaps for smoothed problems and rates for proximally tractable objectives (Tran-Dinh & Cevher, 2015) and

---

[1] Here 'stochastic' refers to randomized selection of the active coordinate.

Table 1. Primal-dual convergence rates of general algorithms applied to (A), for some machine learning and signal processing problems which are examples of our optimization problem formulations (A) and (B). Note that several can be mapped to either (A) or (B). $\lambda > 0$ is a regularization parameter specified by the user. We will discuss the most prominent examples in more detail in Sections 5.2 and 6.

Problem			f/f^*	g/g^*	Convergence
Regularizer	L_1	(A)	$f = \ell(A\boldsymbol{\alpha})$	$g = \lambda\|\boldsymbol{\alpha}\|_1$	Thm 6, Cor 10
	Elastic Net	(A)	$f = \ell(A\boldsymbol{\alpha})$	$g = \lambda\left(\frac{\eta}{2}\|\boldsymbol{\alpha}\|_2^2 + (1-\eta)\|\boldsymbol{\alpha}\|_1\right)$	Thm 2,8, Cor 11
		(B)	$f^* = \lambda\left(\frac{\eta}{2}\|\mathbf{w}\|_2^2 + (1-\eta)\|\mathbf{w}\|_1\right)$	$g^* = \ell(-A^\top\mathbf{w})$	Thm 2,8, Cor 11
	L_2	(A)	$f = \ell(A\boldsymbol{\alpha})$	$g = \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2$	Thm 2,8
		(B)	$f^* = \frac{\lambda}{2}\|\mathbf{w}\|_2^2$	$g^* = \ell(-A^\top\mathbf{w})$	Thm 2,8
	Fused Lasso	(A)	$f = \ell(A\boldsymbol{\alpha})$	$g = \lambda\|M\boldsymbol{\alpha}\|_1$	Thm 6
	Group Lasso	(A)	$f = \ell(A\boldsymbol{\alpha})$	$g = \lambda\sum_k \|\boldsymbol{\alpha}_{\mathcal{G}_k}\|_2$, $\{\mathcal{G}_k\}$ part. of $\{1..n\}$	Thm 6
SVM (hinge loss)		(A)	$f = \frac{\lambda}{2}\|A\boldsymbol{\alpha}\|_2^2$	$g = \sum_i -y_i\alpha_i$ s.t. $y_i\alpha_i \in [0,1]$	Thm 4,9

Loss ℓ	Logistic Regression	$\ell_{\mathsf{LR}}(\mathbf{z}) := \frac{1}{m}\sum_{j=1}^m \log(1+\exp(-y_j\mathbf{z}_j))$	for $\mathbf{z} = A\boldsymbol{\alpha}$ or $\mathbf{z} = -A^\top\mathbf{w}$
	Least Squares	$\ell_{\mathsf{LS}}(\mathbf{z}) := \frac{1}{2}\|\mathbf{z} - \mathbf{b}\|_2^2$	for $\mathbf{z} = A\boldsymbol{\alpha}$ or $\mathbf{z} = -A^\top\mathbf{w}$

also objectives with efficient Fenchel-type operator (Yurtsever et al., 2015). In contrast, our approach preserves all solutions of the original L_1-optimization — it leaves the iterate sequences of existing algorithms unchanged, which is desirable in practice, and allows the reusability of existing solvers. We do not assume proximal or Fenchel tractability.

Distributed algorithms. For L_1-problems exceeding the memory capacity of a single computer, a communication-efficient distributed scheme leveraging this Lipschitzing trick is presented in (Smith et al., 2015; Forte, 2015).

3. Setup and Primal-Dual Structure

In this paper, we consider optimization problems of the following primal-dual structure. As we will see, the relationship between primal and dual objectives has many benefits, including computation of the duality gap, which allows us to have certificates for the approximation quality.

We consider the following pair of optimization problems, which are dual[2] to each other:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left[\; \mathcal{D}(\boldsymbol{\alpha}) := \quad f(A\boldsymbol{\alpha}) + g(\boldsymbol{\alpha}) \quad\right], \quad \text{(A)}$$

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left[\; \mathcal{P}(\mathbf{w}) := \quad f^*(\mathbf{w}) + g^*(-A^\top\mathbf{w}) \;\right]. \quad \text{(B)}$$

The two problems are associated to a given data matrix $A \in \mathbb{R}^{d\times n}$, and the functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are allowed to be arbitrary closed convex functions. Here $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^d$ are the respective variable vectors. The relation of (A) and (B) is called *Fenchel-Rockafellar Duality* where the functions f^*, g^* in formulation (B) are defined as the *convex conjugates*[3] of their corresponding counterparts f, g in (A). The two main powerful features of this general duality structure are first that it includes many more machine learning methods than more traditional du-

---

[2]For a self-contained derivation see Appendix C.
[3]The conjugate is defined as $h^*(\mathbf{v}) := \sup_{\mathbf{u}\in\mathbb{R}^d} \mathbf{v}^\top\mathbf{u} - h(\mathbf{u})$.

ality notions, and secondly that the two problems are fully symmetric, when changing respective roles of f and g. In typical machine learning problems, the two parts typically play the roles of a data-fit (or loss) term as well as a regularization term. As we will see later, those two roles can be swapped, depending on the application.

Optimality Conditions. The first-order optimality conditions for our pair of vectors $\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\alpha} \in \mathbb{R}^n$ in problems (A) and (B) are given as

$$\mathbf{w} \in \partial f(A\boldsymbol{\alpha}), \quad \text{(1a)} \qquad -A^\top\mathbf{w} \in \partial g(\boldsymbol{\alpha}), \qquad \text{(2a)}$$

$$A\boldsymbol{\alpha} \in \partial f^*(\mathbf{w}), \quad \text{(1b)} \qquad \boldsymbol{\alpha} \in \partial g^*(-A^\top\mathbf{w}) \quad \text{(2b)}$$

see, e.g. (Bauschke & Combettes, 2011, Proposition 19.18). The stated optimality conditions are equivalent to $\boldsymbol{\alpha}, \mathbf{w}$ being a saddle-point of the Lagrangian, which is given as $\mathcal{L}(\boldsymbol{\alpha}, \mathbf{w}) = f^*(\mathbf{w}) - \langle A\boldsymbol{\alpha}, \mathbf{w}\rangle - g(\boldsymbol{\alpha})$ if $\boldsymbol{\alpha} \in \text{dom}(g)$ and $\mathbf{w} \in \text{dom}(f^*)$.

Duality Gap. From the definition of the dual problems in terms of the convex conjugates, we always have $\mathcal{P}(\mathbf{w}) \geq \mathcal{P}(\mathbf{w}^\star) \geq -\mathcal{D}(\boldsymbol{\alpha}^\star) \geq -\mathcal{D}(\boldsymbol{\alpha})$, giving rise to the definition of the general duality gap $G(\mathbf{w}, \boldsymbol{\alpha}) := \mathcal{P}(\mathbf{w}) - (-\mathcal{D}(\boldsymbol{\alpha}))$.

For differentiable f, the duality gap can be used more conveniently: Given $\boldsymbol{\alpha} \in \mathbb{R}^n$ s.t. $A\boldsymbol{\alpha} \in \text{dom}(f)$ in the context of (A), a corresponding variable vector $\mathbf{w} \in \mathbb{R}^d$ for problem (B) is given by the first-order optimality condition (1a) as

$$\mathbf{w} = \mathbf{w}(\boldsymbol{\alpha}) := \nabla f(A\boldsymbol{\alpha}). \qquad (3)$$

Under strong duality, we have $\mathcal{P}(\mathbf{w}^\star) = -\mathcal{D}(\boldsymbol{\alpha}^\star)$ and $\mathbf{w}(\boldsymbol{\alpha}^\star) = \mathbf{w}^\star$, where $\boldsymbol{\alpha}^\star$ is an optimal solution of (A). This implies that the suboptimality $\mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - \mathcal{P}(\mathbf{w}^\star)$ is always bounded above by the simpler duality gap function

$$G(\boldsymbol{\alpha}) := \mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - (-\mathcal{D}(\boldsymbol{\alpha})) \geq \mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - \mathcal{P}(\mathbf{w}^\star) \quad (4)$$

which hence acts as a certificate of the approximation quality of the variable vector $\boldsymbol{\alpha}$.

4. Primal-Dual Guarantees for Any Algorithm Solving (A)

In this section we state an important lemma, which will later allow us to transform a suboptimality guarantee of any algorithm into a duality gap guarantee, for optimization problems of the form specified in the previous section.

Lemma 1. *Consider an optimization problem of the form* (A). *Let f be $1/\beta$-smooth w.r.t. a norm $\|.\|_f$ and let g be μ-strongly convex with convexity parameter $\mu \geq 0$ w.r.t. a norm $\|.\|_g$. The general convex case $\mu = 0$ is explicitly allowed, but only if g has bounded support.*

Then, for any $\boldsymbol{\alpha} \in \mathrm{dom}(\mathcal{D})$ and any $s \in [0,1]$, it holds that

$$\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) \geq sG(\boldsymbol{\alpha}) \tag{5}$$
$$+ \tfrac{s^2}{2}\big(\tfrac{\mu(1-s)}{s}\|\mathbf{u}-\boldsymbol{\alpha}\|_g^2 - \tfrac{1}{\beta}\|A(\mathbf{u}-\boldsymbol{\alpha})\|_f^2\big),$$

where $G(\boldsymbol{\alpha})$ is the gap function defined in (4) *and*

$$\mathbf{u} \in \partial g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})). \tag{6}$$

We note that the improvement bound here bears similarity to the proof of (Bach, 2015, Prop 4.2) for the case of an extended Frank-Wolfe algorithm. In contrast, our result here is algorithm-independent, and leads to tighter results due to the more careful choice of $\mathbf{u}$, as we'll see in the following.

4.1. Linear Convergence Rates

In this section we assume that we are using an arbitrary optimization algorithm applied to problem (A). It is assumed that the algorithm produces a sequence of (possibly random) iterates $\{\boldsymbol{\alpha}^{(t)}\}_{t=0}^\infty$ such that there exists $C \in (0,1], D \geq 0$ such that

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star)] \leq (1-C)^t\, D. \tag{7}$$

In the next two theorems, we define $\sigma := \big(\max_{\boldsymbol{\alpha}\neq 0}\|A\boldsymbol{\alpha}\|_f/\|\boldsymbol{\alpha}\|_g\big)^2$, i.e., the squared spectral norm of the matrix A in the Euclidean norm case.

4.1.1. CASE I. STRONGLY CONVEX g

Let us assume g is μ-strongly convex ($\mu > 0$) (equivalently, its conjugate g^* has Lipschitz continuous gradient with a constant $1/\mu$). The following theorem provides a linear convergence guarantee for any algorithm with given linear convergence rate for the suboptimality $\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star)$.

Theorem 2. *Assume the function f is $1/\beta$-smooth w.r.t. a norm $\|.\|_f$ and g is μ-strongly convex w.r.t. a norm $\|.\|_g$. Suppose we are using a linearly convergent algorithm as specified in* (7). *Then, for any*

$$t \geq T := \tfrac{1}{C}\log\frac{D(\frac{\sigma}{\beta}+\mu)}{\mu\epsilon} \tag{8}$$

it holds that $\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \leq \epsilon$.

From (7) we can obtain that after $\frac{1}{C}\log\frac{D}{\epsilon}$ iterations, we would have a point $\boldsymbol{\alpha}^{(t)}$ such that $\mathbb{E}[D(\boldsymbol{\alpha}^{(t)})-D(\boldsymbol{\alpha}^\star)] \leq \epsilon$. Hence, comparing with (8) only few more iterations are needed to get the guarantees for the duality gap. The rate (7) is achieved by most of the first order algorithms, including proximal gradient descent (Nesterov, 2013) or SDCA (Richtárik & Takáč, 2014) with $C \sim \mu$ or accelerated SDCA (Lin et al., 2014) with $C \sim \sqrt{\mu}$.

4.1.2. CASE II. GENERAL CONVEX g (OF BOUNDED SUPPORT)

In this section we will assume that g^* is Lipschitz (in contrast to smooth as in Theorem 2) and show that the linear convergence rate is preserved.

Theorem 3. *Assume that the function f is $1/\beta$-smooth w.r.t. a norm $\|.\|$, g^* is L-Lipschitz continuous w.r.t the dual norm $\|.\|_*$, and we are using a linearly convergent algorithm* (7). *Then, for any*

$$t \geq T := \tfrac{1}{C}\log\frac{2D\max\{1,2\sigma L^2/\epsilon\beta\}}{\epsilon} \tag{9}$$

it holds that $\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \leq \epsilon$.

In (Wang & Lin, 2014), it was proven that feasible descent methods when applied to the dual of an SVM do improve the objective geometrically as in (7). Later, (Ma et al., 2015b) extended this to stochastic coordinate feasible descent algorithms (including SDCA). Using our new Theorem 3, we can therefore extend their results to linear convergence for the duality gap for the SVM application.

4.2. Sub-Linear Convergence Rates

In this case we will focus only on general L-Lipschitz continuous functions g^* (if g is strongly convex, then many existing algorithms are available and converge with a linear rate).

We will assume that we are applying some (possibly randomized) algorithm on optimization problem (A) which produces a sequence (of possibly random) iterates $\{\boldsymbol{\alpha}^{(t)}\}_{t=0}^\infty$ such that

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star)] \leq \tfrac{C}{D(t)}, \tag{10}$$

where $D(t)$ is a function wich has usually a linear or quadratic growth (i.e. $D(t) \sim \mathcal{O}(t)$ or $D(t) \sim \mathcal{O}(t^2)$).

The following theorem will allow to equip existing algorithms with sub-linear convergence in suboptimality, as specified in (10), with duality gap convergence guarantees.

Theorem 4. *Assume the function f is $1/\beta$-smooth w.r.t. the norm $\|.\|$, g^* is L-Lipschitz continuous, w.r.t. the dual norm $\|.\|_*$, and we are using a sub-linearly convergent algorithm as quantified by* (10). *Then, for any $t \geq 0$ such that*

$$D(t) \geq \max\{\tfrac{2C\beta}{\sigma L^2}, \tfrac{2C\sigma L^2}{\beta\epsilon^2}\}, \tag{11}$$

it holds that $\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \leq \epsilon$.

Let us comment on Theorem 4 stated above. If $D(t) \sim \mathcal{O}(t)$ then this shows a rate of $\mathcal{O}(\epsilon^{-2})$. We note two important facts:

1. The guarantee holds for the duality gap of the iterate $\alpha^{(t)}$ and not for some averaged solution.
2. For the SVM case, this rate is consistent with the result of (Hush et al., 2006). Our result is much more general as it holds for any L-Lipschitz continuous convex function g^* and any β-strongly convex f^*.

Let us make one more important remark. In (Shalev-Shwartz & Zhang, 2013) the authors showed that SDCA (when applied on L-Lipschitz continuous g^*) has $D(t) \sim \mathcal{O}(t)$ and they also showed that an averaged solution (over few last iterates) needs only $\mathcal{O}(\epsilon^{-1})$ iterations to have duality gap $\leq \epsilon$. However, as a direct consequence of our Theorem 4 we can show, e.g., that FISTA (Beck & Teboulle, 2009) (aka. accelerated gradient descent algorithm) or APPROX (Fercoq & Richtárik, 2015)[4] (a.k.a. accelerated coordinate descent algorithm) will need $\mathcal{O}(\epsilon^{-1})$ iterations to produce an iterate α such that $G(\alpha) \leq \epsilon$. Indeed, e.g., for APPROX, the function $D(t) = ((t-1)\tau + 2n)^2$ (where τ is the size of a mini-batch) and C is a constant which depends on $\alpha^{(0)}$ and α^* and τ. Hence, to obtain an iterate $\alpha^{(t)}$ such that $\mathbb{E}[G(\alpha^{(t)})] \leq \epsilon$ it is sufficient to choose $\tau \geq 1$ such that $t \overset{(11)}{\geq} 1 - \frac{2n}{\tau} + \max\{\frac{1}{\tau}\sqrt{\frac{2C\beta}{\sigma L^2}}, \frac{L}{\epsilon\tau}\sqrt{\frac{2C\sigma}{\beta}}\}$ is satisfied.

5. Extending Duality to Non-Lipschitz Functions

5.1. Lipschitzing Trick

In this section we present a trick that allows to generalize our results of the previous section from Lipschitz functions g^* to non-Lipschitz functions. The approach we propose, which we call the *Lipschitzing trick*, will make a convex function Lipschitz on the entire domain $\mathbb{R}^d$, by modifying its conjugate dual to have bounded support. Formally, the modification is as follows: Let $\mathcal{B} \subset \mathbb{R}^d$ be some closed, bounded convex set. We modify $g : \mathbb{R}^n \to \mathbb{R}$ by restricting its support to the set $\mathcal{B}$, i.e.

$$\bar{g}(\alpha) := \begin{cases} g(\alpha) & \text{if } \alpha \in \mathcal{B} \\ +\infty & \text{otherwise} \end{cases}. \tag{12}$$

By definition, this function has bounded support, and hence, by Lemma 5, its conjugate function $\bar{g}^*$ is Lipschitz continuous.

Motivation. We will apply this trick to the part g of optimization problems of the form (A) (such that g^* will become Lipschitz). We want that this modification to have no impact on the outcome of any optimization algorithm run-

---

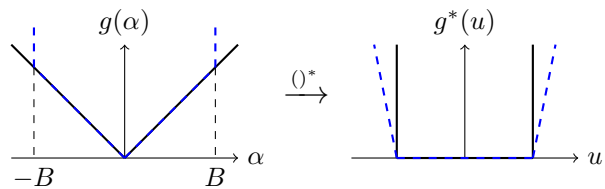[4]APPROX requires g being separable.

Figure 1. Illustration Lipschitzing trick for the scalar function $g(\alpha) = |\cdot|$ with $\mathcal{B} = \{x : |x| \leq B\}$.

ning on (A). Instead, this trick should only affect convergence theory in that it allows us to present a strong primal-dual rate. In the following, we will discuss how this can indeed be achieved by imposing only very weak assumptions on the original problem (A).

The modification is based on the following duality of Lipschitzness and bounded support, as given in Lemma 5 below, which is a generalization of (Rockafellar, 1997, Corollary 13.3.3). We need the following definition:

Definition 1 (B-Bounded Support). A function $g : \mathbb{R}^d \to \mathbb{R}$ has B-*bounded support* if its effective domain is bounded by B w.r.t. a norm $\|.\|$, i.e.,

$$g(\mathbf{u}) < +\infty \implies \|\mathbf{u}\| \leq B. \tag{13}$$

Lemma 5 (Duality between Lipschitzness and L-Bounded Support). *Given a proper convex function g, it holds that g has L-bounded support w.r.t. the norm $\|.\|$ if and only if g^* is L-Lipschitz w.r.t. the dual norm $\|.\|_*$.*

The following Theorem 6 generalizes our previous convergence results, which were restricted to Lipschitz g^*.

Theorem 6. *For an arbitrary optimization algorithm running on problem (A), let $\alpha^{(t)}$ denote its iterate sequence. Assume there is some closed convex set $\mathcal{B}$ containing all these iterates. Then, the same optimization algorithm run on the modified problem — given by Lipschitzing of g^* using $\mathcal{B}$ — would produce exactly the same iterate sequence. Furthermore, Theorem 3 as well as Theorem 4 give primal-dual convergence guarantees for this algorithm (for L such that $\mathcal{B}$ is L-bounded).*

Corollary 7. *Assume the objective of optimization problem (A) has bounded level sets. For $\alpha^{(t)}$ being the iterate sequence of a* monotone *optimization algorithm on problem (A) we denote $\delta_t := \mathcal{D}(\alpha^{(t)})$ and let $\mathcal{B}_t$ be the δ_t-level set of $\mathcal{D}$. Write $B_t > 0$ for a value such that $\mathcal{B}_t$ is B_t-bounded w.r.t. the norm $\|.\|$. Then, at any state t of the algorithm, the set $\mathcal{B}_t$ contains all future iterates and Theorem 6 applies for $L := B_t$.*

5.2. Norm-Regularized Problems

We now focus on some applications. First, we demonstrate how the Lipschitzing trick can be applied to find primal-dual convergence rates for problems regularized by an arbitrary norm. We discuss in particular the Lasso problem

and show how the suboptimality bound can be evaluated in practice. In a second part, we discuss the Elastic Net regularizer and show how it fits into our framework.

We consider a special structure of problem (A), namely

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} f(A\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|. \tag{14}$$

where f is some convex non-negative smooth loss function regularized by any norm $\|.\|$. We choose $\mathcal{B}$ to be the $\|.\|$-norm ball of radius B, such that $\|\boldsymbol{\alpha}\| < B$ for every iterate. The size, B, of this ball can be controled by the amount of regularization λ. Using the Lipschitzing trick with this $\mathcal{B}$, the convergence guarantees of Theorem 6 apply to the general class of norm regularized problems.

Note that for any monotone algorithm (initialized at $\mathbf{0}$) applied to (14) we have $\|\boldsymbol{\alpha}\| \leq \frac{1}{\lambda} f(\mathbf{0})$ for every iterate. Furthermore, at every iterate $\boldsymbol{\alpha}$ we can bound $\|\boldsymbol{\alpha}^+\| \leq \frac{1}{\lambda}(f(A\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|)$ for every future iterate $\boldsymbol{\alpha}^+$. Hence, $B := \frac{1}{\lambda} f(\mathbf{0})$ is a safe choice, e.g. $B := \frac{1}{2\lambda}\|\mathbf{b}\|_2^2$ for least squares loss and $B := \frac{m}{\lambda}\log(2)$ for logistic regression loss.

Duality Gap. For any problem of the form (14) we can now determine the duality gap. We apply the Lipschitzing trick to $g(\boldsymbol{\alpha}) := \lambda\|\boldsymbol{\alpha}\|$ as in (12), then the convex conjugate of $\bar{g}$ is

$$\bar{g}^*(\mathbf{u}) = \begin{cases} 0 & \|\mathbf{u}\|_* \leq \lambda \\ \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\in\mathcal{B}} \mathbf{u}^\top\boldsymbol{\alpha} - \lambda\|\boldsymbol{\alpha}\| & \text{else} \end{cases}. \tag{15}$$

where $\|.\|_*$ denotes the dual norm of $\|.\|$. Hence, using the primal-dual mapping $\mathbf{w}(\boldsymbol{\alpha}) := \nabla f(A\boldsymbol{\alpha})$ we can write the duality gap of the modified problem as

$$\bar{G}(\boldsymbol{\alpha}) = \langle \mathbf{w}(\boldsymbol{\alpha}), A\boldsymbol{\alpha}\rangle + \lambda\|\boldsymbol{\alpha}\| + \bar{g}^*(-A^\top\mathbf{w}(\boldsymbol{\alpha})) \tag{16}$$

Note that the computation of the modified duality gap $\bar{G}(\boldsymbol{\alpha})$ is not harder than the computation of the original gap $G(\boldsymbol{\alpha})$ - it requires one pass over the data, assuming the choice of a suitable set $\mathcal{B}$ in (15). Furthermore, note that in contrast to the unmodified duality gap $G(\boldsymbol{\alpha})$, which is only defined on the set $\{\boldsymbol{\alpha} : \|A^\top\mathbf{w}(\boldsymbol{\alpha})\|_* \leq \lambda\}$, our new gap $\bar{G}(\boldsymbol{\alpha})$ is defined on the entire space $\mathbb{R}^d$.

As an alternative, (Mairal, 2010, Sec. 1.4.1 and App. D.2) has defined a different duality gap by shrinking the dual $\mathbf{w}$ variable until it becomes feasible in $\|A^\top\mathbf{w}\|_* \leq \lambda$.

5.2.1. L_1-REGULARIZED PROBLEMS

The results from the previous section can be ported to the important special case of L_1-regularization:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} f(A\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1. \tag{17}$$

We choose $\mathcal{B}$ to be the L_1-norm ball of radius B and then apply the Lipschitzing trick with $\mathcal{B}$ to the regularization term in (17). An illustration of this modification as well as the impact on its dual are illustrated in Figure 1. Hence, our

theory (Theorem 6) gives primal-dual convergence guarantees for any algorithm applied to the Lasso problem (17). Furthermore, if the algorithm is monotone (initialized at $\boldsymbol{\alpha} := \mathbf{0}$) we know that $\mathcal{B}$ is $\frac{1}{\lambda} f(\mathbf{0})$-bounded.

Duality Gap. The duality gap (16) for the modified Lasso problem can be computed at every iterate $\boldsymbol{\alpha}$ as

$$\bar{G}(\boldsymbol{\alpha}) = \langle \mathbf{w}, A\boldsymbol{\alpha}\rangle + B\left[\|A^\top\mathbf{w}\|_\infty - \lambda\right]_+ + \lambda\|\boldsymbol{\alpha}\|_1 \tag{18}$$

for $\mathbf{w} = \mathbf{w}(\boldsymbol{\alpha})$. For the derivation, see Appendix I.1.

5.2.2. GROUP LASSO, FUSED LASSO AND TV-NORM

Our algorithm-independent, primal-dual convergence guarantees and certificates presented so far are not restricted to L_p-regularized problems, but do in fact directly apply to many more general structured regularizers, some of them shown in see Table 1. This includes group Lasso (L_1/L_2) and other norms inducing structured sparsity (Bach et al., 2012), as well as other penalties such as e.g. the fused Lasso $g(\boldsymbol{\alpha}) := \lambda\|M\boldsymbol{\alpha}\|_1$. The total variation denoising problem is obtained for suitable choice of the matrix M.

5.2.3. ELASTIC NET REGULARIZED PROBLEMS

The second application we will discuss is Elastic Net regularization

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \ell(A\boldsymbol{\alpha}) + \lambda\left(\frac{\eta}{2}\|\boldsymbol{\alpha}\|_2^2 + (1-\eta)\|\boldsymbol{\alpha}\|_1\right), \tag{19}$$

for fixed trade-off parameter $\eta \in (0, 1]$.

Our framework allows two different ways to solve problem (19): Either mapping it to formulation (A) (for $\ell=f$ smooth) as in row 2 of Table 1, or to (B) (for general ℓ) as in row 3 of Table 1. In both scenarios, Theorem 2 gives us a fast linear convergence guarantee for the duality gap, if ℓ is smooth. The other theorems apply accordingly for general ℓ when the problem is mapped to (B).

Whether the choice of a dual or primal optimizer will be more beneficial in practice depends on the case, and will be discussed in more detail for coordinate descent methods in Section 6.

Duality Gap. For the Elastic Net problem (19) mapped to (A), we can compute the duality gap (4) as follows:

$$G(\boldsymbol{\alpha}) = \langle \mathbf{w}, A\boldsymbol{\alpha}\rangle + \frac{1}{2\eta\lambda}\sum_{i=1}^n \left[\left|A_{:i}^\top\mathbf{w}\right| - (1-\eta)\lambda\right]_+^2$$
$$+ \lambda\left(\frac{\eta}{2}\|\boldsymbol{\alpha}\|_2^2 + (1-\eta)\|\boldsymbol{\alpha}\|_1\right)$$

with $\mathbf{w} = \mathbf{w}(\boldsymbol{\alpha}) = \nabla\ell(A\boldsymbol{\alpha})$, see Appendix I.2.

Remark 1. *As $\eta \to 0$ we approach the pure L_1-case and this gap blows up as $G(\boldsymbol{\alpha}) \to \infty$. Comparing this to (18), we see that the Lipschitzing trick allows to get certificates even in cases where the duality gap of the unmodified problem is infinity.*

6. Coordinate Descent Algorithms

We now focus on a very important class of algorithms, that is coordinate descent methods. In this section, we show how our theory implies much more general primal-dual convergence guarantees for coordinate descent algorithms.

Partially Separable Problems. A widely used subclass of optimization problems arises when one part of the objective becomes separable. Formally, this is expressed as $g(\boldsymbol{\alpha}) = \sum_{i=1}^{n} g_i(\alpha_i)$ for univariate functions $g_i : \mathbb{R} \to \mathbb{R}$ for $i \in [n]$. Nicely in this case, the conjugate of g also separates as $g^*(\mathbf{y}) = \sum_i g_i^*(y_i)$. Therefore, the two optimization problems (A) and (B) write as

$$\mathcal{D}(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i) \qquad \text{(SA)}$$

$$\mathcal{P}(\mathbf{w}) := f^*(\mathbf{w}) + \sum_i g_i^*(-A_{:i}^{\top}\mathbf{w}) , \qquad \text{(SB)}$$

where $A_{:i} \in \mathbb{R}^d$ denotes the i-th column of A.

The Algorithm. We consider the coordinate descent algorithm described in Algorithm 1. Initialize $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$ and then, at each iteration, sample and update a random coordinate $i \in [n]$ of the parameter vector $\boldsymbol{\alpha}$ to iteratively minimize (SA). Finally, after T iterations output $\bar{\boldsymbol{\alpha}}$, the average vector over the latest $T - T_0$ iterates. The parameter T_0 is some positive number smaller than T.

Algorithm 1 Coordinate Descent on $\mathcal{D}(\boldsymbol{\alpha})$

1: **Input:** Data matrix A.
   Starting point $\boldsymbol{\alpha}^{(0)} := \mathbf{0} \in \mathbb{R}^n$, $\mathbf{w}^{(0)} = \mathbf{w}(\boldsymbol{\alpha}^{(0)})$.
2: **for** $t = 1, 2, \ldots T$ **do**
3:    Pick $i \in [n]$ randomly
4:    Find $\Delta\alpha_i$ minimizing $\mathcal{D}(\boldsymbol{\alpha}^{(t-1)} + \mathbf{e}_i\Delta\alpha_i)$
5:    $\boldsymbol{\alpha}^{(t)} \leftarrow \boldsymbol{\alpha}^{(t-1)} + \Delta\alpha_i\mathbf{e}_i$
6:    $\mathbf{w}^{(t)} \leftarrow \mathbf{w}(\boldsymbol{\alpha}^{(t)})$
7: **end for**
8: Let $\bar{\boldsymbol{\alpha}} = \frac{1}{T-T_0} \sum_{t=T_0}^{T-1} \boldsymbol{\alpha}^{(t)}$

As we will show in the following section, coordinate descent on $\mathcal{D}(\boldsymbol{\alpha})$ is not only an efficient optimizer of the objective $\mathcal{D}(\boldsymbol{\alpha})$, but also provably reduces the duality gap. Therefore, the same algorithm will simultaneously optimize the dual objective $\mathcal{P}(\mathbf{w})$.

6.1. Primal-Dual Analysis for Coordinate Descent

We first show linear primal-dual convergence rate of Algorithm 1 applied to (SA) for strongly convex g_i. Later, we will generalize this result to also apply to the setting of general Lipschitz g_i. This generalization together with the Lipschitzing trick will allow us to derive primal-dual convergence guarantees of coordinate descent for a much broader class of problems, including the Lasso problem.

For the following theorems we assume that the columns of the data matrix A are scaled such that $\|A_{:i}\| \leq R$ for all $i \in [n]$ and $\|A_{j:}\| \leq P$ for all $j \in [d]$, for some norm $\|.\|$.

Theorem 8. *Consider Algorithm 1 applied to* (SA)*. Assume f is a $1/\beta$-smooth function w.r.t. the norm $\|.\|$. Then, if g_i is μ-strongly convex for all i, it suffices to have a total number of iterations of*

$$T \geq \left(n + \frac{nR^2}{\mu\beta}\right) \log\left(\left[n + \frac{nR^2}{\mu\beta}\right] \frac{\epsilon_D^{(0)}}{\epsilon}\right)$$

to get $\mathbb{E}[G(\boldsymbol{\alpha}^{(T)})] \leq \epsilon$. Moreover, to obtain an expected duality gap of $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \leq \epsilon$ it suffices to have $T > T_0$ with

$$T_0 \geq \left(n + \frac{nR^2}{\mu\beta}\right) \log\left(\left[n + \frac{nR^2}{\mu\beta}\right] \frac{\epsilon_D^{(0)}}{(T-T_0)\epsilon}\right)$$

where $\epsilon_D^{(0)}$ is the initial suboptimality in $\mathcal{D}(\boldsymbol{\alpha})$.

Theorem 8 allows us to upper bound the duality gap, and hence the suboptimality, for every iterate $\boldsymbol{\alpha}^{(T)}$, as well as the average $\bar{\boldsymbol{\alpha}}$ returned by Algorithm 1. In the following we generalize this result to apply to L-Lipschitz functions g_i.

Theorem 9. *Consider Algorithm 1 applied to* (SA)*. Assume f is a $1/\beta$-smooth function w.r.t. the norm $\|.\|$. Then, if g_i^* is L-Lipschitz for all i, it suffices to have a total number of iterations of*

$$T \geq \max\left\{0, n\log\frac{\epsilon_D^{(0)}\beta}{2L^2R^2n}\right\} + n + \frac{20n^2L^2R^2}{\beta\epsilon}$$

to get $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \leq \epsilon$. Moreover, when $t \geq T_0$ with

$$T_0 = \max\left\{0, n\log\frac{\epsilon_D^{(0)}\beta}{2L^2R^2n}\right\} + \frac{16n^2L^2R^2}{\beta\epsilon}$$

we have the suboptimality bound of $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^{\star})] \leq \epsilon/2$, where $\epsilon_D^{(0)}$ is the initial suboptimality.

Remark 2. *Theorem 9 shows that for Lipschitz g_i^*, Algorithm 1 has $\mathcal{O}(\epsilon^{-1})$ convergence in the suboptimality and $\mathcal{O}(\epsilon^{-1})$ convergence in $G(\bar{\boldsymbol{\alpha}})$. Comparing this result to Theorem 4 which suggests $\mathcal{O}(\epsilon^{-2})$ convergence in $G(\boldsymbol{\alpha})$ for $\mathcal{O}(\epsilon^{-1})$ convergent algorithms, we see that averaging the parameter vector crucially improves convergence in the case of non-smooth f.*

Remark 3. *Note that our Algorithm 1 recovers the widely used SDCA setting (Shalev-Shwartz & Zhang, 2013) as a special case, when we choose $f^* := \frac{\lambda}{2}\|.\|_2^2$ in (SB). Furthermore, their convergence results for SDCA are consistent with our results and can be recovered as a special case of our analysis. See Corollaries 16, 18, 19 in Appendix J.*

6.2. Application to L_1 and Elastic Net Regularized Problems

We now apply Algorithm 1 to the L_1-regularized problems, as well as Elastic Net regularized problems. We state improved primal-dual convergence rates which are more tailored to the coordinate-wise setting.

Coordinate Descent on L_1-Regularized Problems. In contrast to the general analysis of L_1-regularized problems in Section 5.2.1, we can now exploit separability of $g(\boldsymbol{\alpha}) := \lambda\|\boldsymbol{\alpha}\|_1$ and apply the Lipschitzing trick coordinate-wise, choosing $\mathcal{B} := \{\alpha : |\alpha| \le B\} \subset \mathbb{R}$. This results in the following stronger convergence results:

Corollary 10. *We can use the Lipschitzing trick together with Theorem 9 to derive a primal-dual convergence result for the Lasso problem (17). We find that the g_i^* are B-Lipschitz after applying the Lipschitzing trick to every g_i, and hence the total number of iterations needed on the Lasso problem to get a duality gap of $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \le \epsilon$ is*

$$T \ge \max\left\{0, n\log\frac{\beta\epsilon_D^{(0)}}{2B^2R^2n}\right\} + n + \frac{20n^2B^2R^2}{\beta\epsilon}$$

Remark 4. *We specify the different parameters of Corollary 10 for least squares loss as well as the logistic regression loss (defined in Table 1). Both are 1-smooth (f^* is 1-strongly convex) and we have $\beta := 1$. The initial suboptimality $\epsilon_D^{(0)}$ can be upper bounded by $\frac{1}{2}\|\mathbf{b}\|_2^2$ for the former and by $m\log(2)$ for the latter. For B we choose $\frac{1}{\lambda}f(\mathbf{0})$.*

Coordinate Descent on Elastic Net Regularized Problems. In Section 5.2.3 we discussed how the Elastic Net problem in (19) can be mapped to our setup. In the first scenario (row 2, Table 1) we note that the resulting problem is partially separable and an instance of (SA).
In the second scenario we map (19) to (B) (row 3, Table 1). Assuming that the loss function ℓ is separable, this problem is an instance of (SB). The convergence guarantees when applying Algorithm 1 on the primal or on the dual are summarized in Corollary 11.

Corollary 11. *Consider Algorithm 1 for an Elastic Net regularized problem (19), running on either the primal or the dual formulation. Then, to obtain a duality gap of $\mathbb{E}[G(\boldsymbol{\alpha}^{(T)})] \le \epsilon$, it suffices to have a total of*

$$T \ge (n + \tfrac{nR^2}{\lambda\eta\zeta})\log([n + \tfrac{nR^2}{\lambda\eta\zeta}]\tfrac{\epsilon_D^{(0)}}{\epsilon})$$

iterations for coordinate descent on the primal (19) and

$$T \ge (d + \tfrac{dP^2}{\lambda\eta\zeta})\log([d + \tfrac{dP^2}{\lambda\eta\zeta}]\tfrac{\epsilon_D^{(0)}}{\epsilon})$$

for coordinate descent on the dual of (19).

According to Corollary 11, the convergence rate is comparable for both scenarios. The constants however depend on the data matrix A – for $d \gg n$ the primal version is beneficial, whereas for $n \gg d$ the dual version is leading.

7. Numerical Experiments

Here we illustrate the usefulness of our framework by showcasing it for two important applications, each one showing two algorithm examples for optimizing (A).
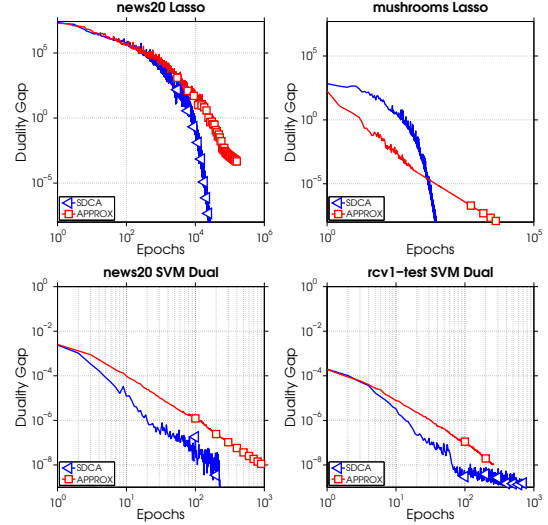

Figure 2. Comparison of CD with its accelerated variant – AP-PROX on Lasso and SVM problems.

Lasso. The top row of Figure 2 shows the primal-dual convergence of Algorithm 1 (CD) as well as the accelerated variant of CD (APPROX, Fercoq & Richtárik (2015)), both applied to the Lasso problem (A). We have applies the Lipschitzing trick as described in Section 5.1. This makes sure that $\mathbf{w}(\boldsymbol{\alpha})$ will be always feasible for the modified dual (B), and hence the duality gap can be evaluated.

SVM. It was shown in (Shalev-Shwartz & Zhang, 2013) that if CD (SDCA) is run on the dual SVM formulation, and we consider an "average" solution (over last few iterates), then the duality gap evaluated at averaged iterates has a sub-linear convergence rate $\mathcal{O}(1/t)$. As a consequence of Theorem 4, we have that the APPROX algorithm (Fercoq & Richtárik, 2015) will provide the same sub-linear convergence in duality gap, but holding for the iterates themselves, not only for an average. On the bottom row of Figure 2 we compare CD with its accelerated variant on two benchmark datasets.[5] We have chosen $\lambda = 1/n$.

8. Conclusions

We have presented a general framework allowing to equip existing optimization algorithms with primal-dual certificates. For future research, it will be interesting to study more applications and algorithms fitting into the studied problem structure, including more cases of structured sparsity, and generalizations to matrix problems.

---

[5] Available from csie.ntu.edu.tw/~cjlin/libsvmtools/datasets.

References

Bach, Francis. Duality Between Subgradient and Conditional Gradient Methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

Bach, Francis, Jenatton, Rodolphe, Mairal, Julien, and Obozinski, Guillaume. Optimization with Sparsity-Inducing Penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

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

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Borwein, J M and Zhu, Q. *Techniques of Variational Analysis and Nonlinear Optimization*. Canadian Mathematical Society Books in Math, Springer, 2005.

Boyd, Stephen P and Vandenberghe, Lieven. *Convex optimization*. Cambridge University Press, 2004.

Chambolle, Antonin and Pock, Thomas. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2010.

Fercoq, Olivier and Richtárik, Peter. Accelerated, Parallel, and Proximal Coordinate Descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Forte, Simone. Distributed Optimization for Non-Strongly Convex Regularizers. Master's thesis, ETH Zürich, 2015.

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

Goldstein, Tom, Li, Min, and Yuan, Xiaoming. Adaptive Primal-Dual Splitting Methods for Statistical Learning and Image Processing. In *NIPS 2015 - Advances in Neural Information Processing Systems 28*, pp. 2080–2088, 2015.

Hsieh, Cho-Jui, Chang, Kai-Wei, Lin, Chih-Jen, Keerthi, S Sathiya, and Sundararajan, Sellamanickam. A dual coordinate descent method for large-scale linear SVM. In *ICML 2008 - Proceedings of the 25th International Conference on Machine Learning*, pp. 408–415, 2008.

Hush, Don, Kelly, Patrick, Scovel, Clint, and Steinwart, Ingo. Qp algorithms with guaranteed accuracy and run time for support vector machines. *The Journal of Machine Learning Research*, 7:733–769, 2006.

Jaggi, Martin. An Equivalence between the Lasso and Support Vector Machines. In *Regularization, Optimization, Kernels, and Support Vector Machines*, pp. 1–26. Chapman and Hall/CRC, 2014.

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

Lacoste-Julien, Simon and Jaggi, Martin. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *NIPS 2015 - Advances in Neural Information Processing Systems 28*, 2015.

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

Lin, Qihang, Lu, Zhaosong, and Xiao, Lin. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. *arXiv:1407.1296*, 2014.

Ma, Chenxin, Smith, Virginia, Jaggi, Martin, Jordan, Michael I, Richtárik, Peter, and Takáč, Martin. Adding vs. averaging in distributed primal-dual optimization. *In ICML 2015 - 32th International Conference on Machine Learning*, 2015a.

Ma, Chenxin, Tappenden, Rachael, and Takáč, Martin. Linear convergence of the randomized feasible descent method under the weak strong convexity assumption. *arXiv:1506.02530*, 2015b.

Mairal, Julien. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, ENS Cachan, 2010.

Necoara, Ion. Linear convergence of first order methods under weak nondegeneracy assumptions for convex programming. *arXiv:1504.06298*, 2015.

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

Nesterov, Yurii. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

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

Qu, Zheng, Richtárik, Peter, and Zhang, Tong. Randomized Dual Coordinate Ascent with Arbitrary Sampling. *arXiv*, 2014.

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

Rockafellar, R Tyrrell. *Convex analysis*. Princeton University Press, 1997.

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

Shalev-Shwartz, Shai. SDCA without duality. *arXiv:1502.06177*, 2015.

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

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

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

Smith, Virginia, Forte, Simone, Jordan, Michael I, and Jaggi, Martin. L1-Regularized Distributed Optimization: A Communication-Efficient Primal-Dual Framework. *arXiv*, 2015.

Takáč, Martin, Bijral, Avleen, Richtárik, Peter, and Srebro, Nathan. Mini-batch primal and dual methods for SVMs. In *In ICML 2013 - 30th International Conference on Machine Learning*, 2013.

Takáč, Martin, Richtárik, Peter, and Srebro, Nathan. Distributed mini-batch SDCA. *arXiv:1507.08322*, 2015.

Tran-Dinh, Quoc and Cevher, Volkan. Constrained convex minimization via model-based excessive gap. In *NIPS 2014 - Advances in Neural Information Processing Systems 27*, 2014.

Tran-Dinh, Quoc and Cevher, Volkan. Splitting the Smoothed Primal-Dual Gap: Optimal Alternating Direction Methods. *arXiv*, 2015.

Wang, Po-Wei and Lin, Chih-Jen. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Yuan, Guo-Xun, Chang, Kai-Wei, Hsieh, Cho-Jui, and Lin, Chih-Jen. A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *Journal of Machine Learning Research*, 11:3183–3234, 2010.

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

Yurtsever, Alp, Dinh, Quoc Tran, and Cevher, Volkan. A Universal Primal-Dual Convex Optimization Framework. In *NIPS 2015 - Advances in Neural Information Processing Systems 28*, pp. 3132–3140, 2015.

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

Appendix

A. Basic Definitions

Definition 2 (L-Lipschitz Continuity). A function $h : \mathbb{R}^d \to \mathbb{R}$ is L-*Lipschitz continuous* w.r.t. a norm $\|.\|$ if $\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we have

$$|h(\mathbf{a}) - h(\mathbf{b})| \leq L\|\mathbf{a} - \mathbf{b}\| \,. \tag{20}$$

Definition' 1 (B-Bounded Support). *A function $h : \mathbb{R}^d \to \mathbb{R}$ has B-bounded support if its effective domain is bounded by B w.r.t. a norm $\|.\|$, i.e.,*

$$h(\mathbf{u}) < +\infty \; \Rightarrow \; \|\mathbf{u}\| \leq L \,. \tag{21}$$

Definition 3. The δ-level set of a function $h : \mathbb{R}^d \to \mathbb{R}$ is defined as $\mathcal{L}(\delta) := \{\mathbf{x} : h(\mathbf{x}) \leq \delta\}$.

Definition 4 (L-Smoothness). A function $h : \mathbb{R}^d \to \mathbb{R}$ is called L-*smooth* w.r.t. a norm $\|.\|$, for $L > 0$, if it is differentiable and its derivative is L-Lipschitz continuous w.r.t. $\|.\|$, or equivalently

$$h(\mathbf{u}) \leq h(\mathbf{w}) + \langle \nabla h(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{w}\|^2 \qquad \forall \mathbf{u}, \mathbf{w} \in \mathbb{R}^d \,. \tag{22}$$

Definition 5 (μ-Strong Convexity). A function $h : \mathbb{R}^d \to \mathbb{R}$ is called μ-*strongly convex* w.r.t. a norm $\|.\|$, for $\mu \geq 0$, if

$$h(\mathbf{u}) \geq h(\mathbf{w}) + \langle \nabla h(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle + \frac{\mu}{2}\|\mathbf{u} - \mathbf{w}\|^2 \qquad \forall \mathbf{u}, \mathbf{w} \in \mathbb{R}^d \,. \tag{23}$$

And analogously if the same holds for all subgradients, in the case of a general closed convex function h.

B. Convex Conjugates

We recall some basic properties of convex conjugates, which we use in the paper.

The convex conjugate of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f^*(\mathbf{v}) := \sup_{\mathbf{u} \in \mathbb{R}^d} \mathbf{v}^\top \mathbf{u} - f(\mathbf{u}) \,. \tag{24}$$

Some useful properties, see (Boyd & Vandenberghe, 2004, Section 3.3.2):

- Double conjugate: $(f^*)^* = f$ if f is closed and convex.

- Value Scaling: (for $\alpha > 0$) $\quad f(\mathbf{v}) = \alpha g(\mathbf{v}) \quad \Rightarrow \quad f^*(\mathbf{w}) = \alpha g^*(\mathbf{w}/\alpha)$.

- Argument Scaling: (for $\alpha \neq 0$) $\quad f(\mathbf{v}) = g(\alpha \mathbf{v}) \quad \Rightarrow \quad f^*(\mathbf{w}) = g^*(\mathbf{w}/\alpha)$.

- Conjugate of a separable sum: $\quad f(\mathbf{v}) = \sum_i \phi_i(v_i) \quad \Rightarrow \quad f^*(\mathbf{w}) = \sum_i \phi_i^*(w_i)$.

Lemma' 5 (Duality between Lipschitzness and L-Bounded Support. A generalization of (Rockafellar, 1997, Corollary 13.3.3)). *Given a proper convex function g, it holds that g has L-bounded support w.r.t. the norm $\|.\|$ if and only if g^* is L-Lipschitz w.r.t. the dual norm $\|.\|_*$.*

Lemma 12 (Duality between Smoothness and Strong Convexity, (Kakade et al., 2009, Theorem 6)). *Given a closed convex function f, it holds that f is μ-strongly convex w.r.t. the norm $\|.\|$ if and only if f^* is $(1/\mu)$-smooth w.r.t. the dual norm $\|.\|_*$.*

Lemma 13 (Conjugates of Indicator Functions and Norms).

i) *The conjugate of the indicator function $\iota_{\mathcal{C}}$ of a set $\mathcal{C} \subset \mathbb{R}^n$ (not necessarily convex) is the support function of the set $\mathcal{C}$, that is*

$$\iota_{\mathcal{C}}^*(\mathbf{x}) = \sup_{\mathbf{s} \in \mathcal{C}} \langle \mathbf{s}, \mathbf{x} \rangle$$

ii) *The conjugate of a norm is the indicator function of the unit ball of the dual norm.*

Proof. (Boyd & Vandenberghe, 2004, Examples 3.24 and 3.26) □

C. Primal-Dual Relationship

The relation of the primal and dual problems (A) and (B) is a special case of the concept of Fenchel Duality. Using the combination with the linear map A as in our case, the relationship is called *Fenchel-Rockafellar Duality*, see, e.g., (Borwein & Zhu, 2005, Theorem 4.4.2) or (Bauschke & Combettes, 2011, Proposition 15.18).

For completeness, we here illustrate this correspondence with a self-contained derivation of the duality.

Starting with the formulation (A), we introduce a helper variable $\mathbf{v} \in \mathbb{R}^d$. The optimization problem (A) becomes:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad f(\mathbf{v}) + g(\boldsymbol{\alpha}) \quad \text{such that } \mathbf{v} = A\boldsymbol{\alpha} . \tag{25}$$

Introducing dual variables $\mathbf{w} = [w_1, \dots, w_d]$, the Lagrangian is given by:

$$L(\boldsymbol{\alpha}, \mathbf{v}; \mathbf{w}) := f(\mathbf{v}) + g(\boldsymbol{\alpha}) + \mathbf{w}^\top (A\boldsymbol{\alpha} - \mathbf{v}) .$$

The dual problem follows by taking the infimum with respect to $\boldsymbol{\alpha}$ and $\mathbf{v}$:

$$\begin{aligned}
\inf_{\boldsymbol{\alpha}, \mathbf{v}} L(\mathbf{w}, \boldsymbol{\alpha}, \mathbf{v}) &= \inf_{\mathbf{v}} \left\{ f(\mathbf{v}) - \mathbf{w}^\top \mathbf{v} \right\} + \inf_{\boldsymbol{\alpha}} \left\{ g(\boldsymbol{\alpha}) + \mathbf{w}^\top A\boldsymbol{\alpha} \right\} \\
&= -\sup_{\mathbf{v}} \left\{ \mathbf{w}^\top \mathbf{v} - f(\mathbf{v}) \right\} - \sup_{\boldsymbol{\alpha}} \left\{ (-\mathbf{w}^\top A)\boldsymbol{\alpha} - g(\boldsymbol{\alpha}) \right\} \\
&= -f^*(\mathbf{w}) - g^*(-A^\top \mathbf{w}) .
\end{aligned} \tag{26}$$

We change signs and turn the maximization of the dual problem (26) into a minimization and thus we arrive at the dual formulation (B) as claimed:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \left[\mathcal{P}(\mathbf{w}) := f^*(\mathbf{w}) + g^*(-A^\top \mathbf{w}) \right] .$$

D. Proof of Lemma 1

The proof is partially motivated by proofs in (Ma et al., 2015a; Shalev-Shwartz & Zhang, 2013) but come with some crucial unique steps and tricks.

Proof of Lemma 1. We have

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) &= \mathcal{D}(\boldsymbol{\alpha}) - \min_{\delta \boldsymbol{\alpha}} \mathcal{D}(\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha}) \\
&= \max_{\delta \boldsymbol{\alpha}} \Big(g(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha}) + f(A\boldsymbol{\alpha}) - f(A(\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha})) \Big) \\
&= \max_{s \in [0,1]} \Big(g(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) - f(A(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha}))) \Big) \\
&\geq g(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) - f(A(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha}))) .
\end{aligned} \tag{27}$$

Here we have chosen $\Delta \boldsymbol{\alpha} := s(\mathbf{u} - \boldsymbol{\alpha})$ with $\mathbf{u}$ defined as in (6) and some $s \in [0, 1]$. Note that for $\mathbf{u}$ 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, while for $\mu = 0$ this follows from the bounded support assumption on g. (For the duality of Lipschitzness and bounded support, see our Lemma 5).

Now, we can use the fact, that function $f : \mathbb{R}^d \to \mathbb{R}$ has Lipschitz continuous gradient w.r.t. some norm $\|.\|_f$, with constant $1/\beta$, to obtain

$$\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) \overset{(27)}{\geq} g(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) - f(A\boldsymbol{\alpha}) - \langle \nabla f(A\boldsymbol{\alpha}), As(\mathbf{u} - \boldsymbol{\alpha}) \rangle - \frac{1}{2\beta} \|As(\mathbf{u} - \boldsymbol{\alpha})\|_f^2$$

$$= g(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha} + s(\mathbf{u} - \boldsymbol{\alpha})) - \langle \nabla f(A\boldsymbol{\alpha}), As(\mathbf{u} - \boldsymbol{\alpha}) \rangle - \frac{1}{2\beta} \|As(\mathbf{u} - \boldsymbol{\alpha})\|_f^2 . \tag{28}$$

Now, we will use a strong convexity property of a function g w.r.t. $\|.\|_g$ to obtain

$$\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) \overset{(28)}{\geq} s \underbrace{\Big(g(\boldsymbol{\alpha}) - g(\mathbf{u}) - \langle \nabla f(A\boldsymbol{\alpha}), A(\mathbf{u} - \boldsymbol{\alpha}) \rangle \Big)}_{\Lambda} + \frac{s^2}{2} \Big(\frac{\mu(1 - s)}{s} \|\mathbf{u} - \boldsymbol{\alpha}\|_g^2 - \frac{1}{\beta} \|A(\mathbf{u} - \boldsymbol{\alpha})\|_f^2 \Big) . \tag{29}$$

Now let us examine the relation of the equation above with duality gap. Therefore we use the definition of the optimization problems (A) and (B), and the definition of the convex conjugates to write the duality gap as:

$$
\begin{aligned}
G(\boldsymbol{\alpha}) = \mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - (-\mathcal{D}(\boldsymbol{\alpha})) &= g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})) + g(\boldsymbol{\alpha}) + f^*(\mathbf{w}(\boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) \\
&= g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})) + g(\boldsymbol{\alpha}) + f^*(\nabla f(A\boldsymbol{\alpha})) + f(A\boldsymbol{\alpha}) \\
&= g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})) + g(\boldsymbol{\alpha}) + \langle \mathbf{w}(\boldsymbol{\alpha}), A\boldsymbol{\alpha} \rangle .
\end{aligned}
\tag{30}
$$

where we have used the mapping $\mathbf{w}(\boldsymbol{\alpha}) = \nabla f(A\boldsymbol{\alpha})$.

Now, let us analyze the expression Λ from (29). We have

$$
\Lambda = g(\boldsymbol{\alpha}) + \langle \mathbf{w}(\boldsymbol{\alpha}), A\boldsymbol{\alpha} \rangle - g(\mathbf{u}) - \langle \mathbf{w}(\boldsymbol{\alpha}), A\mathbf{u} \rangle .
\tag{31}
$$

Now, using the convex conjugate maximal property and (6) we have

$$
g(\mathbf{u}) = \langle \mathbf{u}, -A^\top \mathbf{w}(\boldsymbol{\alpha}) \rangle - g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})).
\tag{32}
$$

Plugging (32), (30) and (31) into (29) gives us

$$
\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) \geq sG(\boldsymbol{\alpha}) + \frac{s^2}{2}\left(\frac{\mu(1-s)}{s}\|\mathbf{u}-\boldsymbol{\alpha}\|_g^2 - \frac{1}{\beta}\|A(\mathbf{u}-\boldsymbol{\alpha})\|_f^2 \right).
$$

and (5) follows. $\qquad\square$

E. Proof of Theorem 2

Proof. Let us upper bound the second term in (5) as given in the main Lemma 1. Recall the definition of the data complexity parameter $\sigma := \left(\max_{\boldsymbol{\alpha}\neq 0} \|A\boldsymbol{\alpha}\|_f / \|\boldsymbol{\alpha}\|_g \right)^2$. We have

$$
\frac{\mu(1-s)}{s}\|\mathbf{u}-\boldsymbol{\alpha}\|_g^2 - \frac{1}{\beta}\|A(\mathbf{u}-\boldsymbol{\alpha})\|_f^2 \geq \left[\frac{\mu(1-s)}{s} - \frac{\sigma}{\beta} \right] \|\mathbf{u}-\boldsymbol{\alpha}\|_g^2.
$$

Now, if we choose $s = \frac{\mu}{\frac{\sigma}{\beta}+\mu}$ then this term vanishes, and therefore

$$
\frac{\mu}{\mu + \frac{\sigma}{\beta}} \mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \overset{(5)}{\leq} \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star)] \overset{(7)}{\leq} (1-C)^t D.
$$

After multiplying the equation above by $\frac{\frac{\sigma}{\beta}+\mu}{\mu}$ and requiring RHS to be $\leq \epsilon$ we will get

$$
\frac{\frac{\sigma}{\beta} + \mu}{\mu}(1-C)^t D \leq \epsilon,
$$

$$
t \geq \frac{\log \frac{\mu\epsilon}{D(\frac{\sigma}{\beta}+\mu)}}{\log(1-C)}
$$

and our claimed convergence bound (8) follows. $\qquad\square$

F. Proof of Theorem 3

Proof. From Lemma 1 and the definition of σ we have that

$$
s\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \overset{(5)}{\leq} \mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star) + \frac{s^2}{2\beta}\|A(\mathbf{u}-\boldsymbol{\alpha})\|_f^2 \right] \overset{(7)}{\leq} (1-C)^t D + \frac{\sigma}{\beta}\frac{s^2}{2}\mathbb{E}[\|\mathbf{u}-\boldsymbol{\alpha}\|_g^2] .
\tag{33}
$$

Now using Lemma 5, because g^* is L-Lipschitz w.r.t. the norm $\|.\|_*$, we have that g is L-bounded w.r.t. the norm $\|.\| = \|.\|_g = \|.\|_f$ and hence for $\boldsymbol{\alpha} \in \text{dom}(g)$ we obtain $\|\boldsymbol{\alpha}\| < L$. From the standard characterization of Lipschitzness as by bounded subgradient norm (see e.g. (Shalev-Shwartz, 2011, Lemma 2.6) or other references) we have for any $\mathbf{u} \in \partial g^*(\mathbf{x})$ that $\|\mathbf{u}\| \leq L$, and hence for any $\mathbf{u}, \boldsymbol{\alpha}$ we must have $\|\mathbf{u}-\boldsymbol{\alpha}\|^2 \leq 2L^2$ by the triangle inequality.

Therefore we conclude that

$$
\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \overset{(33)}{\leq} \frac{1}{s}(1-C)^t D + s\frac{\sigma}{\beta}L^2.
\tag{34}
$$

Now, let us choose $\bar{s} = \min\{1, \frac{\epsilon\beta}{2\sigma L^2}\}$. To have the RHS of (34) $\leq \frac{\epsilon}{2}$ it is enough to choose

$$t \geq T = \frac{\log(\bar{s}\frac{\epsilon}{2D})}{(1-C)}$$

and the claimed convergence bound (9) follows. $\qquad\square$

G. Proof of Theorem 4

Proof. Using Lemma 1 and the bound $\|\mathbf{u} - \boldsymbol{\alpha}\|^2 \leq 2L^2$ derived in the previous section, we have that

$$\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \overset{(5)}{\leq} \frac{1}{s}\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star)] + \frac{s}{2}\frac{\sigma}{\beta}L^2 \overset{(10)}{\leq} \frac{1}{s}\frac{C}{D(t)} + \frac{s}{2}\frac{\sigma}{\beta}L^2. \tag{35}$$

Now, by choosing $s = \sqrt{\frac{C}{D(t)}\frac{2\beta}{\sigma L^2}} \overset{(11)}{\in} [0,1]$ we obtain that

$$\mathbb{E}[G(\boldsymbol{\alpha}^{(t)})] \overset{(35)}{\leq} \sqrt{\frac{2C\sigma L^2}{\beta D(t)}}. \tag{36}$$

We see that the assumption (11) guarantees that the RHS of the above inequality becomes $\leq \epsilon$, as claimed. $\qquad\square$

We note that in the special case of constrained optimization, motivated by the Frank-Wolfe algorithm, (Lacoste-Julien & Jaggi, 2015, Theorem 2) has shown another algorithm-independent bound for the convergence in duality gap, also requiring order of $1/\epsilon^2$ steps to reach accuracy ϵ. On the other hand, for the algorithm-specific result by (Bach, 2015, Prop 4.2) for the case of an extended Frank-Wolfe algorithm, $1/\epsilon$ steps are shown to be sufficient.

H. Proof of Lemma 5

Lemma 5 generalizes the result of (Rockafellar, 1997, Corollary 13.3.3) from the L_2-norm to a general norm $\|.\|$. It can be equivalently formulated as follows:

Lemma 14. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper convex function. In order that $\mathrm{dom}(f^*)$ be L-bounded w.r.t $\|.\|_*$ (a real number $L \geq 0$), it is necessary and sufficient that f be finite everywhere and that*

$$|f(\mathbf{z}) - f(\mathbf{x})| \leq L\|\mathbf{z} - \mathbf{x}\| \quad \forall \mathbf{x}, \mathbf{z}$$

Proof. We follow the line of proof in (Rockafellar, 1997).

It can be assumed that f is closed, as f and its closure $cl(f)$ have the same conjugate, and the Lipschitz condition is satisfied by f if and only if it is satisfied by $cl(f)$. Note that from (Rockafellar, 1997, Theorem 13.3) we know

$$f0^+(\mathbf{y}) := \sup_{\mathbf{x} \in \mathrm{dom}(f)} f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) \tag{37}$$

is the support function of $\mathrm{dom}(f^*)$. Further, from (Rockafellar, 1997, Theorem 10.5) we know: $\mathrm{dom}(f^*)$ is bounded if and only if $f0^+$ is finite everywhere. In more detail, the Lipschitz condition on f, choosing $\mathbf{z} = \mathbf{x} + \mathbf{y}$, is equivalent to having

$$f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) \leq L\|\mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}$$

and by definition (37) of $f0^+$ this is equivalent to

$$f0^+(\mathbf{y}) \leq L\|\mathbf{y}\|, \quad \forall \mathbf{y}$$

Note that $g(\mathbf{y}) := L\|\mathbf{y}\|$ is a finite positively homogenous convex function, we know by (Rockafellar, 1997, Corollary 13.2.2) that it is the support function of a non-empty bounded convex set. Call this set $\mathcal{S}$. We will later show that $\mathcal{S} = L\mathcal{B}$ where $\mathcal{B}$ is the $\|.\|_*$-norm ball. Hence $f0^+(\mathbf{y}) \leq g$ means $cl(\mathrm{dom}\,f^*) \subset \mathcal{S}$, as $f0^+$ is the support function of $\mathrm{dom}(f^*)$. And this shows that the Lipschitz condition holds for some L if and only if $\mathbf{x}^\star \in \mathcal{S} \; \forall \mathbf{x}^\star \in \mathrm{dom}(f^*)$ and hence $\|\mathbf{x}^\star\|_* \leq L$ and f^* is L-bounded for every $\mathbf{x}^\star \in \mathrm{dom}(f^*)$.

It remains to show that $g = L\|\mathbf{y}\|$ is the support function of $\mathcal{S} = L\mathcal{B}$: The support function of a set $\mathcal{C}$ is defined as

$$\delta_{\mathcal{C}}(\mathbf{y}) := \sup_{\mathbf{x} \in \mathcal{C}}\langle \mathbf{y}, \mathbf{x}\rangle.$$

By the definition of the dual norm we find

$$L\|\mathbf{y}\| = L \max_{\mathbf{x}:\|\mathbf{x}\|_* \le 1} \langle \mathbf{y}, \mathbf{x} \rangle = \max_{\mathbf{x}:\|\mathbf{x}\|_* \le L} \langle \mathbf{y}, \mathbf{x} \rangle$$

and hence g is the support function of the set $\mathcal{S} := \{\mathbf{x} : \|\mathbf{x}\|_* \le L\} = L\mathcal{B}$. $\qquad\square$

I. Duality Gaps

I.1. Duality Gap for Lasso (Using Lipschitzing)

Consider the Lasso problem given in (17), which we directly map to our primal optimization problem (A). We use the L_1-norm ball of radius B to apply the Lipschitzing trick, i.e. $\mathcal{B} := \{\mathbf{x} : \|\mathbf{x}\|_1 \le B\}$ and modify the L_1 norm term $g(\boldsymbol{\alpha}) := \lambda \|\boldsymbol{\alpha}\|_1$ as suggested in (12). The convex conjugate hence becomes:

$$\bar{g}^*(\mathbf{u}) = B \left[\|\mathbf{u}\|_\infty - \lambda \right]_+ \tag{38}$$

and using the optimality condition (1a) the gap follows immediately as

$$\begin{aligned}
\bar{G}(\boldsymbol{\alpha}) &= f(A\boldsymbol{\alpha}) + f^*(\mathbf{w}(\boldsymbol{\alpha})) + B \left[\|A^\top \mathbf{w}(\boldsymbol{\alpha})\|_\infty - \lambda \right]_+ + \lambda \|\boldsymbol{\alpha}\|_1 \\
&= \langle \nabla f(A\boldsymbol{\alpha}), A\boldsymbol{\alpha} \rangle + B \left[\|A^\top \mathbf{w}(\boldsymbol{\alpha})\|_\infty - \lambda \right]_+ + \lambda \|\boldsymbol{\alpha}\|_1
\end{aligned}$$

where we used the first-order optimality $f^*(\mathbf{w}(\boldsymbol{\alpha})) = \langle \mathbf{w}(\boldsymbol{\alpha}), A\boldsymbol{\alpha} \rangle - f(A\boldsymbol{\alpha})$ by definition of $\mathbf{w}(\boldsymbol{\alpha})$.

Proof. From (15) we know

$$\bar{g}^*(\mathbf{u}) = \begin{cases} 0 & \|\mathbf{u}\|_\infty \le \lambda \\ \max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_1 \le B} \mathbf{u}^\top \boldsymbol{\alpha} - \lambda \|\boldsymbol{\alpha}\|_1 & \text{else} \end{cases}. \tag{39}$$

It remains to consider the non zero part, where $\|\mathbf{u}\|_\infty > \lambda$. Lets consider the optimization

$$\max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_1 \le B} \mathbf{u}^\top \boldsymbol{\alpha} - \lambda \|\boldsymbol{\alpha}\|_1.$$

Assume the largest element of $\mathbf{u}$ is at index i and we know $[\mathbf{u}]_i > \lambda$, hence the maximizing $\boldsymbol{\alpha}$ in $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_1 \le B\}$ is the vector with a single entry of value B at position i, and (38) follows. $\qquad\square$

I.2. Duality Gap for Elastic Net Regularized Problems

Consider the Elastic Net regularized problem formulation (19). The conjugate dual of the Elastic Net regularizer is given in Lemma 20. We first map (19) to (A), as suggested in row 2 of Table 1. Then its dual counterpart (B) is given by:

$$\mathcal{P}(\mathbf{w}) = \ell^*(\mathbf{w}) + \frac{1}{2\eta\lambda} \sum_{i=1}^n \left(\left[|A_{:i}^\top \mathbf{w}| - (1-\eta)\lambda \right]_+ \right)^2.$$

In light of the optimality condition (1a) we now use the definition of the primal-dual correspondence $\mathbf{w}(\boldsymbol{\alpha}) = \nabla \ell(A\boldsymbol{\alpha})$, and hence obtain

$$G(\boldsymbol{\alpha}) = (A\boldsymbol{\alpha})^\top \nabla \ell(A\boldsymbol{\alpha}) + \frac{1}{2\eta\lambda} \sum_{i=1}^n \left(\left[|A_{:i}^\top \mathbf{w}| - (1-\eta)\lambda \right]_+ \right)^2 + \lambda \left(\frac{\eta}{2} \|\boldsymbol{\alpha}\|_2^2 + (1-\eta)\|\boldsymbol{\alpha}\|_1 \right).$$

Note that when alternatively mapping (19) to (B) (as suggested in row 3 of Table 1) instead of (A), the duality gap function will be similarly structured, but in that case the variable mapping $\mathbf{w}(\boldsymbol{\alpha})$ will be defined using the gradient of the Elastic Net penalty, instead of ℓ.

J. Convergence Results for Coordinate Descent

For the proof of Theorem 8 and 9 and we will need the following Lemma, which we will prove below in Section J.3. This lemma allows us to lower bound the expected per-step improvement in terms of $\mathcal{D}$ for coordinate descent algorithms on $\mathcal{D}$.

Lemma 15. *Consider problem formulation (SB) and (SA). Let f be $1/\beta$-smooth w.r.t. a norm $\|.\|$. Further, let g_i be μ-strongly convex with convexity parameter $\mu \ge 0 \ \forall i \in [n]$. For the case $\mu = 0$ we need the additional assumption of g_i*

having bounded support. Then for any iteration t and any $s \in [0, 1]$, it holds that

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \geq \frac{s}{n} G(\boldsymbol{\alpha}^{t-1}) - \frac{s^2 F^{(t)}}{2}, \tag{40}$$

where

$$F^{(t)} := \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\beta} \|A_{:i}\|^2 - \frac{(1-s)\mu}{s} \right) \mathbb{E}\left[(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 \right], \tag{41}$$

and $u_i^{(t-1)} \in \partial g_i^*(-A_{:i}^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}))$. The expectations are with respect to the random choice of coordinate i in step t.

J.1. Proof of Theorem 8

For the following, we again assume that the columns of the data matrix A are scaled such that $\|A_{:i}\| \leq R$ for all $i \in [n]$ and $\|A_{j:}\| \leq P$ for all $j \in [d]$ and some norm $\|.\|$.

Theorem' 8. *Consider Algorithm 1 applied to (SA). Assume f is a $1/\beta$-smooth function w.r.t. the norm $\|.\|$. Then, if g_i is μ-strongly convex for all i, it suffices to have a total number of iterations of*

$$T \geq \left(n + \frac{nR^2}{\mu\beta} \right) \log \left(\left[n + \frac{nR^2}{\mu\beta} \right] \frac{\epsilon_D^{(0)}}{\epsilon} \right)$$

to get $\mathbb{E}[G(\boldsymbol{\alpha}^{(T)})] \leq \epsilon$. Moreover, to obtain an expected duality gap of $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \leq \epsilon$ it suffices to have $T > T_0$ with

$$T_0 \geq \left(n + \frac{nR^2}{\mu\beta} \right) \log \left(\left[n + \frac{nR^2}{\mu\beta} \right] \frac{\epsilon_D^{(0)}}{(T - T_0)\epsilon} \right)$$

where $\epsilon_D^{(0)}$ is the initial suboptimality in $\mathcal{D}(\boldsymbol{\alpha})$.

Proof. The proof of this theorem is motivated by proofs in (Shalev-Shwartz & Zhang, 2013) but generalizing their result in (Shalev-Shwartz & Zhang, 2013, Theorem 5). To prove Theorem 8 we apply Lemma 15 with $s = \frac{\mu\beta}{R^2 + \mu\beta} \in [0, 1]$. This choice of s implies $F^{(t)} \leq 0 \ \forall t$ as defined in (41). Hence,

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \geq \frac{s}{n} G(\boldsymbol{\alpha}^{(t-1)}) = \frac{s}{n} \mathcal{P}(\mathbf{w}^{(t-1)}) + \mathcal{D}(\boldsymbol{\alpha}^{(t-1)}).$$

Here expectations are over the choice of coordinate i in step t, conditioned on the past state at $t - 1$. Let $\epsilon_D^{(t)}$ denote the suboptimality $\epsilon_D^{(t)} := D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^\star)$. As $\epsilon_D^{(t-1)} \leq P(\mathbf{w}^{(t-1)}) + D(\boldsymbol{\alpha}^{(t-1)})$ and $D(\boldsymbol{\alpha}^{(t-1)}) - D(\boldsymbol{\alpha}^{(t)}) = \epsilon_D^{(t-1)} - \epsilon_D^{(t)}$, we obtain

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \left(1 - \frac{s}{n} \right) \epsilon_D^{(t-1)} \leq \left(1 - \frac{s}{n} \right)^t \epsilon_D^{(0)} \leq \exp(-st/n)\epsilon_D^{(0)} \leq \exp\left(-\frac{\mu\beta t}{n(R^2 + \mu\beta)} \right) \epsilon_D^{(0)}.$$

To finally upper bound the expected suboptimality as $\mathbb{E}[\epsilon_D^{(t)}] \leq \epsilon_D$ we need

$$t \geq \left(\frac{nR^2}{\mu\beta} + n \right) \log \left(\frac{\epsilon_D^{(0)}}{\epsilon_D} \right).$$

And towards obtaining small duality gap as well, we observe that at all times

$$\mathbb{E}[G(\boldsymbol{\alpha}^{(t-1)})] \leq \frac{n}{s} \mathbb{E}[\epsilon_D^{(t-1)} - \epsilon_D^{(t)}] \leq \frac{n}{s} \mathbb{E}[\epsilon_D^{(t-1)}]. \tag{42}$$

Hence with $\epsilon \geq \frac{n}{s} \mathbb{E}[\epsilon_D^{(t)}]$ we must have a duality gap smaller than ϵ. Therefore we require

$$t \geq \left(\frac{nR^2}{\mu\beta} + n \right) \log \left(\left(n + \frac{nR^2}{\mu\beta} \right) \frac{\epsilon_D^0}{\epsilon} \right).$$

This proves the first part of Theorem 8 and the second part follows immediately if we sum (42) over $t = T_0, ..., T - 1$. $\quad\square$

Remark 5. *From Theorem 8 it follows that for $T = 2T_0$ and $T_0 \geq n + \frac{nR^2}{\mu\beta}$ we need*

$$T \geq 2 \left(n + \frac{nR^2}{\mu\beta} \right) \log \left(\frac{\epsilon_D^{(0)}}{\epsilon} \right)$$

Recovering SDCA as a Special Case.

Corollary 16. *We recover Theorem 5 of (Shalev-Shwartz & Zhang, 2013) as a special case of our Theorem 8. We consider their optimization objectives, namely*

$$\mathcal{P}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \Phi_i(\mathbf{x}_i^\top \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{43}$$

$$\mathcal{D}(\boldsymbol{\alpha}) := - \left[\frac{1}{n} \sum_{i=1}^{n} -\Phi_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right\|^2 \right]. \tag{44}$$

where $\mathbf{x}_i$ are the columns of the data matrix X. We assume that $\Phi_i^*(\boldsymbol{\alpha})$ is γ-strongly convex for $i \in [n]$. We scale the columns of X such that $\|\mathbf{x}_i\| \leq 1$. Hence, we find that

$$T_0 \geq \left(n + \frac{1}{\gamma\lambda} \right) \log \left(\left[n + \frac{1}{\gamma\lambda} \right] \frac{1}{\epsilon(T - T_0)} \right)$$

iterations are sufficient to obtain a duality gap of $\mathbb{E}[\mathcal{P}(\bar{\mathbf{w}}) - (-\mathcal{D}(\bar{\boldsymbol{\alpha}}))] \leq \epsilon$.

Proof. We consider (43) and (44) as a special case of the separable problems (SA) and (SB). We set $g_i(\alpha) := \frac{1}{n}\Phi_i^*(-\alpha)$ and $f^*(\mathbf{w}) := \frac{\lambda}{2}\|\mathbf{w}\|^2$. In this case $\mu := \frac{1}{n}\gamma$ and $\beta := \lambda$. Defining $A := \frac{1}{n}X$ and using the assumption $\|\mathbf{x}_i\| \leq 1$ we have $R := \frac{1}{n}$ and using $\Phi_i(0) \leq 1$ we have $\epsilon_D^{(0)} \leq 1$ and applying Theorem 8 to this setting concludes the proof. $\square$

J.2. Proof of Theorem 9

This theorem generalizes the results of (Shalev-Shwartz & Zhang, 2013, Theorem 2).

Theorem' 9. *Consider Algorithm 1 applied to (SA). Assume f is a $1/\beta$-smooth function w.r.t. the norm $\|.\|$. Then, if g_i^* is L-Lipschitz for all i, it suffices to have a total number of iterations of*

$$T \geq \max \left\{ 0, n \log \frac{\epsilon_D^{(0)}\beta}{2L^2R^2n} \right\} + n + \frac{20n^2L^2R^2}{\beta\epsilon}$$

to get $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \leq \epsilon$. Moreover, when $t \geq T_0$ with

$$T_0 = \max \left\{ 0, n \log \frac{\epsilon_D^{(0)}\beta}{2L^2R^2n} \right\} + \frac{16n^2L^2R^2}{\beta\epsilon}$$

we have the suboptimality bound of $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^\star)] \leq \epsilon/2$, where $\epsilon_D^{(0)}$ is the initial suboptimality.

We rely on the following Lemma:

Lemma 17. *Suppose that for all i, g_i^* is L-Lipschitz. Let $F^{(t)}$ be as defined in Lemma 15 (with $\mu = 0$) and assume $\|A_{:i}\| \leq R \ \ \forall i$. Then, $F^{(t)} \leq \frac{4L^2}{\beta}R^2 \ \ \forall t$.*

Proof. We apply the duality of Lipschitzness and bounded support (Lemma 5) to the case of univariate functions g_i. The assumption of g_i^* being L-Lipschitz therefore gives that g_i has L-bounded support, and so $|\alpha_i| \leq L$ for α_i as defined in (41).

Furthermore for u_i, by the equivalence of Lipschitzness and bounded subgradient (see e.g. (Shalev-Shwartz, 2011, Lemma 2.6)) we have $|u_i| \leq L$ and thus, $|\alpha_i - u_i|^2 \leq 4L^2$. Together with $\|A_{:i}\| \leq R$ the claimed bound on $F^{(t)}$ follows. $\square$

Remark 6. *In the case of Lasso we have $g_i(\alpha_i) = \lambda|\alpha_i|$. Using the Lipschitzing trick, we replace $g_i(\cdot)$ by*

$$\bar{g}_i(\alpha) = \begin{cases} \lambda|\alpha| & : \alpha \in [-B, B] \\ +\infty & : otherwise, \end{cases}$$

which has B-bounded support. Hence, its conjugate

$$\bar{g}^*(x) = \begin{cases} 0 & : x \in [-\lambda, \lambda] \\ B(|x| - \lambda) & : otherwise, \end{cases}$$

is B-Lipschitz and Lemma 5 and Lemma 17 apply with $L := B$.

Proof of Theorem 9. Now, to prove Theorem 9, let $F = \max_t F^{(t)}$ and recall that by Lemma 17 we can upper bound F by $\frac{4L^2}{\beta} R^2$.

Furthermore, our main Lemma 15 on the improvement per step tells us that

$$\mathbb{E}\big[\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})\big] \geq \frac{s}{n} G(\boldsymbol{\alpha}^{(t-1)}) - \frac{s^2}{2} F, \tag{45}$$

With $\epsilon_D^{(t)} = D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^\star) \leq G(\boldsymbol{\alpha}^{(t)})$ and $D(\boldsymbol{\alpha}^{(t)}) - D(\boldsymbol{\alpha}^{(t-1)}) = \epsilon_D^{(t-1)} - \epsilon_D^{(t)}$, this implies

$$
\begin{aligned}
\mathbb{E}\big[\epsilon_D^{(t-1)} - \epsilon_D^{(t)}\big] &\geq \frac{s}{n}\epsilon_D^{(t-1)} - \frac{s^2 F}{2} \\
\epsilon_D^{(t-1)} - \mathbb{E}[\epsilon_D^{(t)}] &\geq \frac{s}{n}\epsilon_D^{(t-1)} - \frac{s^2 F}{2} \\
\mathbb{E}[\epsilon_D^{(t)}] &\leq \Big(1 - \frac{s}{n}\Big)\epsilon_D^{(t-1)} + \frac{s^2 F}{2}
\end{aligned}
$$

Expectations again only being over the choice of i in steps t. We next show that this inequality can be used to bound the suboptimality as

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \frac{2Fn^2}{2n + t - t_0} \tag{46}$$

for $t \geq t_0 = \max\left\{0, n\log\left(\frac{2\epsilon_D^{(0)}}{Fn}\right)\right\}$. Indeed, let us choose $s := 1$. Then at $t = t_0$, we have

$$
\begin{aligned}
\mathbb{E}[\epsilon_D^{(t)}] &\leq \Big(1 - \frac{1}{n}\Big)^t \epsilon_D^{(0)} + \sum_{i=0}^{t-1}\Big(1 - \frac{1}{n}\Big)^i \frac{F}{2} \\
&\leq \Big(1 - \frac{1}{n}\Big)^t \epsilon_D^{(0)} + \frac{1}{1 - (1 - 1/n)}\frac{F}{2} \\
&\leq e^{-t/n}\epsilon_D^{(0)} + \frac{nF}{2} \\
&\leq Fn
\end{aligned}
$$

For $t > t_0$ we use an inductive argument. Suppose the claim holds for $t - 1$, therefore

$$
\begin{aligned}
\mathbb{E}[\epsilon_D^{(t)}] &\leq \Big(1 - \frac{s}{n}\Big)\mathbb{E}[\epsilon_D^{(t-1)}] + \frac{s^2 F}{2} \\
&\leq \Big(1 - \frac{s}{n}\Big)\frac{2Fn^2}{2n + (t-1) - t_0} + \frac{s^2 F}{2}
\end{aligned}
$$

choosing $s = \frac{2n}{2n + t - 1 - t_0} \in [0, 1]$ yields

$$
\begin{aligned}
\mathbb{E}[\epsilon_D^{(t)}] &\leq \Big(1 - \frac{2}{2n + t - 1 - t_0}\Big)\frac{2Fn^2}{2n + (t-1) - t_0} + \Big(\frac{2n}{2n + t - 1 - t_0}\Big)^2 \frac{F}{2} \\
&= \Big(1 - \frac{2}{2n + t - 1 - t_0}\Big)\frac{2Fn^2}{2n + (t-1) - t_0} + \Big(\frac{1}{2n + t - 1 - t_0}\Big)\frac{2Fn^2}{2n + t - 1 - t_0} \\
&= \Big(1 - \frac{1}{2n + t - 1 - t_0}\Big)\frac{2Fn^2}{2n + (t-1) - t_0} \\
&= \frac{2Fn^2}{(2n + t - 1 - t_0)}\frac{2n + t - 2 - t_0}{2n + t - 1 - t_0} \\
&\leq \frac{2Fn^2}{(2n + t - t_0)}
\end{aligned}
$$

This proves the bound (46) on the suboptimality. To get a result on the duality gap we sum (45) over the interval $t =$

$T_0 + 1, ..., T$ and obtain

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{(T)})] \geq \frac{s}{n}\mathbb{E}\left[\sum_{t=T_0+1}^{T} P(\mathbf{w}^{(t-1)}) + D(\boldsymbol{\alpha}^{(t-1)})\right] - (T - T_0)\frac{s^2}{2}F,$$

and rearranging terms we get

$$\mathbb{E}\left[\frac{1}{T - T_0}\sum_{t=T_0+1}^{T} P(\mathbf{w}^{(t-1)}) + D(\boldsymbol{\alpha}^{(t-1)})\right] \leq \frac{n}{s(T - T_0)}\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{(T)})] + \frac{sn}{2}F,$$

Now if we choose $\bar{\mathbf{w}}, \bar{\boldsymbol{\alpha}}$ to be the average vectors over $t \in \{T_0 + 1, T\}$, then the above implies

$$\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] = \mathbb{E}[P(\bar{\mathbf{w}}) + D(\bar{\boldsymbol{\alpha}})] \leq \frac{n}{s(T - T_0)}\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{(T)})] + \frac{sn}{2}F,$$

If $T \geq n + T_0$ and $T_0 \geq t_0$, we can set $s = n/(T - T_0)$ and combining this with (46) we obtain

$$
\begin{aligned}
\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] = \quad &\leq \quad \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{(T)})] + \frac{Fn^2}{2(T - T_0)} \\
&\leq \quad \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{\star})] + \frac{Fn^2}{2(T - T_0)} \\
&\leq \quad \frac{2Fn^2}{2n + t - t_0} + \frac{Fn^2}{2(T - T_0)}
\end{aligned}
$$

A sufficient condition to upper bound the duality gap by ϵ is that $T_0 \geq \frac{4Fn^2}{\epsilon} - 2n + t_0$ and $T \geq T_0 + \frac{Fn^2}{\epsilon}$ which also implies $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(T_0)}) - \mathcal{D}(\boldsymbol{\alpha}^{\star})] \leq \epsilon/2$. Since we further require $T_0 \geq t_0$ and $T - T_0 \geq n$, the overall number of required iterations has to satisfy

$$T_0 \geq \max\left\{t_0, \frac{4Fn^2}{\epsilon} - 2n + t_0\right\} \text{ and } T - T_0 \geq \max\left\{n, \frac{Fn^2}{\epsilon}\right\}$$

Using Lemma 17 we can bound the total number of required iterations to reach a duality gap of ϵ by

$$
\begin{aligned}
T \quad &\geq \quad T_0 + n + \frac{n^2}{\epsilon}\frac{4L^2R^2}{\beta} \\
&\geq \quad t_0 + \frac{4n^2}{\epsilon}\frac{4L^2R^2}{\beta} + n + \frac{n^2}{\epsilon}\frac{4L^2R^2}{\beta} \\
&\geq \quad \max\left\{0, n\log\left(\frac{\epsilon_D^{(0)}\beta}{2nR^2L^2}\right)\right\} + n + \frac{20n^2L^2R^2}{\beta\epsilon}
\end{aligned}
$$

which concludes the proof of Theorem 9. $\qquad\square$

Recovering SDCA as a Special Case.

Corollary 18. *We recover Theorem 2 of (Shalev-Shwartz & Zhang, 2013) as a special case of our Theorem 9. We consider the optimization objectives in (43) and (44). We assume that $\Phi_i(\boldsymbol{\alpha})$ is M-Lipschitz for $i \in [n]$ and $\Phi_i(0) \leq 1$. We scale the columns of X such that $\|\mathbf{x}_i\| \leq 1$. Hence, we find that*

$$T \geq T_0 + n + \frac{4M^2}{\lambda\epsilon}$$

$$\geq \max\{0, n\log(0.5n\lambda M^{-2})\} + n + \frac{20M^2}{\lambda\epsilon}$$

iterations are sufficient to obtain a duality gap of $\mathbb{E}[G(\bar{\boldsymbol{\alpha}})] \leq \epsilon$.

Proof. We consider (43) and (44) as a special case (SB) and (SA). We set $g_i(\alpha) := \frac{1}{n}\Phi_i^*(-\alpha)$ and $f^*(\mathbf{w}) := \frac{\lambda}{2}\|\mathbf{w}\|^2$. Hence, $g_i^*(\mathbf{w}^\top\mathbf{a}_i) = \frac{1}{n}\Phi_i(-n\mathbf{w}^\top\mathbf{a}_i)$ is M-Lipschitz and we set $L := M$ and $\beta := \lambda$. We further use Lemma (Shalev-Shwartz & Zhang, 2013, Lemma 20) with $\Phi_i(0) \leq 1$ to bound the primal suboptimality as $\epsilon_D^{(0)} \leq 1$. Finally, by the assumption $\|\mathbf{x}_i\| \leq 1$ and the definition $A := \frac{1}{n}X$ we have $R := \frac{1}{n}$ and applying Theorem 9 to this setting concludes the proof. $\qquad\square$

J.3. Proof of Lemma 15

Proof. The proof of Lemma 15 is motivated by the proof of (Shalev-Shwartz & Zhang, 2013, Lemma 19) but we adapt it to apply to a much more general setting. First note that the one step improvement in the dual objective can be written as

$$\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) = \sum_{i=1}^{n} g_i(\alpha_i^{(t-1)}) + f(A\boldsymbol{\alpha}^{(t-1)}) - \left[\sum_{i=1}^{n} g_i(\alpha_i^{(t)}) + f(A\boldsymbol{\alpha}^{(t)})\right]$$

Note that in a single step of SDCA $\boldsymbol{\alpha}^{(t-1)} \to \boldsymbol{\alpha}^{(t)}$ only one dual coordinate is changed. Without loss of generality we assume this coordinate to be i. Writing $\mathbf{v}(\boldsymbol{\alpha}) = A\boldsymbol{\alpha}$, we find

$$\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) = \underbrace{\left[g_i(\alpha_i^{(t-1)}) + f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)}))\right]}_{(\Gamma)} - \underbrace{\left[g_i(\alpha_i^{(t)}) + f(\mathbf{v}(\boldsymbol{\alpha}^{(t)}))\right]}_{(\Lambda)}.$$

In the following, let us denote the columns of the matrix A by $\mathbf{a}_i$ for $i \in [n]$ for reasons of readability. Then, by definition of the update we have for all $s \in [0, 1]$:

$$\begin{aligned}
(\Lambda) &= g_i(\alpha_i^{(t)}) + f(\mathbf{v}(\boldsymbol{\alpha}^{(t)})) \\
&= \min_{\Delta\alpha_i} \left[g_i(\alpha_i^{(t-1)} + \Delta\alpha_i) + f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)} + \mathbf{e}_i\Delta\alpha_i))\right] \\
&= \min_{\Delta\alpha_i} \left[g_i(\alpha_i^{(t-1)} + \Delta\alpha_i) + f\left(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)}) + \mathbf{a}_i\Delta\alpha_i\right)\right] \\
&\leq \left[g_i(\alpha_i^{(t-1)} + s(u_i^{(t-1)} - \alpha_i^{(t-1)})) + f\left(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)}) + \mathbf{a}_i s(u_i^{(t-1)} - \alpha_i^{(t-1)})\right)\right]
\end{aligned}$$

where we chose $\Delta\alpha_i = s(u_i^{(t-1)} - \alpha_i^{(t-1)})$ with $-u_i^{(t-1)} \in \partial g_i^*(\mathbf{a}_i^\top \mathbf{w}^{(t-1)})$ for $s \in [0, 1]$. As in Section D, in order for $\mathbf{u}$ to be well-defined, we again need the assumption on g to be strongly convex ($\mu > 0$), or to have bounded support. Using μ-strong convexity of g_i, namely

$$\begin{aligned}
g_i(\alpha_i^{(t-1)} + s(u_i^{(t-1)} - \alpha_i^{(t-1)})) &= g_i(su_i^{(t-1)} + (1-s)\alpha_i^{(t-1)}) \\
&\leq sg_i(u_i^{(t-1)}) + (1-s)g_i(\alpha_i^{(t-1)}) - \tfrac{\mu}{2}s(1-s)(u_i^{(t-1)} - \alpha_i^{(t-1)})^2,
\end{aligned}$$

and $\frac{1}{\beta}$-smoothness of f

$$f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)}) + s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i) \leq f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)})) + \langle \nabla f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)})), s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\rangle + \tfrac{1}{2\beta}\|s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\|^2$$

we find that

$$\begin{aligned}
(\Lambda) \leq \; &\left[sg_i(u_i^{(t-1)}) + (1-s)g_i(\alpha_i^{(t-1)}) - \tfrac{\mu}{2}s(1-s)(u_i^{(t-1)} - \alpha_i^{(t-1)})^2\right] \\
&+ \left[f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)})) + \langle \nabla f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)})), s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\rangle + \tfrac{1}{2\beta}\|s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\|^2\right].
\end{aligned}$$

We further note that from the optimality condition (1a) we have $\mathbf{w}(\boldsymbol{\alpha}) = \nabla f(\mathbf{v}(\boldsymbol{\alpha}))$ and rearranging terms yields:

$$\begin{aligned}
(\Lambda) \leq \; &sg_i(u_i^{(t-1)}) - sg_i(\alpha_i^{(t-1)}) - \tfrac{\mu}{2}s(1-s)(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 \\
&+ \underbrace{g_i(\alpha_i^{(t-1)}) + f(\mathbf{v}(\boldsymbol{\alpha}^{(t-1)}))}_{(\Gamma)} + s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}) + \tfrac{1}{2\beta}\|s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\|^2.
\end{aligned}$$

Using this inequality to bound $\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) = (\Gamma) - (\Lambda)$ yields:

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) \geq \; &-sg_i(u_i^{(t-1)}) + sg_i(\alpha_i^{(t-1)}) - su_i^{(t-1)}\mathbf{a}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}) + s\alpha_i^{(t-1)}\mathbf{a}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}) \\
&+ \tfrac{\mu}{2}s(1-s)(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 - \tfrac{1}{2\beta}\|s(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\|^2. \\
\overset{(i)}{\geq} \; &s\Big[g_i^*(-\mathbf{a}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)})) + g_i(\alpha_i^{(t-1)}) + \alpha_i^{(t-1)}\mathbf{a}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}) \\
&+ \tfrac{\mu}{2}(1-s)(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 - \tfrac{s}{2\beta}\|(u_i^{(t-1)} - \alpha_i^{(t-1)})\mathbf{a}_i\|^2\Big].
\end{aligned} \qquad (47)$$

Note that for (i) we used the optimality condition (2b) which translates to $u \in \partial g^*(-\mathbf{a}_i^\top \mathbf{w})$ and yields $g(u) = -u\mathbf{a}_i^\top \mathbf{w} - g^*(-\mathbf{a}_i^\top \mathbf{w})$. Similarly, by again exploiting the primal-dual optimality condition we have $f^*(\nabla f(\mathbf{v})) = \mathbf{v}^\top \nabla f(\mathbf{v}) - f(\mathbf{v})$ and hence we can write the duality gap as:

$$
\begin{aligned}
G(\boldsymbol{\alpha}) = \mathcal{P}(\mathbf{w}) - (-\mathcal{D}(\boldsymbol{\alpha})) &= \sum_{i=1}^{n} g_i^*(-\mathbf{a}_i^\top \mathbf{w}) + f^*(\mathbf{w}) - \left[-\sum_{i=1}^{n} g_i(\alpha_i) - f(A\boldsymbol{\alpha}) \right] \\
&= \sum_{i=1}^{n} \left[g_i^*(-\mathbf{a}_i^\top \mathbf{w}) + g_i(\alpha_i) \right] + f^*(\mathbf{w}) + f(A\boldsymbol{\alpha}) \\
&= \sum_{i=1}^{n} \left[g_i^*(-\mathbf{a}_i^\top \mathbf{w}) + g_i(\alpha_i) \right] + (A\boldsymbol{\alpha})^\top \mathbf{w} \\
&= \sum_{i=1}^{n} \left[g_i^*(-\mathbf{a}_i^\top \mathbf{w}) + g_i(\alpha_i) + \alpha_i \mathbf{a}_i^\top \mathbf{w} \right]
\end{aligned}
$$

using this we can write the expectation of (47) with respect to i as

$$
\mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) \right] \geq s\left(\frac{1}{n} G(\boldsymbol{\alpha}^{(t-1)}) \right) - \frac{s^2}{2} \underbrace{\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 \right] \left(\frac{1}{\beta} \|\mathbf{a}_i\|^2 - \frac{(1-s)\mu}{s} \right) \right]}_{F^{(t)}}.
$$

And we have obtained that

$$
\mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) \right] \geq \frac{s}{n} G(\boldsymbol{\alpha}^{(t-1)}) - \frac{s^2}{2} F^{(t)}.
$$

For the expectation being over the choice of coordinate i in step t. $\qquad\square$

Corollary 19. *We recover Lemma 19 of (Shalev-Shwartz & Zhang, 2013) as a special case of our Lemma 15. We consider their pair of primal and dual optimization objectives, (43) and (44) where $\mathbf{x}_i$ are the columns of the data matrix X. Assume that Φ_i^* is γ-strongly convex for $i \in [n]$, where we allow $\gamma = 0$. Then, for any t, any $s \in [0,1]$ and $-\hat{u}_i^{(t-1)} \in \partial \Phi_i(\mathbf{x}_i^\top \mathbf{w}(\boldsymbol{\alpha}^{(t-1)}))$ we have*

$$
\mathbb{E}[-\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - (-\mathcal{D}(\boldsymbol{\alpha}^{(t-1)}))] \tag{48}
$$

$$
\geq \frac{s}{n} \mathbb{E}[\mathcal{P}(\boldsymbol{\theta}^{t-1}) - (-\mathcal{D}(\boldsymbol{\alpha}^{t-1}))] - \left(\frac{s}{n} \right)^2 \frac{\hat{F}^{(t)}}{2\lambda},
$$

where

$$
\hat{F}^{(t)} := \frac{1}{n} \sum_{i=1}^{n} \left(\|\mathbf{x}_i\|^2 - \frac{\gamma(1-s)\lambda n}{s} \right) \tag{49}
$$

$$
\cdot \mathbb{E}\left[(\hat{u}_i^{(t-1)} - \alpha_i^{(t-1)})^2 \right],
$$

Proof. We set $g_i(\boldsymbol{\alpha}) := \frac{1}{n} \Phi_i^*(-\boldsymbol{\alpha})$ and $f^*(\mathbf{w}) := \frac{\lambda}{2} \|\mathbf{w}\|^2$. From the definition of strong convexity it immediately follows that $\mu = \frac{\gamma}{n}$ and $\beta = \lambda$. As our algorithm works for any data matrix A, we choose $A := \frac{1}{n} X$ and scale the input vectors $\mathbf{x}_i$, i.e. $\mathbf{a}_i = \frac{\mathbf{x}_i}{n}$ before we feed $\mathbf{a}_i$ into the algorithm. To conclude the proof we apply Lemma 15 to this setting and observe that

$$
g_i^*(\mathbf{w}^\top \mathbf{a}_i) = \frac{1}{n} \Phi_i(-n\mathbf{w}^\top \mathbf{a}_i) = \frac{1}{n} \Phi_i(-\mathbf{w}^\top \mathbf{x}_i). \tag{50}
$$

which yields $\hat{u}_i^{(t-1)} = -n u_i^{(t-1)}$. Further note that the conjugate of f^* is given by $f(\mathbf{v}) = \frac{\lambda}{2} \left\| \frac{\mathbf{v}}{\lambda} \right\|^2$ and this leads to the dual-to-primal mapping

$$
\mathbf{w} = \nabla f(\mathbf{v}(\boldsymbol{\alpha})) = \frac{1}{\lambda} \mathbf{v}(\boldsymbol{\alpha}) = \frac{1}{\lambda} \sum_{i=1}^{n} \mathbf{a}_i \alpha_i = \frac{1}{\lambda n} \sum_{i=1}^{n} \mathbf{x}_i \alpha_i.
$$

$\qquad\square$

K. Some Useful Pairs of Conjugate Functions

Elastic Net.

Lemma 20 (Conjugate of the Elastic Net Regularizer, see e.g. (Smith et al., 2015)). *For $\eta \in (0, 1]$, the Elastic Net function $g_i(\alpha) := \frac{\eta}{2}\alpha^2 + (1 - \eta)|\alpha|$ has the convex conjugate*

$$g_i^*(x) := \frac{1}{2\eta}\left(\left[|x| - (1 - \eta)\right]_+\right)^2,$$

where $[.]_+$ is the positive part operator, $[s]_+ = s$ for $s > 0$, and zero otherwise. Furthermore, this g_i^* is $1/\eta$-smooth.

Proof. We start by applying the definition of convex conjugate, that is:

$$\ell(x) = \max_{\alpha \in \mathbb{R}}\left[x\alpha - \eta\frac{\alpha^2}{2} - (1 - \eta)|\alpha|\right].$$

We now distinguish two cases for the optimal: $\alpha^\star \geq 0$, $\alpha^\star < 0$. For the first case we get that

$$\ell(x) = \max_{\alpha \in \mathbb{R}}\left[x\alpha - \eta\frac{\alpha^2}{2} - (1 - \eta)\alpha\right].$$

Setting the derivative to 0 we get $\alpha^\star = \frac{x - (1-\eta)}{\eta}$. To satisfy $\alpha^\star \geq 0$, we must have $x \geq 1 - \eta$. Replacing with $\alpha^\star$ we thus get:

$$\ell(x) = \alpha^\star(x - \tfrac{1}{2}\eta\alpha^\star - (1 - \eta)) = \alpha^\star\left(x - \tfrac{1}{2}(x - (1 - \eta)) - (1 - \eta)\right) =$$

$$\tfrac{1}{2}\alpha^\star\left(x - (1 - \eta)\right) = \tfrac{1}{2\eta}\left(x - (1 - \eta)\right)^2.$$

Similarly we can show that for $x \leq -(1 - \eta)$

$$\ell(x) = \tfrac{1}{2\eta}\left(x + (1 - \eta)\right)^2.$$

Finally, by the fact that $\ell(.)$ is convex, always positive, and $\ell(-(1 - \eta)) = \ell(1 - \eta) = 0$, it follows that $\ell(x) = 0$ for every $x \in [-(1 - \eta), 1 - \eta]$.

For the smoothness properties, we consider the derivative of this function $\ell(x)$ and see that $\ell(x)$ is smooth, i.e. has Lipschitz continuous gradient with constant $1/\eta$, assuming $\eta > 0$. Alternatively, use Lemma 12 for the given strongly convex function g_i. $\square$

Group Lasso. The group Lasso regularizer is a norm on $\mathbb{R}^n$, and is defined as

$$g(\boldsymbol{\alpha}) = \lambda \sum_k \|\boldsymbol{\alpha}_{\mathcal{G}_k}\|_2,$$

for a fixed partition of the indices into disjoint groups, $\{1..n\} = \biguplus_k \mathcal{G}_k$. Here $\boldsymbol{\alpha}_{\mathcal{G}_k} \in \mathbb{R}^{|g|}$ denotes the part of the vector $\boldsymbol{\alpha}$ with indices in the group $g \subseteq [n]$. Its dual norm is $\max_{g \in G} \|\boldsymbol{\alpha}_{(g)}\|_2$. Therefore, by Lemma 13, we obtain the conjugate

$$g^*(\mathbf{y}) = \boldsymbol{\iota}_{\{\mathbf{y} \ | \ \max_{g \in G} \|\boldsymbol{\alpha}_{(g)}\|_2 \leq \lambda\}}(\mathbf{y})$$

See, e.g. (Boyd & Vandenberghe, 2004, Example 3.26).

Logistic Loss.

Lemma 21 (Conjugate and Smoothness of the Logistic Loss). *The logistic classifier loss function f is given as*

$$f(A\boldsymbol{\alpha}) := \sum_{j=1}^d \log\left(1 + \exp\left(-b_j\mathbf{y}_j^\top\boldsymbol{\alpha}\right)\right),$$

Its conjugate f^* is given as:

$$f^*(\mathbf{w}) := \sum_{j=1}^d \left((1 + w_jb_j)\log\left(1 + w_jb_j\right) - w_jb_j\log\left(-w_jb_j\right)\right), \tag{51}$$

with the box constraint $-w_jb_j \in [0, 1]$.

Furthermore, $f^*(\mathbf{w})$ is 1-strongly convex over its domain if the labels satisfy $b_j \in [-1, 1]$.

Proof. See e.g. (Smith et al., 2015) or (Shalev-Shwartz & Zhang, 2013). $\square$

L. More Numerical Experiments