Journal of Machine Learning Research 15 (2014) 1523-1548 Submitted 5/13; Revised 2/14; Published 4/14

Iteration Complexity of Feasible Descent Methods for Convex Optimization

Po-Wei Wang B97058@CSIE.NTU.EDU.TW
Chih-Jen Lin CJLIN@CSIE.NTU.EDU.TW
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan

Editor: S. Sathiya Keerthi

Abstract

In many machine learning problems such as the dual form of SVM, the objective function to be minimized is convex but not strongly convex. This fact causes difficulties in obtaining the complexity of some commonly used optimization algorithms. In this paper, we proved the global linear convergence on a wide range of algorithms when they are applied to some non-strongly convex problems. In particular, we are the first to prove $O(\log(1/\epsilon))$ time complexity of cyclic coordinate descent methods on dual problems of support vector classification and regression.

Keywords: convergence rate, convex optimization, iteration complexity, feasible descent methods

1. Introduction

We consider the following convex optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}), \quad \text{where } f(\boldsymbol{x}) \equiv g(E\boldsymbol{x}) + \boldsymbol{b}^\top \boldsymbol{x}, \tag{1}$$

where $g(\boldsymbol{t})$ is a strongly convex function with Lipschitz continuous gradient, E is a constant matrix, and $\mathcal{X}$ is a polyhedral set. Many popular machine learning problems are of this type. For example, given training label-instance pairs $(y_i, \boldsymbol{z}_i),\ i = 1, \ldots, l$, the dual form of L1-loss linear SVM (Boser et al., 1992) is[1]

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - \mathbf{1}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{w} = E\boldsymbol{\alpha},\ 0 \le \alpha_i \le C,\ i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $E = \left[y_1 \boldsymbol{z}_1, \ldots, y_l \boldsymbol{z}_l\right]$, $\mathbf{1}$ is the vector of ones, and C is a given upper bound. Although $\boldsymbol{w}^\top \boldsymbol{w}/2$ is strongly convex in $\boldsymbol{w}$, the objective function of (2) may not be strongly convex in $\boldsymbol{\alpha}$. Common optimization approaches for these machine learning problems include cyclic coordinate descent and others. Unfortunately, most existing results prove only local linear

1. Note that we omit the bias term in the SVM formulation.

convergence, so the number of total iterations cannot be calculated. One of the main difficulties is that $f(\boldsymbol{x})$ may not be strongly convex. In this work, we establish the global linear convergence for a wide range of algorithms for problem (1). In particular, we are the first to prove that the popularly used cyclic coordinate descent methods for dual SVM problems converge linearly since the beginning. Many researchers have stated the importance of such convergence-rate analysis. For example, Nesterov (2012) said that it is *"almost impossible to estimate the rate of convergence"* for general cases. Saha and Tewari (2013) also agreed that *"little is known about the non-asymptotic convergence"* for cyclic coordinate descent methods and they felt *"this gap in the literature needs to be filled urgently."*

Luo and Tseng (1992a) are among the first to establish the asymptotic linear convergence to a non-strongly convex problem related to (1). If $\mathcal{X}$ is a box (possibly unbounded) and a cyclic coordinate descent method is applied, they proved ϵ-optimality in $O(r_0 + \log(1/\epsilon))$ time, where r_0 is an unknown number. Subsequently, Luo and Tseng (1993) considered a class of feasible descent methods that broadly covers coordinate descent and gradient projection methods. For problems including (1), they proved the asymptotic linear convergence. The key concept in their analysis is a local error bound, which states how close the current solution is to the solution set compared with the norm of projected gradient $\nabla^+ f(\boldsymbol{x})$.

$$\min_{\boldsymbol{x}^* \in \mathcal{X}^*} \|\boldsymbol{x}^r - \boldsymbol{x}^*\| \le \kappa \|\nabla^+ f(\boldsymbol{x}^r)\|, \quad \forall r \ge r_0, \tag{3}$$

where r_0 is the above-mentioned unknown iteration index, $\mathcal{X}^*$ is the solution set of problem (1), κ is a positive constant, and $\boldsymbol{x}^r$ is the solution produced after the r-th iteration. Because r_0 is unknown, we call (3) a local error bound, which only holds near the solution set. Local error bounds have been used in other works for convergence analysis such as Luo and Tseng (1992b). If $r_0 = 0$, we call (3) a global error bound from the beginning, and it may help to obtain a global convergence rate. If $f(\boldsymbol{x})$ is strongly convex and $\mathcal{X}$ is a polyhedral set, a global error bound has been established by Pang (1987, Theorem 3.1). One of the main contributions of our work is to prove a global error bound of the possibly non-strongly convex problem (1). Then we are able to establish the global linear convergence and $O(\log(1/\epsilon))$ time complexity for the feasible descent methods.

We briefly discuss some related works, which differ from ours in certain aspects. Chang et al. (2008) applied an (inexact) cyclic coordinate descent method for the primal problem of L2-loss SVM. Because the objective function is strongly convex, they are able to prove the linear convergence since the first iteration. Further, Beck and Tetruashvili (2013) established global linear convergence for block coordinate gradient descent methods on general smooth and strongly convex objective functions. Tseng and Yun (2009) applied a greedy version of block coordinate descent methods on the non-smooth separable problems covering the dual form of SVM. However, they proved only asymptotic linear convergence and $O(1/\epsilon)$ complexity. Moreover, for large-scale linear SVM (i.e., kernels are not used), cyclic rather than greedy coordinate descent methods are more commonly used in practice.[2] Wright (2012) considered the same non-smooth separable problems in Tseng and Yun (2009) and introduced a reduced-Newton acceleration that has asymptotic quadratic convergence.

2. It is now well known that greedy coordinate descent methods such as SMO (Platt, 1998) are less suitable for linear SVM; see some detailed discussion in Hsieh et al. (2008, Section 4.1).

For L1-regularized problems, Saha and Tewari (2013) proved $O(1/\epsilon)$ complexity for cyclic coordinate descent methods under a restrictive isotonic assumption.

Although this work focuses on deterministic algorithms, we briefly review past studies on stochastic (randomized) methods. An interesting fact is that there are more studies on the complexity of randomized rather than deterministic coordinate descent methods. Shalev-Shwartz and Tewari (2009) considered L1-regularized problems, and their stochastic coordinate descent method converges in $O(1/\epsilon)$ iterations in expectation. Nesterov (2012) extended the settings to general convex objective functions and improved the iteration bound to $O(1/\sqrt{\epsilon})$ by proposing an accelerated method. For strongly convex function, he proved that the randomized coordinate descent method converges linearly in expectation. Shalev-Shwartz and Zhang (2013a) provided a sub-linear convergence rate for a stochastic coordinate ascent method, but they focused on the duality gap. Their work is interesting because it bounds the primal objective values. Shalev-Shwartz and Zhang (2013b) refined the sub-linear convergence to be $O(\min(1/\epsilon, 1/\sqrt{\epsilon}))$. Richtárik and Takáč (2014) studied randomized block coordinate descent methods for non-smooth convex problems and had sub-linear convergence on non-strongly convex functions. If the objective function is strongly convex and separable, they obtained linear convergence. Tappenden et al. (2013) extended the methods to inexact settings and had similar convergence rates to those in Richtárik and Takáč (2014).

Our main contribution is a global error bound for the non-strongly convex problem (1), which ensures the global linear convergence of feasible descent methods. The main theorems are presented in Section 2, followed by examples in Section 3. The global error bound is discussed in Section 4, and the proof of global linear convergence of feasible descent methods is given in Section 5. We conclude in Section 6 while leaving properties of projected gradients in Appendix A.

2. Main Results

Consider the general convex optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \quad f(\boldsymbol{x}), \tag{4}$$

where $f(\boldsymbol{x})$ is proper convex and $\mathcal{X}$ is nonempty, closed, and convex. We will prove global linear convergence for a class of optimization algorithms if problem (4) satisfies one of the following assumptions.

Assumption 2.1 *$f(\boldsymbol{x})$ is σ strongly convex and its gradient is ρ Lipschitz continuous. That is, there are constants $\sigma > 0$ and ρ such that*

$$\sigma\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \leq (\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2))^\top(\boldsymbol{x}_1 - \boldsymbol{x}_2), \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$$

and

$$\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\| \leq \rho\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}.$$

Assumption 2.2 *$\mathcal{X} = \{\boldsymbol{x} \mid A\boldsymbol{x} \leq \boldsymbol{d}\}$ is a polyhedral set, the optimal solution set $\mathcal{X}^*$ is non-empty, and*

$$f(\boldsymbol{x}) = g(E\boldsymbol{x}) + \boldsymbol{b}^\top\boldsymbol{x}, \tag{5}$$

where $g(\boldsymbol{t})$ is σ_g strongly convex and $\nabla f(\boldsymbol{x})$ is ρ Lipschitz continuous. This assumption corresponds to problem (1) that motivates this work.

The optimal set $\mathcal{X}^*$ under Assumption 2.1 is non-empty following Weierstrass extreme value theorem.[3] Subsequently, we make several definitions before presenting the main theorem.

Definition 2.3 (Convex Projection Operator)

$$[\boldsymbol{x}]^+_{\mathcal{X}} \equiv \arg\min_{\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\|.$$

From Weierstrass extreme value theorem and the strong convexity of $\|\boldsymbol{x}-\boldsymbol{y}\|^2$ to $\boldsymbol{y}$, the unique $[\boldsymbol{x}]^+_{\mathcal{X}}$ exists for any $\mathcal{X}$ that is closed, convex, and non-empty.

Definition 2.4 (Nearest Optimal Solution)

$$\bar{\boldsymbol{x}} \equiv [\boldsymbol{x}]^+_{\mathcal{X}^*}.$$

With this definition, $\min_{\boldsymbol{x}^*\in\mathcal{X}^*} \|\boldsymbol{x}-\boldsymbol{x}^*\|$ could be simplified to $\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|$.

Definition 2.5 (Projected Gradient)

$$\nabla^+ f(\boldsymbol{x}) \equiv \boldsymbol{x} - [\boldsymbol{x}-\nabla f(\boldsymbol{x})]^+_{\mathcal{X}}.$$

As shown in Lemma A.6, the projected gradient is zero if and only if $\boldsymbol{x}$ is an optimal solution. Therefore, it can be used to check the optimality. Further, we can employ the projected gradient to define an error bound, which measures the distance between $\boldsymbol{x}$ and the optimal set; see the following definition.

Definition 2.6 *An optimization problem admits a **global error bound** if there is a constant κ such that*

$$\|\boldsymbol{x}-\bar{\boldsymbol{x}}\| \le \kappa\|\nabla^+ f(\boldsymbol{x})\|, \quad \forall \boldsymbol{x}\in\mathcal{X}. \tag{6}$$

*A relaxed condition called **global error bound from the beginning** if the above inequality holds only for $\boldsymbol{x}$ satisfying*

$$\boldsymbol{x}\in\mathcal{X} \ \text{ and } \ f(\boldsymbol{x})-f(\bar{\boldsymbol{x}}) \le M,$$

where M is a constant. Usually, we have

$$M \equiv f(\boldsymbol{x}^0) - f^*,$$

where $\boldsymbol{x}^0$ is the start point of an optimization algorithm and f^* is the optimal function value. Therefore, we called this as a bound "from the beginning."

3. The strong convexity in Assumption 2.1 implies that the sublevel set is bounded (Vial, 1983). Then Weierstrass extreme value theorem can be applied.

The global error bound is a property of the optimization problem and is independent from the algorithms. If a bound holds,[4] then using Lemmas A.5, A.6, and (6) we can obtain

$$\frac{1}{2+\rho}\|\nabla^+ f(\boldsymbol{x})\| \leq \|\boldsymbol{x}-\bar{\boldsymbol{x}}\| \leq \kappa\|\nabla^+ f(\boldsymbol{x})\|, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

This property indicates that $\|\nabla^+ f(\boldsymbol{x})\|$ is useful to estimate the distance to the optimum. We will show that a global error bound enables the proof of global linear convergence of some optimization algorithms. The bound under Assumption 2.1, which requires strong convexity, was already proved in Pang (1987) with

$$\kappa = \frac{1+\rho}{\sigma}.$$

However, for problems under Assumption 2.2 such as the dual form of L1-loss SVM, the objective function is not strongly convex, so a new error bound is required. We prove the bound in Section 4 with

$$\kappa = \theta^2(1+\rho)\left(\frac{1+2\|\nabla g(\boldsymbol{t}^*)\|^2}{\sigma_g} + 4M\right) + 2\theta\|\nabla f(\bar{\boldsymbol{x}})\|, \tag{7}$$

where $\boldsymbol{t}^*$ is a constant vector that equals $E\boldsymbol{x}^*$, $\forall \boldsymbol{x}^* \in \mathcal{X}^*$ and θ is the constant from Hoffman's bound (Hoffman, 1952; Li, 1994).

$$\theta \equiv \sup_{\boldsymbol{u},\boldsymbol{v}} \left\{ \left\| \begin{matrix} \boldsymbol{u} \\ \boldsymbol{v} \end{matrix} \right\| \;\middle|\; \begin{array}{l} \|A^\top \boldsymbol{u} + \left(\begin{smallmatrix} E \\ \boldsymbol{b}^\top \end{smallmatrix}\right)^\top \boldsymbol{v}\| = 1,\ \boldsymbol{u} \geq 0. \\ \text{The corresponding rows of } A, E \text{ to } \boldsymbol{u}, \boldsymbol{v}\text{'s} \\ \text{non-zero elements are linearly independent.} \end{array} \right\}.$$

Specially, when $\boldsymbol{b} = \boldsymbol{0}$ or $\mathcal{X} = \mathbb{R}^l$, the constant could be simplified to

$$\kappa = \theta^2 \frac{1+\rho}{\sigma_g}. \tag{8}$$

Now we define a class of optimization algorithms called the feasible descent methods for solving (4).

Definition 2.7 (Feasible Descent Methods) *A sequence $\{\boldsymbol{x}^r\}$ is generated by a feasible descent method if for every iteration index r, $\{\boldsymbol{x}^r\}$ satisfies*

$$\boldsymbol{x}^{r+1} = [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r) + \boldsymbol{e}^r]^+_{\mathcal{X}}, \tag{9}$$

$$\|\boldsymbol{e}^r\| \leq \beta\|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\|, \tag{10}$$

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^{r+1}) \geq \gamma\|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\|^2, \tag{11}$$

where $\inf_r \omega_r > 0$, $\beta > 0$, and $\gamma > 0$.

4. Note that not all problems have a global error bound. An example is $\min_{x\in\mathbb{R}} x^4$.

The framework of feasible descent methods broadly covers many algorithms that use the first-order information. For example, the projected gradient descent, the cyclic coordinate descent, the proximal point minimization, the extragradient descent, and matrix splitting algorithms are all feasible descent methods (Luo and Tseng, 1993). With the global error bound under Assumption 2.1 or Assumption 2.2, in the following theorem we prove the global linear convergence for all algorithms that fit into the feasible descent methods.

Theorem 2.8 (Global Linear Convergence) *If an optimization problem satisfies Assumption 2.1 or 2.2, then any feasible descent method on it has global linear convergence. To be specific, the method converges Q-linearly with*

$$f(\boldsymbol{x}^{r+1}) - f^* \leq \frac{\phi}{\phi+\gamma}(f(\boldsymbol{x}^r) - f^*), \quad \forall r \geq 0,$$

where κ is the error bound constant in (6),

$$\phi = (\rho + \frac{1+\beta}{\underline{\omega}})(1 + \kappa\frac{1+\beta}{\underline{\omega}}), \quad \text{and} \quad \underline{\omega} \equiv \min(1, \inf_r \omega_r).$$

This theorem enables global linear convergence in many machine learning problems. The proof is given in Section 5. In Section 3, we discuss examples on cyclic coordinate descent methods.

3. Examples: Cyclic Coordinate Descent Methods

Cyclic coordinate descent methods are now widely used for machine learning problems because of its efficiency and simplicity (solving a one-variable sub-problem at a time). Luo and Tseng (1992a) proved the asymptotic linear convergence if sub-problems are solved exactly, and here we further show the global linear convergence.

3.1 Exact Cyclic Coordinate Descent Methods for Dual SVM Problems

In the following algorithm, each one-variable sub-problem is exactly solved.

Definition 3.1 *A cyclic coordinate descent method on a box $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_l$ is defined by the update rule*

$$x_i^{r+1} = \arg\min_{x_i \in \mathcal{X}_i} f(x_1^{r+1}, \ldots, x_{i-1}^{r+1}, x_i, x_{i+1}^r, \ldots, x_l^r), \quad \text{for } i = 1, \ldots, l, \tag{12}$$

where $\mathcal{X}_i$ is the region under box constraints for coordinate i.

The following lemma shows that coordinate descent methods are special cases of the feasible descent methods.

Lemma 3.2 *The cyclic coordinate descent method is a feasible descent method with*

$$\omega_r = 1, \ \forall r, \quad \beta = 1 + \rho\sqrt{l},$$

and

$$\gamma = \begin{cases} \frac{\sigma}{2} & \text{if Assumption 2.1 holds,} \\ \frac{1}{2}\min_i \|E_i\|^2 & \text{if Assumption 2.2 holds with } \|E_i\| > 0, \ \forall i, \end{cases}$$

where E_i is the ith column of E.

Proof This lemma can be directly obtained using Proposition 3.4 of Luo and Tseng (1993). Our assumptions correspond to cases (a) and (c) in Theorem 2.1 of Luo and Tseng (1993), which fulfill conditions needed by their Proposition 3.4. ■

For faster convergence, we may randomly permute all variables before each cycle of updating them (e.g., Hsieh et al., 2008). This setting does not affect the proof of Lemma 3.2.

Theorem 2.8 and Lemma 3.2 immediately imply the following corollary.

Corollary 3.3 *The cyclic coordinate descent methods have global linear convergence if Assumption 2.1 is satisfied or Assumption 2.2 is satisfied with* $\|E_i\| > 0,\ \forall i$.

Next, we analyze the cyclic coordinate descent method to solve dual SVM problems. The method can be traced back to Hildreth (1957) for quadratic programming problems and has recently been widely used following the work by Hsieh et al. (2008). For L1-loss SVM, we have shown in (2) that the objective function can be written in the form of (1) by a strongly convex function $g(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{w}/2$ and $E_i = y_i \boldsymbol{z}_i$ for all label-instance pair $(y_i, \boldsymbol{z}_i)$. Hsieh et al. (2008) pointed out that $\|E_i\| = 0$ implies the optimal α_i^* is C, which can be obtained at the first iteration and is never changed. Therefore, we need not consider such variables at all. With all conditions satisfied, Corollary 3.3 implies that cyclic coordinate descent method for dual L1-loss SVM has global linear convergence. For dual L2-loss SVM, the objective function is

$$\frac{1}{2}\boldsymbol{\alpha}^\top Q \boldsymbol{\alpha} - \mathbf{1}^\top \boldsymbol{\alpha} + \frac{1}{2C}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \tag{13}$$

where $Q_{t,j} = y_t y_j \boldsymbol{z}_t^\top \boldsymbol{z}_j, \forall 1 \le t, j \le l$ and $\mathbf{1}$ is the vector of ones. Eq. (13) is strongly convex and its gradient is Lipschitz continuous, so Assumption 2.1 and Corollary 3.3 imply the global linear convergence.

We move on to check the dual problems of support vector regression (SVR). Given value-instance pairs $(y_i, \boldsymbol{z}_i),\ i = 1, \ldots, l$, the dual form of L1-loss m-insensitive SVR (Vapnik, 1995) is

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^\top \begin{bmatrix} Q & -Q \\ -Q & Q \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} m\mathbf{1} - \boldsymbol{y} \\ m\mathbf{1} + \boldsymbol{y} \end{bmatrix}^\top \boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \le \alpha_i \le C,\ i = 1, \ldots, 2l, \end{aligned} \tag{14}$$

where $Q_{t,j} = \boldsymbol{z}_t^\top \boldsymbol{z}_j, \forall 1 \le t, j \le l$, and m and C are given parameters. Similar to the case of classification, we can also perform cyclic coordinate descent methods; see Ho and Lin (2012, Section 3.2). Note that Assumption 2.2 must be used here because for any Q, the Hessian in (14) is only positive semi-definite rather than positive definite. In contrast, for classification, if Q is positive definite, the objective function in (2) is strongly convex and Assumption 2.1 can be applied. To represent (14) in the form of (1), let

$$E_i = \boldsymbol{z}_i,\ \ i = 1, \ldots, l \ \text{ and } \ E_i = -\boldsymbol{z}_i,\ \ i = l+1, \ldots, 2l.$$

Then $g(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{w}/2$ with $\boldsymbol{w} = E\boldsymbol{\alpha}$ is a strongly convex function to $\boldsymbol{w}$. Similar to the situation in classification, if $\|E_i\| = 0$, then the optimal α_i^* is bounded and can be obtained at the first iteration. Without considering these variables, Corollary 3.3 implies the global linear convergence.

3.2 Inexact Cyclic Coordinate Descent Methods for Primal SVM Problems

In some situations the sub-problems (12) of cyclic coordinate descent methods cannot be easily solved. For example, in Chang et al. (2008) to solve the primal form of L2-loss SVM,

$$\min_{\boldsymbol{w}} \ f(\boldsymbol{w}), \quad \text{where } f(\boldsymbol{w}) \equiv \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{l} \max(1 - y_i \boldsymbol{w}^\top \boldsymbol{z}_i, 0)^2, \tag{15}$$

each sub-problem does not have a closed-form solution, and they approximately solve the sub-problem until a sufficient decrease condition is satisfied. They have established the global linear convergence, but we further show that their method can be included in our framework.

To see that Chang et al. (2008)'s method is a feasible descent method, it is sufficient to prove that (9)-(11) hold. First, we notice that their sufficient decrease condition for updating each variable can be accumulated. Thus, for one cycle of updating all variables, we have

$$f(\boldsymbol{w}^r) - f(\boldsymbol{w}^{r+1}) \geq \gamma \|\boldsymbol{w}^r - \boldsymbol{w}^{r+1}\|^2,$$

where $\gamma > 0$ is a constant. Next, because (15) is unconstrained, if $\boldsymbol{z}_i \in \mathbb{R}^n, \forall i$, we can make

$$\mathcal{X} = \mathbb{R}^n \text{ and } \boldsymbol{e}^r = \boldsymbol{w}^{r+1} - \boldsymbol{w}^r + \nabla f(\boldsymbol{w}^r)$$

such that

$$\boldsymbol{w}^{r+1} = [\boldsymbol{w}^r - \nabla f(\boldsymbol{w}^r) + \boldsymbol{e}^r]^+_{\mathcal{X}}.$$

Finally, from Appendix A.3 of Chang et al. (2008),

$$\|\boldsymbol{e}^r\| \leq \|\boldsymbol{w}^r - \boldsymbol{w}^{r+1}\| + \|\nabla f(\boldsymbol{w}^r)\| \leq \beta \|\boldsymbol{w}^r - \boldsymbol{w}^{r+1}\|,$$

where $\beta > 0$ is a constant. Therefore, all conditions (9)-(11) hold. Note that (15) is strongly convex because of the $\boldsymbol{w}^\top \boldsymbol{w}$ term and $\nabla f(\boldsymbol{w})$ is Lipschitz continuous from (Lin et al., 2008, Section 6.1), so Assumption 2.1 is satisfied. With Theorem 2.8, the method by Chang et al. (2008) has global linear convergence.

3.3 Gauss-Seidel Methods for Solving Linear Systems

Gauss-Seidel (Seidel, 1874) is a classic iterative method to solve a linear system

$$Q\boldsymbol{\alpha} = \boldsymbol{b}. \tag{16}$$

Gauss-Seidel iterations take the following form.

$$\alpha_i^{r+1} = \frac{b_i - \sum_{j=1}^{i-1} Q_{ij}\alpha_j^{r+1} - \sum_{j=i+1}^{l} Q_{ij}\alpha_j^r}{Q_{ii}}. \tag{17}$$

If Q is symmetric positive semi-definite and (16) has at least one solution, then the following optimization problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \quad \frac{1}{2}\boldsymbol{\alpha}^\top Q\boldsymbol{\alpha} - \boldsymbol{b}^\top \boldsymbol{\alpha} \tag{18}$$

has the same solution set as (16). Further, α_i^{r+1} in (17) is the solution of minimizing (18) over α_i while fixing $\alpha_1^{r+1}, \ldots, \alpha_{i-1}^{r+1}, \alpha_{i+1}^{r}, \ldots, \alpha_l^{r}$. Therefore, Gauss-Seidel method is a special case of coordinate descent methods.

Clearly, we need $Q_{ii} > 0, \forall i$ so that (17) is well defined. This condition also implies that

$$Q = E^\top E, \text{ where } E \text{ has no zero column.} \tag{19}$$

Otherwise, $\|E_i\| = 0$ leads to $Q_{ii} = 0$ so the $Q_{ii} > 0$ assumption is violated. Note that because Q is symmetric positive semi-definite, its orthogonal diagonalization $U^T D U$ exists and we choose $E = \sqrt{D}U$. Using (19) and Lemma 3.2, Gauss-Seidel method is a feasible descent method. By Assumption 2.2 and our main Theorem 2.8, we have the following convergence result.

Corollary 3.4 *If*

1. *Q is symmetric positive semi-definite and $Q_{ii} > 0, \forall i$, and*
2. *The linear system (16) has at least a solution,*

then the Gauss-Seidel method has global linear convergence.

This corollary covers some well-known results of the Gauss-Seidel method, which were previously proved by other ways. For example, in most numerical linear algebra textbooks (e.g., Golub and Van Loan, 1996), it is proved that if Q is strictly diagonally dominant (i.e., $Q_{ii} > \sum_{j \neq i} |Q_{ij}|, \ \forall i$), then the Gauss-Seidel method converges linearly. We show in Lemma C.1 that a strictly diagonally dominant matrix is positive definite, so Corollary 3.4 immediately implies global linear convergence.

3.4 Quantity of the Convergence Rate

To demonstrate the relationship between problem parameters (e.g., number of instances and features) and the convergence rate constants, we analyze the constants κ and ϕ for two problems. The first example is the exact cyclic coordinate descent method for the dual problem (2) of L1-loss SVM. For simplicity, we assume $\|E_i\| = 1, \ \forall i$, where E_i denotes the ith column of E. We have

$$\sigma_g = 1 \tag{20}$$

by $g(\boldsymbol{t}) = \boldsymbol{t}^\top \boldsymbol{t}/2$. Observe the following primal formulation of L1-loss SVM.

$$\min_{\boldsymbol{w}} \quad P(\boldsymbol{w}), \ \text{ where } \ P(\boldsymbol{w}) \equiv \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C \sum_{i=1}^{l} \max(1 - y_i \boldsymbol{w}^\top \boldsymbol{z}_i, 0).$$

Let $\boldsymbol{w}^*$ and $\boldsymbol{\alpha}^*$ be any optimal solution of the primal and the dual problems, respectively. By KKT optimality condition, we have $\boldsymbol{w}^* = E\boldsymbol{\alpha}^*$. Consider $\boldsymbol{\alpha}^0 = \mathbf{0}$ as the initial feasible solution. With the duality and the strictly decreasing property of $\{f(\boldsymbol{\alpha}^r)\}$,

$$f(\boldsymbol{\alpha}^r) - f(\boldsymbol{\alpha}^*) \leq f(\mathbf{0}) - f(\boldsymbol{\alpha}^*) = f(\mathbf{0}) + P(\boldsymbol{w}^*) \leq f(\mathbf{0}) + P(\mathbf{0}) \leq 0 + Cl \equiv M. \tag{21}$$

Besides,

$$\frac{1}{2}\boldsymbol{w}^{*\top}\boldsymbol{w}^* \le P(\boldsymbol{w}^*) \le P(\boldsymbol{0}) \le Cl \text{ implies } \|\boldsymbol{w}^*\| = \|E\boldsymbol{\alpha}^*\| \le \sqrt{2Cl}. \tag{22}$$

From (22),

$$\|\nabla f(\bar{\boldsymbol{\alpha}})\| \le \|E\|\|E\boldsymbol{\alpha}^*\| + \|\mathbf{1}\| \le \sqrt{\Sigma_i\|E_i\|^2}\|E\boldsymbol{\alpha}^*\| + \|\mathbf{1}\| \le \sqrt{2Cl} + \sqrt{l}. \tag{23}$$

To conclude, by (7), (20), (21), (22), (23), and $\nabla g(\boldsymbol{w}^*) = \boldsymbol{w}^*$,

$$\begin{aligned}\kappa &= \theta^2(1+\rho)\left(\frac{1+2\|\nabla g(\boldsymbol{w}^*)\|^2}{\sigma_g} + 4M\right) + 2\theta\|\nabla f(\bar{\boldsymbol{\alpha}})\| \\ &\le \theta^2(1+\rho)((1+4Cl)+4Cl) + 2\theta(\sqrt{2Cl}+\sqrt{l}) \\ &= O(\rho\theta^2 Cl).\end{aligned}$$

Now we examine the rate ϕ for linear convergence. From Theorem 2.8, we have

$$\begin{aligned}\phi &= (\rho + \frac{1+\beta}{\underline{\omega}})(1+\kappa\frac{1+\beta}{\underline{\omega}}) \\ &= (\rho + 2 + \rho\sqrt{l})(1+\kappa(2+\rho\sqrt{l})) \\ &= O(\rho^3\theta^2 Cl^2),\end{aligned}$$

where

$$\underline{\omega} = 1, \quad \beta = 1+\rho\sqrt{l}, \quad \gamma = \frac{1}{2} \tag{24}$$

are from Lemma 3.2 and the assumption that $\|E_i\| = 1$, $\forall i$. To conclude, we have $\kappa = O(\rho\theta^2 Cl)$ and $\phi = O(\rho^3\theta^2 Cl^2)$ for the exact cyclic coordinate descent method for the dual problem of L1-loss SVM.

Next we consider the Gauss-Seidel method for solving linear systems in Section 3.3 by assuming $\|Q\| = 1$ and $Q_{ii} > 0$, $\forall i$, where $\|Q\|$ denotes the spectral norm of Q. Similar to (20), we have $\sigma_g = 1$ by $g(\boldsymbol{t}) = \boldsymbol{t}^\top\boldsymbol{t}/2$. Further, $\rho = 1$ from

$$\|\nabla f(\boldsymbol{\alpha}_1) - \nabla f(\boldsymbol{\alpha}_2)\| \le \|Q\|\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\| = \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|.$$

Because the optimization problem is unconstrained, by (8) we have

$$\kappa = \theta^2\frac{1+\rho}{\sigma_g} = 2\theta^2, \tag{25}$$

where θ is defined as

$$\theta \equiv \sup_{\boldsymbol{v}}\left\{\|\boldsymbol{v}\| \,\middle|\, \begin{array}{l}\|E^\top\boldsymbol{v}\| = 1. \\ \text{The corresponding rows of } E \text{ to } \boldsymbol{v}\text{'s} \\ \text{non-zero elements are linearly independent.}\end{array}\right\}, \tag{26}$$

and $E = \sqrt{D}U$ is from the orthogonal diagonalization of Q in (19). To have that the corresponding rows of E to $\boldsymbol{v}$'s non-zero elements are linearly independent, we need

$$v_i = 0 \text{ if } D_{ii} = 0.$$

Therefore, problem (26) becomes to select v_i with $D_{ii} > 0$ such that $E^\top \boldsymbol{v} = 1$ and $\|\boldsymbol{v}\|$ is maximized. Because U's rows are orthogonal vectors and any $D_{ii} > 0$ is an eigen-value of Q, the maximum occurs if we choose $\boldsymbol{\nu} = \boldsymbol{e}_i$ as the indicator vector corresponding to the smallest non-zero eigen-value $\sigma_{\min\text{-nnz}}$. Then,

$$\text{the solution } \boldsymbol{v} \text{ in (26) is } \frac{\boldsymbol{\nu}}{\sqrt{\sigma_{\min\text{-nnz}}}} \text{ and } \theta^2 = \frac{1}{\sigma_{\min\text{-nnz}}}. \tag{27}$$

From Lemma 3.2, $\underline{\omega}$, β, and γ of the Gauss-Seidel method are the same as (24). Thus, Theorem 2.8, (24), (25), and (27) give the convergence rate constant

$$\phi = (3+\sqrt{l})(1+\kappa(2+\sqrt{l})) = (3+\sqrt{l})(1+\frac{4+2\sqrt{l}}{\sigma_{\min\text{-nnz}}}). \tag{28}$$

With (24), (28), and Theorem 2.8, the Gauss-Seidel method on solving linear systems has linear convergence with

$$f(\boldsymbol{\alpha}^{r+1}) - f^* \leq (1 - \frac{\sigma_{\min\text{-nnz}}}{4(6+5\sqrt{l}+l)+(7+2\sqrt{l})\sigma_{\min\text{-nnz}}})(f(\boldsymbol{\alpha}^r) - f^*), \quad \forall r \geq 0.$$

We discuss some related results. A similar rate of linear convergence appears in Beck and Tetruashvili (2013). They assumed f is $\sigma_{\min}$ strongly convex and the optimization problem is unconstrained. By considering a block coordinate descent method with a conservative rule of selecting the step size, they showed

$$f(\boldsymbol{\alpha}^{r+1}) - f^* \leq (1 - \frac{\sigma_{\min}}{2(1+l)})(f(\boldsymbol{\alpha}^r) - f^*), \quad \forall r \geq 0.$$

Our obtained rate is comparable, but is more general to cover singular Q.

4. Proofs of Global Error Bounds

In this section, we prove the global error bound (6) under Assumptions 2.1 or 2.2. The following theorem proves the global error bound under Assumption 2.1.

Theorem 4.1 (Pang 1987, Theorem 3.1) *Under Assumption 2.1,*

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \kappa \|\nabla^+ f(\boldsymbol{x})\|, \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where $\kappa = (1+\rho)/\sigma$.

Proof Because $f(\boldsymbol{x})$ is strongly convex, $\mathcal{X}^*$ has only one element $\bar{\boldsymbol{x}}$. From Lemmas A.4 and A.6, the result holds immediately. ■

The rest of this section focuses on proving a global error bound under Assumption 2.2. We start by sketching the proof. First, observe that the optimal set is a polyhedron by Lemma 4.2. Then $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|$ is identical to the distance of $\boldsymbol{x}$ to the polyhedron. A known technique to bound the distance between $\boldsymbol{x}$ and a polyhedron is Hoffman's bound (Hoffman,

1952). Because the original work uses L1-norm, we provide in Lemma 4.3 a special version of Li (1994) that uses L2-norm. With the feasibility of $\boldsymbol{x}$, there is

$$\|\boldsymbol{x}-\bar{\boldsymbol{x}}\| \leq \theta\left(A, \begin{pmatrix} E \\ \boldsymbol{b}^\top \end{pmatrix}\right) \left\| \begin{matrix} E(\boldsymbol{x}-\bar{\boldsymbol{x}}) \\ \boldsymbol{b}^\top(\boldsymbol{x}-\bar{\boldsymbol{x}}) \end{matrix} \right\|,$$

where $\theta\left(A, \begin{pmatrix} E \\ \boldsymbol{b}^\top \end{pmatrix}\right)$ is a constant related to A, E, and $\boldsymbol{b}$. Subsequently, we bound $\|E(\boldsymbol{x}-\bar{\boldsymbol{x}})\|^2$ and $(\boldsymbol{b}^\top(\boldsymbol{x}-\bar{\boldsymbol{x}}))^2$ in Lemmas 4.4 and 4.5 by values consisting of $\|\nabla^+ f(\boldsymbol{x})\|$ and $\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|$. Such bounds are obtained using properties of the optimization problem such as the strong convexity of $g(\cdot)$. Finally, we obtain a quadratic inequality involving $\|\nabla^+ f(\boldsymbol{x})\|$ and $\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|$, which eventually leads to a global error bound under Assumption 2.2.

We begin the formal proof by expressing the optimal set as a polyhedron.

Lemma 4.2 (Optimal Condition) *Under Assumption 2.2, there are unique $\boldsymbol{t}^*$ and s^* such that $\forall \boldsymbol{x}^* \in \mathcal{X}^*$,*

$$E\boldsymbol{x}^* = \boldsymbol{t}^*, \;\; \boldsymbol{b}^\top \boldsymbol{x}^* = s^*, \;\; \text{and} \;\; A\boldsymbol{x}^* \leq \boldsymbol{d}. \tag{29}$$

Note that A and $\boldsymbol{d}$ are the constants for generating the feasible set $\mathcal{X} = \{\boldsymbol{x} \mid A\boldsymbol{x} \leq \boldsymbol{d}\}$. Further,

$$\boldsymbol{x}^* \text{ satisfies (29)} \Leftrightarrow \boldsymbol{x}^* \in \mathcal{X}^*. \tag{30}$$

Specially, when $\boldsymbol{b} = \boldsymbol{0}$ or $\mathcal{X} = \mathbb{R}^l$,[5]

$$E\boldsymbol{x}^* = \boldsymbol{t}^*, \;\; A\boldsymbol{x}^* \leq \boldsymbol{d} \Leftrightarrow \boldsymbol{x}^* \in \mathcal{X}^*. \tag{31}$$

Proof First, we prove (29). The proof is similar to Lemma 3.1 in Luo and Tseng (1992a). For any $\boldsymbol{x}_1^*, \boldsymbol{x}_2^* \in \mathcal{X}^*$, from the convexity of $f(\boldsymbol{x})$,

$$f((\boldsymbol{x}_1^* + \boldsymbol{x}_2^*)/2) = (f(\boldsymbol{x}_1^*) + f(\boldsymbol{x}_2^*))/2.$$

By the definition of $f(\boldsymbol{x})$ in Assumption 2.2, we have

$$g((E\boldsymbol{x}_1^* + E\boldsymbol{x}_2^*)/2) + \boldsymbol{b}^\top(\boldsymbol{x}_1^* + \boldsymbol{x}_2^*)/2 = (g(E\boldsymbol{x}_1^*) + g(E\boldsymbol{x}_2^*) + \boldsymbol{b}^\top(\boldsymbol{x}_1^* + \boldsymbol{x}_2^*))/2.$$

Cancel $\boldsymbol{b}^\top(\boldsymbol{x}_1^*+\boldsymbol{x}_2^*)/2$ from both sides. By the strong convexity of $g(\boldsymbol{t})$, we have $E\boldsymbol{x}_1^* = E\boldsymbol{x}_2^*$. Thus, $\boldsymbol{t}^* \equiv E\boldsymbol{x}^*$ is unique. Similarly, because $f(\boldsymbol{x}_1^*) = f(\boldsymbol{x}_2^*)$,

$$g(\boldsymbol{t}^*) + \boldsymbol{b}^\top \boldsymbol{x}_1^* = g(\boldsymbol{t}^*) + \boldsymbol{b}^\top \boldsymbol{x}_2^*.$$

Therefore, $s^* \equiv \boldsymbol{b}^\top \boldsymbol{x}^*$ is unique, and $A\boldsymbol{x}^* \leq \boldsymbol{d}$, $\forall \boldsymbol{x}^* \in \mathcal{X}^*$ holds naturally by $\mathcal{X}^* \subseteq \mathcal{X}$. Further,

$$f(\boldsymbol{x}^*) = g(\boldsymbol{t}^*) + s^*, \quad \forall \boldsymbol{x}^* \in \mathcal{X}^*. \tag{32}$$

The result in (29) immediately implies the ($\Leftarrow$) direction of (30). For the ($\Rightarrow$) direction, for any $\boldsymbol{x}^*$ satisfying

$$E\boldsymbol{x}^* = \boldsymbol{t}^*, \;\; \boldsymbol{b}^\top \boldsymbol{x}^* = s^*, \;\; A\boldsymbol{x}^* \leq \boldsymbol{d},$$

we have $f(\boldsymbol{x}^*) = g(\boldsymbol{t}^*) + s^*$. From (32), $\boldsymbol{x}^*$ is an optimal solution.

5. When $\mathcal{X} = \mathbb{R}^l$, we can take zero A and $\boldsymbol{d}$ for a trivial linear inequality.

Now we examine the special cases. If $\boldsymbol{b} = \mathbf{0}$, we have $\boldsymbol{b}^\top \boldsymbol{x} = 0,\ \forall \boldsymbol{x} \in \mathcal{X}$. Therefore, (30) is reduced to (31). On the other hand, if $\mathcal{X} = \mathbb{R}^l$, the optimization problem is unconstrained. Thus,

$$\boldsymbol{x}^* \text{ is optimal} \Leftrightarrow \nabla f(\boldsymbol{x}^*) = \mathbf{0} = E^\top \nabla g(\boldsymbol{t}^*) + \boldsymbol{b}.$$

As a result, $E\boldsymbol{x}^* = \boldsymbol{t}^*$ is a necessary and sufficient optimality condition. ■

Because the optimal set is a polyhedron, we will apply the following Hoffman's bound in Lemma 4.6 to upper-bound the distance to the optimal set by the violation of the polyhedron's linear inequalities.

Lemma 4.3 (Hoffman's Bound) *Let P be the non-negative orthant and consider a non-empty polyhedron*

$$\{\boldsymbol{x}^* \mid A\boldsymbol{x}^* \le \boldsymbol{d},\ \ E\boldsymbol{x}^* = \boldsymbol{t}\}.$$

For any $\boldsymbol{x}$, there is a feasible point $\boldsymbol{x}^*$ such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \le \theta(A, E) \left\| \begin{bmatrix} [A\boldsymbol{x} - \boldsymbol{d}]_P^+ \\ E\boldsymbol{x} - \boldsymbol{t} \end{bmatrix} \right\|, \tag{33}$$

where

$$\theta(A, E) \equiv \sup_{\boldsymbol{u},\boldsymbol{v}} \left\{ \left\| \begin{matrix} \boldsymbol{u} \\ \boldsymbol{v} \end{matrix} \right\| \;\middle|\; \begin{array}{l} \|A^\top \boldsymbol{u} + E^\top \boldsymbol{v}\| = 1,\ \boldsymbol{u} \ge \mathbf{0}. \\ \textit{The corresponding rows of } A,\ E \textit{ to } \boldsymbol{u},\ \boldsymbol{v}\textit{'s} \\ \textit{non-zero elements are linearly independent.} \end{array} \right\}. \tag{34}$$

Note that $\theta(A, E)$ is independent of $\boldsymbol{x}$.

The proof of the lemma is given in Appendix B. Before applying Hoffman's bound, we need some technical lemmas to bound $\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2$ and $(\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2$, which will appear on the right-hand side of Hoffman's bound for the polyhedron of the optimal set.

Lemma 4.4 *Under Assumption 2.2, we have constants ρ and σ_g such that*

$$\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2 \le \frac{1+\rho}{\sigma_g} \|\nabla^+ f(\boldsymbol{x})\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof By $E\bar{\boldsymbol{x}} = \boldsymbol{t}^*$ from Lemma 4.2, the strong convexity of $g(\boldsymbol{t})$, and the definition of $f(\boldsymbol{x})$ in (5), there exists σ_g such that

$$\sigma_g \|E\boldsymbol{x} - \boldsymbol{t}^*\|^2 \le (\nabla g(E\boldsymbol{x}) - \nabla g(E\bar{\boldsymbol{x}}))^\top (E\boldsymbol{x} - E\bar{\boldsymbol{x}}) = (\nabla f(\boldsymbol{x}) - \nabla f(\bar{\boldsymbol{x}}))^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

By Lemma A.3, the above inequality becomes

$$\sigma_g \|E\boldsymbol{x} - \boldsymbol{t}^*\|^2 \le (1+\rho) \|\nabla^+ f(\boldsymbol{x}) - \nabla^+ f(\bar{\boldsymbol{x}})\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|,$$

where ρ is the constant for the Lipschitz continuity of ∇f. Because $\bar{\boldsymbol{x}}$ is an optimal solution, $\nabla^+ f(\bar{\boldsymbol{x}}) = \mathbf{0}$ by Lemma A.6. Thus, the result holds. ■

Next we bound $(\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2$.

Lemma 4.5 *Under Assumption 2.2 and the condition*

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq M, \tag{35}$$

there exists a constant $\rho > 0$ such that

$$\begin{aligned}&(\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2 \\ &\leq 4(1+\rho)M\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + 4\|\nabla f(\bar{\boldsymbol{x}})\|^2\|\nabla^+ f(\boldsymbol{x})\|^2 + 2\|\nabla g(\boldsymbol{t}^*)\|^2\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2.\end{aligned}$$

Proof By $\boldsymbol{b}^\top \bar{\boldsymbol{x}} = s^*$ and $E\bar{\boldsymbol{x}} = \boldsymbol{t}^*$ from Lemma 4.2 and the definition of $f(\boldsymbol{x})$, we have

$$\boldsymbol{b}^\top \boldsymbol{x} - s^* = \nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) - \nabla g(\boldsymbol{t}^*)^\top (E\boldsymbol{x} - \boldsymbol{t}^*).$$

Square both sides of the equality. Then by $(a-b)^2 \leq 2a^2 + 2b^2$,

$$(\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2 \leq 2(\nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}))^2 + 2(\nabla g(\boldsymbol{t}^*)^\top (E\boldsymbol{x} - \boldsymbol{t}^*))^2. \tag{36}$$

Consider the right-hand side in (36). The second term can be bounded by $2\|\nabla g(\boldsymbol{t}^*)\|^2\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2$, and the first term is bounded using the inequalities

$$\begin{aligned}\nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) &\leq \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) \\ &\leq \nabla^+ f(\boldsymbol{x})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}} + \nabla f(\boldsymbol{x}) - \nabla^+ f(\boldsymbol{x})) \\ &\leq \nabla^+ f(\boldsymbol{x})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}} + \nabla f(\boldsymbol{x}) - \nabla f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})) \\ &\leq (1+\rho)\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + \nabla^+ f(\boldsymbol{x})^\top \nabla f(\bar{\boldsymbol{x}}).\end{aligned} \tag{37}$$

The first inequality is by convexity, the second is by Lemma A.1,[6] the third is by $\|\nabla^+ f(\boldsymbol{x})\|^2 \geq 0$, and the last is by the Lipschitz continuity of ∇f. By the optimality of $\bar{\boldsymbol{x}}$,

$$\nabla f(\bar{\boldsymbol{x}})^\top ([\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}} - \boldsymbol{x} + \boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0. \tag{38}$$

Thus, (38), the convexity of $f(\cdot)$, and (35) imply that

$$\nabla f(\bar{\boldsymbol{x}})^\top \nabla^+ f(\boldsymbol{x}) \leq \nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq M. \tag{39}$$

Let

$$a \equiv \nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}),\ u \equiv (1+\rho)\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|,\ v \equiv \nabla f(\bar{\boldsymbol{x}})^\top \nabla^+ f(\boldsymbol{x}).$$

Then we have

$$0 \leq a \leq u + v \text{ from (37) and optimality of } \bar{\boldsymbol{x}},\ a - v \geq 0 \text{ from (39), and } u \geq 0.$$

Therefore, $a^2 \leq au + av \leq au + v(u+v) \leq au + au + v^2 \leq 2au + 2v^2$, and

$$\begin{aligned}&(\nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}))^2 \\ &\leq 2(\nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}))(1+\rho)\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + 2(\nabla f(\bar{\boldsymbol{x}})^\top \nabla^+ f(\boldsymbol{x}))^2 \\ &\leq 2(1+\rho)M\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + 2\|\nabla f(\bar{\boldsymbol{x}})\|^2\|\nabla^+ f(\boldsymbol{x})\|^2.\end{aligned}$$

The last inequality is from (39) and Cauchy's inequality. Together with (36) the result immediately holds. ■

Combining the previous two lemmas, we are now ready to prove the global error bound.

6. Note that we use $([\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}} - \boldsymbol{x} + \nabla f(\boldsymbol{x}))^\top ([\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}} - \boldsymbol{x} + \boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ and $\nabla^+ f(\boldsymbol{x}) = \boldsymbol{x} - [\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}$.

Theorem 4.6 (Error Bound) *Under Assumption 2.2 and any $M > 0$, we have*

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \kappa \|\nabla^+ f(\boldsymbol{x})\|, \quad \forall \boldsymbol{x} \text{ with } \boldsymbol{x} \in \mathcal{X} \text{ and } f(\boldsymbol{x}) - f^* \leq M,$$

where

$$\kappa = \theta^2(1+\rho)\left(\frac{1 + 2\|\nabla g(\boldsymbol{t}^*)\|^2}{\sigma_g} + 4M\right) + 2\theta\|\nabla f(\bar{\boldsymbol{x}})\|,$$

and $\theta \equiv \theta\left(A, \binom{E}{\boldsymbol{b}^\top}\right)$ is defined in Lemma 4.3. Specially, when $\boldsymbol{b} = \boldsymbol{0}$ or $\mathcal{X} = \mathbb{R}^l$,

$$\kappa = \theta(A, E)^2 \frac{1+\rho}{\sigma_g}.$$

Proof Consider the following polyhedron of the optimal solutions,

$$\mathcal{X}^* = \{\boldsymbol{x}^* \mid E\boldsymbol{x}^* = \boldsymbol{t}^*,\ \boldsymbol{b}^\top \boldsymbol{x}^* = s^*,\ A\boldsymbol{x}^* \leq \boldsymbol{d}\},$$

where $\boldsymbol{t}^*$ and s^* are values described in Lemma 4.2. We can then apply Lemma 4.3 to have for any $\boldsymbol{x}$, there exists $\boldsymbol{x}^* \in \mathcal{X}^*$ such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \theta\left(A, \binom{E}{\boldsymbol{b}^\top}\right) \left\| \begin{bmatrix} [A\boldsymbol{x} - \boldsymbol{d}]_P^+ \\ E\boldsymbol{x} - \boldsymbol{t}^* \\ \boldsymbol{b}^\top \boldsymbol{x} - s^* \end{bmatrix} \right\|, \tag{40}$$

where $\theta\left(A, \binom{E}{\boldsymbol{b}^\top}\right)$, independent of $\boldsymbol{x}$, is defined in Lemma 4.3. Denote $\theta\left(A, \binom{E}{\boldsymbol{b}^\top}\right)$ as θ for simplicity. By considering only feasible $\boldsymbol{x}$ and using the definition of $\bar{\boldsymbol{x}}$, (40) implies

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \leq \theta^2(\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2 + (\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2), \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

With Lemmas 4.4 and 4.5, if $f(\boldsymbol{x}) - f^* \leq M$, we can bound $\|E\boldsymbol{x} - \boldsymbol{t}^*\|^2$ and $(\boldsymbol{b}^\top \boldsymbol{x} - s^*)^2$ to obtain

$$\begin{aligned} &\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 \\ &\leq \theta^2(1+\rho)\left(\frac{1 + 2\|\nabla g(\boldsymbol{t}^*)\|^2}{\sigma_g} + 4M\right)\|\nabla^+ f(\boldsymbol{x})\|\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + 4\theta^2\|\nabla f(\bar{\boldsymbol{x}})\|^2\|\nabla^+ f(\boldsymbol{x})\|^2. \end{aligned} \tag{41}$$

Let

$$\begin{aligned} &a \equiv \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \quad c \equiv 2\theta\|\nabla f(\bar{\boldsymbol{x}})\|\|\nabla^+ f(\boldsymbol{x})\|, \quad \text{and} \\ &b \equiv \theta^2(1+\rho)\left(\frac{1 + 2\|\nabla g(\boldsymbol{t}^*)\|^2}{\sigma_g} + 4M\right)\|\nabla^+ f(\boldsymbol{x})\|. \end{aligned} \tag{42}$$

Then we can rewrite (41) as

$$a^2 \leq ba + c^2 \quad \text{with} \quad a \geq 0,\ b \geq 0,\ c \geq 0. \tag{43}$$

We claim that

$$a \leq b + c. \tag{44}$$

Otherwise, $a > b + c$ implies that

$$a^2 > a(b+c) > ba + c^2,$$

a violation to (43). By (42) and (44), the proof is complete.

Now we examine the special case of $\boldsymbol{b} = \mathbf{0}$ or $\mathcal{X} = \mathbb{R}^l$. From (31) in Lemma 4.2, we can apply Lemma 4.3 to have the existence of $\theta(A, E)$ such that $\forall \boldsymbol{x} \in \mathcal{X}$, there is $\boldsymbol{x}^* \in \mathcal{X}^*$ so that

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \theta(A, E)\|E\boldsymbol{x} - \boldsymbol{t}^*\|.$$

With Lemma 4.4, we have

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 \leq \theta(A, E)^2 \frac{1+\rho}{\sigma_g} \|\nabla^+ f(\boldsymbol{x})\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|.$$

After canceling $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|$ from both sides, the proof is complete. ■

5. Proof of Theorem 2.8

The proof is modified from Theorem 3.1 of Luo and Tseng (1993). They applied a local error bound to obtain asymptotic local linear convergence, while ours applies a global error bound to have linear convergence from the first iteration.

By (9) and Lemma A.2, we have

$$\begin{aligned}
&\|\boldsymbol{x}^r - [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\| \\
&\leq \|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\| + \|\boldsymbol{x}^{r+1} - [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\| \\
&= \|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\| + \|[\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r) + \boldsymbol{e}^r]^+_{\mathcal{X}} - [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\| \\
&\leq \|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\| + \|\boldsymbol{e}^r\|. \qquad (45)
\end{aligned}$$

By Lemma A.8, the left-hand side of above inequality could be bounded below by

$$\underline{\omega}\|\boldsymbol{x}^r - [\boldsymbol{x}^r - \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\| \leq \|\boldsymbol{x}^r - [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\|,$$

where $\underline{\omega} = \min(1, \inf_r \omega_r)$. With Theorems 4.1 or 4.6, (45), and (10), we have

$$\|\boldsymbol{x}^r - \bar{\boldsymbol{x}}^r\| \leq \kappa \|\nabla^+ f(\boldsymbol{x}^r)\| \leq \kappa \frac{\|\boldsymbol{x}^r - [\boldsymbol{x}^r - \omega_r \nabla f(\boldsymbol{x}^r)]^+_{\mathcal{X}}\|}{\underline{\omega}} \leq \kappa \frac{1+\beta}{\underline{\omega}} \|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\|, \qquad (46)$$

where $\bar{\boldsymbol{x}}^r$ is the projection of $\boldsymbol{x}^r$ to the optimal set.

$$\bar{\boldsymbol{x}}^r \equiv [\boldsymbol{x}^r]^+_{\mathcal{X}^*}.$$

Next, we bound $f(\boldsymbol{x}^{r+1}) - f(\bar{\boldsymbol{x}}^r)$. Lemma A.1 and the definition of $\boldsymbol{x}^{r+1}$ imply that

$$\left(\boldsymbol{x}^r - \boldsymbol{x}^{r+1} + \boldsymbol{e}^r\right)^\top \left(\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r\right) \geq \omega_r \nabla f(\boldsymbol{x}^r)^\top (\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r). \qquad (47)$$

From the convexity of $f(\boldsymbol{x})$,

$$\begin{aligned}
&f(\boldsymbol{x}^{r+1}) - f(\bar{\boldsymbol{x}}^r) \leq \nabla f(\boldsymbol{x}^{r+1})^\top (\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r) \\
&= (\nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r))^\top (\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r) + \nabla f(\boldsymbol{x}^r)^\top (\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r) \\
&\leq \|\nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r)\| \|\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r\| + \frac{1}{\omega_r} (\boldsymbol{x}^r - \boldsymbol{x}^{r+1} + \boldsymbol{e}^r)^\top (\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r) \qquad (48) \\
&\leq \left(\rho \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\| + \frac{1}{\underline{\alpha}} \|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\| + \frac{1}{\underline{\alpha}} \|\boldsymbol{e}^r\|\right) \|\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r\|. \qquad (49)
\end{aligned}$$

Inequality (48) is from (47), and (49) follows from the Lipschitz continuity of $\nabla f(\boldsymbol{x})$. In addition,

$$\|\boldsymbol{x}^{r+1} - \bar{\boldsymbol{x}}^r\| \leq \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\| + \|\boldsymbol{x}^r - \bar{\boldsymbol{x}}^r\|. \tag{50}$$

From (46), (10), and (50), each term in (49) is bounded by $\|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\|$. Therefore,

$$f(\boldsymbol{x}^{r+1}) - f(\bar{\boldsymbol{x}}^r) \leq \phi\|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\|^2, \quad \text{where } \phi = (\rho + \frac{1+\beta}{\underline{\omega}})(1 + \kappa\frac{1+\beta}{\underline{\omega}}).$$

From (11) and the above inequality,

$$f(\boldsymbol{x}^{r+1}) - f(\bar{\boldsymbol{x}}^r) \leq \frac{\phi}{\phi+\gamma}\left(f(\boldsymbol{x}^r) - f(\bar{\boldsymbol{x}}^r)\right), \quad \forall r.$$

Because $f(\boldsymbol{x})$ is convex, $f(\bar{\boldsymbol{x}}^r)$, $\forall r$ correspond to the same unique optimal function value. Thus the global linear convergence is established.

6. Discussions and Conclusions

For future research, we plan to extend the analysis to other types of algorithms and problems (e.g., L1-regularized problems). Further, the global error bound will be useful in analyzing stopping criteria and the effect of parameter changes on the running time of machine learning problems (for example, the change of parameter C in SVM).

In conclusion, by focusing on a convex but non-strongly convex problem (1), we established a global error bound. We then proved the global linear convergence on a wide range of deterministic algorithms, including cyclic coordinate descent methods for dual SVM and SVR. Consequently, the time complexity of these algorithms is $O(\log(1/\epsilon))$.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant 101-2221-E-002-199-MY3. The authors thank associate editor and anonymous reviewers for valuable comments. We also thank Pinghua Gong for pointing out a mistake in the proof of Lemma 4.5, and Lijun Zhang for a mistake in Section 3.4.

Appendix A. Properties of Projected Gradient

We present some properties of projected gradient used in the proofs. Most of them are known in the literature, but we list them here for completeness. Throughout this section, we assume $\mathcal{X}$ is a non-empty, closed, and convex set.

First, we present a fundamental result used in the paper: the projection theorem to a non-empty closed convex set $\mathcal{X}$. The convex projection in Definition 2.3 is equivalent to the following inequality on the right-hand side of (51). That is, if the inequality holds for any $\boldsymbol{z}$, this $\boldsymbol{z}$ will be the result of the convex projection and vise versa.

Lemma A.1 (Projection Theorem)

$$\boldsymbol{z} = [\boldsymbol{x}]^+_{\mathcal{X}} \Leftrightarrow (\boldsymbol{z} - \boldsymbol{x})^\top(\boldsymbol{z} - \boldsymbol{y}) \leq 0, \ \forall \boldsymbol{y} \in \mathcal{X}. \tag{51}$$

Proof The proof is modified from Hiriart-Urruty and Lemaréchal (2001, Theorem 3.1.1).
From the convexity of $\mathcal{X}$,

$$\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{z} \in \mathcal{X}, \quad \forall \boldsymbol{y} \in \mathcal{X},\ \forall \alpha \in [0,1].$$

By Definition 2.3,

$$\|\boldsymbol{x}-\boldsymbol{z}\|^2 \leq \|\boldsymbol{x} - (\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{z})\|^2, \quad \forall \boldsymbol{y} \in \mathcal{X},\ \forall \alpha \in [0,1].$$

The inequality can be written as

$$0 \leq \alpha(\boldsymbol{z}-\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{z}) + \frac{1}{2}\alpha^2\|\boldsymbol{y}-\boldsymbol{z}\|^2.$$

Divide α from both sides, and let $\alpha \downarrow 0$. Then we have ($\Rightarrow$).

For ($\Leftarrow$), if $\boldsymbol{z} = \boldsymbol{x}$, then $0 = \|\boldsymbol{z}-\boldsymbol{x}\| \leq \|\boldsymbol{y}-\boldsymbol{x}\|$ holds for all $\boldsymbol{y} \in \mathcal{X}$. Thus, $\boldsymbol{z} = [\boldsymbol{x}]_{\mathcal{X}}^+$. If $\boldsymbol{z} \neq \boldsymbol{x}$, then for any $\boldsymbol{y} \in \mathcal{X}$,

$$\begin{aligned} 0 &\geq (\boldsymbol{z}-\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{z}\|^2 + (\boldsymbol{y}-\boldsymbol{x})^\top(\boldsymbol{x}-\boldsymbol{z}) \\ &\geq \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \|\boldsymbol{x}-\boldsymbol{y}\|\|\boldsymbol{x}-\boldsymbol{z}\|. \end{aligned}$$

Divide $\|\boldsymbol{x}-\boldsymbol{z}\| > 0$ from both sides. Because the inequality is valid for all $\boldsymbol{y}$, ($\Leftarrow$) holds. ■

The following lemma shows that the projection operator is Lipschitz continuous.

Lemma A.2 (Lipschitz Continuity of Convex Projection)

$$\|[\boldsymbol{x}]_{\mathcal{X}}^+ - [\boldsymbol{y}]_{\mathcal{X}}^+\| \leq \|\boldsymbol{x}-\boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

Proof The proof is modified from Hiriart-Urruty and Lemaréchal (2001) Proposition 3.1.3.

Let $\boldsymbol{u} = [\boldsymbol{x}]_{\mathcal{X}}^+$ and $\boldsymbol{v} = [\boldsymbol{y}]_{\mathcal{X}}^+$. If $\boldsymbol{u} = \boldsymbol{v}$, then the result holds immediately. If not, with Lemma A.1 we have

$$(\boldsymbol{u}-\boldsymbol{x})^\top(\boldsymbol{u}-\boldsymbol{v}) \leq 0, \tag{52}$$

$$(\boldsymbol{v}-\boldsymbol{y})^\top(\boldsymbol{v}-\boldsymbol{u}) \leq 0. \tag{53}$$

Summing (52) and (53), we have

$$(\boldsymbol{u}-\boldsymbol{v})^\top(\boldsymbol{u}-\boldsymbol{x}-\boldsymbol{v}+\boldsymbol{y}) \leq 0.$$

We could rewrite it as

$$\|\boldsymbol{u}-\boldsymbol{v}\|^2 \leq (\boldsymbol{u}-\boldsymbol{v})^\top(\boldsymbol{x}-\boldsymbol{y}) \leq \|\boldsymbol{u}-\boldsymbol{v}\|\|\boldsymbol{x}-\boldsymbol{y}\|.$$

Cancel $\|\boldsymbol{u}-\boldsymbol{v}\| > 0$ at both sides. Then the result holds. ■

Lemma A.3 *Assume $\nabla f(\boldsymbol{x})$ is ρ Lipschitz continuous. Then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,*

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top(\boldsymbol{x}-\boldsymbol{y}) \leq (1+\rho)\|\nabla^+ f(\boldsymbol{x}) - \nabla^+ f(\boldsymbol{y})\|\|\boldsymbol{x}-\boldsymbol{y}\|.$$

Proof For simplification, we will use $\nabla_{\boldsymbol{x}} \equiv \nabla f(\boldsymbol{x})$ and $\nabla^+_{\boldsymbol{x}} \equiv \nabla^+ f(\boldsymbol{x})$ in this proof. From Lemma A.1,

$$([\boldsymbol{x} - \nabla_{\boldsymbol{x}}]^+_{\mathcal{X}} - \boldsymbol{x} + \nabla_{\boldsymbol{x}})^\top ([\boldsymbol{x} - \nabla_{\boldsymbol{x}}]^+_{\mathcal{X}} - [\boldsymbol{y} - \nabla_{\boldsymbol{y}}]^+_{\mathcal{X}}) \leq 0.$$

With the definition of $\nabla^+ f(\boldsymbol{x})$, this inequality can be rewritten as

$$(\nabla_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{x}})^\top (\boldsymbol{x} - \nabla^+_{\boldsymbol{x}} - \boldsymbol{y} + \nabla^+_{\boldsymbol{y}}) \leq 0.$$

Further, we have

$$\nabla_{\boldsymbol{x}}{}^\top (\boldsymbol{x} - \boldsymbol{y}) \leq \nabla^+_{\boldsymbol{x}}{}^\top (\boldsymbol{x} - \boldsymbol{y}) + \nabla_{\boldsymbol{x}}{}^\top (\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}) - \nabla^+_{\boldsymbol{x}}{}^\top (\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}). \tag{54}$$

Similarly,

$$\nabla_{\boldsymbol{y}}{}^\top (\boldsymbol{y} - \boldsymbol{x}) \leq \nabla^+_{\boldsymbol{y}}{}^\top (\boldsymbol{y} - \boldsymbol{x}) + \nabla_{\boldsymbol{y}}{}^\top (\nabla^+_{\boldsymbol{y}} - \nabla^+_{\boldsymbol{x}}) - \nabla^+_{\boldsymbol{y}}{}^\top (\nabla^+_{\boldsymbol{y}} - \nabla^+_{\boldsymbol{x}}). \tag{55}$$

Summing (54) and (55) leads to

$$\begin{aligned}
&(\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
\leq\ &(\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}})^\top (\boldsymbol{x} - \boldsymbol{y}) + (\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}})^\top (\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}) - \|\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}\|^2 \\
\leq\ &(\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}})^\top (\boldsymbol{x} - \boldsymbol{y}) + (\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}})^\top (\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}).
\end{aligned}$$

With $\nabla f(\boldsymbol{x})$ being ρ Lipschitz continuous, we have

$$\begin{aligned}
(\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}})^\top (\boldsymbol{x} - \boldsymbol{y}) &\leq \|\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}\| (\|\boldsymbol{x} - \boldsymbol{y}\| + \|\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}}\|) \\
&\leq (1 + \rho) \|\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}\| \|\boldsymbol{x} - \boldsymbol{y}\|.
\end{aligned}$$

■

The next two lemmas correspond to the strong convexity and Lipschitz continuity of projected gradient.

Lemma A.4 *If $f(\boldsymbol{x})$ is σ strongly convex and $\nabla f(\boldsymbol{x})$ is ρ Lipschitz continuous,*

$$\frac{\sigma}{1 + \rho} \|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\nabla^+ f(\boldsymbol{x}) - \nabla^+ f(\boldsymbol{y})\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}.$$

Proof With the strong convexity and Lemma A.3,

$$\sigma \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq (\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{y}})^\top (\boldsymbol{x} - \boldsymbol{y}) \leq (1 + \rho) \|\nabla^+_{\boldsymbol{x}} - \nabla^+_{\boldsymbol{y}}\| \|\boldsymbol{x} - \boldsymbol{y}\|.$$

If $\boldsymbol{x} \neq \boldsymbol{y}$, we have the result after canceling $\|\boldsymbol{x} - \boldsymbol{y}\|$ from both sides. For the situation of $\boldsymbol{x} = \boldsymbol{y}$, the result obviously holds. ■

Lemma A.5 (Lipschitz Continuity of Projected Gradient) *If $\nabla f(\boldsymbol{x})$ is ρ Lipschitz continuous, then*

$$\|\nabla^+ f(\boldsymbol{x}) - \nabla^+ f(\boldsymbol{y})\| \leq (2 + \rho) \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}.$$

Proof By the definition of projected gradient and Lemma A.2,

$$\begin{aligned}\|\nabla^+ f(\boldsymbol{x}) - \nabla^+ f(\boldsymbol{y})\| &\leq \|\boldsymbol{x} - \boldsymbol{y}\| + \|[\boldsymbol{x} - \nabla f(\boldsymbol{x})]_{\mathcal{X}}^+ - [\boldsymbol{y} - \nabla f(\boldsymbol{y})]_{\mathcal{X}}^+\| \\ &\leq \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{x} - \boldsymbol{y}\| + \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \\ &\leq (2 + \rho)\|\boldsymbol{x} - \boldsymbol{y}\|.\end{aligned}$$

The last inequality follows from the ρ Lipschitz continuity of $\nabla f(\boldsymbol{x})$. ■

A useful property of projected gradient is to test whether a solution is optimal; see the following lemma.

Lemma A.6 *For any $\boldsymbol{x} \in \mathcal{X}$,*

$$\boldsymbol{x} \text{ is optimal for problem (4)} \Leftrightarrow \nabla^+ f(\boldsymbol{x}) = \boldsymbol{0}.$$

Proof From Lemma A.1 and the definition of $\nabla^+ f(\boldsymbol{x})$,

$$\begin{aligned}&\nabla^+ f(\boldsymbol{x}) = \boldsymbol{0} \Leftrightarrow \boldsymbol{x} = [\boldsymbol{x} - \nabla f(\boldsymbol{x})]_{\mathcal{X}}^+ \\ \Leftrightarrow\ &(\boldsymbol{x} - (\boldsymbol{x} - \nabla f(\boldsymbol{x})))^\top (\boldsymbol{x} - \boldsymbol{y}) \leq 0, \quad \forall \boldsymbol{y} \in \mathcal{X} \\ \Leftrightarrow\ &\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \geq 0, \quad \forall \boldsymbol{y} \in \mathcal{X} \\ \Leftrightarrow\ &\boldsymbol{x} \text{ is optimal.}\end{aligned}$$

The last relation follows from the optimality condition of convex programming problems. ■

The next two lemmas discuss properties of projected gradient defined with different scalars on the negative gradient direction.

Lemma A.7 *$\forall \boldsymbol{x} \in \mathcal{X}$,*

$$\|\boldsymbol{x} - [\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})]_{\mathcal{X}}^+\| \text{ is monotonically increasing for all } \alpha > 0.^{7}$$

Proof Let

$$\boldsymbol{u} = \boldsymbol{x} - \alpha_1 \nabla f(\boldsymbol{x}), \tag{56}$$
$$\boldsymbol{v} = \boldsymbol{x} - \alpha_2 \nabla f(\boldsymbol{x}), \tag{57}$$

where $0 < \alpha_1 < \alpha_2$. By Lemma A.1, we have

$$([\boldsymbol{u}]_{\mathcal{X}}^+ - \boldsymbol{u})^\top ([\boldsymbol{u}]_{\mathcal{X}}^+ - [\boldsymbol{v}]_{\mathcal{X}}^+) \leq 0, \tag{58}$$
$$([\boldsymbol{v}]_{\mathcal{X}}^+ - \boldsymbol{v})^\top ([\boldsymbol{v}]_{\mathcal{X}}^+ - [\boldsymbol{u}]_{\mathcal{X}}^+) \leq 0. \tag{59}$$

Let $\boldsymbol{z} = [\boldsymbol{u}]_{\mathcal{X}}^+ - [\boldsymbol{v}]_{\mathcal{X}}^+$. Expanding the definition of $\boldsymbol{u}$ and $\boldsymbol{v}$ leads to

$$\alpha_1 \nabla f(\boldsymbol{x})^\top \boldsymbol{z} \leq (\boldsymbol{x} - [\boldsymbol{u}]_{\mathcal{X}}^+)^\top \boldsymbol{z} \leq (\boldsymbol{x} - [\boldsymbol{v}]_{\mathcal{X}}^+)^\top \boldsymbol{z} \leq \alpha_2 \nabla f(\boldsymbol{x})^\top \boldsymbol{z}, \tag{60}$$

7. The proof is modified from http://math.stackexchange.com/questions/201168/projection-onto-closed-convex-set .

where the first and the last inequalities are from (58) and (59), respectively, and the second inequality is from $([\boldsymbol{u}]^+_{\mathcal{X}}-[\boldsymbol{v}]^+_{\mathcal{X}})^\top \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{z} \geq 0$. With $0 < \alpha_1 < \alpha_2$, (60) implies $\nabla f(\boldsymbol{x})^\top \boldsymbol{z} \geq 0$ and

$$(\boldsymbol{x} - [\boldsymbol{u}]^+_{\mathcal{X}})^\top \boldsymbol{z} \geq 0.$$

Using this inequality,

$$\|\boldsymbol{x} - [\boldsymbol{v}]^+_{\mathcal{X}}\|^2 = \|\boldsymbol{x} - [\boldsymbol{u}]^+_{\mathcal{X}} + \boldsymbol{z}\|^2 = \|\boldsymbol{x} - [\boldsymbol{u}]^+_{\mathcal{X}}\|^2 + 2(\boldsymbol{x} - [\boldsymbol{u}]^+_{\mathcal{X}})^\top \boldsymbol{z} + \|\boldsymbol{z}\|^2 \geq \|\boldsymbol{x} - [\boldsymbol{u}]^+_{\mathcal{X}}\|^2.$$

Therefore, from (56)-(57),

$$\|\boldsymbol{x} - [\boldsymbol{x} - \alpha_2 \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\| \geq \|\boldsymbol{x} - [\boldsymbol{x} - \alpha_1 \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\|.$$

With $0 < \alpha_1 < \alpha_2$, the proof is complete. ■

Lemma A.8 *$\forall \boldsymbol{x} \in \mathcal{X}$ and $\alpha > 0$, if*

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{x} - [\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}, \\ \boldsymbol{v} &= \boldsymbol{x} - [\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}, \end{aligned}$$

then

$$\min(1, \alpha)\|\boldsymbol{u}\| \leq \|\boldsymbol{v}\| \leq \max(1, \alpha)\|\boldsymbol{u}\|.$$

Proof From Lemma 1 in Gafni and Bertsekas (1984), $\|\boldsymbol{x} - [\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\|/\alpha$ is monotonically decreasing for all $\alpha > 0$. Thus,

$$\alpha\|\boldsymbol{x} - [\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\| \leq \|\boldsymbol{x} - [\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\|, \quad \forall \alpha \leq 1.$$

From Lemma A.7, we have

$$\|\boldsymbol{x} - [\boldsymbol{x} - \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\| \leq \|\boldsymbol{x} - [\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})]^+_{\mathcal{X}}\|, \quad \forall \alpha \geq 1.$$

Therefore, $\min(1, \alpha)\|\boldsymbol{u}\| \leq \|\boldsymbol{v}\|$. A similar proof applies to $\|\boldsymbol{v}\| \leq \max(1, \alpha)\|\boldsymbol{u}\|$. ■

Appendix B. Proof of Hoffman's Bound (Lemma 4.3)

The following proof is a special case of Mangasarian and Shiau (1987) and Li (1994), which bounds the distance of a point to the polyhedron by the violation of inequalities. We begin with an elementary theorem in convex analysis.

Lemma B.1 (Carathèodory's Theorem) *For a non-empty polyhedron*

$$A^\top \boldsymbol{u} + E^\top \boldsymbol{v} = \boldsymbol{y}, \;\; \boldsymbol{u} \geq \boldsymbol{0}, \tag{61}$$

there is a feasible point $(\boldsymbol{u}, \boldsymbol{v})$ such that

The corresponding rows of A, E to $\boldsymbol{u}$, $\boldsymbol{v}$'s non-zero elements are linearly independent. (62)

Proof Let $(\boldsymbol{u}, \boldsymbol{v})$ be a point in the polyhedron, and therefore $E^\top \boldsymbol{v} = \boldsymbol{y} - A^\top \boldsymbol{u}$. If the corresponding rows of E to non-zero elements of $\boldsymbol{v}$ are not linearly independent, we can modify $\boldsymbol{v}$ so that $E^\top \boldsymbol{v}$ remains the same and E's rows corresponding to $\boldsymbol{v}$'s non-zero elements are linearly independent. Thus, without loss of generality, we assume that E is full row-rank. Denote $\boldsymbol{a}_i^\top$ as the ith row of A and $\boldsymbol{e}_j^\top$ as the jth row of E. If the corresponding rows of A, E to non-zero elements of $\boldsymbol{u}, \boldsymbol{v}$ are not linearly independent, there exists $(\boldsymbol{\lambda}, \boldsymbol{\xi})$ such that

1. $(\boldsymbol{\lambda}, \boldsymbol{\xi}) \neq \boldsymbol{0}$.
2. $(\boldsymbol{\lambda}, \boldsymbol{\xi})$'s non-zero elements correspond to the non-zero elements of $(\boldsymbol{u}, \boldsymbol{v})$. That is, $\lambda_i = 0$ if $u_i = 0$, $\forall i$, and $\xi_j = 0$ if $v_j = 0$, $\forall j$.
3. $(\boldsymbol{\lambda}, \boldsymbol{\xi})$ satisfies
$$\sum_{i:\, u_i>0,\, \lambda_i \neq 0} \lambda_i \boldsymbol{a}_i + \sum_{j:\, v_j \neq 0,\, \xi_j \neq 0} \xi_j \boldsymbol{e}_j = \boldsymbol{0}.$$

Besides, the set $\{i \mid u_i > 0,\ \lambda_i \neq 0\}$ is not empty because the rows of E are linearly independent. Otherwise, a contradiction occurs from $\boldsymbol{\lambda} = \boldsymbol{0}$, $\boldsymbol{\xi} \neq \boldsymbol{0}$, and

$$\sum_{j:\, v_j \neq 0,\, \xi_j \neq 0} \xi_i \boldsymbol{e}_j = \boldsymbol{0}.$$

By choosing

$$s = \min_{i:\, u_i>0,\, \lambda_i \neq 0} \frac{u_i}{\lambda_i} > 0,$$

we have

$$A^\top(\boldsymbol{u} - s\boldsymbol{\lambda}) + E^\top(\boldsymbol{v} - s\boldsymbol{\xi}) = A^\top \boldsymbol{u} + E^\top \boldsymbol{v} = \boldsymbol{y} \quad \text{and} \quad \boldsymbol{u} - s\boldsymbol{\lambda} \geq \boldsymbol{0}.$$

This means that $(\boldsymbol{u} - s\boldsymbol{\lambda}, \boldsymbol{v} - s\boldsymbol{\xi})$ is also a member of the polyhedron (61) and has less non-zero elements than $(\boldsymbol{u}, \boldsymbol{v})$. The process could be repeatedly applied until there is a point satisfying the linearly independent condition (62). Thus, if the polyhedron is not empty, we can always find a $(\boldsymbol{u}, \boldsymbol{v})$ such that its non-zero elements correspond to linearly independent rows in (A, E). ■

Now we prove Hoffman's bound (Lemma 4.3) by Carathèodory's theorem and the KKT optimality condition of a convex projection problem.

Proof If $\boldsymbol{x}$ is in the polyhedron, we can take $\boldsymbol{x}^* = \boldsymbol{x}$ and the inequality (33) holds naturally for every positive θ. Now if $\boldsymbol{x}$ does not belong to the polyhedron, consider the following convex projection problem

$$\min_{\boldsymbol{p}} \ \|\boldsymbol{p} - \boldsymbol{x}\|, \quad \text{subject to } A\boldsymbol{p} \leq \boldsymbol{d},\ E\boldsymbol{p} = \boldsymbol{t}. \tag{63}$$

The polyhedron is assumed to be non-empty, so a unique optimal solution $\boldsymbol{x}^*$ of this problem exists. Because $\boldsymbol{x}$ is not in the polyhedron, we have $\boldsymbol{x}^* \neq \boldsymbol{x}$. Then by the KKT optimality

condition, a unique optimal $\boldsymbol{x}^*$ for (63) happens only if there are $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ such that

$$\begin{aligned}&\frac{\boldsymbol{x}^*-\boldsymbol{x}}{\|\boldsymbol{x}^*-\boldsymbol{x}\|}=-A^\top\boldsymbol{u}^*-E^\top\boldsymbol{v}^*,\quad \boldsymbol{u}^*\geq\boldsymbol{0},\\&A\boldsymbol{x}^*\leq\boldsymbol{d},\quad E\boldsymbol{x}^*=\boldsymbol{t},\quad u_i^*(A\boldsymbol{x}^*-\boldsymbol{d})_i=0,\ \forall i=1,\ldots,l.\end{aligned}$$

Denote

$$I=\{i\mid(A\boldsymbol{x}^*-\boldsymbol{d})_i=0\}.$$

Because $u_i^*=0,\forall i\notin I$, $(\boldsymbol{u}_I^*,\boldsymbol{v}^*)$ is a feasible point of the following polyhedron.

$$-A_I^\top\boldsymbol{u}_I-E^\top\boldsymbol{v}=\frac{\boldsymbol{x}^*-\boldsymbol{x}}{\|\boldsymbol{x}^*-\boldsymbol{x}\|},\quad \boldsymbol{u}_I\geq\boldsymbol{0},\tag{64}$$

where A_I is a sub-matrix of A's rows corresponding to I. Then the polyhedron in (64) is non-empty. From Lemma B.1, there exists a feasible $(\hat{\boldsymbol{u}}_I,\hat{\boldsymbol{v}})$ such that

$$\text{The corresponding rows of } A_I,E \text{ to non-zero } \hat{\boldsymbol{u}}_I,\hat{\boldsymbol{v}} \text{ are linearly independent.}\tag{65}$$

Expand $\hat{\boldsymbol{u}}_I$ to a vector $\hat{\boldsymbol{u}}$ so that

$$\hat{u}_i=0,\ \forall i\notin I.\tag{66}$$

Then (65) becomes

$$\text{The corresponding rows of } A,E \text{ to non-zero } \hat{\boldsymbol{u}},\hat{\boldsymbol{v}} \text{ are linearly independent.}\tag{67}$$

By multiplying $(\boldsymbol{x}^*-\boldsymbol{x})^\top$ on the first equation of (64), we have

$$\|\boldsymbol{x}^*-\boldsymbol{x}\|=\hat{\boldsymbol{u}}^\top A(\boldsymbol{x}-\boldsymbol{x}^*)+\hat{\boldsymbol{v}}^\top E(\boldsymbol{x}-\boldsymbol{x}^*)=\hat{\boldsymbol{u}}^\top(A\boldsymbol{x}-\boldsymbol{d})+\hat{\boldsymbol{v}}^\top(E\boldsymbol{x}-\boldsymbol{t}).\tag{68}$$

The last equality is from $E\boldsymbol{x}^*=\boldsymbol{t}$ and (66). Further, by the non-negativity of $\hat{\boldsymbol{u}}$,

$$\hat{\boldsymbol{u}}^\top(A\boldsymbol{x}-\boldsymbol{d})\leq\hat{\boldsymbol{u}}^\top[A\boldsymbol{x}-\boldsymbol{d}]_P^+.\tag{69}$$

From (68) and (69),

$$\|\boldsymbol{x}^*-\boldsymbol{x}\|\leq\hat{\boldsymbol{u}}^\top[A\boldsymbol{x}-\boldsymbol{d}]_P^++\hat{\boldsymbol{v}}^\top(E\boldsymbol{x}-\boldsymbol{t})\leq\left\|\begin{matrix}\hat{\boldsymbol{u}}\\\hat{\boldsymbol{v}}\end{matrix}\right\|\left\|\begin{matrix}[A\boldsymbol{x}-\boldsymbol{d}]_P^+\\E\boldsymbol{x}-\boldsymbol{t}\end{matrix}\right\|.\tag{70}$$

Next we bound $\left\|\begin{smallmatrix}\hat{\boldsymbol{u}}\\\hat{\boldsymbol{v}}\end{smallmatrix}\right\|$. With (64) and (67), we have

$$\|A^\top\hat{\boldsymbol{u}}+E^\top\hat{\boldsymbol{v}}\|=1\text{ and }\left\|\begin{smallmatrix}\hat{\boldsymbol{u}}\\\hat{\boldsymbol{v}}\end{smallmatrix}\right\|\leq\theta(A,E),$$

where $\theta(A,E)$ is defined in (34). Together with (70), the proof is complete. ■

Note that this version of Hoffman's bound is not the sharpest one. For a more complex but tighter bound, please refer to Li (1994).

Appendix C. Strictly Diagonally Dominance and Positive Definiteness

Lemma C.1 *If a symmetric matrix Q is strictly diagonally dominant*

$$Q_{ii} > \sum_{j \neq i} |Q_{ij}|, \quad \forall i, \tag{71}$$

then it is positive definite. The reverse is not true.

Proof The result is modified from Rennie (2005). Because Q is symmetric,

$$Q = RDR^\top, \tag{72}$$

where R is an orthogonal matrix containing Q's eigen-vectors as its columns and D is a real-valued diagonal matrix containing Q's eigen-values. Let $\boldsymbol{u}$ be any eigen-vector of Q. We have $\boldsymbol{u} \neq \mathbf{0}$; otherwise, from (72), the corresponding $Q_{ii} = 0$ and (71) is violated. Let λ be the eigen-value such that $\lambda \boldsymbol{u} = Q\boldsymbol{u}$. Choose $i = \arg\max_j |u_j|$. Because $\boldsymbol{u} \neq \mathbf{0}$, we have either $u_i > 0$ or $u_i < 0$. If $u_i > 0$,

$$Q_{ij} u_j \geq -|Q_{ij}| u_i, \forall j \text{ and } \lambda u_i = \sum_j Q_{ij} u_j \geq (Q_{ii} - \sum_{j \neq i} |Q_{ij}|) u_i. \tag{73}$$

If $u_i < 0$,

$$Q_{ij} u_j \leq -|Q_{ij}| u_i, \forall j \text{ and } \lambda u_i = \sum_j Q_{ij} u_j \leq (Q_{ii} - \sum_{j \neq i} |Q_{ij}|) u_i. \tag{74}$$

By (73) and (74), we have $\lambda \geq Q_{ii} - \sum_{j \neq i} |Q_{ij}| > 0$. Therefore, Q is positive definite.

On the other hand, the following matrix

$$Q = \begin{pmatrix} 2 & 3 \\ 3 & 10 \end{pmatrix}$$

is positive definite but not diagonally dominant. Thus, the reverse is not true. ■

References

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate descent method for large-scale L2-loss linear SVM. *Journal of Machine Learning Research*, 9:1369–1398, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/papers/cdl2.pdf.

Eli M. Gafni and Dimitri P. Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22:936–964, 1984.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Clifford Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4:79–85, 1957.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Verlag, 2001.

Chia-Hua Ho and Chih-Jen Lin. Large-scale linear support vector regression. *Journal of Machine Learning Research*, 13:3323–3348, 2012. URL http://www.csie.ntu.edu.tw/~cjlin/papers/linear-svr.pdf.

Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the Twenty Fifth International Conference on Machine Learning (ICML)*, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/papers/cddual.pdf.

Wu Li. Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs. *SIAM Journal on Control and Optimization*, 32(1):140–153, 1994.

Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region Newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9:627–650, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/papers/logistic.pdf.

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

Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2): 408–425, 1992b.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46:157–178, 1993.

Olvi L. Mangasarian and Tzong-Huei Shiau. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization*, 25(3):583–595, 1987.

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

Jong-Shi Pang. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

Jason D. M. Rennie. Regularized logistic regression is strictly convex. Technical report, MIT, 2005. URL http://qwone.com/~jason/writing/convexLR.pdf.

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

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

Ludwig Seidel. Ueber ein verfahren, die gleichungen, auf welche die methode der kleinsten quadrate führt, sowie lineäre gleichungen überhaupt, durch successive annäherung aufzulösen. *Abhandlungen der Bayerischen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Abteilung*, 11(3):81–108, 1874.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1 regularized loss minimization. In *Proceedings of the Twenty Sixth International Conference on Machine Learning (ICML)*, 2009.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013a.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization, 2013b. arXiv preprint arXiv:1309.2375.

Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: complexity and preconditioning, 2013. arXiv preprint arXiv:1304.5530.

Paul Tseng and Sangwoon Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140:513–535, 2009.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, 1995.

Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Stephen J Wright. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186, 2012.