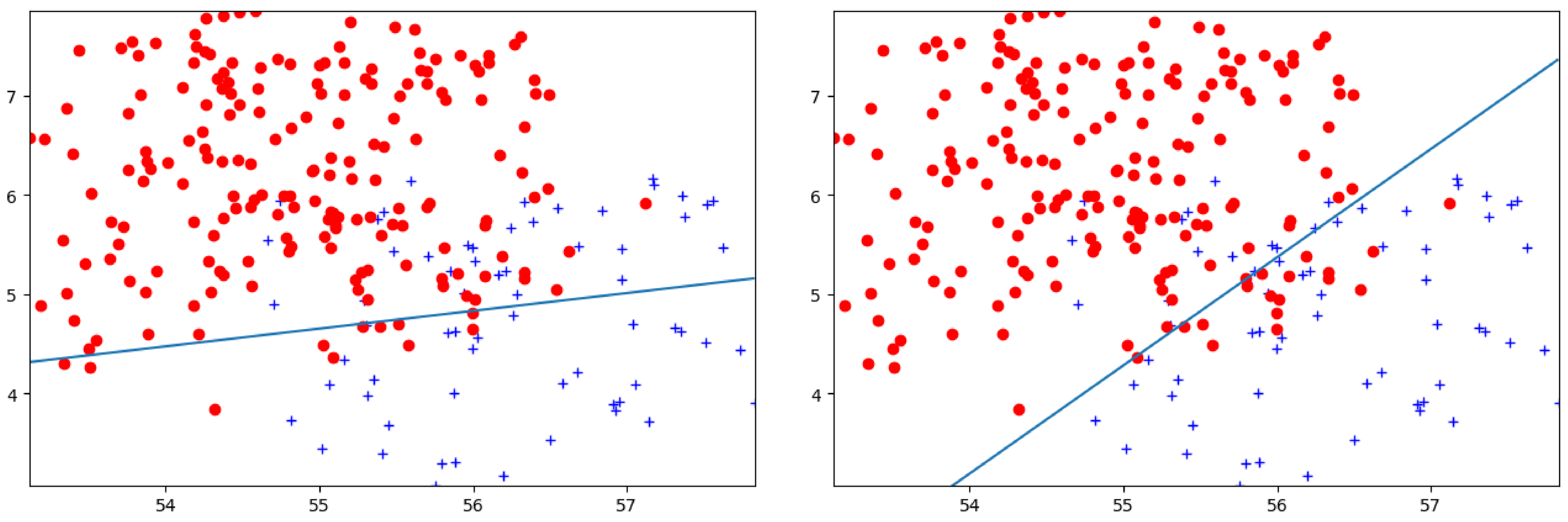
Linear Convergence and Support Vector Identification of Sequential Minimal Optimization

Jennifer She (UBC), Mark Schmidt (UBC)

OVERVIEW: Convergence Analysis of Sequential Minimal Optimization

Motivation:

- Support vector machines (SVMs) are widely used in many applictions.
- Sequential minimal optimization (SMO) has been a popular dual 2-coordinate ascent method for training SVMs for around 20 years.
- SMO can train SVMs with an unregularized bias, which is preferred when there are imbalanced class labels.



(a) regularized bias (b) unregularized bias

- This unregularized bias leads to a linear equality constraint across the dual variables,

$$\sum_{i=1}^{n} \alpha_i x_i = 0 \quad \text{where } \alpha_i \in \{-1, 1\}$$

in addition to the constraints

$$0 \leq x_i \leq c \quad \text{for all } i \in \{1, 2, ..., n\}$$

which complicates analysis of modern stochastic dual coordinate ascent (SDCA) methods.

This Work:

★ New convergence analysis of SMO with uniformly random coordinate selection (rSMO).

→ Show linear convergence of rSMO by generalizing convergence result of block coordinate descent (BCD) with linearly coupled constraints.

→ Previous works give sublinear rates.

→ Show that rSMO identifies the final set of support vectors in a finite number of iterations under mild conditions.

Problem of Interest

- We consider the SVM dual as an instance of the general problem

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad f(x), \tag{1}$$

where $\mathcal{X}$ is a set of the form $\{x \mid l \leq x \leq u, Ax = b\}$.

→ l and u are upper and lower bounds on the variables.

→ A is an $m \times n$ matrix where $m \leq n$ and b is a $m \times 1$ vector defining linear constraints.

- The gradient ∇f is L-Lipschitz continuous,

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \text{for all } x, y \in \mathcal{X}, \tag{2}$$

and the problem satisfies proximal-PL inequality (prox-PL), written in this case as

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(f(x) - f^*), \quad \text{for all } x, y \in \mathcal{X}, \text{ some } \mu > 0, \tag{3}$$

where

$$\mathcal{D}_g(x, \mu) \equiv -2\mu \, \text{argmin}_{y \in \mathcal{X}}\{\langle \nabla f(x), y - x\rangle + \frac{L}{2}||y - x||^2\}.$$

→ Prox-PL is weaker than strong convexity and always holds for SVMs by convexity of f and quadratic growth (QG) property.

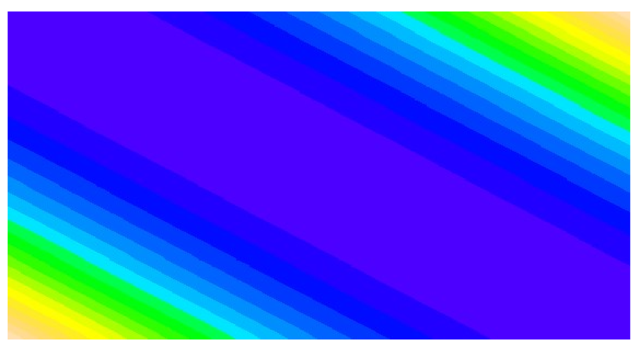


Figure: convex + QG but not strongly convex

Sequential Minimal Optimization

★ On each iteration, SMO chooses a block from the candidate block set

$$B = \{\{i, j\} | i, j \in \{1, 2, ..., n\}, i \neq j\}.$$

★ The iteration update corresponds to

$$x_i^{k+1} = \left\{ x_i - \frac{1}{H_{ii}^k + H_{jj}^k \pm 2H_{ij}^k}[\nabla_i f(x) \pm \nabla_j f(x)] \right\}_{\text{clipped}}$$

$$x_j^{k+1} = \left\{ x_j - \frac{1}{H_{jj}^k + H_{ii}^k \pm 2H_{ij}^k}[\nabla_j f(x) \pm \nabla_i f(x)] \right\}_{\text{clipped}}$$

where $\pm$ is $-\alpha_i \cdot \alpha_j$, and the updates are clipped to stay within the bounds $[\mathcal{L}_i^k, \mathcal{U}_i^k]$:

$$\mathcal{L}_i^k = \begin{cases} \max\{0, x_i^k - (c - x_j^k)\} & \alpha_i = \alpha_j \\ \max\{0, x_i^k - x_j^k\} & \alpha_i \neq \alpha_j \end{cases} \qquad \mathcal{U}_i^k = \begin{cases} \min\{c, x_i^k + x_j^k\} & \alpha_i = \alpha_j \\ \min\{c, x_i^k + (c - x_j^k)\} & \alpha_i \neq \alpha_j \end{cases}$$

and $[\mathcal{L}_j^k, \mathcal{U}_j^k]$ (with the indices swapped) respectively.

- $H^k = \nabla^2 f(x^k)$

→ Avoid $H_{ii}^k + H_{jj}^k \pm 2H_{ij}^k = 0$ without strong convexity by using $H^k = L\mathbb{I}$ instead.

Block Coordinate Descent

- SMO is an instance of general BCD with an iteration update given by

$$x^{k+1} = \underset{\{y_{b^k} \mid y \in \mathcal{X}\}}{\text{argmin}} \{f(x^k) + \langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}||y - x^k||^2_{H^k}\}. \tag{4}$$

- The candidate block set B contains the supports of the elementary vectors of null(A):

Definition

→ For $A \in \mathbb{R}^{m \times n}$ and $m \leq n$, an *elementary vector* of $\text{null}(A)$, is a vector $d \in \text{null}(A)$ such that

$$\forall d' \in \text{null}(A) \text{ that are conformal to } d, \ \text{supp}(d') = \text{supp}(d).$$

→ Let $d, d' \in \mathbb{R}^n$. Then d' is *conformal* to d if

$$\text{supp}(d') \subseteq \text{supp}(d) \text{ and } d'_j d_j \geq 0, \ \forall j = 1, ..., n.$$

$$\text{Let } d = \begin{pmatrix} d_1 \\ 0 \\ d_3 \\ d_4 \\ 0 \end{pmatrix}, \quad \text{then} \quad \underbrace{\begin{pmatrix} d_1 \\ 0 \\ 0 \\ 5 \cdot d_4 \\ 0 \end{pmatrix}}_{\text{conformal}} \quad \text{and} \quad \underbrace{\begin{pmatrix} d_1 \\ 0 \\ 0 \\ -5 \cdot d_4 \\ d_5 \end{pmatrix}}_{\text{not conformal}}.$$

- The matrix H^k satisfies the generalized inequality,

$$\nabla^2 f(x^k) \preceq H^k \preceq L\mathbb{I}. \tag{5}$$

Linear Convergence

Lemma

BCD with uniformly random block selection with updates (4) for problem (1) achieves

$$\mathbb{E}[f(x^k)] - f^* \leq \left(1 - \frac{\mu}{|B|L}\right)^k (f(x^0) - f^*). \tag{6}$$

Proof Outline:

Our argument closely follows the analysis of Necoara & Patrascu.

→ By Lipschitz-continuity of ∇f (2) and the property $H^k \preceq L\mathbb{I}$ (5) ,

$$\mathbb{E}[f(x^{k+1})] \leq f(x^k) + \frac{1}{|B|}\sum_{i=1}^{|B|} \min_{\{d_{b_i} \mid x^k_{b_i} + d_{b_i} \in \mathcal{X}\}} \left\{\langle \nabla f_{b_i}(x^k), d_{b_i}\rangle + \frac{L}{2}||d_{b_i}||^2\right\}. \tag{7}$$

→ By B containing the supports of the elementary vectors and properties of comformality,

$$(7) \leq f(x^k) + \frac{1}{|B|}\min_{\{d \mid x^k + d \in \mathcal{X}\}} \left\{\langle \nabla f(x^k), d\rangle + \frac{L}{2}||d||^2\right\}$$

$$= f(x^k) + \frac{1}{|B|}\min_{y \in \mathcal{X}} \left\{\langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}||y - x^k||^2\right\}. \tag{8}$$

→ By prox-PL inequality (3),

$$(8) \leq f(x^k) - \frac{1}{|B|}\frac{\mu}{L}(f(x^k) - f^*).$$

We can subtract f^* from both sides and rearrange the terms, and apply this recursively to get the linear convergence result (6).

Support Vector Identification

- We additionally require that the active set of the solution set X^* is unique,

Definition

The *active set* for SVMs is the set $\mathcal{Z} = \{i : x_i^* = 0 \text{ or } c \text{ for all } x^* \in X^*\}$.

and the non-degeneracy conditions:

→ $\nabla_i f(x^*) \neq 0$ for all $i \in Z$ and $x^* \in X^*$,

→ $|\nabla_i f(x^*)| \neq |\nabla_j f(x^*)|$ for all $i \in Z$, $j = 1, ..., n$, $i \neq j$ and $x^* \in X^*$.

Lemma

rSMO for problems satisfying the additional requirements above detects the final set of support vectors after some finite iterate $\mathcal{K}$.

Intuition:

→ Use induction on decreasing order of $|\nabla_i f(x^*)|$, for $i \in Z$ to show that rSMO detects the active set after some finite iterate $\mathcal{K}$.

$$x^0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \\ x_4^0 \\ x_5^0 \end{pmatrix} \quad \xrightarrow{\text{after finite } \mathcal{K} \text{ iterations}} \quad x^{\mathcal{K}} = \begin{pmatrix} x_1^{\mathcal{K}} \\ 0 \\ c \\ x_4^{\mathcal{K}} \\ 0 \end{pmatrix}, \quad \text{where} \quad x^* = \begin{pmatrix} x_1^* \\ 0 \\ c \\ x_4^* \\ 0 \end{pmatrix}.$$

→ Follow the argument used in Nutini et. al.'s work.

→ Pick out the indices i that are not on the active set.