Iteration Complexity of Feasible Descent Methods for Convex Optimization

Chih-Jen Lin
Department of Computer Science
National Taiwan University

Joint work with Po-Wei Wang

Talk at SIAM Conference on Optimization, May 2014

Outline

- Introduction
- Feasible descent methods and linear-convergence proof
- Rate of the linear convergence
- Discussions and conclusions

Outline

- Introduction
- Feasible descent methods and linear-convergence proof
- Rate of the linear convergence
- Discussions and conclusions

Problem

$$\min_{\mathbf{x}\in\mathcal{X}} \quad f(\mathbf{x}).$$

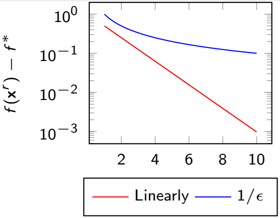
$f(\mathbf{x})$ is convex differentiable, $\mathcal{X}$ is closed and convex.

We want to know

- Iterations to reach $f(\mathbf{x}^r) - f^* \leq \epsilon$

Specially, we investigate algorithms with linear convergence

$$f(\mathbf{x}^{r+1}) - f^* \leq (1 - \frac{1}{c})(f(\mathbf{x}^r) - f^*), \forall r.$$



Motivation

- Dual problem of support vector classification is

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

  $E = \left[y_1\mathbf{z}_1, \ldots, y_l\mathbf{z}_l\right]$ is the data matrix, $(y_i, \mathbf{z}_i)$: label-instance pair, and $\mathbf{1}$ is the vector of ones
- $\mathbf{w}^\top\mathbf{w}/2$ is strongly convex in $\mathbf{w}$, but Hessian may not be strongly convex in $\boldsymbol{\alpha}$
- Coordinate descent method is commonly used, but complexity not very clear

Difficulties

For some convex but not non-strongly convex problems,

Asymptotic Linear Convergence (Luo and Tseng, 1993)

$$\exists r_0 \text{ such that } f(\mathbf{x}^{r+1}) - f^* \leq (1 - \frac{1}{c})(f(\mathbf{x}^r) - f^*), \quad \forall r \geq r_0.$$

Usually we only know the existence of r_0 but not its relation to problem parameters.

To estimate iteration numbers, we hope to have

Global Linear Convergence

$$f(\mathbf{x}^{r+1}) - f^* \leq (1 - \frac{1}{c})(f(\mathbf{x}^r) - f^*), \quad \forall r.$$

Difficulties (Cont'd)

- We also hope to know more about the convergence rate
- That is, how the rate is related to the data
- Properties of the data include range of feature values, number of instances, number of features etc.

Past Studies

- We are interested in deterministic algorithms (e.g., cyclic coordinate descent)
- Interestingly, more studies have been done on the complexity of randomized coordinate descent:
  - Linear convergence for strongly convex $f(\cdot)$ (Nesterov, 2012; Richtárik and Takáč, 2014; Tappenden et al., 2013)
  - Sub-linear convergence for non-strongly convex $f(\cdot)$ (Shalev-Shwartz and Tewari, 2009; Nesterov, 2012; Shalev-Shwartz and Zhang, 2013a,b)

Past Studies (Cont'd)

- Past work on complexity of cyclic coordinate descent:
  - Linear convergence for l2-loss SVM (Chang et al., 2008); smooth and strongly convex $f(\cdot)$ (Beck and Tetruashvili, 2013)
  - Sub-linear convergence for non-strongly convex $f(\cdot)$ (Tseng and Yun, 2009; Saha and Tewari, 2013)

Outline

- Introduction
- Feasible descent methods and linear-convergence proof
- Rate of the linear convergence
- Discussions and conclusions

Framework: Feasible Descent Methods

A sequence $\{\mathbf{x}^r\}$ is generated by a feasible descent method if for all iteration index r, $\{\mathbf{x}^r\}$ satisfies

$$\begin{aligned} \mathbf{x}^{r+1} &= [\mathbf{x}^r - \omega_r \nabla f(\mathbf{x}^r) + \mathbf{e}^r]^+_{\mathcal{X}}, \\ \|\mathbf{e}^r\| &\le \beta \|\mathbf{x}^r - \mathbf{x}^{r+1}\|, \\ f(\mathbf{x}^r) - f(\mathbf{x}^{r+1}) &\ge \gamma \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2, \end{aligned}$$

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

Coordinate descent is a special case

Examples of Feasible Descent Methods for Machine Learning

- Coordinate descent methods for dual Support Vector Classification (SVC)
- Coordinate descent methods for dual Support Vector Regression (SVR)
- Inexact coordinate descent for primal SVC

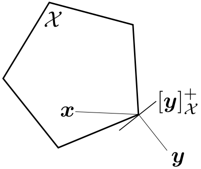
  Inexact: one-variable sub-problem approximately solved
- Gauss-Seidel method for solving linear systems

Projected Gradient

We need the following tools

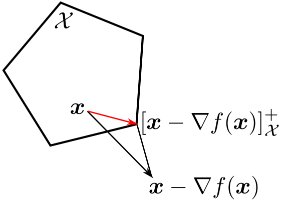
Definition (Convex Projection)

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



Definition (Projected gradient)

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



Lemma (Optimality condition)

$$\nabla^+ f(\mathbf{x}^*) = \mathbf{0} \;\Leftrightarrow\; \mathbf{x}^* \textit{ is optimal.}$$

Existing Techniques to Prove Asymptotic Linear Convergence

In Luo and Tseng (1993), they prove the following error bound

$$\min_{\mathbf{x}^* \in \mathcal{X}^*} \|\mathbf{x}^r - \mathbf{x}^*\| \leq \kappa \|\nabla^+ f(\mathbf{x}^r)\|, \quad \forall r \geq r_0,$$

where $\mathcal{X}^*$ is the set of optimal solutions

We call this a local error bound because of r_0.

We aim at proving a global error bound and knowing more about κ

Existing Techniques to Prove Asymptotic Linear Convergence (Cont'd)

- In a sense you can also say that a local error bound is global. If $\mathcal{X}$ is compact, there exists $\bar{\kappa}$ such that

$$\min_{\mathbf{x}^* \in \mathcal{X}^*} \|\mathbf{x}^r - \mathbf{x}^*\| \leq \bar{\kappa} \|\nabla^+ f(\mathbf{x}^r)\|, \quad \forall r \geq 0$$

- Based on the existence of such bounds, linear convergence has recently been established (e.g., Hong et al., 2014; Kadkhodaie et al., 2014) for problems not covered in (Luo and Tseng, 1993)
- However, we are interested in rate analysis here, so we must know more about κ

Sufficient Condition for Global Linear Convergence

We proved that feasible descent methods have global linear convergence if the following condition holds.

Global Error Bound from the Beginning

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \kappa \|\nabla^+ f(\mathbf{x})\|,$$

for all $\mathbf{x}$ satisfying

$$\mathbf{x} \in \mathcal{X} \text{ and } f(\mathbf{x}) - f^* \leq M,$$

where $\bar{\mathbf{x}}$ is the nearest optimum to $\mathbf{x}$, f^* is the optimal value, and $M \equiv f(\mathbf{x}^0) - f^*$. We will check details of κ

Who Has A Global Error Bound from the Beginning?

Assumption (Strongly Convex)

$f(\mathbf{x})$ is σ strongly convex and ∇f is ρ Lipschitz continuous.

A global error bound has been proved in Pang (1987)

However, recall our goal is to study non-strongly convex problems such as SVM dual

Who Has A Global Error Bound from the Beginning? (Cont'd)

Assumption (Strongly Convex Composition)

$\mathcal{X}$ is a polyhedral set $\{\mathbf{x} \mid A\mathbf{x} \leq \mathbf{d}\}$ and

$$f(\mathbf{x}) = g(E\mathbf{x}) + \mathbf{b}^\top \mathbf{x}, \tag{1}$$

where $g(\cdot)$ is σ_g strongly convex and ∇f is ρ Lipschitz continuous.

Our main result: global error bound for (1)

Then we can prove global linear convergence of feasible descent methods for (1)

Key Ideas in Our Proof

- Optimal solution set is a polyhedral set

$$E\mathbf{x}^* = \mathbf{t}^*, \ \ \mathbf{b}^\top \mathbf{x}^* = s^*, \text{ and } \ A\mathbf{x}^* \le \mathbf{d}.$$

- Using Hoffman's bound (Hoffman, 1952) to bound the distance between $\mathbf{x}$ and a polyhedron. We proved a modified version from Li (1994)

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \le \theta\left(A, \left(\begin{smallmatrix} E \\ \mathbf{b}^\top \end{smallmatrix}\right)\right) \left\| \begin{matrix} E(\mathbf{x} - \bar{\mathbf{x}}) \\ \mathbf{b}^\top(\mathbf{x} - \bar{\mathbf{x}}) \end{matrix} \right\|,$$

where $\theta\left(A, \left(\begin{smallmatrix} E \\ \mathbf{b}^\top \end{smallmatrix}\right)\right)$ is a constant related to A, E, $\mathbf{b}$.

- Finally, we bound $\|E(\mathbf{x} - \bar{\mathbf{x}})\|^2$ and $(\mathbf{b}^\top(\mathbf{x} - \bar{\mathbf{x}}))^2$

Outline

- Introduction
- Feasible descent methods and linear-convergence proof
- Rate of the linear convergence
- Discussions and conclusions

The Error Bound Constants

We proved

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \kappa \|\nabla^{+} f(\mathbf{x})\|$$

with

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

Recall that

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

where $g(\cdot)$ is σ_g strongly convex and ∇f is ρ Lipschitz

If $\mathcal{X} = \mathbb{R}^l$ or $\mathbf{b} = \mathbf{0}$, κ can be simplified to

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

The Convergence Rate

With an error bound, the feasible descent method

$$\mathbf{x}^{r+1} = [\mathbf{x}^r - \omega_r \nabla f(\mathbf{x}^r) + \mathbf{e}^r]^+_{\mathcal{X}},$$
$$\|\mathbf{e}^r\| \leq \beta \|\mathbf{x}^r - \mathbf{x}^{r+1}\|,$$
$$f(\mathbf{x}^r) - f(\mathbf{x}^{r+1}) \geq \gamma \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2,$$

converges linearly with

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

where

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

Examples of the Error Bound Constant

Dual problem of l1-loss support vector classification

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

$E = \left[y_1 \mathbf{z}_1, \ldots, y_l \mathbf{z}_l\right]$ is the data matrix, $(y_i, \mathbf{z}_i)$: label-instance pair, and $\mathbf{1}$ is the vector of ones

If coordinate descent methods are used and each instance is normalized to unit length,

$$\kappa = O(\rho \theta^2 C l),$$

where l is the number of training instances.

Examples of the Convergence Rate

For dual problem of l1-loss support vector classification, the cyclic coordinate descent method has global linear convergence.

$$f(\mathbf{x}^{r+1}) - f^* \leq (1 - \frac{1}{2\phi + 1})(f(\mathbf{x}^r) - f^*), \quad \forall r,$$

where

$$\phi = O(l\rho^2\kappa) = O(\rho^3\theta^2 Cl^2).$$

Outline

Conclusions

- For some non-strongly convex functions, we provide rate analysis of linear convergence for feasible descent methods
- The key idea is to prove an error bound between any point and the optimal solution set
- Our result enables the global linear convergence of optimization methods for some machine learning problems
- Details of the proof can be found at: P.-W. Wang and C.-J. Lin. Iteration complexity of feasible descent methods for convex optimization. Journal of Machine Learning Research, 2014.