ENGG 5501: Foundations of Optimization **2016–17 First Term**

Handout 8: A Quick Tour on Optimization Algorithms

Instructor: Anthony Man–Cho So November 28, 2016

1 Introduction

Up till now we have been focusing on characterizing the optimal solutions of various classes of optimization problems. The theories we developed allow us to gain insights into the structures of those problems, and when combined with specific domain knowledge, those insights can often yield interesting consequences. However, all those theories will be of little value if we cannot find a solution to the optimization problem in question. Thus, we shall now discuss some techniques for solving various classes of optimization problems.

It turns out that many optimization algorithms are *iterative* in nature. By "iterative", we mean that the algorithm generates a sequence of points, each of which is calculated based on the point(s) preceding it. As an illustration, consider a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, and suppose that we are interested in minimizing it; i.e., we would like to find $x^* \in \mathbb{R}^n$ such that $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. An iterative algorithm will start at some initial point x^0, and then generate updates $x^1, x^2, \ldots$ according to some pre–determined rules. Of course, we would like the sequence of points generated by the algorithm to converge to either a global or local minimum of f. An even more desirable feature is that such a convergence occurs in a finite number of steps. In the following sections we will study some iterative algorithms in detail and analyze their convergence properties.

2 Basic Elements of Iterative Algorithms

To fix ideas, let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Consider the unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

In general, it may be too ambitious to find a global minimum of f. Hence, we will just look for a **stationary point** of f; i.e., a point $\bar{x} \in \mathbb{R}^n$ that satisfies $\nabla f(\bar{x}) = \mathbf{0}$. To begin, let $x^0 \in \mathbb{R}^n$ be an initial iterate with $\nabla f(x^0) \neq \mathbf{0}$. In order to achieve progress, we need to proceed in some **search direction** $d^k \in \mathbb{R}^n$. For instance, we can update the iterates according to the following rule:

$$x^{k+1} = x^k + \alpha_k d^k \qquad \text{for } k = 0, 1, \ldots. \tag{2}$$

Here, $\alpha_k > 0$ is called the **step size** and controls how far we proceed in the direction d^k. Note that (2) actually defines a *family* of update rules that are parametrized by the search directions $\{d^k\}_{k \geq 0}$ and step sizes $\{\alpha_k\}_{k \geq 0}$. There are many possibilities in choosing the search directions and step sizes. Below are some common choices.

2.1 Choosing the Search Directions

2.1.1 The Method of Steepest Descent

Roughly speaking, the method of steepest descent is based on minimizing a linear approximation of f at the current iterate $x^k \in \mathbb{R}^n$. Specifically, suppose that the current iterate x^k satisfies $\nabla f(x^k) \neq \mathbf{0}$. Then, we may construct a *linear approximation* of f at x^k, which is given by

$$f^k(x) \equiv f(x^k) + \nabla f(x^k)^T(x - x^k).$$

Now, recall that if there exists a $d \in \mathbb{R}^n$ such that $\nabla f(x^k)^T d < 0$, then there exists an $\alpha_0 > 0$ such that $f(x^k + \alpha d) < f(x^k)$ for all $\alpha \in (0, \alpha_0)$ (Proposition 1 of Handout 7). Thus, in order to guarantee descent, we need to choose $x \in \mathbb{R}^n$ such that $\nabla f(x^k)^T(x - x^k) < 0$. Of course, if $\nabla f(x^k) \neq \mathbf{0}$, then we can make $\nabla f(x^k)^T(x - x^k)$ as negative as possible. Thus, we need to restrict the length of the direction $d = x - x^k$. In particular, we may consider the following:

$$\begin{array}{ll} \text{minimize} & \nabla f(x^k)^T d \\ \text{subject to} & \|d\|_2^2 \leq \|\nabla f(x^k)\|_2^2. \end{array}$$

By the Cauchy–Schwarz inequality, we see that the optimal solution to the above problem is $d^k = -\nabla f(x^k)$. The direction d^k is called the **direction of steepest descent**, and the resulting iterative algorithm

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \qquad \text{for } k = 0, 1, \ldots \tag{3}$$

is called the **method of steepest descent** or simply the **gradient method**.

2.1.2 Newton's Method

Recall that in the method of steepest descent, we first construct a linear approximation of f at each iterate, and then choose a direction that optimizes the local rate of descent. In practice, such a method usually converges very slowly, as it does not necessarily maximize the total decrease in the objective value that can be achieved by moving in a direction. A better alternative is the so–called **Newton's method**. The idea is to minimize in each iteration a *quadratic approximation* of f at the current iterate x^k, which is given by

$$f^k(x) \equiv f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k).$$

Specifically, we first compute

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

Then, the next iterate x^{k+1} is obtained by setting $\nabla f^k(x) = \mathbf{0}$ and solving for x; i.e.,

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) \qquad \text{for } k = 0, 1, \ldots. \tag{4}$$

In particular, the search direction is given by $d^k = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$. The resulting iterative method (4) is the so–called **pure Newton's method**. We can also consider the following more general iterative method:

$$x^{k+1} = x^k - \alpha_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) \qquad \text{for } k = 0, 1, \ldots,$$

where $\alpha_k > 0$ is the step size. Note that in order for the search direction d^k to be a descent direction, it is necessary that $\nabla f(x^k)^T d^k < 0$, or equivalently, $\nabla f(x^k)^T \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) > 0$. The latter condition can be ensured if $\nabla^2 f(x^k) \succ \mathbf{0}$.

An important feature of Newton's method is that it is independent of the choice of coordinates. This should not be very surprising since the quadratic approximation of f should not depend on a particular coordinate system. Specifically, let $T : \mathbb{R}^n \to \mathbb{R}^n$ be given by $T(x) = Ax$, where $A \in \mathbb{R}^{n \times n}$ is non–singular and $b \in \mathbb{R}^n$. Consider the function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ given by $\tilde{f}(x) = f(T(x))$. The quadratic approximation of $\tilde{f}$ at $y^k = T^{-1}(x^k)$ is given by

$$\tilde{f}^k(y) = \tilde{f}(y^k) + \nabla \tilde{f}(y^k)^T (y - y^k) + \frac{1}{2}(y - y^k)^T \nabla^2 \tilde{f}(y^k)(y - y^k).$$

It follows that

$$\nabla \tilde{f}^k(y) = \nabla \tilde{f}(y^k) + \nabla^2 \tilde{f}(y^k)(y - y^k) = \nabla \tilde{f}(y^k) + \nabla^2 \tilde{f}(y^k) A^{-1} A(y - y^k).$$

Upon solving $\nabla \tilde{f}^k(y) = \mathbf{0}$ for y and setting $x = T(y)$, we have

$$x = x^k - \left(\nabla^2 \tilde{f}(y^k) A^{-1}\right)^{-1} \nabla \tilde{f}(y^k).$$

Now, by the Chain Rule, we have

$$\nabla \tilde{f}(y) = A^T \nabla f(T(y)) \qquad \text{and} \qquad \nabla^2 \tilde{f}(y) = A^T \nabla^2 f(T(y)) A.$$

Hence, the search direction $\tilde{d}^k$ at the point $y^k = T^{-1}(x^k)$ is given by

$$\tilde{d}^k = -\left(\nabla^2 \tilde{f}(y^k) A^{-1}\right)^{-1} \nabla \tilde{f}(y^k) = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) = d^k.$$

This shows that Newton's method is independent of the choice of coordinates.

2.2 Choosing the Step Sizes

The simplest choice for the step sizes is to use the same value for all iterations; i.e., set $\alpha_k = \alpha$ for some $\alpha > 0$ and for $k = 0, 1, \ldots$. Unfortunately, an iterative method with constant step size may perform very poorly. On the one hand, if the step size is too large, then we may not be able to guarantee that the descent condition is satisfied, and the method may not converge. On the other hand, if the step size is too small, then the method may converge very slowly. An alternative approach is to perform a **line search** at each iterate. Specifically, given the current iterate $x^k \in \mathbb{R}^n$ and a search direction $d^k \in \mathbb{R}^n$, we compute the step size α_k by

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha d^k). \tag{5}$$

In fact, it is not necessary that we find the exact minimum, as the goal of the line search (5) is simply to achieve a substantial decrease in the objective value by proceeding in the direction d^k. For various implementations of (approximate) line search, we refer the reader to [1, Appendix C].

2.3 Handling Constraints

So far our discussion has focused on the unconstrained optimization problem (1). Let us now turn our attention to constrained optimization problems of the form

$$v^* = \min_{x \in X} f(x), \tag{6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is as before and $X \subset \mathbb{R}^n$ is non–empty and closed. To develop an iterative algorithm for solving (6), a natural idea is to modify the update rule (2) to

$$x^{k+1} = \Pi_X \left(x^k + \alpha_k d^k \right) \qquad \text{for } k = 0, 1, \ldots, \tag{7}$$

where $\Pi_X : \mathbb{R}^n \to X$ is the projection operator onto X. Note that if the projection in (7) is well–defined for each $k \geq 0$ and if $x^0 \in X$, then the sequence of iterates $\{x^k\}_{k \geq 0}$ will all be feasible for (6). In particular, by taking $d^k = -\nabla f(x^k)$, we obtain the **projected gradient method**:

$$x^{k+1} = \Pi_X \left(x^k - \alpha_k \nabla f(x^k) \right) \qquad \text{for } k = 0, 1, \ldots. \tag{8}$$

3 Convergence Analysis of the Gradient Method

As mentioned in the Introduction, an important issue concerning iterative methods is their convergence behavior. In this section we will develop the machinery for analyzing the projected gradient method (8) when it is applied to solve a class of constrained convex optimization problems. Specifically, consider the following assumptions concerning (6):

Assumption 1

(a) The function f is continuously differentiable and strongly convex with parameter $\sigma > 0$, and the gradient ∇f is Lipschitz continuous with parameter $L > 0$; i.e., for all $x, y \in \mathbb{R}^n$,

$$\sigma \|x - y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y), \tag{9}$$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2. \tag{10}$$

(b) The set X is non–empty, convex, and closed.

Under Assumption 1, it can be shown that the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is bounded for any given $x^0 \in \mathbb{R}^n$, and hence the optimal solution set X^* of problem (6) is non–empty. This motivates us to ask whether the sequence of iterates $\{x^k\}_{k \geq 0}$ generated by the projected gradient method (8) converges to an element of X^*. Towards that end, let us first study the convergence behavior of the objective values $\{f(x^k)\}_{k \geq 0}$ associated with the iterates $\{x^k\}_{k \geq 0}$. We begin with the following proposition:

Proposition 1 *For any $x, y \in \mathbb{R}^n$,*

$$\left| f(y) - f(x) - \nabla f(x)^T (y - x) \right| \leq \frac{L}{2} \|x - y\|_2^2.$$

Proof Consider the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(t) = f(x + t(y - x))$. By the Fundamental Theorem of Calculus and the Chain Rule, we have

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)\, dt = \int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt.$$

Upon writing

$$\int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt = \int_0^1 \left[\nabla f(x + t(y - x)) - \nabla f(x)\right]^T (y - x)\, dt + \nabla f(x)^T (y - x)$$

and using the Cauchy–Schwarz inequality together with the fact that ∇f is Lipschitz continuous with constant $L > 0$, we have

$$\begin{aligned} \left| f(y) - f(x) - \nabla f(x)^T (y - x) \right| &\leq \int_0^1 \left| \left[\nabla f(x + t(y - x)) - \nabla f(x)\right]^T (y - x) \right| dt \\ &\leq L\|y - x\|_2^2 \int_0^1 t\, dt \\ &= \frac{L}{2}\|y - x\|_2^2, \end{aligned}$$

as desired. □

An important consequence of Proposition 1 is that the projected gradient method (8) is a descent method when the step sizes $\{\alpha_k\}_{k\geq 0}$ are sufficiently small. Specifically, we have the following corollary:

Corollary 1 *Suppose that Assumption 1 holds for problem (6). Then, the sequence of iterates $\{x^k\}_{k\geq 0}$ generated by the projected gradient method (8) satisfies*

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2}\left(\frac{1}{\alpha_k} - L\right)\|x^k - x^{k+1}\|_2^2. \tag{11}$$

In particular, if $\alpha_k < 1/L$ for all $k \geq 0$, then the sequence $\{f(x^k)\}_{k\geq 0}$ is monotonically decreasing and hence convergent.

Although the sequence $\{f(x^k)\}_{k\geq 0}$ is convergent when the step sizes $\{\alpha_k\}_{k\geq 0}$ are sufficiently small, we do not yet know whether the limit is v^*, or whether the sequence of iterates $\{x^k\}_{k\geq 0}$ converges. To study these issues, we need a measure to quantify the progress of method (8). One natural candidate would be $\mathrm{dist}(\cdot, X^*)$, the distance to the optimal solution set X^*. Despite its intuitive appeal, such a measure is hard to compute or analyze. Alternatively, observe that every optimal solution x^* to problem (6) satisfies the following necessary and sufficient optimality condition:

$$x^* = \Pi_X\left(x^* - \nabla f(x^*)\right). \tag{12}$$

This, together with (8), motivates the use of the residual map $R : \mathbb{R}^n \to \mathbb{R}^n$, where

$$R(x) = x - \Pi_X(x - \nabla f(x)).$$

Roughly speaking, the function R measures how much a solution $x \in \mathbb{R}^n$ violates the optimality condition (12). In particular, x is an optimal solution to (6) if and only if $R(x) = \mathbf{0}$. However, since $\|R(\cdot)\|_2$ is only a surrogate of $\mathrm{dist}(\cdot, X^*)$, we need to establish a relationship between them. This motivates the development of the following **error bound** for problem (12):

Theorem 1 *Suppose that Assumption 1 holds for problem (6). Then, for all $x \in \mathbb{R}^n$,*

$$\mathrm{dist}(x, X^*) \leq \frac{1+L}{\sigma}\|R(x)\|_2.$$

To prove Theorem 1, we need the following result:

Proposition 2 *Under Assumption 1, we have*

$$(\nabla f(x) - \nabla f(y))^T(x-y) \leq (1+L)\|R(x) - R(y)\|_2\|x-y\|_2$$

for all $x, y \in \mathbb{R}^n$. Consequently, we have

$$\|x-y\|_2 \leq \frac{1+L}{\sigma}\|R(x) - R(y)\|_2$$

for all $x, y \in \mathbb{R}^n$.

Proof By Theorem 8 of Handout 2, for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned} 0 &\geq [(x - \nabla f(x)) - \Pi_X(x - \nabla f(x))]^T [\Pi_X(y - \nabla f(y)) - \Pi_X(x - \nabla f(x))] \\ &= (R(x) - \nabla f(x))^T (R(x) - R(y) - (x-y)). \end{aligned}$$

It follows that

$$\nabla f(x)^T(x-y) \leq R(x)^T((x-y) - (R(x) - R(y))) + \nabla f(x)^T(R(x) - R(y)). \tag{13}$$

By interchanging x and y in the above derivation, we obtain

$$\nabla f(y)^T(y-x) \leq R(y)^T((y-x) - (R(y) - R(x))) + \nabla f(y)^T(R(y) - R(x)). \tag{14}$$

Upon adding (13), (14) together and simplifying, we have

$$\begin{aligned} (\nabla f(x) - \nabla f(y))^T(x-y) &\leq (R(x) - R(y))^T(x - y + \nabla f(x) - \nabla f(y)) - \|R(x) - R(y)\|_2^2 \\ &\leq (R(x) - R(y))^T(x - y + \nabla f(x) - \nabla f(y)) \\ &\leq \|R(x) - R(y)\|_2 (\|x-y\|_2 + \|\nabla f(x) - \nabla f(y)\|_2) \\ &\leq (1+L)\|R(x) - R(y)\|_2\|x-y\|_2, \end{aligned}$$

as desired. This, together with the strong convexity of f, implies that

$$\sigma\|x-y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^T(x-y) \leq (1+L)\|R(x) - R(y)\|_2\|x-y\|_2.$$

It follows that

$$\|x-y\|_2 \leq \frac{1+L}{\sigma}\|R(x) - R(y)\|_2,$$

and the proof is completed. □

Proof of Theorem 1 Upon taking $y \in X^*$ in Proposition 2 and noting that $R(y) = \mathbf{0}$, we have

$$\mathrm{dist}(x, X^*) \leq \|x-y\|_2 \leq \frac{1+L}{\sigma}\|R(x)\|_2$$

for any $x \in \mathbb{R}^n$. This completes the proof. □

Armed with the error bound in Theorem 1, let us proceed to establish the convergence of $\{f(x^k)\}_{k\geq 0}$ to v^*, as well as the convergence of $\{x^k\}_{k\geq 0}$ to an element of X^*. In fact, our proof will establish the global rate of convergence as well. To begin, we need the following result, whose proof can be found in [2, Lemma 1] and [8, Lemma A.7]:

Fact 1 *For any given $x \in \mathbb{R}^n$, the function*

$$\alpha \mapsto \frac{1}{\alpha}\|x - \Pi_X(x - \alpha\nabla f(x))\|_2$$

is decreasing in $\alpha > 0$, and the function

$$\alpha \mapsto \|x - \Pi_X(x - \alpha\nabla f(x))\|_2$$

is increasing in $\alpha > 0$.

Consider the sequence of iterates $\{x^k\}_{k\geq 0}$ generated by the projected gradient method (8). Define $\bar{x}^k = \Pi_{X^*}(x^k)$ for $k \geq 0$. Note that $\bar{x}^k$ is well–defined, as it can be easily verified that X^* is a non–empty closed convex set. Furthermore, suppose that $\underline{\alpha}' = \inf_{k\geq 0}\alpha_k > 0$ and $\bar{\alpha} = \sup_{k\geq 0}\alpha_k < 1/L$. Then, by Fact 1, we have

$$\underline{\alpha}\|R(x^k)\|_2 = \underline{\alpha}\left\|x^k - \Pi_X\left(x^k - \nabla f(x^k)\right)\right\|_2 \leq \left\|x^k - \Pi_X\left(x^k - \alpha_k\nabla f(x^k)\right)\right\|_2 = \|x^k - x^{k+1}\|_2,$$

where $\underline{\alpha} = \min\{1, \underline{\alpha}'\}$. This, together with Theorem 1, implies that

$$\mathrm{dist}(x^k, X^*) = \|x^k - \bar{x}^k\|_2 \leq \frac{1+L}{\sigma}\|R(x^k)\|_2 \leq \frac{1+L}{\underline{\alpha}\sigma}\|x^k - x^{k+1}\|_2. \tag{15}$$

Next, we estimate the cost–to–go $f(x^k) - v^*$ of x^k. By Theorem 8 of Handout 2 and the definition of x^{k+1}, we have

$$(x^k - \alpha_k\nabla f(x^k) - x^{k+1})^T(\bar{x}^k - x^{k+1}) \leq 0,$$

or equivalently,

$$(x^k - x^{k+1})^T(\bar{x}^k - x^{k+1}) \leq \alpha_k\nabla f(x^k)^T(\bar{x}^k - x^{k+1}).$$

This implies that

$$\begin{aligned}
f(x^{k+1}) - v^* &= f(x^{k+1}) - f(\bar{x}^k) \\
&\leq \nabla f(x^{k+1})^T(x^{k+1} - \bar{x}^k) \\
&= (\nabla f(x^{k+1}) - \nabla f(x^k))^T(x^{k+1} - \bar{x}^k) + \nabla f(x^k)^T(x^{k+1} - \bar{x}^k) \\
&\leq \|\nabla f(x^{k+1}) - \nabla f(x^k)\|_2\|x^{k+1} - \bar{x}^k\|_2 + \frac{1}{\alpha_k}(x^k - x^{k+1})^T(x^{k+1} - \bar{x}^k) \\
&\leq \left(L + \frac{1}{\underline{\alpha}}\right)\|x^{k+1} - x^k\|_2\|x^{k+1} - \bar{x}^k\|_2.
\end{aligned}$$

Since

$$\|x^{k+1} - \bar{x}^k\|_2 \leq \|x^{k+1} - x^k\|_2 + \|x^k - \bar{x}^k\|_2 \leq \left(\frac{1+L}{\sigma\underline{\alpha}} + 1\right)\|x^k - x^{k+1}\|_2$$

by (15), we conclude that

$$f(x^{k+1}) - v^* \leq \left(L + \frac{1}{\underline{\alpha}}\right)\left(\frac{1+L}{\sigma\underline{\alpha}} + 1\right)\|x^k - x^{k+1}\|_2^2.$$

This, together with (11), implies that

$$f(x^{k+1}) - v^* \leq \kappa(f(x^k) - f(x^{k+1})),$$

or equivalently,

$$f(x^{k+1}) - v^* \leq \frac{\kappa}{1+\kappa}(f(x^k) - v^*) \tag{16}$$

for all $k \geq 0$, where

$$\kappa = 2\left(\frac{1}{\bar{\alpha}} - L\right)^{-1}\left(L + \frac{1}{\underline{\alpha}}\right)\left(\frac{1+L}{\sigma\underline{\alpha}} + 1\right).$$

Note that (16) implies the sequence $\{f(x^k) - v^*\}$ converges to zero at a geometric rate. Moreover, we have

$$f(x^{k+1}) \leq \frac{\kappa}{1+\kappa} f(x^k) + \frac{1}{1+\kappa} v^*,$$

which yields

$$f(x^k) - v^* \leq \left(\frac{\kappa}{1+\kappa}\right)^k f(x^0)$$

for all $k \geq 1$. This, together with (11), implies that

$$\begin{aligned} \|x^k - x^{k+1}\|_2^2 &\leq 2\left(\frac{1}{\bar{\alpha}} - L\right)^{-1}\left[f(x^k) - v^*\right] \\ &\leq 2f(x^0)\left(\frac{1}{\bar{\alpha}} - L\right)^{-1}\left(\frac{\kappa}{1+\kappa}\right)^k \end{aligned}$$

for all $k \geq 1$. Combining this with (15), we obtain

$$\|R(x^k)\|_2 \leq \frac{1}{\underline{\alpha}}\left[2f(x^0)\left(\frac{1}{\bar{\alpha}} - L\right)^{-1}\right]^{1/2}\left(\frac{\kappa}{1+\kappa}\right)^{k/2}$$

for all $k \geq 1$, which shows that $\{x^k\}_{k\geq 0}$ converges to an element of X^* at a geometric rate as well.

4 Further Reading

In this lecture we just barely scratched the surface of the vast field of optimization algorithms. For further reading we refer the reader to [1, 3, 5, 6]. The development in Section 3, as well as its many powerful extensions, can be found in [4, 7, 8].

References

[1] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, second edition, 1999.

[2] E. M. Gafni and D. P. Bertsekas. Two–Metric Projection Methods for Constrained Optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964, 1984.

[3] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*, volume 116 of *International Series in Operations Research and Management Science*. Springer Science+Business Media, LLC, New York, third edition, 2008.

[4] Z.-Q. Luo and P. Tseng. Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[5] J. Renegar. *A Mathematical View of Interior–Point Methods in Convex Optimization*, volume 3 of *MPS–SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2001.

[6] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, Princeton, New Jersey, 2006.

[7] A. M.-C. So. Non–Asymptotic Convergence Analysis of Inexact Gradient Methods for Machine Learning Without Strong Convexity. Preprint, available at `http://www.se.cuhk.edu.hk/~manchoso/papers/inexact_GM_conv.pdf`, 2013.

[8] P.-W. Wang and C.-J. Lin. Iteration Complexity of Feasible Descent Methods for Convex Optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.