Technical Report

Xu Liwei

August 2021

Contents

1 Introduction

This is the technical report of my ICM2A internship in Huawei Belgian Research Center. Huawei is a leading telecom solutions provider, i was working in Huawei BeRC's 3NLab(Neural Networks for Networks). This laboratory is specialized in nonlinear optimization and neural computing methods for the solving of NP-hard optimization problems, nonconvex optimization problems (both in the objective and constraints, etc). During my internship, i tested the performance of different optimization algorithms, combined with distinct conditions and settings. Every friday, i have the chance to go to office and present my process or results physically. Other time, i worked at home and keep a contact with my instructors and colleges online. Since this technical report could be too redundant, i add the following note according to the points of EMSE ICM2A internship requirements.

For the first part "The ability to produce a technical/scientific report", you could find the research scheme: In section 2, i get start from Unconstrained Optimization, which contains the basis and commonly used optimization methods. After i have a good understanding of the classic optimization methods, I encountered a difficulty: Dual Methods. I read several books and did most of the corresponding exercises to understand it. In section 3, I present the duality framework in theory and practical. The duality framework is very important in Optimization theory, since it has the ability to simplify some rather complex problems, like transforming the nonconvex problems to convex problems. Later, i introduce some convex constrained problems and specific methods for constraint, like Penalty and Augmented Lagrangian Methods. In section 5, I present the Alternating Direction Method of Multipliers, which is the main research interest during my internship. You could find some numerical results and corresponding codes. In section 6, I give a few comparison results and analysis as the conclusion. Last, I also put a reference in the end of this report, page 156.

For the second part "Mastery of technical, scientific and methodological skills", the subject of my internship is comparing the performance of ADMM with some other optimization methods on specific problems, combined with some classic acceleration methods and adaptive restart. For the challenges related to the subject, the main challenge is how to understand and use the duality framework to transform nonconvex problems to convex problems. In section 3, you could find the solutions for this challenge. These nonlinear optimization and neural computing methods are competitive against polynomial-time approximations used to be classically involved in network optimization problem solving. For the point "Ability to mobilise knowledge", you could easily find some of the used techniques in Section 2.

For the third part "a sense of perspective and distanced analysis", I added an ending part as a summary for my internship ICM2A by the end of this report.

2 Unconstrained Optimization

2.1 Line Search Methods

Each iteration of a line search method computes a search direction p_k and then decides how far to move along that direction. The iteration is given by:

$$x_{k+1} = x_k + \alpha_k p_k,$$

where the positive scalar α_k is called the step length. The success of a line search method depends on effective choices of both the direction p_k and the step length α_k. Most line search algorithms requires p_k to be a descent direction,*i.e.*, one for $p_k^{\mathrm{T}} \bigtriangledown f_k < 0$ Moreover, the search direction often has the form

$$p_k = -B_k^{-1} \bigtriangledown f_k, \tag{2.1}$$

where B_k is a symmetric and nonsingular matrix. In the steepest descent method, B_k is simply the identity matrix I, while in Newton's method, B_k is the exact Hessian $\bigtriangledown^2 f(x_k)$. In quasi-Newton methods, B_k is an approximation to the Hessian that is updated at every iteration by means of a low-rank formula. When p_k is defined by (2.1) and B_k is positive definite, we have

$$p_k^{\mathrm{T}} \bigtriangledown f_k = - \bigtriangledown f_k^{\mathrm{T}} B_k^{-1} \bigtriangledown f_k < 0,$$

and therefore p_k is a descent direction. To avoid the insufficient reduction in f, we need to enforce a sufficient decrease condition.

2.1.1 Step length: Wolfe conditions and Goldstein conditions

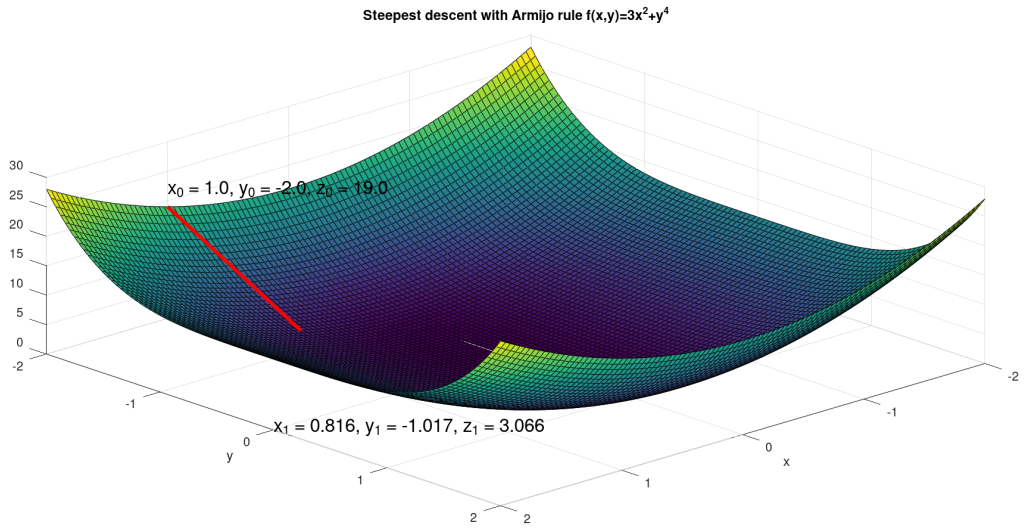
A popular inexact line search condition stipulates that α_k should first of all give sufficient decrease in the objective function f, as measured by the following inequality:

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \bigtriangledown f_k^{\mathrm{T}} p_k, \tag{2.2}$$

for some constant $sigma \in (0, 1)$. In other words, the reduction in f should be proportional to both the step length α_k and the directional derivative $\bigtriangledown f_k^{\mathrm{T}} p_k$, Inequality(2.2) is sometimes called the Armijo condition. In practice, c_1 is chosen to be quite small, say $c_1 = 10^{-4}$. Here, fixed scalars s, β andσ, with $0 < \beta < 1$, and $0 < \sigma < 1$ are chosen, and we set $\alpha^k = \beta^{m_k} s$, where m_k is the first nonnegative interger m for which

$$f(x^k) - f(x^k + \beta^m s p^k) \geq -\sigma \beta^m s \bigtriangledown f(x^k)' p^k. \tag{2.3}$$

Exercise:
Consider the problem of minimization the function of two variables $f(x, y) = 3x^2 + y^4$, Apply one iteration of the steepest descent method with (1,-2) as the starting point and with the stepsize chosen by the Armijo rule with $s = 1, \sigma = 0.1, and\ \beta = 0.5$.
The sufficient decrease condition is not enough by itself to ensure that the algorithm



make reasonable progress, to rule out unacceptably short steps we introduce a second requirement, called the curvature condition, which requires α_k to satisfy:

$$\bigtriangledown f(x_k + \alpha_k p_k)^{\mathrm{T}} p_k \geq c_2 \bigtriangledown f_k^{\mathrm{T}} p_k, \tag{2.4}$$

for some constant $c_2 \in (c_1, 1)$. The sufficient decrease(2.2) and curvature conditions(2.3) are known collectively as the Wolfe conditions. The strong Wolfe conditions require α_k to satisfy

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \bigtriangledown f_k^{\mathrm{T}} p_k,$$
$$\| \bigtriangledown f(x_k + \alpha_k p_k)^{\mathrm{T}} p_k \| \geq c_2 \| \bigtriangledown f_k^{\mathrm{T}} p_k \|,$$

with $0 < c_1 < c_2 < 1$. The only difference with the Wolfe conditions is that we no longer allow the derivative $\phi'(\alpha)$ to be too positive. Hence, we exclude points that are far from stationary points of ϕ. It's not difficult to prove that there exist step lengths that satisfy the Wolfe conditions for every function f that is smooth and bounded below.

Theorem 2.1 *suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable. Let p_k be a descent direction at x_k, and assume that f is bounded below along the ray $\{x_k + \alpha p_k | \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying the Wolfe conditions and the strong Wolfe conditions.*

Like the Wolfe conditions, the Goldstein conditions ensure that the step length α achieves sufficient decrease but is not too short. The goldstein conditions can also be stated as a pair of inequalities, in the following way:

$$f(x_k) + (1 - c)\alpha_k \bigtriangledown f_k^{\mathrm{T}} p_k \leq f(x_k + \alpha p_k) \leq f(x_k) + c\alpha_k \bigtriangledown f_k^{\mathrm{T}} p_k,$$

with $0 < c < 1/2$. The second inequality is the sufficient decrease condition, whereas the first inequality is introduced to control the step length from below. A disadvantage of the Goldstein conditions vis-à-vis the Wolfe conditions is that the first inequality may exclude

all minimizers of ϕ. However, the Goldstein and Wolfe conditions have much in common, and their convergence theories are quite similar. The Goldstein conditions are often used in Newton-type methods but are not well suited for quasi-Newton methods that maintain a positive definite Hessian approximation.

2.1.2 Zoutendijk's Theorem

The angle θ_k between p_k and the steepest descent direction $- \bigtriangledown f_k$, defined by

$$cos\theta_k = \frac{- \bigtriangledown f_k^T p_k}{\| \bigtriangledown f_k \| \| p_k \|}.$$

Theorem 2.2 *Consider any iteration of the form $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a decent direction and α_k satisfies the Wolfe conditions. Suppose that f is bounded below in $\mathbf{R}^n$ and that f is continuously in an open set N containing the level set $L := \{x : f(x) \leq f(x_0)\}$, where x_0 is the starting point of the iteration. Assume also that the gradient $\bigtriangledown f$ is Lipschitz continuous on N, that is, there exists a constant $L > 0$ such that*

$$\| \bigtriangledown f(x) - \bigtriangledown f(\widetilde{x}) \| \leq L \| x - \widetilde{x} \|, \ for \ all \ x, \widetilde{x} \in N.$$

Then

$$\sum_{k \geq 0} cos^2 \theta_k \| \bigtriangledown f_k \|^2 < \infty.$$

The proof of the theorem is not complex, but it has far-reaching consequences. Similar results to this theorem hold when the Goldstein conditions or strong Wolfe conditions are used in place of the Wolfe conditions.

2.1.3 Newton's Method

We consider the Newton iteration, for which the search is given by

$$p_k^N = - \bigtriangledown^2 f_k{}^{-1} \bigtriangledown f_k.$$

Since the Hessian matrix $\bigtriangledown^2 f_k$ may not always be positive definite, p_k^N may not always be a descent direction, so many of the ideas discussed in the part no longer apply. Here we discuss just the local rate-of-convergence properties of Newton's method. We know that for all x in the vicinity of a solution point x^* such that $\bigtriangledown^2 f(x^*)$ is positive definite, the Hessian $\bigtriangledown^2 f(x)$ will also be positive definite. Newton's method will be well defined in this region and will converge quadratically, provided that the step lengths α_k are eventually always 1.

Theorem 2.3 *suppose that f is twice differentiable and the Hessian $\bigtriangledown^2 f(x)$ is Lipschitz continuous in a neighborhood of a solution x^* at which the sufficient conditions are satisfied. Consider the iteration $x_{k+1} = x_k + p_k$, Then:*

1. *if the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*;*
2. *the rate of convergence of $\{x_k\}$ is quadratic;*
3. *the sequence of gradient norms $\{\| \bigtriangledown f_k\|\}$ converges quadratically to zero.*

2.1.4 Quasi-Newton Method

Suppose that we have the search direction has the form

$$p_k = -B_k^{-1} \bigtriangledown f_k,$$

where the symmetric and positive definite matrix B_k is updated at every iteration by a quasi-Newton updating formula. We assume here that the step length α is computed by an inexact line search that satisfies the Wolfe or strong Wolfe conditions, with the same provious mentioned above for Newton's method: The line search algorithm will always try the step length $\alpha = 1$ first, and will accept this value if it satisfies the Wolfe condition.This implementation detail turns out to be crucial in obtaining a fast rate of convergence.

Theorem 2.4 *Suppose that $f : \mathrm{R}^n \to \mathrm{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satifies the Wolfe conditions with $c_1 \leq \frac{1}{2}$. If the sequence x_k converges to a point x^* such that $\bigtriangledown^2 f(x^*)$ is positive definite, and if search direction satisfies*

$$\lim_{k\to\infty} \frac{\| \bigtriangledown f_k + \bigtriangledown^2 f_k p_k\|}{\|p_k\|} = 0, \tag{2.5}$$

then
(i) *the step length $\alpha_k = 1$ is admissible for all k greater than a certain index k_0; and*
(ii) *if $\alpha_k = 1$ for all $k > k_0, \{x_k\}$ converges to x^* superlinearly.*

If p_k is a quasi-Newton search direction of the form $p_k = -B_k^{-1} \bigtriangledown f_k$, then (2.5) is equivalent to

$$\lim_{k\to\infty} \frac{\|(B_k - \bigtriangledown^2 f(x^*))p_k\|}{\|p_k\|} = 0, \tag{2.6}$$

Hence, we have the result that a superlinear convergence rate can be attained even if the sequence of quasi-Newton matrices B_k does not converge to $\bigtriangledown^2 f(x^*)$;

Theorem 2.5 *Suppose that $f : \mathrm{R}^n \to \mathrm{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + p_k$ and that $p_k = -B_k^{-1} \bigtriangledown f_k$. Let us assume also that $\{x_k\}$ converges to a point x^* such that $\bigtriangledown f(x^*) = 0$ and $\bigtriangledown^2 f(x^*)$ is positive definite. Then x_k converges superlinearly if and only if (2.6) holds.*

Proof 2.1 *$p_k^N = - \bigtriangledown^2 f_k^{-1} \bigtriangledown f_k$ is the Newton direction. Assuming that (2.6) holds, we have that*

$$\begin{aligned} p_k - p_k^N &= \bigtriangledown^2 f_k^{-1}(\bigtriangledown^2 f_k p_k + \bigtriangledown f_k) \\ &= \bigtriangledown^2 f_k^{-1}(\bigtriangledown^2 f_k - B_k)p_k \\ &= O(\|(\bigtriangledown^2 f_k - B_k)p_k\|) \\ &= o(\|p_k\|), \end{aligned}$$

This proof need intermediate results in the proof the Theorem 4.3,

2.1.5 Newton's Method with Hessian Modification

Considering the Hessian matrix $\bigtriangledown^2 f(x)$ may not be positive definite, so the Newton direction p_k^N defined by

$$\bigtriangledown^2 f(x_k)p_k^N = - \bigtriangledown f(x_k) \tag{2.7}$$

may not be a descent direction. We now describe an approach to overcome this difficulty when a direct linear algebra technique, such as Gaussian elimination, is used to solve the Newton equations. This approach obtains the step p_k from a linear system identical to (2.7), except that the coefficient matrix is replaced with a positive definite approximation, formed before or during the solution process. The modified Hessian is obtained by adding either a positive diagonal matrix or a full matrix to the true Hessian $\bigtriangledown^2 f(x_k)$.

Algorithm 1 Line Search Newton with Modification

Given initial point x_0;
for k = 0, 1, 2, ...
    Factorize the matrix $B_k = \bigtriangledown^2 f(x_k) + E_k$, where $E_k = 0$ if $\bigtriangledown^2 f(x_k)$
        is sufficiently positive definite; otherwise, E_k is chosen to
        ensure that B_k is sufficiently positive definite;
    Solve $B_k p_k = - \bigtriangledown f(x_k)$;
    Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$, where α_k satisfies the Wolfe, Goldstein, or
        Armijo backtracking condition;
end

Some approaches do not compute E_k explicitly, but rather introduce extra steps and tests into standard factorization procedures, modifying these procedures these procedures "on the fly" so that the computed factors are the factors of a positive definite matrix. Strategies based on modifying a Cholesky factorization and on modifying a symmetric indefinite factorization of the Hessian are described in this section.
Algorithm 1 is a practical Newton method that can be applied from any starting point. We can establish fairly satisfactory global convergence results for it, provided that the strategy for choosing E_k(and hence B_k) satisfies the bounded modified factorization property. This property is that the matrices in the sequence $\{B_k\}$ have bounded condition number whenever the sequence of Hessian $\{\bigtriangledown^2 f(x_k)\}$ is bounded; that is

$$\kappa(B_k) = \|B_k\|\|B_k^{-1}\| \leq C, \ \textit{some } C > 0 \textit{ and all } k = 0, 1, 2, ...$$

If this property holds, the modified line search Newton method global convergence,

2.1.6 Backtracking Line Search Algorithm

In the backtracking line search we assume that $f : \mathbf{R}^n \to \mathbf{R}$ is differentiable and that we are given a direction d of strict descent at the current point x_c, that is $f'(x_c; d) < 0$.
    Initialization: Choose $\gamma \in (0, 1)$, and $c \in (0, 1)$.

Having x_c obtain x_+ as following:
STEP 1 : Compute the backtracking stepsize

$$t^* := max\rho^\nu$$
$$subject\ to\ \nu \in 0, 1, 2, ...\ and$$
$$f(x_c + \rho^\nu d) \leq f(x_c) + c\rho^\nu f'(x_c; d).$$

STEP 2 : Set $x_+ = x_c + t * d$.
The backtracking procedure of Step 1 is easy to program. The pseudo code follows:

$$f_c = f(x_c)$$
$$\Delta f = cf'(x_c; d)$$
$$newf = f(x_c + d)$$
$$t = 1$$
`while` $newf > f_c + t\Delta f$
$$t = \gamma t$$
$$newf = f(x_c + td)$$
`endwhile`

Convergence analysis

Algorithm 2 Global Backtracking

procedure Backtrackingglobal(x^0, σ_1, θ)
$k \to 0$
repeat
Find $d^k \in \mathbf{R}^n$ such that $\Delta f(x^k; d^k) < 0$
if no such d^k **then**
$0 \in \partial f(x_k)$ **return**
end if
$t \leftarrow 1$
while $f(x^k + kd^k) > f(x^k) + \sigma_1 t\Delta f(x^k; d^k)$ **do**
$t \leftarrow \gamma t$
end while
$t_k \leftarrow t$
$x^k \leftarrow x^k + t_k d^k$
$k \leftarrow k + 1$
until
end procedure

Let f be as $[f(x) = h(c(x)] + g(x)]$, $x^0 \in dom(g), 0 < \sigma_! < 1, and 0 < \theta < 1$. Set $L := lev_f(f(x^0))$. Suppose there exists $M > 0$ and $\widetilde{M} > 0$ such that $\|d^k\| \leq M$, $sup_{x \in L} \| \bigtriangledown c(x) \| \leq \widetilde{M}$, and that
(i) $\bigtriangledown c$ is $L_{\bigtriangledown c}$ - Lipschitz on $L + MB_n$;
(ii) h is $L_h - Lipschitz$ on $c(L + MB) + \widetilde{M}MB_m$. Let $\{x^k\}$ be a sequence initialized at x^0 and generated by Algorithm 4: Then one of the following must occur:
(a) the algorithm terminates finitely at a first-order stationary point for f.

(b) $f(x^k) \searrow -\infty$;
(c) $\sum_{k=0}^{\infty} \frac{\Delta f(x^k;d^k)^2}{\|d^k\|_2^2} < \infty$, in particular, $\Delta f(x^k;d^k) \to 0$.
Remark : When h is the identity on R on $g = 0$, we recover the convergence analysis of backtracking for smooth minimization.
Corollary : Let the hypotheses hold, if $0 < \beta < 1$ and the direction $\{d^k\}$ are chosen to satisfy

$$\Delta f(x^k;d^k) \leq \beta \bar{\Delta}_k f < 0,$$

then the occurrence of (c) implies that cluster points of $\{x^k\}$ are first-order stationary for $[f(x) = h(c(x)) + g(x)]$.

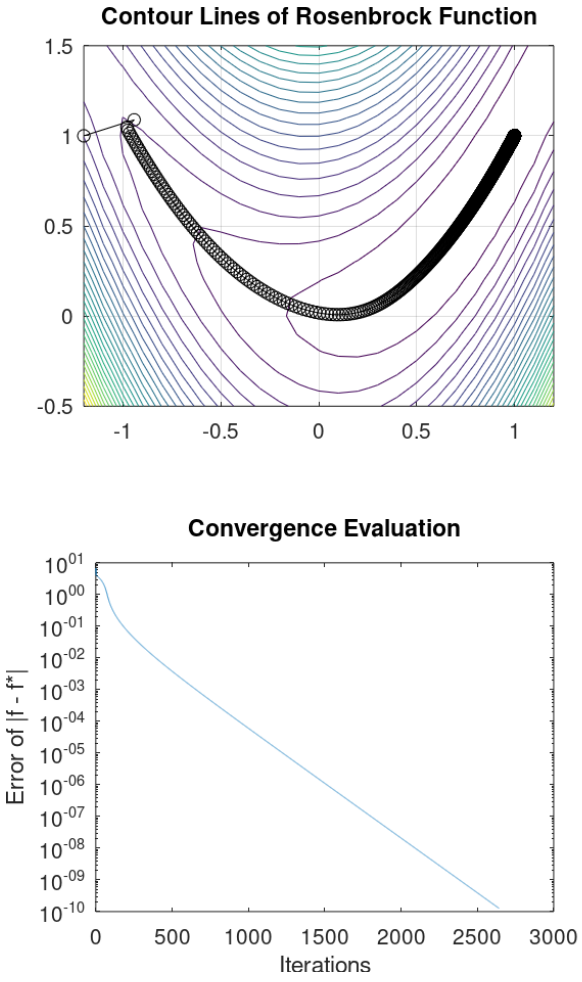


Figure 1: Backtracking method

Algorithm 3 Global Wolfe condition template

procedure Wolfeglobal($x^0, \sigma_1, \sigma_2, \mu$)
$k \leftarrow 0$
repeat
Find $d^k \in \mathrm{R}^n$ such that $\Delta f(x^k; d^k) < 0$
if no such d^k **then**
$0 \in \partial f(x^k)$ **return**
end if
Let t_k be a step size satisfying Conditions(Armijo, weak Wolfe condition, Wolfe condition, strong Wolfe condition.)
if no such t_k **then**
f unbounded below. **return**
end if
$x^k \leftarrow x^k + t_k d^k$
$k \leftarrow k + 1$
until
end procedure

2.1.7 Comparison with different conditions

Type	condition	Properties	Summary
Armijo Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$	sufficient decrease	
Weak Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and $\sigma_2\Delta f(x;d) \leq \frac{f(x+td;\mu d)}{\mu}$, with $0<\sigma_1<\sigma_2<1, \mu>0$.	sufficient decrease and curvature condition with a modification which prevents the line search early termination at "strongly negatige" slopes.	make d^k less of a direction of descent (and possibly a direction of ascent) at the new point.
Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and $\sigma_2 f'(x;d) \leq f'(x+td;d)$, with $0<\sigma_1<\sigma_2<1$.	sufficient decrease and curvature condition	
Strong Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and$\lvert f'(x+td;d)\rvert \leq \sigma_2\lvert f'(x;d)\rvert$, with $0<\sigma_1<\sigma_2<1$.	sufficient decrease and curvature condition with a modification to for t_k to lie in at least a broad neighborhood of a local minimizer or stationary point	don't allow the derivative to be too positive, try to push the directional derivative in direction d^k closer to zero at the new point.

Remark 1. The condition $\sigma_2\Delta f(x;d) \leq \frac{f(x+td;\mu d)}{\mu}$ is a *curvature condition* that parallels the classical weak Wolfe curvature condition for smooth, unconstrained minizaition:

$$\sigma_2 f'(x;d) \leq f'(x+td;d),$$

which prevents the line search early termination at "strongly negative" slopes.
Remark 2. The strong Wolfe condition require $|f'(x+td;d)| \leq -\sigma_2 f'(x;d)$, whenever f is smooth. However, in nonsmooth minimization, kinks and upward cusps at local minimizaters make the condition unworkable. Lemma 5.1 in [6] proved that the set of points satisfying weak Wolfe conditions has nonempty interior.
1.Armijo Condition: with *initial* $t=1$, $\gamma=0.5$, $\sigma_1=0.4$. ($t>0, \gamma\in(0,1), \sigma_1\in(0,0.5)$; it converges faster when $\gamma \searrow 0$, $\sigma_1 \nearrow 0.5$.)
2. Wolfe condition: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$
3. Weak Wolfe Conditions: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$
4. Strong Wolfe Conditions: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$

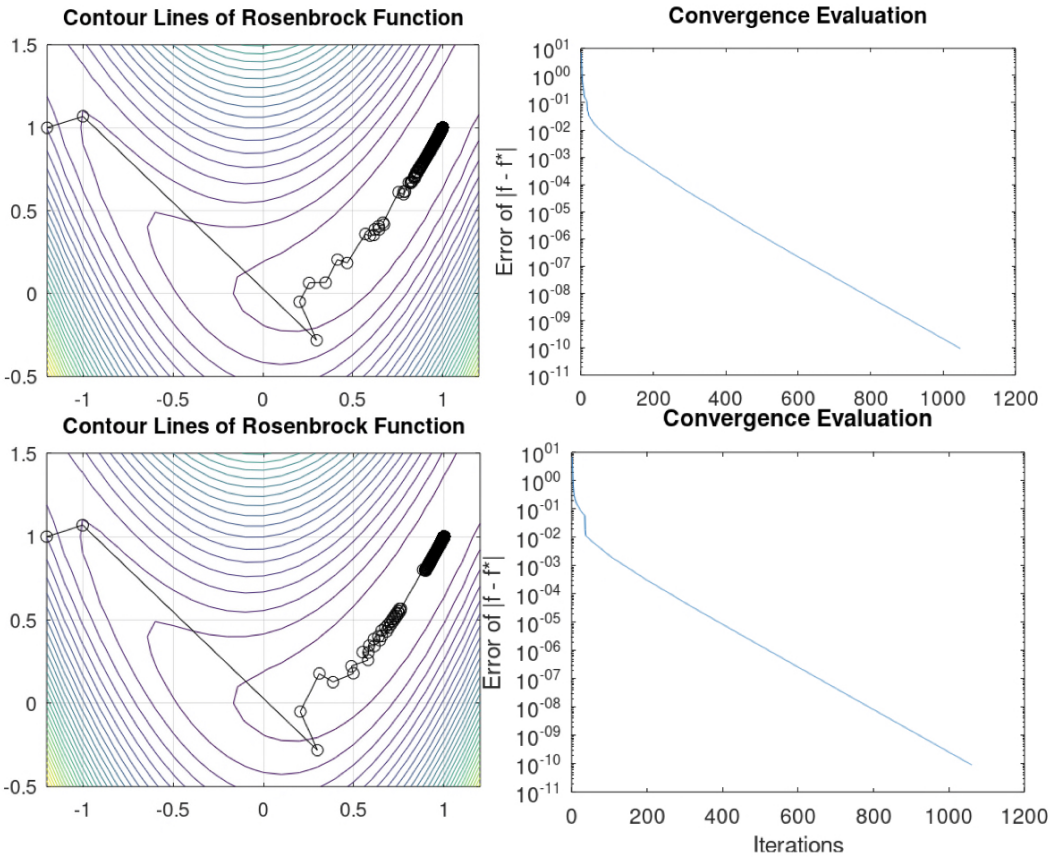


Figure 2: Comparison between Armijo and Wolfe conditions,programming in a native way

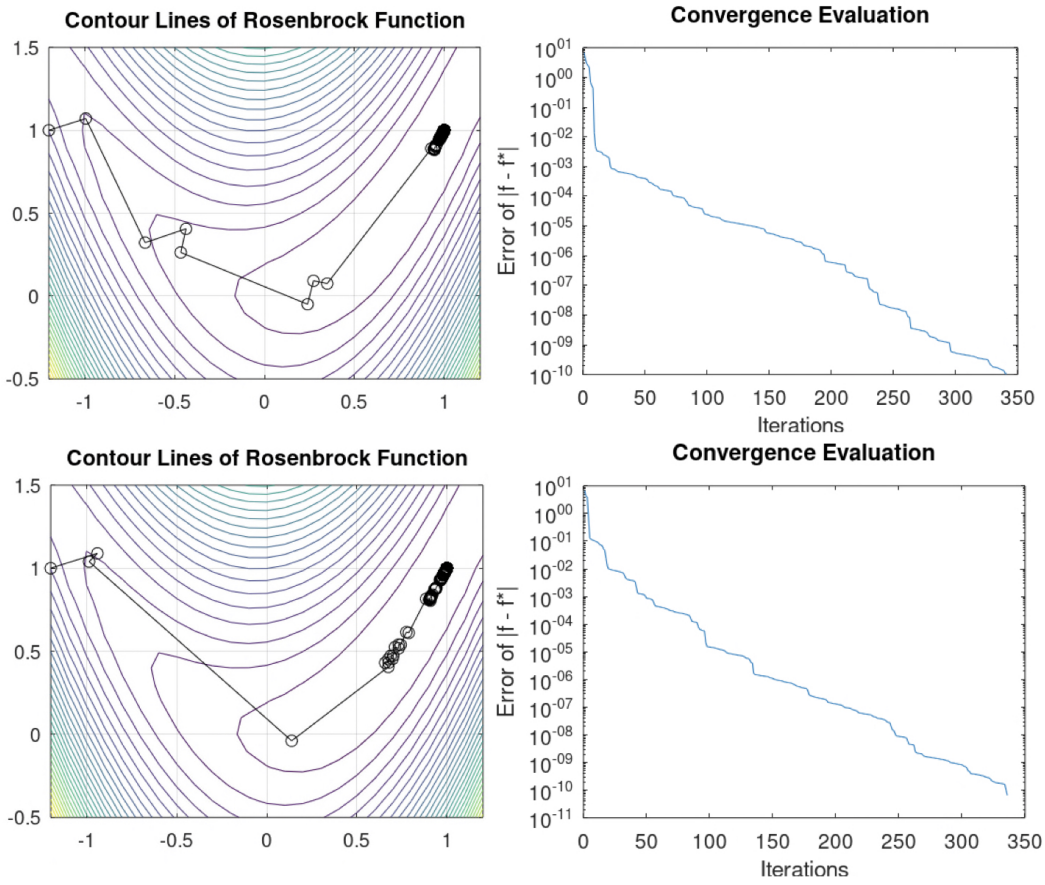


Figure 3: Comparison between Strong Wolfe conditions and Weak Wolfe Conditions, programming with zoom

2.1.8 Barzilai-Borwein step size method

It is a gradient method with modified step sizes, which are motivated by Newton's method but not involves any Hessian. At nearly no extra cost, the method often significantly improves the performance of a standard gradient method. The method is used along with non-monotone line search search as a safeguard.
Let $g^{(k)} = \bigtriangledown f(x^k)$ and $F^{(k)} = \bigtriangledown^2 f(x^{(k)})$. Gradient method: $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$, Newton's method: $x^{k+1} = x^{(k)} - (F^{(k)})^{-1} g^{(k)}$. The Barzilai-Borwein mnethod chooses α_k so that $\alpha_k g^{(k)}$ approximates $(F^{(k)})^{-1} g^{(k)}$ without computing $F^{(k)}$.
Consider

$$\underset{x}{minimize}\, f(x) = \frac{1}{2} x^T A x - b^T x,$$

where $A \succ 0$ is symmetric. Gradient is $g^{(k)} = Ax^{(k)} - b$. Hessian is A. Newton step: $d^{(k)}_{newton} = -A^{-1} g^{(k)}$.Goal: Choose α_k so that $-\alpha_k g^{(k)} = -(\alpha_k^{-1} I)^{-1} g^{(k)}$ approximates $-A^{-1} g^{(k)}$.
Define: $s^{(k-1)} := x^{(k)} - x^{(k-1)}$ and $y^{(k-1)} := g^{(k)} - g^{(k-1)}$. Then A satisfies:

$$As^{(k-1)} = y^{(k-1)}.$$

Therefore, given $s^{(k-1)}$ and $y^{(k-1)}$, how about choose α_k so that

$$(\alpha_k^{-1} I) s^{(k-1)} \approx y^{k-1}$$

Barzilai-Borwein method:

- Least-squares problem: (let $\beta = \alpha^{-1}$)

$$\alpha_k^{-1} = \underset{\beta}{argmin} \frac{1}{2} \|s^{(k-1)}\beta - y^{k-1}\|^2 \Longrightarrow \alpha_k^1 = \frac{(s^{(k-1)})^T s^{(k-1)}}{(s^{(k-1)})^T y^{(k-1)}}$$

- Alternative Least-squares problem:

$$\alpha_k = \underset{\alpha}{argmin} \|s^{(k-1)} - y^{(k-1)}\alpha\|^2 \Longrightarrow \alpha_k^2 = \frac{(s^{(k-1)})^T y^{(k-1)}}{(y^{(k-1)})^T y^{(k-1)}}$$

α_k^1 and α_k^2 are called the Barzilai-Borwein step sizes. Since $x^{(k-1)}$ and $g^{(k-1)}$ and thus $s^{(k-1)}$ and $y^{(k-1)}$ are unavailable at $k = 0$, we apply the standard gradient descent at $k = 0$ and start Barzilai-Borwein method at $k = 1$. We can use α_k^1 or α_k^2 or alternate between them. We can fix $\alpha_k = \alpha_k^1$ or $\alpha_k = \alpha_k^2$ for a few consecutive steps. It performs very well on minimizing quadratic and many other functions. However, f_k and $\| \bigtriangledown f_k \|$ are not monotonic. For quadratic functions, it has R-linear convergence. For 2D quadratic function, it has Q-superlinear convergence. No convergence guarantee for smooth convex problems. On these problems, we pair up Barzilai-Borwein with non-monotone line search.
Nonmonotone Line Search
Some growth in the function value is permitted, sometimes improve the likehood of finding a global optimum. Improve convergence speed when a monotone scheme is forced to creep along the bottom of a narrow curved valley. Comments:

Algorithm 4 Zhang-Hager nonmonotone line search

Initialize $0 < c_1 < c_2 < 1, C_0 \leftarrow f(x^0), Q_0 \leftarrow 1, \eta < 0, k \leftarrow 0$
While not converged do
compute α_k satisfying the modified Wolfe conditions OR
find α_k by backtracking, to satisfy the modified Armijo condition:
sufficient decrease: $f(x^{(k)} + \alpha_k d^{(k)}) \leq C_k + c_1 \alpha_k \bigtriangledown f_k^T d^{(k)}$
$x^{k+1} \leftarrow x^{(k)} + \alpha_k d^{(k)}$
$Q_{k+1} \leftarrow \eta Q_k + 1, C_{k+1} \leftarrow (\eta Q_k C_k + f(x^{k+1}))/Q_{K+1}$.

- If $\eta = 1$, then $C_k = \frac{1}{k+1} \sum_{j=0}^{k} f_j$.
- Since $\eta < 1$, C_k is a weighted sum of all past f_j, more weights on recent f_j.

Convergence: if $f \in C^1$ and bounded below, $\bigtriangledown f_k^T d^{(k)} < 0$, then

- $f_k \leq C_k \leq \frac{1}{k+1} \sum_{j=0}^{(k)} f_j$
- there exists α_k satisfying the modified Wolfe or Armijo conditions

In addition, if $\bigtriangledown f$ is Lipschitz with constant L, then $\alpha_k > C \frac{|\bigtriangledown f_k^T d^{(k)}|}{\|d^{(k)}\|}$ for some constant depending on c_1, c_2, L and the backing factor. Furthermore, if for all sufficiently large k, we have uniform bounds

$$\bigtriangledown f_k^T d^{(k)} \leq -c_3 \| \bigtriangledown f_k \|^2 \; and \; \|d^{(k)}\| \leq c_4 \| \bigtriangledown f_k \|$$

then $lim_{k \to \infty} \bigtriangledown f_k = 0$ Once again, pairing with non-monotone linear search, Barzilai-Borwein gradient methods work very well on general unconstrained differentiable problems.

2.2 Subgradient Methods

In this subsection, i mainly talk about the subgradient methods to handle some non-differential problems. But firstly i would like to introduce the original form: gradient methods by giving a brief recapitulating of it.

2.2.1 Gradient Methods

Gradient methods is the most fundamental method in optimization, but it has lots of important applications in Machine Learning algorithms. A gradient measures how much the output of a function changes if you change the inputs a little bit. And as we have already known, the antigradient is the direction of locally steepest descent of a differentiable function.

$$Choose : x_0 \in R^n.$$
$$Iterate : x_{k+1} = x_k - h_k \bigtriangledown f(x_k), k = 0, 1, ...$$

Function $f(x)$ defined on R^n, $\bigtriangledown f(x)$ is the differential function of $f(x)$. We refer to this scheme as the Gradient Method. The scalar factors for the gradients, h_k, are called the step sizes. They must be positive. There are many variants of this method, which differ one from another by the step-size strategy.

2.2.2 Subgradient

Definition 2.1 *We say a vector $g \in \mathrm{R}^n$ is a subgradient of $f : \mathrm{R}^n \to \mathrm{R}$ at $x \in domf$ if for all $z \in domf$,*

$$f(z) >= f(x) + g^T(z - x).$$

A subgradient can exist even when f is not differentiable at x. One way to interpret a subgradient, a vector g is a subgradient of f at x if the affine function of z $f(x)+g(z-x)$ is a global underestimator of f. Geometrically, g is a subgradient of f at x if $(g, -1)$ support $epif$ at $(x, f(x))$.

A function f is called subdifferentiable at x if there exists at least one subgradient at x. The set of subgradients of f at the point x is called the subdifferential of f at x, and is denoted as $\partial f(x)$. A function f is called subdifferentiable if it is subdifferentiable at all $x \in domf$. The subdifferential $\partial f(x)$ is always a closed convex set, even if f is not convex. This follows from the fact that it is the intersection of an infinite set of halfspaces:

$$\partial f = \bigcap_{z \in domf} \{g | f(z) \geq f(x) + g^T(z - x)\}.$$

In addition, if f is continuous at x, then the subdifferential $\partial f(x)$ is bounded. Indeed, choose some $\epsilon > 0$ such that $\infty < \underline{f} \leq f(y) \leq \bar{f} < \infty$ for all $y \in R^n$ such that $\|y-x\|_2 < \epsilon$.

A point $x^\star$ is a minimizer of a function f (not necessarily convex) if and only if f is subdifferentiable at $x^\star$ and

$$0 \in \partial f(x^\star),$$

i.e., $g = 0$ is a subgradient of f at $x^\star$. This follows directly from the fact that $f(x) \geq f(x^\star)$ for all $x \in domf$. And clearly if f is subdifferentiable at $x^\star$ with $0 \in \partial f(x^\star)$, then $f(x) \geq f(x^\star) + 0^T(x - x^\star)$ for all x.

While this simple characterization of optimality via the subdifferential holds for nonconvex function, it is not particularly useful in that case, since we generally cannot find the subdifferential of a nonconvex function. The condition $0 \in \partial f(x^\star)$ reduces to $\bigtriangledown f(x^\star) = 0$ when f is convex and differentiable at $x^\star$.

There is a somewhat more complex version of the result that $0 \in \partial f(x)$ if and only if x minimizes f for constrained minimization. Consider finding the minimizer of a subdifferentiable function f over a closed convex set X. Then we have $x^\star$ minimizes f if and only if there exists a subgradient $g \in \partial f(x^\star)$ such that

$$g^T(y - x^\star) \geq 0 \ for\ all\ y\ \in X.$$

For convex function f, the directional derivative of f at the point $x \in \mathrm{R}^n$ in the direction v is

$$f'(x;v) := \lim_{t \searrow 0} \frac{f(x+tv) - f(x)}{t}.$$

This quantity always exists for convex f, though it may be $+\infty$ or $-\infty$. The directional derivative $f'(x;v)$ satisfies the following genneral formula for convex f:

$$f'(x;v) = \sup_{g \in \partial f(x)} g^T v.$$

Nonnegative scaling: For $\alpha > 0, \partial(\alpha f)(x) = \alpha \partial f(x)$.
Sum and integral: Suppose $f = f_1 + f_2 + \cdots + f_m$, where $f_1, f_2, ..., f_m$ are convex functions. Then we have

$$\partial f(x) = \partial f_1(x) + \partial f_2(x) + \cdots + \partial f_m(x).$$

This property extends to infinite sums, integrals, and expectations(provided they exist).
Affine transformations of domain: Suppose f is convex, and let $h(x) = f(Ax + b)$. Then $\partial h(x) = A^T \partial f(Ax + b)$.
Pointwise maximum:Suppose f is the pointwise maximum of convex functions $f_1, ..., f_m, i.e.$,

$$f(x) = \max_{i=1,...,m} f_i(x),$$

where the functions f_i are subdifferentiable. We first show how to construct a subgradient of f at x. Let k be any index for which $f_k(x) = f(x)$, and let $g \in \partial f_k(x)$, then $g \in \partial f(x)$. In other words, to find a subgradient of the maximum of the maximum of functions, we can choose one of the functions that achieves the maximum at the point, and choose any subgradient of that function at the point. This follows from

$$f(z) \geq f_k(z) \geq f_k(x) + g^T(z - x) = f(x) + g^T(y - x).$$

More generally, we have

$$\partial f(x) = Co \cup \{\partial f_i(x) | f_i(x) = f(x)\},$$

$i.e.$, the subdifferential of the maximum of functions is the convex hull of the union of subdifferentials of the 'active' functions at x.

2.2.3 Subgradient Methods

The subgradient method is a very simple algorithm for minimizing a nondifferentiable convex function. The method looks very much like the ordinary gradient method for differentiable functions, but with several notable exceptions.

- The subgradient method applies directly to nondifferentiable f.
- The step lengths are not chosen via a line search, as in the ordinary gradient method. In the most common cases, the step lengths are fixed ahead of time.

- Unlike the ordinary gradient method, the subgradient method is not a descent method; the function can (and often does) increase.

By combing the subgradient method with primal or dual decomposition techniques, it is sometimes possible to develop a simple distributed algorithm for a problem.

We start with the unconstrained case, where the goal is to minimize $f : \mathrm{R}^n \to \mathrm{R}$, which is convex and has domain R^n(for now). To do this, the subgradient method uses the simple iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}.$$

Here $x^{(x)}$ is the kth iterate, $g^{(k)}$ is any subgradient of f at $x^{(k)}$, and $\alpha_k > 0$ is the kth step size. Thus, at each iteration of the subgradient method, we take a step in the direction of a negative subgradient.

In the subgradient method the step size selection is very different from the standard gradient method. Many different types of step size rules are used.

- Constant step size. $\alpha = \alpha$ is a positive constant, Independent of k.
- Constant step length. $\alpha_k = \gamma/\|g^{(k)}\|_2$, where $\gamma > 0$. This means that $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$
- Square summable but not summable. The step sizes satisfy

$$\alpha_k \geq 0, \ \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \ \sum_{k=1}^{\infty} \alpha_k = \infty.$$

  One typical example is $\alpha = a/(n + k)$, where $a > 0$ and $b \geq 0$.

- Nonsummable diminishing. The step sizes satisfy

$$\alpha_k \geq 0, \lim_{k\to\infty} \alpha_k = 0, \sum_{k=1}^{\infty} \alpha_k = \infty.$$

  Step sizes that satisfy this condition are called diminishing step size rules. A typical example is $\alpha_k = a/\sqrt{k}$, where $a > 0$.

- Nonsummable diminishing step lengths. The step sizes are chosen as $\alpha_k = \gamma_k/\|g^{(k)}\|_2$, where

$$\gamma_k \geq 0, \lim_{k\to\infty} \gamma_k = 0, \sum_{k=1}^{\infty} \gamma_k = \infty.$$

There are still other choices, and many variations on these choices. The most interesting feature of these choices is that they are determined before the algorithm is run; they do not depend on any data computed during the algorithm. This is very different from the step size rules found in standard descent method, which very much depend on the current point and search direction.

There are many results on convergence of the subgradient method. For constant step size and constant step length, the subgradient algorithm is guaranteed to converge to within some range of the optimal value, *i.e.*, we have

$$\lim_{k\to\infty} f_{best}^{(k)} - f^\star < \epsilon$$

where $f^\star$ denotes the optimal value of the problem. (This implies that the subgradient method finds an ϵ-suboptimal point within a finite number of steps.) The number of ϵ is a function of the step size parameter h, and decreases with it.
For the diminishing step size and step length rules(and therefore also the square summable but not summable step size rule), the algorithm is guaranteed to convergence to the optimal value, we have $lim_{k\to\infty} f(x^{(k)}) = f^\star$. It's remarkable that such a simple algorithm can be used to minimize any convex function for which you can compute a subgradient at each point.

When the function f is differentiable, we can say a bit more about the convergence. In this case, the subgradient method with constant step size yields convergence to the optimal value, provided the parameter α is small enough.

2.2.4 Convergence proof

Here we give a proof of some typical convergence results for the subgradient method. We assume that there is a minimizer of f, say $x^\star$. We also make one other assumption on f: We will assume that the norm of the subgradient is bounded. *i.e.*, there is a G such that $\|g^{(k)}\|_2 \leq G$ for all k. This will be the case if, for example, f satisfies the Lipschitz condition

$$|f(u) - f(v)| \leq G\|u - v\|_2,$$

for all u, v, because then $\|g\|_2 \leq G$ for any $g \in \partial f(x)$, at any x. In fact, some versions of the subgradient method (*eg*, diminishing nonsummable step lengths) work when this assumption doesn't hold. We'll also assume that a number R is known that satisfies $R \geq \|x^{(1)} - x^\star\|_2$. We can interpret R as an upper bound on $\text{dist}(x^{(1)}, X^\star)$, the distance of the initial point to the optimal set.

For the standard gradient descent method, the convergence proof is based on the function value decreasing at each step. In the subgradient method, the key quantity is not the function value(which often increase); it is the **Euclidean distance to the optimal set**. Reacall that $x^\star$ is a point that minimize f, *i.e.*, it is an arbitrary optimal point.

$$\begin{aligned}\|x^{(x+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\ &= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2\|g^{(k)}\|_2^2 \\ &\leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k(f(x^{(k)}) - f^\star) + \alpha_k^2\|g^{(k)}\|_2^2\end{aligned}$$

Applying the inequality above recursively, we have

$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^{k} \alpha_i(f(x^{(x^(i)} - f^\star). + \sum_{i=1}^{k} \alpha_i^2\|g^{(i)}\|_2^2.$$

Since $\|x^{(k+1)} - x^\star\|_2^2 \geq 0$ and $\|x^{(1)} - x^\star\|_2^2 \leq R$ we have

$$2\sum_{i=1}^{k} \alpha_i (f(x^{(i)} - f^\star) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2. \tag{2.8}$$

Combining this with

$$\sum_{i=1}^{k} \alpha_i (f(x^{(i)} - f^\star) \geq (\sum_{i=1}^{k} \alpha_i) \min_{i=1,\ldots,k} (f(x^{(i)} - f^\star) = (\sum_{i=1}^{k} \alpha_i)(f_{best}^{(k)} - f^\star),$$

Then we have the inequality

$$f_{best}^{(k)} - f^\star = \min_{i=1,\ldots,k} f(x^{(i)}) - f^\star \leq \frac{R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{2.9}$$

Finally, using the assumption $\|g^{(k)}\|_2 \leq G$, we obtain the basic inequality

$$f_{best}^{(k)} - f^\star \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{2.10}$$

From this inequality we can read off various convergence results.
Constant step size. When $\alpha_k = \alpha$, we have

$$f^(k)_{best} - f^\star \leq \frac{R^2 + G^2\alpha^2 k}{2\alpha k}.$$

The righthand side convergences to $G^2\alpha/2$ as $k \to \infty$. Thus, for the subgradient method with fixed step size α, $f_{best}^{(k)}$ convergences to within $G^2\alpha/2$ of the optimal. We also find that $f(x^{(k)} - f^\star \leq G^2\alpha$ within at most $R^2/(G^2\alpha^2)$ steps.
Square summable but not summable. Now we have

$$f_{best}^{(k)} - f^\star \leq \frac{R^2 + G^2\|\alpha\|_2^2}{2\sum_{i=1}^{k} \alpha_i}.$$

which converges to zeros as $k \to \infty$, since the numerator converges to $R^2 + G^2\|\alpha\|_2^2$, and this denominator grows without bound. Thus, the subgradient method converges.
Diminishing step size rule. If the sequence α_k converges to zeros and is nonsummabl, then the righthand side of the inequality converges to zero, which implies the subgradient method converges. To show that, let $\epsilon > 0$, then there exists an integer N_1 such that $\alpha_i \leq \epsilon/G^2$ for all $i > N_i$. There exists an inter N_2 such that

$$\sum_{i=1}^{N_2} \alpha_i \geq \frac{1}{\epsilon}(R^2 + G^2 \sum_{i=1}^{n_1} \alpha_i^2),$$

Since $\sum_{i=1}^{\infty} \alpha = \infty$. Let $N = max\{N1, N2\}$. Then for $k > N$, we have

$$\begin{aligned}
\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} &\leq \frac{R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} + \frac{G^2 \sum_{i=N_1+1}^{k} \alpha_i^2}{2\sum_{i=1}^{N_1} \alpha_i + 2\sum_{i=N_1+1}^{k} \alpha_i} \\
&\leq \frac{R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2}{(2/\epsilon)(R^2 + G^2 \sum_{i=1}^{N_1} \alpha_1^2)} + \frac{G^2 \sum_{i=N_1+1}^{k} (\epsilon\alpha_i/G^2)}{2\sum_{i=N_1+1}^{k} \alpha_i} \\
&= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Nonsummable diminishing step lenths. Finally, suppose that $\alpha_k = \gamma/\|g^{(k)}\|_2$, with γ_k nonsummable and converging to zero. The inequality becomes

$$f_{best}^{(k)} - f^\star \leq \frac{R^2 + \sum_{i=1}^k \gamma_k^2}{2\sum_{i=1}^k \alpha_i} \leq \frac{R^2 + \sum_{i=1}^k \gamma_k^2}{(2/G)\sum_{i=1}^k \gamma_i},$$

which converges to zero as $k \to 0$.
A stopping criterion: We can use (2.8) to find a lower bound on $f^\star$ that is sharper than the lower bounds (2.9) and (2.10), and can be used as a stopping criterion. Re-arranging (2.8) and using $R \geq \|x^{(1)} - x^\star\|_2$.we get

$$f^\star \geq l_k = \frac{2\sum_{i=1}^k \alpha_i f(x^{(i)} - R^2 - \sum_{i=1}^K \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^k \alpha_i},$$

which can be computed after the kth step. The sequence $l_1, l_2, ...$ need not increase, so we can keep track of the best lower bound on $f^\star$ found so far,

$$l_{best}^{(k)} = max\{l_1, l_2, ..., l_k\}.$$

We can terminate the algorithm when $f_{best}^{(k)} - l_{best}^{(k)}$ is smaller than some threshold. This bound is better than (2.10), since it doesn't depend on G, but it too goes to zero very slowly. For this reason, the subgradient method is usually used without any formal stopping criterion.

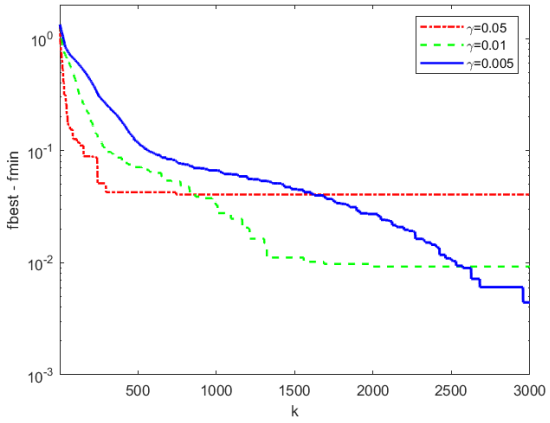


Figure 4: The value of $f_{best}^{(k)} - f^\star$ versus iteration number k, for the subgradient method with constant step length γ

.

2.2.5 Polyak's step length

In this subsection, we describe a subgradient step length choice due to Polyak.
Optimal step size choice when $f^\star$ is known. Polyak suggests a step size that can be

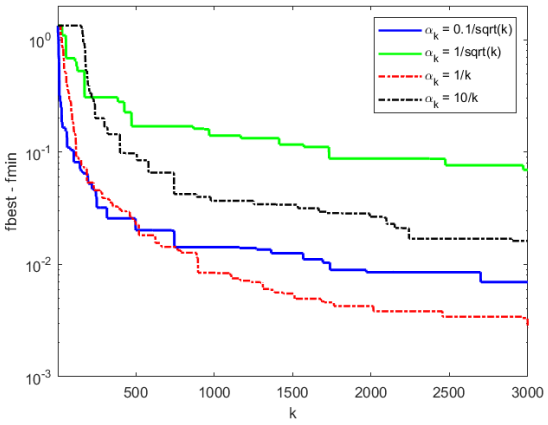


Figure 5: The value of $f^{(k)} - f^\star$ versus iteration number k, for the subgradient method with two diminishing step rules $\alpha_k = 0.1/\sqrt{k}$ and $\alpha_k = 1/\sqrt{k}$, and with two square summable step size rules $\alpha_k = 1/k$ and $\alpha_k = 10/k$.

used when the optimal value $f^\star$ is known, and in some sense optimal. (You might image that $f^\star$ is rarely known, but we will see that's not the case.) The step size is

$$\alpha_k = \frac{f(x^{(k)}) - f^\star}{\|g^{(k)}\|_2^2}.$$

To motivate this step size, imagine that

$$f(X^{(K)} - \alpha g^{(k)}) \approx f(x^{(x)}) + g^{(k)T}(x^{(k)} - \alpha g^{(k)} - x^{(k)}) = f(x^{(k)} - \alpha g^{(k)T} g^{(k)}.$$

(This would be the case if α were samll, and $g^{(k)} = \bigtriangledown f(x^{(k)})$.) Replacing the lefthand side with $f^\star$ and solving for α gives the step length above.

we can give another simple motivation for the step length. The subgradient method starts from the basic inequality.

$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k(f(x^{(k)} - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2.$$

The step size above minimizes the righthand side.
To analyze convergence, we substitute the step size into (2.8), to get

$$2\sum_{i=1}^{k} \frac{(f(x^{x(i)} - f^\star)^2}{\|g^{(i)}\|_2^2} \leq R^2 + \sum_{i=1}^{k} \frac{(f(x^{x(i)} - f^\star)^2}{\|g^{(i)}\|_2^2},$$

so

$$\sum_{i=1}^{k} \frac{(f(x^{x(i)} - f^\star)^2}{\|g^{(i)}\|_2^2} \leq R^2.$$

Using $\|g^{(i)}\|_2 \leq G$ we get

$$\sum_{i=1}^{k} (f(x^{(i)}) - f^\star)^2 \leq R^2 G^2.$$

We conclude that $f(x^{(k)} \to f^\star$. The number of steps needed before we can guarantee suboptimality ϵ is $k = (RG/\epsilon)^2$, which is optimal from our analysis.
Polyak step size choice with estimated f^*. The basic idea is to estimate the optimal value $f^\star$, as $f_{best} - \gamma^k$, where $\gamma^k > 0$ and $r^{(k)} \to 0$. This suggests the step size

$$\alpha_k = \frac{f(x^{(k)} - f_{best}^{(k)} + \gamma_k}{\|g^{(k)}\|_2^2}.$$

We'll also need $\sum_{k+1}^{\infty} \gamma_k = \infty$. Note that γ_k has a simple interpretation: It's our estimate of how suboptimal the current point is. Then we have $f_{best}^{(k)} \to f^*$.
To show this, we substitue α_i into the basic inequality (2.8) to get

$$\begin{aligned} R^2 &\geq \sum_{i=1}^{k} (2\alpha_i^2 \|g^{(i)}\|_2^2) \\ &= \sum_{i=1}^{k} \frac{2(f(x^{(i)}) - f_{best}^{(i)} + \gamma_i)(f(x^{(i)} - f_{best}^{(i)} + \gamma_i)^2}{\|g^{(i)}\|_2^2} \\ &= \sum_{i=1}^{k} \frac{(f(x^{(i)} - f_{best}^{(i)} + \gamma_i)(f(x^{(i)} - f^\star) + (f_{best}^{(i)} - f^\star) - \gamma_i}{\|g^{(i)}\|_2^2}. \end{aligned}$$

Now we can prove convergence. Suppose $f^{(k)} - f^\star \geq \epsilon > 0$. Then for $i = 1, ..., k, f(x^{(i)}) - f^\star \geq \epsilon$. Find N for which $\gamma_i \leq \epsilon$ for $i \geq N$. This implies the second term in the numerator is at least ϵ. And in particular, it is positive. It follows the terms in the sum above for $i \geq N$ are positive. Let S denote the sum above, up to $i = N - 1$. We assume $k \geq N$. We then have

$$\sum_{i=1}^{k} \frac{f(x^{(i)} - f_{best}^{(i)} + \gamma_i)(f(x^{(i)} - f^\star) + (f^{(i)})_{best} - \gamma_i}{\|g^{(i)}\|_2^2} \leq R^2 - S.$$

We get a lower bound on the lefthand inside using $f(x^{(i)} - f_{best}^{(i)} + \gamma_i \geq \gamma_i$, along with the inequality above and $\|g^{(i)}\|_2^2 \leq G$ to get

$$(\epsilon/G^2) \sum_{i=N}^{k} \gamma_i \leq R^2 - S.$$

Since the lefthand side convergence to ∞ and righthand side doesn't depend on k, we see that k cannot not be too large.

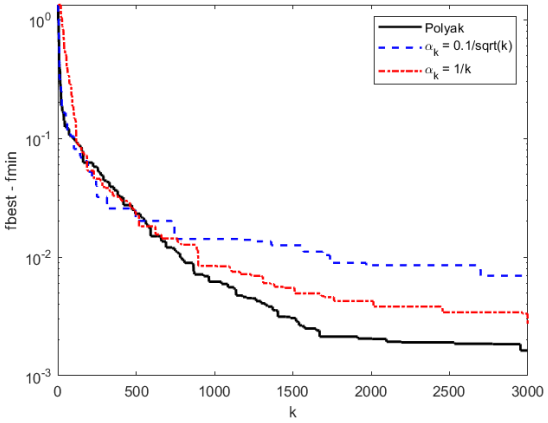


Figure 6: The value of $f_{best}^{(k)} - f^\star$ versus iteration number k, for the subgradient method with Polyak's step size and the subgradient methods with diminishing step sizes considered in previous examples

.

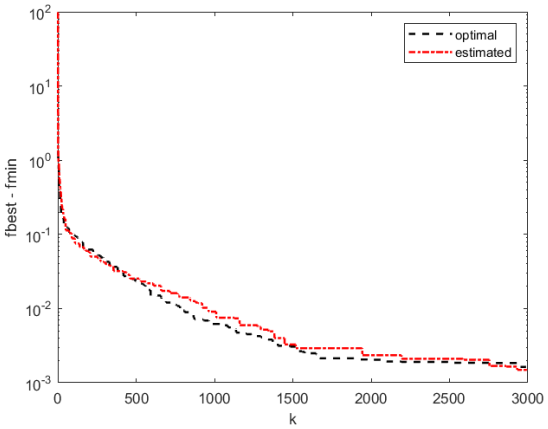


Figure 7: The value of $f_b^{(k)} est - f^\star$ versus iteration number k, for the subgradient method with Polyak's step size and the estimated optimal step size.

2.2.6 Alternating projections

Polyak's step length can be used to derive some versions of the alternating projections method for finding a point in the intersection of convex sets.
Suppose that we want to find a point in

$$C = C_1 \cap \cdots \cap C_m,$$

where $C_1, ..., C_m \subseteq \mathrm{R}^n$ are closed and convex, and we assume that C is nonempty. We can do this by minimizing the function

$$f(x) = max\{dist(x, C_1), ..., dist(x, C_m)\},$$

which is convex, and has minimization value $f^\star = 0$(since C is nonempty).

We explain how to find a subgradient g of f at x. If $f(x) = 0$, we can take $g = 0$ (which in any case means we are done). Otherwise find an index j such that $dist(x, C_j) = f(x), i.e.$, find a set that has a maximum distance to x. A subgradient of f is

$$g = \bigtriangledown dist(x, C_j) = \frac{x - \Pi_{C_j}(x)}{\|x - \Pi_{C_i}(x)\|_2},$$

where Π_{C-j} is Euclidean projection onto $C - j$. Note that $\|g\|_2 = 1$, so we can take $G = 1$.

The subgradient algorithm update, with step size rule and assuming that the index j is one for which $x^{(k)}$ has maximum distance to C_j, is given by

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k g^{(k)} \\ &= x^{(k)} - f(x^{(k)}) \frac{x^{(k)} - \Pi_{C_j}(x^{(k)}}{\|x^{(k)} - \Pi_{C_j}(x^{(k)})\|_2} \\ &= \Pi_{C_j}(x^{(k)}). \end{aligned}$$

Here we use $\|g^{(k)}\|_2 = 1$ and $f^\star = 0$ in the second line, and

$$f(x^{(k)}) = dist(x^{(k)}, C_j) = \|x^{(k)} - \Pi_{C_j}(x^{(k)})\|_2$$

in the thire line.

The algorithm is very simple: at each step, we simply project the current point onto the farthest set. This is an extension of the famous *alternatingprojections* algorithm. (When there are just two sets, then at each step you project the current point onto the other set. Thus the projections simply alternate.)

We are only guaranteed that $f(x^{(k)} \to f^\star = 0$. In the other words, a subsequence of our points approaches a point in C; we are not guaranteed to actually find a point in C (except in the limit). This can be adressed several ways. One way is to run the algorithm using closed sets $\widetilde{C_i} \subseteq intC_i$, so that $x^{(k)} \to \widetilde{C} = \widetilde{C_1} \cap \cdots \cap \widetilde{C_m}$. Then we are guaranteed that $x^{(k)} \in C$ for some (finite) k.

Another method is to do $over - projection$ at each step. Suppose we know the intersection of the sets contains a Euclidean ball of radius ϵ. Its center is a point that is ϵ-deep in all the sets. Then we can over project by ϵ, which roughly speaking means we project the current point to the farthest set, and then keep moving a distance ϵ:

$$x^{(k+1)} = \Pi_{C_j}(x^{(k)}) - \epsilon \frac{x^{(k)} - \Pi_{C_j}(X^{(k)}}{\|x^{(k)} - \Pi_{C_j}(x^{(k)}\|_2}.$$

Alternating projections is usually applied when projection onto the sets is simple. It can be used, of course, in cases where a bit more computation is needed to compute the Euclidean projection, *e.g.*, for a polyhedron(which can be done by solving a QP).
We want to find a point that satisfies $f_i(x) \leq 0, i = 1, ..., m$. (We assume we can find a subgradient of each function, at any point.)

To solve this set of convex inequalities, we can minimize the unconstrained function $f(x) = max_i f_i(x)$ using the subgradient method. If the set of inequalities is strictly feasible, then $f^\star$ is negative, and in a finite number of steps we'll find a point with $f(x) \leq 0, i.e.$, a feasible point. We can also use the step size taht uses knowledge of the optimal value, applied to the function

$$f(x) = max f_1(x), ..., f_m(x), -\epsilon,$$

where $\epsilon > 0$ is a tolerance. Assuming there exists a point with $f_i(x) \leq -\epsilon$, we can use the step length

$$\alpha = \frac{f(x) + \epsilon}{\|g\|_2^2}.$$

We can give a simple interpretation of this step length, taking the case $\epsilon = 0$ for simplicity. Suppose the current point is x, and that $f_i(x) = f(x) > 0$, with $g \in \partial f_i(x)$. Let $x^\star$ be any point with $f_i(x^\star) \leq 0$. Then we have

$$0 \geq f_i(x^\star) \geq f_i(x) + g^T(x^\star - x),$$

$i.e.$, $x^\star$ is in the halfspace

$$H = \{z | 0 \geq f_i(x) + g^T(z - x)\}.$$

The subgradient update at x, using Polyak's step length, is just projection of x onto the halfspace H.

2.2.7 Projected subgradient method

One extension of the subgradient method is the *projected subgradient method*, which solves the constrained convex optimization problem

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ x \in C, \end{aligned}$$

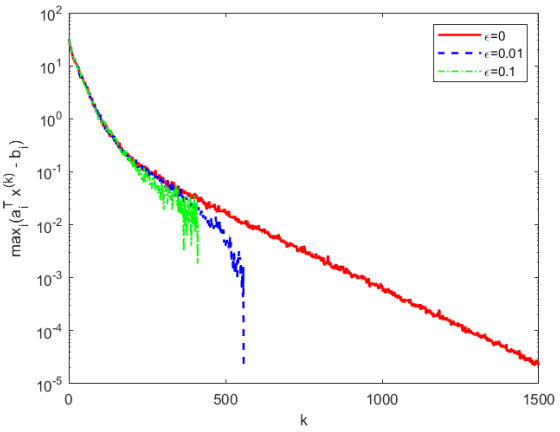


Figure 8: Convergence of the maximum violation for the linear feasibility problem, where we use the subgradient method with Polyak's step and three different value of tolerence ϵ.

where C is a convex set. The projected subgradient method is given by

$$x^{(k+1)} = \Pi(x^{(k)} - \alpha_k g^{(k)}),$$

Where Π is (Euclidean) projection on C, and $g^{(k)}$ is any subgradient of f at $x^{(k)}$. The step size rules described before can be used here, with similar convergence results. Note that $x^{(k)} \in C$,*i.e.*, $x^{(k)}$ is feasible.

The convergence proofs for the subgradient method are readily extended to handle the projected subgradient method. Let $z^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$, *i.e.*, a standard subgradient update, before the projection back onto C. As in the subgradient method, we have

$$\begin{aligned}\|z^{(k+1)} - x^\star\|_2^2 =& \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\ =& \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ \leq& \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2.\end{aligned}$$

Now we observe that

$$\|x^{(k+1)} - x^\star\|_2 = \|\Pi(z^{(k+1)}) - x^\star\|_2 \leq \|z^{(k+1)} - x^\star\|_2^2,$$

when we project a point onto C, we move closer to every point in C, and in particular, any optimal point. Combining this with the inequality above we get

$$\|x^{(k+1)} - x^\star\| \leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2,$$

and the proof proceeds exactly as in the ordinary subgradient method. Also, note that Polyak's step size can be applied here directly, and has the same convergence guarantee.

In some cases we can express the projected subgradient update in an alternative way. When C is affine, *i.e.*, $C = \{x | Ax = b\}$, where A is fat and full rank, the projection operator is affine, and given by

$$\Pi(z) = z - A^T(AA^T)^{-1}(Az - b).$$

In this case, we can simplify the subgradient update to

$$x^{(k+1)} = x^{(k)} - \alpha_k (I - A^T (AA^T)^{-1} A) g^{(k)}, \tag{2.11}$$

where we use $Ax^{(k)} = b$. Thus we simply project the current subgradient onto the nullspace of A, and then update as usual. The update (2.11) is not the same as the projected subgradient update when C is not affine, because in this case the projection operator is not affine.

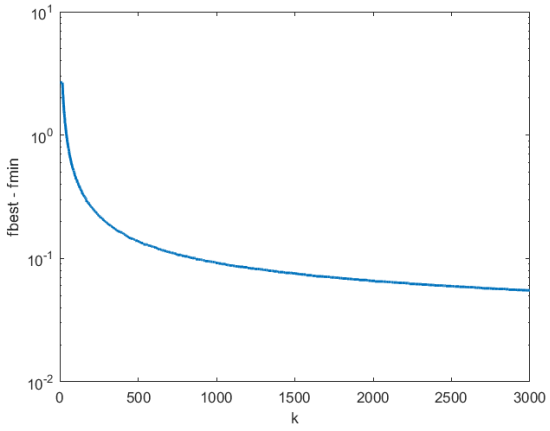


Figure 9: The value of $f_{best}^{(k)} - f^\star$ versus iteration number k, for the subgradient method with the Polyak estimated step size rule $r_k = 0.1/k$

Projected subgradient for dual problem
One famous application of the projected subgradient method is to the dual problem. We start with the (convex) primal problem

$$\begin{aligned} &minimize\ f_0(x) \\ &subject\ to\ f_i(x) \leq 0, i = 1, ..., m \end{aligned}$$

We'll assume, for simplicity, that for each $\lambda \geq 0$, the Lagrangian

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

has a unique minimizer over x, which we denote $x^\star(\lambda)$. The dual function is then

$$g(\lambda) = \inf_x L(x, \lambda) = f_0(x^\star(\lambda)) + \sum_{i=1}^{m} \lambda_i f_i(x^\star(\lambda))$$

(for $\lambda \geq 0$). The dual problem is

$$\begin{aligned} &maximize\ g(\lambda) \\ &subject\ to\ \lambda \geq 0. \end{aligned}$$

We'll assume that Slater's condition holds(again, for simplicity), so we can solve the primal problem by finding an optimal point $\lambda^\star$ of the dual, and then taking $x^\star = x^\star(\lambda^\star)$. We will solve the dual problem using the projected subgradient method,

$$\lambda^{(k+1)} = (\lambda^{(k)} - \alpha_k h)_+, \ h \in \partial(-g)(\lambda^{(k)}).$$

Let's now work out a subgradient of the negative dual function. Since -g is a supremum of a family of affine functions of λ, indexed by x, we can find a subgradient by finding one of these functions that achieves the supremum. But there is just one, and it is

$$-f_0(x^\star(\lambda)) - \sum_{i=1}^m \lambda_i f_i(x^\star(\lambda)),$$

which has gradient (with respect to λ)

$$h = -(f_1(x^\star(\lambda)), ..., f_m(x^\star(\lambda)) \in \partial(-g)(\lambda).$$

(Our assumptions imply that $-g$ has only one element in its subdifferential, which means g is differentiable. Differentiability means that a small enough constant step size will yield convergence. In any case, the projected subgradient method can be used in cases where the dual is nondifferentiable.)

$$x^{(k)} = \underset{x}{argmin}(f_0(x) + \sum_{i=1}^m \lambda_i^{(k)} f_i(x))$$
$$\lambda_i^{(k+1)} = (\lambda_i^{(k)} + \alpha_k f_i(x^{(k)}))_+.$$

In this algorithm, the primal iterates $x^{(k)}$ are not feasible, but become feasible only in the limit. (Sometimes, we can find a method for constructing a feasible, suboptimal $\widetilde{x}^{(k)} from\ x^{(k)}$.) The dual function values $g(\lambda^{(k)})$, as well as the primal function values $f_0(x^{(k)})$, converge to $f^\star = f_0(x^\star)$.

We can interpret λ_i as the price for a "resource" with usage measured by $f_i(x)$. When we calculate $x^\star(\lambda)$, we are finding the x that minimizes the total cost, *i.e.*, the objective plus the total bill(or revenue) for the resources used. The goal is to adjust the prices so that the resource usage is within budget($i.e., f_i(x) \leq 0.$). At each step, we increase the price λ_i if resourece i is over-utilized ($i.e., f_i(x) > 0$), and we decrease the price λ_i if resource i is under-utilized($i.e., f_i(x) < 0$). But we never let prices get negative (which would encourge, rather than discourage, resource usage).

In general, there is no reason to solve the dual instead of the primal. But for specific problems there can be an advantage.

Numerical Example

We consider the problem of minimizing a strictly convex quadratic function over the unit box:

$$minimize\ (1/2)x^T Px - q^T x$$
$$subject\ to\ x_i^2 \leq 1,\ i = 1, ..., n,$$

where $P \succ 0$. The Lagrangian is

$$L(x, \lambda) = (1/2)x^T(P + diag(2\lambda))x - q^T x - \mathbf{1}^T\lambda,$$

so $x^\star = \frac{q}{p+diag(2\lambda^{(k)})}$. The projected subgradient algorithm for the dual is

$$x^{(k)} = (P + diag(2\lambda))^{-1}q,\ \lambda_i^{k+1} = (\lambda_i^{(k)} + \alpha_k((x_i^k)^2 - 1))_+.$$

The dual function is differentible, so we can use a fixed size α (provided it is small enough).

The iterates $x^{(k)}$ are not feasible. But we can construct a nearby feasible $\hat{x}^{(k)}$ as

$$\hat{x}_i^{(k)} = \begin{cases} 1, & x_i^{(k)} > 1 \\ -1, & x_i^{(k)} < -1 \\ x_i^{(k)}, & -1 \le x_i^k \le 1. \end{cases}$$

We consider an instance with $n = 50$. We start the algorithm with $\lambda^{(1)} = 1$, and use a fixed step size $\alpha = 0.1$. Figure shows the convergence of $g(\lambda^{(k)})$ (a lower bound on the optimal value) and $f_0(\hat{x}^{(k)})$(an upper bound on the optimal value), versus iterations.

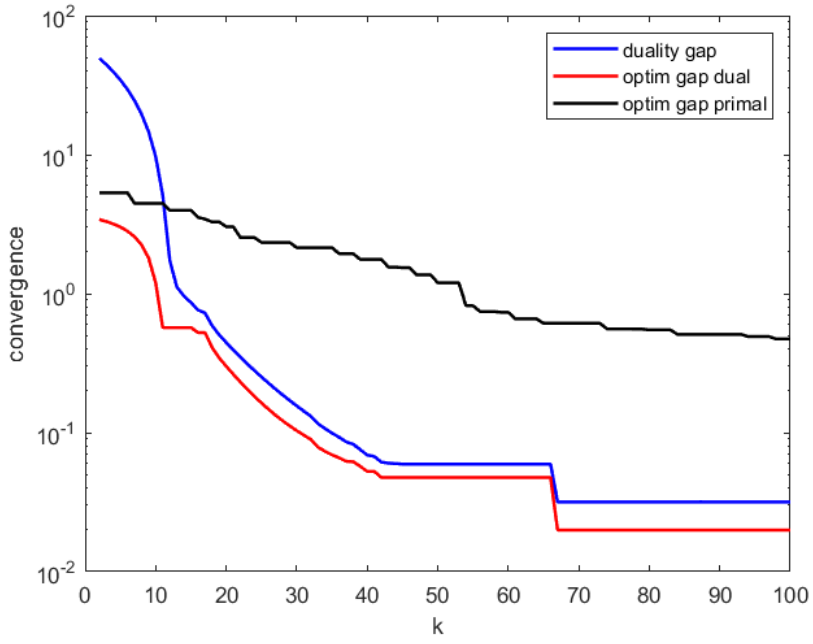


Figure 10: Comparison of different gaps, *duality gap* : $fbest - gbest$; *optim gap dual* : $fbest - f_{min}$; *optim gap primal* : $fpbest - f_{min}$

.

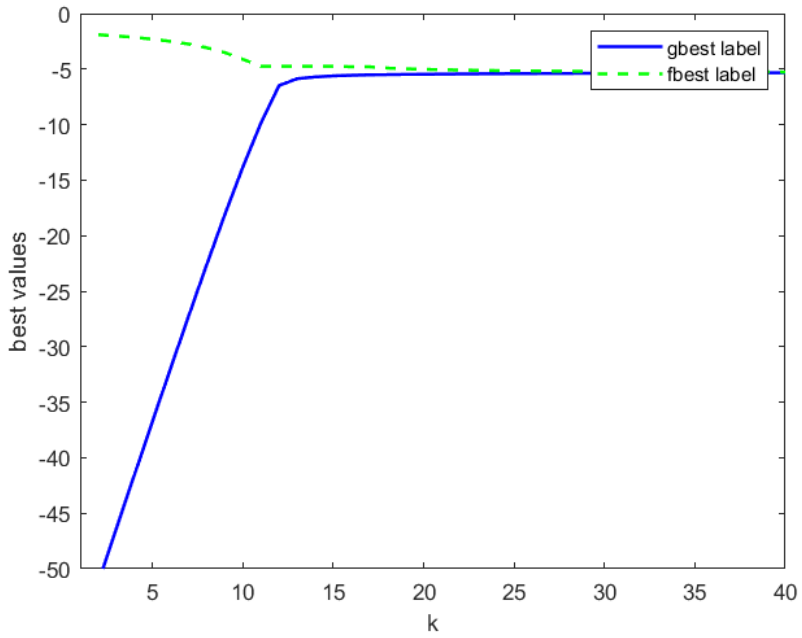


Figure 11: The values of the gbest and fbest, versus the iteration k. We use the fixed step size with $\alpha = 0.1$.

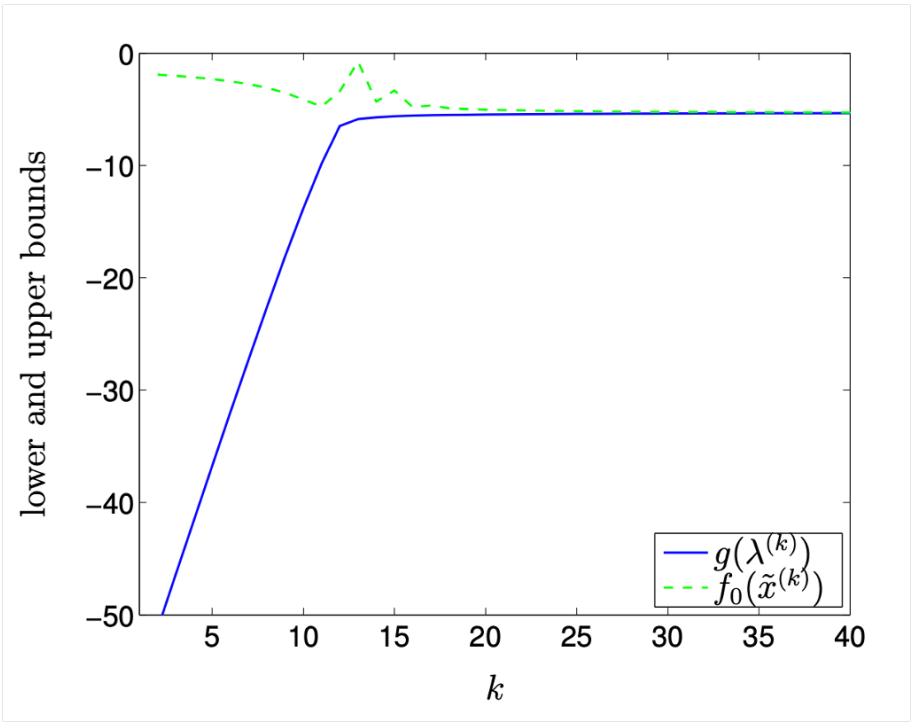


Figure 12: The values of the lower bound $g(\lambda^{(k)})$ and the upper bound $f_0(\hat{x}^{(k)})$, versus the iteration k. We use the fixed step size with $\alpha = 0.1$.

2.2.8 Primal-dual subgradient method

The *primal − dual subgradient method* is an extension of the subgradient method that solves constrained convex optimization problem of the form

$$\begin{aligned} &minimize f_0(x) \\ &subject\ to\ f_i(x) \leq 0,\ i = 1, ..., m \\ &Ax = b, \end{aligned}$$

with variable $x \in \mathrm{R}^n$, where the functions $f_0, f_1, ...f_m : \mathrm{R}^n \to \mathrm{R}$ are convex, not necessarily differentiable, and have domain R^n. In the following sections, we will consider problems that have the equality constraint only, and show the method with proof of its convergence. Extending these ideas to handle problems with both inequality and equality constraints is simple.

Equality constrained problems

In this section, we consider the equality constrained problem

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ Ax = b, \end{aligned}$$

with variable $x \in \mathrm{R}^n$. We assume that $A \in \mathrm{R}^{m \times n}$, *i.e.*, there are m equality constraints. We focus on solving the so-called *augumentedproblem*

$$\begin{aligned} &minimize\ f(x) + (\rho/2)\|Ax - b\|_2^2 \\ &subject\ to\ Ax = b. \end{aligned}$$

Let $L(x, v) = f(x) + v^T(Ax - b) + (\rho/2)\|Ax - b\|_2^2$ denote the lagrangian for the augmented problem, which is also called the augmented Lagrangian. We define a set-valued mapping T by

$$T(x, v) = \begin{bmatrix} \partial_x L(x, v) \\ -\partial_v L(x, v) \end{bmatrix} = \begin{bmatrix} \partial f(x) + A^T v + \rho A^T(Ax - b) \\ b - Ax \end{bmatrix}$$

The optimality condition for the augmented problem(and the original one as well) is

$$0 \in T(x^\star, v^\star).$$

Such a primal-dual pair is a saddle-point of the augmented Lagrangian:

$$L(x^\star, v) \leq L(x^\star, v^\star) \leq L(x, v^\star)$$

for all x and all v.

The primal-dual subgradient method finds a saddle point of the Lagrangian by the simple iteration that resembles the Uzawa iteration:

$$z^{(k+1)} = z^{(k)} - \alpha_k T^{(k)},$$

where $z^{(k)} = (x^{(k)}, v^{(k)}$ is the kth iterate of the primal and dual variables, $T^{(k)}$ is any element of $T(z^{(k)}$, and $\alpha_k > 0$ is the kth step size. By expanding it out, we can also write the method as

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k(g^{(k)} + A^T v^{(k)} + \rho A^T(Ax^{(k)} - b)) \\ v^{(k+1)} &= v^{(K)} + \alpha_k(Ax^{(k)} - b). \end{aligned}$$

Here, $g^{(k)}$ is any subgradient of f at $x^{(k)}$. Notice that $x^{(k)}$ is not necessarily feasible.

Let $z^\star = (x^\star, v^\star)$ be any pair of optimal variables, satisfying

$$Ax^\star = b,\ 0 \in \partial_x L(x^\star, v^\star).$$

We use $p^\star = f(x^\star)$ to denote the optimal value. We prove that the algorithm converges, *i.e.*,

$$\lim_{k\to\infty} f(x^{(k)}) = p^\star,\ \lim_{k\to\infty} \|Ax^{(k)} - b\|_2 = 0,$$

using the step size rule $\alpha_k = \gamma_k / \|T^{(k)}\|_2$, where γ_k is chosen under *square summable but not summable*.

For the convergence proof, we will assume that a number R is known that satisfies $R \geq \|z^{(1)}\|_2$ and $R \geq \|z^\star\|_2$. We will also assume that the norm of the subgradients of f is bounded on compact sets.

$$\|z^{(k+1)} - z^\star\|_2^2 + 2\sum_{i=0}^{k} \gamma_i \frac{T^{(i)T}}{\|T^{(i)}\|_2}(z^{(i)} - z^\star) = \|z^{(i)} - z^\star\|_2^2 + \sum_{i=0}^{k} \gamma_i^2 \leq 4R^2 + S.$$

Since the both terms on the lefthand side are nonnegative, for all k, we have

$$\|z^{(k+1)} - z^\star\|_2^2 \leq 4R^2 + S,\ 2\sum_{i=0}^{k} \gamma_i \frac{T^{(i)T}}{\|T^{(i)}\|_2}(z^{(i)} - z^\star) \leq 4R^2 + S,$$

We assumed that $\|z^\star\|_2$ is bounded, so the first inequality implies that $z^{(k)}$ cannot be too far from the origin, In other words, there exists a number D satisfying $\|z^{(k)}\|_2 \leq D$ for all k, namely $D = R + \sqrt{4R^2 + S}$. By assumption, the norm of subgradients on the set $\|x^{(k)}\|_2 \leq D$ is bounded, so it follows that $\|T^{(k)}\|_2$ is bounded.

$$0 \leq L(x^{(k)}, v^\star) - L(x^\star, v^\star) + (\rho/2)\|Ax^{(k)} - b\|_2^2 \leq T^{(k)T}(z^{(k)} - z^\star),$$

Implies that

$$\lim_{k\to\infty} L(x^{(k)}, v^\star) = L(x^\star, v^\star) = p^*,\ \lim_{k\to\infty} \|Ax^{(k)} - b\|_2 = 0.$$

Finally,

$$p^\star = \lim_{k\to\infty} L(x^{(k)} + \lim_{k\to\infty} v^{\star T}(Ax^{(k)} - b) = \lim_{k\to\infty} f(x^{(k)})$$

Inequality constrained problems

The method from the previous section can be modified to handle inequality constrained problems. Suppose that we want to solve the problem

$$\begin{aligned} &minimize\ f_o(x) \\ &subject\ to\ f_i(x) \leq 0, i = 1, ..., m. \end{aligned}$$

with variable $x \in \mathrm{R}^n$.

Notice that $(\rho/2)\sum_{i=1}^{m}(f_i(x)_+)^2$, with $\rho > 0$, can be added to the objective without changing the optimal value or the set of optimal. Thus, from now on, we focus on solving the augmented problem

$$\begin{aligned} &minimize\ f_0(x) + (\rho/2)\|F(x)\|_2^2 \\ &subject\ to\ F(x) \leq 0, \end{aligned}$$

where F is defined by

$$F(x) = \begin{bmatrix} f_1(x)_+ \\ \vdots \\ f_m(x)_+ \end{bmatrix}.$$

Let $L(x, \lambda) = f_0(x) + \lambda^T F(x) + (\rho/2)\|F(x)\|_2^2$ be the augmented Lagrangian. We define a set-valued mapping T by

$$T(x, \lambda) = \begin{bmatrix} \partial_x L(x, \lambda) \\ -\partial_\lambda L(x, \lambda) \end{bmatrix} = \begin{bmatrix} \partial f_0(x) + \sum_{i=1}^m (\lambda_i + \rho f_i(x)_+)\partial f_i(x)_+ \\ -F(x) \end{bmatrix}.$$

The optimality condition for the augmented problem is then

$$0 \in T(x^\star, \lambda^\star).$$

Such a primal-dual pair is a saddle-point of the augmented Lagrangian:

$$L(x^\star, \lambda) \leq L(x^\star, \lambda^\star) \leq L(x, \lambda^\star)$$

for all x and all λ.

The primal-dual subgradient method can be written as

$$z^{(k+1)} = z^{(k)} - \alpha_k T^{(k)},$$

where $z^{(k)} = (x^{(k)}, \lambda^{(k)})$ is the kth iterate of the primal and dual variables. By expanding it out, we can also also write the method as

$$\begin{aligned} x^{k+1} &= x^{(k)} - \alpha_k (g_0^{(k)} + \sum_{i=1}^m (\lambda_i^{(k)} + \rho f_i(x^{(k)})_+) g_i^{(k)}) \\ \lambda_i^{(k+1)} &= \lambda_i^k + \alpha_k f_i(x^{(k)}),\ i = 1, ..., m. \end{aligned}$$

Here, $g^{(k)}$ is any subgradient of f_i at $x^{(k)}$. The basic condition we set as previous.

$$\|z^{(k+1)} - z^\star\|_2^2 + 2\sum_{i=0}^k \gamma_i \frac{T^{(i)T}}{\|T^{(i)}\|_2}(z^{(i)} - z^\star) = \|z^{(i)} - z^\star\|_2^2 + \sum_{i=0}^k \gamma_i^2 \leq 4R^2 + S.$$

Again, we claim that the sum on the lefthand side is nonnegative. By expanding out,

$$T^{(k)T}(z^{(k)} - z^\star) = (g_0^{(k)} + \sum_{i=1}^m (\lambda_i^{(k)} + \rho f_i(x^{(k)})_+) g_i^{(k)})^T (x^{(k)} - x^\star) - F(x^{(k)})^T (\lambda^{(k)} - \lambda^\star).$$

By definition of subgradient and the constraints, we have

$$\begin{aligned} T^{(k)T}(z^{(k)} - z^\star) &\geq f_0(x^{(k)}) - p^\star + \lambda^{\star T} F(x^{(k)} + \rho \|F(x^{(k)})\|_2^2 \\ &= L(x^{(k)}, \lambda^\star) - L(x^\star, \lambda^\star) + (\rho/2)\|F(x^{(k)}\|_2^2 \\ &\geq 0. \end{aligned}$$

The rest of the proof proceeds exactly the same as in the cse of equality constrained problems.

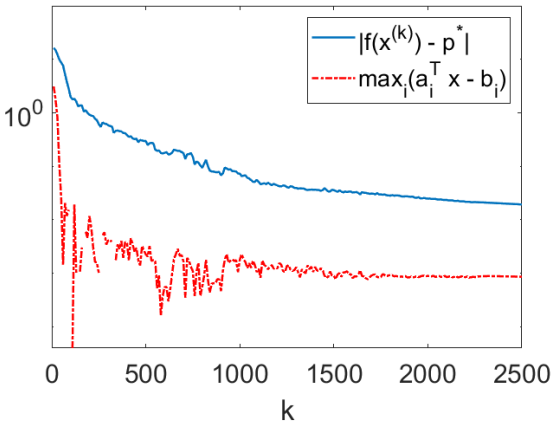


Figure 13: The suboptimality $|f(x^{(k)} - p^*|$, and the maximum violation of the constraints $max_{i=1,\ldots,m}(a_i^T x - b_i)$, versus the iteration number k. In this case, we use the square summable sequence $\gamma_k = 1/k$ to determine the step sizes.

2.2.9 Speeding up subgradient methods

Several general approaches can be used to speed up subgradient methods. *Localizationmethods* such as cutting-plane and ellipsoid methods also require the evaluation of one subgradient per iteration, but require more computation to carry out the update. They are typically much faster than subgradient methods. Some of these methods have real (non-heuristic) stopping criteria.

Another general approach is to base the update on some combination of previously evaluated subgradients. One general class of methods uses an update direction that is a conic combination of the current negative and the last search direction, as in

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

where α_k and β_k are positive. Such algorithm have state, whereas the basic subgradient method is stateless(except for the iteration number). We can interpret the second term as a memory term, or as a momentum term, in the algorithm. Polyak refers to some algorithms of this form as the *heavyballmethod*. *Conjugategradients* methods have a similar form.

We describe two examples of these types of methods , that use a known(or estimated) value of $f^\star$ to determine step lengths. Each has an update of the form

$$x^{(k+1)} = x^{(k)} - \alpha_k s^{(k)}, \ \alpha_k = \frac{f(x^{(k)}) - f^\star}{\|s^{(k)}\|_2^2},$$

where $s^{(k)}$ is a direction to be used in place of a subgradient. In the simple method, $s^{(k)}$ is just a filtered, or smoothed, version of the subgradients:

$$s^{(k)} = (1 - \beta) g^{(k)} + \beta s^{(k-1)},$$

where $0 \leq \beta < 1$ is a (constant) filter parameter that controls how much memory the algorithm has. When $\beta = 0$ we obtain the subgradient method with Polyak's step size.

A more sophisticated method for updating $s^{(k)}$ was proposedby Camerini, Fratta, and Maffioli [CFM75]. Their algorithm has the form

$$s^{(k)} = g^{(k)} + \beta_k s^{(k-1)}, \tag{2.12}$$

where

$$\beta_k = max\{0, -\gamma_k (s^{(k-1)})^T g^{(k)} / \|s^{(k-1)}\|_2^2\}.$$

Here $\gamma_k \in [0, 2]$; they recommend using the constant value $\gamma_k = 1.5$.
They show that

$$\frac{(x^{(k)} - x^\star)^T s^{(k)}}{\|s^{(k)}\|_2^2} \geq \frac{(x^{(k)} - x^\star)^T g^{(k)}}{\|g^{(k)}\|_2^2},$$

i.e., the direction with modified update has a smaller angle towards the optimal set than the negative subgradient. (It follows that the convergence proofs for the subgradient algorithm work for this one as well.)

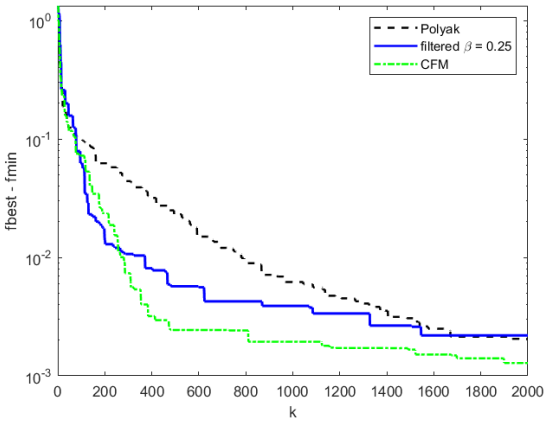


Figure 14: The value of $f_{best}^{(k)} - f^\star$ versus iteration number k, for the subgradient method with two types of Polyak's step sizes, the original update when $\beta = 0$ and a filtered update with $\beta = 0.25$. The plot also shows the subgradient method with CFM step size.

2.3 Proximal Algorithms

2.3.1 Proximal Algorithms

In this section, we discuss the proximal algorithm for minimizing a convex function $f : R^n \to R$ over a closed convex set X. It is given by

$$x^{k+1} = \underset{x \in X}{argmin}\{f(x) + \frac{1}{2c^2}\|x - x^k\|^2\},$$

where x^0 is an arbitrary starting point and c^k is a positive scalar parameter. The algorithm is somewhat different from the feasible direction methods of the preceding sections, because it does not require that f has a gradient. In particular, the entire cost function $f(x)$ is used in the proximal iteration rather than the linear approximation $\bigtriangledown f(x^k)'(x-x^k)$ that is used in the gradient projection method.This makes the algorithm applicable to nondifferentiable convex cost functions f.

The main motivation of the algorithm is regularization: the quadratic term $\|x-x^k\|^2$ makes the function that is minimized in iteration stictly convex with compact level sets, thereby guaranteeing, among others, that x^{k+1} is well-defined as the unique minimum. The creative application of the proximal algorithm can allow the elimination of constraints and nondifferentiablities.

Generally, starting from any nonoptimal point x^k, the cost function value is reduced at each iteration, by setting $x = x^k$, we have

$$f(x^{k+1}) + \frac{1}{2c^k}\|x^{k+1} - x^k\|^2 \leq f(x^k).$$

And the iterate distance to every optimal solution is also reduced. For the proximal algorithm, we have for all k,

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2c^k(f(x^{k+1}) - f(y)) - \|x^k - x^{k+1}\|^2, \forall y \in X.$$

Then, we can gain the convergence results. Let $\{x^k\}$ be a sequence generated by the proximal algorithm. Then, if $\sum_k^\infty = 0 c^k = \infty$, we have

$$f(x^k) \downarrow f^*,$$

and if X^* is nonempty, $\{x^k\}$ convergences to some point in X^*.
(Rate of Convergence) Assume that X^* is nonempty and that for some scalars $\beta > 0, \eta > 0$ and $\gamma \geq 1$, we have

$$f^* + \beta(d(x))^\gamma \leq f(x), \forall x \in X \ with \ d(x) \leq \eta,$$

where

$$d(x) = \inf_{x^* \in X^*} \|x - x^*\|.$$

Let also

$$\sum_{k=0}^{\infty} c^k = \infty,$$

So that the sequence $\{x^k\}$ generated by the proximal algorithm converges to some point in X^*. Then
(a) For all k sufficiently large, we have

$$d(x^{k+1}) + \beta c^k (d(x^{k+1}))^{\gamma - 1} \leq d(x^k), \ if \ \gamma > 1,$$

and

$$d(x^{k+1}) + \beta c^k \leq d(x^k), \ if \gamma = 1 \ and \ x^{k+1} \notin X^*.$$

(b) (Superlinear Convergence) Let $1 < \gamma < 2$ and $x^k \notin X^*$ for all k. Then if $\underset{k\geq 0}{inf}\, c^k > 0$,

$$\underset{k\to\infty}{limsup}\, \frac{d(x^{k+1})}{(d(x^k))^{\frac{1}{(\gamma-1)}}} < \infty$$

(c) (Linear Convergence) Let $\gamma = 2$ and $x^k \notin X^*$ for all k. Then if $\underset{k\to\infty}{lim}\, c^k = c'$ with $c' \in (0, \infty)$,

$$\underset{k\to\infty}{limsup}\, \frac{d(x^{k+1}}{d(x^k)} \leq \frac{1}{1+\beta c'},$$

while if $\underset{k\to\infty}{lim}\, c^k = \infty$,

$$\underset{k\to\infty}{lim}\, \frac{d(x^{k+1}}{d(x^k)} = 0.$$

(d) (Sublinear Convergence) Let $\gamma > 2$. Then

$$\underset{k\to\infty}{limsup}\, \frac{d(x^{k+1})}{d(x^k)^{2/\gamma}} < \infty.$$

The proposition shows that as the growth order γ increases, the rate of convergence becomes slower. An important threshold value is $\gamma = 2$; in this case the distance of the iterates to X^* decreases at least linearly if c^k remains bounded, and decreases even faster if $c^k \to \infty$.Generally, the convergence is accelerated if c^k is increased with k, rather than kept constant; that is illustrated most clearly when $\gamma = 2$. When $1 < \gamma < 2$, the convergence rate is superlinear. When $\gamma > 2$, the convergence rate is generally slower than when $\gamma = 2$.

2.3.2 Proximal Gradient Algorithm

The *proximal gradient algorithm*, which implies to the problem

$$\begin{aligned} &minimize f(x) = g(x) + h(x) \\ &subject\ to\ x \in X \end{aligned}$$

where $g : \mathrm{R^n} \to \mathrm{R}$ is a differentiable convex function, $h : \mathrm{R}^n \to \mathrm{R}$ is a convex function, not necessarily differentiable and X is a closed convex set. This algorithm combines ideas from the gradient projection method and the proximal method. It replaces g with a linear approximation in the proximal minimization.*i.e.*,

$$x^{k+1} \in arg\, \underset{x\in X}{min}\{\bigtriangledown g(x^k)'(x - x^k) + h(x) + \frac{1}{2\alpha^k}\|x - x^k\|^2\},$$

where $\alpha^k > 0$ is a parameter. Thus when g is linear function, we obtain the proximal algorithm for minimizing $g + h$. When h is identically zero, we obtain the gradient projection method. Note tha there is an alternative way to write the algorithm:

$$z^k = x^k - \alpha^k \bigtriangledown g(x^k), \qquad x^{k+1} \in arg\, \underset{x\in X}{min}\{h(x) + \frac{1}{2\alpha^k}\|x - z^k\|^2\},$$

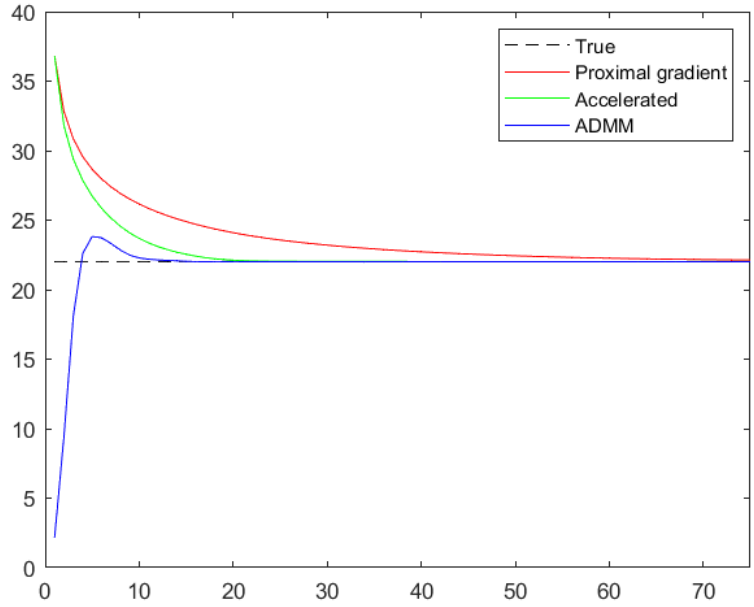


Figure 15: Different applications of Proximal Algorithm on Example Lasso, accelerated method with k/(k+3).

as can be verified by expanding the quadratic

$$\begin{aligned}\|x - z^k\|^2 =& \|x - x^k + \alpha^k \bigtriangledown g(x^k)\|^2. \\ =& \|x - x^k\|^2 + 2\alpha^k \bigtriangledown f(x^k)\|x - x^k\| + (\alpha^k \bigtriangledown f(x^k))^2\end{aligned}$$

Thus the method alternates gradient steps on g with proximal steps on h. The advantage that is method may have over the proximal algorithm is that the proximal step is executed with h rather than $h+g$, and this may be significant if h has simple/favorable structure(e.g., h is l_1 norm or a distance function to a simple constraint set), while g has unfavorable structure. Under relative mild assumptions, it can be shown that the method has a cost function descent property, provided the stepsize α is sufficiently small.
Recall motivation: minimize quadratic approximation to f around x, replace $\bigtriangledown^2 f(x)$ by $\frac{1}{t}I$,

$$x^+ = \underset{z}{argmin}\, f(x) + \bigtriangledown f(x)^T(z - x) + \frac{1}{2t}\|z - x\|_2^2$$

In our case f is not differentiable, but $f = g + h$, g differentiable. We can just make quadratic approximation to g and leave h alone. This is, update

$$\begin{aligned}x^+ &= \underset{z}{argmin}\ \bar{g}_t(z) + h(z) \\ &= \underset{z}{argmin}\ g(x) + \bigtriangledown g(x)^T(z - x) + \frac{1}{2t}\|z - x\|_2^2 + h(z) \\ &= \underset{z}{argmin}\ \frac{1}{2t}\|z - (x - t \bigtriangledown g(x))\|_2^2 + h(z)\end{aligned}$$

The first term stays close to gradient update for g, and the second term also makes h small.
Define proximal mapping:

$$prox_h(x) := \underset{z}{argmin}\{\frac{1}{2}\|z-x\|_2^2+h(z)\}$$

It is well-defined under very general conditions(including nonsmooth convex functions), can be evaluated efficiently for many widely used functions(in particular, regularizers),this abstraction is conceptually and mathematically simple, and covers many well-known optimization algorithms.
Examples:
If $h = 1_C$ is the "indicator" function on set C, then

$$prox_h(x) = arg\min_{z\in C}\|z-x\|_2 \ (Euclidean\ projection)$$

If $h(x) = \lambda\|x\|_1$, then

$$(prox_{\lambda h}(x))_i = \psi_{st}(x_i;\lambda) \ (soft-thresholding)$$

If $f(x) = ag(x) + b$ with $a > 0$, then

$$prox_f(x) = prox_{ag}(x)$$

If $f(x) = g(x) + a^Tx + b$ (affine addition), then

$$prox_f(x) = prox_g(x-a)$$

If $f(x) = g(x) + \frac{\rho}{2}\|x-a\|)_2^2$(quadratic addition), then

$$\begin{aligned}
prox_f(x) &= arg\min_x\{\frac{1}{2}\|z-x\|_2^2+g(z)+\frac{\rho}{2}\|z-\alpha\|_2^2\}\\
&= arg\min_x\{\frac{1+\rho}{2}\|z\|_2^2- < z, x+\rho a > +g(z)\}\\
&= arg\min_x\{\frac{1}{2}\|z\|_2^2-\frac{1}{1+\rho} < z, x+\rho a > +\frac{1}{1+\rho}g(z)\}\\
&= arg\min_x\{\frac{1}{2}\|z-(\frac{1}{1+\rho}x+\frac{\rho}{1+\rho}a)\|_2^2+\frac{1}{1+\rho}g(z)\}\\
&= prox_{\frac{1}{1+\rho}g}(\frac{1}{1+\rho}x+\frac{\rho}{1+\rho}a)
\end{aligned}$$

If $f(x) = g(ax + b)$(scaling and translation) with $a \neq 0$, then

$$prox_f(x) = \frac{1}{a}(prox_{a^2g}(ax+b)-b)$$

If $f(x) = g(Qx)$ with Q orthogonal (orthogonal mapping), then

$$prox_f(x) = Q^Tprox_g(Qx)$$

If $f(x) = g(Qx + b)$ with $QQ^T = \alpha^{-1}I$(orthogonal affine mapping), then

$$prox_f(x) = (I - \alpha Q^T Q)x + \alpha Q^T (prox_{\alpha^{-1}g}(Qx + b) - b)$$

If $f(x) = g(\|x\|_2)$ with $domain(g) = [0, \infty)$,then

$$\begin{aligned} prox_f(x) &= \min_z \{g(\|z\|_2) + \frac{1}{2}\|z\|_2^2 - z^T x + \frac{1}{2}\|x\|_2^2\} \\ &= \min_{\alpha \geq 0} \min_{z:\|z\|_2 = \alpha} \{g(\alpha) + \frac{1}{2}\alpha^2 - z^T x + \frac{1}{2}\|x\|_2^2\} \\ &= \min_{\alpha \geq 0} \{g(\alpha) + \frac{1}{2}\alpha^2 - \alpha\|x\|_2 + \frac{1}{2}\|x\|_2^2\} \\ &= \min_{\alpha \geq 0} \{g(\alpha) + \frac{1}{2}(\alpha - \|x\|_2)^2\} \end{aligned}$$

From the above calculation, we know the optimal point is

$$\begin{aligned} \alpha^* &= \underset{g}{prox}(\|x\|_2) \\ z^* &= \alpha^* \frac{x}{\|x\|_2} = prox_g(\|x\|_2)\frac{x}{\|x\|_2}. \end{aligned}$$

Proximal gradient descent: choose initialize $x^{(o)}$, repeat:

$$x^{(k)} = prox_{h,t_k}(x^{(k-1)} - t_k \bigtriangledown g(x^{(k-1)})), k = 1, 2, 3, ...$$

Algorithm 5 Proximal gradient algorithm

for t = 0,1,... **do**

$x^{(k+1)} = prox_{\eta_t h}(x^t - \eta_t \bigtriangledown f(x^t))$

The algorithm alternates between gradient updates on f and proximal minimization on h. And it is useful if $prox_h$ is inexpensive.
Convergence analysis: will be in terms of the number of iterations, and each iteration evaluates $prox_{h,t}$ once (this can be cheap or expensive, depending on h).
Now consider an example ISTA, given $y \in \mathrm{R}^n, X \in \mathrm{R}^{n \times p}$, recall the lasso criterion:

$$f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Proximal mapping is now

$$\begin{aligned} prox_t(\beta) &= \underset{z}{argmin}\, \frac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\ &= S_{\lambda t}(\beta) \end{aligned}$$

where $S_\lambda(\beta)$ is the soft-thresholding operator, by the property of soft-thresholding operator and recall $\bigtriangledown g(\beta) = -X^T(y - X\beta)$, hence proximal gradient update is:

$$\beta^+ = S_{\lambda t}(\beta + tX^T(y - X\beta))$$

Often called the iterative soft-thresholding algorithm(ISTA).
To make this update step look familiar, can rewrite it as

$$x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})$$

where G_t is the generalized gradient of f,

$$G_t(x) = \frac{x - prox_{h,t}(x - t \bigtriangledown g(x))}{t}$$

Backtracking for proximal gradient descent works similar as in gradient descent, but operates on g and not on f. Choose parameter $0 < \beta < 1$. At each iteration. start at $t = t_{init}$, and while

$$g(x - tG_t(x)) > g(x) - t \bigtriangledown g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2$$

shrink $t = \beta t$, for some $0 < \beta < 1$. Else perform proximal gradient update.
Convergence analysis, for criterion $f(x) = g(x) + h(x)$, we assume:
1. g is convex, differentiable, $dom(g) = \mathrm{R}^n$, and $\bigtriangledown_g$ is Lipschitz continuous with constant $L > 0$
2. h is convex, $prox_t(x) = \underset{z}{argmin}\{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated.

Theorem 2.6 *(convergence of proximal gradeint methods for convex problems) Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then*

$$F(x^{(k)}) - F^{opt} \leq \frac{L\|x^{(0)} - x^\star\|_2^2}{2t}$$

The method achieves better iteration complexity than subgradient method, and it is fast if prox can be efficiently implemented.

Theorem 2.7 *(convergence of proximal gradeint methods for strongly convex problems) Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then*

$$F(x^{(k)}) - F^{opt} \leq (1 - \frac{\mu}{L})^t \|x^0 - x^*\|_2^2$$

Linear convergence: attains ϵ accuracy within $O(log\frac{1}{\epsilon})$ iterations
Let $\tau_L(x) := prox_{(\frac{1}{L}h}(x - \frac{1}{L} \bigtriangledown f(x))$:

Algorithm 6 Backtracking line search for proximal gradient methods

Initialize $\eta = 1, 0 < \alpha < 1/2, 0 < \beta < 1$
while $f(\eta_{Lt}(x^t)) > f(x^t) - < \bigtriangledown f(x^t), x^t - \eta_{L_t}(x^t) > + \frac{L_t}{2}\|\eta_{Lt}(x^t) - x^t\|_2^2$
do
    $L_t \leftarrow \frac{1}{\beta} L_t$

Proximal gradient descent also called composite gradient descent, or generalized gradient descent. This refers to the several cses, when minimizing $f = g + h$:

$$h = 0 : gradient\ descent$$
$$h = I_C : projected\ gradient\ descent$$
$$g = 0 : proximal\ minimization\ algorithm$$

So far, the key requirement in proximal gradient method is the evaluable of prox, but if we can't evaluate prox exactly. In general, not clear what happens if we just minimize this approximately. But if we can precisely control the errors in approximating the prox operator, then we can recover the original convergence rates. In practice, if prox evaluation is done approximately, then it should be done to decently high accuracy.

Turns out we can accelerate proximal gradient descent in order to achieve the optimal $O(1/\sqrt{\epsilon})$ convergence rate.
Nesterov's idea:

$$x^{(t+1)} = y^t - \eta_t \bigtriangledown f(y^t)$$
$$y^{(t+1)} = x^{t+1} + \frac{t}{t+3}(x^{t+1} - x^t)$$

- alternates between gradient updates and proper extrapolation
- each iteration takes nearly the same cost as GD
- not a descent method (i.e. we may not have $f(x^{t+1} \leq f(x^t))$
- one of the most beautiful and mysterious results in optimization

Four ideas by Nesterov:
1. 1983: original acceleration idea for smooth functions
2. 1988: another acceleration idea for smooth functions
3. 2005: smoothing techniques for nonsmooth functions, coupled with original accerlation idea
4. 2007: acceleration idea for composite functions
We follow Beck and Teboulle(2008), an extension of Nesterov(1983) to composite functions
Accelerated proximal gradient method:
We exploit information from the history (*i.e.* past iterates) and add buffers (like momentum) to yield smoother trajectory.
As before, consider

$$\min_x g(x) = h(x)$$

where g is convex, differentiable, and h convex. Accelerated proximal gradient method: choose intial point $x^{(0)} = x(-1) \in \mathrm{R}^n$, repeat:

$$v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$$
$$x^{(k)} = prox_{t_k}(v - t_k \bigtriangledown g(v))$$

for k = 1, 2, 3,...
First step k = 1 is just usual proximal gradient update
After that, $v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$ carries some "momentum" from previous iterations
When $h = 0$ we get accerated gradient method.
Backtracking under with acceleration in different ways. Simple approach: fix $\beta < 1$, $t_0 = 1$. At iteration k, start with $t = t_{k-1}$, and while

$$g(x^+) > g(v) + \bigtriangledown g(v)^T(x^+ - v) + \frac{1}{2t}\|x^+ - v\|_2^2$$

shrink $t = \beta t$, and let $x^+ = prox_t(v - t\bigtriangledown g(v))$. Else keep x^+. Note that this strategy forces us to take decreasing step sizes. Convergence analysis, for criterion $f(x) = g(x) + h(x)$, we assume:
1. g is convex, differentiable, $dom(g) = \mathrm{R}^n$, and $\bigtriangledown_g$ is Lipschitz continuous with constant $L > 0$
2. h is convex, $prox_t(x) = \underset{z}{argmin}\{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated.

Theorem 2.8 *Accelerated proximal gradient method with fixed step size $t \leq \frac{1}{L}$ satisfies*

$$f(x^{(k)}) - f^\star \leq \frac{2\|x^{(0)} - x^\star\|_2^2}{t(k+1)^2}$$

and same result holds for backtracking, with t replaced by β/L.

- iteration complexity: $O(\frac{1}{\sqrt{\epsilon}})$
- much faster than gradient methods

Fast iterative shrinkage-thresholding algorithm(FISTA)

$$x^{t+1} = prox_{\eta_t h}(y^t - \eta_t \bigtriangledown f(y^t))$$
$$y^{t+1} = x^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(x^{t+1} - x^t)$$

where $y^0 = x^0, \theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$
In practice the speedup of using acceleration is diminished in the presence of **warm starts**. For example, suppose want to solve lasso problem for tuning parameters values

$$\lambda_1 > \lambda_2 > \cdots > \lambda_r$$

When solving for λ_1, initialize $x^{(0)} = 0$, record solution $\hat{x}(\lambda_1)$
When solving for λ_j, initialize $x^{(0)} = \hat{x}(\lambda_{j-1})$, the recorded solution for λ_{j-1}
Adaptive restart:
When a certain criterion is met, restart running FISTA with

$$x^0 \leftarrow x^t$$
$$y^0 \leftarrow x^t$$
$$\theta_0 = 1$$

- take the current iterate as a new starting point
- erase all memory of previous iterates and reset the momentum back to zero

Adaptive restart schemes: 1. Function scheme: restart when $f(x^t).f(x^{t-1})$ 2. Gradient scheme: restart when $< \bigtriangledown f(y^{t-1}), x^t - x^{t-1} > 0$

2.4 Conjugate Gradient Methods

The CG method is recommended only for large problems; otherwise, Gaussian elimination or other factorization algorithms such as the singular value decomposition are to be preferred, since they are less sensitive to rounding errors.

The linear conjugate gradient method is an iterative method for solving a linear system of equations

$$Ax = b,$$

where A is an $n \times n$ symmetric positive definite matrix. The problem can be stated equivalently as the following minimization problem:

$$min\ \phi(x) := \frac{1}{2}x^T Ax - b^T x,$$

they have the same unique solution. This equivalence will allow us to interpret the conjugate gradient method either as an algorithm for solving linear systems or as a technique for minimizing convex quadratic functions.

2.4.1 Linear Conjugate Gradient Methods

One of the remarkable properties of the conjugate gradient method is its ability to generate, in a very economical fashion, a set of vectors with a property known as conjugacy. A set of nonzero vectors $\{p_0, p_1, ..., p_l\}$ is said to be conjugate with respect to the symmetric positive define matrix A if

$$p_i^T Ap_j = 0,\ for\ all\ i \neq j.$$

It is easy to show that any set of vectors satisfying this property is also linear independent.

The importance of conjugacy lies in the fact that we can minimize $\phi(x)$ in n steps by successively minimizing it along the individual directions in a conjugate set. Given a starting point $x_0 \in \mathbf{R}^n$ and a set of conjugate directions $\{p_0, p_1, ..., p_{n-1}$, let us generate the sequence $\{x_k\}$ by setting

$$x_{k+1} = x_k + \alpha_k p_k, \tag{2.13}$$

where α_k is the one-dimensional minimizer of the quadratic function ϕ along $x_k + \alpha p_k$, given explicitly by

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T Ap_k}; \tag{2.14}$$

$r_k = Ax_k - b$.
Explanation: by the property of conjugacy, given x_0 and $\{p_i\}_0^{n-1}$, we have

$$x^* - x_0 = \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_{n-1} p_{n-1},$$

In order to calculate α_k, by premultiplying $p_k^T A$, get

$$p_k^T A(x^* - x_0) = p_k^T A(\alpha_k p_k) = \alpha_k (p_k^T A p_k)$$

and

$$(Ax_k - Ax^*) = Ax_k - b = r_k$$
$$x^* - x_0 = (x^* - x_k) + (x_k + x_0),$$

we have

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k};$$

Theorem 2.9 *For any $x_0 \in \mathbf{R}^n$ the sequence $\{x_k\}$ generated by the conjugate direction algorithm (2.16), α_k converges to the solution x^* of the linear system $Ax = b$ in at most n steps.*

Theorem 2.10 *(Expanding Subspace Minimization). Let $x_0 \in \mathbf{R}^n$ be any starting point and suppose that the sequence $\{x_k\}$is generated by the conjugate direction algorithm (2.16), (2.17). Then*

$$r_k^T p_i = 0, \ for \ i = 0, 1, ..., k-1,$$

and x_k is the minimizer of $\phi(x) = \frac{1}{2} x^T A x - b^T x$ over the set

$$\{x | x = x_0 + span\{p_0, p_1, ..., P_{k+1}\}\}.$$

There are many ways to choose the set of conjugate directions, for instance, the eigenvectors $v_1, v_2, \ldots, v_n$ of A are mutually orthogonal as well as conjugate with respect to A, so these could be used as the vectors $\{p_0, p_1, \ldots p_{n-1}\}$. For large-scale applications, however, computation of the complete set of eigenvectors requires an excessive amount of computation. An alternative approach is to modify the Gram-Schmidt orthogonalization process to produce a set pf conjugate direction rather than a set of orthogonal directions. (This modification is easy to produce, since the properties of conjugacy and orthogonality are closely related in spirit).
Explanation: $\{x_1, x_2 \ldots, x_n\}, x_i^T A x_j = 0, i \neq j$. $\{x_i\}_i^n$ conjugate direction. A is positive definite, then $A^{\frac{1}{2}}$ is also positive. We have

$$0 = x_i^T A x_j = x_i^T A^{\frac{1}{2}} A^{\frac{1}{2}} x_j = (A^{\frac{1}{2}} x_i)^T (A^{\frac{1}{2}} x_j)$$
$$= y_i^T y_j = 0$$

we can easily get $\{y_1, y_2, \ldots, y_n\}$ are orthogonal.
Convergence analysis:

Theorem 2.11 *If A has only r distinct eigenvalues, then the CG iteration will terminate at the solution at most r iterations.*

Algorithm 7 Precondition Conjugate Gradient Method

Given x_0; preconditioner M
Set $r_0 \leftarrow Ax_0 - b$;
Solve $My_0 = r_0$ for y_0;
Set $p_0 = -y_0, k \leftarrow 0$;
while $r_k \neq 0$

$$\alpha_k \leftarrow \frac{r_k^T y_k}{p_k^T A p_k};$$
$$x_{k+1} \leftarrow x_k + \alpha_k p_k;$$
$$r_{k+1} \leftarrow r_k + \alpha_k A p_k;$$
$$Solve\ My_{k+1} = r_{k+1};$$
$$\beta_{k+1} \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k};$$
$$p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k;$$
$$k \leftarrow k + 1;$$

Algorithm 8 Conjugate Gradient Method

Given x_0;
Set $r_0 \leftarrow Ax_0 - b, p_0 \leftarrow -r_0, k \leftarrow 0$;
while $r_k \neq 0$

$$\alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k};$$
$$x_{k+1} \leftarrow x_k + \alpha_k p_k;$$
$$r_{k+1} \leftarrow r_k + \alpha_k A p_k;$$
$$\beta_{k+1} \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k};$$
$$p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k;$$
$$k \leftarrow k + 1;$$

end(while)

Theorem 2.12 *If A has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we have that*

$$\|x_{k+1} - x^*\|_A^2 \leq (\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1})^2 \|x_0 - x^*\|_A^2.$$

`Exercise:`
Apply CG method with exact line search to solve

$$min \frac{1}{2} x^T A x + b^T x,$$

starting from $x_0 = (2,1)^T$. Here $A = [4,1;1,3]$ and $b = -(1,2)^T$.
After applying CG method with exact line search with argument A,b, we have:

$$\begin{aligned}
&x_0 = (2,1)^T, r_0 = (8,3)^T, \alpha_0 = 0.2205 \\
&x_1 = (0.2356, 0.3384), r_1 = (0.2810, -0.7492)^T, \\
&\beta_1 = 0.0088, p_1 = (-0.3511, 0.7229)^T, \alpha = 0.4122 \\
&x_2 = (0.0909, 0.6364)^T
\end{aligned}$$

2.4.2 Nonlinear Conjugate Gradient Methods

Algorithm 9 Fletcher-Reeves Method

Given x_0;
Evaluate $f_0 = f(x_0), \bigtriangledown f_0 = \bigtriangledown f(x_0)$;
Set $p_0 \leftarrow - \bigtriangledown f_0, k \leftarrow 0$;
while $\bigtriangledown f_k \neq 0$
Compute α_k and set $x_{k+1} = x_k + \alpha_k p_k$;
Evaluate $\bigtriangledown f_{k+1}$;

$$\begin{aligned}
\beta_{k+1}^{FR} &\leftarrow \frac{\bigtriangledown f_{k+1}^T \bigtriangledown f_{k+1}}{\bigtriangledown f_k^T \bigtriangledown f_k}; \\
p_{k+1} &\leftarrow - \bigtriangledown f_{k+1} + \beta_{k+1}^{FR} p_k; \\
k &\leftarrow k+1;
\end{aligned}$$

end(while)

2.5 Trust-Region Methods

2.5.1 Step

One of the key ingredients in a trust-region algorithm is the strategy for choosing the trust-region radius $\bigtriangledown_k$ at each iteration.

$$f(x_k + p) = f(x_k) + \bigtriangledown f(x_k) p + \frac{1}{2} p^T \bigtriangledown^2 f(x_k + tp) p,$$

By using an approximation B_k to the Hessian in the second-order term, m_k is defined as follows:

$$m_k(p) = f(x_k) + \bigtriangledown f(x_k)^T p + \frac{1}{2} p^T B_k p,$$

wherre B_k is some symmetric matrix. The difference between $m_k(p)$ and $f(x_k + p)$ is $O(\|p\|^2)$. When B_k is equal to the Hessian $\bigtriangledown^2 f(x_k)$, the approximation error in the model function m_k is $O(\|p\|^3)$, so the model is especially accurate when $\|p\|$ is small. To obtain each step, we seek a solution of the subproblem

$$\min_{p \in \mathrm{R}^n} m_k(p) = f(x_k) + \bigtriangledown f^T(x_k) + \frac{1}{2} p^T B_k p \qquad s.t. \|p\| \leq \Delta_k, \tag{2.15}$$

where $\bigtriangledown_k > 0$ is the trust-region radius. The solution p_k^* is the minimizer of m_k in the ball of radius Δ_k. Thus, the trust-region approach requires us to solve a sequence of subproblems in which the objective function and constraint are both quadratic. Given a step p_k we define the ratio

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}; \tag{2.16}$$

the numerator is called the actual reduction, and the denominator is the predicted reduction.Note that since the step p_k is obtained by minimizing the model m_k over a region that includes $p = 0$, the predicted reduction will always be nonnegative. Hence, if ρ is negative, the new objective value $f(x_k + p_k$ is greater than the current value $f(x_k)$, so the step must be rejected. On the other hand, if ρ_k is close to 1, there is good agreement between the model m_k and the function f over this step, so it is safe to expand the trust region for the next iteration. If ρ is positive but significantly smaller than 1, we do not alter the trust region, but if it is close to zero or negative, we shrink the trust region by reducing Δ_k at the next iteration.

Algorithm 10 Trust Region

Given $\hat{\Delta} > 0, \Delta_0 \in (0, \hat{\Delta}$, and $\eta \in [0, \frac{1}{4})$:
for k = 0,1,2,...
    obtain p_k by (approximately) solving (14);
    Evaluate $\rho \; from (15)$
    if $\rho_k < \frac{1}{4}$
        $\Delta_{k+1} = \frac{1}{4}\Delta_k$
    else
        if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$
            $\Delta_{k+1} = min(2\Delta_k, \hat{\Delta})$
        else
            $\Delta_{k+1} = \Delta_k$;
    if $\rho_k > \eta$
        $x_{k+1} = x_k + p_k$
    else
        $x_{k+1} = x_k$;
end(for).

Theorem 2.13 *The vector p^* is a global solution of the trust-region problem*

$$\min_{p\in \mathrm{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \qquad s.t. \|p\| \leq \Delta,$$

if and only if p^* is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied:

$$(B + \lambda I)p^* = -g,$$
$$\lambda(\Delta - \|p^*\|) = 0,$$
$$(B + \lambda I) \text{ is positive semidefinite.}$$

When the solution lies strictly inside the trust region, we must have$\lambda = 0$ and so $Bp^* = -g$ with B positive semidefinite. In the case $\|p^*\| = \Delta$, λ is allowed to take a positive value. We have

$$\lambda p^* = -Bp^* - g = - \bigtriangledown m(p^*).$$

Thus, when $\lambda > 0$, the solution p^* is collinear with the negative gradient of m and normal to its contours.

2.5.2 The Cauchy Point

$$\min_{p\in \mathrm{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \qquad s.t. \|p\| \leq \Delta,$$

Find the vector p_k^s that solves a linear version of $\min_{p\in \mathrm{R}^n} m(p)$, that is

$$p_k^s = arg \min_{p\in \mathrm{R}^n} f_k + g_k^T p \qquad s.t. \|p\| \leq \Delta_k;$$

Calculate the scalar $\tau_k > 0$ that minimizes $m_k(\tau p_k^s)$ subject to satisfying the trust-region bound, that is,

$$\tau_k = arg \min_{\tau \geq 0} m_k(\tau p_k^s) \qquad s.t. \|\tau p_k^s\| \leq \Delta_k;$$

We can easily get (linear version and satisfy $\|p\| \leq \Delta_k$)

$$p_k^s = -\frac{\Delta_k}{\|g_k\|} g_k,$$

Now we consider to obtain τ explicitly

$$f(x) = \begin{cases} 1 & if\ g_k^T B_k g_k \leq 0 \\ min(\|g_k\|^3/(\Delta_k g_k^T B_k g_k), 1) & if\ g_k^T B_k g_k > 0 \end{cases}$$

Explanation: if $g_k^T B_k g_k \leq 0$, the function $m_k(\tau p_k^s)$ decreases monotonically with τ whenever $g_k \neq 0$, so it simply the largest value that satisfies the trust-region bound.For the case $g_k^T B_k g_k \geq 0$, $m_k(\tau p_k^s)$ is a convex quadratic in τ, so τ_k is either the unconstrained minimizer of this quadratic,$\|g_k\|^3/(\Delta_k g_k^T B_k g_k)$, or the boundary value 1.

2.5.3 The Dogleg Method

$$\min_{p\in \mathrm{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \qquad s.t. \|p\| \leq \Delta,$$

When B is positive definite,we have already noted that the unconstrained minimizer of m is $p^B = -B^{-1}g$. When this point is feasible, it is obviously a solution, so we have

$$p^*(\Delta) = p^B, \qquad when\ \Delta \geq \|p^B\|.$$

When Δ is small relative to p^B, the restriction $\|p\| \leq \Delta$ ensures that the quadratic term in m has little effect on the solution. For such Δ, we can get an approximation to p(Δ) by simply omitting the quadratic term and writing

$$p^*(\Delta) \approx -\Delta \frac{g}{\|g\|}, \qquad when\ \Delta\ is\ small.$$

The dogleg method finds an approximate solution by replacing the curved trajectory for $p^*(\Delta)$ with a path consisting of two line segments. The first line segment runs from the origin to the minimize of m along the steepest descent direction, which is

$$p^U = -\frac{g^T g}{g^T B g} g.$$

while the second line segment runs from p^U to p^B. Formally, we denote this trajectory by $\widetilde{p}(\tau)$ for $\tau \in [0, 2]$, where

$$\widetilde{p}(\tau) = \begin{cases} \tau p^U, & 0 \leq \tau \leq 1. \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2. \end{cases}$$

Lemma 2.1 *Let B be positive definite, then 1) $\|\widetilde{p}(\tau)\|$ is an increasing function of τ, 2) $m(\widetilde{p}(\tau))$ is a decreasing function of τ.*

2.5.4 supplement

The pure Newton step is obtained by minimizing over d the second order approximation of f around x^k, given by

$$f^k(d) = f(x^k) + \bigtriangledown f(x^k)' d + \frac{1}{2} d' \bigtriangledown f(x^k) d.$$

We know that $f^k(d)$ is a good approximation of $f(x^k + d)$ when d is in a small neighborhood of zero, but the difficulty is that with unconstrained minimization of $f^k(d)$ one may obtain a step that lies outside the neighborhood. It therefore makes sense to consider a restricted Newton step d^k, which is obtained by minimizing $f^k(d)$ over a suitably small neighborhood of zero, called the trust region:

$$d^k \in arg \min_{\|d\| \leq \gamma^k} f^k(d), \tag{2.17}$$

where γ^k is some positive scalar. An approximate solution of the constrained minimization problem (16) can be obtained quickly using the fact that it has only one constraint. An important observation here is that even if $\bigtriangledown^2 f(x^k)$ is not positive definite or, more generally, even if the pure Newton direction is not a descent direction, the restricted Newton step d^k improves the cost, provided $\bigtriangledown f(x^k) \neq 0$ and γ^k is sufficiently small. The reason is that, in view of (14), $f^k(d^k)$ is smaller than $f(x^k)$[which is equal to $f^k(0)$], and $f(x^k + d^k)$ is very close to its second order expansion $f^k(d^k)$ when $\|d^k\|$ is small.

More specifically, we have for all d with $\|d\| \leq \gamma^k$

$$f(x^k + d) = f^k(d) + o((\gamma^k)^2),$$

so that

$$\begin{aligned} f(x^k + d^k) &= f^k(d^k)) + o((\gamma^k)^2) \\ &= f(x^k) + \min_{\|d\| \leq \gamma^k} \{\bigtriangledown f(x^k)'d + \frac{1}{2} d' \bigtriangledown^2 f(x^k) d\} + o((r^k)^2). \end{aligned}$$

Therefore, denoting

$$\widetilde{d}^k = -\frac{\bigtriangledown f(x^k)}{\| \bigtriangledown f(x^k)\|} \gamma^k,$$

we have

$$\begin{aligned} f(x^k + d^k) &\leq f(x^k) + \bigtriangledown f(x^k)' \widetilde{d}^k + \frac{1}{2} \widetilde{d}^{k\prime} \bigtriangledown^2 f(x^k) \widetilde{d}^k + o((\gamma^k)^2) \\ &= f(x^k) - \gamma^k \| \bigtriangledown f(x^k)\| + \frac{(\gamma^k)^2}{2\| \bigtriangledown f(x^k)\|^2} \bigtriangledown f(x^k)' \bigtriangledown^2 f(x^k) \bigtriangledown f(x^k) + o((\gamma^k)^2). \end{aligned}$$

For γ^k sufficiently small, the negative term $-\gamma^k \| \bigtriangledown f(x^k)\|$ dominates the last two terms on the right-hand side above, showing that

$$f(x^k + d^k) < f(x^k).$$

It can be seen in fact from the precedingg relations that a cost improvement is possible even when $\bigtriangledown f(x^k) = 0$, provided γ^k is sufficiently small and f has a direction of negative curvature at x^k, *i.e.*, $\bigtriangledown^2 f(x^k)$is not positive semidefinite. Thus the preceding procedure will fail to improve the cost only if $\bigtriangledown f(x^k) = 0$ and $\bigtriangledown^2 f(x^k)$ is positive semidefinite, *i.e.*, x^k satisfies the first and second order necessary conditions. In particular, one can make progress even if x^k is a stationary point that is not a local minimum.

We are thus motivated to consider a method of the form

$$x^{k+1} = x^k + d^k,$$

where d^k is the restricted Newton step corresponding to a suitably chosen scalar γ^k as (14). Here, for a given x^k, γ^k should be small enough so that there is cost improvement; one possibility is to start from an initial trial γ^k and successively reduce γ^k by a certain factor as many times as necessary until a cost reduction occurs [$f(x^{k+1} < f(x^k)$]. The choice of the initial trial value for γ^k os crucial here; if it is chosen too large, a large

number of reductions may be necessary before a cost improvement occurs; if it is chosen too small the convergence rate may be poor. In particular, to maintain the superlinear convergence rate of Newton's method, as x^k approaches a nonsingular local minimum, one should select the initial trial value of γ^k sufficiently large so that the restricted Newton step and the pure Newton step coincide.

A reasonable way to adjust the initial trial value for γ^k is to increase this value when the method appears to be progressing well and to decrease this value otherwise. One can measure progress by using the ratio of the actual over predicted cost improvement [based on the approximation $f^k(d)$]

$$\gamma^k = \frac{f(x^k) - f(x^{k+1})}{f(x^k) - f^k(d^k)}$$

In particular, it makes sense to increase the initial trial value for γ ($\gamma^{k+1} > \gamma^k$) if this ratio is close to or above unity, and decrease γ otherwise. The following algorithm is a typical example of such a method. Given x^k and an initial trial value γ^k, it determines x^{k+1} and an initial trial value γ^{k+1} by using two thresholds σ_1, σ_2 with $0 < \sigma_1 \leq \sigma_2 \leq 1$ with two factors β_1, β_2 with $0 < \beta_1 < 1 < \beta_2$ (typical values are $\sigma_1 = 0.2$, $\sigma_2 = 0.8$, $\beta_1 = 0.25$, $\beta_2 = 2$).
Step 1: Find

$$d^k \in arg \min_{\|d\| \leq \gamma^k} f^k(d),$$

If $f^k(d^k) = f(x^k)$ stop (x^k satisfies the first and second order necessary conditions for a local minimum); else go to step 2.
Step 2: If $f(x^k + d^k) < f(x^k)$ set

$$\gamma^k = \frac{f(x^k - f(x^{k+1})}{f(x^k) - f^k(d^k)}$$

and go to Step 3; else set $\gamma^k := \beta_1 \|d^k\|$ and go to Step 1.
Step 3: Set

$$\gamma^{k+1} = \begin{cases} \beta_1 \|d^k\| & if\ \gamma^k\ < \sigma_1, \\ \beta_2 \gamma^k & if \sigma_2 \leq r^k\ and\ \|d^k\| = \gamma^k, \\ \gamma^k & otherwise. \end{cases}$$

Go to the next iteration.

Assuming that f is twice continuously differentiable, it is possible to show that the above algorithm is convergent in the sense that if x^k is a bounded sequence, there exists a limit point of x^k that satisfies the first and second order necessary conditions for optimality. Furthermore, if $\{x^k\}$ converges to a nonsingular local minimum x^*, then asymptotically, the method is identical to the pure form of Newton's method, thereby attaining a superlinear convergence.
Exercise 4.2 page 98 of Numerical optimization
Write a program to implements the dogleg method. Choose B_k to be the exact Hessian. Apply it to solve Rosenbrock's function $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$. Experiment with the update rule for the trust region by changing the constraints, or by designing your own

rules.

$$\bigtriangledown f(x) = \begin{bmatrix} -400x_1x_2 + 400x_1^3 + 2x_1 - 2 \\ -200x_1^2 + 200x_2 \end{bmatrix}$$
$$\bigtriangledown^2 f(x) = \begin{bmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

$x^* = (1,1)^T$, $\bigtriangledown^2 f(x^*)$ is positive definite. Choose $B_k = \bigtriangledown^2 f(x_k), \hat{\Delta} = 1, \Delta_0 = 0.5, \eta = 0.125$ and $x_0 = (0,0)^T$, code in Appendix B.

2.6 Quasi-Newton Methods

The most popular quasi-Newton algorithm is the BFGS method, named for its discoverers Broyden, Fletcher, Glodfarb, and Shanno. The premier iteration is quite similar to the line search Newton method; the key difference is that the appoximate Hessian B_k is used in place of the true Hessian. Instead of computing B_k afresh at every iteration, Davidon proposed to update it in a simple manner to account for the curvature measured during the most recent step. Suppose that we have generated a new iterate x_{K+1} and wish to construct a new quadratic model, of the form

$$m_{k+1}(p) = f_{k+1} + \bigtriangledown f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p.$$

One reasonable requirements is that the gradient of m_{k+1} should match the gradient of the objective function f at the lastest two iterates x_k and x_{k+1}. Since $\bigtriangledown m_{k+1}(0)$ is precisely $\bigtriangledown f_{k+1}$, the second of these conditions is satisfied automatically. The first condition can be written mathematically as

$$\bigtriangledown m_{k+1}(-\alpha_k p_k) = \bigtriangledown f_{k+1} - \bigtriangledown f_k.$$

By rearranging, we obtain

$$B_{k+1}\alpha p_k = \bigtriangledown f_{k+1} - \bigtriangledown f_k. \tag{2.18}$$

To simplify the notation it is useful to define the vectors

$$s_k = x_{k+1} - x_k = \alpha_k p_k,\ y_k = \bigtriangledown f_{k+1} - \bigtriangledown f_k,$$

so that (2.18) becomes

$$B_{k+1} s_k = y_k. \tag{2.19}$$

We refer to this formula as the secant equation.

Given the displacement s_k and the change of gradients y_k, the secant equation requires that the symmetric positive definite matrix B_{k+1} maps s_k into y_k. This will be possible only if s_k and y_k satisfy the curvature condition

$$s_k^T y_k > 0, \tag{2.20}$$

as is easily seen by premultiplying (2.19) by s_k^T. When f is strongly convex, the inequality (2.20) will be satisfied for any two points x_k and x_{k+1}.

Exercise 6.1 page 162 Numerical Optimization
Proof: A function $f(x)$ is strongly convex if all eigenvalues are positive and bounded away from zero. This implies that there exists $\sigma > 0$ such that

$$p^T \bigtriangledown^2 f(x)p \geq \sigma\|p\|^2, \ for\ any\ p.$$

By Taylor's theorem, we have $x_{k+1} = x_k + \alpha_k p_k$, then

$$\bigtriangledown f(x_{k+1}) = \bigtriangledown f(x_k) + \int_0^1 [\bigtriangledown^2 f(x_k + z\alpha_k p_k)\alpha_k p_k]dz.$$

Thus,

$$\begin{aligned} s_k^T y_k = \alpha_k p_k^T &= \alpha p_k^T[\bigtriangledown f(x_{k+1}) - \bigtriangledown f(x_k)] \\ &= \alpha_k^2 \int_0^1 [p_k^T \bigtriangledown^2 f(x_k + z\alpha_k p_k)p_k]dz \\ &\geq \sigma\|p_k\|^2\alpha^2 > 0. \end{aligned}$$

However this condition will not always hold for nonconvex functions, and in this case we need to enforce (21) explicitly, by imposing restrictions on the line search procedure that chooses the step length α. In fact, the condition (21) is guaranteed to hold if we impose the Wolfe or strong Wolfe conditions on the line search. We note form (19) and Wolfe condition that $\bigtriangledown f_{k+1}^T s_k \geq c_2 \bigtriangledown f_k^T s_k$, and therefore

$$y_k^T s_k \geq (c_2 - 1)\alpha_k \bigtriangledown f_k^T p_k.$$

Since $c_2 < 1$ and since p_k is a descent direction, the term on the right is positive, and the curvature condition (21) holds.

Algorithm 11 Framework of Quasi-Newton Method

Given $x_0, \epsilon > 0$, B_0 positive definite $(H_0 = B_0^{-1}), p_0 = -B_0^{-1} \bigtriangledown f_0, k = 0$
`While` $\| \bigtriangledown f_k\| > \epsilon$
    Evaluating $\alpha_k : argmin f(x_k + \alpha_k p_k)$// Wolfe conditions
    Set $x_{k+1} = x_k + \alpha_k p_k$
    Evaluating $B_{k+1} \approx \bigtriangledown^2 f_{k+1}$
    $p_{k+1} = (-B_{k+1}^{-1}) \bigtriangledown f_{k+1}$ $\qquad$ $(p_{k+1} = -H_{k+1} \bigtriangledown f_{k+1})$
    $k = k + 1$
`end(while)`

Rank-one modification on B: $B_{k+1} = B_k + \alpha_k U_k U_k^T$,
Rank-two modification on B: $B_{k+1} = B_k + \alpha_k U_k U_k^T + b_k V_k V_k^T$;
DFP Method:

$$H_{k+1} = H_k + \alpha_k U_k U_k^T + b_k V_k V_k^T // rank - two\ modification\ on\ H$$
$$S_k = H_{k+1} y_k = H_k y_k + \alpha_k U_k (U_k^T y_k) + \beta_k V_k (V_k^T y_k)$$
$$s_k = U_k \implies 1 = \alpha_k (U_k^T y_k) \implies \alpha_k = \frac{1}{U_k^T y_k} = \frac{1}{s_k^T y_k}$$
$$H_k y_k = V_k \implies 1 + b_k V_k^T y_k = 0 \implies b_k = \frac{-1}{V_k^T y_k} = -\frac{1}{y_k^T H_k y_k}$$
$$H_{k+1} = H_k + \frac{S_k S_k^T}{s_k^T y_k} - \frac{H_k V_k V_k^T H_k^T}{y_k^T H_k y_k}$$

BFGS Method:

$$B_{k+1} = B_k + \alpha_k U_k U_k^T + b_k V_k V_k^T // rank - two\ modification\ on B$$
$$y_k = B_{k+1} S_k = B_k S_k + \alpha_k U_k (U_k^T S_k) + b_k V_k (V_k^T S_k)$$
$$y_k = U_k \implies 1 = \alpha_k (U_k^T S_k) \implies \alpha_k = \frac{1}{U_k^T S_k} = \frac{1}{y_k^T S_k}$$
$$B_k S_k = b_k V_k (V_k^T S_k) \implies 1 + b_k (V_k^T S_k) = 0 \implies b_k = -\frac{1}{V_k^T S_k} = \frac{1}{S_k^T B_k S_k}$$
$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T S_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}$$

SR1 Method:

$$B_{k+1} = B_k + \alpha_k U_k U_k^T // rank - one\ modification\ on\ B$$
$$y_k = B_{k+1} s_k = B_k s_k + \alpha_k U_k (U_k^T s_k)$$
$$y_k - B_k s_k = \alpha_k U_k (U_k^T s_k)$$
$$U_k = y_k - B_k s_k \implies \alpha (U_k^T s_k) = 1 \implies \alpha_k = \frac{1}{U_k^T s_k} = \frac{1}{(y_k - B_K s_k)^T s_k}$$
$$B_{k+1} = B_k + \frac{U_k (U_k^T s_k)}{(u_k - B_k s_k)^T s_k} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

3 Duality

3.1 The Lagrange dual function

3.1.1 The Lagrangian

We consider an optimization problem in the standard form:

$$\begin{aligned} &minimize\ f_0(x) \\ &subject\ to\ f_i(x) \leq 0, i = 1, ..., m \\ &\qquad\qquad h_i(x) = 0, i = 1, ..., p, \end{aligned} \tag{3.1}$$

with variable $x \in R^n$. We assume its domain $D = \{\cap_{i=0}^m dom f_i\} \cap \{\cap_{i=1}^p dom h_i\}$ is nonempty, and denote the optimal value of this problem by $p^\star$. We do not assume that the problem is convex.

The basic idea in Lagrangian duality is to take the constraints into account by augmenting the objective function with a weighted sum of the constraint functions. We define the Lagrangian L: $R^m \times R^m \times R^p \to R$ associated with the problem as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

with $domL = D \times R^m \times R^p$. We refer to λ_i as the Lagrange multiplier associated with the ith inequality constraints $f_i(x) \leq 0$; similarly we refer to ν_i as the Lagrange multiplier associated with the ith equality constraint $h_i(x) = 0$. The vectors λ and ν are called the dual variables or Lagrange multiplier vectors associated with the problem.

3.1.2 The Lagrange dual function

We define the Lagrange dual function $g : R^m \times R^p \to R$ as the minimum value of the Lagrangian over x: for $\lambda \in R^m, \nu \in R^p$,

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)).$$

when the Lagrangian is unbounded below in x, the dual function takes on the value $-\infty$. Since the dual function is the pointwise infimum of a family of affine functions of $L(\lambda, \nu)$, it is concave, even when the optimization problem is not convex.

3.1.3 Lower bounds on optimal value

The dual function yields lower bounds on the optimal value $p^\star$ of the problem (3.1): for any $\lambda \succ 0$ and any ν we have

$$g(\lambda, \nu) \leq p^\star.$$

suppose $\widetilde{x}$ is a feasible point for the problem (3.1). *i.e.*, $f_i(\widetilde{x}) \leq 0$ and $h_i(\widetilde{x}) = 0$, and $\lambda \succeq 0$. Then we have

$$\sum_{i=1}^{m} \lambda_i f_i(\widetilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\widetilde{x}) \leq 0,$$

therefore,

$$L(\widetilde{x}, \lambda, \nu) = f_0(\widetilde{x}) + \sum_{i=1}^{m} + \sum_{i=1}^{p} \nu_i h_i(\widetilde{x}) \leq f_0(\widetilde{x}).$$

Hence,

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leq f_0(\widetilde{x}). \tag{3.2}$$

Since $g(\lambda, \nu) \leq f_0(\widetilde{x})$ holds for every feasible point $\widetilde{x}$, Q.E.D.

3.1.4 Linear approximation interpretation

The Lagrangian and lower bound property can be given a simple interpretation, based on a linear approximation of the indicator functions of the sets $\{0\}$ and $-R_+$.

We first rewrite the original problem as an unconstrained problem,

$$minimize\ f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x)),$$

We can think of I_- as a "brick wall" or "infinitely hard" displeasure function; our displeasure rises from zero to infinite as $f_i(x)$ transitions from nonpositive to positive.

Now suppose in the formulation we replace the function $I_-(u)$ with the linear function $\lambda_i u$, where $\lambda_i \geq 0$, and the function $I_0(u)$ with $\nu_i u$. The objective becomes the Lagrangian function $L(x, \lambda, \nu)$, and the dual function value $g(\lambda, \nu)$ is the optimal value of the problem

$$minimize\ L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

In this formulation, we use a linear or "soft" displeasure function in place of I_- and I_0. For an inequality constraint, our displeasure is zero when $f_i(x) = 0$, and is positive when $f_i(x) > 0$(assuming $\lambda_i > 0$); our displeasure gros as the constraint becomes "more violated". Unlike the original formulation, in which any nonposive value of $f_i(x)$ is acceptable, in the soft formulation we actually derive pleasure from constraints that have margin, *i.e.*, from $f_i(x) < 0$.

Clearly the approximation of the indicator function $I_-(u)$ with a linear function $\lambda_i u$ is rather poor. But the linear function is at least an underestimator of the indicator function. Since $\lambda_i u \leq I_-(u)$ and $\nu_i u \leq I_0(u)$ for all u, we see immediately that the dual function yields a lower bound on the optimal value of the original problem.

3.1.5 The Lagrange dual function and conjugate functions

Recall that the conjugate f^* of a function $f : R^n \to R$ is given by

$$f^*(y) = \sup_{x \in dom f} (y^T x - f(x)).$$

The conjugate function and Lagrange dual function are closely related. To see one simple connection, consider the problem

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ x = 0 \end{aligned}$$

This problem has Lagrangian $L(x, \nu) = f(x) + v^T x$, and the dual function

$$g(\nu) = \inf_x (f(x) + \nu^T x) = -\sup_x ((-\nu)^T x - f(x)) = -f^*(-\nu).$$

More generally and more usefully, consider an optimization problem with linear inequality and equality constraints,

$$\begin{aligned} &minimize\ f_0(x) \\ &subject\ to\ cx \succ b \\ &Cx = d. \end{aligned}$$

Using the conjugate of f_0 we can write the dual function for the problem as

$$\begin{aligned} g(\lambda,\nu) &= \underset{x}{inf}(f_0(x) + \lambda^T(Ax-b) + \nu^T(Cx-d)) \\ &= -b^T\lambda - d^T\nu + \underset{x}{inf}(f_0(x) + (A^T\lambda + C^T\nu)^T x) \\ &= -b^T\lambda - d^T\nu - f_0^*(-A^T\lambda - C^T\nu). \end{aligned}$$

The domain of g follows from the domain of f_0^*:

$$domg = \{(\lambda,\nu)| - A^T\lambda - C^T\nu \in domf_0^*\}.$$

3.1.6 Exercises: Basic definitions

3.1 Consider the optimization problem

$$\begin{aligned} &minimize\ x^2 + 1 \\ &subject\ to\ (x-2)(x-4) \leq 0, \end{aligned}$$

with variable $x \in R$.
(a) Analysis of primal problem. Give the feasible set, the optimal value, and the optimal solution.
Since the constraint $(x-2)(x-4) \leq 0$, we can gain the range of x is $[2,4]$. Under the range, we try to minimize the objective function x^2+1. It's obviously that the optimal value is 5, and the optimal solution is $x = 2$.
(b) Lagrangian and dual function. Plot the objective x^2+1 versus x. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x,\lambda)$ versus x for a few positive values of λ. verify the lower bound property ($p^\star \geq inf_x L(x,\lambda) for\ \lambda \geq 0$). Derive and sketch the Lagrange dual function g.

$$L(x,\lambda) = x^2 + 1 + \lambda(x-2)(x-4).$$

We can verify the lower bound property ($p^\star \geq inf_x L(x,\lambda) for \lambda \geq 0$). Since

$$g(x,\lambda) = \underset{x\in[2,4]}{inf} L(x,\lambda) = \underset{x\in[2,4]}{inf}(x^2+1+\lambda(x-2)(x-4)) = \underset{x\in[2,4]}{inf}((1+\lambda)x^2 - 6\lambda x + 8\lambda + 1)$$

By quadratic formulas, we have

$$x = \frac{3\lambda}{1+\lambda}$$

Since $\lambda > 0$, if $\lambda < 2$, we have $\frac{3\lambda}{1+\lambda} < 2$, then we select the $x = 2$. While $\lambda > 2$, we have $\frac{3\lambda}{1+\lambda} \in [2,4]$, then we select $x = \frac{3\lambda}{1+\lambda}$.

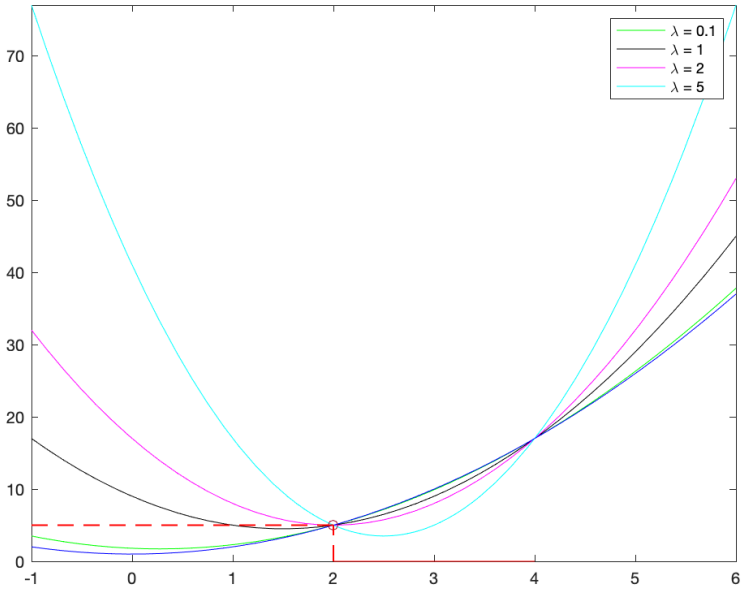


Figure 16

(c) Lagrange dual problem. State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^\star$. Does the strong duality hold?

$$g(\lambda) = \begin{cases} 5, & 0 < \lambda < 2 \\ \dfrac{-\lambda^2 + 9\lambda + 1}{1+\lambda}, & \lambda \geq 2 \end{cases}$$

The dual optimal value and dual optimal solution $\lambda^\star$ is 5 and 2. Thus strong duality holds.
(d) Sensitivity analysis. Let $p^\star(u)$ denote the optimal value of the problem

$$\begin{aligned} &\textit{minimize } x^2 + 1 \\ &\textit{subject to } (x-2)(x-4) \leq u, \end{aligned}$$

as a function of the parameter u. Plot $p^*(u)$. Verify that $dp^\star(0)/du = -\lambda^\star$.
3.2 Weak duality for unbouded and infeasible problems. The weak duality inequality, $d^\star < p^\star$, clearly holds when $d^\star = -\infty$ or $p^\star = \infty$. Show that it holds in the other two cases as well: If $p^\star = -\infty$, then we must have $d^\star = -\infty$, and also, if $d^* = \infty$, then we must have $p^\star = \infty$.
Prove: when $p^\star = -\infty$, then there exists a feasible point with arbitrarily small value of $f_0(x)$. Then we consider the Lagrangian

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

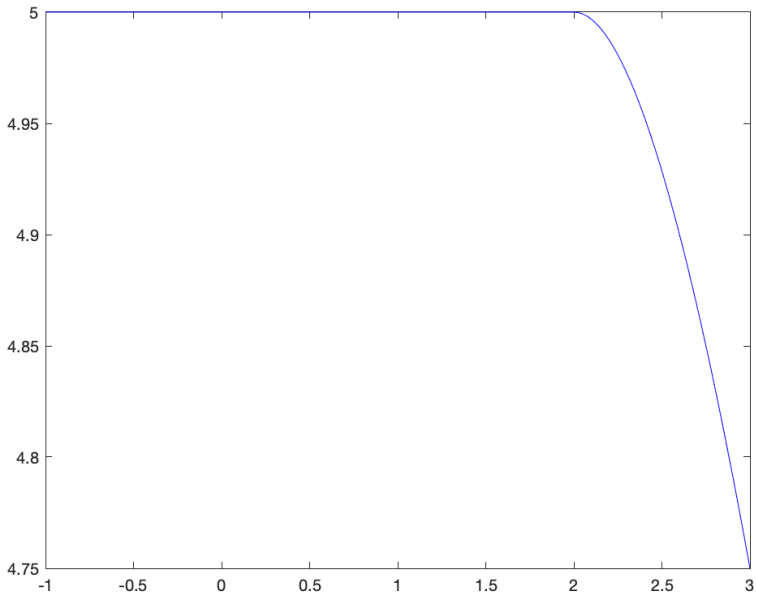


Figure 17

is unbounded below for all $\lambda \succeq 0$, *i.e.*, $g(\lambda) = -\infty$ for $\lambda \succeq 0$. Therefore the dual problem is infeasible, so we have $d^\star = -\infty$.

when $d^\star = \infty$. This is only possible when the primal problem is infeasible. By contradiction, If it were feasible, with $f_i(\widetilde{x}) \leq 0$ for $i = 1, ..., m$, then for all $\lambda \succeq 0$,

$$g(\lambda) = inf(f_0(x) + \sum_i \lambda_i f_i(x)) \leq f_0(\widetilde{x}) + \sum_i \lambda_i f_i(\widetilde{x}),$$

so the dual problem is bounded above. Thus, we gain $p^\star = \infty$.
3.3 Problems with one equality constraint. Express the dual problem of

$$\begin{aligned} &minimize\ c^T x \\ &subject\ to\ f(x) \leq 0, \end{aligned}$$

with $c \neq 0$, in terms of the conjugate f^*. with $c \neq 0$, in terms of the conjugate f^*. Explain why the problem you give is convex. We do not assume f is convex.
Solution: The dual function

$$\begin{aligned} g(\lambda) &= inf(c^T x + \lambda f(x)) \\ &= \lambda inf((c/\lambda)^T x + f(x)) \\ &= -\lambda f_1^*(-c/\lambda), \end{aligned}$$

Then the dual problem is

$$\begin{aligned} &minimize\ -\lambda f_1^\star(-c/\lambda) \\ &subject\ to\ \lambda \geq 0. \end{aligned}$$

The last part is how to show that it is convex.

3.2 The Lagrange dual problem

For each pair (λ, ν) with $\lambda \succeq 0$, the Lagrange dual function gives us a lower bound on the optimal value $p^\star$ of the optimization problem. Thus we have a lower bound that depends on some parameters λ, ν. A natural question is: What is the best lower bound that can be obtained from the Lagrange dual function?

This leads to the optimization problem

$$\begin{aligned} &maximize\ g(\lambda, \nu) \\ &subject\ to\ \lambda \succeq 0. \end{aligned}$$

This problem is called the Lagrange dual problem associated with the problem(27). In this context the original problem is sometimes called the primal problem. This term dual feasible, to describe a pair (λ, ν) with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$, now makes sense. It means, as the name implies, that (λ, ν) is feasible for the dual problem. We refer to (λ^*, ν^*) as dual optimal or optimal Lagrange multipliers if they are optimal for the problem.

The Lagrange dual problem is a convex optimization problem, since the objective to be maximized is concave and constraint is convex. This is the case whether or not the primal problem is convex.

3.2.1 Making dual constraints explicit

The examples above show that it is not uncommon for the domain of the dual function,

$$domg = \{(\lambda, \nu) | g(\lambda, \nu) > -\infty\},$$

to have dimension smaller than $m + p$. In many cases we can identify the affine hull of $domg$, and describe it as a set of linear equality constraints. Roughly speaking, this means we can identify the equality constraints that are "hidden" or "implicit" in the objective g of the dual problem. In this case we can form an equivalent problem, in which these equality constraints are given explicitly as constraints. This following examples demonstrate this idea.

Lagrange dual of standard form LP

Consider an LP in standard form,

$$\begin{aligned} &minimize\ c^T x \\ &subject\ to\ Ax = b \\ &x \succeq 0, \end{aligned}$$

which has inequality constraint functions $f_i(x) = -x_i, i = 1, ..., n$. To form the Lagrangian we introduce multipliers λ_i for the n inequality constraints and multipliers ν_i for the equality constraints, and obtain

$$L(x, \lambda, \nu) = c^T x - \sum_{i=1}^{n} \lambda_i x_i + \nu^T (Ax - b) = -b^T \nu + (c + A^T \nu - \lambda)^T x.$$

The dual function is

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) = -b^T\nu + \inf_x (c + A^T\nu - \lambda)^T x,$$

which is easly determined analytically, since a linear function is bounded below only when it is identically zero. Thus, $g(\lambda,\nu) = -\infty$ except when $c + A^T\nu - \lambda = 0$, in which case it is $-b^T\nu$:

$$g(\lambda,\nu) = \begin{cases} -b^T\nu, & A^T\nu - \lambda + c = 0 \\ -\infty, & otherwise. \end{cases}$$

note that the dual function g is infinite only on a proper affine subset of $R^m \times R^p$. We will see that this is a common occurrence.

The lower bound property is nontrivial only when λ and ν satisfy $\lambda \succ 0$ and $A^T\nu - \lambda + c = 0$. When this occurs, $-p^T\nu$ is a lower bound on the optimal value of the LP. Stictly speaking, the Lagrange dual problem of the standard form LP is to maximize this dual function g subject to $\lambda \succeq 0$, *i.e.*,

$$\begin{aligned} &maximize\ g(\lambda,\nu) = \begin{cases} -b^T\nu, & A^T\nu - \lambda + c = 0 \\ -\infty, & otherwise. \end{cases} \\ &subjece\ to\ \lambda \succeq 0. \end{aligned}$$

Here g is finite only when $A^T\nu - \lambda + c = 0$. We can form an equivalent problem by making these equality constraints explicit:

$$\begin{aligned} maximize\ & -b^T\nu \\ subject\ to\ & A^T\nu - \lambda + c = 0. \\ & \lambda \geq 0. \end{aligned}$$

This problem, can be express as

$$\begin{aligned} maximize\ & -b^T\nu \\ subject\ to\ & A^T\nu + c \succeq 0, \end{aligned}$$

which is an LP in inequality form.

Lagrange dual of inequality form LP

In a similar way we can find the Lagrange dual problem of a linear program in inequality form.

$$\begin{aligned} &minimize\ c^T x \\ &subject\ to\ Ax \preceq b. \end{aligned}$$

The Lagrangian is

$$L(x,\lambda) = c^T x + \lambda^T(Ax - b) = -b^T\lambda + (A^T\lambda + c)^T x,$$

so the dual function is

$$g(\lambda) = \inf_x L(x,\lambda) = -b^T\lambda + \inf_x (A^T\lambda + c)^T x.$$

The infinmum of a linear function is $-\infty$, except in the special case when it is identically zero, so the dual function is

$$g(\lambda, \nu) = \begin{cases} -b^T\nu, & A^T\nu - \lambda + c = 0 \\ -\infty, & otherwise. \end{cases}$$

The dual variable λ is dual feasible if $\lambda \succeq 0$ and $A^T\lambda + c = 0$.

The Lagrange dual of the LP is to maximize g over all $\lambda \succeq 0$. Again we can reformulate this by explicitly including the dual feasible conditions as constraints, as in

$$\begin{aligned} &maximize \ -b^T\lambda \\ &subject\ to\ A^T\lambda + c = 0 \\ &\lambda \succeq 0, \end{aligned}$$

Which is an LP in standard form.

Note the interesting symmetry between the standard and inequality form LPs and their duals: The dual of a standard form LP is an LP with only inequality constraints, and vice versa. One can also verify that the Lagrange dual of this is equivalent to the primal problem.

3.2.2 Weak duality

The optimal value of the Lagrange dual problem, which we denote $d^\star$, is, by defination, the best lower bound on $p^\star$ that can be obtained from the Lagrange dual function. In particular, we have the simple but important inequality

$$d^\star \leq p^\star,$$

which holds even if the original problem is not convex. This property is called weak duality.

We refer to the difference $p^\star - d^\star$ as the optimal duality gap of the original problem, since it gives the gap between the optimal value of the primal problem and the best lower bound on it that can be obtained from the Lagrange dual function. The optimal duality gap is always nonnegative.

The bound can sometimes be used to find a lower bound on the optimal value of a problem that is difficult to solve, since the dual problem is always convex, and in many cases can be solved efficiently, to find $d^\star$.

3.2.3 Strong duality and Slater's constraint qualification

If the equality

$$d^\star = p^\star$$

holds, *i.e.*, the optimal duality gap is zero, then we say that strong duality holds. This means that the best bound that can be obtained from the Lagrange dual function is tight.

Strong duality does not, in general, hold. But if the primal problem is convex, *i.e.*, of the form

$$\begin{aligned} &minimize\ f_0(x) \\ &subject\ to\ f_i(x) \leq 0,\ i = 1, ..., m, \\ &\qquad Ax = b, \end{aligned}$$

with $f_0, ..., f_m$ convex, we usually(but not always) have strong duality. There are many results that establish conditions on the problem, beyond convexity, under which strong duality holds. These conditions are called constraint qualifications.

One simple constraint qualification is Slater's condition: There exists an $x \in relintD$ such that

$$f_i(x) < 0,\ i = 1, ..., m,\ Ax = b.$$

such a point is sometimes called strictly feasible, since the inequality constraints hold with strict inequalities. Slater's theorem states that strong duality holds, if Slater's condition holds and the problem is convex).

Slater's condition can be refined when some of the inequality constraint functions f_i are affine. If the first k constraint functions $f_1, ..., f_k$ are affine, then strong duality holds provided the following weaker condition holds: There exists an $x \in relintD$ with

$$f_i(x) \leq 0, i = 1, ..., k,\ f_i(x) < 0,\ i = k+1, ..., m,\ Ax = b.$$

In other words, the affine inequalities do not need to hold with strict inequality. Note that the refined Slater condition reduces to feasibility when the constraints are all linear equalities and inequalities, and $domf_0$ is open.

Slater's condition not only implies strong duality for convex problems. It also implies that the dual optimal value is attained when $d^\star > -\infty$, *i.e.*, there exists a dual feasible $(\lambda^\star, \nu^\star)$ with $g(\lambda^\star, \nu^\star) = d^\star = p^\star$. We will prove that strong duality obtains, when the primal problem is convex and Slater's condition holds.

A nonconvex quadratic problem with strong duality

On rare occasions strong duality obtains for a nonconvex problem. As an important example, we consider the problem of minimizing a nonconvex quadratic function over the unit ball,

$$\begin{aligned} &minimize\ x^TAx + 2b^Tx \\ &subject\ to\ x^Tx \leq 0, \end{aligned}$$

where $A \in S^n, A \not\succeq 0$, and $b \in R^n$. Since $A \not\succeq 0$, this is not a convex problem. This problem is sometimes called the trust region problem, and arises in minimizing a second-order approximation of a function over the unit ball, which is the region in which the approximation is assumed to be approximately valid.

The Lagrangian is

$$L(x, \lambda) = x^TAx + 2b^Tx + \lambda(x^Tx - 1) = x^T(A + \lambda I)x + 2b^Tx - \lambda,$$

So the Lagrange dual problem is thus

$$\begin{aligned} &maximize \ -b^T(A+\lambda I)^{-1}b - \lambda \\ &subject\ to\ A+\lambda I \succeq 0, b \in R(A+\lambda I), \end{aligned}$$

with variable $\lambda \in R$. Although it is not obvious from this expression, this is a convex optimization problem. In fact, it is readily solved since it can be expressed as

$$\begin{aligned} &maximize \ -\sum_{i=1}^{n}(q_i^T b)^2/(\lambda_i+\lambda) - \lambda \\ &subject\ to\ \lambda \geq -\lambda_{min}(A), \end{aligned}$$

where λ_i and q_i are the eigenvalues and corresponding(orthonormal) eigenvectors of A, and we interpret $(q_i^T b)^2/0$ as 0 if $q_i^T b = 0$ and as ∞ otherwise.

Despite the fact that the original problem is not convex, we always have zero optimal duality gap for this problem: The optimal values of primal problem and the dual problem are always the same. In fact, a more general result holds: strong duality holds for any optimization problem with quadratic objective and one quadratic inequality constraint, provided Slater's condition holds.

3.2.4 Exercises: Examples and applications

3.4 Interpretation of LP dual via relaxed problems. Consider the inequality form LP

$$\begin{aligned} &minimize\ c^T x \\ &subject\ to\ Ax \preceq b, \end{aligned}$$

with $A \in R^{m\times n}$, $b \in R^m$. In this exercise we develop a simple geometric interpretation of the dual LP. Let $w \in R_+^m$. If x is feasible for the LP, *i.e.*, satisfies $Ax \preceq b$, then it also satisfies the inequality

$$w^T Ax \leq w^T b.$$

Geometrically, for any $w \succeq 0$, the halfspace $H_w = \{x | w^T Ax \leq w^T b\}$ contains the feasible set for the LP. Therefore if we minimize the objective c^x over the halfspace H_w we get a lower bound on $p^\star$.
(a) Derive an expression for the minimum value of $c^T x$ over the halfspace H_w(which will depend on the choice of $w \succeq 0$).
(b) Formulate the problem of finding the best such bound, by maximizing the lower bound over $w \succeq 0$.
(c) Relate the results of (a) and (b) to the Lagrange dual of the LP.
Solution:
(a)First of all, the problem is always feasible, the vector c can be decomposed into component parallel to $w^T A$ and a component orthogonal to $w^T A$:

$$c = \lambda A^T w + \hat{c},$$

with $A^T w\hat{c} = 0$.
The optimal value is
(b) We maximize the lower bound by solving

$$\begin{aligned} &maximize\ \lambda w^T b \\ &subject\ to\ c = \lambda A^T w \\ &\qquad\qquad \lambda \leq 0,\ w \succeq 0 \end{aligned}$$

with variables λ and w.
(c) Try to relate (a) and (b), we can just replace $-\lambda w$ as z, we obtain the equivalent problem

$$\begin{aligned} &maximize\ -b^T z \\ &subject\ to\ c = -A^T z \\ &\qquad\qquad z \succeq 0 \end{aligned}$$

This is the dual of the original LP.
3.5 Dual of general LP. Find the dual function of the LP

$$\begin{aligned} &minimize\ c^T x \\ &subject\ to\ Gx \preceq h \\ &\qquad\qquad Ax = b. \end{aligned}$$

Give the dual problem, and make the implicit equality constraints explicit.
Solution: Since the Lagrangian

$$L(x, \lambda, \nu) = c^T x + \lambda^T (Gx - h) + \nu^T (Ax - b)$$

We have the dual Lagrangian function is

$$g(x, \nu) = \inf_x L(x, \lambda, \nu) = \begin{cases} -\lambda^T h - \nu^T b, & c + G^T \lambda + A^T \nu = 0 \\ -\infty, & otherwise \end{cases}$$

And the dual problem is

$$\begin{aligned} &maximize\ g(x, \nu) \\ &subject\ to\ \nu \succeq 0 \end{aligned}$$

After making the implicit equality constraints explict, we obtain

$$\begin{aligned} &maximize\ -\lambda^T h - \nu^T b \\ &subject\ to\ c + G^T \lambda + A^T \nu = 0 \\ &\qquad\qquad \lambda \succeq 0. \end{aligned}$$

3.6 *lower bounds in Chebyshev approximation from least−square.* Consider the Chebyshev or l_∞-norm approximation problem

$$minimize\ \|Ax - b\|_\infty,$$

where $A \in R^{m \times n}$, and $rankA = n$. Let x_{ch} denote an optimal solution(there may be multiple optimal solutions; x_{ch} denotes one of them).

The Chebyshev problem has no closed-form solution, but the corresponding least-squares problem does. Define

$$x_{ls} = argmin\|Ax - b\|_2 = (A^TA)^{-1}A^Tb.$$

We address the following question. Suppose that for a particular A and b we have compute the least-squares solution x_{ls} (but not x_{ch}). How suboptimal is x_{ls} for the Chebyshev problem? In other words, how much larger is $\|Ax_{ls} - b\|_\infty$ than $\|Ax_{ch} - b\|_\infty$?
(a) Prove the lower bound

$$\|Ax_{ls} - b\|_\infty \leq \sqrt{m}\|Ax_{ch} - b\|_\infty,$$

(b) we derived a dual for the general norm approximation problem. Applying the results to the l_∞-norm (and its dual norm, the l_1-norm), we can state the following dual for the chebyshev aooroximation problem:

$$\begin{aligned} &maximize\ b^T\nu \\ &subject\ to\ \|\nu\|_1 \leq 1 \\ &\qquad\qquad A^T\nu = 0. \end{aligned}$$

Any feasible ν corresponds to a lower bound $b^T\nu$ on $\|Ax_{ch} - b\|_\infty$.
Denote the least-squares residual as $r_{ls} = b - Ax_{ls}$. Assuming $r_{ls} \neq 0$, show that

$$\hat{\nu} = -r_{ls}/\|r_{ls}\|_1,\ \widetilde{\nu} = r_{ls}/\|r_{ls}\|_1,$$

as both feasible. By duality $b^T\hat{\nu}$ and $b^T\widetilde{\nu}$ are lower bounds on $\|Ax_{ch} - b\|_\infty$. Which is the better bound? How do these bounds compare with the bound derived in part (a)?
Solution: (a) By the fact that for all $z \in R^m$,

$$\frac{1}{\sqrt{m}}\|z\|_2 \leq \|z\|_\infty \leq \|z\|_2.$$

We have

$$\sqrt{m}\|Ax_{ch} - b\|_\infty \geq \|Ax_{ch} - b\|_2 \geq \|Ax_{ls} - b\|_2 \geq \|Ax_{ls} - b\|_\infty$$

(b)From the expression $x_{ls} = (A^TA)^{-1}$ we note that

$$A^Tr_{ls} = A^T(b - A(A^TA)^{-1}b) = A^Tb - A^Tb = 0.$$

Therefore $A^T\hat{\nu} = 0$ and $A^T\widetilde{\nu} = 0$. Obviously we also have $\|\hat{\nu}\|_1 = 1$ and $\|\widetilde{\nu}\|_1 = 1$, so $\hat{\nu}$ and $\widetilde{\nu}$ are dual feasible.
We can write the dual objective value at $\hat{\nu}$ as

$$b^T\hat{\nu} = \frac{-b^Tr_{ls}}{\|r_{ls}\|_1} = \frac{(Ax_{ls} - b)^Tr_{ls}}{\|r_{ls}\|_1} = -\frac{\|r_{ls}\|_2^2}{\|r_{ls}\|_1}$$

and, similarly,

$$b^T\widetilde{\nu} = \frac{\|r_{ls}\|_2^2}{\|r_{ls}\|_1}.$$

Therefore $\widetilde{\nu}$ gives a better bound than $\widetilde{\nu}$.
Finally, to show that the resulting lower bound is better than the bound in part (a), we have to verify that

$$\frac{\|r_{ls}\|_2^2}{\|r_{ls}\|_1} \geq \frac{1}{\sqrt{m}}\|r_{ls}\|_\infty.$$

This follows from the inequalities

$$\|x\|_1 \leq \sqrt{m}\|x\|_2,\ \|x\|_\infty \leq \|x\|_2$$

which hold for general $x \in R^m$.
3.7 *Piecewise-linear minimization* We consider the convex piecewise-linear minimization problem

$$minimize\ max_{i=1,2,...,m}(a_i^T x + b_i)$$

with variable $x \in R^n$.
(a) Derive a dual problem, based on the Lagrange dual of the equivalent problem

$$\begin{aligned} &minimize\ max_{i=1,...,m} y_i \\ &subject\ to\ y_i = a_i^T x + b_i, i = 1, ..., m \end{aligned}$$

with variables $x \in R^n, y \in R^m$.
(b) Formulate the piecewise-linear minimization problem as an LP, and form the dual of the LP. Relate the LP dual to the dual obtained by in part (a).
(c) Suppose that we approximate the objective function by the smooth function

$$f_0(x) = log(\sum_{i=1}^{m} exp(a_i^T x + b_i)),$$

and solve the unconstrained geometric program.

$$minimize\ log(\sum_{i=1}^{m} exp(a_i^T x + b_i)). \tag{3.3}$$

Show that

$$0 \leq p^\star_{gp} - p^\star_{pwl} \leq logm.$$

(d) Derive similar bounds for the difference between p^*_{pwl} and the optimal value of

$$minimize\ (1/\gamma)log(\sum_{i=1}^{m} exp(\gamma(a_i^T + b_i))),$$

where γ is a parameter. What happens as we increase γ?
Solution:
The Lagrange dual function is

$$g(\lambda) = \inf_{x,y}(\max_{i=1,...,m} y_i + \sum_{i=1}^{m} \lambda_i(a_i^T x + b_i - y_i)). \tag{3.4}$$

For x, it's obvious that only $\sum_{i=1}^{m} = 0$ when $g(\lambda)$ can be finite.
For y, we note that $\underset{y}{inf}(\underset{i}{max}\, y_i - \lambda^T y)$. To analysis this, we first note that if $\lambda \succeq 0, 1^T\lambda = 1$, then

$$\lambda^T y = \sum_j \lambda_j y_j \leq \sum_j \lambda_j \underset{i}{max}\, y_i = \underset{i}{max}\, y_i,$$

we can gain the equality when $y = 0$, so in that case

$$\underset{y}{inf}(\underset{i}{max}\, y_i - \lambda^T y) = 0$$

Then we prove that otherwise, the value will be unbounded blew. Now, assume that $\lambda_j < 0$, then choosing $y_i = 0, i \neq j$, and $y_j = -t$, with $t \geq 0$, and letting t go to infinity, gives

$$\underset{i}{max}\, y_i - \lambda^T y = 0 + t\lambda_k \to -\infty.$$

As a complement, if $1^T\lambda \neq 1$, choosing $y = t1$, gives

$$\underset{i}{max}\, y_i - \lambda^T y = t(1 - 1^T\lambda) \to -\infty,$$

Thus by summing up, we have

$$g(\lambda) = \begin{cases} b^T\lambda & \sum_i \lambda_i a_i = 0,\ \lambda \succeq 0,\ 1^T\lambda = 1 \\ -\infty & otherwise. \end{cases}$$

The resulting dual problem is

$$\begin{aligned} &maximize\ b^T\lambda \\ &subject\ to\ A^T\lambda = 0 \\ &\qquad\qquad 1^T\lambda = 1 \\ &\qquad\qquad \lambda \succeq 0. \end{aligned}$$

(b)

$$minimize\ max_{i=1,2,\ldots,m}(a_i^T x + b_i)$$

The problem is equivalent to the LP

$$\begin{aligned} &minimize\ t \\ &subject\ to\ Ax + b \preceq t1. \end{aligned}$$

the Lagrangian is

$$L(t, x, \lambda) = t + \lambda^T(Ax + b - t1)$$

then the dual function is

$$\begin{aligned} g(\lambda) &= \underset{t,x}{inf}(t + \lambda^T(Ax + b - t1)) \\ &= \underset{t,x}{inf}(t(1 - 1^T\lambda) + A^T\lambda x + b^T\lambda) \end{aligned}$$

then we gain the dual problem is

$$\begin{aligned} &maximize\ b^T\lambda \\ &subject\ to\ A^T\lambda = 0,\ 1^T z = 1,\ \lambda \succeq 0, \end{aligned}$$

which is identical to the dual derived in (a).
(c) Suppose $z^\star$ is dual optimal for the dual GP,

$$\begin{aligned} maximize\ b^T z - \sum_{i=1}^m z_i log z_i \\ subject\ to\ 1^T z = 1 \\ A^T z = 0. \end{aligned}$$

Then $z^\star$ is also feasible for the dual of the piecewise-linear formulation, with objective value

$$b^T z = p^\star_{gp} + \sum_{i=1}^m z_i^\star log z_i^\star.$$

This provides a lower bound on $p^\star_{pwl}$:

$$p^\star_{pwl} \geq p^\star_{gp} + \sum_{i=1}^m z_i^\star log z_i^\star \geq p^\star_{gp} - log m.$$

The bound follows from

$$\max_i (a_i^T x + b_i) \leq log \sum_i exp(a_i^T x + b_i)$$

for all x, and therefore $p^\star_{pwl} \leq p^\star_{gp}$.
In conclusion,

$$p^\star_{gp} - log_m \leq p^\star_{pwl} \leq p^\star_{gp}.$$

(d) We first reformulate the problem as

$$\begin{aligned} minimize\ (1/\gamma) log \sum_{i=1}^m exp(\gamma y_i) \\ subject\ to Ax + b = y. \end{aligned}$$

The Lagrangian is

$$L(x, y, z) = \frac{1}{\gamma} log \sum_{i=1}^m exp(\gamma y_i) + z^T (Ax + b - y).$$

L is bounded below as a function of x only if $A^T z = 0$. To find the optimum over y, we set the gradient equal to zero:

$$\frac{e^{\gamma y_i}}{\sum_{i=1}^m e^{\gamma y_i}} = z_i.$$

This is solvable for y_i if $1^T z = 1$ and $z \succeq 0$. The Lagrange dual function is

$$g(z) = b^T z - \frac{1}{\gamma} \sum_{i=1}^m z_i log z_i,$$

and the dual problem is

$$\begin{aligned} &maximize\ b^T z - (1/\gamma)\sum_{i=1}^{m} z_i log z_i \\ &subject\ to\ A^T z = 0 \\ &\qquad\qquad 1^T z = 1. \end{aligned}$$

Let $p^\star_{gp}(\gamma)$ be the optimal value of the GP. Following the same argument as above, we can conclude that

$$p^\star_{gp}(\gamma) - \frac{1}{\gamma} log m \leq p^\star_{pwl} \leq p^\star_{gp}(\gamma).$$

In other words, $p^\star_{gp}(\gamma)$ approaches $p^\star_{pwl}$ as γ increases.

3.8 Relate the two dual problems derived in example on page 128.

Solution: Suppose for example that ν is feasible in (6.9). Then choosing $\lambda_1 = (A^T\nu + c)^-$ and $\lambda_2 = (A^T\nu + c)^+$, yields a feasible solution in (6.8), with the same objective value. Cpnversely, suppose ν, λ_1 and λ_2 are feasible in (6.8). The equality constraint implies that

$$\lambda_1 = (A^T\nu + c)^- + \nu,\ \lambda_2 = (A^T\nu + c)^+ + \nu,$$

for some $v \succeq 0$. Therefore, we can write (6.8) as

$$\begin{aligned} &maximize\ \ -b^T\nu - u^T(A^T\nu + c)^- + l^T(A^T\nu + c)^+ - (u-l)^T v \\ &subject\ to\ v \succeq 0, \end{aligned}$$

and it is clear that at the optimum $u = 0$. Therefore the optimum ν in (3.8) is also optimal in (3.9).

3.9 *Suboptimality of a simple covering ellipsoid.* Recall the problem of determining the minimimum volume ellipsoid, centered at the origin, that contains the points $\alpha_1, ..., \alpha_m \in R^n$:

(a) Show that the matrix

$$X_{sim} = (\sum_{k=1}^{m} a_k a_k^T)^{-1},$$

is feasible.

(b) Now we establish a bound on how suboptimal the feasible point X_{sim} is, via the dual problem,

$$\begin{aligned} &maximize\ logdet(\sum_{i=1}^{m} \lambda_i a_i a_i^T) - 1^T\lambda + n \\ &subject\ to\ \lambda \succeq 0 \end{aligned}$$

with the implicit constraint $\sum_{i=1}^m \lambda_i a_i a_i^T \succ 0$. To derive a bound, we restrict our attention to dual variables of the form $\lambda = t\mathbf{1}$, where $t > 0$. Find the optimal value of t, and evaluate the dual objective at this λ. Use this to prove that the volume of the ellipsoid $\{u | U^T X_{sim} u \leq 1\}$ is no more than a factor $(m/n)^{n/2}$ more than the volume of the minimum volume ellipsoid.

Solution:
(a)

$$\begin{bmatrix} \sum_{k=1}^m a_k a_k^T & a_k \\ a_i^T & 1 \end{bmatrix} = \begin{bmatrix} \sum_{k\neq i}^m a_k a_k^T & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} a_i \\ 1 \end{bmatrix} \begin{bmatrix} a_i \\ 1 \end{bmatrix}^T$$

is the sum of two positive semidefinite matrices, hence positive semidefinite. The Schur complement of the 1,1 block of this matrix is therefore also positive semidefinite:

$$1 - a_i^T (\sum_{k=1}^m a_k a_k^T)^{-1} a_i \geq 0,$$

which is the desired conclusion.
(b) The dual function evaluated at $\lambda = t\mathbf{1}$ is

$$g(\lambda) = log\ det(\sum_{i=1}^m a_i a_i^T) + nlogt - mt + n.$$

Now we'll maximize over $t > 0$ to get the best lower bound. Setting the derivative with respect to t equal to zero yields the optimal value $t = n/m$. Using this λ we get the dual objective value

$$g(\lambda) = logdet(\sum_{i=1}^m a_i a_i^T) + nlog(n/m).$$

The primal objective value for $X = X_{sim}$ is given by

$$-logdet(\sum_{i=1}^m a_i a_i^T)^{-1},$$

so the duality gap associated with X_{sim} and λ is $nlog(m/n)$. (Recall that $m \geq n$, by our assumption that $a_1, ..., a_m$ span R^n.) It follows that, in terms of the objective function, X_{sim} is no more than $nlog(m/n)$ suboptimal. The volume V of the ellipsoid $varepsilon$ associated with the matrix X is given by $V = exp(-O/2)$, where O is the associated objective function, $O = -logdetX$. The bound follows.
3.10 *Optimal experiment design.* The following problems arise in experiment design
(a) D-optimal design.

$$\begin{aligned} &minimize\ logdet(\sum_{i=1}^p x_i v_i v_i^T)^{-1} \\ &subject\ to\ x \succeq 0,\ 1^T x = 1. \end{aligned}$$

(b) A-optimal design.

$$\begin{aligned} &minimize tr(\sum_{i=1}^p x_i v_i v_i^T)^{-1} \\ &subject\ to\ x \succeq 0, 1^T x = 1. \end{aligned}$$

The domain of both problems is $\{x | \sum_{i=1}^P x_i v_i v_i^T \succ 0\}$. The variable is $x \in R^p$; the vectors $v_1, ..., v_p \in R^n$ are given.
Derive dual problems by first introducing a new variable $X \in S^n$ and an equality constraint

$X = \sum_{i=1}^{p} x_i v_i v_i^T$, and then applying Lagrange duality, Simplify the dual problems as much as you can.

Solution:

(a) D-optimal design.

$$\begin{aligned} &minimize\ logdet(X^{-1}) \\ &subject\ to\ X = \sum_{i=1}^{p} x_i v_i v_i^T \\ &\qquad x \succeq 0, \mathbf{1}^T x = 1. \end{aligned}$$

The Lagrangian is

$$\begin{aligned} L(x, Z, z, \nu) &= logdet(X^{-1}) + tr(ZX) - \sum_{i=1}^{p} x_i v_i^T Z v_i - z^T x + \nu(1^T x - 1) \\ &= logdet(X^{-1}) + tr(ZX) + \sum_{i=1}^{p} x_i(-v_i^T - z_i + \nu) - \nu. \end{aligned}$$

The minimum over x_i is bounded below only if $-\nu - v_i^T Z v_i = z_i$. Setting the gradient with respect to $\mathcal{X}$ equal to zero gives $\mathcal{X}^{-1} = Z$. We obtain the dual function

$$g(Z, z) = \begin{cases} logdetZ + n - \nu & \nu - v_i^T Z v_i = z_i, i = 1, ..., p \\ -\infty & otherwise. \end{cases}$$

The dual problem is

$$\begin{aligned} &maximize\ logdetZ + n - \nu \\ &subject\ to\ v_i^T Z v_i \leq \nu,\ i = 1, ..., p, \end{aligned}$$

with domain $S_{++}^n \times R$. We can eliminate nu by first making a change of variables $W = (1/\nu)Z$, which gives

$$\begin{aligned} &maximize\ logdetW + n + nlog\nu - \nu \\ &subject\ to\ v_i^T \hat{W} v_i \leq 1,\ i = 1, ..., p, \end{aligned}$$

Finally, we note that we can easily optimize $nlog\nu - \nu$ over ν. The optimum is $\nu = n$, and substituting gives

$$\begin{aligned} &maximize\ logdetW + nlogn \\ &subject\ to\ v_i^T W v_i \leq 1,\ i = 1, ..., p. \end{aligned}$$

(b) A-optimal design

$$\begin{aligned} &minimize\ tr(X^{-1}) \\ &subject\ to\ X = (\sum_{i=1}^{p} x_i v_i v_i^T)^{-1} \\ &\qquad x \succeq 0, 1^T x = 1. \end{aligned}$$

The Lagrangian is

$$L(X,Z,z,\nu) = tr(X^{-1}) + tr(ZX) - \sum_{i=1}^{p} x_i v_i^T Z v_i - z^T x + \nu(1^T x - 1)$$
$$= tr(X^{-1}) + tr(ZX) + \sum_{i=1}^{p} x_i(-v_i^T Z v_i - z_i + \nu) - \nu.$$

3.11 Derive a dual problem for

$$minimize \ \sum_{i=1}^{N} \|A_i x + b_i\|_2 + (1/2)\|x - x_0\|_2^2.$$

The problem data are $A_I \in R^{m_i \times n}, b_i \in R^{m_i}$, and $x_0 \in R^n$. First introduce new variables $y_i \in R^{m_i}$ and equality constraints $y_i = A_i x + b_i$.
Solution: First, we give the Lagrangian

$$L(x,y,\lambda) = \sum_{i=1}^{N} \|A_i x + b_i\|_2 + (1/2)\|x - x_0\|_2^2 - \lambda(Y - Ax - b)$$
$$= \sum_{i=1}^{N} \|y_i\|_2 + (1/2)\|x - x_0\|_2^2 - \sum_{i=1}^{N} z_i^T (y_i - A_i x - b_1).$$

Then the dual function

$$g(\lambda) = \inf_{x,y} (\sum_{i=1}^{N} \|y_i\|_2 + (1/2)\|x - x_0\|_2^2 - \sum_{i=1}^{N} z_i^T (y_i - A_i x - b_1))$$

We first minimize over y_i. We have

$$\inf_{y_i} (\|y_i\|_2 + z_i^T y_i)$$

By Cauchy-Schwarz inequlity, if $\|z_i\|_2 \leq 1$,the minimum is 0. and the minimum is reached when $y_i = 0$. It's obvious that otherwise, it will be unbounded below.
Then we minimize over x. By differential, we get

$$x = x_0 - \sum_{i=1}^{N} A_i^T z.$$

Finally, we get the dual problem is

$$maximize \ \sum_{i=1}^{N} (A_i x_0 + b_i)^T z_i - \frac{1}{2} \| \sum_{i=1}^{N} A_I^T z_i \|^2$$
$$subject \ to \ \|z_i\|_2 \leq 1, \ i = 1, ..., N.$$

3.12 *Analytic centering.* Derive a dual problem for

$$minimize \ -\sum_{i=1}^{m} log(b_i - a_i^t x)$$

with domain $\{x|a_i^T < b_i, i = 1, ..., m\}$. First introduce new variables y_i and equality constraints $y_i = b_i - a_i^T x$.
Solution: We derive the dual of the problem

$$\begin{aligned} &minimize \ -\sum_{i=1}^{m} logy_i, \\ &subject \ to \ y = b - Ax, \end{aligned}$$

where $A \in R^{m \times n}$ has a_i^T as its ith row. The Lagrangian is

$$L(x, y, \nu) = -\sum_{i=1}^{m} logy_i + \nu^T(y - b + Ax)$$

and the dual function is

$$g(\lambda) = \underset{x,y}{inf}(-\sum_{i=1}^{m} log(y_i) + \nu^T(y - b + Ax))$$

We can find get the dual problem

$$\begin{aligned} &maximize \ \sum_{i=1}^{m} log(\nu_i) - b^T\nu + m \\ &subject \ to \ A^T\nu = 0. \end{aligned}$$

3.3 Geometric interpretation

3.3.1 Weak and strong duality via set of values

We can give a simple geometric interpretation of the dual function in terms of the set

$$g = \{(f_1(x), ..., f_m(x), h_1(x), ..., h_p(x), f_0(x)) \in R^m \times R^p \times R | x \in D\},$$

which is the set of values taken on by the constraint and objective functions. The optimal value $p^\star$ of (27) is easily expressed in terms of g as

$$p^\star = inf\{t|(u, \nu, t) \in g, \nu \preceq 0, \nu = 0\}.$$

To evalute the dual function at (λ, ν), we minimize the affine function

$$(\lambda, \nu, 1)^T(u, v, t) = \sum_{i=1}^{m} \lambda_i u_i + \sum_{i=1}^{p} \nu_i v_i + t,$$

over $(u, v, t) \in g$, *i.e.*, we have

$$g(\lambda, \nu) = inf\{(\lambda, \nu, 1)^T(u, v, t)\}$$

In particular, we see that if the infimum is finite, then the inequality

$$(\lambda, \nu, 1)^T(u, v, t) \geq g(\lambda, \nu)$$

defines a supporting hyperplane to g. This is sometimes referred to as a nonvertical supporting hyperplane, because the last component of the normal vector is nonzero.

Now suppose $\lambda \succeq 0$. Then, obviously, $t \geq (\lambda, \nu, 1)^T(u, \nu, t)$ if $u \preceq 0$ and $\nu = 0$. Therefore

$$\begin{aligned} p^\star &= inf\{t|(u, \nu, t) \in g, u \preceq 0, v = 0\} \\ &\geq inf\{(\lambda, \nu, 1)^T(u, v, t)|(u, \nu, t) \in g, \nu \preceq 0, \nu = 0\}. \\ &\geq inf\{(\lambda, \nu, 1)^T(u, v, t)|(u, \nu, t) \in g\} \\ &= g(\lambda, v), \end{aligned}$$

i.e., we have weak duality.

3.3.2 Epigraph variation

In this section we describe a variation on the geometric interpretation of duality in terms of g, which explains why stong duality obtains for most convex problems. We define the set $A \subseteq R^m \times R^p \times R$ as

$$A = \{(u, v, t)|\exists x \in D, f_i(x) \leq u_i, i = 1, ..., m, h_i(x) = v_i, i = 1, ..., p, f_0(x) \leq t\},$$

we can think of A as a sort of epigraph form of g, since A includes all the points in g, as well as points that are 'worse', *i.e.*, those with larger objective or inequality constaint function values.

We can express the optimal value in terms of A as

$$p^\star = inf\{t|(0, 0, t) \in A\}.$$

To evaluate the dual function at a point (λ, ν) with $\lambda \geq 0$, we can minimize the affine function $(\lambda, \nu, 1)^T(u, v, t)$ over A : If $\lambda \succeq 0$, then

$$g(\lambda, \nu) = inf\{(\lambda, \nu, 1)^T(u, v, t)|(u, v, t) \in A\}.$$

If the infimum is finite, then

$$(\lambda, \nu, 1)^T(u, v, t) \geq g(\lambda, nu)$$

defines a nonvertical supporting hyperplane to A.

In particular, since $(0, 0, p^\star) \in bdA$, we have

$$p^\star = (\lambda, \nu, 1)^T(0, 0, p^\star) \geq g(\lambda, nu), \tag{3.5}$$

the weak duality lower bound. Strong duality holds if and only if the weak duality lower bound. Strong duality holds if and only if we have equality in (29) for some dual feasible (λ, ν), *i.e.*, there exists a nonvertical supporting hyperplane to A and its boundary point $(0, 0, p^\star)$.

3.3.3 Proof of strong duality under constraint qualification

In this section we prove that Slater's constraint qualification guarantees strong duality (and that the dual optimum is attained) for a convex problem. We consider the primal problem with $f_0, ..., f_m$ convexm and assume Slater's condition holds: There exists $\widetilde{x} \in relintD$ with $f_i(\widetilde{x}) < 0, i = 1, ..., m$ and $A\widetilde{x} = b$. In order to simplify the proof, we make two additional assumptions: first that D has nonempty interior(hence, relint$D$ =int$D$) and second, that rank $A = p$. We assume that $p^\star$ is finite.

The set A defined is readily shown to be convex if the underlying problem is convex. We define a second convex set B a s

$$B = \{(0, 0, s) \in R^m \times R^p \times R | s < p^\star\}.$$

The sets A and B do not intersect. To see this, suppose $(u, v, t) \in A \cap B$. Since $(u, v, t) \in B$ we have $u = 0, v = 0$, and $t < p^\star$. Since $(u, v, t) \in A$, there exists an x with $f_i(x) \leq 0, i = 1, ..., m, Ax - b = 0$, and $f_0(x) \leq t \leq p^\star$, which is impossible since $p^\star$ is the optimal value of the primal problem.

By the separating hyperplane theorem, there exists $(\widetilde{\lambda}, \widetilde{\nu}, \mu) \neq 0$ and α such that

$$(u, v, t) \in A \implies \widetilde{\lambda}^T u + \widetilde{v}^T \nu + \mu t \geq \alpha,$$

and

$$(u, v, t) \in B \implies \widetilde{\lambda}^T u + \widetilde{v}^T \nu + \mu t \leq \alpha.$$

Thus, we can get $\widetilde{\lambda} \succeq 0$. The second condition simply means that $\mu t \leq \alpha$ for all $t < p^\star$, and hence, $\mu p^\star \leq \alpha$. Together with the first condition we conclude that for any $x \in D$,

$$\sum_{i=1}^{m} \widetilde{\lambda}_i f_i(x) + \widetilde{\lambda}^T (Ax - b) + \mu f_0(x) \geq \alpha \geq \mu p^\star.$$

The geometric idea behind the proof is illustrated in figure, for a simple problem with one inequality constraint. The hyperplane separating A and B defines a supporting hyperplane to A at $(0, p^\star)$. Slater's constraint qualification is used to establish that the hyperplane must be nonvertical.(*i.e.*, has a normal vector of the form $(\lambda^\star, 1)$).

3.3.4 Multicriterion interpretation

There is a natural connection between Lagrange duality for a problem without equality constraints,

$$\begin{aligned} &minimize\ f_0(x) \\ subject\ to\ &f_i(x) \leq 0,\ i = 1, ..., m, \end{aligned}$$

and the scalarization method for the unconstrained multicriterion problem

$$minimize\ (w.r.t.R_+^{m+1})\ F(x) = (f_1(x), ..., f_m(x), f_0(x))$$

In this scalarization, we choose a positive vector $\widetilde{\lambda}$, and minimize the scalar function $\widetilde{\lambda}^T F(x)$; any minimizer is guaranteed to Pareto optimal. Since we can scale $\widetilde{\lambda}$ by a positive constant, without affecting the minimizers, we can, without loss of generality, take $\widetilde{\lambda} = (\lambda, 1)$. Thus, in scalarization we minimize the function

$$\widetilde{\lambda}^T F(x) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x),$$

which is exactly the Lagrangian for the problem. To estabilish that every Pareto optimal point of a convex multicriterion problem minimizes the function $\widetilde{\lambda}^T F(x)$ for some nonnegative weight $\widetilde{\lambda}$, we considered the set A,

$$A = \{t \in R^{m+1} | \exists x \in D, f_i(x) \leq t_i, i = 0, ..., m\},$$

Here too we constructed the required weight vector as a supporting hyperplane to the set, at any arbitrary Pareto optimal point. In multicriterion optimzation, we interpret the components of the weight vector as giving the relative weights between the objective functions. When we fix the last component of the weight vector (associated with f_0) to be one, the other weights have the interpretation of the cost relative to $f_0, i.e$, the cost relative to the objective.

3.4 Saddle-point interpretation

In this section we give several interpretations of Lagrange duality.

3.4.1 Max-min characterization of weak and strong duality

It is possible to express the primal and the dual optimization problems in a form that is more symmetric. To simplify the discussion we assume there are no equality constraints; the results are easily extended to cover them.

First note that

$$\sup_{\lambda \succeq 0} L(x, \lambda) = \sup_{\lambda \succeq 0} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x))$$

Indeed, suppose x is not feasible, and $f_i(x) > 0$ for some i. Then $sup_{\lambda \succeq 0} L(x, \lambda) = \infty$, as can be seen by choosing $\lambda_j = 0, j \neq i$, and $\lambda_i \to \infty$. On the other hand, if $f_i(x) \leq 0, i = 1, ..., m$, then the optimal choice of $lambda$ is $\lambda = 0$ and $sup_{\lambda \succeq 0} L(x, \lambda) = f_0(x)$. This means that we can express the optimal value of the primal problem as

$$p^\star = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda).$$

By the definition of the dual function, we also have

$$d^\star = \sup_{\lambda \succeq 0} \inf_x L(x, \lambda).$$

Thus, weak duality can be expressed as the inequality

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda \succeq 0} L(x,\lambda),$$

and strong duality as the equality

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) = \inf_x \sup_{\lambda \succeq 0} L(x,\lambda),$$

Strong duality means that the order of the minimization over x and the maximization over $\lambda \succeq 0$ can be switched without affecting the result.

In fact, the inequality does not depend on any properties of L: We have

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda \succeq 0} L(x,\lambda)$$

for any $f : R^n \times R^m \to R$(and any $W \subseteq R^n$ and $Z \subseteq R^m$). This general inequality is called the max-min inequality. When equality holds, *i.e.*,

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) = \inf_x \sup_{\lambda \succeq 0} L(x,\lambda),$$

we say that f satisfy the strong max-min property or the saddle-point property.

3.4.2 Saddle-point interpretation

We refer to a pair $\widetilde{w} \in W, \widetilde{z} \in Z$ as a saddle-point for f if

$$f(\widetilde{w}, z) \leq f(\widetilde{w}, \widetilde{z}) \leq f(w, \widetilde{z})$$

for all $w \in W$ and $z \in Z$. In other words, $\widetilde{w}$ minimizes $f(w, \widetilde{z})$ and $\widetilde{z}$ maximizes $f(\widetilde{w}, z)$:

$$f(\widetilde{w}, \widetilde{z}) = \inf_{w \in W} f(w, \widetilde{z}), \;\; f(\widetilde{w}, \widetilde{z}) = \sup_{z \in Z} f(\widetilde{w}, z).$$

This implies that the strong max-min property holds, and that the common value is $f(\widetilde{w}, \widetilde{z})$.

Returning to our discussion of Lagrange duality, we see that if $x^\star$ and $\lambda^\star$ are primal and dual optimnal points for a problem in which strong duality obtains, they form a saddle-point for the Lagrangian. The converse is also true: If (x, λ) is a saddle-point of the Lagrangian, then x is primal optimal, λ is dual optimal, and the optimal duality gap is zero.

3.4.3 Exercises: Strong duality and Slater's condition

6.21 *A convex problem in which strong duality fails.* Consider the optimization problem

$$\begin{aligned} &\textit{minimize } e^{-x} \\ &\textit{subject to } x^2/y \leq 0 \end{aligned}$$

with variables x and y, and the domain $D = \{(x, y)|y > 0\}$.
(a) Verify that this is a convex optimization problem. Find the optimal value.
(b)Give the Lagrange dual problem, and find the optimal solution λ^* and the optimal value $f^\star$ of the dual problem. What is the optimal duality gap?
(c)Does Slater's condition hold for this problem.
(d)What is the optimal value $p^\star(u)$ of the perturbed problem

$$\begin{aligned} &minimize\ e^{-x} \\ &subject\ to\ x^2/y \leq u \end{aligned}$$

as a function of u? Verify that the global sensitivity inequality

$$p^\star(u) \geq p^\star(0) - \lambda^\star u$$

does not hold.
Solution:
(a) Since the domain$D = \{(x, y)|y > 0\}$ and constraint $x^2/y \leq 0$, we can gain that $x = 0$ is the only feasible point for this problem. Thus the optimal value is 1.
(b) the dual function

$$g(\nu) = \inf_{x,y}(e^{-x} + \nu x^2/y)$$

with the domain $D = \{(x, y)|y > 0\}$ Then when $y \to \infty$, the second term tends to zero. Finally, we can get the optimal value of the dual problem is 0. The optimal duality gap is 1.
(c) No.
(d) When $u < 0$, the optimal value $p^\star(u)$ will be $+\infty$. When $u > 0$, the optimal value $p^\star(u)$ will be 0. Finally, we can check that the global sensitivity inequality does not hold.
6.22 *Weak max-min inequality*. Show that the weak max-min inequality

$$\sup_{z\in Z}\inf_{w\in W} f(w, z) \leq \inf_{w\in W}\sup_{z\in Z} f(w, z)$$

always holds, with no assumptions on $f : R^n \times R^m \to R, W \subseteq R^n$, or $Z \subseteq R^m$.
Solution: When W and Z are empty, the inequality reduces to $-\infty \leq \infty$.
When W is nonempty, with $\widetilde{w} \in W$, we have

$$\inf_{w\in W} f(w, z) \leq f(\widetilde{w}, z)$$

for all $z \in Z$. Taking the supremum over $z \in Z$ on both sides we get

$$\sup_{z\in Z}\inf_{w\in W} f(w, z) \leq \sup_{z\in Z} f(\widetilde{w}, z)$$

then taking the inf over $\widetilde{w} \in W$ we get the max-min inequality. The proof for nonempty Z is similar.

3.5 Optimality conditions

3.5.1 Certificate of suboptimality and stopping criteria

If we can find a dual feasible (λ, ν), we establish a lower bound on the optimal value of the primal problem: $p^\star \geq g(\lambda, \nu)$. Thus a dual feasible point (λ, ν) provides a proof or certificate that $p^\star \geq g(\lambda, \nu)$. Strong duality means there exist arbitrarily good certificates.

Dual feasible points allow us to bound how suboptimal a given feasible point is, without knowing the exact value of $p^\star$. Indeed, if x is primal and (λ, ν) is dual feasible, then

$$f_0(x) - p^\star \leq f_0(x) - g(\lambda, \nu).$$

In particular, this establishes that x is ϵ-suboptimal, with $\epsilon = f_0(x) - g(\lambda, \nu)$. (It also establishes that (λ, ν) is ϵ-suboptimal for the dual problem.)

We refer to the gap between primal and dual objectives,

$$f_0(x) - g(\lambda, \nu),$$

as the duality gap associated with the primal feasible point x and dual feasible point (λ, ν). A primal dual feasible pair x, (λ, ν) localizes the optimal value of the primal and dual problems to an interval:

$$p^\star \in [g(\lambda, \nu), f_0(x)],\ d^\star \in [g(\lambda, \nu), f_0(x)],$$

the width of which is the duality gap.

If the duality gap of the primal dual feasible pair x, (λ, ν) is zero, *i.e.*, $f_0(x) = g(\lambda, \nu)$, then x is primal optimal and (λ, ν) is dual optimal. We can think of (λ, ν) as a certificate that proves x is optimal (and, similarly, we can think of x as a certificate that proves (λ, ν) is dual optimal).

These observations can be used in optimization algorithms to provide nonheuristic stopping criteria. Suppose an algorithm produces a sequence of primal feasible $x^{(x)}$ and dual feasible $(\lambda^k, \nu^{(k)})$, for $k = 1, 2....$, and $\epsilon_{abs} > 0$ is a given required absolute accuracy. Then the stopping criterion (*i.e.*, the condition for terminating the algorithm)

$$f_0(x^{(k)}) - g(\lambda^{(k)}, \nu^{(k)}) \leq \epsilon_{abs}$$

guarantees that when the algorithm terminates, ϵ_{abs}-suboptimal. Indeed, $(\lambda^{(k)}, \nu^{(k)})$ is a certificate that proves it. (Of course strong duality must hold if this method is to work for arbitrarily small tolerances ϵ_{abs}.)
A similar condition can be used to guarantee a given relative accuracy $\epsilon_{rel} > 0$.
If

$$g(\lambda^{(k)}, \nu^{(k)}) > 0,\ \frac{f_0(x^{(k)} - g(\lambda^{(k)}, \nu^{(k)}}{g(\lambda^{(k)}, \nu^{(k)}} \leq \epsilon_{rel}$$

holds, or

$$f_0(x^{(k)} < 0,\ \frac{f_0(x^{(k)} - g(\lambda^{(k)}, \nu^{(k)}}{-f_0(x^{(k)}} \leq \epsilon_{rel}$$

holds, then $p^\star \neq 0$ and the relative error

$$\frac{f_0(x^{(k)}) - p^\star}{|p^\star|}$$

is guaranteed to be less than or equal to ϵ_{rel}.

3.5.2 Complementary slackness

Suppose that the primal and dual optimal values are attained and equal. Let $x^\star$ be a primal optimal and $(\lambda^\star, \nu^\star)$ be a dual optimal point. This means that

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \inf_x (f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \\ &\leq f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\ &\leq f_0(x^\star) \end{aligned}$$

This first line states that the optimal duality gap is zero, and the second line is definition of the dual function. The third line follows since the infimum of the Lagrangian over x is less than or equal to its value at $x = x^\star$. The last inequality follows from $\lambda_i^\star \geq 0$, $f_i(x^\star) \leq 0, i = 1, ..., m$ and $h_i(x^\star) = 0, i = 1, ..., p$. We conclude that the two inequalities in this chain hold with equality.

We can draw several interesting conclusions from this. Since the inequality in the third line is an equality, we conclude that $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$ over x.(The Lagrangian $L(x, \lambda^\star, \nu^\star)$ can have other minimizers; $x^\star$ is simply a minimizer.)

Another important conclusion is that

$$\sum_{i=1}^m \lambda_i^\star f_i(x^\star) = 0.$$

Since each term in this sum is nonpositive, we conclude

$$\lambda_i^\star f_i(x^\star) = 0, i = 1, ..., m.$$

This condition is known as complementary slackness; it holds for any primal optimal $x^\star$ and any dual optimal $(\lambda^\star, \nu^\star)$ when strong duality holds). We can express the complementary slackness condition as

$$\lambda_i^\star > 0 \implies f_i(x^\star) = 0,$$

or, equivalently,

$$f_i(x^\star) < 0 \implies \lambda_i^\star = 0.$$

Roughly speaking, this means the ith optimal Lagrange multiplier is zero unless the ith constraint is active at the optimum.

3.5.3 KKT optimality conditions

We now assume that the function $f_0, ..., f_m, h_1, ..., h_p$ are differentiable (and therefore have open domains), but we make no assumptions yet about convexity.

KKT conditions for nonconcex provblems

As above, let $x^\star$ and $(\lambda^\star, \nu^\star)$ be any primal and dual optimal points with zero duality gap. Since $x^\star$ minimize $L(x, \lambda^\star, \nu^\star)$over x, it follows that its gradient mush vanish at $x^\star, i.e.,$

$$\bigtriangledown f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \bigtriangledown f_i(x^\star) + \sum_{i=1}^{p} \bigtriangledown h_i(x^\star) = 0.$$

Thus we have

$$\begin{aligned} f_i(x^\star) &\leq 0,\ i = 1, ..., m \\ h_i(x^\star) &= 0,\ i = 1, ..., p \\ \lambda_i^\star &\geq 0,\ i = 1, ..., m \\ \lambda_i^\star f_i(x^\star) &= 0,\ i = 1, ..., m \end{aligned}$$

$$\bigtriangledown f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \bigtriangledown f_i(x^\star) + \sum_{i=1}^{p} \nu_{i=1}^\star \bigtriangledown h_i(x^\star) = 0,$$

which are called the Karush-Kuhn-Trucker(KKT) condition.

To summarize, for any optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions.

KKT conditions for convex problems

When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal. In other words, if f_i are convex and h_i are affine, and $\widetilde{x}, \widetilde{\lambda}, \widetilde{\nu}$ are any points that satisfy the KKT conditions

$$\begin{aligned} f_i(\widetilde{x}) &\leq 0,\ i = 1, ..., m \\ h_i(\widetilde{x}) &= 0,\ i = 1, ..., p \\ \widetilde{\lambda}_i &\geq 0,\ i = 1, ..., m \\ \widetilde{\lambda}_i f_i(x^\star) &= 0,\ i = 1, ..., m \end{aligned}$$

$$\bigtriangledown f_0(\widetilde{x}) + \sum_{i=1}^{m} \widetilde{\lambda}_i \bigtriangledown f_i(\widetilde{x}) + \sum_{i=1}^{p} \widetilde{\nu}_{i=1} \bigtriangledown h_i(\widetilde{x}) = 0,$$

then $\widetilde{x}$ and $(\widetilde{\lambda}, \widetilde{\nu})$ are primal and dual optimal, with zero duality gap.

To see this, note that the first two conditions state that $\widetilde{x}$ is primal feasible. Since $\widetilde{\lambda}_i \geq 0, L(x, \widetilde{\lambda}, \widetilde{\nu})$ is convex in x; the Last KKT condition states that its gradient with respect to x vanishes at $x = \widetilde{x}$, so it follows that $\widetilde{x}$ minimizes $L(x, \widetilde{\lambda}, \widetilde{\nu})$ over x. From this we conclude that

$$\begin{aligned} g(\widetilde{\lambda}, \widetilde{\nu}) &= L(\widetilde{x}, \widetilde{\lambda}, \widetilde{\nu}) \\ &= f_0(\widetilde{x}) + \sum_{i=1}^{m} \widetilde{\lambda}_i f_i(\widetilde{x}) + \sum_{i=1}^{p} \widetilde{\nu}_i h_i(\widetilde{x}) \\ &= f_0(\widetilde{x}), \end{aligned}$$

where in the last line we use $h_i(\widetilde{x}) = 0$ and $\widetilde{\lambda}_i f_i(\widetilde{x}) = 0$. This shows that $\widetilde{x}$ and $(\widetilde{\lambda}, \widetilde{\nu})$ have zero duality gap, and therefore are primal and dual optimal. In summary, for any convex optimization problem with differentiable objective and constraint functions, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap.

If a convex optimization problem with differentiable objective and constraint functions satisfies Slater's condition, then the KKT conditions provide necessary and sufficient conditions for optimality: Slater's condition implies that the optimal duality gap is zero and the dual optimum is attained, so x is optimal if and only if there are (λ, ν) that, together with x, satisfy the KKT conditions.

The KKT conditions play an important role in optimization. In a few special cases it is possible to solve the KKT conditions(and therefore, the optimization problem) analytically. More generally, many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.

Example 5.1 *Equality constrained convex quadratic minimization.* We consider the problem

$$\begin{aligned} &\textit{minimize } (1/2)x^T Px + q^T x + r \\ &\textit{subject to } Ax = b, \end{aligned}$$

where $P \in S_+^n$. The KKT conditions for this problem are

$$Ax^\star = b,\ Px^\star + q + A^T v^\star = 0,$$

which we can write as

1

Solving this set of $m + n$ equations in the $m + n$ variables $x^\star, \nu\star$ gives the optimal primal and dual variables.

3.5.4 Solving the primal problem via the dual

We mentioned that if strong duality holds and a dual optimal solution $(\lambda^\star, \nu^\star)$ exists, then any primal optimal point is also a minimizer of $L(x, \lambda^\star, \nu^\star)$. This fact sometimes allows us to compute a primal optimal solution from a dual optimal solution.

More precisely, suppose we have strong duality and an optimal $(\lambda^\star, \nu^\star)$ is known. Suppose that the minimizer of $L(x, \lambda^\star, \nu^\star), i.e.$, the solution of

$$\textit{minimize } \bigtriangledown f_0(x) + \sum_{i=1}^{m} \lambda_i^\star \bigtriangledown f_i(x) + \sum_{i=1}^{p} \bigtriangledown h_i(x), \tag{3.6}$$

is unique. (For a convex problem this occurs, for example, if $L(x, \lambda^\star, \nu^\star)$ is a strictly convex function of x.) Then if the solution is primal feasible, it must be primal optimal; if it is not primal feasible, then no primal optimal point can exist. $i.e.$, we conclude that the primal optimum is not attained. The observation is interesting when the dual problem is easier to solve than the primal problem, for example, because it can be solved analytically, or has some special structure that can be exploited.

Example *Minimizing a separable function subject to an equality constraint.* We consider the problem

$$\begin{aligned} &minimize\ f_0(x) = \sum_{i=1}^{n} f_i(x_i) \\ &subject\ to\ a^T x = b, \end{aligned}$$

where $a \in R^n, b \in R$, and $f_i : R \to R$ are differentiable and strictly convex. The objective function is called separable since it is a sum of functions of the individual variables $x_1, ..., x_n$. We assume that the domain of f_0 interesects the constraint set, *i.e.*, there exists a point $x_0 \in dom f_0$ with $a^T x_0 = b$, This implies the problem has a unique optimal point $x^\star$.
The lagrangian is

$$L(x, \nu) = \sum_{i=1}^{n} f_i(x_i) + \nu(a^T x - b) = -b\nu + \sum_{i=1}^{n}(f_i(x_i) + \nu a_i x_i),$$

which is also separable, so the dual function is

$$\begin{aligned} g(\nu) &= -b\nu + \underset{x}{inf}(\sum_{[}^{n} i = 1](f_i(x_i) + \nu a_i x_i)) \\ &= -b\nu + \sum_{i=1}^{n} \underset{x_i}{inf}(f_i(x_i) + \nu a_i x_i) \\ &= -b\nu - \sum_{i=1}^{n} f_i^\star(-\nu a_i). \end{aligned}$$

The dual problem is thus

$$maximize\ -b\nu - \sum_{i=1}^{n} f_i^\star(-\nu a_i),$$

with variable $nu \in R$. Now suppose we have found an optimal dual variable $\nu^\star$. Since f_i is strictly convex, the function $L(x, \nu^\star)$ is strictly convex in x, and so has a unique minimizer $\widetilde{x}$. But we also know that $x^\star$ minimizes $L(x, \nu^\star)$, so we must have $\widetilde{x} = x^\star$. We can recover $x^\star$ from $\bigtriangledown_x L(x, \nu^\star) = 0$, *i.e.*, by solving the equations $f_i'(x_i^\star) = -\nu^\star a_i$.

3.5.5 Exercises: Optimality conditions

3.23 Consider the QCQP

$$\begin{aligned} &minimize\ x_1^2 + x_2^2 \\ &subject\ to\ (x_1 - 1)^2 + (x_2 - 1)^2 \le 1 \\ &\qquad\qquad (x_1 - 1)^2 + (x_2 + 1)^2 \le 1 \end{aligned}$$

with variable $x \in R^2$.

(a) Sketch the feasible set and level sets of the objective. Find the optimal point $x^\star$ and optimal value $p^\star$.

(b) Give the KKT conditions. Do there exist Lagrange multipliers $\lambda_1^\star$ and $\lambda_2^\star$ that prove that $x^\star$ is optimal?

(c) Derive and solve the Lagrange dual problem. Does strong duality hold?

Solution:

(a)The figure shows the feasible set, and the contour lines of the objective function. There is only one feasible point, $(0, 1)$, so it is optimal for the primal problem, and we have $p^\star = 1$.

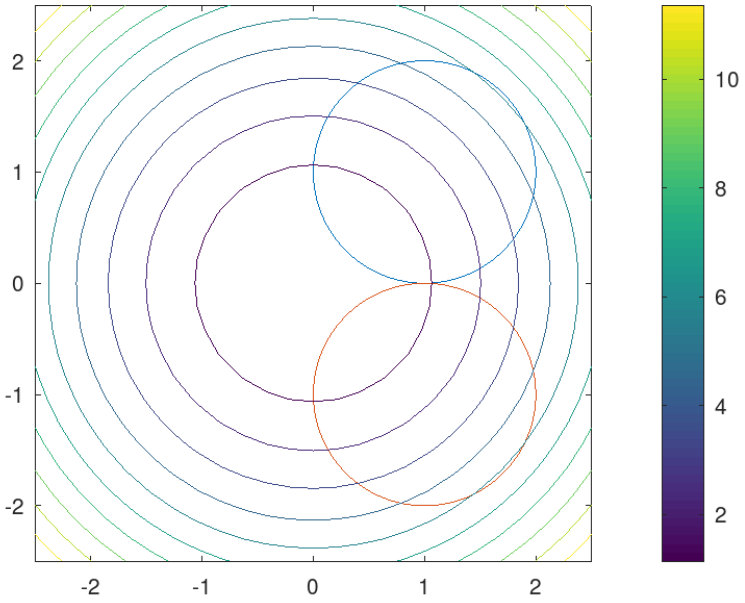


Figure 18

The KKT condition are

$$(x_1^\star - 1)^2 + (x_2^\star - 1)^2 \le 1,\ (x_1^\star - 1)^2 + (x_2^\star + 1)^2 \le 1,$$
$$\lambda_1^\star \ge 0,\ \lambda_2^\star \ge 0$$
$$2x_1^\star + 2\lambda_1^\star(x_1^\star - 1) + 2\lambda_2^\star(x_1^\star - 1) = 0$$
$$2x_2^\star + 2\lambda_1^\star(x_2^\star - 1) + 2\lambda_2^\star(x_2^\star + 1) = 0$$
$$\lambda_1^\star((x_1^\star - 1)^2 + (x_2^\star - 1)^2 - 1) = \lambda_2^\star((x_2^\star + 1)^2 + (x_2^\star + 1)^2 - 1) = 0.$$

But the only feasible point is $(1, 0)$, these conditions reduce to

$$\lambda_1^\star \ge 0,\ \lambda_2^\star \ge 0,\ 2 = 0,\ -2\lambda_1^\star + 2\lambda_2^\star = 0,$$

which have no solution.

(c) The Lagrange dual function is given by

$$g(\lambda_1, \lambda_2) = \inf_{x_1, x_2} L(x_1, x_2, \lambda_1, \lambda_2)$$

where

$$\begin{aligned}
&L(x_1, x_2, \lambda_1, \lambda_2) \\
&= x_1^2 + x_2^2 + \lambda_1((x_1-1)^2 + (x_2-1)^2 - 1) + \lambda_2((x_1-1)^2 + (x_2+1)^2 - 1) \\
&= (1 + \lambda_1 + \lambda_2)x_1^2 + (a + \lambda_1 + \lambda_2)x_2^2 - 2(\lambda_1 + \lambda_2)x_1 - 2(\lambda_1 - \lambda_2)x_2 + \lambda_1 + \lambda_2.
\end{aligned}$$

so L reches its minimum for

$$x_1 = \frac{\lambda_1 + \lambda_2}{1 + \lambda_1 + \lambda_2},\ x_2 = \frac{\lambda_1 - \lambda_2}{1 + \lambda_1 + \lambda_2},$$

And we find

$$g(\lambda_1, \lambda_2) = \begin{cases} -\frac{(\lambda_1+\lambda_2)^2+(\lambda_1-\lambda_2)^2}{(1+\lambda_1+\lambda_2)} + \lambda_1 + \lambda_2 & 1 + \lambda_1 + \lambda_2 \geq 0 \\ -\infty & \textit{otherwise.} \end{cases}$$

And then we can deduce that the optimum of this function is $\lambda_1 = \lambda_2$, the quadratic function and the symmetric property. The dual function then simplifies to

$$g(\lambda_1, \lambda_2) = \frac{2\lambda_1}{2\lambda_1 + 1}$$

We see that $g(\lambda_1, \lambda_2)$ tends to 1 as $\lambda_1 \to \infty$. We have $d^\star = p^\star = 1$, but the dual optimum is not attained.
Recall that the KKT conditions only hold if

(1) Strong duality holds,

(2) Primal optimum is attained,

(3) Dual optimum is attained.

In this example, the KKT conditions fail because the dual optimum is not attained.
3.24 *Equality constrained least-squares.* Consider the equality constrained least-squares problem

$$\begin{aligned}
&\textit{minimize } \|Ax - b\|_2^2 \\
&\textit{subject to } Gx = h
\end{aligned}$$

where $A \in R^{m \times n}$ with $rank\ A = n$, and $G \in R^{p \times n}$ with $rank\ G = p$.
Give the KKT conditions, and derive expressions for the primal solution $x^\star$ and the dual solution $\nu^\star$.
Solution:
The Lagrangian is

$$\begin{aligned}
L(x, \nu) =& \|Ax - b\|_2^2 + \nu^T(Gx - h) \\
=& x^T A^T A x + (G^T\nu - 2A^T b)^T x - \nu^T h,
\end{aligned}$$

with minimizer $x = -(1/2)(A^T A)^{-1}(G^T\nu - 2A^T b)$. The dual function is

$$g(\nu) = -(1/4)(G^T\nu - 2A^T b)^T (A^T A)^{-1}(G^T\nu - 2A^T b) - \nu^T h$$

The optimality conditions are

$$2A^T(Ax^\star - b) + G^T\nu^\star = 0,\ Gx^\star = h.$$

Thus the expression for the primal solution $x^\star$

$$x^\star = (A^TA)^{-1}(A^Tb - (1/2)G^T\nu^\star).$$

By KKT conditions $Gx^\star = h$, we have the expression for the dual solution

$$\nu^\star = -2(G(A^TA)^{-1}G^T)^{-1}(h - G(A^TA)^{-1}A^Tb).$$

3.25 The problem

$$\begin{array}{l} minimize\ -3x_1^2 + x_2^2 + 2x_3^2 + 2(x_1 + x_2 + x_3) \\ subject\ to\ x_1^2 + x_2^2 + x_3^2 = 1, \end{array}$$

is a special case of (), so strong duality holds even though the problem is not convex. Derive the KKT conditions. Find all solutions x, ν that satisfy the KKT conditions. Which pair corresponds to the optimum?

Solution:

The KKT conditions are

$$x_1^2 + x_2^2 + x_3^2 = 1,\ (-3+\nu)x_1 + 1 = 0,\ (1+\nu)x_2 + 1 = 0,\ (2+\nu)x_3 + 1 = 0.$$

Then we have$x_1 = \frac{1}{(-3+\nu)}$, $x_2 = \frac{1}{(1+\nu)}$, $x_3 = \frac{1}{2+\nu}$. We can therefore eliminate x and reduce the KKT conditions

$$\frac{1}{(-3+\nu)^2} + \frac{1}{(1+\nu)^2} + \frac{1}{(2+\nu)^2} = 1$$

By solving the function, we have four solutions:

$$\nu = -3.15,\ \nu = 0.22,\ \nu = 1.89,\ \nu = 4.04$$

corresponding to

$$x = (0.16, 0.47, -0.87),\ x = (0.36, -0.82, 0.45), x = (0.90, -0.35, 0.26),\ x = (-0.97, -0.20, 0.17).$$

By computing their corresponding objective value, $\nu^\star = 4.04$.

We can also evaluate the dual objective at the four candidate values for ν. Finally we can note that we must have

$$\bigtriangledown^2 f_0(x^\star) + \nu^\star \bigtriangledown^2 f_i^\star(x^\star) \succeq 0,$$

because $x^\star$ is a minimizer of $L(x, \nu^\star)$, therefore $\nu^\star \geq 3$.

3.26*Supporting hyperplane interpretation of KKT conditions.* Consider a convex problem with no equality constraints,

$$\begin{array}{l} minimize\ f_0(x) \\ subject\ to\ f_i(x) \leq 0,\ i = 1, ..., m. \end{array}$$

Assume that $x^\star \in R^n$ and $\lambda^\star \in R^m$ satisfy the KKT conditions

$$\begin{aligned} f_i(x^\star) &\leq 0,\ i = 1, ..., m \\ \lambda_i^\star &\geq 0,\ i = 1, .., m \\ \lambda_i^\star f_i(x^\star) &= 0,\ i = 1, ..., m \\ \bigtriangledown f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \bigtriangledown f_i(x^\star) &= 0. \end{aligned}$$

show that

$$\bigtriangledown f_0(x^\star)^T (x - x^\star) \geq 0$$

for all feasible x.
Solution: Suppose x is feasible. Since f_i are convex and $f_i(x) \leq 0$ we have

$$0 \geq f_i(x) \geq f_i(x^\star) + \bigtriangledown f_i(x^\star)^T (x - x^\star),\ i = 1, ..., m.$$

Using $\lambda_i^\star \geq 0$, we conclude that

$$\begin{aligned} 0 &\geq \sum_{i=1}^{m} \lambda_i^\star (f_i(x^\star) + \bigtriangledown f_i(x^\star)^T (x - x^\star)) \\ &= \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \bigtriangledown f_i(x^\star)^T (x - x^\star) \\ &= - \bigtriangledown f_0(x^\star)^T (x - x^\star). \end{aligned}$$

In the last line, we use the complementary slackness condition $\lambda_i^\star f_i(x^\star) = 0$, and the last KKT condition. This shows that $\bigtriangledown f_0(x^\star)^T (x - x^\star) \geq 0$, *i.e.*, $\bigtriangledown f_0(x^\star)$ defines a supporting hyperplane to the feasible set at $x^\star$.

3.6 Perturbation and sensitivity analysis

When the strong duality obtains, the optimal dual variables give very useful information about the sensitivity of the optimal value with respect to perturbations of the constraints.

3.6.1 The perturbed problem

We consider the following perturbed version of the original optimization problem:

$$\begin{aligned} &minimize\ f_0(x) \\ &subject\ to\ f_i(x) \leq u_i,\ i = 1, ..., m \\ &\qquad h_i(x) = v_i, i = 1, ..., p, \end{aligned}$$

with variable $x \in R^n$. This problem coincides with the original problem when $u = 0, v = 0$. When u_i is positive it means that we have relaxed the ith inequality constraint; when u_i is negative, it means that we have tightened the constraint, thus the perturbed problem

results from the original problem by tightening or relaxing each inequality constraint by u_i, and changing the righthand side of the equality constraints by v_i.

We define $p^\star(u, v)$ as the optimal vlaue of the perturbed problem:

$$p^\star(u, v) = inf\{f_0(x)|\exists x \in D, f_i(x) \leq u_i, i = 1, ..., m, h_i(x) = v_i, i = 1, ..., p\}.$$

We can have $p^\star(u, v) = \infty$, which corresponds to perturbations of the constraints that result in infeasibility. Note that $p^\star(0, 0) = p^\star$, the optimal value of the unperturbed problem.

When the original problem is convex, the function $p^\star$ is a convex function of u and v; indeed, its epigraph is precisely the closure of the set A.

3.6.2 A global inequality

Now we assume that strong duality holds, and that the dual optimum is attained. (This is the case if the original problem is convex, and Slater's condition is satisfied). Let $(\lambda^\star, \nu^\star)$ be optimal for the dual of the unperturbed problem. Then for all u and v we have

$$p^\star(u, v) \geq p^\star(0, 0) - \lambda^{\star T} u - \nu^{\star T} v.$$

To establish this inequality, suppose that x is any feasible point for the perturbed problem, *i.e.*, $f_i(x) \leq u_i$, for $i = 1, ..., m$, and $h_i(x) = v_i$, for $i = 1, ..., p$. Then we have, by strong duality,

$$\begin{aligned} p^\star(0, 0) = g(\lambda^\star, \nu^\star) &\leq f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \\ &\leq f_0(x) + \lambda^{\star T} u + \nu^{\star T} v. \end{aligned}$$

We can conclude that for any x feasible for the perturbed problem, we have

$$f_0(x) \geq p^\star(0, 0) - \lambda^{\star T} u - \nu^{\star T} v,$$

Sensitivity interpretations
When strong duality holds, various sensitivity interpretations of the optimal Lagrange variables follow directly from the inequality. Some of the conclusions are

If $\lambda_i^\star$ is large and we tighten the ith constraint (*i.e.*, choose $u_i < 0$), then the optimal value $p^\star(u, v)$ is guaranteed to increase greatly.

If $\nu_i^\star$ is large and positive and we take $v_i < 0$, or if $\nu_i^\star$ is large and negative and we take $v_i > 0$, then the optimal value $p^\star(u, v)$ is guaranteed to increase greatly.

If $\lambda_i^\star$ is small, and we loosen the ith constraint ($u_i > 0$), then the optimal value $p^\star(u, v)$ will not decrease too much.

If $\nu_i^\star$ is small and positive, and $v_i > 0$, or if $\nu_i^\star$ is small and negative and $v_i < 0$, then the optimal value $p^\star(u, v)$ will not decrease too much.

The inequality and the conclusions listed above, give a lower bound on the perturbed optimal value, but no upper bound. For this reason, the results are not symmetric with respect to loosening or tightening a constraint. For example, suppose that $\lambda_i^\star$ is large, and we loosen the ith constraint a bit. In this case, the ith constraint a bit. In this case teh inequality is not useful; it does not, for example, imply that the optimal value will decrease considerably.

3.6.3 Local sensitivity analysis

Suppose now that $p^\star(u, v)$ is differentiable at $u = 0, v = 0$. Then, provided strong duality holds, the optimal dual variables $\lambda^\star, \nu^\star$ are related to the gradient of $p^\star$ at $u = 0, v = 0$:

$$\lambda_i^\star = -\frac{\partial p^\star(0,0)}{\partial u_i},\ \nu_i^\star = -\frac{\partial p^\star(0,0)}{\partial v_i}.$$

Thus when $p^\star(u, v)$ is differentiable at $u = 0, v = 0$, and strong duality holds, the optimal Lagrange multipliers are exactly the local sensitivities of the optimal value with respect to constraint perturbations. In contrast to the nondifferentiable case, this interpretation is symmetric: Tightening the ith inequality constraint a small amount yields an increase in $p^\star$ of approximately $-\lambda_i^\star u_i$; loosening the ith constraint a small amount yields a decrease in $p^\star$ of approximately $\lambda_i^\star u_i$.

To show the equations, suppose $p^\star(u, v)$ is differentiable and strong duality holds. For the perturbation $u = te_i$, $v = 0$, where e_i is the ith unit vector, we have

$$\lim_{t\to 0} \frac{p^\star(te_i, 0) - p^\star}{t} = \frac{\partial p^\star(0,0)}{\partial u_i}$$

And the previous inequality states that for $t > 0$,

$$\frac{p^\star(te_i, 0) - p^\star}{t} \geq -\lambda_i^\star$$

while for $t < 0$ we have the opposite inequality. Taking the limit $t \to 0$, with $t > 0$ yields

$$\frac{\partial p^\star(0,0)}{\partial u_i} \geq -\lambda_i^\star,$$

while taking the limit with $t < 0$ yields the opposite inequality, so we conclude that

$$\frac{\partial p^\star(0,0)}{\partial u_i} = -\lambda_i^\star,$$

The same method can be used to estabilish

$$\frac{\alpha p^\star(0,0)}{\partial v_i} = -v_i^\star.$$

The local sensitivity result gives us a quantitative measure of how active a constraint is at the optimum $x^\star$. If $f_i(x^\star), 0$, then teh constraint is inactive, and it follows that the constraint can be tightened or loosened a small amount without affecting the optimal

value. By complementary slackness, teh associated optimal Lagrange multiplier must be zero. But now suppose that $f_i(x^\star) = 0$,*i.e.*, the ith constraint is active at the optimum. The ith optimal Lagrange multiplier tells us how active the constraint is: If $\lambda_i^\star$ is small, it means that the constraint can be loosened or tightened a bit without much effect on the optimal value; if $\lambda_i^\star$ is large, it means that if the constraint is loosened or tightened a bit, the effect on the optimal value will be great.

3.6.4 Exercises: Perturbation and sensitivity analysis

3.27 *Optimal value of perturbed problem.* Let $f_0, f_1, ..., f_m : R^n \to R$ be convex. Show that the function

$$p^\star(u, v) = inf\{f_0(x) | \exists x \in D, f_i(x) \leq u_i, i = 1, ..., m, Ax - b = v\}$$

is convex. This function is the optimal cost of the perturbed problem, as a function of the perturbations u and v.
Solution:
Define the function

$$g(x, u, v) = \begin{cases} f_0(x) & f_i(x) \leq u_i,\ i = 1, ..., m,\ Ax - b = v \\ \infty & otherwise. \end{cases}$$

Since $f_0, f_1, ..., f_m : R^n \to R$ are convex. Then G is convex on its domain

$$domG = \{(x, u, v) | x \in D, f_i(x) \leq u_i, i = 1, ..., m, Ax - b = v\},$$

Therefore,

$$p^\star(u, v) = \inf_x G(x, u, v)$$

is convex.

3.7 Examples

In this section we show by example that simple equivalent reformulations of a problem can lead to very different dual problems.

3.7.1 Introducing new variables and equality constraints

Consider an unconstrained problem of the form

$$minimize\ f_0(Ax + b).$$

Its Lagrange dual function is the constant $p^\star$. So while we do have strong duality, *i.e.*, $p^\star = d^\star$, the Lagrangian dual is neither useful or interesting.

Now let us reformulate the problem as

$$\begin{aligned} &minimize\ f_0(y) \\ &subject\ to\ Ax + b = y. \end{aligned}$$

Here we have introduced new variables y, as well as new equality constraints $Ax + b = y$. The two problems are clearly equivalent.

The Lagrangian of the reformulated problem is

$$L(x, y, \nu) = f_0(y) + \nu^T(Ax + b - y).$$

To find the dual function we minimize L over x and y. Minimizing over x we find that $g(\nu) = -\infty$ unless $A^T\nu = 0$, in which case we are left with

$$g(\nu) = b^T\nu + \underset{y}{inf}(f_0(y) - \nu^T y) = b^T\nu - f_0^*(\nu),$$

where f_0^* is the conjugate of f_0. The dual problem can therefore be expressed as

$$\begin{aligned} &minimize\ b^T\nu - f_0^*(\nu) \\ &subject\ to\ A^T\nu = 0. \end{aligned}$$

Thus, the dual of the reformulated problem is considerably more useful than the dual of the original problem.

Example *Unconstrained geometric problem* Consider the unconstrained geometric problem

$$minimize\ log(\sum_{i=1}^{m} exp(a_i^T x + b_i)).$$

We first reformulate it by introducing new variables and equality constraints:

$$\begin{aligned} &minimize\ f_0(y) = log(\sum_{i=1}^{m} exp(y_i) \\ &subject\ to\ Ax + b = y, \end{aligned}$$

where a_i^T are the rows of A. The conjugate of the log-sum-exp function is

$$f_0^*(\nu) = \begin{cases} b^T\nu - \sum_{i=1}^m \nu_i log\nu_i & \nu \succeq 0,\ 1^T\nu = 1 \\ \infty & otherwise. \end{cases}$$

so the dual of the reformulated problem can be expressed as

$$\begin{aligned} maximize & b^T\nu - \sum_{i=1}^{m} \nu_i log\nu_i \\ subject\ to\ & q^T\nu = 1 \\ & A^T\nu = 0 \\ & \nu \succeq 0, \end{aligned}$$

which is an entropy maximization problem.
Example *Norm approximation problem.* We consider the constrained norm approximation problem

$$minimize\ \|Ax - b\|,$$

Here too the Lagrange dual function is constant, equal to the optimal value of the problem, and therefore not useful.

Once again we reformulate the problem as

$$\begin{aligned} &minimize\ \|y\| \\ &subject\ to\ Ax - b = y. \end{aligned}$$

The Lagrange dual problem is, following

$$\begin{aligned} &maximize b^T\nu \\ &subject\ to\ \|\nu\|_\star \leq 1 \\ &\qquad A^T\nu = 0, \end{aligned}$$

where we use the fact that the conjugate of a norm is the indicator function of a dual norm unit ball.

The idea of introducing new equality constraints can be applied to the constraint functions as well. Consider, for example, the problem

$$\begin{aligned} &minimize\ f_0(A_0x + b_0) \\ &subject\ to\ f_i(A_ix + b_i) \leq 0,\ i = 1, ..., m, \end{aligned}$$

where $A_i \in R^{k_i \times n}$ and $f_i : R^{k_i} \to R$ are convex. We introduce a new variable $y_i \in R^{k_i}$, for $i = 0, ..., m$, and reformulate the problem as

$$\begin{aligned} &minimize\ f_0(y_0) \\ &subject\ to\ f_i(y_i) \leq 0,\ i = 1, ..., m \\ &\qquad A_ix + b_i = y_i,\ i = 0, ..., m \end{aligned}$$

The Lagrangian for this problem is

$$L(x, y_0, ..., y_m, \nu_0, ..., \nu_m) = f_0(y_0) + \sum_{i=1}^{m} \lambda_i f_i(y_i) + \sum_{i=0}^{m} \nu_i^T(A_ix + b_i - y_i).$$

To find the dual function we minimize over x and y_i. The minimium over x is $-\infty$ unless

$$\sum_{i=0}^{m} A_i^T\nu_i = 0,$$

In which case we have, for $\lambda \succ 0$,

$$\begin{aligned}
&g(\lambda, \nu_0, ..., \nu_m) \\
&= \sum_{i=0}^{m} \nu_i^T b_i + \inf_{y_0,...,y_m} (f_0(y_0) + \sum_{i=1}^{m} \lambda_i f_i(y_i) - \sum_{i=0}^{m} \nu_i^T y_i) \\
&= \sum_{i=0}^{m} \nu_i^T b_i + \inf_{y_0} f(f_0(y_0) - \nu_0^T y_0) + \sum_{i=1}^{m} \lambda_i \inf_{y_i} f(f_i(y_i) - (\nu_i/\lambda_i)^T y_i) \\
&= \sum_{i=1}^{m} \nu_i^T b_i - f_i^*(\nu_i) - \sum_{i=1}^{m} \lambda_i f_i^\star(\nu_i/\lambda_i).
\end{aligned}$$

The last expression involves the perspective of the conjugate function, and is therefore concave in the dual variables. Finally, we address the question of what happens when $\lambda \succeq 0$, but some λ_i are zero. If $\lambda_i = 0$ and $\nu_i \neq 0$, then the dual function is $-\infty$. If $\lambda_i = 0$ and$\nu_i = 0$, however, the terms involving y_i, ν_i, and λ_i are all zero. Thus, the expression above for g is valid for all $\lambda \succeq 0$, if we take $\lambda_i f_i^*(\nu_i/\lambda_i) = 0$wehn $\lambda_i = 0$ and $\nu_i = 0$, and $\lambda_i f_i^*(\nu_i/\lambda_i) = \infty$ when $\lambda_i = 0$ and $\nu_i \neq 0$.

3.7.2 Transforming the objective

If we replace the objective f_0 by an increasing function of f_0, the resulting problem is clearly equivalent. The dual of this equivalent problem, however, can be very different form the dual of the original problem.
Example. we consider again the minimum norm problem

$$minimize \ \|Ax - b\|,$$

We reformulate this problem as

$$\begin{aligned}
&minimize \ (1/2)\|y\|^2 \\
&subject \ to \ Ax - b = y.
\end{aligned}$$

Here we have introduced new variables, and replaced the objective by half its square. Evidently it is equivalent to the original problem.
The dual of the reformulated problem is

$$\begin{aligned}
&maximize \ -(1/2)\|\nu\|_*^2 + b^T\nu \\
&subject \ to \ A^T\nu = 0,
\end{aligned}$$

Note the dual problem is not the same as the dual problem derived earlier.

3.7.3 Implicit constraints

The next simple reformulation we study is to include some of the constraints in the objective function, by modifying the objective function to be infinite when the constraint is

violated.
Example *Linear program with box constraints.* We consider the linear program

$$\begin{aligned} &\textit{minimize } c^T x \\ &\textit{subject to } Ax = b \\ &\qquad\qquad l \preceq x \preceq u \end{aligned}$$

where $A \in R^{p\times n}$. The constraints $l \preceq x \preceq u$ are sometimes called *boxconstraints* or *variablebounds*.
We can, of course, derive the dual of this linear program. The dual will have a Lagrange moltiplier ν associated with the equality constraint. λ_1 associated with the inequality constraint $x \preceq u$ and λ_2 associated with the inequality constraint $l \preceq x$. The dual is

$$\begin{aligned} &\textit{maximize} - b^T\nu - \lambda_1^T u + \lambda_2^T l \\ &\textit{subject to } A^T\nu + \lambda_1 - \lambda_2 + c = 0 \\ &\qquad\qquad \lambda_1 \succeq 0,\ lam_2 \succeq 0. \end{aligned} \tag{3.7}$$

Instead, let us first reformulate the problem as

$$\begin{aligned} &\textit{minimize } f_0(x) \\ &\textit{subject to } AX = b, \end{aligned}$$

where we define

$$f_0(x) = \begin{cases} c^T x & l \preceq x \preceq u \\ \infty & otherwise. \end{cases}$$

Then the two problems are clearly equivalent; we have merely made the explicit box constraints implicit. Thus the dual function is

$$\begin{aligned} g(\nu) &= \inf_{l\preceq x\preceq u} (c^T x + \nu^T(Ax - b)) \\ &= -b^T v - u^T(A^T v + c)^- + l^T(A^T v + c)^+ \end{aligned} \tag{3.8}$$

where $y_i^+ = max\{y_i, 0\}$, $y^- + i = max\{-y_i, 0\}$. So here we are able to derive an analytical formula for g, which is a concave piecewise-linear function.

3.7.4 Numerical examples

We consider the problem of minimizing a strictly convex quadratic function over the unit box:

$$\begin{aligned} &\textit{minimize } (1/2)x^T Px - q^T x \\ &\textit{subject to } x_i^2 \le 1,\ i = 1, ..., n, \end{aligned}$$

where $P \succ 0$. The Lagrangian is

$$\begin{aligned} L(x, \lambda) &= (1/2)x^T px - q^T x + \lambda_i(x_i^2 - 1) \\ &= (1/2)x^T(P + diag(2\lambda))x - q^T x - \mathbf{1}^T\lambda, \end{aligned}$$

By the property of quadratic function, we have $x^\star = \frac{q}{p+diag(2\lambda^{(k)})}$. Then we can have

$$g(\lambda) = -(1/2)q^T x^\star - 1^T\lambda.$$

The projected subgradient algorithm for the dual is

$$x^{(k)} = (P + diag(2\lambda))^{-1}q,\ \lambda_i^{k+1} = (\lambda_i^{(k)} + \alpha_k((x_i^k)^2 - 1))_+.$$

The dual function is differentible, so we can use a fixed size α (provided it is small enough).

The iterates $x^{(k)}$ are not feasible. But we can construct a nearby feasible $\hat{x}^{(k)}$ as

$$\hat{x}_i^{(k)} = \begin{cases} 1, & x_i^{(k)} > 1 \\ -1, & x_i^{(k)} < -1 \\ x_i^{(k)}, & -1 \le x_i^k \le 1. \end{cases}$$

We consider an instance with $n = 50$. We start the algorithm with $\lambda^{(1)} = 1$, and use a fixed step size $\alpha = 0.1$. Figure shows the convergence of $g(\lambda^{(k)})$ (a lower bound on the optimal value) and $f_0(\hat{x}^{(k)})$(an upper bound on the optimal value), versus iterations.

3.8 Theorems of alternatives

3.8.1 Weak alternatives via the dual function

In this section we apply Lagrange duality theory to the problem of determining feasibility of a system of inequalities and equalities

$$f_i(x) \le 0,\ i = 1, ..., m\ h_i(x) = 0,\ i = 1, ..., p. \tag{3.9}$$

We assume the domain of the inequality system, $D = (\cap_{i=1}^m domf_i) \cap (\cap_{i=1}^p domh_i)$, is nonempty. We can think of the system of inequalities and equalities as the standard problem, with objective $f_0 = 0$, *i.e.*,

$$\begin{aligned} &minimize\ 0 \\ &subject\ to\ f_i(x) \le 0,\ i = 1, ..., m \\ &\qquad\qquad h_i(x) = 0,\ i = 1, ..., p. \end{aligned} \tag{3.10}$$

This problem has optimal value

$$p^\star = \begin{cases} 0 & (3.10)\ is\ feasible \\ \infty & (3.10)\ is\ infeasible, \end{cases} \tag{3.11}$$

So solving the optimization problem(3.11) is the same as solving the inequality system(3.10).

The dual function

We associate with the inequality system (3.9) the dual function

$$g(\lambda, \nu) = \inf_{x \in D}(\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)),$$

which is the same as the dual function for the optimization problem(3.10). Since $f_0 = 0$, the dual function is positive homogeneous in (λ, ν) : For $\alpha > 0, g(\alpha\lambda, \alpha\nu) = \alpha g(\lambda, \nu)$. The dual problem associated with (3.10) is to maximize $g(\lambda, \nu)$ subject to $\lambda \succeq 0$. Since g is homogeneous, the optimal value of this dual problem is given by

$$d^\star = \begin{cases} \infty & \lambda \succeq 0, g(\lambda, \nu) > 0 \; is \; feasible \\ 0 & \lambda \succeq 0, g(\lambda, \nu) > 0 \; is \; infeasible, \end{cases} \tag{3.12}$$

Weak duality tells us that $d^\star \leq p^\star$. Combining this fact with (3.11) and (6.12) yields the following: If the equality system

$$\lambda \succeq 0, \; g(\lambda, \nu) > 0$$

is feasible (which means $d^\star = \infty$), then the inequality system (6.9) is infeasible (since we then have $p^\star = \infty$). Indeed, we can interpret any solution (λ, ν) of the inequalities(3.10) as a proof or certificate of infeasiblity of the system(3.9).

We can restate this implication in terms of feasibility of the original system: If the original inequality system (3.9) is feasible, then the inequality system (3.10) must be infeasible. We can interpret an x which satisfies (3.9) as a certificate establishing infeasibility of teh inequality system(3.12).

Two systems of inequalities (and equalities) are called *weak alternatives* is at most one of the two is feasible. Thus the system (3.9) and (3.12) are *weak alternatives*. This is true whether or not the inequalities (3.9) are convex. moreover, the alternative inequality system (3.12) is always convex (*i.e.*, g is concave and the constraints $\lambda_i \geq 0$ are convex).

Strict inequalities

We can also study feasiblity of the strict inequality system

$$f_i(x) < 0, \; i = 1, ..., m, \; h_i(x) = 0, i = 1, ...p. \tag{3.13}$$

with g defined as for the nonstrict inequality system, we have the alternative inequality system

$$\lambda \succeq 0, \; \lambda \neq 0, \; g(\lambda, \nu) \geq 0. \tag{3.14}$$

We can show directly that (3.13) and (3.14) are weak alternatives. Suppose there exists an $\widetilde{x}$ with $f_i(\widetilde{x}) < 0$, $h_i(\widetilde{x} = 0$. Then for any $\lambda \succeq 0$, $\lambda \neq 0$, and ν,

$$\lambda_1 f_1(\widetilde{x}) + \cdots + \lambda_m f_m(\widetilde{x}) + \nu_i h_i(\widetilde{x}) + \cdots v_p h_p(\widetilde{x}) < 0$$

It follows that

$$\begin{aligned} g(\lambda, \nu) &= \underset{x \in D}{inf}(\sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x))) \\ &\leq \sum_{i=1}^{m} \lambda_i f_i(\widetilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\widetilde{x}) \\ &< 0. \end{aligned}$$

Therefore, feasible of (3.13) implies that there does not exist (λ, ν) satisfying (3.14).

Thus, we can prove infeasiblity of (3.13) by producing a solution of the system (3.14); we can prove infeasibility of (3.14) by producing a solution of the system(3.13).

3.8.2 Strong alternatives

When the original inequality is convex, *i.e.*, f_i are convex and h_i are affine, and some type of constraint qualification holds, then the pairs of weak alternatives described above are strong alternatives, which means that exactly one of the two alternatives holds. In other words, each of the inequality systems is feasible if and only if the other is infeasible.

In this subsection we assume that f_i are convex and h_i are affine, so the inequality system can be expressed as

$$f_i(x) \leq 0,\ i = 1, ..., m,\ Ax = b,$$

where $A \in R^{p \times n}$.

Strict inequalities

We first study the strict inequality system

$$f_i(x) \leq 0,\ i = 1, ..., m,\ Ax = b, \tag{3.15}$$

and its alternative

$$\lambda \succeq 0,\ \lambda \neq 0,\ g(\lambda, \nu) \geq 0. \tag{3.16}$$

We need one technical condition: There exists an $x \in relintD$ with $Ax = b$. In other words we not only assume that the linear equality constraints are consistent, but also that they have a solution in relint D. (Very often $D = R^n$, so the condition is satisfied if the equality constraints are consistent.) Under this condition, exactly one of the inequality systems (3.15) and (3.16) is feasible. In other words, the inequality systems (3.15) and (3.16) are strong alternatives.

We will establish this result by considering the related optimization problem

$$\begin{array}{l} minimize\ s \\ subject\ to\ f_i(x) - s \leq 0,\ i = 1, ..., m \\ \qquad\qquad Ax = b \end{array} \tag{3.17}$$

with variables x, s, and domain $D \times R$. The optimal value $p^\star$ of this problem is negative if and only if there exists a solution to the strict inequality system(3.15).

The Lagrange dual function for the problem (3.17) is

$$\inf_{x \in D, s} (s + \sum_{i=1}^{m} \lambda_i (f_i(x) - s) + \nu^T (Ax - b)) = \begin{cases} g(\lambda, \nu) & 1^T \lambda = 1 \\ -\infty & otherwise. \end{cases}$$

Therefore we can express the dual problem of (3.17) as

$$\begin{array}{l} maximize\ g(\lambda, \nu) \\ subject\ to\ \lambda \succeq 0,\ 1^T \lambda = 1. \end{array}$$

Now we observe that Slater's condition holds for the problem (3.17). By the hypoothesis there exists an $\widetilde{x} \in relintD$ with $A\widetilde{x} = b$. Choosing any $\widetilde{s} > max_i f_i(\widetilde{x})$ yields a point $(\widetilde{\lambda}, \widetilde{s})$

which is strictly feasible for (3.17). Therefore we have $d^\star = p^\star$, and the dual optimum $d^\star$ is attained. In other words, there exist $(\lambda^\star, \nu^\star)$ such that

$$g(\lambda^\star, \nu^\star) = p^\star, \ \lambda^\star \succeq 0, \ 1^T\lambda^\star = 1. \tag{3.18}$$

Now suppose that the strict inequality system (3.15) is infeasible, which means that $p^\star \geq 0$. Then $(\lambda^\star, \nu^\star)$ from (3.18) satisfy the alternate inequality system (3.16). Similarly, if the alternate inequality system (3.16) is feasible, then $d^\star = p^\star \geq 0$, which shows that the strict inequality system (3.15) is infeasible. Thus, the inequality systems (3.15) and (3.16) are strong alternatives; each is feasible if and only if the other is not.

Nonstrict inequalities

We now consider the nonstrict inequality system

$$f_i(x) \leq 0, \ i = 1, .., m, \ Ax = b, \tag{3.19}$$

and its alternative

$$\lambda \succeq 0, \ g(\lambda, \nu) > 0. \tag{3.20}$$

We will show these are strong alternatives, provided the following conditions hold: There exists an $x \in relintD$ with $Ax = b$, and the optimal value $p^\star$ of (3.17) is attained. This holds, for example, if $D = R^n$ and $max_i f_i(x) \to \infty$ as $x \to \infty$. With these assumptions we have, as in the strict case, that $p^\star = d^\star$, and that both the primal and dual optimal values are attained. Now suppose that the nonstrict inequality system (3.20) is infeasible, which means that $p^\star > 0$. (Here we use the assumption that the primal optimal value is attained.) Then $(\lambda^\star, \nu^\star)$ from (3.18) satisfy the alternate inequality system (3.20). Thus, the inequality system (3.19) and (3.20) are strong alternatives; each is feasible if and only if the other is not.

3.8.3 Example

Linear inequalities

Consider the system of linear inequalities $Ax \preceq b$. The dual function is

$$g(\lambda) = \inf_x \lambda^T(Ax - b) = \begin{cases} -b^T\lambda & A^T\lambda = 0 \\ -\infty & otherwise. \end{cases}$$

The alternative inequality system is therefore

$$\lambda \succeq 0, \ A^T\lambda = 0, \ b^T\lambda < 0.$$

These are, in fact, strong alternatives. This follows since the optimum in the related problem (6.18) is achieved, unless it is unbounded below.

We now consider the system of strict linear inequalities $Ax \preceq b$, which has the strong alternative system

$$\lambda \subseteq 0, \ \lambda \neq 0, \ A^T\lambda = 0, \ b^T\lambda \leq 0.$$

Farka's lemma In this section we describe a pair of strong alternatives for a mixture of strict and nonstrict linear inequalities, known as Farkas's lemma: The system of inequalities

$$Ax \preceq 0, \ c^T x < 0,$$

where $A \in R^{m\times n}$ and $c \in R^n$, and the system of equalities and inequalities

$$A^T y + c = 0,\ y \succeq 0,$$

are strong alternatives.

we can prove Farkas's lemma directly, using LP duality. Consider the LP

$$\begin{array}{l} minimize\ c^T x \\ subject\ to\ Ax \preceq 0, \end{array}$$

and its dual

$$\begin{array}{c} maximize 0 \\ subject\ to\ A^T y + c = 0 \\ y \succeq 0. \end{array}$$

The primal LP is homogeneous, and so has optimal value 0, if it is not feasible, and optimal value $-\infty$, if it is feasible. The dual LP has optimal value 0, if is feasible, and optimal value is $-\infty$, if it is infeasible.

Since $x = 0$ is feasible, we can rule out the one case in which strong duality can fail for LPs, so we must have $p^\star = d^\star$.

Example *Arbitrage-free bounds on price.* We consider a set of n assets, with prices at the beginning of an investment period $p_1, ..., p_n$, respectively. At the end of the investment period, the value of the assets is $v_1, ..., v_n$. If $x_1, ..., x_n$ represents the initial investment is $p^T x$, and the final value of the investment is $v^T x$.

The value of the assets at the end of the investment period, v, is uncertain. We will assume that only m possible scenarios, or outcomes, are possible. If outcome i occurs, the final value of the assets is $v^{(i)}$, and therefore, the overall value of the investments is $v^{(i)T} x$.

If there is an investment vector x with $p^T x < 0$, and in all possible scenarios, the final value is nonnegative, *i.e.*, $v^{(i)T} x \geq 0$ for $i = 1, ..., m$, then an *arbitrage* is said to exist. The condition $p^T x < 0$ means you are paid to accept the investment mix, and the condition $v^{(i)T} x \geq 0$ for $i = 1, ..., m$ means that no matter what outcome occurs, the final value is nonnegative, so an arbitrage corresponds to a guaranteed money-making investment strategy. It is generally assumed that the prices and values are such that no arbitrage exists. This means that the inequality system

$$Vx \succeq 0,\ p^T x < 0$$

is feasible, where $V_{ij} = v_j^{(i)}$.

Using Farkas' lemma, we have no arbitrage if and only if there exists y such that

$$-V^T y + p = 0,\ y \succeq 0.$$

4 Convex constrained problems

4.1 Equality Constraints

In this section, we consider problems with equality constraints of the form

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ h_i(x) = 0,\ i = 1, ..., m. \end{aligned}$$

We assume that $f : R^n \to R, h_i : R^n \to R, i = 1, ..., m$, are continuously differentiable functions.

Our basic Lagrange multiplier theorem states that for a given local minimum x^*, there exist scalars $\lambda_1, ..., \lambda_m$, called Lagrange multipliers, such that

$$\bigtriangledown f(x^*) + \sum_{i=1}^{m} \lambda_i \bigtriangledown h_i(x^*) = 0.$$

There are two ways to interpret this equation: The cost gradient $\bigtriangledown f(x^*)$ belongs to the subspace spanned by the constraint gradients at x^*.
The cost gradient $\bigtriangledown f(x^*)$ is orthogonal to the subspace of first order feasible variations

$$V(x^*) = \{\Delta x | \bigtriangledown h_i(x^*)'\Delta x = 0, i = 1, ..., m\}.$$

This is the subspace of variations Δx for which the vector $x = x^* + \Delta x$ satisfies the constraint $h(x) = 0$ up to first order. Thus, according to the Lagrange multiplier condition, at the local minimum x^*, the first order cost variation $\bigtriangledown f(x^*)'\Delta x$ is zero for all variations Δx in this subspace. This statement is analogous to the zero gradient condition $\bigtriangledown f(x^*) = 0$ of constrained optimization.

Theorem 4.1 *Lagrange Multiplier Theorem (necessary condition)*
Let x^* be a local minimum of f subject to $h(x) = 0$, and assume that x^* is regular. Then there exists a unique vector $\lambda^* = (\lambda_1^*, ..., \lambda_m^*)$, called a Lagrange multiplier vector, such that

$$\bigtriangledown f(x^*) + \sum_{i=1}^{m} \lambda_i \bigtriangledown h_i(x^*) = 0.$$

If in addition f and h are twice continuously differentiable, we have

$$y'(\bigtriangledown^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \bigtriangledown^2 h_i(x^*))y \geq 0,\ forally \in V(x^*),$$

where $V(x^*)$ is the subspace of first order feasible variations

$$V(x^*) = \{\Delta x | \bigtriangledown h_i(x^*)'\Delta x = 0, i = 1, ..., m\}.$$

Some times it is convenient to write our necessary conditions in terms of the Lagrangian function $L : R^{n+m} \to R$ defined by

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x).$$

Then if x^* is a local minimum that is regular, the Lagrange multiplier condition of previous proposition together with the equation $h(x^*) = 0$ are written compactly as

$$\begin{aligned} \bigtriangledown_x L(x^*, \lambda^*) &= 0, \\ \bigtriangledown_\lambda L(x^*, \lambda^*) &= 0, \\ y' \bigtriangledown^2_{xx} L(x^*, \lambda^*) y &\geq 0, \ for\ all\ y \in V(x^*) \end{aligned}$$

The first order necessary condition may be satisfied by both local minima and local maxima (and possible other vectors). The second order necessary condition is useful in narrowing down the field of candidates for local minima. To guarantee that a given vector is a local minimum, we need sufficient conditions for optimality, which are given by the following proposition.

Theorem 4.2 *(Second Order Sufficiency Conditions)*
Assume that f and h are twice continuously differentiable, and let $x^* \in R^n$ and $\lambda^* \in R^m$ satisfy

$$\begin{aligned} \bigtriangledown_x L(x^*, \lambda^*) &= 0, \\ \bigtriangledown_\lambda L(x^*, \lambda^*) &= 0, \\ y' \bigtriangledown^2_{xx} L(x^*, \lambda^*) y &> 0, \ for\ all\ y \neq 0\ with\ \bigtriangledown h(x^*)' y = 0. \end{aligned}$$

Then x^* is a strict local minimum of f subject to $h(x) = 0$. In fact, there exist scalars $\gamma > 0$ and $\epsilon > 0$ such that

$$f(x) \geq f(x^*) + \frac{\gamma}{2} \|x - x^*\|^2, \ \forall x\ with\ h(x) = 0\ and\ \|x - x^*\| < \epsilon.$$

Note that the above sufficient conditions do not include regularity of the vector x^*.

Theorem 4.3 *(Sensitivity Theorem)*
Let x^* and λ^* be a local minimum and Lagrange multiplier, respectively, satisfying the second order sufficiency condition, and assume that x^* is a regular point. Consider the family of problems

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ h(x) = u, \end{aligned}$$

parameterized by the vector $u \in R^m$, Then there exists an open sphere S centered at $u = 0$ such that for every $u \in S$, there is an $x(u) \in R^n$ and a $\lambda(u) \in R^m$, which are a local minimum-Lagrange multiplier pair of the problem. Furthermore, $x(\dots)$ and $\lambda(\dots)$ are continuously differentiable functions within S and we have

$$x(0) = x^*, \ \lambda(0) = \lambda^*.$$

In addition, for all $u \in S$, we have

$$\bigtriangledown p(u) = -\lambda(u),$$

where $p(u)$ is the optimal cost parameterized by u, i.e.,

$$p(u) = f(x(u)).$$

There are multiple constraints $a_i'x = b_i, i = 1, ..., m$ the preceding argument can be appropriately modified. In particular, we have

$$\begin{aligned} \Delta cost &= f(x^* + \Delta x) - f(x^*) \\ &= \bigtriangledown f(x^*)'\Delta x + o(\|x\|) \\ &= -\sum_{i=1}^{m} \lambda_i^* a_i' \Delta x + o(\|\Delta x\|), \end{aligned}$$

and $a_i'\Delta x = \Delta b_i$ for all i, so we obtain $\Delta cost = -\sum_{i=1}^{m} \lambda_i^* \Delta b_i + o(\|\Delta x\|)$.

4.2 Ineuqlity Constraints

We now consider the following problem, which involves both equality and inequality constraints:

$$\begin{aligned} &minimize f(x) \\ &subject to h_1(x) = 0, ..., h_m(x) = 0, \\ &\qquad g_1(x) \leq 0, ..., g_r(x) \leq 0, \end{aligned}$$

where f, h_i, g_i are continuously differentiable function from R^n to R. We will first use a simple approach to this problem that relies on the theory for equality constraints of the preceding sections. For any feasible point x, the set of active inequality constraints is denoted by

$$A(x) = \{j | g_j(x) = 0\}$$

If $j \notin A(x)$, we say that the jth constraint is inactive at x. We note that if x^* is a local minimum of the inequality constrained problem, then x^* is also a local minimum for a problem identical to inequality constrained problem except that the inactive constraints at x^* have been discarded. Thus, in effect, inactive constraints at x^* don't matter; they can be ignored in the statement of optimality conditions.

On the other hand, at a local minimum, active inequality constraints can be treated to a large extent as equalities. In particular, if x^* is a local minimum of the inequality constrained problem, then x^* is also a local minimum for the equality constrained problem

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ h_1(x) = 0, ..., h_m = 0,\ g_j(x) = 0, \forall j \in A(x^*). \end{aligned}$$

Thus, if x^* is regular for the latter problem, there exist Lagrange multipliers $\lambda_1^*, ..., \lambda_m^*$, and $\mu_j^*, j \in A(x^*)$ such that

$$\bigtriangledown f(x^*) + \sum_{i=1}^{m} \lambda_i^* \bigtriangledown h_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \bigtriangledown g_j(x^*) = 0$$

Assigning zero Lagrange multipliers to the inactive constraints, we obtain

$$\bigtriangledown f(x^*) + \sum_{i=1}^{m} \lambda_i^* \bigtriangledown h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \bigtriangledown g_j(x^*) = 0, \ \mu_j^* = 0, \ \forall j \notin A(x^*),$$

which can be viewed as an analog of the first order optimality condition for the equality constrained problem. There is one more important fact about the Lagrange multipliers μ_j^*: they are nonnegative.

Theorem 4.4 *Karush-Kuhn-Tucker Necessary Conditions*
Let x^* be a local minimum of problem (ICP), and assume that x^* is regular. Then there exist unique Lagrange multiplier vectors $\lambda^* = (\lambda_1^*, ..., \lambda_m^*), \mu^* = (\mu_1^*, ..., \mu_r^*)$, such that

$$\begin{aligned} \bigtriangledown_x L(x^*, \lambda^*, \mu^*) = 0, \\ \mu_j^* \geq 0, \ j = 1, ..., r, \\ \mu_j^* = 0, \ \forall j \notin A(x^*), \end{aligned}$$

where $A(x^*)$ is the set of active constraints at x^*. If in addition f, h and g are twice continuously differentiable, there holds

$$y' \bigtriangledown_{xx}^2 L(x^*, \lambda^*, \mu^*) y \geq 0,$$

for all $y \in R^n$ such that

$$\bigtriangledown h_i(x^*)' y = 0, \forall i = 1, ..., m, \ \bigtriangledown g_i(x^*)' y = 0, \ \forall j \in A(x^*).$$

Our approach for using necessary conditions to solve inequality constrained problems is to consider separately all the possible combinations of constraints being active or inactive.

Theorem 4.5 *Second Order Sufficiency Conditions*
Consider problem(ICP), assume that f, h and g are twice continuously differentiable, and let $x^* \in R^n, \lambda^* \in R^m$, and $\mu^* \in R^r$ satisfy

$$\begin{aligned} \bigtriangledown_x L(x^*, \lambda^*, \mu^*) = 0, \ h(x^*) = 0, \ g(x^*) \leq 0, \\ \mu_j^* \geq 0, \ j = 1, ..., r, \\ \mu_j^* = 0, \ \forall j \notin A(x^*), \\ y'(\bigtriangledown_{xx}^2 L(x^*, \lambda^*, \mu^*) y > 0, \end{aligned}$$

for all $y \neq 0$ such that

$$\bigtriangledown h_i(x^*)' y = 0, \ \forall i = 1, ..., m, \ \bigtriangledown g_i(x^*)' y = 0, \ \forall j \in A(x^*).$$

Assume also that

$$\mu_j^* > 0, \forall j \in A(x^*).$$

Then x^* is a strict local minimum of f subject to

$$h(x) = 0 \ g(x) \leq 0.$$

Theorem 4.6 *(Strengthened Karush-Kuhn-Tucker Conditions)*
Let x^* be a local minimum of problem (ICP), and assume that x^* is regular. Then there exist Lagrange multipliers $\lambda_1^*, ..., \lambda_m^*$ and $\mu_1^*, ..., \mu_r^*$, satisfying the following conditions:
(i) We have

$$\bigtriangledown f(x^*) + \sum_{i=1}^{m} \lambda_i^* \bigtriangledown h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \bigtriangledown g_j(x^*) = 0.$$

(ii) $\lambda_j^* \geq 0$ for all $j = 1, ..., r$
(iii) In every neighborhood N of x^* there is an $x \in N$ such that $\lambda_i^* h_i(x) > 0$ for all i with $\lambda_i^* \neq 0$ and $\mu_i^* \neq 0$ and $\mu_j^* g_j(x) > 0$ for all j with $\mu_j^* \neq 0$. Moreover, if $(\lambda^*, \mu^*) \neq (0, 0)$ this x can be chosen so that $f(x) < f(x^*)$.

A sharper necessary and sufficient condition can be derived, involving minimization of a Lagrangian function. This minimization may involve any subset of the inequality constraints, while the remaining constraints are taken into account using Lagrange multipliers, are shown in the following proposition. The flexibility of assigning Lagrange multipliers to only some of the constraints, while dealing with the other constraints explicitly is often very useful.

Theorem 4.7 *Optimality Condition for Convex Cost and Linear Constraints*
Consider the problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } e_i'x = d_i, \ i = 1, ..., m, \\ &\qquad a_j'x \leq b_j, j = 1, ..., r, \end{aligned}$$

where e_i, a_j and d_i, b_j are given vectors and scalars, respectively, and $f : R^n \to R$ is convex and continuously differentiable. Let I be a subset of the index set $\{1, ..., m\}$, and J be a subset of the index set $\{1, ..., r\}$. Then x^* is a global minimum if and only if x^* is feasible and there exist scalars $\lambda_i^*, i \in I$, and $\mu_j^*, j \in J$, such that

$$\begin{aligned} &\mu_j^* \geq 0, \ j \in J, \\ &\mu_j^* = 0, \ \forall j \in J \text{ with } j \notin A(x^*), \\ &x^* \in \underset{e_i'x = d_i, \ i \notin I; a_j'x \leq b_j, \ j \notin J}{argmin} \{f(x) + \sum_{i \in I} \lambda_i^*(e_i'x - d_i) + \sum_{j \in J} \mu_j^*(a_j'x - b_j)\}. \end{aligned}$$

4.3 Duality Theory

A simpele form for linear constraints, consider the problem

$$\begin{aligned} &minimize f(x) \\ &subject to e_i'x = d_i,\ i = 1, ..., m, \\ &\qquad a_j'x \leq b_j,\ j = 1, ..., r,\ x \in X, \end{aligned}$$

here e_i, a_j and d_i, b_j are given vectors and scalars, respectively, and $f : R^n \to R$ is convex and continuously differentiable function, and X is a polyhedral set, *i.e.*, a set specified by a finite collection of linear equality and inequality constraints. We refer to problem(P) as the primal problem.

Define the Lagrangian function

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i (e_i'x - d_i) + \sum_{j=1}^{r} \mu_j (a_j'x - b_j).$$

Consider also the dual function defined by

$$q(\lambda, \mu) = \inf_{x \in X} L(x, \lambda, \mu).$$

The dual problem is

$$\begin{aligned} &maximize\ q(\lambda, \mu) \\ &subject\ to\ \lambda \in R^m, \mu \geq 0. \end{aligned}$$

Note that if the polyhedral set X is bounded, then it is also compact(every polyhedral set is closed since it is the intersection of closed subspaces). Thus if X is bounded, the dual function takes real values. In general, however, $q(\lambda, \mu)$ can take the value $-\infty$. Thus in effect, the constraint set of the dual problem(D) is the set

$$Q = \{(\lambda, \mu) | \mu \geq 0, q(\lambda, \mu) > -\infty\}.$$

Theorem 4.8 *(Duality Theorem - Differentiable Convex Cost and Linear COnstraints)*
(a) If the primal problem(P) has an optimal solution, the dual problem(D) also has an optimal solution and the corresponding optimal values are equal.
(b) In order for x^* to be an optimal primal solution and (λ^*, μ^*) to be an optimal dual solution, it is necessary and sufficient that x^* is primal feasible, $\mu^* \geq 0, \mu_j^* = 0$ for all $j \neq A(x^*)$, and

$$x^* \in arg \min_{x \in X} L(x, \lambda^*, \mu^*).$$

Geometric Multipliers
A vector $\mu^* = (\mu_1^*, ..., \mu_r^*)$ is said to be a geometric multiplier vector (or simply geometric multiplier) for the primal problem if

$$\mu_j^* \geq 0, j = 1, ..., r,$$

and

$$f^* = \inf_{x \in X} L(x, \mu^*).$$

As indicated by the preceding discussion, there is a connection between geometric and Lagrange multipliers in the case where X, f, and g_j are convex, and f and g_j are continuously differentiable. Then, it can be shown that given an optimal solution x^*, the set of Lagrange multipliers corresponding to x^* is equal to the set of geometric multipliers. However, even under convexity assumptions, it is possible that the problem has no optimal solution and hence no Lagrange multipliers, while the set of geometric multipliers may be nonempty.

To visualize the definition of a geometric multiplier, as well as other concepts related to duality, it is useful to consider hyperplanes in the space of constraint-cost pairs $(g(x), f(x))$ (viewed as vectors in R^{r+1}).

The Weak Duality Theorem

We introduce the dual function q, which is defined for $\mu \in R^r$ by

$$q(\mu) = \inf_{x \in X} L(x, \mu).$$

The dual problem is

$$\begin{aligned} &\textit{maximize } q(\mu) \\ &\textit{subject to } \mu \geq 0, \end{aligned}$$

and corresponds to finding the maximum point of interception, over all hyperplanes with normal $(\mu, 1)$ where $\mu \geq 0$.

Note that $q(\mu)$ may be equal to $-\infty$ for some u. In this case, we effectively have the additional constraint $\mu \in D$ in the dual problem, where D, called the domain of q, is the set of μ for which $q(\mu)$ is finite:

$$D = \{\mu | q(\mu) > -\infty\}.$$

In fact, we may have $q(\mu) = -\infty$ for all $\mu \geq 0$, in which case the dual optimal value

$$q^* = \sup_{\mu \geq 0} q(\mu),$$

is equal to $-\infty$.

Regardless of the structure of the cost and constraints of the primal problem, the dual problem has nice convexity properties, as shown by the following proposition.

Proposition 4.1 *The domain D of the dual function q is convex and q is concave over D.*

Proof For any $x, \mu, \bar{\mu}$, and $\alpha \in [0, 1]$, we have

$$L(x, \alpha\mu + (1 - \alpha)\bar{\mu}) = \alpha L(x, \mu) + (1 - \alpha) L(x, \bar{\mu}).$$

Taking the infimum over all $x \in X$, we obtain

$$q(\alpha\mu + (1-\alpha)\bar{\mu}) \geq \alpha q(\mu) + (1-\alpha)q(\bar{\mu}).$$

Therefore if μ and $\bar{\mu}$ belong to D, the same is true for $\alpha\mu + (1-\alpha)\bar{\mu}$, so D is convex. Furthermore, q is concave over D. **Q.E.D.**

The concavity of q can also be verified by observing that q is defined as the infimum over $x \in X$ of the collection of the concave functions $L(x, \cdot)$;

Another important property is that the optimal dual value is always an underestimate of the optimal primal value.

Theorem 4.9 *Weak Duality Theorem*
We have

$$q^* \leq f^*.$$

Proof For all $\mu \geq 0$, and $x \in X$ with $g(x) \leq 0$, we have

$$q(\mu) = \inf_{z \in X} L(z, \mu) \leq f(x) + \sum_{j=1}^{r} \mu_j g_j(x) \leq f(x),$$

so

$$q^* = \sup_{\mu \geq 0} q(\mu) \leq \inf_{x \in X, g(x) \leq 0} f(x) = f^*.$$

Q.E.D.

If $q^* = f^*$ we say that there is no duality gap and if $q^* < f^*$ we say that there is a duality gap. Note that if there exists a geometric multiplier μ^*, the weak duality theorem ($q^* \leq f^*$) and the definition of a geometric multiplier [$f* = q(\mu^*) \leq q^*$] imply that there is no duality gap. However, the converse is not true. In particular, it is possible that no geometric multiplier exists even though there is no duality gap. In this case the dual problem does not have an optimal solution, as implied by the following proposition.

Proposition 4.2 *(a) If there is no duality gap, the set of geometric multipliers is equal to the set of optimal dual solutions.*
(b) If there is a duality gap, the set of geometric multipliers is empty.

Duality theory is most useful when there is no duality gap. To guarantee that there is no duality gap and that a geometric multiplier exists, it is typically necessary to impose various types of convexity conditions on the cost and the constraints of the primal problem.

One is interested in find lower bounds that are as tight as possible, so the usual approach is to start with some dual feasible solution and iteratively improve it by using some algorithm. A major difficulty here is that the dual function $q(\mu)$ is typically nondifferentiable, so the methods developed so far cannot be used. We will develop special methods for optimization of nondifferentiable cost functions.

Primal and Dual Optimal Solutions

There are powerful characterizations of primal and dual optimal solution pairs, given in the following two propositions. Note, however, that these characterizations are useful only if there is no duality gap, since otherwise there is no geometric multiplier, even if the dual problem has an optimal solution.

Proposition 4.3 *(Optimality Conditions) A pair (x^*, μ^*) is an optimal solution-geometric multiplier pair if and only if*

$$x^* \in X,\ g(x^*) \leq 0,\ (PrimalFeasibility),$$
$$\mu^* \geq 0,\ (DualFeasibility),$$
$$x^* \in arg\min_{x \in X} L(x, \mu^*),\ (LagrangianOptimality)$$
$$\mu_j^* g_j(x^*) = 0,\ j = 1, ..., r,\ (ComplementarySlackness).$$

Proof: If (x^*, μ^*) is an optimal solution-geometric multiplier pair, then x^* is primal feasible and μ^* is dual feasible. Conversely, using the equations, we obtain

$$f^* \leq f(x^*) = L(x^*, \mu^*) = \min_{x \in X} L(x, \mu^*) = q(\mu^*) \leq q^*.$$

Using the weak duality, we see that equality holds throughout in the preceding relation. It follows that x^* is primal optimal and μ^* is dual optimal, while there is no duality gap. **Q.E.D.**

Theorem 4.10 *Saddle Point Theorem*
A pair (x^*, μ^*) is an optimal solution-geometric multiplier pair if and only if $x^* \in X$, $\mu^* \geq 0$, and (x^*, μ^*) is a saddle point of the Lagrangian, in the sense that

$$L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*),\ \forall x \in X, \mu \geq 0$$

Mathematical programming duality is a broadly applicable subject with a rich theory.
(a). *Conditions under which there is no duality gap and there exists a geometric multiplier.* Convexity of the cost function and the constraints is essentially a prerequisite for this. The constraints $g_j(x) \leq 0$ are linear and the constraint set X is polyhedral.

Theorem 4.11 *Strong Duality Theorem - Linear Constraints*
Let the cost function f be convex over R^n and let the set X be polyhedral. Assume that the optimal value f^* is finite. Then there is no duality gap and there exists at least one geometric multiplier.

Note that convexity of f over X is not enough for Strong Duality Theorem holds, it is essential that f be convex over the entire space R^n, as the following example shows. Consider the two-dimensional problem

$$minimize\ e^{-\sqrt{x_1 x_2}},\ \forall x \in X,$$
$$subject\ to\ x_1 = 0,\ x \in X = \{x | x \geq 0\},$$

and $f(x)$ is arbitrarily defined for $x \notin X$. Here it can be verified that f is convex over X(its Hessian is positive definite in the interior of X). Since for feasiblity, we must have $x_1 = 0$, we see that $f^* = 1$. On the other hand, for all $\mu \geq 0$ we have

$$q(\mu) = \inf_{x \geq 0} \{e^{-\sqrt{x_1 x_2}} + \mu x_1\} = 0,$$

since the expression in braces is nonnegative for $x \geq 0$ and can approach zero by taking $x_1 \to 0$ and $x_1 x_2 \to \infty$, It follows that $q^* = 0$. Thus, there is a duality gap, $f^* - q^* = 1$. The difficulty here is that $f(x)$ is not defined as a convex function over R^2.

Proposition 4.4 *(Linear and Quadratic Programming Duality)*
Let the cost function f be convex quadratic and let the set X be polyhedral. Assume that the optimal value f^* is finite. Then, the primal and dual problem have optimal solutions, and there is no duality gap.

If C is a convex subset of R^n, and $aff(C)$ is the affine hull of C (the intersection of all linear manifolds containing C), the relative interior of C, denoted $ri(C)$, is the set of all $x \in C$ for which there exists an $\epsilon > 0$ such that all $z \in aff(C)$ with $\|z - x\| < \epsilon$ are contained in C, i,e., $ri(C)$ is the interior of C relative to $aff(C)$.Every nonempty convex set has a nonempty relative interior.

Proposition 4.5 *(Nonlinear Farkas' Lemma) Let C be a convex subset of R^n and let $F : C \to R$ be a function that is convex over C. Let also $a_j \in R^n$ and $b_j \in R, j = 1, ..., r$, be given vectors and scalars, respectively. Assume that the set*

$$S = \{x | a_j' x \leq b_j, j = 1, ..., r\}$$

contains a vector in the relative interior of C, and that

$$F(x) \geq 0, \ \forall x \in S \cap C.$$

Then, there exist scalars $\mu_j \geq 0, j = 1, ..., r$, such that

$$F(x) + \sum_{j=1}^{r} \mu_j (a_j' x - b_j) \geq 0, \ \forall x \in C.$$

Proof of Strong Duality Theorem:
Without loss of generality, we assume that there are no equality constraints, so we are dealing with the problem

$$\begin{aligned} &\textit{minimize } f(x) \\ &\textit{subject to } x \in X, \ a_j' - b_j \leq 0, j = 1, ..., r, \end{aligned}$$

Let X be expressed in terms of linear inequalities as

$$X = \{x | a_j' - b_j \leq 0, j = r + 1, ..., p\},$$

where p is an integer with $p > r$. By applying the nonlinear Farkas' lemma with

$$C = R^n,\ S = \{x | a_j' x - b_j \leq 0, i = 1, ..., p\},\ F(x) = f(x) - f^*,$$

we see that there exist $\mu_1, ..., \mu_p$ with $\mu_j \geq 0$ for all j, such that

$$f^* \leq f(x) + \sum_{j=1}^{p} \mu_j (a_j' x - b_j),\ \forall x \in R^n.$$

Moreover for $x \in X$, we have

$$\mu_j (a_j' x - b_j) \leq 0,\ \forall j = r+1, ..., p,$$

so the preceding two relations yield

$$f^* \leq f(x) + \sum_{j=1}^{r} \mu_j (a_j' x - b_j),\ \forall x \in X,$$

from which

$$f^* \leq inf\, x \in X L(x, \mu) = q(\mu) \leq q^*.$$

Using the weak duality theorem, it follows that μ is a geometric multiplier and that there is no duality gap. Q.E.D.

The constraints $g_j(x) \leq 0$ are convex and possess a common interior point $\bar{x} \in X$, satisfying $g_j(\bar{x}) < 0$ for all j. We now consider the nonlinearly constrained problem*

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ x \in X,\ g_j(x) \leq 0,\ j = 1, ..., r, \end{aligned}$$

under convexity assumptions. In particular, we will assume the following.
(Convexity and Interior Point) The problem is feasible and its optimal value f^* is finite. Furthermore, the set X is a convex subset of R^n and the functions $f : R^n \to R$, $g_j : R^n \to R$ are convex over X. In addition, there exists a vector $\bar{x} \in X$ such that

$$g_j(\bar{x}) < 0,\ \forall j = 1, ..., r.$$

Theorem 4.12 *(Strong Duality Theorem - Inequality Constraints)*
Let the previous assumption for problem hold, then there is no duality gap and there exists at least one geometric multiplier.*

Mixed Convex and Linear Constraints

$$\begin{aligned} &minimize\ f(x) \\ &subject\ to\ x \in X,\ e_i' x - d_i = 0,\ i = 1, ..., m, \\ &\qquad g_j(x) \leq 0, j = 1, ..., \bar{r},\ a_j' x - b_j \leq 0, j = \bar{r}+1, ..., r. \end{aligned}$$

4.4 Penalty and Augmented Lagrangian Methods

The basic idea in penalty methods is to eliminate some or all of the constraints and add to the cost function a penalty term that prescribes a high cost of infeasible points. Associated with these methods is a penalty parameter c that determines the severity of the penalty and as a consequence, the extent to which the resulting unconstrained problem approximates the original constrained problem.As c takes higher values, the approximation becomes increasingly accurate. We focus attention primarily on the popular quadratic penalty function.

Consider first the equality constrained problem

$$\begin{aligned} &minimize f(x) \\ &subject\ to h(x) = 0,\ x \in X, \end{aligned}$$

where $f : R^n \to R, h : R^n \to R^m$ are given functions, and X is a given subset of R^n. Much of our analysis in this section will focus on the case where $X = R^n$, and x^* together with a Lagrange multiplier vector λ^* satisfies the sufficient optimality conditions. At the center of our development is the augmented Lagrangian function $L_c : R^n \times R^m \to R$,

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2,$$

where c is a positive penalty parameter.
There are two mechanisms by which unconstrained minimization of $L_c(\cdot, \lambda)$ can yield points closed to x^*:
(a) By taking λ close to λ^*. Indeed, if c is higher than a certain threshold, then for some $\lambda > 0$ and $\epsilon > 0$ we have

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2,\ \forall x\ with\ \|x - x^*\| < \epsilon,$$

so that x^* is a strict unconstrained local minimum of the augmented Lagrangian $L_c(\cdot, \lambda^*)$ corresponding to λ^*. This suggests that if λ is closed to λ^*, a good approximation to x^* can be found by unconstrained minimization of $L_c(\cdot, \lambda)$.
(b) By taking c large. Indeed for high c, there is high cost for infeasibility, so the unconstrained minima of $L_c(\cdot, \lambda)$ will be nearly feasible. Since $L_c(x, \lambda) = f(x)$ for feasible x, we expect that $L_c(x, \lambda) \approx f(x)$ for nearly feasible x Therefore, we can also expect to obtain a good approximation to x^* by unconstrained minimization of $L_c(\cdot, \lambda)$ when c is large.

A first update formula for λ^k in the quadratic penalty method is

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

If the generated sequence $\{x^K\}$ converges to a local minimum x^* that is regular, then $\{\lambda^k + c^k h(x^k)\}$ converges to the corresponding Lagrange multiplier λ^*.
Example: A Nonconvex problem
Consider the problem

$$\begin{aligned} &minimize\ \frac{1}{2}(-x_1^2 + x_2^2) \\ &subject\ to\ x_1 = 1 \end{aligned}$$

with optimal solution $x^* = (1, 0)$ and Lagrange multiplier $\lambda^* = 1$. The augmented Lagrangian is given by

$$L_c(x, \lambda) = \frac{1}{2}(-x_1^2 + x_2^2) + \lambda(x_1 - 1) = \frac{c}{2}(x_1 - 1)^2.$$

By differential, the vector x^k minimizing $L_{c^k}(x, \lambda^k)$ is given by

$$x^k = (\frac{c^k - \lambda^k}{c^k - 1}, 0).$$

For this formula to be correct, however, it is necessary that $c^k > 1$; for $c^k < 1$ the augmented Lagrangian has no minimum, and the same is true for $c^k = 1$ unless $\lambda^k = 1$. The multiplier updating formula can be written as

$$\lambda^{k+1} = \lambda^k + c^k(\frac{c^k - \lambda^k}{c^k - 1} - 1) = -\frac{\lambda^k}{c^k - 1} + \frac{c^k}{c^k - 1},$$

or by introducing the Lagrange multiplier $\lambda^* = 1$,

$$\lambda^{k+1} - \lambda^* = -\frac{\lambda^k - \lambda^*}{c^k - 1}.$$

From the iteration, it can be seen that similar conclusions to those of the preceding example can be drawn. In particular, it is not necessary to increase c^k to ∞ to obtain convergence, although doing so results in a better convergence rate. However, there is a difference: whereas in the preceding example, convergence was guaranteed for all positive sequence $\{c^k\}$, in the present example, the minimizing points exist only if $c^k > 1$. Hence, c^k plays a convexification role; once it exceeds the threshold value of 1 the penalty term convexifies the augmented Lagrangian, thus compensating for the nonconvexity of the cost function. Moreover, it can be seen that to obtain a convergence, the penalty parameter c^k must eventually exceed 2, so that the scalar

$$\frac{-1}{c^k - 1}$$

multiplying λ^k has absolute value less that one. The need for c^k exceed twice the value of the convexification threshold is a fundamental characteristic of multiplier methods when applied to nonconvex problems.

Computational Aspects - Choice of Parameters: In addition to addressing the problem of ill-conditioning, an important practical question in the method of multipliers is how to select the initial multiplier λ_0 and the penalty parameter sequence. Clearly, in view of the interpretations given earlier, any prior knowledge should be exploited to select λ^0 as close as possible to λ^*. The main considerations to be kept in mind for selecting the penalty parameter sequence are the following: (a) c^k should eventually become larger than the threshold level necessary to bring to bear the positive features of the multiplier iteration.
(b) The initial parameter c^0 should not be too large to the point where it causes ill-conditioning at the first unconstrained minimization.
(c) c^k should not be increased too fast to the point where too much ill-conditioning is forced upon the unconstrained minimization routine too early.
(d) c^k should not be increased too slowly, at least in the early minimizations, to the extent that the multiplier iterations has poor convergence rate.

4.5 Conjugate Method

In this section we consider the problem

$$\begin{aligned} &\textit{minimize } f_1(x) - f_2(x) \\ &\textit{subject to } x \in X_1 \cap X_2, \end{aligned}$$

where f_1 and f_2 are real-valued function on R^n, and X_1 and X_2 are subsets of R^n. We assume throughout that this problem has a feasible solution and a finite value, denoted f^*.

We will explore a classical form of duality, which is useful in many contexts. This duality was developed by W.Fenchel, a Danish mathematician who pioneered in the 40s the use of geometric convexity methods in optimizaton. One way to derive Fenchel duality is to convert problem to the following problem in the variables $y \in R^n$ and $z \in R^n$

$$\begin{aligned} &\textit{minimize } f_1(y) - f_2(z) \\ &\textit{subject to } z = y, y \in X_1,\ z \in X_2, \end{aligned}$$

and to dualize the constant $z = y$. The dual function is

$$\begin{aligned} q(\lambda) &= \inf_{y \in X_1, z \in X_2} \{f_1(y) - f_2(z) + (z - y)'\lambda\} \\ &= \inf_{z \in X_2} \{z'y - f_2(z)\} + \inf_{y \in X_1} \{f_1(y) - y'\lambda\} \end{aligned}$$

or, equivalently,

$$q(\lambda) = g_2(\lambda) - g_1(\lambda),$$

where the functions $g_1 : R^n \to (-\infty, \infty]$ and $g_2 : R^n \to [-\infty, \infty)$ are defined by

$$\begin{aligned} g_1(\lambda) &= \sup_{x \in X_1} \{x'\lambda - f_1(x)\}, \\ g_2(\lambda) &= \inf_{x \in X_2} \{x'\lambda - f_2(x)\}. \end{aligned}$$

The function g_1 is known as the conjugate convex function corresponding to the pair (f_1, X_1), while the function g_2 is known as the conjugate concave function corresponding to (f_2, X_2). It's straightforward to show the convexity of the sets Λ_1 and Λ_2 over which g_1 and g_2 are finite,

$$\lambda_1 = \{\lambda | g_1(\lambda) < \infty\},\ \lambda_2 = \{\lambda | g_2(\lambda) > -\infty\},$$

Furthermore, g_1 is convex over Λ_1 and g_2 is concave over Λ_2.

The dual problem is given by

$$\begin{aligned} &\textit{minimize } g_2(\lambda) - g_1(\lambda) \\ &\textit{subject to } \lambda \in \Lambda_1 \cap \Lambda_2. \end{aligned}$$

Proposition 4.6 *(Fenchel Optimality Conditions)*
A pair (x^*, λ^*) is an optimal solution pair for the problems if and only if

$$x^* \in X_1 \cap X_2, \ (primal\ feasibility),$$
$$\lambda^* \in \Lambda_1 \cap \Lambda_2, (dual\ feasibility),$$
$$x^* \in arg\max_{x \in X_1}\{x'\lambda^* - f_1(x)\},\ x^* \in arg\min_{x \in X_2}\{x'\lambda^* - f_2(x)\}, (Lagrangian\ optimality).$$

Since there is no duality gap, we need some convexity assumptions. The sets X_1, X_2 are convex. Furthermore, the function f_1 is convex over X_1, and the function f_2 is concave over X_2.

Proposition 4.7 *(Primal Fenchel Duality Theorem)*
Let assumption hold, then the dual problem has an optimal solution and there is no duality gap, i.e., we have

$$\inf_{x \in X_1 \cap X_2}\{f_1(x) - f_2(x)\} = \max_{\lambda \in \Lambda_1 \cap \Lambda_2}\{g_2(\lambda) - g_1(\lambda)\},$$

(1) The relative interiors of X_1 and X_2 have nonempty intersection.
(2) X_1 and X_2 are polyhedral and f_1 and f_2 are convex and concave over R^n, respectively, (rather than just over X_1 and X_2 respectively).

Proposition 4.8 *(Dual Fenchel Duality Theorem)*
Assume that the epigraphs $\{(x, w)|x \in X_1, f_1(x) \leq w\}$ and $\{(x, w)|x \in X_2, f_2(x) \geq w\}$ are closed convex sets. Then there exists an optimal primal solution and there is no duality gap if one of the following two conditions holds:
(1) The relative interiors of Λ_1 and Λ_2 have nonempty intersection.
(2) Λ_1 and Λ_2 are polyhedral, and g_1 and g_2 can be extended to real-valued convex and concave function, respectively, which are defined over the entire space R^n.

Note, that similar to the case of Fenchel Optimality Conditions, we have that (x^*, λ^*) are a primal and dual optimal solution pair if and only if the primal and dual feasibility conditions hold, together with the alternative Lagrangian optimality condition

$$\lambda^* \in arg\max_{\lambda \in \Lambda_1}\{\lambda' x^* - g_1(\lambda)\},\ \lambda^* \in arg\min_{\lambda \in \Lambda_2}\{\lambda' x^* - g_2(\lambda)\}.$$

Indeed, the Lagrangian optimality conditions are equivalent, since theu are both equivalent to the condition

$$x^{*\prime}\lambda^* = f_1(x^*) + g_1(\lambda^*) = f_2(x^*) + g_2(\lambda^*).$$

Moreover, if $\Lambda_1 = R^n$ and g_1 is differentiable, the Lagrangian optimality condition shows that

$$x^* = \bigtriangledown g_1(\lambda^*),$$

while if $\Lambda_2 = R^n$ and g_2 is differentiable, it shows that

$$x^* = \bigtriangledown g_2(\lambda^*).$$

4.6 Linear and Quadratic Problem

The standard form quadratic program(QP) is

$$\begin{aligned} &minimize\ \frac{1}{2}x^{\mathrm{T}}Px + q^{\mathrm{T}}x \\ &subject\ to\ Ax = b, x \geq 0, \end{aligned}$$

with variable$x \in \mathbb{R}^n$; we assume that $P \in S_+^n$. When $P = 0$, this reduces to the standard form linear program(LP).
We express it in ADMM form as

$$\begin{aligned} &minimize f(x) + g(z) \\ &subject\ to\ x - z = 0, \end{aligned}$$

where

$$f(x) = \frac{1}{2}x^{\mathrm{T}}Px + q^{\mathrm{T}}x,\ domf = \{x|Ax = b\}$$

is the original objective with restricted domain and g is the indicator function of the nonnegative orthant R_+^n. The scaled form of ADMM consists of the iterations

$$\begin{aligned} x^{k+1} &:= \underset{x}{argmin}(f(x) + \frac{\rho}{2}\|x - z^k + u^k\|_2^2) \\ z^{k+1} &:= (x^{k+1} + u^k)_+ \\ u^{k+1} &:= u^k + x^{k+1} - z^{k+1} \end{aligned}$$

The x-update is an equality-constrained least squares problem with optimality conditions

$$\begin{bmatrix} P + \rho I & A^{\mathrm{T}} \\ A & 0 \end{bmatrix} \begin{bmatrix} x^{k+1} \\ \nu \end{bmatrix} + \begin{bmatrix} q - \rho(z^k - u^k) \\ -b \end{bmatrix} = 0.$$

when $P = 0$, we gain the x-update for the linear programming.

4.7 Least Absolute Deviations

A simple variant on least squares fitting is least absolute deviations, in which we minimize $\|Ax - b\|_2^2$. Least absolute deviations provides a more robust fit than least squares when the data contains large outliers, and has been used extensively in statistics and econometrics. In ADMM form, the problem can be written as

$$\begin{aligned} &minimize\ \|z\|_1 \\ &subject\ to\ Ax - z = b, \end{aligned}$$

so $f = 0$ and $g = \|\|_1$. Exploiting the special form of f and g, and assuming $A^{\mathrm{T}}A$ is invertible, ADMM can be expressed as

$$\begin{aligned} x^{k+1} &:= (A^{\mathrm{T}}A)^{-1}A^T(b + z^k - u^k) \\ z^{k+1} &:= S_{\frac{1}{\rho}}(Ax^{k+1} - b + u^k) \\ u^{k+1} &:= u^k + Ax^{k+1} - z^{k+1} - b, \end{aligned}$$

where the soft thresholding operator is interpreted elementwise. The x-update step is the same as carrying out a least squares fit with coefficient matrix A and righthand side $b + z^k + u^k$. Thus ADMM can be interpreted as a method for solving a least absolute deviations problems by iteratively solving the associated least squares problem with a modified righthand side; the modification is then updated using soft thresholding. With factorization caching, the cost of subsequent least squares iterations is much smaller than the initial one, often making the time required to carry out least absolute deviations very nearly the same as the time required to carry out least squares.

4.8 Basis pursuit

Basis pursuit is the equality-constrained l_1 minimization problem

$$\begin{aligned} &minimize\ \|x\|_1 \\ &subject\ to\ Ax = b, \end{aligned}$$

with $x \in \mathbb{R}^n,\ A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$, with $m < n$. In ADMM form, basis pursuit can be written as

$$\begin{aligned} &minimize f(x) + \|z\|_1 \\ &subject\ to\ x - z = 0, \end{aligned}$$

where f is the indicator function of $\{x \in \mathbb{R}^n \mid Ax = b\}$. The ADMM algorithm is

$$\begin{aligned} x^{k+1} &:= (I - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A)(z^k - u^k) + A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}b \\ z^{k+1} &:= S_{\frac{1}{\rho}}(x^{k+1} + u^k) \\ u^{k+1} &:= u^k + x^{k+1} - z^{k+1}, \end{aligned}$$

The x-update involves solving a linearly-constrained minimum Euclidean norm problem. It can also be written as $x^{k+1} := \Pi(z^k - u^k)$, where Π is projection onto $\{x \in \mathbb{R}^n | Ax = b\}$.

4.9 Lasso

General $\ell 1$ Regularized Loss Minimization

$$minimize\ l(x) + \lambda\|x\|_1,$$

where l is any convex loss function.
An import special case of general $l1$ regularized loss minimization is $l1$ regularized linear regression, also called the *lasso*. This involves solving

$$minimize\ \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

Where $\lambda > 0$. In ADMM form, the lasso problem can be written as

$$\begin{aligned} &minimize\ f(x) + g(z) \\ &subject\ to\ x - z = 0, \end{aligned}$$

The ADMM algorithm is

$$\begin{aligned}
x^{k+1} &:= \underset{x}{argmin}(f(x) + \frac{1}{2}\|x - z^k + u^k\|_2^2) \\
&= \underset{x}{argmin}(\frac{1}{2}\|Ax - b\|_2^2 + \frac{1}{2}\|x - z^k + u^k\|_2^2) \\
&= (A^{\mathrm{T}}A + \rho I)^{-1}(A^{\mathrm{T}}b + \rho(z^k - u^k)) \\
z^{k+1} &:= \underset{z}{argmin}(g(z) + \frac{1}{2}\|x^{k+1} - z + u^k\|_2^2) \\
&= \underset{z}{argmin}(\lambda\|x\|_1 + \frac{1}{2}\|x^{k+1} - z + u^k\|_2^2) \\
&= S_{\lambda/\rho}(x^{k+1} + u^k) \\
u^{k+1} &:= u^k + x^{k+1} - z^{k+1}.
\end{aligned}$$

Note that $(A^{\mathrm{T}}A + \rho I)$is always invertible, since $\rho > 0$. The x-update is essentially a ridge regression.(i.e., quadratically regularized least squares) computation.When using a direct method, we can cache an initial factorizaion to make subsequent iterations much cheaper.

I implemented Lasso without computing the gradient because i had substituted $\underset{x}{argmin}(\frac{1}{2}\|Ax - b\|_2^2 + \frac{1}{2}\|x - z^k + u^k\|_2^2)$ with $(A^{\mathrm{T}}A + \rho I)^{-1}(A^{\mathrm{T}}b + \rho(z^k - u^k))$ in Octave programming.

5 Alternating Direction Method of Multipliers

5.1 Base of Alternating Direction Method of Multipliers

In order to gain a deep understanding of the ADMM, let's start from its composited elements firstly: Augmented Lagrangians Method and Dual Method.

5.1.1 Augmented Lagrangians Method

We describe this approach in the context of the equality-constrained problem

$$\min_x f(x), \quad \textit{subject to } c_i = 0, i \in \mathcal{E}.$$

The augmented lagrangians function$\mathcal{L}_A(x, \lambda; \mu)$ for this formulation is

$$\mathcal{L}_A(x, \lambda; \mu) \triangleq f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} {c_i}^2(x),$$

we see that the augumented Lagrangian function $\mathcal{L}_A(x, \lambda; \mu)$ differs from the standard Lagrangian by the presence of the squared terms, while it differs from the quadratic penalty function in the presence of the summation term involving λ. In this sense, it is a combination of the Lagrangian function and the quadratic penalty function.

5.1.2 Dual Method

Consider the equality-constrained convex optimization problem

$$minimize\ f(x), \quad subject\ to\ Ax = b$$

with variable $x \in \mathbb{R}^n$,where $A \in \mathbb{R}^m \times n$, and $f : \mathbb{R}^n \to \mathbb{R}$. The Lagrangian for problem is

$$\mathcal{L}(x, y) = f(x) + y^{\mathrm{T}}(Ax - b)$$

and the dual function is

$$g(y) = \inf_x L(x, y) = -f^*(-A^{\mathrm{T}}y) - b^{\mathrm{T}}y,$$

where y is the dual variable or Lagrange multiplier, and f^* is the convex conjugate of f. The dual problem is

$$maximize\ g(y),$$

with variable $y \in \mathbb{R}^m$.Assuming that the strong duality holds, the optimal values of the primal and dual problems are the same. We can recover a primal optimal point $x^\star$ from a dual optimal point $y^\star$ as

$$x^\star = \operatorname*{argmin}_x L(x, y^\star)$$

In the dual ascent method, we solve the dual problem using gradient ascent. Assuming that g is differentiable, the gradient $\bigtriangledown g(y)$ can be evaluated as $\bigtriangledown g(y) = Ax^+ - b$ where $x^+ = argmin_x L(x, y)$. The dual ascent method consists of iterating the updates

$$x^{k+1} := \operatorname*{argmin}_x L(x, y^k)$$
$$y^{k+1} := y^k + \alpha^k(Ax^k + 1 - b)$$

where $\alpha^k > 0$ is a step size, and the superscript is the iteration counter. The first step is an x-minimization step and the second step is a dual variable update. The major benefit of the dual ascent method is that it can lead to a decentralized algorithm in some cases. Suppose that the objective f is separable(with respecet to a partition or splitting of the variable into subvectors), meaning that

$$f(x) = \sum_{i=1}^{N} f_i(x_i),$$

where$x = (x_1, x_2..., x_N)$ and the variables $x_i \in \mathbb{R}^{n_i}$ are subvectors of x. Partitioning the matrix A conformably as

$$A = [A_1, A_2..., A_N],$$

so $Ax = \sum_{i=1}^{N} A_i x_i$, the Lagrangian can be written as

$$L(x, y) = \sum_{i=1}^{N} L_i(x_i, y) = \sum_{i=1}^{N} (f_i(x_i) + y^{\mathrm{T}} A_i x_i - (1/N) y^{\mathrm{T}} b),$$

which is also separable in x. This means that the x-minimization step splits into N separate problems that can be solved in parallel. Explicitly, the algorithm is

$$x_i^{k+1} := \underset{x_i}{argmin}\, L_i(x_i, y^k)$$
$$y^{k+1} := y^k + \alpha^k(Ax^k + 1 - b)$$

The x-minimization step is carried out independently, in parallel, for each $i = 1, ..., N$. In this case, we refer to the dual ascent method as dual decomposition.

5.2 Alternating Direction Method of Multipliers

ADMM is an algorithm that is intended to blend the decomposability of dual ascent with the superior convergence properties of the method of multipliers.

$$minimize\ f(x) + g(z), \quad subject\ to\ Ax + Bz = c$$

with variables $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, where $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$, and $c \in \mathbb{R}^p$. The optimal value of the problem will be denoted by

$$p^\star = inf\{f(x) + g(z) \mid Ax + Bz = c\}.$$

As in the method of multipliers, we form the augmented Lagrangian

$$\mathcal{L}_\rho(x, z, y) = f(x) + g(z) + y^{\mathrm{T}}(Ax + Bz - c) + (\frac{\rho}{2})\|Ax + Bz - c\|_2^2.$$

ADMM consists of the iterations

$$x^{k+1} := \underset{x}{argmin}\, \mathcal{L}_\rho(x, z^k, y^k)$$
$$z^{k+1} := \underset{z}{argmin}\, \mathcal{L}_\rho(x^{k+1}, z, y^k)$$
$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

where $\rho > 0$.The algorithm consists of an x-minimization step,a z-minimization step and a dual variable update. By combining the linear and quadratic terms in the augmented Lagrangian and scaling the dual variable, ADMM can be written in a different form, which is often more convennient, we called it Scaled Form

$$y^{\mathrm{T}}r + \frac{\rho}{2}\|r\|_2^2 = \frac{\rho}{2}\|r + \frac{1}{\rho}y\|_2^2 - \frac{\rho}{2}\|y\|_2^2 = \frac{\rho}{2}\|r + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2$$

Where the residual $r = Ax + Bz - c, u = (\frac{1}{\rho})y$ is the scaled dual variable.We can express ADMM as

$$x^{k+1} := \underset{x}{argmin}(f(x) + \frac{\rho}{2}\|Ax + Bz^k - c + u^k\|_2^2$$
$$z^{k+1} := \underset{z}{argmin}(g(z) + \frac{\rho}{2}\|Ax^{k+1} + Bz - c + u^k\|_2^2$$
$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c$$

Here the $(k + 1)$th iterate u^{k+1} is just given by a running sum of residuals:

$$u^{k+1} = u^0 + \sum_{i=1}^{k+1}(Ax^i + Bz^i - c),$$

5.2.1 Convergence

Here is a basic but very general convergence result for ADMM.
Assumption 1. The (extended-real-valued) functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, proper and convex.
Assumption 2. The unaugmented Lagrangian L_0 has a saddle point.
Under assumptions 1 and 2, the ADMM iterates satisfy the following:
Residual convergence. $r^k \to 0$ as $k \to \infty$, *i.e.*, the iterates approach feasibility.
Objective convergence. $f(x^k) + g(z^k) \to p^\star$ as$k \to \infty$, *i.e.*, the objective function of the iterates approaches the optimal value.
Dual variable convergence. $y^k \to y^\star$ as $k \to \infty$, where $y^\star$ is a dual optimal point.
We do not generically get primal convergence now, but this is true under more assumptions.
Convergence rate: not known in general, but roughly it behaves like a first-order method.
In practice, ADMM usually obtains a relatively accurate solution in a handful of iteration, but it requires a large number of iteration for a highly accurate solution. The choice of ρ can greatly influence practical convergence of ADMM.

5.2.2 Duality theory and saddle-points

Duality theory has recently received an elegant formulation, based on the consideration of a family of perturbed problems $(\mathcal{P}_p)$ associated to a problem $(\mathcal{P})$:Inf$\phi(v)$, with $v \in V$. We consider a general bifunction $\phi : V \times Y \to]-\infty, +\infty]$ such that

$$\Phi(v, 0) = \phi(v)$$

and the problem depending on a perturbation $p \in Y$

$$(\mathcal{P}_p) \ \underset{v \in V}{Inf}\, \Phi(v, p).$$

In this framework, we define a Lagrangian function:$\Lambda : V \times Y' \to [-\infty, +\infty]$ associated to $(\mathcal{P}_p)$ by the relation

$$\Lambda(v, \lambda) = \underset{p \in Y}{Inf}\{\Phi(v, p) - (\lambda, p)\}.$$

We can verify that

$$\underset{\lambda \in Y'}{Sup}\, \Lambda(v, \lambda) = \Phi(v, 0);$$

Hence the problem

$$\underset{v \in V}{Inf}\, \underset{\lambda \in Y'}{Sup}\, \Lambda(v, \lambda)$$

is the same with the original problem $(\mathcal{P})$. Parallel to the duality theory for convex programming in terms of mini-max, we define a dual problem to $(\mathcal{P})$

$$(\mathcal{D}) \ \underset{\lambda \in Y'}{Sup}\, \underset{v \in V}{Inf}\, \lambda(v, \lambda).$$

Let

$$(\mathcal{P}) \ \underset{v \in V}{Inf}\{f(Av) - < b, v >\} \ with \ b \in V'.$$

We assume that f is the sum of two lower semi-continuous convex functions f_1 and f_2 from Y into $]-\infty,+\infty]$,

$$f = f_1 + f_2$$

f_1 is $C^1.Gateaux - differentiable$ and its gradient ∇f_1 is weakly continuous on finite dimensional subspaces of V and storngly monotone: i.e. there exists $\gamma > 0$ such that

$$\forall y, z \in y, (\nabla f_1(y) - \nabla f_1(z), y - z) \geq \gamma |y - z|^2. \tag{5.1}$$

And the interior in Y of $domf_2$ is non empty. We assume moreover that the operator $A'A$ is an isomorphism from V onto V'; i.e. there exists $\alpha > 0$ such that

$$|Av|^2 \geq \alpha^2 \|v\|^2 \tag{5.2}$$

We consider the specific perturbation bifunction

$$\Phi(v, p) = f(Av + p) - < b, v > .$$

The dual problem is defined by

$$\underset{\lambda \in Y'}{Sup}\, \underset{v \in V}{Inf} \{- < b, v > + \underset{p \in Y}{Inf} [f(Av + p) - (\lambda, p)]\}$$

After the change of the variable $y = Av + p$

$$\underset{\lambda \in Y'}{Sup} \quad \underset{(v,y) \in V \times Y}{Inf} \{f(y) + (\lambda, Av - y) - < b, v >)\}, \tag{5.3}$$

PROPOSITION: Under (5.1) and (5.2), there exists a unique solution v^* to $(\mathcal{P})$.

Theorem 5.1 *Any saddle point $(v^*, y^*; \lambda^*)$ of $L(v, y; \lambda)$ over $(V \times Y) \times Y'$ satisfies: v^* is solution of $(\mathcal{P})$, $y^* = Av^*$ and $\lambda^* \in \partial f(Av^*)$ with $A'\lambda^* = b$.*

Although $(\mathcal{D})$ has been shown to have a solution λ^*, the inner minimization problem in (5.3) may not have a bounded solution for every $\lambda \in Y'$. For this reason, we switch to the augmented Lagrangian.

$$\Phi_r = \Phi(v, p) + \frac{r}{2}|p|^2 = g(v) + f(Av + p) + \frac{r}{2}|p|^2.$$

it's the same method to arrive:

$$\underset{\lambda \in Y'}{Sup} \quad \underset{(v,y) \in V \times Y}{Inf} \{f(y) + g(v) + (\lambda, Av - y) + \frac{r}{2}|Av - y|^2\}, \tag{5.4}$$

Theorem 5.2 *if $r > 0$, any saddle point of the augmented Lagrangian L_r is saddle point of the standard Lagrangian L and conversely.*

We assuume that the hypotheses (5.1), (5.2) and the interior in Y of $domf_2$ is nonempty. To solve (5.4), we can maximize the differentiable function $d_r(\lambda)$ using a gradient algorith,. This forms the basis for Uzawa's algorithm.
By replacing y^{n+1} by y^n, the resulting system is much simpler to solve and provides an approximation to the minimization in Step 1.
this part is based on the paper: D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," Computers and Mathematics with Applications, vol. 2, pp. 17–40, 1976.

Algorithm 12 Uzawa's algorithm

Let $\lambda^0 \in Y$. By induction λ^n being given
Step 1 Find v^{n+1},y^{n+1} minimizing on $V \times Y$

$$L_r(v, y; \lambda^n) = f(y) + g(v) + (\lambda, Av - y) + \frac{r}{2}|Av - y|^2.$$

step 2 Make

$$\lambda^{n+1} = \lambda^n + \rho(Av^{n+1} - y^{n+1}).$$

5.2.3 Proximal Operator

First, consider the simple case where $A = I$ in x^+, then the x-update is

$$x^+ = \underset{x}{argmin}(f(x) + \frac{1}{2C_k}\|x - \nu\|_2^2).$$

As a function of ν, the righthand side is denoted $prox_f, \rho(\nu)$ and is called the proximity operator of f with penalty ρ. In variational analysis,

$$f\tilde{(}\nu) = \underset{x}{inf}(f(x) + \frac{1}{2C_k}\|x - \nu\|_2^2)$$

is known as the Moreau-Yosida regularization of f. The x-minimization in the proximity operator is generally referred to as proximal minimization. While these observations don't by themselves allow us to improve the efficiency of ADMM, it does tie the x-minimization step to other well known ideas.
To deal with the case where f is convex but nondifferentiable, we provide a proof that is based on a hyperplane separation argument. For the proximal algorithm, we have for all k

$$f(y) \geq f(x^{k+1}) + \frac{1}{c^k}(x^k - x^{k+1})'(y - x^{k+1}), \forall y \in X.$$

Proof 5.1 *Consider the following two convex subsets of* R^{n+1}*:*

$$C_1 = (x, w)|f(x) < w, x \in X,$$
$$c_2 = (x, w)|w \leq \gamma^k - \frac{1}{2c^k}\|x - x^k\|^2, x \in R^n,$$

where γ^k *is given by*

$$\gamma^k = f(x^{k+1}) - \frac{1}{2c^k}\|x^{k+1} - x^k\|;$$

From the definition of x^{k+1} *via the proximal iteration,*$x^{k+1} = \underset{x \in X}{argmin}\{f(x) + \frac{1}{2c^k}\|x - x^k\|^2\}$*, the two sets are disjoint, so there exists a separating hyperplane* H *passing through the point* $(x^{k+1}, f(x^{k+1}))$*. Since* C_2 *is defined by a quadratic function, the normal of this hyperplane is defined by the corresponding gradient, and a simple geometric arguement shows that the vector*

$$(\frac{x^k - x^{k+1}}{c^k}, 1)$$

is normal to H. Since C_1 must lie on the opposite side of H, we obtain

$$f(y) + \frac{1}{c^k}(x^k - x^{k+1})'y \geq f(x^{k+1}) + \frac{1}{c^k}(x^k - x^{k+1})'x^{k+1}, \qquad \forall y \in R^n,$$

which is equivalent to the desired relation. QED

Generally, starting from any nonoptimal point x^k, the cost function value is reduced at each iteration, by setting $x = x^k$, we have

$$f(x^{k+1}) + \frac{1}{2c^k}\|x^{k+1} - x^k\|^2 \leq f(x^k).$$

The following proposition provides an inequality, which among others shows that the iterate distance to every optimal solution is also reduced.
For the proximal algorithm, we have for all k,

$$\|x^{k+1} - y\| \leq \|x^k - y\|^2 - 2c^k(f(x^{k+1}) - f(y)) - \|x^k - x^{k+1}\|^2, \forall y \in X.$$

Proof 5.2 *We have*

$$\begin{aligned}\|x^k - y\|^2 =& \|x^k - x^{k+1} + x^{k+1} - y\|^2 \\ =& \|x^k - x^{k+1}\|^2 + 2(x^k - x^{k+1})'(x^{k+1} - y) + \|x^{k+1} - y\|^2.\end{aligned}$$

By combing this relation with $f(y) \geq f(x^{k+1}) + \frac{1}{c^k}(x^k - x^{k+1})'(y - x^{k+1}), \forall y \in X$. the result follows.

Let us denote by f^* the optimal value

$$f^* = \inf_{x \in X} f(x),$$

(which may be $-\infty$) and by X^* the set of minima of f(which may be empty),

$$X^* = \operatorname*{argmin}_{x \in X} f(x).$$

(Convergence) Let $\{x^k\}$ be a sequence generated by the proximal algorithm. Then, if $\sum_{k=0}^{\infty} = \infty$, we have

$$f(x^k) \downarrow f^*,$$

and if X^* is nonempty, $\{x^k\}$ converges to some point in X^*.

5.3 Some Basic Explanations on ADMM Algorithm

5.3.1 Optimality Conditions

The necessary and sufficient optimality conditions for the ADMM problem are primal feasibility,

$$Ax^\star + Bz^\star - c = 0,$$

and dual feasibility,

$$0 \in \partial f(x^\star) + A^{\mathrm{T}} y^\star$$
$$0 \in \partial g(z^\star) + B^{\mathrm{T}} y^\star.$$

Here, ∂ denotes the subdifferential operator; Since z^{k+1} minimizes $L_p(x^{k+1}, z, y^k)$ by definition, we have that

$$\begin{aligned} 0 &\in \partial g(z^{k+1}) + B^{\mathrm{T}} y^k + \rho B^{\mathrm{T}}(Ax^{k+1} + Bz^{k+1} - c) \\ &= \partial g(z^{k+1}) + B^{\mathrm{T}} y^k + \rho B^{\mathrm{T}} r^{k+1} \\ &= \partial g(z^{k+1}) + B^{\mathrm{T}} y^{k+1}. \end{aligned}$$

This means that z^{k+1} and y^{k+1} always satisfy the second dual feasibility, so attaining optimality comes down to satisfying the first dual feasibility and the primal feasibility. Since x^{k+1} minimizes $L_\rho(x^{k+1}, z, y^k)$ by definition, we have that

$$\begin{aligned} 0 &\in \partial f(x^{k+1}) + A^{\mathrm{T}} y^k + \rho A^{\mathrm{T}}(Ax^{k+1} + Bz^k - c) \\ &= \partial f(z^{k+1}) + A^{\mathrm{T}}(y^k + \rho r^{k+1} + \rho B(z^k - z^{k+1})) \\ &= \partial f(z^{k+1}) + A^{\mathrm{T}} y^{k+1} + \rho A^{\mathrm{T}} B(z^k - z^{k+1}). \end{aligned}$$

equivalently,

$$\rho A^{\mathrm{T}} B(z^{k+1} - z^k) \in \partial f(x^{k+1}) + A^{\mathrm{T}} y^{k+1}.$$

This means that the quantity

$$s^{k+1} = \rho A^{\mathrm{T}} B(z^{k+1} - z^k)$$

can be viewed as a residual for the first dual feasibility condition. We will refer to s^{k+1} as the dual residual at iteration k+1, and to $r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$ as the primal residual at iteration k+1.
In summary, the optimality conditions for the ADMM problem consist of thtree condition. The last condition always holds for $(x^{k+1}, z^{k+1}, y^{k+1})$; the residuals for the other two are the primal and dual residuals r^{k+1} and s^{k+1}, respectively. These two residuals converge to zero as ADMM proceeds. (In fact, the convergence proof in appendix shows $B(z^{k+1} - z^k)$ convergence to zero, which implieds s^k converges to zero.)

5.3.2 Stopping Criteria

The residuals of the optimality conditions can be related to a bound on the objective suboptimality of the current point, *i.e.*, $f(x^k) + g(z^k) - p^\star$. As shown in the convergence proof in appendix, we have

$$f(x^{k+1}) + g*(z^k) - p^\star \leq -(y^k)^{\mathrm{T}} r^k + (x^k - x^\star)^{\mathrm{T}} s^k.$$

This shows that when the residuals r^k and s^k are samll, the objective suboptimality also must be small. We cann't use this inequality directly in a stopping criterion, however, since we don't know $x^\star$. But if we guess or estimate that $\|x^k - x^\star\|_2 \leq d$, we have that

$$f(x^k) + g(z^k) - p^\star \leq -(y^k)^{\mathrm{T}} r^k + d\|s^k\|_2 \leq \|y^k\|_2 \|r^k\|_2 + d\|s^k\|_2.$$

The middle or righthand terms can be used as an approximate bound on the objective suboptimality (which deepends on our guess of d). This suggests that a reasonable termination criterion is that the primal and dual residuals must be small, *i.e.*,

$$\|r^k\|_2 \leq \epsilon^{pri} \text{ and } \|s^k\|_2 \leq \epsilon^{dual},$$

where ϵ^{pri} and ϵ^{dual} are feasibility tolerances for the primal and dual feasibility conditions. These tolerances can be chosen using an absolute and relative criterion, such as

$$\epsilon^{pri} = \sqrt{p}\epsilon^{abs} + \epsilon^{rel} max\{\|Ax^k\|_2, \|Bz^k\|_2, \|c\|_2\},$$
$$\epsilon^{dual} = \sqrt{n}\epsilon^{abs} + \epsilon^{rel}\|A^{\mathrm{T}}y^k\|_2,$$

where $\epsilon^{abs} > 0$ is an absolute tolerance and $\epsilon^{rel} > 0$ is a relative tolerance. (The factors $\sqrt{p}$ and $\sqrt{n}$ account for the fact that the l_2 norms are in $\mathrm{R^p}$ and $\mathrm{R^n}$, respectively.) A reasonable value for the relative stopping criterion might be $\epsilon = 10^{-3}$ or 10^{-4}, depending on the application. The choice of absolute stopping criterion depends on the scale of the typical variable values.

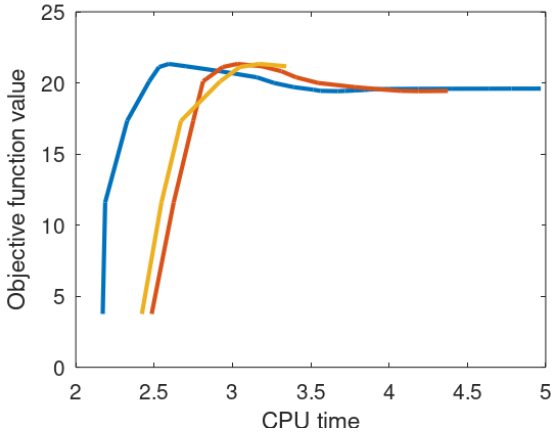


Figure 19: Objective function of Lasso, with different ϵ. The yellow, red and blue line are corresponding to $\epsilon^{abs} = 10^{-3}$ $\epsilon^{rel} = 10^{-1}, \epsilon^{abs} = 10^{-4}$ $\epsilon^{rel} = 10^{-2}, \epsilon^{abs} = 10^{-5}$ $\epsilon^{rel} = 10^{-3}$.

In practice, ADMM usually obtians a relatively accurate solution, but it requires a large number of iterations(or CPU time) for a highly accurate solution(like first-order method).Obviously, with the decreasing of ϵ^{abs} and ϵ^{rel}, the accuracy of the convergence becomes higher.

5.3.3 Varing Penalty Parameter

A standard extension is to use possibly different penalty parameter ρ^k for each iteration, with the goal of improving the convergence in practice, as well as making performance less dependent on the initial choice of the penalty parameter. Though it can be difficult to prove the convergence of ADMM when ρ varies by iteration, the fixed ρ theory still

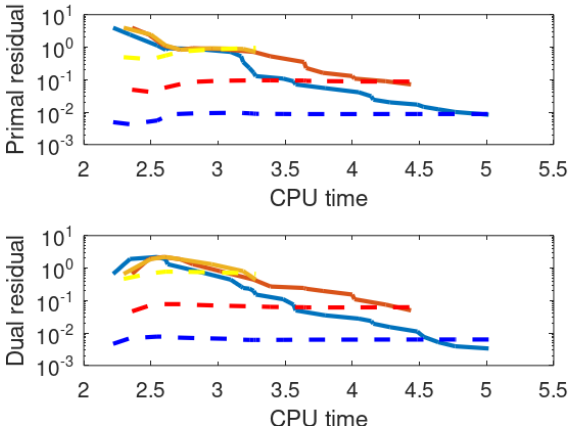


Figure 20: Solid line: $\|r^k\|_2$ and $\|s^k\|_2$, Dotted line: ϵ^{pri} and ϵ^{dual}. The yellow, red and blue lines are corresponding to $\epsilon^{abs} = 10^{-3}$ $\epsilon^{rel} = 10^{-1}, \epsilon^{abs} = 10^{-4}$ $\epsilon^{rel} = 10^{-2}$ *and* $\epsilon^{abs} = 10^{-5}$ $\epsilon^{rel} = 10^{-3}$.

applies if one just assumes that ρ becomes fixed after a finite number of iterations. A simple scheme that often works well is:

$$f(x) = \begin{cases} \tau^{incr}\rho^k & if\ \|r^k\|_2 > \mu\|s^k\|_2 \\ \rho^k/\tau^{decr} & if\ \|s^k\|_2 > \mu\|r^k\|_2 \\ \rho^k & otherwise, \end{cases}$$

where $\mu > 1, \tau^{incr} > 1$, and $\tau^{decr} > 1$ are parameters. Typical choices might be $\mu = 10$ and $\tau^{incr} = \tau^{decr} = 2$. The idea behind this penalty parameter update is to try to keep the primal and dual residual norms within a factor of μ of one another as they both converge to zero. The ADMM update equations suggest that large values of ρ place a large penalty on violations of primal feasibility and so tend to produce small primal residuals. Conversely, the definition of s^{k+1} suggests that small values of ρ tend to reduce the dual residual, but at the expense of reducing the penalty on primal feasibility, which may result in a larger primal residual.The adjustment schema inflates ρ by τ^{incr} when the primal residual appears large compared to the dual residual, and deflates ρ by τ^{decr} when the primal residual seems too small relative to the dual residual.Note: the scaled dual variable $u^k = (\frac{1}{\rho})y^k$ must be rescaled after updating ρ.
Parameter ρ adjusts the trade-off between minimization the objective function and the feasibility of constraints. Choice of ρ can greatly influence practical convergence of ADMM:

$$\rho\ too\ large\ \rightarrow not\ enough\ emphasis\ on\ minimizing\ f + g$$
$$\rho\ too\ small\ \rightarrow not\ enough\ emphasis\ on\ feasibility$$

(Role of the Penalty Parameter in ADMM)
The purpose of this exercise is to illustrate by example the difficulty of selecting a good value of c in ADMM.(c is the value of ρ.)Consider the problem of minimizing $\frac{1}{2}x^2 + \frac{1}{2}(ax)^2$ over $x \in R$, which is in the ADMM format with the notational identifications $f_1(x) = \frac{1}{2}x^2$, $f_2(z) = \frac{1}{2}z^2$, and $A = a$. Here a, x, and z are scalars with $a \neq 0$. Consider also the

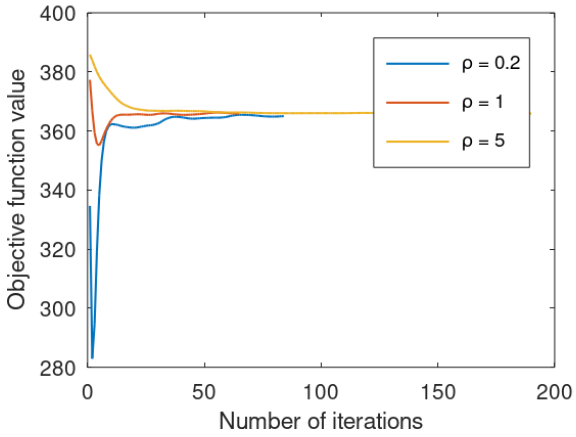


Figure 21: Objective function of Linprog, with different ρ

following algorithm with $c > 0$:

$$x^{k+1} = \frac{a(cz^k - \lambda^k)}{1 + ca^2},\ z^{k+1} = \frac{\lambda^k + cax^{k+1}}{1 + c},\ \lambda^{k+1} = \lambda^k + c(ax^{k+1} - z^{k+1}).$$

(a) Verify that this is the ADMM for the problem.

$$L_c(x, z, \lambda) = \frac{1}{2}x^2 + \frac{1}{2}z^2 + \lambda'(Ax - z) + \frac{c}{2}\|Ax - z\|^2.$$

$$\begin{aligned} x^{k+1} &\in \underset{x \in X}{argmin}\, L_c(x, z^k, \lambda^k) \\ &= \underset{x \in X}{argmin}\{\frac{1}{2}x^2 + \frac{1}{2}(z^k)^2 + (\lambda^k)'(Ax - z^k) + \frac{c}{2}\|Ax - z^k\|^2\} \end{aligned}$$

Since $A = a$, by differential, $x^{k+1} = \frac{a(cz^k - \lambda^k)}{1+ca^2}$.

$$\begin{aligned} z^{k+1} &\in \underset{z \in Z}{argmin}\, L_c(x^{k+1}, z, \lambda^k), \\ &= \underset{z \in Z}{argmin}\{\frac{1}{2}(x^{k+1})^2 + \frac{1}{2}z^2 + (\lambda^k)'(Ax^{k+1} - z) + \frac{c}{2}\|Ax^{k+1} - z\|^2\}, \end{aligned}$$

Since $A = a$, by differential, $z^{k+1} = \frac{\lambda^k + cax^{k+1}}{1+c}$.
For λ, $\lambda^{k+1} = \lambda^k + c(ax^{k+1} - z^{k+1})$. (b) Show that for all $k \geq 1$, we have $\lambda^k = z^k$ and that an equivalent form of the algorithm is

$$x^{k+1} = \frac{a(c-1)}{1 + ca^2}z^k,\ z^{k+1} = \frac{1 + c^2a^2}{(1 + ca^2)(1 + c)}z^k.$$

For all $k \geq 1$,

$$\begin{aligned} z^{k+1} - \lambda^{k+1} &= \frac{\lambda^k + cax^{k+1}}{1+c} - \lambda^k - c(ax^{k+1} - z^{z+1}) \\ &= \frac{1}{1+c}(\lambda^k + cax^{k+1} - (1+c)\lambda^k - (1+c)c(ax^{k+1} - z^{k+1})) \\ &= \frac{1}{1+c}(-c\lambda^k + cax^{k+1} - cax^{k+1} + cz^{k+1} - c^2ax^{k+1} + c^2z^{k+1}) \\ &= \frac{c}{1+c}(-\lambda - acx^{k+1} + z^{k+1} + cz^{k+1}) \\ &= \frac{c}{1+c}((1+c)z^{k+1} - (\lambda^k + acx^{k+1})) \end{aligned}$$

Since $z^{k+1} = \frac{\lambda^k + cax^{k=1}}{1+c}, c > 0$, we get for all $k \geq 1, z^{k+1} - \lambda^{k+1} = 0, Q.E.D.$
Substitute $\lambda^k = z^k$ into $x^{k+1} = \frac{a(cz^k - \lambda^k}{1+ca^2}$, we have $x^{k+1} = \frac{a(c-1)}{1+ca^2}z^k$, and then use $x^{k+1} = \frac{a(c-1)}{1+ca^2}z^k$ with $\lambda^k = z^k$ to get $z^{k+1} = \frac{1+c^2a^2}{(1+ca^2)(1+c)}z^k$.
(c) plot the ratio of linear convergence

$$\beta(c) = \frac{1+c^2a^2}{(1+ca^2)(1+c)},$$

as a function of c, and verify that $\beta(c) \in (0,1)$ for all $c > 0$, that $\beta(c) \to 1$ as either $c \downarrow 0$, or $c \uparrow \infty$, and that $\beta(c)$ is minimized over $c \in (0,\infty)$ at $c = \frac{1}{|a|}$.
Since $c > 0$, we can get $\beta(c) > 0$. And $1 + c^2a^2 < 1 + c^2a^2 + c + ca^2 = (1+ca^2)(1+c)$, so $\beta(c) \in (0,1)$ for all $c > 0$

$$\beta(c) = \frac{1+c^2a^2}{(1+ca^2)(1+c)} = \frac{1+c^2a^2}{1+c^2a^2+c+ca^2}$$

It's clear that $\beta(c) \to 1$ as either $c \downarrow 0$ or $c \uparrow \infty$.
In order to minimize $\beta(c)$,

$$\beta(c) = \frac{1+c^2a^2}{1+c^2a^2+c+ca^2} = \frac{\frac{1}{ca}}{\frac{1}{ca}+ca+\frac{1}{a}+a} = \frac{1}{1+\frac{\frac{1}{a}+a}{\frac{1}{ca}+ca}},$$

Since $\frac{1}{a} + a$ is a constant, we need to minimize $\frac{1}{ca} + ca$ to minimize $\beta(c)$, with $c \in (0,\infty)$, thus $c = \frac{1}{|a|}$.

5.3.4 Over-relaxation

In the $z-$ and $y-$updates, the quantity Ax^{k+1} can be replaced with

$$\alpha^k Ax^{k+1} - (1-\alpha^k)(Bz^k - c),$$

where $\alpha^k \in (0,2)$ is a relaxation parameter; when $\alpha^k > 1$, this technique is called over-relaxation, and when $\alpha^k < 1$, it is called under-relaxation. Some experiments suggest that over-relaxation with $\alpha \in [1.5, 1.8]$ can improve convergence.

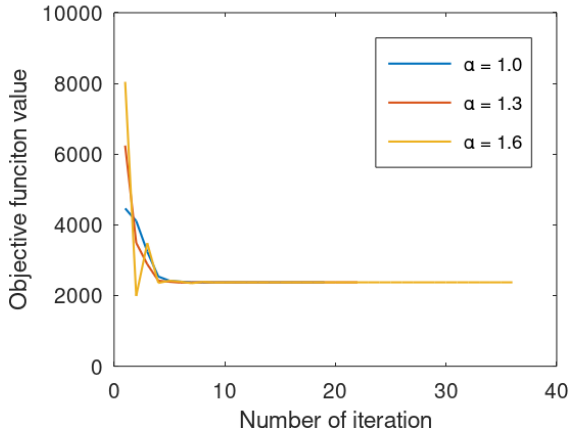


Figure 22: Objective function of Least Abolute Deviations, with different α

5.3.5 Warm Start

A standard trick is to initialize the iterative method used in the x-update at the solution x^k obtained in the previous iteration, this is called a warm start. The previous ADMM iterate often gives a good enough apporximation to result in far fewer iterations(of the iterative method used to compute the update x^{k+1}) than if the iterative method were started at zero or some other default initialization. This is especially the case when ADMM has almost converged, in which case the updates will not change significantly from their previous values. Unfinished part

Algorithm 13 Basic ADMM Algorithm

Initialize x_0, z_0 and y_0, set max-iteration
While k = 1 : max-iteration
Repeat

$$\begin{aligned} x^k &= \underset{x}{argmin}(f(x) + \frac{\rho}{2}\|Ax + Bz^{k-1} - c + u^{k-1}\|_2^2 \\ z^k &= \underset{z}{argmin}(g(z) + \frac{\rho}{2}\|Ax^k + Bz - c + u^{k-1}\|_2^2 \\ u^k &= u^{k-1} + Ax^k + Bz^k - c \end{aligned} \tag{5.5}$$

If satisfying stopping criteria:

$$\|r^k\| \leq \epsilon^{pri} \;\&\&\; \|s^k\| \leq \epsilon^{dual}, \tag{5.6}$$

break

Here in formula (5.5) i would like to make a direct connection with scaled form of ADMM. But in practice, we can just use x, z and u, ignoring the superscript. Concerning its stopping criteria (5.6), i given a brief explanation later. But the derivation of the stopping criteria is based on the particular convergence result in this report. When we need to solve some problems not satisfying the assumption 5.1 and 5.2, we may need to consider another stopping criteria.

5.3.6 Lipschitz Condition

Definition 5.1 *f is Lipschitz continuous on Q with constant L, if for all $x, y \in Q$, we have*

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

If $\bigtriangledown f$ is Lipschitz continuous, then function f is Lipschitz continuous gradient. Similarly, if $\bigtriangledown^2 f$ is Lipschitz continuous, then function f is Lipschitz continuous Hessian.

Theorem 5.3 *If f is Lipschitz continuous gradient on* R^n. *Then for any $x, y \in R^n$ we have*

$$|f(y) - f(x) - \langle f'(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|^2$$

Proof 5.3 *For all $x, y \in R^n$ we have*

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle f'(x + \tau(y - x)), y - x\rangle d\tau \\ &= f(x) + \langle f'(x), y - x\rangle + \int_0^1 \langle f'(x + \tau(y_x)) - f'(x), y - x\rangle d\tau. \end{aligned}$$

Therefore

$$\begin{aligned} |f(y) - f(x) - \langle f'(x), y - x\rangle| &= |\int_0^1 \langle f'(x + \tau(y_x)) - f'(x), y - x\rangle d\tau| \\ &\leq \int_0^1 |\langle f'(x + \tau(y - x)) - f'(x), y - x\rangle| d\tau \\ &\leq \int_0^1 \|f'(x + \tau(y - x)) - f'(x)\| \cdot \|y - x\| d\tau \\ &\leq \int_0^1 \tau L\|y - x\|^2 d\tau = \frac{L}{2}\|y - x\|^2. \end{aligned}$$

Theorem 5.4 *If f is Lipschitz continuous hessian on* R^n. *Then for any $x, y \in R^n$ we have*

$$|f(y) - f(x) - \langle f'(x), y - x\rangle - \frac{1}{2}\langle f''(x)(y - x), y - x\rangle| \leq \frac{L}{6}\|y - x\|^3$$

5.3.7 Complexity

Analytical complexity: The number of calls of the oracle which is necessary to solve problem P up to accuracy ϵ.
Arithmetical complexity: The total number of arithmetic operations(including the work of oracle and work of method), which is necessary for solving problem P up to accuracy ϵ.

Algorithm 14 General Alternating Minimization Algorithm

Input: Objective function $f : \mathcal{X} \times \mathcal{Y} \to R$
Output: A point $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ with near-optimal objective value
1: $(x^1, y^1) \leftarrow INITALIZE()$
2: **for** t = 1, 2, ..., T **do**
3: $\quad x^{t+1} \leftarrow argmin_{x \in \mathcal{X}} f(x, y^t)$
4: $\quad y^{t+1} \leftarrow argmin_{y \in \mathcal{Y}} f(x^{t+1}, y)$
end for
return(x^T, y^T)

5.4 Alternating Minimization Algorithm

The alternating minimization algorithm is outlined in Algorithm 3 for an optimization problem on two variables constrained to the sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. The procedure can be easily extended to functions with more variables, or have more complicated constraint sets of the form $\mathcal{Z} \subset \mathcal{X} \times \mathcal{Y}$. After an initialization step. AMA alternately fixes one of the variables and optimizes over the other.

The approach of solving several intermediate marginal optimization problems instead of a single big problem is the key to the practical success of AMA. Alternating minimization is mostly used when these marginal problems are easy to solve. Later, we will see that there exist simple, often closed form solutions to these marginal problems for applications such that as matrix completion, robust learning etc.

Definition 5.2 *Marginally Optimum Coordinate. Let f be a function of two variables constrained to be in the sets $\mathcal{X}, \mathcal{Y}$ respectively. For any point $y \in \mathcal{Y}$, we say that $\hat{x}$ is a marginally optimal coordinate with respect to $\boldsymbol{y}$, and use the shorthand $\hat{x} \in mOPT_f(y)$, if $f(\hat{x}, y) \leq f(x, y)$ for all $x \in \mathcal{X}$. Similarly for any $x \in \mathcal{X}$, we say $\hat{y} \in mOPT_f(x)$ if $\hat{y}$ is a marginally optimal coordinate with respect to x.*

Definition 5.3 *Bistable Point. Given a function f over two variables constrained within the sets $\mathcal{X}, \mathcal{Y}$ respectively, a point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is constrained a bistable point if $y \in mOPT_f(x)$ and $x \in mOPT_f(y)$ i.e. , both coordinates are marginally optimal with respect to each other.*

It is easy to see that the optimum of the optimization problem must be a bistable point. The reader can also verify that the AMA procedure must stop after it has reached a bistable point. However, two questions arise out of this. First, how fast does AMA approach a bistable point and second, even if it reaches a bistable point, is that point guaranteed to be globally optimal?

The first question will be explored in detail later. It is interesting to note that the AMA procedure has no parameters, such as step length. This can be interpreted as a benefit as well as a drawback. While it relieves the end-user from spending time tweaking parameters, the convergence eof the AMA procedure is totally dependent on structural properties of the optimization problem.

5.4.1 Convergence Guarantee for AMA

We will investigate what happens when the AMA is executed with a non-convex function. Note that the previous results crucially uses convexity and will not extend here. Moreover, for non-convex functions, there is no assurance that all bistable points are global minima. Instead we will have to fend for fast convergence to a bistable point, as well as show that the function is globally minimized there.

Doing the above will require additional structure on the objective function. In the following, we will denote $f^* = min_{x,y} f(x,y)$ to be the optimum value of the objective function. We will fix (x^*, y^*) to be any point such that $f(x^*, y^*) = f^*$(there may be several). We will also let $Z^* \subset R^p \times R^q$ denote the set of all bistable points for f.

First of all, notice that if a continously differentiable function $f : R^p \times R^q \to R$ is marginally convex (strongly or otherwise) in both its variables, then its bistable points are exactly its stationary points.

Lemma 5.1 *A point (x,y) is bistable with respect to a continuously differentiable function $f : R^p \times R^q$ that is marginally convex in both its variables iff $\bigtriangledown f(x,y) = 0$.*

Proof. It is easy to see that partial derivatives must vanish at a bistable point since the function is differentiable and thus we get $\bigtriangledown f(x,y) = [\bigtriangledown_x f(x,y), \bigtriangledown_y f(x,y)] = 0$. Arguing the other way round, if the gradient, and by extension the partial derivatives, vanish at (x,y), then by marginal convexity, for any x'

$$f(x',y) - f(x,y) \geq\rangle \bigtriangledown_x f(x,y), x' - x\langle = 0$$

Similarly, $f(x,y') \geq f(x,y)$ for any y'. Thus (x,y) is bistable.

The above tells us that Z^* is also the set of all stationary points of f. However, not all points in Z^* may be global minima. Addressing this problem requires careful initialization and problem-specific analysis, that we carry out for problems such as matrix completion. For now, we introduce a generic robust bistability property that will be very useful in the analysis. Similar properties are frequently used in the analysis of AMA-style algorithms.

Definition 5.4 *(Robust Bistability Property) A function $f : R^p \times R^q \to R$ satisfies the C-robust bistability property if for some $C > 0$, for every $(x,y) \in R^p \times R^q, \hat{y} \in mOPT_f(x)$ and $\hat{x} \in mOPT_f(y)$, we have*

$$f(x,y^*) + f(x^*,y) - 2f^* \leq C \cdot (2f(x,y) - f(X,\hat{y}) - f(\hat{x},y)).$$

The right hand expression captures how much one can reduce the function value locally by performing marginal optimizations. The property suggests that if not much local improvement can be made (i.e., if $f(x,\hat{y}) \approx f(x,y) \approx f(\hat{x},y)$) then we are close to the optimum. This has a simple corollary that all bistable points achieves the globally optimal function value. We now present a convergence analysis for AMA.

Theorem 5.5 *Let $f : R^p \times R^q \to R$ be a continuously differentiable (but possible non-convex) function that, within the region $S_0 = \{x, y : f(x, y) \leq f(0, 0)\} \in f(0, 0)\} \in R^{p+q}$, satisfies the properties of $\alpha-$MSC, β-MSS in both its variables, and C-robust bistability. Let Algorithm 3 be executed with the initialization $(x^1, y^1) = (0, 0)$. Then after at most $T = O(log\frac{1}{\epsilon})$ steps, we have $f(x^T, y^T) \leq f^* + \epsilon$.*

Note that the MSC/MSS and robust bistability properties need only hold within the sublevel set S_0. This again underlines the importance of proper initialization. Also note that AMA offers rapid convergence despite the non-convexity of the objective. In order to prove the result, the following consequence of C-robust bistability will be useful.

Lemma 5.2 *Let f satisfy the properties mentioned in Theorem 4.1. Then for any $(x, y) \in R^p \times R^q$, $\hat{y} \in mOPT_f(x)$ and $\hat{x} \in mOPT_F(y)$,*

$$\|x - x^*\|_2^2 + \|y - y^*\|_2^2 \leq \frac{C\beta}{\alpha}(\|x - \hat{x}\|_2^2 + \|y - \hat{y}\|_2^2)$$

Proof. Applying MSC/MSS repeatedly gives us

$$f(x, y^*) + f(x^*, y) \geq 2f^* + \frac{\alpha}{2}(\|x - x^*\|_2^2 + \|y - y^*\|_2^2)$$
$$2f(x, y) \leq f(x, \hat{y}) + f(\hat{x}, y) + \frac{\beta}{2}(\|x - \hat{x}\|_2^2 + \|y - \hat{y}\|_2^2)$$

Applying the robust stability then proves the result.

It is noteworthy that Lemma 4.2 relates local convergence to global convergence and assures us that reaching an almost bistable point is akin to converging to the optimum. Such a result can be crucial, especially for non-convex problems. Indeed, similar properties are used in other proofs concerning coordinate minimization as well.
Proof of Theorem 4.1. We will use $\Phi_t = f(x^t, y^t) - f^*$ as the potential function. Since the intermediate steps in AMA are marginal optimizations and not gradient steps, we will actually find it useful to apply marginal strong convexity at a local level, and apply marginal strong smoothness at a global level instead.
Apply Marginal Strong Smoothness As $\bigtriangledown_x f(x^*, y^*) = 0$, applying MSS gives us

$$f(x^{t+1}, y^*) - f(x^*, y^*) \leq \frac{\beta}{2}\|x^{t+1} - x\|_2^2.$$

Further, the AMA updates ensure $y^{t+1} \in mOPT_f(x^{t+1})$, which gives

$$\Phi_{t+1} = f(x^{t+1}, y^{t+1}) - f^* \leq f(x^{t+1}, y^*) - f^* \leq \frac{\beta}{2}\|x^{t+1} - x^*\|_2^2,$$

Apply Marginal Strong Convexity Since $\bigtriangledown_x f(x^{t+1}, y^t) = 0$,

$$f(x^t, y^t) \geq f(x^{t+1}, y^t) + \frac{\alpha}{2}\|x^{t+1} - x^t\|_2^2$$
$$\geq f(x^{t+1}, y^{t+1}) + \frac{\alpha}{2}\|x^{t+1} - x^t\|_2^2,$$

5.5 Practical Examples

This chapter we consider the problem

$$\begin{aligned} &minimize\ H(u) + G(v) \\ &subject\ to\ Au + Bv = b \end{aligned}$$

over variables $u \in R^{N_u}$ and $v \in R^{N_v}$, where $H : R^{N_u} \to (-\infty, \infty]$ and $G : R^{N_v} \to (-\infty, \infty]$ are closed convex functions, $A \in R^{N_b \times N_u}$ and $B \in R^{N_b \times N_v}$ are linear operators, and $b \in R^{N_b}$ is a vector of data.
Recall:

Algorithm 15 ADMM.

Require: $\nu_0 \in R^{N_\nu}$, $\lambda_0 \in R_b^N$, $\tau > 0$
1: **for** k = 0,1,2... **do**
2: $u_{k+1} = argmin_u H(u) + \langle \lambda_k, -Au \rangle + \frac{\tau}{2}\|b - Au - Bv_k\|^2$
3: $v_{k+1} = argmin_v G(v) + \langle \lambda_k, -Bv \rangle + \frac{\tau}{2}\|b - Au_{k+1} - Bv\|^2$
$\lambda_{k+1} = \lambda_k + \tau(b - Au_{k+1} - Bv_{k+1})$
end for

Problem 1:Box-constrained Quadratic problem

$$\begin{aligned} &minimize\ f(x) = (1/2)x^T Px + q^T x + r, \\ &subject\ to\ lb \leq x \leq ub \end{aligned}$$

Step 1: we separate the minimization function as the form, $H(x) = (1/2)x^T Px + q^T x + r$, $G(y) = 0$, the constraint is $lb - x \leq 0$ and $x - ub \leq 0$.
Step 2: the update of x involves solving the KKT(Karush-Kuhn-Tucker) system, since there is no equality constraint in this problem, we can get the x update easily,

$$(P + \rho I)x^{k+1} = \rho(z^k - u^k) - q$$

```
for k = 1:MAX_ITER
    if k > 1
        x = R \ (R' \ (rho*(z - u) - q));
    else
        R = chol(P + \rho*eye(n));
        x = R \ (R' \ (rho*(z - u) - q));
    end
```

the update of z based on the box-constraint $lb \leq x \leq ub$

```
z = min(ub, max(lb, x_hat + u));
```

and the update of u is

```
u = u + (x_hat - z);
```

Problem 2:Basis pursuit
Basis pursuit is the equality-constrained l_1 minimization problem

$$\begin{aligned} &minimize\ \|x\|_1 \\ &subject\ to\ Ax = b, \end{aligned}$$

with variable $x \in R^n$, data $A \in R^{m \times n}$, $b \in R^m$, with $m < n$. Basis pursuit is often used as a heuristic for finding a parse solution to an underdetermined system of linear equations. It plays a central role in modern statistical signal processing, particularly the theory of compressed sensing.
In ADMM form, basis pursuit can be written as

$$\begin{aligned} &minimize\ f(x)+\|z\|_1 \\ &subject\ to\ x - z = 0, \end{aligned}$$

where f is the indicator function of $\{x \in R^n | Ax = b\}$. The ADMM algorithm is then

$$\begin{aligned} x^{k+1} &:= \prod(z^k - u^k) \\ z^{k+1} &:= S_{1/\rho}(x^{k+1} + u^k) \\ u^{k+1} &:= u^k + x^{k+1} - z^{k+1}, \end{aligned}$$

where $\prod$ is projection onto $\{x \in R^n | Ax = b\}$. The x-update, which involves solving a linearly-constrained minimum Euclidean norm problem, can be written explicitly as

$$x^{k+1} := (I - A^T(AA^T)^{-1}A)(z^k - u^k) + A^T(AA^T)^{-1}b.$$

```
% precompute static variables for x-update (projection on to Ax=b)
AAt = A*A';
P = eye(n) - A' * (AAt \ A);
q = A' * (AAt \ b);

for k = 1:MAX_ITER
    % x-update
    x = P*(z - u) + q;
```

```
 % z-update with relaxation
    z = shrinkage(x + u, 1/rho);
```

```
    u = u + (x - z);
```

Problem 3:Lasso
Lasso is an important special case of *general* l_1 *regularized loss minimization*,

l_1 regularized linear regression. The lasso has been widely applied, particularly in the analysis of biological data, where only a small fraction of a huge number of possible factors are actually predictive of some outcome of interest. This involves solving

$$minimize\ (1/2)\|Ax-b\|_2^2+\lambda\|x\|_1,$$

Step 1: we separate the minimization function as the form, $f(x)=(1/2)\|Ax-b\|_2^2$ and $g(z)=\lambda\|z\|_1$.
Step 2: for $f(x)=(1/2)\|Ax-b\|_2^2$, we get

$$f(x)=(1/2)x^TA^TAx-A^Tbx+(1/2)b^Tb$$

By the Quadratic objective terms, we achieve

$$x^+=(A^TA+\rho I)^{-1}(A^Tb+\rho(z^k-u^k))$$

```
for k = 1:MAX_ITER

    % x-update
    q = Atb + rho*(z - u);    % temporary value
    if( m ≥ n )    % if skinny
       x = U \ (L \ q);
    else            % if fat
       x = q/rho - (A'*(U \ ( L \ (A*q) )))/rho^2;
    end
```

Step 3: for $g(z)=\lambda\|z\|_1$, by soft thresholding, we have

$$z^{k+1}:=S_{\lambda/\rho}(x^{k+1}+u^k)$$

where the soft thresholding operator S is defined as

$$S_k(a)=\begin{cases}a-k & a>k\\ 0 & |a|\le k\\ a+k & a<-k,\end{cases}$$

```
z = shrinkage(x + u, lambda/rho);

function z = shrinkage(x, kappa)
    z = max( 0, x - kappa ) - max( 0, -x - kappa );
end
```

Similarly, the update of u is

```
u = u + (x_hat - z);
```

In summary

$$x^{k+1} := (A^T A + \rho I)^{-1}(A^T b + \rho(z^k - u^k))$$
$$z^{k+1} := S_{\lambda/\rho}(x^{k+1} + u^k)$$
$$u^{k+1} := u^k + x^{k+1} - z^{k+1}.$$

6 Conclusion

Most of the results obtained are in the corrsponding section, here i give a summary of the representative results as the conclusion of this technical report.

6.1 Line Search: Comparison with Different Conditions

Algorithm 16 Global Wolfe condition template

procedure Wolfeglobal($x^0, \sigma_1, \sigma_2, \mu$)
  $k \leftarrow 0$
  repeat
    Find $d^k \in \mathrm{R}^n$ such that $\Delta f(x^k; d^k) < 0$
    if no such d^k **then**
      $0 \in \partial f(x^k)$ **return**
    end if
    Let t_k be a step size satisfying Conditions(Armijo, weak Wolfe condition, Wolfe condition, strong Wolfe condition.)
    if no such t_k **then**
      f unbounded below. **return**
    end if
    $x^k \leftarrow x^k + t_k d^k$
    $k \leftarrow k + 1$
  until
end procedure

Type	condition	Properties	Summary
Armijo Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$	sufficient decrease	
Weak Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and $\sigma_2\Delta f(x;d) \leq \frac{f(x+td;\mu d)}{\mu}$, with $0<\sigma_1<\sigma_2<1, \mu>0$.	sufficient decrease and curvature condition with a modification which prevents the line search early termination at "strongly negatige" slopes.	make d^k less of a direction of descent (and possibly a direction of ascent) at the new point.
Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and $\sigma_2 f'(x;d) \leq f'(x+td;d)$, with $0<\sigma_1<\sigma_2<1$.	sufficient decrease and curvature condition	
Strong Wolfe Condition	$f(x+td) \leq f(x)+\sigma_1 t\Delta f(x;d)$ and$\lvert f'(x+td;d)\rvert \leq \sigma_2\lvert f'(x;d)\rvert$, with $0<\sigma_1<\sigma_2<1$.	sufficient decrease and curvature condition with a modification to for t_k to lie in at least a broad neighborhood of a local minimizer or stationary point	don't allow the derivative to be too positive, try to push the directional derivative in direction d^k closer to zero at the new point.

Remark 1. The condition $\sigma_2\Delta f(x;d) \leq \frac{f(x+td;\mu d)}{\mu}$ is a *curvature condition* that parallels the classical weak Wolfe curvature condition for smooth, unconstrained minizaition:

$$\sigma_2 f'(x;d) \leq f'(x+td;d),$$

which prevents the line search early termination at "strongly negative" slopes.
Remark 2. The strong Wolfe condition require $|f'(x+td;d)| \leq -\sigma_2 f'(x;d)$, whenever f is smooth. However, in nonsmooth minimization, kinks and upward cusps at local minimizaters make the condition unworkable. Lemma 5.1 in [6] proved that the set of points satisfying weak Wolfe conditions has nonempty interior.
1.Armijo Condition: with *initial* $t=1$, $\gamma=0.5$, $\sigma_1=0.4$. ($t>0, \gamma\in(0,1), \sigma_1\in(0,0.5)$; it converges faster when $\gamma\searrow 0$, $\sigma_1\nearrow 0.5$.)
2. Wolfe condition: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$
3. Weak Wolfe Conditions: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$
4. Strong Wolfe Conditions: with initial t = 1, $\sigma_1=0.25, \sigma_2=0.75$

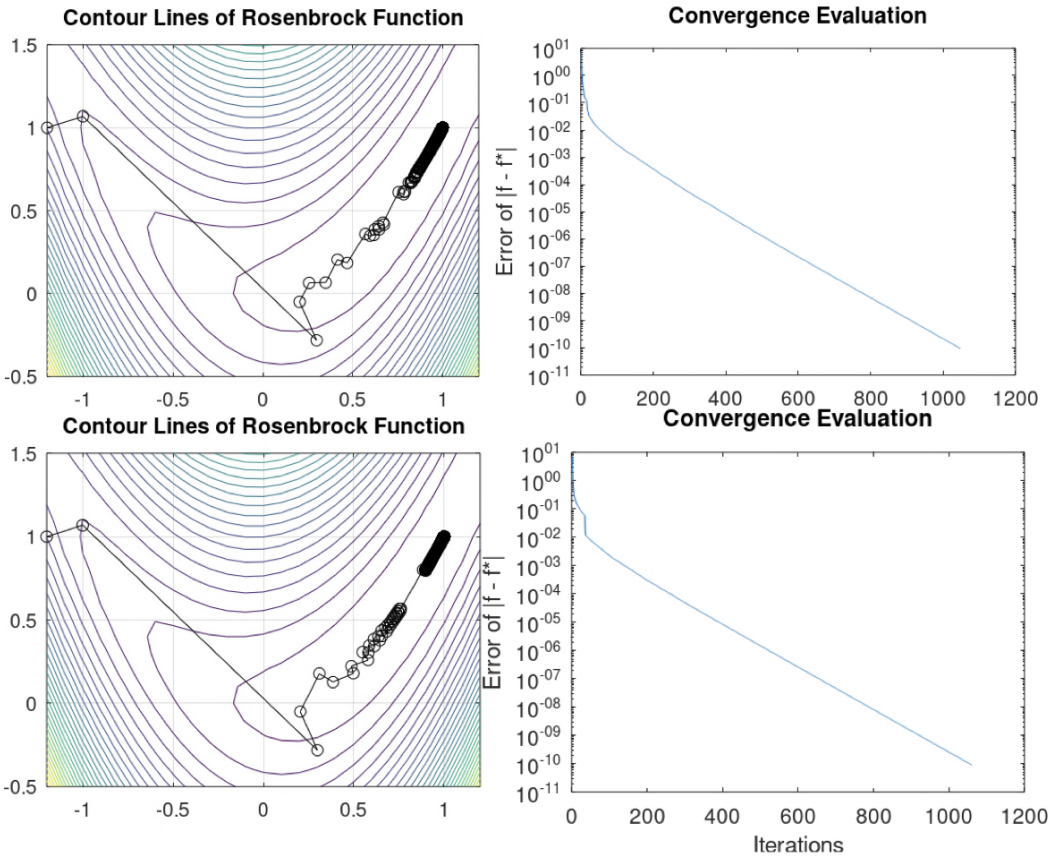


Figure 23: Comparison between Armijo and Wolfe conditions,programming in a native way

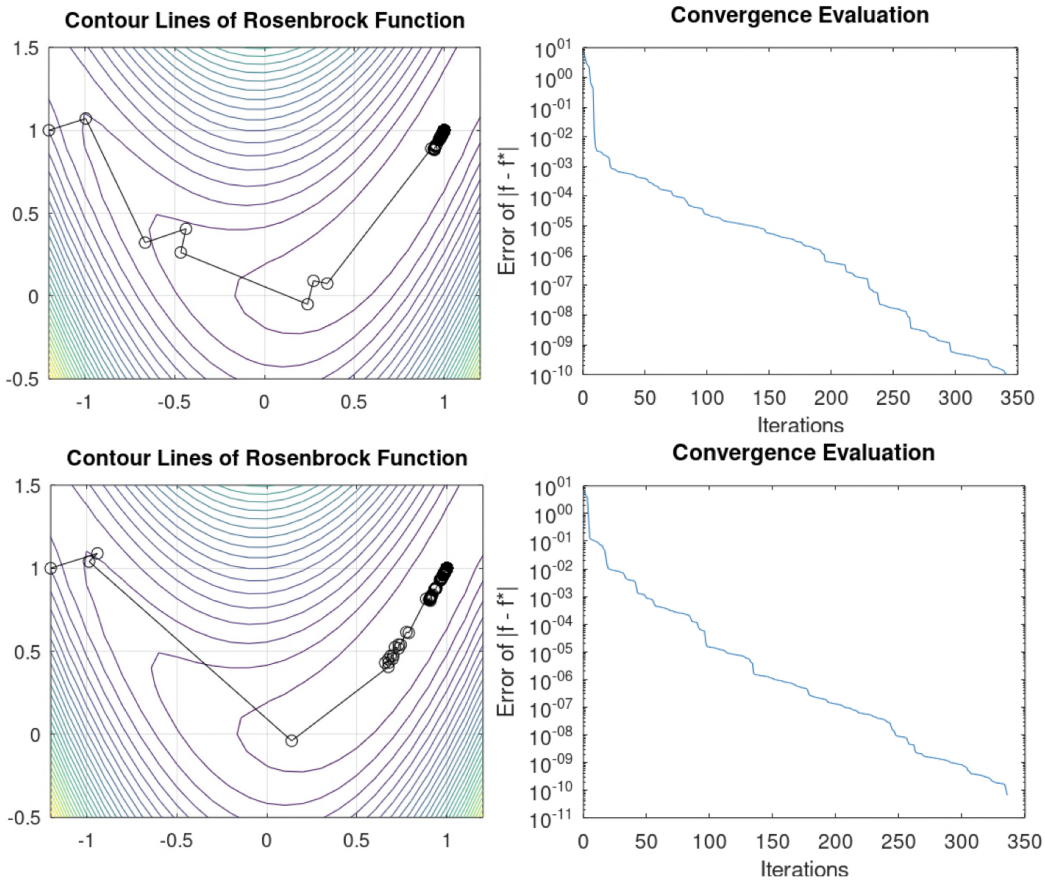


Figure 24: Comparison between Strong Wolfe conditions and Weak Wolfe Conditions, programming with zoom

6.2 Proximal Algorithm: Comparison the performances under different settings

In subsection 2.3.2 of this technical report, we introduced the Proximal Algorithms, combined with several acceleration method. Here i give two results to compare the performances under different settings, in order to gain a direct understanding of these methods. Given $b \in R^n, A \in R^{n \times p}$, the lasso criterion is given by

$$\begin{aligned} f(x) &= \frac{1}{2}\|Ax - b\|^2 + \nu\|x\|_1 \\ &= g(x) + h(x) \end{aligned}$$

Then the proximal mapping for $h(x) = \nu\|x\|_1$ is

$$prox_{h,t}(x) = argmin_z\{\frac{1}{2t}\|x - z\|_2^2 + \nu\|z\|_1\}$$

The solution to this problem is: $x^\star = S_\nu(x)$, where S_ν is the soft-thresholding operator:

$$S_\nu(x) = \begin{cases} x_i - \nu, & x_i > \nu \\ 0, & -\nu < x_i < \nu \\ x_i - \nu, & x_i < -\nu \end{cases}$$

The proximal update is then given by:

$$\begin{aligned} x^+ &= S_\nu(x - \alpha \bigtriangledown g(x)) \\ &= S - \nu(x - \alpha A^T(Ax - b)) \end{aligned}$$

This is called the iterative soft thresholding algorithm(ISTA).

Fast iterative shrinkage-thresholding algorithm(FISTA):

$$\begin{aligned} x^{k+1} &= prox_{\alpha_t h}(y^k - \alpha_k \bigtriangledown f(y^k)) \\ y^{k+1} &= x^{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(x^{k+1} - x^k) \end{aligned}$$

where $y^0 = x^0, \theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$, adopt the momentum coefficients originally proposed by Nesterov'83.

In my opinion, these methods are ingenious and progressive. Heavy-ball method is pysically intuitive, Nesterov's mehod split the structure of Heavy-ball method and find the magic 3 (which is the smallest constant that guarantes $O(\frac{1}{\tau^2})$ convergence) to achieve an acceleration. Beck and Teboulle equip the GD part of Nesterov's method with proximal gradient descent (we have illustrated that proximal gradient method get a faster convergence than GD). These methods and modifications are natural, but the search for the feasible coefficients is complex.

Let $\tau_\alpha(x) := prox_{\alpha g}(x - \alpha \bigtriangledown f(x))$.

Algorithm 17 Algorithm Backtracking line search proximal gradient method

Initialize $0 < \beta < 1$

while $f(\tau_{\alpha_t}(x^t)) > f(x^t) - < \bigtriangledown f(x^t), x^t - \tau_{\alpha_t}(x^t) > + \frac{1}{2\alpha_t}\|\tau_{\alpha_t}(x^t) - x^t\|_2^2$

do $\alpha_t \leftarrow \beta\alpha_t$

$$x^{k+1} = prox_{\alpha_t h}(y^k - \alpha_k \bigtriangledown f(y^k))$$
$$y^{k+1} = x^{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(x^{k+1} - x^k)$$

where $y^0 = x^0, \theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$

Gradient scheme: restart when $< \bigtriangledown f(y^{t-1}), x^t - x^{t-1} >$ greater than 0, momentum lead us towards a bad direction.

$$x^0 \leftarrow x^t$$
$$y^0 \leftarrow y^t$$
$$\theta_o = 1$$

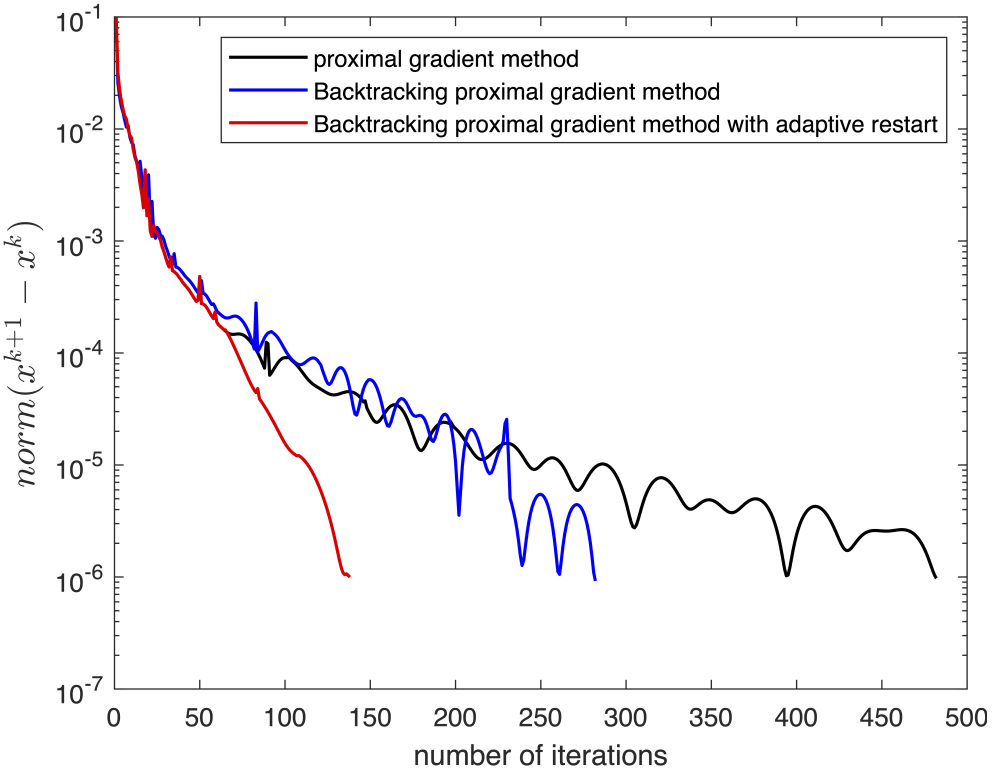


Figure 25: Comparison the performance of Proximal gradient method on Lasso Problem, combined with different settings

7 Ending

It is a great experience to do an internship in such a laboratory. The internship is not easy, but i was hard-working and tried my best to do it well. This internship helped me i)learned and explored new techniques ii)improved my language ability iii)also makes some friends. Benefit from this internship experience, i know how to choose the optimization algorithm for distinct problems in the future work. Besides, I have a deeper understanding of Machine Learning, since optimization algorithm is the core of Machine Learning. I think the biggest acquisition is that I gain an interest in math, programming and algorithms. I learned the mathematical verification of some algorithms and implemented them manually, combined with some modifications or new techniques. I found the importance of math, programming and performance of related algorithms for the future industry. Due to this, I wish i could learn more about programming and algorithms. I have formed my future career plan and for this semester ICM3A, I will learn major computer science. Last but not least, I would like to express my gratitude to my school instructor and company instructor. Without you instruction and support, i could not have completed my internship. Thank you.

A Convergence Proof

The unaugmented Lagrangian L_0 has a saddle point.Let $(x^\star, z^\star, y^\star)$ be the saddle point for L_0, and define

$$V^k = (1/\rho)\|y^k - y^\star\|_2^2 + \rho\|B(z^k - z^\star)\|_2^2,$$

We will see that $V^k$is a $Lyapunovfunction$ for the algorithm. Decomposed the proof into three inequalities:

$$\begin{aligned}
&p^\star - p^{k+1} \leq y^{\star\mathrm{T}} r^{k+1},\\
&p^{k+1} - p^\star \leq -(y^{k+1})^{\mathrm{T}} r^{k+1} - \rho(B(z^{k+1} - z^k))^{\mathrm{T}}(-r^{k+1} + B(z^{k+1} - z^\star)),\\
&V^{k+1} \leq V^k - \rho\|r^{k+1}\|_2^2 - \rho\|B(z^{k+1} - z^k)\|_2^2.
\end{aligned}$$

Proof of the first inequality: Since $(x^\star, z^\star, y^\star)$ a saddle point for L_0, we have

$$L_0(x^\star, z^\star, y^\star) \leq L_0(x^{k+1}, z^{k+1}, y^\star).$$

Bring the formula $L_0(x, z, y) = f(x)+g(z)+y^{\mathrm{T}}(Ax-b)$,and using the constrained condition $Ax^\star + Bz^\star = c$,, this can be written as

$$p^\star - p^{k+1} \leq y^{\star\mathrm{T}} r^{k+1},$$

Proof of the second inequality: based on assumption 1, f and g are closed, proper and convex. Thus f , g and the L_ρ are subdifferentiable. By diffinition, x^{k+1} minimizes $L_p(x, z^k, y^k)$.The necessary and sufficient optimality condition is

$$0 \in \partial L_p(x^{k+1}, z^k, y^k) = \partial f(x^{k+1}) + A^{\mathrm{T}} y^k + \rho A^T(Ax^{k+1} + Bz^k - c).$$

Since $y^{k+1} = y^k + \rho r^{k+1}$, we can plug in $y^k = y^{k+1} - \rho r^{k+1}$ and rearrange to obtain

$$0 \in \partial f(x^{k+1}) + A^T(y^{k+1} - \rho B(z^{k+1} - z^k)).$$

This implies that x^{k+1} minimizes $f(x) + (y^{k+1} - \rho B(z^{k+1} - z^k))^{\mathrm{T}} Ax$. A similar argument shows that z^{k+1}minimizes $g(z) + y^{(k+1)\mathrm{T}} Bz$.. It follows that

$$f(x^{k+1} + (y^{k+1} - \rho B(z^{k+1} - z^k))^{\mathrm{T}} Ax^{k+1} \leq f(x^\star) + (y^{k+1} - \rho B(z^{k+1} - z^k))^{\mathrm{T}} Ax^\star$$

and that

$$g(z^{k+1}) + y^{(k+1)\mathrm{T}} Bz^{k+1} \leq g(z^\star) + y^{(k+1)\mathrm{T}} Bz^\star.$$

Adding the two inequalities, the second inequality proved. The third inequality is deduced from the first two inequalities. By adding them and multiplying through by two, we have

$$2(y^{k+1} - y^\star)^{\mathrm{T}} r^{k+1} - 2\rho(b(z^{k+1} - z^k))^{\mathrm{T}} r^{k+1} + 2\rho(B(z^{k+1} - zk))^{\mathrm{T}}(B(z^{k+1} - z^\star)) \leq 0.$$

We begin by rewriting the first term. Substituting $y^{k+1} = y^k + \rho r^{k+1}$gives

$$2(y^k - y^\star)^{\mathrm{T}} r^{k+1} + \rho\|r^{k+1}\|_2^2 + \rho\|r^{k+1}\|_2^2,$$

and substituting $r^{k+1} = (1/\rho)(y^{k+1} - y^k)$ in the first two terms gives

$$\frac{2}{\rho}(y^k - y^\star)^{\mathrm{T}}(y^{k+1} - y^k) + (\frac{1}{\rho}\|y^{k+1} - y^k\|_2^2 + \rho\|r^{k+1}\|_2^2.$$

Since $y^{k+1} - y^k = (y^{k+1} - y^\star) - (y^k - y^\star)$, this can be written as

$$\frac{1}{\rho}(\|y^{k+1} - y^\star\|_2^2 - \|y^k - y^\star\|_2^2) + \rho\|r^{k+1}\|_2^2.$$

Then the remaining terms,i.e.,

$$\rho\|r^{k+1}\|_2^2 - - 2\rho(b(z^{k+1} - z^k))^{\mathrm{T}}r^{k+1} + 2\rho(B(z^{k+1} - zk))^{\mathrm{T}}(B(z^{k+1} - z^\star)),$$

Similarly, we substitute $z^{k+1} - z^k = (z^{k+1} - z^\star) - (z^k - z^\star)$, in the last term gives

$$\rho\|r^{k+1} - B(z^{k+1} - z^k)\|_2^2 + \rho\|B(z^{k+1} - z^k)\|_2^2 + 2\rho(B(z^{k+1} - z^k))^{\mathrm{T}}(B(z^k - z^\star)),$$

and substituting $z^{k+1} - z^k = (z^{k+1} - z^\star) - (z^k - z^\star)$ in the last two terms, we get

$$\rho\|r^{k+1} - B(z^{k+1} - z^k)\|_2^2 + \rho(\|B(z^{k+1} - z^\star)\|_2^2 - \|B(z^k - z^\star\|_2^2).$$

With the previous step, this implies that the adding can be written as

$$V^k - V^{k+1} \geq \rho\|r^{k+1} - B(z^{k+1} - z^k)\|_2^2.$$

It now suffices to show that the middle term $-2\rho r^{(k+1)T}(B(z^{k+1} - z^k))$ is positive. To see this, recall that z^{k+1} minimizes $g(z) + y^{(k+1)T}Bz$ and z^k minimizes $g(z) + y^{k\mathrm{T}}Bz$ and z^k minimizes $g(z) + y^{k\mathrm{T}}Bz$, so we can add

$$g(z^{k+1}) + y^{(k+1)T}Bz^{k+1} \leq g(z^k) + y^{(k+1)\mathrm{T}}Bz^k$$

and

$$g(z^k) + y^{k\mathrm{T}}bz^k \leq g(z^{k+1}) + y^{k\mathrm{T}}Bz^{k+1}$$

adding them to get

$$(y^{k+1} - y^k)^{\mathrm{T}}(B(z^{k+1} - z^k)) \leq 0.$$

Substituting $y^{k+1} - y^k = \rho r^{k+1}$ gives the result, since $\rho > 0$, Q.E.D.

B Primary Octave Code

B.1 Section 5.3.1

```
function alpha = armijo(s, sigma, beta, x_k, y_k, F, GF, d, maxit = 10)
        m = 0;
        while m < maxit
            f1 = F(x_k, y_k);
            f2 = F(x_k + beta^m * s * d(1), y_k + beta^m * s * d(2));
            diff = f1 - f2;
            lowerbound = -sigma * beta^m * s * GF(x_k, y_k)' * d;
            if diff >= lowerbound
                alpha = beta^m * s;
                break
            end
            m += 1;
        end
        alpha = beta^m * s;
end
function x_{k+1}= steepestdescent(x_k, alpha_k, GF_k)
        x_{k+1} = x_k - alpha_k * GF_k;
end
```

B.2 Section 5.3.2

```
function x=TrustRegion(Delta,delta,eta,x0)
        x=x0;
        g=[-400*x(1)*x(2) + 400 * x(1)^3 + 2 * x(1) - 2; -200 * x(1)^2 + 200*x(2)];
        B=[1200*x(1)^2-400*x(2)+2 -400*x(1); -400*x(1) 200 ];
        while g' * g > 1e - 8
            x'
            B=[1200*x(1)^2 - 400 * x(2)+2 -400*x(1); -400*x(1) 200 ];
            p=Dogleg(B,g,delta);
            rho=((100*(x(2)-x(1)^2)^2 + (1 - x(1))^2) -...
                    (100*(x(2)+p(2)-(x(1)+p(1))^2)^2 + (1 - x(1) - p(1))^2))/...
                    (-g' * p - 0.5 * p' * B * p);
            if rho<0.25
                    delta=0.25*delta;
            else
                    if rho > 0.75 && abs(p' * p - delta^2) <1e-8
                        delta=min(2*delta, Delta);
                    end
            end
            if rho > eta
```

```
            x=x+p;
        end
        g=[-400*x(1)*x(2) + 400*x(1)^3 + 2 * x(1) - 2; -200 * x^2(1) + 200*x(2)];
    end
end
function p=Dogleg(B, g, delta)
    p=-B\g;
    if p' * p - delta^2 >1e-8
        pu=-(g' * g)/(g'*B*g)*g;
        if pu'*pu-delta^2 >1e-8
            p=pu/(pu' * pu)^2 * delta;
        else
            a=(p-pu)'*(p-pu);
            b=2*(p-pu)' * pu - 2 * (p - pu)'*(p-pu);
            c=pu' * pu + (p - pu)' * (p - pu) - 2 * (p - pu)' * pu - delta^2;
            tau=(-b+sqrt(b^2 - 4 * a * c))/(2*a);
            p=pu+(tau-1)*(p-pu);
        end
    end
end
```

B.3 Section 5.3.3

```
function x=CG(A, b, x0)
    x=x0
    r0=A*x-b
    p=-r0; k=0; [n,n]=size(A);
    while r0' * r0 > 1e - 8&&k < n
        ap=A*p;
        alpha = r0' * r0/(p' * ap)
        x=x+alpha*p
        r=r0+alpha*ap
        beta=(r' * r)/(r0' * r0)
        p=-r+beta*p
        k=k+1;
        r0=r;
    end
end
```

B.4 Section 5.3.4

```
function x=DFP(B0,x0,eps,Gf)
        k=0;gf0=Gf(x0);B=B0;
        H0=B0;
        while gf0' * gf0 > eps^2
                p=-H0*gf0;
                x=x0 + -p' * gf0/(p' * B0 * p) * p;
                gf=Gf(x);
                s=x-x0; y=gf-gf0; rho = 1/(y' * s);
                B0 = B0 - B0 * rho * s * y';
                B=B0; B0 = s' * B0; B0=rho*y*B0;
                B=B - B0 + rho * y * y';
                H=H0 - (H0 * y) * (y' * H0)/(y' * H0 * y) + s * s'/(y' * s);
                k=k+1;
                gf0=gf;
                x0=x;
                x'
                B0=B;
                H0=H;
        end
        k
end
function x=BFGS(H0,x0,eps,Gf)
        k=0;gf0=Gf(x0);H=H0; B0=H0;
        while gf0' * gf0 > eps^2
                p=-H0*gf0;
                x=x0 + -p' * gf0/(p' * B0 * p) * p;
                gf=Gf(x);
                s=x-x0; y= gf - gf0; rho = 1/(y' * s);
                H0=H0 - H0 * rho * y * s';
                H=H0; H0=y' * H0; H0=rho*s*H0;
                H=H - H0 + rho * s * s';
                B=B0 - (B0 * s) * (s' * B0)/(s' * B0 * s) + y * y'/(y' * s);
                k=k+1;
                gf0=gf;
                x0=x;
                x'
                H0=H;
                B0=B;
        end
        k
end
```

References

[1] S.Boyd, N.Parikh, E.Chu, B.Peleato and J.Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," Foundations and Trends in Machine Learning (2010).

[2] Numerical optimization: 2nd Edition by Jorge Nocedal Stephen J. Wright, Springer, 2006

[3] Nonlinear Programming: 3rd Edition by Dimitri P. Bertsekas, Athena Scientific, 2016

[4] Lectures on Convex Optimization: 2rd Edition by Yurii Nesterov, Springer, 2018

[5] Convex Optimization: by Stephen Boyd and Lieven Vandenberghe, Cambridge University Press, 2004

[6] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," Computers and Mathematics with Applications, vol. 2, pp. 17–40, 1976.

[7] Jean Jacques Moreau, "Functions convexes duales et points proximaux dans un espace hilbertien" Comptes rendus hebdomadaires des séances de l'académie des sciences, Elsevier, 1962, 225, pp.2897-2899. hal-01867195