SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Tuning Linear Programming Solvers for Query Optimization

Sarra Ben Mohamed

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Tuning Linear Programming Solvers for Query Optimization

Anpassung von Linear Programming Solvern für Anfrageoptimierung

Author:	Sarra Ben Mohamed
Supervisor:	Prof. Dr. Thomas Neumann
Advisor:	Altan Birler
Submission Date:	15/10/2023

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15/10/2023 Sarra Ben Mohamed

Acknowledgments

Abstract

Contents

1 Introduction

Our aim with this project is to investigate and compare different methods and techniques to solve small linear programming problems representing among others the problem of cardinality estimation. A way to estimate realistic and useful upper bounds of query sizes is through linear programming. Studies have shown that cardinality estimation is the major root of many issues in query optimization [Ngo22], which is why we want a practical estimate to choose the best from data plans to run efficient queries. For this purpouse, we will introduce a formal description of the cardinality estimation problem, represent it in the form of a packing linear programming problem with the intention of maximizing the size of the query under some constraints. The result is hundreds of relatively small LP that we collect in datasets and solve them with different methods and algorithms. We then draw conclusions based on the results of our experiments, benchmarks and the previous work done on similar packing LP problems. This should guide us into constructing a thorough analysis of the particularities of these LP problems, what's unique about their structure and if their solution process follows any patterns. We then discuss and draw hypotheses on the ways this analysis can be exploited to further optimize the solution process: which methods or combination of methods deliver the best time and memory complexity.

2 Related work

2.1 Background

In this chapter we will talk about optimization, in particular the field of linear programming. We will elaborate on the most widely used algorithms and techniques to tackle this problem, and we present some use cases and benchmarks. A major use case of linear programming solvers is cardinality estimation, which is a crucial step in the pipeline of query optimization. We will present the background and related work needed to understand our contribution.

2.1.1 Linear Programming

Informally, Linear Programming (LP) is a method to calculate the best possible outcome from a given set of requirements. A concrete real-world application of such a method is for instance aiming to maxmize profit in a business, given some constraints on your variables like raw material availability, labor hours, etc.

Formally, LP is a mathematical modeling technique in which a linear function (called the objective function) $z : \mathbb{R}^n \to \mathbb{R}$ is maximized or minimized when subject to a set of linear constraints or inequalities. A maximization LP problem is then defined as:

$$\begin{aligned} \text{Maximize} \quad & z = \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{2.1}$$

Where n is the number of decision variables and m is the number of constraints: $\mathbf{x} \in \mathbb{R}^n$ is the column vector of decision variables. $\mathbf{c} \in \mathbb{R}^n$ is the column vector of coefficients in the objective function. $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the coefficient matrix in the constraints. $\mathbf{b} \in \mathbb{R}^m$ is the column vector of the right-hand sides of the constraints. In the following sections, we focus on LP problems that are maximization problems and we primarily use the matrix representation of the problem.

To derive the setting for our contribution, we also explore a special instance of LP problems called packing LP.

Packing LP

One LP problem class that we are dealing with is called the packing LP problem. It is a special instance where: $\mathbf{c} = \mathbf{b} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}$. Our specific problem is then expressed as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{n} x_j \\ \text{subject to} \quad & \\ & \mathbf{Ax} \le \mathbf{1}_m, && \qquad (2.2) \\ & x_i \ge 0, \qquad\qquad i = 1, \dots, n && \qquad (2.3) \end{aligned}$$

Where $\mathbf{1}_m = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

This specific class of LPs has a simple structure that we can exploit, see Chapter **??**, to further optimize our implementation.

2.1.2 Duality

The duality theorem is an interesting result in linear programming, that states that very instance of maximization problem has a corresponding minimization problem called its dual problem. The two problems are linked in an interesting way: if one problem has an optimal solution, then so does the other, and their optimal solutions are equal.

For instance, consider the primal-dual pair LP:

$$\begin{array}{rlcrl} \text{maximize} & \mathbf{c}^T\mathbf{x} & & \text{minimize} & \mathbf{b}^T\mathbf{y} \\ \text{subject to} & \mathbf{Ax} \le \mathbf{b} & \longrightarrow & \text{subject to} & \mathbf{A}^T\mathbf{y} \ge \mathbf{c} \\ & \mathbf{x} \ge 0 & & & \mathbf{y} \ge 0 \end{array}$$

2.1.3 Geometric Interpretation

In this part let's assume for simplicity that we have two decision variables in our LP problem, i.e. $\mathbf{x} \in \mathbb{R}^2$ is a two dimensional vector. This assumption will allow us to plot our problem on a 2D plane. An example is shown in figure **??**. The linear programming problem 2.1 can be understood geometrically as follows: Each inequality in the set of constraints is represented by a line. Therefore, the feaible region χ of any LP problem can be described as the intersection of a finite number of hal-spaces, or lines, which is

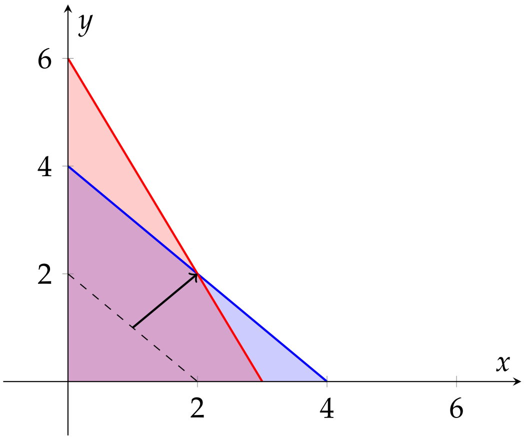

Figure 2.1: Feasible region of the LP problem

called a polyhedron, or in our 2D case, a polytope. The extreme points (corners) of a polytope are called vertices. They lie in the intersection of at most n constraints, that define χ. If more than n lines, it is called degenerate.