Distributed Optimization

Christoffer Sloth

This lecture note is based on [4] that describes dual decomposition and [2] (primarily Chapter 3 and Chapter 7) that presents the distributed optimization method called Alternating Direction Method of Multipliers (ADMM). Additionally, [3, 1] are used for the preliminaries in optimization. Two different distributed optimization methods are presented and subsequently compared. Both algorithms have been used in recent research. The algorithm ADMM was invented in the 1970s, but has received tremendous interest in recent years, see e.g. [5, 6]. After reading this note, you must

- have knowledge in convex optimization,
- have knowledge in methods of multipliers,
- have knowledge in distributed optimization,
- be capable of applying *dual decomposition* and *alternating direction method of multipliers* on a given optimization problem, and distribute its solution.

1 Introduction to Optimization

This section gives an introduction to optimization, and presents the methods that underlay dual decomposition and ADMM. In particular, the section gives a problem formulation, and describes duality and the method of multipliers.

1.1 Problem Setup

We consider the following optimization problem, called the primal problem.

Problem 1 (Primal Problem)**.**

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

with variable $x \in \mathcal{D} \subseteq \mathbb{R}^n$*, where* $\mathcal{D} = \cap_{i=1}^m \mathrm{dom}(f_i) \bigcap \cap_{i=1}^p \mathrm{dom}(h_i) \neq \emptyset$.

We denote the optimal value of (1) by $p^* \in \mathbb{R}$, and recall that we search for a value of x such that $f_0(x)$ is minimal.

Problem 1 is a very general optimization problem, which may be difficult to solve. Therefore, several classes of optimization problems have been defined, including convex optimization problems. Problem 1 is a convex optimization problem if

1. The functions f_i are convex for $i = 0, \ldots, m$.

2. The functions h_i are affine for $i = 1, \ldots, p$.

Recall the following definition of a convex function.

Definition 1 (Convex Function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for all $x_1, x_2 \in \mathbb{R}^n$*

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \qquad \forall \alpha \in [0,1].$$

Problem 1 is often reformulated as an optimization problem on the following form

$$\begin{aligned} &\text{minimize } f_0^*(x) \\ &\text{subject to } h_i(x) = 0, \qquad i = 1, \ldots, p \end{aligned} \tag{2}$$

The two optimization problems look very similar, except that (1) contains inequality constraints. However, inequality constraints can be incorporated into the cost. This is shown in the next example.

Example 1. *We consider the following very simple constraint optimization problem*

$$\begin{aligned} &\text{minimize } x \\ &\text{subject to } x \in \begin{bmatrix} -1 & 1 \end{bmatrix}. \end{aligned}$$

It is seen that the constraints on x can also be phrased as

$$\begin{aligned} \text{minimize } & x \\ \text{subject to } & -x - 1 \leq 0 \\ & x - 1 \leq 0. \end{aligned}$$

We define the indicator function $h_+ : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$

$$h_+(x) = \begin{cases} 0 & \textit{if } x \in \begin{bmatrix} -1 & 1 \end{bmatrix} \\ \infty & \textit{otherwise.} \end{cases}$$

Consider the optimization problem

$$\text{minimize } x + h_+(x).$$

This optimization problem has the same optimal solution as the original problem. This is also clear from Figure 1.

1.2 Duality

This section gives a short introduction to duality and is based on Chapter 5 in [3]. The essence of the section is that a dual problem can be associated to a general optimization problem. A solution to the dual problem gives a lower bound on the solution of the primal problem. In certain cases, the gap between the primal and dual solutions is zero (the duality gap is zero); thus, the optimal solution to the dual problem is also the optimal solution to the primal problem.

We associate a Lagrangian to the optimization problem (1).

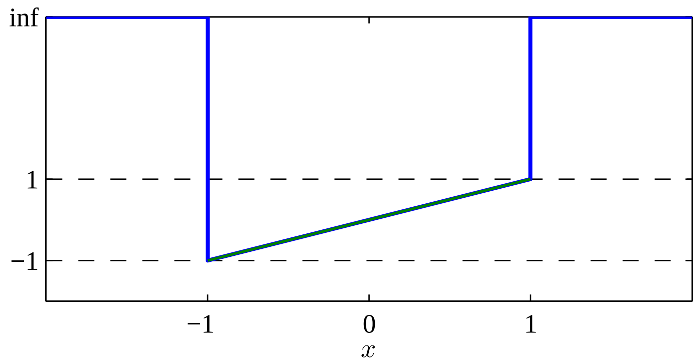

Figure 1: Illustration of the objective of the two optimization problems. The green line is only drawn between -1 and 1, because the constrained problem is infeasible outside this range.

Definition 2 (Lagrangian). *The Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ *associated to* (1) *is*

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

where λ_i *is the* Lagrange multiplier *associated with the* i^{th} *inequality constraint* $f_i(x) \leq 0$, *and* ν_i *is the Lagrange multiplier associated with the* i^{th} *equality constraint* $h_i(x) = 0$. *The vectors* λ *and* ν *are called the* dual variables *of problem* (1).

For fixed values of the Lagrange multipliers, we associate a so-called Lagrange dual function.

Definition 3 (Lagrange Dual Function). *The Lagrange dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ *is defined as the minimum of the Lagrangian over* x *for given values of* $\lambda \in \mathbb{R}^m$ *and* $\nu \in \mathbb{R}^p$,

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

It is seen that g is a concave function, since it is the point-wise minimum of a family of affine functions. The dual function $g(\lambda, \nu)$ gives a lower bound on the optimal value p^* of problem (1).

Lemma 1. *For any* $\lambda \in \mathbb{R}^m_{\geq 0}$ *and any* $\nu \in \mathbb{R}^p$ *we have*

$$g(\lambda, \nu) \leq p^*, \tag{3}$$

where g *is the dual function for an optimization problem* (1) *which has optimal value* p^*.

Proof. Suppose $\bar{x}$ is a feasible point of problem (1). Then

$$\sum_{i=1}^{m} \underbrace{\lambda_i}_{\lambda \geq 0} \underbrace{f_i(\bar{x})}_{\leq 0} + \sum_{i=1}^{p} \nu_i \underbrace{h_i(\bar{x})}_{=0} \leq 0.$$

This implies that

$$L(\bar{x}, \lambda, \nu) = f_0(\bar{x}) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(\bar{x}) + \sum_{i=1}^{p} \nu_i h_i(\bar{x})}_{\leq 0} \leq f_0(\bar{x}).$$

Thus,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq L(\bar{x}, \lambda, \nu) \leq f_0(\bar{x}).$$

From this, (3) follows directly. $\square$

The pair (λ, ν) with $\lambda \geq 0$ and $(\lambda, \nu) \in \mathrm{dom}(g)$ is said to be dual feasible. We are of course interested in obtaining the best (maximum) lower bound on the optimal value p^*. To obtain this, we define the Lagrange dual problem.

Definition 4 (Lagrange Dual Problem)**.** *Let g be the Lagrange dual function associated to* (1)*, and $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^p$ be Lagrange multipliers. Then the Lagrange dual problem associated to problem* (1) *is defined as*

$$\begin{aligned} &\text{maximize } g(\lambda, \nu) \\ &\text{subject to } \lambda \geq 0. \end{aligned} \tag{4}$$

We denote the optimal value of the Lagrange dual problem by d^*.

The property

$$d^* \leq p^*$$

is referred to as weak duality.

Definition 5 (Optimal Duality Gap)**.** *Let p^* be the optimal solution to the primal problem* (1) *and d^* be the optimal solution to the dual problem* (4)*. Then the optimal duality gap is defined as*

$$p^* - d^*.$$

It is of course desired that the primal and dual problems attain the same optimal solution. This property is called strong duality.

Definition 6 (Strong Duality)**.** *Let p^* be the optimal solution to the primal problem* (1) *and d^* be the optimal solution to the dual problem* (4)*. We say that strong duality holds if*

$$d^* = p^*.$$

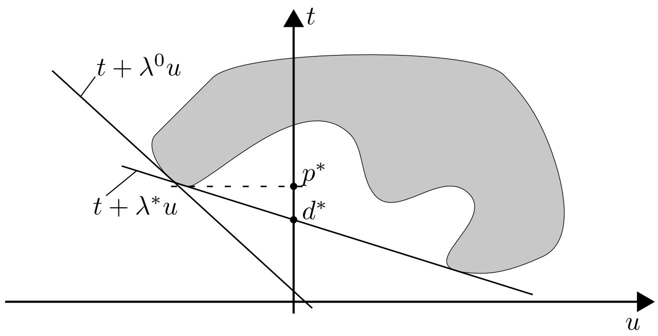

Figure 2: Illustration of duality, where t is the cost and u must be less than zero.

We say that the primal problem (1) is strictly feasible if there exists $x \in \mathrm{relint}(D)$ (point in the relative interior) such that

$$\begin{aligned} f_i(x) < 0, &\qquad i = 1, \ldots, m \\ h_i(x) = 0, &\qquad i = 1, \ldots, p. \end{aligned}$$

Slater's condition is a sufficient condition for strong duality.

Theorem 1 (Slater's Condition)**.** *If the primal problem is convex and strictly feasible. Then we have strong duality.*

1.3 Method of Multipliers

The method of multipliers is a method for solving an optimization problem by the use of an augmented Lagrangian function rather than the Lagrantian introduced in Definition 2. In the following, we consider the optimization problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } Ax = b, \end{aligned} \tag{5}$$

where $x \in \mathbb{R}^n$ is the variable, $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, $A \in \mathbb{R}^{p \times n}$, and $b \in \mathbb{R}^p$.

The augmented Lagrangian for (5) is

$$L_\rho(x, \lambda) = f(x) + \lambda^T(Ax - b) + \frac{\rho}{2}||Ax - b||_2^2$$

where $\rho \in \mathbb{R}_{>0}$. Notice that the "normal" Lagrangian is $L_0(x, \lambda)$.

One could question the validity of augmenting the Lagrangian; thus, we pose the optimization problem

$$\begin{aligned} &\text{minimize } f(x) + \frac{\rho}{2}||Ax - b||_2^2 \\ &\text{subject to } Ax = b. \end{aligned} \tag{6}$$

This optimization problem has the same optimal value as (5), since for any feasible point $\bar{x}$, $A\bar{x} - b = 0$. Therefore, it is concluded that the augmentation of the Lagrangian is valid. Next, we write an algorithm for solving the optimization problem in (5).

The Lagrange dual function $g : \mathbb{R}^p \to \mathbb{R}$ associated to the augmented Lagrangian $L_\rho(x, \lambda)$ is defined as

$$\begin{aligned} g(\lambda) &= \inf_x L_\rho(x, \lambda) \\ &= \inf_x \left(f(x) + \lambda^T (Ax - b) + \frac{\rho}{2} ||Ax - b||_2^2 \right). \end{aligned}$$

Finally, the Lagrange dual problem becomes

$$\text{maximize } g(\lambda). \tag{7}$$

An optimization problem is solved via method of multipliers by the following algorithm.

Algorithm 1 (Method of Multipliers).
Input: *Cost function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, specification of equality constraints $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$, initial value for Lagrange multiplier λ_{init}, and update step $\rho \in \mathbb{R}_{>0}$.*
Output: *x^* and λ^*.*
Procedure:

0. *Initialization: Let $k := 0$ and $\lambda^k = \lambda_{\text{init}}$.*
1. *Find primal variable $x^{k+1} := \arg\min_x L_\rho(x, \lambda^k)$.*
2. *Update dual variable $\lambda^{k+1} := \lambda^k + \rho(Ax^{k+1} - b)$.*
3. *If $|L_\rho(x^{k+1}, \lambda^k) - f(x^{k+1})| > \epsilon$ then $k := k + 1$, and go to step 1.*
4. *Output x^{k+1} and λ^k.*

It is clear that the optimization problem (5) is primal feasible if a minimizer x^* satisfies

$$Ax^* - b = 0.$$

Dual feasibility comes from the definition of the dual function, and is

$$\nabla_x f(x^*) + A^{\mathrm{T}} \lambda^* = 0.$$

Per definition x^{k+1} minimizes the augmented Lagrangian, i.e.,

$$\begin{aligned} 0 &= \nabla_x L_\rho(x^{k+1}, \lambda^k) \\ &= \nabla_x \left(f(x^{k+1}) + (\lambda^k)^{\mathrm{T}} (Ax^{k+1} - b) + \frac{\rho}{2} ||Ax^{k+1} - b||_2^2 \right) \\ &= \nabla_x f(x^{k+1}) + A^{\mathrm{T}} \lambda^k + \frac{\rho}{2} ((Ax^{k+1} - b)^{\mathrm{T}} A + A^{\mathrm{T}} (Ax^{k+1} - b)) \\ &= \nabla_x f(x^{k+1}) + A^{\mathrm{T}} (\lambda^k + \rho(Ax^{k+1} - b)) \\ &= \nabla_x f(x^{k+1}) + A^{\mathrm{T}} \lambda^{k+1}. \end{aligned}$$

It is seen that the choice of step length ρ in the algorithm ensures dual feasibility. It should also be noted that this problem is not separable; hence, it is not appropriate for distributed optimization. However, it has good robustness and convergence properties. This is illustrated in the next example.

Example 2. *We consider the following optimization problem*

$$\begin{aligned} &\text{maximize } 2x + 3y \\ &\text{subject to } x^2 + 2y^2 - 34 = 0. \end{aligned} \tag{8}$$

The problem is first solved using the Lagrangian and then with the augmented Lagrangian. The Lagrangian for the problem is

$$L(x, y, \lambda) = -2x - 3y + \lambda(x^2 + 2y^2 - 34).$$

The dual problem is solved with initial Lagrange multiplier $\lambda^0 = 1$ *and the algorithm is terminated when the difference between the primal and dual solutions is less than* $10 \cdot 10^{-3}$*. Figure 3 illustrates the Lagrange multiplier and residual* $(x^2 + 2y^2 - 34)$ *for each step. The augmented Lagrangian for the problem is*

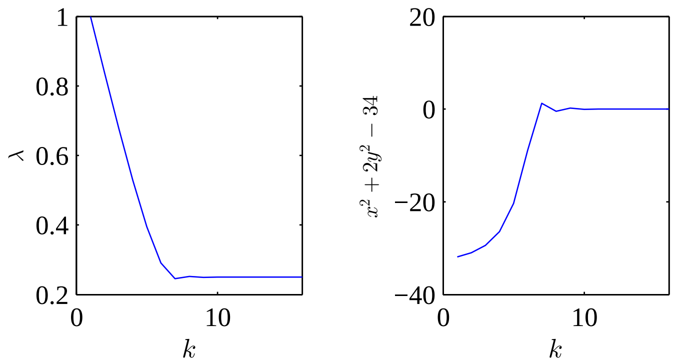

Figure 3: Lagrange multiplier and residual for each step of the optimization problem using "normal" Lagrangian.

$$L(x, y, \lambda) = -2x - 3y + \lambda(x^2 + 2y^2 - 34) + \frac{\rho}{2}(x^2 + 2y^2 - 34)^2.$$

The dual problem is again solved with initial Lagrange multiplier $\lambda^0 = 1$*,* $\rho = 1$*, and the algorithm is terminated when the difference between the primal and dual solutions is less than* $10 \cdot 10^{-3}$*. Figure 4 illustrates the Lagrange multiplier and residual for each step.*

It is seen from Figure 5 that both algorithms solve the optimization problem.

2 Distributed Optimization Methods

This section presents the two distributed optimization methods: Dual decomposition and Alternating Direction Method of Multipliers. The general idea is to split an optimization problem into smaller subproblems that can be solved individually after only little global information has been distributed. The key assumption that allows an optimization problem to be distributed is that the cost is additively separable. To simplify the presentation, it is assumed that the optimization problem is distributed into two subproblems (in general they

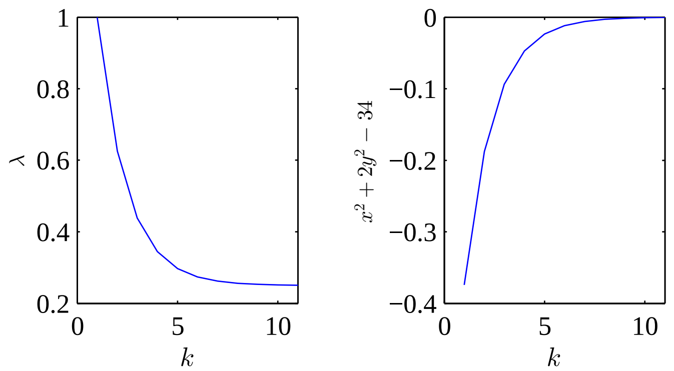

Figure 4: Lagrange multiplier and residual for each step of the optimization problem using augmented Lagrangian.

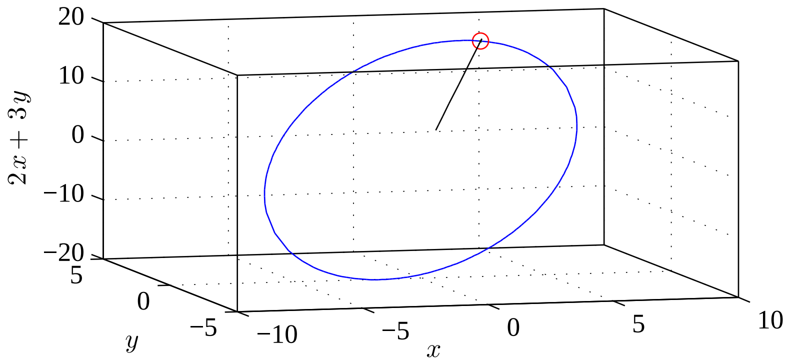

Figure 5: Illustration of the feasible set (blue line) projected to (x, y)-plane, optimal value (red circle), and solutions at each step of the algorithm (black and green lines).

may be distributed into an arbitrary number of subproblems), i.e., we consider optimization problems on the following form

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } f_1(x) + f_2(z) \leq 0 \\ &\qquad\qquad Ax + Bz = c \end{aligned} \tag{9}$$

where $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$ are variables, and $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$, $f_1 : \mathbb{R}^n \to \mathbb{R}$, $f_2 : \mathbb{R}^m \to \mathbb{R}$, $A \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times m}$, and $c \in \mathbb{R}^p$ are given.

2.1 Dual Decomposition

This subsection presents the dual decomposition method, and is based on [4]. To decompose an optimization problem, dual decomposition can be utilized.

We consider the following optimization problem

$$\begin{aligned} \text{minimize } & f(x) + g(z) \\ \text{subject to } & x \in \mathcal{X},\ z \in \mathcal{Z} \\ & f_1(x) + f_2(z) \leq 0. \end{aligned} \tag{10}$$

where $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$ are variables, $\mathcal{X}$ and $\mathcal{Z}$ are convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Finally, $f_1 : \mathbb{R}^n \to \mathbb{R}^p$ and $f_2 : \mathbb{R}^m \to \mathbb{R}^p$ are maps with convex coordinate functions.

To do dual decomposition, we form the Lagrangian

$$L(x, z, \lambda) = f(x) + g(z) + \lambda^T (f_1(x) + f_2(z)),$$

where $\lambda \in \mathbb{R}^p$ is a vector of Lagrange multipliers. The Lagrangian can be rewritten as

$$L(x, z, \lambda) = \big(f(x) + \lambda^T f_1(x)\big) + \big(g(z) + \lambda^T f_2(z)\big).$$

For a fixed value of λ, the dual function can be written as $\varphi(\lambda) = \varphi_1(\lambda) + \varphi_2(\lambda)$. The dual function can be found in two separate subproblem. To find $\varphi_1(\lambda)$ we solve the following problem

$$\begin{aligned} \text{minimize } & f(x) + \lambda^T f_1(x) \\ \text{subject to } & x \in \mathcal{X}. \end{aligned} \tag{11}$$

To find $\varphi_2(\lambda)$ we solve the following problem

$$\begin{aligned} \text{minimize } & g(z) + \lambda^T f_2(z) \\ \text{subject to } & z \in \mathcal{Z}. \end{aligned} \tag{12}$$

To update the Lagrange multipliers λ in connection with maximizing the dual problem, we need subgradients.

Definition 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$. A vector $g \in \mathbb{R}^n$ is a subgradient of f at $x \in \operatorname{dom}(f)$ if*

$$f(z) \geq f(x) + g^T (z - x) \qquad \forall z \in \operatorname{dom}(f).$$

Remark that if f is differentiable, then its gradient at x is a subgradient.

Now the dual decomposition follows the following algorithm.

Algorithm 2 (Dual Decomposition)**.**
Input: *Cost functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, specification of equality constraints $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$ and $c \in \mathbb{R}^p$, initial value for Lagrange multiplier λ_{init}, ϵ_{primal}, ϵ_{dual}, and update step $\rho \in \mathbb{R}_{>0}$.*
Output: *p^*, and λ^*.*
Procedure:

0. *Initialization: Let $k := 0$ and $\lambda^k = \lambda_{\text{init}}$.*

1. *Solve the subproblems*

$$\begin{aligned} \text{minimize } & f(x) + (\lambda^k)^T f_1(x) \\ \text{subject to } & x \in \mathcal{X}. \end{aligned}$$

and

$$\begin{aligned} &\text{minimize } g(z) + (\lambda^k)^T f_2(z) \\ &\text{subject to } z \in \mathcal{Z}. \end{aligned}$$

and denote the minimizers by $\bar{x}$ and $\bar{z}$.

2. *Update dual variables $\lambda^{k+1} := \lambda^k + \alpha_k(f_1(\bar{x}) + f_2(\bar{z}))$, where α_k is the step size.*

This procedure should be repeated until an optimum is attained.

2.2 Alternating Direction Method of Multipliers

We consider the following type of optimization problem

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } Ax + Bz = c, \end{aligned} \tag{13}$$

where $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$ are variables, and $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$, $A \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times m}$, and $c \in \mathbb{R}^p$ are given. We denote the optimal solution to the primal optimization problem (13) by p^*.

Notice that the cost is additively separable in x and z. Also we impose a convexity assumption to solve the problem.

Assumption 1. *The functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ are assumed to be convex.*

Recall from the previous chapter that the considered convex (primal) problem has the following form

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0, \qquad i = 1, \ldots, m \\ &\qquad\qquad\quad Ax = b, \end{aligned} \tag{14}$$

with $f_0, \ldots, f_m$ convex. The alternating direction method of multipliers is close related to the method of multipliers, where an augmented Lagrangian is utilized.

For the optimization problem (13), the augmented Lagrangian is

$$L_\rho(x, z, \lambda) = f(x) + g(z) + \lambda^T(Ax + Bz - c) + \frac{\rho}{2}||Ax + Bz - c||_2^2. \tag{15}$$

Instead of finding the Lagrange dual function by jointly minimizing the augmented Lagrangian over x and z, the Lagrangian is minimized by alternating between minimization over x and z. Thus the method applies the following

$$\begin{aligned} \bar{x} &= \arg\inf_x L_\rho(x, z, \lambda) \\ &= \arg\inf_x (f(x) + g(z) + \lambda^T(Ax + Bz - c) + \frac{\rho}{2}||Ax + Bz - c||_2^2) \\ &= \arg\inf_x (f(x) + \lambda^T Ax + \frac{\rho}{2}||Ax + Bz - c||_2^2) \\ \bar{z} &= \arg\inf_z (g(z) + \lambda^T Bz + \frac{\rho}{2}||A\bar{x} + Bz - c||_2^2). \end{aligned}$$

This implies that ADMM follows the next algorithm.

Algorithm 3 (Alternating Direction Method of Multipliers)**.**
Input: *Cost functions* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *and* $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$*, specification of equality constraints* $A \in \mathbb{R}^{p\times n}$*,* $B \in \mathbb{R}^{p\times m}$ *and* $c \in \mathbb{R}^p$*, initial value for Lagrange multiplier* λ_{init}*, initial value for* z*-variable* z_{init}*,* ϵ_{primal}*,* ϵ_{dual}*, and update step* $\rho \in \mathbb{R}_{>0}$*.*
Output: p^**, and* λ^**.*
Procedure:

0. *Initialization: Let* $k := 0$*,* $z^k = z_{\text{init}}$*, and* $\lambda^k = \lambda_{\text{init}}$*.*
1. *Find first part of primal variable* $x^{k+1} := \arg\min_x L_\rho(x, z^k, \lambda^k)$*.*
2. *Find second part of primal variable* $z^{k+1} := \arg\min_z L_\rho(x^{k+1}, z, \lambda^k)$*.*
3. *Update dual variable* $\lambda^{k+1} := \lambda^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$*.*
4. *If* $||r^{k+1}||_2 > \epsilon_{\text{primal}}$ *or* $||s^{k+1}||_2 > \epsilon_{\text{dual}}$ *then* $k := k + 1$ *and go to step 1.*
5. *Output* $f(x^{k+1}) + g(z^{k+1})$*, and* λ^{k+1}*.*

The algorithm converges under the following conditions.

Proposition 1 (Convergence)**.** *If*

1. *the 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 and*
2. *the unaugmented Lagrangian* L_0 *has a saddle point,*

then the ADMM iterates satisfy the following

- *Residual convergence.* $r^k \to 0$ *as* $k \to \infty$*, i.e., the iterates approach feasibility.*
- *Objective convergence.* $f(x^k) + g(z^k) \to p^*$ *as* $k \to \infty$*, i.e., the objective function of the iterates approaches the optimal value.*
- *Dual variable convergence.* $\lambda^k \to \lambda^*$ *as* $k \to \infty$*, where* λ^* *is a dual optimal point.*

Proposition 2 (Optimality)**.** *Alternating direction method of multipliers attain the optimal solution of optimization problem* (13) *if and only if it is primal feasible*

$$Ax^* + Bz^* - c = 0 \tag{16a}$$

and dual feasible

$$0 \in \partial f(x^*) + A^T\lambda^* \tag{16b}$$
$$0 \in \partial g(z^*) + B^T\lambda^*. \tag{16c}$$

Next, we show that (16c) is satisfied in every step. As z^{k+1} minimizes $L_\rho(x^{k+1}, z, \lambda^k)$, we have

$$\begin{aligned}
&0 \in \partial g(z^{k+1}) + B^T\lambda^k + \rho B^T(Ax^{k+1} + Bz^{k+1} - c)\\
&0 \in \partial g(z^{k+1}) + B^T\lambda^k + \rho B^T r^{k+1}\\
&0 \in \partial g(z^*) + B^T\lambda^{k+1}.
\end{aligned}$$

The second dual feasibility condition (16b) is note automatically satisfied

$$
\begin{aligned}
&0 \in \partial f(z^{k+1}) + A^T\lambda^k + \rho A^T(Ax^{k+1} + Bz^k - c) \\
&0 \in \partial f(z^{k+1}) + A^T\lambda^k + \rho A^T(Ax^{k+1} + Bz^k - c) + \underbrace{(\rho A^T B z^{k+1} - \rho A^T B z^{k+1})}_{=0} \\
&0 \in \partial f(z^{k+1}) + A^T(\lambda^k + \rho r^{k+1} + \rho B(z^k - z^{k+1})) \\
&0 \in \partial f(z^*) + A^T\lambda^{k+1} + \rho A^T B(z^k - z^{k+1}).
\end{aligned}
$$

Thus the dual residual (deviation from dual feasibility) is

$$s^{k+1} = \rho A^T B(z^{k+1} - z^k)$$

and the primal residual is

$$r^{k+1} = Ax^{k+1} + Bz^{k+1} - c.$$

The stopping criteria for ADMM are therefore

$$||r^{k+1}||_2 \leq \epsilon_{\text{primal}} \qquad ||s^{k+1}||_2 \leq \epsilon_{\text{dual}}. \tag{17}$$

2.2.1 Scaled Alternating Direction Method of Multipliers

Most papers utilize a scaled version of ADMM. Therefore, it is presented in the following.

The last two terms

$$\lambda^T(Ax + Bz - c) + \frac{\rho}{2}||Ax + Bu - c||_2^2 \tag{18}$$

of the augmented Lagrangian $L_\rho(x, z, \lambda)$ (see (15)) are often scaled, since this leads to shorter expressions. The scaled version of ADMM is derived in the following. First, we define the residual $r \equiv Ax + Bu - c$ that is the deviation from satisfying the equality constraints. Then (18) can be written as

$$
\begin{aligned}
\lambda^T r + \frac{\rho}{2}||r||_2^2 &= \frac{\rho}{2}\left(r + \frac{1}{\rho}\lambda\right)^T\left(r + \frac{1}{\rho}\lambda\right) - \frac{1}{2\rho}\lambda^T\lambda \\
&= \frac{\rho}{2}||r + \frac{1}{\rho}\lambda||_2^2 - \frac{1}{2\rho}||\lambda||_2^2 \\
&= \frac{\rho}{2}||r + \bar{\lambda}||_2^2 - \frac{\rho}{2}||\bar{\lambda}||_2^2
\end{aligned}
$$

where $\bar{\lambda} = \frac{1}{\rho}\lambda$ is the scaled dual variable.

Algorithm 4 (Scaled Alternating Direction Method of Multipliers)**.**
Input: *Cost functions* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *and* $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$*, specification of equality constraints* $A \in \mathbb{R}^{p \times n}$*,* $B \in \mathbb{R}^{p \times m}$ *and* $c \in \mathbb{R}^p$*, initial value for Lagrange multiplier* λ_{init}*, initial value for z-variable* z_{init}*,* ϵ_{primal}*,* ϵ_{dual}*, and update step* $\rho \in \mathbb{R}_{>0}$*.*
Output: p^**, and* λ^**.*
Procedure:

0. *Initialization: Let* $k := 0$ *and* $\bar{\lambda}^k = \lambda_{\text{init}}$*.*

1. *Find primal variable*

   (a) $x^{k+1} := \arg\min_x \left(f(x) + \frac{\rho}{2}||Ax + Bz^k - c + \bar{\lambda}^k||_2^2\right)$.

   (b) $z^{k+1} := \arg\min_z \left(g(z) + \frac{\rho}{2}||Ax^{k+1} + Bz - c + \bar{\lambda}^k||_2^2\right)$.

2. *Update dual variable* $\bar{\lambda}^{k+1} := \bar{\lambda}^k + Ax^{k+1} + Bz^{k+1} - c$.

3. *If* $||r^{k+1}||_2 > \epsilon_{\text{primal}}$ *or* $||s^{k+1}||_2 > \epsilon_{\text{dual}}$ *then* $k := k + 1$ *and go to step 1.*

4. *Output* $f(x^{k+1}) + g(z^{k+1})$, *and* λ^{k+1}.

2.2.2 Consensus and Alternating Direction Method of Multipliers

If we have an optimization problem with an additive separable cost, then ADMM can be utilized to solve the problem in a partly distributed manner. Such a problem is on the following form

$$\text{minimize } \sum_{i=1}^{N} f_i(x),$$

where $x \in \mathbb{R}^n$ is the variable and $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex. This type of optimization problem could occur if we have several inequality constraints, each giving rise to an indicator function being added to the cost function.

A way to solve the previous problem is to pose it as a consensus problem. In optimization the next problem is called consensus problem

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{N} f_i(x_i) \\ &\text{subject to } x_i - z = 0, \qquad i = 1, \dots, N \end{aligned}$$

where $x_i \in \mathbb{R}^n$ is a vector of local variables and z is a vector of global variables. Now there are N copies of the variable x that must all equal z.

To solve the consensus problem, we form the augmented Lagrangian

$$L_\rho(x, \dots, x_N, z, \lambda) = \sum_{i=1}^{N} \left(f_i(x_i) + \lambda_i^T(x_i - z) + \frac{\rho}{2}||x_i - z||_2^2 \right).$$

In the following, we show that the x- and λ-update steps can be distributed; however, this is not the case for the z-update. The steps in the ADMM algorithm becomes

$$x_i^{k+1} := \arg\min_{x_i} \left(f_i(x_i) + (\lambda_i^k)^T(x_i - z^k) + \frac{\rho}{2}||x_i - z^k||_2^2 \right),$$

which can be accomplished independently for each i. The z-update step is

$$\begin{aligned} z^{k+1} &:= \arg\min_z \left(\sum_{i=1}^{N} (f_i(x_i^{k+1}) + (\lambda_i^k)^T(x_i^{k+1} - z) + \frac{\rho}{2}||x_i^{k+1} - z||_2^2) \right) \\ &:= \arg\min_z \left(\sum_{i=1}^{N} (-(\lambda_i^k)^T z + \frac{\rho}{2}||x_i^{k+1} - z||_2^2) \right). \end{aligned}$$

Per definition z^{k+1} is a minimizer of the expression; hence, the gradient with respect to z must be zero. This implies that

$$
\begin{aligned}
0 &= \sum_{i=1}^{N}(-(\lambda_i^k)^T - \rho(x_i^{k+1} - z^{k+1})) \\
z^{k+1} &= \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{\rho}(\lambda_i^k)^T + x_i^{k+1}), \qquad (19)
\end{aligned}
$$

which is the update for z.

Knowing z^{k+1} the update of dual variables can also be accomplished distributed as

$$\lambda_i^{k+1} := \lambda_i^k + \rho(x_i^{k+1} - z^{k+1}).$$

The ADMM algorithm can be simplified further, by realizing that (19) is just addition of average values. Thus,

$$z^{k+1} = \frac{1}{\rho}(\bar{\lambda}^k)^T + \bar{x}^{k+1},$$

where $\bar{x}$ denotes the average value of x_i. From this, it follows that

$$\bar{\lambda}^{k+1} := \bar{\lambda}^k + \rho(\bar{x}^{k+1} - z^{k+1}),$$

and for $k \geq 0$, $\bar{\lambda}^{k+1} = 0$, i.e., $z^k = \bar{x}^k$.

Algorithm 5 (Alternating Direction Method of Multipliers)**.**
Input: *Cost functions* $f_{:}\mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, *initial value for Lagrange multiplier* λ_{init}, *initial value for* $\bar{x}$*-variable* $\bar{x}_{\text{init}}$, ϵ_{primal}, ϵ_{dual}, *and update step* $\rho \in \mathbb{R}_{>0}$.
Output: p^*, *and* λ^*.
Procedure:

0. *Initialization: Let* $k := 0$, $\bar{x}^k = \bar{x}_{\text{init}}$, *and* $\lambda^k = \lambda_{\text{init}}$.

1. *Find primal variable*

$$x_i^{k+1} := \arg\min_x \left(f_i(x_i) + (\lambda_i^k)^T(x_i - \bar{x}^k) + \frac{\rho}{2}||x_i - \bar{x}^k||_2^2 \right).$$

2. *Update dual variable* $\lambda^{k+1} := \lambda^k + \rho(x_i^{k+1} - \bar{x}^{k+1})$.

3. *If* $||r^{k+1}||_2 > \epsilon_{\text{primal}}$ *or* $||s^{k+1}||_2 > \epsilon_{\text{dual}}$ *then* $k := k+1$ *and go to step 1.*

4. *Output* $f(x^{k+1}) + g(z^{k+1})$, *and* λ^{k+1}.

2.2.3 Distributed Alternating Direction Method of Multipliers

The ADMM algorithm can be distributed if we consider an optimisation problem with additively separable cost

$$
\begin{aligned}
&\text{minimize } \sum_{i}^{N} f_i(x_i), \\
&\text{subject to } x_i - G_i z = 0, \qquad i = 1, \dots, N
\end{aligned}
$$

where $x_i \in \mathbb{R}^{n_i}$ are local variables and $z \in \mathbb{R}^n$ is a global variable, and $G_i \in \mathbb{R}^{n_i \times n}$ is a signature matrix.

This optimization problem represents a situation where each function f_i only depends on a subset of the variables - hopefully $n_i << n$. This is a consensus problem that can be distributed. To ease notation, we define

$$\tilde{z}_i \equiv G_i z.$$

With the new notation, we have

$$\begin{aligned} &\text{minimize } \sum_i^N f_i(x_i), \\ &\text{subject to } x_i - \tilde{z}_i = 0 \qquad i = 1, \ldots, N. \end{aligned}$$

The augmented Lagrangian for this problem is

$$L_\rho(x, \ldots, x_N, z, \lambda) = \sum_{i=1}^N \left(f_i(x_i) + \lambda_i^T (x_i - \tilde{z}_i) + \frac{\rho}{2} ||x_i - \tilde{z}_i||_2^2 \right).$$

In the ADMM algorithm, the x-update is similar to the consensus-case

$$x_i^{k+1} := \arg\min_{x_i} \left(f_i(x_i) + (\lambda_i^k)^T x_i + \frac{\rho}{2} ||x_i - \tilde{z}_i^k||_2^2 \right).$$

Similar to the consensus problem, the z-update is an average, but now only including information among the subproblems that include a particular global variable z_i; thus, the algorithm can be distributed. To calculate the average, we define the matrix

$$G_{\text{inv}} = \begin{bmatrix} G_1^T & \cdots & G_N^T \end{bmatrix},$$

and denote the vector of row sums of G_{inv} by g_{inv}, then

$$z_i^{k+1} := \frac{1}{g_{\text{inv},i}} \pi_i G_{\text{inv}} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix},$$

where π_i is the projection onto the ith coordinate.

3 Comparison of Presented Methods

The previous section presented two method for distributing an optimization problem. However, the attentive reader may have noticed that the dual decomposition algorithm is not really distributed. The solution of the dual problem can be distributed, but not the update of the dual variables (λ). This computation relies on knowledge of the global state of the system; thus, the algorithm can be depicted as shown in Figure 6. Here a coordinator calculates the dual variable based on information from all computing entities.

In the solution of very large optimization problems and in the deployment of many distributed devices, it is desirable to exploit algorithms that do not contain a central coordinator, but are truly distributed. In distributed optimization, devices only communicate with their neighbors. The communication in a distributed algorithm is illustrated in Figure 7. The ADMM method has this particular property.

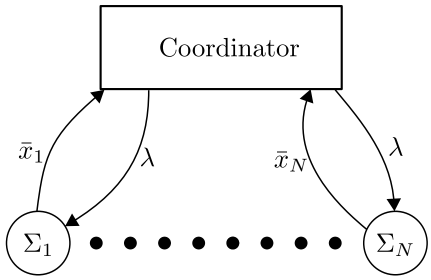

Figure 6: Illustration of computations in dual decomposition that requires a coordinator with global knowledge.

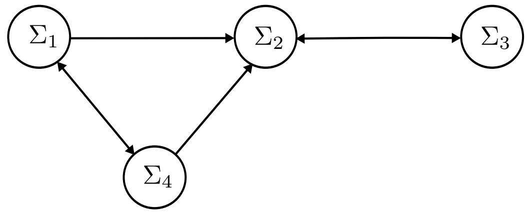

Figure 7: Communication in distributed algorithm.

References

[1] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Athena Scientific, 1996.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[4] S. Boyd, L. Xiao, A. MutaPcic, and J. Mattingley. Notes on decomposition methods. Notes for EE364B, Stanford University, Winter 2006-07, 2008.

[5] F. Lin, M. Fardad, and M. R. Jovanović. Design of optimal sparse feedback gains via the alternating direction method of multipliers. *IEEE Trans. Automat. Control*, 58(9):2426–2431, September 2013.

[6] J. Mota, J. Xavier, P. Aguiar, and M. Puschel. Distributed ADMM for model predictive control and congestion control. In *2012 IEEE 51st Annual Conference on Decision and Control (CDC)*, pages 5110–5115, 2012.