# Optimization for Data Science

F. Rinaldi[1]

1 Università degli Studi di Padova | Dipartimento Matematica

Padova
2020

# Outline

**Optimization for Data Science**

1 Duality in Nonlinear Programming

2 Distributed Optimization and Learning Using the ADMM

# Duality Theory

- It gives very important results both from a theoretical and an algorithmic point of view.
- It enables us to relate every constrained problem with a dual problem (similar to what we have seen when analyzing FW).
- Under some specific assumptions the dual has a strong connection with the primal and, usually, a structure that can be better exploited from a computational point of view.

# Lagrangian Function

## Starting Problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in X \end{array} \tag{1}$$

with $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^p$ and $X \subseteq \mathbb{R}^n$.
Hence we have:

$$C = \{x \in X : \quad h(x) = 0, \quad g(x) \leq 0\}. \tag{2}$$

## Lagrangian Function

we can build the *Lagrangian function* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ given as:

$$L(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x). \tag{3}$$

# How to Build the Lagrangian Problem

## Lagrangian Problem

In connection with the previous problem (1) We can thus define the *Lagrangian dual problem* related to the primal problem described in (1):

$$\begin{array}{ll} \max & \varphi(\lambda, \mu) \\ \text{s.t.} & \lambda \geq 0 \end{array} \tag{4}$$

where $\varphi : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is given as follows:

$$\varphi(\lambda, \mu) = \inf_{x \in X} \left\{ f(x) + \lambda^\top g(x) + \mu^\top h(x) \right\} = \inf_{x \in X} L(x, \lambda, \mu). \tag{5}$$

- Notice that for some $(\lambda, \mu)$ function $\varphi(\lambda, \mu)$ might be $-\infty$.
- Hence, in some cases, it might be useful to consider the set:

$$\Delta = \{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^p : \varphi(\lambda, \mu) > -\infty\}. \tag{6}$$

# Weak Duality

**Theorem [Weak Duality]**

Let $f \in C(\mathbb{R}^n)$ , $g_i \in C(\mathbb{R}^n)$, for all $i = 1, \ldots, m$, and $h_j \in C(\mathbb{R}^n)$, for all $j = 1, \ldots, p$. For any feasible point $x \in \mathbb{R}^n$ of the primal problem (1), that is $x \in X$, $g(x) \leq 0$ and $h(x) = 0$, and for any feasible point $(\lambda, \mu)$ of the dual problem (4), that is $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^p$ and $\lambda \geq 0$, we have:

$$\varphi(\lambda, \mu) \leq f(x). \tag{7}$$

**Proof.**

- From definition of $\varphi$ and from $x \in X$, $g(x) \leq 0$ e $h(x) = 0$, $\lambda \geq 0$, we have

$$\begin{aligned} \varphi(\lambda, \mu) &= \inf_{x \in X} \left\{ f(x) + \lambda^\top g(x) + \mu^\top h(x) \right\} \\ &\leq f(x) + \lambda^\top g(x) + \mu^\top h(x) \leq f(x). \end{aligned} \tag{8}$$

Thus we proved our result.

# A Useful Result

## Corollary

Let $f \in C(\mathbb{R}^n)$ and $g_i \in C(\mathbb{R}^n)$, for all $i = 1, \ldots, m$, and $h_j \in C(R^n)$, for all $j = 1, \ldots, p$. The following properties hold:

i)
$$\max_{\lambda \geq 0} \varphi(\lambda, \mu) \leq \min_{x \in C} f(x);$$

ii) if a $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^m \times \mathbb{R}^p$ with $\bar{\lambda} \geq 0$ and a point $\bar{x} \in X$ with $g(\bar{x}) \leq 0$ and $h(\bar{x}) = 0$, are such that
$$\varphi(\bar{\lambda}, \bar{\mu}) = f(\bar{x}),$$
then $(\bar{\lambda}, \bar{\mu})$ is an optimal solution for the dual and $\bar{x}$ is an optimal solution for the primal;

iii) if the primal is unbounded, then
$$\varphi(\lambda, \mu) = -\infty,$$
for all $(\lambda, \mu) \in R^m \times R^p$ with $\lambda \geq 0$;

iv) if the dual is unbounded, then the primal is unfeasible.

# Comments

- From $i)$ we have that for an $x^*$ optimal solution of the primal and a pair $(\lambda^*, \mu^*)$ optimal solution of the dual, the following inequality holds:

$$\varphi(\lambda^*, \mu^*) \leq f(x^*).$$

- If

$$\varphi(\lambda^*, \mu^*) < f(x^*),$$

  we have a *duality gap*.
- In case

$$\varphi(\lambda^*, \mu^*) = f(x^*),$$

  we have a zero duality gap.

# Identifying Optimal Solutions

## Optimality Conditions

We say that $(x^*, \lambda^*, \mu^*)$ satisfy optimality conditions for the primal if the following are satisfied:

- Dual feasibility:
$$x^* \in \operatorname*{Argmin}_{x \in X} L(x, \lambda^*, \mu^*),$$
$\lambda^* \geq 0;$
- Primal feasibility:
$$g(x^*) \leq 0, \quad h(x^*) = 0, \quad x^* \in X;$$
- Complementary slackness:
$$\lambda^{*\top} g(x^*) = 0.$$

- When functions are continuously differentiable and some other convexity assumptions are satisfied, we can equivalently write in place of first dual feasibility condition,

$$\nabla f(x^*) + \nabla g(x^*)^\top \lambda^* + \nabla h(x^*)^\top \mu^* = 0.$$

# The Quadratic Case

## Convex Quadratic Problems

Now we focus on convex quadratic problems of the form:

$$\begin{array}{ll} \min & \frac{1}{2}x^\top Qx + c^\top x \\ \text{s.t.} & Ax \leq b, \end{array}$$

with $x \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

## Lagrangian Dual

We consider the Lagrangian dual related to the above problem:

$$\begin{array}{ll} \max & \varphi(\lambda) \\ \text{s.t.} & \lambda \geq 0, \end{array} \tag{9}$$

with

$$\varphi(\lambda) = \inf_{x\in\mathbb{R}^n} \left\{ L(x,\lambda) = \frac{1}{2}x^\top Qx + c^\top x + \lambda^\top (Ax - b) \right\}.$$

# Comments

- We assume here that the primal admits an optimal solution.
- The function $L(x, \lambda)$, for any fixed value $\lambda \geq 0$, is a convex quadratic function.
- It is thus bounded from below if and only if its minimum is achieved, which in turn can be true if and only if the gradient of $L$ with respect to $x$ vanishes at some point (see optimality conditions).
- Thus, if $\varphi(\lambda) = -\infty$, there is no $x$ satisfying

$$\nabla_x L(x, \lambda) = Qx + c + A^\top \lambda = 0.$$

- Otherwise we can rewrite

$$\begin{aligned} L(x, \lambda) &= -\frac{1}{2} x^\top Q x + x^\top (Qx + c + A^\top \lambda) - \lambda^\top b \\ &= -\frac{1}{2} x^\top Q x - \lambda^\top b, \end{aligned}$$

## Dual Problem

$$\begin{array}{ll} \max & L(x, \lambda) = -\frac{1}{2} x^\top Q x - \lambda^\top b \\ \text{s.t.} & Qx + c + A^\top \lambda = 0 \\ & \lambda \geq 0. \end{array} \tag{10}$$

# A Useful Result

**Proposition [Strong duality for quadratic problems]**

Let $x^*$ be optimal for the primal, then there exists a vector $\lambda^* \geq 0$ such that $(x^*, \lambda^*)$ is optimal for the dual and the two extremal values are equal. Furthermore, if $(\tilde{x}, \tilde{\lambda})$ is optimal for the dual, then some $x^*$ satisfying

$$Q(x^* - \tilde{x}) = 0, \tag{11}$$

$$\tilde{\lambda}^\top (Ax^* - b) = 0 \tag{12}$$

and

$$Ax^* \leq b$$

is optimal for the primal and the two extremal values are equal.

# SVM training

- We now can apply this result to the SVM training problems.
- Let us start with the linearly separable case.
- The Lagrangian function for the problem is

$$L(w, \theta, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{P} \lambda_i \left[ y^i (w^\top x^i + \theta) - 1 \right].$$

## Dual Problem

$$\begin{aligned} \max \quad & \frac{1}{2}\|w\|^2 - \sum_{i=1}^{P} \lambda_i \left[ y^i (w^\top x^i + \theta) - 1 \right] \\ \text{s.t.} \quad & \nabla_w L(w, \theta, \lambda) = w - \sum_{i=1}^{P} \lambda_i y^i x^i = 0 \\ & \nabla_\theta L(w, \theta, \lambda) = \sum_{i=1}^{P} \lambda_i y^i = 0 \\ & \lambda \geq 0. \end{aligned} \tag{13}$$

- We now use the equivalence

$$\max_{x \in C} f(x) \equiv -\min_{x \in C} -f(x).$$

- Using the first equality constraint we can rewrite the problem as follows

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} y^i y^j (x^i)^\top x^j \lambda_i \lambda_j - \sum_{i=1}^{P} \lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{P} \lambda_i y^i = 0 \\ & \lambda \geq 0. \end{aligned} \tag{14}$$

## Equivalent Formulation

$$\begin{aligned} \min \quad & \frac{1}{2} \lambda^\top X^\top X \lambda - e^\top \lambda \\ \text{s.t.} \quad & \sum_{i=1}^{P} \lambda_i y^i = 0 \\ & \lambda \geq 0, \end{aligned} \tag{15}$$

with $X = [y^1 x^1 \ldots y^P x^P]$. Thus we get a convex quadratic problem with simple constraints.

# Building Up the Primal Solution

- Using the result reported in Theorem 12, we have that the primal optimal solution $(w^*, \theta^*)$ can be built starting from the dual solution $(\tilde{w}, \tilde{\theta}, \tilde{\lambda})$, where

$$\tilde{w} = \sum_{i=1}^{P} \tilde{\lambda}_i y^i x^i \quad \text{and} \quad \tilde{\theta} \in \mathbb{R}.$$

- Indeed, since equality (11) holds, we have

$$w^* = \tilde{w} = \sum_{i=1}^{P} \tilde{\lambda}_i y^i x^i.$$

- Those vectors that have a $\tilde{\lambda}_i > 0$ are called support vectors.
- Furthermore, since Strong Duality conditions hold, we can write

$$\tilde{\lambda}_i [y^i ({w^*}^\top x^i + \theta^*) - 1] = 0, \quad i = 1, \ldots, P$$

and for all $\tilde{\lambda}_i > 0$, we have $y^i({w^*}^\top x^i + \theta^*) = 1$. Hence, we can calculate $\theta^*$ by using any of those equations.

# Nonlinearly Separable Case

## Same trick applies...

Following the same reasoning as before, we get

$$\begin{aligned} \min \quad & \frac{1}{2}\lambda^\top X^\top X\lambda - e^\top\lambda \\ \text{s.t.} \quad & \sum_{i=1}^{P} \lambda_i y^i = 0 \\ & 0 \leq \lambda \leq C, \end{aligned} \tag{16}$$

with $X = [y^1x^1 \ldots y^Px^P]$. Thus we have a convex quadratic problem with simple constraints (Take a look at the notes for further details).

# Distributed optimization and learning using the Alternating Direction Method of Multipliers

- Many problems of recent interest in statistics and machine learning can be posed in the framework of convex optimization.
- Due to the high dimension and complexity of modern datasets, it is really important to solve problems with a very huge number of features or training examples.
- As a result, both the decentralized storage of these datasets as well as the development of distributed methods are desirable.
- We describe the *Alternating Direction Method of Multipliers*(ADMM), first introduced by Douglas and Rachford (1956).
- This approach is well suited to distributed convex optimization and, in particular, to huge-scale problems arising in data science.

# ADMM

## Problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{17}$$

with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable convex function.

- We rewrite the problem in the equivalent form

$$\begin{aligned} \min \quad & f(x) + \frac{\rho}{2}\|Ax - b\|^2 \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{18}$$

where $\rho > 0$ is a *penalty parameter*.

# Lagrangian Dual

### Lagrangian Function related to Problem (17)

Consider the Lagrangian function related to this equivalent reformulation:

$$L_\rho(x, \mu) = f(x) + \mu^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|^2.$$

- Thanks to the new term the dual function

$$\varphi(\mu) = \inf_{x \in \mathbb{R}^n} L_\rho(x, \mu)$$

has nice properties (like, e.g., continuous differentiability under mild conditions).

# Calculating the Gradient of $\varphi(\mu)$

## How to Calculate the Gradient $g(\mu)$

For a given $\bar{\mu}$

- minimize over $x$:

$$\bar{x} \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} L_\rho(x, \bar{\mu}).$$

- evaluate the equality constraint residual:

$$g(\bar{\mu}) = A\bar{x} - b.$$

# Dual Ascent Method

- Applying a gradient like approach (*dual ascent method*) to the dual problem

$$\max_{\mu \in \mathbb{R}^m} \varphi(\mu)$$

yields the algorithm that everybody knows as *method of multipliers*.

- Keep in mind that we are maximizing here, then we want to get a ascent direction.
- Easy to check that the best ascent direction is the gradient.

**Method of Multipliers (Iteration $k$)**

$$x_{k+1} = \operatorname*{Argmin}_{x \in \mathbb{R}^n} L_\rho(x, \mu_k)$$

and

$$\mu_{k+1} = \mu_k + \rho(Ax_{k+1} - b).$$

# Augmented Lagrangian and the Method of Multipliers

- It is possible to prove that the choice of $\rho$ as a stepsize guarantees dual feasibility.
- Checking dual feasibility of $(x_{k+1}, \mu_{k+1})$ is very easy. Indeed, from minimization on $x$, we get

$$
\begin{aligned}
0 &= \nabla_x L_\rho(x_{k+1}, \mu_k) = \nabla_x f(x_{k+1}) + A^\top(\mu_k + \rho(Ax_{k+1} - b)) \\
&= \nabla_x f(x_{k+1}) + A^\top \mu_{k+1}.
\end{aligned}
$$

- Furthermore, as the method of multipliers proceeds, the primal residual $Ax_{k+1} - b$ converges to zero, thus giving optimality.

# Alternating Direction Method of Multipliers

## A Structured Problem

$$\begin{array}{ll} \min & f(x) + g(z) \\ \text{s.t.} & Ax + Cz = b, \end{array} \tag{19}$$

With $A \in \mathbb{R}^{m \times n_1}$, $C \in \mathbb{R}^{m \times n_2}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}_1^n \to \mathbb{R}$ and $g : \mathbb{R}_2^n \to \mathbb{R}$ continuously differentiable convex functions.

## Augmented Lagrangian Function Related to the Problem

$$L_\rho(x, z, \mu) = f(x) + g(z) + \mu^\top (Ax + Cz - b) + \frac{\rho}{2} \|Ax + Cz - b\|^2,$$

with $\rho > 0$.

# Alternating Direction Method of Multipliers II

- ADMM consists of three different steps:

$$x_{k+1} = \operatorname*{Argmin}_{x \in \mathbb{R}_1^n} L_\rho(x, z_k, \mu_k),$$

$$z_{k+1} = \operatorname*{Argmin}_{z \in \mathbb{R}_2^n} L_\rho(x_{k+1}, z, \mu_k),$$

and

$$\mu_{k+1} = \mu_k + \rho(Ax_{k+1} + Cz_{k+1} - b).$$

- In ADMM, $x$ and $z$ are updated in an alternating or sequential fashion, which accounts for the term alternating direction.
- ADMM can be hence viewed as a version of the method of multipliers where a single Gauss-Seidel step over $x$ and $z$ is used instead of the usual joint minimization.
- Splitting the minimization over $x$ and $z$ into two steps is precisely what allows for decomposition when f or g are separable.

# Algorithmic Scheme

**Algorithm 1** `Alternating Direction Method of Multipliers`

Choose points $x_1 \in \mathbb{R}^{n_1}$, $z_1 \in \mathbb{R}^{n_2}$, $\mu_1 \in \mathbb{R}^m$ and $\rho > 0$

For $k = 1, \ldots$

Set

$$x_{k+1} = \operatorname*{Argmin}_{x \in \mathbb{R}_1^n} L_\rho(x, z_k, \mu_k)$$

Set

$$z_{k+1} = \operatorname*{Argmin}_{z \in \mathbb{R}_2^n} L_\rho(x_{k+1}, z, \mu_k)$$

Set

$$\mu_{k+1} = \mu_k + \rho(Ax_{k+1} + Cz_{k+1} - b)$$

End for

# Comments

- Convergence of the method can be proved under standard assumptions.
- The rate is in general *sublinear*.
- Improving convergence: use different penalty parameters $\rho_k$ for each iteration.
- ADMM converges even with approximate minimizations w.r.t. $x$ and $z$ (provided certain conditions are satisfied) [Eckstein and Bertsekas].
- This modification is important when iterative methods are needed to get the $x$ or $z$ updates.
- IDEA: Solve the minimizations only approximately at first, and then more accurately as the iterations go on.

# Consensus Optimization

- There has recently been interest in coordination of networks consisting of multiple agents.
- GOAL: Collectively optimize a global objective.
- Motivated by the emergence of large-scale networks ( e.g., mobile ad hoc networks and wireless-sensor networks).
- Networks characterized by the lack of centralized access to information and time-varying connectivity.
- Optimization algorithms deployed in such networks should be
  - completely distributed, relying only on local observations and information;
  - robust against unexpected changes in topology, such as link or node failures;
  - scalable in the size of the network.
- We describe two variants of the *consensus problem* and distributed ADMM-based methods for solving them.

# Global Consensus Problem

- In consensus, we consider a multiagent network model, where $P$ agents exchange information over a connected network.
- Each agent $i$ has a "local function" $f_i(x)$.
- The vector $x \in \mathbb{R}^n$ is a global decision vector that the agents need to collectively determine.

**GOAL**

Agents need to cooperatively optimize a global-objective function, that means solving the following problem:

$$\min_{x \in \mathbb{R}^n} \quad \sum_{i=1}^{P} f_i(x), \tag{20}$$

with $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, P$ continuously differentiable convex functions.

# Global Consensus Problem

## Formulation of the Global Consensus Problem

Last problem can be rewritten as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^{P} f_i(x^i), \\ \text{s.t.} \quad & x^i = z, \quad i = 1, \ldots, P \end{aligned} \tag{21}$$

with the auxiliary variables $z \in \mathbb{R}^n$, $x^i \in \mathbb{R}^n$, $i = 1, \ldots, P$. Notice that the constraints are such that all the local variables should agree, i.e., be equal.

## Augmented Lagrangian Function

The augmented Lagrangian function is in this case

$$L_\rho(x^1, \ldots, x^P, z, \mu) = \sum_{i=1}^{P} \left[ f_i(x^i) + {\mu^i}^\top (x^i - z) + \frac{\rho}{2} \|x^i - z\|^2 \right],$$

with $\rho > 0$.

# ADMM Approach

**Generic iteration of ADMM for the problem**

$$x_{k+1}^i = L_\rho(x^i, x_k^{-i}, z_k, \mu_k) = \operatorname*{Argmin}_{x^i \in \mathbb{R}^n} f_i(x^i) + {\mu_k^i}^\top (x^i - z_k) + \frac{\rho}{2}\|x^i - z_k\|^2, \quad i = 1, \ldots, P,$$

$$z_{k+1} = \operatorname*{Argmin}_{z \in \mathbb{R}^n} L_\rho(x_{k+1}^1, \ldots, x_{k+1}^P, z, \mu_k) = \frac{1}{P}\sum_{i=1}^{P}\left(x_{k+1}^i + \frac{\mu_k^i}{\rho}\right),$$

and

$$\mu_{k+1}^i = \mu_k^i + \rho(x_{k+1}^i - z_{k+1}), \quad i = 1, \ldots, P.$$

We indicate with $x^{-i}$ the set of all $x^j$, such that $j \neq i$.

- The $x^i$ and $\mu^i$ calculations are carried out independently for each $i = 1, \ldots, P$.
- In the literature, the processing element that handles the global variable $z$ is usually called *central collector* or *fusion center*.

# Global Consensus Problem

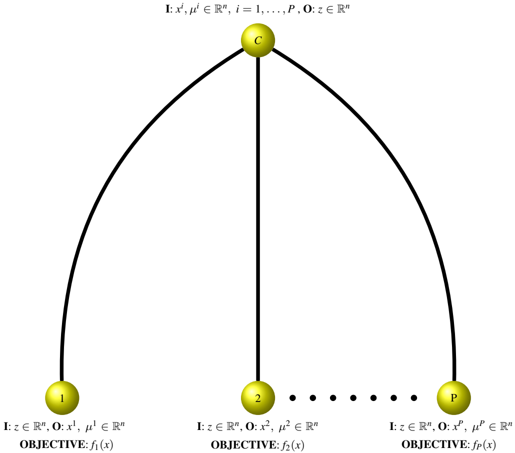


Figure: Global consensus problem.

# General Global Consensus Problem

- It is possible to consider a more general form of the consensus problem, in which each agent $i$ has a "local function" $f_i(x^i)$.
- The vector $x^i \in \mathbb{R}^{n_i}$, is a selection of the components of the global vector $z \in \mathbb{R}^n$ that the agents need to collectively determine.

## Formulation of the Problem

This problem can be written as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^{P} f_i(x^i), \\ \text{s.t.} \quad & x^i = z^i, \quad i = 1, \ldots, P, \end{aligned} \tag{22}$$

where $z^i \in \mathbb{R}^{n_i}$ is a subvector of the global vector $z$.

## For further details...

Take a look at the notes