Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Mathematische Optimierung

# First Order Primal Dual Optimization Methods

Michael Ulbrich

October 2013

# Contents

# 1 Introduction

Due to the strong need of optimization methods for solving very large scale problems that often involve (structured) nonsmoothness (like $l_1$- or TV-regularizations) and/or (structured) constraints, current research in optimization has developed significant interest in optimization methods that are (1) simple and (2) allow for a decomposition of the overall problem into comparably easy, sometimes independent (i.e., parallelizable) subproblems. Prototypes of such "simple" schemes are first order methods that only involve first derivatives and often can be viewed as variants of (projected) steepest descent methods. On the other hand, the removal of constraints that couple subblocks of the optimization variables can be achieved by Lagrange multipliers and dualization. This course focusses on both aspects – primal-dual techniques and their combination with first order methods.

Although certain extensions exist, the major part of available work on such methods is devoted to convex optimization problems and convex-concave saddle point problems. Therefore, we will focus on the convex case, too.

We start by introducing two important primal-dual approaches of the form that we will investigate in detail in this lecture.

## 1.1 A first order primal dual method

Consider a problem of the form

$$\min_{x\in C}\max_{y\in K}\ y^TAx + g^Tx - h^Ty \tag{1}$$

with closed convex sets $C \subset \mathbb{R}^n$ and $K \subset \mathbb{R}^m$, $g \in \mathbb{R}^n$, $h \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$. Associated to this, we also consider

$$\max_{y\in K}\min_{x\in C}\ y^TAx + g^Tx - h^Ty.$$

Later in this course, we will generalize the setting sigbnificantly by replacing the linear functions $g^Tx$ and $h^Ty$ by lower semicontinuous convex extended functions.

We want to find a saddle point of the underlying convex-concave function

$$q(x, y) := y^TAx + g^Tx - h^Ty$$

on $C \times K$. This, as we will see in section 2, is the same as finding a solution $\bar{x}$ of the first problem and a solution $\bar{y}$ of the second problem that attain the same optimal value ("min max = max min"). This optimal value is called *saddle value*.

**Definition 1.1** *A point $(\bar{x}, \bar{y}) \in C \times K$ is a (min-max) saddle point of $q$ on $C \times K$ if*

$$q(\bar{x}, y) \le q(\bar{x}, \bar{y}) \le q(x, \bar{y}) \quad \forall\, x \in C,\ y \in K.$$

We start with the observation that "min max $\geq$ max min" always holds, which forms the core of weak duality results:

**Lemma 1.2** *Let $V$, $W$ be sets and let $l : V \times W \to \mathbb{R}$ be a function. Then*

$$\sup_{w\in W} \inf_{v\in V} \; l(v,w) \leq \inf_{v\in V} \sup_{w\in W} \; l(v,w).$$

**Proof**: Let $(\tilde{v},\tilde{w}) \in V \times W$ be arbitrary. Then there holds

$$\inf_{v\in V} l(v,\tilde{w}) \leq l(\tilde{v},\tilde{w}) \leq \sup_{w\in W} l(\tilde{v},w)$$

Applying $\sup_{\tilde{w}\in W}$ yields

$$\sup_{\tilde{w}\in W} \inf_{v\in V} l(v,\tilde{w}) \leq \sup_{w\in W} l(\tilde{v},w)$$

und taking the $\inf_{\tilde{v}\in V}$ results in

$$\sup_{\tilde{w}\in W} \inf_{v\in V} l(v,\tilde{w}) \leq \inf_{\tilde{v}\in V} \sup_{w\in W} l(\tilde{v},w)$$

□

Saddle point problems arise in a variety of situations, e.g, on a general level in the context of Lagrange and Fenchel duality, but also in the context of reformulating nonsmooth convex functions.

**Example 1.3** Consider the constrained $l_1$-approximation problem

$$\min_{x\in C} \; \|Ax-h\|_1$$

with $C$, $A$, and $h$ as above and $\|v\| = \sum_i |v_i|$. This is a nonsmooth problem. One can resolve the nonsmooth absolute value function by introducing an additional vector $w \in \mathbb{R}^m$ and writing

$$\min_{x,w} \; \sum w_i \quad \text{s.t.} \quad -w \leq Ax-h \leq w, \quad x \in C.$$

Alternatively, one can use

$$\|v\|_1 = \max_{\|y\|_\infty \leq 1} \; y^T v,$$

where $\|y\|_\infty = \max_i \; |y_i|$. Then, setting $K = \{y\,;\, \|y\|_\infty \leq 1\}$, we have

$$\|Ax-h\|_1 = \max_{y\in K} \; y^T(Ax-h).$$

Thus, the approximation problem can be written as

$$\min_{x\in C} \max_{y\in K} \; y^T Ax - h^T y$$

This has exactly the form (1) with $g = 0$.

The first primal dual algorithm [CP11,PCBC09] that we will consider takes projected gradient steps alternating between $x$ and $y$ and combines the results in a clever way: Given $x^0 \in C$ and $y^0 \in K$, we set $\tilde{x}^0 := x^0$ and iterate as follows:

$$\begin{aligned} y^{k+1} &:= P_K(y^k + \sigma(A\tilde{x}^k - h)), \\ x^{k+1} &:= P_C(x^k - \tau(A^T y^{k+1} + g)), \\ \tilde{x}^{k+1} &:= 2x^{k+1} - x^k. \end{aligned}$$

Here, $P_C$ and $P_K$ are the projections onto $C$ and $K$, respectively, and $\tau, \sigma > 0$ are step sizes satisfying $\tau\sigma\|A\|^2 < 1$.

Choosing the function $q(x, y) := y^T Ax + g^T x - h^T y$ introduced above, we have

$$\nabla_y q(x, y) = Ax - h, \quad \nabla_x q(x, y) = A^T y + g.$$

Thus, the first step is a projected steepest ascent step for $q$ in the $y$-variable at $y^k$ with $x$ fixed at $\tilde{x}^k$. The step size is $\sigma$. The second step is a steepest descent step for $q$ in the $x$-variable at $x^k$ with $y$ fixed at $y^{k+1}$. The step size is $\tau$. Then $\tilde{x}^{k+1}$ is chosen as the step from $x^k$ along the vector $x^{k+1} - x^k$ with step size 2:

$$\tilde{x}^{k+1} := x^k + 2(x^{k+1} - x^k)$$

It might be more intuitive to avoid $\tilde{x}^k$, which corresponds to maintaining $\tilde{x}^k = x^k$, which is the same as choosing step size 1 in the update formula of $\tilde{x}^{k+1}$:

$$\tilde{x}^{k+1}_{AH} := x^k + (x^{k+1} - x^k) = x^{k+1}.$$

In deed, this results in the classical Arrow-Hurwicz algorithm.

It turns out that introducing $\tilde{x}^k$ and updating it with a step size larger than 1 (canonically 2) improves the properties of the algorithm.

We will discuss the properties of this simple, yet efficient method in a significantly more general setting than the introductory problem considered here.

## 1.2 Alternating direction method of multipliers (ADMM)

A second method that we will consider is the alternating direction method of multipliers (ADMM), sometimes also called alternating direction method (ADM) [BPCPE10].

### 1.2.1 Augmented Lagrangian method

Before we can describe the ADMM method, we need some background on augmented Lagrangian methods (also called methods of multipliers) [Ber96].

Consider

$$\min_{x\in C} f(x) \quad \text{s.t.} \quad Ax = b. \tag{2}$$

with $C \subset \mathbb{R}^n$ closed and convex, $f : U \to \mathbb{R}$ convex and continuously differentiable on an open neighborhood $U$ of $C$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.

Augmented Lagrangian methods solve subproblems of the form

$$\min_{x\in C} L^a(\gamma; x, \lambda), \tag{3}$$

with the penalty parameter $\gamma \geq 0$ and the multiplier estimate $\lambda \in \mathbb{R}^m$ fixed and the augmented Lagrangian function

$$L^a(\gamma; x, \lambda) = L(x, \lambda) + \frac{\gamma}{2}\|Ax - b\|^2 = f(x) + \lambda^T(Ax - b) + \frac{\gamma}{2}\|Ax - b\|^2,$$

where $L(x, \lambda) = f(x) + \lambda^T(Ax - b)$ is the usual Lagrange function. The augmented Lagrangian method solves this problem repeatedly while updating $\lambda$ and $\gamma$.

If we set $\lambda = 0$ then this is just the quadratic penalty method.

Further, if $C = \mathbb{R}^n$ and $(\bar{x}, \bar{\lambda})$ is a KKT point, then there holds

$$\nabla_x L(\bar{x}, \bar{\lambda}) = 0, \quad A\bar{x} = b.$$

This implies

$$\nabla_x L^a(\gamma; \bar{x}, \bar{\lambda}) = \nabla_x L(\bar{x}, \bar{\lambda}) + \gamma A^T(A\bar{x} - b) = 0.$$

Hence, for the choice $\lambda = \bar{\lambda}$, $\bar{x}$ is a stationary point of $L^a(\gamma; \cdot, \bar{\lambda})$. This shows the advantage of augmented Lagrange approach over the quadratic penalty method: If $\lambda$ is chosen appropriately, then $\gamma$ does not have to be driven to $\infty$ to achieve convergence.

In the more general case where $C$ is not the whole space, then, under a constraint qulification (CQ), e.g., Slater's condition ($\exists\ \hat{x} \in \text{int}(C)$, $A\hat{x} = b$) the optimality conditions of (2) are

$$\bar{x} \in C, \quad \nabla_x L(\bar{x}, \bar{\lambda})^T(x - \bar{x}) \geq 0 \quad \forall\, x \in C, \tag{4}$$

$$A\bar{x} = b. \tag{5}$$

Furthermore, the optimality condition for (3) reads

$$\bar{x} \in C, \quad \nabla_x L^a(\gamma; \bar{x}, \lambda)^T(x - \bar{x}) \geq 0 \quad \forall\, x \in C. \tag{6}$$

Also here, if we consider a KKT-pair $(\bar{x}, \bar{\lambda})$ of (2) and choose $\lambda = \bar{\lambda}$, then there holds

$$\nabla_x L^a(\gamma; \bar{x}, \bar{\lambda}) = \nabla_x L(\bar{x}, \bar{\lambda}) + \gamma A^T(A\bar{x} - b) = \nabla_x L(\bar{x}, \bar{\lambda}).$$

Therefore, $\bar{x}$ satisfies $(6)|_{\lambda=\bar{\lambda}}$.

A central question is how to update $\lambda$ and $\gamma$. Let $x^k$ denote the current iterate and let $\lambda^k$ and $\gamma_k$ be the current choices of $\lambda$ and $\gamma$. Then $x^{k+1}$ is obtained by solving

$$\min_{x\in C} \; L^a(\gamma_k; x, \lambda^k). \tag{7}$$

We then have

$$\begin{aligned}\nabla_x L^a(\gamma_k; x^{k+1}, \lambda^k) &= \nabla f(x^{k+1}) + A^T\lambda^k + \gamma A^T(Ax^{k+1} - b)\\ &= \nabla f(x^{k+1}) + A^T[\lambda^k + \gamma_k(Ax^{k+1} - b)]\\ &= \nabla_x L(x^{k+1}, \lambda^{k+1}),\end{aligned}$$

where we made the choice $\lambda^{k+1} := \lambda^k + \gamma_k(Ax^{k+1} - b)$. This makes sense since then $(x^{k+1}, \lambda^{k+1})$ satisfy the first part (4) of the KKT-conditions for (2):

$$x^{k+1} \in C, \quad \nabla_x L(x^{k+1}, \lambda^{k+1})^T(x - x^{k+1}) \geq 0 \quad \forall\, x \in C.$$

Note that the $\lambda$-update can also be written as follows:

$$\lambda^{k+1} := \lambda^k + \gamma_k \nabla_\lambda L^a(\gamma_k; x^{k+1}, \lambda^k).$$

Improving feasibility w.r.t. the constraint $Ax = b$ is achieved by a suitable $\gamma$-update:

With fixed parameters $\beta \in (0, 1)$, $\rho > 1$ (e.g., $\rho = 10$), choose

$$\gamma_{k+1} := \rho\gamma_k \;\text{ if } \|Ax^{k+1} - b\| \geq \beta\|Ax^k - b\|, \quad \gamma_{k+1} := \gamma_k \;\text{ otherwise.}$$

The $\lambda$-update can be viewed as a steepest ascent step for the augmented Lagrangian dual function. Details on duality will follow later, we just give a short sketch:

If $(\bar{x}, \bar{\lambda})$ is a KKT-pair, i.e., satisfies (4),(5), then there also holds (6) and thus $\bar{x}$ is a minimum of $L^a(\gamma; \cdot, \bar{\lambda})$ on $C$. Furthermore, (5) means $\nabla_\lambda L^a(\gamma; \bar{x}, \bar{\lambda}) = 0$ and thus $\bar{\lambda}$ is a maximum of $L^a(\gamma; \bar{x}, \cdot)$ on $\mathbb{R}^m$. Thus, $(\bar{x}, \bar{\lambda})$ is a (min-max) saddle point of $L^a(\gamma; \cdot, \cdot)$ on $C \times \mathbb{R}^m$. Further, there holds

$$\sup_{\lambda\in\mathbb{R}^m} \; L^a(\gamma; x, \lambda) = \begin{cases} f(x) & (Ax = b),\\ \infty & (Ax \neq b).\end{cases}$$

In this sense (2) is equivalent to

$$\inf_{x\in C} \sup_{\lambda\in\mathbb{R}^m} \; L^a(\gamma; x, \lambda). \tag{8}$$

The augmented Lagrangian dual problem is given by

$$\sup_{\lambda\in\mathbb{R}^m} \inf_{x\in C} L^a(\gamma; x, \lambda). \tag{9}$$

We know that the value of (9) is always majorized by the value of (8), see Lemma 1.2. Since $(\bar{x}, \bar{\lambda})$ is a saddle point, we have even equality of the primal and dual value and it is attained at $(\bar{x}, \bar{\lambda})$. More details will follow later. The function

$$d^a(\lambda) = \inf_{x\in C} L^a(\gamma; x, \lambda)$$

is the augmented Lagrangian dual function. For convenience, we have dropped the dependence on $\gamma$ here.

# References

[Ber96] D. P. Bertsekas: *Constrained optimization and Lagrange multiplier methods*, Athena Scientific, Belmont, 1996.

[BPCPE10] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein: *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Foundations and Trends in Machine Learning Vol. 3, No. 1 (2010), 1122.

[CP11] A. Chambolle, T. Pock: *A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging*, Journal of Mathematical Imaging and Vision, Vol. 40, No. 1 (2011), 120-145.

[PCBC09] T. Pock, D. Cremers, H. Bischof, A. Chambolle: *An Algorithm for Minimizing the Piecewise Smooth Mumford-Shah Functional*. In: IEEE International Conference on Computer Vision (ICCV), 2009.