Proposal for Qualifying Exam

Will Wright

Draft Date: June 26, 2019

Exam Committee

Committee Chairperson:

Prof. Matthias Koeppe

Committee Members:

Prof. Michael Friedlander

Prof. Roger Wets

Prof. Roland Freund

Prof. Ilias Tagkopoulos

Exam Logistics

Date:

Thursday, February 18, 2016

Time:

12:00pm - 3:00pm

Location:

3240 MSB

Proposed Research

Title: Smooth Exact Penalty Functions for Constrained Optimization

Abstract: Many methods in constrained optimization involve passing the constraints into the objective, forming a smooth exact penalty function. We consider two common smoothing methods, the augmented Lagrangian method and the proximal point method. Though these methods differ superficially, both have been used recently to create an active set Newton method with a superior convergence rate on least squares problems with box or l_1-regularization constraints. This method has been constructed as the Newton system of both a particular augmented Lagrangian [Fletcher, 2015] and a modified Moreau envelope [Patrinos, 2014]. We show the equivalence of these penalty functions under first orthant constraints (x ≥ 0), and propose a generalized augmented Lagrangian (defined using conjugate functions) that we conjecture results in a single variable function identical to the modified Moreau envelope. Finally, we propose strategies for extending the smooth exact penalty function to handle equality constraints.

1 Introduction and Notation

Many methods in constrained optimization involve reformulating the problem as an unconstrained problem by passing the constraint(s) into the objective. We will first discuss two of the most common current methods: the augmented Lagrangian method (Section 2) and the proximal point method (Section 3). In the process, we will motivate the construction of both methods and highlight parallels between both methods. These two methods will lead us to a discussion of a recent active set method with a convergence rate superior to all other methods for quadratic programs with simple constraints (such as box bounds or l_1-regularization). This method (Section 4) has been constructed as the Newton system of both a particular augmented Lagrangian [10] and a modified Moreau envelope [17]. We will proceed (Section 5) by showing the equivalence of these value functions under first orthant constraints ($x \geq 0$). Finally, we will close with a proposal to use duality theory, and in particular conjugate functions in the augmented Lagrangian, to create a generalized augmented Lagrangian with first order conditions that we conjecture will result in an exact penalty function identical to the modified Moreau envelope. We also discuss the possibility of extending the augmented Lagrangian to handle equality constraints.

Before we proceed, we will first establish notation conventions and offer remarks to the reader. All unlabeled norms $||\cdot||$ refer to the $2-$norm. A function f is said to be *smooth* if it is continuously differentiable ($f \in \mathcal{C}^1$). If it is non-smooth, it may still admit a generalized differential ∂f know as a *subdifferential* (see Section 5). A function is said to be *proper* if $f(x) < \infty$ for some x, and *lower semi-continuous* if f attains all lower limits, i.e. $\liminf_{u \to x} f(u) \geq f(x)$ for all x. In a single iteration the update term is x^+; and a minimum of a convex optimization problem is x^*. A symmetric positive definite matrix Q is denoted $Q \succ 0$, and $x \geq 0$ refers to x with all coordinates non-negative. L_f will refer to the Lipschitz constant of f, which we will always assume exists. Optimization texts alternate between $\frac{1}{\mu}$ and L when referring to the descent and update parameter for the methods we will be discussing. In this context $L = ||Q||_2 = \lambda_{\max}(Q)$, the Lipschitz constant of the quadratic function $\frac{1}{2}x^TQx + q^Tx$. However, to maintain consistency and highlight the equivalence of the methods discussed below, we have chosen $\frac{1}{\mu}$ as the update parameter throughout the paper (at the cost of occasional ease of reading). We will always assume $\mu \in (0, 1/L_f)$.

2 Background: The Augmented Lagrangian Method

We first consider inequality-equality constrained optimization problems (IECP) of the form

$$\begin{array}{ll} \min & f(x) \\ \text{st} & g(x) \geq 0 \\ & h(x) = 0 \end{array} \tag{1}$$

where f is sufficiently smooth ($f \in \mathcal{C}^1$ or $\mathcal{C}^2$), $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^p$, and all f, g, h are proper, lower semi-continuous. If $g \equiv 0$, then we recover the equality constrained problem (ECP), minimizing $f(x)$ subject to $h(x) = 0$, which we will briefly discuss to motivate how the augmented Lagrangian may be applied to (1).

One class of methods for solving (1) involves reformulating the ECP as an unconstrained optimization problem by embedding the constraint into the objective and solving the resulting unconstrained problem iteratively. One classical unconstrained reformulation involves adding a quadratic penalty function to obtain

$$Q(x; \mu) = f(x) + \frac{1}{2\mu}||h(x)||^2 \tag{2}$$

Conceptually this construction with penalty parameter $\mu > 0$ allows the user to approach x^* by iteratively solving for the x^{k+1} which minimizes $Q(x^k; \mu^k)$; as k increases we have $\mu^k \to 0$ to drive (x^k) into the feasible region. In practice however, the Hessian of $Q(x^k; \mu^k)$ grows increasingly ill-conditioned as μ^k approaches 0. To see this, note that the Hessian of the penalty term, $\nabla^2_{xx} \frac{1}{2\mu}||h(x)||^2 = \frac{1}{\mu}\big(\sum_{i=1}^p h_i(x)\nabla^2 h_i(x) + \nabla h(x)\nabla h(x)^T\big)$, has a spectrum with magnitude dependent on μ.

To avoid this ill-conditioning as we approach x^*, another classical unconstrained reformulation of (1) uses the Lagrangian

$$L(x, z) = f(x) - z^T h(x) \tag{3}$$

In this case we require that first order (KKT) conditions be met: $\nabla_x L(x^*, z^*) = 0$ and $\nabla_z L(x^*, z^*) = 0$. Additionally, if we reintroduce the inequality constraint g, we have the Lagrangian

$$L(x, y, z) = f(x) - y^T g(x) - z^T h(x)$$

with the KKT conditions

$$\begin{aligned} \nabla_x L(x^*, y^*, z^*) &= 0 \\ g(x^*) &\geq 0 \\ h(x^*) &= 0 \\ y^* &\geq 0 \\ y_i^* g_i(x^*) &= 0 \quad \forall i \in [m] \end{aligned}$$

To solve (1) using the Lagrangian we may locally approximate f, g and h quadratically and obtain a sequence $x^k \to x^*$ by either 1) solving the KKT system at (x^k, y^k, z^k) directly, or 2) splitting L and obtaining x^k, y^k, and z^k separately. One problem with this strategy is that the KKT system in general involves solving a non-positive definite system since L is not convex in (x, y, z), thus requiring a factorization with $\mathcal{O}(n^3)$ flops.

We have seen that the quadratic penalty function, while potentially convex with positive-definite Hessian, requires a parameterization that creates ill-conditioned systems. In contrast, the Lagrangian might offer a nicely conditioned Newton system, but one which is not positive definite. The classical solution to these two problems, first discussed by Hestenes in 1969 and later studied in depth by Rockafellar, is to combine both of these methods by adding both penalty terms to form the augmented Lagrangian (or method of multipliers). This well-conditioned, convex penalty function is typically used to create an effective primal-dual iterative method.

To motivate the construction of the augmented Lagrangian, we first set $g \equiv 0$ and examine the ECP, which has the augmented Lagrangian

$$\mathcal{L}(x, z; \mu) = f(x) - z^T h(x) + \frac{1}{2\mu} ||h(x)||^2 \tag{4}$$

A typical iteration of the augmented Lagrangian method involves an *inner* iteration where the multiplier z^k is fixed and x is updated by minimizing $\mathcal{L}$, then an *outer* iteration which updates the Lagrange multiplier estimate, and finally a new parameter $\mu^{k+1} \geq \mu^k$ is selected. During each iteration, a termination criterion such as $||\nabla_x \mathcal{L}(x^k, z^k; \mu^k)|| \leq \tau^k$ is tested after the inner iteration against some tolerance τ^k.

Since the inner iteration is an unconstrained minimization of a sum of functions f and h, we may use x^{k-1} as a guess for x^k, and descend using gradient descent if $f, h \in C^1$ or a Newton / quasi-Newton method if $f, h \in C^2$. The outer iteration has the standard update

$$z^{k+1} = z^k - \frac{1}{\mu^k} h(x^k) \tag{5}$$

The intuition behind this update comes from the first order optimality condition on $\mathcal{L}$. Using x^k from the inner iteration, we require that

$$\nabla_x \mathcal{L}(x^k, z^k; \mu^k) = \nabla f(x^k) - \left[z^k - \frac{1}{\mu^k} h(x^k)\right]^T \nabla_x h(x^k) \approx 0$$

From the view of the standard Lagrange method (3), we have the Lagrange multiplier $z^* \approx z^k - \frac{1}{\mu^k} h(x^k)$, hence the update (5).

In contrast to the quadratic penalty method (2), the penalty parameter sequence $\{\mu^k\}$ typically decreases monotonically such that $\frac{1}{\mu_k} \to k < \infty$, and generally with k fairly small. For sufficiently small μ, $\mathcal{L}(x, z; \mu)$ will be convex in a neighborhood of x^*. Furthermore, $\mathcal{L}$ is an exact penalty function; if the user were to know the exact multiplier term z^*, then $\mathcal{L}(x, z^*; \mu)$ will have x^* as a strict minimizer for sufficiently large μ (see [14], p517). Hence the augmented Lagrangian method allows the user to update μ^k conservatively based on how well z^k approximates z^*, and thus maintain a well conditioned inner iteration.

If we reintroduce the inequality constraint $g(x) \geq 0$, there are a few variants of the augmented Lagrangian method used to solve (1). One method uses slack variables $s_i \geq 0$ to insert the equality constraint $g(x) - s = 0$ into the ECP formulation just discussed. This bound-constrained Lagrangian method is implemented by the widely used LANCELOT software package (part of the GALAHAD library). Another variant linearlizes the constraints at each iteration and solves a sequence of approximations F^k to (4). In this linearly constrained Lagrangian method there are a few choices for F_k, one of which is implemented in the widely used MINOS software package. A third method, initially discussed by Rockafellar in 1974 and of primary importance to us, leads to an unconstrained formulation of (1). This method relies on defining the augmented Lagrangian as a piecewise continuous function which splits depending on whether each constraint $g_i(x^k) \geq 0$ is active or inactive at the k^{th} iteration. To simplify the construction, we will set $h \equiv 0$ giving an inequality constrained problem (ICP). As a first step toward this construction, we consider an unconstrained formulation of the ICP:

$$\min_{x\in\mathbb{R}^n} \max_{z\geq 0} \left\{ f(x) - z^T g(x) \right\} \tag{6}$$

Like the standard Lagrangian function, this non-smooth (and impractical) construction can be smoothed by adding a quadratic penalty term

$$\min_{x\in\mathbb{R}^n} \max_{z\geq 0} \left\{ f(x) - z^T g(x) - \frac{\mu}{2}||z - z^k||^2 \right\} \tag{7}$$

which includes multiplier approximations z^k and penalizes steps away from these multipliers. This formulation is coordinate-wise separable in z, giving the explicit solution

$$z_i = \begin{cases} z_i^k - \frac{1}{\mu} g_i(x) & \text{if} \quad g_i(x) \leq \mu z_i^k \qquad \text{(active coordinate)} \\ 0 & \text{if} \quad g_i(x) \geq \mu z_i^k \qquad \text{(inactive coordinate)} \end{cases} \tag{8}$$

to the inner maximization problem. Substituting (8) into (7) gives the problem

$$\min_{x\in\mathbb{R}^n} F(x, z^k, \mu^k) \tag{9}$$

where we have Rockafellar's unconstrained smooth augmented Lagrangian for inequality constraints

$$F(x, z; \mu) = f(x) + \sum_{i=1}^{n} \begin{cases} -g_i(x) z_i + \frac{1}{2\mu} g_i(x)^2 & \text{if} \quad g_i(x) \leq \mu z_i \\ -\frac{\mu}{2} z_i^2 & \text{if} \quad g_i(x) \geq \mu z_i \end{cases} \tag{10}$$

Note the resemblance of (10) to the equality-constrained augmented Lagrangian (4). Yet unlike the bound-constrained and linearly constrained methods for solving (1), this unconstrained method did not have any published implementations until very recently. In 2015 Roger Fletcher first applied (10) to the subproblem of (1) with inequality constraint $x \geq 0$, and derived an active set Newton method with superior convergence rate [10]. We will discuss this implementation in more detail in Section 4. First let us develop another method for solving (1) which appears somewhat different from the augmented Lagrangian, yet has also generated this active set Newton method under more general constraints.

3 Background: The Moreau Envelope and Proximal Methods

In this section we again consider the general minimization problem (1). In contrast to the augmented Lagrangian, we now consider the reformulation of (1) using a different class of exact penalty functions known as envelopes. The classical such value function is the *Moreau envelope*, first discussed by Minty in 1962 and then formalized by Moreau in 1965:

$$e_{\mu P}(x) := \inf_u \left\{ P(u) + \frac{1}{2\mu}||u - x||^2 \right\} \tag{11}$$

where the function P is proper and lower semi-continuous (so minimizers are attainable). In practice P is often a penalty function based on a given constraint set or requirement (e.g. $P(x) = ||x||_1$). As was the case with the augmented Lagrangian, the use of the 2-norm term in the Moreau envelope serves to smooth P, making $e_{\mu P}$ continuously differentiable. Additionally, $e_{\mu P}(x) \leq P(x)$ for all x, with equality at x only if this location is a local minimum of P. So if P is convex then P and its Moreau envelope share the same minimizers. Conceptually, e smooths or regularizes P by defining an envelope bounding P from below. Along with smoothing P, the Moreau envelope extends dom(P) to all of $\mathbb{R}^n$.

Extensive literature has been written on the properties of the Moreau envelope (see for instance [1], [18]). However, our primary concern is a particular use of the Moreau envelope in constructing iterative methods for solving (1), where P corresponds to constraints. This leads to the construction of the vector-valued *proximal map*:

$$\text{prox}_{\mu P}(x) := \underset{u}{\text{argmin}} \left\{ P(u) + \frac{1}{2\mu}||u-x||^2 \right\} \tag{12}$$

If P is strongly convex then the proximal map is a contraction mapping, i.e. Lipschitz continuous with constant less than 1. For general P, the proximal map is firmly nonexpansive, meaning that for all x, y in $\mathbb{R}^n$

$$||\text{prox}_{\mu P}(x) - \text{prox}_{\mu P}(y)||^2 \leq (x-y)^T(\text{prox}_{\mu P}(x) - \text{prox}_{\mu P}(y)) \tag{13}$$

Since firmly nonexpansive operators are a special class of nonexpansive operators (those that are Lipschitz continuous with constant 1), prox may thus be use to develop iterative methods for solving (1). To see the intuition of a proximal iteration $x^+ = \text{prox}_{\mu P}(x)$, consider the simple example of projecting $x \in \mathbb{R}^n$ onto a closed, convex set $C \subset \mathbb{R}^n$. If we let P be the indicator function on C, i.e.

$$P(x) = \delta_C(x) := \begin{cases} 0 & \text{if} \quad x \in C \\ +\infty & \text{else} \end{cases} \tag{14}$$

then

$$\text{prox}_{\mu P}(x) = \underset{u}{\text{argmin}} \left\{ \delta_C(x) + \frac{1}{2\mu}||u-x||^2 \right\} = \underset{u \in C}{\text{argmin}} \left\{ \frac{1}{2\mu}||u-x||^2 \right\} = \Pi_C(x) \tag{15}$$

and the Moreau envelope evaluates $e_{\mu\delta_C}(x) = \frac{1}{2\mu}\text{dist}_2(x, C)$, the $\mu-$scaled Euclidean distance from x to C.

The proximal map may be applied to constrained optimization problems like (1) to produce a first order method known as *proximal splitting* or the *forward-backward method.* Similar to the augmented Lagrangian, we form an unconstrained optimization problem as a sum of functions expressing the objective and the constraints. Here the forward-backward method solves (1) by optimizing the composite function $F(x) = f(x) + P(x)$, where P represents a penalty on violating the constaints. For instance, in (1) if f is C^1, h is linear, and g is concave, then $C = \{x \mid h(x) = 0 \;\&\; g(x) \geq 0\}$ is convex and we may "split" each iteration by first performing gradient descent on f and then projecting onto C. Thus the forward-backward method

$$\text{prox}_{\mu P}(x - \mu\nabla f(x)) = \Pi_C(x - \mu\nabla f(x)) \tag{16}$$

generalizes standard gradient projection. In the next section we will introduce a linearlization of the objective function f into the Moreau envelope, resulting in a particular envelope used by Patrinos, Stella, and Bemporad in [17], to develop the active set Newton method (19) below.

4 An Active Set Newton Method Based on Two Constructions: A Modified Envelope and an Augmented Lagrangian

The augmented Lagrangian method and the proximal mapping method both approach constrained optimization problems with the same general strategy: add a 2-norm term to the objective function f to introduce the constraints in a smoothed manner, giving an unconstrained exact penalty function which also serves as a merit function for determining stepsize and convergence (1). As with the augmented Lagrangian, the ability of the proximal map to generalize constrained optimization problems unites a variety of optimization techniques under the large and well established theory of proximal mapping and the Moreau envelope (see [1], [16], or [18]).

Yet in contrast to the augmented Lagrangian, proximal methods do not have a multiplier variable in their definition (i.e. formulations here are only dependent on variable x and parameter μ). There appears to be a fundamental difference between the two-step inner and outer iteration structure of the standard augmented Lagrangian method and the single-step (albeit split) iteration of the forward-backward method. But in mathematics in general, and optimization in particular, we often find that two seemingly disparate methods for solving a class of problems are in fact equivalent under certain circumstances. Toward this end, two authors have recently introduced exact penalty functions for solving (1), each of which leads to an active-set Newton method with superior speed of convergence. Their two respective value functions, an augmented Lagrangian based on (10) and a Moreau-type envelope, are shown in Section 5 to be equivalent under first orthant constraints ($x \geq 0$). Furthermore, we propose the equivalence is much larger; but first let us discuss the active-set Newton method itself and each of its derivations.

To simplify our initial discussion and focus on intuition, we begin by considering quadratic programs with first orthant constraints, i.e. the specific case of (1) where we have $Q \succ 0$ and seek

$$\begin{array}{ll} \min & f(x) := \frac{1}{2}x^TQx + q^Tx \\ \text{st} & \qquad\quad x \geq 0 \end{array} \tag{17}$$

Without the first orthant constraint we could solve this problem accurately using the standard Newton method with update $x^+ = x + \mu d$, where d satisfies the Newton system

$$\nabla^2 f(x)d = -\nabla f(x) \tag{18}$$

and μ satisfies an appropriate stepsize inequality.

Once we introduce the first orthant constraint, a similar Newton system can be developed which maintains the spirit of (18). While unconstrained gradient descent would have the update $x^+ = x - \mu\nabla f(x)$, if we use the notation $+ = \{x \mid x \geq 0\}$ then the forward-backward method, or projected gradient descent, has the update $x^+ = \Pi_+(x - \mu\nabla f(x))$. We use this projected gradient in place of the standard $-\mu\nabla f(x)$ in (18) to arrive at the projected Newton system

$$\begin{bmatrix} I_{AA} & 0 \\ \mu Q_{JA} & \mu Q_{JJ} \end{bmatrix} \begin{bmatrix} d_A \\ d_J \end{bmatrix} = \begin{bmatrix} \Pi_+(x - \mu\nabla f(x))_A - x_A \\ \Pi_+(x - \mu\nabla f(x))_J - x_J \end{bmatrix} \tag{19}$$

The subscripts A and J refer respectively to whether constraints on the righthand side projection subproblem are active or inactive; so in each iteration the coordinates are split accordingly. If a particular coordinate i is active (i.e. $i \in A$) then (19) gives a Newton step $d_i = [\Pi_+(x - \mu\nabla f(x)) - x]_i$, which results in the update $x_i^+ = x_i + d_i = [\Pi_+(x - \mu\nabla f(x))]_i$. In other words, the i^{th} coordinate of the Newton system was found to be minimized on the edge of the bounds where the projection $\Pi_+(x - \mu\nabla f(x))$ is active. In this sense, (19) eliminates the unnecessary solving of this portion of the Newton system by setting d_A equal to the righthand side of (19). On the other hand, if a particular coordinate i is inactive then we see $[\Pi_+(x - \mu\nabla f(x)) - x]_i$ in (19) is equal to $[-\mu\nabla f(x)]_i$ in 18. In this case (19) requires the solving of a Newton system of size $|J| \times |J|$, giving the inactive Newton step coordinates d_J.

This construction leads to a logical iteration structure. First we find $\Pi_+(x - \mu\nabla f(x))$, giving us active A and inactive J which partition $[n]$. We then set $d_A = [\Pi_+(x - \mu\nabla f(x)) - x]_A$ for substitution into (19) to solve the $|J| \times |J|$ Newton subsystem

$$\mu Q_{JJ}d_J = [\Pi_+(x - \mu\nabla f(x)) - x]_J - \mu Q_{JA}d_A \tag{20}$$

Due to its intuitive and easily implemented structure, this active set Newton method has been used to solve (1) since as early as 1998 (e.g. [9]). However, the problem for such implementations was that there did not exist a reliable, general merit function like the augmented Lagrangian (4) or the Moreau envelope (11) on which the Newton system (19) was clearly based. This problem was overcome independently by Patrinos, Stella, and Bemporad in [17] using a Moreau-type envelope and by Fletcher in [10] using the augmented Lagrangian.

The envelope merit function of [17] involves a simple modification of the Moreau envelope (11). In Section 3 we defined the Moreau envelope $e_{\mu P}$ and discussed implementations involving P being a constraint penalty function (e.g. $P = \delta_+$). If we modify this definition by adding a linearlization of the objective function $f(x)$ to $P(u)$, i.e. $P'(u) := f(x) + \nabla f(x)^T(u - x) + P(u)$, then we have what Patrinos defines as the *forward-backward envelope* (FBE)

$$F_\mu(x) := e_{\mu P'}(x) = \min_u \Big\{ f(x) + \nabla f(x)^T(u - x) + P(u) + \frac{1}{2\mu}||u - x||^2 \Big\} \tag{21}$$

The properties of the FBE are discussed in detail in [17]. In particular, F_μ is continuously differentiable with gradient

$$\nabla F_\mu(x) = \frac{1}{\mu}(I - \mu\nabla^2 f(x))(x - \mathrm{prox}_{\mu P}(x - \mu\nabla f(x))) \tag{22}$$

In the typical case that penalty function P is non-smooth, ∇F_μ is also non-smooth. However we may consider a generalized Hessian of the FBE, called a *linear Newton approximation*

$$\partial^2 F_\mu(x) := \left\{\frac{1}{\mu}(I - \mu\nabla^2 f(x))(I - J(I - \mu\nabla^2 f(x))) \mid J \in \partial_{\mathbf{C}}(\mathrm{prox}_{\mu P})(x - \mu\nabla f(x))\right\} \tag{23}$$

where $\partial_{\mathbf{C}}$ is the generalized Clarke Jacobian. A detailed second-order analysis of F_μ is provided in [17, 3]. However, we will move on toward the derivation of (19). In applying equations (22) and (23) to (17) to develop a Newton system, we first note that when $+ = \{x \mid x \geq 0\}$, we have $\mathrm{prox}_{\mu P}(x) = \Pi_+(x)$ which is separable. In this case $J \in \partial_{\mathbf{C}}(\mathrm{prox}_{\mu P})(u) = \partial_{\mathbf{C}}(\Pi_+)(u)$, the Clarke Jacobian in (23), will be diagonal with

$$J_{ii} = \begin{cases} 1 & \text{if} \quad u_i > 0 \quad \text{(inactive coordinate)} \\ 0 & \text{if} \quad u_i < 0 \quad \text{(active coordinate)} \\ 1 \text{ or } 0 & \text{if} \quad u_i = 0 \end{cases} \tag{24}$$

So if we take $H(x) \in \partial^2 F_\mu(x)$ then the FBE gives us the Newton system $H(x)d = -\nabla F_\mu(x)$ which reduces to

$$(I - J(I - \mu\nabla^2 f(x)))d = \Pi_+(x - \mu\nabla f(x)) - x \tag{25}$$

or in the case of the quadratic program

$$(I - J + \mu JQ)d = \Pi_+(x - \mu\nabla f(x)) - x \tag{26}$$

For the set of active coordinates, (26) reduces to $d_A = [\Pi_+(x - \mu\nabla f(x)) - x]_A$. The remaining inactive coordinates give the system $[\mu Qd]_J = [\Pi_+(x - \mu\nabla f(x)) - x]_J$. Thus, if we split the system accordingly, we arrive at (19).

Now we consider the augmented Lagrangian function of [10]. In Section 2 we closed by discussing Rockafellar's unconstrained smooth augmented Lagrangian (10). If we apply this construction to (17), and use the substitution $y = \mu z$ on the Lagrange multipliers of (10), we arrive at

$$F(x, y; \mu) = f(x) + \frac{1}{\mu}\sum_{i=1}^{n} \begin{cases} -x_i y_i + \frac{1}{2}x_i^2 & \text{if} \quad x_i \leq y_i \\ -\frac{1}{2}y_i^2 & \text{if} \quad x_i \geq y_i \end{cases} \tag{27}$$

To simplify differentiation of F with respect to the multiplier variable, we next substitute $w = y - x$, use the notation $(w_-)_i = \min(w_i, 0)$, and clear μ from the sum, giving

$$\begin{aligned} F(x, w; \mu) &= \mu f(x) + \textstyle\sum_{i=1}^{n} \begin{cases} -x_i w_i - \frac{1}{2}x_i^2 & \text{if} \quad w_i \geq 0 \\ -\frac{1}{2}(w_i + x_i)^2 & \text{if} \quad w_i \leq 0 \end{cases} \\ &= \mu f(x) - w^T x - \tfrac{1}{2}x^T x - \tfrac{1}{2}w_-^T w_- \end{aligned} \tag{28}$$

Fixing w and setting $\nabla_x F(x, w; \mu) = 0$ in the inner iteration forces

$$\begin{aligned} w(x) &= \mu\nabla f(x) - x \\ &= \mu(Qx + q) - x \end{aligned} \tag{29}$$

making (28) a single variable penalty function

$$F(x, w(x); \mu) = -\frac{\mu}{2}x^T Qx + \frac{1}{2}x^T x - \frac{1}{2}w(x)_-^T w(x)_- \tag{30}$$

The final form in (30) of the augmented Lagrangian has gradient

$$\nabla_x F(x, w(x); \mu) = (I - \mu Q)x - (\mu Q - I)Jw \tag{31}$$

and Hessian

$$\begin{array}{rcl} \nabla^2_{xx} F(x, w(x); \mu) & = & (I - \mu Q) - (I - \mu Q)J(I - \mu Q) \\ & = & (I - \mu Q)(I - J(I - \mu Q)) \end{array} \tag{32}$$

leading to the Newton system

$$\begin{array}{rcl} \nabla^2_{xx} F(x) d & = & -\nabla^2_x F \\ (I - J + \mu JQ)d & = & -x - Jw \end{array} \tag{33}$$

To complete the equivalence between (33) and (26), we handle $-Jw$ by noting

$$\begin{array}{rcl} -[Jw]_i & = & -\begin{cases} [\mu \nabla f(x) - x]_i & \text{if } w_i < 0 \\ 0 & \text{if } w_i > 0 \end{cases} \\ & = & -\begin{cases} [-(x - \mu \nabla f(x))]_i & \text{if } w_i < 0 \\ 0 & \text{if } w_i > 0 \end{cases} \\ & = & [\Pi_+(x - \mu \nabla f(x))]_i \end{array} \tag{34}$$

Thus $(I - J + \mu JQ)d = -x - Jw = [\Pi_+(x - \mu \nabla f(x))] - x$ and we arrive at (19), the same Newton system as with the FBE.

5 The Equivalence of the FBE and the Augmented Lagrangian and Proposal for Generalizing

In this section we begin by demonstrating explicitly the equivalence of the augmented Lagrangian with the FBE when applied to a quadratic program with first orthant constraints (17). Given that the FBE (21) is currently a more general construction than the augmented Lagrangian (7, 27-30), one goal of our work is to generalize fully this equivalence by defining an augmented Lagrangian using convex analysis tools such as conjugate functions and support functions. Such a generalization of the augmented Lagrangian would have the advantage of combining the broad body of knowledge about the augmented Lagrangian with the broad set of applications of the FBE as discussed in [17]. We will discuss a few preliminary thoughts for extending this equivalence and close with a remark regarding how we may also incorporate equality constraints into this exact penalty function (which is not an implementable constraint type using the current constructions).

We begin by showing a baseline equivalence of the two exact penalty functions.

Proposition 1. *Consider the quadratic program with first orthant constraints (17), $\mu \in (0, 1/L_f)$, and first order conditions on the augmented Lagrangian (30). Then the forward-backward envelope (21) is equivalent to the augmented Lagrangian (30) over all $\mathbb{R}^n$. That is,*

$$L(x, w, \mu) = F^\mu(x) \tag{35}$$

Proof. First, note that the FBE with first orthant constraints can also be written

$$F_\mu(x) = f(x) - \frac{\mu}{2}||\nabla f(x)||^2 + e_{\mu\delta_+}(x - \mu \nabla f(x)) \tag{36}$$

From Section 4, recall that if we fix (w, μ) in $L(x, w, \mu)$ and require $\nabla_x L(x, w, \mu) = 0$, then we find

$$w(x) = \mu \nabla f(x) - x$$

If we define $(w_-)_i = \min(w_i, 0)$ and w_+ similarly, then for any $v \in \mathbb{R}^n$, $v = v_- + v_+$, $v^T v_- = v_-^T v_-$, and $-v_+ = (-v)_-$. Insomuch, we find that

$$\begin{array}{rcl} L(x, w, \mu) & = & f(x) - \frac{1}{\mu}(\mu \nabla f(x) - x)^T x - \frac{1}{2\mu} x^T x - \frac{1}{2\mu}(\mu \nabla f(x) - x)_-^T (\mu \nabla f(x) - x)_- \\ & = & f(x) - \frac{\mu}{2}||\nabla f(x)||^2 + \frac{1}{2\mu}||x||^2 + \frac{\mu}{2}||\nabla f(x)||^2 - x^T \nabla f(x) \\ & & -\frac{1}{\mu}||(\mu \nabla f(x) - x)_-||^2 + \frac{1}{2\mu}||(\mu \nabla f(x) - x)_-||^2 \\ & = & f(x) - \frac{\mu}{2}||\nabla f(x)||^2 + \frac{1}{2\mu}||x - \mu \nabla f(x)||^2 + \frac{1}{2\mu}||(\mu \nabla f(x) - x)_-||^2 \\ & & +\frac{1}{\mu}(x - \mu \nabla f(x))^T (\mu \nabla f(x) - x)_- \\ & = & f(x) - \frac{\mu}{2}||\nabla f(x)||^2 + \frac{1}{2\mu}||(x - \mu \nabla f(x)) + (\mu \nabla f(x) - x)_-||^2 \\ & = & f(x) - \frac{\mu}{2}||\nabla f(x)||^2 + e_{\mu\delta_+}(x - \mu \nabla f(x)) \\ & = & F_\mu(x) \end{array}$$

□

As we see, the equivalence between these two exact penalty functions under first orthant constraints is fairly straight-forward. In seeking to generalize this equivalence, we will first state a few potentially helpful convex analysis tools.

Definition 1. *Given $f : dom\ f \subset \mathbb{R}^n \to \mathbb{R}$, the subdifferential of f at $x \in dom\ f$ is given by*

$$\partial f(x) = \{y \mid f(z) \geq f(x) + y^T(z - x) \text{ for all } z \in dom\ f\} \tag{37}$$

Definition 2. *Given $D \subset \mathbb{R}^n$, the support function of D is given by $\sigma_D : \mathbb{R}^n \to \overline{\mathbb{R}}$ where*

$$\sigma_D(y) = \sup_{x \in D} y^T x \tag{38}$$

Definition 3. *Given $f : \mathbb{R}^n \to \mathbb{R}$, the conjugate of f is given by $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ where*

$$f^*(y) = \sup_x \{y^T x - f(x)\} \tag{39}$$

Using conjugacy, if we apply the general definition of Rockafellar's augmented Lagrangian to the ICP, we have

$$\begin{array}{rcl} L(x, z^k; \mu) & = & f(x) + \sup\limits_{z \geq 0} \{-z^T g(x) - \frac{\mu}{2}||z - z^k||^2\} \\ & = & f(x) - \frac{\mu}{2}||z^k||^2 + \sup\limits_{z} \big\{z^T(-\frac{\mu}{2}z + \mu z^k - g(x)) \ - \delta_+(z)\big\} \\ & = & f(x) - \frac{\mu}{2}||z^k||^2 + (\delta_+(\cdot) + \frac{\mu}{2}|| \cdot ||^2)^*(\mu z^k - g(x)) \\ & = & f(x) - \frac{\mu}{2}||z^k||^2 + \sup\limits_{z} \big\{z^T(g(x) - \mu z^k) - \frac{\mu}{2}||z||^2 - \delta_+(-z)\big\} \\ & = & f(x) - \frac{\mu}{2}||z^k||^2 + (\delta_+^*(\frac{1}{\mu}\cdot) + \frac{1}{2\mu}|| \cdot ||^2)^*(\frac{1}{\mu}g(x) - z^k) \end{array} \tag{40}$$

where z^k is the current Lagrange multiplier estimate. The last line in this expression seems to offer some insight into how to extend the augmented Lagrangian beyond indicator functions on inequality bounds and toward general penalty functions. Our first remark is that δ_+ in the conjugate comes from requirement $z \geq 0$ in the augmented Lagrangian. Thus δ_+ is tethered to the constraint inequality $g(x) \geq 0$. This begs the question: how might a more general constraint function impact the operator $(\delta_+ + \frac{\mu}{2}|| \cdot ||^2)^*(\cdot)$?

Also note that the term $\mu z^k - g(x)$ closely resembles the Lagrange multiplier update as discussed in (5). This leads us to another question: What would be the Lagrange multiplier z^k for more general constraints? To shed some light, note that the 1st orthant QP has standard Lagrangian $L(x, z) = f(x) - z^T x$ giving the multiplier estimate $z^k = \nabla f(x^k)$. Applying this term to (40) gives an expression very similar to the FBE. For general separable constraints $g(x) = \sum_{i=1}^n g_i(x_i) \geq 0$, we find the multiplier estimate $z_i^k = \frac{[\nabla f(x)]_i}{\partial_i g_i(x_i)}$.

Generally, there are a few other properties of these convex analysis tools which appear useful for our goals. If f is strictly convex and the maximizer at y exists, then we have $\nabla f^*(y) = \text{argmax}_x\{y^T x - f(x)\}$. It can also be show that $(I + \partial\delta_+)^{-1} = \Pi_+$, and more generally $\text{prox}_{\lambda f} = (I + \lambda\partial f)^{-1}$ [16].

Finally, we seek to extend the smooth exact penalty method(s) above so they might handle equality constraints in addition to inequality constraints. In its current form, the FBE cannot take on equality constraits in an implementable way. If we apply $(h(x) = 0)$ to F_μ, this constraint takes the form of an indicator function and when implemented requires projection onto the nullspace of h, which is typically prohibitive. However, Michael Friedlander has suggest a method he and a colleague developed which offers a smooth exact penalty function for the ECP and is efficiently implementable. We will be examining this method and determining how it might be incorporated in the method(s) discussed above.

Proposed Exam Syllabus

Topic #1: Analysis [12]

- Metric and Banach Spaces
  - Compactness in finite-dimensions
  - Convergence in the uniform topology
  - Linear functionals and bounded linear maps
  - The kernel and range of linear maps
  - Convergence in the space of bounded linear operators
  - Dual spaces
- Hilbert Spaces
  - Inner products, orthogonality and projections
  - Self-adjoint operators and applications to matrix decomposition
- Fourier Analysis
  - The Fourier basis, definitions, and properties
  - Fourier series of differentiable functions
  - Wavelets
- Bounded Linear Operators on Hilbert Space
  - Orthogonal projections
  - Dual space of Hilbert space and representation theorems
  - Weak convergence in Hilbert space
  - Banach-Alaoglu Theorem
- Spectrum of Bounded Linear Operators
  - Diagonalization of matrices
  - Spectral theorem for compact, self-adjoint operators
  - Fredholm Alternative Theorem
  - Functions of operators
- Differential Calculus and Variational Methods
  - Derivatives of maps on Banach spaces
  - Fréchet and Gâteaux derivatives

Topic #2: Measure Theory

- Measures
  - σ-algebras and measure spaces
  - Construction of measure spaces from premeasures on algebras
  - Lebesgue-Stieltjes measures on $\mathbb{R}$
- Integration
  - Motivation of Lebesgue measure
  - The Lebesgue integral and associated convergence theorems
  - Convergence in measure, Egoroff's theorem
  - Product measures and the Fubini-Tonelli theorem
  - Lebesgue measures on $\mathbb{R}^n$
- Fundamental Theorem
  - Signed and complex measures
  - Radon-Nikodym theorem

Topic #3: Algebra [8]

- Modules
  - Basic definitions and constructions
  - Free and projective modules
  - Bilinear forms and structure of modules over principal ideal rings
- Tensor Products
  - Definition and basic properties
  - Tensor products of algebras and symmetric algebras
- Representation theory
  - Basic definitions of representations and characters
  - Representations of finite abelian groups
  - Class functions and orthogonality of characters

Topic #4: Penalty and Augmented Lagrangian Methods [14], [18]

- Duality
- Equality and Inequality Constraints
- Generalizations on Penalty Functions [2]
- Generalizations on Constraint Functions [1]
- Current Newton Methods [10]
- Relevant Textbook Chapters:
  - Numerical Optimization, Ch 12: Theory of Constrained Optimization
  - Numerical Optimization, Ch 16: Quadratic Programming
  - Numerical Optimization, Ch 17: Penalty and Augmented Lagrangian Methods
  - Variational Analysis, Ch 11: Dualization
  - Constrained Optimization and Lagrange Multiplier Methods, Ch 3: Inequality Constrained and Nondifferentiable Optimization
  - Constrained Optimization and Lagrange Multiplier Methods, Ch 5: Nonquadratic Penalty Functions - Convex Programming
  - Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Ch 19: Duality in Convex Optimization

Topic #5: Smooth Exact Penalty Functions and Extensions

- Moreau Envelope [1], [18]
- Proximal Methods
  - Proximal operator properties and interpretations [6], [16]
  - Elementary convergence properties [5], [7]
  - Alternating direction method of multipliers (ADMM) [16]
- Forward-Backward Envelope and Resulting Newton Methods [17]
- Subderivatives and Subgradients
- Relevant Textbook Chapters:
  - Variational Analysis, Ch 1: Max and Min
  - Variational Analysis, Ch 8: Subderivatives and Subgradients
  - Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Ch 12: Infimal Convolution
  - Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Ch 13: Conjugation

  - Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Ch 14: Further Conjugation Results
  - Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Ch 15: Fenchel-Rockafellar Duality

Topic #6: Numerical Methods and Additional Topics

- Numerical Methods
  - Conditioning, floating-point arithmetic, and stability of algorithms
  - QR decomposition
  - Singular Value Decomposition
  - Sparse LU/Cholesky factorization
  - Conjugate gradient method [14]
- Optimality Conditions [14]
  - First order conditions
  - Second order conditions
- Quasi-Newton Methods and Implementation Strategies
  - Accelerated proximal gradient algorithms [19]
  - Methods for adaptive line search and strong convexity estimation [11]
  - Proximal Newton-type methods [13]
- Applications
  - l_1-regularization in signal and image processing [15]
  - Support vector machines [3]
  - Clustering [3]

References

[1] Bauschke, Heinz H., and Patrick L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. New York: Springer, 2011.

[2] Bertsekas, Dimitri P. Constrained Optimization and Lagrange Multiplier Methods. New York: Academic, 1982.

[3] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York: Springer. 2006.

[4] Boyd, Stephen P., and Lieven Vandenberghe. Convex Optimization. Cambridge, UK: Cambridge University Press. 2004.

[5] Byrne, Charles L. "An Elementary Proof of Convergence for the Forward-Backward Splitting Algorithm." Journal of Nonlinear and Convex Analysis. 2013.

[6] Combettes, Patrick L., Jean-Christophe Pesquet. "Proximal Splitting Methods in Signal Processing." 2010.

[7] Combettes, Patrick L., Valerie R. Wajs. "Signal Recovery by Proximal Forward-Backward Splitting." SIAM Multiscale Modeling and Simulation 4.4, 2005.

[8] Dummit, David S., and Richard M. Foote. Abstract Algebra. New Dehli: Wiley, 2014.

[9] Facchinei, Francisco., Joaquim Judice, and Joao Soares. "An Active Set Newton Algorithm for Large-Scale Nonlinear Programs with Box Constraints." SIAM Journal on Optimization. 1998.

[10] Fletcher, Roger. "Augmented Lagrangians, Non-negative QP and Extensions." Preprint. 2015.

[11] Gonzaga, Clovis, Elizabeth Karas, Diane Rossetto. "An Optimal Algorithm for Constrained Differentiable Convex Optimization." SIAM Journal on Optimization 23.4, 2013.

[12] Hunter, John K., and Bruno Nachtergaele. Applied Analysis. Singapore: World Scientific, 2001.

[13] Lee, Jason D., Yuekai Sun, Michael A. Saunders. "Proximal Newton-Type Methods for Minimizing Composite Functions." SIAM Journal on Optimization 24.3, 2014.

[14] Nocedal, Jorge, and Stephen J. Wright. Numerical Optimization. New York: Springer, 1999.

[15] Palomar, Daniel P., and Yonina C. Eldar. Convex Optimization in Signal Processing and Communications. Cambridge: Cambridge University Press, 2010.

[16] Parikh, Neal, Boyd, Stephen. Proximal Algorithms. Now Pub, 2014.

[17] Patrinos, Panagiotis, Lorenzo Stella, Alberto Bemporad. "Forward-Backward Truncated Newton Methods for Convex Composite Optimization." arXiv preprint: 1402.6655, 2014.

[18] Rockafellar, R. Tyrrell, and Roger J. Wets. Variational Analysis. Berlin: Springer, 1998.

[19] Tseng, Paul. "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization." Technical Report. 2008.