Notes for 2017-04-26

1 Computing with constraints

Recall that our basic problem is

$$\text{minimize } \phi(x) \text{ s.t. } x \in \Omega$$

where the feasible set Ω is defined by equality and inequality conditions

$$\Omega = \{x \in \mathbb{R}^n : c_i(x) = 0, i \in \mathcal{E} \text{ and } c_i(x) \le 0, i \in \mathcal{I}\}.$$

In the last lecture, we described three different ways to formulate constrained optimization problem that allow us to build on techniques we previously explored from unconstrained optimization and equation solving:

1. *Constraint elimination* (for equality constraints): Find a parameterization $g : \mathbb{R}^{n-m} \to \Omega$ formulations and minimize $\phi(g(y))$ without constraints. This requires that the constraints be simple (e.g. affine equality constraints).

2. *Barriers and penalties:* Add a term to the objective function depending on some parameter μ. This term penalizes x values that violate the constraint (penalty methods) or that come close to $\partial\Omega$ from the inside (barrier methods). As $\mu \to 0$, the unconstrained minimum of the modified problems converges to the constrained minimum of the original.

3. *Lagrange multipliers*: Add new variables (multipliers) corresponding to "forces" needed to enforce the constraints. The *KKT conditions* are a set of nonlinear equations in the original unknowns and the multipliers that characterize constrained stationary points.

Our goal now is to sketch how modern constrained optimization algorithms incorporate these different ways of looking at the problem. A full treatment is well beyond the scope of the class, but we hope to give you at least the keywords you will need should you encounter them in a textbook, paper, or a cocktail party[1]. Ideally, knowing something about what happens in

---

[1] If you must, replace "cocktail party" with "job interview" — but if you do, I think you should seek more interesting cocktail parties. No, I do not get invited to many cocktail parties myself.

the algorithms will also help you think about which of various equivalent formulations of an optimization problem will be more (or less) friendly to solvers. The plan is to first give a "lay of the land" of different families of algorithms, then to give a more detailed treatment with the running example of linearly constrained quadratic programs.

2   Lay of the Land

As we mentioned before, problems with *inequality* constraints tend to be more difficult than problems with *equality* constraints alone, because it involves the combinatorial subproblem of figuring out which constraints are *active* (a constraint $c_i(x) \leq 0$ is active if $c_i(x) = 0$ at the optimum). Once we have figured out the set of active constraints, we can reduce an inequality-constrained problem to an equality-constrained problem. Hence, the purely equality-constrained case is an important subproblem for inequality-constrained optimizers, as well as a useful problem class in its own right.

For problems with only equality constraints, there are several standard options:

- *Null space methods* deal with linear equality constraints by reducing to an unconstrained problem in a lower-dimensional space.

- *Projected gradient methods* deal with simple equality constraints by combining a (scaled) gradient step and a projection onto a constraint set.

- *Penalty methods* approximately solve an equality-constrained problem through an unconstrained problem with an extra term that penalizes proposed soutions that violate the constraints. That is, we use some constrained minimizer to solve

$$\text{minimize } \phi(x) + \frac{1}{\mu} \sum_{i \in \mathcal{E}} c_i(x)^2.$$

  As $\mu \to 0$, the minimizers to these approximate problems approach the true minimizer, but the Hessians that we encounter along the way become increasingly ill-conditioned (with condition number proportional to μ^{-1}).

- *KKT solvers* directly tackle the first-order optimality conditions (the KKT conditions), simultaneously computing the constrained minimizer and the associated Lagrange multipliers.

- *Augmented Lagrangian* methods combine the advantages of penalty methods and the advantages of the penalty formulation. In an augmented Lagrangian solver, one finds critical points for the augmented Lagrangian

$$\mathcal{L}(x, \lambda; \mu) = \phi(x) + \frac{1}{\mu} \sum_{i \in \mathcal{E}} c_i(x)^2 + \lambda^T c(x)$$

  by alternately adjusting the penalty parameter μ and the Lagrange multipliers.

In the inequality-constrained case, we have

- *Active set methods* solve (or approximately solve) a sequence of equality-constrained subproblems, shuffling constraints into and out of the proposed working set along the way. These methods are particularly attractive when one has a good initial estimate of the active set.

- *Projected gradient methods* deal with simple inequality constraints by combining a (scaled) gradient step and a projection onto a constraint set.

- *Barrier methods* and *penalty methods* add a term to the objective function in order to penalize constraint violations or near-violations; as in the equality-constrained case, a parameter μ governs a tradeoff between solution quality and conditioning of the Hessian matrix.

- *Interior point methods* solve a sequence of barrier subproblems using a continuation strategy, where the barrier or penalty parameter μ is the continuation parameter. This is one of the most popular modern solver strategies, though active set methods may show better performance when one "warm starts" with a good initial guess for the solution and the active set of constraints.

As with augmented Lagrangian strategies in the equality-constrained case, state-of-the art strategies for inequality-constrained problems often combine

approaches, using continuation with respect to a barrier parameters as a method of determining the active set of constraints in order to get to an equality-constrained subproblem with a good initial guess for the solution and the Lagrange multipliers.

The *sequential quadratic programming* (SQP) approach for nonlinear optimization solves a sequence of linearly-constrained quadratic optimization problems based on Taylor expansion of the objective and constraints about each iterate. This generalizes simple Newton iteration for unconstrained optimization, which similarly solves a sequence of quadratic optimization problems based on Taylor expansion of the objective. Linearly-constrained quadratic programming problems are hence an important subproblem in SQP solvers, as well as being an important problem class in their own right.

3   Quadratic programs with equality constraints

We begin with a simple case of a quadratic objective and linear equality constraints:

$$\phi(x) = \frac{1}{2}x^T H x - x^T d$$
$$c(x) = A^T x - b = 0,$$

where $H \in \mathbb{R}^{n \times n}$ is symmetric and positive definite *on the null space of A^T* (it may be indefinite or singular overall), $A \in \mathbb{R}^{n \times m}$ is full rank with $m < n$, and $b \in \mathbb{R}^m$. Not only are such problems useful in their own right, solvers for these problems are also helpful building blocks for more sophisticated problems — just as minimizing an unconstrained quadratic can be seen as the starting point for Newton's method for unconstrained optimization.

3.1   Constraint elimination (linear constraints)

As discussed last time, we can write the space of solutions to the constraint equations in terms of a (non-economy) QR decomposition of A:

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

where Q_2 is a basis for the null space of A. The set of solutions satisfying the constraints $Ax = b$ is

$$\Omega = \{u + Q_2 y : y \in \mathbb{R}^{(n-m)}, u = Q_1 R_1^{-T} b\};$$

here u is a *particular solution* to the problem. If we substitute this parameterization of Ω into the objective, we have the unconstrained problem

$$\text{minimize } \phi(u + Q_2 y).$$

While we can substitute directly to get a quadratic objective in terms of y, it is easier (and a good exercise in remembering the chain rule) to compute the stationary equations

$$0 = \nabla_y \phi(u + Q_2 y) = \left(\frac{\partial x}{\partial y} \right)^T \nabla_x \phi(u + Q_2 y)$$

$$= Q_2^T(H(Q_2 y + u) - d) = (Q_2^T H Q_2)y - Q_2^T(d - Hu).$$

In general, even if A is sparse, Q_2 may be dense, and so even if H is dense, we find that $Q_2^T H Q_2$ is dense.

3.2   Barriers, penalties, and conditioning

Now consider a penalty formulation of the same equality-constrained optimization function, where the penalty is quadratic:

$$\text{minimize } \phi(x) + \frac{1}{2\mu}\|A^T x - b\|^2.$$

In fact, the augmented objective function is again quadratic, and the critical point equations are

$$(H + \mu^{-1} A A^T)x = d + \mu^{-1} A b.$$

We can analyze this more readily by changing to the Q basis from the QR decomposition of A that we saw in the constraint elimination approach:

$$\begin{bmatrix} Q_1^T H Q_1 + \mu^{-1} R_1 R_1^T & Q_1^T H Q_2 \\ Q_2^T H Q_1 & Q_2^T H Q_2 \end{bmatrix} (Q^T x) = \begin{bmatrix} Q_1^T d + \mu^{-1} R_1 b \\ Q_2^T d \end{bmatrix}$$

Taking a Schur complement, we have

$$(\mu^{-1} R_1 R_1^T + F)(Q_1^T x) = \mu^{-1} R_1 b - g$$

where

$$F = Q_1^T H Q_1 - Q_1^T H Q_2 (Q_2^T H Q_2)^{-1} Q_2^T H Q_1$$
$$g = [I - Q_1^T H Q_2 (Q_2^T H Q_2)^{-1} Q_2^T] d$$

As $\mu \to 0$, the first row of equations is dominated by the μ^{-1} terms, and we are left with

$$R_1 R_1^T (Q_1^T x) - R_1 b \to 0$$

i.e. $Q_1 Q_1^T x$ is converging to $u = Q_1 R_1^{-T} b$, the particular solution that we saw in the case of constraint elimination. Plugging this behavior into the second equation gives

$$(Q_2^T H Q_2)(Q_2^T x) - Q_2^T (d - Hu) \to 0,$$

i.e. $Q_2^T x$ asymptotically behaves like y in the previous example. We need large μ to get good results if the constraints are ill-posed or if $Q_2^T H Q_2$ is close to singular. But in general the condition number scales like $O(\mu^{-1})$, and so large values of μ correspond to problems that are numerically unattractive.

3.3   Lagrange multipliers and KKT systems

The KKT conditions for our equality-constrained problem say that the gradient of

$$L(x, \lambda) = \phi(x) + \lambda^T (A^T x - b)$$

should be zero. In matrix form, the KKT system (saddle point system)

$$\begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} d \\ b \end{bmatrix}.$$

If A and H are well-conditioned, then so is this system, so there is no bad numerical behavior. The system also retains whatever sparsity was present in the original system matrices H and A. However, adding the Lagrange multipliers not only increases the number of variables, but the extended system lacks any positive definiteness that H may have.

The KKT system is closely related to the penalty formulation that we saw in the previous subsection, in that if we use Gaussian elimination to remove the variable λ in

$$\begin{bmatrix} H & A \\ A^T & -\mu I \end{bmatrix} \begin{bmatrix} \hat{x} \\ \lambda \end{bmatrix} = \begin{bmatrix} d \\ b \end{bmatrix},$$

we have the Schur complement system

$$(H + \mu^{-1} A A^T)\hat{x} = d + \mu^{-1} A b,$$

which is identical to the stationary point condition for the quadratically penalized objective.

3.4   Augmented Lagrangian

From a solver perspective, the block 2-by-2 structure of the KKT system looks highly attractive. Alas, we do *not* require that H be positive definite, nor even that it be nonsingular; to have a unique global minimum, we only need positive definiteness of the projection of H onto the null space (i.e. $Q_2^T H Q_2$ should be positive definite). This means we cannot assume that (for example) H will admit a Cholesky factorization.

The augmented Lagrangian approach can be seen as solving the constrained system

$$\text{minimize } \frac{1}{2} x^T H x - d^T x + \frac{1}{2\mu} \|A^T x - b\|^2 \text{ s.t. } A^T x = b.$$

The term penalizing nonzero $\|A^T x - b\|$ is, of course, irrelevant at points satisfying the constraint $A^T x = b$. Hence, the constrained minimum for this augmented objective is identical to the constrained minimum of the original objective. However, if the KKT conditions for the modified objective take the form

$$\begin{bmatrix} H + \mu^{-1} A A^T & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} d + \mu^{-1} A b \\ b \end{bmatrix}.$$

Now we do not necessarily need to drive μ to zero to obtain a good solution; but by choosing μ small enough, we can ensure that $H + \mu^{-1} A A^T$ is positive definite (assuming that the problem is convex subject to the constraint).