15.084J Recitation Handout 5

Fifth Week in a Nutshell:

- When is KKT Necessary
- Sufficient Conditions
- Steepest Descent for Constrained Problems

New Definitions

- x is a slater point for a set of constraints if x is *strictly* feasible for all inequalities, and feasible for all equalities.
- f is pseudoconvex if for all $x, y \in X$ $\nabla f(x)^t(y-x) \geq 0 \rightarrow f(y) \geq f(x)$. The important thing is that if you move in an increasing direction from x, the value of the function will never be lower than the value at x.
- f is quasiconvex if for all $0 \leq \lambda \leq 1$, and all $x, y \in X$, $f(\lambda x + (1-\lambda)y) \leq \max(f(x), f(y))$. The important thing is that the level sets are convex; if $f(x) \leq \alpha, f(y) \leq \alpha$ then the entire line between x and y has values of f that are at most α

Any differentiable convex function is pseudoconvex.
Any convex function, and any pseudoconvex function, is quasiconvex.
Pseudo and quasi concave have the obvious definitions.

When is KKT a Necessary or Sufficient Condition?

We know that the Fritz-John conditions are necessary. Unfortunately, it allows points where the constraints are doing something strange without regard to the function; sometimes these are optimal, but often we want to rule them out to get the stronger KKT conditions.

Thus, we'd like to say "If *blah*, then KKT is a necessary condition". What can we fill in for *blah*? Alternately, can we say "KKT+*blah* is sufficient"?

Necessary:

- The gradients of constraints at $\bar{x}$ are linearly independent – this rules out the possibility of a combination of them equalling zero, but unfortunately may depend on $\bar{x}$.
- Slater condition: A slater point exists, g_i are pseudo-convex, and ∇h_i are linearly independent.
- All constraints are linear. This rules out bad things happening at the intersection of constraints.

Sufficient:

- Convexity: In unconstrained optimization, $\nabla f = 0$ is necessary, and combined with f convex was sufficient. Now, KKT is sufficient when combined with f is convex, g is convex, and both h and $-h$ are convex (so h is linear). This makes our feasible region convex, ruling out the bad cases of the F-J conditions.
- Weak Convexity 1: We don't actually need g to be convex; all we need is for $g(x) \leq 0$ to be convex. This happens when g is quasiconvex. And sometimes just when that level set is nice.
- Weak Convexity 2: We don't actually need f convex; pseudo-convex works, because it guarantees that increasing directions stay increasing.
- Weak Convexity 3: We don't actually need h linear; if you have a nonlinear h, you might be able to change variables to reduce the dimension and remove h from the problem (but that may affect convexity of f and g).
- Hessian: In unconstrained, we required $dHd \geq 0$ for all d. Now we require it for all *feasible* d, and use the hessian of the Lagrangian. This is usually sufficiently nasty as to be pointless.

Steepest Descent for Constrained Problems

The basic idea behind steepest descent: move in the most improving direction. Keep going until it isn't improving anymore. What to do with constrained problems? How about moving in the most improving *feasible* direction, and keep going until it isn't improving anymore?

What is the most improving feasible direction? $min_d f(x)d, ||d|| \leq 1, d\texttt{feasible}$. And how do we express the feasibility of d? If we've got linear equality constraint $Ax = b$, and we're at a feasible point, we need $Ad = 0$.

Also, what if we want to use a different metric than the euclidean ball? Pick a $Q \succ 0$, and the constraint is $d^t Q d \leq 1$.

Now we have $min_d \nabla f(x)^t d$ subject to $d^t Q d \leq 1, Ad = 0$

This can be solved via a couple pages of algebra (see notes), in closed form; thus in practice, it's not meaningfully harder than in the unconstrained case.

Performance? Same as unconstrained case. Convergence? Always hits a KKT point.

Why Q? $Q = H(x)$ gives you Newton's Method! Sadly, it also forces you to calculate and invert H, and gives you the bad convergence when far from an optimum of Newton. $Q = H(x) + \delta I$, with δ shrinking over some iterations, gives you roughly steepest descent at first (when far from optimum), and Newton later, when close to optimum – best of both worlds! If your problem is poorly conditioned, a properly chosen Q can improve the conditioning, and make for better convergence (but such Q is potentially hard to find).