Advanced Mathematical Programming
IE417

Lecture 16

Dr. Ted Ralphs

Reading for This Lecture

- Sections 8.6-8.8

Method of Steepest Descent

- Up until now, we discussed methods that use only function evaluations.
- As before, if the objective function is differentiable, we can use the derivative to guide the search.
- Recall the direction of steepest descent at x^* is $-\nabla f(x)$.
- Method of steepest descent: Iteratively perform line searches in the direction of steepest descent.
- Because this is a line search algorithm, it will converge as long as f is continuous and differentiable.

Problems with this Algorithm

- This algorithm can have problems if the Hessian is ill-conditioned.
- This is essentially because the linear approximation is not good when the gradient is near zero.
- In this case, the error term in the approximation begins to dominate.
- In the worst case, the search path can zigzag wildly.

Convergence Rate

- Suppose the Hessian has a condition number α.
- If $\alpha >> 1$, this means that the second-order approximation to the function is highly non-circular.
- This is what causes the zigzagging.
- Under mild conditions, if f is continuous and twice-differentiable, it can be shown that the convergence rate of this algorithm is linear and bounded by $(\alpha - 1)^2/(\alpha + 1)^2$.

Armijo's Rule

- Substitute for exact line search.
- Driven by two parameters, $0 < \varepsilon < 1$ and $\alpha > 1$.
- We define

$$\Theta(\lambda) = f(x^* + \lambda d) \text{ and } \Theta'(\lambda) = \Theta(0) + \lambda\varepsilon\nabla\Theta(0).$$

- A step length λ^* is considered *acceptable* if

$$\Theta(\lambda^*) \leq \Theta'(\lambda^*) \text{ and } \Theta(\alpha\lambda^*) > \Theta'(\alpha\lambda^*)$$

Convergence of Steepest Descent

Definition 1. *A function f is* Lipschitz continuous *with constant G if*

$$\|f(x) - f(y)\| \leq G\|x - y\|$$

- Because this is a line search algorithm, it will converge as long as f is continuous and differentiable and we use exact line search.
- A version using Armijo's Rule is also guaranteed to converge as long as $\nabla f(x)$ is Lipschitz continuous with constant $G > 0$.

Newton's Method

- Essentially the same as in one-dimensional search.
- Use a quadratic approximation of the function.
- Take the derivative and set it to zero.
- Then, $x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x)$.
- Note that this can be interpreted as a steepest descent method with affine scaling.
- In essence, we are reversing the effect of an ill-conditioned Hessian.
- This method will converge in one step for quadratics.

Comments on Newton's Method

- $H(x_k)$ must have full rank.
- This implies that we can only converge to local optima with positive definite Hessians.
- Note that if $cond(H(x_k)) >> 1$, then finding the next iterate is an ill-conditioned problem.
- As long as our starting solution is "close enough" to a local optima x^* with $H(x^*)$ positive definite, this method will converge at least quadratically.
- Proof is using Theorem 7.2.3 with $\alpha(x) = \|x - x^*\|$.

Modifying Newton's Method

- Problems with the method
  - May not be defined if $H(x_k)$ does not have full rank.
  - Step size is fixed and may not give descent in f.
  - Not globally convergent.
- Levenberg-Marquardt Methods
  - For $\delta > 0$, choose $\varepsilon \geq 0$ to be the smallest scalar such that all the eigenvalues of the matrix $\varepsilon I + H$ are $\geq \delta$.
  - Perform line search in the direction $-B\nabla f(x)$ where $B = (\varepsilon I + H)^{-1}$.

Comments on Levenberg-Marquardt Methods

- By construction, the direction $-B\nabla f(x)$ is a descent direction and hence f is a descent function in Theorem 7.2.3. Hence, these methods are globally convergent.
- Implementation
  - Given x_k, try to find the Cholesky factorization of $\varepsilon_k I + H_k$.
  - If unsuccessful, increase ε_k and repeat.
  - Otherwise, solve $LL^T(x_k - x_{k+1}) = -\nabla f(x)$.
  - Compute R_k, the ratio of predicted to actual descent.
  - If $R_k < 0.25$, increase ε. If $R_k > 0.75$, decrease ε.

Trust Region Methods

- Similar to the implementation of L-M methods just described.
- Define $\Omega_k = \{x : \|x - x_k\| \leq \Delta_k\}$, a *trust region* over which the quadratic approximation to f is "good."
- At each step, solve $\min\{q(x) : x \in \Omega_k\}$ where q is the quadratic approximation.
- If R_k, ratio of actual to predicted decrease, is less than 0.25, then decrease Δ. If $R_k > 0.75$, increase Δ.
- The dog-leg trajectory is another similar method.