THE UNIVERSITY OF CHICAGO

11. Nonlinear Equations Short

Example Nonlinear Equations: Quasi-Likelihoods

- (Good reference: chapter 9 in McCullagh and Nelder, Generalized Linear Models)
- Assume that we have a vector of responses, **Y**, which are independent with mean **m** and variance function that depends on the mean, e.g $\sigma^2 V(\mathbf{m})$ (but I do not know the distribution itself, i.e the likelihood).
- The function $U(\mathbf{m}) = \frac{Y - \mathbf{m}}{\sigma^2 V(\mathbf{m})}$ behaves like a gradient of the log-likelihood w.r.t m, e.g. $\mathrm{E}(U(\mathbf{m})) = 0$

Maximum Likelihood Estimation (MLE)

Consider you want to maximize the Likelhood of a Gaussian process

$$\max_{\theta} \quad -\frac{1}{2} y^T K^{-1} y - \frac{1}{2}\log(\det K) - \frac{n}{2}\log 2\pi$$

- The classical solutions to compute log(det) require factorization of matrix K . But for 1B data points and dense K, you need 8*10^18 bytes to store the matrix
- Idea use the score equations (plus a random UE of trace)

$$\frac{1}{2} y^T K^{-1}(\partial_j K) K^{-1} y - \frac{1}{2}\operatorname{tr}\left[K^{-1}(\partial_j K)\right] = 0$$

- Nonlinear equations appear in estimation many times without the accompanying optimization problem

Nonlinear Equations

- Solve the equation (nonlinear system of equations) :

$$r(x) = 0, \quad r : \mathbb{R}^n \to \mathbb{R}^n$$

- Here r is the vector function

$$r(x) = \left[r_1(x), r_2(x), \ldots, r_n(x) \right]^T$$

- A nonlinear system of equations can have none, one, or multiple solutions

$$x^2 = -1; \quad x_1 = 1; \quad \sin(x_1) = 0.5$$

- We cannot answer the global questions again as in optimization, but we aim to find one solution as fast as we can.

Connections with least squares:

- Note that *we can write the problem as a nonlinear least squares problem*.

$$\min_x f(x) := \sum_{i=1}^{n} r_i(x)^2$$

- Some some techniques will replicate the ones from least squares (though m=n).
- But there are also differences.
  - Once continuous differentiability of r_i is sufficient.
  - Quasi-Newton methods are not as efficient here (if n is larger, density ..)
  - There is no natural minimization fun, there are many (f is just one of them) but none is “ideal”.

11.1 Local methods.Newton's method

Algorithm 11.1 (Newton's Method for Nonlinear Equations).

Choose x_0;
for $k = 0, 1, 2, \ldots$
Calculate a solution p_k to the Newton equations

$$J(x_k)p_k = -r(x_k);$$

$x_{k+1} \leftarrow x_k + p_k$;
end (for)

- Thm 11.2: If r(x) is continuously differentiable and

$J(x^*)$ is nonsingular $J(x) = \nabla_x r(x)$ then, if starting Algorithm 11.1 sufficiently close to x^* then x_k converges superlinearily to x^*

- If r(x) is Lipschitz continuously differentiable, the the convergence is Q-quadratic.
- If the Newton system is singular, you have arrived at stationary point for the associated nonlinear least squares.

11.2 Practical Line Search Methods

- Based on the observation that the Newton direction for nonlinear equations is a descent direction for f(x)!
- Indeed: $J(x_k)p_k = -r(x_k)$ implies that

$$p_k^T \nabla f(x_k) = -p_k^T J_k^T r_k = -\|r_k\|^2 < 0$$

- We get global convergence from the equivalent of Zoutendijk's theorem applied here Theorem 11.6.
- This implies that $J_k^T r_k \to 0$
- If the limit point is not degenerate (the Jacobian is not singular) then we obtain a solution of the problem.
- Similarly to Line Search convergence method/newton method we also get superlinear/quadratic convergence of this method.
- We can do this with backtracking, or Wolfe.