11.1

Unconstrained Minimization in R^n

Katta G. Murty, IOE 611 Lecture slides

Consider: $\min \theta(x)$ over $x \in R^n$

We consider **Descent Methods** first. Each iteration in these methods consists of 2 steps:

1. Find a search direction (a descent direction) $\bar{y}$ at current point $\bar{x}$.

2. Use a line search method to find the step length $\bar{\lambda}$. $\bar{x} + \bar{\lambda}\bar{y}$ is the new point, terminate with that as the best point if practical termination conds. are met. Otherwise go to the next iteration with the new pt. as current pt.

Steepest Descent Method: Cauchy (1847). A gradient method.

Search direction is steepest descent direction with I as the metric matrix, it is $-(\nabla\theta(x))^T$ at x.

Method globally convergent even with inexact line searches. Far from optimum the method works well, but in the nbhd. of a stationary pt. method very slow, taking small nearly orthogonal steps (zigzagging).

If Hessian at opt. is PD and its condition no. = $\frac{\text{its largest eigen value}}{\text{its smallest eigen value}} = \alpha$; rate of convergence of method becomes increasingly slower as $\alpha \to \infty$ depending on initial sol. x^0. Convergence linear with rate bounded above by $\frac{(\alpha-1)^2}{(\alpha+1)^2}$.

Newton's Method for Unconstrained Minimization

When $\bar{x}$ is current point, method approximates $\theta(x)$ around $\bar{x}$ by quadratic function $\theta(\bar{x}) + \nabla\theta(\bar{x})(x - \bar{x})$ $+\frac{1}{2}(x - \bar{x})^T \nabla^2_{xx}\theta(\bar{x})(x - \bar{x})$.

1st order nec. conds. for minimum of quad. appro. is that $y = x - \bar{x}$ satisfy $\nabla^2_{xx}\theta(\bar{x})y = -(\nabla\theta(\bar{x}))^T$. This gives the iteration:

$$x^{r+1} = x^r + y^r$$

where $\nabla^2_{xx}\theta(x^r)y^r = -(\nabla\theta(x^r))^T$.

If Hessian is PD, y^r is unique and method well defined. The direction y^r called **Newton direction** for $\theta(x)$ at x^r. Traditional Newton step length is always 1.

Variable metric property.

Theorem: Suppose $H(x) = \nabla^2_{xx}\theta(x) = (h_{ij}(x))$ is PD and each $h_{ij}(x)$ is Lipschitz cont. with constant γ. Let $\bar{x}$ satisfy $\nabla\theta(\bar{x}) = 0$. If x^0 is close to $\bar{x}$, $\{x^r\}$ converges to $\bar{x}$ at 2nd order rate.

Levenberg-Marquardt Modified Newton Method

Problems with Newton's method:

1. $H(x)$ may not be PD (then Newton direction may not be descent direction), may be singular (then Newton direction not defined).
2. Step length of 1 may not give desceny in $\theta(x)$. Fixed by choosing step length by a line search routine.
3. Not globally convergent.

To avoid 1, 3, use as search direction $-(\epsilon I + H(x))^{-1}\nabla\theta(x)$ where $\epsilon > 0$ choosen so that $\epsilon I + H(x)$ is PD.

If ϵ too small, $\epsilon I + H(x)$ may be near singular. If ϵ too large, $\epsilon I + H(x)$ becomes diagonally dominant & method behaves like steepest descent with linear convergence rate.

To choose ϵ start with some $\epsilon > 0$ & ascertain PD of $\epsilon I + H(x)$ by attempting to construct its Cholesky factorization. If unsuccessful, multiply ϵ by 4 and repeat until such a factorization is available.

Model-Trust Region Methods

Another group ensuring global convergence while ensuring fast local convergence.

The N. & L. M. M. N. methods used a quad. model of function to determine search direction, and then a line search technique to determine step length in that direction. The line search technique does not use the Hessian or the full dimensional quad. model of the function. 2nd difficulty with these methods is that *region of trust* within which quad. approx. at current pt. is sufficiently reliable may not include next pt. x^{r+1}.

T. R. methods circumvent these problems by 1st choosing a trial step length Δ_r (within which quad. approx. is considered good), and then uses the quad. model to select the best step of at most length Δ_r by solving:

$$\begin{aligned} \min Q(x) &= \theta(x^r) + \nabla\theta(x^r)s + \frac{1}{2}s^T H s \\ \text{s. to} \quad ||s|| &\leq \Delta_r \end{aligned}$$

where $H = \nabla^2_{xx}\theta(x^r)$ is the Hessian and $s = x - x^r$. Δ_r is an estimate of how far we can trust quad. model, so it is called **trust radius**.

Case 1: If H PD and $||H^{-1}(\nabla(\theta(x^r))^T|| \leq \Delta_r$, $s^r = H^{-1}(\nabla(\theta(x^r))^T$ is unique sol.

Case 2: Otherwise, sol. s^r satisfies $||s^r|| = \Delta^r$ and $(H + \mu_r I)s^r = -(\nabla\theta(x^r))^T$ where $\mu_r \geq 0$ is s. th. $H + \mu_r I$ is at least PSD.

In this case if H is PD, sol. given by $s^r = -(H+\mu_r I)^{-1}(\nabla\theta(x^r))^T$ where $\mu_r > 0$ is s. th. $||s^r|| = \Delta_r$.

If H indefinite, let $\lambda_1 =$ its smallest eigen value and let $v^1 \in R^n$ denote its corresponding eigen vector. Then:

Either $H + \mu_r I$ is PD and $s^r = -(H + \mu_r I)^{-1}(\nabla\theta(x^r))^T$ for the unique $\mu_r > \max\{0, -\lambda_1\}$ for which $||s^r|| = \Delta_r$,

Or $\mu_r = -\lambda_1$ and $s^r = -(H + \mu_r I)^+(\nabla\theta(x^r))^T + \omega v^1$, where $\omega \in R^1$ is choosen so that $||s^r|| = \Delta_r$, and $(H + \mu_r I)^+$ is the Moore-Penrose pseudoinverse.

In practice, μ_r calculated approx. by an iterative process with each iteration requiring the Cholesky factorization of a matrix of form $(H + \mu I)$ as described above.

Approx. sol. given by following **dog-leg step**. CP = Cauchy pt., NP = Newton pt., x^{r+1} is intersection of line segment joining CP and NP with sphere if intersection exists, or take x^{r+1} as NP itself.

Procedure for Selecting Δ_r

After x^{r+1} obtained with trial value of Δ_r, compute

$$R_r = \frac{\theta(x^r) - \theta(x^{r+1})}{Q(x^r) - Q(x^{r+1})}$$

If :

$0 < R_r < 0.25$, make $\Delta_{r+1}\frac{\Delta_r}{4}$

$R_r > 0.75$ and $||x^{r+1} - x^r|| = \Delta_r$, make $\Delta_{r+1} = 2\Delta_r$.

$R_r \leq 0$, i.e., $\theta(x)$ did not improve in this iteration, reject x^{r+1}, keep x^r as new pt. and repeat this step with $\Delta_{r+1} = \frac{\Delta_r}{4}$.

Otherwise keep $\Delta_{r+1} = \Delta_r$ and continue with x^{r+1} as new pt.

Quasi-Newton Methods

Methods based on building up the Hessian through the computed values of $\nabla\theta(x)$ & $\theta(x)$. Also called **Secant Methods**.

B_r = approximation to Hessian at rth step

Methods usually begin with $B_0 = I$, and many of these methods maintain B_r symmetric and PD.

If x^r is current pt. & B_r the current approx., then the search direction at x^r is $s^r = -B_r^{-1}(\nabla\theta(x^r))^T$; and a line search is performed to get the next point.

Variable metric property.

Quasi-Newton Condition: From Taylor series expansion we have $(\nabla\theta(x^{r+1}))^T \approx (\nabla\theta(x^r))^T + \nabla^2_{xx}\theta(x^r)(x^{r+1} - x^r)$. The condition requires that B_{r+1} satisfy

$$(\nabla\theta(x^{r+1}) - \nabla\theta(x^r))^T = B_{r+1}(x^{r+1} - x^r)$$

Updating formulae generally also have **hereditary symmetry, hereditary PD** properties.

Most successful of these methods is the BFGS method, which uses the updating formula:

$$B_{r+1} = B_r + \frac{y^r(y^r)^T}{(y^r)^T s^r} - \frac{B_r s^r (s^r)^T B_r}{(s^r)^T B_r s^r}$$

where $s^r = x^{r+1} - x^r$, $y^r = (\nabla\theta(x^{r+1}) - \nabla\theta(x^r))^T$.

In implementing this method, instead of updating B_r, the Cholesky factor of B_r is directly updated.

Method also called **PD Secant method**.

Method usually reset after n iterations.

Conjugate Direction (Gradient) Methods

Introduced by Hestenes & Stiefel (1952), originally for solving $Ax = b$, square nonsingular system, through min $(Ax - b)^T(Ax - b)$.

These methods use only 1st order derivatives, and do not need storing or updating a square matrix.

First consider min $f(x) = cx + \frac{1}{2}x^T Ax$ where A is PD Symmetric. These methods developed originally to solve this problem using at most n line searches.

Conjugacy wrt A (PD Symmetric): Set of nonzero vectors $\{p_{.1}, \dots, p_{.n}\}$ is said to be conjugate wrt A iff $p_{.i}Ap_{.j} = 0 \quad \forall i \neq j$.

Let $P_{n\times n}$ be s. th. its set of col vectors is conjugate wrt A. Then the linear transformation $x = Pz$ diagonalizes $f(x)$ into $F(z) = cPz + \frac{1}{2}P^T APz$. $P^T AP$ is diagonal because of the conjugacy condition. Hence $F(z)$ is separable in the z variables, i.e., $F(z) = \Sigma_{j=1}^{n} F_j(z_j)$, so minimizing $F(z)$ can be carried out through n line searches by the alternating variable method. So, in the x-space, $f(x)$ can be minimized through n line searches, once in each of the directions $\{P_{.1}, \ldots, P_{.n}\}$ in any order.

The C. G. methods generate the C. directions one after the other, so that each is a descent direction at current pt. at the time that direction is generated.

General C. G. Method: Step 1: Initiate with any pt. x^0. Search direction in this step is the steepest descent direction $y^0 = -(\nabla f(x^0))^T$. Do a line search.

General Step: Let x^r be current pt. Search direction is $y^r = -(\nabla f(x^r))^T + \beta_r y^{r-1}$. Do a line search. to get next pt. x^{r+1}. If $r + 1 = n$, this pt. is optimal, terminate. Otherwise go to the next step.

Different C. G. methods use different formula for β_r. These are:

$$\beta_r = \begin{cases} \frac{||\nabla f(x^r)||^2}{||\nabla f(x^{r-1})||^2} & \text{Fletcher \& Reeves method} \\ \frac{(\nabla f(x^r) - \nabla f(x^{r-1}))(\nabla f(x^r))^T}{||\nabla f(x^{r-1})||^2} & \text{Polak, Ribiere, Polyak method} \\ \frac{-||\nabla f(x^r)||^2}{(\nabla f(x^{r-1}))y^r} & \text{ C. descent method} \end{cases}$$

Each direction is descent direction at current pt. if all line searches are carried exactly. For quad. function $f(x)$ all above formulae give same value to β_r if all line searches carried exactly.

To min general nonlinear function $\theta(x)$ apply same method. Won't guarantee min in exactly n steps. Method usually reset after every n steps, or whenever generated direction not descent at current pt. When n large, 2nd method seems better.

Practical Termination Conds.

Terminate in Step r, when some or all of these quantities are small:

$$|\theta(x^r) - \theta(x^{r-1})|, \; ||\nabla\theta(x^r)||, \; ||x^r - x^{r-1}||.$$

The Simplex Direct Search Method for NLP

Nelder & Mead (1965). A pattern search technique useful for problems of low dimension. Not good when dimension is high. Aim is to obtain a small simplex containing a minimum. In the end either the best vertex is taken as the best pt., or the best pt. obtained by some interpolation in final simplex.

Four types of moves. Each iteration begins with a simplex $< x^1, \ldots, x_{n+1} >$ with its vertices sorted so that $\theta(x^i) \leq \theta(x^{i+1})$ $\forall i$.

1. Method tries to replace worst vertex x^{n+1} by a better pt. Best facet of current simplex is $< x^1, \ldots, x^n >$ (it excludes worst vertex), its centroid is $\bar{x} = \frac{1}{n} \Sigma_{i=1}^n x^i$. Reflexion x^r of x^{n+1} through $\bar{x}$ in this facet is $x^r = 2\bar{x} - x^{n+1}$.

1.1. If $\theta(x^r) < \theta(x^1)$, very successful. Now try expanding simplex in same direction. Let $x^e = 2x^r - x_{n+1}$.

Expansion successful if $\theta(x^e) < \theta(x^1)$. In this case, the new simplex is $< x^e, x^1, \ldots, x^n >$, go to next iteration.

If expansion unsuccessful (i.e., $\theta(x^e) \geq \theta(x^1)$), new simplex is $< x^r, x^1, \ldots, x^n >$, go to next iteration.

1.2. If $\theta(x^1) \leq \theta(x^r) \leq \theta(x^n)$, new simplex is $< x^1, \ldots, x^n, x^r >$, sort its vertices & go to next iteration.

1.3. If $\theta(x^r) \geq \theta(x^{n+1})$, try to contract simplex internally along reflection direction. Let $x^c = \frac{1}{2}(\bar{x} + x^{n+1})$.

Contraction successful if $\theta(x^c) < \theta(x^{n+1})$, new simplex is $< x^1, \ldots, x^n, x^c >$, sort vertices & go to next iteration.

Contraction failed if $\theta(x^c) \geq \theta(x^{n+1})$, now shrink simplex (we tacitly assume that we are close enough to minimizer to need smaller moves for improvement). New simplex is $< x^1, \frac{x^1+x^2}{2}, \ldots, \frac{x^1+x^{n+1}}{2} >$, sort vertices & go to next iteration.

1.4. If $\theta(x^n) \leq \theta(x^r) \leq \theta(x^{n+1})$, define the shadow contraction pt. $x^{sc} = \frac{1}{2}(x^r + x^{n+1})$, carry out Step 1.3 with x^{sc} replacing x^c there.