CSL361 Problem set 10: Steepest descent and Conjugate gradients

April 17, 2017

1 Necessary and sufficient conditions for relative minimum

1. (First-order necessary condition for a minimum) Let $\Omega \subseteq \mathbb{R}^n$ and let f defined on Ω be continuous and differentiable. If x^* is a local minimum point of f over Ω, then for any $d \in \mathbb{R}^n$ that is a feasible direction at x^*, show that $\nabla f(x^*)d \geq 0$. Also, show that if x^* is an interior point of Ω, then $\nabla f(x^*)d = 0$.

2. (Second-order necessary condition for a minimum) Let $\Omega \subseteq \mathbb{R}^n$ and let f defined on Ω be twice continuously differentiable. If x^* is a local minimum point of f over Ω, then for any $d \in \mathbb{R}^n$ that is a feasible direction at x^*, show that

   (a) $\nabla f(x^*)d \geq 0$.

   (b) if $\nabla f(x^*)d = 0$, then $d^T \nabla^2 f(x^*)d = d^T H(f(x^*))d \geq 0$.

   Also, show that if x^* is an interior point of Ω, then $\nabla f(x^*)d = 0$, and $\forall d$, $d^T \nabla^2 f(x^*)d \geq 0$.

3. (Second-order sufficient condition) Let f be twice continuously differentiable in a region in which x^* is an interior point. Suppose also that

   (a) $\nabla f(x^*)d = 0$.

   (b) $d^T \nabla^2 f(x^*)d = d^T H(f(x^*))d \geq 0$.

   Then x^* is a local minimum of f.

2 Convex sets and convex functions

1. Let f_1 and f_2 be convex functions defined on a convex set Ω. Show that

   (a) $f_1 + f_2$ is convex on Ω

   (b) af_1 is convex on Ω for any $a \geq 0$.

2. Let f be a convex function defined on a convex set Ω. Show that the set $\Gamma_c = \{x : x \in \Omega, f(x) \leq x\}$ is convex for every real number c.

3. Suppose f is continuous and differentiable. Show that f is convex on a convex set Ω *if and only if* $f(y) \geq f(x) + \nabla f(x)(y - x)$, $\forall x, y \in \Omega$.

4. Let f be twice continuously differentiable. Then show that f is a convex function over a convex set Ω containing an interior point *if and only if* the Hessian matrix of f is positive semidefinite throughout Ω.

5. Let f be a convex function defined on a convex set Ω. Show that the set Γ where f attains its minimum is convex, and any local minimum of f is a global minimum.

6. Let f be a convex function defined on a convex set Ω. If there is a point $x^* \in \Omega$, such that $\forall y \in \Omega$, $\nabla f(x^*)(y - x^*) \geq 0$, then show that x^* is a global minimum of f over Ω,

7. Let f be a convex function defined on a bounded, closed, convex set Ω. Show that if f has a maximum in Ω, it is achieved at an extreme point of Ω.

3 Steepest descent

1. Suppose f is a continuous and differentiable convex function defined over a convex set Ω. If $r(x) = -\nabla f(x)$, then show that the steepest descent iteration defined by $x_{k+1} = x_k + \alpha_k r(x_k)$, where α_k is a non-negative scalar chosen to minimize $f(x_k + \alpha_k r(x_k))$, is guaranteed to find a global minimum of f over Ω if $x_0 \in \Omega$.

2. Consider the quadratic form $f(x) = \frac{1}{2}x^T Ax - b^T x + c$. Show that

   (a) $f'(x) = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$

   (b) If A is symmetric $f'(x) = Ax - b$.

(c) $\frac{1}{2}(A + A^T)$ is always symmetric.

(d) If A is positive definite and symmetric, then the solution of the linear system $x^* = A^{-1}b$ gives a global minimum of f.

3. If A is postive definite and symmetric), then show that for the gradient descent procedure for the quadratic form

(a) $r(x) = f'(x) = b - Ax$

(b) x^* that minimizes f, also minimizes $E(x) = \frac{1}{2}(x - x^*)^T A(x - x^*)$ and vice-versa.

(c) $\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}$

(d) The iterative process satisfies

$$E(x_{k+1}) = \left(1 - \frac{(r_k^T r_k)^2}{(r_k^T A^{-1} r_k)(r_k^T A r_k)}\right) E(x_k)$$

(e) If λ_1 and λ_n are the largest and smallest eigenvalues of A, then, for any vector x

$$\frac{(x^T x)^2}{(x^T A^{-1} x)(x^T A x))} \geq \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}$$

(f) Conclude that for all steps k

$$E(x_{k+1}) \leq \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 E(x_k)$$

(g) Conclude that if $\lambda_1 = \lambda_n$ then steepest descent converges in one step. But, on the other hand, if $\lambda_1 >> \lambda_n$, then the convergence can be extremely slow. Try to visualize the two situations geometrically. (It may be worth the effort to program the two dimensional case in Matlab, and visualize the iterations along with contour plots of A for different A's).

(h) Show that the successive search direction r_{i+1} and r_i are always orthogonal.

(i) Defining $e_i = x_i - x*$, show that if e_i is an eigenvector of A, then so is $r_i = -Ae_i = -\lambda e_i$, and the iteration converges in one step.

(j) For a more general case, express e_i as a linear combination of eigenvectors of A and relate the steepest descent iteration with *power iteration* and *Rayleigh quotient*.

(k) Show that the steepest descent recurrence for the quadratic case can be given as

$$\begin{aligned} r_0 &= b - Ax_0 \\ \alpha_k &= \frac{r_k^T r_k}{r_k^T A r_k} \\ r_{k+1} &= r_k - \alpha_k A r_k \end{aligned}$$

Argue that for sparse matrices the above iteration can be implemented easily if there is an efficient procedure for matrix-vector multiplication.

4. Suppose f is defined on $\mathbb{R}^n$, has continuous second partial derivatives, and has a local minimum at x^*. Also assume that the Hessian matrix of f at x^* has smallest eigenvalue λ_n and largest eigenvalue λ_1. If $\{x_k\}$ is a sequence generated by steepest descent and it converges to x^*, then show that the sequence of objective values $\{f(x_k)\}$ converges to $f(x^*)$ linearly with a convergence rate no greater than $\left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2$.

4 Conjugate directions

Consider the quadratic, positive definite case. Suppose $d_0, d_1, \ldots, d_{n-1}$ are a set of n directions that are A-orthogonal, or *conjugate*, such that $d_i^T A d_j = 0$. Show the following:

1. If the iteration is $x_{i+1} = x_i + \alpha_i d_i$, then the value of α_i which satisfies the requirement $d_i^T A e_{i+1} = 0$ (the error e_{i+1} has no component aligned to d_i) is
$$\alpha_i = -\frac{d_i^T A e_i}{d_i^T A d_i} = \frac{d_i^T r_i}{d_i^T A d_i}$$

2. The orthogonality requirement $d_i^T A e_{i+1} = 0$ is equivalent to finding the minimum point along the direction d_i; and to the condition $-r_{i+1}^T d_i = 0$.

3. If $e_0 = \sum_{j=0}^{n-1} \delta_j d_j$, then $\delta_k = -\alpha_k$. Thereby conclude that the *conjugate directions iteration* computes the solution x^* in exactly n steps.

4. Also, the error after i steps, $e_i = \sum_{j=i}^{n-1} \delta_j d_j$

5. If we start with n linearly independent vectors $u_0, u_1, \ldots, u_{n-1}$ vectors, then the following *conjugate Gram-Schmidt process* can generate the required *conjugate vectors* $d_0, d_1, \ldots, d_{n-1}$

$$\begin{aligned} d_0 &= u_0 \\ d_i &= u_i + \sum_{k=0}^{i-1} \beta_{ik} d_k, \ i > k \\ \beta_{ij} &= -\frac{u_i^T A d_j}{d_i^T A d_j}, \ i > j \end{aligned}$$

6. Convince yourself that in the above procedure all the old search vectors must be kept in memory to construct each new one, and $O(n^3)$ operations are required to generate the whole set.

7. If we define $\mathcal{D}_i = \text{span}\{d_0, d_1, \ldots, d_{i-1}\}$, then show that the residual r_i is orthogonal to $\mathcal{D}_i$, i.e., $d_i^T r_j = u_i^T r_j = 0$ for $i < j$.

8. Also, show that $d_i^T r_i = u_i^T r_i$.

9. Show that $r_{i+1} = r_i - \alpha_i A d_i$.

5 Conjugate gradients

If in the *conjugate directions* procedure, we make the setting $u_i = r_i$, then show the following:

1. $r_i^T r_j = 0$ for $i \neq j$; and

$$\begin{aligned} \mathcal{D}_i &= \text{span}\{r_0, r_1, \ldots, r_{i-1}\} \\ &= \text{span}\{d_0, A d_0, A^2 d_0, \ldots, d_0\} \\ &= \text{span}\{r_0, A r_0, A^2 r_0, \ldots, r_0\} \end{aligned}$$

2. Convince yourself that because $d_{i-1} \in \mathcal{D}_i$, each new subspace $\mathcal{D}_{i+1}$ is formed from the union of the previous subspace $\mathcal{D}_i$ and $A\mathcal{D}_i$.

3. Also convince yourself that r_{i+1} orthogonal to $\mathcal{D}_{i+1}$ ($d_i^T r_j = u_i^T r_j = 0$ for $i < j$) implies that r_{i+1} is A-othogonal to $\mathcal{D}_i$. Gram-Schimdt conjugation becomes easy, because r_{i+1} is already A-othogonal to all the previous search directions except d_i.

4. Use the expressions $\beta_{ij} = -\frac{r_i^T A d_j}{d_i^T A d_j}$ and $r_{j+1} = r_j - \alpha_j A d_j$ to show that

$$r_i^T A d_j = \begin{cases} \frac{1}{\alpha_i} r_i^T r_i & \text{if } i = j \\ -\frac{1}{\alpha_{i-1}} r_i^T r_i & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases}$$

and, hence

$$\beta_{ij} = \begin{cases} \frac{1}{\alpha_{i-1}} \frac{r_i^T r_i}{d_{i-1}^T A d_{i-1}} & \text{if } i = j+1 \\ 0 & \text{if } i > j+1 \end{cases}$$

5. Substituting $\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}$ and $d_i^T r_i = u_i^T r_i$, conclude that

$$\beta_{i,i-1} = \frac{r_i^T r_i}{r_{i-1}^T r_{i-1}}$$

6. Combining everything, conclude that the *conjugate gradient* procedure is

$$\begin{aligned} d_0 = r_0 &= b - Ax_0 \\ \alpha_i &= \frac{r_i^T r_i}{d_i^T A d_i} \\ x_{i+1} &= x_i + \alpha_i d_i \\ r_{i+1} &= r_i - \alpha_i A d_i \\ \beta_{i+1} &= \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} \\ d_{i+1} &= r_{i+1} + \beta_{i+1} d_i \end{aligned}$$

7. Argue that the time compexity per-iteration is reduced from $O(n^2)$ to $O(m)$ where m is the number of non-zero entries in A.

6 Newton and Levenberg-Marquardt algorithms

1. Contrast the Newton and Levenberg-Marquardt algorithms discussed in the class with Steepest Descent and Conjugate Gradients.