Problem 1

最小化$f(x)$等价于求$f'(x)=0$的根，所以

(a)向前误差：

$$|x^* - x_{test}|$$

(b)向后误差：

$$|f'(x^*) - f'(x_{test})|$$

(c)条件数：

$$\begin{aligned}\frac{|x^* - x_{test}|}{|f'(x^*) - f'(x_{test})|} &\approx \frac{|x^* - x_{test}|}{|f''(x_{test})(x^* - x_{test})|} \\ &= \frac{1}{|f''(x_{test})|}\end{aligned}$$

Problem 2

(a)ϵ相当于相对误差，相比于$(x\Diamond y)+\epsilon$的形式，$(1+\epsilon)(x\Diamond y)$可以更方便的比较不同测量值的误差大小。

(b)证明：存在性是显然的，所以我们只需证明

$$0 \leq |\epsilon| < \epsilon_{\max}$$

左边的不等号显然，只需考虑右边的不等号，利用反证法，假设

$$|\epsilon| \geq \epsilon_{\max}$$

如果$\epsilon \geq \epsilon_{\max}$，那么

$$\prod_{i=1}^{k}(1+\epsilon_i) = (1+\epsilon)^k \geq (1+\epsilon_{\max})^k > 1$$

所以必存在ϵ_j，使得

$$\begin{aligned}1+\epsilon_j &\geq 1+\epsilon_{\max} \\ \epsilon_j &\geq \epsilon_{\max}\end{aligned}$$

这就产生了矛盾，因此$\epsilon \geq \epsilon_{\max}$不可能发生。

如果$\epsilon \leq -\epsilon_{\max}$，那么

$$\prod_{i=1}^{k}(1+\epsilon_i) = (1+\epsilon)^k \leq (1-\epsilon_{\max})^k < 1$$

所以必存在ϵ_j，使得

$$
\begin{aligned}
1+\epsilon_j &\le 1-\epsilon_{\max} \\
\epsilon_j &\le -\epsilon_{\max}
\end{aligned}
$$

这就产生了矛盾，因此$\epsilon \le -\epsilon_{\max}$不可能发生。

所以

$$
|\epsilon| < \epsilon_{\max}
$$

(c)这里要注意一点，我们的运算也会产生误差，所以

(i)

$$
\begin{aligned}
\overline{\bar{x}.\bar{y}} &= (1+\epsilon_1)(1+\epsilon_2)xy.(1+\epsilon_3) \\
&= (1+\epsilon)^3 xy \\
&= (1+3\epsilon+O(\epsilon^2))xy
\end{aligned}
$$

其中$0 \le |\epsilon| < \epsilon_{\max}$，第二个等号是因为(b)。由上式可得，我们的误差上界为

$$
3\epsilon_{\max} + O(\epsilon_{\max}^2)
$$

(ii)

$$
\begin{aligned}
\frac{\bar{x}}{\bar{y}} &= \frac{1+\epsilon_1}{1+\epsilon_2}\frac{x}{y}.(1+\epsilon_3) \\
&= \frac{(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)}{(1+\epsilon_2)(1+\epsilon_2)}\frac{x}{y} \\
&= \frac{(1+\epsilon)^3}{(1+\epsilon_2)^2}\frac{x}{y} \\
&\le (1+\epsilon)^3\frac{x}{y} \\
&= (1+3\epsilon+O(\epsilon^2))\frac{x}{y}
\end{aligned}
$$

其中$0 \le |\epsilon| < \epsilon_{\max}$，第三个等号是因为(b)。有上式可得，我们的误差上界为

$$
3\epsilon_{\max} + O(\epsilon_{\max}^2)
$$

(d)注意到

$$
\begin{aligned}
\overline{\bar{x}-\bar{y}} &= \big((1+\epsilon_1)x-(1+\epsilon_2)y\big).(1+\epsilon_3) \\
&= (1+\epsilon_3)(x+\epsilon_1 x - y - \epsilon_2 y) \\
&= x-y+(\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3)x-(\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3)y
\end{aligned}
$$

计算相对误差可得：

$$
\begin{aligned}
\frac{|\overline{\bar{x}-\bar{y}}-(x-y)|}{|x-y|} &= \frac{|(\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3)x-(\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3)y|}{|x-y|} \\
&= |\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3-\frac{(\epsilon_1-\epsilon_2+\epsilon_1\epsilon_3-\epsilon_2\epsilon_3)y}{x-y}|
\end{aligned}
$$

如果x, y非常接近，那么不难看出上式趋于无穷大，因此减法的相对误差无法估计。

(e)考虑带误差的递推式：

$$
\begin{aligned}
\bar{s}_k &= (\bar{s}_{k-1} + (1+\epsilon_0)x).(1+\epsilon_{k-1}) \\
&= \bar{s}_{k-1}(1+\epsilon_{k-1}) + x(1+\epsilon_0)(1+\epsilon_{k-1}) \\
&= \Big(\bar{s}_{k-2}(1+\epsilon_{k-2}) + x(1+\epsilon_0)(1+\epsilon_{k-2})\Big)(1+\epsilon_{k-1}) + x(1+\epsilon_0)(1+\epsilon_{k-1}) \\
&= \bar{s}_{k-2}(1+\epsilon_{k-2})(1+\epsilon_{k-1}) + x(1+\epsilon_0)\Big((1+\epsilon_{k-2})(1+\epsilon_{k-1}) + (1+\epsilon_{k-1})\Big) \\
&= \ldots \\
&= \bar{s}_1 \prod_{i=1}^{k-1}(1+\epsilon_i) + x(1+\epsilon_0)\left(\sum_{j=1}^{k-1}\prod_{i=j}^{k-1}(1+\epsilon_j)\right) \\
&= x(1+\epsilon)^{k-1} + x(1+\epsilon_0)\left(\sum_{j=1}^{k-1}(1+\epsilon_j')^{k-j}\right) \\
&= x\big(1+(k-1)\epsilon\big) + x(1+\epsilon_0)\left(\sum_{j=1}^{k-1}(1+(k-j)\epsilon_j')\right) + O(\epsilon_{\max}^2) \\
&= x + (k-1)\epsilon x + x(1+\epsilon_0)\left(k-1+\sum_{j=1}^{k-1}(k-j)\epsilon_j'\right) + O(\epsilon_{\max}^2) \\
&= x + (k-1)\epsilon x + (k-1)x + \left((k-1)\epsilon_0 + \sum_{j=1}^{k-1}(k-j)\epsilon_j'\right)x + O(\epsilon_{\max}^2) \\
&= kx + \left((k-1)\epsilon + (k-1)\epsilon_0 + \sum_{j=1}^{k-1}(k-j)\epsilon_j'\right)x + O(\epsilon_{\max}^2)
\end{aligned}
$$

对大括号的式子进行估计：

$$
\begin{aligned}
|(k-1)\epsilon + (k-1)\epsilon_0 + \sum_{j=1}^{k-1}(k-j)\epsilon_j'| &\le |\epsilon_{\max}||2k-2+\frac{(k-1)k}{2}| \\
&= |\epsilon_{\max}||(k-1)\frac{k+4}{2}|
\end{aligned}
$$

计算相对误差可得

$$
\begin{aligned}
|\frac{\bar{s_k}-s_k}{s_k}| &= \left|\frac{\left((k-1)\epsilon + (k-1)\epsilon_0 + \sum_{j=1}^{k-1}(k-j)\epsilon_j'\right)x + O(\epsilon_{\max}^2)}{kx}\right| \\
&\le |\epsilon_{\max}||\frac{(k+4)(k-1)}{2k}| + O(\epsilon_{\max}^2) \\
&= \frac{k}{2}|\epsilon_{\max}| + O(\epsilon_{\max}^2)
\end{aligned}
$$

(f)考虑带误差的递推式：

$$
\begin{aligned}
\bar{q}_k &= (\bar{q}_{k-1} + \bar{q}_{k-1}).(1+\epsilon_{k-1}) \\
&= 2\bar{q}_{k-1}(1+\epsilon_{k-1}) \\
&= 2^2\bar{q}_{k-2}(1+\epsilon_{k-1})(1+\epsilon_{k-2}) \\
&= 2^k\bar{q}_0\prod_{i=0}^{k-1}(1+\epsilon_i) \\
&= 2^k x(1+\epsilon)^k \\
&= q_k(1+k\epsilon+O(\epsilon^2))
\end{aligned}
$$

因为

$$
|k\epsilon| = |\epsilon|.\log_2 n \le |\epsilon_{\max}|.\log_2 n
$$

所以相对误差的上界约等于

$$
|\epsilon_{\max}|.\log_2 n
$$

Kahan求和的方法比较复杂，可以参考计算机程序设计艺术（第2卷）中文版第235页和598页，这里只给出结果：

$$
\begin{aligned}
\hat{S}_n &= \sum_{i=1}^{n}(1+\mu_i)x_i \\
|\mu_i| &\le 2u + O(nu^2)
\end{aligned}
$$

这个结果说明Kahan求和产生的误差和计算次数无关。

Problem 3

(a)假设$A, B \in \mathbb{R}^{n\times n}$是上三角矩阵，即

$$
当 i > j 时，a_{ij} = b_{ij} = 0
$$

记$C = AB$，考虑$b_{ij}(i > j)$

$$
\begin{aligned}
b_{ij} &= \sum_{k=1}^{n} a_{ik}b_{kj} \\
&= \sum_{k=1}^{j} a_{ik}b_{kj} + \sum_{k=j+1}^{n} a_{ik}b_{kj}
\end{aligned}
$$

对前一项来说，因为$i > j \ge k$，所以$a_{ik} = 0$；对于后一项来说，$k > j$，所以$b_{kj} = 0$，因此对于$i > j$，我们有

$$
c_{ij} = 0
$$

这说明$C = AB$是上三角矩阵。

(b)直接计算特征多项式即可：

$$\begin{vmatrix} \lambda - u_{11} & \dots & \dots & \dots \\ 0 & \lambda - u_{22} & \dots & \dots \\ 0 & 0 & \dots & \dots \\ 0 & 0 & 0 & \lambda - u_{nn} \end{vmatrix} = \prod_{i=1}^{n} (\lambda - u_{ii})$$

所以上三角阵的特征值为其对角元。

下面证明$\{\vec{v}_1, \dots, \vec{v}_k\}$线性无关，假设

$$\sum_{i=1}^{k} \alpha_i \vec{v}_i = 0$$

两边左乘U^m可得

$$\sum_{i=1}^{k} \alpha_i U^m \vec{v}_i = 0$$

$$\sum_{i=1}^{k} \alpha_i u_{ii} U^{m-1} \vec{v}_i = 0$$

$$\dots$$

$$\sum_{i=1}^{k} \alpha_i u_{ii}^m \vec{v}_i = 0$$

对$m = 0, \dots, k-1$，将这些等式写成矩阵形式可得：

$$(\alpha_1 \vec{v}_1, \dots, \alpha_k \vec{v}_k) \begin{pmatrix} 1 & u_{11} & \dots & u_{11}^{k-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & u_{kk} & \dots & u_{kk}^{k-1} \end{pmatrix} = (0, \dots, 0)$$

记

$$A = \begin{pmatrix} 1 & u_{11} & \dots & u_{11}^{k-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & u_{kk} & \dots & u_{kk}^{k-1} \end{pmatrix}$$

A的行列式为范德蒙行列式，因为u_{ii}互不相同，所以$|A| \neq 0$，从而A可逆，因此

$$(\alpha_1 \vec{v}_1, \dots, \alpha_k \vec{v}_k) = (0, \dots, 0)$$

所以

$$\alpha_i \vec{v}_i = 0$$

因为$\vec{v}_i \neq 0$，所以$\alpha_i = 0$，从而$\{\vec{v}_1, \dots, \vec{v}_k\}$线性无关。

(c)假设$A \in \mathbb{R}^{n \times n}$是下三角矩阵，即

$$\text{当} i < j \text{时，} a_{ij} = 0$$

假设$B \in \mathbb{R}^{n\times n}$是$A$的逆，即

$$AB = BA = I_n$$

下面证明B是下三角矩阵，首先由A可逆，我们知道A的对角元$a_{ii} \neq 0, i = 1, \ldots, n$，接着考虑$AB$第$i$行$i+1$列到第$n$列的元素，由$AB = I_n$可知，这些元素都为0。

首先考虑AB的第1行

$$\sum_{k=1}^{n} a_{1k} b_{kj} = a_{11} b_{1j} = 0$$
$$j = 2, \ldots, n$$

所以

$$b_{1j} = 0, j = 2, \ldots, n$$

接下来用数学归纳法证明$b_{ij} = 0, j = i+1, \ldots, n$，我们对$i$做数学归纳法，基本情形$i = 1$已证明，假设$i = k$时结论成立，我们来推出$i = k+1$时结论也成立。

考虑$AB = I_n$的第$k+1$行第j个元素，其中$j \geq k+2$，显然该元素为0，所以

$$\sum_{s=1}^{n} a_{k+1,s} b_{s,j} = \sum_{s=1}^{k} a_{k+1,s} b_{s,j} + a_{k+1,k+1} b_{k+1,j} + \sum_{s=k+2}^{n} a_{k+1,s} b_{s,j} = 0$$

因为$j \geq k+2$，所以由归纳假设

$$b_{s,j} = 0, s = 1, \ldots k$$

其次当$s \geq k+2$时，由下三角矩阵的定义可知

$$a_{k+1,s} = 0$$

因此

$$\sum_{s=1}^{n} a_{k+1,s} b_{s,j} = a_{k+1,k+1} b_{k+1,j} = 0$$

因为$a_{k+1,k+1} \neq 0$，所以

$$b_{k+1,j} = 0, j \geq k+2$$

因此$i = k+1$时结论也成立。综上

$$b_{ij} = 0, j > i$$

所以B是下三角矩阵。