Solution to Homework #1

1. (#2.1) For $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ we have

$$\nabla f(x) = \begin{pmatrix} 400(x_1^3 - x_1 x_2) + 2(x_1 - 1) \\ 200(x_2 - x_1^2) \end{pmatrix},$$

and

$$\nabla^2 f(x) = \begin{pmatrix} 400(3x_1^2 - x_2) + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}.$$

To be any minimizer of this function a point must satisfy $\nabla f(x) = 0$ and this only occurs when $400(x_1^3 - x_1 x_2) + 2(x_1 - 1) = 0$ and $200(x_2 - x_1^2) = 0$. The second equality holds only when $x_2 = x_1^2$. Plugging this result into the first equality, we get that $x_1 = 1$ Thus the only possible minimizer (or maximizer) is $(1, 1)$. Thus if (1,1) is a minimizer, it is the only local minimizer. At $x = (1, 1)$ we have

$$\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix},$$

and so for any vector $v = (v_1, v_2) \neq (0, 0)$, we have

$$v^T \nabla^2 f(1, 1) v = 802v_1^2 - 800v_1 v_2 + 200v_2^2 = 200(v_2 - 2v_1)^2 + 2v_1^2 > 0.$$

Thus $\nabla^2 f(1, 1)$ is positive definite so $(1, 1)$ is a minimizer, and thus the only local minimizer. Equivalently, we could find that the eigenvalues of $\nabla^2 f(1, 1)$ are ≈ 0.4 and ≈ 1000, both positive and thus it is positive definite.

2. (#2.3) For $f_1(x) = a^T x = \sum_{i=1}^n a_i x_i$:

$$\frac{\partial f_1}{\partial x_i} = a_i \text{ and } \frac{\partial^2 f_1}{\partial x_i \partial x_j} = 0.$$

Thus $\nabla f_1(x) = a$ and $\nabla^2 f_1(x) = 0$.

For $f_2(x) = x^T A x = \sum_{i,j=1}^n a_{ij} x_i x_j$:

$$\frac{\partial f_2}{\partial x_i} = a_{ij} x_j + a_{ji} x_j = 2a_{ij} x_j \text{ and } \frac{\partial^2 f_2}{\partial x_i \partial x_j} = 2a_{ij},$$

(since A is symmetric). Thus $\nabla f_2(x) = 2Ax$ and $\nabla^2 f_2(x) = 2A$.

3. (#2.8) For $f(x) = (x_1 + x_2^2)^2$, we have $\nabla f(x) = [2(x_1 + x_2^2), 4x_2(x_1 + x_2^2)]^T$. At $x = [1, 0]^T$, $\nabla f(x) = [2, 0]$ and since $p^T \nabla f(x) = -2 < 0$, $p = [-1, 1]^T$ is a descent direction. Why? Because to be a descent direction p must satisfy $f(x+tp) < f(x)$ for $0 < t < t_0$ for some $t_0 > 0$. Thus if f is smooth $\lim_{t \to 0} \frac{1}{t}(f(x+tp) - f(x)) \leq 0$. But $\lim_{t \to 0} \frac{1}{t}(f(x+tp) - f(x)) = \nabla f(x)^T p$.

Let $g(\alpha) = f(x + \alpha p)$. We wish to find all the (local) minimizers of g for $\alpha > 0$. To be a minimizer, we must have $g'(\alpha) = 0$ or $\nabla f(x + \alpha p)^T p = 0$. With $x = [1, 0]^T$ and $p = [-1, 1]^T$ this reduces to

$$\begin{aligned} 0 &= [2(1 - \alpha + \alpha^2), 4\alpha(1 - \alpha + \alpha^2)] \cdot [-1, 1] \\ &= -2(1 - \alpha + \alpha^2) + 4\alpha(1 - \alpha + \alpha^2) \\ &= (4\alpha - 2)(1 - \alpha + \alpha^2). \end{aligned}$$

Thus the only possible minimizer occur when $\alpha = \frac{1}{2}$. Since p is a descent direction and clearly $f(x + \alpha p)$ is large for large α, this must be the only minimizer of $g(\alpha)$.

4. (#2.14) To judge convergence rates we look at the ratio

$$\left|\frac{x_{k+1}}{x_k}\right| = \frac{\frac{1}{(k+1)!}}{\frac{1}{k!}} = \frac{1}{k+1}.$$

This clearly goes to 0 as $k \to \infty$, so we at least have Q-superlinear convergence. For quadratic convergence, we look at

$$\left|\frac{x_{k+1}}{x_k^2}\right| = \frac{\frac{1}{(k+1)!}}{\frac{1}{(k!)^2}} = \frac{k!}{k+1}.$$

In this case, this ratio is unbounded as $k \to \infty$, so we do not have Q-quadratic convergence. Thus the convergence is superlinear, but not quadratic.

5. First some basics. The eigenpairs of A are $(1, [1,1]^T)$ and $(3, [1,-1]^T)$.

   (a) Since A has all positive eigenvalues, it is positive definite (and obviously symmetric), thus the minimizer occurs at $\hat{x} = A^{-1}b = [1,-1]^T$.

   (b) The graph of ϕ is a concave bowl with the center at $x = \hat{x}$, thus the set of x such that $\phi(x) = 0$ should be a circle-like object. Computing $\phi(\hat{x}+z) = \phi(\hat{x}) + \frac{1}{2}z^TAz$ we get for $\phi(x) = 0$, $\frac{1}{2}z^TAz = 3$ or, multiplying it out, $z_1^2 - z_1z_2 + z_2^2 = 3$. We can rewrite this as $\frac{1}{4}(z_1+z_2)^2 + \frac{3}{4}(z_1-z_2)^2 = 3$ which is the equation of an ellipse (rotated 45° from horizontal). So the set of all x's with $\phi(x) = 1$ is a rotated ellipse centered at $\hat{x}$.

   (c) Take $c = \hat{x} = [1,-1]^T$ and then there are 8 choices for T. The basic form is $T = \sqrt{2}V\sqrt{D}^{-1}$ where V is the matrix constructed from the eigenvectors (choice of order), normalized so that $V^TV = I$ and $\sqrt{D}$ is a diagonal matrix with square roots of the eigenvalues (choice of sign on each one). One example is

$$T = \sqrt{2}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{\sqrt{3}} \\ 1 & -\frac{1}{\sqrt{3}} \end{bmatrix}.$$

   With such a T and c, we have $\phi(x) = w_1^2 + w_2^2 - 3$ when $x = Tw + c$. Let $z = Tw$ then for $x = Tw + c = z + \hat{x}$, we have $\phi(x) = \phi(\hat{x}) + \frac{1}{2}z^TAz$. $\phi(\hat{x}) = -3$ and, using $A = VDV^T$, we have

$$\begin{aligned} \frac{1}{2}z^TAz &= \frac{1}{2}w^TT^TVDV^TTw \\ &= w^T(\sqrt{D}^{-T}V^TVDV^TV\sqrt{D}^{-1})w \\ &= w^T(\sqrt{D}^{-1}D\sqrt{D})w \\ &= w^Tw = w_1^2 + w_2^2. \end{aligned}$$

6. The exact answer is $W(3) = 0$ and from the expansion we see that we have to work with some large numbers. So to get a result of 0 there must be lots of cancellation. This is a trouble spot in floating point arithmetic and thus we do not expect to compute $W(3)$ as 0.

   To get an estimate, I looked at each term (with $x = 3$) and figured that the 6th term ($-46710x^5$) is the largest giving a value of -11350530. Since this and all the other terms will be rounded to 4 significant digits, I'll lose 4 significant digits thus at first glance I'd expect an answer at best in the range $-5000 < W(3) < 5000$. If the roundings are balanced, i.e. some

round down and some round up, we could do better. If the roundings are all round downs (or ups), then we could do worse.

Another approach is to use the round-off error estimates. This gets pretty complex but the key to understanding can be found in just a part of the process: to calculate $a-b$ where a and b are both positive real numbers. Using our knowledge of rounding we'd expect the answer to be computed as $z = (a(1+\delta_1) - b(1+\delta_2))(1+\delta_3)$ where $|\delta_i| \leq \epsilon$. Then

$$|(a-b) - z| = |a\delta_1 - b\delta_2 + \delta_3(a-b)| \leq \epsilon(|a| + |b| + |a-b|).$$

Note if a and b are similar in size then $|a-b|$ will be small, but $a+b$ could be quite large. From this result, we can argue that the computed value of $W(3)$ could have an error roughly the size of $\epsilon = 5 \times 10^{-4}$ times $|W|(3)$, where $|W|$ means taking the absolute value of each term before adding, so that $|W|(3) \approx 4.5 \times 10^7$. Thus we might expect an error of the size of about 22000. With this sort of worse case analysis, the error estimate is often much larger than the actual error as it is based on all the rounding errors being in the worse possible direction, while in practice they don't.

Finally, I wrote a program to actually compute the value, doing all the rounding. I got an answer of 3640. See the graph of the computed $W(x)$ compared to the original form below.