Homework 3 solution

1. (Exercises from Beck Ch 2)

   a. (Beck 2.17) Here $f(x_1, x_2) = 2x_2^3 - 6x_2^2 + 3x_1^2 x_2$,

      i. The gradient is

      $$\nabla f(x) = \begin{bmatrix} 6x_1 x_2 \\ 6x_2^2 - 12x_2 + 3x_1^2 \end{bmatrix},$$

      There are two stationary points: $x^* = (0,0)$ and $x^* = (0,2)$.

      ii. The Hessian is

      $$\nabla^2 f(x) = \begin{bmatrix} 6x_2 & 6x_1 \\ 6x_1 & 12x_2 - 12 \end{bmatrix}$$

      At $x^* = (0,0)$,

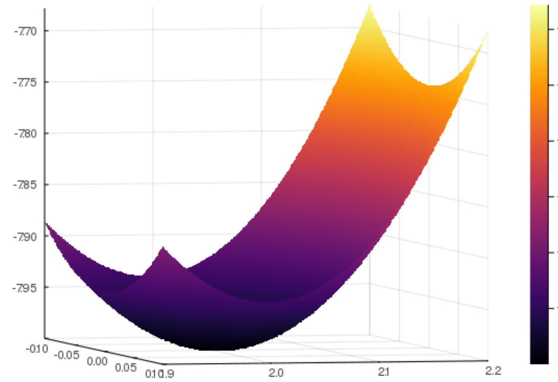
      $$\nabla^2 f([\,0\;0\,]) = [\,0 \quad 0\;0 \quad -12\,]$$

      which is negative semidefinite. (Diagonal with all non-positive values, and one 0.) We cannot tell if it is a local minimum or maximum or saddle point. At $x^* = (0,2)$,

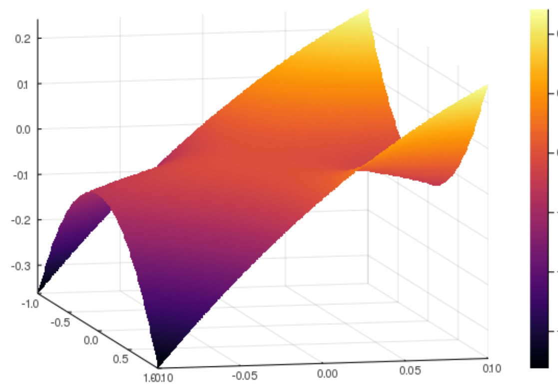
      $$\nabla^2 f([\,0\;2\,]) = [\,12 \quad 0\;0 \quad 12\,]$$

      which is positive definite. (Diagonal with all strictly positive values). This point must be a strict local minima.

      iii. At $x = (0,2)$, it is clearly a strict local minimum.



At $x = (0, 0)$, it is a saddle point.



b. (Beck 2.19) We first show that forward implication. Since $\nabla^2 f(x) = A \succeq 0$ for all x, if we find any point in which $\nabla f(x) = 0$, then we have found a global minimizer of this function. If $b \in \mathbf{range}(A)$, then there exists a y where $Ay = b$. Taking $x^* = -y$ gives the stationary point we need.

Now assume that $b \notin \mathbf{range}(A)$. Then, this means that $b = u + v$ where $u \in \mathbf{range}(A)$ and $v \in \mathbf{null}(A^T) = \mathbf{null}(A)$ where $v \neq 0$. Now take any $x = \gamma v$ for any scalar c. Then

$$f(\gamma v) = \frac{\gamma^2}{2} \underbrace{v^T A v}_{=0} + \gamma \underbrace{b^T v}_{=v^T v} + c = \gamma \|v\|_2^2 + c.$$

Picking $\gamma \to -\infty$ shows that $f(\gamma v) \to -\infty$ is unbounded below.

2. To compute $\mathbf{tr}(A^T B)$, we must first form the matrix product $A^T B$ which requires $O(n^2 m)$ flops and $O(n^2)$ storage. Then extracting the trace is an

additional $O(n)$ flops and $O(1)$ storage. So, in total, $O(n^2m+n)$ flops (or $O(n^2m)$ as the dominating term) and $O(n^2+1)$ storage (or just $O(n^2)$).

To compute the right and side, we do not need any additional storage, and just require $O(mn)$ flops.

Now if $m \gg n$, this is a significant reduction in storage, and if n is large is a significant reduction in flops. The key takeaway is that, for proper scalability, though many things are equivalent, how you implement it matters.

3. Here, $f:\mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function that has L-Lipschitz gradient.

   a. The directional derivative of ∇f at x in the direction v is

$$\nabla^2 f(x)v = \lim_{t\searrow 0} \frac{\nabla f(x+tv)-\nabla f(x)}{t}. \tag{1}$$

   So,

$$\begin{aligned}
\|\nabla^2 f(x)v\|_2 &= \|\lim_{t\searrow 0} \frac{\nabla f(x+tv)-\nabla f(x)}{t}\|_2 &(2)\\
&= \lim_{t\searrow 0} \|\frac{\nabla f(x+tv)-\nabla f(x)}{t}\|_2 &(3)\\
&\le \lim_{t\searrow 0} \frac{L\|tv\|}{t} &(4)\\
&= L\|v\|_2 &(5)
\end{aligned}$$

   where second line follows from continuity of norms and third line follows from L-Lipschitz of gradient.

   b. From above, we have that any fixed x satisfies the inequality $\|\nabla^2 f(x)v\| \le L\|v\|_2$ for all v.

   Fix x and let (λ_+, v_+) be the maximal eigen-pair of the matrix $\nabla^2 f(x)$. So, $\|\nabla^2 f(x)v\| \le L\|v\|_2$ for all v gives $\lambda_+ \le L$. Thus, all eigenvalues of

$\nabla^2 f(x)$ is bounded from above by L. As x is arbitrary, we get that for all x, the eigenvalues of $\nabla^2 f(x)$ is bounded from above by L.

c. Using Taylor's remainder theorem, we get

$$f(v) = f(w) - \nabla f(w)^\intercal (v-w) + \frac{1}{2}(v-w)^\intercal \nabla^2 f(\xi)(v-w),$$

where $v, w \in \mathbb{R}^n$ and $\xi \in [v, w]$. Since $\|\nabla^2 f(x)v\| \leq L\|v\|_2$ for all v and x, we also have $v^\intercal \nabla^2 f(x) v \leq L\|v\|_2^2$ all v and x. Thus,

$$f(v) = f(w) + \nabla f(w)^\intercal (v-w) + \frac{L}{2}\|v-w\|_2^2.$$

d. A gradient descent step is $x_{k+1} = x_k - \alpha \nabla f(x_k)$. Substituting $v = x_{k+1}$ and $w = x_k$, we get

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \nabla f(x_k)^\intercal (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 && (6 \\ &= f(x_k) - \alpha \nabla f(x_k)^\intercal \nabla f(x_k) + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|_2^2 && (7 \\ &= f(x_k) - \alpha\|\nabla f(x_k)\|_2^2 \left(1 - \frac{L\alpha}{2}\right) && (8 \end{aligned}$$

Note that $\alpha\|\nabla f(x_k)\|_2^2(1 - \frac{L\alpha}{2}) > 0$ if x_k is not a stationary point and $0 < \alpha < \frac{2}{L}$