Problem 1

(a) From multivariable calculus and the assumption that A being symmetric, we know for $f(x) = x^TAx - 2b^Tx$, the gradient[1] is

$$\nabla f(x) = (A + A^T)x - 2b = 2Ax - 2b.$$

In addition, we know when the multivariable function reaches minimum (global minimum in this context, as the function is convex), the gradient is 0. Therefore we know at the optimum point $x*$, $\nabla f(x^*) = 0$, which leads to $2Ax^* - 2b = 0$ and lets us solve $Ax = b$.

The gradient being 0 only makes sense when the solution is feasible. In order words, we need to show the minimum of $f(x)$ can only be finite. This can be shown by contradiction. Suppose the minimum is $-\infty$, it can only occurs when $x \to -\infty$ or $x \to \infty$, because for any finite x, $f(x)$ is finite. Since A is positive semidefinite we know for any x, x^TAx is non-negative. In order words, all the eigenvalue of A is non-negative. For orthogonal eigenvectors $x_1, x_2, ..., x_n$, and eigenvalues $\gamma_1, \gamma_2, ..., \gamma_n$, and $x = c_1x_1 + c_2x_2, + ... + c_nx_n$, we have $x^TAx = \gamma_1c_1^2x_1^2 + \gamma_2c_2^2x_2^2 + ... + \gamma_nc_n^2x_n^2$. Since all γ's are non-negative, this sum is overwhelmly larger than $b_1x_1 + b_2x_2 + ... + b_nx_n$, when $x \to \pm\infty$, which contradicts with the assumption that the sum of the two can go to $-\infty$.

The only exception in the last example is that some eigenvalue can be 0 can we only extend that part of eigenvector to be ∞, then the sum of the matrix A part will not overwhelm bx. Notice that in that case $Ax = \sum_{\gamma_i=0} \gamma_ix_i = 0$ and $Ax = b$ would not be solvable except the trivial case $b = 0$. □

(b) Write out the eigenvalues of A in the descending order $\gamma_1, \gamma_2, ..., \gamma_n$, and corresponding orthogonal eigenvectors $x_1, x_2, ..., x_n$. Suppose $x = c_1x_1 + c_2x_2, + ... + c_nx_n$. Then we have

$$\frac{||Ax||_2}{||x||_2} = \frac{||\gamma_1c_1x_1 + \gamma_2c_2x_2 + ... + \gamma_nc_nx_n||_2}{||c_1x_1 + c_2x_2 + ... + c_nx_n||_2} \leq \frac{||\gamma_1c_1x_1 + \gamma_1c_2x_2 + ... + \gamma_1c_nx_n||_2}{||c_1x_1 + c_2x_2 + ... + c_nx_n||_2} = \gamma_1.$$

Notice that the above inequality can saturate when we pick $x = x_1$, i.e., the eigenvector corresponding to the largest eigenvalue. Therefore, we conclude that $(||Ax||_2/||x||_2)_{\max} = \gamma_1$.

[1] A quick way of deriving this is by observing $(x+\delta)^TA(x+\delta) - x^TAx = x^TA\delta + \delta^TAx + O(|\delta|^2)$. By dividing (elementwise in δ) it with δ and let $\delta \to 0$, we got the derivative $(A + A^T)x$. $\nabla b^Tx = b$ is straightforward from the definition.

(c) Express $r^{(t)}$ in terms of $r^{(t-1)}$ as

$$\begin{aligned} r^{(t)} &= Ax^{(t)} - b \\ &= A\left[x^{(t-1)} - \eta\nabla f(x^{(t-1)})\right] - b \\ &= r^{(t-1)} - \eta A\left[2A^T x^{(t-1)} - 2b\right] \\ &= (1 - 2\eta A)r^{(t-1)} \end{aligned}$$

Now use $\eta = 1/(2||A||_2)$ and take ℓ_2 norm

$$\begin{aligned} ||r^{(t)}||_2 &= ||(1-2\eta A)r^{(t-1)}||_2 \\ &= \|\Big| \left(1 - \frac{A}{||A||_2}\right) r^{(t-1)}\Big|\Big|_2 \\ &\le \Big|\Big| \left(1 - \frac{A}{||A||_2}\right) \Big|\Big|_2 \Big|\Big| r^{(t-1)} \Big|\Big|_2 \end{aligned}$$

To use the condition number of A, we observe

$$\begin{aligned} \Big|\Big| \left(1 - \frac{A}{||A||_2}\right) \Big|\Big|_2 &= \max_{x\neq 0} \frac{\Big|\Big| \left(1 - \frac{A}{||A||_2}\right) x \Big|\Big|_2}{||x||_2} \\ &= \max_{x\neq 0} \frac{\Big|\Big| \left(1 - \frac{A}{||A||_2}\right) (c_1x_1 + c_2x_2 + ... + c_nx_n) \Big|\Big|_2}{||c_1x_1 + c_2x_2 + ... + c_nx_n||_2} \\ &= \max_{x\neq 0} \frac{\Big|\Big| c_1x_1 + c_2x_2 + ... + c_nx_n - \frac{1}{||A||_2}(\gamma_1c_1x_1 + \gamma_2c_2x_2 + ... + \gamma_nc_nx_n) \Big|\Big|_2}{||c_1x_1 + c_2x_2 + ... + c_nx_n||_2} \\ &\le \max_{x\neq 0} \frac{\Big|\Big| c_1x_1 + c_2x_2 + ... + c_nx_n - \frac{1}{||A||_2}(\gamma_nc_1x_1 + \gamma_nc_2x_2 + ... + \gamma_nc_nx_n) \Big|\Big|_2}{||c_1x_1 + c_2x_2 + ... + c_nx_n||_2} \\ &= 1 - \frac{\gamma_n}{\gamma_1} \\ &= 1 - \frac{1}{\kappa(A)}, \end{aligned}$$

where the inequality actually can stature when picking $x = x_n$. Therefore we reach the desire conclusion. □

(d) To show $||A\tilde{x} - b|| < \epsilon$, we essentially need to find a t such that $||r^{(t)}|| < \epsilon$. Notice that from (c) we have

$$\begin{aligned} ||r^{(t)}||_2 &\le \left(1 - \frac{1}{\kappa(A)}\right) ||r^{(t-1)}||_2 \\ &\le \left(1 - \frac{1}{\kappa(A)}\right)^2 ||r^{(t-2)}||_2 \\ &\le ... \\ &\le \left(1 - \frac{1}{\kappa(A)}\right)^t ||r^{(0)}||_2 \end{aligned}$$

Now if we set $t = log_{1-\frac{1}{\kappa(A)}} \frac{\epsilon}{||r^{(0)}||_2} = log_{\frac{\kappa(A)}{\kappa(A)-1}} \frac{||r^{(0)}||_2}{\epsilon}$, we will have $||r^{(t)}||_2 \le \epsilon$. Suppose other operations are in the same order of outputting the vector, $O(n)$. Therefore the running time for a good enough solution is in the order of $O(n\, log_{\frac{\kappa(A)}{\kappa(A)-1}}(||r^{(0)}||_2/\epsilon))$.

In fact, we can pick the initial $x^{(0)} = 0$ such that $||r^{(0)}||_2 = ||b||_2$. Therefore the running time we got is $O(n\, log_{\frac{\kappa(A)}{\kappa(A)-1}}(||r^{(0)}||_2/\epsilon))$.

(e) From (d) we know for running time in the scale of $O(log(1/\epsilon))$ we can achieve $||Ax - b||_2 < \epsilon$. From (a) we know $Ax^* = b$. Then we have

$$||(Ax - b)||_2 = ||(Ax - b) - (Ax^* - b)||_2 = |(A(x - x^*)||_2 \le \epsilon$$

Since we know A is invertible, we can set ϵ to be $\epsilon/||A^{-1}||_2$, multiplying $||A^{-1}||_2$ on both end of the above inequality, we have

$$||x - x^*||_2 = ||A^{-1}A(x - x^*)||_2 \le ||A^{-1}||_2||A(x - x^*)||_2 \le ||A^{-1}||_2 \frac{\epsilon}{||A^{-1}||_2}$$

Hence, we need $O(nlog_{\frac{\kappa(A)}{\kappa(A)-1}}(||A^{-1}||_2||r^{(0)}||_2/\epsilon))$ running time to reach $||x - x^*||_2 \le \epsilon$.

(f) The intuition is that when Ax is close to b, x is close to a x^* that solves $Ax - b$. Part (a) still works as it does not involve the assumption of A being invertible. Similarly, the argument in (b) still works. For lateral parts, notice the following fact of non-invertible matrix A. A has an invertible part A^+ and a null part A^0. $Ax = b$ only include the invertible part of A, i.e., $A^+x = b$ and $A^0x = 0$. Then the whole problem above can still work when we only consider the invertible part. In part (c), the new κ only uses the minimum eigenvalue of the invertible part, which is the minimum nonzero eigenvalue of A. For part (d) and (e) the time bound analysis still apply following (c).[2]

[2] Received the hint from TA afterwards, the approach here is essentially taking the pseudoinverse A^+.

Problem 2

(a) To show the problem is of the desired form, it is sufficient to show $Ax \geq 0$ is a convex set, denoted as S. This can be shown by checking any $x, y \in S$. This means $Ax \geq 0$ and $Ay \geq 0$. Then for any $\lambda \in [0,1]$, $A(\lambda x + (1-\lambda)y) = \lambda Ax + (1-\lambda)Ay \geq 0$. This is because $\lambda \geq 0$ and $(1-\lambda) \geq 0$. Therefore $(\lambda x + (1-\lambda)y) \in S$ as well. Hence $Ax \geq 0$ is a convex set. □

(b) Denote the contraction coefficient for g and h to be w_g, w_h, respectively. Then for any $x, y \in \mathbb{R}^n$,

$$||g(h(x)) - g(h(y))||_2 \leq w_g ||h(x) - h(y)||_2 \leq w_g w_h ||x - y||_2$$

Therefore, $w_g w_h$, the product of contraction coeffiction of g and h, is suitable for all x, y. The actual contraction coefficient for $g \cdot h$ would be at most $w_g w_h$. □

(c) Noted.

(d) For x, y we write out

$$\begin{aligned} ||g(x) - g(y)||_2 &= ||(x - \eta \nabla f(x)) - (y - \eta \nabla f(y))||_2 \\ &= ||(x - y) - \eta \left(\nabla f(x) - \nabla f(y) \right) ||_2 \\ &= ||(x - y) - \eta Z \left(x - y \right) ||_2 \\ &\leq (1 - \eta \gamma_n) ||x - y||_2 \end{aligned}$$

where γ_n is the smallest eigenvalue[3] of Z. Notice that η is selected as $1/\beta$. Recall the definition of β-smoothness $||\nabla f(x) - \nabla f(y)|| \leq \beta ||x - y||$. Then we have for *any* x and y,

$$||Z(x - y)||_2 = ||\nabla f(x) - \nabla f(y)||_2 \leq \beta ||x - y||_2,$$

in order words, $||Z(x - y)||_2 / ||x - y||_2 \leq \beta$. We know β is therefore the largest eigenvalue[4] of Z.

Hence we reach

$$||g(x) - g(y)||_2 \leq \left(1 - \frac{\gamma_n}{\gamma_1}\right) ||x - y||_2 = \left(1 - \frac{1}{\kappa(Z)}\right) ||x - y||_2 \leq \left(1 - \frac{1}{\kappa(f)}\right) ||x - y||_2$$

Therefore, the contraction coefficient is at most $1 - \frac{1}{\kappa(f)}$. □

[3] The inequality is shown in 1(c).

[4] Proof is shown in 1(b).

(e) We prove this by contradiction. Suppose there are two different points y, z in the convex set having same minimum ℓ_2 distance to x. Then we know $\lambda y + (1-\lambda)z$ is also in the set S. Also,

$$\begin{aligned}||\lambda y + (1-\lambda)z - x||_2 &= ||\lambda y - \lambda x + (1-\lambda)z - (1-\lambda)x||_2 \\ &< ||\lambda y - \lambda x||_2 + ||(1-\lambda)z - (1-\lambda)x||_2 \\ &= \lambda||y-x||_2 + (1-\lambda)||z-x||_2 \\ &= ||y-x||_2\end{aligned}$$

The inequality is from the triangle inequality. Notice that vector $(y-x)$ and $(z-x)$ are in different direction (if they are in the same direction, then there can only be closest point to x on a single direction), the ℓ_2 norm of the sum two different direction vector is *straightly* smaller than sum of the ℓ_2 norm of each one of them. This reaches the contradiction with the assumption that $||y-x||_2$ is minimum for all points in the convex set S. □

(f) For any x and y, we utilize what we derived in (d) and

$$\begin{aligned}\prod_S(x) - \prod_S(y) &= \text{argmin}_z||z-x||_2 - \text{argmin}_w||w-y||_2 \\ &\leq ||w-x||_2 - ||w-y||_2 \\ &\leq ||(w-x)-(w-y)||_2 \\ &= ||x-y||_2.\end{aligned}$$

where w is taken two be the *argmin* w.r.t. y above. Hence $\prod_S$ is a contraction operation. □

(g) From part (d) we know the contraction coefficient for $g(x) = x - \eta\nabla f(x)$ is at most $1 - \frac{1}{\kappa(f)}$. From part (f) we know $\prod_S$ is a contraction operation, therefore its contraction coefficient is at most 1. From part (b) we know the composition of two contraction operation is also a contraction operation, with the contraction coefficient at most the product of each. Hence we know $\prod_S(g(x))$ is also a contraction operation, with the contraction coefficient at most $1 - \frac{1}{\kappa(f)}$.

Now, consider the optimal solution x^*. At step t, the ℓ_2 distance for $x^{(t)}$ to the optimal solution is $||x^{(t)} - x^*||_2$. Then we apply the method in part (e), then x will be transformed to $x^{(t+1)} = \prod_S(g(x^{(t)}))$. If now we can show $\prod_S(g(x^*))$ is mapped to x^*, from the contraction operation and the corresponding coefficient above, we know at each step x will be brought closer to optimal x^* by a factor of $1 - \frac{1}{\kappa(f)}$ or better.

To prove $\prod_S(g(x^*)) = x^*$, consider first the special case where the gradient at x^* is 0. In this case, $\prod_S(g(x^*)) = \prod_S(x^*)$. Since x^* is in S, we know $\prod_S(x^*)$ returns x^*.

Then consider the case where the gradient is nonzero. Specifically, from the hint, consider the hyperplane that contains x^* and is orthogonal to the gradient at x^*. We need to prove that S must lie entirely on one side (inclusive) of this hyperplane. This is because if S does not hold above, then we can take a very small step in some direction that decreases the value of f, contradicting optimality. More formally, since $f(x+\delta)-f(x) \leq [\nabla f(x)]^T\delta+\beta/2||\delta||^2$, we can choose a direction of δ in the direction that results in a *negative* $[\nabla f(x)]^T\delta$ (this is feasible because part of the hyperplane containing x is not orthogonal to $\nabla f(x)$). Then we make the δ small enough, such that the negative linear term predominant the quadratic term. Thus, we construct a way of decreasing the value of f, contradicting optimality.

(h) At each time step, we do a gradient descent and then take a projection step, results in a $1-\frac{1}{\kappa(f)}$ factor improvement to the optimal x^*. Therefore, after t steps, we will have the ℓ_2 distance to optimal solution

$$\left(1-\frac{1}{\kappa(f)}\right)^t ||x_0-x^*||_2$$

We need to set

$$t=\log_{1-\frac{1}{\kappa(f)}}\frac{\epsilon}{||x_0-x^*||_2}$$

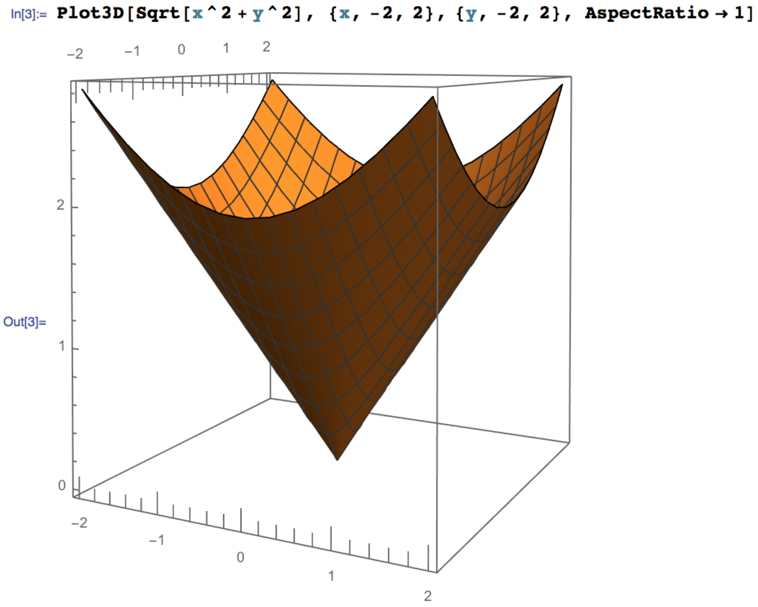
to make the ℓ_2 distance ϵ.

In each step, the projection of gradient takes at least $O(n)$ time, where n is the number of entries in the vector x. Therefore, the running time is in the order of $O\left(n\log_{\frac{\kappa(A)}{\kappa(A)-1}}(||x_0-x^*||_2/\epsilon)\right)$.

Problem 3

(a) Strongly convex says there exists $\alpha > 0$, such that for any x, y, $y^T \nabla^2 f(x) y \geq \alpha ||y||^2$. In order words, the smallest eigenvalue of $\nabla^2 f(x)$ is always larger than α. In the class, we have an equivalent condition for α-strong convexity, An equivalent condition is $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|y - x\|_2^2$. This also graphically means strongly convex function is lower bounded by a quadratic bowl $\frac{\alpha}{2}\|y - x\|_2^2$.

Therefore, we can construct a convex cone as $f(x) = ||x||$. It is convex because $||\lambda x + (1 - \lambda) y|| < \lambda ||x|| + (1 - \lambda) ||y||$ following the triangle inequality. The unique minimum is $x = 0$. $f(x)$ is not strongly convex because for any $\alpha > 0$, there is an x such the the quadratic term overwhelms the linear term, hence $f(x)$ cannot be always above any quadratic bowl.



```
In[3]:= Plot3D[Sqrt[x^2 + y^2], {x, -2, 2}, {y, -2, 2}, AspectRatio → 1]
```

(b) For symmetric matrix A and B, the minimum eigenvalue of the sum of them can be written as $\min_{x \neq 0} \frac{||(A+B)x||_2}{||x||_2}$, where the actual value can be achieved by setting

x to be the eigenvector corresponding to the minimum eigenvalue[5]. Also, we have

$$\begin{aligned}
\frac{||(A+B)x||_2}{||x||_2} &= \frac{||Ax+Bx||_2}{||x||_2} \\
&= \frac{||(c_1^A\gamma_1^A x_1^A + c_2^A\gamma_2^A x_2^A + ... + c_n^A\gamma_n^A x_n^A) + (c_1^B\gamma_1^B x_1^B + c_2^B\gamma_2^B x_2^B + ... + c_n^B\gamma_n^B x_n^B)||_2}{||x||_2} \\
&\geq \frac{||(c_1^A\gamma_n^A x_1^A + c_2^A\gamma_n^A x_2^A + ... + c_n^A\gamma_n^A x_n^A) + (c_1^B\gamma_n^B x_1^B + c_2^B\gamma_n^B x_2^B + ... + c_n^B\gamma_n^B x_n^B)||_2}{||x||_2} \\
&= \frac{||\gamma_n^A x + \gamma_n^B x||_2}{||x||_2} \\
&= \gamma_n^A + \gamma_n^B
\end{aligned}$$

where γ^A's and γ_B's are in descending order. Therefore the minimum eigenvalue of the sum of a pair of symmetric matrices is at least the sum of their minimum eigenvalues. □

(c) Suppose the component for vector x is $\xi_1, \xi_2, ..., \xi_n$, then $||x||^2 = (\xi_1^2 + \xi_2^2 + ... + \xi_n^2)$. Notice that $\partial||x||^2/\partial\xi_i\partial\xi_j = 2$ if $i = j$ and $\partial||x||^2/\partial\xi_i\partial\xi_j = 0$ otherwise. Hence the Hessian of $||x||^2$ is a diagonal matrix with entries on the diagonal 2, and 0 elsewhere. Therefore, the smallest eigenvalue of $\nabla^2\alpha||x||^2$ is 2α, which is the lower bound of strong convexity of $h(x)$.

(d) Because x^* minimizes $f(x)$, $f(x^*) \leq f(y^*)$ follows naturally. For the other part of the inequality we have $f(y^*) \leq f(y^*) + \alpha||y^*||_2^2 \leq f(x^*) + \alpha||x^*||_2^2$, as y^* minimizes $h(x) = f(x) + \alpha||x||_2^2$.

(e) To find a y^*, such that $||f(y^*) - f(x^*)|| \leq \epsilon$, from (d), we know $f(x^*) \leq f(y^*) \leq f(x^*) + \alpha||x^*||_2^2$, therefore it is sufficient to reach $||f(x^*) + \alpha||x^*||_2^2 - f(x^*)|| \leq \epsilon$, which is to have $\alpha||x^*||^2 \leq \epsilon$. Since we already know $||x^*|| \leq R$, we can set $\alpha = \epsilon/R^2$. Then it suffices to look for y^* that minimizes $h_{\alpha=\epsilon/R^2}(x)$.

From part (c) we know the $h(x)$ is bounded by the smallest eigenvalue of the Hessian, which is $2\alpha = 2\epsilon/R^2$. The largest eigenvalue for $f(x)$ is $\beta + 2\alpha$.

From problem 2, we can start $x_0 = 0$ and $||x_0 - x^*|| = ||x^*|| \leq R$, the running time is bounded by $O\left(n \log_{\frac{\kappa}{\kappa-1}}(R/\epsilon)\right)$, where $\kappa = (\beta + 2\alpha)/2\alpha$ and $2\alpha = 2\epsilon/R^2$.

[5] Full derivation is similar to Problem 1(b).

Problem 4

(a) Since the optimal distance d is known in a priori, we can build clusters iteratively. In each iteration, pick any point that has not been assigned to a cluster yet, and pick any free-to-assign point within d distance to that point, and form a new cluster. The iterations terminate when no points all assigned to some cluster.

From triangle inequality, since we know the points in a cluster are all within distance d to the center (the one we pick first in each iteration), each point is at most within $2d$ distance to any other point in the cluster. This forms a 2-approximation.

The iterating algorithm has to terminate before or on the k-th iteration. This is because we know there are k clusters, each point that is not assigned to the current new cluster must has been assigned to some other cluster (within distance d to that center). After k iterations, the points to all k clusters (in optimal case) must be assigned to some (actual) clusters. Therefore the algorithm terminates in k iterations.

(b) In each iteration, the point chosen is the *maximum* distance to the centers being chosen already. In the optimal clusters, any point in larger distance than d is in a different cluster. Therefore in the algorithm the point being selected must be from another cluster (if it exists, its distance is the largest to all centers). Since we know there are k clusters in the optimal clustering, after k iterations, the maximum distance of points to all centers will be less than d. The 2-approximation follows by the similar triangular inequality argument, where all points in a cluster have distance at most $2d$.