Pr. 1. (sol/hs064)

(a) We know that $\|\boldsymbol{A}\|_2^2 = \sigma_1^2$ and $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = \sum_i \sigma_i^2$. So clearly $\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_{\mathrm{F}}^2$. Taking square roots yields the desired solution.

(b) Since $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ we have $\sum_i \sigma_i^2 \leq \sum_{i=1}^r \sigma_1^2 = r\sigma_1^2$, so $\|\boldsymbol{A}\|_{\mathrm{F}}^2 \leq r\,\|\boldsymbol{A}\|_2^2$. Taking square roots yields the desired solution.

(c) We know that $\|\boldsymbol{A}\|_*^2 = \left(\sum_i \sigma_i\right)^2 = \sum_i \sigma_i^2 + 2\sum_{i\neq j}\sigma_i\sigma_j = \|\boldsymbol{A}\|_{\mathrm{F}}^2 + 2\sum_{i\neq j}\sigma_i\sigma_j \geq \|\boldsymbol{A}\|_{\mathrm{F}}^2$.

(d) One way is to use the identity $(a-b)^2 \geq 0 \Rightarrow a^2 + b^2 \geq 2ab$ for any a, b. Applying this to the summation above yields $2\sum_{i\neq j}\sigma_i\sigma_j \leq (r-1)\sum_i \sigma_i^2$ because there are $r-1$ terms in the $i \neq j$ summation. (If this is not clear, try multiplying it out for some small r.) Thus $\|\boldsymbol{A}\|_*^2 = \left(\sum_i \sigma_i\right)^2 = \sum_i \sigma_i^2 + 2\sum_{i\neq j}\sigma_i\sigma_j \leq r\sum_i \sigma_i^2 = r\,\|\boldsymbol{A}\|_{\mathrm{F}}^2$.

Another way to see this is to use convexity of the function $f(x) = x^2$, as hinted, *i.e.*, $f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$ for all x_1, x_2 and for any $0 \leq \theta \leq 1$. So we have $\left(\frac{1}{r}\sum_i \sigma_i\right)^2 \leq \frac{1}{r}\sum_i \sigma_i^2$. Multiplying both sides by r^2 yields $\left(\sum \sigma_i\right)^2 \leq r\sum \sigma_i^2$, or $\|\boldsymbol{A}\|_*^2 \leq r\,\|\boldsymbol{A}\|_{\mathrm{F}}^2$.

(e) For a vector $\boldsymbol{y} \in \mathbb{R}^n$,

$$\|\boldsymbol{y}\|_\infty = \max_i |y_i| = \sqrt{\max_i |y_i|^2} \leq \sqrt{\sum_{i=1}^n |y_i|^2} = \|\boldsymbol{y}\|_2.$$

Similarly,

$$\|\boldsymbol{y}\|_2^2 = \sum_{i=1}^n |y_i|^2 \leq n \times \max_i |y_i|^2 = n \times \left(\max_i |y_i|\right)^2 = n\|\boldsymbol{y}\|_\infty^2.$$

Hence,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{y}\|_\infty \leq \|\boldsymbol{y}\|_2.$$

Then, noting that $\|\boldsymbol{A}\|_\infty = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}$, we have

$$\|\boldsymbol{A}\|_\infty = \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_\infty} \leq \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_2} = \sqrt{n}\|\boldsymbol{A}\|_2,$$

as desired. Essentially, we have proved the relevant statement about vectors, and then used the fact that $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{x}$ are vectors.

(f) First show the analogous statement about vectors easily:

$$\boldsymbol{y} \in \mathbb{R}^m \Rightarrow \|\boldsymbol{y}\|_2^2 = \sum_{i=1}^m |y_i|^2 \leq m \times \left(\max_i |y_i|^2\right) = m \times (\max_i |y_i|)^2 = m\|\boldsymbol{y}\|_\infty^2 \Rightarrow \|\boldsymbol{y}\|_2 \leq \sqrt{m}\|\boldsymbol{y}\|_\infty.$$

Now the proof for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is simple:

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leq \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\sqrt{m}\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_2} \leq \sqrt{m}\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \sqrt{m}\|\boldsymbol{A}\|_\infty,$$

where the middle inequality uses the following lower bound that is a special case ($n = 1$) of part (e) above:

$$\|\boldsymbol{x}\|_\infty^2 = (\max_i |x_i|)^2 = \max_i |x_i|^2 \leq \sum_i |x_i|^2 = \|\boldsymbol{x}\|_2^2 \Rightarrow \|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2.$$

(g) As stated in the notes (proving it would be a useful exercise): $\|\boldsymbol{A}\|_1 = \max_j \sum_i |A_{ij}|$, *i.e.*, the largest absolute sum column sum. Thus, $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}^T\|_\infty$. Combining this identity with the results of parts (e) and (f) with m and n exchanged yields: $\frac{1}{\sqrt{m}}\|\boldsymbol{A}\|_1 \leq \|\boldsymbol{A}\|_2 \leq \sqrt{n}\|\boldsymbol{A}\|_1$.

Pr. 2. (sol/hs070)
Here we have that $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ so $\boldsymbol{W}\boldsymbol{r} = \boldsymbol{W}\boldsymbol{b} - \boldsymbol{W}\boldsymbol{A}\boldsymbol{x}$. Let $\widetilde{\boldsymbol{r}} = \boldsymbol{W}\boldsymbol{r}$, $\widetilde{\boldsymbol{A}} = \boldsymbol{W}\boldsymbol{A}$ and $\widetilde{\boldsymbol{b}} = \boldsymbol{W}\boldsymbol{b}$. Then minimizing $\|\boldsymbol{W}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{W}\boldsymbol{b}\|_2$ is equivalent to minimizing $\|\widetilde{\boldsymbol{A}}\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|_2$. Thus the desired solution is simply $\widehat{\boldsymbol{x}} = \widetilde{\boldsymbol{A}}^\dagger \widetilde{\boldsymbol{b}} = \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Sigma}}^\dagger \widetilde{\boldsymbol{U}}^T \boldsymbol{W}\boldsymbol{b}$, where $\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^T$ is an SVD of the weighted matrix $\widetilde{\boldsymbol{A}} = \boldsymbol{W}\boldsymbol{A}$. (The SVD of $\boldsymbol{A}$ itself is *not* useful here.)

Pr. 3. (sol/hs029)
(This solution considers a more general case with some $\theta > 0$. Note that $\theta = 1$ in the problem statement.) An eigenvalue λ of $\boldsymbol{B} = \boldsymbol{A} + \theta \boldsymbol{x}\boldsymbol{x}'$ satisfies $\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = \det(\boldsymbol{A} + \theta\boldsymbol{x}\boldsymbol{x}' - \lambda\boldsymbol{I}) = 0$. Note that when $\lambda \neq A_{ii}$ for any i then $\boldsymbol{A} - \lambda\boldsymbol{I}$ is invertible. Thus

$$\begin{aligned}\det(\boldsymbol{A} + \theta\boldsymbol{x}\boldsymbol{x}' - \lambda\boldsymbol{I}) &= \det((\boldsymbol{A} - \lambda\boldsymbol{I}) \cdot (\boldsymbol{I} + (\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}\theta\boldsymbol{x}\boldsymbol{x}') = \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \cdot (\boldsymbol{I} + \underbrace{\theta(\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}\boldsymbol{x}}_{\triangleq \boldsymbol{w}}\boldsymbol{x}') \\ &= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \cdot (1 + \theta\boldsymbol{x}' \underbrace{(\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}\boldsymbol{x}}_{=\boldsymbol{w}}). \qquad \text{(From from HW 1, } \det(\boldsymbol{I} + \boldsymbol{w}\boldsymbol{x}') = 1 + \boldsymbol{x}'\boldsymbol{w}.)\end{aligned}$$

Thus an eigenvalue of $\boldsymbol{B}$ that is not an eigenvalue of $\boldsymbol{A}$ must satisfy the equation

$$\boldsymbol{x}'(\lambda\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{x} = \frac{1}{\theta}.$$

Since $\boldsymbol{A}$ is diagonal, a_{ii} are the eigenvalues of $\boldsymbol{A}$, and we obtain the (implicit) relationship

$$\frac{x_1^2}{\lambda - a_{11}} + \frac{x_2^2}{\lambda - a_{22}} + \ldots + \frac{x_n^2}{\lambda - a_{nn}} = \frac{1}{\theta}. \tag{1}$$

Because $x_i \neq 0$ is given, this equation will have n solutions whenever $\boldsymbol{A}$ does not have repeated eigenvalues. When $\boldsymbol{A}$ has repeated eigenvalues, some of the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ will coincide because the degree of the above polynomial will not be n. The problem statement says $\boldsymbol{A}$ has distinct entries so this technicality is avoided.
Alternatively, one can use properties 16, 17 in Section 1.4 of Laub (see HW1, Question 3), as follows:

$$\begin{aligned}\det(\boldsymbol{A} - \lambda\boldsymbol{I} + \theta\boldsymbol{x}\boldsymbol{x}') &= \det\left(\begin{bmatrix} \boldsymbol{A} - \lambda\boldsymbol{I} & -\theta\boldsymbol{x} \\ \boldsymbol{x}' & 1 \end{bmatrix}\right) \\ &= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \det(1 + \theta\boldsymbol{x}'(\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}\boldsymbol{x}).\end{aligned}$$

Here is another way. Let z be an eigenvalue of $\boldsymbol{B}$ associated with the eigenvector $\boldsymbol{v}$. Then

$$\begin{aligned}&\boldsymbol{B}\boldsymbol{v} = z\boldsymbol{v} \Rightarrow (\boldsymbol{A} + \theta\boldsymbol{x}\boldsymbol{x}')\boldsymbol{v} = z\boldsymbol{v} \Rightarrow (z\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} = \theta\boldsymbol{x}\boldsymbol{x}'\boldsymbol{v} \\ &\Rightarrow \boldsymbol{v} = \theta(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{v} \qquad (\text{for } z \neq \lambda(\boldsymbol{A})) \\ &\Rightarrow (\boldsymbol{x}'\boldsymbol{v}) = \theta\boldsymbol{x}^T(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{x}(\boldsymbol{x}'\boldsymbol{v}) \\ &\Rightarrow 1 = \theta\boldsymbol{x}^T(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{x} \qquad (\text{assuming that the scalar } \boldsymbol{x}'\boldsymbol{v} \neq 0).\end{aligned}$$

The assumption that $\boldsymbol{x}'\boldsymbol{v} \neq 0$ implies that $\boldsymbol{v} \notin \mathcal{N}(\boldsymbol{x})$. More work would be needed to complete this argument, although the equivalence with the previous derivations shows that this assumption is valid whenever $x_i \neq 0$ when considering the eigenvalues of $\boldsymbol{B}$ that are not equal to the eigenvalues of $\boldsymbol{A}$.

Pr. 4. (sol/hs082)

(a) Note that $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \delta^2 \|\boldsymbol{x}\|_2^2 = \| \underbrace{\begin{bmatrix} \boldsymbol{A} \\ \delta \boldsymbol{I} \end{bmatrix}}_{=:\widetilde{\boldsymbol{A}}} \boldsymbol{x} - \underbrace{\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}}_{=:\widetilde{\boldsymbol{b}}} \|_2^2$. Thus we can use the standard least squares result to get the solution $\hat{\boldsymbol{x}} = \widetilde{\boldsymbol{A}}^+ \widetilde{\boldsymbol{b}} = (\boldsymbol{A}'\boldsymbol{A} + \delta \boldsymbol{I})^{-1} \boldsymbol{A}'\boldsymbol{b}$, where we use the fact that $(\boldsymbol{A}'\boldsymbol{A} + \delta \boldsymbol{I})$ is always invertible whenever $\delta > 0$.

(b) As $\delta \to \infty$ we have $\hat{\boldsymbol{x}}(\delta) \to \boldsymbol{0}$ which makes sense because the $\delta^2 \|\boldsymbol{x}\|_2^2$ term dominates the cost function.

(c) The iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mu \widetilde{\boldsymbol{A}}'(\widetilde{\boldsymbol{A}}\boldsymbol{x}_k - \widetilde{\boldsymbol{b}}) = \boldsymbol{x}_k - \mu \left(\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) + \delta^2 \boldsymbol{x}_k \right)$, will converge to $\hat{\boldsymbol{x}}$ whenever $\mu < 2/\sigma_1(\widetilde{\boldsymbol{A}})^2$.

(d) Note that

$$\sigma_1(\widetilde{\boldsymbol{A}}) = \sqrt{\sigma_1^2(\boldsymbol{A}) + \delta^2},$$

and we can use this equality to determine the range of allowable step sizes: $0 < \mu < \frac{2}{\sigma_1^2(\boldsymbol{A}) + \delta^2}$.

Pr. 5. (sol/hsx01)
Grader: full credit for any proposed exam problem that looks like a sincere attempt, iff it is submitted on time in the proper format (pdf) and location (Canvas). You do not need to check accuracy of the answer given.

Pr. 6. (sol/hs103)

Let $\boldsymbol{a}_i := \mathrm{vec}(\boldsymbol{A}_i) = \begin{bmatrix} \boldsymbol{A}_i(:,1) \\ \boldsymbol{A}_i(:,2) \\ \vdots \\ \boldsymbol{A}_i(:,n) \end{bmatrix} \in \mathbb{R}^{mn}$. Similarly let $\boldsymbol{b} = \mathrm{vec}(\boldsymbol{B})$.

For any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F = \|\mathrm{vec}(\boldsymbol{A})\|_2$, since the squared Frobenius norm is the sum of squares of each element within the matrix, which is equal to the sum of squares of each element in the vec of the matrix. Therefore,

$$\begin{aligned} \underset{x_1,\ldots,x_k}{\arg\min} \left\| \sum_{i=1}^k x_i \boldsymbol{A}_i - \boldsymbol{B} \right\|_{\mathrm{F}} &= \underset{x_1,\ldots,x_k}{\arg\min} \left\| \mathrm{vec}\left(\sum_{i=1}^k x_i \boldsymbol{A}_i - \boldsymbol{B} \right) \right\|_2 = \underset{x_1,\ldots,x_k}{\arg\min} \left\| \sum_{i=1}^k x_i \mathrm{vec}(\boldsymbol{A}_i) - \mathrm{vec}(\boldsymbol{B}) \right\|_2 \\ &= \underset{\boldsymbol{x}}{\arg\min} \| \underbrace{\begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_k \end{bmatrix}}_{:=\widetilde{\boldsymbol{A}}} \boldsymbol{x} - \boldsymbol{b} \|_2. \end{aligned}$$

The problem is now reduced to the ordinary least squares formulation, and the solution is given by $\hat{\boldsymbol{x}} = \left(\widetilde{\boldsymbol{A}} \right)^\dagger \boldsymbol{b}$.

Pr. 7. (sol/hs087)

(a) A possible `Julia` implementation is

```
function lsngd(A, b; x0 = zeros(size(A,2)), nIters = 200, mu = 0)
#
# Syntax:       x = lsngd(A, b; mu, x0, nIters)
#
# Inputs:       A is a m x n matrix
#
#               b is a vector of length m
#
#               mu is the step size to use, and must satisfy
#               0 < mu <= 1 / sigma_1(A)^2 to guarantee convergence
#               where sigma_1(A) is the first (largest) singular value.
#               A default value for mu will be explained in Ch.5.
#
#               x0 is the initial starting vector (of length n) to use.
#               Its default value is all zeros for simplicity.
#
#               nIters is the number of iterations to perform (default 200)
#
# Outputs:      x is a vector of length n containing the approximate solution
#
# Description:  Performs Nesterov-accelerated gradient descent
#               to solve the least squares problem
#               \argmin_x \| b - A x \|_2
#

    if (mu == 0) # use the following default value:
        mu = 1. / (maximum(sum(abs.(A),dims=1)) * maximum(sum(abs.(A),dims=2)))
    end

    # Parse inputs
    b = vec(b)
    x0 = vec(x0)

    # Nesterov-accelerated gradient descent
    t = 0
    xLast = x0
    x = x0
    for _ in 1:nIters
        # t update
        tLast = t
        t = 0.5 * (1 + sqrt(1 + 4 * t^2))

        # z update (momentum)
        z = x + ((tLast - 1) / t) * (x - xLast)

        # x update
        xLast = x
        x = z - mu * (A' * (A * z - b))
    end

    return x
end
```

(b) Figure 1 compares sequences of $\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|$ versus k for standard gradient descent and Nestorov-accelerated gradient descent.

(c) For step sizes $\mu = \{0.25, 0.5, 0.75, 1\}/\sigma_1^2(A)$, clearly Nesterov-accelerated gradient descent converges faster than standard gradient descent, although the convergence is not necessarily monotone.

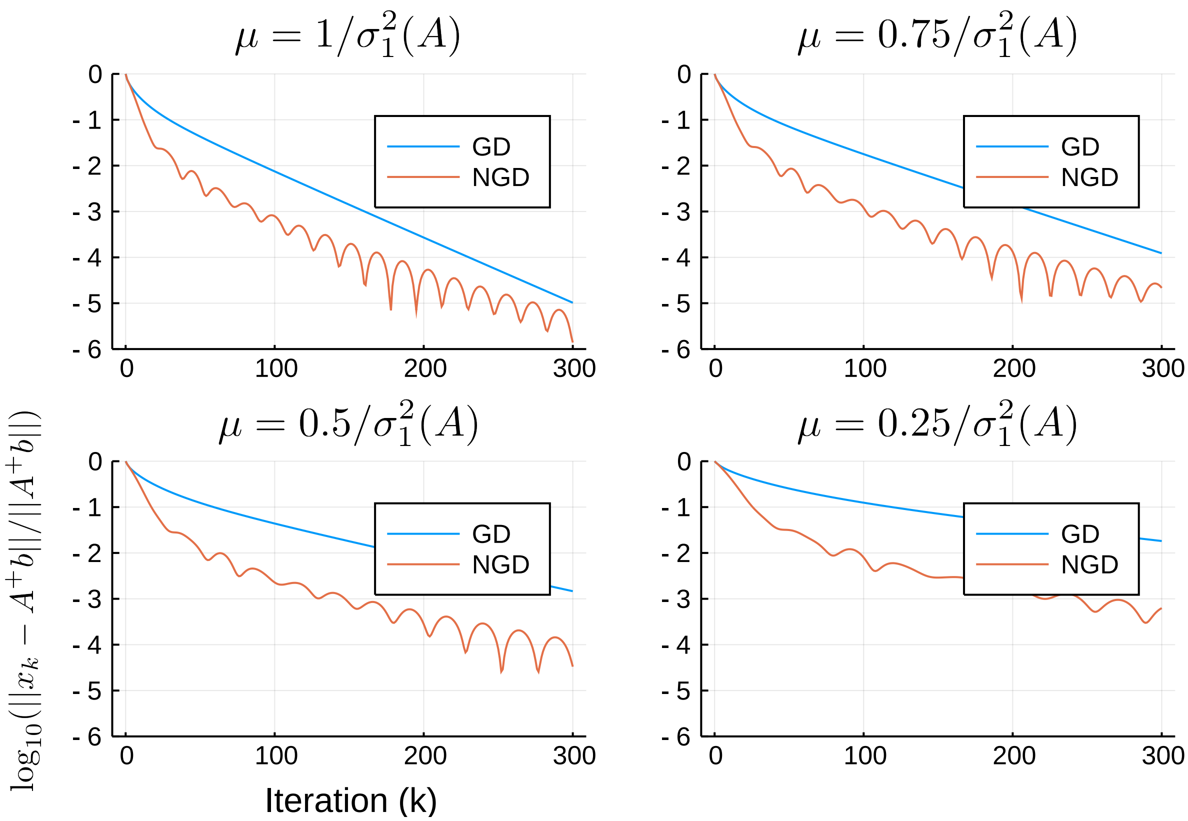

Figure 1: Nesterov-accelerated vs. standard gradient descent for a least squares problem, for four values of step size μ.

Optional problem(s) below

Pr. 8. (sol/hsj41)

(a) From a previous problem, the general solution is $\hat{\boldsymbol{x}} = (\boldsymbol{A}'\boldsymbol{A} + \delta\boldsymbol{I})^{-1}\boldsymbol{A}'\boldsymbol{b}$.

Using the push-through identity, $\hat{\boldsymbol{x}} = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}' + \delta\boldsymbol{I})^{-1}\boldsymbol{b}$.

Here, because $\boldsymbol{A}$ is a Parseval tight frame and $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{z}$, we have $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$ so $\hat{\boldsymbol{x}} = \boldsymbol{A}'(\boldsymbol{I} + \delta\boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{z} = \frac{1}{1+\delta}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{z}$.

In general, no further simplification seems possible.

(b) If $\delta = 0$ then $\hat{\boldsymbol{x}} = \boldsymbol{A}'\boldsymbol{A}\boldsymbol{z}$, and $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}\boldsymbol{z}$, so both $\hat{\boldsymbol{x}}$ and $\boldsymbol{z}$ here provide equivalent signal representations

(c) The singular values of $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$ which are all 1, so $\kappa(\boldsymbol{A}) = 1$, which is the best possible condition number.
