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 $\|\boldsymbol{A}\|_2^2 \le \|\boldsymbol{A}\|_{\mathrm{F}}^2$. Taking square roots yields the desired inequality.

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

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

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

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

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

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i| = \sqrt{\max_i |x_i|^2} \le \sqrt{\sum_{i=1}^n |x_i|^2} = \|\boldsymbol{x}\|_2.$$

Similarly,

$$\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^n |x_i|^2 \le n \times \max_i |x_i|^2 = n \times \left(\max_i |x_i|\right)^2 = n\|\boldsymbol{x}\|_\infty^2.$$

Hence,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2.$$

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

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

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) It follows from the previous part that $\boldsymbol{y} \in \mathbb{F}^m \Rightarrow \|\boldsymbol{y}\|_2 \le \sqrt{m}\|\boldsymbol{y}\|_\infty$ and $1/\|\boldsymbol{x}\|_2 \le 1/\|\boldsymbol{x}\|_\infty$. Now the proof for $\boldsymbol{A} \in \mathbb{F}^{m\times n}$ is simple:

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\ne\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \le \max_{\boldsymbol{x}\ne\boldsymbol{0}} \frac{\sqrt{m}\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_2} \le \sqrt{m}\max_{\boldsymbol{x}\ne\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \sqrt{m}\|\boldsymbol{A}\|_\infty.$$

(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 column sum. Thus, $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}'\|_\infty$. Combining this identity with the results of parts (e) and (f) with m and n exchanged yields: $\frac{1}{\sqrt{m}}\|\boldsymbol{A}\|_1 \le \|\boldsymbol{A}\|_2 \le \sqrt{n}\|\boldsymbol{A}\|_1$.

Optional challenge.
$\boldsymbol{A} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ leads to equality in (a) (b) (c) (d) and the second inequality in (g)
$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \end{bmatrix}$ leads to equality in (f) and the first inequality in (g)

Pr. 2. (sol/hsj62)
If either $\boldsymbol{A}$ or $\boldsymbol{B}$ is $\boldsymbol{0}$ then the problem is trivial, so assume they are both nonzero, with $r = \operatorname{rank}(\boldsymbol{A})$ and $s = \operatorname{rank}(\boldsymbol{B})$. Denote compact SVDs of $\boldsymbol{A}$ and $\boldsymbol{B}$ by $\boldsymbol{A} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r'$ and $\boldsymbol{B} = \boldsymbol{X}_s \boldsymbol{\Omega}_s \boldsymbol{Y}_s'$. Here $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Omega}_s$ are square and symmetric and invertible (but possibly different sizes if $\boldsymbol{A}$ and $\boldsymbol{B}$ have different ranks).
Now $\boldsymbol{A}\boldsymbol{B}' = \boldsymbol{0}_{M\times M}$ means $\boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r' \boldsymbol{Y}_s \boldsymbol{\Omega}_s \boldsymbol{X}_s' = \boldsymbol{0}_{M\times M}$. Multiplying both sides on the left by $\boldsymbol{\Sigma}_r^{-1} \boldsymbol{U}_r'$ and on the right by $\boldsymbol{X}_s \boldsymbol{\Omega}_s^{-1}$ shows that $\boldsymbol{V}_r' \boldsymbol{Y}_s = \boldsymbol{0}_{r\times s}$. So $\begin{bmatrix} \boldsymbol{V}_r & \boldsymbol{Y}_s \end{bmatrix}$ is a matrix with $r+s$ orthonormal columns.
Likewise $\boldsymbol{A}'\boldsymbol{B} = \boldsymbol{0}$ leads to $\boldsymbol{U}_r' \boldsymbol{X}_s = \boldsymbol{0}_{r\times s}$, so $\begin{bmatrix} \boldsymbol{U}_r & \boldsymbol{X}_s \end{bmatrix}$ is a matrix with $r+s$ orthonormal columns.
Therefore, the following decomposition is a valid compact SVD of $\boldsymbol{A}+\boldsymbol{B}$, to within permutations for sorting the singular values:

$$\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r' + \boldsymbol{X}_s \boldsymbol{\Omega}_s \boldsymbol{Y}_s' = \begin{bmatrix} \boldsymbol{U}_r & \boldsymbol{X}_s \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_s \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_r & \boldsymbol{Y}_s \end{bmatrix}'.$$

(This "SVD of a sum" may be useful itself when the conditions hold.) Thus

$$\|\boldsymbol{A}+\boldsymbol{B}\|_* = \|\boldsymbol{\Sigma}_r\|_* + \|\boldsymbol{\Omega}_s\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*.$$

Now there is one more subtle point we must address here. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are both $M \times N$, we need to be sure that $\operatorname{size}(\boldsymbol{\Sigma}_r) + \operatorname{size}(\boldsymbol{\Omega}_s) \leq \min(M,N)$. Specifically, if $\boldsymbol{U}_r$ is $M \times r$ and $\boldsymbol{X}_s$ is $M \times s$, then we need $r + s \leq M$ for the above compact SVD of $\boldsymbol{A}+\boldsymbol{B}$ to be valid. This inequality is assured by the condition $\boldsymbol{U}_r' \boldsymbol{X}_s = \boldsymbol{0}_{r\times s}$ because $\boldsymbol{U}_r$ and $\boldsymbol{X}_s$ are each orthonormal bases for subspaces in $\mathbb{F}^M$, so if the sum of their dimensions were to exceed M then their spans would have a nontrivial intersection which would contradict $\boldsymbol{U}_r' \boldsymbol{X}_s = \boldsymbol{0}$. Likewise for $\boldsymbol{V}_r$ and $\boldsymbol{Y}_s$.

Pr. 3. (sol/hsj5i)

(a) Yes. For $\boldsymbol{A} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r'$, the Moore-Penrose pseudo-inverse is $\boldsymbol{A}^+ = \boldsymbol{V}_r \boldsymbol{\Sigma}_r^{-1} \boldsymbol{U}_r'$ which has nonzero singular values $1/\sigma_r, \ldots, 1/\sigma_1$ and hence spectral norm $\|\boldsymbol{A}^+\|_2 = 1/\sigma_r$. And $\boldsymbol{G} = \boldsymbol{A}^+$ is a valid generalized inverse.

(b) Yes. For $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r'$, define $\boldsymbol{G} = \boldsymbol{V} \begin{bmatrix} \boldsymbol{\Sigma}_r^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & 7\boldsymbol{I} \end{bmatrix} \boldsymbol{U}'$ then one can verify that $\boldsymbol{AGA} = \boldsymbol{A}$ and $\|\boldsymbol{G}\|_2 = \max(1/\sigma_1, \ldots, 1/\sigma_r, 7) = 7$ because $1 = \sigma_r \leq \ldots \leq \sigma_1$.

Pr. 4. (sol/hsj5s)
Claim: If $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal projection matrices, then $\rho(\frac{1}{2}\boldsymbol{A} + \frac{1}{2}\boldsymbol{B}) \leq 1$.
$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{U}'$ and $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}'$ for some matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ having orthonormal columns, so $\|\boldsymbol{A}\|_2, \|\boldsymbol{B}\|_2 \leq 1$,
and $\rho\left(\frac{1}{2}\boldsymbol{A} + \frac{1}{2}\boldsymbol{B}\right) = \left\|\frac{1}{2}\boldsymbol{A} + \frac{1}{2}\boldsymbol{B}\right\|_2 \leq \frac{1}{2}\|\boldsymbol{A}\|_2 + \frac{1}{2}\|\boldsymbol{B}\|_2 \leq 1$.
Arguments of the form " $\rho\left(\frac{1}{2}\boldsymbol{A} + \frac{1}{2}\boldsymbol{B}\right) \leq \frac{1}{2}\rho(\boldsymbol{A}) + \frac{1}{2}\rho(\boldsymbol{B})$" are incorrect because in general it is not the case that $\rho(\boldsymbol{A}+\boldsymbol{B}) \leq \rho(\boldsymbol{A}) + \rho(\boldsymbol{B})$, because spectral radius is not a norm. In this problem $\boldsymbol{A}$ and $\boldsymbol{B}$ are given to be orthogonal projection matrices, so they are symmetric, hence $\rho(\boldsymbol{A}) = \|\boldsymbol{A}\|_2$. But still an explanation that uses "$\rho(\boldsymbol{A}+\boldsymbol{B}) \leq \rho(\boldsymbol{A}) + \rho(\boldsymbol{B})$" or such does not earn full credit.
There is no "triangle inequality" for spectral radius in general.

Pr. 5. (sol/hs120)

(a) Here $f(\boldsymbol{x}) = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$, so that

$$\begin{aligned} \|f(\boldsymbol{x}) - f(\boldsymbol{y})\| &= \|\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b})\| = \|\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})\| \\ &\leq \|\boldsymbol{A}'\boldsymbol{A}\|_2 \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sigma_1^2(\boldsymbol{A})\|\boldsymbol{x} - \boldsymbol{y}\|_2 \qquad (\text{since } \|\boldsymbol{A}'\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A}'\boldsymbol{A}) = \sigma_1^2(\boldsymbol{A})), \end{aligned}$$

and so $\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$ is a Lipschitz function with Lipschitz constant $\sigma_1^2(\boldsymbol{A})$.

(b) Now let $f(\boldsymbol{X}) = \sigma_1(\boldsymbol{X})$. Apply the reverse triangle inequality $|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$ as follows:

$$|\sigma_1(\boldsymbol{X}) - \sigma_1(\boldsymbol{Y})| = |\|\boldsymbol{X}\|_2 - \|\boldsymbol{Y}\|_2| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|_2 = \sigma_1(\boldsymbol{X} - \boldsymbol{Y}),$$

Thus $\sigma_1(\cdot)$ is a 1-Lipschitz function with respect to the spectral norm.

(c) A previous HW problem showed that $\sigma_1(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_{\mathrm{F}}$ so the above inequalities also lead to $|\sigma_1(\boldsymbol{X}) - \sigma_1(\boldsymbol{Y})| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}$. Thus $\sigma_1(\cdot)$ is also a 1-Lipschitz function with respect to the Frobenius norm.

Pr. 6. (sol/hs077)

(a) A possible `Julia` implementation is

```
"""
    `x = nnlsgd(A, b ; mu=0, x0=zeros(size(A,2)), nIters::Int=200)`

Performs projected gradient descent to solve the least squares problem:
``\\argmin_{x \\geq 0} 0.5 \\| b - A x \\|_2`` with nonnegativity constraint.

In:
- `A` `m x n` matrix
- `b` vector of length `m`

Option:
- `mu` step size to use, and must satisfy ``0 < mu < 2 / \\sigma_1(A)^2``
to guarantee convergence,
where ``\\sigma_1(A)`` is the first (largest) singular value.
Ch.5 will explain a default value for `mu`
- `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)

Out:
`x` vector of length `n` containing the approximate LS solution
"""
function nnlsgd(A, b ; mu::Real=0, x0=zeros(size(A,2)), nIters::Int=200)

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

    x = x0
    for _ in 1:nIters
        x -= mu * (A' * (A * x - b)) # gradient descent step
        x = max.(x, 0) # project onto non-negative orthant
    end

    return x
end
```

(b) After 100 iterations, the first three values are
`x[1:3] = [0.8891, 0.9788, 1.20026]`

(c) `xnnls[5:7]` is 1.058567, 0.37206, 0.90699

(d) Figure 1 shows one realization of the system. Since $\mu \in (0, 2/\sigma_1^2(\boldsymbol{A}))$, the iterates converge to the NNLS solution.

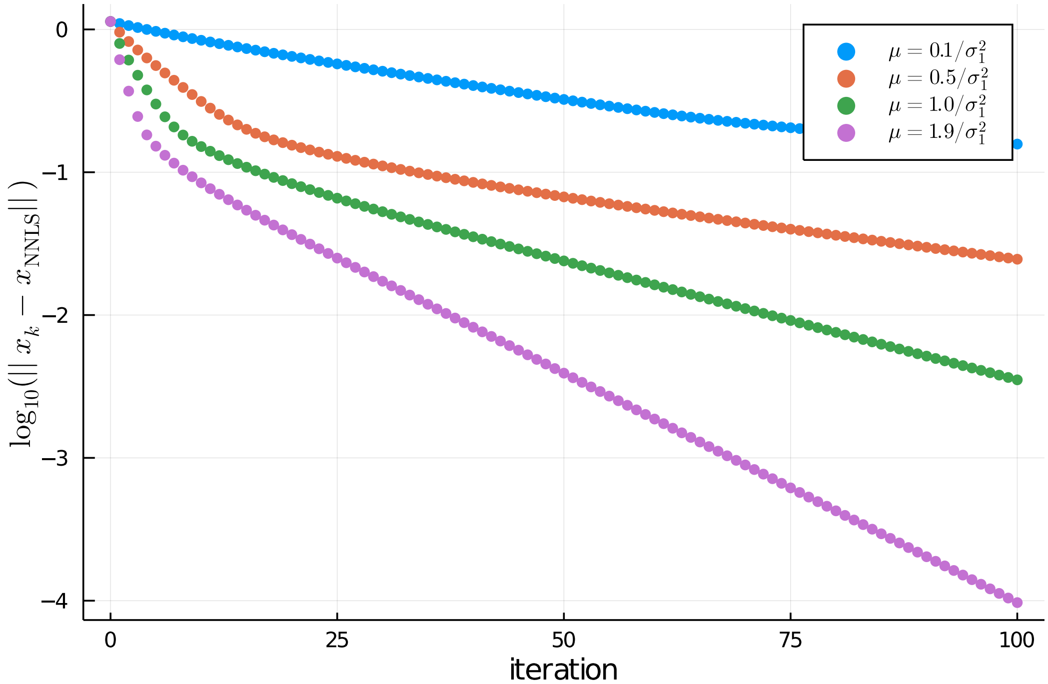


Figure 1: Convergence of projected gradient sequence $\boldsymbol{x}_k$ to the NNLS minimizer $\boldsymbol{x}_{\mathrm{NNLS}}$.

Pr. 7. (sol/hs043)

(a) For fixed α and $\boldsymbol{Q}$, we have a least squares problem of the form

$$\arg\min_{\boldsymbol{\mu}} \|\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n'\|_{\mathrm{F}}^2 ,$$

where $\boldsymbol{Z} \triangleq \boldsymbol{B} - \alpha\boldsymbol{Q}(\boldsymbol{A} - \boldsymbol{\mu}_A\mathbf{1}_n')$. There are a few ways to solve this, all of which require a bit of deriving.

1. Convert Frobenius norm to a trace:

$$\begin{aligned}\|\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n'\|_{\mathrm{F}}^2 &= \mathrm{trace}((\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n')'(\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n')) = \mathrm{trace}(\boldsymbol{Z}'\boldsymbol{Z}) - 2\mathrm{trace}(\boldsymbol{Z}'\boldsymbol{\mu}\mathbf{1}_n') + \mathrm{trace}(\boldsymbol{\mu}'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{\mu}) \\ &= \mathrm{trace}(\boldsymbol{Z}'\boldsymbol{Z}) - 2\boldsymbol{\mu}'\boldsymbol{Z}\mathbf{1}_n + n\,\|\boldsymbol{\mu}\|_2^2 .\end{aligned}$$

Zeroing the gradient w.r.t. $\boldsymbol{\mu}$ yields $\mathbf{0} = -2\boldsymbol{Z}\mathbf{1}_n + 2\boldsymbol{\mu} \Rightarrow \boldsymbol{\mu}_* = \boldsymbol{Z}\mathbf{1}_n$.

2. Convert the Frobenious norm squared into a sum of row Euclidean norms squared:

$$\|\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n'\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \left\|\boldsymbol{Z}_{i,:}^T - \mathbf{1}_n\mu_i\right\|_2^2 \Rightarrow \hat{\mu}_i = \mathbf{1}_n^+\boldsymbol{Z}_{i,:}^T = \frac{1}{n}\mathbf{1}_n'\boldsymbol{Z}_{i,:}^T = \frac{1}{n}\boldsymbol{Z}_{i,:}\mathbf{1}_n \Rightarrow \boldsymbol{\mu}_* = \frac{1}{n}\boldsymbol{Z}\mathbf{1}_n.$$

3. Convert the Frobenious norm into a Euclidean norm using vec and use vec trick twice:

$$\begin{aligned}&\|\boldsymbol{Z} - \boldsymbol{\mu}\mathbf{1}_n'\|_{\mathrm{F}}^2 = \|\mathrm{vec}(\boldsymbol{Z}) - \mathrm{vec}(\boldsymbol{\mu}\mathbf{1}_n')\|_2^2 = \|\mathrm{vec}(\boldsymbol{Z}) - (\mathbf{1}_n \otimes \boldsymbol{I})\boldsymbol{\mu}\|_2^2 \Rightarrow \\ &\qquad \boldsymbol{\mu}_* = (\mathbf{1}_n \otimes \boldsymbol{I})^+\mathrm{vec}(\boldsymbol{Z}) = (\mathbf{1}_n^+ \otimes \boldsymbol{I})\mathrm{vec}(\boldsymbol{Z}) = \frac{1}{n}(\mathbf{1}_n' \otimes \boldsymbol{I})\mathrm{vec}(\boldsymbol{Z}) = \frac{1}{n}\mathrm{vec}(\boldsymbol{I}\boldsymbol{Z}\mathbf{1}_n) = \frac{1}{n}\boldsymbol{Z}\mathbf{1}_n\end{aligned}$$

Thus by any of these derivations the optimal $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu}_* = \frac{1}{n}\boldsymbol{Z}\mathbf{1}_n = \frac{1}{n}\boldsymbol{B}\mathbf{1}_n - \frac{\alpha}{n}\boldsymbol{Q}\left(\boldsymbol{A} - \boldsymbol{\mu}_A\mathbf{1}_n'\right)\mathbf{1}_n = \boldsymbol{\mu}_B - \alpha\boldsymbol{Q}\left(\boldsymbol{\mu}_A - \boldsymbol{\mu}_A\right) = \boldsymbol{\mu}_B.$$

In turn, with $\boldsymbol{\mu} = \boldsymbol{\mu}_*$ and fixed α, we have the problem

$$\arg\min_{\boldsymbol{Q}:\ \boldsymbol{Q}'\boldsymbol{Q}=\boldsymbol{I}_d} \|\boldsymbol{B}_0 - \alpha\boldsymbol{Q}\boldsymbol{A}_0\|_{\mathrm{F}} .$$

The course notes show that the solution to the above problem is given by $\boldsymbol{Q}_* = \boldsymbol{U}\boldsymbol{V}'$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'$ is an SVD of $\boldsymbol{B}_0(\alpha\boldsymbol{A}_0') = \alpha\boldsymbol{B}_0\boldsymbol{A}_0'$. However, note that the singular vectors of $\alpha\boldsymbol{B}_0\boldsymbol{A}_0'$ are the same as the singular vectors of $\boldsymbol{B}_0\boldsymbol{A}_0'$ when $\alpha > 0$ (to within sign changes that do not affect the $\boldsymbol{U}\boldsymbol{V}'$ product), so the optimal $\boldsymbol{Q}$ is in fact independent of $\alpha > 0$. Thus $\boldsymbol{Q}_* = \boldsymbol{V}\boldsymbol{U}'$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are left and right singular vectors, respectively, of $\boldsymbol{B}_0\boldsymbol{A}_0'$.

Finally, with $\boldsymbol{\mu} = \boldsymbol{\mu}_*$ and $\boldsymbol{Q} = \boldsymbol{Q}_*$, we have the problem

$$\arg\min_{\alpha \geq 0} \Phi(\alpha), \qquad \Phi(\alpha) \triangleq \left\| \boldsymbol{B}_0 - \alpha \underbrace{\boldsymbol{Q}_* \boldsymbol{A}_0}_{\triangleq \tilde{\boldsymbol{A}}_0} \right\|_F^2$$

$$= \operatorname{Tr}((\boldsymbol{B}_0 - \tilde{\boldsymbol{A}}_0\alpha)(\boldsymbol{B}_0 - \tilde{\boldsymbol{A}}_0\alpha)') = \alpha^2 \operatorname{Tr}(\tilde{\boldsymbol{A}}_0\tilde{\boldsymbol{A}}_0') - 2\alpha \operatorname{Tr}(\boldsymbol{B}_0\tilde{\boldsymbol{A}}_0') + \operatorname{Tr}(\boldsymbol{B}_0\boldsymbol{B}_0').$$

Differentiating and zeroing yields

$$0 = \dot{\Phi}(\alpha_*) = 2\alpha_* \operatorname{Tr}(\tilde{\boldsymbol{A}}_0\tilde{\boldsymbol{A}}_0') - 2\operatorname{Tr}(\boldsymbol{B}_0\tilde{\boldsymbol{A}}_0')$$

$$\Rightarrow \alpha_* = \frac{\operatorname{Tr}(\boldsymbol{B}_0\tilde{\boldsymbol{A}}_0')}{\operatorname{Tr}(\tilde{\boldsymbol{A}}_0\tilde{\boldsymbol{A}}_0')} = \frac{\operatorname{Tr}(\boldsymbol{B}_0\boldsymbol{A}_0'\boldsymbol{Q}_*')}{\operatorname{Tr}(\boldsymbol{Q}_*\boldsymbol{A}_0\boldsymbol{A}_0'\boldsymbol{Q}_*')} = \frac{\operatorname{Tr}(\boldsymbol{B}_0\boldsymbol{A}_0'\boldsymbol{Q}_*')}{\operatorname{Tr}(\boldsymbol{A}_0\boldsymbol{A}_0')}.$$

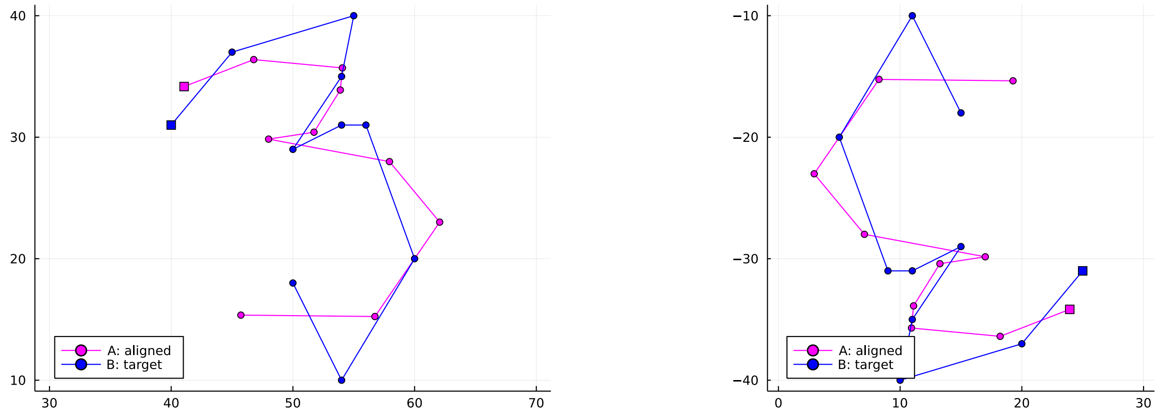
One can verify that $\alpha_* \geq 0$.

(b) A possible `Julia` implementation is

```
using LinearAlgebra: svd, tr, norm
using Statistics: mean

"""
    Aa = procrustes(B, A ; center::Bool=true, scale::Bool=true)

In:
* `B` and `A` are `d × n` matrices

Option:
* `center=true/false` : consider centroids?
* `scale=true/false` : optimize alpha or leave scale as 1?
Your solution needs only to consider the defaults for these.

Out:
* `Aa` `d × n` matrix containing `A` Procrustes-aligned to `B`

Returns `Aa = alpha * Q * (A - muA) + muB`, where `muB` and `muA` are
the `d × n` matrices whose rows contain copies of the centroids of
`B` and `A`, and `alpha` (scalar) and `Q` (`d × d` orthogonal matrix) are
the solutions to the Procrustes + centering / scaling problem

`\\argmin_{alpha, muA, muB, Q: Q'Q = I} \\| (B - muB) - alpha * Q (A - muA) \\|_F`
"""
function procrustes(B, A ; center::Bool=true, scale::Bool=true)

    # Center data
    if center
        muB = mean(B, dims=2)
        muA = mean(A, dims=2)
        B0 = B .- muB
        A0 = A .- muA
    else
        B0 = B
        A0 = A
        muB = 0
    end

    # Procrustes rotation
    U, _, V = svd(B0 * A0')
    Q = U * V'

    # Optimal scaling
    if scale
        alpha = tr(B0 * A0' * Q') / norm(A0)^2
    else
```

```
        alpha = 1
    end

    # Align data
    Aa = alpha * (Q * A0) .+ muB
    return Aa
end
```

(c) Figure 2 depicts the aligned digits for each dataset.



(a) Dataset 1 (b) Dataset 2

Figure 2: Aligned digits.

(d) For both datasets $\|\hat{\boldsymbol{A}} - \boldsymbol{B}\|_{\mathrm{F}} \approx 12.804$

Pr. 8. (sol/hsj9m)
Thank you for your feedback via the course evaluations.

Non-graded problem(s) below

Pr. 9. (sol/hs134)

(a) Rewrite the problem as

$$\arg\min_{\boldsymbol{X}} \|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_{\mathrm{F}} = \arg\min_{\boldsymbol{X}} \|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_{\mathrm{F}}^2 = \arg\min_{\boldsymbol{X}} \sum_{k=1}^{K} \|\boldsymbol{A}\boldsymbol{X}[:,k] - \boldsymbol{B}[:,k]\|_2^2.$$

The minimization may be performed independently over each term in the summation: each term is a least squares problem on its own! Hence, we have that $\hat{\boldsymbol{X}}[:,k] = \boldsymbol{A}^{+}\boldsymbol{B}[:,k]$, or equivalently that $\hat{\boldsymbol{X}} = \boldsymbol{A}^{+}\boldsymbol{B}$.

(b) If we only needed to solve for a single column of $\hat{\boldsymbol{X}}$, then backslash would be the most efficient:

```
Xh[:,1] = A \ B[:,1]
```

But when $\boldsymbol{B}$ is very wide, it is more efficient to precompute the pseudo-inverse and apply it to all columns:

```
Xh = pinv(A) * B
```

Pr. 10. (sol/hs034)
Here $\boldsymbol{A} \in \mathbb{R}_8^{19\times 48}$ has rank $r = 8$. We are looking for the number of linearly independent solutions to the homogeneous linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. All vectors $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ must belong to the nullspace of $\boldsymbol{A}$. Thus, the number of linearly independent solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ is precisely the dimension of the nullspace of $\boldsymbol{A}$. From the rank-plus-nullity theorem (Corollary 3.18), we have that $n = \dim \mathcal{N}(A) + \dim \mathcal{R}(A)$. Here $n = 48$, $r = \mathrm{rank}(\boldsymbol{A}) = \dim \mathcal{R}(A) = 8$ so that we must have $\dim \mathcal{N}(A) = 48 - 8 = 40$. Thus there are 40 linearly independent solutions of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$.

Pr. 11. (sol/hs038)
Here is the solution for the general case. If $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{B} \in \mathbb{R}^{N\times M}$, then

$$\mathrm{vec}(\boldsymbol{A}^T) = \begin{bmatrix} A(1,:)^T \\ A(2,:)^T \\ \vdots \\ A(M,:)^T \end{bmatrix} \quad \text{and} \quad \mathrm{vec}(\boldsymbol{B}) = \begin{bmatrix} B(:,1) \\ B(:,2) \\ \vdots \\ B(:,M) \end{bmatrix} \Rightarrow \mathrm{vec}(\boldsymbol{A}^T)^T \mathrm{vec}(\boldsymbol{B}) = \sum_{m=1}^{M} A(m,:)B(:,m).$$

Furthermore,

$$\begin{aligned}
\mathrm{trace}(\boldsymbol{A}\boldsymbol{B}) &= \mathrm{trace}\left(\begin{bmatrix} A(1,:) \\ A(2,:) \\ \vdots \\ A(M,:) \end{bmatrix} \begin{bmatrix} B(:,1) & B(:,2) & \dots & B(:,M) \end{bmatrix} \right) \\
&= \mathrm{trace}\left(\begin{bmatrix} A(1,:)B(:,1) & A(1,:)B(:,2) & \dots & A(1,:)B(:,M) \\ A(2,:)B(:,1) & A(2,:)B(:,2) & \dots & A(2,:)B(:,M) \\ & & \ddots & \\ A(M,:)B(:,1) & A(M,:)B(:,2) & \dots & A(M,:)B(:,M) \end{bmatrix} \right) \\
&= \sum_{m=1}^{M} A(m,:)B(:,m) = \mathrm{vec}(\boldsymbol{A}^T)^T \mathrm{vec}(\boldsymbol{B}),
\end{aligned}$$

using the preceding equality. This is the general property.
When $\boldsymbol{A}$ is symmetric (and $\boldsymbol{B}$ is square with the same size), then $\boldsymbol{A} = \boldsymbol{A}^T$, so $\mathrm{trace}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^T \mathrm{vec}(\boldsymbol{B})$.

Pr. 12. (sol/hsj51)
Show that the weighted Euclidean norm $\|\boldsymbol{x}\|_{\boldsymbol{W}}$ is a valid norm iff $\boldsymbol{W}$ is a positive definite matrix.
First we show sufficiency: if $\boldsymbol{W}$ is positive definite, then we show that $\|\boldsymbol{x}\|_{\boldsymbol{W}} = \sqrt{\boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}}$ is a valid norm.

- Clearly $\boldsymbol{x} = \boldsymbol{0} \Rightarrow \|\boldsymbol{x}\|_{\boldsymbol{W}} = 0$
- By definition of a positive definite matrix, $\|\boldsymbol{x}\|_{\boldsymbol{W}} = 0 = \sqrt{\boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}} \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
- $\|\alpha\boldsymbol{x}\|_{\boldsymbol{W}} = \sqrt{(\alpha^*)\boldsymbol{x}'\boldsymbol{W}(\alpha\boldsymbol{x})} = |\alpha| \sqrt{\boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}} = |\alpha| \, \|\boldsymbol{x}\|_{\boldsymbol{W}}$
- Because $\boldsymbol{W}$ is (Hermitian) positive definite, it has a unitary eigendecomposition of the form $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$ where the eigenvalues in $\boldsymbol{\Lambda}$ are all positive. Let $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}'$ so that $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{S} = \boldsymbol{S}'\boldsymbol{S}$. Then $\|\boldsymbol{x}\|_{\boldsymbol{W}} = \sqrt{\boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}} = \sqrt{\boldsymbol{x}'\boldsymbol{S}'\boldsymbol{S}\boldsymbol{x}} = \|\boldsymbol{S}\boldsymbol{x}\|$ so $\|\boldsymbol{x}+\boldsymbol{y}\|_{\boldsymbol{W}} = \|\boldsymbol{S}(\boldsymbol{x}+\boldsymbol{y})\| \leq \|\boldsymbol{S}\boldsymbol{x}\| + \|\boldsymbol{S}\boldsymbol{y}\| = \|\boldsymbol{x}\|_{\boldsymbol{W}} + \|\boldsymbol{y}\|_{\boldsymbol{W}}$

Now we show necessity. If $\|\boldsymbol{x}\|_{\boldsymbol{W}}$ is a valid norm, then for all $\boldsymbol{x}$ $0 \leq \|\boldsymbol{x}\|_{\boldsymbol{W}}^2 = \boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}$ and for $\boldsymbol{x} \neq \boldsymbol{0}$: $0 < \|\boldsymbol{x}\|_{\boldsymbol{W}}^2 = \boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}$. These are the two conditions for a (Hermitian) symmetric matrix to be positive definite.

Pr. 13. (sol/hsj52)
Using the compact SVD $\boldsymbol{A} = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r'$ we have

$$\operatorname{trace}((\boldsymbol{A}\boldsymbol{A}')^2) = \operatorname{trace}((\boldsymbol{U}_r\boldsymbol{\Sigma}_r^2\boldsymbol{U}_r')^2) = \operatorname{trace}(\boldsymbol{U}_r\boldsymbol{\Sigma}_r^4\boldsymbol{U}_r') = \operatorname{trace}(\boldsymbol{\Sigma}_r^4) = \sum_{k=1}^{r} \sigma_k^4 = \|\boldsymbol{A}\|_{S,4}^4 .$$

Pr. 14. (sol/hsj53)

(a) Solution 1.

$$\left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix}' \right\|_2 = \|\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{B}\boldsymbol{B}'\|_2 \leq \|\boldsymbol{A}\boldsymbol{A}'\|_2 + \|\boldsymbol{B}\boldsymbol{B}'\|_2 = \|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2$$

Solution 2. For any $\boldsymbol{s}, \boldsymbol{t} \neq \boldsymbol{0}$:

$$\begin{aligned} \left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \right\|_2^2 &\leq \frac{\left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{t} \end{bmatrix} \right\|_2^2}{\left\| \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{t} \end{bmatrix} \right\|_2^2} = \frac{\|\boldsymbol{A}\boldsymbol{s} + \boldsymbol{B}\boldsymbol{t}\|_2^2}{\|\boldsymbol{s}\|_2^2 + \|\boldsymbol{t}\|_2^2} \leq \frac{(\|\boldsymbol{A}\boldsymbol{s}\|_2 + \|\boldsymbol{B}\boldsymbol{t}\|_2)^2}{\|\boldsymbol{s}\|_2^2 + \|\boldsymbol{t}\|_2^2} \leq \frac{(\|\boldsymbol{A}\|_2 \|\boldsymbol{s}\|_2 + \|\boldsymbol{B}\|_2 \|\boldsymbol{t}\|_2)^2}{\|\boldsymbol{s}\|_2^2 + \|\boldsymbol{t}\|_2^2} \\ &= (\|\boldsymbol{A}\|_2 \cos\phi + \|\boldsymbol{B}\|_2 \sin\phi)^2 \leq \|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2 , \end{aligned}$$

where $\cos\phi = \frac{\|\boldsymbol{s}\|_2^2}{\|\boldsymbol{s}\|_2^2 + \|\boldsymbol{t}\|_2^2}$.

(b) For any $\boldsymbol{A} \in \mathbb{F}^{M\times N}$ with compact SVD $\boldsymbol{A} = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r'$, define $\boldsymbol{B} = \boldsymbol{u}_1 b \boldsymbol{v}'$ for any unit-vector $\boldsymbol{v} \in \mathbb{F}^K$.
Then $\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{U}_r \operatorname{Diag}\left(\sigma_1^2 + b^2, \sigma_2, \ldots, \sigma_r\right) \boldsymbol{U}_r'$, so $\|\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B}\|_2 = \sigma_1^2(\boldsymbol{A}) + b^2 = \|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2$.

(c) Using the first part and the invariance of the spectral norm to matrix transpose:

$$\left\| \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} \boldsymbol{A}' & \boldsymbol{B}' \end{bmatrix} \right\|_2 \leq \sqrt{\|\boldsymbol{A}'\|_2^2 + \|\boldsymbol{B}'\|_2^2} = \sqrt{\|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2}.$$
