CS/ECE/ME 532
Homework 5

1. the face emotion classification problem from HW 2. Design and compare the performances of the classifiers proposed in **a** and **b**, below. In each case, divide the dataset into 8 equal sized subsets (e.g., examples $1-16$, $17-32$, etc). Use 6 sets of the data to estimate $\boldsymbol{w}$ for each choice of the *regularization parameter*, select the best value for the regularization parameter by estimating the error on one of the two remaining sets of data, and finally use the $\boldsymbol{w}$ corresponding to the best value of the regularization parameter to predict the labels of the remaining "hold-out" set. Compute the number of mistakes made on this hold-out set and divide that number by 16 (the size of the set) to estimate the error rate. Repeat this process 56 times (for the 8×7 different choices of the sets used to select the regularization parameter and estimate the error rate) and average the error rates to obtain a final estimate.

a) Truncated SVD solution. Use the pseudo-inverse $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$, where $\boldsymbol{\Sigma}^{-1}$ is computed by inverting the k largest singular values and setting others to zero. Here, k is the regularization parameter and it takes values $k = 1, 2, \ldots, 9$; i.e., compute 9 different solutions, $\widehat{\boldsymbol{w}}_k$.

b) Regularized LS. Let $\widehat{\boldsymbol{w}}_\lambda = \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$, for the following values of the regularization parameter $\lambda = 0, 2^{-1}, 2^0, 2^1, 2^2, 2^3$, and 2^4. Show that $\widehat{\boldsymbol{w}}_\lambda$ can be computed using the SVD and use this fact in your code.

SOLUTION: Running the code below, we find an average error rate of `0.1116` for the truncated SVD and an average error rate of `0.0480` for regularized least-squares (RLS). So RLS appears to have the lowest error

```
load face_emotion_data
[~,m] = size(X);

error_RLS = zeros(8,7);  % error rate for regularized least squares
error_SVD = zeros(8,7);  % error rate for svd truncation

% SVD parameters to test
kvals = 1:9;
err_mini_SVD = zeros(numel(kvals),1);

% RLS parameters to test
lamvals = 2.^[-inf -1:4];
err_mini_RLS = zeros(numel(lamvals),1);

% LOOP OVER FIRST HOLD-OUT SET
for h = 1:8
   ih = 16*h-15:16*h;              % the set of hold-out indices
   it = setdiff(1:8*16,ih);        % the set of training indices
```

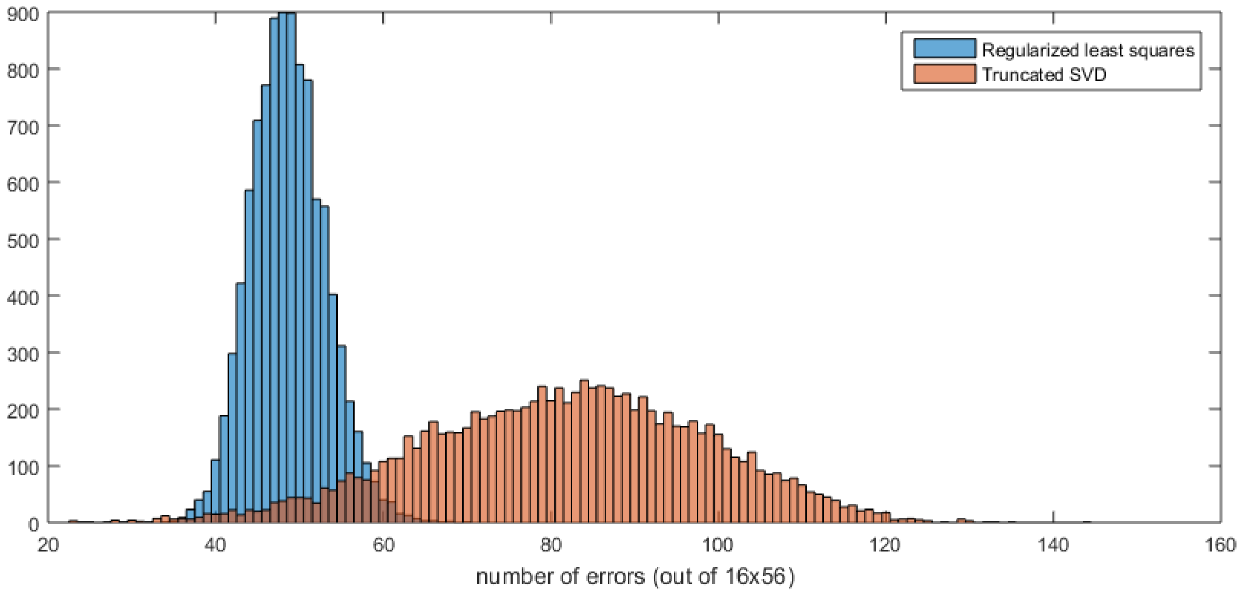
```
    % first holdout set and training set
    Xhh = X(ih,:); yhh = y(ih);
    Xtt = X(it,:); ytt = y(it);

    % LOOP OVER SECOND HOLD-OUT SET
    for j = 1:7
        jh = 16*j-15:16*j;        % inner set of hold-out indices
        jt = setdiff(1:7*16,jh);  % inner set of training indices

        % second holdout set and training set
        Xh = Xtt(jh,:); yh = ytt(jh);
        Xt = Xtt(jt,:); yt = ytt(jt);

        [U,S,V] = svd(Xt,'econ');  % u will be skinny; s and v square

        % TEST SVD TRUNCATION
        wt = zeros(m,numel(kvals));
        for i = 1:numel(kvals)
            k = kvals(i);
            % compute estimate by using only first k singular vectors
            wt(:,i) = V(:,1:k)*diag( 1./diag(S(1:k,1:k)) )*(U(:,1:k)'*y
            yp = sign(Xh*wt(:,i));   % error on 2nd holdout
            err_mini_SVD(i) = mean( yp ~= yh );
        end
        [~,ind] = min(err_mini_SVD); % choose best based on 2nd holdout
        ypp = sign(Xhh*wt(:,ind));   % compute error based on 1st holdo
        error_SVD(h,j) = mean( ypp ~= yhh );

        % TEST REGULARIZED LEAST-SQUARES
        wt = zeros(m,numel(lamvals));
        for i = 1:numel(lamvals)
            lambda = lamvals(i);
            % compute estimate by using regularization
            wt(:,i) = V*diag( diag(S)./(diag(S).^2 + lambda) )*(U'*yt);
            yp = sign(Xh*wt(:,i));   % error on 2nd holdout
            err_mini_RLS(i) = mean( yp ~= yh );
        end
        [~,ind] = min(err_mini_RLS); % choose best based on 2nd holdout
        ypp = sign(Xhh*wt(:,ind));   % compute error based on 1st holdo
        error_RLS(h,j) = mean( ypp ~= yhh );
    end
end

% compute average error rate over all trials
avg_err_rate_SVD = mean(error_SVD(:))
```

```
avg_err_rate_RLS = mean(error_RLS(:))
```

c) Use the original dataset to generate 3 new features for each face, as follows. Take the 3 new features to be random linear combination of the original 9 features. This can be done with the Matlab command `X*randn(9,3)` and augmenting the original matrix $\boldsymbol{X}$ with the resulting 3 columns. Will these new features be helpful for classification? Why or why not? Repeat the experiments in (a) and (b) above using the 12 features.

SOLUTION: The code is very similar to the code above so we omit it. This time, we add three random features (extra columns) to X that are linear combinations of the existing columns. The plot below shows the error distribution over 10,000 trials. The average error rates over all these trials is 0.0911 ± 0.0188 for the truncated SVD and 0.0543 ± 0.0050 for RLS. So RLS still outperforms SVD in this case.

It is not surprising that adding random features does not improve the classifier. The new random columns are linear combinations of the existing columns, so range(X) does not change. There is still variability however, because of (1) the nonlinearity inherent in our classification method (the `sign` function), (2) the fact that $\|w\|_2$ can actually be reduced when we add more columns (even though $\|Xw - y\|_2$ can't be improved), so the λ values we test get shifted and the weighting changes.



2. Many sensing and imaging systems produce signals that may be slightly distorted or blurred (e.g., an out-of-focus camera). In such situations, algorithms are needed to deblur the data to obtain a more accurate estimate of the true signal. The Matlab code `blurring.m` generates a random signal and a blurred and noisy version of it, similar to the example shown below. The code simulates this equation:

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{w} \;+\; \boldsymbol{e}\;,$$

where $\boldsymbol{y}$ is the blurred and noisy signal, $\boldsymbol{X}$ is a matrix that performs the blurring operation, $\boldsymbol{w}$ is the true signal, and $\boldsymbol{e}$ is a vector of errors/noise. The goal is to estimate $\boldsymbol{w}$ using $\boldsymbol{y}$ and $\boldsymbol{X}$.

a) Implement the standard LS, truncated SVD, and regularized LS methods for this problem.

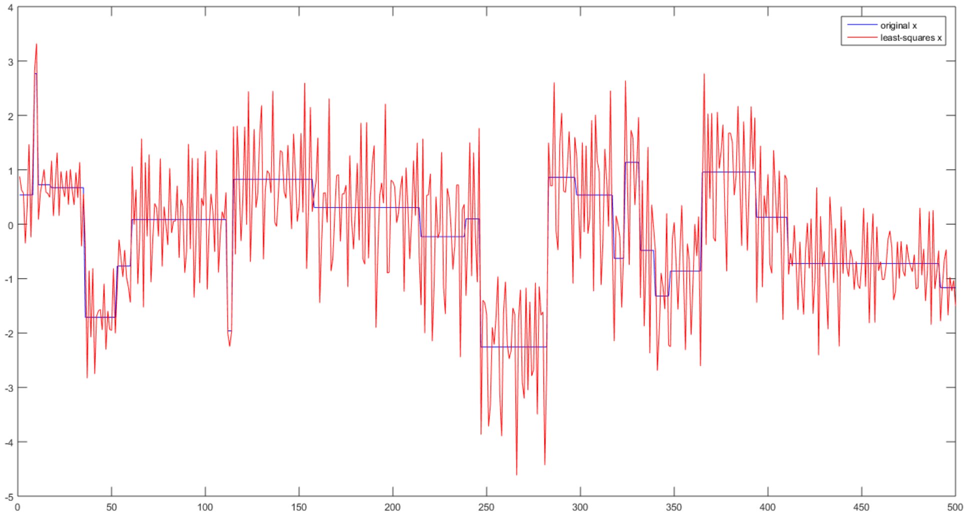
SOLUTION: One easy way to implement each estimator is to start with $\boldsymbol{A}$ and $\boldsymbol{b}$ and compute the following for example:

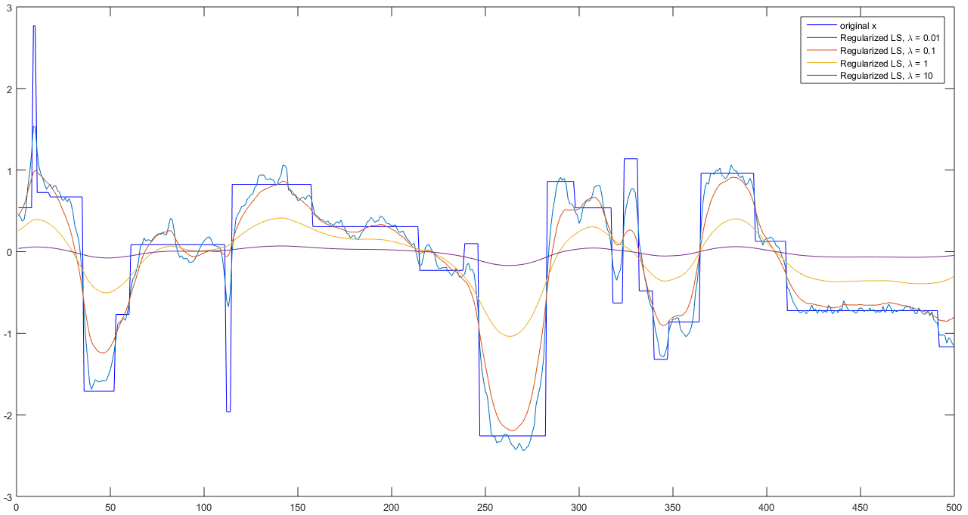
```
[U,S,V] = svd(X,'econ');

w_LS = V*diag( diag(S).^(-1) )*(U'*y);

% (loop over k values)
w_SVD = V(:,1:k)*diag( diag(S(1:k,1:k)).^(-1) )*(U(:,1:k)'*y);

% (loop over L (lambda) values)
w_RLS = V*diag( diag(S).*( (diag(S).^2+L).^(-1) ) )*(U'*y);
```

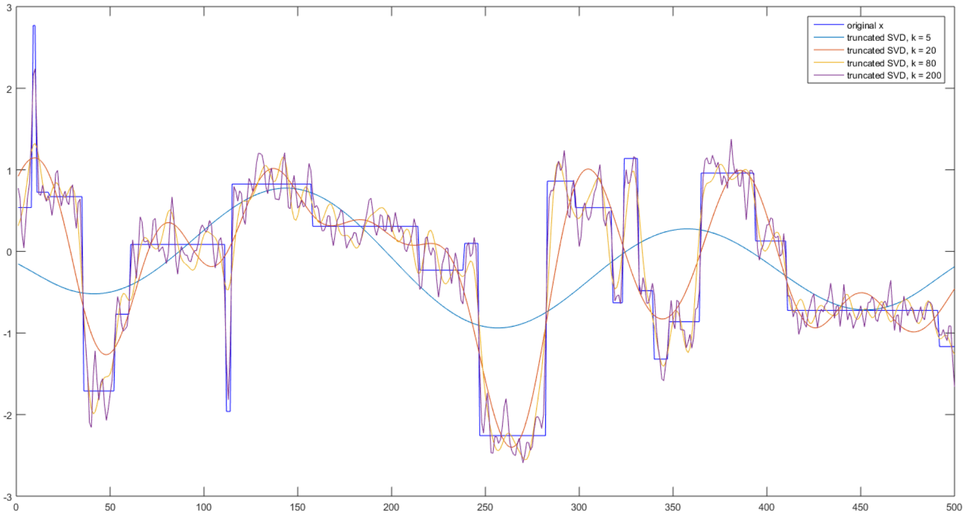
Note that I used `diag( diag(S).^(-1) )` rather than `inv(S)` since we know S is diagonal—it's more efficient to take the diagonal elements, invert them individually, and then rebuild a diagonal matrix.

b) Experiment with different averaging functions (i.e., different values of k in the code) and with different noise levels (σ in the code). How do the blurring and noise level affect the value of the regularization parameters that produce the best estimates?

SOLUTION: See the following page for sample estimators with default noise:




3
2
1
0
-1
-2
-3
0
50
100
150
200
250
300
350
400
450
500
original x
Regularized LS, λ = 0.01
Regularized LS, λ = 0.1
Regularized LS, λ = 1
Regularized LS, λ = 10


3
2
1
0
-1
-2
-3
0
50
100
150
200
250
300
350
400
450
500
original x
truncated SVD, k = 5
truncated SVD, k = 20
truncated SVD, k = 80
truncated SVD, k = 200

A great deal can be said about these estimators. Here are some examples of observations.

- The LS estimator is very noisy. This is because every singular value of $\mathbf{X}$ gets inverted. The smallest singular values are quite small, so when they get inverted they become large. These large values multiply the noisy $\boldsymbol{y}$ vector and ultimately amplify the noise. Even though the noise on $\boldsymbol{y}$ is relatively small, it gets amplified and this is evident in the estimate $\widehat{\boldsymbol{w}}$.
- The RLS estimator adds a regularization *before* inverting. So rather than computing $1/\sigma$ for each singular value of $\mathbf{X}$, we compute $\sigma/(\sigma^2+\lambda)$, which is smaller than $1/\sigma$. This moves σ away from zero before inverting, which causes the noise to be amplified less. However, as λ gets larger the inverse is driven to zero. In the plot, we see that the limit $\lambda \to 0$ recovers the noisy LS estimate, and as λ gets larger the estimates are simultaneously de-noised and driven towards zero.
- The SVD estimator simply truncates the small singular values rather than inverting them. So as in the RLS case we are reducing the effect of noise, but without driving the estimate to zero. The result is akin to a Fourier series approximation, except the basis used is not the Fourier basis—it is a basis determined by the singular vectors. An important twist is that there is an additional trade-off: including more singular values increases the quality of the estimate, but also increases susceptibility to noise.

3. Landweber convergence – REQUIRED FOR GRADUATE STUDENTS ONLY. Consider the Landweber iteration for solving a standard least-squares problem with $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, and $\boldsymbol{X}$ has full column rank. Recall that this iteration begins with some initial $\boldsymbol{w}_0$ and then:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \tau \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y}) \qquad \text{for } k = 0, 1, \dots \tag{1}$$

a) We expect the algorithm to converge to $\boldsymbol{w}_\star = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. Define the *error* as $\boldsymbol{e}_k := \boldsymbol{w}_k - \boldsymbol{w}_\star$. Show how to rewrite (1) in the form $\boldsymbol{e}_{k+1} = \boldsymbol{P}\boldsymbol{e}_k$. What is the matrix $\boldsymbol{P}$?

SOLUTION: Substitute $\boldsymbol{w}_k \mapsto \boldsymbol{e}_k + \boldsymbol{w}_\star$ and $\boldsymbol{w}_{k+1} \mapsto \boldsymbol{e}_{k+1} + \boldsymbol{w}_\star$ and obtain:

$$\boldsymbol{e}_{k+1} = \boldsymbol{e}_k - \tau\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{e}_k + \boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{y})$$

Then, using the fact that $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}_\star = \boldsymbol{X}^\top\boldsymbol{y}$, the expression simplifies to:

$$\boldsymbol{e}_{k+1} = (\boldsymbol{I} - \tau\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{e}_k$$

Therefore $\boldsymbol{P} = \boldsymbol{I} - \tau\boldsymbol{X}^\top\boldsymbol{X}$.

b) Define the *residual* $\boldsymbol{r}_k := \boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y}$. Show how to rewrite (1) in the form $\boldsymbol{r}_{k+1} = \boldsymbol{Q}\boldsymbol{r}_k$. What is the matrix $\boldsymbol{Q}$?

SOLUTION: Compute the residual at time $k+1$ and obtain:

$$\begin{aligned}
\boldsymbol{r}_{k+1} &= \boldsymbol{X}\boldsymbol{w}_{k+1} - \boldsymbol{y} \\
&= \boldsymbol{X}(\boldsymbol{w}_k - \tau\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y})) - \boldsymbol{y} \\
&= (\boldsymbol{I} - \tau\boldsymbol{X}\boldsymbol{X}^\top)(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y}) \\
&= (\boldsymbol{I} - \tau\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{r}_k
\end{aligned}$$

Therefore $\boldsymbol{Q} = \boldsymbol{I} - \tau \boldsymbol{X}\boldsymbol{X}^\top$.

c) Let $\{\sigma_i\}$ be the singular values of $\boldsymbol{X}$. Prove that when $0 < \tau < \frac{2}{\sigma_1^2}$, we have $\lim_{k\to\infty} \boldsymbol{e}_k = \mathbf{0}$. **Hint:** substitute the SVD of $\boldsymbol{X}$ into your expression for $\boldsymbol{P}$.

SOLUTION: Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{X}$ be the full SVD of $\boldsymbol{X}$. The expression for $\boldsymbol{e}_{k+1}$ becomes:

$$\boldsymbol{e}_{k+1} = (\boldsymbol{I} - \tau \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top)\boldsymbol{e}_k$$

Multiply by $\boldsymbol{V}^\top$ on both sides and define $\boldsymbol{V}^\top\boldsymbol{e}_k = \boldsymbol{q}_k$ and rewrite as:

$$\boldsymbol{q}_{k+1} = (\boldsymbol{I} - \tau\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})\boldsymbol{q}_k$$

Since $\boldsymbol{V}^\top$ is orthogonal, $\boldsymbol{e}_k \to \mathbf{0}$ if and only if $\boldsymbol{q}_k \to \mathbf{0}$. And this, in turn, is equivalent to each component of $\boldsymbol{q}_k$ going to zero separately (because $\boldsymbol{I} - \tau\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ is diagonal). Moreover, $\boldsymbol{X}$ has full column rank, so $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ is invertible—it is diagonal with entries $\sigma_1 \geq \cdots \geq \sigma_n > 0$. The equations corresponding to individual components of $\boldsymbol{q}_k$ look like: $q_{k+1} = (1 - \tau\sigma^2)q_k$. We have convergence to zero if and only if $|1 - \tau\sigma_i^2| < 1$ for all i. Rearranging, this is equivalent to: $0 < \tau < \frac{2}{\sigma_i^2}$. Since this must hold for all i, then because $\sigma_1 \geq \cdots \geq \sigma_n$, it must be the case that $0 < \tau < \frac{2}{\sigma_1^2}$.

d) Prove that if $\boldsymbol{X}$ is rank-deficient and $\boldsymbol{w}_0 = \mathbf{0}$, then the Landweber iteration converges to the minimum norm solution. **Hint:** redo part **a)** using $\boldsymbol{w}_\star = \boldsymbol{X}^\dagger\boldsymbol{y}$ and see how this affects part **c)**.

SOLUTION: The minimum-norm solution is $\boldsymbol{w}_\star = \boldsymbol{X}^\dagger\boldsymbol{y}$, which also satisfies $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}_\star = \boldsymbol{X}^\top\boldsymbol{y}$. So the proof from part **a)** still holds here. Repeating the arguments from the proof of **c)**, we conclude as before that

$$\boldsymbol{q}_{k+1} = (\boldsymbol{I} - \tau\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})\boldsymbol{q}_k$$

The only difference is that this time, $\boldsymbol{X}$ is rank-deficient. Therefore, $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ will have diagonal components $\sigma_1, \ldots, \sigma_r, 0, \ldots, 0$. For the first r components of $\boldsymbol{q}_k$, the same result holds as before; we have convergence to zero if and only if $0 < \tau < \frac{2}{\sigma_1^2}$. For the other components, we have: $q_{k+1} = q_k = \cdots = q_0$. Therefore, the only way these remaining components can go to zero is if they are always zero. Taking a closer look at the last $n - r$ components of $\boldsymbol{q}_0$,

$$\begin{aligned}
\begin{bmatrix}\mathbf{0} & \boldsymbol{I}\end{bmatrix}\boldsymbol{q}_0 &= \begin{bmatrix}\mathbf{0} & \boldsymbol{I}\end{bmatrix}\boldsymbol{V}^\top\left(\boldsymbol{w}_0 - \boldsymbol{X}^\dagger\boldsymbol{y}\right)\\
&= \begin{bmatrix}\mathbf{0} & \boldsymbol{I}\end{bmatrix}\boldsymbol{V}^\top\left(\boldsymbol{w}_0 - \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^\top\boldsymbol{y}\right)\\
&= \boldsymbol{V}_2^\top\left(\boldsymbol{w}_0 - \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^\top\boldsymbol{y}\right)\\
&= \boldsymbol{V}_2^\top\boldsymbol{w}_0
\end{aligned}$$

where we used in the last step that $\boldsymbol{V}_2^\top\boldsymbol{V}_1 = \mathbf{0}$. Since $\boldsymbol{w}_0 = 0$ by assumption, we have that $\begin{bmatrix}\mathbf{0} & \boldsymbol{I}\end{bmatrix}\boldsymbol{q}_0 = \mathbf{0}$. In other words, the components of $\boldsymbol{q}_k$ that don't shrink (τ has no effect on them) are zero anyway! This shows that $\boldsymbol{w}_k \to \boldsymbol{X}^\dagger\boldsymbol{y}$, as required. Note that $\boldsymbol{w}_0 = \mathbf{0}$ is a stronger condition than what's actually required for the convergence result. We will have $\boldsymbol{w}_k \to \boldsymbol{X}^\dagger\boldsymbol{y}$ if and only if $\boldsymbol{w}_0 = \boldsymbol{V}_1\boldsymbol{w}$ for some $\boldsymbol{w}$. Put another way, the result will be true if and only if $\boldsymbol{w}_0 \in \text{null}(\boldsymbol{X})^\perp$. In particular, it's true when $\boldsymbol{w}_0 = \mathbf{0}$.