

1(a) (9 points)

$$X^T X = (U D V^T)^T (U D V^T) \quad [\because X = U D V^T]$$

$$= V D^T U^T U D V^T$$

$$= V D^T D V^T \quad [\because U^T U = I_p]$$

$$= V \tilde{D} V^T$$

$$\therefore X^T X = V \tilde{D} V^T \quad [\text{eigendecomposition of } X^T X]$$

1(b) (9 points)

$$X^T X + \lambda I = V \tilde{D} V^T + \lambda I_d$$

$$= V \tilde{D} V^T + \lambda V V^T \quad [\text{replacing } I_d = V V^T]$$

$$= (V \tilde{D} + \lambda V) V^T$$

$$= V (\tilde{D} + \lambda I_d) V^T$$

$$\therefore X^T X + \lambda I = V (\tilde{D} + \lambda I_d) V^T \quad [\text{eigendecomposition of } X^T X + \lambda I]$$

1(c) (9 points)

$$(X^T X + \lambda I)^{-1} = (V (\tilde{D} + \lambda I_d) V^T)^{-1}$$

$$= V (\tilde{D} + \lambda I_d)^{-1} V^T$$

So, eigendecomposition of $(X^T X + \lambda I)^{-1}$ is $V (\tilde{D} + \lambda I_d)^{-1} V^T$

1(d) (8 points)

$$\begin{aligned}\beta^\lambda &= (X^T X + \lambda I)^{-1} X^T Y \\ &= V(D^T D + \lambda I_d)^{-1} V^T V D^T U^T Y \\ &= V(D^T D + \lambda I_d)^{-1} D^T U^T Y \quad [\because V^T V = I] \\ &= V M U^T Y\end{aligned}$$

where $M = (D^T D + \lambda I_d)^{-1} D^T \in \mathbb{R}^{d \times d}$ is diagonal

where $m_{jj} = \frac{d_j}{d_j^2 + \lambda}$ for $j=1, \dots, d$

1(e)

$$\begin{aligned}\hat{\beta}^{(\lambda)} &= \arg \min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \|y_i - x_i \beta\|_2^2 + \lambda \|\beta\|_2^2 \\ &= \arg \min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - x_i \beta)^T (y_i - x_i \beta) + \lambda \beta^T \beta \quad \text{--- (1)}\end{aligned}$$

Let f be the objective function at (1). Since, f is convex and differentiable, we solve $\nabla f(\beta^{(\lambda)}) = 0$ for β^λ , which is

$$\begin{aligned}-2X^T Y + 2X^T X \beta^{(\lambda)} + 2\lambda \beta^{(\lambda)} &= 0 \\ \Rightarrow (X^T X + \lambda I_d) \beta^{(\lambda)} &= X^T Y \\ \Rightarrow \beta^{(\lambda)} &= (X^T X + \lambda I_d)^{-1} X^T Y\end{aligned}$$

\therefore So, $\beta^{(\lambda)}$ is the minimizer at the equation (1) which is known as ridge regression [Shown]

1(f) (8 points)

the mean squared error (MSE) of the ridge estimator is

$$\text{MSE}(\hat{\beta}_\lambda | X) = E[\|\hat{\beta}_\lambda - \beta\|^2 | X]$$

$$= \text{trace}(\text{Var}[\hat{\beta}_\lambda | X]) + \|\text{bias}(\hat{\beta}_\lambda | X)\|^2$$

The OLS estimator has zero bias (since, $\lambda=0$), so its MSE is

$$\text{MSE}(\hat{\beta} | X) = E[\|\hat{\beta} - \beta\|^2 | X]$$

$$= \text{trace}(\text{Var}[\hat{\beta} | X])$$

So, training MSE is minimized with $\lambda=0$

By the Gauss-Markov theorem, the OLS estimator has the lowest variance (and the lowest MSE) among the estimators that are unbiased, there exists a biased estimator (a ridge estimator) whose MSE is lower than that of OLS.

$X^T X$ is a real symmetric matrix with real eigenvalues.

It is also a positive semidefinite matrix, so all eigenvalues are nonnegative. Let us denote these real nonnegative eigenvalues by $\{\nu_i\}_{i=1}^d$.

Now, any eigenvector ν_i of $X^T X$ corresponding to eigenvalue ν_i is also an eigenvector of $(X^T X + \lambda I)$ with eigenvalue $(\nu_i + \lambda)$. Since $(X^T X + \lambda I)\nu_i = X^T X \nu_i + \lambda \nu_i = (\nu_i + \lambda)\nu_i$

Since, $\lambda > 0$, all eigenvalues of $(X^T X + \lambda I)$ are positive, so it is a full rank matrix and invertible.

Apart from that, if β_j 's are unconstrained, they can explode (can grow large and can cause overfitting).

Hence, they are susceptible to very high variance. To control variance, we need to regularize the coefficients. Here, λ works as a shrinkage parameter, λ controls the size of the coefficients and amount of regularization.

1(a) (9 points)

$$\beta^\lambda = (X^T X + \lambda I)^{-1} X^T Y \quad \text{--- (1)}$$

Solving for β^λ using the closed-form solution in (1) is less efficient. If we wish to compute $\beta^{(\lambda)}$ for multiple values of λ , this becomes computationally expensive and inefficient, especially when d is large. Since, we will want to select the best value for λ , it is more efficient to take advantage of the singular value decomposition. That is why it is useful to construct β_λ in terms of the eigen decomposition used in (d).

Cross Validating Polynomial Ridge Regression (50 Total Points)

Here we will explore cross validation to pick two hyperparameters (the order, m , and penalty, λ) to polynomial ridge regression:

$$\operatorname{argmin}_{\beta} \sum_{i=1}^N \left(\left(\sum_{j=0}^m x_i^j \beta_j \right) - y_i \right)^2 + \lambda \|\beta\|_2^2,$$

where $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$ are the 1d input and output, respectively.

Let's import additional packages of use.

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import time
from sklearn import preprocessing
```

Generate data.

Generate N data samples of input (1d) features X and outputs Y , which shall depend non-linearly on X .

```
d = 1
N = 100

def f(x):
    return np.exp(0.5*x)+2.0*x*np.sin(4.0*x)

def make_data(N, sigma=0.5):
    X = 5.0*np.random.rand(N, 1)-2.5
    Y = f(X) + sigma*np.random.randn(N,1)
    return X, Y
```

Visualize the data.

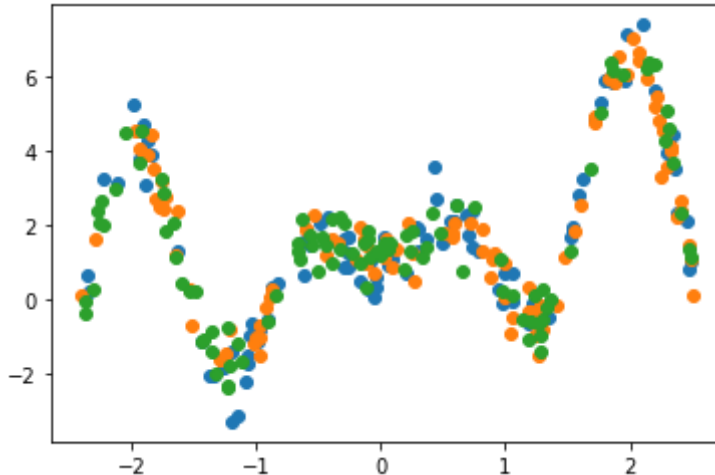
```
#np.random.seed(321)
np.random.seed(786)

X, Y = make_data(N)
plt.scatter(X, Y)
```

```
X, Y = make_data(N)
plt.scatter(X, Y)
```

```
X, Y = make_data(N)
plt.scatter(X, Y)
```

<matplotlib.collections.PathCollection at 0x7fbc7363ec50>



▼ Direct Linear Regression Cross Validation (17 Points)

Helper regression functions.

```
def standardize(X):
    return preprocessing.scale(X)

def poly_feats(X, order):
    """
    Args:
        X: N x 1 matrix of 1d input features
        order: max power of polynomial terms
    Returns:
        polyX: N x order+1 matrix of features 1, x, x**2, ..., x**order
    """
    return np.concatenate([X**j for j in range(order+1)], -1)

def linear_coefs(X, Y, ridge_penalty=0.0):
    """
    Args:
        X: N x d matrix of input features
        Y: N x 1 matrix (column vector) of output response
        ridge_penalty: scalar>0 penalty
    Returns:
        Beta: d x 1 matrix of linear coefficients
    """
```

```

"""
d = X.shape[1]
beta = np.linalg.solve(
    np.matmul(X.T, X)+ridge_penalty*np.eye(d),
    np.matmul(X.T, Y))
return beta

```

(3 Points) First we shall split our dataset into folds.

```

def fold_indices(N, K):
    """
    Return an array of elements where each element indicates what hold out fold
    the corresponding instance is in.
    Args:
        N: integer, number of training instance
        K: integer, number of folds
    Returns:
        folds: N length numpy array where folds[i] \in {1, ..., K} is the fold that
        instance i is in
    """
    # TODO (Hint: should be a single line)
    folds = None
    folds = np.arange(N) % K
    return folds

```

(10 Points) Implement model selection with cross validation.

#remove later

```

def cv_model_select(X, Y, lambdas, poly_orders, folds, K=None, plot=False,
                    figsize=(14, 12)):
    """
    Perform model selection to select the order of the polynomial and the l2 ridge
    penalty to use for regression.
    Args:
        X: N x d matrix of input features
        Y: N x 1 matrix (column vector) of output response
        lambdas: nlambdas length array of ridge regression penalties to select from
        poly_orders: npoly_order length array of polynomial orders to select from
        folds: N length numpy array where folds[i] \in {1, ..., K} is the fold that
        instance i is in
        K: integer, number of folds (if not given it is computed from given folds)
        plot: boolean, visualize models for folds
        figsize: tuple of figure size to plot
    Returns:
        mli: integer, where lambdas[mli] is selected lambda
        mpoi: integer, where poly_orders[mpoi] is selected polynomial order
    """

```

```

errs: nlambda x npoly_order matrix of average error using respective models
      across folds. i.e. errs[i, j] uses lambdas[i] ridge penalty and an order
      of poly_orders[j] to get the average test error across the folds.
"""
if K is None:
    K = np.max(folds)+1
gridX = np.reshape(np.linspace(-2.5, 2.5, 100), (100, 1))
N = X.shape[0]
nlambda = len(lambdas)
npoly_order = len(poly_orders)
errs = np.zeros((nlambda, npoly_order))

# {for k in range(K): | for poi in range(npoly_order): | for li in range(nlambda):}
# {for k in range(K): | for poi in range(npoly_order): | for li in range(nlambda):}
# {for k in range(K): | for poi in range(npoly_order): | for li in range(nlambda):}
for poi in range(npoly_order):
    for li in range(nlambda):
        MSE_CV = []
        for k in range(K):
            X_hold_out = []
            X_train = []
            Y_hold_out = []
            Y_train = []
            for index in range(N):
                if folds[index]==k:
                    X_hold_out.append(X[index])
                    Y_hold_out.append(Y[index])
                else:
                    X_train.append(X[index])
                    Y_train.append(Y[index])

            X_hold_out = np.array(X_hold_out)
            X_train = np.array(X_train)
            Y_hold_out = np.array(Y_hold_out)
            Y_train = np.array(Y_train)

            X_poly = poly_feats(X_train, poly_orders[poi])
            beta_poly = linear_coefs(X_poly, Y_train, lambdas[li])
            #print(X_hold_out.shape, beta_poly.shape)

            X_hold_out_poly = poly_feats(X_hold_out, poly_orders[poi])
            Y_hat = np.matmul(X_hold_out_poly, beta_poly)
            MSE = (Y_hold_out-Y_hat)**2
            MSE = np.sum(MSE, axis=0)/Y_hat.shape[0]
            MSE_CV.append(MSE)

        beta_notk = None # TODO: coefs for kth fold
        beta_notk = beta_poly
        if plot:
            plt.subplot(nlambda, npoly_order, li*npoly_order+poi+1)
            plt.title('l: {}, o: {}'.format(lambdas[li], poly_orders[poi]))

```



```

gridX_poly = poly_feats(gridX, poly_orders[poi])
gridY = np.matmul(gridX_poly, beta_notk) # predict Y here
plt.plot(gridX, gridY) # TODO: Predictions on gridX
trainX = X_train
trainY = Y_train
plt.scatter(trainX, trainY, marker='+') # Plot training from fold
testX = X_hold_out
testY = Y_hold_out
plt.scatter(testX, testY, marker='+') # Plot testing from fold
plt.ylim((np.min(Y), np.max(Y)))
MSE_CV = np.array(MSE_CV)
errs[li, poi] = np.mean(MSE_CV)

errs = errs/N
mli, mpoi = np.unravel_index(np.argmin(errs), (nlambdas, npoly_order))
return mli, mpoi, errs

```

Run on the folds.

```

lambdas = [2**j for j in range(-10, 4, 1)]
orders = [j for j in range(1, 8)]
X_standardized = standardize(X)
mli, mpoi, errs = cv_model_select(
    X_standardized, Y, lambdas, orders, fold_indices(N, 5), 5, plot=True)

```

https://colab.research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=yhl78EfwrnGc&printMode=true

https://colab.research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=yhl78EfwrnGc&printMode=true

https://colab.research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=yhl78EfwrnGc&printMode=true

https://colab.research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=yhI78EfwrnGc&printMode=true

research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=vhl78EfwrnGc&printMode=true

https://colab.research.google.com/drive/1TPWBETA_efwLKrrjM4YfFbhiscC_I-TK#scrollTo=yhl78EfwrnGc&printMode=true

```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib

```

Plot errors.

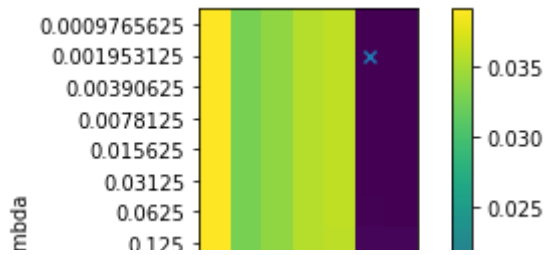
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib

plt.imshow(errs, vmax=np.min([10.0*np.min(errs), np.max(errs)]))
plt.xticks(ticks=range(len(orders)), labels=orders)
plt.xlabel('polynomial order')
plt.yticks(ticks=range(len(lambdas)), labels=lambdas)
plt.ylabel('lambda')
plt.colorbar()
plt.scatter(mpoi, mli, marker='x')

```


<matplotlib.collections.PathCollection at 0x7fbc7301ba50>



(3 Points) Return (and plot) a model using the selected hyper-parameters.

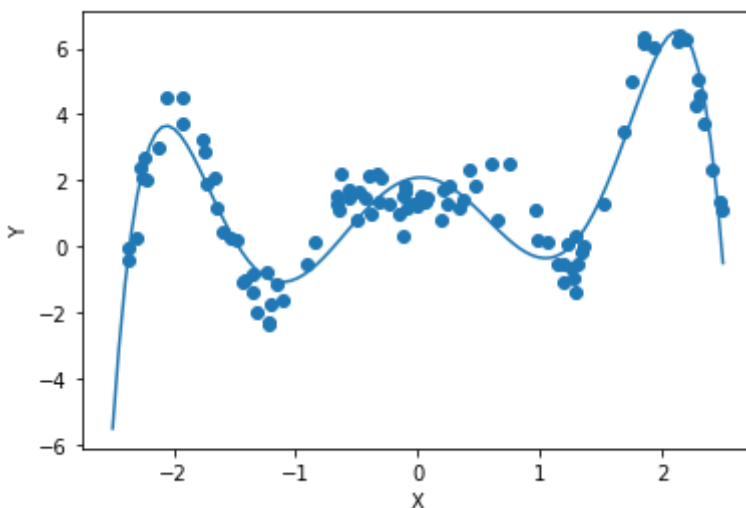
λ 0.015

```
# TODO: return a model corresponding to selected hyperparameters
gridX = np.reshape(np.linspace(-2.5, 2.5, 100), newshape=(100, 1))
```

```
X_poly = poly_feats(X, orders[mpoi])
beta_poly = linear_coefs(X_poly, Y, lambdas[mli])
gridX_poly = poly_feats(gridX, orders[mpoi])
Y_poly = np.matmul(gridX_poly, beta_poly)
```

```
est_gridY = None # TODO: estimate using this model on gridX
est_gridY = Y_poly
plt.plot(gridX, est_gridY)
plt.scatter(X, Y)
plt.xlabel('X')
plt.ylabel('Y')
```

Text(0, 0.5, 'Y')



▼ Estimated Generalization Error (10 Points)

Next we will compare the cross validation estimate of test error, with the actually test error using unseen data.

(4 points) Use the `cv_model_select` estimate the test error for 14 th order polynomial regression with ordinary least squares regression (no l2 penalty), and for a 7 th order polynomial regression with ridge regression (with $\lambda=1.25$) with 5 fold cross validation

#remove later

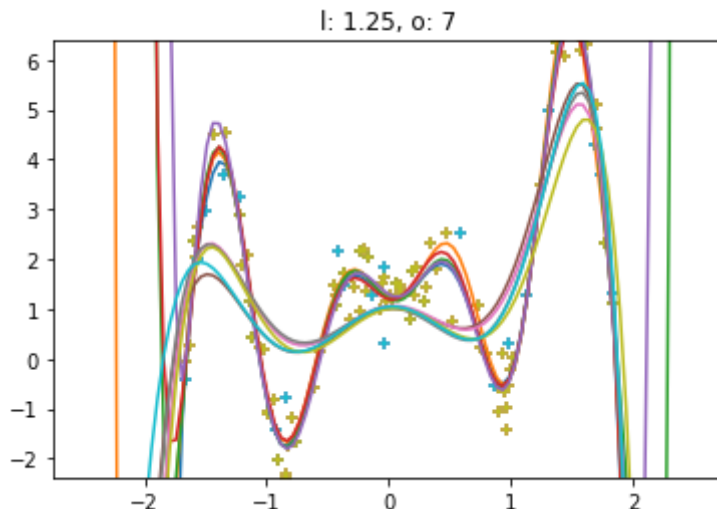
```
X_standardized = standardize(X)

mli, mpoi, errs = cv_model_select(
    # TODO,
    X_standardized, Y, [0.0], [14], fold_indices(N, 5), K=5,
    plot=True, figsize=(3,3))
cverr_000_14 = errs[mli, mpoi] # TODO error for 14th order polynomial OLS

mli, mpoi, errs = cv_model_select(
    # TODO,
    X_standardized, Y, [1.25], [7], fold_indices(N, 5), K=5,
    plot=True, figsize=(3,3))
cverr_125_7 = errs[mli, mpoi]# TODO error for 7th order polynomial with 1.25 ridge pe

print('CV Error for 14th order polynomial with no ridge penalty: {}'.format(
    cverr_000_14))
print('CV Error for 7th order polynomial with 1.25 ridge penalty: {}'.format(
    cverr_125_7))
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:68: Matplotlib
CV Error for 14th order polynomial with no ridge penalty: 0.002863182244478
CV Error for 7th order polynomial with 1.25 ridge penalty: 0.01823647653008
```



(4 Points) Compare the estimated CV errors to using 1000 new instances as a test set.

```
Xtest, Ytest = make_data(1000)

Xtest_poly = poly_feats(Xtest, 14)
X_poly = poly_feats(X, 14)

beta_poly = linear_coefs(X_poly, Y, 0.0)
Y_hat = np.matmul(Xtest_poly, beta_poly)
err_000_14 = None # TODO: Error using Xtest, and Ytest with 14th order poly OLS
err_000_14 = np.sum((Ytest-Y_hat)**2, axis=0)/1000

poly_X_7 = poly_feats(X, 7)
Xtest_poly = poly_feats(Xtest, 7)
beta_poly = linear_coefs(poly_X_7, Y, 1.25)
Y_hat = np.matmul(Xtest_poly, beta_poly)
err_125_7 = None # TODO: Error using Xtest, and Ytest w 7th order and 1.25 pen.
err_125_7 = np.sum((Ytest-Y_hat)**2, axis=0)/1000

print('Test Error for 14th order polynomial with no ridge penalty: {} (CV: {})\n'
      'Test Error for 7th order polynomial with 1.25 ridge penalty: {} (CV: {})'
      .format(err_000_14, cverr_000_14, err_125_7, cverr_125_7))

Test Error for 14th order polynomial with no ridge penalty: [0.26117134] (CV: 0.1)
Test Error for 7th order polynomial with 1.25 ridge penalty: [0.97018458] (CV: 0.1)
```

(2 points) What are some takeaways from the above?

TODO: respond The ridge problem penalizes the large regression coefficients, and the larger the parameter lambda, the larger the penalty. That is why the beta's was shrank too much in the second model and made bad predictions for test data.

▼ Faster Linear Regression Cross Validation (13 Points)

Next, we shall use the derived coefficients in terms of SVD to speed up cross validation.

(5 Points) Use the SVD of X to compute the ridge regression linear coefficients.

```
#remove later

def svd2beta(U, S, Vt, SUtY=None, Y=None, ridge_penalty=0.0):
    """
```

Produce the beta coefficients from SVD decomposition.

Args:

- U: N x d unitary matrix (from U, S, Vt = svd(X))
- S: d length array of singular values
- Vt: d x d unitary matrix (from U, S, Vt = svd(X))
- SUtY: d x 1 vector of cached product of $\text{diag}(S) U^T Y$
- Y: N x 1 vector of training responses (used if SUtY is None)
- ridge_penalty: real, lambda penalty to ridge regression

Returns:

- beta: d x 1 vector of coefficients for ridge regression

"""

```
D = np.diag(S)
p = D.shape[0]
if SUtY is None:
    SUtY = None # TODO
    SUtY = D.T @ U.T @ Y
M = (D.T @ D) + (ridge_penalty * np.eye(p))
M = np.linalg.inv(M)
beta = ( Vt.T @ M @ SUtY )
return beta # TODO
```

(3 Points) Compare the beta coefficients found using SVD vs. directly

```
def svd_linear_coefs(X_poly, Y, ridge_penalty):
    U, s, VT = np.linalg.svd(X_poly, full_matrices=False)
    beta_svd = svd2beta(U = U, S = s, Vt = VT, SUtY = None, Y = Y, ridge_penalty = ridge_penalty)
    return beta_svd
```

Hint: make sure to set full_matrices correctly in np.linalg.svd

```
X_poly = poly_feats(X, orders[mpoi])
beta_poly = linear_coefs(X_poly, Y, lambdas[mli])
```

```
U, s, VT = np.linalg.svd(X_poly, full_matrices=False)
beta_svd = None # TODO
beta_svd = svd2beta(U = U, S = s, Vt = VT, SUtY = None, Y = Y, ridge_penalty = lambdas[mli])
```

```
beta_direct = None # TODO
beta_direct = beta_poly
print(np.concatenate([beta_svd, beta_direct], -1))
print('Max differences in betas {}'.format(np.max(np.abs(beta_direct-beta_svd))))
```

```
[[1.45431495 1.45431495]
 [0.48180842 0.48180842]]
Max differences in betas 0.0
```

(5 Points) Implement cross validation, but using SVD for speed. (*Hint: think carefully about the order in which you compute things in order to speed up computation.*)


```

def cv_svd_model_select(X, Y, lambdas, poly_orders, folds, K=None, plot=False,
                        figsize=(14, 12)):
    """
    Perform model selection to select the order of the polynomial and the l2 ridge
    penalty to use for regression.
    *Computes beta coefficients using svd for speed.*
    Args:
        X: N x d matrix of input features
        Y: N x 1 matrix (column vector) of output response
        lambdas: n lambdas length array of ridge regression penalties to select from
        poly_orders: npoly_order length array of polynomial orders to select from
        folds: N length numpy array where folds[i] \in {1, ..., K} is the fold that
            instance i is in
        K: integer, number of folds (if not given it is computed from given folds)
        plot: boolean, visualize models for folds
        figsize: tuple of figure size to plot
    Returns:
        mli: integer, where lambdas[mli] is selected lambda
        mpoi: integer, where poly_orders[mpoi] is selected poly nomial order
        errs: n lambdas x npoly_order matrix of avarage error using respective models
            across folds. i.e. errs[i, j] uses lambdas[i] ridge penalty and an order
            of poly_orders[j] to get the average test error accross the folds.
    """
    if K is None:
        K = np.max(folds)+1
    gridX = np.reshape(np.linspace(-2.5, 2.5, 100), (100, 1))
    N = X.shape[0]
    nlambdas = len(lambdas)
    npoly_order = len(poly_orders)
    errs = np.zeros((nlambdas, npoly_order))

    # Hint: the order really matters here.
    #for li in range(nlambdas):
    #    for poi in range(npoly_order):
    for poi in range(npoly_order):
        for li in range(nlambdas):
            MSE_CV = []
            for k in range(K):
                X_hold_out = []
                X_train = []
                Y_hold_out = []
                Y_train = []
                for index in range(N):
                    if folds[index]==k:
                        X_hold_out.append(X[index])
                        Y_hold_out.append(Y[index])
                    else:
                        X_train.append(X[index])
                        Y_train.append(Y[index])

            X_hold_out = np.array(X_hold_out)

```

```

X_train = np.array(X_train)
Y_hold_out = np.array(Y_hold_out)
Y_train = np.array(Y_train)

X_poly = poly_feats(X_train, poly_orders[poi])
beta_svd = svd_linear_coefs(X_poly, Y_train, lambdas[li])
beta_poly = beta_svd

#linear_coefs(X_poly, Y_train, lambdas[li])
#print(X_hold_out.shape, beta_poly.shape)

X_hold_out_poly = poly_feats(X_hold_out, poly_orders[poi])
Y_hat = np.matmul(X_hold_out_poly, beta_poly)
MSE = (Y_hold_out-Y_hat)**2
MSE = np.sum(MSE, axis=0)/Y_hat.shape[0]
MSE_CV.append(MSE)

beta_notk = None # TODO: coefs for kth fold
beta_notk = beta_poly
if plot:
    plt.subplot(nlambdas, npoly_order, li*npoly_order+poi+1)
    plt.title('l: {}, o: {}'.format(lambdas[li], poly_orders[poi]))
    gridX_poly = poly_feats(gridX, poly_orders[poi])
    gridY = np.matmul(gridX_poly, beta_notk) # predict Y here
    plt.plot(gridX, gridY) # TODO: Predictions on gridX
    trainX = X_train
    trainY = Y_train
    plt.scatter(trainX, trainY, marker='+') # Plot training from fold
    testX = X_hold_out
    testY = Y_hold_out
    plt.scatter(testX, testY, marker='+') # Plot testing from fold
    plt.ylim((np.min(Y), np.max(Y)))
MSE_CV = np.array(MSE_CV)
errs[li, poi] = np.mean(MSE_CV)
errs = errs/N
mli, mpoi = np.unravel_index(np.argmin(errs), (nlambdas, npoly_order))
return mli, mpoi, errs

```

Let's time both the direct and SVD based CV model selection.

```

lambdas = [2**j for j in range(-10, 4, 1)]
orders = [j for j in range(1, 8)]
N_time = 1000
K_time = 5
finds= fold_indices(N_time, K_time)
X_time, Y_time = make_data(N_time)

start_dir = time.process_time()

```

```

mli_dir, mpoi_dir, errs_dir = cv_model_select(
    X_time, Y_time, lambdas, orders, finds, K_time, plot=False)
end_dir = time.process_time()

start_svd = time.process_time()
mli_svd, mpoi_svd, errs_svd = cv_svd_model_select(
    X_time, Y_time, lambdas, orders, finds, K_time, plot=False)
end_svd = time.process_time()

dir_time = end_dir-start_dir
svd_time = end_svd-start_svd
print('SVD is {} times faster! ({} vs. {})'.format(dir_time/svd_time, svd_time, dir_time))

SVD is 0.28727039191935827 times faster! (3.842880377 vs. 1.1039457520000013)

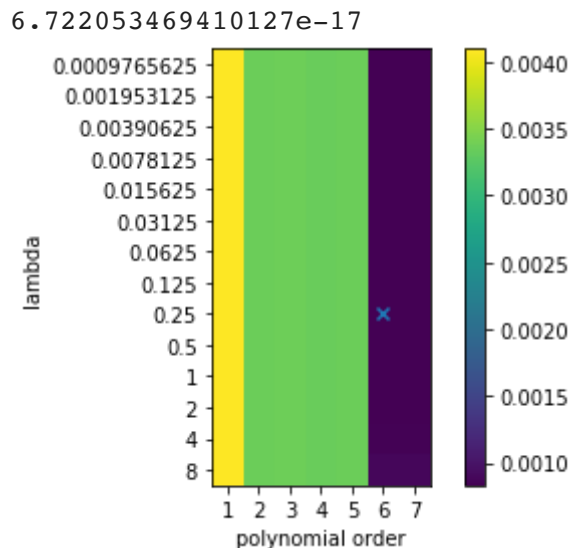
```

Check that the CV errors we got are the same.

```

plt.imshow(errs_svd, vmax=np.min([10.0*np.min(errs_svd), np.max(errs_svd)]))
plt.xticks(ticks=range(len(orders)), labels=orders)
plt.xlabel('polynomial order')
plt.yticks(ticks=range(len(lambdas)), labels=lambdas)
plt.ylabel('lambda')
plt.colorbar()
plt.scatter(mpoi_svd, mli_svd, marker='x')
np.max(np.abs(errs_dir-errs_svd))

```



✓ 0s completed at 8:17 PM

● ✕