CHECKPOINT Questions

- K-fold cross validation
  - K chunks of the data
  - Train how many times?
K-fold cross validation

- K chunks of the data
- Train how many times?
- K times!
- Cross-validation error is the average of the validation errors over K folds
Clever idea: Use many small validation sets without losing too much training data.

Still need to break off our test set like before. After doing so, break the training set into chunks.

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chunk1</td>
<td>Chunk2</td>
</tr>
</tbody>
</table>

For a given model complexity, train it times. Each time use all but one chunk and use that left out chunk to determine the validation error.

80-20 split. Take the 80% split and do a k-fold cross-validation on it.
The process generally goes

```python
chunk_1, ..., chunk_k, test = split_data(dataset)
for each model complexity p:
    for i in [1, k]:
        model = train_model(model_p, chunks - i)
        val_err = error(model, chunk_i)
        avg_val_err = average val_err over chunks
        keep track of p with smallest avg_val_err
return model trained on train with best p + error(model, test)
```
- **K-fold cross validation**
  - K chunks of the data
  - Train how many times?
  - K times!
  - Cross-validation error is the average of the validation errors over K folds
  - E.g. K = 4 (split train data into 4)
    - Train on Chunk 1,2,3 and validate on 4 -> val1
    - Train on Chunk 1,3,4 and validate on 2 -> val2
    - Train on Chunk 1,2,4 and validate on 3 -> val3
    - Train on Chunk 2,3,4 and validate on 1 -> val4
    - Cross validation error - (val1 + val2 + val3 + val4)/4
  - For M models - M*K trainings need to happen!
Questions

- **Validation vs Test Set - When to use what??**
  - Training done on train data
  - Test error is an approximation of the true error
  - So why do we need the validation data set?
  - Validation data used to pick the right model

- **Generalization error**
  - Measures performance on unseen data
  - What is the unseen data here?
  - Why unseen data?
  - Shopping example
Q2 - Using the regularizer
- Impact of regularizer on bias/model-complexity?
- What happens when you cube the weights and sum them up?
- Same as what happens when you sum the weights!!
CHECKPOINT
Questions

- Q2 - Using the regularizer
  - Impact of regularizer on bias/model-complexity?
  - What happens when you cube the weights and sum them up?
  - Same as what happens when you sum the weights!!
  - What parameters minimize the sum of weights?
  - What parameters minimize the sum of squares of weights?
Notations

- **What? w-hat?**
  - Hat notation used for predicted parameters/coefficients

- **Weights vs coefficients vs parameters**

- **Input data vs features**

- **Regression vs Regularization**
  - Regularization is an objective function added on to regular regression that reduces over-fitting of the model!
  - Lambda needs to be picked - How?
How should we choose the best value of $\lambda$?

- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega})$ on the **training set**
- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega})$ on the **test set**
- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega})$ on the **validation set**
- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega}) + \lambda \|\hat{\omega}\|_2^2$ on the **training set**
- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega}) + \lambda \|\hat{\omega}\|_2^2$ on the **test set**
- Pick the $\lambda$ that has the smallest $RSS(\hat{\omega}) + \lambda \|\hat{\omega}\|_2^2$ on the **validation set**
- Pick the $\lambda$ that results in the smallest coefficients
- Pick the $\lambda$ that results in the largest coefficients
- None of the above
RECAP

- Ridge Regression
- Choosing lambda
- What does high/low lambda mean?
Choosing $\lambda$

For any particular setting of $\lambda$, use Ridge Regression objective

$$\hat{w}_{ridge} = \min_w RSS(w) + \lambda \|w_{1:D}\|_2^2$$

If $\lambda$ is too small, will overfit to training set. Too large, $\hat{w}_{ridge} = 0$.

How do we choose the right value of $\lambda$? We want the one that will do best on future data. This means we want to minimize error on the validation set.

Don’t need to minimize $RSS(w) + \lambda \|w_{1:D}\|_2^2$ on validation because you can’t overfit to the validation data (you never train on it).

Another argument is that it doesn’t make sense to compare those values for different settings of $\lambda$. They are in different “units” in some sense.
Choosing $\lambda$

The process for selecting $\lambda$ is exactly the same as we saw with using a validation set or using cross validation.

```
for $\lambda$ in $\lambda$s:
    Train a model using Gradient Descent
    \[
    \hat{\omega}_{\text{ridge}}(\lambda) = \min_{\omega} \text{RSS}_{\text{train}}(\omega) + \lambda \|1_1: D \|^2
    \]
    Compute validation error
    \[
    \text{validation\_error} = \text{RSS}_{\text{val}}(\hat{\omega}_{\text{ridge}}(\lambda))
    \]
    Track $\lambda$ with smallest validation\_error
    Return $\lambda^*$ & estimated future error $\text{RSS}_{\text{test}}(\hat{\omega}_{\text{ridge}}(\lambda^*))$
```

There is no fear of overfitting to validation set since you never trained on it! You can just worry about error when you aren’t worried about overfitting to the data.
What exactly is Gradient Descent??
What exactly is Gradient Descent??
- Moving in the direction of negative gradient
Gradient Descent

- What exactly is Gradient Descent??
- Moving in the direction of negative gradient
GRADIENT DESCENT

Consider the quadratic function in the previous figure. Which direction (left or right) does the gradient point to for a) a point on the left side of the global minimum and for b) a point on the right side of the global minimum.

a) Left, Right
b) Right, Right
c) Left, Left
d) Right, Left
Gradient Descent

- What exactly is Gradient Descent??
- Moving in the direction of negative gradient
- Gradient is a generalization of slope

![Gradient Descent Diagram](image.png)
Gradient Descent - 2 dimensions
If you have $K$ features and a linear regression model fit to these $K$ features - What will be the dimension of the gradient?
If you have $K$ features and a linear regression model fit to these $K$ features - What will be the dimension of the gradient?

- $K+1$
- Gradient is a vector (generalization of slope which is a scalar)
Gradient Descent - Non-convex functions
Feature Selection and All Subsets
Benefits

Why do we care about selecting features? Why not use them all?

Complexity
Models with too many features are more complex. Might overfit!

Interpretability
Can help us identify which features carry more information.

Efficiency
Imagine if we had MANY features (e.g. DNA). \( \hat{w} \) could have \( 10^{11} \) coefficients. Evaluating \( \hat{y} = \hat{w}^T h(x) \) would be very slow!

If \( \hat{w} \) is sparse, only need to look at the non-zero coefficients

\[
\hat{y} = \sum_{\hat{w}_j \neq 0} \hat{w}_j h_j(x)
\]
Might have many features to potentially use. Which are useful?

Lot size
Single Family
Year built
Last sold price
Last sale price/sqft
Finished sqft
Unfinished sqft
Finished basement sqft
# floors
Flooring types
Parking type
Parking amount
Cooling
Heating
Exterior materials
Roof type
Structure style

Dishwasher
Garbage disposal
Microwave
Range / Oven
Refrigerator
Washer
Dryer
Laundry location
Heating type
Jetted Tub
Deck
Fenced Yard
Lawn
Garden
Sprinkler System

...
How happy are you? What part of the brain controls happiness?
Best Model
Size 0

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

\[ \text{Note only: } y_i = \phi \]
Best Model
Size 1

Features
- # bathrooms
- # bedrooms
- sq.ft. living
- sq.ft lot
- floors
- year built
- year renovated
- waterfront

\[ RSS_{\text{train}}(\hat{w}) \]

# of features
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

RSS_{\text{train}}(\hat{\omega})

# of features

0 1
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$R^2_{\text{train}}(\hat{w})$
Best Model
Size 1

Features
- # bathrooms
- # bedrooms
- sq.ft. living
- sq.ft lot
- floors
- year built
- year renovated
- waterfront

For each feature:
- $R_{SS_{train}}(\hat{w})$ vs. # of features

|$w|$ $0$ $1$
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

\[ \text{RSS}_{\text{train}}(\hat{w}) \]

# of features
Best Model
Size 1

Features
- # bathrooms
- # bedrooms
- sq.ft. living
- sq.ft lot
- floors
- year built
- year renovated
- waterfront

\[ \text{RSS}_{\text{train}}(\hat{\omega}) \]

\[ 0 \quad 1 \]

# of features
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$R_{SS_{train}}(\hat{w})$ vs. # of features
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

\( R_{SS_{train}}(\hat{w}) \)

# of features

0 1
Best Model
Size 1

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$RSS_{\text{train}}(\hat{\theta})$ vs. # of features
Best Model
Size 2

Not necessarily nested!
Best Model – Size 1: sq.ft living
Best Model – Size 2:
  # bathrooms & # bedrooms

Features
  # bathrooms
  # bedrooms
  sq.ft. living
  sq.ft lot
  floors
  year built
  year renovated
  waterfront

\( RSS_{\text{train}}(\hat{\omega}) \)

# of features

0 1 2
Best Model
Size 3

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$\min_{W} \sum_{i} (y_i - \sum_{j} W(j)^T x_i)^2$
Best Model
Size 4

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$RSS_{train}(\hat{w})$

# of features
Best Model
Size 5

Features:
- # bathrooms
- # bedrooms
- sq.ft. living
- sq.ft lot
- floors
- year built
- year renovated
- waterfront

Graph:
- $R^2_{\text{train}}$ vs. # of features
- X-axis: # of features (0-5)
- Y-axis: $R^2_{\text{train}}$ (values range from 0 to 1)
- Data points represent the performance of the model with varying feature sets.
Best Model
Size 6

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront
Best Model
Size 7

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront

$RSS_{\text{train}}(\hat{w})$

# of features
Best Model
Size 8

Features
# bathrooms
# bedrooms
sq.ft. living
sq.ft lot
floors
year built
year renovated
waterfront
Choose Num Features?

Option 1
Assess on a validation set

Option 2
Cross validation

Option 3+
Other metrics for penalizing model complexity like BIC
Efficiency of All Subsets

How many models did we evaluate?
\[ \hat{y}_i = \epsilon_i \]
\[ \hat{y}_i = w_0 h_0(x) + \epsilon_i \]
\[ \hat{y}_i = w_1 h_1(x) + \epsilon_i \]

... 
\[ \hat{y}_i = w_0 h_0(x) + w_1 h_1(x) + \epsilon_i \]

... 
\[ \hat{y}_i = w_0 h_0(x) + w_1 h_1(x) + ... + w_D h_D(x) \epsilon_i \]

If evaluating all subsets of 8 features only took 5 seconds, then

- 16 features would take 21 minutes
- 32 features would take almost 3 years
- 100 features would take almost 7.5*10^{20} years
  - 50,000,000,000x longer than the age of the universe!
Knowing it’s impossible to find exact solution, approximate it!

**Forward stepwise**
Start from model with no features, iteratively add features as performance improves.

**Backward stepwise**
Start with a full model and iteratively remove features that are the least useful.

**Combining forward and backwards steps**
Do a forward greedy algorithm that eventually prunes features that are no longer as relevant

*And many many more!*
Example
Greedy
Algorithm

Start by selecting number of features $k$

$$S_0 \leftarrow \{\}$$
for $i \leftarrow 1..k$:

- Find feature $f_i$ not in $S_{i-1}$, that when combined with $S_{i-1}$, minimizes the loss the most.

$$S_i \leftarrow S_{i-1} \cup \{f_i\}$$
Return $S_k$

Called greedy because it makes choices that look best at the time.
GREEDY ALGORITHM

How many times would we need to train the model to follow the greedy forward procedure assuming there are N possible features and we stop the algorithm at K features?

a) O(K^2)
b) O(N^2)
c) O(NK)
d) O(NK^2)
e) O(N^2K)
Option 2
Regularization
Recap: Regularization

Before, we used the quality metric that minimized loss
\[ \hat{w} = \min_w L(w) \]

Change quality metric to balance loss with measure of overfitting
- \[ L(w) \] is the measure of fit
- \[ R(w) \] measures the magnitude of coefficients

\[ \hat{w} = \min_w L(w) + R(w) \]

How do we actually measure the magnitude of coefficients?
Recap: Magnitude

Come up with some number that summarizes the magnitude of the coefficients in $w$.

Sum?

$$R(w) = \sum_{j=0}^{D} w_j$$

Sum of absolute values?

$$R(w) = \sum_{j=0}^{D} |w_j| \triangleq \|w\|_1$$

L1 norm (come back Wed!)

Sum of squares?

$$R(w) = \sum_{j=0}^{D} w_j^2 \triangleq \|w\|_2$$

L2 norm (today)
We saw that Ridge Regression shrinks coefficients, but they don’t become 0. What if we remove weights that are sufficiently small?
Instead of searching over a discrete set of solutions, use regularization to reduce coefficient of unhelpful features.

Start with a full model, and then “shrink” ridge coefficients near 0. Non-zero coefficients would be considered selected as important.
Ridge for Feature Selection

Look at two related features #bathrooms and # showers. Our model ended up not choosing any features about bathrooms!
What if we had originally removed the # showers feature?

- The coefficient for # bathrooms would be larger since it wasn’t “split up” amongst two correlated features
- Instead, it would be nice if there were a regularizer that favors sparse solutions in the first place to account for this…
LASSO Regression

Change quality metric to minimize

is a tuning parameter that changes how much the model cares about the regularization term.

What if?

\[ \hat{\omega} = \min_{\omega} \text{RSS}(\omega) \implies \hat{\omega}_{LS} \]

What if?

\[ \hat{\omega} = 0 \]

in between?

\[ 0 \leq \| \hat{\omega}_{LASSO} \|_1 \leq \| \hat{\omega}_{LS} \|_1 \]
Ridge Coefficient Paths

Coefﬁcients vs. \( \lambda \):

- Many small, but non-zero coefﬁcients.
LASSO Coefficient Paths

sparse = many 0 coefs.

- Coefficients vs. \( \lambda \)
- One non-zero
- Not exactly 0

Graph showing coefficient paths for different values of \( \lambda \).
Example from Google’s Machine Learning Crash Course
There is no poll to answer for this question. This is an open-ended question.

Why might the shape of the L1 penalty cause more sparsity than the L2 penalty?
When using the L1 Norm ($) as a regularizer, it favors solutions that are **sparse**. Sparsity for regression means many of the learned coefficients are 0.

This has to do with the shape of the norm
Another way to visualize why LASSO prefers sparse solutions
Brain Break
Choosing $\lambda$

Exactly the same as Ridge Regression :)

This will be true for almost every hyper-parameter we talk about.

A hyper-parameter is a parameter you specify for the model that influences which parameters (e.g. coefficients) are learned by the ML algorithm.

---

Hyper-parameter tuning:

- For each setting of HPs:
  - Train model with current HP setting
  - Validate predictor w/ validation set / cross-val
  - Track HP w/ lowest val error
  - Return best HP and estimate of test error.
A very common usage of LASSO is in feature selection. If you have a model with potentially many features you want to explore, you can use LASSO on a model with all the features and choose the appropriate to get the right complexity.

Then once you find the non-zero coefficients, you can identify which features are the most important to the task at hand.
LASSO adds bias to the Least Squares solution (this was intended to avoid the variance that leads to overfitting)

- Recall Bias-Variance Tradeoff

It’s possible to try to remove the bias from the LASSO solution using the following steps

1. Run LASSO to select the which features should be used (those with non-zero coefficients)
2. Run regular Ordinary Least Squares on the dataset with only those features

Coefficients are no longer shrunk from their true values
Issues with LASSO

1. Within a group of highly correlated features (e.g. # bathroom and # showers), LASSO tends to select amongst them arbitrarily.
   - Maybe it would be better to select them all together?
2. Often, empirically Ridge tends to have better predictive performance

**Elastic Net** aims to address these issues

\[
\hat{w}_{\text{ElasticNet}} = \min_w RSS(w) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2
\]

Combines both to achieve best of both worlds!
Be careful when interpreting results of feature selection or feature importances in Machine Learning!

- Selection only considers features included
- Sensitive to correlations between features
- Results depend on the algorithm used!
Recap

**Theme:** Use regularization to do feature selection

**Ideas:**
- Describe “all subsets” approach to feature selection and why it’s impractical to implement.
- Formulate LASSO objective
- Describe how LASSO coefficients change as hyper-parameter is varied
- Interpret LASSO coefficient path plot
- Compare and contrast LASSO and ridge