Section 03: Solutions

1. Bias-Variance Trade-off

Consider a simple statistical learning setting, in which we assume that there is some unknown function relating two random variables X and Y (e.g. $Y = 2X$). Let us denote this function by $Y = \eta(X)$; however, we don't know specifically what this function $\eta(\cdot)$ is. Our goal is as follows. Given X, we want to predict Y with the smallest possible error, in expectation. We formalize this notion below.

(a) Find the function η that minimizes the expected squared error $\mathbb{E}[(Y - \eta(X))^2]$. **Hint:** Observe that $\mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[\mathbb{E}[(Y - \eta(X))^2|X = x]]$ (The "Tower Rule").

Solution:

To determine the best $\eta(X)$, we compute the derivative of hint with respect to $\eta(X)$ and set it to zero, as below.

$$\begin{aligned} 0 &= \frac{d}{d\eta(X)}\mathbb{E}[(Y - \eta(X))^2|X = x] \\ &= \mathbb{E}[\frac{d}{d\eta(X)}(Y - \eta(X))^2|X = x] \\ &= \mathbb{E}[-2(Y - \eta(X))|X = x] \\ &= -2\mathbb{E}[Y|X = x] + 2\eta(X) \end{aligned}$$

Rearranging, we conclude that the optimal function $\eta(x)$ is $\mathbb{E}[Y|X = x]$.

(b) While ideally we want η to be what we computed above, in reality, however, we are restricted to our training data and a function class, the best we can do is $\hat{f}_D = \arg\min_{f \in F} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$, where $D = \{(x_i, y_i)\}$. Here, (x_i, y_i) is a sample from distribution P_{XY}. To account for the prediction error (i.e. quality of our estimator $\hat{f}_D$), we need to calculate

$$\mathbb{E}[\mathbb{E}_D[(Y - \hat{f}_D(x))^2]|X = x]]$$

We can break the expectation into

$$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]] + \mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$$

$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]]$ is called **irreducible error** — the error incurred even in ideal situation.

$\mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$ is called **learning error** — the error incurred by the learning setting (e.g. insufficient data, the chosen model class F is not expressive enough etc.)

Express the **learning error** in terms of

- bias — $(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])$
- and variance — $\mathbb{E}_D[(\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2]$

and explain why there is a trade-off.

Hint: $\eta(x) = \theta$, $\hat{f}_D(x) = \hat{\theta}$ and $\mathbb{E}[\hat{f}_D(x)] = \theta^*$

Solution:

Note that (given some distribution D) θ and θ^* are numbers and hence $\mathbb{E}[\theta] = \theta$ and $\mathbb{E}[\theta^*] = \theta^*$.

$$\begin{aligned}
\mathbb{E}[(\eta(x) - \hat{f}_D(x))^2] &= \mathbb{E}[(\theta - \hat{\theta})^2] \\
&= \mathbb{E}[((\theta - \theta^*) + (\theta^* - \hat{\theta}))^2] \\
&= (\theta - \theta^*)^2 + 2(\theta - \theta^*)\mathbb{E}[\theta^* - \hat{\theta}] + \mathbb{E}[(\theta^* - \hat{\theta})^2] \\
&= (\theta - \theta^*)^2 + \mathbb{E}[(\theta^* - \hat{\theta})^2]
\end{aligned}$$

Note that we can do the last step because $\mathbb{E}[\hat{\theta}] = \theta^*$.
The right term is the variance and the left term is the bias squared.

As complexity of F goes up, the bias is decreasing, while the variance is increasing. Thus, we want to find the sweet spot that both of them are reasonably low. This is called bias-variance tradeoff.

2. Generalized Least Squares Regression

We already saw linear regression in class and the ridge regression will be covered in week three. Here we consider a problem that generalizes both of these. As a reminder, in linear regression, we seek a model that captures a linear relationship between input data and output data. The general case we consider imposes additional structure on the model.

Consider an experiment in which you have n data points $x_i \in \mathbb{R}^d$ and corresponding n observations y_i. We wish to come up with a model $\omega \in \mathbb{R}^d$ that satisfies the following properties: first, the error $\sum_{i=1}^n (x_i^\top \omega - y_i)^2$ should be small; second, we don't want small changes in training data resulting in large changes in solution; third, we want to put different weights in controlling the magnitude of different coordinates of ω. We therefore define

$$\widehat{\omega}_{\text{general}} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2.$$

Here, D is a diagonal matrix, with positive entries on the diagonal. Observe that when D is the identity matrix, we recover ridge regression, and when $\lambda = 0$, we recover least squares regression. Different weights on D_{ii} cause the magnitudes of ω_i to be controlled differently.

2.1. Closed form in the general case

Deduce the closed form solution for $\widehat{\omega}_{\text{general}}$. You should be comfortable with proofs in the "coordinate" form as well as the "matrix" form.

Solution:

We first give the proof using "matrix" notation. The objective function can be expressed as

$$\begin{aligned} f(\omega) &= \|X\omega - y\|_2^2 + \lambda\omega^\top D\omega \\ &= (X\omega - y)^\top (X\omega - y) + \lambda\omega^\top D\omega \\ &= (X\omega)^\top X\omega - (X\omega)^\top y - y^\top X\omega + y^\top y + \lambda\omega^\top D\omega \\ &= \omega^\top X^\top X\omega - 2\omega^\top X^\top y + y^\top y + \lambda\omega^\top D\omega \\ &= \omega^\top (X^\top X + \lambda D)\omega - 2\omega^\top X^\top y + y^\top y \end{aligned}$$

The gradient of f is

$$\begin{aligned} \nabla f(\omega) &= \nabla_\omega(\omega^\top (X^\top X + \lambda D)\omega - 2\omega^\top X^\top y + y^\top y) \\ &= \nabla_\omega(\omega^\top (X^\top X + \lambda D)\omega) - 2\nabla_\omega(\omega^\top X^\top y) + \nabla_\omega(y^\top y) \\ &= 2(X^\top X + \lambda D)\omega - 2X^\top y \end{aligned}$$

Here note that $X^\top X + \lambda D$ is a symmetric matrix, which explains the factor 2 in the gradient term. Setting the gradient $\nabla f(\omega)$ to zero, we can conclude that

$$(X^\top X + \lambda D)\widehat{\omega}_{\text{general}} = X^\top y$$

If $X^\top X + \lambda D$ is full rank then we can get a unique solution:

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

Since D is already given to be a diagonal matrix with strictly positive entries on the diagonal, any strictly positive λ will make the matrix $X^\top X + \lambda D$ invertible.

Solution:

We now give a solution in the "coordinate" form. The objective, when written in coordinate form, is $f(\omega) = \sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2$. As in the previous proof, we first simplify it as follows and then set it zero:

$$\begin{aligned}
\nabla_\omega \left[\sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2\right] &= \nabla_\omega \sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \nabla_\omega \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2 \\
&= \sum_{i=1}^{n} \nabla_\omega (y_i - x_i^\top \omega)^2 + 2\lambda D\omega \\
&= -\sum_{i=1}^{n} 2x_i(y_i - x_i^\top \omega) + 2\lambda D\omega \\
&= -\sum_{i=1}^{n} 2x_i y_i + \sum_{i=1}^{n} 2x_i x_i^\top \omega + 2\lambda D\omega \\
&= -2\sum_{i=1}^{n} x_i y_i + 2\left(\sum_{i=1}^{n} x_i x_i^\top + \lambda D\right)\omega \\
&= 0 \ \ \text{(set it to be 0)}
\end{aligned}$$

$$\widehat{\omega}_{\text{general}} = \left(\sum_{i=1}^{n} x_i x_i^\top + \lambda D\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right)$$

Note that, as expected, this exactly matches the answer we got from the previous approach (because x_i's are all the rows of X, and therefore $\sum_i x_i y_i = X^\top y$, and $\sum_i x_i x_i^\top = X^\top X$).

2.2. Special cases: linear regression and ridge regression

(a) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double all the values of y_i?

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling y doubles ω as well. This makes sense intuitively as well because if the observations are scaled up, the model should also be.

(b) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double the data matrix $X \in \mathbb{R}^{n \times d}$?

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling X halves ω. This also makes sense intuitively because the error we are trying to minimize is $\|X\omega - y\|_2^2$, and if the X has doubled, while y has remained unchanged, then ω must compensate for it by reducing by a factor of 2.

(c) Suppose $D = I$ (that is, it is the identity matrix). That is, this is the *ridge* regression setting. Explain why $\lambda > 0$ ensures a "well-conditioned" setting.

Solution:

The solution is $\widehat{\omega} = (X^\top X + \lambda I)^{-1} X^\top y$. We already saw in a previous part that $X^\top X$ is always positive semidefinite, that is, its eigenvalues are at least zero. Adding λI, where $\lambda > 0$, ensures that $X^\top X + \lambda I$ is in fact positive *definite*. This helps us have a good condition number.

3. Biased Test Error

Is the test error unbiased for these programs? If not, how can we fix the code so it is?

3.1. Program 1

```
1  # Given dataset of 1000-by-50 feature
2  # matrix X, and 1000-by-1 labels vector
3
4  mu = np.mean(X, axis=0)
5  X = X - mu
6
7  idx = np.random.permutation(1000)
8  TRAIN = idx[0:900]
9  TEST = idx[900::]
10
11 ytrain = y[TRAIN]
12 Xtrain = X[TRAIN, :]
13
14 # solve for argmin_w ||Xtrain*w - ytrain||_2
15 w = np.linalg.solve(np.dot(Xtrain.T, Xtrain), np.dot(Xtrain.T, ytrain))
16
17 b = np.mean(ytrain)
18
19 ytest = y[TEST]
20 Xtest = X[TEST, :]
21
22 train_error = np.dot(np.dot(Xtrain, w)+b - ytrain,
23                np.dot(Xtrain, w)+b - ytrain ) / len(TRAIN)
24 test_error = np.dot(np.dot(Xtest, w)+b - ytest,
25                np.dot(Xtest, w)+b - ytest ) / len(TEST)
26
27 print('Train error = ', train_error)
28 print('Test error = ', test_error)
```

Solution:

The error is at the beginning of the program on lines 4 and 5. Notice how μ is a function of both the train and test data. By de-meaning the entire dataset before splitting, we are intertwining the train and test data. The correct procedure is:

- Split into train and test
- Compute the mean of the train data, μ_{train}
- De-mean both the train and test data with μ_{train}

3.2. Program 2

```
# Given dataset of 1000-by-50 feature
# matrix X, and 1000-by-1 labels vector

def fit(Xin, Yin, _lambda):
    mu = np.mean(Xin, axis=0)
    Xin = Xin - mu
    w = np.linalg.solve(np.dot(Xin.T, Xin) + _lambda * np.eye(Xin.shape[1]), np.dot(Xin.T, Yin))
    b = np.mean(Yin) - np.dot(w, mu)
    return w, b

def predict(w, b, Xin):
    return np.dot(Xin, w) + b

idx = np.random.permutation(1000)
TRAIN = idx[0:800]
VAL = idx[800:900]
TEST = idx[900::]

ytrain = y[TRAIN]
Xtrain = X[TRAIN, :]
yval = y[VAL]
Xval = X[VAL, :]

# use cross validation to pick the best hyper-parameter to use
lambdas = [10 ** -5, 10 ** -4, 10 ** -3, 10 ** -2]
err = np.zeros(len(lambdas))

for idx, _lambda in enumerate(lambdas):
    w, b = fit(Xtrain, ytrain, _lambda)
    yval_hat = predict(w, b, Xval)
    err[idx] = np.mean((yval_hat - yval)**2)

lambda_best = lambdas[np.argmin(err)]

Xtot = np.concatenate((Xtrain, Xval), axis=0)
ytot = np.concatenate((ytrain, yval), axis=0)

w, b = fit(Xtot, ytot, lambda_best)

ytest = y[TEST]
Xtest = X[TEST, :]

ytot_hat = predict(w, b, Xtot, lambda_best)
train_error = np.mean((ytot_hat - ytot) **2)
ytest_hat = predict(w, b, Xtest, lambda_best)
test_error = np.mean((ytest_hat - ytest) **2)

print('Train error = ', train_error)
print('Test error = ', test_error)
```

Solution:

We are adding the validation data back into training (creating `Xtot`), and then retraining the whole model on this data. However, optimal value of λ **depends** on size of the training dataset, so by adding more data we are using incorrect value in final `fit` call. In general, models get better the more data you give them, but only add the validation set back in if you are confident the hyperparameter doesn't depend on the number of elements, and that you aren't allowing your model access to the test set.

4. Extra: Stein's Paradox

In this problem, we'll use bias-variance tradeoff to find a non-obvious way of estimating the mean of unrelated distributions.

So far in class, we've always been trying to learn a function – given a bunch of features, understand how they predict the single-number output. In this problem, we're trying to do something a little different. We have n completely unrelated probability distributions. We're going to get one sample from each of the distributions, and attempt to predict each of their means. For some examples, our distributions might be: high temperature in Chicago on January 1st, low temperature in Seattle on December 1st, and your friend's score on the midterm.

More formally, let $\theta \in \mathbb{R}^n$ be the (unknown) true means of our n distributions. We will get a vector X where each $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$. We're assuming that every distribution has the same variance, but our means could be very different. Our job is to report $\hat{\theta}$ to minimize our expected error: $\mathbb{E}\left[||\hat{\theta} - \theta||_2^2\right]$.

4.1. The Natural Estimator

The most natural estimator is to just guess X (i.e. set $\hat{\theta} = X$). It doesn't seem like we have any other viable strategy. We'll use bias-variance tradeoff to show that there's actually a better estimator.

(a) Split the error into bias^2 and variance. I.e. show

$$\mathbb{E}[||\hat{\theta} - \theta||_2^2] = ||\mathbb{E}[\hat{\theta}] - \theta||_2^2 + \mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2]$$

Hint: add and subtract $\mathbb{E}[\hat{\theta}]$.

Solution:

We'll show two versions of this calculation. They're identical, but in one we use summations, and in the other we use vector notation.

$$\begin{aligned}
\mathbb{E}[||\hat{\theta} - \theta||_2^2] &= \mathbb{E}\left[\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right] = \sum_{i=1}^n \mathbb{E}\left[(\hat{\theta}_i - \theta_i)^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i] + \mathbb{E}[\hat{\theta}_i] - \theta_i)^2\right] \\
&= \sum_{i=1}^n \mathbb{E}[(\mathbb{E}[\hat{\theta}_i] - \theta_i)^2] + \sum_{i=1}^n \mathbb{E}[(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])^2] + \sum_{i=1}^n \mathbb{E}\left[2(\mathbb{E}[\hat{\theta}_i] - \theta_i)(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[(\mathbb{E}[\hat{\theta}_i] - \theta_i)^2\right] + \sum_{i=1}^n \mathbb{E}\left[(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])^2\right] + \sum_{i=1}^n \mathbb{E}[2(\mathbb{E}[\hat{\theta}_i] - \theta_i)(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])] \\
&= \mathbb{E}||\mathbb{E}[\hat{\theta}] - \theta||_2^2 + \mathbb{E}||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2 + \mathbb{E}[2(\mathbb{E}[\hat{\theta}_i] - \theta_i)(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])]
\end{aligned}$$

Where we used linearity of expectation repeatedly. It is now enough to show that the final term is equal to 0. Indeed, note that $\mathbb{E}[\hat{\theta}_i] - \theta_i$ is just a number, so we can move the expectation inside to get the last

term is: $2(\mathbb{E}[\hat{\theta}_i] - \theta_i) \cdot \mathbb{E}\left[(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right]$ By linearity of expectation, the last factor in the product is 0 Thus the whole term is 0 as required.

Using vector notation:

$$\begin{aligned}\mathbb{E}\left[||\hat{\theta} - \theta||_2^2\right] &= \mathbb{E}\left[||\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta||_2^2\right] \\ &= \mathbb{E}\left[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2\right] + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta})^T(\mathbb{E}[\hat{\theta}] - \theta)\right] + ||\mathbb{E}[\hat{\theta}] - \theta||_2^2 \\ &= \mathbb{E}\left[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2\right] + ||\mathbb{E}[\hat{\theta}] - \theta||_2^2\end{aligned}$$

Where the last step is a result of $\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right] = \mathbb{E}[\hat{\theta}] - \mathbb{E}[\mathbb{E}[\hat{\theta}]] = 0$.

(b) What is the variance of the estimator $\hat{\theta} = X$? Hint: Remember that for a random variable Z, $\text{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$

Solution:

$$\begin{aligned}\mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2] &= \mathbb{E}\left[\sum_{i=1}^{n}\left(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right)^2\right] \\ &= \sum_{i=1}^{n}\left(\sigma^2\right) \\ &= n\sigma^2\end{aligned}$$

Where the third line follows from knowing that $\text{Var}(\hat{\theta}_i) = \sigma^2$.

(c) What is the bias^2 of the estimator $\hat{\theta} = X$?

Solution:

$$
\begin{aligned}
||\mathbb{E}[\hat{\theta}] - \theta||_2^2 &= \sum_{i=1}^{n} (\mathbb{E}[\hat{\theta}_i] - \theta_i)(\mathbb{E}[\hat{\theta}_i] - \theta_i]) \\
&= \sum_{i=1}^{n} (\theta_i - \theta_i)(\theta_i - \theta_i) \\
&= 0
\end{aligned}
$$

4.2. A Different Estimator

The Bias-Variance Tradeoff says that since our error is just the sum of the bias^2 and the variance, if we can find a way to "tradeoff" bias for variance, we can affect our error. With our previous estimator, the two sources of error are quite imbalanced. None of our error is from bias, it all comes from variance. Can we think of a way to reduce variance (even if it means increasing the bias)?

Normally, the way we would reduce variance would be to sample the random variables again and take the average of the samples. But we can't do that for this problem (it would take us a whole year to get another high temperature on January 1st). Another way to decrease the variance is to "scale down" the random variable. E.g. say we'll have $\frac{9}{10}$ of our estimator come from the random object, and the remaining $\frac{1}{10}$ come from somewhere else. What else can we use? Let's just use 0. This kind of estimator is sometimes called a "shrinkage estimator" because we're pulling the results toward 0.

Our estimator is going to be $\frac{9}{10}X$. We've certainly decreased the variance. But that should sound crazy – we're biasing ourselves. We're intentionally guessing something we **know** is a biased estimator. But our hope is that we will decrease the variance enough to more than cancel out the increase in bias. Let's see.

(a) We've changed the estimator, does the error still break down neatly into bias and variance? Or do we have to change some math from part a of the last question?

Solution:

We never used any facts about $\hat{\theta}$ in the calculation! None of the work needs to change. Bias-Variance tradeoff is a function of the way we define our error, not of the particular prediction we make.

(b) What is the variance of the estimator $\hat{\theta} = \frac{9}{10}X$?

Solution:

$$\begin{aligned}\mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2] &= \mathbb{E}\left[\sum_{i=1}^{n}\left(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right)^2\right] \\ &= \sum_{i=1}^{n}\left(\frac{81}{100}\sigma^2\right) \\ &= \frac{81n}{100}\sigma^2\end{aligned}$$

Where the third line follows from $\text{Var}\left(\frac{9}{10}\hat{\theta}_i\right) = \left(\frac{9}{10}\right)^2 \text{Var}(\hat{\theta}_i)$

(c) What is the bias^2 of the estimator $\hat{\theta} = \frac{9}{10}X$?

Solution:

$$\begin{aligned}||\mathbb{E}[\hat{\theta}] - \theta||_2^2 &= \sum_{i=1}^{n}(\mathbb{E}[\hat{\theta}_i] - \theta_i)(\mathbb{E}[\hat{\theta}_i] - \theta_i]) \\ &= \sum_{i=1}^{n}\left(\frac{1}{10}\theta_i\right)\left(\frac{1}{10}\theta_i\right) \\ &= \frac{1}{100}\sum_{i=1}^{n}\theta_i^2\end{aligned}$$

(d) Suppose you know that the variance of our samples is quite a bit. Specifically assume $\sigma^2 > 1/10\theta_i^2$ for all i. (For the temperature examples we gave, this is pretty reasonable. At least if we use Celsius temperatures. The average Celsius high in Chicago is about 4 degrees, so a variance of about 1.6 degrees Celsius suffices) Have we improved the estimator?

Solution:

If $\frac{19n}{100}\sigma^2 > \frac{1}{100}||\theta||^2$ (i.e. the decrease in variance was more than the increase in the bias^2) then we have improved the estimator. But by our assumption, $\frac{1}{100}||\theta||^2 < \frac{10}{100}n\sigma^2 < \frac{19n}{100}\sigma^2$, so we have improved our estimator!

What's the message here? This is a very weird estimator. It's so weird that when estimators like this work, it's referred to as "Stein's Paradox." What fundamentally happens in this example is that because we're using squared error, if we can shrink the error of the worst estimate (even if we correspondingly increase the error of the other estimates) we'll shrink the overall error, because squared error very heavily punishes large errors, but only moderately punishes moderate errors. One way of interpreting this paradox is that

the squared error isn't the only error you might care about. You might really care about the sum of the absolute errors (instead of their squares) in which case, you wouldn't see a similar effect.

Regardless of the interpretation, this result is counter-intuitive. We just decided to guess smaller values than our samples gave us. On its face the idea is crazy – we only would have come up with such an idea if we had broken down our error into bias and variance and could see where the error was coming from. This is the real takeaway of this calculation – if you can understand where your error is coming from, you might be able to generate new ideas for what to do about it.

4.3. Thinking More about the Estimators

The estimator we came up with in the last problem is unintuitive for more reasons than we've already seen. Suppose we scaled all our data points (i.e. instead of X we got $aX + b$), e.g. we were expecting to get our data in Celsius, but it came to us in Fahrenheit. We would expect that scaling our old estimate, i.e. now reporting $a\hat{\theta} + b$ would give us the same answer as if we did our estimate afresh knowing we were getting data in Fahrenheit.

(a) Is the "natural estimator" scale invariant?

Solution:

Yes, we're guessing $aX + b$ either way.

(b) Is the "shrinkage estimator" $\frac{9}{10}X$ scale invariant?

Solution:

No! If we estimate first, then scale we get $a\frac{9}{10}X + b$ but if we scale first then estimate we get $\frac{9}{10}(aX + b)$. Our estimator is shrinking toward 0. But when we changed from Fahrenheit to Celsius, we changed what 0 was, so we got a different prediction!

Here's another interpretation of Stein's Paradox. We might implicitly want a few things for our estimators: accuracy (in the sense of low-mean-squared-error), unbiasedness, and scale-invariance. It turns out if you're willing to give up two of those three, you can slightly improve the error. Whether that tradeoff makes sense or not depends on what you want – do you care more that your estimator makes a lot of intuitive sense, and behaves "nicely" or do you want to scrape every last bit of error you can out of your estimator? Neither decision is always right or always wrong, but you should at least consider what you're giving up in either case.

It turns out we can do even better than estimating $\frac{9}{10}X$– our idea was to shrink X toward 0 by a constant ratio (i.e. multiply everything by $9/10$). If we instead shrink it in a way that depends on σ^2 and $||X||_2^2$, we'll be able to come up with an estimator that has less error than X, regardless of the relationship between σ^2 and θ.

The "James-Stein Estimator, $\hat{\theta} = \left(1 - \frac{(n-2)\sigma^2}{||X||_2^2}\right) X$, always has less error than X.

4.4. A Harder Calculation

Want more practice with bias-variance tradeoff? Here's another version of Stein's Paradox. Instead of shrinking toward 0, shrink toward $\overline{X}$, the mean of all of your data points, i.e. $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If your data points are coming from similar sources (say each θ_i is a different baseball player's true batting average), you can think of this as reflecting a belief that all the means should be generally similar. (Though this shrinking still works even if the sources are very different, and the result is still counter-intuitive). Let $\mathbf{1}$ be the $n \times 1$ vector of all 1's, and let λ be a real number between 0 and 1. In this problem we'll find the way to choose λ to make $\hat{\theta} = (1-\lambda)X + \lambda\overline{X}\mathbf{1}$ as good of an estimator as possible.

(a) What is the variance of the estimator $\hat{\theta} = (1-\lambda)X + \lambda\overline{X}\mathbf{1}$?

Solution:

We'll first do an intermediate calculation that we'll need later:

$$\begin{aligned}
\mathbb{E}[\overline{X} - \overline{\theta}] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - \theta_i)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \theta_i)(X_j - \theta_j)\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \theta_i)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Now let's calculate the variance.

$$\begin{aligned}
\mathbb{E}\left[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2\right] &= \mathbb{E}\left[||(1-\lambda)(X-\theta) + \lambda(\overline{X} - \overline{\theta})\mathbf{1}||_2^2\right] \\
&= (1-\lambda)^2\mathbb{E}\left[||X-\theta||_2^2\right] + 2\lambda(1-\lambda)\mathbb{E}\left[(X-\theta)^T\mathbf{1}(\overline{X}-\overline{\theta})\right] + \lambda^2 n\mathbb{E}\left[(\overline{X}-\overline{\theta})^2\right] \\
&= (1-\lambda)^2 n\sigma^2 + 2\lambda(1-\lambda)n\mathbb{E}\left[(\overline{X}-\overline{\theta})^2\right] + \lambda^2 n\mathbb{E}\left[(\overline{X}-\overline{\theta})^2\right] \\
&= (1-\lambda)^2 n\sigma^2 + 2\lambda(1-\lambda)\sigma^2 + \lambda^2\sigma^2 \\
&= (1-\lambda)^2 n\sigma^2 + (1-\lambda+\lambda)^2\sigma^2 - (1-\lambda)^2\sigma^2 \\
&= (1-\lambda)^2(n-1)\sigma^2 + \sigma^2
\end{aligned}$$

(b) What is the bias^2 of the estimator?

Solution:

$$\begin{aligned}||\mathbb{E}[\hat{\theta}]||_2^2 &= ||(1-\lambda)\theta + \lambda\overline{\theta}\mathbf{1} - \theta||_2^2 \\ &= \lambda^2||\overline{\theta}\mathbf{1} - \theta||_2^2 \\ &= \lambda^2 \sum_{i=1}^{n}(\theta_i - \overline{\theta})^2\end{aligned}$$

(c) What value of λ will result in the best estimator?

Solution:

As always, our strategy is to take a derivative and set it equal to 0. Our objective is $(1-\lambda)^2(n-1)\sigma^2 + \sigma^2 + \lambda^2\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2$, so we want to solve

$$-2(1-\lambda)(n-1)\sigma^2 + 2\lambda\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2 = 0$$

Solving for λ gives

$$\lambda = \frac{\sigma^2}{\sigma^2 + \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \theta)^2}$$