Assignment #1

Due: 11:59pm February 2, 2018

# Homework 1: Linear Regression

## Introduction

This homework is on different forms of linear regression and focuses on loss functions, optimizers, and regularization. Linear regression will be one of the few models that we see that has an analytical solution. These problems focus on deriving these solutions and exploring their properties.

If you find that you are having trouble with the first couple problems, we recommend going over the fundamentals of linear algebra and matrix calculus. We also encourage you to first read the Bishop textbook, particularly: Section 2.3 (Properties of Gaussian Distributions), Section 3.1 (Linear Basis Regression), and Section 3.3 (Bayesian Linear Regression). (Note that our notation is slightly different but the underlying mathematics remains the same :).

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page.

**Problem 1** (Priors and Regularization,15pts)

In this problem we consider a model of Bayesian linear regression. Define the prior on the parameters as,

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \tau_w^{-1}\mathbf{I}),$$

where $\tau_w$ is as scalar precision hyperparameter that controls the variance of the Gaussian prior. Define the likelihood as,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mathbf{w}^\mathsf{T}\mathbf{x}_i, \tau_n^{-1}),$$

where $\tau_n$ is another fixed scalar defining the variance.

(a) Using the fact that the posterior is the product of the prior and the likelihood (up to a normalization constant), i.e.,

$$\arg\max_{\mathbf{w}} \ln p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \arg\max_{\mathbf{w}} \ln p(\mathbf{w}) + \ln p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}).$$

Show that maximizing the log posterior is equivalent to minimizing a regularized loss function given by $\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})$, where

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2$$
$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w}$$

Do this by writing $\ln p(\mathbf{w} \mid \mathbf{y}, \mathbf{X})$ as a function of $\mathcal{L}(\mathbf{w})$ and $\mathcal{R}(\mathbf{w})$, dropping constant terms if necessary. Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})$ for a $\lambda$ expressed in terms of the problem's constants.

(b) Notice that the form of the posterior is the same as the form of the ridge regression loss

$$\mathcal{L}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}.$$

Compute the gradient of the loss above with respect to $\mathbf{w}$. Simplify as much as you can for full credit. Make sure to give your answer in vector form.

(c) Suppose that $\lambda > 0$. Knowing that $\mathcal{L}$ is a convex function of its arguments, conclude that a global optimizer of $\mathcal{L}(\mathbf{w})$ is

$$\mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{1}$$

For this part of the problem, assume that the data has been centered, that is, pre-processed such that $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$.

(d) What might happen if the number of weights in $\mathbf{w}$ is greater than the number of data points $N$? How does the regularization help ensure that the inverse in the solution above can be computed?

**Solution**

(a) We can treat the prior as a multivariate Normal distribution and the likelihood as a summation of independent Normal distributions after applying natural log. The summation can be pulled out because of the properties of log in line (3). By multiplying by $-1$, argmax is transformed into argmin in line (5). Lastly, we can drop constant terms as they do not impact the result when solving for the maximum/minimum value in lines (3) and (9).

$$\begin{aligned}
\arg\max_{\mathbf{w}} \ln p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) &= \arg\max_{\mathbf{w}} \ln p(\mathbf{w}) + \ln p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) && (2)\\
&= \arg\max_{\mathbf{w}} \ln(\det(2\pi\tau_w^{-1})^{-1/2}) + \ln(e\{-\frac{1}{2}\mathbf{w}^\top \tau_w I \mathbf{w}\}) && (3)\\
&\quad + \ln(\frac{1}{\sqrt{2\pi}\tau_n^{-1}}) + \ln(\sum_i -e\{\frac{(y_i - \mathbf{w}^\top x_i)^2}{2\tau_n^{-1}}\})\\
&= \arg\max_{\mathbf{w}} \ln(e\{-\frac{1}{2}\mathbf{w}^T \tau_w I \mathbf{w}\}) - \sum \ln(e\{\frac{(y_i - \mathbf{w}^\top x_i)^2}{2\tau_n^{-1}}\}) && (4)\\
&= \arg\max_{\mathbf{w}} -\frac{\tau_w}{2}\mathbf{w}^\top\mathbf{w} - \sum \frac{\tau_n(y_i - \mathbf{w}^\top x_i)^2}{2} && (5)\\
&= \arg\min_{\mathbf{w}} \frac{\tau_w}{2}\mathbf{w}^\top\mathbf{w} + \sum \frac{\tau_n(y_i - \mathbf{w}^\top x_i)^2}{2} && (6)\\
&= \arg\min_{\mathbf{w}} \tau_n(\frac{\tau_w}{2\tau_n}\mathbf{w}^\top\mathbf{w} + \sum \frac{(y_i - \mathbf{w}^\top x_i)^2}{2}) && (7)\\
&= \arg\min_{\mathbf{w}} \tau_n(\lambda\mathcal{R}(\mathbf{w}) + \mathcal{L}(\mathbf{w})) && (8)\\
&= \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w}) && (9)
\end{aligned}$$

where $\lambda$ is $\tau_w/\tau_n$ in line (9).

(b)

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= (\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}\\
&= (\mathbf{y}^\top - \mathbf{w}^\top\mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}\\
&= (\mathbf{y}^\top - \mathbf{w}^\top\mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}\\
&= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}\\
&= \mathbf{y}^\top\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}\\
\nabla\mathcal{L}(\mathbf{w}) &= -2\mathbf{X}^\top\mathbf{y} + (\mathbf{X}^\top\mathbf{X} + (\mathbf{X}^\top\mathbf{X})^\top)\mathbf{w} + 2\lambda\mathbf{w}\\
&= -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}\\
&= 2(-\mathbf{X}^\top\mathbf{y} + \mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w})
\end{aligned}$$

(c) Because the data is pre-processed with mean zero, this helps the model because we know that the data has mean zero and do not need to include an offset term such as $w_0$. Furthermore, this is beneficial for computation as most models expect data to be in this format and can return more accurate fits.

The global optimizer of $\mathcal{L}(\mathbf{w})$ occurs when $\nabla\mathcal{L}(\mathbf{w}) = 0$.

$$\begin{aligned}
\nabla\mathcal{L}(\mathbf{w}) = -\mathbf{X}^\top\mathbf{y} + \mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w} &= 0\\
(\mathbf{X}^\top\mathbf{X} + \lambda)\mathbf{w} &= \mathbf{X}^\top\mathbf{y}\\
\mathbf{w} &= (\mathbf{X}^\top\mathbf{X} + \lambda)^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}$$

(d) If the number of weights in $\mathbf{w}$ is greater than the number of data points, then we may observe overfitting to the dataset since the many features can finely tune to the few amount of data points. Furthermore, $\mathbf{X}$ may not be full rank, meaning that the matrix is singular. In other words, the columns of $\mathbf{X}^\top\mathbf{X}$ may be linearly dependent due to the high number of features, leading to a determinant of zero.

We can address this with the regularization term because we know that $\mathbf{X}^\top\mathbf{X}$ is symmetric. We can choose $\lambda$ such that the eigenvalues are all positive. This results in a positive definite matrix that is guarenteed to be invertible.

## 2. Modeling Changes in Congress

The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file `data/year-sunspots-republicans.csv` contains the data you will use for this problem. It has three columns. The first one is an integer that indicates the year. The second is the number of sunspots. The third is the number of Republicans in the Senate. The data file looks like this:

```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
1968,105.9,43
1970,104.5,44
```

and you can see plots of the data in Figures 1 and 2.

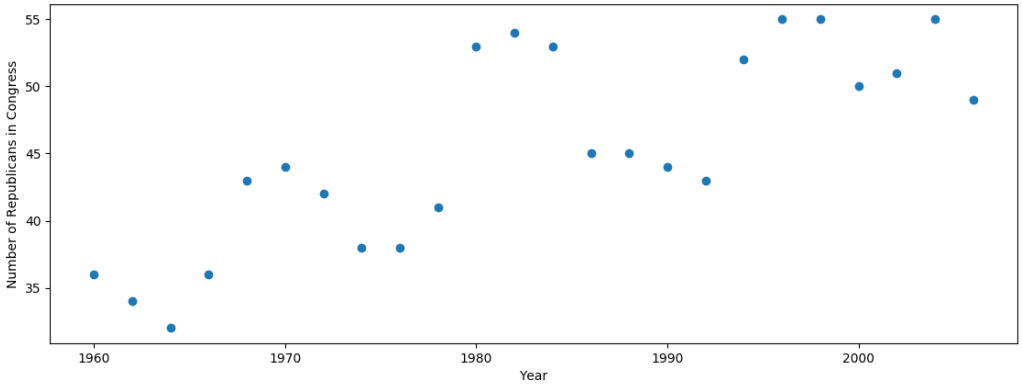



Figure 1: Number of Republicans in the Senate. The horizontal axis is the year, and the vertical axis is the number of Republicans.

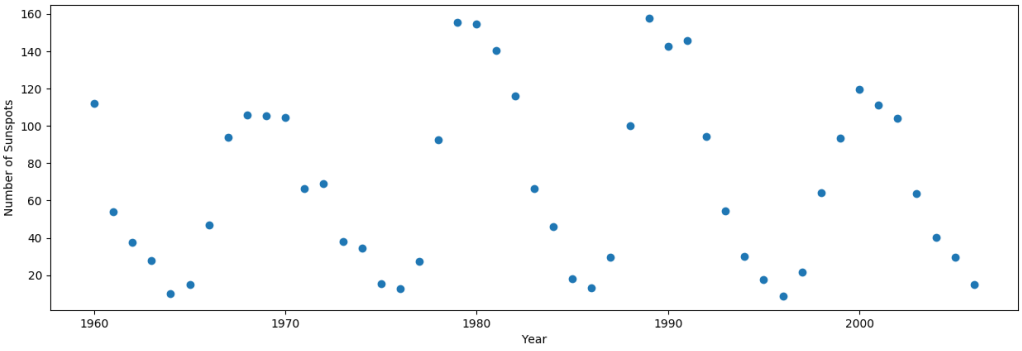



Figure 2: Number of sunspots by year. The horizontal axis is the year, and the vertical axis is the number of sunspots.

Data Source: http://www.realclimate.org/data/senators_sunspots.txt

**Problem 2** (Modeling Changes in Republicans and Sunspots, 15pts)

Implement basis function regression with ordinary least squares for years vs. number of Republicans in the Senate. Some sample Python code is provided in `linreg.py`, which implements linear regression. Plot the data and regression lines for the simple linear case, and for each of the following sets of basis functions:

(a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$

(b) $\phi_j(x) = \exp \frac{-(x-\mu_j)^2}{25}$ for $\mu_j = 1960, 1965, 1970, 1975, \ldots 2010$

(c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 5$

(d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 25$

In addition to the plots, provide one or two sentences for each with numerical support, explaining whether you think it is fitting well, overfitting or underfitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.

Next, do the same for the number of sunspots vs. number of Republicans, using data only from before 1985. What bases provide the best fit? Given the quality of the fit, would you believe that the number of sunspots controls the number of Republicans in the senate?

**Solution**

Mean squared error (MSE) was used to quantify the fit of the model on the dataset. The variables and MSE value are reported in the chart titles.

Years versus number of Republicans in the Senate:

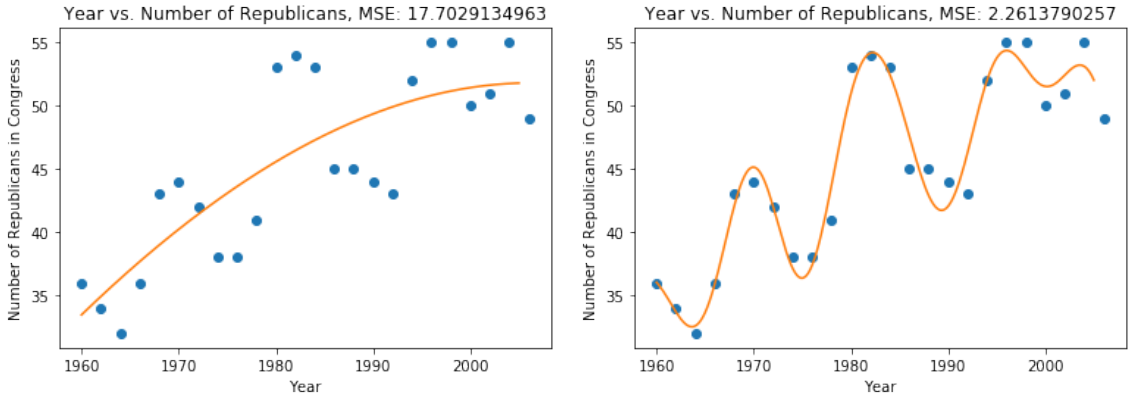



Figure 3: Basis functions a (left) and b (right)

Basis function (a) has a high MSE of 17.70 with a simple curve that seems to be underfitting the data with lower bias/higher variance as it does not capture the curvature pattern. Basis function (b) performs better with a MSE of 2.26 and a model that seems to fit the data well by representing the overall shape.

Basis function (c) has the higest MSE in the comparison of years vs. number of republicans at 45.12. The curve is heavily underfitting the data, likely limited by the lack of detail captured with $j$ capped at 5. Basis function (d) seems to fit the data well with a MSE of 1.62 and captures the general wave-like trend of the data.

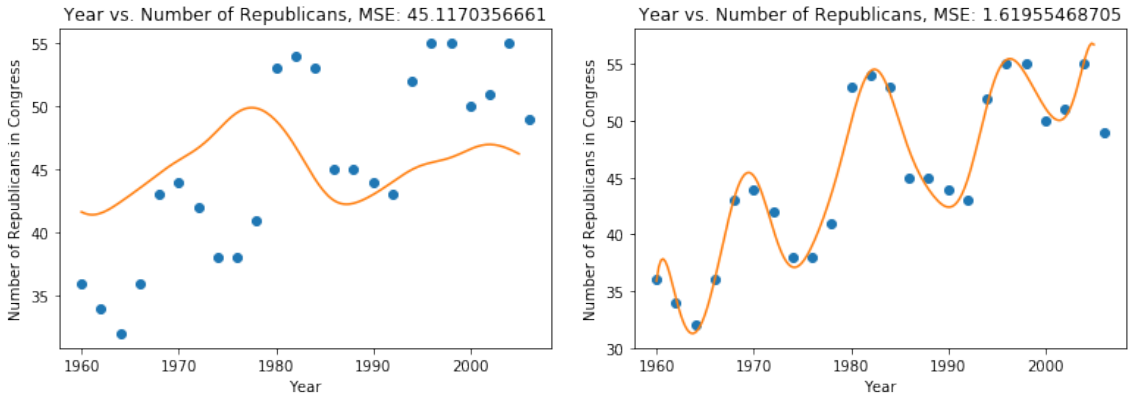



Figure 4: Basis functions c (left) and d (right)

Number of Sunspots versus number of Republicans in the Senate

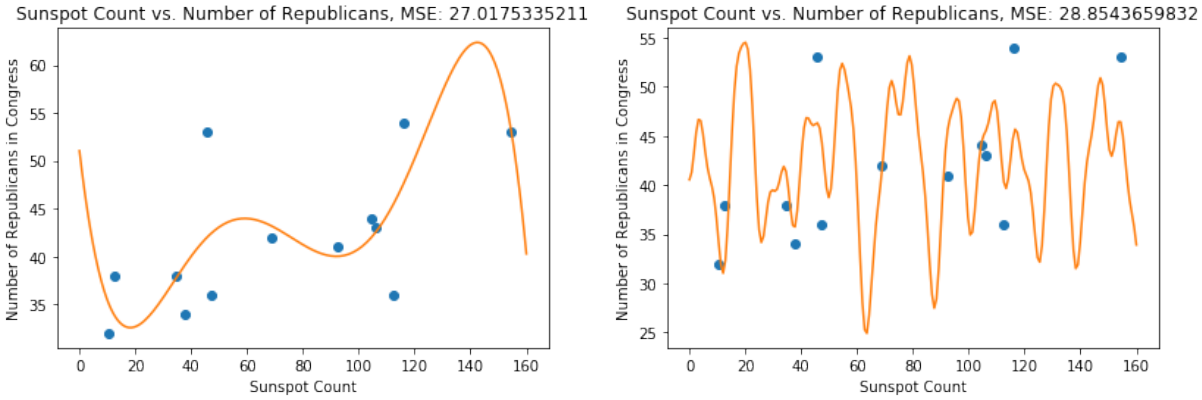



Figure 5: Basis functions a (left) and c (right)

Basis function (a) is underfitting the data with MSE of 27.02. This model does not resemble the general trend of the data and yields a high MSE. Basis function (c) is a poor fit of the data, yielding a jerky curve that is not representative of the general trend, and due to the limitations of the basis function, results in a poor MSE value of 28.85. The basis function produces a curve that is both high variance and high bias.

Basis function (d) is heavily overfitting the dataset with MSE of 9.74e-24 or 0. The model is running close through all of the data points with extremely high bias/low variance. Given basis function (d), it seems just looking at the numbers that the number of sunspots could control the number of Repbulicans in the Senate. Of course, common sense tells us that this is illogical. We observe that sunspot count versus number of Republicans was only run on data prior to 1985. This shows the ability to identify false positive correlations within a dataset. If we used the full dataset, the high precision fit (albeit overfitting) we observe in (d) would not exist. Instead, we would yield a poor MSE of 25.92 if we used the full sunspot/republican dataset.

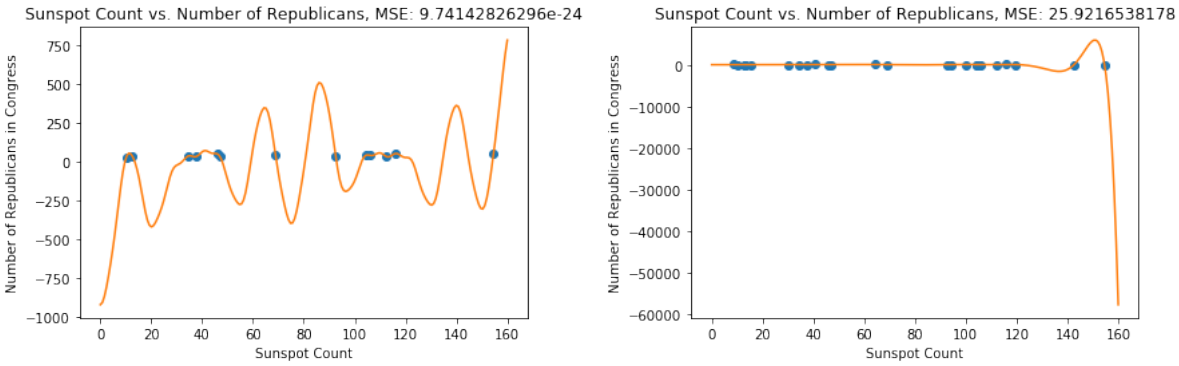



Figure 6: Basis function d (prior-1985 data) and basis function d (all data)

**Problem 3** (BIC, 15pts)

Adapted from *Biophysics : Searching for Principles* by William Bialek.

Consider data $\mathcal{D} = \{(x_i, y_i)\}$ where we know that

$$y_i = f(x_i; \mathbf{a}) + \epsilon_i$$

where $f(x_i; \mathbf{a})$ is a function with parameters $\mathbf{a}$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an additive noise term. We assume that $f$ is a polynomial with coefficients $\mathbf{a}$. Consider the class of all polynomials of degree $K$. Each of these polynomials can be viewed as a generative model of our data, and we seek to choose a model that both explains our current data and is able to generalize to new data. This problem explores the use of the Bayesian Information Criterion (BIC) for model selection. Define the $\chi^2$ (*chi-squared*) error term for each model of the class as:

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left( y_i - \sum_{j=0}^{K} a_j x_i^j \right)^2$$

Using this notation, a formulation of the BIC can be written as:

$$-\ln P(x_i, y_i | \text{model class}) \approx \frac{N}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{N} \ln p(x_i) + \frac{1}{2}\chi^2_{min} + \frac{K+1}{2} \ln N$$

where $\chi^2_{min}(K)$ denote the minimum value of the $\chi^2$ over the set of polynomial models with $K$ parameters. Finally, assume that $x_i \sim Unif(-5, 5)$ and that each $a_j \sim Unif(-1, 1)$. Let $K_{true} = 10$.

(a) Write code that generates $N$ data points in the following way:

1. Generate a polynomial $f(x) = \sum_{j=0}^{K_{true}} a_j x^j$
2. Sample $N$ points $x_i$
3. Compute $y_i = f(x_i) + \epsilon_i$ where $\epsilon$ is sampled from $\mathcal{N}(0, \sigma^2 = \frac{\max_i f(x_i) - \min_i f(x_i)}{10})$.

(b) For a set of $y_i$ generated above and a given $K$, write a function that minimizes $\chi^2$ for a polynomial of degree $K$ by solving for $\mathbf{a}$ using numpy `polyfit`. Check for $N = 20$ that $\chi^2_{min}(K)$ is a decreasing function of $K$.

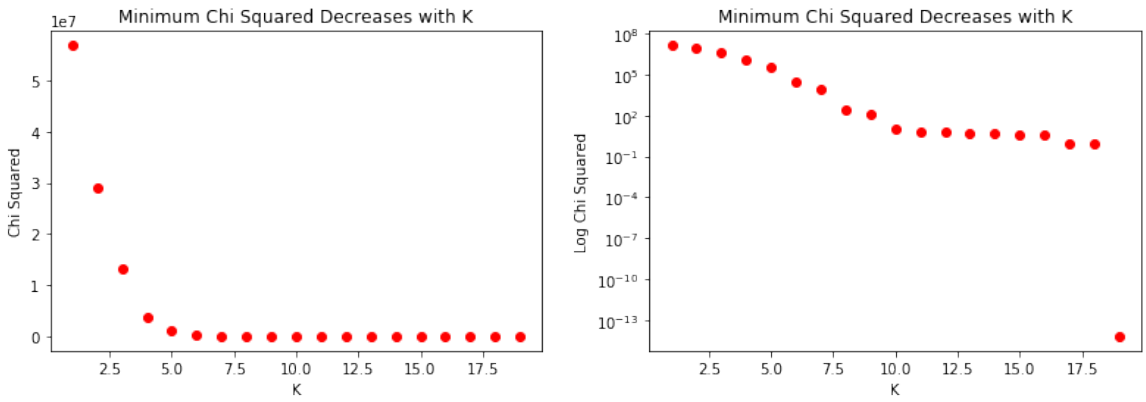
(c) For $N = 20$ samples, run 500 trials. This involves generating a new polynomial for each trial, then from that polynomial, 20 sample data points $\{(x_i, y_i)\}$. For each trial, we can calculate the optimal $K$ by minimizing BIC. Compute the mean and variance of the optimal $K$ over 500 trials.

(d) For $N$ ranging from 3 to $3 \cdot 10^4$ on a log scale (you can use the function $3 * np.logspace(0, 4, 40)$ as your values of $N$), compute the mean and variance of the optimal $K$ over 500 trials for each $N$. Plot your results, where the x-axis is the number of samples ($N$) on a log-scale, and the y-axis is the mean value of the optimal $K$ with error bars indicating the variance over 500 trials. Verify that minimizing the BIC controls the complexity of the fit, selecting a nontrivial optimal $K$. You should observe that the optimal K is smaller than $K_{true}$ for small data sets, and approaches $K_{true}$ as you analyze larger data sets.

**Solution**

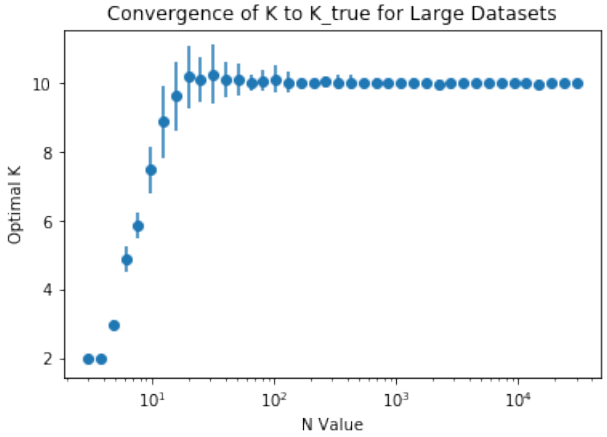
(a) See attached Python file.

(b) We can verify that $\chi^2_{\min}(K)$ is a decreasing function of $K$ in the plots below (left: raw values, right: log Y scale representing $\chi^2_{\min}$ values).




(c) The mean of optimal $K$ is 10.07 and the variance is 0.464. See code in attached Python file.

(d) In the plot below, we see $K$ tending toward $K_{true}$ which shows that minimizing BIC penalizes overly complex polynomials. The first two terms of BIC are constants and invariant. The third term of BIC represents the $\chi^2_{\min}(K)$ value. We observed in part (b) that this value decreases as a function of $K$, suggesting for larger $K$ values. The fourth term of BIC adds $\frac{K+1}{2} \ln N$ to the score and penalizes high $K$ value models, creating a trade off. As $K$ tends to $K_{true}$ for larger datasets, we observe a large drop off in $\chi^2_{\min}(K)$ at $K = 10$. At this point, BIC begins to penalize more complex models and $K > K_{true}$ values start to yield higher BIC scores. This results in the pattern below with optimal $K < K_{true}$ for small data sets and $K$ approaching $K_{true}$ for larger datasets. There is initially high variance tending toward $K_{true}$ but this decreases with the additional of more data points.



**Problem 4** (Calibration, 1pt)
Approximately how long did this homework take you to complete?

**Answer:** 15 hours
**Name:** Christine Zhang
**Email:** christinezhang@college.harvard.edu
**Collaborators:** David Yang