1 Progress

We have assembled two large datasets for the purposes of the project:

- **Pennsylvania voter file:** This file contains a row for every registered voter in Pennsylvania, as well as their party registration, limited demographic information (age, gender), and their voting participation over a set of recent primaries and general elections.
- **Precinct election results:** This file, sourced from the Open Election project, gives the total number of votes received by each candidate in the 2016 presidential election.

Because these files were sourced from distinct locations, there are some challenges in matching up every precinct in every county. For the milestone, we have focused on four of Pennsylvania's 67 counties: Adams, Allegheny, Bedford, and Chester. These counties are diverse: Allegheny and Chester were won by Clinton, while Bedford and Chester were won by Trump. Together, they represent 1,639 precincts and about 950,000 total votes for Clinton and Trump.

We have written code to build and train simple models over this dataset. To compute our objective – the Poisson Binomial log likelihood – at a specific value of β, we use the following procedure for a precinct k:

- Compute the number of Trump and Clinton votes within k (the average precinct having around 600 votes)
- Look up all the voters in k who voted in the 2016 presidential election. Using their data from the voter file, construct the design matrix X_k.
- For a given value of β, compute a vector of probabilities $p_k = \sigma(X_k\beta)$ where $\sigma(\cdot)$ is the sigmoid function. Compute the log-likelihood of the Poisson Binomial using this vector of probabilities p_k and the true number of Trump/Clinton votes.

These values are summed over all precincts to get an overall log-likelihood.

Note that because the data is imperfect – and because individuals may vote for third party candidates or fill out a ballot but only vote in other races – the Clinton/Trump vote total sum is not exactly equal to the number of voters who participated in a given precinct. A quick visual check confirmed that the numbers are within 10% in most cases. To get around this issue, we take the percentage of Clinton/Trump votes that went to Clinton, multiply this by the length of p_k, and round the resulting number to estimate the number of voters in our file who cast a ballot for Clinton or would have done so if they "had to" choose one of the candidates.

We want the model to learn the correct values of β so as to maximize the overall likelihood of the observed data. But we have a problem: the Poisson Binomial likelihood involves sums over all possible configurations of votes – i.e. if Clinton received 200 out of 500 total votes in a precinct, then the likelihood involves a sum over $\binom{500}{200}$ configurations. While this can be estimated well using a discrete Fourier transform (as the `poibin` package we use does), the gradient is computationally intractable.

To address this problem, we make use of the Lyapunov CLT, to observe that the asymptotic distribution of d_k, the number of Democratic votes in precinct k, is given by:

$$d_k \xrightarrow{d} N\left(\sum_i p_{k,i}, \sum_i p_{k,i}(1-p_{k,i})\right)$$

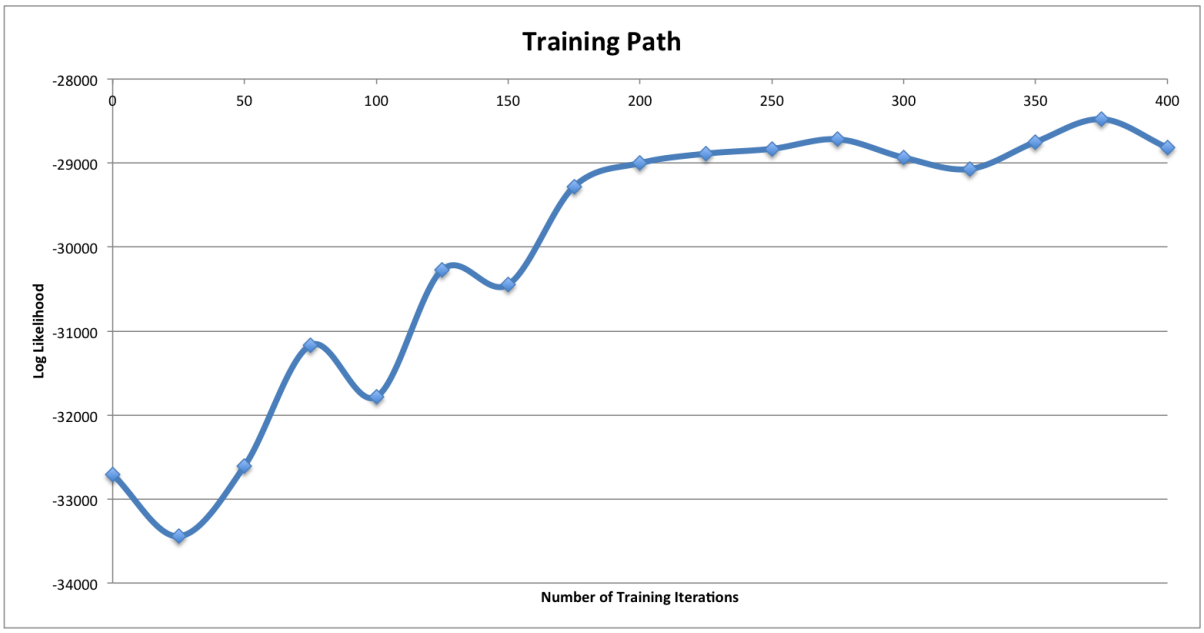
where $p_{k,i}$ is the i^{th} entry of p_k. This result is proven in the appendix. It allows us to *estimate* the likelihood with a much simpler function of β. In this case, the contribution of precinct k to the overall log-likelihood is approximately:

$$\ell_k = -log\left(\mu_k\right) + \frac{1}{\phi_k^2}\left(d_k - \mu_k\right)^2$$

where $\mu_k = \sum_i p_{k,i}(1-p_{k,i}), \phi_k^2 = \sum_i p_{k,i}(1-p_{k,i})$, and $p_{k,i} = \sigma(\beta^T x_{k,i})$. This yields a gradient of the form:

$$\nabla_\beta \ell_k = \frac{1}{2}\left(\frac{(d_k-\mu_k)^2}{\phi^4} - \frac{1}{\phi^2}\right)\left(\sum_i (2p_{k,i}-1)p_{k,i}^2 x_{k,i}\right) + \frac{1}{\phi^2}(d_k-\mu_k)\left(\sum_i p_{k,i}(1-p_{k,i})x_{k,i}\right)$$

Using stochastic gradient descent, we fit a simple model to the four counties where we predict the probability to vote for Clinton given county, party registration, primary participation, gender, and age. Computing the likelihood every 25 iterations, we see a modest improvement over the first 400 iterations (after which the likelihood essentially plateaus):



Moreover, the coefficients all appear to have the correct sign in the model after 400 iterations (noting that positive indicates more likely to vote for Clinton over Trump):

Coefficient	Fitted Value
is Chester resident?	0.61
is Adams resident?	-0.42
is Bedford resident?	-0.78
is Allgeheny resident?	0.39
is registered Democrat?	0.30
is registered Republican?	-0.52

Coefficient	Fitted Value
voted in Democratic primary?	0.18
voted in Republican primary?	-0.30
is Female?	0.03
is Male?	-0.05
Age	-0.26

2 Next Steps

(*NITIN CLEAN THIS UP *)

- Fit models to more data and fit more complex models with more predictors
- Evaluate as a prediction task – fit model to 70% of the precincts and test on the remaining 30% of the precincts, using the total number of Trump and Clinton votes in the precinct as the outcome variable.
  - We're particularly curious if this approach (modeling at the individual level) can yield better predictions than modeling at the precinct level (e.g. by aggregating demographics to the precinct level and then fitting a linear model).
  - This is where we can make comparisons to more conventional modeling approaches: linear regression, regression trees, and neural nets, all fit at the precinct level
- Explore other methods for maximizing the log likelihood and see if they perform better in practice. May want to explore numerical gradients and expectation maximization.

3 CLT Proof

Define $d_k = \sum_{i=1}^{n} d_{k,i}$ to be the number of Democratic votes in precinct k, where $d_{k,i}$ is an indicator variable denoting whether person i in precinct k voted Democrat. We know that d_k follows a Poisson Binomial distribution with success probabilities $p_k = (p_{k,1}, \ldots, p_{k,n})$. Define:

$$s_k^2 = \sum_{i=1}^{n} p_{k,i}(1 - p_{k,i})$$

We check the Lyapunov CLT condition for the fourth moment:

$$\lim_{n\to\infty} \frac{1}{s_k^4} \sum_{i=1}^{n} E\left((d_{k,i} - p_{k,i})^4\right) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} p_{k,i}(1 - p_{k,i})\left(3p_{k,i}^2 - 3p_{k,i} + 1\right)}{\left(\sum_{i=1}^{n} p_{k,i}\left(1 - p_{k,i}\right)\right)^2} \stackrel{?}{=} 0$$

Observe that $3p_{k,i}^2 - 3p_{k,i} + 1 \in (0, 1)$ if $p_{k,i} \in (0, 1)$. Hence, the numerator is strictly less than $\sum_{i=1}^{n} p_{k,i}(1 - p_{k,i})$. Hence, if we can guarantee the numerator grows without bound, then this limit is 0 and the Lyapunov CLT applies. We can do so using a simple condition, like enforcing that there is some $\epsilon > 0$ such that $\epsilon < \bar{p}_i < 1 - \epsilon$ for all i (i.e. the mean probability of voting Democrat in a precinct never falls below some low threshold ϵ or above some high threshold $1 - \epsilon$.

The Lypaunov CLT now tells us that:

$$\frac{d_k - \sum_{i=1}^{n} p_{k,i}}{s_k} \xrightarrow{d} N(0, 1)$$