Erfan Sobhaei
402208568
Due Date: 1402/8/30

Assignment 1

Sharif University of Technology
Deep Learning - CS22860
Fall 2023

Problem 1

There is a coin-tossing experiment with probability p for getting a head. We run this experiment n times. Getting a head in each experiment could be modeled as a *Bernoulli* random variable. Therefore, if we show the total number of heads in n tosses with X, we have $X \sim Binomial(n, p)$.

We know that, $X = r$ so we have:

$$Likelihood = l_X(\theta) = \mathbb{P}(X = r \mid \theta)$$
$$M.L.E : \hat{p} = \underset{p}{\operatorname{argmax}} \{l_X(p)\} = \underset{p}{\operatorname{argmax}} \left\{ \binom{n}{r} p^r (1-p)^{n-r} \right\}$$

In the expression we want to minimize, $\binom{n}{k}$ is irrelevant to p so we will find $\underset{r}{\operatorname{argmax}}\{p^r(1-p)^{n-r}\}$. To find the $argmax$ of this function, we will apply log_e and find its $argmax$ instead. Therefore, we have:

$$\hat{p} = \underset{p}{\operatorname{argmax}} \{r \log(p) + (n-r)\log(1-p)\}$$
$$\frac{d}{dp}\left(r\log(p) + (n-r)\log(1-p)\right) = \frac{r}{p} + \frac{r-n}{1-p} = \frac{r - rp + rp - pn}{p(1-p)} = \frac{r-np}{p(1-p)}$$
$$\frac{d}{dp}\log(l_X(p)) = 0 \implies r - np = 0 \implies \hat{p} = \frac{r}{n}$$

For calculating the bias and variance of this estimator, we conclude as follows:

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \implies Bias(\hat{p}) = \mathbb{E}(\hat{p}) - p = \mathbb{E}(\frac{r}{n}) - p = \frac{r}{n} - p$$
$$Var(\hat{\theta}) = \mathbb{E}\left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2\right) \implies Var(\hat{p}) = \mathbb{E}\left((\hat{p}_n - \mathbb{E}(\hat{p}_n))^2\right) = \mathbb{E}\left((\frac{r}{n} - \frac{r}{n})^2\right) = 0$$

In conclusion, the estimator that we achieved from M.L.E is a biased indicator with zero variance.

Problem 2

We consider two random variables as $X \sim Binomial(n, p)$ and $\theta \sim Beta(\alpha, \beta)$. For finding posterior probability, we use *Bayes theorem* as follows:

$$f_\theta(\theta = p \mid X = r) = \frac{\mathbb{P}(X = r \mid \theta = p) f_\theta(p)}{\mathbb{P}(X = r)} = \frac{\binom{n}{r} p^r (1-p)^{n-r} c p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 \binom{n}{r} p^r (1-p)^{n-r} c p^{\alpha-1}(1-p)^{\beta-1} dp} = \frac{p^{r+\alpha-1}(1-p)^{n+\beta-r-1}}{C}$$

Therefore, *posterior probability* is just $p^{r+\alpha-1}(1-p)^{n+\beta-r-1}$ over a normalization constant. For performing Maximum A Posteriori (MAP), we will find the $argmax$ of this function. Similar to Problem 1, we will apply the natural logarithm and find the point where the derivative of the function is zero.

$$\hat{\theta} = \hat{p} = \underset{p}{\operatorname{argmax}} \left\{p^{r+\alpha-1}(1-p)^{n+\beta-r-1}\right\} = \underset{p}{\operatorname{argmax}} \{(r+\alpha-1)\log(p) + (n+\beta-r-1)\log(1-p)\}$$

$$\frac{d}{dp}\left(r+\alpha-1\right)\log(p)+\left(n+\beta-r-1\right)\log(1-p))=\frac{p(2-\alpha-\beta-n)+(r+\alpha-1)}{p(1-p)}$$

$$\frac{d}{dp}\log(\mathbb{P}(\theta=p\mid X=r)=0 \implies p(2-\alpha-\beta-n)+(r+\alpha-1)=0 \implies p(2-\alpha-\beta-n)=(-r-\alpha+1)$$

$$\implies \hat{p}=\frac{r+\alpha-1}{n+\alpha+\beta-2}$$

To prevent this value from becoming negative, $n \geq 2$.

In comparison to MLE, there is an additional term in the nominator $(\alpha-1)$, and one in the denominator $(\alpha-1+\beta-1)$. The role of n and r is the same as MLE. By increasing β, the estimated value becomes smaller (and vice versa). Changing α closely depends on the other values. There are some example values for this experiment $n=100,\ r=45$.

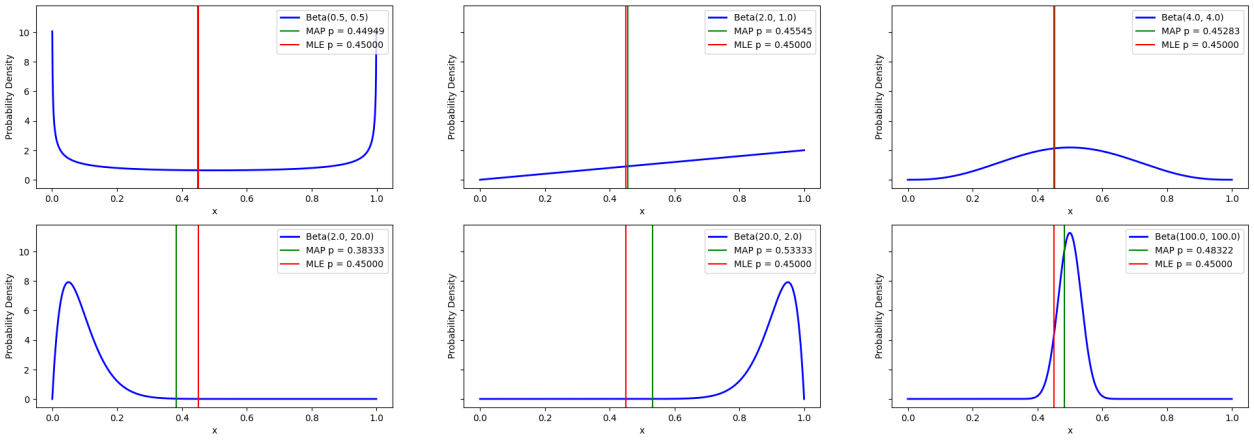

Figure 1: Probability density function for MLE and MAP estimator

As we can see in Figure 1, based on the prior distribution, the estimated value for MAP moves closer to the expected value of the distribution rather than the MLE estimated value. Also, when the density of the expected value of the distribution is more significant, the move is more notable.

Another interesting viewpoint is that the Beta distribution is a conjugate prior, which means if we assume our prior distribution to be Beta, the posterior distribution is another Beta distribution with updated parameters. To recall, the calculated posterior probability was $cp^{r+\alpha-1}(1-p)^{n+\beta-r-1}$. Its distribution is $Beta(r+\alpha, n+\beta-r)$. We know that our updated belief of the value of probability is from this distribution, and one of the good ways to estimate it is the mode of the distribution. In this case, the mode of the Beta distribution is $\frac{(r+\alpha)-1}{(r+\alpha)+(n+\beta-r)-2}$ which is in fact the calculated $\hat{p}$. From this, we conclude that the role of parameters is just to update our parameters of Beta distribution and adjust the shape of its distribution.

Please note the code generated Figure 1 is available on my GitHub.

Problem 3

a) To prove the equation, we start with the following expression:

$$(x_i - \hat{\mu}_i)^2 = ((x_i - \mu) - (\hat{\mu}_i - \mu))^2 = (x_i - \mu)^2 + (\hat{\mu}_i - \mu)^2 - 2(x_i - \mu)(\hat{\mu}_i - \mu)$$
$$(\hat{\mu}_i - \mu)^2 = (x_i - \hat{\mu}_i)^2 - (x_i - \mu)^2 + 2(x_i - \mu)(\hat{\mu}_i - \mu)$$

We take the expectation of the above equation:

$$\mathbb{E}(\|\hat{\mu} - \mu\|^2) = \mathbb{E}[(x - \hat{\mu})^2 - (x - \mu)^2 + 2(x - \mu)(\hat{\mu} - \mu)]$$
$$\mathbb{E}(\|\hat{\mu} - \mu\|^2) = \mathbb{E}(\|(x - \hat{\mu})\|^2) + \mathbb{E}(\|(x - \mu)\|^2) + \mathbb{E}(2(x - \mu)(\hat{\mu} - \mu))$$
$$\mathbb{E}(\|\hat{\mu} - \mu\|)^2) = \mathbb{E}(\|(x - \hat{\mu})\|^2) - n(1) + 2\sum \mathrm{Cov}(x_i, \hat{\mu}_i)$$

b) In order to prove $\mathrm{Cov}(x_i, \hat{\mu}_i) = \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial x_i}\right)$, we subtract μ_i from x for getting $(x - \mu) \sim \mathcal{N}(0, 1)$ so we have:

$$\mathrm{Cov}(\hat{\mu}_i(x), x_i) = \mathbb{E}(\hat{\mu}_i(x)x_i) - \mathbb{E}(\hat{\mu}_i(x))\mathbb{E}(x_i) = \mathbb{E}(\hat{\mu}_i(x)(x_i - \mathbb{E}(x_i)))$$
$$= \mathbb{E}(\hat{\mu}_i(x)(x_i - \mu_i)) = \int \hat{\mu}_i(x)(x_i - \mu_i) \prod_{j=1}^{n} \varphi(x_j - \mu_j)\, dx$$
$$= \int \hat{\mu}_i(x) \cdot \Big[-\frac{\partial}{\partial x_i} \prod_{j=1}^{n} \varphi(x_j - \mu_j)\Big] dx = \int \frac{\partial}{\partial x_i} \hat{\mu}_i(x) \cdot \prod_{j=1} \varphi(x_j - \mu_j) dx = \mathbf{E}\frac{\partial \hat{\mu}_i}{\partial x_i}.$$

c) From the equation derived in part a, we can replace the $\hat{\mu}$ with $\hat{\mu}_{JS}$:

$$\mathbb{E}(\|\hat{\mu}_{JS} - \mu\|)^2) = \mathbb{E}(\|(x - \hat{\mu})\|^2) - n(+2\sum \mathrm{Cov}(x_i, \hat{\mu}_{JS_i})$$
$$= \left\|x - \left(1 - \frac{n-2}{\|x\|^2}\right)x\right\|^2 - n + 2\sum \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial x_i}\right) = \left\|\left(\frac{n-2}{\|x\|^2}\right)x\right\|^2 - n + 2\sum \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial x_i}\right)$$
$$= \frac{(n-2)^2}{\|x\|^2} - n + 2\sum \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial x_i}\right)$$

Also, we compute the partial derivative of the James-Stein Estimator w.r.t x:

$$\frac{\partial \hat{\mu}_i}{\partial x_i} = \left(1 - \frac{n-2}{\|x_i\|^2}\right)(1) + \left(\frac{n-2}{\|x_i\|^4} 2x_i\right)(x_i) \implies \sum \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial x_i}\right) = n - \frac{(n-2)^2}{\|x\|^2}$$

With back substituting the last term in the equation we get:

$$\mathbb{E}(\|\hat{\mu}_{JS} - \mu\|)^2) = \frac{(n-2)^2}{\|x\|^2} - n + 2\left(n - \frac{(n-2)^2}{\|x\|^2}\right) = n - \frac{(n-2)^2}{\|x\|^2}$$

In the last expression, the second term is a positive term for $n \geq 3$, therefore we are subtracting a positive term from $\mathbb{E}(\|x - \mu\|^2) = n$. When the order of $\|x\|^2$ is relatively smaller than n^2, the second term will go to zero, so the error will become n. In other cases, however, the value is smaller than n. Therefore, the James-Stein estimator is not worse than the maximum likelihood estimator anywhere.

My answer to this problem is highly inspired by [1] and [2].

Assignment 1

Problem 4

a) For random vector $x \in \mathbb{R}^d$ where $d \geq 2$, we know the general form of the joint probability density function (pdf) is the product of all individual probability density functions.

$$X \sim \mathcal{N}(0, \sigma^2) \implies f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

$$f(x_1, x_2, \ldots, x_d) = f(x_1) f(x_2) ... f(x_d)$$

$$f(x_1, x_2, \ldots, x_d) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^d} e^{-\frac{x_1^2 + x_2^2 + \ldots + x_d^2}{2\sigma^2}}$$

Let us define the distance of the origin in the d-dimensional space as $r^2 = x_1^2 + x_2^2 + \ldots + x_d^2$. It is possible to consider $f(x_1, x_2, \ldots, x_d)$ as a function of r only. We have

$$f(x_1, x_2, \ldots, x_d) = f(r) = Ae^{-Br^2}$$

where $A = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^d}$ and $B = \frac{1}{2\sigma^2}$.

We can infer that the joint probability density function is a *radial* function of r. Therefore, we should look for r^2 instead of elements of x. As long as this function is only related to the distance, it almost surely has the spherical symmetry property for large d.

b) When x has a radial joint density function, we can decompose and consider it as a linear combination d orthogonal vectors. The general form of function is

$$f(r) = Ae^{-Br^2}$$

Where r^2 could be written as $x_1^2 + x_2^2 + \ldots + x_d^2$, so we have

$$f(x_1^2 + x_2^2 + \ldots x_d^2) = Ae^{-B(x_1^2 + x_2^2 + \ldots + x_d^2)} = A\left(e^{-Bx_1^2} \cdot e^{-Bx_2^2} \cdot ... \cdot e^{-Bx_d^2}\right)$$

By decomposing r to d orthogonal vectors, we showed that $\mu = 0$, because if that was not the case it could be projected onto orthogonal lines, and algebraically it had terms that were multiple of elements of x. For variance, the restriction here is that A and B were the same value for each pdf as they are factored out. It is impossible to have different variances and come up with a radial function, so the variances of all components are actually the same.

Problem 5

a) To prove given equality, we will write the residuals sum of squared errors, calculate its derivative, and set it to zero.

$$\begin{aligned} RSS &= \|y - \hat{y}\|^2 \\ &= (y - \hat{y})^T (y - \hat{y}) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= (y^T - \hat{\beta}^T X^T)(y - X\hat{\beta}) \\ &= y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta} \end{aligned}$$

From linear algebra derivation formulas we know $\frac{\partial x^T A}{\partial x} = A^T$ and $\frac{\partial x^T A^T A x}{\partial x} = 2x^T A$, so

$$\begin{aligned} \frac{\partial SSE}{\partial \hat{\beta}} = 0 &\implies 0 - y^T X - (X^T y)^T + 2\hat{\beta}^T X^T X = 0 \\ &\implies 2\hat{\beta}^T X^T X = 2y^T X \implies \hat{\beta}^T = y^T X (X^T X)^{-1} \\ &\implies \hat{\beta} = (X^T X)^{-1} X^T y \end{aligned}$$

Now we have $\hat{\beta}$, we substitute it in $\hat{y} - y$:

$$\begin{aligned} \hat{y} - y &= X\hat{\beta} - (X\beta + \epsilon) \\ &= X\hat{\beta} - X\beta - \epsilon \\ &= X[(X^T X)^{-1} X^T y] - X\beta - \epsilon \\ &= X[(X^T X)^{-1} X^T (X\beta - \epsilon)] - X\beta - \epsilon \\ &= X[(X^T X)^{-1} X^T] X\beta - X(X^T X)^{-1} X^T \epsilon - X\beta - \epsilon \\ &= X[I]\beta - X(X^T X)^{-1} X^T \epsilon - X\beta - \epsilon \\ &= (X(X^T X)^{-1} X^T - I)\epsilon \end{aligned}$$

b) From part a, the value of $\hat{\beta}$ is calculated, so we have:

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X(X^T X)^{-1} X^T y \\ &= Py \end{aligned}$$

We know that $X(X^T X)^{-1} X^T$ is an orthogonal projection matrix of rank $n - k$. We call it P such that $\hat{y} = Py$, which means that we project matrix y onto the column space of X. By that, we try to find the best possible projection on column space of X, given y. We continue as follows:

$$\begin{aligned} \|y - \hat{y}\|^2 &= (y - \hat{y})^T (y - \hat{y}) \\ &= (y - Py)^T (y - Py) \\ &= ((I - P)y)^T ((I - P)y) \\ &= y^T (I - P)^T (I - P) y \\ &= y^T (I - P)^2 y \\ &= y^T (I - P) y \end{aligned}$$

Next, we need to get the expected value of $y^T(I-P)y$:

$$\begin{aligned}\mathbb{E}\left(\|y-\hat{y}\|^2\right) &= \mathbb{E}\left(y^T(I-P)y\right)\\ &= tr(\sigma^2(I-P)) + E(y)(I-P)E(y^T)\\ &= \sigma^2(n-k)+0\\ \implies \frac{\mathbb{E}\left(\|\hat{y}-y\|^2\right)}{n} &= \frac{\sigma^2(n-p)}{n}\end{aligned}$$

c) The conclusion of the expressions in part a is that, even though we have a closed-form formula for calculating the best fitting regression model, it will always have a ϵ called irreducible error. We never can reduce this error due to the nature of the problem and our measurement noises. From part b, we can infer the relation between the prediction error, number of features, and variance of errors of labels. Therefore, when p is sufficiently large related to n, the regression model will learn the noise of instances instead of the real patterns.

My answer to this problem is highly inspired by [3], [4] and [5].

Problem 6

To show that any $\lambda > 0$, there exists a $t > 0$ such that the answer for $\text{argmin}_\beta \|X\beta - y\|^2$ s.t. $\|\beta\|_p \leq t$ equals to the answer of $\text{argmin}_\beta \|X\beta - y\|^2 + \lambda\|\beta\|_p$. We use Lagrangian for the first optimization problem to change the form of constraints, so we have:

$$g(\alpha) = \min_\beta \mathcal{L}(\beta, \alpha; X, y) = \min_\beta\{\|X\beta - y\|^2 + \alpha\left(\|\beta\|_p - t\right)\}$$

We know that the constrained objective function and its constraint are both convex, so we have strong equality. Thus, we have:

$$\begin{aligned}\|X\beta^* - y\|^2 &= g(\alpha^*)\\ &= \min_\beta\{\|X\beta - y\|^2 + \alpha^*\left(\|\beta\|_p - t\right)\}\\ &\leq \|X\beta^* - y\|^2 + \alpha^*\left(\|\beta^*\|_p - t\right)\\ &\leq \|X\beta^* - y\|^2\end{aligned}$$

The last inequality because that we assume $\alpha \geq 0$ and $\|\beta\|_p \leq t$; We see that all the expression above are equal because of the equality between first and last terms. In order to find the relation between t and λ we have:

$$\begin{aligned}\min_\beta\{\|X\beta - y\|^2 + \alpha^*\left(\|\beta\|_p - t\right)\} &= \|X\beta^* - y\|^2 + \alpha^*\left(\|\beta^*\|_p - t\right)\\ \min_\beta\{\|X\beta - y\|^2 + \alpha^*\left(\|\beta\|_p\right)\} &= \|X\beta^* - y\|^2 + \alpha^*\left(\|\beta^*\|_p\right)\end{aligned}$$

From this we reach the conclusion that optimal β is in fact β^* and the answer of aforementioned two equations which are constrained, and unconstrained problems are the same. Therefore, we can conclude that $\alpha^* = \lambda$ and there is no difference between $\|\beta\|_p \leq t$ constraint or $\lambda\|\beta\|$ term in unconstrained optimization.

My answer to this problem is highly influenced by [6].

Problem 7

a) For the best subset selection method, we can use residual sum of squared errors as the objective function as follows:

$$\begin{aligned} J(\hat{\beta}) &= \{\|y - X\hat{\beta}\|^2\} \\ &= \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right) \\ &= y^T y - 2\hat{\beta}X^T y + \hat{\beta}^T \hat{\beta} \end{aligned}$$

From $X^T X = I$, we know that X is orthogonal. Therefore, from $y = X\beta$, we conclude that $\beta = X^T y$, then we have:

$$J(\hat{\beta}) = y^T y - 2\hat{\beta}(\beta) + \hat{\beta}^2$$

Then we take the partial derivative w.r.t. $\hat{\beta}$ and set it to zero:

$$\frac{\partial J}{\partial \hat{\beta}} = -2\beta + 2\hat{\beta} = 0$$

In the result above, we conclude that estimators for parameters are the same as parameters, but due to the nature and definition of the best subset selection method, we should only choose M elements with the largest magnitude from β, so we show this by an indicator function multiplier. In Figure 2, the selected features for some values of M are shown.

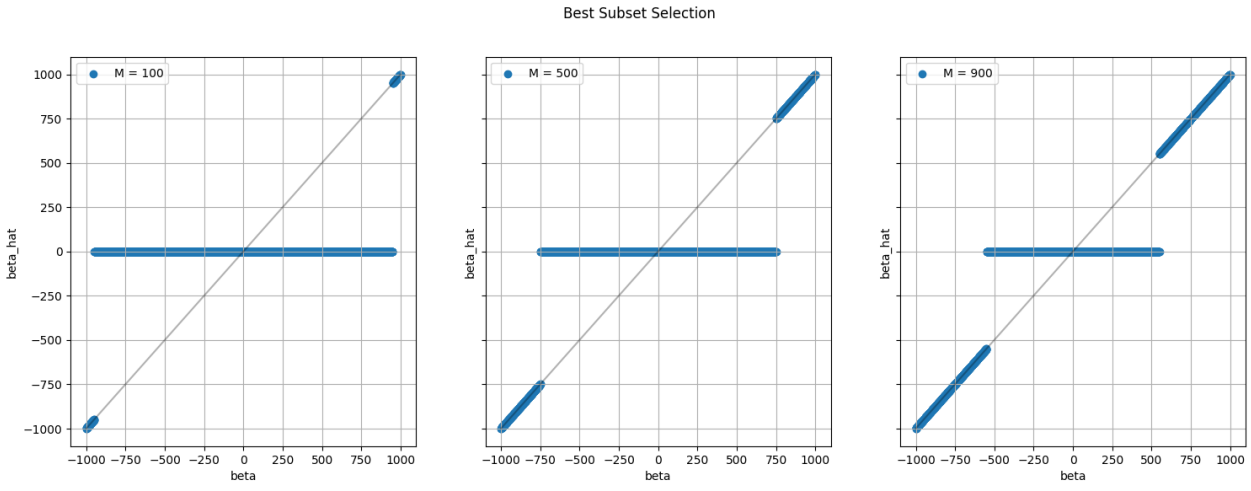

Figure 2: Selected features by best subset selection against default features

b) Like part a, we write the residuals sum of squared error, but in this case with regularization term:

$$\begin{aligned} J(\hat{\beta}) &= \{\|y - X\hat{\beta}\|^2\} + \lambda\|\hat{\beta}\|^2 \\ &= y^T y - 2\hat{\beta}X^T y + \hat{\beta}^2 + \lambda\|\hat{\beta}\|^2 \end{aligned}$$

Then we take the partial derivative w.r.t. $\hat{\beta}$ and set it to zero:

$$\begin{aligned}\frac{\partial J}{\partial \hat{\beta}} &= -2\beta + 2\hat{\beta} + 2\lambda\hat{\beta} = 0 \\ &\implies \hat{\beta}(1+\lambda) = \beta \\ &\implies \hat{\beta} = \frac{(1+\lambda)}{\beta}\end{aligned}$$

So all $\hat{\beta}_i$s have a coefficient, so the line on which they lay, changes a little bit in terms of slope. In Figure 3, the estimated features for some values of λ are shown.

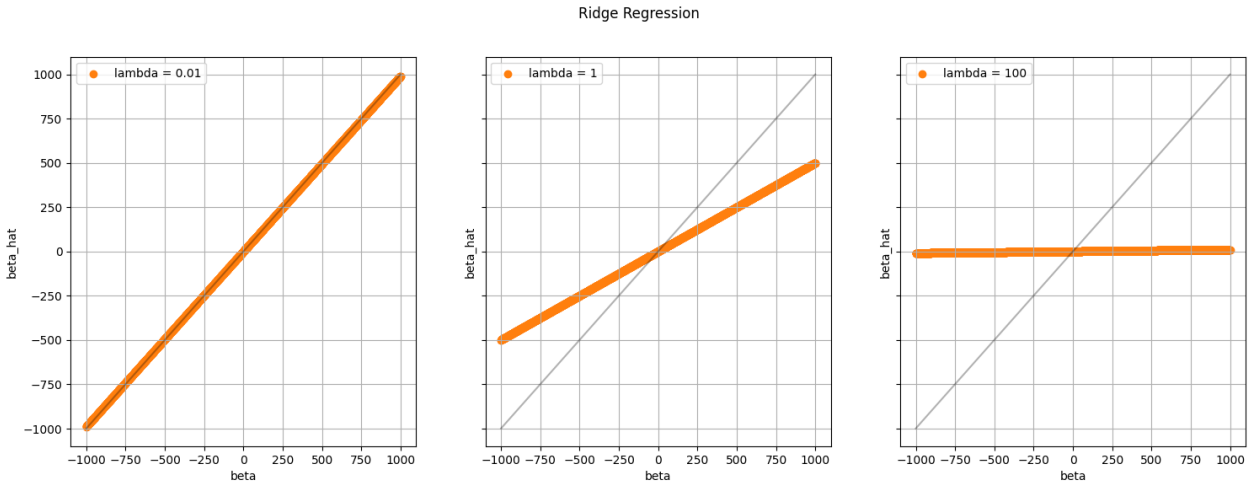

Figure 3: Regularized features by Ridge against default features

c) In the Lasso case, we use the residual sum of error plus the first norm, as the objective function. Therefore, we have:

$$\begin{aligned}J(\hat{\beta}) &= \{\|y - X\hat{\beta}\|^2\} + \lambda|\beta| \\ &= y^T y - 2\hat{\beta}X^T y + \hat{\beta}^2 + \lambda|\beta| \\ &= y^T y - 2\hat{\beta}\beta + \hat{\beta}^2 + \lambda|\beta|\end{aligned}$$

Then we take the partial derivative w.r.t. $\hat{\beta}$ and set it to zero considering the fact that the derivation of absolute value function is sign function, so:

$$\begin{aligned}\frac{\partial J}{\partial \hat{\beta}} &= -2\beta + 2\hat{\beta} + \text{sign}(\beta)\lambda = 0 \\ &\implies \hat{\beta} = \beta - \text{sign}(\beta)\lambda \\ &\implies \hat{\beta} = \text{sign}(\beta)|\beta| - \text{sign}(\beta)\lambda \\ &\implies \hat{\beta} = \text{sign}(\beta)(|\beta| - \lambda)_+\end{aligned}$$

As it can be seen easily, depending on the sign of the β the estimator will be shifted vertically by λ units like the given figure. In Figure 4, the estimated and selected features for some values of λ are shown.

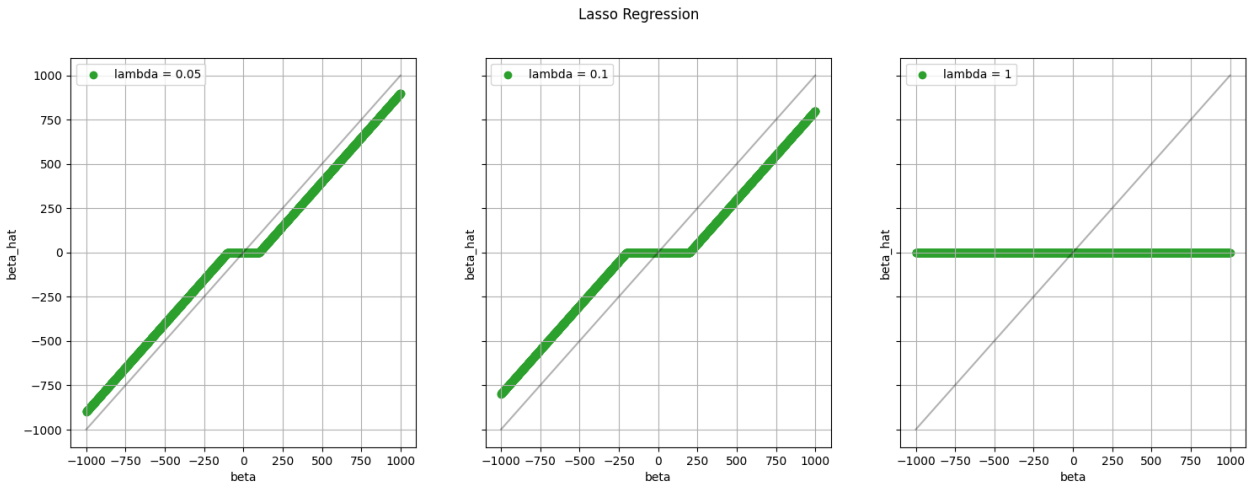

Figure 4: Regularized features by Lasso against default features

As an additional method, I tried the Elastic Net which is basically the mixture of the Ridge and the Lasso, and adds both L1 and L2 norms to the objective function. The result is depicted in the .ipynb notebook.

Please note that in my solution I used the symbol β for the parameters of the ordinary least squares problem, but in the question, it is denoted as $\hat{\beta}$, so the derived formulas are essentially the same things with slightly different notation.

Problem 8

a) From the question, we know that $x_i \sim \mathcal{N}(\mu, \sigma_i^2)$, and all estimators are unbiased, so it makes sense to assume that estimators whose σ_i^2 are smaller, contribute more effectively to estimate the true mean. On the other hand, estimators with high variance could yield values that are very far from the true mean. Therefore, the importance of x_i is $\frac{1}{\sigma_i^2}$. With this intuition, we calculate a weighted mean as the proposed estimator denoted as $\hat{\mu}$.

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\ &= \frac{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\end{aligned}$$

In one extreme case, suppose for all estimators the variance is the same ($\forall i\ \sigma_i = \sigma$).

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{1}{\sigma^2} x_i}{\sum_{i=1}^{n} \frac{1}{\sigma^2}} = \frac{\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i}{\frac{n}{\sigma^2}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

The proposed estimator in this case equals to the empirical mean. It actually is a reasonable choice because when all variances are the same, no estimator is more likely to have additional information about the true mean, so we average them with the same weight.

On the other extreme, suppose one estimator has a bounded variance, and the others have unbounded ($\sigma_k = \sigma, \forall i \neq k \ \sigma_i = \lim_{x\to\pm\infty} x = \pm\infty$).

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} = \frac{\frac{1}{\sigma} x_k + (n-1)0}{\frac{1}{\sigma} + (n-1)0} = x_k$$

The estimator is a good choice in this case as well. Because it used the only available information that we had about the true mean, which was x_k, and ignored the other ones. Another important fact about this estimator is that its expected value equals the true mean, and it still is an unbiased estimator for μ.

b) We assume the parameters as a vector such as $\theta = [\mu, \sigma_1, ..., \sigma_n]$, and for the maximum likelihood we have:

$$\begin{aligned}
\underset{\theta}{\operatorname{argmax}} \{f(X \mid \theta)\} &= \underset{\theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{\frac{-(x_i-\mu)^2}{\sigma_i^2}} \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{\frac{-(x_i-\mu)^2}{\sigma_i^2}} \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log \left(\frac{e^{\frac{-(x_i-\mu)^2}{\sigma_i^2}}}{\sqrt{2\pi}\sigma_i} \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \left(-\log(\sqrt{2\pi}) - \log(\sigma_i) + \frac{-(x_i-\mu)^2}{\sigma_i^2} \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \left(-\log(\sigma_i) + \frac{-(x_i-\mu)^2}{\sigma_i^2} \right) \right\}
\end{aligned}$$

To find the critical points of a function $f' : \mathbb{R}^{n+1} \to \mathbb{R}$, we take the partial derivative w.r.t. each input variable separately, set them to zero, and then test their second partial derivative.

$$\begin{aligned}
\frac{\partial f'}{\partial \mu} &= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \left(-\log(\sigma_i) + \frac{-(x_i-\mu)^2}{\sigma_i^2} \right) \\
&= \sum_{i=1}^{n} \left(0 + \left(\frac{2(x_i-\mu)}{\sigma_i^2} \right) \right) \\
&\implies 0 = \sum_{i=1}^{n} \frac{2(x_i-\mu)}{\sigma_i^2} \\
&\implies n\mu = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \implies \mu = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{n}
\end{aligned}$$

And for σ_i's:

$$\begin{aligned}
\frac{\partial f'}{\partial \sigma_i} &= \frac{\partial}{\partial \sigma_i} \sum_{i=1}^{n} \left(-\log(\sigma_i) + \frac{-(x_i - \mu)^2}{\sigma_i^2} \right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{\sigma_i} + \frac{2(x_i - \mu)^2}{\sigma_i^3} \right) \\
&\implies 0 = \sum_{i=1}^{n} \left(-\frac{1}{\sigma_i} + \frac{2(x_i - \mu)^2}{\sigma_i^3} \right) = \sum_{i=1}^{n} \left(\frac{-\sigma_i^2 + 2(x_i - \mu)^2}{\sigma_i^3} \right) \\
&\implies 0 = \sum_{i=1}^{n} \left(-\sigma_i^2 + 2(x_i - \mu)^2 \right) \\
&\implies \sum_{i=1}^{n} \sigma_i^2 = \sum_{i=1}^{n} 2(x_i - \mu)^2
\end{aligned}$$

There is no proper analytical solution for σ_i. We should try numerical methods to estimate their values, so in the Problem 8 section of the PS1.ipynb on my GitHub, I used stochastic gradient descent to numerically find the maximum likelihood for an instance of this problem.

c) For this part we use the Bayes formula as follows. Please note that we consider the normalization constant, namely $f(x)$ to be k.

$$\begin{aligned}
f(\theta \mid X) &= \frac{f(X \mid \theta) f(\theta)}{f(X)} \\
&= \frac{f(X \mid \theta) f(\mu) f(\sigma)}{k} \\
&= \frac{\prod_{i=1}^{n} \frac{\lambda_i}{\sqrt{2\pi}} exp\left\{ -\lambda_i (x_i - \mu)^2 \right\} f(\sigma) f(\mu)}{k} \\
&= \frac{\prod_{i=1}^{n} \frac{\lambda_i}{\sqrt{2\pi}} exp\left\{ -\lambda_i (x_i - \mu)^2 \right\} c\lambda_i^{a-1} exp\left\{ -b\lambda_i \right\} f(\mu)}{k} \\
&= \frac{\prod_{i=1}^{n} c\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} f(\mu)}{k.(2\pi)^{\frac{n}{2}}} \\
&= \frac{\prod_{i=1}^{n} \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \right) f(\mu)}{\frac{1}{c^n} k.(2\pi)^{\frac{n}{2}}} \\
&= \frac{\prod_{i=1}^{n} \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \cdot \text{Uniform}(-\infty, +\infty) \right)}{\frac{1}{c^n} k.(2\pi)^{\frac{n}{2}}} \\
&= \frac{\prod_{i=1}^{n} \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \right)}{C}
\end{aligned}$$

A Gaussian distribution with unbounded variance is just a Uniform distribution on the whole $\mathbb{R}$. It did not depend on i, so we took it out of the production and assumed all the constants in the denominator as C. We find the argmax of this function, so we take the logarithm as below.

$$
\begin{aligned}
\underset{\theta}{\operatorname{argmax}} \{f(\theta \mid X)\} &= \underset{\theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \log \left(\prod_{i=1}^{n} \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \right) \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log \left(\lambda_i^a exp\left\{ -\lambda_i (x_i - \mu)^2 - b\lambda_i \right\} \right) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} a \log(\lambda_i) + \left(-\lambda_i (x_i - \mu)^2 - b\lambda_i \right) \right\}
\end{aligned}
$$

In this part, I wrote a function to numerically compute the optimal value for μ and σ_i instead of algebraically calculating the formula for them.

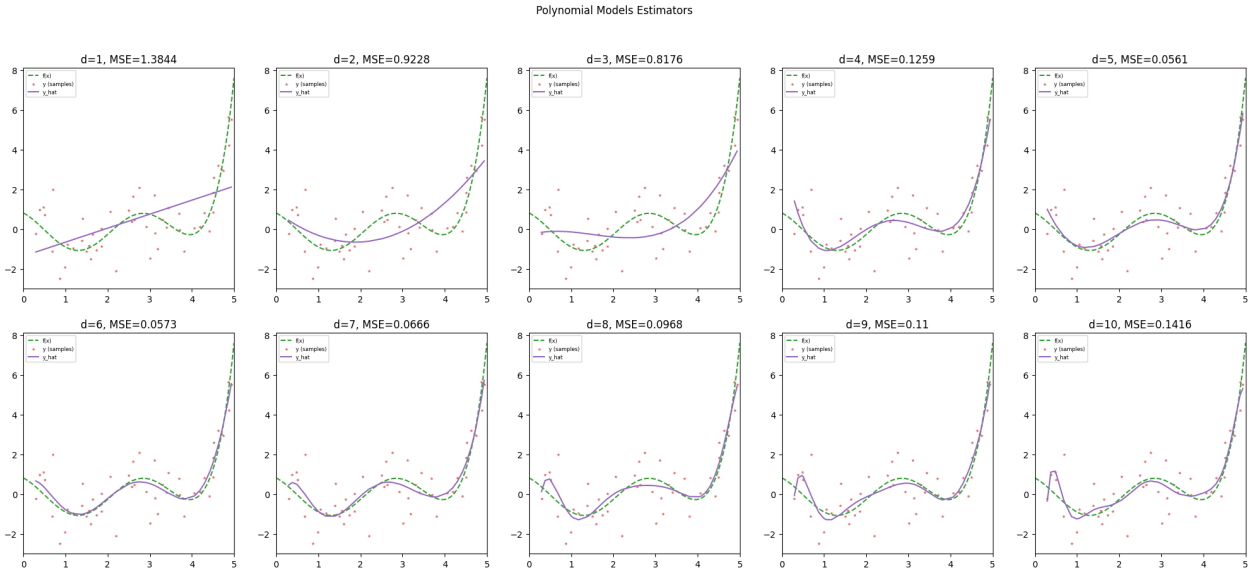

Figure 5: Polynomial Models for Estimating $f(x)$

Problem 9

The following result is produced by the Problem 9 section of the PS1.ipynb on GitHub.

The function that is used as $f(x)$ is $\frac{\frac{1}{2}e^{2x} - 1000\sin(2x-1) - e^{x+3}}{1000} + \mathcal{N}(0,1)$ which we look at from 0 to 1.

a) In Figure 5, we can see the fitted model with polynomial features for $d \in \{1, 2, ..., 10\}$. In lower degrees, the model has high bias and underfit related to the actual function. On the other extreme, the very high-degree models, have a high variance and are prone to overfit the samples instead of estimating the actual function. Based on the mean squared error between $f(x)$ and $\hat{f}(x)$, the best model is $d = 5$, however, this is not the general case and with other configurations, it could change; In conclusion, it is reasonable that assume for this function degrees around 5 can generate a good estimator.

b) In Figure 6, we used trigonometric functions for fitting the model, namely $a\sin(dx) + b\cos(dx)$. In this case, the best error value is achieved by $d = 2$ which consists of powers up to two of sin and cos functions.

c) To evaluate a proper measure for different values of variable d, we simulate the model fitting experiments with different start points, 10000 times. Then, is calculated and shown for all d's. The results of polynomial and trigonometric polynomial models are shown in Figure 7

For the polynomial models, the best value of d is 7. Generally, all models have roughly the same order of errors on this dataset. On the other hand, the best value for trigonometric polynomial models is $d = 4$. However, for $d = 10$ the error is extremely high which shows us that the model highly overfitted the data. Another considerable point is that for one experiment that we did in Figure 6 the best value of d was 2, but in general, $d = 4$ could be a better choice.

Problem 10

My code for this problem is available on the Problem 10 section of the PS1.ipynb on my GitHub.

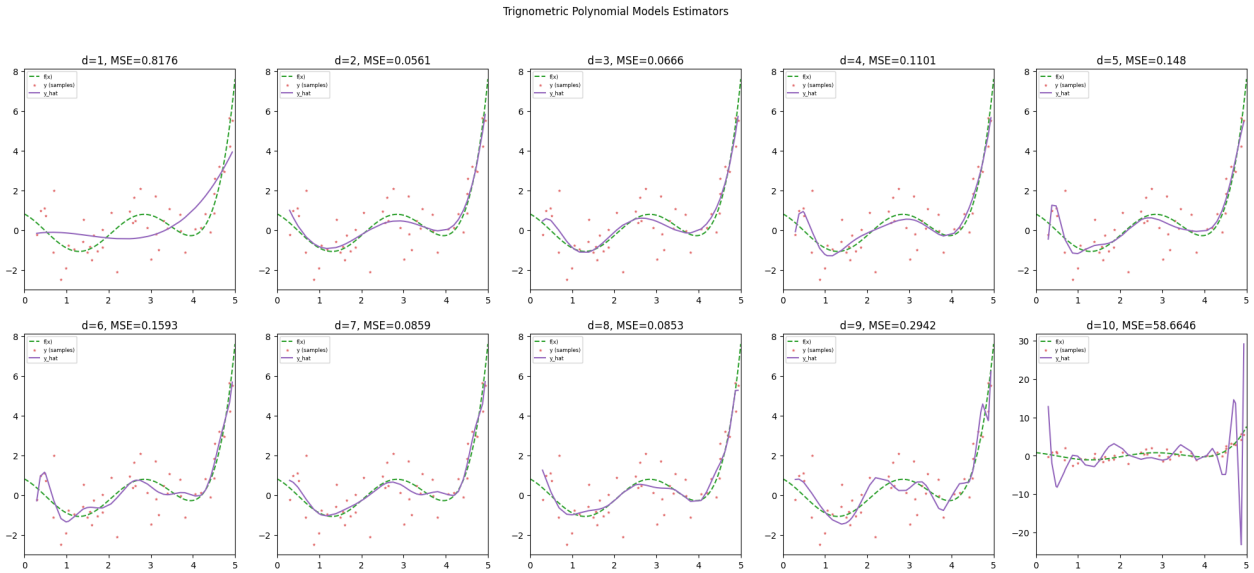

Figure 6: Trigonometric Polynomial Models for Estimating $f(x)$

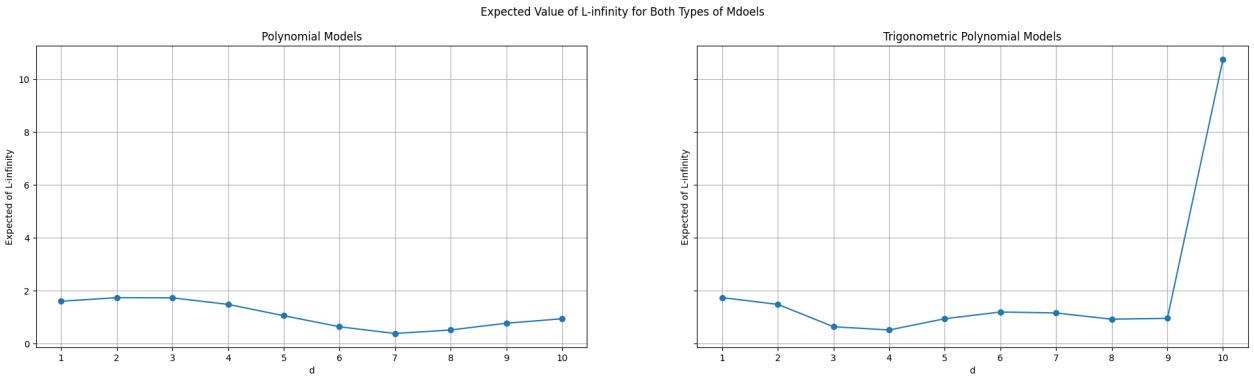

Figure 7: $\mathbb{E}(\|\hat{f} - f\|_{\text{inf}})$ for both models for $d \in \{1, 2, ..., 10\}$