1. (a)

I did this homework with Luning Zhao.

We work on our homework seperately, but we discuss when meeting problems.

Homework is fine.

(b) I certify that all solutions are entirely in my words. and that I have not looked at another student's solutions. I have credited all external sources in this write up.

2. (a) Rewrite $\min_w \|y - Xw\|^2$ to unconstrained problem:

subject to $\|w\|^2 \le \beta^2$

$$L(w, \varepsilon) = \min_w \left(\|y - Xw\|^2 + \varepsilon \left(\|w\|^2 - \beta^2 \right) \right)$$

(b) Increasing β means $\|w\|^2$ can go to larger value,

In the ridge regression context, this means λ can go smaller.

(c) change in the $w^T x$ is $w^T(x+\varepsilon) - w^T x = w^T \varepsilon \le \|w\| \|\varepsilon\|$

(d) $f(w) = \|y - Xw\|^2 + \lambda \|w\|^2$

$$= (y - Xw)^T (y - Xw) + \lambda w^T w$$

$$= y^T y - y^T X w - w^T X^T y + w^T X^T X w + \lambda w^T w$$

$$\frac{\partial f(w)}{\partial w} = -y^T X - y^T X + w^T X^T X + w^T X^T X + \lambda w^T + \lambda w^T = 0$$

$$-2 y^T X + 2 w^T X^T X + 2\lambda w^T = 0$$

$$w^T (X^T X + \lambda I) = y^T X$$

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

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

(e) If σ^2 is eval for $X^T X$, then $(X^T X)\vec{z} = \sigma^2 \vec{z}$.

For $(X^T X + \lambda I)$, $(X^T X + \lambda I)\vec{z} = X^T X \vec{z} + \lambda I \vec{z}$

$$= (\sigma^2 + \lambda)\vec{z}.$$

so $\sigma^2 + \lambda$ is eval for $(X^T X + \lambda I)$.

The evals for $(X^T X + \lambda I)$ will be $\sigma_1^2 + \lambda, \cdots, \sigma_d^2 + \lambda$.

If some eval for $X^T X$ $\sigma_i^2 << 0$, then it's numerically difficult to calculate $(X^T X)^{-1}$.

but $\sigma_i^2 + \lambda$ will not be much smaller than 0. so it's much easier to calculate $(X^T X + \lambda I)^{-1}$

(f) I think $\lambda_2 = 0.5$ is a better choice.

First, with $\lambda = 0.5$, all eigenvalues for $(X^TX + \lambda I)$ will not be closer to zero, so we can easily calculate $(X^TX + \lambda I)^{-1}$

Second, with $\lambda = 100$, $\|w\|^2$ will be restricted to a small region, and $(X^TX + \lambda I)^{-1}$ will be very small.

(g) because $X \in R^{4\times5}$, X must have nullspace w_0 so that $Xw_0 = 0$.

The reason is the 5 columns cannot be linear independent when they only have 4 numbers in each column.

So any vector $\vec{w}(\alpha)$ can be written as $\vec{w}(\alpha) = \vec{w_0} + X^T\alpha$.

where $\vec{w}$ is in X nullspace and $X^T\alpha$ is not. $(\vec{w_0}^T X^T\alpha = (X\vec{w_0})^T\alpha = 0)$

For OLS, $f(w) = \|Xw - y\|^2 = (Xw-y)^T(Xw-y)$

$$= w^TX^TXw - w^TX^Ty - y^TXw + y^Ty$$

$$= (w_0 + X^T\alpha)^T X^TX (w_0 + X^T\alpha) - (w_0 + X^T\alpha)^T X^Ty$$
$$- y^TX(w_0 + X^T\alpha) + y^Ty$$

$$= \cancel{w_0^TX^TXw_0} + \alpha^TX\cancel{X^TXw_0} + \cancel{w_0^TX^TX}X^T\alpha + \alpha^TXX^TXX^T\alpha$$
$$- \cancel{w_0^TX^T}y - \alpha^TXX^Ty - \cancel{y^TXw_0} - y^TXX^T\alpha + y^Ty$$

$$= \alpha^TXX^TXX^T\alpha - \alpha^TXX^Ty - y^TXX^T\alpha + y^Ty.$$

$f(w)$ is independent of w_0, so $\frac{\partial f(w)}{\partial w_0} = 0 \Rightarrow w_0$ is arbitrary.

so OLS has infinite number of solutions.

For ridge regression, $f(w) = \|Xw - y\|^2 + \lambda\|w^2\|$

$$= \alpha^TXX^TXX^T\alpha - \alpha^TXX^Ty - y^TXX^T\alpha + y^Ty$$
$$+ \lambda[w_0^Tw_0 + \alpha^T\cancel{X}w_0 + w_0^T\cancel{X^T\alpha} + \alpha^TXX^T\alpha]$$

$$= \alpha^T X X^T X X^T \alpha - \alpha^T X X^T y - y^T X X^T \alpha + y^T y$$
$$+ \lambda w_0^T w_0 + \lambda \alpha^T X X^T \alpha.$$

$$\frac{\partial f(w)}{\partial w_0} = \lambda w_0^T + \lambda w_0^T = 2\lambda w_0^T = 0 \Rightarrow w_0 = 0.$$

$\Rightarrow w = X^T \alpha$ is the unique solution for ridge regression.

(h) If $\lambda \to 0$, ridge regression will be the same as OLS.

(i) $f(w) = \frac{1}{2} \|y - Xw\|^2 + \lambda \|\Gamma w\|^2$

$$= \frac{1}{2} [(y-Xw)^T (y-Xw)] + \lambda [(\Gamma w)^T \Gamma w]$$

$$= \frac{1}{2} (y^T y - y^T X w - w^T X^T y + w^T X^T X w) + \lambda (w^T \Gamma^T \Gamma w)$$

$$\frac{\partial f(w)}{\partial w} = \frac{1}{2} (-y^T X - y^T X + w^T X^T X + w^T X^T X) + \lambda (w^T \Gamma^T \Gamma + w^T \Gamma^T \Gamma)$$

$$= -y^T X + w^T X^T X + 2\lambda w^T \Gamma^T \Gamma = 0$$

$$w^T (X^T X + 2\lambda \Gamma^T \Gamma) = y^T X$$

$$w^T = y^T X (X^T X + 2\lambda \Gamma^T \Gamma)^{-1}$$

$$w = (X^T X + 2\lambda \Gamma^T \Gamma)^{-1} X^T y$$

3. (a) $F = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$ $\quad x_1 \neq x_2 \Rightarrow \text{rank}(F) = 2$

(b) $F' = \begin{bmatrix} 1 & x_1 & \cdots & x_1^D \\ 0 & x_2 - x_1 & \cdots & x_2^D - x_1^D \\ \vdots & & & \\ 0 & x_n - x_1 & \cdots & x_n^D - x_1^D \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ -1 & 0 & 1 \cdots & 0 \\ -1 & 0 & 0 \cdots & 1 \end{bmatrix}}_{A} F$

$\det(F') = \det(AF) = \det(A)\det(F)$

$\det(A) = 1 \cdot I - 0 = 1 \Rightarrow \det(F') = \det(F)$

(c) $\begin{bmatrix} 1 & x_1 & x_1^2 \cdots & x_1^D \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 \cdots & x_2^D - x_1^D \\ \vdots & & & \\ 0 & x_n - x_1 & x_n^2 - x_1^2 \cdots & x_n^D - x_1^D \end{bmatrix} \overset{P_0}{\Rightarrow}$ i) $\Rightarrow \overset{P_1}{\begin{bmatrix} 1 & x_1 & \cdots & x_1^{D-1} & 0 \\ 0 & x_2 - x_1 & & x_2^{D-1} - x_1^{D-1} & x_2^{D-1}(x_2 - x_1) \\ \vdots & \vdots & & \vdots & \\ 0 & x_n - x_1 & & x_n^{D-1} - x_1^{D-1} & x_n^{D-1}(x_n - x_1) \end{bmatrix}}$

ii) $\Rightarrow \begin{bmatrix} 1 & x_1 & \cdots & 0 & 0 \\ 0 & x_2 - x_1 & & x_2^{D-2}(x_2 - x_1) & x_2^{D-1}(x_2 - x_1) \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & x_n - x_1 & & x_n^{D-2}(x_n - x_1) & x_n^{D-1}(x_n - x_1) \end{bmatrix}$

$\Rightarrow \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & x_2 - x_1 & x_2^2 - x_2 x_1 & & x_2^{D-1}(x_2 - x_1) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & x_n - x_1 & x_{2n}^2 - x_n x_1 & & x_n^{D-1}(x_n - x_1) \end{bmatrix}$

$P_1 = P_0 \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 1 & -x_1 \\ & & & & 1 \end{bmatrix}}_{A}$ $\quad \det(A) = 1 \Rightarrow \det(F'') = \det(F')$.

(d) If $A\vec{x}_1 = \lambda_1 \vec{x}_1$, then $\begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix}\begin{bmatrix} 0 \\ \vec{x}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ A\vec{x}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_1 \vec{x}_1 \end{bmatrix} = \lambda_1 \begin{bmatrix} 0 \\ \vec{x}_1 \end{bmatrix}$

So $\lambda_1 \cdots \lambda_d$ is evals for B.

$\begin{bmatrix} 1 & 0^T \\ 0 & A \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ so 1 is the other eval for B.

$\det(B) = 1 \det(A) - 0 = \det(A)$.

(e) $\det(F') = \det \begin{bmatrix} x_2 - x_1 & x_2(x_2 - x_1) & \cdots & x_2^{D-1}(x_2 - x_1) \\ \vdots & & & \\ x_n - x_1 & x_n(x_n - x_1) & \cdots & x_n^{D-1}(x_n - x_1) \end{bmatrix}$

$= \prod_{i=2}^{n} (x_i - x_1) \det \begin{bmatrix} 1 & x_2 & \cdots & x_2^{D-1} \\ \vdots & & & \\ 1 & x_n & \cdots & x_n^{D-1} \end{bmatrix}$

$= \prod_{i=2}^{n} (x_i - x_1) \cdot \det [P_{D-1}(x_2), P_{D-1}(x_3) \cdots P_{D-1}(x_n)]^T$

$= \prod_{1 \le i < j \le n} (x_j - x_i) \neq 0 \Rightarrow \det(F') \neq 0 \Rightarrow$ full rank.

(f) Using stars and bars argument, we first make $D+\ell$ $*$s.

Then we mark ℓ bars, the number of $*$s after each bar is the exponents for x_i.

For example, $\ell = 2$, $D = 3$.

$\times \boxed{\times} \times \boxed{\times} \times \Rightarrow x_1 x_2$

$\boxed{\times} \times \boxed{\times} \times \times \Rightarrow x_1 x_2^2$

So size of $P_D(x)$ is $\binom{D+\ell}{\ell}$.

(g) $F_\ell = \begin{bmatrix} 1 & x_1 & x_1 & \cdots & x_1 & x_1^2 & \cdots & x_1^2 & \cdots & x_1^D & x_1^D & \cdots & x_1^D \\ 1 & x_2 & x_2 & \cdots & x_2 & x_2^2 & \cdots & x_2^2 & \cdots & x_2^D & \cdots & & \\ \vdots & & & & & & & & & & & & \end{bmatrix}$

$Col_2 = Col_3 \Rightarrow F_\ell$ always has linearly dependent columns.

When $D=1$, F_ℓ is more likely linear independent compared with large D.

(h) To make F_ℓ full rank, we should try to make x_i different from each other.

4. (a) i) $e^x = e^0 + f'(x)\,x = 1 + e^0 \cdot x = 1 + x$

$e^x = e^0 + \frac{f'(x)}{1}x + \frac{f''(x)}{2!}x^2 = 1 + x + \frac{1}{2}x^2$

$e^x = e^0 + \frac{f'(x)}{1}x + \frac{f''(x)}{2!}x^2 + \frac{f'''(x)}{3!}x^3 = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3$

$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4$

ii) $\sin x = \sin 0 + \cos 0 \cdot x = x$

$\sin x = x + (-\sin 0)/2!\, x^2 = x$

$\sin x = x - \frac{\cos 0}{3!}x^3 = x - \frac{1}{6}x^3$

$\sin x = x - \frac{1}{6}x^3 + \frac{\sin 0}{4!}x^4 = x - \frac{1}{6}x^3$

(b) see plot in next page.

- When $m=2$, $\|f - \phi_m\|_\infty = \sup_{x \in I} |e^x - 1 - x - \frac{1}{2}x^2|$

  From the plot we can see that $\|f - \phi_m\|$ gets maximum when $x = 3$

  so $\|f - \phi_m\|_\infty = |e^3 - 1 - 3 - \frac{1}{2}9| = |e^3 - 8.5| \approx 11.6$

- For non-zero integer m,

  $|f - \phi_m| = |e^x - \sum_{i=0}^{m} \frac{1}{i!}x^i| \le \frac{T|x|^{m+1}}{(m+1)!}$

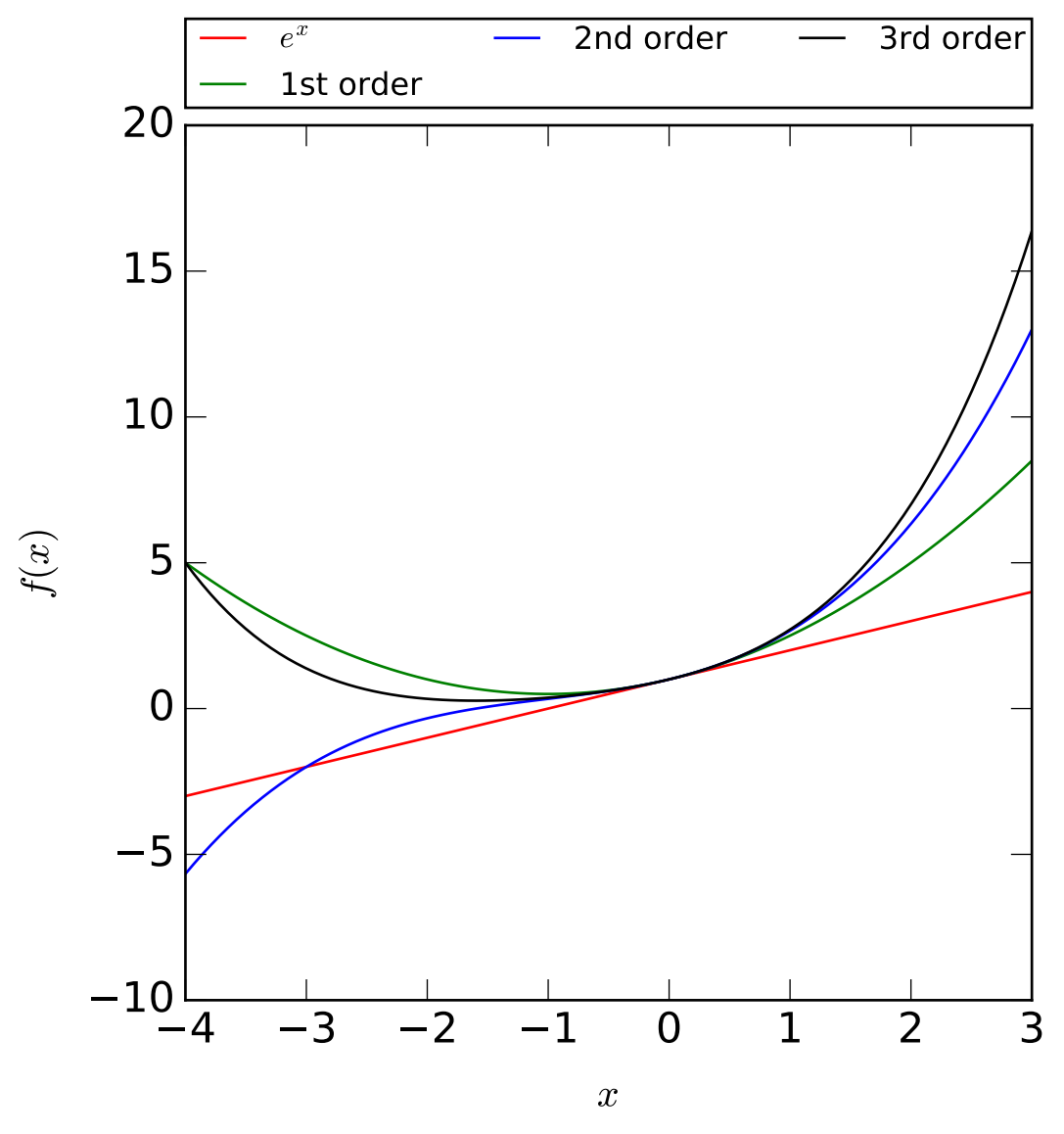
  where $T = \max |f^{(m)}x| = \max |e^x| = e^3$.

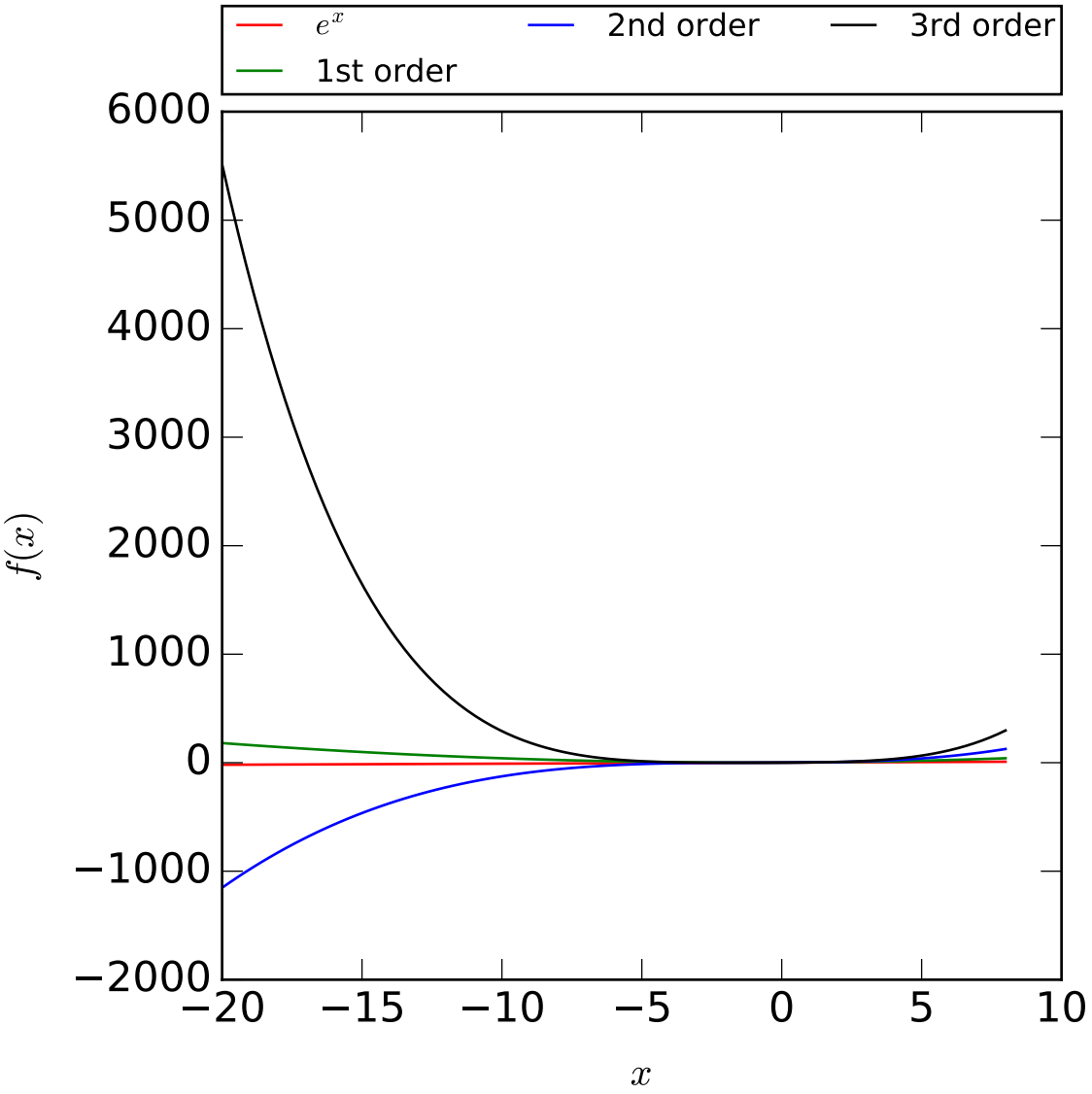
  So $|f - \phi_m| \le \frac{e^3 \cdot 4^{m+1}}{(m+1)!}$

- $\lim_{m\to\infty} \frac{e^3 \cdot 4^{m+1}}{(m+1)!} = \lim_{m\to\infty} \frac{e^3 \cdot 4^{m+1}}{\sqrt{2\pi(m+1)}\left(\frac{m+1}{e}\right)^{m+1}} = \lim_{m\to\infty} \frac{e^3}{\sqrt{2\pi(m+1)}}\left(\frac{4e}{m+1}\right)^{m+1} = 0$

- approximation error becomes very large outside the bounded interval I.

[ADD PIC]


e^x
2nd order
3rd order
1st order
20
15
10
5
0
−5
−10
−4
−3
−2
−1
0
1
2
3
$f(x)$
x


e^x
2nd order
3rd order
1st order
6000
5000
4000
3000
2000
1000
0
−1000
−2000
−20
−15
−10
−5
0
5
10
$f(x)$
x

(c) For $f(x) = e^x$, $|f(x) - \phi_D(x)| \le \frac{e^3 \cdot 4^{D+1}}{(D+1)!} \le \varepsilon$

$\Rightarrow \quad 3 + (D+1)\ln 4 - \ln\sqrt{2\pi(D+1)}\left(\frac{D+1}{e}\right)^{D+1} \le \ln\varepsilon$

$\Rightarrow -3 - (D+1)\ln 4 + \ln\sqrt{2\pi(D+1)} + (D+1)\ln(D+1) - D + 1 \ge \ln\frac{1}{\varepsilon}$

$\Rightarrow \quad (D+1)\ln(D+1) - (\ln 4 + 1)(D+1) + \frac{1}{2}\ln(D+1) \ge O\left(\ln\frac{1}{\varepsilon}\right)$

$\Rightarrow \quad D\ln D + \frac{1}{2}\ln D - 2D \ge O\left(\ln\frac{1}{\varepsilon}\right)$

when $\varepsilon << 1$, $\ln\left(\frac{1}{\varepsilon}\right) >> 1$,

If $D > O\left(\ln\frac{1}{\varepsilon}\right)$, the $D\ln D \ge O\left(\ln\frac{1}{\varepsilon}\right)$.

For $f(x) = \sin x$, $|f(x) - \phi_D(x)| \le \frac{1 \cdot 4^{D+1}}{(D+1)!} \le \varepsilon$,

Similar as $f(x) = e^x$, we can prove if $D > O(\log\frac{1}{\varepsilon})$, $|f(x) - \phi_D(x)| \le \varepsilon$.

(d) $\|f - \phi_D\|_\infty \le \frac{T|x - x_0|^{D+1}}{(D+1)!} \le \varepsilon$.

$\Rightarrow \quad \frac{T|x - x_0|^{D+1} \cdot e^{D+1}}{\sqrt{2\pi(D+1)}\,(D+1)^{D+1}} = \frac{T}{\sqrt{2\pi}} \frac{(e|x - x_0|)^{D+1}}{(D+1)^{D+\frac{3}{2}}} \le \varepsilon$.

because T, x, x_0 are bounded, when $D \to \infty$, $\frac{T}{\sqrt{2\pi}} \frac{(e|x - x_0|)^{D+1}}{(D+1)^{D+\frac{1}{2}}} \to 0$.

so for any $\varepsilon > 0$, there is a $D \ge 1$ s.t $\|f - \phi_D\|_\infty < \varepsilon$

(e) $f(\vec{v}) = f(\vec{v}_0) + \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} +$ [illegible]

$$+ \frac{1}{2!}\begin{bmatrix} x - x_0 & y - y_0 \end{bmatrix} \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$$

$= f(\vec{v}_0) + \nabla f^T(\vec{v} - \vec{v}_0) + \frac{1}{2!}(\vec{v} - \vec{v}_0)^T \nabla^2 f (\vec{v} - \vec{v}_0)$

Here, $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$ $\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$

· For $f(\vec{v}) = e^x y^2$, $\frac{\partial f}{\partial x} = e^x y^2$, $\frac{\partial f}{\partial y} = 2e^x y$

$\frac{\partial^2 f}{\partial x^2} = e^x y^2$ $\frac{\partial^2 f}{\partial x \partial y} = 2e^x y$ $\frac{\partial^2 f}{\partial y \partial x} = 2e^x y$ $\frac{\partial^2 f}{\partial y^2} = 2e^x$

$\Rightarrow f(\vec{v}) = e^{x_0} y_0^2 + \begin{bmatrix} e^{x_0} y_0^2 & 2e^{x_0} y_0 \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$

$+ \frac{1}{2} \begin{bmatrix} x - x_0 & y - y_0 \end{bmatrix} \begin{bmatrix} e^{x_0} y_0^2 & 2e^{x_0} y_0 \\ 2e^{x_0} y_0 & 2e^{x_0} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$

(f) $\vec{v}(t) = t\vec{v} = t \begin{bmatrix} x \\ y \end{bmatrix}$

$g(t) = f(\vec{v}(t)) = e^{\cancel{t x_0}} (t y_0)^2 + \begin{bmatrix} e^{\cancel{t x_0}} (t y_0)^2 & 2e^{\cancel{t x_0}} \cancel{t y_0} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$

$+ \frac{1}{2} \begin{bmatrix} x - x_0 & y - y_0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$

$= \frac{2}{2} \begin{bmatrix} 0 & y - y_0 \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} = (y - y_0)^2 = t^2 y_1^2$

5. (a) Jaina's problem can be written as:

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^D \\ 1 & x_2 & x_2^2 & \cdots & x_2^D \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^D \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}}_{w} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{y}.$$

$$\arg\min_w \| Xw - y \|^2$$

(b) see output of b.py

(c) The average training error decreases with D, because with higher poly order, we have more model parameters, so we can fit data better.
If our model has n+1 parameters, $X \in R^{n \times n+1}$, so we can't find the inverse of $(X^T X)^{-1}$, In the end, we are over fitting with n poly order.

(d) Test error is always larger than training error. [plot see output of c.py]
Because w is minimized using train data, so bias will be minimized for training data.

(e) We should use the order when testing error becomes flat, which is poly order = 4.
(plot from d)

(f) See code

(g) See code.

7.

Q: Prove Sherman - Morrison formula.

Suppose $A \in R^{n \times n}$ is an invertible square matrix, and, $u, v \in R^n$ are column vectors. If $A + uv^T$ is in invertible, the

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

A: $(A + uv^T)\left(A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}\right)$

$$= AA^{-1} - \frac{uv^TA^{-1}}{1 + v^TA^{-1}u} + uv^TA^{-1} - \frac{uv^TA^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

$$= I + uv^TA^{-1} - \frac{u(1 + v^TA^{-1}u)v^TA^{-1}}{1 + v^TA^{-1}u}$$

$$= I + uv^TA^{-1} - uv^TA^{-1}$$

$$= I$$