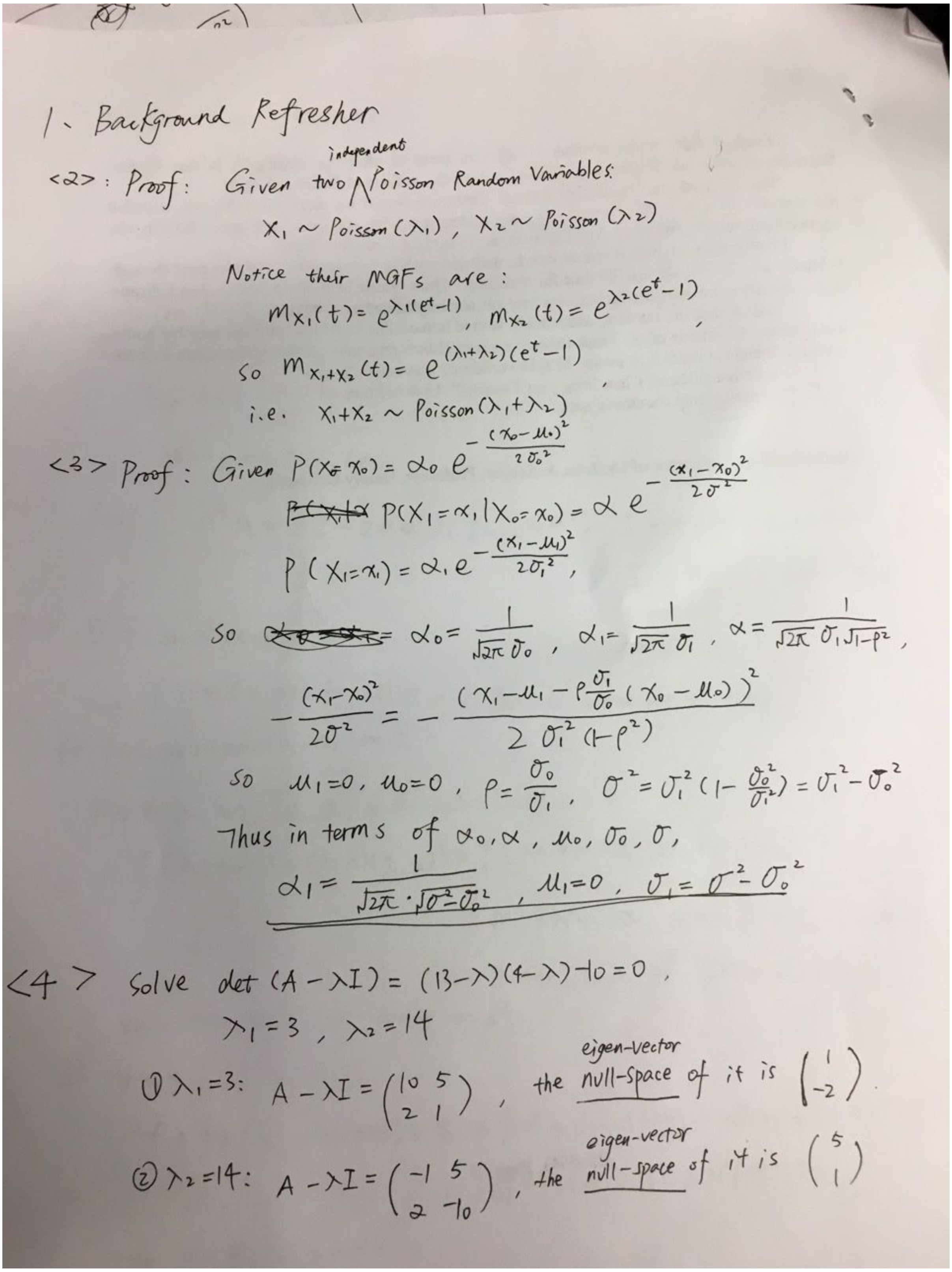
COMP540 HW1 Yue Pan NetIP:yp15 Zhong Lin NetID:zl44

Background Warm-up:



1、Background Refresher

<2> : Proof: Given two independent Poisson Random Variables:

$X_1 \sim Poisson(\lambda_1)$, $X_2 \sim Poisson(\lambda_2)$

Notice their MGFs are :

$m_{X_1}(t) = e^{\lambda_1(e^t-1)}$, $m_{X_2}(t) = e^{\lambda_2(e^t-1)}$,

so $m_{X_1+X_2}(t) = e^{(\lambda_1+\lambda_2)(e^t-1)}$,

i.e. $X_1+X_2 \sim Poisson(\lambda_1+\lambda_2)$

<3> Proof : Given $P(X_0 = x_0) = \alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}}$

~~$P(x_1|$~~ $P(X_1 = x_1 | X_0 = x_0) = \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}$

$P(X_1 = x_1) = \alpha_1 e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}$,

so $\alpha_0 = \frac{1}{\sqrt{2\pi}\sigma_0}$, $\alpha_1 = \frac{1}{\sqrt{2\pi}\sigma_1}$, $\alpha = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}$,

$$-\frac{(x_1-x_0)^2}{2\sigma^2} = -\frac{\left(x_1-\mu_1-\rho\frac{\sigma_1}{\sigma_0}(x_0-\mu_0)\right)^2}{2\sigma_1^2(1-\rho^2)}$$

so $\mu_1 = 0$, $\mu_0 = 0$, $\rho = \frac{\sigma_0}{\sigma_1}$, $\sigma^2 = \sigma_1^2(1-\frac{\sigma_0^2}{\sigma_1^2}) = \sigma_1^2 - \sigma_0^2$

Thus in terms of $\alpha_0, \alpha, \mu_0, \sigma_0, \sigma$,

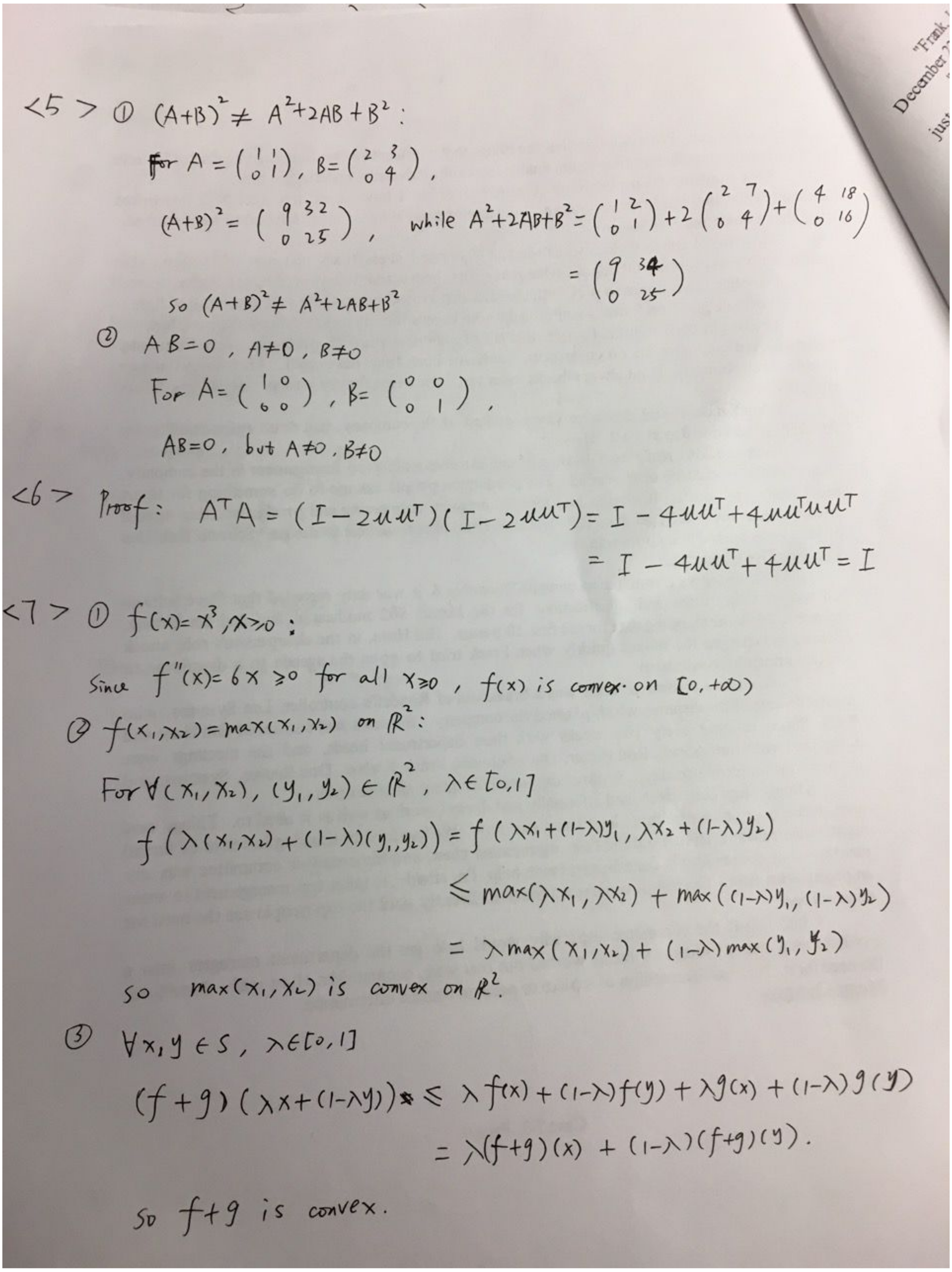
$$\underline{\underline{\alpha_1 = \frac{1}{\sqrt{2\pi}\cdot\sqrt{\sigma^2-\sigma_0^2}}, \ \mu_1 = 0, \ \sigma_1 = \sigma^2 - \sigma_0^2}}$$

<4> Solve $\det(A-\lambda I) = (13-\lambda)(4-\lambda)-10 = 0$,

$\lambda_1 = 3$, $\lambda_2 = 14$

① $\lambda_1 = 3$: $A - \lambda I = \begin{pmatrix} 10 & 5 \\ 2 & 1 \end{pmatrix}$, the $\overset{\text{eigen-vector}}{\underline{\text{null-space}}}$ of it is $\begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

② $\lambda_2 = 14$: $A - \lambda I = \begin{pmatrix} -1 & 5 \\ 2 & -10 \end{pmatrix}$, the $\overset{\text{eigen-vector}}{\underline{\text{null-space}}}$ of it is $\begin{pmatrix} 5 \\ 1 \end{pmatrix}$



<5> ① $(A+B)^2 \neq A^2+2AB+B^2$:

For $A=\begin{pmatrix}1&1\\0&1\end{pmatrix}$, $B=\begin{pmatrix}2&3\\0&4\end{pmatrix}$,

$(A+B)^2=\begin{pmatrix}9&32\\0&25\end{pmatrix}$, while $A^2+2AB+B^2=\begin{pmatrix}1&2\\0&1\end{pmatrix}+2\begin{pmatrix}2&7\\0&4\end{pmatrix}+\begin{pmatrix}4&18\\0&16\end{pmatrix}$

$=\begin{pmatrix}9&34\\0&25\end{pmatrix}$

so $(A+B)^2 \neq A^2+2AB+B^2$

② $AB=0$, $A\neq 0$, $B\neq 0$

For $A=\begin{pmatrix}1&0\\0&0\end{pmatrix}$, $B=\begin{pmatrix}0&0\\0&1\end{pmatrix}$,

$AB=0$, but $A\neq 0$, $B\neq 0$

<6> Proof: $A^TA=(I-2uu^T)(I-2uu^T)=I-4uu^T+4uu^Tuu^T$

$=I-4uu^T+4uu^T=I$

<7> ① $f(x)=x^3, x\geq 0$:

Since $f''(x)=6x\geq 0$ for all $x\geq 0$, $f(x)$ is convex. on $[0,+\infty)$

② $f(x_1,x_2)=\max(x_1,x_2)$ on $\mathbb{R}^2$:

For $\forall (x_1,x_2),(y_1,y_2)\in\mathbb{R}^2$, $\lambda\in[0,1]$

$f(\lambda(x_1,x_2)+(1-\lambda)(y_1,y_2))=f(\lambda x_1+(1-\lambda)y_1, \lambda x_2+(1-\lambda)y_2)$

$\leq \max(\lambda x_1,\lambda x_2)+\max((1-\lambda)y_1,(1-\lambda)y_2)$

$=\lambda\max(x_1,x_2)+(1-\lambda)\max(y_1,y_2)$

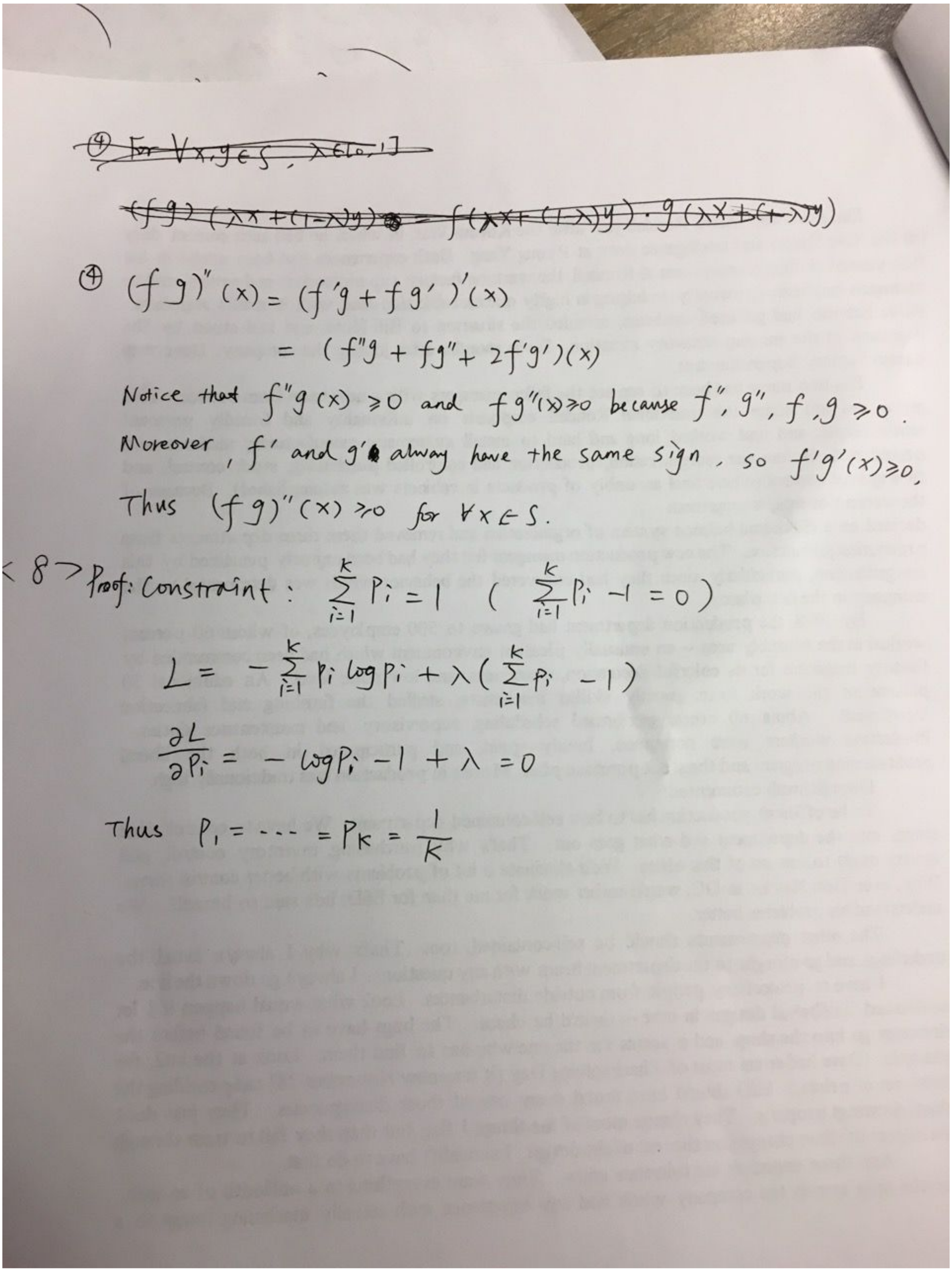
so $\max(x_1,x_2)$ is convex on $\mathbb{R}^2$.

③ $\forall x,y\in S$, $\lambda\in[0,1]$

$(f+g)(\lambda x+(1-\lambda y))\leq \lambda f(x)+(1-\lambda)f(y)+\lambda g(x)+(1-\lambda)g(y)$

$=\lambda(f+g)(x)+(1-\lambda)(f+g)(y)$.

so $f+g$ is convex.



~~④ for $\forall x, y \in S$, $\lambda \in [0,1]$~~

~~$(fg)(\lambda x+(1-\lambda)y) = f(\lambda x+(1-\lambda)y)\cdot g(\lambda x+(1-\lambda)y)$~~

④ $$(fg)''(x) = (f'g + fg')'(x)$$
$$= (f''g + fg'' + 2f'g')(x)$$

Notice that $f''g(x) \geq 0$ and $fg''(x) \geq 0$ because $f'', g'', f, g \geq 0$.

Moreover, f' and g' alway have the same sign, so $f'g'(x) \geq 0$.

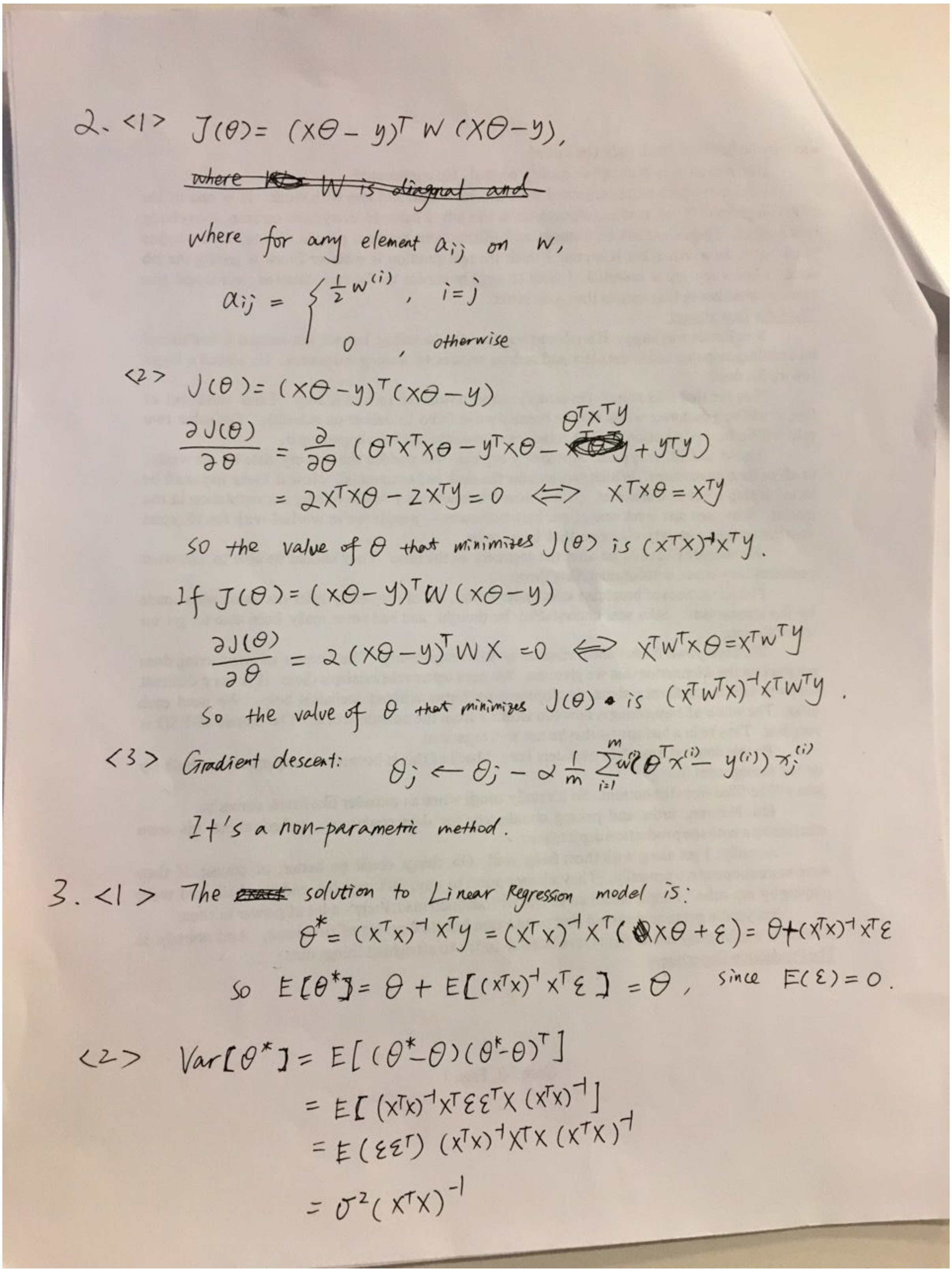
Thus $(fg)''(x) \geq 0$ for $\forall x \in S$.

<8> Proof: Constraint: $\sum_{i=1}^{k} P_i = 1$ $\left(\sum_{i=1}^{k} P_i - 1 = 0 \right)$

$$L = -\sum_{i=1}^{k} p_i \log p_i + \lambda\left(\sum_{i=1}^{k} p_i - 1\right)$$

$$\frac{\partial L}{\partial P_i} = -\log P_i - 1 + \lambda = 0$$

Thus $P_1 = \cdots = P_K = \frac{1}{K}$



2. <1> $J(\theta) = (X\theta - y)^T W (X\theta - y)$,

~~where W is diagnal and~~

where for any element a_{ij} on W,

$$a_{ij} = \begin{cases} \frac{1}{2} w^{(i)}, & i=j \\ 0, & \text{otherwise} \end{cases}$$

<2> $J(\theta) = (X\theta - y)^T (X\theta - y)$

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} (\theta^T X^T X\theta - y^T X\theta - \overset{\theta^T X^T y}{\cancel{}}y + y^T y)$$

$$= 2X^T X\theta - 2X^T y = 0 \iff X^T X\theta = X^T y$$

so the value of θ that minimizes $J(\theta)$ is $(X^T X)^{-1} X^T y$.

If $J(\theta) = (X\theta - y)^T W (X\theta - y)$

$$\frac{\partial J(\theta)}{\partial \theta} = 2(X\theta - y)^T W X = 0 \iff X^T W^T X\theta = X^T W^T y$$

So the value of θ that minimizes $J(\theta)$ is $(X^T W^T X)^{-1} X^T W^T y$.

<3> Gradient descent: $\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}$

It's a non-parametric method.

3. <1> The ~~exact~~ solution to Linear Regression model is:

$$\theta^* = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\theta + \varepsilon) = \theta + (X^T X)^{-1} X^T \varepsilon$$

so $E[\theta^*] = \theta + E[(X^T X)^{-1} X^T \varepsilon] = \theta$, since $E(\varepsilon) = 0$.

<2> $Var[\theta^*] = E[(\theta^* - \theta)(\theta^* - \theta)^T]$

$$= E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}]$$

$$= E(\varepsilon \varepsilon^T) (X^T X)^{-1} X^T X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

Part 1:

Problem 3.1.A3:

Predict value for lower status percentage of 5%:

```
pred_cost = linear_reg.predict(np.array([[1,5]])) * 10000
```

And we get the result:

```
For lower status percentage = 5, we predict a median home value of [
298034.49412207]
```

Predict values for lower status percentage of 50%:

```
pred_cost = linear_reg.predict(np.array([[1,50]])) * 10000
```

And we get the result:

```
For lower status percentage = 50, we predict a median home value of
[-129482.12889799]
```

Problem 3.1.B5:

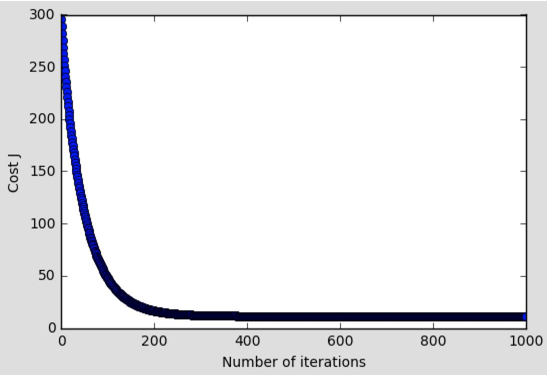


fig1 alpha = 0.01

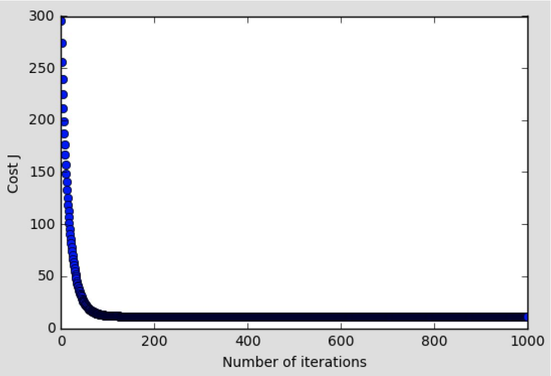


fig2 alpha = 0.03

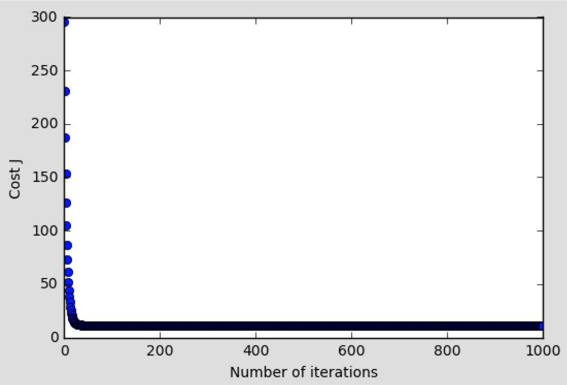


fig3 alpha = 0.1

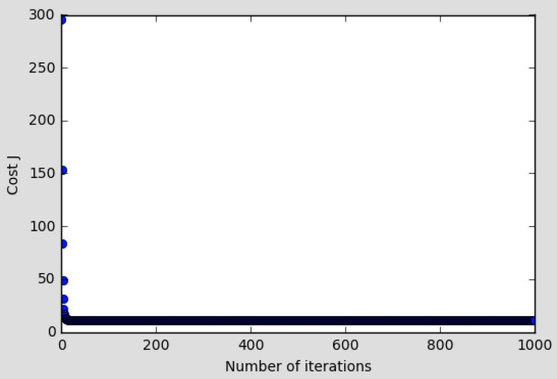


fig4 alpha = 0.3

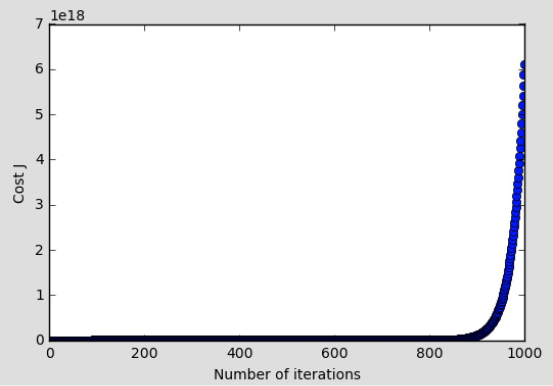


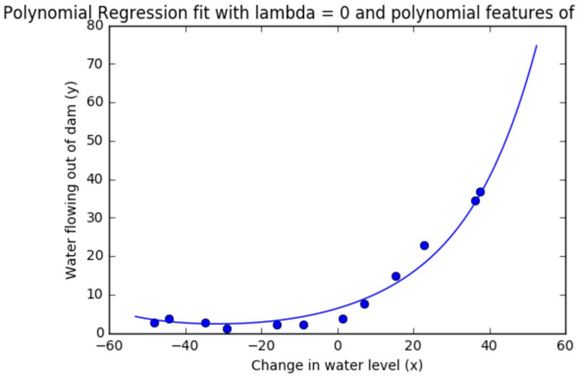
fig5   alpha = 0.33

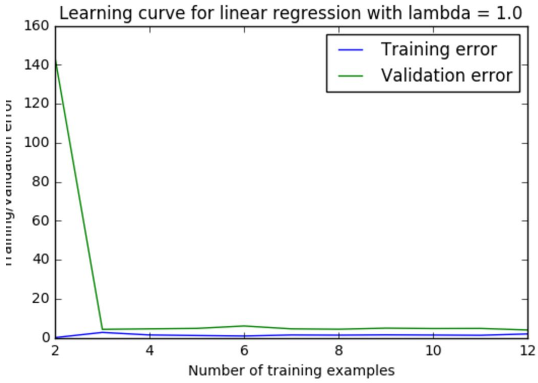
When alpha(learning rate) is at a certain low range, increasing the alpha value will significantly speed up the convergence rate. However, when alpha is larger than a certain value, the loss value is not converging anymore, it diverges instead. In this specific problem, we notice that the best alpha value should end up with around 0.3.

part2:

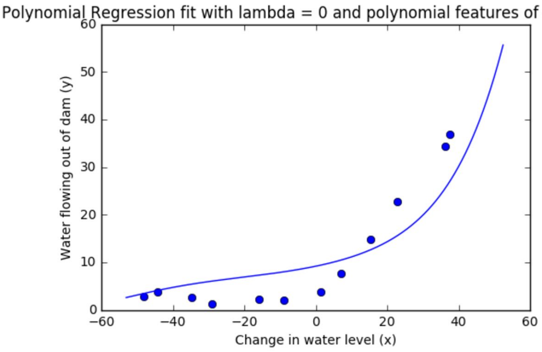
A4: Adjusting the regularization parameter

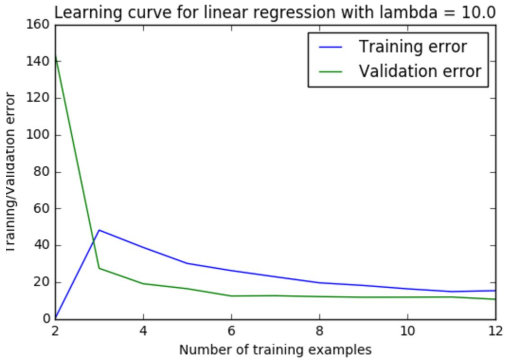
1) lambda= 1



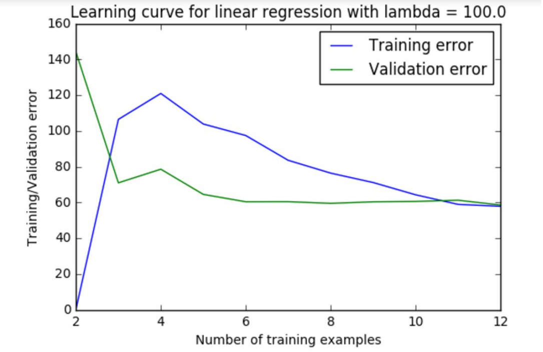


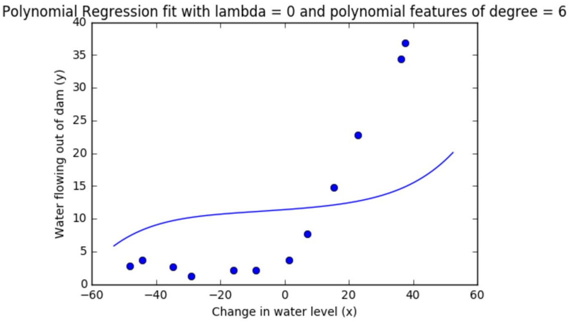
2) lambda= 10





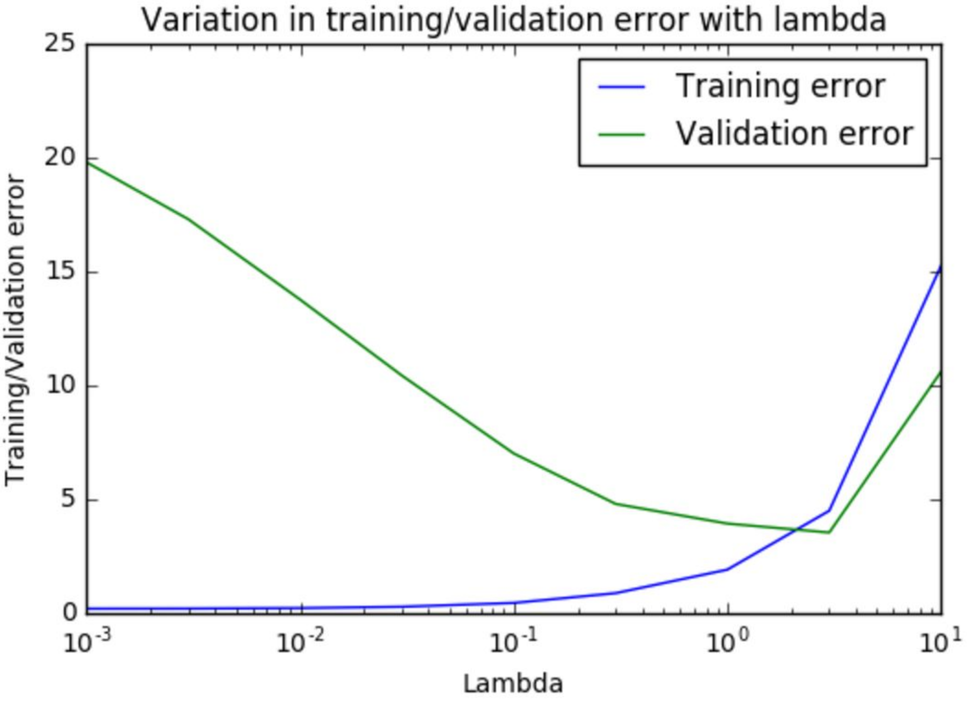
3) lambda= 100





Conclusion: As we can see above, when lambda equal to 1, we have best polynomial regression fit and learning curve.

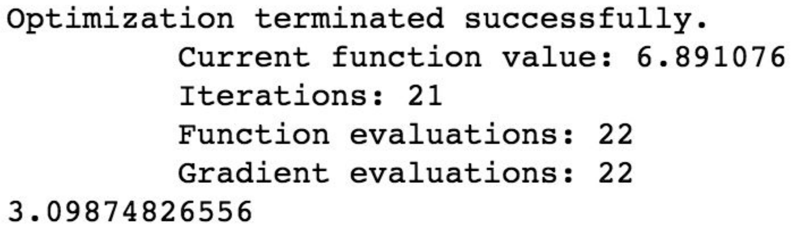
A5: Selecting lambda using a validation set



Conclusion: Due to randomness, the cross validation error can sometimes be lower than the training error. Therefore, when lambda equal to 3 we have best choice for this problem.

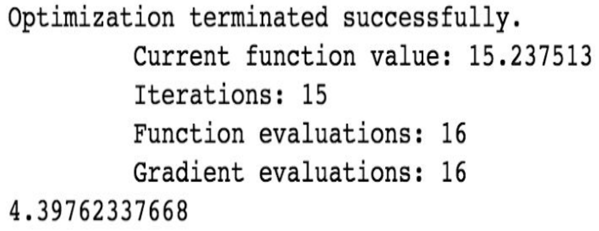
A6: Computing test set error

The error of the best model that we found is shown below:
lambda= 1



```
Optimization terminated successfully.
         Current function value: 6.891076
         Iterations: 21
         Function evaluations: 22
         Gradient evaluations: 22
3.09874826556
```

lambda= 3



```
Optimization terminated successfully.
         Current function value: 15.237513
         Iterations: 15
         Function evaluations: 16
         Gradient evaluations: 16
4.39762337668
```