1. (1) $P(y;\lambda)=\dfrac{e^{-\lambda}\lambda^{y}}{y!}=\dfrac{1}{y!}e^{y\log\lambda-\lambda} \Rightarrow P(y;\lambda)$ 属于指数族，$b(y)=\dfrac{1}{y!}$，

$\eta=\log\lambda$，$T(y)=y$，$a(\eta)=e^{\eta}=\lambda$

(2) $\mathrm{Var}\ y=\mathbb{E}(y)=\lambda$，$g(\eta):=\mathbb{E}(T(y))=\mathbb{E}(y)=\lambda=e^{\eta}$.

(3) $\log P(y^{(i)}|x^{(i)};\theta)=y^{(i)}\log\lambda^{(i)}-\lambda^{(i)}-\log(y^{(i)}!)$. 由 $\lambda^{(i)}=e^{\theta^T x^{(i)}}$，

$\log P(y^{(i)}|x^{(i)};\theta)=y^{(i)}\theta^T x^{(i)}-e^{\theta^T x^{(i)}}-\log(y^{(i)}!)$

$\Rightarrow \dfrac{\partial}{\partial\theta_j}\log P(y^{(i)}|x^{(i)};\theta)=y^{(i)}x_j^{(i)}-e^{\theta^T x^{(i)}}x_j^{(i)}$

梯度下降规则为 $\theta_j\leftarrow\theta_j+\alpha(y^{(i)}-e^{\theta^T x^{(i)}})x_j^{(i)}$

2. (1) $1=\int_\Omega P(y;\eta)dy=\int_\Omega b(y)e^{\eta y-a(\eta)}dy \Rightarrow e^{a(\eta)}=\int_\Omega b(y)e^{\eta y}dy$

$=e^{a(\eta)}\dfrac{\partial a(\eta)}{\partial\eta}=\dfrac{\partial}{\partial\eta}\int_\Omega b(y)e^{\eta y}dy=\int_\Omega\dfrac{\partial}{\partial\eta}(b(y)e^{\eta y}dy)=\int_\Omega y\,b(y)e^{\eta y}dy.$

$\Rightarrow \dfrac{\partial a(\eta)}{\partial\eta}=e^{-a(\eta)}\int_\Omega y\,b(y)e^{\eta y}dy=\int_\Omega y\,b(y)e^{\eta y-a(\eta)}dy=\int_\Omega y\,P(y;\eta)dy=\mathbb{E}(Y;\eta).$

(2) $e^{a(\eta)}\dfrac{\partial a(\eta)}{\partial\eta}=\int_\Omega y\,b(y)e^{\eta y}dy \Rightarrow e^{a(\eta)}\left(\dfrac{\partial^2 a(\eta)}{\partial\eta^2}+\left(\dfrac{\partial a(\eta)}{\partial\eta}\right)^2\right)=\int_\Omega y^2 b(y)e^{\eta y}dy$

$\Rightarrow \dfrac{\partial^2 a(\eta)}{\partial\eta}=\int_\Omega y^2 b(y)e^{\eta y-a(\eta)}dy-\left(\dfrac{\partial a(\eta)}{\partial\eta}\right)^2=\int_\Omega y^2 P(y;\eta)dy-\mathbb{E}(Y)^2$

$=\mathbb{E}(Y^2)-\mathbb{E}(Y)^2=\mathrm{Var}(Y)$

(3) $\ell(\theta)=-\log p(y;\theta)=-\log b(y)-\eta y+a(\eta)=-\log b(y)-y\theta^T x+a(\theta^T x)$

$\nabla_\theta\ell(\theta)=\dfrac{\partial a(\theta^T x)}{\partial\theta}-yx=x\dfrac{\partial a(\eta)}{\partial\eta}-yx=x\cdot\mathbb{E}(Y)-yx$

$\nabla_\theta^2\ell(\theta)=\dfrac{\partial^2 a(\theta^T x)}{\partial\theta}=\dfrac{\partial^2 a(\eta)}{\partial\eta^2}\cdot xx^T=\mathrm{Var}(Y)\cdot xx^T$.

由 $\mathrm{Var}(Y)\geqslant 0$，$xx^T\succ 0$ 有 $\nabla_\theta^2\ell(\theta)\succeq 0\Rightarrow\ell(\theta)$ 凸.

3. (1) $(X\Theta-Y)_{ij}=(\Theta^T x^{(i)})_j-y_j^{(i)}\Rightarrow\frac{1}{2}\sum_{i,j}((\Theta^T x^{(i)})_j-y_j^{(i)})^2=\frac{1}{2}\sum_{i,j}(X\Theta-Y)_{ij}^2=\frac{1}{2}\|X\Theta-Y\|_F^2$

由 $\|A\|_F^2=\mathrm{tr}(A^TA)$ 有 $J(\Theta)=\frac{1}{2}\mathrm{tr}((X\Theta-Y)^T(X\Theta-Y))$

(2) $\dfrac{\partial J(\Theta)}{\partial\Theta}=X^T(X\Theta-Y)=0\Rightarrow X^TX\Theta=X^TY.$

对 X^TX 可逆，$\Theta=(X^TX)^{-1}X^TY$，对 X^TX 不可逆，$\Theta\leftarrow\eta X^T(X\Theta-Y)$.

(3) 多变量最小二乘可把 Θ 拆成多个行：$J(\Theta)=\frac{1}{2}\sum_j(\sum_i(\Theta_j x^{(i)}-y_j^{(i)})^2)$

每一个 j 可视为独立的单变量最小二乘，相互之间无关，可以逐一优化至 min

此时多变量最小二乘可优化至 min 且相同.

4. (1) $P(t^{(i)}=1|y^{(i)}=1)=\dfrac{P(y^{(i)}=1|t^{(i)}=1)P(t^{(i)}=1)}{P(y^{(i)}=1|t^{(i)}=1)P(t^{(i)}=1)+\underline{P(y^{(i)}=1|t^{(i)}=0)P(t^{(i)}=0)}}=1.$ $\leftarrow 0$

(2) $P(y^{(i)}=1)=P(y^{(i)}=1|t^{(i)}=1)P(t^{(i)}=1)+P(y^{(i)}=1|t^{(i)}=0)P(t^{(i)}=0)=\alpha P(t^{(i)}=1)$

$\Rightarrow P(t^{(i)}=1)=\frac{1}{\alpha}P(y^{(i)}=1)$

(3) $h(x^{(i)})=P(y^{(i)}=1|t^{(i)}=1)P(t^{(i)}=1)+P(y^{(i)}=1|t^{(i)}=0)P(t^{(i)}=0)=\alpha P(t^{(i)}=1)$

又 $y^{(i)}=1\Rightarrow t^{(i)}=1$，故 $h(x^{(i)})=\alpha$.