1. a

The negative logarithm is a strictly convex function, and by Jensen's inequality we have:

$$D_{KL}(P\|Q) = \sum_{x \in X} P(x) \cdot \log \frac{Q(x)}{P(x)}$$

$$\geq -\log\left(\sum_{x \in X} P(x) \frac{Q(x)}{P(x)}\right)$$

$$= -\log\left(\sum_{x \in X} Q_x\right)$$

$$= 0$$

Also equality holds iff $\frac{Q(x)}{P(x)}$ is constant with probability 1, i.e $P = Q$.

1. b.

$$D_{KL}(P(x,y)\|Q(x,y))$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x)\,Q(y|x)}{Q(x)\,Q(y|x)}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_x \sum_y P(x,y) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_y \sum_{x \neq y} P(y) P(x|y) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_y P(y) \sum_x P(x|y) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \mathcal{D}_{KL}(P(X) \| Q(X)) + \mathcal{D}_{KL}(P(Y|X) \| Q(Y|X))$$

1.c

$$D_{KL}(\hat{P} \| P_\theta) = \sum_{x \in X} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)}$$

$$= \sum_{x \in X} \hat{P}(x) \log \hat{P}(x) - \sum_{x \in X} \hat{P}(x) \log P_\theta(x)$$

$$= \sum_{x \in X} \hat{P}(x) \log \hat{P}(x) - \sum_{x \in X} \frac{1}{n} \sum_{i=1}^{n} 1\{x^{(i)} = x\} \log P_\theta(x)$$

$$= \sum_{x \in X} \hat{P}(x) \log \hat{P}(x) - \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in X} 1\{x^{(i)} = x\} \log P_\theta(x)$$

$$= \sum_{x \in X} \hat{P}(x) \log \hat{P}(x) - \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)})$$

In the last line sum does not depend on θ, so we find

$$\arg\min_\theta D_{KL}(\hat{P} \| P_\theta) = \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)})$$

4.a

To derive the closed-form solution for $\hat{\beta}_\lambda$, we start with ridge regression objective function:

$$J_\lambda(\beta) = \frac{1}{2}\|X\beta - y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2.$$

To find the closed-form solution, we minimize $J_\lambda(\beta)$ w.r.t β.
Taking the derivative of $J_\lambda(\beta)$ w.r.t β we get

$$\frac{\partial J_\lambda(\beta)}{\partial(\beta)} = -2X^T(y - X\beta) + 2\lambda\beta$$

Setting this derivative to zero, we obtain:

$$X^TX\beta + \lambda\beta = X^Ty$$

Factoring out β, we get

$$(X^TX + \lambda I_{d\times d})\beta = X^Ty$$

Multiplying both sides by the inverse of $(X^TX + \lambda I_{d\times d})$, we get:

$$\hat{\beta}_\lambda = (X^TX + \lambda I_{d\times d})^{-1} X^Ty$$

This is a closed-form solution for $\hat{\beta}_\lambda$ when $\lambda > 0$.

When $\lambda = 0$, the matrix $(X^TX + \lambda I_{d\times d})$ is no longer guaranteed to be invertable, and there may be more than one solution that minimize $J_0(\beta)$. In this case we define $\hat{\beta}_0$ as:

$$\hat{\beta}_0 = (X^TX)^+ X^Ty$$

where $(X^TX)^+$ is Moore-Penrose pseudo-inverse of X^TX. This definition ensures that $\hat{\beta}_0$ minimize $J_0(\beta)$ and has the minimum possible L2 norm among all solutions that minimize $J_0(\beta)$.