HW #5

1.

(a) $$\left\|x_i - \sum_{j=1}^{k} z_{ij} v_j\right\|^2 = x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j$$

We know:

$$\left\|x_i - \sum_{j=1}^{k} z_{ij} v_j\right\|_2^2 = \left(x_i - \sum_{j=1}^{k} z_{ij} v_j\right)^T \left(x_i - \sum_{j=1}^{k} z_{ij} v_j\right)$$

$$= x_i^T x_i - \sum_{j=1}^{k} z_{ij} v_j^T x_i - x_i^T \sum_{j=1}^{k} z_{ij} v_j + \left(\sum_{j=1}^{k} z_{ij} v_j\right)^T \left(\sum_{j=1}^{k} z_{ij} v_j\right)$$

$\vdots$

$$= x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j$$

QED

(b) By definition,

$$J_k = \frac{1}{n} \sum_{i=1}^{n} \left(x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{k} v_j^T \frac{1}{n} \left(\sum_{i=1}^{n} x_i x_i^T \right) v_j$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{k} v_j^T \Sigma v_j$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{k} \lambda_j$$

QED

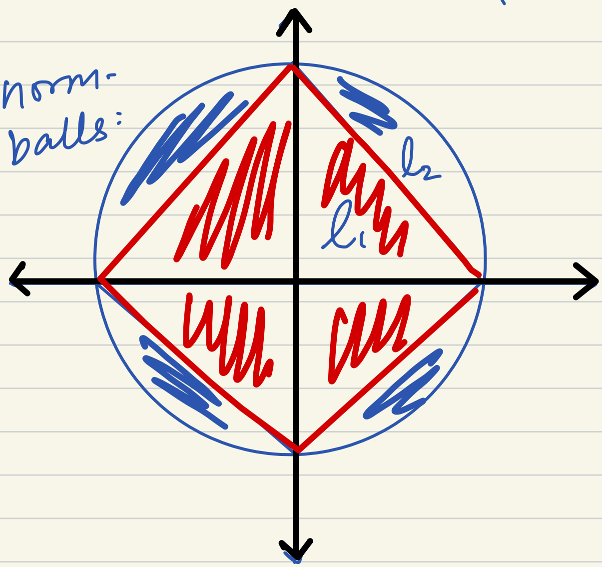
(c) Since $J_d = 0$, we know $\sum_{j=1}^{d} \lambda_j = \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i$

$$J_k = \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{d} \lambda_j + \sum_{j=k+1}^{d} \lambda_j$$

$$= \sum_{j=k+1}^{d} \lambda_j$$

Reconstruction error using PCA proj = sum of e-values thrown out

2. norm-balls:



Minimize : $f(x)$
Subj to : $\|x\|_p \leq K$
is equivalent to

from class notes,

$$\inf_{x} \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) = \inf_{x} \sup_{\lambda \geq 0} f(x) + \lambda(\|x\|_p - k)$$

$$\sup_{\lambda \geq 0} \inf_{x} f(x) + \lambda(\|x\|_p - k) = \sup_{\lambda \geq 0} g(\lambda)$$

Since minimizing $f(x) + \lambda(\|x\|_p - k)$ over x is equivalent to minimizing value of $f(x) + \lambda\|x\|_p$, optimizing x will solve

$$\text{minimize} : f(x) + \lambda\|x\|_p \quad \text{for some } \lambda \geq 0.$$

Considering this, we can consider l_1 regularization as projecting the actual optimal solution onto some suitably sized l_1 norm ball. Since l_1 norm ball has sharper edges, Pr(landing on edge, not face) >> than that for l_2 ball. This is due to rotation invariance of l_2.
$\therefore$ We can see that l_1 penalty will have more weights $= 0$ than l_2 ball.

QED.

3. EXTRA CREDIT

We know that

$$\text{Maximize} : P(\theta|D) = \frac{P(D|\theta)\, P(\theta)}{P(D)}$$

is equivalent to maximizing $\log P(\theta|D)$ given monotonicity of $\log(x)$.

This gives

maximize:

$$\log P(\theta|D) = \log P(D|\theta) + \log P(\theta) - \log P(D)$$

$P(D)$ is constant so can be dropped.

maximize: $\log P(D|\theta) + \log P(\theta)$

$$= \log P(D|\theta) + \ln\left(\exp\left(-\frac{|\theta_i|}{b}\right)\right)$$

$$= \log P(D|\theta) + \log\left(\prod_i \exp\left(-\frac{|\theta_i|}{b}\right)\right)$$

$$= \log P(D|\theta) + \sum_i \left(-\frac{|\theta_i|}{b}\right)$$

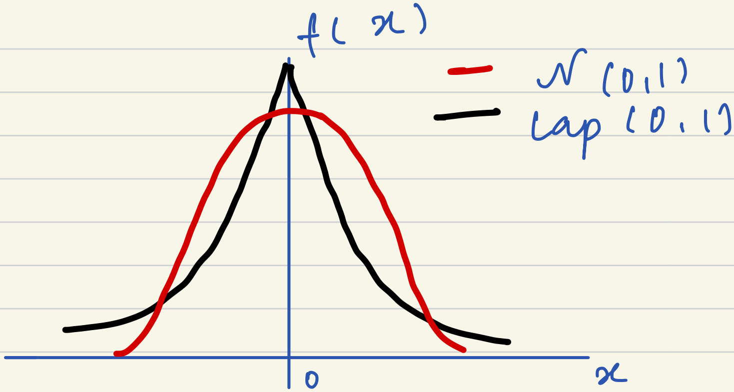
$$= \log P(D|\theta) - \frac{1}{b}\sum_i |\theta_i|$$

$$= \log P(D|\theta) - \lambda\|\theta\|_1 \quad , \lambda = \frac{1}{b}$$

This is the same as

minimize: $-\log P(D|\theta) + \lambda\|\theta\|_1$

which is of the same form as l_1 regularization.



We can see that Laplace dist. has a sharp peak at $x=0$ which gives it a higher probability of being $=0$ than normal dist.

$\therefore$ The weights are more likely to be 0 and $\therefore$ more likely to be sparse.