Henry Trinh

HW5

1. Deriving the Residual Error for PCA:

a. Prove that $\|\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j\|^2 = \vec{x}_i^T x_i - \sum_{j=1}^{k} \vec{v}_j^T \vec{x}_i \vec{x}_i^T \vec{v}_j$

We prove this algebraically:

$\|\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j\|^2 = (\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j)^T(\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j)$. By expansion:

$= \vec{x}_i^T\vec{x}_i - (\sum_{j=1}^{k} z_{ij}v_j)^T\vec{x}_i - \vec{x}_i^T(\sum_{j=1}^{k} z_{ij}\vec{v}_j) + (\sum_{j=1}^{k} z_{ij}\vec{v}_j)^T(\sum_{j=1}^{k} z_{ij}\vec{v}_j)$. Since $z_{ij} = \vec{x}_i^T\vec{v}_j$

$= \vec{x}_i^T\vec{x}_i - 2\sum_{j=1}^{k} z_{ij}\vec{v}_j^T\vec{x}_i + \underline{\sum_{j=1}^{k}\sum_{i=1}^{k} z_{ij}\vec{v}_j^T z_{ij}\vec{v}_j} \rightarrow = \sum_{j=1}^{k}\vec{v}_j^T(\sum_{i=1}^{k} z_{ij}z_{ij})\vec{v}_j$.

Since $\vec{v}_i^T\vec{v}_j = 1$ iff $i=j$: $\vec{x}_i^T\vec{x}_i - 2\sum_{j=1}^{k} z_{ij}\vec{v}_j^T\vec{x}_i + \sum_{j=1}^{k}\vec{v}_j^T\vec{x}_i\vec{x}_i^T v_j$. Since $z_{ij} \in \mathbb{R}$:

$\Rightarrow \vec{x}_i^T\vec{x}_i - 2\sum_{j=1}^{k} z_{ij}\vec{v}_j^T\vec{x}_i + \sum_{j=1}^{k}\vec{v}_j^T\vec{x}_i\vec{x}_i^T\vec{v}_j$. Hence, by simplification:

$\therefore \boxed{\|\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j\|^2 = \vec{x}_i^T x_i - \sum_{j=1}^{k}\vec{v}_j^T\vec{x}_i\vec{x}_i^T\vec{v}_j}$ as desired. //

b. Show that $J_K = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\vec{v}_j^T\vec{x}_i\vec{x}_i^T\vec{v}_j) = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\lambda_j$.

Since $\vec{v}_j^T\Sigma\vec{v}_j = \lambda_j\vec{v}_j^T\vec{v}_j = \lambda_j$, we show the following algebraically:

$J_K = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\vec{v}_j^T\vec{x}_i\vec{x}_i^T\vec{v}_j) = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\vec{v}_j^T(\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i\vec{x}_i^T)\vec{v}_j$

$= \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\sum_{i=1}^{n}\vec{v}_j \Rightarrow \boxed{\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\lambda_j}$, as desired. //

c. If $K=d$, there is no truncation, so $J_D = \varnothing$. Use this to show that the error from only using $K<d$ terms is given by the following:

$$J_K = \sum_{j=K+1}^{d}\lambda_j \quad (1).$$

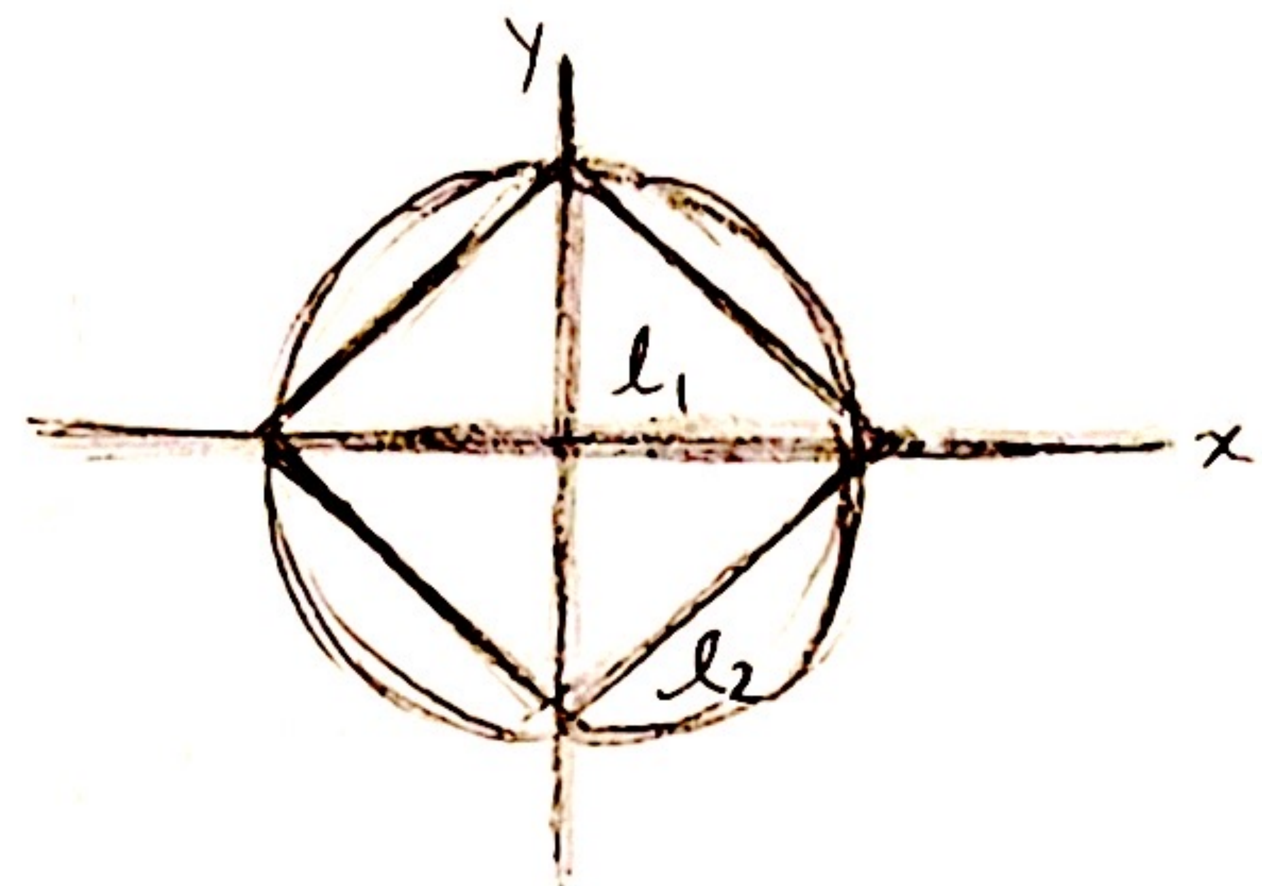
Since we can partition the sum $\sum_{j=1}^{d}\lambda_j$ into $\sum_{j=1}^{k}\lambda_j$ and $\sum_{j=K+1}^{d}\lambda_j$, then:

$J_K = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{d}\lambda_j + \sum_{j=K+1}^{d}\lambda_j = \boxed{\sum_{j=K+1}^{d}\lambda_j}$, as desired. //

2. ℓ_1 - Regularization:

Consider the ℓ_1 norm of a vector $x \in \mathbb{R}^n$: $\|\vec{x}\|_1 = \sum_i |\vec{x}_i|$.

Draw the norm-ball $B_K = \{\vec{x} : \|\vec{x}\|_1 \leq K\}$ for $k=1$. On the same plot, draw the Euclidean norm-ball $A_K = \{\vec{x} : \|\vec{x}\|_2 \leq K\}$ for $k=1$ behind B_K.



Now, show that the optimization problem:

$$\text{minimize } f(x), \text{ subject to } \|\vec{x}\|_p \leq k$$

is equivalent to: minimize $f(x) + \lambda\|\vec{x}\|_p$. Then, argue why using ℓ_1-regularization (adding a $\lambda\|\vec{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2-regularization.

We re-write our original problem as $\inf_{x} \sup_{\lambda \geq 0} L(x,\lambda) = \inf_{x} \sup_{\lambda\geq 0} f(x) + \lambda(\|\vec{x}\|_p - k)$.
In its dual, we can "flip" the infimum and supremum, such that:

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

Since the minimizing value of $f(x) + \lambda(\|\vec{x}\|_p - k)$ over x is equivalent to the minimizing value of $f(x) + \lambda\|\vec{x}\|_p$, and $(-\lambda k)$ does not depend on x, we know that the optimizing x will solve "minimize $f(x) + \lambda\|\vec{x}\|_p$" for some value $\lambda \geq 0$. Hence, in tandem with our plot, ℓ_1-regularization is the projection of our actual optimal solution onto some well-defined ℓ_1 norm-ball. As our ℓ_1 ball has sharper edges, the probability of landing on an edge and not on the face [where both elements of the vector are non-zero] is infinitely larger than the ℓ_2 ball. Specifically, this is due to the rotation invariance of the ℓ_2, which does not hold for the ℓ_1 ball. Furthermore, if we were to then generalize this to higher dimensions, the ℓ_1-penalty would encourage more weights to be zero, compared to the ℓ_2 ball, as desired.