Homework 5

Hoang Chu

Problem 1.

Solution:

1. We have:

$$\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|_2^2 = \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^\top \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i - \mathbf{x}_i^\top \sum_{j=1}^{k} z_{ij}\mathbf{v}_j + \left(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^\top \left(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i + \left(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^\top \left(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^k \qquad \text{(bringing } \mathbf{x}_i^\top \text{ into sum)}$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^\top z_{ij}^\top z_{ij}\mathbf{v}_j$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}_j$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}_j + \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}_j$$

$$= \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}_j$$

□

2. We have:

$$
\begin{aligned}
J_k &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{v}_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top \boldsymbol{\Sigma}\mathbf{v}_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}\lambda_j
\end{aligned}
$$

$\square$

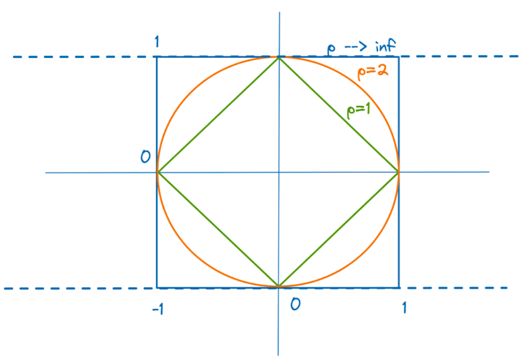
3. Since $J_d = 0$ we know $\sum_{j=1}^{d}\lambda_j = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \mathbf{x}_i$. Then:

$$
J_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{d}\lambda_j + \sum_{j=k+1}^{d}\lambda_j = \sum_{j=k+1}^{d}\lambda_j
$$

$\square$

Problem 2.

Solution:



The norm ball:

We know the optimization problem minimize: $f(\mathbf{x})$ subj. to: $\|\mathbf{x}\|_p \le k$ is equivalent to:

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

In its dual we can flip the inf and sup.

$$
\sup_{\lambda\ge 0}\inf_{\mathbf{x}} f(\mathbf{x}) + \lambda\left(\|\mathbf{x}\|_p - k\right) = \sup_{\lambda\ge 0} g(\lambda)
$$

Since the minimizing value of $f(\mathbf{x}) + \lambda\left(\|\mathbf{x}\|_p - k\right)$ over $\mathbf{x}$ is equivalent to the minimizing value of $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p(-\lambda k$ doesn't depend on $\mathbf{x}$), we know the the optimizing $\mathbf{x}$ will minimize: $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$ for some suitable value of $\lambda \geq 0$. Looking at the plot and this result, we can consider ℓ_1 regularization as projecting the actual optimal solution of your problem onto some suitably sized ℓ_1 norm ball.

Since the 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 nonzero) is infinitely larger than the ℓ_2 ball. This is due to the rotation invariance of the ℓ_2 that certainly doesn't hold for the ℓ_1 ball.

Generalizing to higher dimensions, we can see that the ℓ_1 penalty will encourage more weights to be zero compared to the ℓ_2 ball, as desired.

Problem 3.

Solution: We know the Maximum-a-Posteriori problem of maximize:

$$\mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})}$$

is equivalent to maximizing $\log \mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D})$ given the monotonicity of $\log(x)$. This gives maximize:

$$\log \mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D}) = \log \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \log \mathbb{P}(\boldsymbol{\theta}) - \log \mathbb{P}(\mathcal{D})$$

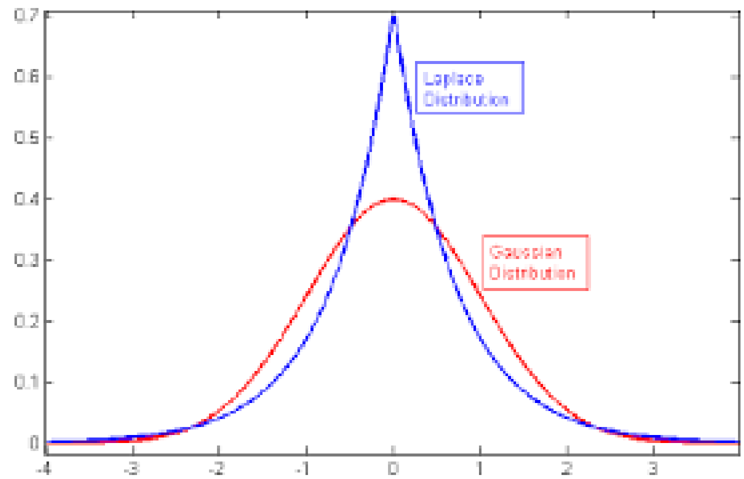
Since $\mathbb{P}(\mathcal{D})$ is a constant not dependent on $\boldsymbol{\theta}$, we can drop that term from the problem and flip into a minimization problem, giving minimize: $-\log \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) - \log \mathbb{P}(\boldsymbol{\theta})$.

Given a prior $\boldsymbol{\theta}_i \sim \mathrm{Lap}(0, b)$,

$$\begin{aligned}
-\log \mathbb{P}(\boldsymbol{\theta}) &= -\log \prod_i \exp\left(-\frac{|\boldsymbol{\theta}_i|}{b}\right) + Z \\
&= \frac{1}{b}\sum_i |\boldsymbol{\theta}_i| + Z \\
&= \lambda \|\boldsymbol{\theta}\|_1 + Z.
\end{aligned}$$

It follows that our original problem is equivalent to: minimize: $-\log \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1$,

or a ℓ_1 regularized maximum likelihood estimate, as desired. Note the plots of the Standard Normal and Laplace Densities.


Laplace
Distribution
Gaussian
Distribution
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3