Nicholas Trieu
MATH 189r
Homework 2 Part 1
October 3, 2016

1. (Conditioning a Gaussian) Note that from Murphy page 113. "Equation 4.69 is of such importance in this book that we have put a box around it, so you can easily find it." That equation is important. Read through the proof of the result. Suppose we have a distribution over random variables $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ that is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \boldsymbol{\mu}_2 = 5, \ \boldsymbol{\Sigma}_{11} = \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}, \ \boldsymbol{\Sigma}_{21}^\top = \boldsymbol{\Sigma}_{12} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}, \ \boldsymbol{\Sigma}_{22} = \begin{bmatrix} 14 \end{bmatrix}.$$

Compute

(a) The marginal distribution $p(\mathbf{x}_1)$. Plot the density in $\mathbb{R}^2$.

(b) The marginal distribution $p(\mathbf{x}_2)$. Plot the density in $\mathbb{R}^1$.

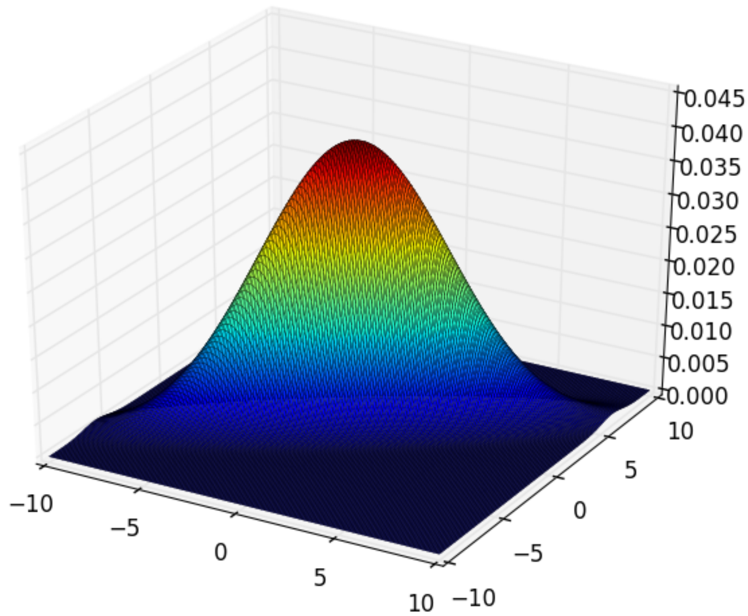
(c) The conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$

(d) The conditional distribution $p(\mathbf{x}_2|\mathbf{x}_1)$

Answers to 1:

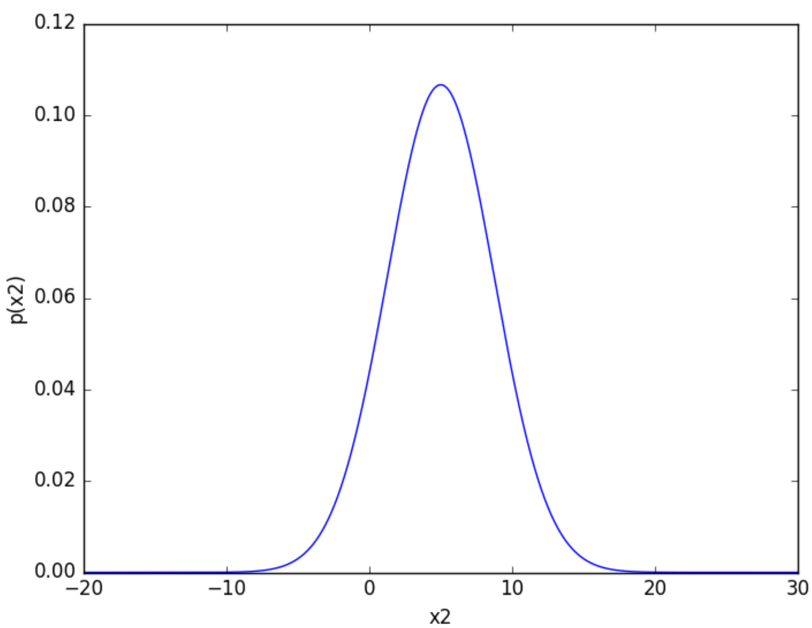
(a) The marginal distribution $p(\mathbf{x}_1)$ is given by

$$\begin{aligned} p(\mathbf{x}_1) &= N(\mathbf{x}_1 | \boldsymbol{\mu}_1, \Sigma_{11}) \\ &= \frac{1}{(2\pi) \cdot \sqrt{14}} \exp\left[-\frac{1}{2} \mathbf{x}_1^\top \begin{bmatrix} 13/14 & -4/7 \\ -4/7 & 3/7 \end{bmatrix} \mathbf{x}_1 \right] \end{aligned}$$



(b) The marginal distribution $p(\mathbf{x}_2)$ is given by

$$\begin{aligned} p(\mathbf{x}_2) &= N(\mathbf{x}_2|\boldsymbol{\mu}_2, \Sigma_{22}) \\ &= \frac{1}{\sqrt{2\pi} \cdot \sqrt{14}} \exp\left[-\frac{1}{2}(x_2 - 5) \cdot \frac{1}{14} \cdot (x_2 - 5)\right] \end{aligned}$$



(c) The conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is given by

$$\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) &= N(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \Sigma_{1|2}) \\
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \begin{bmatrix} 59/14 & 57/14 \\ 57/14 & 61/14 \end{bmatrix}
\end{aligned}$$

so the conditional distribution is

$$\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) = &\frac{1}{(2\pi)\cdot\sqrt{25/14}} \exp\left[-\frac{1}{2}(\mathbf{x}_1 - \begin{bmatrix} 5/14 \\ 11/14 \end{bmatrix} \cdot (x_2 - 5))^\top \begin{bmatrix} 61/25 & -57/25 \\ -57/25 & 59/25 \end{bmatrix} \right. \\
&\left. \times (\mathbf{x}_1 - \begin{bmatrix} 5/14(\mathbf{x}_2 - 5) \\ 11/14(\mathbf{x}_2 - 5) \end{bmatrix}) \right]
\end{aligned}$$

(d) The conditional distribution $p(\mathbf{x}_2|\mathbf{x}_1)$ is given by

$$\begin{aligned}
p(\mathbf{x}_2|\mathbf{x}_1) &= N(\mathbf{x}_2|\boldsymbol{\mu}_{2|1}, \Sigma_{2|1}) \\
\boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
\Sigma_{2|1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= 25/14
\end{aligned}$$

so the conditional distribution is

$$p(\mathbf{x}_2|\mathbf{x}_1) = \frac{1}{\sqrt{2\pi}\cdot\sqrt{14/25}} \exp\left[-\frac{7}{25}(x_2 - 5 + \begin{bmatrix} -23/14 & 13/7 \end{bmatrix} \mathbf{x}_1)^2 \right]$$

2. (ℓ_1-Regularization) Consider the ℓ_1 norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|.$$

Plot the norm-ball $B_k = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq k\}$ for $k = 1$. On the same plot, plot the Euclidean norm-ball $A_k = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq k\}$ for $k = 1$ behind the first plot. Show that the optimization problem

$$\begin{aligned} &\text{minimize: } f(\mathbf{x}) \\ &\text{subj. to: } \|\mathbf{x}\|_p \leq k \end{aligned}$$

is equivalent to

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$$

(hint: create the Lagrangian). With this knowledge, and the plots given above, argue why using ℓ_1 regularization (adding a $\lambda\|\mathbf{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2 regularization for suitably large λ.

Answers to 2:
Consider the Lagrangian

$$L = f(\mathbf{x}) - \lambda(k - \|\mathbf{x}\|_p).$$

Thus, minimizing $f(\mathbf{x})$ is equivalent to the problem

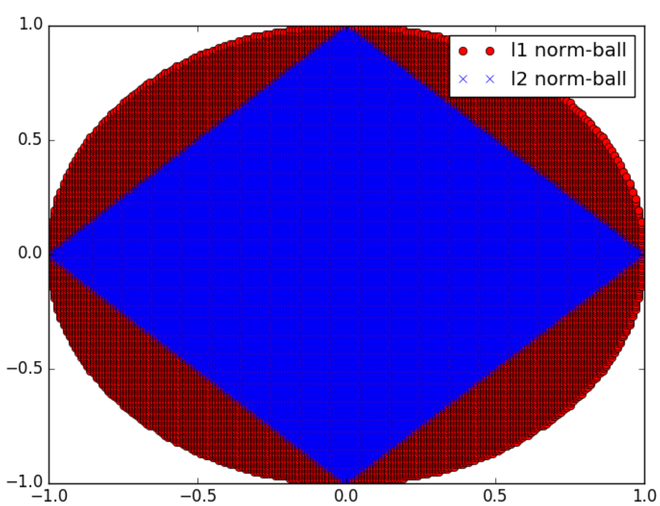
$$\text{minimize: } f(\mathbf{x}) - \lambda(k - \|\mathbf{x}\|_p).$$

Now note that λk is independent of $\mathbf{x}$. Thus our problem is equivalent to

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$$

as desired.

Plot: Using l_1 reg will give sparser solutions because it allows huge disparity in the values of $\mathbf{x}$, whereas l_2 punishes large values of $\mathbf{x}$ by squaring. This is shown by the plot, where the viable solutions for the l_1 norm are more concentrated around 0 for each value of $\mathbf{x}$.



3. (Lasso) Show that placing an equal zero-mean Laplace prior on each element of the weights $\boldsymbol{\theta}$ of a model is equivalent to ℓ_1 regularization in the Maximum-a-Posteriori estimate

$$\text{maximize: } \mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})}.$$

Note the form of the Laplace distribution is

$$\text{Lap}(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

where μ is the location parameter and $b > 0$ controls the variance. Plot the density $\text{Lap}(x|0,1)$ and the standard normal $\mathcal{N}(x|0,1)$ and suggest why this would lead to sparser solutions than a Gaussian prior on each elements of the weights (which correspond to ℓ_2 regularization).

Answers to 3:
Recall that maximizing the probability estimate is equivalent to maximizing its log. Note that the log of the Laplace distribution is

$$\log \text{Lap}(x|\mu, b) = -\frac{|x-\mu|}{b} + \text{constant term}$$

which corresponds to l_1 regularization, since we can scale by b and ignore the constant term in the maximization.

Plot: Using the Laplace prior leads to sparser solutions than using the Gaussian prior because it corresponds with l_1 instead of l_2 regularization, which leads to more 0 coefficients (as explained in problem 2). The plot verifies this effect because we see that the Laplace distribution is more concentrated around 0 than the Gaussian distribution.