ECEN 649 Pattern Recognition – Fall 2022
Problem Set 1 Solutions

Chapter 2

2.1. Equation (2.3):

$$\begin{aligned}
P(\mathbf{X} \in E, Y=0) &= \sum_{\mathbf{x}^k \in E} P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0)\,,\\
P(\mathbf{X} \in E, Y=1) &= \sum_{\mathbf{x}^k \in E} P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)\,,
\end{aligned}$$

for all $E \subseteq D$.

Equation (2.4):

$$\begin{aligned}
P(Y=0 \mid \mathbf{X}=\mathbf{x}^k) &= \frac{P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0)}{P(\mathbf{X}=\mathbf{x}^k)}\\
&= \frac{P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0)}{P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0)+P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)}\,,\\
P(Y=1 \mid \mathbf{X}=\mathbf{x}^k) &= \frac{P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)}{P(\mathbf{X}=\mathbf{x}^k)}\\
&= \frac{P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)}{P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0)+P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)}\,,
\end{aligned}$$

for all $\mathbf{x}^k \in D$.

Equation (2.8):

$$\begin{aligned}
\varepsilon^0[\psi] &= P(\psi(\mathbf{X})=1 \mid Y=0) = \sum_{\mathbf{x}^k:\psi(\mathbf{x}^k)=1} P(\mathbf{X}=\mathbf{x}^k \mid Y=0)\,,\\
\varepsilon^1[\psi] &= P(\psi(\mathbf{X})=0 \mid Y=1) = \sum_{\mathbf{x}^k:\psi(\mathbf{x}^k)=0} P(\mathbf{X}=\mathbf{x}^k \mid Y=1)\,.
\end{aligned}$$

Equation (2.9):

$$\begin{aligned}
\varepsilon[\psi] &= P(\psi(\mathbf{X}) \neq Y) = P(\psi(\mathbf{X})=1, Y=0) + P(\psi(\mathbf{X})=0, Y=1)\\
&= P(\psi(\mathbf{X})=1 \mid Y=0)P(Y=0) + P(\psi(\mathbf{X})=0 \mid Y=1)P(Y=1)\\
&= P(Y=0)\,\varepsilon^0[\psi] + P(Y=1)\,\varepsilon^1[\psi]\\
&= \sum_{\mathbf{x}^k:\psi(\mathbf{x}^k)=1} P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0) + \sum_{\mathbf{x}^k:\psi(\mathbf{x}^k)=0} P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)\,.
\end{aligned}$$

Equation (2.11):

$$\varepsilon[\psi] = E[\varepsilon[\psi \mid \mathbf{X}=\mathbf{x}^k]] = \sum_{\mathbf{x}^k \in D} \varepsilon[\psi \mid \mathbf{X}=\mathbf{x}^k]\, P(\mathbf{X}=\mathbf{x}^k)\,.$$

Equation (2.30):

$$\begin{aligned}\varepsilon^* &= \sum_{\mathbf{x}^k \in D} (I_{\eta(\mathbf{x}^k) \leq 1-\eta(\mathbf{x}^k)}\, \eta(\mathbf{x}^k) + I_{\eta(\mathbf{x}^k) > 1-\eta(\mathbf{x}^k)}(1-\eta(\mathbf{x}^k)))P(\mathbf{X} = \mathbf{x}^k) \\ &= E[\min\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\}]\,.\end{aligned}$$

Equation (2.34):

$$\begin{aligned}\varepsilon^* &= P(Y=0)\,\varepsilon^0[\psi^*] \,+\, P(Y=1)\,\varepsilon^1[\psi^*] \\ &= \sum_{\mathbf{x}^k : P(Y=1)P(\mathbf{X}=\mathbf{x}^k|Y=1) > P(Y=0)P(\mathbf{X}=\mathbf{x}^k|Y=0)} P(Y=0)P(\mathbf{X}=\mathbf{x}^k \mid Y=0) \\ &+ \sum_{\mathbf{x}^k : P(Y=1)P(\mathbf{X}=\mathbf{x}^k|Y=1) \leq P(Y=0)P(\mathbf{X}=\mathbf{x}^k|Y=0)\}} P(Y=1)P(\mathbf{X}=\mathbf{x}^k \mid Y=1)\,.\end{aligned}$$

Equation (2.34):

$$E[\eta(\mathbf{X})] \,=\, \sum_{\mathbf{x}^k \in D} P(Y=1 \mid \mathbf{X}=\mathbf{x}^k)P(\mathbf{X}=\mathbf{x}^k) \,=\, P(Y=1)\,.$$

2.3. (a) From (2.32),

$$\varepsilon^* = 0 \;\Leftrightarrow\; E[|2\eta(\mathbf{X})-1|] = 1 \;\Leftrightarrow\; \eta(\mathbf{X}) = 0 \text{ or } 1 \;\text{ with probability 1.}$$

(b) If $\varepsilon^* = P(Y \neq \psi^*(\mathbf{X})) = 0$ then $P(Y = \psi^*(\mathbf{X})) = 1$, i.e., $f = \psi^*$ is a function such that $Y = f(\mathbf{X})$ with probability 1. On the other hand, if there is a function f s.t. $Y = f(\mathbf{X})$ with probability 1. i.e., $P(Y = f(\mathbf{X})) = 1$, then $\varepsilon^* \leq \varepsilon[f] \,=\, P(Y \neq f(\mathbf{X})) = 0$ so that $\varepsilon^* = 0$.

(c) We show the contrapositive:

$$\varepsilon^* > 0 \;\Leftrightarrow\; P(p(\mathbf{X} \mid Y=0)p(\mathbf{X} \mid Y=1) > 0) > 0\,.$$

The condition on the right of this equation is equivalent to there being a closed and bounded region $A \in R^d$ with $P(A) > 0$, over which $p(\mathbf{x} \mid Y=0) > 0$ and $p(\mathbf{x} \mid Y=1) > 0$. Since $P(Y=0)P(Y=1) \neq 0$, it folows from (2.4) that $P(Y=0 \mid \mathbf{x})P(Y=1 \mid \mathbf{x}) = (1-\eta(\mathbf{x}))\eta(\mathbf{x}) > 0$ over A. This is equivalent to the condition $\min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\} > 0$ over A. Now let $c = \inf_{\mathbf{x}\in A} \min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\}$. Since A is closed and bounded, we have $c > 0$. Then,

$$\varepsilon^* \,=\, E[\min\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\}] \geq E[cI_A] = cP(A) > 0\,,$$

as required.

2.4. (a) We have that

$$\begin{aligned}P(\psi(\mathbf{X}) \neq Y \mid \mathbf{X}=\mathbf{x}) &= \sum_{i=0}^{c-1} P(\psi(\mathbf{X}) = i, Y \neq i \mid \mathbf{X}=\mathbf{x}) \\ &= \sum_{i=0}^{c-1} I_{\psi(\mathbf{X})=i}\, P(Y \neq i \mid \mathbf{X}=\mathbf{x}) \,=\, \sum_{i=0}^{c-1} I_{\psi(\mathbf{X})=i}\,(1-\eta_i(\mathbf{x})) \\ &= 1 \,-\, \sum_{i=0}^{c-1} I_{\psi(\mathbf{x})=i}\,\eta_i(\mathbf{x}) \,=\, 1-\eta_{\psi(\mathbf{x})}(\mathbf{x})\,.\end{aligned}$$

(b) Directly from the previous item,

$$\varepsilon \,=\, \int P(\psi(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \,=\, 1 \,-\, \sum_{i=0}^{c-1} \int I_{\psi(\mathbf{x})=i}\, \eta_i(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

$$= 1 \,-\, \sum_{i=0}^{c-1} \int_{\{\mathbf{x} \mid \psi(\mathbf{x})=i\}} \eta_i(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}\,.$$

(c) Again using the result of item (a),

$$P(\psi(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x}) - P(\psi^*(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x})$$

$$= \eta_{\psi^*(\mathbf{x})}(\mathbf{x}) - \eta_{\psi(\mathbf{x})}(\mathbf{x}) \,=\, \left[\max_{i=0,1,\ldots,c-1} \eta_i(\mathbf{X})\right] \,-\, \eta_{\psi(\mathbf{x})}(\mathbf{x}) \geq 0\,,$$

by definition of $\psi^*(\mathbf{x})$. Integration over the feature space yields

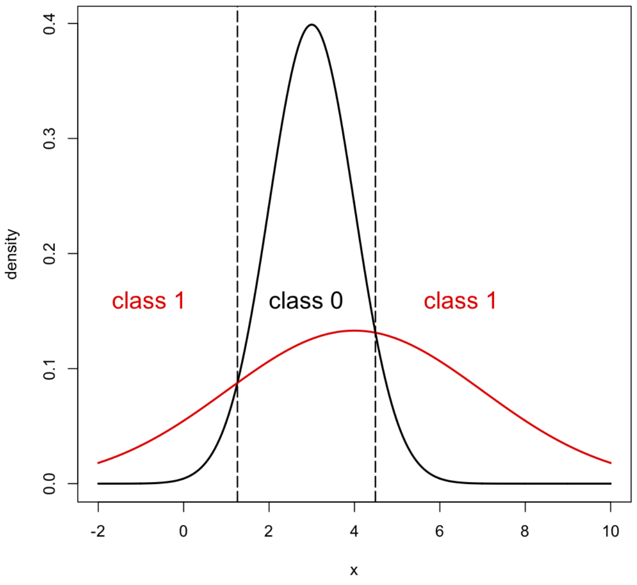
$$\varepsilon - \varepsilon^* \,=\, E\left[P(\psi(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x}) - P(\psi^*(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x})\right] \,\geq\, 0\,.$$

(d) Using the result of item (a) and the definition of ψ^*,

$$\varepsilon^* \,=\, E\left[P(\psi^*(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x})\right] \,=\, 1 - E\left[\eta_{\psi^*(\mathbf{x})}(\mathbf{x})\right] \,=\, 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(\mathbf{X})\right]\,.$$

(e) If if $z_1, \ldots, z_c$ are nonnegative numbers that add up to one, then the minimum value that $\max\{z_1, \ldots, z_c\}$ can take is $1/c$, when $z_1 = \cdots = z_c = 1/c$. From part (d), the Bayes error is maximum when $\max\{\eta_1(\mathbf{x}), \ldots, \eta_c(\mathbf{x})\}$ is minimum at each value of $\mathbf{x} \in R^d$ (or, at least, over a set of probability one). This means that $\eta_i(\mathbf{x})$ must be equal to $1/c$, for $i = 1, \ldots, c$, and the maximum Bayes error is $1 - E[1/c] = 1 - 1/c$.

2.7. (a) We can ignore the prior probabilities, since they are equal, and compare the Gaussian densities directly. As seen in the following plot, this produces a two-point decision boundary (the classifier is not linear in this heteroskedastic case).



(b) The decision boundary consists of the two points where the densities equal each other. One thus needs to solve the equation:

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-3)^2}{2}\right) \;=\; \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{(x-4)^2}{18}\right)$$

Taking logs on both sides and simplifying leads to be following second-degree equation:

$$x^2 \;-\; \frac{23}{4}\,x \;+\; \frac{65-18\ln 3}{8} \;=\; 0\,.$$

The two solutions of this equation are $x_1 \approx 1.259$ and $x_2 \approx 4.491$. The Bayes classifier is therefore:

$$\psi^*(x) \;=\; \begin{cases} 0, & 1.259 < x < 4.491, \\ 1, & \text{otherwise.} \end{cases}$$

(c) We have that

$$\begin{aligned} \varepsilon^1[\psi^*] \;&=\; \int_{\{x \mid \psi^*(x)=0\}} p(x \mid Y=1)\,dx \;=\; P(1.259 < X < 4.491 \mid Y=1) \\ &=\; P\left(\frac{1.259-4}{3} < \frac{X-4}{3} < \frac{4.491-4}{3} \;\middle|\; Y=1\right) \\ &=\; P\left(\frac{1.259-4}{3} < Z < \frac{4.491-4}{3}\right) \\ &=\; P\left(-0.913\bar{6} < Z < 0.163\bar{6}\right) \\ &=\; \Phi(0.163\bar{6}) - \Phi(-0.913\bar{6}) \;\approx\; 0.385\,, \end{aligned}$$

where Z denotes a standard Gaussian random variable with CDF $\Phi(x)$. In completely analogous fashion, one computes:

$$\varepsilon^0[\psi^*] \;=\; P(X < 1.259 \mid Y=0) + P(X > 4.491 \mid Y=0) \;\approx\; 0.109\,.$$

Therefore,

$$\begin{aligned} \text{sensitivity} \;&=\; 1 - \varepsilon^1[\psi] \;\approx\; 61.5\%, \\ \text{specificity} \;&=\; 1 - \varepsilon^0[\psi] \;\approx\; 89.1\%. \end{aligned}$$

(d) The Bayes error is simply

$$\varepsilon[\psi^*] \;=\; (1-c)\varepsilon^0[\psi^*] + c\varepsilon^1[\psi^*] \;=\; \frac{0.385+0.109}{2} \;=\; 0.243\,,$$

where $c = P(Y=1) = 0.5$.

2.9. (a) The discriminant is given by

$$D^*(x_1,x_2) \;=\; [x_1\; x_2]\, A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;+\; \mathbf{b}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;+\; c$$

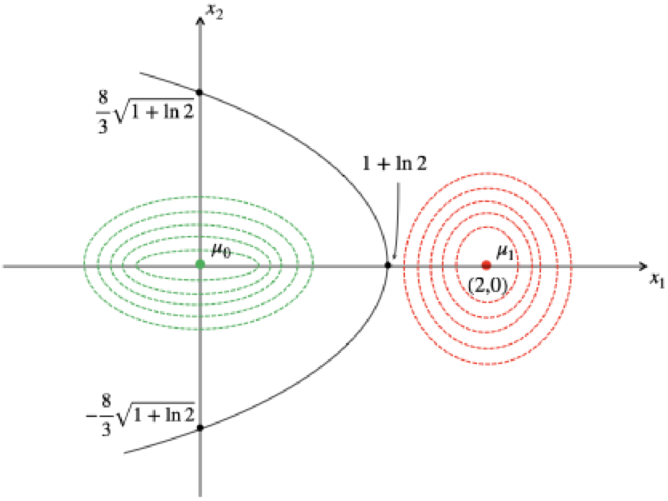
where

$$\begin{aligned} A \;&=\; \frac{1}{2}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)\,, \\ \mathbf{b} \;&=\; \Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_0^{-1}\boldsymbol{\mu}_0\,, \\ c \;&=\; \frac{1}{2}(\boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}\ln\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\,. \end{aligned}$$

For part (a), this gives

$$D^*(x_1, x_2) \;=\; \frac{3}{8}x_2^2 \;+\; x_1 - (1 + \ln 2)\,.$$

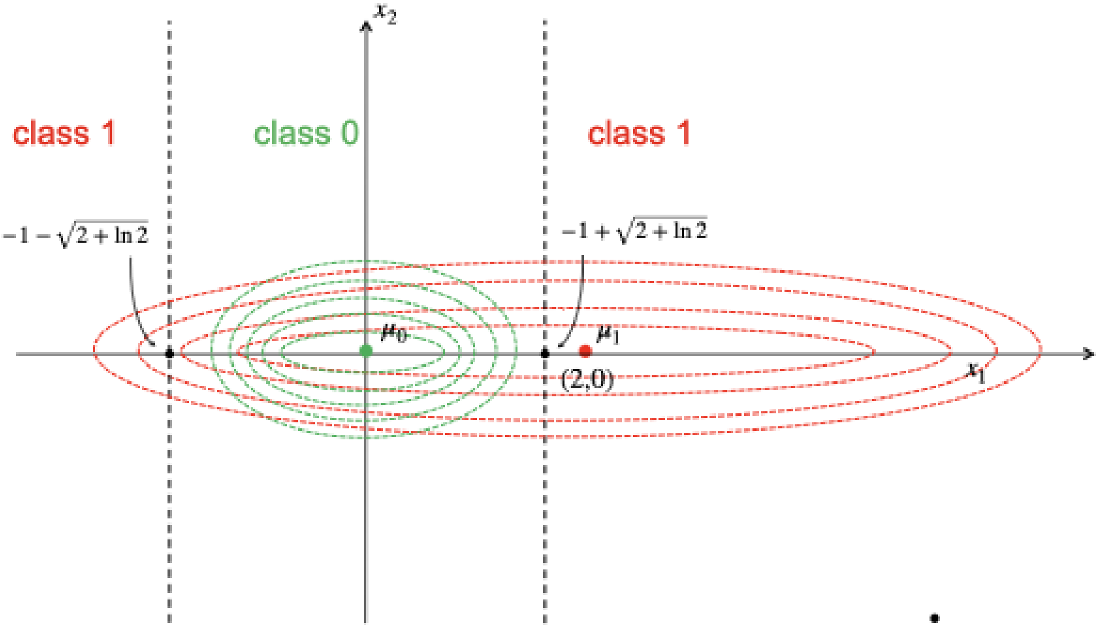
Setting this to zero produces a parabolic optimal decision boundary, displayed below.



(b) Here,

$$D^*(x_1, x_2) \;=\; \frac{1}{8}x_1^2 + \frac{1}{2}x_1 - \frac{1}{2}(1 + \ln 2)\,,$$

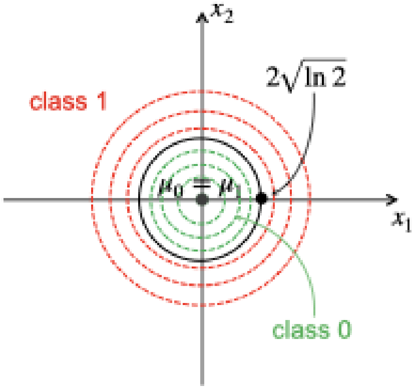
which is a function of x_1 only. Setting this to zero produces an optimal decision boundary consisting of two vertical lines, located at $x_1 = -1 \pm \sqrt{2 + \ln 2}$, which is displayed below.



(c) Here, the class means coincide, and discrimination is afforded by the difference in variances only. We have

$$D^*(x_1, x_2) = \frac{1}{4}x_1^2 + \frac{1}{4}x_2^2 - \ln 2\,.$$

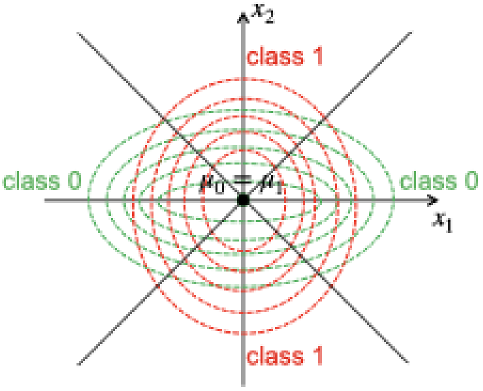
Setting this to zero produces a circular optimal decision boundary, centered at the origin with radius $2\sqrt{\ln 2}$, displayed below.



(d) Similarly, discrimination here is possible only due to the difference in variances. In this case,

$$D^*(x_1, x_2) = -\frac{1}{4}x_1^2 + \frac{1}{4}x_2^2\,.$$

Setting this to zero produces two intersecting, perpendicular lines going through the origin, as displayed below.



2.17. (a) The inverse covariance matrix is

$$\Sigma^{-1}_{d\times d} = \begin{bmatrix} \Sigma^{-1}_{l_1\times l_1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1}_{l_2\times l_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1}_{l_k\times l_k} \end{bmatrix}, \tag{1}$$

where, using the hint,

$$\Sigma_{l_i\times l_i}^{-1} \;=\; \frac{1}{\sigma_i^2(1-\rho)(1+(l-1)\rho)}\begin{bmatrix} 1+(l-2)\rho & -\rho & \cdots & -\rho \\ -\rho & 1+(l-2)\rho & \cdots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \cdots & 1+(l-2)\rho \end{bmatrix}. \tag{2}$$

The squared Mahalanobis distance between the class centers is $\delta^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$. Here, $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0 = (1,\ldots,1)$. Premultiplying and postmultiplying any matrix by a vector of 1's produces the sum of all the elements in the matrix. After some simplification, one obtains

$$\delta^2 \;=\; \sum_{i=1}^{k}\frac{l_i}{\sigma_i^2[1+(l_i-1)\rho_i]}\,. \tag{3}$$

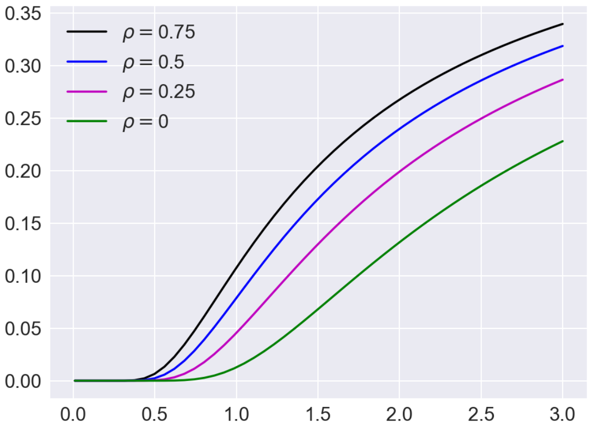
Finally, the Bayes error is given by

$$\varepsilon^* \;=\; \Phi\left(-\frac{\delta}{2}\right) \;=\; \Phi\left(-\frac{1}{2}\sqrt{\sum_{i=1}^{k}\frac{l_i}{\sigma_i^2[1+(l_i-1)\rho_i]}}\right). \tag{4}$$

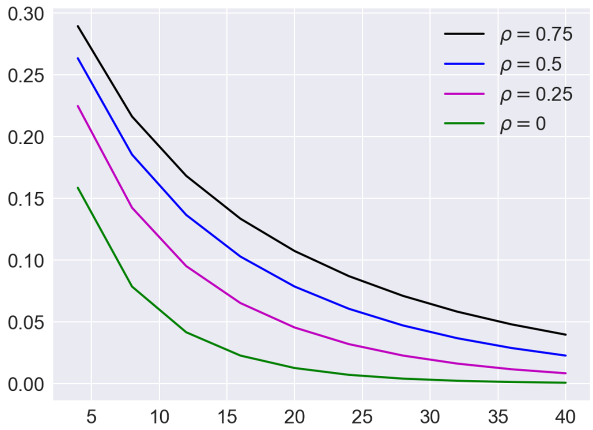
(b) The previous formula becomes:

$$\varepsilon^*(d,l,\sigma,\rho) \;=\; \Phi\left(-\frac{1}{2\sigma}\sqrt{\frac{d}{1+(l-1)\rho}}\right). \tag{5}$$

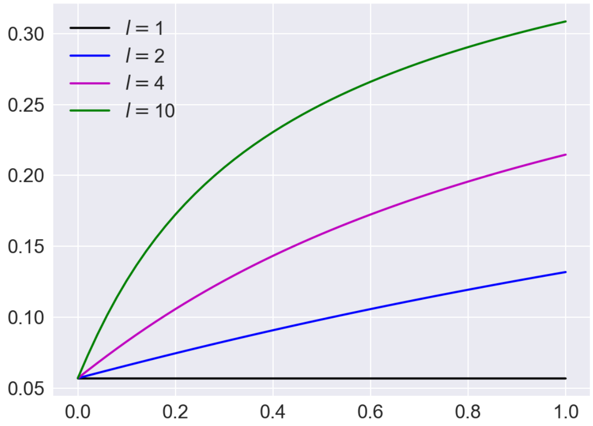
i. We can see in the plot below that indeed the Bayes error increases with σ and ρ. The error with $\rho = 0$ (uncorrelated features) is significantly smaller than in any of the correlated cases.



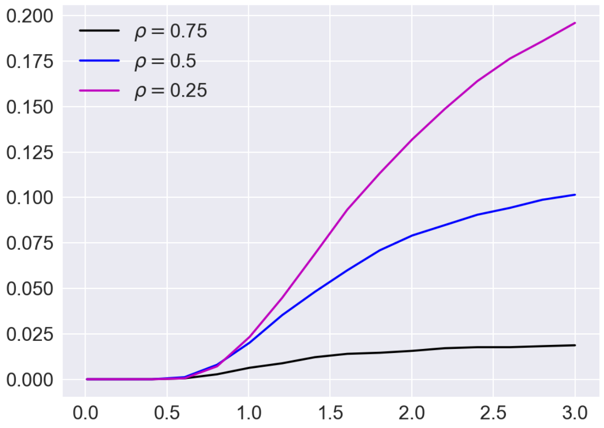
ii. We can see in the plot below that indeed the Bayes error decreases with increasing d, even with correlated features. This is typical behavior for the Bayes error, as more features bring additional discriminatory information.



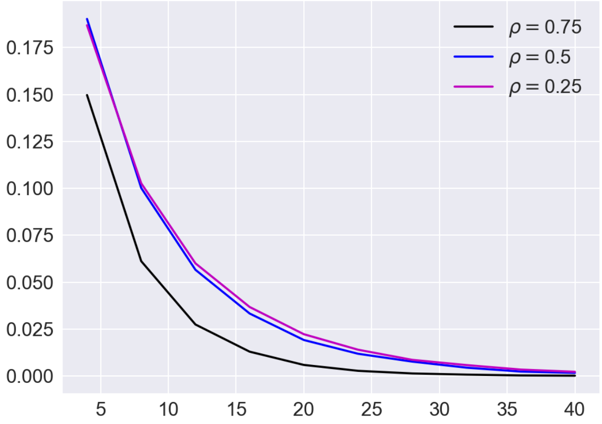
iii. We can see in the plot below that indeed the Bayes error increases with increasing correlation ρ.



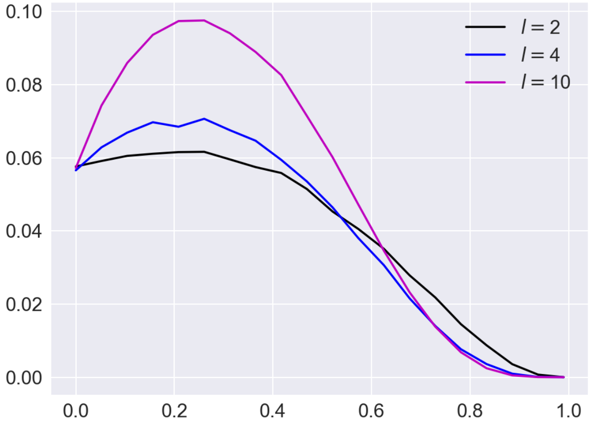
(c) i. We can see in the plot below that in the heteroskedastic case the Bayes error still increases with increasing σ. However, unlike the homoskedastic case, the Bayes error is smaller with larger correlation between features in class 1. This is because more correlation makes the Gaussian density of class 1 more distinct than that of class 0, which is spherical (uncorrelated). This is a peculiarity of the heteroskedastic case. Notice the numerical noise in the curves, which is produced by the approximate Monte-Carlo computation of the Bayes error.



ii. We can see in the plot below that the Bayes error still decreases with increasing d in the heteroskedastic case. We can see again the apparently contradictory fact that larger correlation in class 1 results in a smaller Bayes error; indeed, it is uniformaly smaller as d varies. Once again, this is due to the fact that larger correlation in class 1 adds discrimnatory information with respect to class 0, where the features are uncorrelated.



iii. This is the most interesting plot of the three. Here we can see that at smaller correlations, smaller block sizes produce smaller Bayes error, which is the same behavior observed in the homoskedastic case. However, as the correlation in class 1 increases, larger block size produces a (slightly) better Bayes error. As correlation increases, the problem becomes more nonlinear and behavior becomes more distinct than in the homoskedastic case.


$l = 2$
$l = 4$
$l = 10$
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.2
0.4
0.6
0.8
1.0