Exercice 4. On considère $X = (T, U, V)$ où T, U, V sont des variables aléatoires réelles i.i.d. de loi exponentielle standard. On pose $Y = \mathbb{1}_{\{T+U+V<\theta\}}$ où $\theta \in \mathbb{R}_+^*$ est fixé.

1) i) Rappeler la densité f_1 et la fonction de répartition F_1 d'une loi exponentielle standard.

   ii) Calculer la densité f_2 et la fonction de répartition F_2 de la variable aléatoire $T + U$.

   iii) Calculer la densité f_3 et la fonction de répartition F_3 de la variable aléatoire $T + U + V$.

2) Calculer le classifieur de Bayes $(t, u) \in \mathbb{R}_+^2 \mapsto g_1^*(t, u) \in \{0, 1\}$ lorsque V n'est pas observée. Calculer le risque 0-1 associé à ce classifieur. En donner une approximation numérique lorsque $\theta = 9$.

3) On suppose à présent que seule T est observée. Reprendre les calculs précédents puis comparer les risques 0-1 obtenus lorsque $\theta = 9$.

4) Proposer un classifieur lorsque X n'a aucune composante qui soit observée. Calculer son risque 0-1 et en donner une approximation numérique lorsque $\theta = 9$. Qu'en concluez-vous ?

Solution.

1) Toutes les fonctions explicitées ci-dessous sont illustrées à la Figure 5.

   i) La densité commune de T, U et V est $f_1 : x \in \mathbb{R} \mapsto e^{-x}\, \mathbb{1}_{\{x\geq 0\}}$. Sa fonction de répartition est $F_1 : x \in \mathbb{R} \mapsto (1 - e^{-x})\, \mathbb{1}_{\{x\geq 0\}}$.

   ii) La variable aléatoire $T + U$ est une somme de deux variables indépendantes de même densité f_1. Elle admet donc une densité $f_2 = f_1 * f_1$, qui pour tout $x \in \mathbb{R}$ vaut

$$f_2(x) = \int_{\mathbb{R}} f_1(x-u)\, f_1(u)\, du = \int_{\mathbb{R}} e^{u-x}\, e^{-u}\, \mathbb{1}_{\{x-u\geq 0\}}\, \mathbb{1}_{\{u\geq 0\}}\, du = \left(\int_0^x e^{-x}\, du\right) \mathbb{1}_{\{x\geq 0\}}$$
$$= x\, e^{-x}\, \mathbb{1}_{\{x\geq 0\}}.$$

   Avec une intégration par parties on obtient sa fonction de répartition : pour tout $x \in \mathbb{R}$,

$$F_2(x) = \left(1 - (x+1)\, e^{-x}\right) \mathbb{1}_{\{x\geq 0\}}.$$

   iii) De la même manière, $T + U + V$ est la somme de $T + U$ et V, indépendantes, de densités respectives f_2 et f_1. Elle admet donc une densité $f_3 = f_2 * f_1$ qui pour tout $x \in \mathbb{R}$ vaut

$$f_3(x) = \int_{\mathbb{R}} f_1(x-u)\, f_2(u)\, du = \int_{\mathbb{R}} u\, e^{-u}\, \mathbb{1}_{\{u\geq 0\}}\, e^{u-x}\, \mathbb{1}_{\{x-u\geq 0\}}\, du = \left(\int_0^x u\, e^{-x}\, du\right) \mathbb{1}_{\{x\geq 0\}}$$

$$= \frac{1}{2} x^2 e^{-x} \mathbb{1}_{\{x \geq 0\}}.$$

En intégrant par parties on obtient sa fonction de répartition : pour tout $x \in \mathbb{R}$,

$$\mathrm{F}_3(x) = \left(1 - \left(\frac{x^2}{2} + x + 1\right)\right) e^{-x} \mathbb{1}_{\{x \geq 0\}}.$$

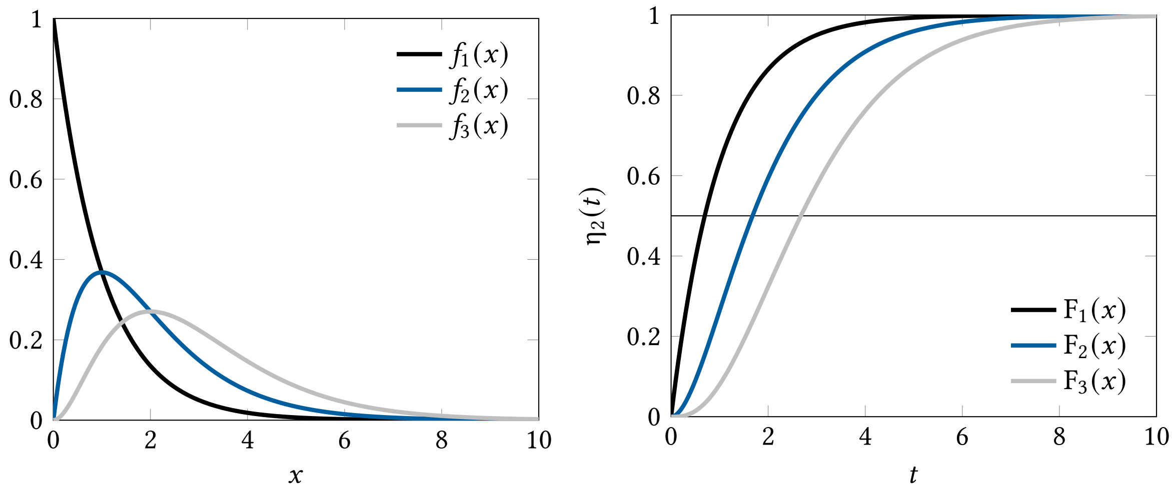

FIGURE 5 – Densités (gauche) et fonctions de répartition (droite) des variables aléatoires T, T + U et T + U + V.

2) Notons $\eta_1 : (t, u) \in \mathbb{R}_+^2 \mapsto \mathbb{P}\left(\mathrm{Y} = 1 \mid (\mathrm{T}, \mathrm{U}) = (t, u)\right)$. Par construction, on a

$$g_1^* : (t, u) \in \mathbb{R}_+^2 \mapsto \mathbb{1}_{\left\{\eta_1(t,u) > \frac{1}{2}\right\}}.$$

Soient $t, u \in \mathbb{R}_+$. Alors

$$\begin{aligned} \eta_1(t, u) &= \mathbb{P}\left(\mathrm{Y} = 1 \mid (\mathrm{T}, \mathrm{U}) = (t, u)\right) = \mathbb{P}\left(\mathrm{T} + \mathrm{U} + \mathrm{V} < \theta \mid (\mathrm{T}, \mathrm{U}) = (t, u)\right) \\ &= \mathbb{P}\left(\mathrm{V} < \theta - t - u\right) \quad \text{(indépendance)} \\ &= \left(1 - e^{t+u-\theta}\right) \mathbb{1}_{\{t+u \leq \theta\}}. \quad \text{(V de loi exponentielle standard)} \end{aligned}$$

Ainsi, $\eta_1(t, u) > \frac{1}{2}$ ssi

$$\left| \begin{array}{l} t + u \leq \theta \\ 1 - e^{t+u-\theta} > \frac{1}{2} \end{array} \right. \quad \Leftrightarrow \quad \left| \begin{array}{l} t + u \leq \theta \\ t + u < \theta - \ln(2) \end{array} \right. \quad \Leftrightarrow \quad t + u < \theta - \ln(2).$$

On en conclut que $g_1^*(t, u) = \mathbb{1}_{\{t+u<\theta-\ln(2)\}}$.

Son risque 0-1, illustré à la Figure 6, est

$$
\begin{aligned}
\mathrm{L}\left(g_1^*\right) &= \mathbb{E}\left(\eta_1(\mathrm{T},\mathrm{U})\, \mathbb{1}_{\left\{\eta_1(\mathrm{T},\mathrm{U}) \leq \frac{1}{2}\right\}} + (1-\eta_1(\mathrm{T},\mathrm{U}))\, \mathbb{1}_{\left\{\eta_1(\mathrm{T},\mathrm{U}) > \frac{1}{2}\right\}}\right) \\
&= \mathbb{E}\left(\left(1-e^{\mathrm{T}+\mathrm{U}-\theta}\right) \mathbb{1}_{\{\mathrm{T}+\mathrm{U}\leq\theta\}}\, \mathbb{1}_{\{\mathrm{T}+\mathrm{U}\geq\theta-\ln(2)\}}\right) \\
&\quad + \mathbb{E}\left(\left(1-\left(1-e^{\mathrm{T}+\mathrm{U}-\theta}\right) \mathbb{1}_{\{\mathrm{T}+\mathrm{U}\leq\theta\}}\right) \mathbb{1}_{\{\mathrm{T}+\mathrm{U}<\theta-\ln(2)\}}\right) \\
&= \mathbb{E}\left(\left(1-e^{\mathrm{T}+\mathrm{U}-\theta}\right) \mathbb{1}_{\{\theta-\ln(2)\leq\mathrm{T}+\mathrm{U}\leq\theta\}}\right) \\
&\quad + \mathbb{E}\left(e^{\mathrm{T}+\mathrm{U}-\theta} \mathbb{1}_{\{\mathrm{T}+\mathrm{U}<\theta-\ln(2)\}}\right) \qquad \left(1 = \mathbb{1}_{\{\mathrm{T}+\mathrm{U}\leq\theta\}} + \mathbb{1}_{\{\mathrm{T}+\mathrm{U}>\theta\}}\right) \\
&= \int_0^{+\infty} \left(1-e^{x-\theta}\right) \mathbb{1}_{\{\theta-\ln(2)\leq x\leq\theta\}}\, x\, e^{-x}\, dx \\
&\quad + \int_0^{+\infty} e^{x-\theta}\, \mathbb{1}_{\{x<\theta-\ln(2)\}}\, x\, e^{-x}\, dx. \qquad (\mathrm{T}+\mathrm{U} \text{ de densité } f_2)
\end{aligned}
$$

Si $\theta \geq \ln(2)$, cela donne

$$
\begin{aligned}
\mathrm{L}\left(g_1^*\right) &= \int_{\theta-\ln(2)}^{\theta} \left(1-e^{x-\theta}\right) x\, e^{-x}\, dx + \int_0^{\theta-\ln(2)} e^{-\theta}\, x\, dx \\
&= \int_{\theta-\ln(2)}^{\theta} x\, e^{-x}\, dx - \int_{\theta-\ln(2)}^{\theta} e^{-\theta}\, x\, dx + \int_0^{\theta-\ln(2)} e^{-\theta}\, x\, dx \\
&= \mathrm{F}(\theta) - \mathrm{F}(\theta-\ln(2)) + e^{-\theta}\left(\left[\frac{x^2}{2}\right]_0^{\theta-\ln(2)} - \left[\frac{x^2}{2}\right]_{\theta-\ln(2)}^{\theta}\right) \\
&= -(\theta+1)\, e^{-\theta} + (\theta-\ln(2)+1)\, e^{\ln(2)-\theta} + \frac{e^{-\theta}}{2}\left(2\,(\theta-\ln(2))^2-\theta^2\right) \\
&= e^{-\theta}\left(-\theta+2\,(\theta-\ln(2))-1+2+(\theta-\ln(2))^2-\frac{\theta^2}{2}\right) \\
&= e^{-\theta}\left((\theta+1-\ln(2))^2-\theta\left(1+\frac{\theta}{2}\right)\right),
\end{aligned}
$$

et si $\theta < \ln(2)$ on obtient

$$
\begin{aligned}
\mathrm{L}\left(g_1^*\right) &= \int_0^{\theta} \left(1-e^{x-\theta}\right) x\, e^{-x}\, dx = \int_0^{\theta} x\, e^{-x}\, dx - \int_0^{\theta} e^{-\theta}\, x\, dx \\
&= \mathrm{F}(\theta) - e^{-\theta}\left[\frac{x^2}{2}\right]_0^{\theta} = 1-(\theta+1)\, e^{-\theta} - e^{-\theta}\,\frac{\theta^2}{2} \\
&= 1-e^{-\theta}\left(1+\theta\left(1+\frac{\theta}{2}\right)\right).
\end{aligned}
$$

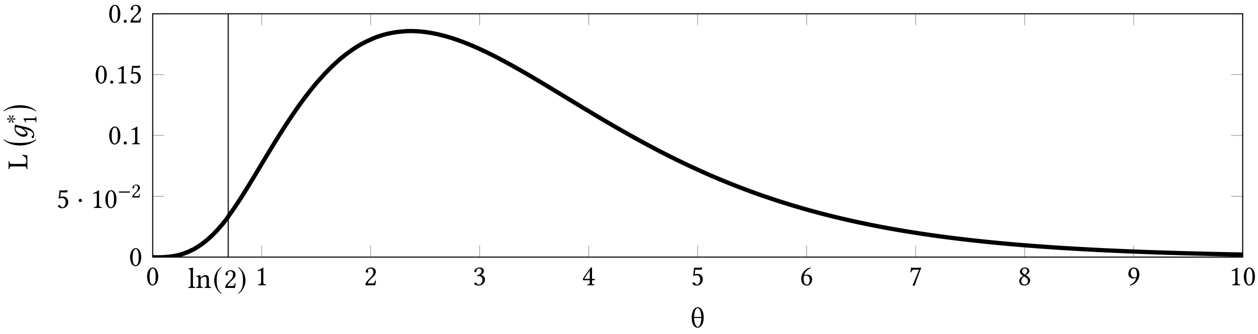

FIGURE 6 – Tracé du risque de Bayes en fonction de θ.

Lorsque $\theta = 9$, une approximation numérique donne $L\left(g_1^*\right) \approx 0.0046$.

3) Notons $\eta_2 : t \in \mathbb{R}_+ \mapsto \mathbb{P}\left(Y = 1 \mid T = t\right)$. Par construction, le classifieur de Bayes lorsque seule T est observée est

$$g_2^* : t \in \mathbb{R}_+ \mapsto \mathbb{1}_{\left\{\eta_2(t) > \frac{1}{2}\right\}}.$$

Soit $t \in \mathbb{R}_+$. Comme $U + V$ admet la même fonction de répartition F_2 que $T + U$, on a

$$\begin{aligned}\eta_2(t) &= \mathbb{P}\left(Y = 1 \mid T = t\right) = \mathbb{P}\left(T + U + V < \theta \mid T = t\right) \\ &= \mathbb{P}\left(V + U < \theta - t\right) \quad \text{(indépendance)} \\ &= \left(1 - (\theta - t + 1)\ e^{t-\theta}\right) \mathbb{1}_{\{\theta \geq t\}}.\end{aligned}$$

Malheureusement, nous ne pouvons trouver de solution analytique à l'équation $\eta_2(t) > \frac{1}{2}$. On peut remarquer en revanche que η_2 est continue strictement décroissante sur $[0, \theta]$, puis constante égale à 0 sur $]\theta, +\infty[$. Elle vaut $1 - (\theta + 1)\ e^{-\theta} = F_2(\theta)$ en 0, et 0 en θ. La fonction F_2 est quant à elle strictement croissante sur $\mathbb{R}_+$, vaut 0 en 0 et tend vers 1 quand $\theta \to +\infty$. Il existe donc $\theta_0 \in \mathbb{R}_+^*$ tel que $F_2(\theta_0) = \frac{1}{2}$ (TVI). Avec une approximation numérique on trouve $\theta_0 \approx 1.6783$. Par conséquent, si $\theta \leq \theta_0$, alors $\eta_2 \leq \frac{1}{2}$ et g_2^* affecte systématiquement le label 0 à toutes les observations. Si $\theta \geq \theta_0$, il existe en revanche $t_\theta \in]0, \theta[$ tel que $\eta_2(t) > \frac{1}{2}$ ssi $t < t_\theta$ (TVI) et pour tout $t \in \mathbb{R}_+$ on a $g_2^*(t) = \mathbb{1}_{\{t < t_\theta\}}$. Ces résultats sont illustrés à la Figure 7.

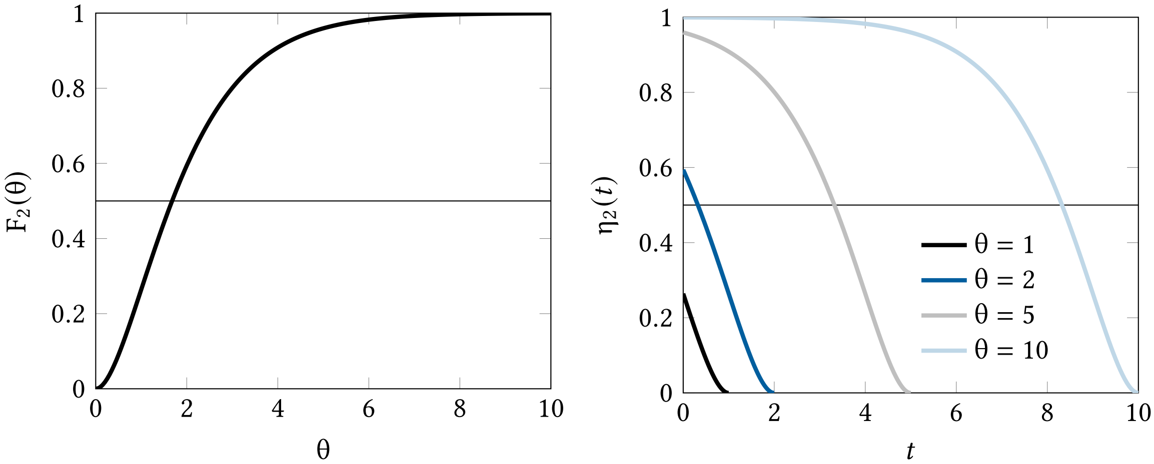

FIGURE 7 – Tracé de $\theta \in \mathbb{R}_+ \mapsto \mathrm{F}_2(\theta)$ (gauche) et de $t \in \mathbb{R}_+ \mapsto \eta_2(t)$ selon θ (droite).

Si $\theta \leq \theta_0$, le risque 0-1 de g_2^* est alors simplement

$$\mathrm{L}\left(g_2^*\right) = \mathbb{P}\left(\mathrm{Y}=1\right) = \mathbb{P}\left(\mathrm{T}+\mathrm{U}+\mathrm{V}<\theta\right) = \mathrm{F}_3(\theta) = 1 - \left(\frac{\theta^2}{2}+\theta+1\right) e^{-\theta}.$$

Si $\theta > \theta_0$, on obtient

$$\begin{aligned}
\mathrm{L}\left(g_2^*\right) &= \mathbb{E}\left(\eta_2(\mathrm{T})\, \mathbb{1}_{\left\{\eta_2(\mathrm{T}) \leq \frac{1}{2}\right\}} + \left(1-\eta_2(\mathrm{T})\right)\, \mathbb{1}_{\left\{\eta_2(\mathrm{T}) > \frac{1}{2}\right\}}\right) \\
&= \mathbb{E}\left(\left(1-(\theta-\mathrm{T}+1)\, e^{\mathrm{T}-\theta}\right) \mathbb{1}_{\{\theta \geq \mathrm{T}\}}\, \mathbb{1}_{\{\mathrm{T} \geq t_\theta\}}\right) \\
&\quad + \mathbb{E}\left(\left(1-\left(1-(\theta-\mathrm{T}+1)\, e^{\mathrm{T}-\theta}\right) \mathbb{1}_{\{\theta \geq \mathrm{T}\}}\right) \mathbb{1}_{\{\mathrm{T} < t_\theta\}}\right) \\
&= \mathbb{E}\left(\left(1-(\theta-\mathrm{T}+1)\, e^{\mathrm{T}-\theta}\right) \mathbb{1}_{\{t_\theta \leq \mathrm{T} \leq \theta\}}\right) \\
&\quad + \mathbb{E}\left((\theta-\mathrm{T}+1)\, e^{\mathrm{T}-\theta}\, \mathbb{1}_{\{\mathrm{T} < t_\theta\}}\right) \qquad \left(1 = \mathbb{1}_{\{\theta \geq \mathrm{T}\}} + \mathbb{1}_{\{\theta < \mathrm{T}\}}\right) \\
&= \int_0^{+\infty} \left(1-(\theta-t+1)\, e^{t-\theta}\right) \mathbb{1}_{\{t_\theta \leq t \leq \theta\}}\, e^{-t}\, dt \\
&\quad + \int_0^{+\infty} (\theta-t+1)\, e^{t-\theta}\, \mathbb{1}_{\{t < t_\theta\}}\, e^{-t}\, dt \qquad (\mathrm{T} \text{ de densité } f_1) \\
&= \int_{t_\theta}^{\theta} e^{-t}\, dt - \int_{t_\theta}^{\theta} (\theta-t+1)\, e^{-\theta}\, dt + \int_0^{t_\theta} (\theta-t+1)\, e^{-\theta}\, dt \\
&= \mathrm{F}_1(\theta) - \mathrm{F}_1(t_\theta) + e^{-\theta}\left((\theta+1)\,(t_\theta-\theta+t_\theta) + \int_{t_\theta}^{\theta} t\, dt - \int_0^{t_\theta} t\, dt\right) \\
&= e^{-t_\theta} + e^{-\theta}\left((\theta+1)\,(2\,t_\theta-\theta) - t_\theta^2 + \frac{\theta^2}{2} - 1\right).
\end{aligned}$$

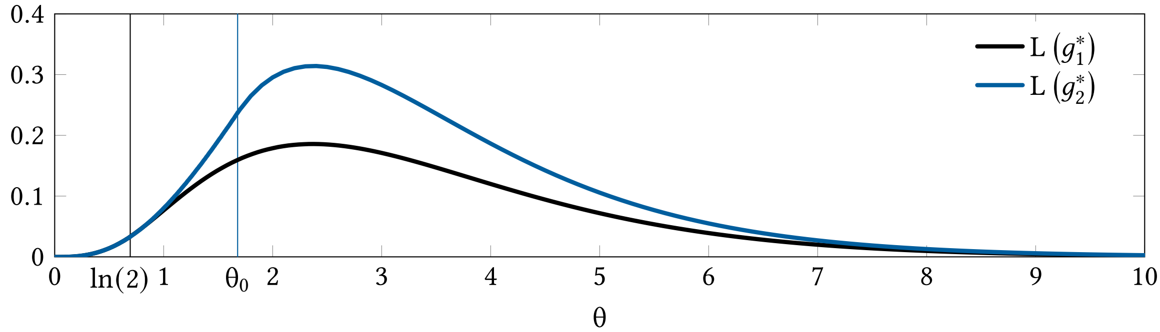

FIGURE 8 – Risques de Bayes en fonction de θ (approximation numérique pour $L\left(g_2^*\right)$).

Lorsque $\theta = 9$, une approximation numérique donne $L\left(g_2^*\right) \approx 0.0059$. On remarque que cette quantité est strictement supérieure à celle trouvée pour g_1^* lorsque $\theta = 9$. Il semble donc (de manière logique) qu'avec moins d'information à disposition, le risque d'erreur augmente. La Figure 8 confirme cette intuition pour tout $\theta > 0$.

4) En l'absence de toute information sur T, U et V, on peut toujours affecter systématiquement le label le plus probable. Nous avons vu à la question précédente que

$$\mathbb{P}\left(Y = 1\right) = F_3(\theta) = 1 - \left(\frac{\theta^2}{2} + \theta + 1\right) e^{-\theta}.$$

Il s'agit d'une fonction strictement croissante de θ, valant 0 en 0 et tendant vers 1 quand $\theta \longrightarrow +\infty$, il existe donc $\theta_1 \in \mathbb{R}_+^*$ tel que $F_3(\theta_1) = \frac{1}{2}$ (TVI). Une approximation numérique donne $\theta_1 \approx 2.6741$. Par conséquent, pour tout $\theta \leq \theta_1$, le label 0 est le plus probable, et inversement pour tout $\theta > \theta_1$. On peut donc prendre pour classifieur la constante $g_3^* = \mathbb{1}_{\{\theta > \theta_1\}}$.

Son erreur de classification est

$$\begin{aligned} L\left(g_3^*\right) &= \mathbb{P}\left(Y = 1\right) \mathbb{1}_{\{\theta \leq \theta_1\}} + \mathbb{P}\left(Y = 0\right) \mathbb{1}_{\{\theta > \theta_1\}} \\ &= \mathbb{1}_{\{\theta \leq \theta_1\}} + \left(\mathbb{1}_{\{\theta > \theta_1\}} - \mathbb{1}_{\{\theta \leq \theta_1\}}\right) \left(\frac{\theta^2}{2} + \theta + 1\right) e^{-\theta}. \end{aligned}$$

Pour $\theta = 9$ une approximation numérique donne $L\left(g_3^*\right) \approx 0.0062$. A nouveau, avec moins d'information, on augmente le risque de classification. La Figure 9 confirme ce constat pour tout $\theta > 0$.

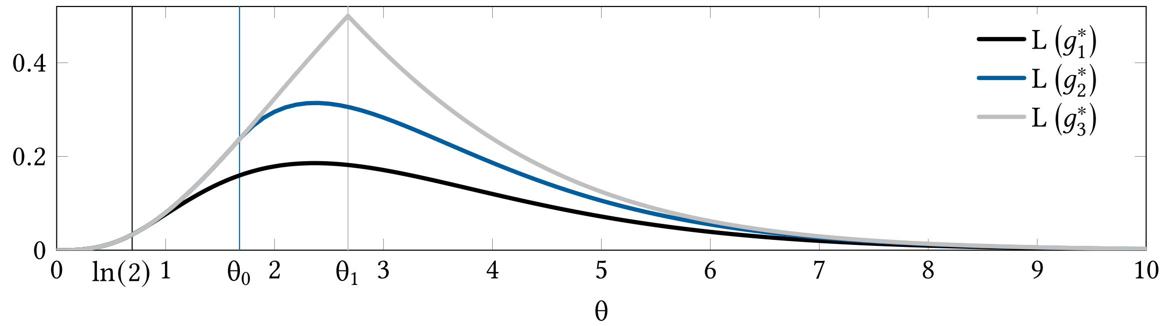

FIGURE 9 – Risques de Bayes en fonction de θ (approximation numérique pour $\mathrm{L}\left(g_2^*\right)$).