EXERCISE: INFERENCE FOR BINOMIAL SAMPLING

ERIK QUAEGHEBEUR AND GERT DE COOMAN

1. REPRESENTING THE DATA

Only part of the emails Gert receives are interesting; those he labels with a I. The others he labels with a U. Within a busy hour, he receives a sequence of mails and labels them as follows:

$$I \qquad U \qquad U \qquad U \qquad I \qquad I \qquad I \qquad U \qquad U \qquad I$$

We see that there are 4 interesting and 6 uninteresting mails. The data is completely described in a necessary and sufficient way by either

(1) the count vector of interesting and uninteresting mails: $n = (n_I, n_U) = (4,6)$,
(2) the total number of mails, $N = 10$, and the number (or relative frequency) of interesting mails, $n_I = 4$ ($f_I = \frac{2}{5}$), or
(3) the total number of mails, $N = 10$, and the number (or relative frequency) of uninteresting mails, $n_U = 6$ ($f_U = \frac{3}{5}$).

The the frequency vector $f = (f_I, f_U) = (\frac{2}{5}, \frac{3}{5})$ together with the total number of mails, $N = 10$ is also sufficient.

The frequency vector determines a probability mass function

$$p(z) = \begin{cases} \frac{2}{5}, & z = I, \\ \frac{3}{5}, & z = U \end{cases}$$

that can be used as a predictive inference model.

The imprecise probabilistic (predictive) inference model we obtain by mixing it with a vacuous model with mixing coefficient $\varepsilon \in [0,1]$ is characterized by the credal set $\{(1-\varepsilon)p + \varepsilon q : q \in \Sigma_{I,U}\}$. The corresponding lower and upper probability mass functions are

$$\underline{p}(z) = \begin{cases} (1-\varepsilon)\frac{2}{5}, & z = I, \\ (1-\varepsilon)\frac{3}{5}, & z = U, \end{cases} \qquad \overline{p}(z) = \begin{cases} (1-\varepsilon)\frac{2}{5} + \varepsilon, & z = I, \\ (1-\varepsilon)\frac{3}{5} + \varepsilon, & z = U. \end{cases}$$

The concrete values obtained when we consider the linear-vacuous models generated by adding one pseudocount ($\varepsilon = \frac{1}{11}$) or five pseudocounts ($\varepsilon = \frac{1}{15}$) are

$$\underline{p}(z) = \begin{cases} \frac{4}{11}, & z = I, \\ \frac{6}{11}, & z = U, \end{cases} \quad \overline{p}(z) = \begin{cases} \frac{5}{11}, & z = I, \\ \frac{7}{11}, & z = U. \end{cases} \qquad \underline{p}(z) = \begin{cases} \frac{4}{15}, & z = I, \\ \frac{6}{15}, & z = U, \end{cases} \quad \overline{p}(z) = \begin{cases} \frac{9}{15}, & z = I, \\ \frac{11}{15}, & z = U. \end{cases}$$

2. BRINGING IN THE SAMPLING MODEL

Now assume that the category of the emails is determined by a Bernoulli process: iid repetitions of, conceptually, a coin flip with probability $\theta_I = \vartheta \in [0,1]$ of turning out interesting and probability $\theta_U = 1 - \vartheta$ of not turning out interesting. So the probability mass function for each email is

$$p_{\mathrm{Br}}(z \,|\, \vartheta) = \begin{cases} \vartheta, & z = I, \\ 1 - \vartheta, & z = U. \end{cases}$$

Date: 24 July 2014.

Note that, in contrast to the case of Poisson sampling, it is always possible to find a value for ϑ that makes this probability mass function identical to the one derived from the frequency vector. Here, this value is $\frac{2}{5}$.

The likelihood function for a single-sample Bernoulli process is then

$$L_z(\vartheta) = \begin{cases} \vartheta, & z = I, \\ 1-\vartheta, & z = U. \end{cases}$$

Because of the iid assumption, this straightforwardly generalizes to the N-sample case:

$$\begin{aligned} L_x(\vartheta) &= \vartheta^{n_I}(1-\vartheta)^{n_U} = \vartheta^{n_I}(1-\vartheta)^{N-n_I} \\ &= \vartheta^{Nf_I}(1-\vartheta)^{Nf_U} = \vartheta^{Nf_I}(1-\vartheta)^{N-Nf_I}. \end{aligned}$$

It can be shown that the relative frequency f_I of interesting emails is the maximum likelihood estimate of ϑ. Fixing the sample size N, the f_I is—amongst others—a sufficient statistics.

3. Bringing in the prior

The conjugate prior for Bernoulli sampling is the Beta distribution, with density

$$p_{\mathrm{Be}}(\vartheta \,|\, \alpha, \beta) \coloneqq \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}.$$

For later convenience, we will use another parametrization, with $s \coloneqq \alpha+\beta$ and $t \coloneqq (t_I, t_U) \coloneqq (\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta})$, so $st_I = \alpha$, $st_U = \beta$ and

$$p_{\mathrm{Be}}(\vartheta \,|\, s, t) \coloneqq \frac{\Gamma(s)}{\Gamma(st_I)\Gamma(st_U)} \vartheta^{st_I-1}(1-\vartheta)^{st_U-1} = \frac{\Gamma(s)}{\Gamma(st_I)\Gamma(s(1-t_I)} \vartheta^{st_I-1}(1-\vartheta)^{s(1-t_I)-1}.$$

Recall that the (prior) mean for ϑ is given by $\frac{\alpha}{\beta} = t_I$ and the (prior) variance by $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{t_I t_U}{(s+1)}$.

Again due to the compatibility of the expressions of likelihood and conjugate prior, we can directly write down the expression for the posterior:

$$\begin{aligned} p(\vartheta \,|\, x) &\propto L_x(\vartheta) p_{\mathrm{Be}}(\vartheta \,|\, s, t) \\ &\propto \vartheta^{Nf_I}(1-\vartheta)^{Nf_U} \vartheta^{st_I-1}(1-\vartheta)^{st_U-1} \\ &= \vartheta^{Nf_I+st_I-1}(1-\vartheta)^{Nf_U+st_U-1} \\ &\propto p_{\mathrm{Be}}\left(\vartheta \,\middle|\, N+s, \frac{Nf+st}{N+s}\right). \end{aligned}$$

So, by normalization, the posterior is of the same conjugate form as the prior, now with parameters $N+s$ and $\frac{Nf_I+st}{N+s}$.

Gert is bored with our questions about his email experiences, so we have no idea what he thinks. We remain totally ignorant about the value of ϑ.

In the classical literature, we find many priors proposed for such a situation:

- Bayes's (uniform) prior $p_{\mathrm{Be}}(\vartheta \,|\, 2, (\frac{1}{2}, \frac{1}{2}))$. The corresponding distributions are visualized in Figure 1.
- Haldane's (improper) prior $p_{\mathrm{Be}}(\vartheta \,|\, 0, (t_I, t_U))$. The corresponding distributions are visualized in Figure 2.
- Jeffrey's (reparametrization invariant) prior $p_{\mathrm{Be}}(\vartheta \,|\, 1, (\frac{1}{2}, \frac{1}{2}))$. The corresponding distributions are visualized in Figure 3.

In the imprecise probabilistic literature, the *imprecise Beta model* and its multi-category variant the *imprecise Dirichlet model* are often used to model prior ignorance. Instead of relying on a specific shape of a single density—special choices for s and t—, the ignorance is expressed by considering all possible

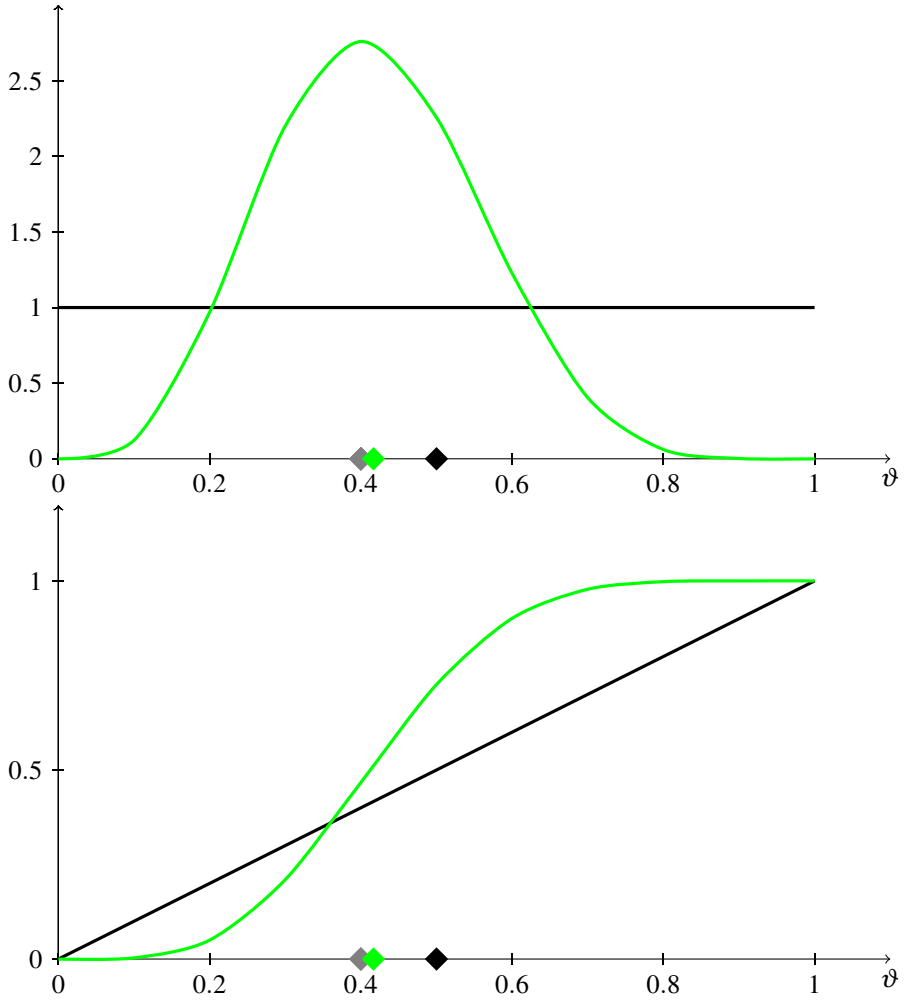

FIGURE 1. Probability density function plot and cumulative distribution function plot for Laplace's prior (in black) and the corresponding posterior (in green).

values for t; s can be chosen based on learning rate considerations (it is also possible to let it vary in some interval). We visualize the corresponding probabilistic inference models in Figure 4 and 5.

4. FROM PARAMETRIC TO PREDICTIVE INFERENCE

We can again derive predictive inference models from the parametric ones, i.e., to do inference about a sequence Y of M future observations. Again, if the parameter of the sampling model—now a Bernoulli process with probability ϑ—is known, the model for the sequence follows by the iid assumption; to wit, the joint probability mass function with expression

$$p(y \mid \vartheta) \coloneqq \prod_{i=1}^{M} p_{\mathrm{Br}}(y_i \mid \vartheta).$$

Because of the uncertainty about the parameter value, we need to take our model for ϑ into account. Using a single prior or posterior conjugate distribution, we obtain the predictive probability mass function by

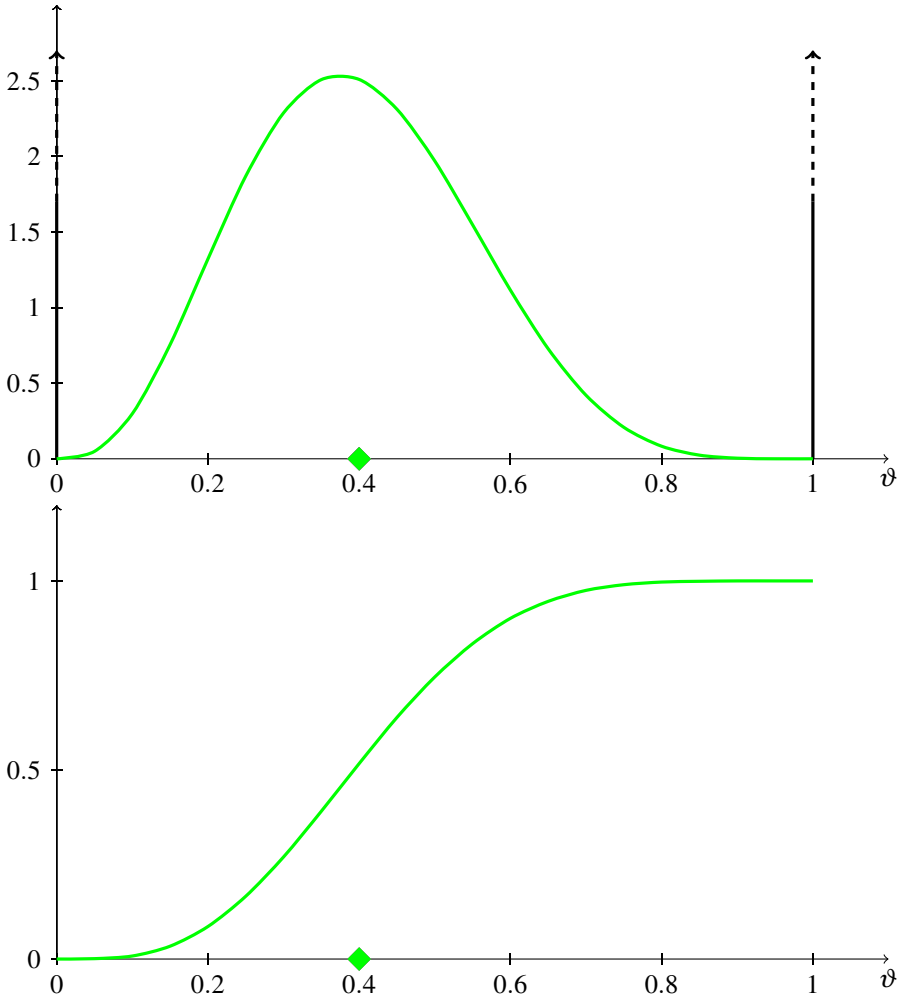

FIGURE 2. Probability density function plot and cumulative distribution function plot for Haldane's prior (in black) and the corresponding posterior (in green).

using this conjugate distribution to average over all possible values for ϑ:

$$\begin{aligned} p(y\,|\,\tilde{s},\tilde{t}) &= \int_0^1 p(y\,|\,\vartheta)\,p(\vartheta\,|\,\tilde{s},\tilde{t})\,\mathrm{d}\vartheta \\ &= \frac{\Gamma(\tilde{s})}{\Gamma(\tilde{s}\tilde{t}_I)\Gamma(\tilde{s}\tilde{t}_U)}\,\frac{\Gamma(m_I+\tilde{s}\tilde{t}_I)\Gamma(m_U+\tilde{s}\tilde{t}_U)}{\Gamma(M+\tilde{s})}\int_0^1 p\left(\vartheta\,\middle|\,M+\tilde{s},\frac{m+\tilde{s}\tilde{t}}{M+\tilde{s}}\right)\mathrm{d}\vartheta \\ &= \frac{\frac{\Gamma(m_I+\tilde{s}\tilde{t}_I)}{\Gamma(\tilde{s}\tilde{t}_I)}\,\frac{\Gamma(m_U+\tilde{s}\tilde{t}_U)}{\Gamma(\tilde{s}\tilde{t}_U)}}{\frac{\Gamma(M+\tilde{s})}{\Gamma(\tilde{s})}} \\ &= \frac{m_I!\,m_U!}{M!}\,\frac{\binom{m_I+\tilde{s}\tilde{t}_I-1}{m_I}\binom{m_U+\tilde{s}\tilde{t}_U-1}{m_U}}{\binom{M+\tilde{s}-1}{M}} = \frac{1}{\binom{M}{m_I}}\,\frac{\binom{m_I+\tilde{s}\tilde{t}_I-1}{m_I}\binom{m_U+\tilde{s}\tilde{t}_U-1}{m_U}}{\binom{M+\tilde{s}-1}{M}}, \end{aligned}$$

where $\tilde{s}$ and $\tilde{t}$ refer to either the prior or the posterior parameters and we made use of the fact that $p(y\,|\,\lambda) = L_y(\lambda)$. This is the expression of the (compound) Beta-binomial distribution's probability mass

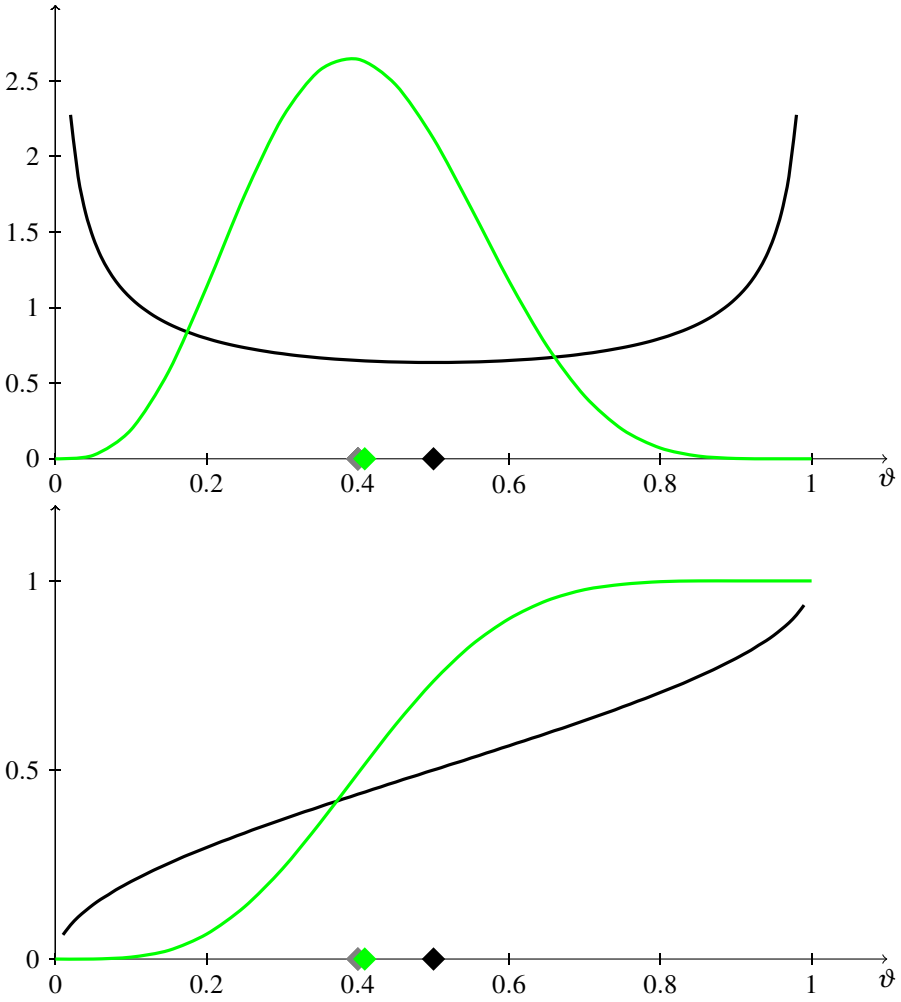

FIGURE 3. Probability density function plot and cumulative distribution function plot for Jeffrey's prior (in black) and the corresponding posterior (in green).

function, with t_I the probability that the mail is interesting. (In case of observation of counts and not of sequences, i.e., of *count* binomial sampling, the expression must be multiplied by $\binom{M}{m_I}$.) Again, in general, this expression cannot be simplified further and needs to be evaluated numerically in this form.

Now focus on prediction of the next observation Z, so with $M := 1$. Then the predictive probability mass function's expression becomes:

$$p(z \,|\, \tilde{s}, \tilde{t}) := \frac{\binom{m_I + \tilde{s}\tilde{t}_I - 1}{m_I} \binom{m_U + \tilde{s}\tilde{t}_U - 1}{m_U}}{\binom{1+\tilde{s}}{1}} = \begin{cases} \tilde{t}_I, & z = I, \\ \tilde{t}_U, & z = U. \end{cases}$$

This is the expression of the (compound) Beta-Bernoulli distribution's probability mass function, with t_I the probability that the mail is interesting. (Actually, it coincides with the Bernoulli distribution.) Below, we give expressions and values for the immediate predictive models corresponding to the parametric models given above (Laplace, Haldane, Jeffrey, and then $t_I \in [0,1]$ for $s := 1$ and $s := 5$):

Laplace: prior: $s = 2$, $t = (\frac{1}{2}, \frac{1}{2})$, posterior: $s + N = 12$, $t = (\frac{5}{12}, \frac{7}{12})$ or $t_I \approx 0.417$;

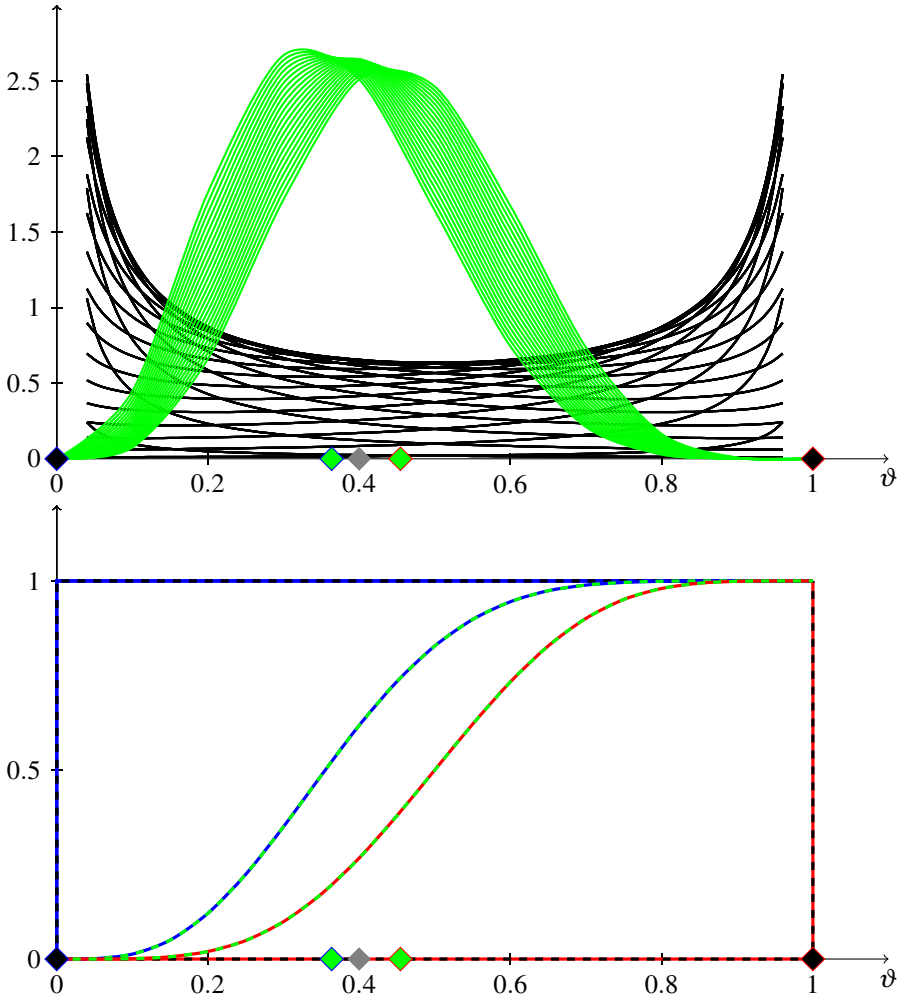

FIGURE 4. Lower and upper probability density function plot and pbox plot for the imprecise Beta model with $s := 1$: prior (in black) and the corresponding posterior (in green).

Haldane: prior: $s = 0$, $t = (t_I, t_U$, posterior: $s + N = 10$, $t = (\frac{2}{5}, \frac{3}{5})$ or $t_I = 0.4$;
Jeffrey: prior: $s = 1$, $t = (\frac{1}{2}, \frac{1}{2})$, posterior: $s + N = 11$, $t = (\frac{9}{22}, \frac{13}{22})$ or $t_I \approx 0.409$;
Imprecise:
prior: $s = 1$, $t_I \in [0,1]$, posterior: $s = 11$, $t_I \in [\frac{4}{11}, \frac{5}{11}] \approx [0.363, 0.455]$;
prior: $s = 5$, $t_I \in [0,1]$, posterior: $s = 15$, $t_I \in [\frac{4}{15}, \frac{3}{5}] \approx [0.266, 0.6]$.

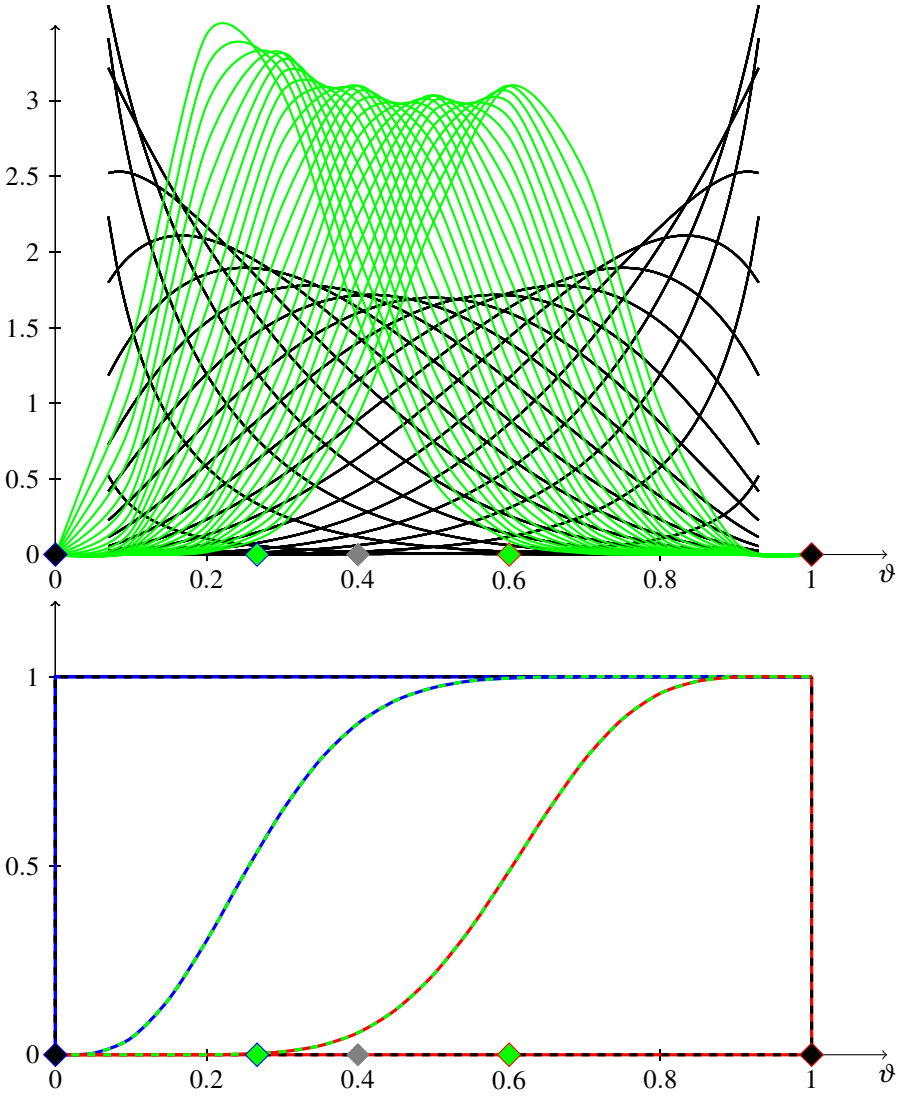

FIGURE 5. Lower and upper probability density function plot and pbox plot for imprecise Beta models with $s := 5$: prior (in black) and the corresponding posterior (in green).