Abstract This paper proposes a model of prior ignorance about a scalar variable based on a set of distributions \( \mathcal{M} \). In particular, a set of minimal properties that a set \( \mathcal{M} \) of distributions should satisfy to be a model of prior ignorance without producing vacuous inferences is defined. In the case the likelihood model is in the one-parameter exponential family of distributions, it is shown that the above minimal properties are equivalent to a special choice of the domains for the parameters of the conjugate exponential prior. This makes it possible to define the largest (that is, the least-committal) set of conjugate priors \( \mathcal{M} \) that satisfies the above properties. The obtained set \( \mathcal{M} \) is a model of prior ignorance w.r.t. the functions (queries) that are commonly used for statistical inferences; it is easy to elicit and, because of conjugacy, tractable; it encompasses frequentist and "objective" Bayesian inferences with improper priors. An application of the model to a problem of inference with count data is presented.

1 Introduction

Scientific experimental results generally consist of sets of data \( \{y_1, \ldots, y_n\} \). Statistical methods are then typically used to derive conclusion on both the nature of the process, which has produced those observations, and on the expected behaviour at future instances of the same process. A central element of any statistical analysis is the specification of a probability model: the likelihood. This is assumed to describe the
mechanism that has generated the observations as a function of a parameter $w \in \mathbb{R}$ (in the following we assume that $w$ is scalar), the so-called state-of-nature, about which only limited information (if any) is available. In the Bayesian paradigm, this information is modelled by means of a probability distribution (a prior) too, where the probabilistic model over $w$ is not a description of the variability of $w$ (parameters are typically fixed unknown quantities) but a description of the uncertainty about their values before observing the data.

An important problem in Bayesian analysis is how to define the prior distribution. If any prior information about the parameter $w$ is available, it should be incorporated in the prior distribution. On the other hand, in the case (almost) no prior information is available on $w$, the prior should be selected so as to reflect such state of ignorance. The search for a prior distribution representing “ignorance” constitutes a fascinating chapter in the history of Bayesian Statistics [1]. There are two main avenues to represent ignorance.

The first assumes that ignorance can be modelled satisfactorily by a single “non-informative” prior density such as for instance Laplace’s prior, Jeffreys’ prior, or the reference prior by Bernardo (see [1, Sec. 5.6.2] for a review). This view has been questioned on diverse grounds. Noninformative priors are typically improper and may lead to an improper posterior. Moreover, even if the posterior is proper, it can be inconsistent with the likelihood model (i.e., incoherent in the subjective interpretation of probability [2, Ch. 7], as it will be shown with an example in Section 2). On our view, however, the most important criticism to noninformative priors is that they are not enough expressive to represent ignorance (this will become more precise later in this Introduction).

An alternative is to use a set of prior distributions, $\mathcal{M}$, rather than a single distribution, to model prior ignorance about statistical parameters. Each prior distribution in $\mathcal{M}$ is updated by Bayes’ rule, producing a set of posterior distributions. In fact there are two distinct approaches of this kind, which have been compared by Walley [2]. The first approach, known as Bayesian sensitivity analysis or Bayesian robustness ([3, 7]), assumes that there is an ideal prior distribution $\pi_0$ which could, ideally, model prior uncertainty. It is assumed that we are unable to determine $\pi_0$ accurately because of limited time or resources. The criterion for including a particular prior distribution $\pi$ in $\mathcal{M}$ is that $\pi$ is a plausible candidate to be the ideal distribution $\pi_0$. However, this approach can be unsatisfactory when there is little prior information or the information is of doubtful relevance. Then there is no ideal prior distribution, because no single prior distribution could adequately model the limited prior information.

The second approach, known as the theory of imprecise probabilities or coherent lower (and upper) previsions, was developed by Walley [2] from earlier ideas [4, 7]. This approach generalizes Bayesian sensitivity analysis by directly emphasizing the upper and lower expectations (also called previsions) that are generated by $\mathcal{M}$. The upper and lower expectations of a bounded function (we call it a gamble) $g : \mathcal{W} \to \mathbb{R}$, denoted by $\mathcal{E}_U(g)$ and $\mathcal{E}_L(g)$, are respectively the supremum and infimum of the expectations $\mathcal{E}_P(f)$ over the probability measures $P$ in $\mathcal{M}$ (if $\mathcal{M}$ is assumed to be closed and convex, it is fully determined by all the upper and lower expectations). The up-

---

1 Closed and convex in the weak* topology, see [2, Sec. 3.6] for more details.
per and lower expectations have a behavioural interpretation (explained in Section 2),
but, contrary to the robust Bayesian approach, there is no special commitment to the
individual probability distributions in $\mathcal{M}$.

In choosing a set $\mathcal{M}$ to model prior ignorance, the main aim is to generate upper
and lower expectations with the property that $E(g) = \inf_w g(w)$ and $\overline{E}(g) = \sup_w g(w)$
for suitable gambles $g$. This means that the only information about $E(g)$ is that it belongs to $[\inf_w g(w), \sup_w g(w)]$, which is equivalent to state a condition of complete prior ignorance about the value of $g$ (this is the reason why we said that a single, however “noninformative”, prior cannot model prior ignorance).

For instance, assume that $w$ is a location parameter and that $\mathcal{W} = \mathbb{R}$, then in
the case there is no prior information about $w$, one expects that for $g = I_{(-\infty,a]}$ for
any finite $a \in \mathbb{R}$, the lower and upper expected values of $g$ being $E(I_{(-\infty,a]}) =$
0 and $\overline{E}(I_{(-\infty,a]}) = 1$, where $I_{(-\infty,a]}$ denotes the indicator function over the set $(-\infty,a]$, i.e., $I_{(-\infty,a]}(w) = 1$ if $w$ belongs to $(-\infty,a]$, and $I_{(-\infty,a]}(w) = 0$ otherwise.
Since $E(I_{(-\infty,a]})$ is equal to the cumulative distribution of $w$, $E(I_{(-\infty,a]}) = 0$ and
$\overline{E}(I_{(-\infty,a]}) = 1$ state that the only knowledge about the cumulative distribution of $w$ is that it is between 0 and 1, which is a state of complete ignorance.

Modeling a state of prior ignorance on $w$ is not the only requirement for $\mathcal{M}$, it
should also lead to non-vacuous posterior inferences. Posterior inferences are vacuous if the lower and upper expectations of all gambles $g$ coincide with the infimum and, respectively, the supremum of $g$. Notice that, in the case $\mathcal{M}$ includes all the possible prior distributions, any inference about $w$ is vacuous, i.e., the set of posterior distributions obtained by applying Bayes’ rule to a given likelihood and to each distribution in the set of priors includes again all the possible distributions. This means that our prior beliefs do not change with experience (i.e., there is no learning from data). This point is clearly stated in [5], where the authors define a some properties that a general set $\mathcal{M}$ of distributions should have to model a state of prior ignorance about $w$. In particular, $\mathcal{M}$ should produce self-consistent probabilistic models, represent the state of prior ignorance, give non-vacuous inferences, satisfy certain invariance properties, be easy to elicit and tractable [5].

In this paper, we follow the second approach to model prior ignorance about statistical parameters by using Walley’s theory of coherent lower previsions [2]. Let us summarize the main contributions of this paper.

In Section 2, inspired by the work in [5], we define some minimal properties that a set $\mathcal{M}$ of distributions should satisfy to be a model a prior ignorance that does not lead to vacuous inferences. These minimal properties are obtained by relaxing and generalizing the properties in [5]. The new set of properties now captures also the case in which the likelihood is not perfectly known.

Next, we focus on the exponential family. We give some preliminaries about such a family in Section 3. In Section 4, we consider the case that the likelihood model is in the one-parameter exponential family and $\mathcal{M}$ includes the corresponding conjugate exponential priors. We show that there exists a parametrization of $\mathcal{M}$ which equivalently satisfies the properties defined in Section 2, and which is unique up

\footnote{In the following, we use the notation $g$ without argument to denote the gamble, while $g(w)$ is used to denote the value of the gamble $g$ at $w$.}
to the choice of its size which determines the degree of robustness (or caution) of
the inferences. Stated differently, we prove that the set of priors \( \mathcal{M} \) satisfying the
above properties can be uniquely obtained by letting the parameters of the conjugate
exponential prior vary in suitable sets. The obtained set \( \mathcal{M} \) has the following charac-
teristics, which could make of it an appealing alternative to noninformative priors to
model prior ignorance:

- \( \mathcal{M} \) produces self-consistent (or coherent) probabilistic models;
- \( \mathcal{M} \) is a model of prior ignorance w.r.t. the functions (queries) that are commonly
  used for statistical inferences (i.e., lower/upper expectations of such functions are
  vacuous a-priori);
- it is easy to elicit and, because of conjugacy, tractable;
- the inferences drawn with the model \( \mathcal{M} \) encompass, with a suitable choice of the
  size of \( \mathcal{M} \), the frequentist inferences and “objective” Bayesian inferences with
  improper priors, which makes of \( \mathcal{M} \) a naturally robust approach to inference.

We compare the obtained set of priors \( \mathcal{M} \) with other models of prior ignorance
expressed through set of distributions in Section 4.1. In Section 5, we discuss the
choice of the parameters governing the robustness of the inferences that are obtained
through \( \mathcal{M} \), while in Section 6 we show the application of these ideas to a problem
of inference with count data.

2 Properties for near-prior ignorance

The aim of this section is to define which minimal properties the set of priors \( \mathcal{M} \)
should satisfy in the case where there is (almost) no prior information about \( w \in \mathcal{W} \subseteq \mathbb{R} \). Before listing these properties, we discuss the interpretation of \( \mathcal{M} \) by briefly
introducing the behavioural interpretation of upper and lower expectations.

behavioural interpretation of \( \mathcal{M} \) in terms of buying and selling prices. Let us briefly
sketch how this is done.

By regarding a gamble \( g : \mathcal{W} \rightarrow \mathbb{R} \) as a random reward, which depends on the a
priori unknown value of \( w \), the expectation (also called prevision) of \( g \) w.r.t. \( w \), i.e.,
\( E(g) \), represents a subject’s fair price for the function \( g \). This means that he should
be disposed to accept the uncertain rewards \( g - E(g) + \epsilon \) (i.e., to buy \( g \) at the price
\( E(g) - \epsilon \)) and \( E(g) - g + \epsilon \) (i.e., to sell \( g \) at the price \( E(g) + \epsilon \)) for every \( \epsilon > 0 \).

More generally, the supremum acceptable buying price and the infimum accept-
able selling prices for \( g \) need not coincide, meaning that there may be a range of prices
\([a, b]\) for which our subject is neither disposed to buy nor to sell \( g \) at a price \( k \in [a, b] \).
His supremum acceptable buying price for \( g \) is then his lower expectation \( E(g) \), and it
holds that the subject is disposed to accept the uncertain reward \( g - E(g) + \epsilon \) for
every \( \epsilon > 0 \); and his infimum acceptable selling price for \( g \) is his upper prevision
\( \overline{E}(g) \), meaning that he is disposed to accept the reward \( \overline{E}(g) - g + \epsilon \) for every \( \epsilon > 0 \).
A consequence of this interpretation is that \( E(g) = -\overline{E}(-g) \) for every function \( g \).

Under this behavioural interpretation, a state of ignorance about a gamble \( g \) is
modelled by setting \( E(g) = \inf_w g(w) \) and \( \overline{E}(g) = \sup_w g(w) \). This means that our
subject is neither disposed to buy nor to sell $g$ at any price $k \in (\inf_\omega g(\omega), \sup_\omega g(\omega))$.  
In other words, our subject is disposed to buy (sell) $g$ only at a price less (greater) or equal than the minimum (maximum) reward that he would gain from $g$. This means that the available information on $\omega$ does not allow our subject to set any meaningful buying or selling price for $g$, which is equivalent to state that our subject is in a state of ignorance.

In [2], it is proven that a closed and convex set of probability distributions can be equivalently characterized by the lower (or upper) expectation functional that it generates as the lower (upper) envelope of the expectations obtained from the distributions in such a set. Vice versa, given a functional $E(g)$ that satisfies some regularity properties [2, Ch. 2], it is possible to define a family $\mathcal{M}$ of probability distributions that generates the lower expectation $E(g)$ for any $g$. This establishes a one-to-one correspondence between closed convex sets of probability distributions and lower expectations.

In the case the available prior information is scarce, it seems thus more natural to define $\mathcal{M}$ according to the behavioural interpretation, i.e., in terms of the upper and lower expectations it generates [5]. For instance, in problems where there is (almost) no prior information one would expect the set $\mathcal{M}$ to be “large” in the sense that its generated upper and lower expectations are far apart (vacuous or almost vacuous).

Modelling a state of prior ignorance on $\omega$ is not the only requirement for $\mathcal{M}$, it must also produce non-vacuous posterior inferences (otherwise it is useless in practice). Hereafter, inspired by the work in [5], we define a set of minimal properties that $\mathcal{M}$ or, equivalently, the lower and upper expectations it generates, should satisfy to be a model of prior ignorance and produce consistent and meaningful posterior inferences. Then, in the next sections, we specialize these requirements to the case of the exponential family. The first requirement for $\mathcal{M}$ is coherence.

(A.1) Coherence. Inferences based on $\mathcal{M}$ should be strongly coherent [2, Sec. 7.1.4(b)].

Under the behavioural interpretation, this means that we should not be able to raise the lower expectation (supremum acceptable buying price) of a given gamble $g$ taking into account the acceptable transactions implicit in the other lower expectation models.

In practice, strong coherence imposes joint constraints on the prior, likelihood and posterior lower expectation models, in the sense that, when considered jointly, they should not imply inconsistent assessments. In [2, Sec. 7.8.1], it is proven that, in the case the prior and likelihood lower expectation models are obtained as lower envelopes of standard expectations w.r.t. sets of proper density functions and the posterior set of densities is obtained from these sets by element-wise application of Bayes’ rule for density functions, then strong coherence of the respective lower expectation models is satisfied. The following example shows that, when this is not the case, the inference model can be incoherent [2, Sec. 7.4.4].

Example 1. Consider the following pair of Normal models: likelihood

$$\prod_{i=1}^{n} \mathcal{N}(y_i; w, \sigma^2)$$

with variance $\sigma^2$ and posterior $\mathcal{N}(w; \bar{y}_n, \sigma^2/n)$ with mean $\bar{y}_n$

---

3 This holds under standard assumptions on the existence of density functions and applicability of Bayes’ rule.
To see that the inferences are incoherent, consider the event $A$ that $|w| \leq |\hat{y}_n|$ and, thus, the gamble $g = I_A$. Then, from the likelihood model it follows that $P(A|w) = E[I_A|w] = \frac{1}{2} + \Phi(-2|w|/\sqrt{\sigma}) > \frac{1}{2}$, for each real $w$, where $\Phi(\cdot)$ denotes the standard Normal distribution. By considering the posterior model, one has $P(A|y_1, \ldots, y_n) = E[I_A|y_1, \ldots, y_n] = \frac{1}{2} - \Phi(-2|\hat{y}_n|/\sqrt{\sigma}) < \frac{1}{2}$, for each $y_1, \ldots, y_n$. Thus the likelihood and posterior models are inconsistent. In fact, since $\Phi(-2|\hat{y}_n|/\sqrt{\sigma}) \to 0$ as $n \to \infty$, it follows that $P(A|w) \to 0.5$ (no matter the value of $w$). Notice that the posterior $\mathcal{M}(w; \hat{y}_n, \sigma^2/n)$ can be obtained, via Bayes’ rule, from the likelihood and the improper prior $p(w) = 1$. A criticism of the improper prior is, in fact, that the posterior distribution it generates is often not coherent with the likelihood model for finite $n$. Finally, observe that, for $n \to \infty$, it follows that $P(A|w), P(A|y_1, \ldots, y_n) \to 0.5$ and, thus, the incoherence vanishes in the limit.

Besides coherence, other requirements for the set $\mathcal{M}$ are that it should represent the state of prior ignorance on $w$, but without producing vacuous posterior inferences (posterior inferences are vacuous if the lower and upper expectations of all gambles $g$ coincide with the infimum and, respectively, the supremum of $g$). In the case $\mathcal{M}$ includes all the possible prior distributions, any inference about $w$ is vacuous, i.e., the set of posterior distributions obtained by applying Bayes’ rule to a given likelihood and to each distribution in the set of priors includes again all the possible distributions. This means that our prior beliefs do not change with experience (there is no learning from data). Thus, $\mathcal{M}$ should be large enough to model a state of prior ignorance w.r.t. a set of suitable gambles (i.e., a set of gambles of interest $\mathcal{G}_0$ w.r.t. which we assess our state of prior ignorance), but no too large to prevent learning from taking place. These two contrasting requirements are captured by the following two properties for $\mathcal{M}$.

(A.2) $\mathcal{G}_0$-prior ignorance. The prior upper and lower expectations of some suitable set of gambles $\mathcal{G}_0$ under $\mathcal{M}$ are vacuous, i.e., $E[g] = \inf_w g(w)$ and $\bar{E}[g] = \sup_w g(w)$ for all $g \in \mathcal{G}_0$.

(A.3) $\mathcal{G}$-learning. For a chosen set of gambles $\mathcal{G} \supseteq \mathcal{G}_0$ and for each $g \in \mathcal{G}$ satisfying $\bar{E}[g] - E[g] > 0$, there exists a finite $\delta > 0$ (possibly dependent on $g$) such that for each $n \geq \delta$ and non-empty set of observations $\{y_1, \ldots, y_n\}$, at least one of these two conditions is satisfied:

$$E[g|y_1, \ldots, y_n] \neq \inf_w g(w), \quad \bar{E}[g|y_1, \ldots, y_n] \neq \sup_w g(w).$$

where $E[g|y_1, \ldots, y_n]$ and $\bar{E}[g|y_1, \ldots, y_n]$ denote the posterior lower and upper expectations of $g$ after having observed $\{y_1, \ldots, y_n\}$. Furthermore, for each $g \in \mathcal{G}_0$, (1) must hold for any $\delta > 0$.

Property (A.2) states that $\mathcal{M}$ should be vacuous a priori w.r.t. some set of gambles $\mathcal{G}_0$, i.e., the lower and upper expectations of $g \in \mathcal{G}_0$ coincides with the infimum and,
respectively, the supremum of \( g \). In the case \( \mathcal{M} \) includes all possible distributions then (A.2) holds for any function \( g \). Here, conversely, we require that (A.2) is satisfied for some subset of gambles \( \mathcal{G}_0 \). The subset of gambles \( \mathcal{G}_0 \) used in (A.2) should include the gambles \( g \) w.r.t. which we state our condition of prior ignorance. Furthermore, the set \( \mathcal{G}_0 \) should be as large as possible to guarantee that also \( \mathcal{M} \) is as large as possible, but no too large to be incompatible with the requirement (A.3) of learning. In fact, that property (A.3) states that \( \mathcal{M} \) should be non-vacuous a posteriori for any gamble \( g \in \mathcal{G} \supset \mathcal{G}_0 \), which is a condition for learning from the observations. The set of gambles \( \mathcal{G} \) used in (A.3) should include the gambles \( g \) w.r.t. which we are interested in computing expectations (i.e., making inferences). The fact that \( \mathcal{G} \) must include \( \mathcal{G}_0 \) is the only constraint on \( \mathcal{G} \), meaning that (A.3) requires that \( \mathcal{M} \) is not vacuous w.r.t. all the gambles for which the near-prior ignorance has been imposed. Moreover, for these gambles, it is required that (1) holds for any \( \delta > 0 \), i.e., after one observation the condition of prior ignorance must already be left.

Since \( \mathcal{M} \) is a model of prior ignorance, it is also desirable that the influence of \( \mathcal{M} \) on the posterior inferences vanishes increasing the number of observations \( n \). This is captured by the following property.

(A.4) Convergence. For each gamble \( g \in \mathcal{G} \) and non-empty set of observations \( \{y_1, \ldots, y_n\} \), the following conditions are satisfied for \( n \to \infty \):

\[
\begin{align*}
E[g|y_1, \ldots, y_n] & \to E^*[g|y_1, \ldots, y_n], \\
E[g|y_1, \ldots, y_n] & \to E^*[g|y_1, \ldots, y_n],
\end{align*}
\]

where \( E^*[g|y_1, \ldots, y_n] \) are the posterior lower and upper expectations obtained as lower envelopes of standard expectations w.r.t. the posterior densities derived, via Bayes’ rule, from the likelihood model and the improper prior density \( p(w) = 1 \).

Property (A.4) states that, for \( n \to \infty \), \( \mathcal{M} \) should give the same lower and upper expectations of \( g \in \mathcal{G} \) as those obtained from the improper prior density \( p(w) = 1 \). The fact that \( E^*[g|y_1, \ldots, y_n] < E^*[g|y_1, \ldots, y_n] \) accounts for the general case in which the likelihood model is described by a set of likelihoods (for a single likelihood it would be \( E^*[g|y_1, \ldots, y_n] = E^*[g|y_1, \ldots, y_n] \)). Although improper priors produce posteriors which are often incoherent with the likelihood model, (A.4) does not conflict with the requirement of coherence in (A.1). In fact (A.4) is a limiting property that holds only for \( n \to \infty \) (furthermore in Example 1, incoherence vanishes at the limit). In order to better understand properties (A.1)–(A.4), in Section 4 we will show their instantiation to the case of the exponential family. Before discussing these results, in the next section we introduce the exponential family of distributions and review its main properties [1, Ch. 5].

3 Exponential Family

Consider a sampling model where i.i.d. samples of a random variable \( Z \) are taken from a sample space \( \mathcal{X} \) that is distributed according to an exponential family.
Definition 1. A probability density \( p(z|x) \), parametrized by \( x \in \mathcal{X} \subseteq \mathbb{R} \), is said to belong to the one-parameter exponential family if it is of the form
\[
p(z|x) = f(z)g(x)\exp(c\phi(x)h(z)), \quad z \in \mathcal{Z}
\]
where, given \( f, h, \phi \) and \( c \), it results that
\[
[g(x)]^{-1} = \int_{z \in \mathcal{X}} f(z)\exp(c\phi(x)h(z)) \, dz < \infty.
\]

Sometimes it is more convenient to rewrite (3) in a different form.

Definition 2. The probability density
\[
p(y|w) = k(y)\exp(yw - b(w)), \quad y \in \mathcal{Y},
\]
derived from (3) via the transformations \( y = h(z) \) and \( w = c\phi(x) \), is called the canonical form of representation of the exponential family; \( w \) is called the natural (or canonical) parameter.

The canonical form has some useful properties. The mean and variance of \( y \) are given by
\[
E[Y|w] = \frac{\partial}{\partial w}b(w),
\]
\[
E[(Y - E[Y|w])^2|w] = \frac{\partial^2}{\partial w^2}b(w),
\]
where it has been assumed that \( \frac{\partial^2}{\partial w^2}b(w) > 0 \) and \( \frac{\partial}{\partial w}b(w) \in \text{Int}(\mathcal{Y}) \) for each \( w \), with \( \text{Int}(\mathcal{Y}) \) denoting the interior of \( \mathcal{Y} \) [7]. Second, in the case \( n \) observations \( y_i = h(z_i) \) are available, it follows that
\[
\prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i)\exp(n\hat{y}_n w - b(w)),
\]
where \( \hat{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i \) is the sample mean of \( y_i \) which, thus, is a sufficient statistic for \( w \). Furthermore, by interpreting the density function in (7) as a likelihood function \( L(w) \), with \( y = [y_1, \ldots, y_n]^T \), we can define the corresponding conjugate prior.

Definition 3. A probability density \( p(w|n_0, y_0) \), parametrized by \( n_0 \in \mathbb{R}^+ \) and \( y_0 \in \text{Int}(\mathcal{Y}) \), is said to be the canonical prior of (4) if
\[
p(w|n_0, y_0) = k(n_0, y_0)\exp(n_0(y_0 w - b(w))), \quad w \in \mathcal{W},
\]
where \( n_0 \) is the so-called number of pseudo-observations and \( \gamma_0 \) is the so-called pseudo-observation.

When \( \mathcal{W} = \mathbb{R}, 0 < n_0 < \infty \) and \( \gamma_0 \in \text{Int}(\mathcal{W}), \) (8) is a proper density [7]. Some examples of conjugate densities belonging to the one-parameter exponential (canonical) family and defined in \( \mathcal{W} = \mathbb{R} \) follow.

Gaussian with known variance \( x \in \mathcal{X} = \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \),

\[
p(x|\mu, \sigma^2) \propto \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \propto \exp \left( \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} \right),
\]

by setting \( n_0 = 1/\sigma^2, \gamma_0 = \mu, w = x \) and \( b(w) = x^2/2 \).

Beta \( x \in \mathcal{X} = (0, 1), t \in (0, 1), s \in \mathbb{R}^+ \),

\[
p(x|s, t) \propto x^{s-1}(1-x)^{t-1}
= \exp((st - 1) \ln(x) + (s(1-t) - 1) \ln(1-x))
= \exp((st \ln(x) + (s(1-t)) \ln(1-x)) \exp(-\ln(x(1-x)))
= (1-x)^s \exp \left( st \ln \left( \frac{x}{1-x} \right) \right) \frac{1}{x(1-x)}
= \exp(n_0(y_0 w - b(w)),
\]

where the last equality has been obtained by setting \( n_0 = s, y_0 = t, w = \ln(x/(1-x)), b(w) = -\ln(1-x) = \ln(1+\exp(w)) \) and considering the change of variable \( dx = \exp(w)/(1+\exp(w))^2 dw \).

The pair likelihood and conjugate prior in the canonical exponential family satisfy a set of interesting properties; most of them are particularly useful to represent the nature of the Bayesian “learning” process. A list of such properties is given in the following lemma, whose proof is omitted (see [1, Ch. 5]).

**Lemma 1.** For a pair of likelihood and conjugate prior in the canonical exponential family, it holds that:

(i) the posterior density for \( w \) is:

\[
p(w|n', y') = k(n', y') \exp(n' w - b(w)),
\]

where \( n' = n + n_0 \) and \( y' = \frac{n_0 y_0 + n y}{n + n_0} \);

(ii) the predictive density for future observations \( \{y_{n+1}, \ldots, y_{n+m}\} \) is

\[
p(y_{n+1}, \ldots, y_{n+m}|y_1, \ldots, y_n) = \prod_{j=1}^{m} k(y_{n+j}) \frac{k \left( n_0 + m, \frac{n_0 y_0 + n y}{n + n_0} \right)}{k \left( n_0 + n + m, \frac{n_0 y_0 + (n + m) y}{n + m + n_0} \right)}.
\]
Lemma 2. Suppose that the canonical conjugate prior family is such that \( p(w|n_0, y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \). Then the prior mean of the function \( \frac{\partial b(w)}{\partial w} \) is zero. The second term is equal to \( y_0 \), and the posterior mean:

\[
E_w \left[ \frac{\partial b(w)}{\partial w} | n_0, y_0 \right] = y_0 \quad \text{and the posterior mean:}
\]

\[
E_w \left[ \frac{\partial b(w)}{\partial w} | n', y' \right] = \frac{n_0 y_0 + n y_0}{n + n_0} . \tag{11}
\]

Proof: First of all, it holds that

\[
E_w \left[ \frac{\partial b(w)}{\partial w} | n_0, y_0 \right] = \frac{\int_{w \in \mathcal{W}} \frac{\partial b(w)}{\partial w} \exp(n_0(y_0w - b(w)))dw}{\int_{w \in \mathcal{W}} \exp(n_0(y_0w - b(w)))dw} . \tag{12}
\]

Consider for the moment only the numerator, then by integration by parts applied to the two functions \( \frac{\partial b(w)}{\partial w} \exp(-n_0b(w)) = -\frac{\partial}{\partial w} \exp(-n_0b(w)) \) and \( \exp(n_0y_0w) \), one gets

\[
\int_{w \in \mathcal{W}} \frac{\partial b(w)}{\partial w} \exp(n_0(y_0w - b(w)))dw = \int_{w \in \mathcal{W}} \frac{\partial b(w)}{\partial w} \exp(-n_0b(w)) \exp(n_0y_0w)dw
\]

\[
= -\frac{1}{n_0} \exp(-n_0b(w)) \exp(n_0y_0w) \sup_{w \in \mathcal{W}} w + \int_{w \in \mathcal{W}} \frac{n_0y_0}{n_0} \exp(-n_0b(w)) \exp(n_0y_0w)dw
\]

\[
\propto -\frac{1}{n_0} P(w|n_0, y_0) \sup_{w \in \mathcal{W}} + y_0 \int_{w \in \mathcal{W}} P(w|n_0, y_0)dw . \tag{13}
\]

Since, by hypothesis, \( P(w|n_0, y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \), the first term is zero. The second term is equal to \( y_0 \), since the integral cancels with the denominator. Finally, result (11) follows immediately from Lemma 1.

Notice that \( P(w|n_0, y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \) holds for any canonical priors such that \( \mathcal{W} = \mathbb{R} \), but in general it is not true for truncated priors, i.e., in the case \( \mathcal{W} \subset \mathbb{R} \). This is one of the reasons why it has been assumed that \( \mathcal{W} = \mathbb{R} \). In (5), it has been shown that \( \frac{\partial b(w)}{\partial w} \) is the mean of \( y \). Hence, \( \frac{\partial b(w)}{\partial w} \) is the quantity about which we will have prior beliefs before seeing the data \( y \) and posterior beliefs after observing the data. Hence, the results in Lemma 2 are particularly important, because they provide us with a closed formula for the prior and posterior mean of \( \frac{\partial b(w)}{\partial w} \). For sampling models such that \( \frac{\partial b(w)}{\partial w} = x \) (e.g., Gaussian, Beta and Gamma density), Lemma 2 gives thus a closed formula for the prior and posterior mean of \( x \).
Lemma 3. Suppose that the canonical conjugate prior family is such that \( p(w|n_0,y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \) and that \( \frac{1}{\sigma_w^2} b(w) = x \), then \( y_0 \) and (11) are the prior and, respectively, posterior mean of \( x \).

Hereafter, we give the instantiation of the above lemmas in the case of the conjugate Gaussian and Binomial model.

Gaussian with known variance. It has already been shown that
\[
\mathcal{N}(x; \mu, \sigma_0^2) = p(w|n_0,y_0) = k(n_0,y_0) \exp(n_0 (y_0 w - b(w))),
\]
with \( n_0 = 1/\sigma_0^2 \), \( y_0 = \mu \), \( w = x \) and \( b(w) = x^2/2 \). By considering the likelihood,
\[
L(y|w) = \prod_{i=1}^{n} \mathcal{N}(y_i;x,\sigma^2) = \prod_{i=1}^{n} k(y_i) \exp(n(\hat{y}_i w - b(w)))
\]
and assuming that the variance \( \sigma^2 \) is known, and redefining \( n = n/\sigma^2 \), then \( \mathcal{N}(x;\mu,\sigma_0^2) \) and \( L(y|w) \) are conjugate and, thus, the posterior density is
\[
\mathcal{N}(x;\hat{\mu},\hat{\sigma}^2) = k(n',y') \exp(n'(y' w - b(w)))
\]
where
\[
\hat{\mu} = y' = \frac{n_0 y_0 + n y}{n + n_0} = \sigma^2 \left( \frac{\mu}{\sigma_0^2} + \frac{n y}{\sigma^2} \right)
\]
\[
\hat{\sigma}^2 = \frac{1}{n'} = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2/n}
\]
Since \( \frac{1}{\sigma_w^2} b(w) = w = x \) and since \( p(w|n_0,y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \), then from Lemma 3 it follows that \( \hat{\mu} \) is also the posterior mean of \( x \). It is straightforward to verify that \( \hat{\sigma}^2 \) is also the variance of \( x \).

Beta. For Beta density it has been shown that
\[
\text{Beta}(x|s,t) = k(n_0,y_0) \exp(n_0 (y_0 w - b(w)))
\]
with \( n_0 = s \), \( y_0 = t \), \( w = \ln(x/(1-x)) \) and \( b(w) = -\ln(1-x) = \ln(1+\exp(w)) \). By considering as likelihood a binomial density,
\[
\text{Bi}(y|x) = \prod_{i=1}^{n} k(y_i) \exp(n(\hat{y}_i w - b(w)))
\]
then \( \text{Beta}(x|s,t) \) and \( \text{Bi}(y|x) \) are conjugate and, thus, the posterior density is
\[
\text{Beta} \left( x | s + n, \frac{st + \sum_{i=1}^{n} y_i}{s + n} \right) = k(n',y') \exp(n'(y' w - b(w)))
\]
where \( n' = n + s \) and \( y' = \frac{st + \sum_{i=1}^{n} y_i}{s + n} \). In this case, the posterior mean and variance are
\[
\hat{\mu} = \frac{st + \sum_{i=1}^{n} y_i}{s + n}, \quad \hat{\sigma}^2 = \frac{\hat{\mu}(1-\hat{\mu})}{s + n + 1}.
\]
Since \( \frac{1}{\sigma_w^2} b(w) = w = x \) and since \( p(w|n_0,y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \), then from Lemma 3 it follows again that \( \hat{\mu} \) is also the posterior mean of \( x \). It is straightforward to verify that \( \hat{\sigma}^2 \) is also the variance of \( x \).
4 Family of priors in the exponential family

Consider the problem of statistical inference about the real-valued parameter \( w \) from noisy measurements \( \{y_1, \ldots, y_n\} \) and assume that the likelihood is completely described by the following probability density function (PDF) belonging to the exponential family:

\[
\prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i) \exp(n(\hat{\gamma}_n w - b(w))),
\]

where the parameters of the likelihood, i.e., sample mean \( \hat{\gamma}_n = \sum_{i=1}^{n} y_i \) and \( n \in \mathbb{R}^+ \), are known (the likelihood can be modelled by a single PDF). By conjugacy and following a Bayesian approach, as prior for \( w \) we may consider the PDF \( p(w|n_0, y_0) \) defined in (8) for a given value of the parameter \( y_0 \) and \( n_0 \).

In the case there is not enough information about \( w \) to uniquely determine the values of the parameters \( y_0 \) and \( n_0 \), we can consider the family of priors \( p(w|n_0, y_0) \) obtained by letting \( y_0 \) vary in \( \mathcal{Y}' \subseteq \text{Int}(\mathcal{Y}) \) and \( n_0 \) in some set \( \mathcal{M}_0 \subseteq \mathbb{R}^+ \), which could depend on \( y_0 \). The question to be addressed is whether such family of priors satisfies the properties (A.1)–(A.4) discussed in Section 2. The answer to this question is given in the next theorem.

**Theorem 1.** Consider as set of priors \( \mathcal{M} \) the family of conjugate priors \( p(w|n_0, y_0) \) with:

- \( y_0 \) spanning the set \( \mathcal{Y}' \subseteq \text{Int}(\mathcal{Y}) \),
- \( n_0 \) spanning the set \( \mathcal{M}_0 \subseteq \mathbb{R}^+ \), with \( \mathcal{M}_0 \) possibly dependent on \( y_0 \),

where \( \mathcal{Y} \) is assumed to be convex and \( \mathcal{W} = \mathbb{R} \). If and only if the following conditions hold:

(a) For each \( y_0 \in \mathcal{Y}' \) and \( n_0 \in \mathcal{M}_0 \), it holds that \( p(w|n_0, y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \);

(b) \( \mathcal{Y}' = \text{Int}(\mathcal{Y}) \);

(c) \( \mathcal{M}_0 \) satisfies the following constraints: \( 0 < \inf \mathcal{M}_0 \subseteq \mathcal{W} \), \( \sup \mathcal{M}_0 \leq \min(\bar{\mathcal{M}}, \frac{\bar{\mathcal{M}}}{\gamma_0}) \)

then, given the parameters \( \bar{n}_0 \) and \( c \), \( \mathcal{M} \) is the largest set which satisfies properties (A.1)–(A.4), with \( \mathcal{Y}_0 = \{ \frac{\partial h(w)}{\partial w} \} \) and \( \mathcal{W} \) including all the gambles integrable with respect to the exponential family density functions with support in \( \mathcal{W} \).

**Proof:** The proof is organized as follows. First we prove the necessity of the conditions (a)–(c) for (A2)–(A4). Second we prove their sufficiency. Then we show that \( \mathcal{M} \) is the largest set which satisfies these properties. Finally, we prove (A.1).

**Property (A.2):** prior ignorance, in case \( g(w) = \frac{\partial h(w)}{\partial w} \). Consider the prior \( p(w|n_0, y_0) \) and the function \( \frac{\partial h(w)}{\partial w} \). Since by hypothesis \( p(w|n_0, y_0) \to 0 \) for \( w \to \sup \mathcal{W} \) and \( w \to \inf \mathcal{W} \), from Lemma 2 it follows that \( E_w \left[ \frac{\partial h(w)}{\partial w} | n_0, y_0 \right] = y_0 \). Since the domain
of \( \frac{\partial b(w)}{\partial w} \) is \( \text{Int}(\mathcal{Y}) \), because of (5), a necessary condition for (A.2) to be satisfied is \( \mathcal{Y}' = \text{Int}(\mathcal{Y}) \).

In fact, in this case, since
\[
E_w \left[ \frac{\partial b(w)}{\partial w} | n_0, y_0 \right] = y_0,
\]
it follows that, for \( g(w) = \frac{\partial b(w)}{\partial w} \),
\[
E[g] = \inf \mathcal{Y}' = \inf g \quad \text{and} \quad \overline{E}[g] = \sup \mathcal{Y}' = \sup g.
\]

This proves that \( \mathcal{Y}' = \text{Int}(\mathcal{Y}) \) is a necessary condition for (A.2) to hold in the case \( g(w) = \frac{\partial b(w)}{\partial w} \).

Property (A.3): learning. To prove this property, we exploit the fact that the posterior density in (9) belongs to the exponential family and, thus, is fully described by the parameters \( n' = n + n_0 \) and \( y' = \frac{n_0 y_0 + n y}{n + n_0} \). For the proof, we distinguish three cases
\[
\mathcal{Y} = \mathbb{R}, \quad \mathcal{Y} = [a, \infty) \quad \text{(or} \quad \mathcal{Y} = (-\infty, a) \text{)} \quad \text{with} \quad a \in \mathbb{R}, \quad \text{and} \quad \mathcal{Y} \subset \mathbb{R} \text{ finite}. \]

In the last two cases w.l.o.g. it can be assumed that \( \mathcal{Y} = [0, \infty) \) (or \( \mathcal{Y} = (-\infty, 0) \)) and, respectively, \( \mathcal{Y} = [0, 1] \) (by shifting and scaling \( \mathcal{Y} \)); since \( \mathcal{Y} \) has been assumed to be convex, these three cases account for all the possibilities.

Consider the posterior density in (9), i.e.,
\[
p \left( w | n', y' \right) = k(n', y') \exp \left( w \left( y' w - b(w) \right) \right), \quad w \in \mathcal{Y}, \quad (15)
\]
where \( n' = n + n_0 \) and \( y' = \frac{n_0 y_0 + n y}{n + n_0} \). Then, a necessary condition for (A.3) to hold is that \( n_0 \leq \pi_0 < \infty \). In fact, assume that this does not hold and consider the gamble \( g = \frac{\partial b(w)}{\partial w} \) in \( \mathcal{Y}_0 \). Since
\[
E_w \left[ \frac{\partial b(w)}{\partial w} | n_0, y_0 \right] = y_0,
\]
it results that for \( n_0 \to \infty \), then \( y' \to y_0 \), which means no learning for any value of \( y_0 \) and \( \hat{y}_n \). In the case \( n_0 \leq \pi_0 < \infty \), condition (A.3) can still be violated if \( \mathcal{Y} = [0, \infty) \) (or \( \mathcal{Y} = (-\infty, 0) \)) or \( \mathcal{Y} = (-\infty, +\infty) \). For \( \mathcal{Y} = [0, +\infty) \), assume that \( \hat{y}_n = 0 \), then
\[
0 \leq y' = \frac{n_0 y_0 + n \hat{y}_n}{n + n_0} < \infty,
\]
where the left bound has been obtained for \( y_0 = 0 \) and the right bound for \( n_0 y_0 \to \infty \).

In the case \( \mathcal{Y} = (-\infty, +\infty) \), then, for any \( \hat{y}_n \),
\[
-\infty \leq y' = \frac{n_0 y_0 + n \hat{y}_n}{n + n_0} < \infty,
\]
where the right and left bounds have been obtained for \( n_0 y_0 \to \pm\infty \). Therefore, \( n_0 y_0 \leq c \) for some \( 0 < c < \infty \) is also a necessary condition for learning. Since there cannot be convergence without learning, the conditions \( n_0 \leq \pi_0 < \infty \) and \( n_0 |y_0| \leq c \) are also necessary for (A.4).

Consider now the sufficiency. For (A.2), the condition \( \mathcal{Y}' = \text{Int}(\mathcal{Y}) \) is clearly also sufficient. Consider (A.3). Since for \( n_0 \leq \pi_0 < \infty \), \( n_0 |y_0| \leq c \) and for \( n \to \infty \) (i.e., the number of observations goes to infinity), it results that \( n' \approx n \) and \( y' \approx \hat{y}_n \). Then, since it has been assumed a priori that \( \overline{E}[g] - \underline{E}[g] > 0 \), both conditions in (1) cannot be false at the same time. Assume, by contradiction, that there exists a gamble \( g \in \mathcal{Y} \) w.r.t. which the lower and upper prior and posterior expectations are, respectively, always equal for each \( n \). Since for \( n \to \infty \), it results that \( y' \to \hat{y}_n \) and \( n' \to n \), then the lower and upper posterior expectations converge, i.e., \( \underline{E}[g | y_1, \ldots, y_n] \to \underline{E}[g | \hat{y}_n, \ldots, \hat{y}_n] \). This implies that \( \underline{E}[g] = \overline{E}[g] \), which is a contradiction because by

\footnote{Actually it is enough that \( \mathcal{Y}' \) include neighbourhoods of the extremes of the set \( \mathcal{Y} \).}

\footnote{The limit \( n_0 y_0 \to \infty \) means that the increasing of \( y_0 \) is faster than the decreasing of \( n_0 \).}
hypothesis we have assumed that $\mathbb{E}[g] - \mathbb{E}[\hat{g}] > 0$. A consequence is that there exists a $\delta > 0$ such that for all $n > \delta$ either $\mathbb{E}[g|y_1, \ldots, y_n] \neq \mathbb{E}[\hat{g}]$ or $\mathbb{E}[g|y_1, \ldots, y_n] \neq \mathbb{E}[\hat{g}]$.

To prove the second part of (A.3), consider the gamble $g \in \mathcal{G}_0 = \{ \frac{\partial h(w)}{\partial w} \}$. In this case the left and right bound of $E[g|y_1, \ldots, y_n]$ are:

$$\frac{n\hat{y}_n}{n + \bar{n}_0} \leq y' = \frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0} \leq \frac{\bar{n}_0 + n\hat{y}_n}{n + \bar{n}_0}, \quad (18)$$

for $\mathcal{Y} = [0, 1]$:

$$\frac{n\hat{y}_n}{n + \bar{n}_0} \leq y' = \frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0} \leq \frac{c + n\hat{y}_n}{n}, \quad (19)$$

for $\mathcal{Y} = [0, +\infty)$ (or $\mathcal{Y} = (-\infty, 0]$);\(^6\)

$$\min\left(\frac{-c + n\hat{y}_n}{n + \bar{n}_0}, \frac{c + n\hat{y}_n}{n}\right) \leq y' = \frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0} \leq \max\left(\frac{c + n\hat{y}_n}{n + \bar{n}_0}, \frac{c + n\hat{y}_n}{n}\right), \quad (20)$$

for $\mathcal{Y} = (-\infty, +\infty)$; then, in all three cases, for $n > 0$ and independently of the value of $\hat{y}_n$, at least one between the lower and upper bound differs from its a priori value $\mathbb{E}[g]$ or, respectively, $\mathbb{E}[\hat{g}]$. This is obvious for the right bound of (19) and the left and right bounds of (20), since a priori $\mathbb{E}[g] = -\infty$ and $\mathbb{E}[\hat{g}] = \infty$. For the left and right bounds of (18), it follows that if $\hat{y}_n = 0$ then $\mathbb{E}[g|y_1, \ldots, y_n] = \mathbb{E}[g] = 0$ but $\mathbb{E}[\hat{g}|y_1, \ldots, y_n] \neq \mathbb{E}[\hat{g}] = 1$ and, vice versa, if $\hat{y}_n = 1$ then $\mathbb{E}[g|y_1, \ldots, y_n] = \mathbb{E}[g] = 1$ but $\mathbb{E}[\hat{g}|y_1, \ldots, y_n] \neq \mathbb{E}[\hat{g}] = 0$. Thus, the second part of (1) holds for any $\delta > 0$. This proves that $n_0 \leq \bar{n}_0 < \infty$ and $n_0|y_0| \leq c$ are necessary and sufficient for (A.3) to be satisfied.

Since for $n_0 \leq \bar{n}_0 < \infty$, $n_0|y_0| \leq c$ and $n \to \infty$ it results that $y' = \frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0}$ and $n' = n + n_0$ do not depend on $n_0$ and $y_0$ but only on $n$ and $\hat{y}_n$, the sufficiency for (A.4) follows also straightforwardly. In fact it can be verified that, for $n \to \infty$, it results that $\mathbb{E}[g|y_1, \ldots, y_n], \mathbb{E}[\hat{g}|y_1, \ldots, y_n] \to E^*|g|y_1, \ldots, y_n]$, where $E^*|g|y_1, \ldots, y_n]$ is the posterior expectation obtained from the improper prior density $p(w) = 1$.

To sum up, necessary and sufficient conditions for (A2)–(A4) are $\mathcal{Y}' = Int(\mathcal{Y})$ and $0 < n_0 < min(\bar{n}_0, c|y_0|)$ and the corresponding set of priors $\mathcal{M}$ is also the largest set which satisfies (A2)–(A4). This proves the theorem for (A2)–(A4). Consider now property (A.1), coherence. Notice that, for each fixed value of the parameters $0 < n_0 < min(\bar{n}_0, c|y_0|)$ and $y_0 \in Int(\mathcal{Y})$, the set of priors $\mathcal{M}$ includes only proper densities [7]. Thus, since the set of posterioris is obtained by applying Bayes' rule to each pair likelihood-prior in $\mathcal{M}$, the strong coherence of priors, likelihood and posteriors follows by the application of the lower envelope theorem [2, Theorem 7.8.1].\(^7\)

\(^6\) Since $\hat{y}_n \geq 0$, it results that $\frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0} \geq \frac{n\hat{y}_0}{n + n_0}$ and, thus, $\frac{n_0\hat{y}_0 + n\hat{y}_n}{n + n_0}$ is a right bound for $y'$. In the case, $\mathcal{Y} = (-\infty, 0]$, since $\hat{y}_n \leq 0$, it results that $\frac{n\hat{y}_0}{n + n_0} \leq \frac{n\hat{y}_0}{n + n_0}$ and, thus, $\frac{n\hat{y}_0}{n + n_0}$ is a left bound for $y'$.

\(^7\) Walley's theory is defined only for bounded gambles. A question to be addressed in future is whether the strong coherence of the model in Theorem 1 extends to the unbounded case.
Some remarks on Theorem 1.

1. In order to better understand the intuition behind the theorem, consider the Gaussian case, i.e., \( \mathcal{Y} = (-\infty, +\infty) \), \( y_0 \) the prior mean and \( \sigma_0^2 = 1/n_0 \) the prior variance. In this case, the set of priors \( \mathcal{M} \) is equal to:

\[
\left\{ \mathcal{N} \left( w; y_0, \sigma_0^2 \right) : y_0 \in (-\infty, +\infty), \ max(1/\pi_0, |y_0|/c) < \sigma_0^2 < \infty \right\}.
\]

(21)

Hence, \( \mathcal{M} \) includes all the Gaussian densities with mean free to vary in \( \mathbb{R} \) and variance lower bounded by \( 1/\pi_0 \) but linearly increasing with \( |y_0| \). Notice, in fact, that if \( |y_0| > c/\pi_0 \), then \( \sigma_0^2 \geq |y_0|/c \). Hence, considering the likelihood \( \mathcal{N}(y_i; w, \sigma^2) \) for \( i = 1, \ldots, n \), the corresponding set of posteriors is equal to:

\[
\left\{ \mathcal{N} \left( w; y_p, \sigma_p^2 \right) : y_p = \sigma_p^2 \left( \frac{y_0}{\sigma_0^2} + \frac{n y_n}{\sigma^2} \right), \ \sigma_p^2 = \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}, \right.
\]
\[
y_p \in (-\infty, +\infty), \ max(1/\pi_0, |y_0|/c) < \sigma_p^2 < \infty \right\}.
\]

(22)

For \( 1/n_0 = \sigma_0^2 \) fixed and \( |y_0| \to \infty, |y_p| = |n_0 y_0 + n y_n|/(n + n_0) = |y_0| \to \infty \). Thus, the intuition behind the dependence of \( \sigma_0^2 \) on \( |y_0| \) is to prevent that, for \( |y_0| \to \infty \), also \( |y_p| = |n_0 y_0 + n y_n|/(n + n_0) \to \infty \). That is, the contribution of \( y_0 \) to \( y_p \) must decrease as \( |y_0| \to \infty \), otherwise the observations do not contribute to \( y_p \) (learning cannot take place). This is essentially the meaning of the constraint \( max(1/\pi_0, |y_0|/c) < \sigma_0^2 \), i.e., the variance of the Gaussians in \( \mathcal{M} \) cannot be smaller than \( |y_0|/c \).

2. The family of priors \( \mathcal{M} \) defined in Theorem 1 is completely determined by the two parameters \( c > 0 \) and \( \pi_0 > 0 \) (actually just by \( \pi_0 > 0 \) in the case \( \mathcal{Y} = [0, 1] \)). The larger are these parameters the larger is the family of priors \( \mathcal{M} \) and, thus, the more conservative are the posterior inferences. The choice of these parameters will be discussed in Section 5.

3. In the case the observations are binary, i.e., \( \mathcal{Y} = [0, 1] \), the set of priors \( \mathcal{M} \) transformed back to the original parameter space \( \mathcal{X} \) reduces to the Imprecise Beta Model discussed in [2, Section 5.3.1], [8]:

\[
\mathcal{M} = \left\{ \text{Beta}(x; st, s(1-t)) : t \in [0, 1] \right\},
\]

(23)

where \( x \in (0, 1) \), \( n_0 = s = \pi_0 \) is a positive fixed value and \( \text{Beta}(x; \alpha, \beta) \) is the Beta density with parameters \( \alpha \) and \( \beta \). The Imprecise Beta Model (23) and his multidimensional extension [9] have been applied effectively in classification [10] and system reliability problems [11].

In the case the observations belong to the real line then, for \( \pi_0 \) suitably small, the set of priors \( \mathcal{M} \) reduces to the family of Gaussian priors with infinite variance.
discussed in [5, Section 3.3] and the bounds in (20) become approximately equal to:

\[-c + n\hat{y}_n \leq \frac{n_0y_0 + n\hat{y}_n}{n} \leq \frac{c + n\hat{y}_n}{n}\]

(24)

The main difference is that the family of priors defined in Theorem 1 has been proved to be strongly coherent, while no proof of coherence is given for the model in [5, Section 3.3]; the coherence of this model is still an open problem.

4.1 Additional properties for near-prior ignorance

It is worth comparing properties (A.1)–(A.4) with the properties for near-prior ignorance discussed in [5]. Hereafter, for the convenience of the reader, we summarize these properties.

(B) Near-prior Ignorance The following conditions on the upper and lower expectations generated by \( M \) are necessary for \( M \) to be sufficiently large.

(B1) If \( A = [a - \zeta, a + \zeta] \subseteq \mathbb{R} \) is a finite interval for some \( a \in \mathbb{R} \) and \( \zeta > 0 \), then \( \mathbb{E}(I_{[A]}) = 0 \) and \( \mathbb{E}(I_{[A]}) \to 1 \) as \( 2\zeta \to \infty \).

(B2) If \( A = [a, +\infty) \) or \( A = (-\infty, a] \) is a semi-infinite interval, then \( \mathbb{E}(I_{[A]}) = 0 \) and \( \mathbb{E}(I_{[A]}) = 1 \).

(B3) The upper and lower mean under \( M \) are \( \mathbb{E}(w) = -\infty \) and \( \mathbb{E}(w) = +\infty \).

(C) Translation invariance When there is no prior information on the value of \( w \), the set \( M \) should be invariant under translations of the scale on which measurements and \( w \) are made. Equivalently, upper and lower expectations generated by \( M \) should be translation-invariant.

(D) Dependence on sample size Upper and lower probabilities of the standard \( \gamma \)-intervals should converge to the nominal probability \( \gamma \) as \( n \to \infty \).

8 Other properties for \( M \) defined in [5] are: easy elicitation, tractability, variety of shapes and weak coherence of posteriors and likelihood.

Concerning the coherence property (A.1) defined in Section 4, this is more restrictive than the property of weak coherence defined in [5]. (A.1) implies weak coherence, while the converse is not true. We point the reader to [2, Ch. 7] for more details. Concerning properties (B1)–(B3), they can be obtained from (A.2) for different choices of \( g \) (the meaning of the limit in (B1) will be discussed later). Hence, (A.2) includes (B1)–(B3) as special cases. However, notice that, the functions \( g \) that we use to model our state of near-prior ignorance about \( w \) are usually those listed in (B1)–(B3), i.e., \( g(w) = w \) and \( g(w) = I_{[A]}(w) \), where \( A \) can be a finite interval or a semi-infinite interval (for \( A = (-\infty, a] \), \( E[I_{[A]}] \) coincides the cumulative distribution of \( w \). Thus, (B2) and (B3) are exactly the property (A.2) defined in Section 2 in the case \( g(w) = w \) and \( g(w) = I_{[A]}(w) \). Concerning property (B1), if we remove the limit condition in the upper, it would again coincide with (A.2) by setting \( g(w) = I_{[A]}(w) \),

8 For Gaussian densities, the standard \( \gamma \)-intervals are \([\mu - z_\gamma \sigma^2, \mu + z_\gamma \sigma^2]\), where \( \mu, \sigma^2 \) are the posterior mean and variance and \( z_\gamma \in \mathbb{R}^+ \) depends on \( \gamma \). For sets of densities, we can define the \( \gamma \)-interval as the smaller interval whose probability of enclosing \( w \) is at least \( \gamma \).
i.e., $E(I(A)) = \inf I(A)(w) = 0$ and $E(I(A)) = \sup I(A)(w) = 1$ for any $\zeta$. However, if (A.2) holds for $g(w) = I(A)(w)$ with $A$ arbitrary, then (A.3) cannot be satisfied. In fact, the requirement $E(I(A)) = 1$ implies that $\mathcal{M}$ must include a proper density with support in $A$. However, since $A$ can be arbitrarily small, this density can be arbitrarily close to a Dirac delta centered in $A$. Thus, $\mathcal{M}$ must include densities approaching all possible Dirac’s delta in the real line. This clearly implies that, no matter the observations, the posterior upper and lower expectations of $g(w) = I(A)(w)$ would coincide with the prior upper and lower expectations and, thus, (A.3) cannot be satisfied. (B1) is thus a way to relax the requirement of prior ignorance w.r.t. $g(w) = I(A)(w)$ with $A = [a_0 - \zeta, a + \zeta]$, which is also compatible with (A.3).

**Corollary 1.** Under the hypotheses of Theorem 1, the family of priors $\mathcal{M}$ satisfies (B1) and (B2).

**Proof:** Concerning (B1), consider the case in which $A = [w_0 - \zeta, w_0 + \zeta]$, for some $w_0 \in \mathcal{W}$ and $\zeta < \infty$. The expected value w.r.t. $w$ of the indicator $I_{[w_0 - \zeta, w_0 + \zeta]}(w)$ is given by

$$P(A) = \int_{w \in \mathcal{W}} I_{[w_0 - \zeta, w_0 + \zeta]}(w)p(w|n_0, y_0)dw = \int_{w_0 - \zeta}^{w_0 + \zeta} k(n_0, y_0)\exp(n_0(y_0w - b(w)))dw.$$ 

It is clear that for $\zeta \to \infty$, $\mathcal{E}_w[I_{[w_0 - \zeta, w_0 + \zeta]}] \to 1$, being $p(w|n_0, y_0)$, for some $0 < n_0 \leq n_\theta$ and $y_0 \in \text{Int}(\mathcal{Y})$, a proper density. In order to prove that $E_w[I_{[w_0 - \zeta, w_0 + \zeta]}] = 0$ for any $\zeta < \infty$, consider the first and second derivatives of $p(w|n_0, y_0)$, i.e.,

$$\frac{\partial p(w|n_0, y_0)}{\partial w} = n_0 \left( y_0 - \frac{\partial b(w)}{\partial w} \right) p(w|n_0, y_0),$$

$$\frac{\partial^2 p(w|n_0, y_0)}{\partial w^2} = -\frac{\partial^2 b(w)}{\partial w^2} n_0 p(w|n_0, y_0) + n_0^2 \left( y_0 - \frac{\partial b(w)}{\partial w} \right)^2 p(w|n_0, y_0).$$

(25)

Since for $\frac{\partial b(w)}{\partial w} = y_0$ the first derivative is zero and the second derivative is negative (being $p(w|n_0, y_0) > 0$ and $\frac{\partial^2 b(w)}{\partial w^2} > 0$), the value $w^*$ such that $\frac{\partial b(w)}{\partial w} = y_0$ is a maximum of $p(w|n_0, y_0)$. Since $\frac{\partial b(w)}{\partial w}, y_0 \in \text{Int}(\mathcal{Y})$ and since $\frac{\partial^2 b(w)}{\partial w^2} > 0$, the function $\frac{\partial b(w)}{\partial w}$ is always increasing in $\mathbb{R}$ and obtains its infimum $\inf \mathcal{Y}$ and supremum $\sup \mathcal{Y}$ for $w \to \pm \infty$. Thus, by moving $y_0$ between $\inf \mathcal{Y}$ and $\sup \mathcal{Y}$, $w^*$ can be shifted arbitrarily in $\mathbb{R}$. Then assume, by contradiction, that

$$\int_{w_0 - \zeta}^{w_0 + \zeta} k(n_0, y_0)\exp(n_0(y_0w - b(w)))dw \geq \delta$$

(26)
for all \( w_0 \), where \( \delta > 0 \) can be fixed arbitrarily small. Then, in the case \( w^* = w_0 + 3\zeta \), it follows that

\[
\int_{-\infty}^{\infty} k(n_0, y_0) \exp(n_0(y_0w - b(w))) dw \geq 2\delta,
\]

since \( w^* = w_0 + 3\zeta \) is a maximum of \( p(w|n_0, y_0) \), \( p(w|n_0, y_0) \) is increasing in \( (-\infty, w^*) \) and, thus,

\[
\int_{w_0 - \zeta}^{w_0 + \zeta} k(n_0, y_0) \exp(n_0(y_0w - b(w))) dw \leq \int_{w_0 - \zeta}^{w_0 + 3\zeta} k(n_0, y_0) \exp(n_0(y_0w - b(w))) dw.
\]

Hence, in general if \( w^* = w_0 + (2r + 1)\zeta \) with \( r = 1, 2, 3, \ldots \), it follows that

\[
\int_{-\infty}^{\infty} k(n_0, y_0) \exp(n_0(y_0w - b(w))) dw \geq (r + 1)\delta.
\]

Thus, being \( \delta > 0 \), if (26) holds and since \( w^* \) is free to vary in \( \mathbb{R} \), we can always find a value of \( w^* \) such that \( \int_{-\infty}^{\infty} p(w|n_0, y_0) > 1 \), which contradicts \( p(w|n_0, y_0) \) being a PDF. Therefore, for \( y_0 \to \sup \mathcal{Y} \), it must hold that \( \delta \to 0 \) and thus \( p(w|n_0, y_0) \to 0 \) in \([w_0 - \zeta, w_0 + \zeta] \). A similar conclusion can be derived by moving \( y_0 \) towards \( \inf \mathcal{Y} \).

This implies that \( P(A) = 0 \) for \( y_0 \to \sup \mathcal{Y} \) (or \( y_0 \to \inf \mathcal{Y} \)), which proves (B1) w.r.t. \( w \). For the lower the same result can be proven for any semi-infinite interval such as \( A = [w_0, +\infty) \) or \( A = (-\infty, w_0] \). Furthermore, since \( E[I_{\{A\}}] = 0 \) implies \( E[I_{\{\mathbb{R} \setminus A\}}] = 1 \) \cite{12}, (B2) follows straightforwardly.

To be invariant to re-parametrizations \cite[Sec. 5]{1} of the parameter space is a desirable property for a model of prior ignorance. Since, in this paper, we consider the one-parameter exponential family, we have formulated the near-prior ignorance property w.r.t. the natural parameter \( w \). This means that our near-prior ignorance property holds only for a specific parametrization of the parameter space, i.e., the one which transforms the original parameter \( x \in \mathcal{X} \subseteq \mathbb{R} \) into the natural parameter \( w \in \mathcal{W} = \mathbb{R} \).

The advantage of this approach is that likelihood and prior, under this parametrization, belong to the natural exponential family and this considerably simplifies the computations of the posterior inferences (e.g., the expectations of some functions of interest can be computed in closed form). Furthermore, since it has been assumed that \( \mathcal{W} = \mathbb{R} \), we are just considering parametrizations that transform \( x \) in a location parameter (in the Gamma density case, \( x \) is a scale parameter \( x > 0 \)) while in the Beta density case \( x \) is an unknown chance \( x \in [0, 1] \), but for both cases \( \mathcal{W} = \mathbb{R} \). Hence, since \( w \) is defined in \( \mathbb{R} \), the invariance property that we would like to respect is translation invariance. This means that for all \( g : \mathcal{W} \to \mathbb{R} \), \( w_0 \in \mathcal{W} \) and \( g'(w) = g(w - w_0) \), the family \( \mathcal{W} \) is translation invariant if \( E[g] = E[g'] \). Since this property must hold for each function \( g \), it is satisfied if for each \( w_0 \), \( n_0 \) and \( y_0 \) there exist \( n'_0, y'_0 \) such that:

\[
\exp(n_0(y_0(w - w_0) - b(w - w_0))) \propto \exp(n'_0(y'_0w - b(w))).
\]  

(27)
The validity of the above equality depends on the functional form of $b(w)$ and, thus, on the particular member of the exponential family under consideration. In particular, (27) can be satisfied if the member of the exponential family has a location parameter [12]. From the results in [12] for the one-parameter exponential family, it follows that the densities which have a location parameter are those that can be written as Gamma densities (also the Gaussian density is included in this family). For this family, translation invariance can be satisfied only if the parameters of the densities $y_0$ and $n_0$ are free to vary independently. This is in fact the only way to satisfy (27) for each pair $n_0$ and $y_0$. In the case considered in the present paper, since $y_0$ and $n_0$ are not free to vary independently (because of the constraint $n_0 \leq \min(n_0, |y_0|)$) translation invariance cannot be guaranteed in general.

However, it can be noticed that, at the decreasing of $\pi_0$, the left and right side member in (27) become closer and closer and, thus, $M$ gets closer and closer to a translation invariant model (since the densities in $M$ get closer to uniform densities). As observed in Section 4, in the case $\mathcal{Y} = (-\infty, +\infty)$ and for $\pi_0 \to 0$, the set of priors $M$ reduces to the family of Gaussian priors with infinite variance discussed in [5, Section 3.3], which is indeed translation invariant. Also in the case $\mathcal{Y} = [0, +\infty)$, the set of priors $M$ reduces to a translation invariant model for $\pi_0 \to 0$. The same happens for the case $\mathcal{Y} = [0, 1]$ even if in this case, for $\pi_0 \to 0$, the set of priors $M$ collapses to a single improper prior (since $n_0 \to 0$ is not contrasted by $|y_0| \to \infty$).

As discussed in [5, Section 3.3], an important consequence of translation invariance is that the imprecision (that is, the size) of the posterior set of densities does not depend on the sample mean $\hat{y}_n$; the only effect of shifting $\hat{y}_n$ is to shift the posterior set by the same amount. Translation invariance is a sufficient condition for the independence of the sample mean $\hat{y}_n$ of the imprecision of the posterior set of densities, but it is not necessary. In fact, there are families $M$ that are not translation invariant but for which this independence holds as, for instance, the family $M$ in Theorem 1 in the case $\mathcal{Y} = [0, 1]$ and $\pi_0 > 0$. This independence is used in Section 5 to choose the value of the parameter $c$ of the set of priors $M$ based on the desirable imprecision of the posterior set of densities.

Concerning property (D), this is a condition for learning, convergence of the lower towards the upper probability and convergence to the nominal probability $\gamma$. It states that, at increasing $n$, the effects of the lack of prior information should be less and less important (learning) and that lower and upper probabilities should converge to the nominal probability $\gamma$ (convergence). Property (D) is thus similar to the properties (A.3)–(A.4) defined in Section 2. Notice that there cannot be convergence without learning. Thus, (D) requires (A.3). The convergence defined in (A.4) is more general than that in (D). In [5], it is in fact assumed that there is a single likelihood distribution which describes the observation model and, thus, at the increasing of the number of observations we expect that lower and upper posterior expectations get closer and closer. The convergence to the nominal probability $\gamma$ is then a consequence of the central limit theorem for distributions. Conversely, property (A.4) only states that the influence of the prior set of distributions should vanish with increasing number of observations. However, in the case of the exponential family discussed in Section 4,
it can easily be proved that not only (A.4) but also (D) holds.

**Corollary 2.** Under the hypotheses of Theorem 1, the family of priors \(\mathcal{M}\) satisfies (D).

**Proof:** Consider the quantity

\[ y' = \frac{n_0 y_0 + n \hat{y}_n}{n + n_0}. \]

Assuming the constraint \( n_0 < \min(\bar{n}_0, \frac{1}{\log}) \), then \( y' \to \hat{y}_n \) and \( n' = n + n_0 \to n \) as \( n \to \infty \). Thus, the family of posterior densities converges to

\[ p(w | n, \hat{y}_n) \approx k(n, \hat{y}_n) \exp(n(\hat{y}_n w - b(w))). \quad (28) \]

Hence, for any function \( g \), the convergence of the lower towards the upper expectation of \( g \) follows straightforward. The convergence of the upper and lower probabilities of the standard \( \gamma \)-intervals to the nominal probability \( \gamma \) is then a consequence of the central limit theorem.

It is also interesting to compare the set of priors \(\mathcal{M}\) in Theorem 1 with another model for ignorance, the Bounded Derivative Model (BDM) [13]. In the BDM, \(\mathcal{M}_{BDM}\) includes all continuous proper probability density functions for which the derivative of the log-density is bounded by a positive constant, i.e., \[ \left| \frac{\partial}{\partial w} \log p(w) \right| \leq c. \]

It can be verified that BDM satisfies all the properties (A1)–(A4), with \( \bar{n}_0 \) and \( \gamma \) defined as in Theorem 1. BDM is a non-parametric model and, in this sense, is more general than the model resulting from Theorem 1 that is restricted to the one-parameter exponential family only. A drawback of this generality is that inferences with BDM can in general be difficult to compute [13, Sec. 6], while this is not the case for the model resulting from Theorem 1 because of conjugacy.

Conversely, a model for statistical inferences based on set of densities belonging to the exponential family is presented in [14]. The main difference with respect to the present work is that the model in [14] is not a model of prior ignorance, as pointed out by the authors, i.e., the set \(\mathcal{V}'\) in Theorem 1 is chosen in [14] to reflect the prior information on \( y_0 \) and, thus, the posterior inferences depend on this information. Since no constraint between \( n_0 \) and \( y_0 \) is assumed, the model in [14] can also violate (A.3)–(A.4) in the case \(\mathcal{V}' = \text{Int}(\mathcal{V})\), and hence it can produce vacuous inferences.

## 5 How to choose \(\bar{n}_0\) and \(c\)

From Theorem 1, it follows that the family of priors \(\mathcal{M}\) is completely determined by the two parameters \(\bar{n}_0\) and \(c\). The aim of this section is to give guidelines for the choice of these parameters. First of all, notice that larger values of \(\bar{n}_0\) and \(c\) imply larger sets of priors \(\mathcal{M}\) and, thus, more robustness to the lack of prior information in
the posterior inferences. Thus, measures of robustness can be used to select $\pi_0$ and $c$ such as [2]: (a) the convergence rate of the lower and/or upper expectations to suitable limits; (b) the convergence rate of the posterior imprecision, i.e., the difference between upper and lower expectations. Here the expectations are computed w.r.t. some function of interest $g$ and the convergence is defined w.r.t. the number of samples $n$.

Another possible requirement for the choice of $\pi_0$ and $c$ is that the family of priors $\mathcal{M}$ should be large enough to encompass frequentist or objective Bayesian inferences, but not too large to avoid obtaining too weak inferences.

As in the previous sections we distinguish three cases $\mathcal{Y} = [0, 1]$: $\mathcal{Y} = (-\infty, \infty)$ and $\mathcal{Y} = (0, \infty)$ (or $\mathcal{Y} = (-\infty, 0)$). Consider the case $\mathcal{Y} = [0, 1]$. Since $\mathcal{Y}$ is finite, it follows that $c/|y_0|$ is bounded below by $c$. Then, by selecting $c > \pi_0$, the set of priors $\mathcal{M}$ depends only on $\pi_0$ (remember in fact that in this case $0 < n_0 \leq \pi_0 < \infty$ is a necessary and sufficient condition for learning and convergence). Thus, only the parameter $\pi_0$ must be specified. As already discussed in Section 4, in the case $\mathcal{Y} = [0, 1]$ the set of priors $\mathcal{M}$ coincides with the Imprecise Beta Model. The choice of the parameter $\pi_0$ for this model is widely discussed in [8],[9]. Several arguments support the choice of $\pi_0 = 2$. In fact, in this case, the Imprecise Beta Model encompasses the three Beta densities that are commonly used by Bayesians to model prior ignorance about an unknown chance: the uniform prior ($n_0 = 2$ and $y_0 = 1/2$), Jeffreys’ prior ($n_0 = 1$ and $y_0 = 1/2$) and Haldane’s prior ($n_0 = 0$). Furthermore, for $\pi_0 = 2$, it is also guaranteed that the 95% one- and two-sided credible intervals for $x$ are also (at least) 95% confidence intervals in the frequentist sense [9].

By considering $g(w) = \frac{\partial b(w)}{\partial w}$, we can also use the posterior imprecision, i.e., $\mathcal{E}[g] - \mathcal{E}[g]$, which is equal to $\pi_0/(n + \pi_0)$, to select $\pi_0$. Thus, by selecting for instance $\pi_0 = 1$, we can impose that the posterior imprecision reduces to $1/2$ its initial value after $n = 1$ observations (for $\pi_0 = 2$ the imprecision reduces to $2/3$ its initial value after $n = 1$ observations, etc.). Smaller values of $\pi_0$ produces faster convergence and stronger conclusions, whereas larger values produces more cautious inferences.

In the case $\mathcal{Y} = (-\infty, \infty)$, from (20), it follows that for $g(w) = \frac{\partial b(w)}{\partial w}$ the posterior imprecision is equal to:

$$
\begin{align*}
\begin{cases}
\frac{2c}{n} & \text{if } |n\hat{y}_n| < c, \\
\frac{2c}{n} + \frac{\pi_0}{n + \pi_0} \left(\hat{y}_n - \frac{c}{n}\right) & \text{if } n\hat{y}_n - c > 0, \\
\frac{2c}{n} + \frac{\pi_0}{n + \pi_0} \left(-\hat{y}_n - \frac{c}{n}\right) & \text{if } n\hat{y}_n + c < 0.
\end{cases}
\end{align*}
$$

(29)

Thus, apart from the first case, the imprecision does in general depend on the observations through $\hat{y}_n$. A way to overcome this problem is by selecting an arbitrarily small value for $\pi_0$. In fact, when $\pi_0$ approaches zero, the part of the imprecision that depends on $\hat{y}_n$ vanishes and, thus, the imprecision reduces to $2c/n$. Hence, as in the case $\mathcal{Y} = [0, 1]$, we can select a-priori the width of the posterior imprecision by fixing a suitable value of $c$. In the Gaussian case, $n$ depends also on $\sigma^2$ (see the Gaussian
case in Section 3) and, thus, the value of $c$ should be selected based on both the number of measurements and their variance. Notice that in the case $\mathcal{Y} = (-\infty, \infty)$, both the frequentist and noninformative Bayesian inferences are encompassed by $\mathcal{M}$ for each $c > 0$. In fact, the inferences obtained by Jeffreys’ (improper uniform) prior are included in the set of posteriors inferences obtained by $\mathcal{M}$ for each $c > 0$.

In general, it can be noticed that, by setting $\pi_0 > 0$, we include in $\mathcal{M}$ all the Gaussians with mean $\gamma_0 = 0$ and variance $\pi_0 = \sigma^2 < \infty$. The posterior corresponding to the prior with mean $\gamma_0 = 0$ and variance $n_0 = \pi_0 > 0$ is a Gaussian with mean $(n + \pi_0)^{-1}n\hat{y}_n$ and variance $(n + \pi_0)^{-1}$. The posterior mean corresponds to the so-called ridge-estimator or Tikhonov-regularization. This is the most commonly used method of regularization to overcome the problem of multicollinearity in linear regression. Since we are dealing with scalar variables, we do not face the problem of multicollinearity. However, it is worth pointing out that the ridge-estimator is included in the set of posterior estimates generated by $\mathcal{M}$. The choice of $\pi_0 > 0$ thus guarantees that $\mathcal{M}$ can encompass also some of the “regularized frequentist inferences” (this can be interesting for a future extension of the model to the multivariate case).

In the case $\mathcal{Y} = [0, \infty)$, from (19) it follows that the lower and upper bounds of the expectation of $g(w) = \frac{\partial b(w)}{\partial w}$ are

$$E[g] = \frac{n\hat{y}_n}{n + \pi_0}, \quad \mathcal{P}[g] = \frac{c + n\hat{y}_n}{n}. \quad (30)$$

Thus, the posterior imprecision is equal to

$$\frac{c}{n + \pi_0} \hat{y}_n. \quad (31)$$

In this case, the posterior imprecision depends on $\hat{y}_n$ apart from the case $\pi_0 \to 0$ where it goes to zero. However, for $\pi_0 \to 0$, the lower expectation in (30) reduces to $\hat{y}_n$ and, thus, the left bound is always equal to $\hat{y}_n$. Thus, in the case $\pi_0 \to 0$, the posterior imprecision is strongly asymmetric w.r.t. $\hat{y}_n$. However, it is interesting to observe that, for $\pi_0 \to 0$ and $c = 1$, the set of posteriors obtained by $\mathcal{M}$ includes the two posterior densities that are obtained by the two priors which are commonly used by Bayesians to model prior ignorance about a Poisson parameter: the positive uniform density ($n_0 = 0$ and $n_0\gamma_0 = 1$) and Jeffreys’ prior ($n_0 = 0$ and $n_0\gamma_0 = 1/2$). As in the case $\mathcal{Y} = (-\infty, \infty)$, by selecting $\pi_0 > 0$, $\mathcal{M}$ can encompass also some of the “regularized frequentist inferences”.

To sum up, in the case $\mathcal{Y} = (-\infty, \infty)$, for inferences in problems where there is almost no prior information about the parameters, we can consider the set of posteriors that one obtains for $\pi_0 > 0$ but suitably small. In fact, in this case, the set of priors $\mathcal{M}$

\[\pi_0 \to 0, \text{the priors in } \mathcal{M} \text{ becomes improper and, thus, we cannot use the lower envelope theorem to prove (A.1) as shown in the proof of Theorem 1. This means that inferences obtained for } \pi_0 \to 0 \text{ may be incoherent. A question to be addressed in future is whether the posterior model obtained for } \pi_0 \to 0 \text{ and likelihood model are coherent.} \]
is close to a translation invariant model (as discussed in Section 4.1), encompasses frequentist and objective Bayesian inferences and produces self-consistent inferences (i.e., it satisfies (A.1)). The same holds in the case \( \mathcal{Y} = [0, \infty) \) if \( \pi \) is selected to encompass the two noninformative priors which are commonly used by Bayesians to model prior ignorance, i.e., \( c > 1 \). However, in this case, also a non-negligible value of \( \pi_0 \) seems to be convenient. In fact, a value of \( \pi_0 \) that is not too small avoids that the lower of (30) reduces to \( \hat{y}_n \), guaranteeing thus more robust inferences w.r.t. the maximum likelihood estimator.

6 Application: inferences for a Poisson model

The Poisson distribution is used to count the number of occurrences of rare events which are occurring randomly through time (or space) at a constant rate. Suppose we have a random sample of \( n \) independent observations from a Poisson distribution with parameter \( \lambda \). The Poisson density is:

\[
p(y|x) = \prod_i x^{y_i} \exp(-x) / y_i! \propto x^{n \hat{y}_n} \exp(-nx),
\]

where \( y_i \) is the number of occurrences of an event in the \( i \)-th observation and \( n \hat{y}_n = \sum y_i \). It is well known that the conjugate prior of a Poisson density is the Gamma density \( p(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta \lambda) \) with \( \alpha, \beta \in \mathbb{R}^+ \), which belongs to the one-parameter exponential family. Hence, likelihood and prior can be rewritten in the canonical form

\[
p(y|w) \propto \exp(n (\hat{y}_n w - b(w))), \quad p(w|n_0, y_0) \propto \exp(n_0 (y_0 w - b(w))), \quad (32)
\]

where \( n_0 = \beta, y_0 = \alpha/\beta, w = \ln(x) \) and \( b(w) = \exp(w) \).

In the case no prior information is available on \( w \), the following two densities are used by Bayesians to model prior ignorance about the Poisson parameter: the positive uniform density \( (\beta = n_0 = 0 \text{ and } \alpha = n_0y_0 = 1) \) and Jeffreys’ prior \( (\beta = n_0 = 0 \text{ and } \alpha = n_0y_0 = 1/2) \). Although both densities are improper, the posteriors obtained from these two priors are proper.

In the following we compare the frequentist and Bayesian inferences (the latter obtained using the above noninformative priors) with the inferences produced by the model in Section 4, which considers the following set of priors:

\[
\mathcal{M} = \left\{ k(n_0, y_0) \exp(n_0 (y_0 w - b(w))) : y_0 \in (0, +\infty), \quad 0 < n_0 \leq \min(\pi_0, c/y_0) \right\}.
\]

Observe that the set \( \mathcal{M} \) includes only proper priors and is really a model of prior ignorance w.r.t. the functions that are commonly used for statistical inferences.

Mean: Prior lower and upper expectations of \( \lambda \) are zero and, respectively \( \infty \). This follows from Theorem 1, since \( E[\frac{\partial b(w)}{\partial w}] = y_0 \).
**Variance:** Prior lower and upper variance are defined as follows [2, Appendix G]:

\[
V(x) = \inf_{\mu \in \mathbb{R}^+} E \left[ \left( \frac{\partial b(w)}{\partial w} - \mu \right)^2 \right],
\]

\[
\nabla(x) = \inf_{\mu \in \mathbb{R}^+} E \left[ \left( \frac{\partial b(w)}{\partial w} - \mu \right)^2 \right].
\]

For a single Gamma density, it results that:

\[
\inf_{\mu \in \mathbb{R}^+} E \left[ \left( \frac{\partial b(w)}{\partial w} - \mu \right)^2 \right] = \inf_{\mu \in \mathbb{R}^+} E \left[ \left( \frac{\partial b(w)}{\partial w} - y_0 - \mu \right)^2 \right] = \inf_{\mu \in \mathbb{R}^+} \frac{y_0}{n_0} + (y_0 - \mu)^2 = \frac{y_0}{n_0},
\]

(36)

where \( V(x) = y_0/n_0 \) is the variance of a Gamma density. Therefore, \( \min_{\mu \in \mathbb{R}^+} E \left[ \left( \frac{\partial b(w)}{\partial w} - \mu \right)^2 \right] \) is just an alternative expression for the variance and Equations (34)–(35) generalize (36) to the case of a set of densities. In particular, for the set of densities in (33), from (34)–(36) it follows that:

\[
V(x) = \inf_{\mu \in \mathbb{R}^+} \inf_{y_0 \in (0, +\infty), \ 0 < n_0 \leq \min\{y_0, \psi'/|y_0|\}} \frac{y_0}{n_0} + (y_0 - \mu)^2 = 0,
\]

(37)

\[
\nabla(x) = \inf_{\mu \in \mathbb{R}^+} \sup_{y_0 \in (0, +\infty), \ 0 < n_0 \leq \min\{y_0, \psi'/|y_0|\}} \frac{y_0}{n_0} + (y_0 - \mu)^2 = \infty.
\]

(38)

The lower variance is obtained for \( \mu = y_0 \); the upper variance for \((y_0 - \mu)^2 \to \infty\). Notice that (34)–(35) is again a state of complete ignorance.\(^{10}\)

**One-sided hypothesis testing:** Prior lower and upper probabilities of the hypotheses

- \( H_0 \): the Poisson population parameter \( x \leq x_0 \);
- \( H_1 \): the Poisson population parameter \( x > x_0 \).

are \( P(H_i) = 0, \ \overline{P}(H_i) = 1, \ i = 0, 1 \). This follows from the property (B2) discussed in Section 4.1 and reflects our state of prior ignorance about the probability of \( H_i \).

**Credible interval:** A 100\((1 - \gamma)\)% credible interval is the smallest interval \( \mathcal{X} \) that has at least probability \((1 - \gamma)\) of including the true \( x \), i.e., \( E[I_{\{x \in \mathcal{X}\}}] > (1 - \gamma) \). For the model \( \mathcal{M} \), it can be verified that \( \mathcal{X} = (0, \infty) \) for any \( \gamma > 0 \), which is again a state of complete ignorance. This follows from the same arguments used to prove property (B1) discussed in Section 4.1. Credible intervals are also used for two-sided hypothesis testing:

- \( H_0 \): the Poisson population parameter \( x = x_0 \);
- \( H_1 \): the Poisson population parameter \( x \neq x_0 \).

\(^{10}\) The fact that \( \mathcal{M} \) is a model of prior ignorance for the variance is true in the specific case of Gamma-like densities, but it is not true in general, e.g., for Gaussian densities.
The hypothesis $H_0$ is rejected with probability $(1 - \gamma)$ if $x_0$ is not included in the $100(1 - \gamma)$% credible interval. Since $\mathcal{X} = (0, \infty)$, no hypothesis $x = x_0$ can be rejected a-priori for any $\gamma > 0$.

The set of posteriors resulting from (33) is:

$$\mathcal{M}_p = \left\{ k(n', y') \exp\left( n'\left(y'w - b(w)\right)\right) : y' = (n_0y_0 + n\hat{y}_n)/(n + n_0), n' = n + n_0, y_0 \in (0, +\infty), 0 < n_0 \leq \min(\bar{n}_0, c/|y_0|) \right\}. \quad (39)$$

This set of posteriors can be used to derive inferences as described in the following.

**Mean:** Posterior expectations of $x$ are included in the interval $[n\hat{y}_n/(n + \bar{n}_0), (n\hat{y}_n + c)/n]$. The lower expectation is less than the maximum likelihood estimator of $x$ for any $\bar{n}_0 > 0$, while the upper, for $c > 1$, is greater than the Bayesian posterior means obtained with the positive uniform and Jeffreys’ improper priors.

**Variance:** Posterior lower and upper variances are:

$$\mathbb{V}(x) = \inf_{\mu \in \mathbb{R}^+} \inf_{y_0 \in (0, +\infty)} \frac{y'}{n'} + (y' - \mu)^2 = \frac{n\hat{y}_n}{(n + \bar{n}_0)^2}$$

$$\mathbb{V}(x) = \inf_{\mu \in \mathbb{R}^+} \sup_{y_0 \in (0, +\infty)} \frac{y'}{n'} + (y' - \mu)^2 = \frac{n\hat{y}_n + c}{n}, \quad (41)$$

which are, for $c > 1$, bounds for the Bayesian variances with noninformative priors.

**One-sided hypothesis testing:** Lower and upper posterior probabilities of $H_0$ are:

$$P(H_0|y) = \min_{\alpha \in [0, \bar{n}_0]} \int_{-\infty}^{\ln(y_0)} k \left( n + \alpha, \frac{c + n\hat{y}_n}{n + \alpha} \right) \exp \left( \frac{c + n\hat{y}_n}{n + \alpha} \left( y - b(w) \right) \right) dw$$

$$P(H_0|y) = \int_{-\infty}^{\ln(y_0)} k(n''', y''') \exp(n'''(y'''w - b(w))) dw,$$

where $y''' = \frac{n\hat{y}_n}{n + \bar{n}_0}, n''' = n + \bar{n}_0$. Figure 1 shows the value of $P(H_0|y)$ as a function of $n$ in the case $\hat{y}_n = 0.75, \bar{n}_0 \approx 0, x_0 = 1$ and for three different values of $c$, i.e., $\{0.5, 1, 2\}$. The lower probabilities $P(H_0|y)$ for $c = 0.5$ and $c = 1$ coincide with the Bayesian probabilities $P(H_0|y)$ obtained with Jeffreys and, respectively, uniform priors. It can be noticed that, based on Bayesian inferences, the hypothesis $H_1$ can be rejected after just 2 observations, being $P(H_0|y) > 0.5$ for $n \geq 2$ (for the uniform prior, while for Jeffreys’ prior this already happens at $n = 1$). Conversely, by using the method proposed in this paper, the hypothesis $H_1$ can be rejected only after $n \geq 7$ observations in the case $c = 2$, i.e., when $P(H_0|y) > 0.5$. Figure 2 shows the value of $P(H_0|y)$ as a function of $y$ in the case $n = 1, \bar{n}_0 \approx 0, x_0 = 1$ and for the previous three values of $c$. It can be noticed again that, based on Bayesian inferences, the hypothesis $H_0$ is rejected after $n = 1$ observation if $y > 1$ (for Jeffreys’ prior, while for the uniform prior this already happens at $y = 0.5$). Conversely, by using the method proposed in this paper,
\[ H_0 \text{ is rejected for } y > 1.5, \text{ i.e., when } P(H_0|y) < 0.5. \]

**Credible interval:** The posterior 100(1 - \(\gamma\))% credible interval for \(x\) is

\[
\mathcal{X} = \bigcup \left\{ [w_1, w_2] : \int_{w_1}^{w_2} p(w|n', y') \, dw = 1 - \gamma, \right. \\
y' = (n_0y_0 + n\hat{y}_n) / (n + n_0), \, n' = n + n_0, \\
y_0 \in (0, +\infty), \, 0 < n_0 \leq \min(\bar{n}_0, c/|y_0|) \left\} \right.,
\]

where, for each \(n', y'\), only the intervals with minimum size are considered in (42).

In the case \(\bar{n}_0 \approx 0\), the left extreme of \(\mathcal{X}\) is approximatively equal to the value \(w_1\) obtained for \(y' = \hat{y}_n\), \(n' = n\) and the upper extreme to the value \(w_2\) obtained for \(y' = c\hat{y}_n + c\), \(n' = n\). Again for \(c > 1\), \(\mathcal{X}\) includes the 100(1 - \(\gamma\))% Bayesian credible interval computed with the noninformative priors. For two-sided hypothesis testing, \(H_0\) can be rejected with posterior probability 1 - \(\gamma\) if \(x_0 / \in \mathcal{X}\). Figure 3 compares the width of 95% credible interval \(\mathcal{X}\), for different values of \(n\), \(\gamma = 0.05\), \(\hat{y}_n = 1\), \(\bar{n}_0 \approx 0\), and \(c \in \{0.5, 1, 2\}\), with the width of the 95% Bayesian credible interval obtained with the improper uniform. It can be noticed that the credible interval \(\mathcal{X}\) includes the Bayesian one for any \(c > 1\) and, that the difference in width between the Bayesian and the proposed credible interval is particular large for small \(n\).

Summing up the previous results:

- \(\mathcal{M}\) is really a model of prior ignorance w.r.t. the functions (queries) that are commonly used for statistical inferences, since the lower and upper expectations of such functions are vacuous a-priori;
- because of conjugacy, the computation of the posterior set of densities and, thus, of the posterior inferences is straightforward (as in the Bayesian case);
- the inferences drawn with the model \(\mathcal{M}\) encompass, for any \(\bar{n}_0 > 0\) and for \(c > 1\), the frequentist inferences and “objective” Bayesian inferences with improper priors;
- inferences drawn with the model \(\mathcal{M}\) for \(c > 1\) are more robust than noninformative priors, when only few observations are available.

Thus, \(\mathcal{M}\) can be regarded as a model of prior ignorance that can be used alternatively to the noninformative priors.

**7 Conclusions**

In this paper, we have proposed a model of prior ignorance about a scalar variable based on a set of distributions \(\mathcal{M}\). In particular, we have defined some minimal properties that a set \(\mathcal{M}\) of distributions should satisfy to be a model a prior ignorance without producing vacuous inferences. When the likelihood model is in the one-parameter exponential family of distributions, we have shown that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors \(\mathcal{M}\) that is the largest one which is equivalent to imposing the above properties. The obtained set \(\mathcal{M}\) is a model of prior ignorance w.r.t. the functions
Fig. 1 Lower and upper probability of $H_0$ as a function of $n$ for different values of $c$.

Fig. 2 Lower and upper probability of $H_0$ as a function of $y$ for different values of $c$.

Fig. 3 Credible intervals as a function of $n$ for different values of $c$.

(queries) that are commonly used for statistical inferences; it is easy to elicit and, because of conjugacy, tractable; it encompasses frequentist and “objective” Bayesian inferences with improper priors. Future work will address the following issues: (1) application of the model to real data; (2) extension of the minimal properties to the multivariate case; (3) extension of the one-parameter exponential family set of priors to the $k$-parameter exponential family.
References