Electronic Journal of Statistics
Vol. 11 (2017) 1858–1889
ISSN: 1935-7524
DOI: 10.1214/17-EJS1278

Data-driven nonlinear expectations for statistical uncertainty in decisions*

Samuel N. Cohen

Mathematical Institute, Woodstock Rd, Oxford
e-mail: samuel.cohen@maths.ox.ac.uk

Abstract: In stochastic decision problems, one often wants to estimate the underlying probability measure statistically, and then to use this estimate as a basis for decisions. We shall consider how the uncertainty in this estimation can be explicitly and consistently incorporated in the valuation of decisions, using the theory of nonlinear expectations.

MSC 2010 subject classifications: Primary 62F86, 62F25; secondary 62A86, 91G70, 90B50.
Keywords and phrases: Statistical uncertainty, robustness, nonlinear expectation.

Received September 2016.

Example 1. Consider the following problem. Let $\{X_n\}_{n\in\mathbb{N}}$ be identical independent Bernoulli random variables, with unknown parameter $p = P(X_n = 1) = 1 - P(X_n = 0)$, i.e. independent tosses of the same, possibly unfair, coin. You observe $\{X_n\}_{n=1}^N$, and then need to draw a conclusion about the likely behaviour of an iid trial X.

In a classical frequentist framework, this is straightforward: the estimator of p (either from MLE or moment matching) is given by $\hat{p} = S_N/N$, where $S_N = \sum_{n=1}^N X_N$; this estimate has sampling variance $p(1-p)/N \approx \hat{p}(1-\hat{p})/N$.

Suppose we need to evaluate a wager on X. Given a loss function ϕ, we would then usually calculate the expected loss $E[\phi(X)]$, where the expectation is based on the estimated parameters. Without loss of generality, we can assume $\phi(0) = 0$, so the inferred expectation is simply given by

$$\hat{E}[\phi(X)] = \hat{p}\phi(1).$$

This leads to a surprising conclusion: the precision of the estimate of p has no impact on our assessment of the wager. To see this, consider a sample based on $N' \gg N$ observations, but with the same value of $\hat{p}$. Then the precision of the estimate (as indicated by the reciprocal of the sampling variance) is much higher, but the expected loss of the wager remains identical. Consequently, when considering this wager, this approach concludes that you are indifferent between the settings when p is known precisely or imprecisely. For example, suppose there were two coins, the first was thrown 3 times with 2 heads, the second 3000 times with 2000 heads. The estimated-expected-loss criterion then states that you are

*Research supported by the Oxford–Nie Financial Big Data Laboratory and the Oxford–Man Institute for Quantitative Finance.

indifferent in choosing which coin to bet on, which is contrary to experience. Note that this conclusion is not changed by the presence of the loss function ϕ.

Some may argue that this is a particular flaw in the frequentist point-estimate approach, as the error of the estimate of p is not part of the probabilistic framework we use when calculating the expectation. So, let's take a Bayesian approach and put a prior on p. The choice of prior used is immaterial, as the behaviour is determined by (writing $\mathcal{F}_N$ for the σ-algebra generated by our observations)

$$E[\phi(X)|\mathcal{F}_N] = E\big[E[\phi(X)|p, \mathcal{F}_N]\big|\mathcal{F}_N\big] = E[p\phi(1)|\mathcal{F}_N] = E[p|\mathcal{F}_N]\phi(1)$$

so only the posterior mean value of p has any impact, not its posterior variance (or any other measure of uncertainty). Even if we extend beyond taking an expected payoff, for example to considering a posterior mean-variance criterion, we would find that the posterior variance of $\phi(X)$ is

$$E\Big[\Big(\phi(X) - E[\phi(X)|\mathcal{F}_N]\Big)^2\Big|\mathcal{F}_N\Big] = E[p|\mathcal{F}_N](1 - E[p|\mathcal{F}_N])\phi(1)^2$$

which still only depends on the posterior mean of p. The same conclusion will be reached for any criterion which depends only on the posterior law of $\phi(X)$.

From this, we can conclude both the frequentist and Bayesian expected loss approaches fail to incorporate uncertainty in p in our decision making, in this simple setting[1].

The unusual behaviour of this type of example has been noticed before. For example, Keynes remarks (using the term 'evidential weight' to indicate a concept similar to the precision of probabilities):

> For in deciding on a course of action, it seems plausible to suppose that we ought to take account of the weight as well as the probability of different expectations. —J.M. Keynes, A Treatise on Probability[2], 1921 [13, p.76]

Knight [14] argues that ignoring this uncertainty is not descriptive of people's actions – we do, generally, have a strict preference for knowledge of the probabilities of outcomes (see also the more general criticism of Allais [1]). This leads him to distinguish between the concepts of 'risk', which is associated with the outcome of X given p, and 'uncertainty'[3], which is associated with our lack of knowledge of p.

[1] The mathematical reason for this is that a mixture of Bernoulli random variables is again a Bernoulli random variable. Therefore, at the level of the marginal distribution of X, every hierarchical model is equivalent to a non-hierarchical model, and a Bayesian approach adds little mathematically. The simplicity of this setting may seem contrived, but demonstrates that one cannot, in general, claim that a Bayesian posterior expected loss approach is sufficient to deal with all forms of uncertainty.

[2] This idea is discussed at length in Keynes' treatise, but is not pursued as a principle in statistics, as is shown by the next sentence: "But it is difficult to think of any clear example of this, and I do not feel sure that the theory of 'evidential weight' has much practical significance." In some sense, the aim of this paper is to address this lack of examples in a concrete mathematical fashion, and to propose practical solutions based on classical statistical methods.

[3] This is a significant simplification of Knight's argument, which also looks at the question of estimating probabilities of future events, which by their very nature, are not the same as events which have already occurred. Nevertheless, the terminology of 'Knightian uncertainty' has become common as referring to lack of knowledge of probabilities, so we retain this usage.

Within either of the two classical frameworks considered above, there is a natural and classical way to deal with this issue. For a frequentist, instead of using the point estimate $\hat{p}$, one could consider building a confidence interval for p, and then comparing wagers by their worst expectation among parameters within the confidence interval. As the sample size increases, the confidence interval typically shrinks, and so (for a fixed value of $\hat{p}$) the value of the wager increases. Similarly for a Bayesian, using a credible interval in the place of the confidence interval. While well known and sensible, this is (at least on the surface) an ad hoc fix. In more complex settings, where the parameter p is replaced by a multidimensional parameter and we are interested in comparing the values of a variety of random outcomes (whose expectations are generally nonlinear functions of the parameters), confidence sets become less natural, so a more general and rigorous approach seems to be needed.

1. Uncertain valuations

As Example 1 shows, to fully incorporate our statistical uncertainty, we cannot simply estimate the (posterior) distribution of the outcome. Instead, we need to retain some knowledge of how accurate that estimate is, and feed that additional knowledge into our decision making.

Instead of simply dealing with a single probability, we will study the effect of using the likelihood function (which indicates how well a model fits our observations) to generate a 'convex expectation', closely related to the risk measures often studied in mathematical finance. The theory of these nonlinear expectations is explored in detail in Föllmer and Schied [8] (up to some changes of sign) and gives a mathematically rigorous way to deal with 'Knightian uncertainty'. In economics, this is closely linked to Gilboa and Schmeidler's model of multiple priors [10]. However, little work has been done on connecting nonlinear expectations with statistics.

For Example 1 above, our proposal amounts to the following. Instead of working with the expected loss $E[\phi(X)]$ under one particular estimated measure, consider the quantity

$$\mathcal{E}(\phi(X)) = \sup_{q\in[0,1]} \left\{ E_q[\phi(X)] - (k^{-1}\alpha(q))^\gamma \right\}$$

for a fixed uncertainty aversion parameter $k > 0$ and exponent $\gamma \geq 1$. Here E_q denotes the expectation with probability q and α is the negative log-likelihood of our observations, shifted to have minimal value zero, that is (for $\hat{p} = S_N/N$ as above),

$$\alpha(q) = N\left(\hat{p}\log\left(\frac{\hat{p}}{q}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-q}\right)\right) \approx \frac{N}{\hat{p}(1-\hat{p})}(q-\hat{p})^2, \qquad (1)$$

where the approximation is for large N, in a sense to be explored later (see Section 3.2). The operator $\mathcal{E}$ gives an 'upper' expectation for the loss, depending on the certainty of our parameter estimate given the sample. In effect, we are considering all possible values for p, and using our data to determine how reasonable we think they are (as indicated by $-(k^{-1}\alpha(p))^{\gamma}$).

If we were to use $\mathcal{E}$ to choose between a family of wagers ϕ_i, we would obtain a classical minimax or 'robust optimization' problem (see for example Ben-Tal, El Ghaoui and Nemirovski [5]),

$$\min_i \mathcal{E}(\phi_i(X)) = \min_i \sup_{q\in[0,1]} \big\{E_q[\phi_i(X)] - (k^{-1}\alpha(q))^{\gamma}\big\}.$$

The expectation $\mathcal{E}$ can be thought of as an 'upper' expectation, and is convex. The corresponding 'lower' expectation $-\mathcal{E}(-\xi)$ can also be defined, and is concave. This leads naturally to

$$\Big[-\mathcal{E}(-\xi), \mathcal{E}(\xi)\Big]$$

as an interval prediction for ξ. Comparing with more familiar quantities, such as (frequentist) confidence intervals, (Bayesian) credible intervals and upper and lower probabilities in Dempster–Schafer theory, we see that an interval estimate is a natural object to study when describing uncertainty in parameters. We shall see that confidence intervals (in particular, likelihood intervals) arise as a special case of our approach.

We proceed as follows: First, we give a summary of some of the basic properties of nonlinear expectations. Secondly, we consider the effect of using the log-likelihood as the basis for a penalty function and the corresponding "divergence-robust nonlinear expectations", and their connection to relative entropy. Using this, we tease out generic large-sample approximations, in both parametric and non-parametric settings. Finally, we consider the connection between divergence-robust expectations and robust statistics (in particular M-estimates).

1.1. Nonlinear expectations

In this section we introduce the concepts of nonlinear expectations and convex risk measures, and discuss their connection with penalty functions on the space of measures. These objects provide a technical foundation with which to model the presence of uncertainty in a random setting. This theory is explored in some detail in Föllmer and Schied [8] and Frittelli and Rosazza-Gianin [9], among many others. We here present, without proof, the key details of this theory as needed for our analysis.

Definition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $L^\infty(\mathcal{F})$ denote the space of P-essentially bounded $\mathcal{F}$-measurable random variables. A nonlinear expectation on $L^\infty(\mathcal{F})$ is a mapping*

$$\mathcal{E} : L^\infty(\mathcal{F}) \to \mathbb{R}$$

satisfying the assumptions,

- *Strict Monotonicity: for any* $\xi_1, \xi_2 \in L^\infty(\mathcal{F})$*, if* $\xi_1 \geq \xi_2$ *a.s. then* $\mathcal{E}(\xi_1) \geq \mathcal{E}(\xi_2)$*, and if in addition* $\mathcal{E}(\xi_1) = \mathcal{E}(\xi_2)$ *then* $\xi_1 = \xi_2$ *a.s.*
- *Constant triviality: for any constant* $k \in \mathbb{R}$, $\mathcal{E}(k) = k$.
- *Translation equivariance: for any* $k \in \mathbb{R}$, $\xi \in L^\infty(\mathcal{F})$, $\mathcal{E}(\xi + k) = \mathcal{E}(\xi) + k$.

A 'convex' expectation in addition satisfies

- *Convexity: for any* $\lambda \in [0,1]$, $\xi_1, \xi_2 \in L^\infty(\mathcal{F})$,

$$\mathcal{E}(\lambda\xi_1 + (1-\lambda)\xi_2) \leq \lambda\mathcal{E}(\xi_1) + (1-\lambda)\mathcal{E}(\xi_2).$$

If $\mathcal{E}$ *is a convex expectation, then the operator defined by* $\rho(\xi) = \mathcal{E}(-\xi)$ *is called a* convex risk measure*. A particularly nice class of convex expectations is those which satisfy*

- *Lower semicontinuity: For a sequence* $\{\xi_n\}_{n\in\mathbb{N}} \subset L^\infty(\mathcal{F})$ *with* $\xi_n \uparrow \xi \in L^\infty(\mathcal{F})$ *pointwise,* $\mathcal{E}(\xi_n) \uparrow \mathcal{E}(\xi)$.

Definition 2. *Let* $\mathcal{M}_1$ *denote the space of all probability measures on* $(\Omega, \mathcal{F})$ *absolutely continuous with respect to* P.

The following theorem (which was expressed in the language of risk measures) is due to Föllmer and Scheid [8] and Frittelli and Rosazza-Gianin [9].

Theorem 1. *Suppose* $\mathcal{E}$ *is a lower semicontinuous convex expectation. Then there exists a 'penalty' function* $\alpha : \mathcal{M}_1 \to [0,\infty]$ *such that*

$$\mathcal{E}(\xi) = \sup_{Q\in\mathcal{M}_1} \big\{E_Q[\xi] - \alpha(Q)\big\}.$$

Provided $\alpha(Q) < \infty$ *for some* Q *equivalent to* P*, we can restrict our attention to measures in* $\mathcal{M}_1$ *equivalent to* P *without loss of generality.*

Remark 1. The convex expectation $\mathcal{E}$ is defined above as an operator on L^∞. However, given the equivalent representation

$$\mathcal{E}(\xi) = \sup_{\{Q\in\mathcal{M}_1:\alpha(Q)<\infty\}} \big\{E_Q[\xi] - \alpha(Q)\big\},$$

we can clearly define $\mathcal{E}(\xi)$ for a wider class of random variables. In particular, $\mathcal{E}(\xi)$ is well defined (but may be infinite) for all random variables ξ such that $E_Q[|\xi|] < \infty$ for some $Q \in \mathcal{M}_1$ with $\alpha(Q) < \infty$.

Given a convex nonlinear expectation $\mathcal{E}$, there is a natural class of 'acceptable' random variables for a decision problem, namely (given we evaluate losses) the convex level set

$$\mathcal{A} = \{\xi : \mathcal{E}(\xi) \leq 0\}.$$

One can also use a nonlinear expectation as a value to be optimized; in this setting the convexity of the operator is of significant interest. Finally, one can use a nonlinear expectation to give a robust point estimate of ξ, given a loss function ϕ, by choosing the value $\hat{\xi} \in \mathbb{R}$ which minimizes the loss $\mathcal{E}(\phi(\xi - \hat{\xi}))$ (cf. Wald [21]).

2. Penalties and likelihood

The general framework of nonlinear expectations is well suited to modelling Knightian uncertainty, but is not usually connected with statistical estimation. We would like to have a general principle for treating our uncertainty, which is closely tied to classical statistics. Rather than continuing to take an abstract axiomatic approach, we shall consider a concrete proposal, based on the following classical concepts.

Definition 3. *For a model $Q \in \mathcal{M}_1$, let $L(Q|\mathbf{x})$ denote the likelihood of $\mathbf{x}$ under Q, that is the density of $\mathbf{x}$ with respect to a reference measure (which we shall take to be Lebesgue measure on $\mathbb{R}^N$ for simplicity).*

Let $\mathcal{Q} \subseteq \mathcal{M}_1$ be a set of models under consideration (for example, a parametric set of distributions). We then define the "$\mathcal{Q}|\mathbf{x}$-divergence" to be the negative log-likelihood ratio

$$\alpha_{\mathcal{Q}|\mathbf{x}}(Q) := -\log\big(L(Q|\mathbf{x})\big) + \sup_{\tilde{Q}\in\mathcal{Q}}\Big\{\log\big(L(\tilde{Q}|\mathbf{x})\big)\Big\}. \tag{2}$$

Remark 2. The right hand side of (2) is well defined whether or not a maximum likelihood estimator[4] exists. Given a $\mathcal{Q}$-MLE $\hat{Q}$, we would have the simpler representation

$$\alpha_{\mathcal{Q}|\mathbf{x}}(Q) := -\log\Big(\frac{L(Q|\mathbf{x})}{L(\hat{Q}|\mathbf{x})}\Big).$$

Definition 4. *For fixed observations $\mathbf{x}$, for an uncertainty aversion parameter $k > 0$ and exponent $\gamma \in [1, \infty]$, we define the convex expectation*

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi) := \sup_{Q\in\mathcal{Q}}\Big\{E_Q[\xi(\omega,\mathbf{x})] - \Big(\frac{1}{k}\alpha_{\mathcal{Q}|\mathbf{x}}(Q)\Big)^{\gamma}\Big\}, \tag{3}$$

where we adopt the convention $x^\infty = 0$ for $x \in [0,1]$ and $+\infty$ otherwise. Here $\xi : \Omega \times \mathbb{R}^N \to \mathbb{R}$ is a Borel measurable function with respect to $\mathbf{x}$.

We call $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}$ the "$\mathcal{Q}|\mathbf{x}$-divergence robust expectation" (with parameter k, γ), or simply the "DR-expectation[5]".

Remark 3. We have defined the DR-expectation separately for each fixed observation vector $\mathbf{x}$. For this to be meaningful, we formally first consider $\mathbf{x}$ as random, and choose a version of the divergence $\alpha_{\mathcal{Q}|\cdot}(Q)$ for every measure $Q \in \mathcal{Q}$. Using this approach, it is difficult to show that the map $\mathbf{x} \to \mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi)$ is measurable, particularly given the presence of the supremum in the definition. For reasonable choices of $\mathcal{Q}$ (which give continuous versions of the divergence and conditional expectation), this will nevertheless typically be the case (cf. Remark 11).

[4] Recall that a $\mathcal{Q}$-MLE (maximum likelihood estimator) is a map $\mathbf{x} \to \hat{Q} \in \mathcal{Q}$ such that $L(\hat{Q}|\mathbf{x}) \geq L(Q|\mathbf{x})$ for all $Q \in \mathcal{Q}$. We say that a quantity Y is a $\mathcal{Q}$-MLE for $E_Q[\xi]$ if $Y = E_{\hat{Q}}[\xi]$ where $\hat{Q}$ is a $\mathcal{Q}$-MLE.

[5] This acronym could also stand for 'Data-driven Robust expectation', which may be a preferable emphasis.

In the definition of the DR-expectation, we have allowed ξ to depend (explicitly) on the observations $\mathbf{x}$. We shall now focus our attention on the following special case, which allows analytic tractability. Unless otherwise stated, we take this assumption as given from this point onwards.

Assumption 1. *Suppose* $\mathbf{x} = \{X_n\}_{n=1}^N$ *and, under each* $Q \in \mathcal{Q}$*, we know* $X, \{X_n\}_{n\in\mathbb{N}}$ *are iid random variables and* $\xi = \phi(X)$ *for some Borel function* ϕ*.*

Note that, under Assumption 1, we can write $E_Q[\xi(\omega,\mathbf{x})] = E_Q[\xi] = E_Q[\phi(X)]$ for all $Q \in \mathcal{Q}$. Our attention will mainly be on the two extremal cases $\gamma = 1$ and $\gamma = \infty$, where

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\xi) := \sup_{Q\in\mathcal{Q}} \Big\{ E_Q[\xi] - \frac{1}{k}\alpha_{\mathcal{Q}|\mathbf{x}}(Q)\Big\}, \quad \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(\xi) := \sup_{\{Q:\alpha_{\mathcal{Q}|\mathbf{x}}(Q)<k\}} \Big\{E_Q[\xi]\Big\}.$$

The intervening cases are natural interpolations between these. The statement $1^\infty = 0$ is natural from a convex analytic perspective, as it implies $|x|^q$ is proportional to the convex dual of $|x|^p$, whenever $p^{-1}+q^{-1} = 1$, for $p \in [1,\infty]$.

Remark 4. In Example 1 above[6], $\mathcal{Q}$ corresponds to the set of measures such that $X, \{X_n\}_{n=1}^N$ are iid Bernoulli with parameter $p \in [0,1]$. In this example, we did not consider all measures in $\mathcal{M}_1$ (this would include, for example, models where $\{X_n\}_{n=1}^N$ and X come from completely unrelated distributions), but neither did we restrict our attention to a single $Q \in \mathcal{Q}$.

Typically the operator $\mathcal{E}$ cannot be evaluated by hand, instead numerical optimization or approximation is needed. In the setting of Example 1 above, if $\gamma = 1$ then a closed form representation can be obtained, however is quite inelegant (the optimal q is the solution to a quadratic equation, but the resulting equation for $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi)$ does not simplify). A simple example where closed form quantities can be derived is when the data are assumed to be Gaussian with unknown mean.

Example 2. Suppose $\mathbf{x} = (X_1, X_2, ..., X_N)$ and $\mathcal{Q}$ corresponds to those measures under which $X, \{X_n\}_{n=1}^N$ are iid $N(\mu, 1)$ random variables, where μ is unknown. Then, if $\bar{X} = N^{-1}\sum_{n=1}^N X_n$ denotes the sample mean, for any constant $\beta > 0$, consider $\xi = \beta X$. Simple calculus can be used to derive

$$\begin{aligned}
\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\beta X) &= \sup_{\mu\in\mathbb{R}} \Big\{ \beta\mu - \Big(\frac{1}{2k}\Big(\sum_{n=1}^N (\mu - X_n)^2 - \sum_{n=1}^N (\bar{X} - X_n)^2\Big)\Big)^\gamma \Big\} \\
&= \sup_{\mu\in\mathbb{R}} \Big\{ \beta\mu - \Big(\frac{N}{2k}(\mu-\bar{X})^2\Big)^\gamma\Big\} \\
&= \beta\bar{X} + \beta^{\frac{2\gamma}{2\gamma-1}}\Big(\frac{2k}{N}\Big)^{\frac{\gamma}{2\gamma-1}}(2\gamma)^{\frac{-2\gamma}{2\gamma-1}}(2\gamma-1).
\end{aligned}$$

[6]Of course, this example does not satisfy our assumption that all measures in $\mathcal{Q}$ are absolutely continuous with respect to Lebesgue measure, but one can observe directly that the DR-expectation is well defined (and measurable) in this case.

In particular, when $\gamma = 1$, we have

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X) = \beta\bar{X} + \frac{\beta^2 k}{2N} = \beta\bar{X} + \frac{k}{2}\operatorname{Var}(\beta\bar{X})$$

and, from the definition,

$$\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(\beta X) = \beta\bar{X} + \beta\sqrt{\frac{2k}{N}} = \beta\bar{X} + \sqrt{2k}\operatorname{Sd}(\beta\bar{X}).$$

In the latter case, taking $k \approx 2$, we obtain the upper bound of the classical 95% confidence interval for $E[\beta X]$.

The corresponding lower expectations are given by the symmetric quantities

$$-\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(-\beta X) = \beta\bar{X} - \frac{\beta^2 k}{2N}, \qquad -\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(-\beta X) = \beta\bar{X} - \beta\sqrt{\frac{2k}{N}}.$$

From this example we can observe a few phenomena, which we will discuss more generally below. First, for $\gamma = \infty$, $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}$ is positively homogeneous, that is, $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\beta X) = \beta\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(X)$ for all $\beta > 0$, and there is a close relationship between $\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(X)$ and the classical confidence interval for $E[X]$. On the other hand, this is not satisfied when $\gamma \neq \infty$.

Secondly, for any γ, as the ratio of the uncertainty parameter and the sample size $k/N \to 0$, the DR-expectation converges to the (unique) $\mathcal{Q}$-MLE $\beta\bar{X}$. This convergence is of the order $(k/N)^{\frac{\gamma}{2\gamma-1}}$.

In this setting we can also calculate, for $\beta > 0$,

$$\begin{aligned}\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X^2) &= \sup_{\mu\in\mathbb{R}}\Big\{\beta(1+\mu^2) - \frac{N}{2k}(\mu - \bar{X})^2\Big\} \\ &= \begin{cases}\beta + \beta\Big(\frac{N(N-2k\beta^2)}{(N-2k\beta)^2}\Big)\bar{X}^2 & \beta < N/2k \\ +\infty & \beta \geq N/2k\end{cases}\end{aligned}$$

whereas

$$\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(\beta X^2) = \beta\Big(1 + \bar{X} + \sqrt{2k/N}\Big)^2,$$

which is always finite. Explosions in $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}$ will be considered in more detail in Section 4. Notice that again, as $k/N \to 0$,

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\beta X^2) \to \beta(1 + \bar{X}^2),$$

which is the $\mathcal{Q}$-MLE for $E[\beta X^2]$.

We have noticed above, in the Gaussian case, that our nonlinear expectation is positively homogeneous only in the case $\gamma = \infty$. This is a general fact, as shown by the following proposition.

Proposition 1. *In the case $\gamma = \infty$ (and only in this case, provided the likelihood is finite and varying for a nontrivial subset of $\mathcal{Q}$), our nonlinear expectation is positively homogeneous, that is*

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\beta\xi) = \beta\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi) \text{ for all } \beta > 0 \text{ iff } \gamma = \infty.$$

Proof. It is classical (see for example Föllmer and Schied [8]) that a convex nonlinear expectation is positively homogeneous if and only if the penalty takes only the values $\{0, \infty\}$. Given the likelihood is finite and varying on a nontrivial subset of $\mathcal{Q}$, this is not the case for any $\gamma < \infty$, but is the case for $\gamma = \infty$ by definition. □

2.1. Dynamic consistency

Within the theory of nonlinear expectations, much attention has been paid to questions of dynamic consistency. If we have a family $\{\mathcal{E}_s\}_{s\geq 0}$ of 'conditional' nonlinear expectations relative to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$, then dynamic consistency requires, for every ξ and all $s \leq t$, that we have (i) the recursivity relationship $\mathcal{E}_s(\mathcal{E}_t(\xi)) = \mathcal{E}_s(\xi)$ and (ii) the regularity condition $\mathcal{E}_s(I_A\xi) = I_A\mathcal{E}_s(\xi)$ for all $A \in \mathcal{F}_t$. This concept is generally not appropriate for our approach, as the expectations we define are typically not consistent. This can be seen from the following easy extension of Example 2.

Example 3. In the context of Example 2, write $\mathbf{x}_N = \{X_1, ..., X_N\}$, so $\mathcal{F}_N = \sigma(\mathbf{x}_N)$. We have

$$\begin{aligned}\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_1}(\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_2}(X)) &= \mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_1}\Big(\frac{X_1+X_2}{2} + \frac{k}{4}\Big) = \frac{X_1}{2} + \frac{k}{4} + \mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_1}\Big(\frac{X_2}{2}\Big)\\ &= \frac{X_1}{2} + \frac{k}{4} + \frac{X_1}{2} + \frac{k}{8} = X_1 + \frac{3k}{8}\\ &\neq X_1 + \frac{k}{2} = \mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_1}(X)\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_1}(\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_2}(X)) &= \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_1}\Big(\frac{X_1+X_2}{2} + \sqrt{\frac{2k}{2}}\Big) = \frac{X_1}{2} + \sqrt{k} + \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_1}\Big(\frac{X_2}{2}\Big)\\ &= \frac{X_1}{2} + \sqrt{k} + \frac{X_1}{2} + \sqrt{\frac{k}{2}} = X_1 + (1+2^{-1/2})\sqrt{k}\\ &\neq X_1 + \sqrt{2k} = \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_1}(X).\end{aligned}$$

So in either case, the nonlinear expectation $\{\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}_N}\}_{N\in\mathbb{N}}$ is not recursive.

In effect, our problem differs from the dynamically consistent one in the following (closely related) key ways:

- In a dynamically consistent setting, the penalty is prescribed while the observations lead to conditional expectations appearing in the nonlinear expectation. In our setting, the penalty is *determined by the observations* (through the $\mathcal{Q}|\mathbf{x}$-divergence), and so only the family of models $\mathcal{Q}$ and the constants k, γ need to be specified. In this way, the observations will inform our understanding of the real-world probabilities directly, rather than simply replacing them with conditional probabilities.

- In a dynamically consistent setting, the underlying models are typically required to be stable under pasting through time. Conceptually, this implies that there is no significant link assumed between the 'true' model governing our observations at different times[7]. Conversely, in our setting, we typically assume that the underlying model is constant through time (e.g. through the assumption that our observations are iid), and hence repeated observations *can* inform our view of the 'true' model.

2.2. *Exponentials and entropy*

It will not be a surprise that there is a connection between the convex expectation we propose and a more traditional quantity in risk-averse decision making, namely the certainty equivalent under exponential utility.

Definition 5. *For a random variable ξ, under a reference measure P, the certainty equivalent under exponential utility has definition*

$$\mathcal{E}^k_{\exp}(\xi) = \frac{1}{k} \log E_P[\exp(k\xi)]$$

where $k > 0$ is a risk-aversion parameter. Defining the relative entropy (or Kullback–Liebler divergence)

$$D_{\mathrm{KL}}(Q||P) = E_Q\Big[\log\Big(\frac{dQ}{dP}\Big)\Big] = E_P\Big[\frac{dQ}{dP}\log\Big(\frac{dQ}{dP}\Big)\Big]$$

we have the representation (see, for example, [8])

$$\mathcal{E}^k_{\exp}(\xi) = \sup_{Q\in\mathcal{M}_1}\Big\{E_Q[\xi] - \frac{1}{k}D_{\mathrm{KL}}(Q||P)\Big\}.$$

Replacing expectations by conditional expectations, we obtain the conditional certainty equivalent.

Remark 5. It is useful to consider the relative entropy of the law of X separately from the other observations $\{X_n\}_{n\in\mathbb{N}}$. We therefore define

$$D_{\mathrm{KL}|X}(Q||P) = \int \log\Big(\frac{f(x,Q)}{f(x,P)}\Big) f(x,Q)dx.$$

[7] In particular, new observations cannot affect our opinions on the measure which was active at an earlier time. As previously mentioned, this connects more generally with the concerns of Knight [14], who discusses the problem that observations at different times may be from different models. The difficulty lies in the fact that, without some presumption of homogeneity in nature, statistical inference is impossible.

Assuming $\hat{Q} \approx P$, where P is the 'real world' probability measure, in light of the law of large numbers we hope for a simple connection, at least asymptotically, between the scaled deviance

$$\frac{1}{N}\alpha_{\mathcal{Q}|\mathbf{x}}(Q) = -\frac{1}{N}\sum_{n=1}^{N}\log\Big(\frac{f(X_n,Q)}{f(X_n,\hat{Q})}\Big) \approx -E_P\Big[\log\Big(\frac{f(X,Q)}{f(X,\hat{Q})}\Big)\Big] \approx D_{\mathrm{KL}|X}(P||Q)$$

and the penalty in the exponential utility, that is, $D_{\mathrm{KL}}(Q||P)$. In general, this is made more difficult by the fact we have an infinite family of measures $\mathcal{Q}$, and by the lack of symmetry in the relative entropy, as $D_{\mathrm{KL}}(Q||P) \neq D_{\mathrm{KL}}(P||Q)$. We shall pursue this connection in the coming section.

Remark 6. One extension of our approach is to change the penalty function to include a entropy term taken in the 'other' direction, that is, to use the penalty

$$\alpha^{k,\beta}(Q) = \inf_{Q'\in\mathcal{Q}}\Big\{\frac{1}{\beta}D_{\mathrm{KL}}(Q||Q') + \frac{1}{k}\alpha_{\mathcal{Q}|\mathbf{x}}(Q')\Big\}$$

for some $\beta > 0$. This is particularly of interest where we wish to include both uncertainty aversion and risk aversion (as measured using exponential utility[8]). This is well defined for all measures $Q \in \mathcal{M}_1$, and gives the expectation:

$$\sup_{Q\in\mathcal{M}_1}\big\{E_Q[\xi] - \alpha^{k,\beta}(Q)\big\} = \frac{1}{\beta}\log\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(e^{\beta\xi}).$$

3. Large-sample theory

In this section, we shall seek to study the large-sample theory of the nonlinear expectation $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}$. In practice, this is particularly useful to give approximations and qualitative descriptions of its behaviour.

Throughout this section, we shall assume that we have observations $\{X_n\}_{n\in\mathbb{N}}$, and a family of measures $\mathcal{Q}$ under which $X, \{X_n\}_{n\in\mathbb{N}}$ are iid random variables with corresponding densities $f(x;Q)dx$. We write $\mathbf{x}_N = (X_1, ..., X_N)$. We shall be interested in determining the behaviour, for large N, of $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X))$, where ϕ is a bounded function. For simplicity, we shall assume that the MLE exists (however our results can be extended to remove this assumption, with an increase in notational complexity). We write $\hat{Q}_N$ for the $\mathcal{Q}$-MLE based on observations $\mathbf{x}_N$.

Given the lack of positive homogeneity, it is useful to consider the behaviour of $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(c\xi)$ when varying c. The following lemma connects variation of c with variation in k.

Lemma 1. *For any $c, k > 0$, any $\gamma < \infty$, any random variable ξ,*

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(c\,\xi) = c\,\mathcal{E}^{k',\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi),$$

where $k' = c^{1/\gamma}k$.

[8] Replacing entropy with a different penalty would allow for other risk aversion functionals to be considered, if desired. This is a form of inf-convolution, as considered by Barrieu and El Karoui [3].

Proof.

$$\mathcal{E}^{1,\gamma}_{\mathcal{Q}|\mathbf{x}}(c\xi) = \sup_{Q\in\mathcal{Q}}\Big\{E_Q[c\xi] - \big(\alpha_{\mathcal{Q}|\mathbf{x}}(Q)\big)^\gamma\Big\}$$
$$= c\sup_{Q\in\mathcal{Q}}\Big\{E_Q[\xi] - \Big(\frac{1}{c^{1/\gamma}k}\alpha_{\mathcal{Q}|\mathbf{x}}(Q)\Big)^\gamma\Big\} = c\mathcal{E}^{k',\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi). \qquad \square$$

To enable a simple description of our asymptotic results, we recall the following definition. Here P^* is the outer measure associated with P, to deal with any potential lack of measurability.

Definition 6. *Consider sequences $f = \{f_n\}_{n\in\mathbb{N}}$ and $g = \{g_n\}_{n\in\mathbb{N}}$ of functions $\Omega \to \mathbb{R}$.*

(i) We write $f = O_P(g)$ whenever f_n/g_n is stochastically bounded, that is, $P^*(|f_n/g_n| > M) \to 0$ as $M \to \infty$ for each n,
(ii) We write $f = o_P(g)$ whenever $\lim_{n\to\infty} P^*(|f_n/g_n| > \epsilon) = 0$ for all $\epsilon > 0$.

Note that this depends on the choice of measure P.

Remark 7. In the above analysis, we have assumed that $\mathcal{Q}$ is a family of measures absolutely continuous with respect to a given measure P. For a finite observation vector $\mathbf{x}_N = \{X_1, ..., X_N\}$, this assumption is not overly restrictive, as we only made use of the behaviour of the measures on $\sigma(\mathbf{x}_N, X)$ (cf. Assumption 1). When we start considering the limiting behaviour of our expectations, this becomes problematic – Assumption 1 implies that, for every $Q \in \mathcal{Q}$, the restrictions of Q and P to $\sigma(\mathbf{x}_N, X)$ are equivalent, for every finite N, but will generally be singular for $N = \infty$. This weaker assumption implies that the statements $f = O_P(g)$ and $f = O_Q(g)$ are no longer equivalent.

3.1. Nonparametric results

We now give some results for a general $\mathcal{Q}$. We restrict our attention to bounded random variables $\xi = \phi(X)$. Given we will take a supremum over a family of densities, we need a uniform version of the law of large numbers. For this reason, we make the following definition.

Definition 7. *We say a family $\mathcal{Q}$ is a Glivenko–Cantelli–Donsker family of measures (or GCD family) if, for any $P \in \mathcal{Q}$,*

$$\sup_{Q\in\mathcal{Q}}\Big\{\frac{\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q)}{N} - D_{\mathrm{KL}|X}(P||Q)\Big\} = O_P(N^{-1/2})$$

Remark 8. The reason for the name (Glivenko–Cantelli–Donsker) is simply because, if we have a uniform weak Glivenko–Cantelli theorem when indexing the empirical distribution by the family of log-likelihoods, then the term in brackets converges in probability to 0. If we also have a uniform Donkser theorem, then we know that $\sqrt{N}\big(\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q)/N - D_{\mathrm{KL}}(P||Q)\big)$ converges (in some sense) to a finite-valued Gaussian process, which implies it is of the order stated.

It is easy to show, given consistency of the MLE and some integrability, that a finite family $\mathcal{Q}$ is always a GCD family. Clearly a subset of a GCD family is also a GCD family. The following lemma gives a sufficient condition for a family of measures to be GCD, in terms of the smoothness of the densities.

Lemma 2. *Suppose $\mathcal{Q}$ is a family of measures such that $\{X_n\}_{n\in\mathbb{N}}$ are iid with respective densities $\{f(\cdot;Q)\}_{Q\in\mathcal{Q}}$ which satisfy*

i) there is a compact set K, and a uniform constant $\epsilon > 0$ such that for every $P \in \mathcal{Q}$, $P(X \in K) = 1$, and $f(\cdot;P) > \epsilon$ on K,
ii) there is $C < \infty$ and $\rho > 1/2$ such that, for all $P, Q \in \mathcal{Q}$, the likelihood ratios $f(\cdot, Q)/f(\cdot, P)$ take values in $[C^{-1}, C]$ and are uniformly ρ-Hölder continuous with norm C, that is, writing $L(x) = f(x, Q)/f(x, P)$,

$$\sup_{x,y} \frac{|L(x) - L(y)|}{|x - y|^\rho} \leq C.$$

Then $\mathcal{Q}$ is a GCD family.

Proof. See Appendix. □

Example 4. For fixed $m \in \mathbb{N}$, $\bar{\sigma} > 0$ and $\epsilon > 0$, suppose we have observations $\{X_i\}_{i=1}^N$ modelled using mixtures of Gaussian distributions (with a small regularizing diffuse component). In particular, suppose the density for each observation can be written

$$f_X(x) = \epsilon\varphi(x; 0, \bar{\sigma}^2) + (1 - \epsilon)\sum_{j=1}^{m} \pi_j \varphi(x; \mu_j, \sigma_j^2)$$

where $\varphi(\cdot; \mu_j, \sigma_j^2)$ is the density of the $N(\mu_j, \sigma_j^2)$ distribution. A measure $Q \in \mathcal{Q}$ is then determined by the parameters (μ, σ, π), where μ is a vector in $[-\epsilon^{-1}, \epsilon^{-1}]^m$, σ a vector in $[\epsilon, \bar{\sigma}/2]^m$ and π is a probability vector in $\mathbb{R}^m$.

Defining $\tilde{X}_i = \Phi(X_i/\bar{\sigma})$, where Φ is the standard normal cdf, one can check that (for every $m, \bar{\sigma}, \epsilon$) the conditions of Lemma 2 are satisfied for the density of $\tilde{X}_i$. As Φ is invertible, X_i and $\tilde{X}_i$ generate the same divergence and their distributions have the same entropy. We conclude that $\mathcal{Q}$ is a GCD family.

This example shows that assuming $\mathcal{Q}$ is a GCD family is a technical requirement, rather than being a significant restriction on modelling (as sending $m, \bar{\sigma} \to \infty, \epsilon \to 0$ we obtain a weakly dense family of densities for X).

Remark 9. Given the generality of the assumptions we have placed on our family of measures, obtaining a rigorous proof is difficult. It will prove convenient to allow the uncertainty aversion parameter k to depend on the sample size N. We prove two main results:

- As the sample size becomes large, provided $k_N = o(N)$ the DR expectation is consistent. (Note that this includes the case when k is constant.)
- As the sample size becomes large, provided $k_N = o(N)$ and $k_N > O(N^{1/2})$, we obtain a second-order estimate for the expectation, corresponding to the Gaussian case. Provided the family $\mathcal{Q}$ is sufficiently rich, this bound is attained (this forms a 'central limit'-type result).

By Lemma 1, this second case can also be interpreted as being a result for 'asymptotically large' random variables, as taking $\delta \in (1/2, 1)$ we have the identity $N^{-\delta}\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(N^\delta \xi) = \mathcal{E}^{N^\delta k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi)$.

We begin with the case $\gamma = 1$.

Theorem 2. *Suppose $\mathcal{Q}$ is a GCD family of measures and $k_N = o(N)$. Consider a random variable $\xi = \phi(X)$, where ϕ is a bounded measurable function and $X, \{X_n\}_{n\in\mathbb{N}}$ are iid under every $Q \in \mathcal{Q}$.*

(i) $\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}$ is a consistent estimator, that is

$$\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) = E_P[\xi] + o_P(1)$$

as $N \to \infty$ for every $P \in \mathcal{Q}$.

(ii) We have the asymptotic behaviour (as $N \to \infty$, for each $P \in \mathcal{Q}$)

$$\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) \leq E_P[\xi] + \frac{k_N}{N}\frac{1}{2}\mathrm{Var}_P(\xi) + O\Big(\Big(\frac{k_N}{N}\Big)^2\Big) + O_P\Big(\frac{N^{1/2}}{k_N}\Big)$$

with equality whenever P is such that, for all N sufficiently large, the measure $\tilde{P}$ with density

$$f(x;\tilde{P}) = \frac{f(x;P)}{\lambda - \frac{k_N}{N}\phi(x)}$$

(where $\lambda > \sup_x \phi(x)$ is chosen to ensure this is a probability density) is also in $\mathcal{Q}$.

Remark 10. Given the error of the expectation based on the $\mathcal{Q}$-MLE is asymptotically of the order of $1/\sqrt{N}$, the requirement implied by (ii) that k_N grows faster than $\sqrt{N}$, is unsurprising, as this is what is needed to ensure that the risk aversion term $\frac{k_N}{2N}\mathrm{Var}(\xi)$ asymptotically dominates the statistical error of the estimation of $E_P[\xi]$.

Proof. We begin by proving (ii). As $\mathcal{Q}$ is a GCD family, we know that,

$$\Big|\frac{1}{N}\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q) - D_{\mathrm{KL}|X}(P||Q)\Big| \leq O_P(N^{-1/2}),$$

with error bounded independently of Q. Hence, uniformly in Q,

$$\Big|\frac{1}{k_N}\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q) - \frac{N}{k_N}D_{\mathrm{KL}|X}(P||Q)\Big| \leq O_P(N^{1/2}/k_N).$$

Calculating $\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi)$, we have

$$\begin{aligned}\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) &= \sup_{Q\in\mathcal{Q}}\Big\{E_Q[\xi] - \frac{1}{k_N}\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q)\Big\} \\ &= \sup_{Q\in\mathcal{Q}}\Big\{E_Q[\xi] - \frac{N}{k_N}D_{\mathrm{KL}|X}(P||Q)\Big\} + O_P(N^{1/2}/k_N).\end{aligned}$$

We shall now focus on solving the problem under the assumption that the penalty is given by $(N/k_N)D_{\mathrm{KL}}(P||Q)$.

For fixed N, we can try and solve this simplified problem directly. Assuming the optimum will be attained with a measure denoted Q^g, this corresponds to finding the density $g = f(\cdot, Q^g)$. Calculus of variations yields

$$\phi + \frac{N}{k_N}\Big(\frac{g}{f(\cdot,P)} + \lambda\Big) = 0,$$

or equivalently

$$g = \frac{f(\cdot,P)}{\lambda - \frac{k_N}{N}\phi},$$

where λ is chosen to ensure g is a density, that is, $E_P[(\lambda - \frac{k_N}{N}\xi)^{-1}] = 1$. This requires $\lambda > \frac{k_N}{N}\sup_x \phi(x)$ (this is the reason we have assumed $\xi = \phi(X)$ is bounded). As the map $\lambda \mapsto (\lambda - \frac{k_N}{N}\phi(x))^{-1}$ is monotone, we also know that the corresponding value of λ is unique and

$$\lambda \in \Big[1 + \frac{k_N}{N}\inf_x \phi(x), \quad 1 + \frac{k_N}{N}\sup_x \phi(x)\Big].$$

This avoids inconsistency with the requirement $\lambda > \frac{k_N}{N}\sup_x \phi(x)$ whenever N is large enough that $\frac{N}{k_N} > 2\sup_x |\phi(x)|$. For every fixed large N, we have a compact set of values for λ. Therefore, we can assume $(\lambda - \frac{k_N}{N}\xi)^{-1}$ is uniformly approximated by its Taylor series in λ around $\lambda = 1 + \frac{k_N}{N}E_P[\xi]$. Furthermore, we immediately see the first approximation

$$\lambda = 1 + \frac{k_N}{N}E_P[\xi] + O(k_N/N).$$

Expanding the Taylor series of $(\lambda - \frac{k_N}{N}\xi)^{-1}$, we have

$$1 = E_P\Big[1 - \Big(\lambda - 1 - \frac{k_N}{N}\xi\Big) + \Big(\lambda - 1 - \frac{k_N}{N}\xi\Big)^2 + \dots\Big]$$

or equivalently

$$\lambda = 1 + \frac{k_N}{N}E_P[\xi] + E_P\Big[\Big(\lambda - 1 - \frac{k_N}{N}\xi\Big)^2\Big] + O\Big(E_P\Big[\Big(\lambda - 1 - \frac{k_N}{N}\xi\Big)^3\Big]\Big). \quad (4)$$

Substituting our first approximation of λ on the right hand side of (4), we have

$$\lambda = 1 + \frac{k_N}{N}E_P[\xi] + \Big(\frac{k_N}{N}\Big)^2 \mathrm{Var}_P[\xi] + O\Big(\Big(\frac{k_N}{N}\Big)^2\Big).$$

Substituting this second approximation back into (4), we observe that the error can be taken to be $O((k_N/N)^3)$, rather than $O((k_N/N)^2)$.

We can now approximate our convex expectation. We know that

$$\begin{aligned} E_{Q^g}[\xi] &= E_P\Big[\frac{\xi}{\lambda - \frac{k_N}{N}\xi}\Big] = E_P\Big[\xi\Big(1 - \frac{k_N}{N}(E_P[\xi] - \xi) + O((k_N/N)^2)\Big)\Big] \\ &= E_P[\xi] + \frac{k_N}{N}\mathrm{Var}_P[\xi] + O((k_N/N)^2) \end{aligned}$$

and similarly

$$E_P\Big[\log\Big(\frac{dP}{dQ^g}\Big)\Big] = E_P\Big[\log\Big(\lambda - \frac{k_N}{N}\xi\Big)\Big] = \frac{1}{2}\Big(\frac{k_N}{N}\Big)^2 \mathrm{Var}_P[\xi] + O\Big(\Big(\frac{k_N}{N}\Big)^3\Big).$$

Hence we can calculate the desired approximation

$$\begin{aligned}\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) &= \sup_{Q\in\mathcal{Q}}\Big\{E_Q[\xi] - \frac{N}{k_N}D_{\mathrm{KL}|X}(P||Q)\Big\} + O_P\Big(\frac{N^{1/2}}{k_N}\Big)\\ &\leq E_{Q^g}[\xi] - \frac{N}{k_N}E_P\Big[\log\Big(\frac{dP}{dQ^g}\Big)\Big] + O_P\Big(\frac{N^{1/2}}{k_N}\Big)\\ &= E_P[\xi] + \frac{k_N}{2N}\mathrm{Var}_P(\xi) + O\Big(\Big(\frac{k_N}{N}\Big)^2\Big) + O_P\Big(\frac{N^{1/2}}{k_N}\Big).\end{aligned} \tag{5}$$

with equality whenever $Q^g \in \mathcal{Q}$, as stated in (ii).

We now seek to reduce to the assumptions of (i). As increasing k_N will only increase the (nonnegative) differences

$$\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) - E_{\hat{Q}_N}[\xi], \qquad E_{\hat{Q}_N}[\xi] + \mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(-\xi)$$

and we know that $E_{\hat{Q}_N}[\xi]$ is consistent, we can assume that $N^{1/2}/k_N \to 0$ without loss of generality. Under this assumption, the right hand side of (5) converges to $E_P[\xi]$, and hence we verify that $\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi) \to_P E_P[\xi]$ as desired. □

We will now consider the case $\gamma = \infty$. It is easy to check that the interval

$$\mathcal{I}_N(\xi) = \Big[-\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(-\xi), \quad \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\xi)\Big]$$

is a likelihood interval for $E[\xi]$, that is, it corresponds to the range of expectations under the measures in $\mathcal{Q}$ with likelihood at least e^{-k}. Such intervals are commonly used as generalizations of confidence intervals (see for example Hudson [12], drawing on the well known results of Neyman and Pearson [17]). In this context, we shall see that a stronger property holds, as the confidence region is uniform in ϕ. (See also Theorem 7.)

We here only use the assumption that $k_N = o(N)$ (which includes the case where k is constant).

Theorem 3. *Suppose $\mathcal{Q}$ is a GCD family and $X, \{X_n\}_{n\in\mathbb{N}}$ are iid under each $Q \in \mathcal{Q}$. Then if $k_N = o(N)$, the nonlinear expectation with $\gamma = \infty$ is a* uniformly *consistent estimator, that is,*

$$\sup_{\phi:|\phi|\leq 1}\Big\{\mathcal{E}^{k_N,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) - E_P[\phi(X)]\Big\} = o_P(1) \quad \textit{for all } P \in \mathcal{Q}.$$

Proof. Observe that

$$\mathcal{E}^{k_N,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = \sup_{\{Q:\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q)\leq k_N\}}\{E_Q[\phi(X)]\}.$$

As $\mathcal{Q}$ is a GCD family, we know that for any $P \in \mathcal{Q}$,

$$\frac{1}{N}\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q) = D_{\mathrm{KL}|X}(P||Q) + O_P(N^{-1/2})$$

and so, provided $k_N = o(N)$,

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q) \leq k_N \qquad \Leftrightarrow \qquad D_{\mathrm{KL}|X}(P||Q) \leq \frac{k_N}{N} + O_P(N^{-1/2}) = o_P(1)$$

with the terminal error uniform in Q. From Pinsker's inequality, looking only at the marginal law of X, we know that the total variation norm satisfies

$$\int |f(x;P) - f(x;Q)|dx = \big\|P|_{\sigma(X)} - Q|_{\sigma(X)}\big\|_{\mathrm{TV}} \leq \sqrt{2D_{\mathrm{KL}|X}(P||Q)}.$$

Therefore,

$$\begin{aligned}
&\sup_{\phi:|\phi|\leq 1}\Big\{\mathcal{E}^{k_N,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) - E_P[\phi(X)]\Big\}\\
&= \sup_{\phi:|\phi|\leq 1}\sup_{\{Q:\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q)\leq k_N\}}\Big\{E_Q[\phi(X)] - E_P[\phi(X)]\Big\}\\
&= \sup_{\phi:|\phi|\leq 1}\sup_{\{Q:D_{\mathrm{KL}|X}(P||Q)\leq o_P(1)\}}\Big\{E_Q[\phi(X)] - E_P[\phi(X)]\Big\}\\
&\leq \sup_{\{Q:D_{\mathrm{KL}|X}(P||Q)\leq o_P(1)\}}\Big\{\big\|P|_{\sigma(X)} - Q|_{\sigma(X)}\big\|_{\mathrm{TV}}\Big\}\\
&\leq o_P(1).
\end{aligned}$$

It follows that the nonlinear expectation is a uniformly consistent estimator. □

By a simple comparison, we also obtain consistency for all other $\gamma \in [1,\infty]$.

Corollary 1. *If $\mathcal{Q}$ is a GCD family, $k_N = o(N)$ and $\gamma \in [1,\infty]$, the nonlinear expectation $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}_N}(\phi(\xi))$ is a consistent estimator of $E_P[\phi(\xi)]$, for any bounded Borel-measurable ϕ.*

Proof. We know that the two extreme cases $\gamma = 1$ and $\gamma = \infty$ are both consistent, as is the MLE $E_{\hat{Q}_N}[\phi(\xi)]$ (this follows, for example, from the fact $E_{\hat{Q}_N}[\xi] \in \mathcal{I}_N$, where $\mathcal{I}_N$ is as in Theorem 3). Furthermore, for any γ, as $|x|^\gamma \geq \min\{|x|, |x|^\infty\}$, it is easy to check from the definition that

$$E_{\hat{Q}_N}[\xi] \leq \mathcal{E}^{k_N,\gamma}_{\mathcal{Q}|\mathbf{x}_N}(\xi) \leq \max\Big\{\mathcal{E}^{k_N,1}_{\mathcal{Q}|\mathbf{x}_N}(\xi), \mathcal{E}^{k_N,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\xi)\Big\}.$$

The result follows. □

3.2. Parametric results

We now suppose that $\mathcal{Q}$ is a family of measures coming from a 'nice' parametric family. In this setting, we can obtain more precise asymptotics by considering the divergence as a function of the parameter, rather than as a function of the abstract space of probability measures. For simplicity, we shall consider an exponential family of measures, which is general enough for many applications, but gives sufficient structure to obtain tight results. We shall also assume throughout that, for every $Q \in \mathcal{Q}$, $X, \{X_n\}_{n\in\mathbb{N}}$ are iid with density $f(\cdot; Q)$.

Definition 8. *A distribution is said to come from an exponential family (in natural parameters) if the density can be written*

$$f(x;Q) = h(x) \exp\Big\{\langle \theta, T(x)\rangle - A(\theta)\Big\}.$$

Here θ is the parameter of Q, and is in an open subset Θ of $\mathbb{R}^d$ for some d. The exponential family is then defined by the quantities Θ, T and A, where T is a (vector of) sufficient statistics, A is the log-partition function and h is a normalization function (which can be deduced from T, A).

We write θ_Q for the parameters of Q, Q^θ for the measure associated with θ, and E_θ for E_{Q^θ}, etc...

For any exponential family, it is standard that A is convex and C^3. We shall here make a slightly stronger assumption.

Assumption 2. *For a given exponential family (in particular a given log-partition function A and sufficient statistic T), we assume that*

- *(i) $\mathcal{Q}$ corresponds to the family of measures with parameters in an open set $\Theta \subseteq \mathbb{R}^d$.*
- *(ii) The Hessian $\mathfrak{I}_\theta = \partial^2 A(\theta)$ (commonly known as the information matrix) is (strictly) positive definite at every point of Θ.*
- *(iii) The $\mathcal{Q}$-MLE exists and is consistent, with probability tending to 1 as $N \to \infty$ (that is, for every $Q \in \mathcal{Q}$, a maximizer $\hat{Q}_N$ exists with Q-probability approaching 1 and $\hat{\theta}_N = \theta_{\hat{Q}_N} \to_Q \theta_Q$).*

These assumptions can be justified using weak assumptions on the family considered, see for example Berk [6, Theorem 3.1], Silvey [19] or the more general discussion of Lehmann [15] (see also [16]). For more advanced discussion of the theory of likelihood in exponential families, see Barndorff-Nielsen [2]. Many standard distribution families are exponential families satisfying these assumptions (e.g. Gaussian, log-Gaussian, exponential, gamma, beta and Dirichlet distributions, with their usual parameterizations).

Under this assumption, whenever the $\mathcal{Q}$-MLE $\hat{\theta}_N$ exists, we observe that the divergence is given by

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) = -\sum_{n=1}^{N} \langle \theta - \hat{\theta}_N, T(X_i)\rangle + N\big(A(\theta) - A(\hat{\theta}_N)\big),$$

using the natural abuse of notation $\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) := \alpha_{\mathcal{Q}|\mathbf{x}_N}(Q^\theta)$. Given a first order condition will hold at the MLE, we can simplify to remove dependence on the observations (except through the MLE)

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) = N\big(A(\theta) - A(\hat{\theta}_N) - \langle \theta - \hat{\theta}_N, \partial A(\hat{\theta}_N)\rangle\big).$$

Remark 11. Under Assumption 2, $\theta \mapsto \alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta)$ is continuous and measurable. If we also know $\theta \mapsto E_\theta[\xi]$ is upper semicontinuous (which is easy to check for bounded ξ), then the supremum in (3) can be taken over a countable dense subset of Θ, which implies immediately that the DR-expectation is a measurable function of $\mathbf{x}$.

The following result will allow us to get a tight asymptotic approximation of the penalty, as it will allow us to focus our attention on a small ball around the MLE.

Lemma 3. *Let $\rho > 0$ be a constant and let $\hat{\theta}_N$ denote the MLE of θ (when it exists, and an arbitrary point in Θ otherwise). Then, for each $P \in \mathcal{Q}$, there exist constants c_1, c_2 independent of N such that, writing*

$$R = \frac{c_1 \rho}{N} \vee \sqrt{\frac{c_2 \rho}{N}} = O(N^{-1/2}),$$

we have that

$$P\big(\alpha_{\mathcal{Q}|\mathbf{x}}(\theta) > \rho \text{ for all } \theta : \|\theta - \hat{\theta}_N\| > R\big) \to 1.$$

In other words, with high probability, we know $\alpha_{\mathcal{Q}|\mathbf{x}}(\theta) > \rho$ whenever $\|\theta - \hat{\theta}\| > R = O(N^{-1/2})$.

Proof. See Appendix. □

Remark 12. The previous result will mainly be used to show that, when we consider bounded random variables, for any $P \in \mathcal{Q}$ we can approximate the divergence by

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) = \frac{N}{2}(\theta - \hat{\theta}_N)^\top \Big[\mathfrak{I}_{\hat{\theta}_N} + O_P(N^{-1/2})\Big](\theta - \hat{\theta}_N).$$

This is itself an interesting and useful result, particularly when we use the DR-expectation approach as a first step in a larger problem. For example, when we use a DR-expectation to capture the uncertainty in calibration of a model, which we then wish to use in a variety of settings this result shows that it is enough (to first order) to penalize using the observed information matrix, rather than repeatedly calculating the likelihood function. This is the approximation we made in (1).

As the approximation is a quadratic, the optimization needed to calculate $\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}_N}$ is straightforward (particularly for linear or quadratic functionals of the parameters), which can have significant numerical advantages (see for example Ben-Tal and Nemirovski [4]).

We now use this approximation to give asymptotic estimates for the DR-expectation. This can be seen as an analogue to the central limit theorem (cf. Example 2). Note that, unlike in the nonparametric case, we do not need to scale the risk aversion parameter k as $N \to \infty$ to obtain a second order approximation. It is convenient to make the following definition.

Definition 9. *Let ϕ be a bounded function such that the map $\tilde{\phi} : \theta \mapsto E_\theta[\phi(X)]$ is differentiable. We write*

$$V(\phi, \hat{\theta}) := (\partial\tilde{\phi}|_{\hat{\theta}})^\top (\mathfrak{I}_{\hat{\theta}}^{-1})(\partial\tilde{\phi}|_{\hat{\theta}}).$$

Remark 13. Observe that, by classical arguments, if ϕ can be written as a linear function of the sufficient statistics then

$$V(\phi, \hat{\theta}) = \mathrm{Var}_{\hat{\theta}}(\phi(X)).$$

If $\hat{\theta}_N$ has the variance appearing in the central limit theorem, that is[9], $\mathrm{Var}(\hat{\theta}_N) \approx N^{-1}\mathfrak{I}_{\theta_P}^{-1}$, then (given an appropriate array of integrability and continuity assumptions) we have the approximate variance of the MLE-expectation

$$\frac{1}{N} V(\phi, \hat{\theta}_N) \approx \mathrm{Var}_P(E_{\hat{\theta}_N}[\phi(X)]).$$

Theorem 4. *Let ϕ be a bounded function such that the map $\tilde{\phi} : \theta \mapsto E_\theta[\phi(X)]$ is twice differentiable. Then for all $P \in \mathcal{Q}$,*

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = E_{\hat{\theta}_N}[\phi(X)] + \frac{k}{2N} V(\phi, \hat{\theta}_N) + O_P(N^{-3/2}).$$

Proof. Fix $P \in \mathcal{Q}$. For simplicity, we write $\hat{\theta}$ for $\hat{\theta}_N$. To begin, observe that

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = \sup_{\theta \in \Theta} \Big\{ E_{Q^\theta}[\phi(X)] - \frac{1}{k}\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) \Big\}$$

and as ϕ is bounded, we need only consider those measures

$$\Theta_N = \Big\{ \theta \in \Theta : \alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) \le k \sup_x |\phi(x)| \Big\}.$$

From Lemma 3, we know that

$$P\Big(\sup_{\theta \in \Theta_N} \|\theta - \hat{\theta}\| > O(N^{-1/2}) \Big) \to 0.$$

We know $\hat{\theta} \to_P \theta_P$ and $\tilde{\phi}$ is twice differentiable at θ_P, so for $\theta \in \Theta_N$,

$$\begin{aligned} E_{Q^\theta}[\phi(X)] &= \tilde{\phi}(\hat{\theta}) + \Big\langle \theta - \hat{\theta}, \partial\tilde{\phi}|_{\hat{\theta}} + O_P(\|\theta - \hat{\theta}\|) \Big\rangle \\ &= \tilde{\phi}(\hat{\theta}) + \Big\langle \theta - \hat{\theta}, \partial\tilde{\phi}|_{\hat{\theta}} + O_P(N^{-1/2}) \Big\rangle \end{aligned}$$

We also know that $\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta)$ is smooth, convex and minimized at $\hat{\theta}$, so for $\theta \in \Theta_N$,

$$\begin{aligned} \alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) &= \frac{N}{2}(\theta - \hat{\theta})^\top \Big[\mathfrak{I}_{\hat{\theta}} + O_P(\|\theta - \hat{\theta}\|) \Big] (\theta - \hat{\theta}) \\ &= \frac{N}{2}(\theta - \hat{\theta})^\top \Big[\mathfrak{I}_{\hat{\theta}} + O_P(N^{-1/2}) \Big] (\theta - \hat{\theta}). \end{aligned}$$

[9]See Lehmann [15, Section 7.7] for one set of sufficient conditions under which this holds.

Substituting these, we have the approximate DR-expectation

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = \tilde{\phi}(\hat{\theta}) + \sup_{\theta\in\Theta_N}\Big\{\Big\langle \theta-\hat{\theta}, \partial\tilde{\phi}|_{\hat{\theta}} + O_P(N^{-1/2})\Big\rangle - \frac{N}{2k}(\theta-\hat{\theta})^\top\Big[\mathfrak{I}_{\hat{\theta}} + O_P(N^{-1/2})\Big](\theta-\hat{\theta})\Big\}.$$

The term in braces has optimizer

$$\theta^* = \hat{\theta} + \frac{k}{N}\Big(\mathfrak{I}_{\hat{\theta}} + O_P(N^{-1/2})\Big)^{-1}\Big(\partial\tilde{\theta}|_{\hat{\theta}} + O_P(N^{-1/2})\Big),$$

where we know that, as $\hat{\theta}\to\theta_P$ and $\mathfrak{I}_{\theta_P}$ is positive definite, with P-probability approaching 1 the matrix $\mathfrak{I}_{\hat{\theta}} + O_P(N^{-1/2})$ is nonsingular. Substituting, we have the desired approximation

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = \tilde{\phi}(\hat{\theta}) + \frac{k}{2N}(\partial\tilde{\phi}|_{\hat{\theta}})^\top(\mathfrak{I}^{-1}_{\hat{\theta}})(\partial\tilde{\phi}|_{\hat{\theta}}) + O_P(N^{-3/2}). \qquad \square$$

We now consider the case $\gamma=\infty$.

Theorem 5. *Let ϕ be a bounded function such that the map $\tilde{\phi}:\theta\mapsto E_{Q^\theta}[\phi(X)]$ is twice differentiable. Then for all $P\in\mathcal{Q}$,*

$$\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = E_{\hat{\theta}_N}[\phi(X)] + \sqrt{\frac{2k}{N}V(\phi,\hat{\theta}_N)} + O_P(N^{-3/4}).$$

Proof. The proof follows much in the same way as the case $\gamma=1$ and we use the same notation. We know that

$$\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\phi(X)) = E_{\hat{\theta}}[\phi(X)] + \sup_{\{\theta:\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta)\le k\}}\Big\langle\theta-\hat{\theta}, \partial\tilde{\phi}|_{\hat{\theta}} + O_P(\|\theta-\hat{\theta}\|)\Big\rangle.$$

We see that

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(\theta) = \frac{N}{2}(\theta-\hat{\theta})^\top\Big[\mathfrak{I}_{\hat{\theta}} + O_P(\|\theta-\hat{\theta}\|)\Big](\theta-\hat{\theta})$$

and from Lemma 3, with probability approaching 1, it is enough to consider $\Theta_N = \{\theta : \|\theta-\hat{\theta}\| < O_P(N^{-1/2})\}$. Standard optimization then yields

$$\begin{aligned}
&\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}}(\phi(X)) - E_{\hat{\theta}}[\phi(X)]\\
&= \sqrt{\frac{k}{2N}}\Big((\partial\tilde{\phi}|_{\hat{\theta}} + O_P(N^{-1/2}))^\top\Big[\mathfrak{I}_{\hat{\theta}} + O_P(N^{-1/2})\Big]^{-1}(\partial\tilde{\phi}|_{\hat{\theta}} + O_P(N^{-1/2}))\Big)^{1/2}\\
&= \sqrt{\frac{k}{2N}}\Big((\partial\tilde{\phi}|_{\hat{\theta}})^\top\mathfrak{I}^{-1}_{\hat{\theta}}(\partial\tilde{\phi}|_{\hat{\theta}})\Big)^{1/2} + O_P(N^{-3/4}).
\end{aligned}$$

The result follows. $\square$

Remark 14. The cases $\gamma\in(1,\infty)$ can also be treated using the approximation implied by Lemma 3 (in the way suggested by Remark 12), and are left as a tedious exercise for the reader.

For a more general result, we recall the following theorem (named for Wilks [22]) from classical likelihood theory:

Theorem 6 (Wilks' theorem)**.** *Suppose $P \in \mathcal{Q}$, the family of densities corresponding to $\mathcal{Q}$ satisfies appropriate regularity conditions*[10] *and $\mathcal{Q}$ is parameterized by an open subset of $\mathbb{R}^d$. Then*

$$\frac{1}{2}\alpha_{\mathcal{Q}|\mathbf{x}}(P) \to_{P\text{-Dist}} \chi^2_d$$

where P-Dist refers to convergence, in distribution, under P.

Theorem 7. *Suppose the MLE is consistent and Wilks' theorem holds under every $P \in \mathcal{Q}$. Then, for a random variable ξ (which may be unbounded and depend on the observations $\mathbf{x}_N$)*

$$\mathcal{I}_N(\xi) = \Big[-\mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(-\xi), \quad \mathcal{E}^{k,\infty}_{\mathcal{Q}|\mathbf{x}_N}(\xi)\Big]$$

is a likelihood interval for $E[\xi(\omega, \mathbf{x}_N)]$, with the uniform asymptotic property

$$\lim_{N\to\infty} P\Big(E_P[\xi(\omega,\mathbf{x}_N)] \in \mathcal{I}_N \textit{ for all } \xi\Big) \geq F_{\chi^2_d}(2k).$$

Proof. That $\mathcal{I}_N(\xi)$ corresponds to a likelihood interval is trivial, as $\gamma = \infty$ implies we are considering expectations under those measures where the log likelihood (relative to the MLE) is at least k. Wilks' theorem then determines the asymptotic behaviour of the relative log likelihood, in particular, we know

$$P\big(\alpha_{\mathcal{Q}|\mathbf{x}_N}(P) \leq k\big) \to F_{\chi^2_d}(2k) \qquad \text{for all } P \in \mathcal{Q},$$

where $F_{\chi^2_d}$ is the cdf of the χ^2_d-distribution. Clearly $\alpha_{\mathcal{Q}|\mathbf{x}_N}(P) \leq k$ implies $E_P[\xi|\mathbf{x}_N] \in \mathcal{I}_N(\xi)$ for all ξ. We then obtain the desired result,

$$P\Big(E_P[\xi(\omega,\mathbf{x}_N)] \in \mathcal{I}_N \text{ for all } \xi\Big) \geq P\big(\alpha_{\mathcal{Q}|\mathbf{x}_N}(P) \leq k\big) \to F_{\chi^2_d}(2k). \qquad \square$$

4. Robustness and models

In this section, we shall consider the behaviour of the divergence-robust expectation for *unbounded* random variables, and its relationship with 'robust' statistical estimates. We shall regard the sample size N as fixed. The following theorem complements our earlier asymptotic results (which were generally for bounded outcomes), to demonstrate that without any parametric structure most unbounded random variables do not have finite DR-expectations.

[10] Conditions for Wilks' theorem are related to those needed for the MLE to satisfy a central limit theorem (as in the Wald test), and are typically based on integrability and differentiability assumptions on the densities (which do not need to come from an exponential family). These hold for the majority of classical statistical models. See Lehmann [15, Section 7.7] for full details (in particular, this result is a special case of [15, Theorem 7.7.4]).

Theorem 8. *Let $\mathcal{Q}$ be a family of measures such that $\mathcal{Q}$ is closed under taking finite mixtures (i.e. finite convex combinations of measures). Then for any random variable ξ such that $\sup_{Q\in\mathcal{Q}}\{E_Q[\xi]\} = \infty$, for any $\gamma \in [1,\infty]$,*

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi) = \infty.$$

Proof. For any $\epsilon > 0$, let $Q^\epsilon \in \mathcal{Q}$ be a measure such that $E_{Q^\epsilon}[\xi] > \epsilon^{-2}$. Fix $P \in \mathcal{Q}$ such that $\alpha_{\mathcal{Q}|\mathbf{x}}(P) \leq k$. We define the mixture distribution $P(\epsilon) = (1-\epsilon)P + \epsilon Q^\epsilon$. It follows that $P(\epsilon) \in \mathcal{Q}$ and, provided $E_P[\xi] > -\infty$,

$$E_{P(\epsilon)}[\xi] = (1-\epsilon)E_P[\xi] + \epsilon E_{Q^\epsilon}[\xi] \geq (1-\epsilon)E_P[\xi] + \epsilon^{-1} \to \infty$$

as $\epsilon \to 0$. Also, we know

$$L(P(\epsilon)|\mathbf{x}) = (1-\epsilon)L(P|\mathbf{x}) + \epsilon L(Q^\epsilon|\mathbf{x}) > (1-\epsilon)L(P|\mathbf{x}), \tag{6}$$

so (assuming for notational simplicity that the $\mathcal{Q}$-MLE $\hat{Q}$ exists)

$$\alpha_{\mathcal{Q}|\mathbf{x}}(P(\epsilon)) = -\log\Big(\frac{L(P(\epsilon)|\mathbf{x})}{L(\hat{Q}|\mathbf{x})}\Big) < -\log\Big(\frac{(1-\epsilon)L(P|\mathbf{x})}{L(\hat{Q}|\mathbf{x})}\Big) \to \alpha_{\mathcal{Q}|\mathbf{x}}(P) \leq k.$$

It follows that, as $\epsilon \to 0$,

$$\mathcal{E}^{k,\gamma}_{\mathcal{Q}|\mathbf{x}}(\xi) \geq (1-\epsilon)E_P[\xi] + \epsilon^{-1} - \Big(\frac{1}{k}\log\Big(\frac{(1-\epsilon)L(P|\mathbf{x})}{L(\hat{Q}|\mathbf{x})}\Big)\Big)^\gamma \to \infty. \qquad \square$$

Remark 15. The above assumes $\mathcal{Q}$ is closed under finite mixtures of measures. If we assume that $\mathcal{Q}$ is such that $\{X_n\}_{n\in\mathbb{N}}$ are iid, then this is not the case. However, for $N < \infty$, an almost identical proof holds whenever $\mathcal{Q}$ corresponds to a family of densities $f(\cdot;Q)$, and this family of densities is closed under taking finite mixtures. (The only significant change is that we obtain the inequality $L(P(\epsilon)|\mathbf{x}) > (1-\epsilon)^N L(P|\mathbf{x})$ in the place of (6).)

This result highlights the importance of parametric structure for estimation of unbounded random variables, in terms of restricting the class of probability measures that can be considered. This restriction can be thought of in terms of restricting the probabilities of very large (positive or negative) values of ξ, and hence ensuring enough integrability that finite expectations arise. Without these restrictions, unlikely events (which by their very nature will generally not be seen in the data, so are not penalized) result in unbounded expectations.

Given the importance of parametric families, it is then of interest to consider how the 'statistical robustness' of the parametric estimation problem interacts with the 'robustness' of the expectations considered. Before giving general results, we consider a simple setting.

Example 5. Consider $X, \{X_n\}_{n=1}^N$ iid observations from a Laplace (or double exponential) distribution with known scale 1 and unknown mean μ. That is, X_n

has density

$$f(x) = \frac{1}{2}\exp(-|x-\mu|).$$

Let $\mathcal{Q}$ denote the corresponding family of measures and write Q^μ for the measure with mean μ. For simplicity, assume N is odd, so the MLE is uniquely given by $Q^{\mathbf{m}}$, where $\mathbf{m}$ is the sample median. This is known to be 'statistically robust', see Huber and Ronchetti [11], as it does not depend on extreme observations, and is therefore unaffected by outliers.

The $\mathcal{Q}|\mathbf{x}_N$-divergence is then given by

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q^\mu) = \sum_{n=1}^N \Big(|X_n-\mu| - |X_n - \mathbf{m}|\Big).$$

For X an iid observation from the same distribution as X_n and $\beta > 0$ (the case $\beta < 0$ is symmetric) we have

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X) = \sup_\mu \Big\{\beta\mu - \frac{1}{k}\sum_{n=1}^N \Big(|X_n-\mu| - |X_n-\mathbf{m}|\Big)\Big\}.$$

A first observation which can be drawn is that $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X)$ is generally infinite, unless $\beta < N/k$. To see this, observe that if $\beta > N/k$, then the function to maximize is linear and increasing for $\mu > \max_{n\leq N} X_n$.

Assuming that $\beta < N/k$, the function to maximize is piecewise linear, concave and asymptotically decreasing (for both positive and negative μ), so a finite solution exists. Except at points where $\mu = X_n$ for some n, we can differentiate to obtain the equation

$$0 = \beta - \frac{1}{k}\sum_{n=1}^N (I_{\{X_n<\mu\}} - I_{\{X_n>\mu\}}) =: \beta - \frac{N}{k}G(\mu).$$

As we are looking for the maximal solution, we can generally state that the solution will be

$$\mu^* = \inf\Big\{\mu : \beta - \frac{N}{k}G(\mu) > 0\Big\}.$$

We can also write

$$G(\mu) = (1-F(\mu)) - F(\mu-)$$

where $F(y) = \frac{1}{N}\sum_{n=1}^N I_{\{X_n\leq y\}}$ is the empirical cdf of our observations. Assuming N is moderately large, this is well approximated by a continuous increasing function (so all quantiles are uniquely defined), and we obtain

$$\mu^* \approx F^{-1}\Big(\frac{1}{2} + \frac{\beta k}{2N}\Big).$$

It follows that the optimizing choice of μ^* is given by the empirical $\frac{1}{2} + \frac{\beta k}{2N}$ quantile.

Introducing this back into our equation for $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$, we obtain

$$\begin{aligned}
\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X) &= \beta\mu^* - \frac{1}{k}\sum_{n=1}^{N}\Big(|X_n-\mu^*| - |X_n - \mathbf{m}|\Big) \\
&= \beta\mu^* - (\mu^* - \mathbf{m})\sum_{n=1}^{N}(I_{\{X_n\leq \mathbf{m}\}} - I_{\{X_n>\mu^*\}}) \\
&\qquad + \sum_{n=1}^{N} I_{\{\mathbf{m}<X_n\leq\mu^*\}}\Big(\frac{\mu^*+\mathbf{m}}{2} - X_n\Big) \\
&\approx \beta\mu^* - \Big(\mu^* - \mathbf{m} - \frac{\mu^*+\mathbf{m}}{2}\Big)\frac{k\beta}{2N} - \sum_{n=1}^{N} I_{\{\mathbf{m}<X_n\leq\mu^*\}}X_n \\
&\approx \Big(1-\frac{k}{4N}\Big)\beta\mu^* + \frac{3k}{4N}\beta\mathbf{m} - \frac{k}{2N}\frac{\sum_{n=1}^{N} I_{\{\mathbf{m}<X_n\leq\mu^*\}}(\beta X_n)}{\sum_i I_{\{\mathbf{m}<X_n\leq\mu^*\}}}.
\end{aligned}$$

We see that the divergence-robust estimate depends on a weighted combination of the median $\beta\mathbf{m}$, an upper quantile $\beta\mu^*$, and the mean value taken between these two bounds[11]. Therefore, this quantity can still be robustly estimated, as it still does not depend on the tail behaviour beyond the $\frac{1}{2}+\frac{\beta k}{2N}$ quantile. More formally, the breakdown point of this estimator (the proportion of the data which can be made arbitrarily large without affecting the estimate) is $\frac{1}{2}(1-\beta\frac{k}{N})$.

It is easy to see (as $E_{Q^\mu}[\beta X^2] = \beta\mu^2 + 2\beta$) that

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X^2) = \infty$$

for all $N, k, \beta > 0$. For negative β, a finite answer can be obtained, but even its approximate closed-form representation is inelegant.

Comparing this example with the normal example (Example 2), we can see that, when considering a likelihood model, there is a delicate relationship between the 'statistical' robustness in the classical estimation problem and the 'parameter uncertainty' robustness embedded in $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}$. We now seek to make this behaviour more precise, for the general setting of a family of measures $\mathcal{Q}$ describing an uncertain 'location parameter'.

Assumption 3. *Suppose that under $Q \in \mathcal{Q}$, we know $X, \{X_n\}_{n=1}^N$ are iid observations from a distribution with density $\exp(\Psi(x-\mu_Q))$, and so $\mathcal{Q}$ is parameterized by $\mu_Q = E_Q[X] \in \mathbb{R}$. Suppose Ψ has monotone increasing derivative ψ (which may be discontinuous) and*

$$-\infty \leq \lim_{x\to-\infty}\psi(x) < 0 < \lim_{x\to+\infty}\psi(x) \leq \infty.$$

Note that the MLE parameter (assuming it exists, for simplicity) is given by the solution μ to $\sum_n \psi(X_n-\mu) = 0$.

[11] This estimate can then be compared with the various perturbations of value-at-risk considered by Cont, Deguest and Scandolo [7]. However, it is important to note that this closed-form is only for the random variables βX, not for a general random variable.

Definition 10. *An estimator $\hat{\theta}$ defined by an implicit equation of the form*

$$\sum_{n=1}^{N} \psi(X_i; \hat{\theta}) = 0$$

is called an M-estimate.

The following result is given by Chapter 3 of Huber and Ronchetti [11], in particular Theorem 3.6, and addresses the 'statistical robustness' of an M-estimate.

Theorem 9. *In the setting of Assumption 3, the following are equivalent.*

(i) The MLE parameter has a breakdown point above zero (that is, some fraction of the observations can be made arbitrarily large or small without making the MLE arbitrarily large or small),
(ii) The MLE parameter is weakly continuous with respect to the empirical cdf of observations, for any empirical cdf where the MLE parameter is uniquely defined,
(iii) ψ is bounded.

Remark 16. The classic example where this result holds is Example 5, where the MLE parameter is given by the sample median.

Theorem 10. *In the setting of Assumption 3, the conditions of Theorem 9 are equivalent to*

(iv) For any fixed k, N, for all $\beta \in \mathbb{R}$ sufficiently large (in absolute value), for X an iid copy of the observations,

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X) \notin \mathbb{R}.$$

Proof. We seek to show Theorem 9(iii) and Theorem 10(iv) are equivalent. First, if (iii) holds, then Ψ is of linear growth. Let $\beta > N \sup_x |\psi(x)|$. We can then calculate

$$\mathcal{E}_{\mathcal{Q}|\mathbf{x}}(\beta X) = \sup_{\mu_Q} \Big\{ \beta\mu_Q - \sum_n \Big(\Psi(X_n - \mu_Q) - \Psi(X_n - \mu_{\mathrm{MLE}}) \Big) \Big\}.$$

As β is larger than the maximal derivative of $\sum_n \Psi(X_n - \mu_Q)$, we can see that the term in brackets is unbounded, so (iv) holds. A similar result holds if $\beta < -N \sup_x |\psi(x)|$.

To show (iv) implies (iii), we first observe that (iv) implies that for all β sufficiently large,

$$\sup_{\mu_Q} \Big\{ \beta\mu_Q - \sum_n \Big(\Psi(X_n - \mu_Q) - \Psi(X_n - \mu_{\mathrm{MLE}}) \Big) \Big\} = \infty.$$

In other words, Ψ is bounded above by a linear function. As $\psi = \Psi'$ is monotone increasing, this implies that ψ is bounded above. A similar argument shows that ψ is bounded below. $\square$

Conversely, we have the following.

Theorem 11. *In the setting of Assumption 3, write*

$$\psi(\infty) = \lim_{x\to\infty} \psi(x), \qquad \psi(-\infty) = \lim_{x\to-\infty} \psi(x).$$

If $\psi(-\infty) < -|\beta|/N$ and $|\beta|/N < \psi(\infty)$, then the nonlinear expectation $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ is finite, and has breakdown fraction at least

$$\delta = \min\Big\{\frac{\psi(\infty) - |\beta|/N}{\psi(\infty) - \psi(-\infty)}, \frac{-|\beta|/N - \psi(-\infty)}{\psi(\infty) - \psi(-\infty)}\Big\}.$$

That is, for any $m < \delta N$, at least m observations can be made arbitrarily large or small while $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ remains bounded.

Proof. Consider the function

$$\lambda^\beta(\mu, \mathbf{x}) := \frac{\beta}{N} - \frac{1}{N}\sum_{n=1}^N \psi(X_n - \mu) \tag{7}$$

From a first order condition, the value of $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X)$ is given by

$$\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X) = \beta\mu^\beta(\mathbf{x}) - \sum_{n=1}^N \Big(\Psi(X_n - \mu^\beta(\mathbf{x})) - \Psi(X_n - \mu_{\mathrm{MLE}}(\mathbf{x}))\Big)$$

where $\mu^\beta(\mathbf{x})$ is the solution to $\lambda^\beta(\mu, \mathbf{x}) \risingdotseq 0$ (where $\risingdotseq$ indicates either equality or a change of sign, if λ^β jumps over zero) and $\mu_{\mathrm{MLE}}(\mathbf{x})$ is the MLE based on $\mathbf{x}$.

Now observe that λ^β is monotone increasing with respect to μ and, as we have assumed $\psi(-\infty) < \beta/N < \psi(\infty)$, we know $\lim_{\mu\to\infty} \lambda^\beta(\mu, \mathbf{x}) > 0$ and $\lim_{\mu\to-\infty} \lambda^\beta(\mu, \mathbf{x}) < 0$. Therefore, there is exactly one (finite) solution to $\lambda^\beta(\mu, \mathbf{x}) \risingdotseq 0$. It follows that $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X)$ exists and is real.

We now need to determine the breakdown fraction. For M a set of indices, let $\mathbf{x}(M, y)$ denote the set of observations, with X_i replaced by y_i for $i \in M$. Suppose $|M| = m$ and $m/N < \delta$. We wish to show that there is a bound on $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ which is uniform in y. From the definition and nonnegativity of the penalty function, it is easy to see that

$$\beta\mu^{-|\beta|}(\mathbf{x}(M, y)) \le \mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X) \le \beta\mu^{|\beta|}(\mathbf{x}(M, y)),$$

it follows that it is enough for us to prove that $\mu^{\pm|\beta|}(\mathbf{x}(M, y))$ is uniformly bounded in y.

As ψ is monotone, we observe that

$$\begin{aligned}
&\frac{\beta}{N} - \frac{1}{N}\sum_{n\notin M} \psi(X_n - \mu) - \frac{m}{N}\psi(\infty) \\
&\le \lambda^\beta(\mu, \mathbf{x}(M, y)) \\
&\le \frac{\beta}{N} - \frac{1}{N}\sum_{n\notin M} \psi(X_n - \mu) - \frac{m}{N}\psi(-\infty).
\end{aligned}$$

By monotonicity, it is enough to show that the terms on the right and left have roots for finite values of μ (as these will not depend on y). Considering the lower bound first, we see that as $\mu \to \infty$, we obtain

$$\frac{\beta}{N} - \frac{1}{N}\sum_{n\notin M}\psi(X_n-\mu) - \frac{m}{N}\psi(\infty) \to \frac{\beta}{N} - \Big(1-\frac{m}{N}\Big)\psi(-\infty) - \frac{m}{N}\psi(\infty) > 0,$$

and as $\mu \to -\infty$,

$$\frac{\beta}{N} - \frac{1}{N}\sum_{n\notin M}\psi(X_n-\mu) - \frac{m}{N}\psi(\infty) \to \frac{\beta}{N} - \psi(\infty) < 0.$$

Therefore, there is a finite root for the lower bound on $\lambda^\beta(\mu, \mathbf{x}(M,y))$. A similar argument applies to the upper bound, and when replacing β with $-\beta$. By monotonicity, we conclude that $\mu^{\pm|\beta|}(\mathbf{x}(M,y))$ and hence $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ are uniformly bounded in y, as desired. $\square$

To conclude, we observe that the non-existence of finite values for $\mathcal{E}^{k,1}_{Q|\mathbf{x}}(X)$ can also manifest itself in surprising ways, as we can see from the following extension of Example 2.

Example 6. Consider the case where $X, \{X_i\}_{i=1}^N$ are iid $N(\mu,\sigma^2)$, where both μ and σ^2 are unknown. The divergence penalty is then (writing the MLE estimate of variance $\hat\sigma^2 = \frac{1}{N}\sum_{n=1}^N (X_n - \bar X)^2$)

$$\begin{aligned}\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q^{\mu,\sigma^2}) &= \frac{N}{2}\log(\sigma^2) + \sum_{n=1}^N \frac{(X_n-\mu)^2}{2\sigma^2} - \frac{N}{2}\log(\hat\sigma^2) - \sum_{n=1}^N\frac{(X_N-\bar x)^2}{2\hat\sigma^2}\\ &= \frac{N}{2}\Big(\log(\sigma^2/\hat\sigma^2) + \frac{\frac{1}{N}\sum_{n=1}^N (X_n-\mu)^2}{\sigma^2} - 1\Big).\end{aligned}$$

If we attempt to calculate $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}}(\beta X)$, we obtain

$$\begin{aligned}\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X) &= \sup_{\mu,\sigma^2}\Big\{\beta\mu - \frac{N}{2k}\Big(\log(\sigma^2/\hat\sigma^2) + \frac{\frac{1}{N}\sum_{n=1}^N (X_n-\mu)^2}{\sigma^2} - 1\Big)\Big\}\\ &= \sup_{\sigma^2}\Big\{\beta\bar X + \frac{\beta^2 k}{2N}\sigma^2 - \frac{N}{2k}\Big(\log(\sigma^2/\hat\sigma^2) - 1 + \frac{\hat\sigma^2}{\sigma^2}\Big)\Big\}.\end{aligned}$$

This causes a problem, as the term on the right is unbounded above with respect to σ^2. Looking more closely, this function typically has a local maximum for $\sigma^2 \approx \hat\sigma^2$, but for very large values of σ the $\frac{\beta^2 k}{2N}\sigma^2$ term will dominate. Therefore, there is no way that, even for large samples, a finite value of $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ can be obtained.

One possible way to deal with this is to modify our approach slightly, either by including a prior distribution[12], or by adding an additional regularizing term

[12] In this case, to ensure a finite answer, the prior distribution would need to be asymptotically exponentially small as $\sigma^2 \to \infty$, which is not the case for the conjugate inverse-Gamma distributions for σ^2. This renders explicit calculation difficult.

to ensure the supremum chooses values close to the statistical parameters. For example, taking the penalty

$$\tilde{\alpha}_{\mathcal{Q}|\mathbf{x}_N}(Q^{\mu,\sigma^2}) = \frac{N}{2}\Big(\log(\sigma^2/\hat{\sigma}^2) + \frac{\frac{1}{N}\sum_{n=1}^N (X_n-\mu)^2}{\sigma^2} - 1\Big) + \epsilon(\sigma^2 - \hat{\sigma}^2)$$

for some $\epsilon > 0$, results in a finite value for $\mathcal{E}^{k,1}_{\mathcal{Q}|\mathbf{x}_N}(\beta X)$ whenever $\frac{\beta^2 k}{2N} < \frac{\epsilon}{k}$, in particular one obtains consistent estimates as $N \to \infty$.

Appendix

Proof of Lemma 2. We know that

$$\alpha_{\mathcal{Q}|\mathbf{x}_N}(Q) = -\sum_{n=1}^N \log\Big(\frac{f(X_n;Q)}{f(X_n;P)}\Big) + \sum_{n=1}^N \log\Big(\frac{f(X_n;\hat{Q}_N)}{f(X_n;P)}\Big).$$

Considering the first term, by translation and scaling, we can assume that $K = [0,1]$. For any $P \in \mathcal{Q}$, write $F_P(x) = \int_0^x f(y;P)dy$ for the distribution function associated with P. We know that F_P^{-1} is C^1 with a norm on its derivative independent of P, by assumption (ii). Next observe that the natural logarithm is C^∞ on $[C^{-1}, C]$, so by standard results on composition of functions, the map

$$u \mapsto \ell(u;Q,P) := \log\Big(\frac{f(F_P^{-1}(u),Q)}{f(F_P^{-1}(u),P)}\Big)$$

is ρ-Hölder continuous, with a norm independent of P, Q. Working under P, we note that $U_n = F_P(X_n)$ are independent and uniformly distributed on $[0,1]$, and $E_P[\ell(U;Q,P)] = D_{\mathrm{KL}|X}(P||Q)$.

By rescaling, we can assume, without loss of generality, that for every $P, Q \in \mathcal{Q}$,

$$\ell(\cdot;Q,P) \in \mathbf{F}_\rho = \{g : |g(u)-g(v)| \leq |u-v|^\rho \text{ for all } u,v \in [0,1]\}.$$

It is enough, therefore, to prove a uniform convergence rate for functions in $\mathbf{F}_\rho$.

We can now appeal to Corollary 17.3.3 and the proof of Theorem 17.3.1 of Shorack and Wellner [18, p.633] (itself based on Strassen and Dudley [20]) to see that, writing

$$\sqrt{N}\Big(\frac{1}{N}\sum_n g(U_n) - E[g(U)]\Big) =: Z_N(g),$$

we know that for any $\eta > 0$, there is M sufficiently large (independent of N) that

$$P\Big(\sup_{g\in\mathbf{F}_\rho} \|Z_N(g)\| > M\Big) \leq \eta.$$

(The usual purpose of this is as a step towards showing that Z_N converges weakly to a Gaussian process, which is a form of Donsker's theorem.) By rearranging, it follows that

$$\gamma(N) := \sup_{Q\in\mathcal{Q}} \Big\{\Big| -\frac{1}{N}\sum_n \log\Big(\frac{f(X_n;Q)}{f(X_n;P)}\Big) - D_{\mathrm{KL}|X}(P||Q)\Big|\Big\} = O_P(N^{-1/2}) \quad (8)$$

In particular, we know that $\hat{Q}_N$ takes values in $\mathcal{Q}$, so,

$$\Big| -\frac{1}{N}\sum_{n=1}^{N} \log\Big(\frac{f(X_n;\hat{Q}_N)}{f(X_n;P)}\Big) - D_{\mathrm{KL}|X}(P||\hat{Q}_N)\Big| = O_P(N^{-1/2}). \quad (9)$$

From the definition of $\hat{Q}_N$ we see that

$$\begin{aligned}
0 &= \frac{1}{N}\sum_{n=1}^{N} \log\Big(\frac{f(X_n;P)}{f(X_n;P)}\Big)\\
&\leq \frac{1}{N}\sum_{n=1}^{N} \log\Big(\frac{f(X_n;\hat{Q}_N)}{f(X_n;P)}\Big) = \sup_{Q\in\mathcal{Q}} \frac{1}{N}\sum_{n=1}^{N} \log\Big(\frac{f(X_n;Q)}{f(X_n;P)}\Big)\\
&\leq -D_{\mathrm{KL}|X}(P||\hat{Q}_N) + \gamma(N).
\end{aligned}$$

Therefore,

$$0 \leq D_{\mathrm{KL}|X}(P||\hat{Q}_N) \leq \gamma(N) = O_P(N^{-1/2}). \quad (10)$$

The result then follows from using (8), (9) and (10) with the triangle inequality. □

Proof of Lemma 3. Our proof depends on three facts: that α is locally a quadratic to second order (via Taylor's theorem), that the MLE is consistent (allowing us to bound the third derivative with high probability), and that α is convex (which controls its global behaviour). We write $\hat{\theta}$ for $\hat{\theta}_N$ for notational simplicity.

As the MLE is consistent (and exists with high probability), as $N \to \infty$, for any radius $C > 0$, we know

$$P(\|\hat{\theta} - \theta^P\| < C/2) \to 1. \quad (11)$$

We also know that, for some constant k (which will in general depend on P and on C being sufficiently small, but is independent of N), we have the bound $\|\partial^3 A(\theta)\| \leq k$ for all θ with $\|\theta - \theta^P\| < C$. Combining these, for all θ with $\|\theta - \hat{\theta}\| < C/2$, from Taylor's theorem

$$\alpha_{\mathcal{Q}|\mathbf{x}}(\theta) \geq N\Big(\frac{1}{2}(\theta - \hat{\theta})^\top \mathfrak{I}_{\hat{\theta}}(\theta - \hat{\theta}) - k\|\theta - \hat{\theta}\|^3\Big).$$

As we know that $\mathfrak{I}_{\hat{\theta}}$ is not degenerate (uniformly in a neighbourhood of θ^P), we can also assume that (making k sufficiently large)

$$(\theta - \hat{\theta})^\top \mathfrak{I}_{\hat{\theta}}(\theta - \hat{\theta}) \geq \frac{1}{k}\|\theta - \hat{\theta}\|^2.$$

Therefore, taking $C \leq k^{-2}$, on the set $\|\theta - \hat{\theta}\| \leq C/2$ we have

$$\begin{aligned} \alpha_{\mathcal{Q}|\mathbf{x}}(\theta) &\geq N\Big(\frac{1}{k}\|\theta - \hat{\theta}\|^2 - k\|\theta - \hat{\theta}\|^3\Big) \\ &\geq N\Big(\frac{1}{k} - \frac{kC}{2}\Big)\|\theta - \hat{\theta}\|^2 = \frac{N}{k}\Big(1 - \frac{k^2 C}{2}\Big)\|\theta - \hat{\theta}\|^2 \\ &\geq \frac{N}{2k}\|\theta - \hat{\theta}\|^2 \end{aligned} \tag{12}$$

Note that k and C do *not* depend on N, so (11) remains valid.

We now need to extend the bound of (12) to all θ. We know that $\alpha_{\mathcal{Q}|\mathbf{x}}$ is convex and $\alpha_{\mathcal{Q}|\mathbf{x}}(\hat{\theta}) = 0$. For any point θ such that $\|\theta - \hat{\theta}\| > C/2$, its projection on the ball of radius $C/2$ around $\hat{\theta}$ is given by

$$\theta_\pi = \hat{\theta} + \lambda(\theta - \hat{\theta}) := \hat{\theta} + \frac{C}{2\|\theta - \hat{\theta}\|}(\theta - \hat{\theta}).$$

Hence, from (12), we know that

$$\alpha_{\mathcal{Q}|\mathbf{x}}(\theta) \geq \frac{1}{\lambda}\alpha_{\mathcal{Q}|\mathbf{x}}(\theta_\pi) \geq \frac{1}{\lambda}\frac{N}{2k}\|\theta_\pi - \hat{\theta}\|^2 = \frac{N}{Ck}\|\theta - \hat{\theta}\|. \tag{13}$$

Combining (12) and (13), we know that

$$\alpha_{\mathcal{Q}|\mathbf{x}}(\theta) \geq \Big(\frac{N}{2k}\|\theta - \hat{\theta}\|^2\Big) \wedge \Big(\frac{N}{Ck}\|\theta - \hat{\theta}\|\Big).$$

Now consider the set $\{\theta : \alpha_{\mathcal{Q}|\mathbf{x}}(\theta) \leq \rho\}$. We know that for all θ in this set,

$$\Big(\frac{N}{2k}\|\theta - \hat{\theta}\|^2\Big) \wedge \Big(\frac{N}{Ck}\|\theta - \hat{\theta}\|\Big) \leq \rho$$

which implies

$$\|\theta - \hat{\theta}\| \leq \frac{Ck\rho}{N} \vee \sqrt{\frac{2k\rho}{N}} =: R. \qquad \square$$

References

[1] Maurice Allais. Le comportement de l'homme rationel devant le risque: critique des postulats et axiomes de l'ecole Americaine. *Econometrica*, 21(4):503–546, 1953. MR0058952

[2] Ole E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley and Sons, 1978. (reprint 2014). MR0489333

[3] Pauline Barrieu and Nicole El Karoui. Inf-convolution of risk measures and optimal risk transfer. *Finance and Stochastics*, 9(2):269–298, 2005. MR2211128

[4] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4), 1998. MR1662410

[5] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009. MR2546839

[6] Robert H. Berk. Consistency and asymptotic normality of MLE's for exponential models. *The Annals of Mathematical Statistics*, 43(1):193–204, 1972. MR0298810

[7] Rama Cont, Romain Deguest, and Giacomo Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10(6):593–606, 2010. MR2676786

[8] Hans Föllmer and Alexander Schied. *Stochastic Finance: An introduction in discrete time*. Studies in Mathematics 27. de Gruyter, Berlin-New York, 2002. MR1925197

[9] Marco Frittelli and Emanuela Rosazza Gianin. Putting order in risk measures. *Journal of Banking & Finance*, 26(7):1473–1486, 2002.

[10] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989. MR1000102

[11] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2009. MR2488795

[12] D.J. Hudson. Interval estimation from the likelihood function. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(2):256–262, 1971. MR0315832

[13] John Maynard Keynes. *A Treatise on Probability*. Macmillan and Co., 1921. (Reprint by BN Publishing, 2008).

[14] Frank Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin, 1921. (Reprint by Dover publications 2006).

[15] E.L. Lehmann. *Elements of Large-Sample Theory*. Springer, 1999. MR1663158

[16] E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998. MR1639875

[17] J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. A*, 231:289–337, 1933.

[18] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. John Wiley and Sons, 1986. (Reprint by SIAM Classics in Applied Mathematics 2009). MR0838963

[19] S.D. Silvey. The Lagrangian multiplier test. *Ann. Math. Statist*, 30(2):389–407, 1959. MR0104307

[20] V. Strassen and R.M. Dudley. *The central limit thoerem and epsilon-entropy*, volume 89 of *Lecture Notes in Mathematics*, pages 224–231. Springer-Verlag, 1969. MR0279872

[21] Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945. MR0012402

[22] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 1938.