Measuring Distribution Model Risk*

Thomas Breuer Imre Csiszár

21 January 2013

Abstract

We propose to interpret distribution model risk as sensitivity of expected loss to changes in the risk factor distribution, and to measure the distribution model risk of a portfolio by the maximum expected loss over a set of plausible distributions defined in terms of some divergence from an estimated distribution. The divergence may be relative entropy, a Bregman distance, or an f-divergence. We give formulas for the calculation of distribution model risk and explicitly determine the worst case distribution from the set of plausible distributions. We also give formulas for the evaluation of divergence preferences describing ambiguity averse decision makers.

Keywords: multiple priors, model risk, ambiguity aversion, multiplier preferences, divergence preferences, stress tests, relative entropy, f-divergence, Bregman distance, maximum entropy principle, exponential family

JEL classification: D81, C44, C60, G01, G32, M48
AMS classification: 62C20, 90B50, 91B30, 94A17

*Thomas Breuer, PPE Research Centre, FH Vorarlberg, `thomas.breuer@fhv.at`. Imre Csiszár, Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, `csiszar.imre@mta.renyi.hu`. This work has been supported by the Hungarian National Foundation for Scientific Research under Grant K76088 and by the Austrian Forschungsförderungsgesellschaft in its Josef Ressel Center for Optimisation under Uncertainty. The second author has benefited from discussions with František Matúš.

1 The problem of model risk

Financial risk measurement, pricing of financial instruments, and portfolio selection are all based on statistical models. If the model is wrong, risk numbers, prices, or optimal portfolios are wrong. Model risk quantifies the consequences of using the wrong models in risk measurement, pricing, or portfolio selection.

The two main elements of a statistical model in finance are a risk factor distribution and a pricing function. Given a portfolio (or a financial instrument), the first question is: On which kind of random events does the value of the portfolio depend? The answer to this question determines the state space Ω.[1] A point $r \in \Omega$ is specified by a collection of possible values of the risk factors. The state space codifies our lack of knowledge regarding all uncertain events affecting the value of a given instrument or portfolio. The specification of a distribution class and some parameter estimation procedure applied to historical data determines some best guess risk factor distribution, call it $\mathbb{P}_0$. The second central element is a pricing function $X : \Omega \to \mathbb{R}$ describing *how* risk factors impact the portfolio value at some given future time horizon. We work in a one-stage set-up. Often modellers try to use only risk factors which are (derived from) prices of basic financial instruments. Describing the price of the portfolio as a function of the prices of these basic instruments is a modelling exercise, which is prone to errors. It involves asset pricing theories of finance with practically non-trivial assumptions on no arbitrage, complete markets, equilibrium, etc. Together the risk factor distribution and the pricing function determine the profit loss distribution. In a last step, a risk measure associates to the profit loss distribution a risk number describing a capital requirement.

Corresponding to the two central elements of a statistical model we distinguish two kinds of model risk: distribution model risk and pricing model risk. This paper is concerned with *distribution model risk*.[2] (For an interesting approach to pricing model risk we refer to Cont [2006].) Although $\mathbb{P}_0$ is a best guess of the risk factor distribution, one is usually aware that due to model specification errors or estimation errors the data generating process might be different from $\mathbb{P}_0$. Distribution model risk should quantify

[1]It is possible to choose a larger state space including variables which do not affect the value of the given portfolio. This could allow to compare different portfolios, which do not all depend on the same risk factors. Typically modellers try to keep the number of risk factors small and therefore use a smaller state space. With various techniques they try model some risk factors as a function of a smaller set of risk factors. Thus the number of risk factors actually used in the model, although it may go into the thousands, is typically much smaller than the number of variables influencing the loss.

[2]Gibson [2000] uses the term model risk for what we call distribution model risk. For a first classification of model risks we refer to Crouhy et al. [1998]. Distribution model risk encompasses both estimation risk and misspecification risk in the sense of Kerkhof et al. [2010], but here we do not need to distinguish the two.

the consequences of working with $\mathbb{P}_0$ instead of the true but unknown data generating process. We propose to measure distribution model risk by

$$\text{MR} := -\inf_{\mathbb{P}\in\Gamma} E_{\mathbb{P}}(X) \tag{1}$$

where Γ is some set of plausible alternative risk factor distributions. So MR is the negative of the worst expected value which could result if the risk factor distribution is some unknown distribution in Γ. We propose to choose for Γ balls of distributions, defined in terms of some divergence, centered at $\mathbb{P}_0$:

$$\Gamma = \{\mathbb{P} : D(\mathbb{P}\,||\,\mathbb{P}_0) \le k\}, \tag{2}$$

where the divergence D could be the relative entropy (synonyms: Kullback-Leibler distance, I-divergence), some Bregman distance, or some f-divergence. Γ contains all risk factor distributions $\mathbb{P}$ whose divergence from $\mathbb{P}_0$ is smaller than some radius $k > 0$. The parameter k has to be chosen by hand and describes the degree of uncertainty about the risk factor distribution. For larger values of k the set of plausible alternative distributions is larger, which is appropriate for situations in which there is more model uncertainty. In Section 3 we give the definitions of various divergences and discuss the choice of divergence D.

In a previous paper (Breuer and Csiszár [2012]) we have addressed the problem (1) for the special case where D is the relative entropy, assuming some regularity conditions which ensure the worst case distribution solving (1) is from some exponential family. The present paper first extends those results, giving the solution for the pathological cases when these regularity conditions are not met (Section 5). Second, as main mathematical result, we provide the solution to Problem (1), including the characterization of the minimiser when it exists, for Γ of the form (2) defined in terms of a convex integral functional (Section 6). The special cases of Bregman balls and f-divergence balls are treated in Section 7. Finally, in Section 8 we will address the related, mathematically simpler, problem

$$W := \inf_{\mathbb{P}} \left[E_{\mathbb{P}}(X) + \lambda D(\mathbb{P}\,||\,\mathbb{P}_0)\right], \quad \lambda > 0. \tag{3}$$

Decision makers with divergence preferences rank alternatives X by this criterion. We apply the methods of Section 6 to derive an explicit solution for the divergence preference problem (3).

Mathematically, our approach will be to exploit the relationship of Problem (1) to that of minimizing convex integral functionals (and specifically relative entropy) under moment constraints. The tools we need do not go beyond convex duality for $\mathbb{R}$ and $\mathbb{R}^2$, and many results directly follow from known ones about the moment problem. While in the literature attention is frequently restricted to essentially bounded X, here the $\mathbb{P}_0$-integrability of X suffices.

2 Relation to the literature

Problem (1) has been addressed in the literature in two related contexts: coherent risk measures and ambiguity. Law-invariante risk measures assign to a profit loss distribution a number interpreted as risk capital. Artzner et al. [1999] and Föllmer and Schied [2004] formulated requirements for risk measures and coined the terms 'coherent' resp. 'convex' for risk measures fulfulling them. *Every* coherent risk measure can be represented as (1) for some closed convex set Γ of probabilities.[3] The risk capital required for a portfolio is the worst expected loss over the set Γ.

In the context of risk measurement, the model risk measure (1) is yet another coherent risk measure. Defining the risk measure by the set Γ via the representation (1) is natural when addressing distribution model risk. Risk measures defined in terms of the profit loss distribution, like Value at Risk or Expected Shortfall, rely on a specific distribution model, which may be misspecified or misestimated. For a fixed portfolio, represented by a pricing function X, a different risk factor distribution gives rise to a different profit loss distribution, and therefore to a different risk capital requirement. Expression (1) measures exactly this model dependence.[4]

On the other hand, Problem (1) describes ambiguity averse preferences: A widely used class of preferences allowing for ambiguity aversion are the multiple priors preferences, also known as maxmin expected utility preferences, axiomatised by Gilboa and Schmeidler [1989].[5] (Another description of ambiguity aversion are the divergence preferences (3).) Agents with multiple priors preferences choose acts X with higher worst expected utility, where the worst case is taken over a closed convex set set Γ of finitely additive probabilities. The set Γ is interpreted as a set of priors held by the agent, and ambiguity is reflected by the multiplicity of the priors. Interpreting the choice of a portfolio as an act, the risk measure representation

[3] The representation theorem is due to Artzner et al. [1999] for finite sample spaces, for general probability spaces see Delbaen [2002] or Föllmer and Schied [2002]. Its formal statement is not needed for our purposes.

[4] One could object that Expected Shortfall is coherent and therefore can be represented by eq. (1) as a maximum expected loss over some set Γ of alternative distribution models. The set Γ equals $\{\mathbb{P} = \mathbb{P}_0[.|A] : \mathbb{P}_0(A) \geq \alpha\}$, which contains distributions so different from $\mathbb{P}_0$ that they are hardly plausible to arise from the same historical data by estimation or specification errors. Or, one could represent expected shortfall by eq. (1) with Γ as in (2), taking $D = D_f$ as in (7) below, with the pathological convex function f equal to 0 in the interval $[0, 1/\alpha]$ and $+\infty$ otherwise [Föllmer and Schied, 2004, Theorem 4.47]. But this f does not meet the assumptions in Section 3 and the corresponding D_f is not a divergence in our sense.

[5] Gilboa and Schmeidler [1989] worked in the setting of Anscombe and Aumann [1963] using lottery acts. Casadesus-Masanell et al. [2000] translated their approach to Savage acts. In the Gilboa-Schmeidler theory the utility of outcomes occurs separately, whereas in our notation the utility is part of the function X, which we would interpret as the utility of outcomes.

(1) and the multiple priors preference representation agree, see Föllmer and Schied [2002]. A decision maker who ranks portfolios by lower values of some coherent risk measure displays multiple priors preferences. And vice versa, a decision maker with multiple priors preferences acts as if she were minimising some coherent risk measure.

In the context of the Gilboa-Schmeidler theory, our results provide explicit expressions for the decision criterion of ambiguity averse decision makers, in the special case that the priors set Γ is given by (2). Choosing the same Γ for all agents may be at odds with a descriptive view of real agents' preferences. But from a normative point of view our choice of Γ in (2) is motivated by general arguments (Section 3). Our results can serve as a starting point for the further analysis of portfolio selection and contingent claim pricing under model uncertainty, extending, among others, work of Avellaneda and Paras [1996], Friedman [2002a,b], Calafiore [2007].

In the present context, the choice of Γ by (2) with $D(\mathbb{P}||\mathbb{P}_0)$ equal to relative entropy, has been proposed by Hansen and Sargent [2001], see also Ahmadi-Javid [2011] and Breuer and Csiszár [2012]. Friedman [2002a] also used relative entropy balls as sets of possible models. Hansen and Sargent [2001, 2007, 2008], Barillas et al. [2009] and others have used a relative entropy-based set of alternative models. Their work is set in a multiperiod framework. It deals with questions of optimal choice, whereas we take the portfolio X as given. Maccheroni et al. [2006] presented a unified framework encompassing both the multiple priors preference (1) and the divergence preferences (3). They proposed to use weighted f-divergences, which are also covered in our framework. Ben-Tal and Teboulle [2007, Theorem 4.2] showed that their optimised certainty equivalent for a utility function u can be represented as divergence preference (3) with D equal to the f-divergence with the function f satisfying $u(x) = -f^*(-x)$. For both, the worst case solution is a member of the same generalised exponential family. This paper makes clear the reasons.

Finally but importantly, the work of Ahmadi-Javid [2011] has to be cited for solutions of (1) and (3), in case of relative entropy and of f-divergences, in the form of convex optimization formulas involving two real variables (one in the case of relative entropy). The relationship of these results to ours will not be discussed here but we mention that in Ahmadi-Javid [2011] the pathological cases for relative entropy treated in Section 5 were not addressed, and the results for f-divergences were obtained under the assumptions that f is cofinite and X is essentially bounded.

3 Measures of plausibility of alternative risk factor distributions

We define divergences between non-negative functions on the state space Ω, which may be any set equipped with a σ-algebra not mentioned in the sequel, and with some measure μ on that σ-algebra. Here μ may or may not be a probability measure. Then the divergence between distributions (probability measures on Ω) absolutely continuous with respect to μ is taken to be the divergence between the corresponding density functions. In our terminology, a divergence is non-negative and vanishes only for identical functions or distributions. (Functions which are equal μ-a.e. are regarded as identical.) A divergence need not be a metric, may be non-symmetric, and the divergence balls need not form a basis for a topology in the space of probability distributions.

The relative entropy of two non-negative functions p, p_0 is defined as

$$I(p||\, p_0) := \int_\Omega [p(r) \log \frac{p(r)}{p_0(r)} - p(r) + p_0(r)] d\mu(r).$$

If p, p_0 are μ-densities of probability distributions $\mathbb{P}, \mathbb{P}_0$ this reduces to the original definition of Kullback and Leibler [1951],

$$I(\mathbb{P}\,||\,\mathbb{P}_0) = \int \log \frac{d\mathbb{P}}{d\mathbb{P}_0}(r) d\mathbb{P}(r) \quad \text{if } \mathbb{P} \ll \mathbb{P}_0.$$

If a distribution $\mathbb{P}$ is not absolutely continuous with respect to $\mathbb{P}_0$, take $I(\mathbb{P}\,||\,\mathbb{P}_0) = +\infty$.[6]

Bregman distances, introduced by Bregman [1967], and f-divergences, introduced by Csiszár [1963, 1967], and Ali and Silvey [1966], are classes of divergences parametrised by convex functions $f : (0, \infty) \to \mathbb{R}$, extended to $[0, \infty)$ by setting $f(0) := \lim_{t \to 0} f(t)$. Below, f is assumed strictly convex but not necessarily differentiable.

The Bregman distance of non-negative (measurable) functions p, p_0 on Ω, with respect to a (finite or σ-finite) measure μ on Ω is defined by

$$B_{f,\mu}(p, p_0) := \int_\Omega \Delta_f(p(r), p_0(r)) \mu(dr), \tag{4}$$

where, for s, t in $[0, +\infty)$,

$$\Delta_f(s,t) := \begin{cases} f(s) - f(t) - f'(t)(s-t) & \text{if } t > 0 \text{ or } t = 0,\ f(0) < +\infty \\ s \cdot (+\infty) & \text{if } t = 0 \text{ and } f(0) = +\infty. \end{cases} \tag{5}$$

[6]Note that $I(\mathbb{P}\,||\,\mathbb{P}_0)$ is a less frequent notation for relative entropy than $D(\mathbb{P}\,||\,\mathbb{P}_0)$, it has been chosen here because we use the latter to denote any divergence.

If the convex function f is not differentiable at t, the right or left derivative is taken for $f'(t)$ according as $s > t$ or $s < t$.

The Bregman distance of distributions $\mathbb{P} \ll \mu, \mathbb{P}_0 \ll \mu$ is defined by

$$B_{f,\mu}(\mathbb{P}, \mathbb{P}_0) := B_{f,\mu}\left(\frac{d\mathbb{P}}{d\mu}, \frac{d\mathbb{P}_0}{d\mu}\right). \tag{6}$$

Clearly, $B_{f,\mu}$ is a bona fide divergence whenever f is strictly convex in $(0, +\infty)$. For $f(s) = s \log s - s + 1$, B_f is the relative entropy I. For $f(s) = -\log s$, B_f is the Itakura-Saito distance. For $f(s) = s^2$, B_f is the squared L^2-distance.

The f-divergence between non-negative (measurable) functions p and p_0 is defined, when f additionally satisfies $f(s) \geq f(1) = 0$,[7] by

$$D_f(p||p_0) := \int_\Omega f\left(\frac{p(r)}{p_0(r)}\right) p_0(r)\, \mu(dr). \tag{7}$$

At places where $p_0(r) = 0$, the integrand by convention is taken to be $p(r) \lim_{s\to\infty} f(s)/s$. The f-divergence of distributions $\mathbb{P} \ll \mu, \mathbb{P}_0 \ll \mu$, defined as the f-divergence of the corresponding densities, does not depend on μ and is equal to

$$D_f(\mathbb{P}||\mathbb{P}_0) := \int_\Omega f\left(\frac{d\mathbb{P}_a}{d\mathbb{P}_0}\right) d\mathbb{P}_0 + \mathbb{P}_s(\Omega) \lim_{s\to\infty} \frac{f(s)}{s}, \tag{8}$$

where $\mathbb{P}_a$ and $\mathbb{P}_s$ are the absolutely continuous and singular components of $\mathbb{P}$ with respect to $\mathbb{P}_0$. Note that if f is cofinite, i.e., if the limit in (8) is $+\infty$, then $\mathbb{P} \ll \mathbb{P}_0$ is a necessary condition for the finiteness of $D_f(\mathbb{P}||\mathbb{P}_0)$, while otherwise not.

For $f(s) = s \log s - s + 1$, D_f is the relative entropy. For $f(s) = -\log s + s + 1$, D_f is the reversed relative entropy. For $f(s) = (\sqrt{s} - 1)^2$, D_f is the squared Hellinger distance. For $f(s) = (s-1)^2/2$, D_f is the relative Gini concentration index. For more details about f-divergences see Liese and Vajda [1987].

Relative entropy appears the most versatile divergence measure for probability distributions or non-negative functions, extensively used in diverse fields including statistics, information theory, statistical physics, see e.g. Kullback [1959], Csiszár and Körner [2011], Jaynes [1957]. For its applications in econometrics, see Golan et al. [1996] or Grechuk et al. [2009]. In the context of this paper, Hansen and Sargent [2001] have used expected value minimization over relative entropy balls. Arguments for (2) with any f-divergence in the role of D, or more generally with a weighted f-divergence involving a (positive) weight function $w(r)$ in the integral in (7), have been

[7] This makes sure that (7) indeed defines a divergence between any non-negative functions; if attention is restricted to probability densities resp. probability distributions, it suffices to assume that $f(1) = 0$.

put forward by Maccheroni et al. [2006]. Results of Ahmadi-Javid [2011] indicate advantages of relative entropy over other f-divergences also in this context. In another context, Grunwald and Dawid [2004] argue that distances between distributions might be chosen in a utility dependent way. Relative entropy is natural only for decision makers with logarithmic utility. Picking up this idea, for decision makers with non-logarithmic utility one might define the radius in terms of some utility dependent distance. We are unaware of references employing (2) with Bregman distances, although this would appear natural, particularly as Bregman distances have a beautiful interpretation as measuring the expected utility losses due to the convexity of f.

In the context of inference, the method of maximum entropy (or relative entropy minimization) is distinguished by axiomatic considerations. Shore and Johnson [1980], Paris and Vencovská [1990], and Csiszár [1991] showed that it is the only method that satisfies certain intuitively desirable postulates. Still, relative entropy cannot be singled out as providing the only reasonable method of inference. Csiszár [1991] determined what alternatives (specifically, Bregman distances and f-divergences) come into account if some postulates are relaxed. In the context of measuring risk or evaluating preferences under ambiguity aversion, axiomatic results distinguishing relative entropy or some other divergence are not available.

An objection against the choice of the set Γ in (2) with D equal to relative entropy or a related divergence should also be mentioned. It is that all distributions in this set are absolutely continuous with respect to $\mathbb{P}_0$. In the literature of the subject, even if not working with divergences, it is a rather common assumption that the set of feasible distributions is dominated; one notable exception is Cont [2006]. Sometimes the assumption that Γ is dominated is hard to justify. For example, in a multiperiod setting where Ω is the canonical space of continuous paths and Γ is a set of martingale laws for the canonical process, corresponding to different scenarios of volatilities, this Γ is typically not dominated (see Nutz and Soner [2012]). Or, if we use a continuous default distribution $\mathbb{P}_0$, can we always be sure that the data generating process is not discrete? And should it not be possible to approximate in some appropriate sense a continuous distribution by discrete ones?

If an f-divergence with a non-cofinite f is used, then the set Γ of alternative distributions is not dominated, see (8). But since all distributions singular to $\mathbb{P}_0$ have the same f-divergence from $\mathbb{P}_0$, even f-divergences with non-cofinite f are not appropriate to describe the approximation of a continuous distribution by discrete distributions. Bregman distances have a similar shortcoming. In practice, this objection does not appear a serious obstacle, for the set Γ of theoretical alternatives may be extended by distributions close to them in an appropriate sense involving closeness of expectations, which negligibly changes the theoretical risk value (1).

4 Intuitive relation of worst case risk and maximum entropy inference

The purpose of this section is to develop intuition on the relation between Problem (1) and the maximum entropy problem. Let us consider the mathematically simplest case of Problem (1), when Γ is a sufficiently small relative entropy ball. Then Problem (1) requires the evaluation of

$$\inf_{\mathbb{P}:I(\mathbb{P}\,||\,\mathbb{P}_0)\leq k} E_{\mathbb{P}}(X) =: V(k), \tag{9}$$

for sufficiently small k. We follow Breuer and Csiszár [2012], using techniques familiar in the theory of exponential families, see Barndorff-Nielsen [1978], and large deviations theory, see Dembo and Zeitouni [1998]. The meaning of 'sufficiently small' will be made precise later in this section. The cases when the relative entropy ball is not 'sufficiently small' will be treated in Section 5.

Observe that Problem (1) with Γ a relative entropy ball is "inverse" to a problem of maximum entropy inference. If an unknown distribution $\mathbb{P}$ had to be inferred when the available information specified only a feasible set of distributions, and a distribution $\mathbb{P}_0$ were given as a prior guess of $\mathbb{P}$, the maximum entropy[8] principle would suggest to infer the feasible distribution $\mathbb{P}$ which minimizes $I(\mathbb{P}\,||\,\mathbb{P}_0)$. In particular, if the feasible distributions were those with $E_{\mathbb{P}}(X) = b$, for a constant b, we would arrive at the problem

$$\inf_{\mathbb{P}:E_{\mathbb{P}}(X)=b} I(\mathbb{P}\,||\,\mathbb{P}_0). \tag{10}$$

Note that the objective function of problem (1) is the constraint in the maximum entropy problem (10), and vice versa (Fig. 1). It is therefore intuitively expected that (taking k and b suitably related) both problems are solved by the same distribution $\overline{\mathbb{P}}$,

$$\arg\min_{\mathbb{P}:I(\mathbb{P}\,||\,\mathbb{P}_0)\leq k} E_{\mathbb{P}}(X) = \arg\min_{\mathbb{P}:E_{\mathbb{P}}(X)=b} I(\mathbb{P}\,||\,\mathbb{P}_0) =: \overline{\mathbb{P}}, \tag{11}$$

see Fig. 1. The literature on the maximum entropy problem establishes that (under some regularity conditions) the solution $\overline{\mathbb{P}}$ is a member of the exponential family of distributions $\mathbb{P}(\theta)$ with canonical statistic X, which have a $\mathbb{P}_0$-density

$$\frac{d\mathbb{P}(\theta)}{d\mathbb{P}_0}(r) := \frac{e^{\theta X(r)}}{\int e^{\theta X(r)} d\mathbb{P}_0(r)} = e^{\theta X(r)-\Lambda(\theta)}, \tag{12}$$

where $\theta \in \mathbb{R}$ is a parameter and the function Λ is defined as

$$\Lambda(\theta) := \log \int e^{\theta X(r)} d\mathbb{P}_0(r). \tag{13}$$

[8]This name refers to the special case when $\mathbb{P}_0$ is the uniform distribution; then minimising $I(\mathbb{P}\,||\,\mathbb{P}_0)$ is equivalent to maximising the Shannon differential entropy of $\mathbb{P}$.

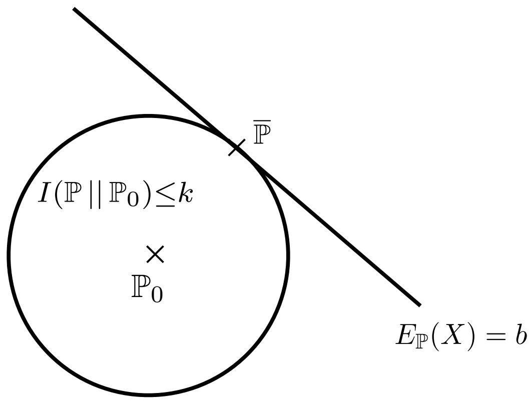

Figure 1: **Relation of Worst Case and Maximum Entropy problem.** What is the objective function in problem (1) is the constraint in the maximum entropy problem (10), and vice versa.

Among actuaries the distributions from the exponential family are often referred to as Esscher transforms.

For members $\mathbb{P}(\theta)$ of the exponential family, the expected profit can be written as

$$E_{\mathbb{P}(\theta)}(X) = \int X(r)\exp(\theta X(r) - \Lambda(\theta))d\mathbb{P}_0(r) = \Lambda'(\theta), \qquad (14)$$

and the relative entropy to $\mathbb{P}_0$ is

$$\begin{aligned} I(\mathbb{P}(\theta)\,||\,\mathbb{P}_0) &= \int \log\frac{d\mathbb{P}(\theta)}{d\mathbb{P}_0}(r)d\mathbb{P}(\theta)(r) = \int (\theta X(r) - \Lambda(\theta))d\mathbb{P}(\theta)(r) \\ &= \theta E_{\mathbb{P}(\theta)}(X) - \Lambda(\theta) = \theta\Lambda'(\theta) - \Lambda(\theta). \qquad (15) \end{aligned}$$

If the identity (11) holds and the solution of Problem (9) is from the exponential family, then one can determine which member of the exponential family solves the problem, by solving the equation

$$\theta\Lambda'(\theta) - \Lambda(\theta) = k \qquad (16)$$

for θ. Typically, (16) has both a positive and a negative solution, and the corresponding $\mathbb{P}(\theta)$ is the maximiser resp. minimiser of $E_{\mathbb{P}}(X)$ subject to $I(\mathbb{P}\,||\,\mathbb{P}_0) \leq k$. Call the negative solution $\overline{\theta}$. The solution to Problem (9) can then be expressed in terms of the Λ-function:

$$\inf_{\mathbb{P}:I(\mathbb{P}\,||\,\mathbb{P}_0)\leq k} E_{\mathbb{P}}(X) = \inf_{\theta:\theta\Lambda'(\theta)-\Lambda(\theta)\leq k} \Lambda'(\theta) = \Lambda'(\overline{\theta}),$$

(The last equality follows from the convexity of Λ.) This solution is illustrated in Fig. 2. The worst expected profit $V(k)$ is the slope of the tangent to

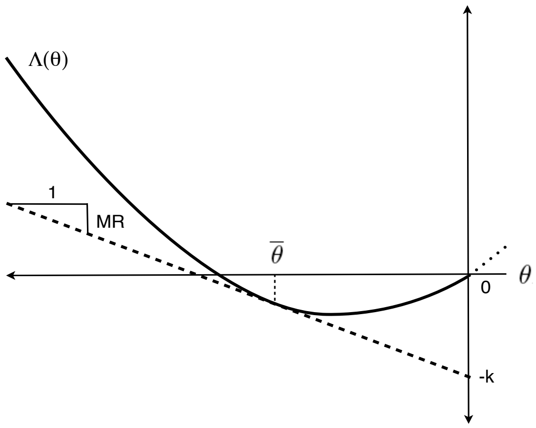

Figure 2: **Solution of the worst case problem problem from the Λ-function.** The optimal value achieved for Problem (9) is the slope of the tangent to the curve $\Lambda(\theta)$ passing through $(0, -k)$. $\overline{\theta}$ is the θ-coordinate of the tangent point.

the curve $\Lambda(\theta)$ passing through $(0, -k)$. $\overline{\theta}$ is the θ-coordinate of the tangent point. From the figure it is obvious that $\overline{\theta}\Lambda'(\overline{\theta}) - \Lambda(\overline{\theta}) = k$.

So far the intuition about the solution in what one could call the generic case. It requires two important assumptions: Identity (11) should hold and the equation (16) should have a (unique) negative solution $\overline{\theta}$. Breuer and Csiszár [2012] give precise conditions under which the solution is indeed of the generic form above. The first condition is relevant when X is essentially bounded below, the other two when it is not:

(i) If $\operatorname{ess\,inf}(X)$ is finite, $k < k_{\max} := -\log \mathbb{P}_0(\{r : X(r) = \operatorname{ess\,inf}(X)\})$.

(ii) $\theta_{\min} := \inf\{\theta : \Lambda(\theta) < +\infty\} < 0$,

(iii) If $\theta_{\min}$, $\Lambda(\theta_{\min})$, and $\Lambda'(\theta_{\min})$ are all finite then $k \leq \theta_{\min}\Lambda'(\theta_{\min}) - \Lambda(\theta_{\min})$.

The concepts used above are in close analogy to statistical mechanics. The risk factor vector r is the counterpart of the phase space points. The pricing function X is the counterpart of the energy function. Λ is the counterpart of the logarithm of the partition function Z. θ is the counterpart of the inverse temperature parameter $\beta = 1/kT$. The worst case distribution (12) is the counterpart of the canonical distribution.

5 Maximum Loss over relative entropy balls: The pathological cases

Now let us turn to the solution of Problem (9) in the pathological case where Γ is a large relative entropy ball, so that one of the conditions (i)-(iii) is violated.

First consider the case that assumption (i) above is violated, where the loss is essentially bounded and the sphere is not "sufficiently small". Long bond portfolios are examples for this case. In this case equation (16) has no negative solution. The shape of the Λ-function is displayed in Fig. 3.

Proposition 1. *If* $\operatorname{ess\,inf}(X)$ *is finite, and* $k \geq k_{\max}$ *(defined in (i), Section 4) then the solution to Problem* (9) *is* $V(k) = \operatorname{ess\,inf}(X)$. *The worst case distribution* $\overline{\mathbb{P}}$ *has the* $\mathbb{P}_0$*-density*

$$\frac{d\overline{\mathbb{P}}}{d\mathbb{P}_0}(r) := \begin{cases} 1/\beta & \text{if } X(r) = \operatorname{ess\,inf}(X) \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta = \mathbb{P}_0(\{r : X(r) = \operatorname{ess\,inf}(X)\})$.

Proof. The distribution $\overline{\mathbb{P}}$ satisfies

$$I(\overline{\mathbb{P}} \,||\, \mathbb{P}_0) = \int \log \frac{d\overline{\mathbb{P}}}{d\mathbb{P}_0} d\overline{\mathbb{P}} = -\log \beta,$$

hence $I(\overline{\mathbb{P}} \,||\, \mathbb{P}_0) \leq k$ if $k \geq -\log \beta$. Then $V(k) \leq E_{\overline{\mathbb{P}}}(X)$. Trivially $V(k) \geq \operatorname{ess\,inf}(X)$. The claim $V(k) = \operatorname{ess\,inf}(X)$ follows. $\square$

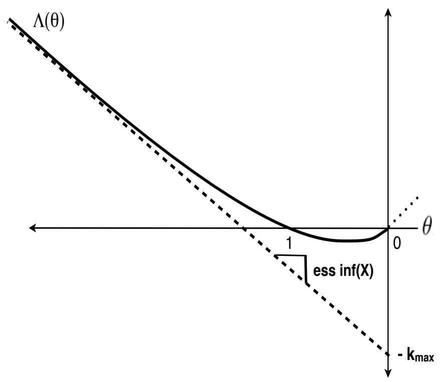

Figure 3: The pathological case of Proposition 1.

Next consider the pathogolical case that assumption (ii) above is violated so that $\theta_{\min} = 0$, and thus $\Lambda(\theta) = +\infty$ for all $\theta < 0$.

Proposition 2. *If $\theta_{\min}$ defined in (ii) of Section 4 equals zero, then the solution to Problem* (9) *is $V(k) = -\infty$ for all $k > 0$.*

Proof. Let $\beta_{m,n} := \mathbb{P}_0(\{r : -n \le X(r) \le m\})$ and consider the measures $\mathbb{P}_{m,n} \ll \mathbb{P}_0$ with

$$\frac{d\mathbb{P}_{m,n}}{d\mathbb{P}_0}(r) := \begin{cases} 1/\beta_{m,n} & \text{if } -n \le X(r) \le m \\ 0 & \text{otherwise.} \end{cases}$$

Obviously $I(\mathbb{P}_{m,n}||\mathbb{P}_0) = -\log\beta_{m,n}$. For any $\mathbb{P} \ll \mathbb{P}_{m,n}$,

$$I(\mathbb{P}\,||\,\mathbb{P}_0) = \int \log(\frac{d\mathbb{P}}{d\mathbb{P}_{m,n}}\frac{d\mathbb{P}_{m,n}}{d\mathbb{P}_0})d\mathbb{P} = I(\mathbb{P}\,||\,\mathbb{P}_{m,n}) - \log\beta_{m,n}$$

is arbitrarily close to $I(\mathbb{P}\,||\,\mathbb{P}_{m,n})$ if m and n are sufficiently large. Hence to prove that $V(k) = -\infty$ for all $k > 0$, it suffices to find to any given m and sufficiently large n distributions $\mathbb{P} \ll \mathbb{P}_{m,n}$ with $I(\mathbb{P}\,||\,\mathbb{P}_{m,n})$ arbitrarily close to zero and $E_{\mathbb{P}}(X)$ arbitrarily low.

In the rest of this proof, m is fixed and n will go to $+\infty$. Define $\mathbb{P}$ and $\Lambda_{m,n}$ by

$$\frac{d\mathbb{P}}{d\mathbb{P}_{m,n}}(r) := \frac{e^{\theta X(r)}}{\int e^{\theta X(r)}d\mathbb{P}_{m,n}(r)} =: e^{\theta X(r) - \Lambda_{m,n}(\theta)}$$

for any $\theta < 0$. $\mathbb{P}$ and $\Lambda_{m,n}$ depend on θ. As in (14), $E_{\mathbb{P}}(X) = \Lambda'_{m,n}(\theta)$ and $I(\mathbb{P}||\mathbb{P}_{m,n}) = \theta\Lambda'_{m,n}(\theta) - \Lambda_{m,n}(\theta)$ for any $\theta < 0$. For each θ,

$$-\theta m \ge \Lambda_{m,n}(\theta) = \int_\theta^0 \Lambda'_{m,n}(\xi)d\xi \ge -\theta\Lambda'_{m,n}(\theta), \tag{17}$$

since $\Lambda'_{m,n}$ is increasing. For fixed $\theta < 0$, $\Lambda_{m,n}(\theta) \to \infty$ as $n \to \infty$ since $\Lambda(\theta) = \infty$ by assumption. By (17) it follows that $\Lambda'_{m,n}(\theta) \to -\infty$ as $n \to \infty$, and hence there exists a sequence $\theta_n \uparrow 0$ such that $\Lambda'_{m,n}(\theta_n) \to -\infty$ and $\theta_n\Lambda'_{m,n}(\theta_n) \to 0$ as $n \to \infty$. By inequality (17), this implies $|\Lambda_{m,n}(\theta_n)| \to 0$ and hence $I(\mathbb{P}||\mathbb{P}_{m,n}) \to 0$ as $n \to \infty$. This completes the proof that, for $\mathbb{P}$ defined with $\theta = \theta_n$, $E_{\mathbb{P}}(X)$ will be arbitrarily low and $I(\mathbb{P}\,||\,\mathbb{P}_{m,n})$ arbitrarily small but positive. □

Finally consider the case that both $\Lambda(\theta_{\min})$ and $\Lambda'(\theta_{\min})$ are finite, but the sphere is not "sufficiently small". The shape of the Λ-function is displayed in Fig. 4.

Proposition 3. *If $-\infty < \theta_{\min} < 0$, ($\theta_{\min}$ is defined in (ii) of Section 4), and both $\Lambda(\theta_{\min})$ and $\Lambda'(\theta_{\min})$ are finite, and additionally $k > \theta_{\min}\Lambda'(\theta_{\min}) - \Lambda(\theta_{\min})$, then*

$$V(k) = (k + \Lambda(\theta_{\min}))/\theta_{\min}, \tag{18}$$

but there is no distribution achieving this value; the infimum in Problem (9) *is not a minimum.*

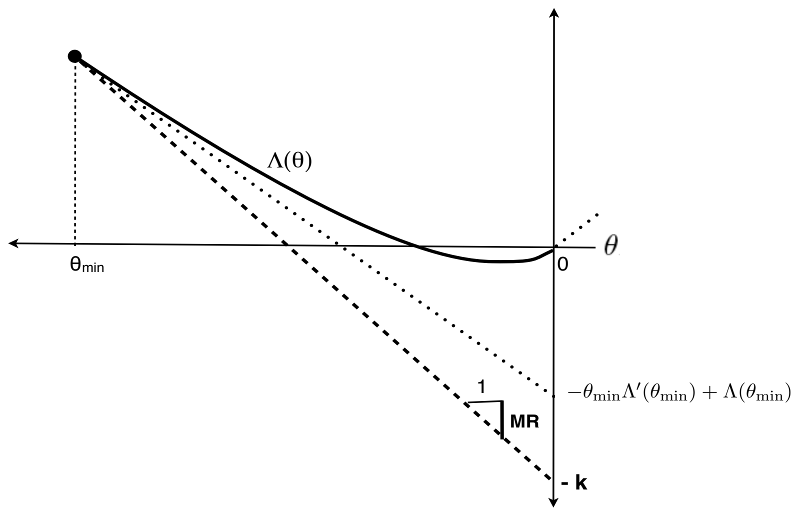

Figure 4: The pathological case of Proposition 3.

Proof. Define $\mathbb{P}(\theta_{\min})$ as in (12) with $\theta_{\min}$ in the place of θ. Then

$$\begin{aligned} I(\mathbb{P}\|\mathbb{P}_0) &= \int \log\left(\frac{d\mathbb{P}}{d\mathbb{P}(\theta_{\min})}\frac{d\mathbb{P}(\theta_{\min})}{d\mathbb{P}_0}\right) d\mathbb{P} \\ &= I(\mathbb{P}\|\mathbb{P}(\theta_{\min})) + \int \log(\exp(\theta_{\min}X(r) - \Lambda(\theta_{\min}))d\mathbb{P}(r) \\ &= I(\mathbb{P}\|\mathbb{P}(\theta_{\min})) + \theta_{\min}E_{\mathbb{P}}(X) - \Lambda(\theta_{\min}) \qquad (19) \end{aligned}$$

for all $\mathbb{P} \ll \mathbb{P}(\theta_{\min})$. Hence, if $I(\mathbb{P}\,||\,\mathbb{P}_0) \le k$ then (using $\theta_{\min} < 0$)

$$E_{\mathbb{P}}(X) \ge (k + \Lambda(\theta_{\min}) - I(\mathbb{P}\|\mathbb{P}(\theta_{\min})))/\theta_{\min}, \qquad (20)$$

proving that $V(k) \ge (k+\Lambda(\theta_{\min}))/\theta_{\min}$. To show that equality holds, apply the result of Proposition 2 to $\mathbb{P}(\theta_{\min})$ in the role of $\mathbb{P}_0$, then the role of $\Lambda(\theta)$ is played by

$$\overline{\Lambda}(\theta) := \log \int e^{\theta X(r)} d\mathbb{P}(\theta_{\min})(r) = \Lambda(\theta + \theta_{\min}) - \Lambda(\theta_{\min}).$$

Clearly, $\overline{\Lambda}(\theta) = \infty$ for all $\theta < 0$, hence by Proposition 2 there exist distributions $\mathbb{P}'$ with $I(\mathbb{P}'\|\mathbb{P}(\theta_{\min}))$ arbitrarily small but positive and $E_{\mathbb{P}'}(X)$ arbitrarily low. Then, for any small $\epsilon > 0$, a suitable linear combination $\mathbb{P}$ of $\mathbb{P}'$ and $\mathbb{P}(\theta_{\min})$ satisfies $E_{\mathbb{P}}(X) = (k + \Lambda(\theta_{\min}) - \epsilon)/\theta_{\min}$ and $I(\mathbb{P}\|\mathbb{P}(\theta_{\min})) < \epsilon$. For this $\mathbb{P}$, eq. (19) implies that $I(\mathbb{P}\,||\,\mathbb{P}_0) \le k$ and the claim $V(k) \le (k + \Lambda(\theta_{\min}))/\theta_{\min}$ follows. This proves that $V(k) = (k + \Lambda(\theta_{\min}))/\theta_{\min}$.

(20) implies that in Problem (9) the supremum is not attained because $I(\mathbb{P}\,||\,\mathbb{P}(\theta_{\min}))$ is strictly positive when $E_{\mathbb{P}}(X) < \Lambda'(\theta_{\min}) = E_{\mathbb{P}(\theta_{\min})}(X)$. $\square$

Remark 1. Consider the convex conjugate of Λ defined by

$$\Lambda^*(x) := \sup_{\theta}(\theta x - \Lambda(\theta)), \tag{21}$$

which is a convex, lower semicontinuous function on $\mathbb{R}$. Clearly,

$$\Lambda^*(x) = \theta x - \Lambda(\theta) \quad \text{if} \quad x = \Lambda'(\theta). \tag{22}$$

However, for some x perhaps no θ satisfies $x = \Lambda'(\theta)$. In the generic case, when the assumptions (i)-(iii) of Section 4 are met, the optimal value attained in Problem (1) is equal to $x = \Lambda'(\overline{\theta})$ for $\overline{\theta}$ satisfying (16), which x is the unique solution of

$$\Lambda^*(x) = k \text{ and } x < E_{\mathbb{P}_0}(X). \tag{23}$$

The proof of Proposition 3 establishes that $V(k)$ always equals the solution of (23) when it exists, even if (16) does not have a solution.

6 A more general framework

Now we construct a unified framework that covers the choices of Γ in (2) when D is an f-divergence or a Bregman distance, as well as others. In this framework, Γ is chosen as a set of probability measures $P \ll \mu$ (where μ is a given measure on Ω, finite or σ-finite) of the form

$$\Gamma = \{\mathbb{P} \ll \mu : \ p = d\mathbb{P}/d\mu \ \text{ satisfies } \ H(p) \leq k\}, \tag{24}$$

where H is a convex integral functional defined as

$$H(p) := \int_{\Omega} \beta(r, p(r))\mu(dr), \tag{25}$$

for measurable, non-negative functions p on Ω. Here $\beta : \Omega \times (0, +\infty) \to \mathbb{R}$ is a mapping such that $\beta(r, s)$ is a measurable function of r for each $s \in (0, +\infty)$ and a strictly convex function of s for each $r \in \Omega$. The definition of β is extended to $s \leq 0$ by

$$\beta(r, 0) := \lim_{s \downarrow 0} \beta(r, s), \quad \beta(r, s) := +\infty \text{ if } s < 0. \tag{26}$$

No differentiability assumptions are made about β but the convenient notations $\beta'(r, 0)$ and $\beta'(r, +\infty)$ will be used for the common limits of the left and right derivatives of $\beta(r, s)$ by s as $s \downarrow 0$ resp. $s \uparrow +\infty$. Note that

$$\beta'(r, +\infty) = \lim_{s \uparrow +\infty} \frac{\beta(r, s)}{s}. \tag{27}$$

With the understandings (26), the mapping $\beta : \Omega \times \mathbb{R} \to (-\infty, +\infty]$ is a convex normal integrand in the sense of Rockafellar and Wets [1997], which ensures the measurability[9] of the function $\beta(r, p(r))$ in (25) and of similar functions later on, as in (36) and (38).

Depending on the choice of β, $H(p)$ will be relative entropy to $\mathbb{P}_0$, some Bregman distance, some f-divergence, or some other divergence, as in Section 7 below. Our general assumption about the relation of β and the best guess distribution $\mathbb{P}_0$, always satisfied in the above cases, will be that the minimum of $H(p)$ among probability densities p is attained for p_0, the density of $\mathbb{P}_0$; without any loss of generality, this minimum is supposed to be 0, thus

$$H(p) \geq H(p_0) = 0 \text{ whenever } \int p d\mu = 1. \tag{28}$$

In addition, we assume that $E_{\mathbb{P}_0}(X) = \int X p_0 d\mu$ exists and

$$m := \mu\text{-ess inf}(X) \ < b_0 := E_{\mathbb{P}_0}(X) < \ M := \mu\text{-ess sup}(X). \tag{29}$$

Relation of the model risk problem and the moment problem The distribution model risk (1) with Γ as in (24) is evaluated by solving the worst case problem

$$\inf_{p: \int p d\mu = 1, H(p) \leq k} \int X p d\mu =: V(k) \tag{30}$$

and then taking $\text{MR} = -V(k)$. Our goal is to determine $V(k)$, and also the minimiser (the density of the worst case scenario in Γ), if $V(k)$ is finite and the minimum in (30) is attained. If this minimiser exists, it is unique, by strict convexity of β.

Problem (30) is related to the the moment problem

$$\inf_{p: \int p d\mu = 1, \int X p d\mu = b} H(p) =: F(b) \tag{31}$$

in analogy to the relation between problem (9) and the maximum entropy problem (10) described in Section 4. Denote

$$k_{\max} := \lim_{b \downarrow m} F(b).$$

Proposition 4. *Supposing*

$$0 < k < k_{\max}, \tag{32}$$

there exists a unique b with $m < b < b_0$ and

$$F(b) = k, \tag{33}$$

[9] Measurability issues will not be entered below. For the measurability of functions we deal with, see references in Csiszár and Matúš [2012] to the book of Rockafellar and Wets [1997].

and then the solution to problem (30) *has the value*

$$V(k) = b. \tag{34}$$

The minimum in (30) *is attained if and only if that in* (31) *is attained (for this* b*), in which case the same* p *attains both minima.*

Proof. As the convex function F attains its minimum 0 at b_0, the assumption (32) trivially implies the existence of a unique b satisfying (33). Moreover, then each $t \in (b, b_0)$ satisfies $F(t) < k$, hence there exist functions p with $\int p d\mu = 1$, $\int X p d\mu = t$ such that $F(t) > k$. This proves that $V(k) \leq b$. On the other hand, $F(t) > k$ if $t \in (m, b)$ (hence also $F(m) > k$ if m is finite), which means that the conditions $\int p d\mu = 1$ and $\int X p d\mu = t$ imply $H(p) \geq F(t) > k$ for each $t \in (-\infty, b)$. Since $\int X p d\mu > -\infty$ if $H(p) < \infty$, as verified later (Corollary 3 of Theorem 2), this proves that $V(k) \geq t$. The last assertion of the Proposition follows obviously. □

Remark 2. The condition (32) in Proposition 4 covers all interesting values of k. Indeed, one easily sees that if $k > k_{\max}$ or $k \geq k_{\max} > 0$ then $V(k) = m$, while clearly $V(0) = b_0$. This also means that the functional H can be suitable for assigning model risk only if $k_{\max} > 0$. A necessary and sufficient condition for $k_{\max} > 0$, analogous to condition (ii) in Section 4, will be given in Corollary 2 of Theorem 2. Note that if $m = -\infty$ then $k_{\max} > 0$ implies $k_{\max} = \infty$, in which case each $k > 0$ meets condition (32).

For technical reasons, it will be convenient to regard $F(b)$ as the instance $a = 1$ of the function

$$J(a, b) := \inf_{p: \int p d\mu = a, \int X p d\mu = b} H(p), \quad (a, b) \in \mathbb{R}^2. \tag{35}$$

Problem (35) is a special case of minimising convex integral functionals under moment constraints, which has an extensive literature. For references, see the recent work of Csiszár and Matúš [2012], relied upon here also for results that date back much earlier, perhaps under less general conditions. The results in Csiszár and Matúš [2012] will be used (without further mentioning this) with the choice $\phi : r \to (1, X(r))$ of the moment mapping when the “value function” there reduces to the function J here. Many results in that reference need a condition called dual constraint qualification which, however, always holds in the current setting, namely, the set Θ defined in (39) is non-empty (see the passage following (39)).

The role of the function Λ in Section 4 will be played by the function

$$K(\theta_1, \theta_2) := \int \beta^*(r, \theta_1 + \theta_2 X(r)) \mu(dr), \quad (\theta_1, \theta_2) \in \mathbb{R}^2, \tag{36}$$

where β^* is the convex conjugate of β with respect to the second variable,

$$\beta^*(r,x) := \sup_{s\in\mathbb{R}} \left(xs - \beta(r,s)\right), \quad x \in \mathbb{R}. \tag{37}$$

The properties of β imply that $\beta^*(r,x)$ is a convex function of x which is finite, non-decreasing, and differentiable in the interval $(-\infty, \beta'(r,+\infty))$, see (27). At $x = \beta'(r,+\infty)$, if finite, $\beta^*(r,x)$ may be finite or $+\infty$. The derivative $(\beta^*)'(r,x)$ equals zero for $x \leq \beta'(r,0)$, is positive for $\beta'(r,0) < x < \beta'(r,+\infty)$, and grows to $+\infty$ as $x \uparrow \beta'(r,+\infty)$.

The following functions on Ω will play the role of the exponential family, but are parametrised by two variables and need not integrate to 1:

$$p_\theta(r) := (\beta^*)'(r, \theta_1 + \theta_2 X(r)), \quad \theta = (\theta_1,\theta_2) \in \Theta \tag{38}$$

where[10]

$$\Theta := \left\{\theta : K(\theta_1,\theta_2) < +\infty,\ \ \theta_1 + \theta_2 X(r) < \beta'(r,+\infty)\ \mu\text{-a.e.}\right\}. \tag{39}$$

The properties of β^* stated above imply for any (θ_1,θ_2) in the effective domain $\operatorname{dom} K := \{(\theta_1,\theta_2) : K(\theta_1,\theta_2) < +\infty\}$ of K that $(\overline{\theta_1},\theta_2) \in \Theta$ for each $\overline{\theta_1} < \theta_1$. In particular, Θ contains the interior of $\operatorname{dom} K$. If $\beta'(r,+\infty) = +\infty$ μ-a.e. then $\Theta = \operatorname{dom} K$. As verified later, see Remark 4, the default density p_0 is equal to $p_{(\theta_0,0)}$ for some θ_0 with $(\theta_0,0) \in \Theta$.

The function K is equal to the convex conjugate of J:

$$K(\theta_1,\theta_2) = J^*(\theta_1,\theta_2) := \sup_{(a,b)\in\mathbb{R}^2} \left(\theta_1 a + \theta_2 b - J(a,b)\right), \tag{40}$$

see [Csiszár and Matúš, 2012, Theorem 1.1]. In particular, K is a lower semicontinuous proper[11] convex function. Also, K is differentiable in the interior of $\operatorname{dom} K$, and

$$\nabla K(\theta) = \left(\int p_\theta d\mu, \int X p_\theta d\mu\right), \quad \theta = (\theta_1,\theta_2) \in \operatorname{int}\operatorname{dom} K, \tag{41}$$

see [Csiszár and Matúš, 2012, Corollary 3.8].

Main results We calculate b satisfying (33), which by (34) amounts to solving problem (30), by evaluating instead of J the function K^*, using the identity $J^* = K$ which implies (Rockafellar [1970, Theorem 12.2])

$$J(a,b) = K^*(a,b), \quad (a,b) \in \text{int dom } J. \tag{42}$$

[10]The definition (39) makes sure that the derivative in (38) exists for μ-a.e. $r \in \Omega$ if $(\theta_1,\theta_2) \in \Theta$. For all other $r \in \Omega$, if any, one may set by definition $p_{\theta_1,\theta_2} = 0$.

[11]I.e., it never equals $-\infty$ and is not identically $+\infty$.

K^* is the convex conjugate of K,

$$K^*(a,b) := \sup_{(\theta_1,\theta_2)\in\mathbb{R}^2} \left(\theta_1 a + \theta_2 b - K(\theta_1,\theta_2)\right), \quad (a,b)\in\mathbb{R}^2, \tag{43}$$

and the interior of the effective domain of J is, by Csiszár and Matúš [2012, Lemma 6.6]

$$\text{int dom } J = \{(a,b) : a > 0,\ am < b < aM\}. \tag{44}$$

Proposition 4 and (42), (44) imply for $0 < k < k_{\max}$ the analogue of Remark 1: A (unique) b satisfies

$$K^*(1,b) = k \quad \text{and} \quad b < b_0 = E_{\mathbb{P}_0}(X), \tag{45}$$

and then $V(k) = b$. This already provides a recipe for computing $V(k)$. In regular cases, a more explicit solution is available, based on the following key result about Problem (35), see [Csiszár and Matúš, 2012, Lemma 4.4, Lemma 4.10]:

Lemma 1. *If $\theta = (\theta_1,\theta_2)\in\Theta$ satisfies*

$$\int p_\theta \, d\mu = a, \quad \int X p_\theta \, d\mu = b \tag{46}$$

then it attains the maximum in (43). Moreover, in case $(a,b)\in$ int dom J, the existence of $\theta\in\Theta$ satisfying (46) is necessary and sufficient for the attainment of the minimum in (35), and then $p = p_\theta$ is the (unique) minimiser.

Theorem 1. *Assuming (28), (29), (32), if*

$$\overline{\theta}_2 < 0, \quad \int p_{\overline{\theta}} \, d\mu = 1, \quad \overline{\theta}_1 + \overline{\theta}_2 \int X p_{\overline{\theta}} \, d\mu - K(\overline{\theta}) = k \tag{47}$$

for some $\overline{\theta} = (\overline{\theta}_1, \overline{\theta}_2)\in\Theta$ then the value of the inf in (30) is

$$V(k) = \int X p_{\overline{\theta}} \, d\mu. \tag{48}$$

Essential smoothness[12] *of K is a sufficient condition for the existence of such $\overline{\theta}$. Further, a necessary and sufficient condition for p to attain the minimum in (30) is $p = p_{\overline{\theta}}$ for the $\overline{\theta}\in\Theta$ satisfying (47).*

[12] A lower semicontinuous proper convex function is essentially smooth if its effective domain has nonempty interior, the function is differentiable there, and at non-interior points of the effective domain the directional derivatives in directions towards the interior are $-\infty$. The latter trivially holds if the effective domain is open.

Corollary 1. *If the equations*

$$\frac{\partial}{\partial\theta_1}K(\theta) = 1, \quad \theta_1 + \theta_2\frac{\partial}{\partial\theta_2}K(\theta) - K(\theta) = k \tag{49}$$

have a solution $\overline{\theta} = (\overline{\theta}_1, \overline{\theta}_2) \in$ *int dom* K *with* $\overline{\theta}_2 < 0$ *then* $\overline{\theta}$ *satisfies* (47) *and the solution to Problem* (30) *equals*

$$V(k) = \left.\frac{\partial K(\theta)}{\partial\theta_2}\right|_{\theta=\overline{\theta}}. \tag{50}$$

The Corollary follows from the Theorem because, for $\overline{\theta} \in \text{int dom } K$, the equations in (47) are equivalent to those in (49), by (41). However, if K is not essentially smooth, $\overline{\theta} \in \text{int dom } K$ is not a necessary condition for (47).

Proof. By Lemma 1, if $\theta = (\theta_1, \theta_2) \in \Theta$ satisfies

$$\int p_\theta \, d\mu = 1, \int Xp_\theta \, d\mu = b \tag{51}$$

then it attains the maximum in (43). It follows, using (42), that (47) implies for $b := \int Xp_{\overline{\theta}} \, d\mu$, if it satisfies $m < b < M$, that

$$F(b) = J(1, b) = \overline{\theta}_1 + \overline{\theta}_2 b - K(\overline{\theta}) = k. \tag{52}$$

Due to Proposition 4, to prove (48) it remains to show that $m < b < b_0$. Clearly, $k < k_{\max}$ implies $m < b$. Further, (43) and (52) imply

$$F(t) = K^*(1, t) \geq \overline{\theta}_1 + \overline{\theta}_2 t - K(\overline{\theta}_1, \overline{\theta}_2) = F(b) + \overline{\theta}_2(t - b), \quad t \in (m, M). \tag{53}$$

Since $\overline{\theta}_2 < 0$, this shows that $F(t) > F(b)$ if $t \in (m, b_0)$, completing the proof of (48).

Suppose next that K is essentially smooth. Then to b in (33) there exists $\overline{\theta} \in \text{int dom } K$ with

$$(1, b) = \nabla K(\overline{\theta}), \tag{54}$$

because $(1, b) \in \text{int dom } J$ and the gradient vectors of the essentially smooth K cover int dom $K^* =$ int dom J, see Rockafellar [1970, Corollary 26.4.1]. Clearly, (54) implies that $\overline{\theta}$ attains the maximum in (43), hence it satisfies (52). This means by (54) that $\overline{\theta}$ satisfies the equations in (49), equivalent to those in (47). It remains to show that $\overline{\theta}_2 < 0$, but this follows from (53) applied to $t = b_0$.

Finally, the last assertion of Theorem 1 follows from Proposition 4 and Lemma 1. □

Conditions for $k_{\max} > 0$. In Proposition 4 and Theorem 1 the condition $k_{\max} > 0$ has been assumed. In this subsection we give a necessary and sufficient condition for this to hold. We begin with a remark.

Remark 3. A simpler instance of [Csiszár and Matúš, 2012, Lemma 4.10] than Lemma 1, namely with the constant mapping $r \to 1$ taken for the moment mapping ϕ, gives the following: the necessary and sufficient condition for p to minimise $H(p)$ subject to $\int p\, d\mu = a$ $(a > 0)$ is that $p(r) = (\beta^*)'(r, \theta)$ for some $\theta \in \mathbb{R}$ with $\beta^*(r, \theta)$ μ-integrable, and then the minimum is equal to $a\theta - \int \beta^*(r, \theta)\, d\mu(r)$. This establishes the claim that the default density p_0, minimising $H(p)$ subject to $\int p\, d\mu = 1$, equals $p_{(\theta_0, 0)}$ for some θ_0 with $(\theta_0, 0) \in \Theta$; this θ_0 also satisfies $\theta_0 - \int \beta^*(r, \theta_0)\, d\mu(r) = H(p_0) = 0$.

Theorem 2. *Assuming* (28), (29), *for $b < b_0$ we have $F(b) > 0$ if and only if*

$$\text{there exists } \theta = (\theta_1, \theta_2) \in \operatorname{dom} K \text{ with } \theta_2 < 0. \tag{55}$$

Proof. To prove the necessity of (55), we may assume $m < b < b_0$. Then $(1, b) \in \operatorname{int} \operatorname{dom} J$, see (44), hence the convex function J has nonempty subgradient at $(1, b)$ [Rockafellar, 1970, Theorem 23.4]. As $J^* = K$, if $\overline{\theta} = (\overline{\theta}_1, \overline{\theta}_2)$ belongs to that subgradient then

$$F(b) = J(1, b) = \overline{\theta}_1 + \overline{\theta}_2 b - K(\overline{\theta}) \tag{56}$$

by [Rockafellar, 1970, Theorem 23.5], which implies as in the proof of Theorem 1 that this $\overline{\theta}$ also satisfies (53). In turn, (53) with $t = b_0$ implies that $\overline{\theta}_2 \leq 0$, with the strict inequality if $F(b) > 0$. This proves the necessity of (55).

For sufficiency, suppose that $F(b) = 0$ for some $b \neq b_0$, $m < b < M$. By Remark 3, then $F(b) = 0 = \theta_0 - K(\theta_0, 0)$, hence $\theta^* := (\theta_0, 0)$ is a maximiser of $g(\theta) := \theta_1 + \theta_2 b - K(\theta)$, see (42), (43). It follows that for no $\bar{\theta} \in \operatorname{dom} K$ can the directional derivative $g'(\theta^*; \bar{\theta} - \theta^*)$ be positive. By [Csiszár and Matúš, 2012, Lemma 3.6, Remark 3.7], this directional derivative is equal to

$$(\bar{\theta}_1 - \theta_0) + \bar{\theta}_2 b - \int (\bar{\theta}_1 - \theta_0 + \bar{\theta}_2 X) p_{\theta^*}\, d\mu = \bar{\theta}_2 (b - b_0).$$

Thus, the existence of $\bar{\theta} \in \operatorname{dom} K$ with $\bar{\theta}_2 < 0$ rules out $b < b_0$, proving the sufficiency part of the Theorem. □

Corollary 2. *Condition* (55) *is necessary and sufficient for $k_{\max} > 0$. Sufficient conditions are the finiteness of m or the essential smoothness of K.*

Proof. If m is finite then each $\theta_2 < 0$ satisfies condition (55) with some θ_1. Indeed, since $\theta_1 + \theta_2 X \leq \theta_1 + \theta_2 m$ μ-a.e., if the right hand side is less than θ_0 in Remark 3 then $(\theta_1, \theta_2) \in \operatorname{dom} K$. If K is essentially smooth then condition (55) holds because int dom K contains $\theta^* = (\theta_0, 0)$. Indeed, otherwise the directional derivatives of K at θ^* in directions towards interior points were equal to $-\infty$, and θ^* could not maximize $\theta_1 + \theta_2 b_0 - K(\theta)$. □

Corollary 3. *If $k_{\max} > 0$ then $\int p d\mu = 1, H(p) < +\infty$ imply $\int X p d\mu > -\infty$.*

Proof. Substitute in the Fenchel inequality $xs \leq \beta(r,s) + \beta^*(r,x)$ (a consequence of (37)) $x := \theta_1 + \theta_2 X(r)$, $s := p(r)$ and integrate. It follows that if $(\theta_1, \theta_2) \in \operatorname{dom} K$ and p satisfies the hypotheses then

$$\theta_1 + \theta_2 \int X p d\mu \leq H(p) + K(\theta_1, \theta_2) < +\infty.$$

Taking (θ_1, θ_2) as in (55), the assertion follows. □

7 MaxLoss over Bregman balls and f-divergence balls

We now come back to the more specific choices (2), where Γ is a ball of distributions in terms of some divergence D, centered at some $\mathbb{P}_0$.

Relative entropy balls Let us briefly check how the unified framework leads, in the special case of relative entropy balls, to the results of Breuer and Csiszár [2012, Theorem 1] reported in Section 4.

Set $\mu = \mathbb{P}_0$ and take $\beta(r,s) := f(s) := s \log s - s + 1$. Then $\beta^*(r,x) = f^*(x) = \exp(x) - 1$ and $\beta^{*\prime}(r,x) = (f^*)'(x) = \exp(x)$. Hence, using (13) and (36),

$$K(\theta_1, \theta_2) = \int (\exp(\theta_1 + \theta_2 X) - 1)\, d\mathbb{P}_0 = \exp(\theta_1 + \Lambda(\theta_2)) - 1$$

and $\Theta = \operatorname{dom} K = \mathbb{R} \times \operatorname{dom} \Lambda$. The functions p_θ, $\theta \in \Theta$ of (38) are of form $\exp(\theta_1 + \theta_2 X(r))$, and integrate to 1 if and only if $\theta_1 = -\Lambda(\theta_2)$. Then p_θ is the $\mathbb{P}_0$-density of $\mathbb{P}(\theta_2)$ in the exponential family (12).

The first equation in (47) requires p_θ to be a density, thus $\theta_1 = -\Lambda(\theta_2)$. Then $\int X p_\theta d\mathbb{P}_0 = \Lambda'(\theta_2)$, see (14), and the second equation in (47) reads $-\Lambda(\theta_2) + \theta_2 \Lambda'(\theta_2) = k$, which is (16). Thus Theorem 1 gives the result in Section 4 that if (14) has a negative solution $\overline{\theta}$ then $V(k) = \Lambda'(\overline{\theta})$, a worst case scenario exists, and its density is $p_{\overline{\theta}}$.

f-divergence balls Setting $\mu = \mathbb{P}_0$ again, take now any autonomous integrand for β given by a convex function f as in Section 3, and let

$$H(p) := \int f(p)\, d\mathbb{P}_0. \tag{57}$$

Then the set Γ of distributions given by (24) is equal to the f-divergence ball $\{\mathbb{P} : D_f(\mathbb{P}||\mathbb{P}_0) \leq k\}$ if f is cofinite, while if $f'(+\infty) := \lim_{s\to\infty} f(s)/s$ is finite, Γ is a proper subset of that ball. We will focus on Γ defined by (24) anyway.

If f is not cofinite then $f^*(x) = +\infty$ for $x > f'(+\infty)$, hence

$$K(\theta_1, \theta_2) = \int f^*(\theta_1 + \theta_2 X)\, d\mathbb{P}_0$$

is infinite when $\theta_2 < 0$, unless $m := \operatorname{ess\,inf}(X)$ is finite. By Corollary 3 of Theorem 2, this means that the functional (57) can be adequate for assigning model risk only if f is cofinite or if X is essentially bounded below. In the latter case, (θ_1, θ_2) with $\theta_2 < 0$ belongs to $\operatorname{int}\operatorname{dom} K$ if and only if $\theta_1 + \theta_2 m < f'(+\infty)$.

The most poular f-divergences are the *power divergences*, defined by

$$f_\alpha(s) := [s^\alpha - \alpha(s-1) - 1]/[\alpha(\alpha-1)], \quad \alpha \in \mathbb{R}.$$

Formally, f_α is undefined if $\alpha = 0$ or $\alpha = 1$, but the definition is commonly extended by limiting, thus

$$f_0(s) := \log s + s - 1, \qquad f_1(s) := s \log s - s + 1.$$

This means that also $D_{f_0}(\mathbb{P}||\mathbb{P}_0) = I(\mathbb{P}_0||\mathbb{P})$ and $D_{f_1}(\mathbb{P}||\mathbb{P}_0) = I(\mathbb{P}||\mathbb{P}_0)$ are regarded as power divergences. Note that the function f_α is cofinite if and only if $\alpha \geq 1$, and $f'_\alpha(+\infty) = 1/(1-\alpha)$ if $\alpha < 1$.

Let us determine the family of functions

$$p_\theta(r) = (f^*_\alpha)'(\theta_1 + \theta_2 X(r)), \quad \theta = (\theta_1, \theta_2) \in \Theta \tag{58}$$

that contains the worst case densities in power divergence balls, more exactly, in (24) with $f = f_\alpha$. Since $f'_\alpha(s) = [s^{\alpha-1} - 1]/(\alpha - 1)$ grows from $-\infty$ to $1/(1-\alpha)$ if $\alpha < 1$ or from $1/(1-\alpha)$ to $+\infty$ if $\alpha > 1$, as s runs over $(0, +\infty)$. In the interval $(-\infty, 1/(1-\alpha))$ or $(1/(1-\alpha), +\infty)$, respectively, $(f^*_\alpha)'$ is the inverse function of f'_α, thus

$$(f^*_\alpha)'(x) = [x(\alpha-1)+1]^{1/(\alpha-1)} \quad \text{if} \quad \alpha < 1,\ x < 1/(1-\alpha) \quad \text{or} \quad \alpha > 1,\ x > 1/(1-\alpha).$$

Clearly, $(f^*_\alpha)'(x)$ does not exist if $\alpha < 1$ and $x \geq 1/(1-\alpha)$, while if $\alpha > 1$ and $x \leq 1/(1-\alpha)$ then $(f^*_\alpha)'(x) = 0$. This gives a simple formula for the functions p_θ in (58). Unlike for the relative entropy case, however, no explicit condition is available for $\int p_\theta d\mathbb{P}_0 = 1$, and the two equations in Theorem 1 cannot be reduced to one.

Bregman balls In the special case $\mu = \mathbb{P}_0$, the Bregman distance (6) reduces to f-divergence: If f is a non-negative convex function with $f(1) = 0$ and differentiable at $s = 1$ then $\Delta_f(s, 1) = f(s)$, consequently

$$B_{f,\mathbb{Q}}(\mathbb{P}, \mathbb{Q}) = D_f(\mathbb{P}||\mathbb{Q}) \text{ for } \mathbb{P} \ll \mathbb{Q}.$$

Hence, in this subsection, μ is taken different from $\mathbb{P}_0$; for simplicity, f is assumed differentiable. To obtain for D in (2) resp. H in (25) the Bregman distance $B_{f,\mu}$ of (4), we choose the non-autonomous integrand

$$\beta(r, s) = f(s) - f(p_0(r)) - f'(p_0(r))(s - p_0(r)).$$

To make sure that this meets the assumptions on β, in case $f'(0) = -\infty$ we assume that the default density p_0 is μ-a.e. positive; this assumption is not needed if $f'(0) > -\infty$.

By Csiszár and Matúš [2012, Lemma 2.6], the convex conjugate of β with respect to s equals

$$\beta^*(r, x) = f^*(x + f'(p_0(r))) - f^*(f'(p_0(r))).$$

The function K from (36) equals

$$K(\theta) := \int_\Omega \left[f^*(\theta_1 + \theta_2 X(r) + f'(p_0(r))) - f^*(f'(p_0(r))\right] d\mu(r).$$

The family $\{p_\theta(r) : \theta \in \Theta\}$ is formed by the (non-negative) functions

$$p_\theta(r) = \beta^{*\prime}(\theta_1 + \theta_2 X(r)) = f^{*\prime}[\theta_1 + \theta_2 X(r) + f'(p_0(r))].$$

Note that while the case of Bregman balls is covered by our general results, it is not apparent that the current special form of β would substantially simplify their application.

8 Evaluation of divergence preferences

Finally, we briefly address divergence preferences, i.e., the problem (3) which, in the framework of Section 6, is simpler than the minimization of $H(p)$ over the set (24). Divergence preferences include as special case the multiplier preferences of Hansen and Sargent [2001], when we choose the relative entropy I for D. Maccheroni et al. [2006] choose for D the more general weighted f-divergences

$$D_f^w(\mathbb{P}, \mathbb{P}_0) := \begin{cases} \int_\Omega w(r) f\left(\frac{d\mathbb{P}}{d\mathbb{P}_0}(r)\right) d\mathbb{P}_0(r) & \text{if } \mathbb{P} \ll \mathbb{P}_0, \\ +\infty & \text{otherwise,} \end{cases} \tag{59}$$

where w is a normalised, non-negative weight function.

Below, more generally, the role of D is given to any convex functional as in (25). Introducing a new convex integrand and intergal functional by

$$\tilde{\beta}(r,s) := X(r)s + \lambda\beta(r,s), \quad \tilde{H}(p) := \int \tilde{\beta}(r,p(r))d\mu(r),$$

(where $\lambda > 0$ is fixed), we can write

$$W := \inf_{p:\int pd\mu=1} \left[\int Xpd\mu + \lambda H(p)\right] = \inf_{p:\int pd\mu=1} \tilde{H}(p). \tag{60}$$

Thus, the problem is to minimize the functional $\tilde{H}(p)$ under the single constraint $\int pd\mu = 1$.

In analogy to (35), consider

$$\tilde{J}(a) := \inf_{p:\int pd\mu=a} \tilde{H}(p), \quad a \in \mathbb{R}.$$

Note that $\tilde{\beta}$ meets the basic assumptions on β (though (28) does not hold for $\tilde{H}$), and that

$$(\tilde{\beta})^*(r,x) = \sup_s [xs - X(r)s - \lambda\beta(r,s)] = \lambda\beta^*\left(r, \frac{x - X(r)}{\lambda}\right).$$

It follows by [Csiszár and Matúš, 2012, Theorem 1.1] that the convex conjugate of $\tilde{J}$ equals

$$\tilde{K}(\theta) := \int (\tilde{\beta})^*(r,\theta)d\mu(r) = \lambda \int \beta^*\left(r, \frac{\theta - X(r)}{\lambda}\right) d\mu(r), \quad \theta \in \mathbb{R},$$

or, with the notation (36),

$$\tilde{J}^*(\theta) = \tilde{K}(\theta) = \lambda K(\frac{\theta}{\lambda}, -\frac{1}{\lambda}), \quad \theta \in \mathbb{R}.$$

As the interior of dom $\tilde{J}$ is $(0,+\infty)$, it follows that $\tilde{J}(a) = \tilde{K}^*(a)$ for each $a > 0$. In particular,

$$\begin{aligned} W = \tilde{J}(1) = \tilde{K}^*(1) &= \sup_{\theta\in\mathbb{R}} (\theta - \tilde{K}(\theta)) = \sup_{\theta\in\mathbb{R}} \left[\theta - \lambda K\left(\frac{\theta}{\lambda}, -\frac{1}{\lambda}\right)\right] \\ &= \lambda \sup_{\theta_1\in\mathbb{R}} \left[\theta_1 - K\left(\theta_1, -\frac{1}{\lambda}\right)\right]. \end{aligned} \tag{61}$$

Proposition 5. *The necessary and sufficient condition for $W > -\infty$ in* (60) *is the existence of $\theta_1 \in \mathbb{R}$ with*

$$(\theta_1, -1/\lambda) \in \text{dom } K, \tag{62}$$

and then

$$W = \lambda \sup_{\theta_1} \left[\theta_1 - K(\theta_1, -1/\lambda)\right]. \tag{63}$$

If for some $\theta = (\theta_1, -1/\lambda)$ as in (62) the function p_θ in (38) has integral equal to one, then θ_1 attains the maximum in (63), and $p = p_\theta$ attains the minimum in (60). Otherwise, among the numbers θ_1 satisfying (62) there exists a largest one $\theta_{1\,\max}$, and p_θ with $\theta = (\theta_{1\,\max}, -1/\lambda)$ has integral less than one; then $\theta_1 = \theta_{1\,\max}$ attains the maximum in (63).

Proof. Clearly, $W = \tilde{J}(1) > -\infty$ if and only if $\tilde{J}$ never equals $-\infty$, thus its conjugate $\tilde{K}$ is not identically $+\infty$; by the formula for $\tilde{K}$, this proves the first assertion. The second assertion follows from (61). As the supremum in (63) is the same as the supremum defining $\tilde{K}^*(1)$ in (61) (with θ/λ substituted by θ_1), the next assertion follows from the simple instance of [Csiszár and Matúš, 2012, Lemma 4.10] used in Remark 3 (note that the function $(\beta^*)'(r, \theta)$ there, replacing β by $\tilde{\beta}$ and θ by $\theta_1\lambda$, gives the function p_θ in the Proposition). For the last assertion, recall that the maximum in the definition of $\tilde{K}^*(1)$, and therefore in (63), is always attained, because $a = 1$ is in the interior of $\operatorname{dom} \tilde{K}^*$ (as in Remark 3). Then the (left) derivative by θ_1 of $K(\theta_1, -1/\lambda)$ at the maximiser, say θ_1^*, has to be ≤ 1, and the strict inequality can hold only if $\theta^* = \theta_{1\,\max}$. As the mentioned derivative equals the integral of p_{θ^*} with $\theta^* = (\theta_1^*, -1/\lambda)$, this completes the proof. $\square$

Evaluation of multiplier preferences As an example apply Proposition 5 to reproduce a result of Hansen and Sargent [2001]. We evaluate the objective function of an agent with multiplier preferences (3) choosing for D the relative entropy. This corresponds to the choice $\beta(r, s) = s \log s - s + 1$, and $\mu = \mathbb{P}_0$. In this case, the condition for $W > -\infty$ in Proposition 5 becomes $-1/\lambda \in \operatorname{dom} \Lambda$. Under that condition, the function p_θ with

$$\theta = (\overline{\theta}_1, -1/\lambda), \quad \overline{\theta}_1 = -\Lambda(-\frac{1}{\lambda})$$

has integral equal to one, hence the Proposition gives that this p_θ, namely the member $\mathbb{P}(-1/\lambda)$ of the exponential family with parameter value $-1/\lambda$, attains the minimum in the definition (60) of W. It also follows that

$$W = \lambda\overline{\theta}_1 = -\lambda\Lambda(-\frac{1}{\lambda}).$$

References

A. Ahmadi-Javid. Entropic Value at Risk: a new coherent risk measure. *Journal of Optimizaton Theory and Applications*, 155(3):1105–1123, 2011.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society Ser. B*, 28:131–142, 1966.

F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205, 1963.

P. Artzner, F. Delbaen, J.-M. Ebner, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

M. Avellaneda and A. Paras. Managing the volatility risk of portfolios of derivative securities: the lagrangian uncertain volatility model. *Applied Mathematical Finance*, 3:21–52, 1996.

F. Barillas, L. P. Hansen, and T. J. Sargent. Doubts or variability? *Journal of Economic Theory*, 144:2388–2418, 2009.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory (Wiley series in probability & mathematical statistics)*. Wiley, 1978.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17:449–476, 2007.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

T. Breuer and I. Csiszár. Systematic stress tests with entropic plausibility constraints. *Journal of Banking and Finance*, 2012. In press. `http://dx.doi.org/10.1016/j.jbankfin.2012.04.013`.

G. C. Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal of Optimization*, 18:853–877, 2007.

R. Casadesus-Masanell, P. Klibanoff, and E. Ozdenorem. Maxmin expected utility of Savage acts with a set of priors. *Journal of Economic Theory*, 92:35–65, 2000.

R. Cont. Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, 16:519–547, 2006.

M. Crouhy, D. Galai, and R. Mark. Model risk. *Journal of Financial Engineering*, 7:267–288, 1998.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 8: 85–108, 1963.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics*, 19 (4):2032–2066, 1991.

I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011. First edition at Academic Press 1981.

I. Csiszár and F. Matúš. On minimization of entropy functionals under moment constraints. *Kybernetika*, 48:637–689, 2012.

F. Delbaen. Coherent risk measures on general probability spaces. In K. Sandmann and P. J. Schönbucher, editors, *Advances in Stochastics and Finance: Essays in Honour of Dieter Sondermann*, pages 1–37. Springer, 2002.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics*. Springer, 2nd edition, 1998.

H. Föllmer and A. Schied. Robust preferences and convex measures of risk. In K. Sandmann and P. Schonbucher, editors, *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*. Springer, 2002.

H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*, volume 27 of *de Gruyter Studies in Mathematics*. Walter de Gruyter, 2nd edition, 2004.

C. Friedman. Confronting model misspecification in finance: Tractable collections of scenario probability measures for robust financial optimization problems. *International Journal of Theoretical and Applied Finance*, 5: 33–54, 2002a.

C. Friedman. Conditional value-at-risk in the presence of multiple probability measures. *Journal of Risk*, 4:69–92, 2002b.

R. Gibson. *Model Risk: Concepts, Calibration and Pricing*. Risk Books, 2000.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

A. Golan, G. G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, 1996.

B. Grechuk, A. Molyboha, and M. Zabarankin. Maximum entropy principle with general deviation measures. *Mathematics of Operations Research*, 34:445–467, 2009.

P. D. Grunwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 32 (4):1367–1433, 2004.

L. Hansen and T. Sargent. Robust control and model uncertainty. *American Economic Review*, 91:60–66, 2001.

L. P. Hansen and T. Sargent. *Robustness*. Princeton University Press, 2008.

L. P. Hansen and T. J. Sargent. Recursive robust estimation and control without commitment. *Journal of Economic Theory*, 136:1–27, 2007.

E. T. Jaynes. Information theory and statistical mechanics i. *Physical Review*, 106:620–630, 1957.

J. Kerkhof, B. Melenberg, and H. Schumacher. Model risk and capital reserves. *Journal of Banking and Finance*, 34:267–279, 2010.

S. Kullback. *Information Theory and Statistics*. Wiley, 1959.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, 1987.

F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74:1447–1498, 2006.

M. Nutz and H. M. Soner. Superheding and dynamic risk measures under volatility uncertainty. *SIAM Journal of Control and Optimization*, 50: 2065–2089, 2012.

J. B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal in Inexact Reasoning*, 4:183–223, 1990.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 1997.

J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, 1980. Correction IT-29 (1983), 942–943.