Incorporating views in Mathematical Models using Entropy Approach

Santanu Dey
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai, India - 400 005
dsantanu@tsc.tifr.res.in

Sandeep Juneja
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai, India - 400 005
juneja@tifr.res.in

ABSTRACT

A mathematical model based on historical data or general past experience may at times be an unsatisfactory model for the future. One way to come up with a more accurate model is to explicitly incorporate in it views that are believed to better reflect the future. We address this issue by letting μ denote the original probability measure of a mathematical model. We then search for a probability measure ν that minimizes a distance measure with respect to μ and satisfies certain user specified views or constraints. We consider Kullbach- Liebler distance as well as other f-divergences as measures of distance between the probability measures. We show that under the KL distance, our optimization problem may lack a closed form solution when views involve fat tailed distributions. This drawback may be corrected if another 'polynomial f-divergence' is used. On a small example, we compare the nature of the optimal probability measure when the objective function corresponds to total variation distance, KL distance and polynomial f-divergence. We also discuss the optimal solution structure under these distances when the views are on probabilities of underlying events. In particular, we discuss using a popular example, how tail event views, so important in current environment, may be efficiently incorporated using appropriate importance sampling techniques.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic Algorithms (including Monte-Carlo) Stochastic Processes

; I.6.1 [**Computing Methodologies**]: Simulation Theory

General Terms

economics, management, theory

1. INTRODUCTION

Mathematical modeling under uncertainty is an imprecise science. While it may be difficult to come up with models that accurately represent the future, one can make efforts to avoid grossly inaccurate models. For instance, building mathematical models based solely on past data or practices may misrepresent future in a fast changing environment. One way to improve the quality of a mathematical model is to modify it by incorporating subjective views of experts. This requires a reasonable and tractable methodology to arrive at the modification. For instance, if a model assigns an expectation of 10 to a particular random variable, when an accurate view based on experience of experts suggests that this should be 15, then there is a need to come up with a reasonable and simple modification of the model that achieves this mean.

In this paper we let μ denote a probability measure that represents the original or the 'prior' probability model. We search for a new model ν that is constrained to lie in a set associated with the views held by experts, and that is closest to μ in the sense that it minimizes a certain notion of distance between probability measures.

This line of reasoning is motivated from the evolving literature in finance based on the pioneering work of Black and Litterman[1]. They consider investment models where the subjective views of portfolio managers are used to update models of the market using ideas from Bayesian analysis. There work focussed on Gaussian framework with views restricted to linear combinations on expectations of returns from different securities. Since then a number of variations and improvements have been suggested (see, e.g., [8], [9],[10]). More recently, Meucci [7] proposed entropy pooling (EP) where the original model can involve general distributions and views can be on any characteristic of the probability model. He proposed that the new probability measure be selected that minimizes the Kulbach-Leibler entropy with respect to the original measure subject to the views modeled as constraints. He however does not characterize the optimal solution in such settings. Instead, he focuses on approximating the original distribution by a discrete one generated via Monte-Carlo sampling (or from data). Then, through solving a convex programming problem, he adjusts weights to these samples so that they minimize the Kulbach-Leibler distance (also known as I-divergence) from the original sampled distribution while satisfying the view constraints.

Clearly, this approach is applicable more broadly to mathematical models outside the field of finance. In this paper we build upon ideas proposed by Meucci [7]. We note that

for a broad class of views, the optimal solution to the KL distance minimization can be characterized as a probability measure obtained by suitably exponentially twisting the original measure. This measure is known in literature as the Gibbs measure and our analysis is based on the well known ideas involved in Gibbs conditioning principle (see, for instance, [4]). We further note that such a characterization may fail when the underlying distributions do not have appropriate exponential moments. This is an important case in practice as fat-tailed distributions do not have exponential moments and are increasingly used in modeling in diverse areas such as communications networks, insurance and finance. We show that one reasonable way to get a good change of measure that incorporates views in this setting is by replacing KL distance by a suitable 'polynomial f-divergence' as an objective in our optimization problem. Our main result is the characterization of the optimal solution under this new objective. We focus on a class of f-divergence functions and heuristically argue that KL distance may be considered a limiting case of this class.

Note that the total variation distance can be another measure of distance between probabilities that we may consider as an objective in our optimization problem. We show through a simple example some qualitative differences in solutions corresponding to different objective functions that may be used in our optimization problem.

Finally, we focus on the case where the views are expressed as constraints on the probability values of disjoint sets. Here, we note that the optimal solution is the same whether the objective is KL distance or f-divergence. Furthermore, it has a simple representation in terms of the original probability measure.

We further specialize to the case where the views are associated with tail events and the performance measure of interest is also a tail event. There, through an example associated with the large deviation probability of a random walk, we argue that naive simulation will do a poor job of estimating the performance measure of interest, while appropriate importance sampling performs this estimation asymptotically efficiently.

In Section 2 we discuss how views may be incorporated when the objective is to minimize KL distance. In Section 3, we extend this to the case where the objective is to minimize polynomial f-divergence. In Section 4, we consider different objectives on a simple example and discuss qualitative differences in the resulting optimal probability measures. In Section 5, we consider the specialized problem where the views are on probabilities of disjoint sets. In Section 6, we discuss through a popular example how tail views may be incorporated and the resulting computational issues. Finally, we end with a brief conclusion in Section 7.

2. INCORPORATING VIEWS USING KL DISTANCE

In this section we characterize the optimal probability measure that minimizes the KL distance with respect to the original probability measure subject to the views expressed as moment constraints of specified functions.

Some notation and basic concepts are needed to support our analysis. Let $(\Omega, \mathcal{F}, \mu)$ denote the underlying probability space. Let $\mathcal{P}$ be the set of all probability measures on $(\Omega, \mathcal{F})$. For any $\nu \in \mathcal{P}$ the relative entropy of ν w.r.t μ or equivalently the KL distance of ν w.r.t μ is defined as

$$H(\nu \mid \mu) := \int \log\{\frac{d\nu}{d\mu}\}\, d\nu$$

if ν is absolutely continuous with respect to μ and $log(\frac{d\nu}{d\mu})$ is integrable. The KL distance $H(\nu \mid \mu) = +\infty$ otherwise. See, for instance [2], for concepts related to relative entropy.

Let $\mathcal{P}(\mu)$ be the set of all probability measures which are absolutely continuous w.r.t. μ. Let $\psi : \Omega \to \mathbb{R}$ be a measurable function such that $\int |\psi| e^{\psi} d\mu < \infty$. Let

$$\Lambda(\psi) := \log \int e^{\psi}\, d\mu \ \in (-\infty, +\infty]$$

denote the logarithmic moment generating function of ψ w.r.t μ.

Then, it is well known that

$$\Lambda(\psi) = \sup_{\nu \in \mathcal{P}(\mu)} \{\int \psi\, d\nu - H(\nu \mid \mu)\}.$$

Furthermore, this supremum is attained at $\nu*$ given by:

$$\frac{d\nu*}{d\mu} = \frac{e^{\psi}}{\int e^{\psi}\, d\mu}. \tag{1}$$

(see for instance, [5]).

In our optimization problem we look for a probability measure $\nu \in \mathcal{P}(\mu)$ that minimizes the KL distance w.r.t. μ. We restrict our search to probability measures that satisfy moment constraints such as

$$\int g_i\, d\nu \geq c_i,$$

and/or

$$\int g_i\, d\nu = c_i,$$

where each g_i is a measurable function.

For instance, views on probability of certain sets can be modeled by setting g_i's as indicator functions of those sets. If our underlying space supports random variables $(X_1, \ldots, X_n)$ under the probability measure μ, each g_i may denote a constraint on the moments of the random variables.

Formally, our optimization problem $\mathbf{O_1}$ is:

$$\min_{\nu \in \mathcal{P}(\mu)} \int log(\frac{d\nu}{d\mu})\, d\nu \tag{2}$$

subject to the constraints

$$\int g_i\, d\nu \geq c_i, \tag{3}$$

for $= i = 1, \ldots, k_1$ and

$$\int g_i\, d\nu = c_i, \tag{4}$$

for $= i = k_1 + 1, \ldots, k$. Here k_1 can take any value between 0 and k.

The solution to this is characterized by the following assumption:

ASSUMPTION 1. *There exist $\lambda_i \geq 0$ for $i = 1, \ldots, k_1$, and $\lambda_{k_1+1}, ..., \lambda_k \in \mathbb{R}$ such that*

$$\int e^{\sum_i \lambda_i g_i} \, d\mu < \infty$$

and the probability measure ν^0 given by

$$\nu^0(A) = \int_A \frac{e^{\sum_i \lambda_i g_i} \, d\mu}{\int e^{\sum_i \lambda_i g_i} \, d\mu} \tag{5}$$

for all $A \in \mathcal{F}$ satisfies the constraints (3) and (4). Furthermore, the complementary slackness conditions

$$\lambda_i (c_i - \int g_i \, d\nu) = 0,$$

hold for $= i = 1, \ldots, k_1$.

The following theorem follows:

THEOREM 1. *Under Assumption (1), ν^0 is an optimal solution to $\mathbf{O_1}$.*

Proof: $\mathbf{O_1}$ is equivalent to maximizing $-H(\nu \mid \mu) = -\int log(\frac{d\nu}{d\mu}) \, d\nu$ subject to the constraints (3) and (4). The Lagrangian for the above maximization problem is:

$$\begin{aligned} \mathcal{L} &= \sum_i \lambda_i \int g_i \, d\mu + (-H(\nu \mid \mu)) \\ &= \sum_i \int \psi \, d\mu - H(\nu \mid \mu) \end{aligned}$$

where we have put $\psi = \sum_i \lambda_i g_i$. Then by (1) and the preceding discussion, it follows that ν^0 maximizes $\mathcal{L}$. By Lagrangian duality, due to Assumption 1, ν^0 also solves $\mathbf{O_1}$. □

EXAMPLE 1. Suppose that under μ, random variables $X = (X_1, \ldots, X_n)$ have a multivariate Gaussian distribution $N(a, \Sigma)$, that is, with mean $a \in \Re^n$ and variance covariance matrix $\Sigma \in \Re^{n \times n}$. If a view corresponds to their mean vector equalling $\hat{a}$, then this can achieved by a new probability measure ν^0 obtained by exponentially twisting μ by a vector $\lambda \in \Re^n$ such that

$$\lambda = (\Sigma^{-1})(\hat{a} - a).$$

Then, under ν^0, X is $N(\hat{a}, \Sigma)$ distributed.

REMARK 1. Note that the constraints (3) and (4) do not, for instance, capture views on variance and correlations of the underlying variables. We discuss one way to incorporate such constraints. As an illustration, consider the case where X is a rv under μ and there is a single constraint that under the new measure, its variance equals s. This may be incorporated by setting

$$\int x d\nu(x) = a$$

and

$$\int x^2 d\nu(x) = a^2 + s$$

Then along with λ one also searches for a such that Assumption 1 holds.

3. INCORPORATING VIEWS USING F-DIVERGENCE

In this section, we first note through an example involving a fat-tailed distribution, that Assumption 1 may not hold in certain settings. This motivates the use of other f-divergences as an objective to our problem. We then characterize the optimal solution under polynomial f-divergence. This may exist even when Assumption 1 does not hold.

EXAMPLE 2. Suppose that under μ, non-negative random variable X has a Pareto distribution with probability density function

$$f(x) = \frac{\alpha - 1}{(1 + x)^\alpha}$$

for $x \geq 0$. The mean under this pdf equals $1/(\alpha - 2)$. Suppose the view is that the mean should be

$$\geq 2/(\alpha - 2).$$

It is well known and easily checked that

$$\frac{\int x e^{\lambda x} f(x) dx}{\int e^{\lambda x} f(x) dx}$$

is an increasing function of λ that equals ∞ for $\lambda > 0$.

Hence, Assumption 1, does not hold for Example 2. This motivates the need for another objective function that helps find a probability distribution close to μ that satisfies constraints such as (3) and (4) and does not require that the optimal solution be obtained using exponential twisting. We now address this issue using polynomial f-divergence. We first define general f-divergence as introduced in [3].

DEFINITION 1. *Let $f : (0, \infty) \to \mathbb{R}$ be strictly convex function. The f-divergence of a probability measure ν w.r.t another probability measure μ equals*

$$I_f(\nu \mid \mu) := \int f\left(\frac{d\nu}{d\mu}\right) d\mu$$

if ν is absolutely continuous and $f(\frac{d\nu}{d\mu})$ is integrable w.r.t. μ. Otherwise, $I_f(\nu \mid \mu)$ equals ∞.

Note that KL distance corresponds to the case $f(u) = u \log u$. This, as we mentioned earlier, is also known as I-divergence. Other popular examples of f include

$$f(u) = -\log u, \;\; f(u) = u^{\beta+1}, \;\; \beta > 0, \;\; f(u) = e^u.$$

In this section we consider $f(u) = u^{\beta+1},\ \beta > 0$ and refer to it as *polynomial f-divergence*. That is, we focus on

$$I_f(\nu \mid \mu) = \int \left(\frac{d\nu}{d\mu}\right)^{\beta} d\nu = \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu.$$

It is easy to see using Jensen's inequality that

$$\min_{\nu \in \mathcal{P}(\mu)} \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu$$

is achieved by $\nu = \mu$.

Our optimization problem $\mathbf{O_2}$ may be stated as:

$$\min_{\nu \in \mathcal{P}(\mu)} \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu \tag{6}$$

subject to (3) and (4).

In the following assumption we specify the solution form:

ASSUMPTION 2. *There exist $\lambda_i \geq 0$ for $i = 1, \ldots, k_1$, and $\lambda_{k_1+1}, ..., \lambda_k \in \mathbb{R}$ such that $\sum_i \lambda_i g_i + 1 \geq 0$,*

$$\int (\sum_i \lambda_i g_i + 1)^{1/\beta}\, d\mu < \infty,$$

and the probability measure ν^1 given by

$$\nu^1(A) = \int_A \frac{(\sum_i \lambda_i g_i + 1)^{1/\beta}\, d\mu}{\int (\sum_i \lambda_i g_i + 1)^{1/\beta}\, d\mu} \tag{7}$$

for all $A \in \mathcal{F}$ satisfies the constraints (3) and (4). Furthermore, the complementary slackness conditions

$$\lambda_i (c_i - \int g_i\, d\nu) = 0,$$

hold for $= i = 1, \ldots, k_1$.

THEOREM 2. *Under Assumption(2), ν^1 is an optimal solution to $\mathbf{O_2}$.*

Let $\xi = \int (\sum_i \lambda_i g_i + 1)^{1/\beta}\, d\mu$. and $\hat{\lambda}_i = (\beta+1)\lambda_i / \xi^{\beta}$.

Proof: Consider the Lagrangian $\mathcal{L}(\nu)$ for $\mathbf{O_2}$ defined as

$$\int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu - \sum_i \hat{\lambda}_i (\int g_i\, d\nu - c_i). \tag{8}$$

We first argue that $\mathcal{L}(\nu)$ is a convex function of ν. Given that $\lambda_i \int g_i\, d\nu$ are linear in ν, it suffices to show that $\int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu$ is a convex function of ν.

Note that for $0 \leq s \leq 1$,

$$\int \left(\frac{d(s\nu_1 + (1-s)\nu_2)}{d\mu}\right)^{\beta+1} d\mu$$

equals

$$\int \left(s\frac{d\nu_1}{d\mu} + (1-s)\frac{d\nu_2}{d\mu}\right)^{\beta+1} d\mu,$$

which in turn is dominated by

$$\int \left(s(\frac{d\nu_1}{d\mu})^{\beta+1} + (1-s)(\frac{d\nu_2}{d\mu})^{\beta+1}\right) d\mu,$$

which equals

$$s\int \left(\frac{d\nu_1}{d\mu}\right)^{\beta+1} d\mu + (1-s)\int \left(\frac{d\nu_2}{d\mu}\right)^{\beta+1} d\mu.$$

Therefore, the Lagrangian $\mathcal{L}(\nu)$ is a convex function of ν.

We now prove that $\mathcal{L}(\nu)$ is minimized at ν^1. For this, all we need to show is that we cannot improve by moving away from ν^1 in any feasible direction. Since, ν^1 satisfies all the constraints, the result then follows. We now show this.

Let f denote $\frac{d\nu}{d\mu}$ and $f^1 = \frac{d\nu^1}{d\mu}$. Note that (8) may be re-expressed as

$$\int (f^{\beta+1} - \sum_i \hat{\lambda}_i g_i f)\, d\mu.$$

For any $\nu \in \mathcal{P}(\mu)$ and $t \in [0,1]$ consider the function

$$G_\nu(t) = \mathcal{L}((1-t)\nu^1 + t\nu).$$

This in turn equals

$$\int [\{(1-t)f^1 + tf\}^{\beta+1} - \sum_i \hat{\lambda}_i g_i \{(1-t)f^1 + tf\}]\, d\mu.$$

We now argue that $\frac{d}{dt}_{t=0} G_\nu(t) = 0$. Then from this and convexity of $\mathcal{L}$, the result follows.

To see this, note that $\frac{d}{dt}_{t=0} G_\nu(t)$ equals

$$\int \{(\beta+1)(f^1)^{\beta} - \sum_i \hat{\lambda}_i g_i\}(f - f^1)\, d\mu. \tag{9}$$

Due to the definition of f_1 and $\hat{\lambda}_i$, it follows that the term inside the braces in the integrand in (9) is a constant. Since both ν^1 and ν are probability measures, and $\int (f - f^1)\, d\mu = 0$, therefore $\frac{d}{dt}_{t=0} G_\nu(t) = 0$ and the result follows. □

3.1 Convergence of polynomial f-divergence to KL distance

In this section we briefly and heuristically discuss the convergence of solution of $\mathbf{O_2}$ to that of $\mathbf{O_1}$ as $\beta \to 0$ in a simple setting. A rigorous proof in a general setting will appear in a more elaborate version of this paper.

Consider a probability space where a random variable X has pdf f. We search for a pdf $\hat{f}$ that satisfies a single constraint

$$\int g(x)\hat{f}(x)dx = c \tag{10}$$

for a non-decreasing function g and a constant $c > \int g(x)f(x)dx$.

Let pdf

$$h_\lambda(x) = \frac{\exp(\lambda g(x))f(x)}{\int \exp(\lambda g(x))f(x)dx}$$

be well defined for all λ and let E_h denote the expectation operator under any pdf h. It is well known and easily

checked that $E_{h_\lambda} g(X)$ is a non-decreasing function of λ. We suppose that it is strictly increasing and we let $\lambda^* > 0$ denote the solution to

$$E_{h_\lambda} g(X) = c.$$

Note that h_{λ^*} equals $\hat{f}$ that minimizes the KL-distance with f subject to the constraint (10).

Define the probability density function

$$\psi_{\lambda,\alpha}(x) = \frac{(1+\lambda g(x)/\alpha)^\alpha f(x)}{\int (1+\lambda g(x)/\alpha)^\alpha f(x) dx}.$$

Note that $E_{\psi_{\lambda,\alpha}} g(X)$ is a non-decreasing function of λ. To see this, differentiate

$$E_{\psi_{\lambda,\alpha}} g(X) = \frac{\int g(x)(1+\lambda g(x)/\alpha)^\alpha f(x) dx}{\int (1+\lambda g(x)/\alpha)^\alpha f(x) dx}$$

w.r.t. λ to get the difference of

$$E_{\psi_{\lambda,\alpha}} \left(\frac{g(X)^2}{1+\lambda g(X)/\alpha} \right)$$

and

$$E_{\psi_{\lambda,\alpha}} g(X) E_{\psi_{\lambda,\alpha}} \left(\frac{g(X)}{1+\lambda g(X)/\alpha} \right).$$

This is the covariance of positively associated random variables and hence is non-negative. We assume that it is strictly positive for each (λ, α) so that $E_{\psi_{\lambda,\alpha}} g(X)$ is a strictly increasing function of λ for each α.

Let $\lambda(\alpha)$ uniquely solve the equation

$$E_{\psi_{\lambda,\alpha}} g(X) = c.$$

Note that under mild conditions,

$$\lim_{\alpha \to \infty} \psi_{\lambda,\alpha}(x) = h_\lambda(x)$$

for all x. We now argue that

$$\lambda(\alpha) \to \lambda^* \tag{11}$$

as $\alpha \to \infty$. It then follows that

$$\psi_{\lambda(\alpha),\alpha}(x) \to h_{\lambda^*}(x)$$

for all x, indicating that the solution to $\mathbf{O_2}$ converges to that of $\mathbf{O_1}$ as $\beta = 1/\alpha \to 0$.

To see (11), first note that under mild conditions,

$$\lim_{\alpha \to \infty} E_{\psi_{\lambda,\alpha}} g(X) = E_{h_\lambda} g(X).$$

This implies that for any $\epsilon > 0$, there exists α_ϵ such that for all $\alpha \geq \alpha_\epsilon$,

$$|E_{\psi_{\lambda^*,\alpha}} g(X) - E_{\psi_{\lambda(\alpha),\alpha}} g(X)| < \epsilon.$$

(Since, $E_{\psi_{\lambda(\alpha),\alpha}} g(X) = E_{h_{\lambda^*}} g(X) = c$.)

Now for each α, $E_{\psi_{\lambda,\alpha}} g(X)$ is a continuous and strictly increasing function of λ, (11) must hold.

4. COMPARING DIFFERENT OBJECTIVES

Given that in many examples one can use KL distance as well as polynomial f-divergence as an objective function for arriving at an updated probability measure, it is natural to compare the optimal solutions in these cases. Note that the total variation distance between two probability measures μ and ν equals

$$\sup\{\mu(A) - \nu(A) | A \in \mathcal{F}\}.$$

This may also serve as an objective function in our search for a reasonable probability measure that incorporates expert views and is close to the original probability measure. This has an added advantage of being a metric (e.g., it satisfies the triangular inequality).

We now compare these three different types of objectives to get a qualitative flavor of the differences in the tail distribution of the corresponding optimal solutions.

Suppose that the rv X is exponentially distributed with rate α. Then its pdf equals

$$f(x) = \alpha e^{-\alpha x},\ x \geq 0.$$

Now suppose that our view is that the expectation under the posterior measure is $\tilde{E}(X) = \int x \tilde{f}(x)\, dx = 1/\gamma$. We assume $1/\gamma > 1/\alpha$. This will often be the case with conservative views.

KL Distance: When the objective function is to minimize KL distance, the optimal solution is obtained as an exponentially twisted distribution that satisfies the desired constraint. It is easy to see that exponentially twisting an exponential distribution with rate α by an amount θ leads to another exponential distribution with rate $\beta - \theta$ (assuming that $\theta < \beta$). Therefore, in our case

$$\tilde{f}(x) = \gamma e^{-\gamma x},\ x \geq 0.$$

satisfies the given constraint and is a solution to this problem. Note here that the tail distribution function equals $\exp(-\gamma x)$ and is heavier than $\exp(-\alpha x)$, the tail distribution of the original distribution of X.

Polynomial f-divergence: Now consider the case where the objective corresponds to a polynomial f-divergence with parameter equal to β, i.e, it equals

$$\int \left(\frac{\tilde{f}(x)}{f(x)} \right)^{\beta+1} f(x) dx.$$

Under this objective, the optimal pdf

$$\tilde{f}(x) = \frac{(\lambda x + 1)^{1/\beta} \alpha e^{-\alpha x}}{\int (\lambda x + 1)^{1/\beta} \alpha e^{-\alpha x}\, dx}$$

where $\lambda > 0$ is chosen so that the mean under $\tilde{f}$ equals $1/\gamma$.

While this may not have a closed form solution, it is clear that on a logarithmic scale, $\tilde{f}(x)$ is asymptotically similar to $\exp(-\alpha x)$ as $x \to \infty$ and hence has a lighter tail than the solution under the KL distance.

Total Variation Distance: Under total variation distance as an objective, we show that given any ε, we can find a new density function $\tilde{f}$ so that the mean under the new distribution equals $1/\gamma$ while the total variation distance is less than ε. Thus the optimal value of the objective function is zero, although there may be no pdf that attains this value.

To see this, consider,

$$\tilde{f}(x) = \varepsilon/2\frac{I_{(a-\delta,a+\delta)}}{2\delta} + (1-\varepsilon/2)\alpha e^{-\alpha x},\ x \geq 0.$$

Then,

$$\tilde{E}(X) = \int x\tilde{f}(x)\,dx = (\varepsilon/2)a + \frac{1-\varepsilon/2}{\alpha}.$$

Thus, given any ε if we select

$$a = \frac{1/\gamma - 1/\alpha}{\varepsilon/2} + 1/\alpha$$

we see that

$$\tilde{E}(X) = (\varepsilon/2)a + \frac{1-\varepsilon/2}{\alpha} = 1/\gamma.$$

We now show that total variation distance between f and $\tilde{f}$ is less than ε. To see this, note that

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/2)P(A)$$

for any set A disjoint from $(a-\delta, a+\delta)$, where the probability P corresponds to the density f. Furthermore, letting $L(S)$ denote the Lebesgue measure of set S,

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/4\delta)L(A) + (\varepsilon/2)P(A)$$

for any set $A \subset (a-\delta, a+\delta)$. Thus, for any set $A \subset (0,\infty)$

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq$$

$$(\varepsilon/4\delta)L(A\bigcap(a-\delta, a+\delta)) + (\varepsilon/2)P(A).$$

Therefore,

$$\sup_A |\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/2)P(A) + (\varepsilon/4\delta)2\delta < \varepsilon.$$

This also illustrates that it may be difficult to have an elegant characterization of solutions under the total variation distance, making the other two as more attractive measures from this viewpoint.

5. VIEWS ON PROBABILITY OF DISJOINT SETS

In this brief section, we consider the case where the views correspond to consistent probability assignments to mutually exclusive and exhaustive set of events. We note that in this setting, objective functions associated with KL distance, polynomial f-divergence and total variation distance give identical results.

Suppose that our views correspond to:

$$\nu(B_i) = \alpha_i,\ i = 1, 2, ...k \text{ where}$$

$$B_i's \text{ are disjoint, } \cup B_i = \Omega \text{ and } \sum_{i=1}^{k} \alpha_i = 1.$$

KL Distance: Then, under the KL distance setting, for any event A, the optimal

$$\nu(A) = \frac{\int_A e^{\sum_i \lambda_i I(B_i)}\,d\mu}{\int e^{\sum_i \lambda_i I(B_i)}\,d\mu} = \frac{\sum_i e^{\lambda_i}\mu(A\cap B_i)}{\sum_i e^{\lambda_i}\mu(B_i)}.$$

Select λ_i so that $e^{\lambda_i} = \alpha_i/\mu(B_i)$. Then it follows that the specified views hold and

$$\nu(A) = \sum_i \alpha_i \mu(A\cap B_i)/\mu(B_i). \tag{12}$$

Polynomial f-divergence: The analysis remains identical when we use polynomial f-divergence with parameter β. Here, we see that optimal

$$\nu(A) = \frac{\sum_i(\lambda_i+1)^{1/\beta}\mu(A\cap B_i)}{\sum_i(\lambda_i+1)^{1/\beta}\mu(B_i)}$$

Again, by setting $(\lambda_i+1)^{1/\beta} = \alpha_i/\mu(B_i)$, (12) holds.

Total Variation Distance: If the objective is the total variation distance, then clearly, the objective function is $\geq \max_i |\mu(B_i) - \alpha_i|$. We now show that ν defined by (12) achieves this lower bound.

To see this, note that

$$\begin{aligned}
|\nu(A) - \mu(A)| &= |\sum_i (\nu(A\cap B_i) - \mu(A\cap B_i))\,| \\
&\leq \sum_i |\nu(A\cap B_i) - \mu(A\cap B_i)| \\
&\leq \sum_i \frac{\mu(A\cap B_i)}{\mu(B_i)}|\alpha_i - \mu(B_i)| \\
&\leq \max_i |\mu(B_i) - \alpha_i|.
\end{aligned}$$

6. RARE EVENT ANALYSIS

In this section we outline through a representative example the computational issues that may arise when views associated with probability of rare events need to be incorporated and when the performance measure of interest may also be a rare event.

Consider a setting where $(X_i : i \geq 1)$ are i.i.d. random variables under μ. Let $S_n = \sum_{i=1}^n X_i$ denote the associated random walk. Suppose that the performance measure of interest is the rare event probability of the event $\{S_n/n \geq a\}$ for $a > EX_i$.

Further suppose that there is a single view from an expert familiar with the first half of the random variables that their tail probability under the posterior measure,

$$\nu(S_{n/2}/(n/2) \geq b) = \beta_n$$

for $b > EX_i$. Then, letting $B = \{S_{n/2}/(n/2) \geq b\}$, and B^c, the complement of B, from Section 5, it follows that for any

set A,

$$\nu(A) = \beta_n \frac{\mu(A \cap B)}{\mu(B)} + (1 - \beta_n) \frac{\mu(A \cap B^c)}{\mu(B^c)}$$

In particular, our interest is in $A = \{S_n/n \geq a\}$. Then, estimating $\nu(A)$ brings up interesting rare event simulation challenges in terms of efficiently estimating the rare event probabilities $\mu(S_n/n \geq a, S_{n/2}/(n/2) \geq b), \mu(S_{n/2}/(n/2) \geq b)$ and $\mu(S_n/n \geq a, S_{n/2}/(n/2) < b)$. Note that using naive simulation to estimate these probabilities can be computationally prohibitive if n is large. Also note that efficient estimation techniques for $\mu(S_{n/2}/(n/2) \geq b)$ are well known, see for instance, [6]. However, efficient estimation techniques for probabilities such as

$$\mu(S_n/n \geq a, S_{n/2}/(n/2) \geq b) \tag{13}$$

and

$$\mu(S_n/n \geq a, S_{n/2}/(n/2) < b) \tag{14}$$

require further analysis.

Here we outline the insights on how rare events may happen in these cases and reasonable importance sampling techniques to efficiently estimate these probabilities. Detailed analysis of the proposed algorithms as well as numerical support would be provided elsewhere.

First we review an efficient importance sampling technique to estimate $\mu(S_n/n \geq a)$. Suppose that F denotes the distribution function of each X_i and

$$\Lambda(\theta) = \log \int e^{\theta x} dF(x)$$

denotes the associated log-moment generating function. Let

$$F_\theta(x) = e^{\theta x - \Lambda(\theta)} dF(x)$$

denote the distribution obtained by exponentially twisting F by an amount θ. Let θ_y denote the exponential twist that makes the mean of X_i equal to y under F_θ. Then, importance sampling that asymptotically efficiently estimates $\mu(S_n/n \geq a)$ corresponds to generating i.i.d. samples of $(X_i : i \geq 1)$ using the distribution F_{θ_a}, whenever such a distribution exists.

Now consider the problem of estimating (13). When $b < a$, this probability can be seen to asymptotically similar to $\mu(S_n/n \geq a)$. This is because, the most likely way the event $\{S_n/n \geq a\}$ happens is that

$$\{S_{n/2}/(n/2) \approx a, (S_n - S_{n/2})/(n/2) \approx a\}.$$

This can be inferred from large deviations analysis in, for instance, [4]. Hence, the importance sampling technique for estimating $\mu(S_n/n \geq a)$ can be shown to be asymptotically optimal for estimating (13) in this case. As discussed earlier, this involves generating samples of $(X_i : i \geq 1)$ using the distribution F_{θ_a}.

When $b \geq a + (a - EX_i)$, then it can be shown that

$$\mu(S_n/n \geq a, S_{n/2}/(n/2) \geq b) \sim \mu(S_{n/2}/(n/2) \geq b).$$

Here, the most likely way the joint event happens is that $S_{n/2}/(n/2) \approx b$ and the remaining random variables evolve in the usual way with mean EX_i. Therefore, importance sampling technique to efficiently estimate $\mu(S_{n/2}/(n/2) \geq b)$ can also be used to efficiently estimate (13), with the caveat that the importance sampling is turned off after generating $S_{n/2}$. Thus, $(X_i : i \leq n/2)$ are generated using the distribution F_{θ_b} and the remaining $X_i's$ are generated using the original distribution function F.

An interesting case corresponds to $a < b < a + (a - EX_i)$. The most likely path for this can be shown to correspond to the intersection of two rare events: $S_{n/2}/(n/2) \approx b$ and $(S_n - S_{n/2})/(n/2) \approx 2a - b$. An importance sampling technique for this corresponds to generating $(X_i : i \leq n/2)$ using the distribution F_{θ_b} and the remaining $X_i's$ are generated using the distribution function $F_{\theta_{2a-b}}$.

Similarly, it is easy to outline the most likely rare event paths and the associated importance sampling techniques for estimating (14) for different relationships between a and b.

7. CONCLUSION

In this paper we built upon an existing methodology to add views by experts to an existing model to arrive at a more accurate model. This problem was addressed in an optimization framework where we searched for a probability measure close to the one corresponding to the existing model while satisfying the specified views modeled as constraints. We characterized the optimal solution when the objective function corresponded to minimizing KL-distance between the two probability measures as well as when it corresponded to polynomial f-divergence. The latter has an advantage of having a closed form solution in settings involving fat-tailed distributions where the KL distance may not have a nice characterization of optimal solution.

We qualitatively compared the solution to the optimization problem in a simple setting when the objective function was KL distance, polynomial f-divergence and total variation distance between two probability measures. Our analysis indicated that in certain settings, KL distance may put more mass in tails compared to polynomial f-divergence, which may penalize tail deviations more.

We also discussed the optimal solution structure under these divergences when the views are on probabilities of underlying events. In particular, we discussed through a popular example, how tail event views may be computationally efficiently incorporated using appropriate importance sampling techniques.

The analysis is presented in this paper through simple examples. These ideas will be presented in a more general framework in a separate paper.

8. REFERENCES

[1] F. Black and R. Litterman. Asset allocation:combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 1990.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Wiley series in Telecommunications, 1999.

[3] I. Csiszar. A class of measure of informitivity of observation channels. *Periodica Mathematica Hungerica.*, 2(1-4):191–213, 1972.

[4] A. Dembo and O. Zeitouni. *Large deviation techniques and application*. Springer, Application of mathematics-38, 1998.

[5] P. Dupuis and R. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley, Wiley series in probability and statistics, 1986.

[6] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. *Handbook in Operations Research and Mangement Sciences, Volume 13: Simulation*, pages 291–350, 2006.

[7] A. Meucci. Fully flexible views:theory and practice. *Risk Magazine*, 21(10):97–102, 2008.

[8] J. Mina and J. Xiao. Return to riskmetrics:the evolution of a standard. *RiskMatrics publications*, 2001.

[9] J. Pazier. Global portfolio optimization revisited:a least discrimination alternative to black-litterman. *ICMA Centre Discussion Papers in Finance*, July 2007.

[10] E. Qian and S. Gorman. Conditional distribution in portfolio theory. *Financial analyst journal.*, 57(2):44–51, March-April 1993.