Allocation scoring rules as a generalization of quantile scoring rules

2023-06-09

Quantile Scoring Rules

A forecaster is asked to recommend, that is, *predict*, the appropriate supply x of some good or investment in some precautionary measure intended to satisfy a random future demand or need Y. As a running example, we will consider the provision of a limited supply of a limited health care resource such as oxygen or ventilators. Suppose there is an incremental loss $O \geq 0$ incurred when over-prediction leads to unused supply and an incremental loss $U > 0$ incurred when under-prediction leads to unmet demand or need. Let g be a non-decreasing function that expresses a utility associated with the good or misfortune to which Y refers. This function can be thought of as specifying the rate at which costs accrue depending on the scale of need, through a mechanism that will be made precise below.

For notation, we will let

- $\mathscr{F}_0$ be the class of probability measures on the Borel-Lebesgue sets of $\mathbb{R}$
- a probability measure $P \in \mathscr{F}_0$ induced by a random variable X be identified by its cumulative distribution function F_X
- F denote some forecast distribution for the level of future need Y which is known entirely or up to some parameters by its forecaster
- G denote the unknowable distribution of Y
- $Q_F := F^{-1} : [0,1] \rightrightarrows \mathbb{R}$ be the (set-valued) quantile function which maps a probability level $\alpha \in [0,1]$ to the corresponding quantile set $\left[q^-_{\alpha,F}, q^+_{\alpha,F}\right]$ of F where $q^-_{\alpha,F} := \sup\{x \mid F(x) < \alpha\}$ and $q^+_{\alpha,F} := \sup\{x \mid F(x) \leq \alpha\}$
- $Q_\alpha : \mathscr{F}_0 \rightrightarrows \mathbb{R}$ be the functional which maps a distribution F to its α-level quantile set
- $q_{\alpha,F}$ be a (usually unique) element of $Q_F(\alpha) = Q_\alpha(F)$
- $q_{\alpha,Y}$ be an element of the unknowable set $Q_G(\alpha) = Q_\alpha(G)$

Let us measure the forecaster's performance when the realized need is y with the scoring function

$$s_{O,U}(x,y) = O(g(x) - g(y))_+ + U(g(y) - g(x))_+ \tag{1}$$

where $u_+ := u\mathbb{1}\{u > 0\}$ and $\mathbb{1}\{A\}$ is the indicator function of the event A. This is a *negatively oriented* scoring function. That is, a smaller score indicates better performance corresponding to a smaller cost due to under- or over-prediction of the realized y.

The scoring function $s_{O,U}$ has the special property that by giving the forecaster prior notice that $s_{O,U}$ will be used to measure performance, we *elicit* an α-level quantile $q_{\alpha,Y}$ for Y, where $\alpha = U/(U+O)$. That is, we create a situation where a forecaster believing that $Y \sim F$ and trying only to minimize their expected score

$$\overline{s}_{F,O,U}(x) := E_F[s_{O,U}(x,Y)] = OE_F[(g(x) - g(Y))_+] + UE_F[(g(Y) - g(x))_+] \tag{2}$$

will report their quantile $q_{\alpha,F}$. To see this, notice that in order to minimize this expectation, they must forecast a solution of the first order equation

$$\begin{aligned}
0 = \frac{d}{dx}\overline{s}_{F,O,U}(x) &= OE_F\left[\frac{d}{dx}(g(x)-g(Y))_+\right] + UE_F\left[\frac{d}{dx}(g(Y)-g(x))_+\right] \\
&= OE_F[g'(x)\mathbb{1}\{Y < x\}] + UE_F[-g'(x)\mathbb{1}\{Y \geq x\}] \qquad (3) \\
&= Og'(x)\mathbb{P}_F(Y < x) - Ug'(x)\mathbb{P}_F(Y \geq x) \qquad (4) \\
&= g'(x)\left(OF(x) - U(1-F(x))\right) \qquad (5) \\
&= (O+U)g'(x)\left(F(x)-\alpha\right), \qquad (6)
\end{aligned}$$

that is, a quantile $q_{\alpha,F} \in Q_F(\alpha)$. (See appendix for an alternative derivation in terms of a density.) The critical point $q_{\alpha,F}$ is actually a minimum since (6) is non-positive on $\{x < \inf(Q_F(\alpha))\}$ and non-negative on $\{x > \sup(Q_F(\alpha))\}$. With the assumptions that $g'(q) > 0$ and that $Q_F(\alpha)$ is the singleton $\{q_{\alpha,F}\}$, this minimum is unique since

$$\begin{aligned}
\frac{d^2}{dx^2}\overline{s}_{F,O,U}(x)\Big|_{x=q_{\alpha,F}} &= (O+U)\left(g'(x)f(x) + g''(x)\left(F(x)-\alpha\right)\right)\Big|_{x=q_{\alpha,F}} \qquad (7) \\
&= (O+U)(g'(q_{\alpha,F})f(q_{\alpha,F}) + g''(q_{\alpha,F})\cdot 0) > 0. \qquad (8)
\end{aligned}$$

By reparametrizing to $\alpha = U/(U+O)$ and $\kappa = U+O$, we can express s in several alternative forms:

$$\begin{aligned}
s_{O,U}(x,y) &= \kappa((1-\alpha)(g(x)-g(y))_+ + \alpha(g(y)-g(x))_+) \qquad (9) \\
&= \kappa(\mathbb{1}\{y \leq x\} - \alpha)(g(x)-g(y)) \qquad (10) \\
&= -\kappa\left[\alpha g(x) + \mathbb{1}\{y \leq x\}(g(y)-g(x)) + h_1(y)\right] \qquad (\text{with} \quad h_1(y) = -\alpha g(y)) \qquad (11) \\
&= \kappa\left[(1-\alpha)g(x) + \mathbb{1}\{y > x\}(g(y)-g(x)) + h_2(y)\right] \qquad (\text{with} \quad h_2(y) = (\alpha-1)g(y)) \qquad (12) \\
&:= s_{\kappa,\alpha}(x,y) := \kappa s_\alpha(x,y). \qquad (13)
\end{aligned}$$

Since h_1 and h_2 are functions of only y, they do not enter the first order equation (6) and can be regarded as nuisance functions carrying information about the application irrelevant to quantile elicitation. The forms (10) and (11) appear as equations (41) and (40) in (Gneiting and Raftery 2007).[1] The form (12) is often used in operations research and meteorology literature, and is discussed later in the section Meteorologist parameters.

Note from (10) that $\frac{d}{dx}\overline{s}_{F,\alpha}(x) = \frac{d}{dx}E_F[s_\alpha(x,Y)] = k(x)E_F[V_\alpha(x,Y)]$, where $k(x) = \kappa g'(x)$ and

$$V_\alpha(x,y) = \mathbb{1}\{y \leq x\} - \alpha. \tag{14}$$

By virtue of $E_F[V_\alpha(Q_F(\alpha),Y)] = 0$, V_α is said to be an *identification function* for the α quantile. The fact that, generally speaking, any elicitable functional (such as a quantile) has an identification function is known as *Osband's principle*.

Forecasting the minimizer of $\overline{s}_{F,\kappa,\alpha}(x) = E_F[s_{\kappa,\alpha}(x,Y)]$—that is, the forecaster quantile $q_{\alpha,F}$—is known as the *Bayes act* a_F under s_α for the forecaster. Assuming a forecaster is informed, rational, and risk-neutral enough to take a_F, we can evaluate F, implicitly, as a distributional forecast, by the *scoring rule* S_α induced by s_α

$$S_\alpha(F,y) := s_\alpha(Q_F(\alpha),y). \tag{15}$$

It follows from this definition that

$$S_\alpha(F,\tilde{F}) := E_{\tilde{F}}[S_\alpha(F,Y)] \geq S_\alpha(F,F) \quad \text{for all} \quad F,\tilde{F} \in \mathcal{F}_0. \tag{16}$$

A scoring rule $S(F,y)$ is said to be *proper* when it satisfies (16) and *strictly proper* when the inequality is strict. S_α is not strictly proper since (16) is an equality for any F and $\tilde{F}$ sharing an α quantile. S_α is

[1] Or rather the negatives of these equations since Gneiting and Raftery use positively oriented scoring functions. Also note that their x plays the role of our y.

however sometimes said to be strictly proper for the α quantile since the inequality becomes strict whenever $q_{\alpha,F} \neq q_{\alpha,\tilde{F}}$, (see e.g., (Jose and Winkler 2009)). Another way to refer to this property is to call the scoring function s_α *strictly consistent* for the α quantile functional (Gneiting 2011). Having a strictly consistent scoring function makes the α quantile *elicitable.*

If y is generated by $Y \sim G$ then $S_\alpha(F, y)$ is an estimate of

$$S_\alpha(F, G) := E_G[S_\alpha(F, Y)] = E_G[s_\alpha(Q_F(\alpha), Y)], \tag{17}$$

the expected value under s_α of the forecaster's α quantile under the *actual* data generating process G. More abstractly, this estimand is the value at F of the functional $\overline{S}_{G,\alpha}$ on $\mathcal{F}_0$ given by

$$\overline{S}_{G,\alpha}(F) := \overline{s}_{G,\alpha}(Q_\alpha(F)) \tag{18}$$

where Q_α is the α quantile functional.

Remark. Relaxing either of the assumptions that $g'(q_{\alpha,F}) > 0$ or that $Q_F(\alpha)$ is a singleton set would require adopting additional criteria for whether $q_{\alpha,F}$ is optimal. One possibility would be to take $q_{\alpha,F} = \min\{x \mid F_X(x) \geq \alpha\}$, and in case $g'(q_{\alpha,F}) = 0$ or fails to exist, choose the prediction x to be the infimum of the region around $q_{\alpha,F}$ where g is flat. An example of such a g is the *power curve* of a wind turbine which becomes constant for wind speeds above the turbine's maximum operating speed. Here x would be a forecast of future wind speed Y at time t and $g(x)$ would be the power a turbine operator commits to supplying the electric grid at t. Taking g to be constant below a certain point could also serve to express the futility or lack of meaning of forecasts below that point, such as when a forecast user is only able to sell a commodity for which demand can become negative, turning into supply.

Examples of forecaster expected scores

Corresponding to the two expressions of an expectation $E_F[h(y)\mathbb{1}\{a < y < b\}]$ as either side of the Riemann-Stieltjes integration-by-parts formula $\int_a^b h(y)dF(y) = h(y)F(y)\big|_a^b - \int_a^b F(y)dh(y)$, the expected score (2) can be written as either

$$\begin{aligned}
\overline{s}_{F,O,U}(x) &= O\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}\{Y \leq u < x\}dg(u)dF(y) + U\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}\{x \leq u < Y\}dg(u)dF(y) && (19)\\
&= O\int_{\mathbb{R}} \mathbb{1}\{u < x\}F(u)dg(u) + U\int_{\mathbb{R}} \mathbb{1}\{u \geq x\}(1 - F(u))dg(u) && (20)\\
&= O\int_{-\infty}^{x} F(u)dg(u) + U\int_{x}^{\infty}(1 - F(u))dg(u), && (21)
\end{aligned}$$

or

$$\begin{aligned}
\overline{s}_{F,O,U}(x) &= O\int_{-\infty}^{x}(g(x) - g(y))dF(y) + U\int_{x}^{\infty}(g(y) - g(x))dF(y) \\
&= O\left(g(x)F(x) - \int_{-\infty}^{x} g(y)dF(y)\right) + U\left(\int_{x}^{\infty} g(y)dF(y) - g(x)(1 - F(x))\right) && (22)\\
&= g(x)\left((O + U)F(x) - U\right) - (O + U)\int_{-\infty}^{x} g(y)dF(y) + U\int_{\mathbb{R}} g(y)dF(y) && (23)\\
&= \kappa\left[g(x)(F(x) - \alpha) - \overline{g}_F(x) + \alpha\overline{g}_F\right] && (24)
\end{aligned}$$

where

$$\begin{aligned}
\overline{g}_F(x) &= E_F[g(Y)\mathbb{1}\{Y \leq x\}] && (25)\\
\overline{g}_F &= E_F[g(Y)]. && (26)
\end{aligned}$$

The function $\overline{g}_F(x)$ sometimes goes by the name of "partial expectation" (Schlaifer Raiffa, p. 109) and can often be made more explicit for certain F and g.

Taking, for example, $g(x) = x$ and $F = \text{Exp}(1/\sigma)$, we have

$$\overline{g}_F(x) = \int_0^{\max(0,x)} y \frac{e^{-y/\sigma}}{\sigma} dy = \mathbb{1}\{x \geq 0\} \left[\sigma - e^{-x/\sigma}(\sigma + x)\right] \tag{27}$$

so that (24) is

$$\begin{align} \overline{s}_{\text{Exp}(1/\sigma),\kappa,\alpha}(x) &= \kappa \left[x(\mathbb{1}\{x \geq 0\} \left[1 - e^{-x/\sigma}\right] - \alpha) - \mathbb{1}\{x \geq 0\} \left[\sigma - e^{-x/\sigma}(\sigma + x)\right] + \alpha\sigma\right] \tag{28} \\ &= \kappa \left[\mathbb{1}\{x \geq 0\}\sigma e^{-x/\sigma} + (\mathbb{1}\{x \geq 0\} - \alpha)(x - \sigma)\right]. \tag{29} \end{align}$$

And if $F = F_{\mu,\sigma}$ lies in a location-scale family, that is, $F(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$, we have

$$\begin{align} \int_{-\infty}^{x} y dF(y) &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} (\mu + \sigma z) dF_0(z) \tag{30} \\ &= \mu F(x) + \sigma \int_{-\infty}^{\frac{x-\mu}{\sigma}} z dF_0(z). \tag{31} \end{align}$$

Then (24) becomes

$$\begin{align} \overline{s}_{F,\kappa,\alpha}(x) &= \kappa \left[x(F(x) - \alpha) - \mu F(x) - \sigma \int_{-\infty}^{\frac{x-\mu}{\sigma}} z dF_0(z) + \alpha E_F[Y]\right] \tag{32} \\ &= \kappa \left[F(x)(x - \mu) - \alpha(x - E_F[Y]) - \sigma \int_{-\infty}^{\frac{x-\mu}{\sigma}} z dF_0(z)\right] \tag{33} \end{align}$$

For example, if $F = U[a, b]$ is uniform we can take $F_0 = U[0, 1], \mu = a, \sigma = b - a$, and (33) is

$$\begin{align} \overline{s}_{F,\kappa,\alpha}(x) &= \begin{cases} -\alpha\kappa \left[x - \frac{a+b}{2}\right] & x \leq a \\ \kappa \left[\frac{x-a}{b-a}(x - a) - \alpha\left(x - \frac{a+b}{2}\right) - (b - a)\frac{1}{2}\left(\frac{x-a}{b-a}\right)^2\right] & a < x < b \\ \kappa \left[x - a - \alpha\left(x - \frac{a+b}{2}\right) - \frac{b-a}{2}\right] & b \leq x \end{cases} \tag{34} \\ &= \begin{cases} -\kappa\alpha \left[x - \frac{a+b}{2}\right] & x \leq a \\ \frac{\kappa}{2} \left[\frac{1}{b-a}(x - (a + \alpha(b - a)))^2 + \alpha(1 - \alpha)(b - a)\right] & a < x < b \\ \kappa(1 - \alpha) \left[x - \frac{a+b}{2}\right] & b \leq x \end{cases} \tag{35} \end{align}$$

Thus, outside (a, b), $\overline{s}_{U[a.b],\kappa,\alpha}(x)$ has the form of pinball loss centered on the mean $m_F = \frac{a+b}{2}$ and is interpolated on (a, b) by a quadratic centered at the quantile $a + \alpha(b - a)$ with slopes matching the linear tails at a and b (see figure 1).

Or if $F = \Phi_{\mu,\sigma}$ is normal with location $\mu = E_F[Y]$ and scale $\sigma = \sqrt{\text{Var}_F(Y)}$, we have

$$\int_{-\infty}^{\frac{x-\mu}{\sigma}} z dF_0(z) = -\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} = -\sigma\varphi_{\mu,\sigma}(x) \tag{36}$$

so that

$$\overline{s}_{F,\kappa,\alpha}(x) = \kappa \left[(\Phi_{\mu,\sigma}(x) - \alpha)(x - \mu) + \sigma^2\varphi_{\mu,\sigma}(x)\right]. \tag{37}$$

We finally note that for non-trivial g's it may be helpful to make the substitution $v = g(u)$ in (21) to get

$$\overline{s}_{F,O,U}(x) = O \int_{-\infty}^{g(x)} F(g^{-1}(u)) du + U \int_{g(x)}^{\infty} (1 - F(g^{-1}(u))) du \tag{38}$$

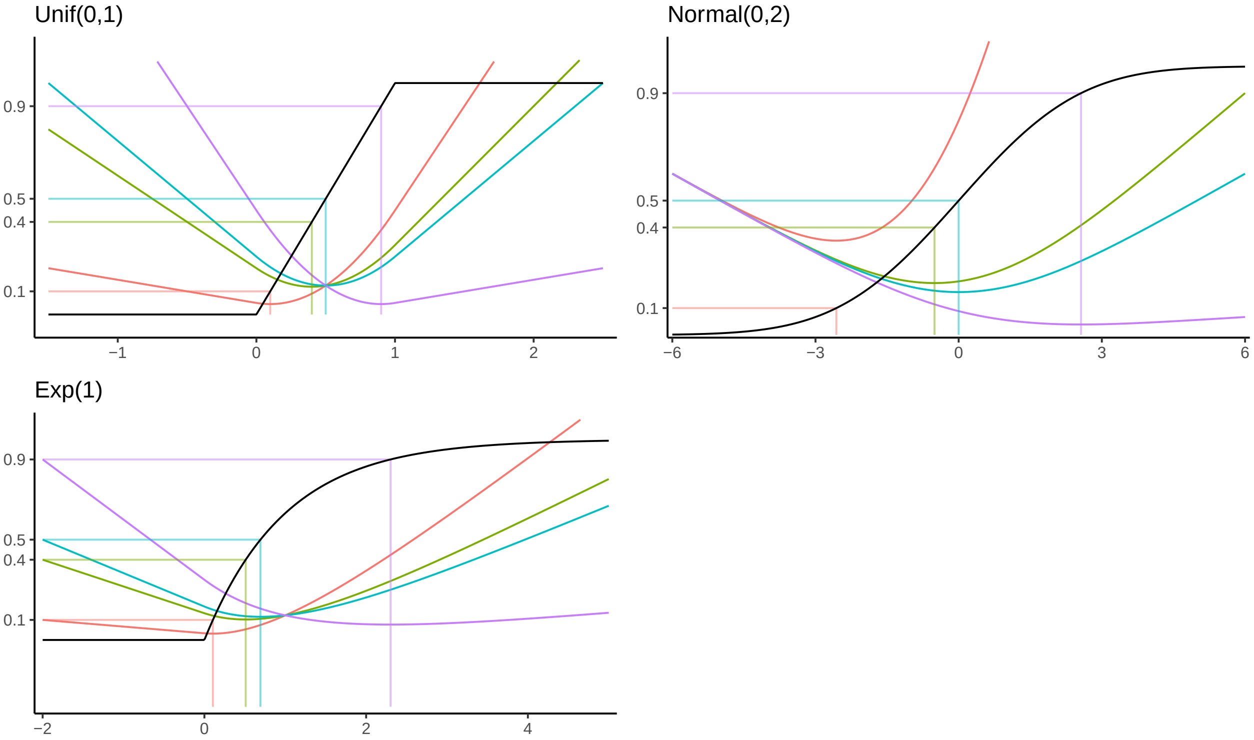

Figure 1: Plots of CDFs and expected quantile scores at levels .1, .4, .5, and .9

and in (25) to get

$$\overline{g}_F(x) = \int_{-\infty}^{g(x)} v d(F \circ g^{-1})(v). \tag{39}$$

Entropy and Divergence

Let F, F_i be general distributions such as forecaster beliefs or reports about Y, and G be the true distribution of Y. For a general scoring rule S, the S-entropy of F is defined as

$$e_S(F) := S(F,F) = E_F[S(F,Y)]. \tag{40}$$

Applying this to S_α, we can define the g, α-entropy of F as

$$e_\alpha(F) := S_\alpha(F,F) = E_F[s_\alpha(q_{F,\alpha}, Y)]. \tag{41}$$

If a forecaster believes that $Y \sim F$ then they expect to receive a score of $e_\alpha(F)$ by reporting their α quantile. According to (21) and (24),

$$e_\alpha(F) = (1-\alpha)\int_{-\infty}^{q_{\alpha,F}} F(u)dg(u) + \alpha\int_{q_{\alpha,F}}^{\infty}(1-F(u))dg(u) \tag{42}$$

$$= \alpha\overline{g}_F - \overline{g}_F(x). \tag{43}$$

Similarly, we define the g, α-divergence between two distributions as

$$d_\alpha(F_1,F_2) := S_\alpha(F_1,F_2) - S_\alpha(F_2,F_2) = S_\alpha(F_1,F_2) - e_\alpha(F_2). \tag{44}$$

Again from (21) we get

$$d_\alpha(F_1,F_2) = (1-\alpha)\int_{-\infty}^{q_{\alpha,F_1}} F_2(u)dg(u) + \alpha\int_{q_{\alpha,F_1}}^{\infty}(1-F_2(u))dg(u) \tag{45}$$

$$- (1-\alpha)\int_{-\infty}^{q_{\alpha,F_2}} F_2(u)dg(u) - \alpha\int_{q_{\alpha,F_2}}^{\infty}(1-F_2(u))dg(u) \tag{46}$$

$$= (1-\alpha)\int_{q_{\alpha,F_2}}^{q_{\alpha,F_1}} F_2(u)dg(u) + \alpha\int_{q_{\alpha,F_1}}^{q_{\alpha,F_2}}(1-F_2(u))dg(u). \tag{47}$$

And from (24),

$$d_\alpha(F_1,F_2) = g(q_{\alpha,F_1})(F_2(q_{\alpha,F_1})-\alpha) - \overline{g}_{F_2}(q_{\alpha,F_1}) + \alpha\overline{g}_{F_2} - \alpha\overline{g}_{F_2} + \overline{g}_{F_2}(q_{\alpha,F_2}) \tag{48}$$

$$= g(q_{\alpha,F_1})(F_2(q_{\alpha,F_1})-\alpha) + \int_{q_{\alpha,F_1}}^{q_{\alpha,F_2}} g(y)dF_2(y) \tag{49}$$

To emphasize the reality of G we write $e_\alpha(G)$ as $e_\alpha(Y)$, which with (44) gives

$$S_\alpha(F,G) = d_\alpha(F,G) + e_\alpha(Y). \tag{50}$$

This expresses the true expected score of F (i.e., not that calculated by the forecaster) as the sum of the failure of F to match G and the inherent uncertainty of Y.

We can also replace G in (50) by $G|F$, the true distribution of Y *given* that the forecaster believes $Y \sim F$—a belief presumably based on some knowledge of leading indicators for Y. This gives

$$S_\alpha(F,G|F) = d_\alpha(F,G|F) + e_\alpha(Y|F) \tag{51}$$

Using

$$d_\alpha(G,G|F) = S_\alpha(G,G|F) - e_\alpha(Y|F) \tag{52}$$

this can be expanded to

$$S_\alpha(F,G|F) = d_\alpha(F,G|F) - d_\alpha(G,G|F) + S_\alpha(G,G|F). \tag{53}$$

Now view F explicitly as a realization of a distribution-valued random variable $\mathcal{F}$ associated with the forecaster. To measure the forecaster's perfomance in general and not just on the occasion when they form the belief $Y \sim F$, we can take the $\mathcal{F}$ expectation of (53) to get a decomposition

$$E_{\mathcal{F}}[S_\alpha(F,G|F)] = E_{\mathcal{F}}[d_\alpha(F,G|F)] - E_{\mathcal{F}}[d_\alpha(G,G|F)] + e_\alpha(Y). \tag{54}$$

The first two terms on the RHS are known as the *unreliability* and the *resolution* of the forecaster.

Newsvendor parameters

In the inventory management literature, minimization of an expectation of the form of $\overline{s}_F(x)$ for a given F arises in the *newsvendor problem*. This involves the ordering decision faced by a retailer of a perishable good (such as newpapers) when customer demand is uncertain. Because the focus here is on revenues and expenditures, terms other than just the over- and underprediction penalties in (1) often appear, giving a retailer's scoring function as

$$s_{\mathbf{n}}(x,y) = cx - p\min(x,y) - u(x-y)_+ + r(y-x)_+ \tag{55}$$

where $\mathbf{n} = \{c,p,u,r\}$. This score is netting the monetary transfers that occur when where x newspapers are ordered at price c, $\min(y,x)$ are sold at retail price p, $(x-y)_+$ are sold at salvage price u, and $r(y-x)_+$ of monetarily valued customer "goodwill" is lost. Using $x = y + (x-y)_+ - (y-x)_+$ and $\min(x,y) = y - (y-x)_+$ (55) can be rewritten as

$$s_{\mathbf{n}}(x,y) = (c-u)(x-y)_+ + (p-c+r)(y-x)_+ - (p-c)y \tag{56}$$

giving the retailer's objective function under a demand distribution F as

$$\overline{s}_{F,\mathbf{n}}(x) = (c-u)E_F[(x-Y)_+] + (p-c+r)E_F[(Y-x)_+] - (p-c)E[Y]. \tag{57}$$

This differs by a constant from our $\overline{s}_{F,O,U}(x)$ in (2) with $g(x) = x$, $O = c-u$ and $U = p-c+r$, and so has the same minimizer

$$Q = F^{-1}\left(\frac{U}{O+U}\right) = F^{-1}\left(\frac{p-c+r}{p-u+r}\right). \tag{58}$$

Note, for reference, that in the literature $\overline{s}_{F,\mathbf{n}}(x)$ is often encountered in the form

$$(O+U)\int_0^x (x-y)dF(y) - Ux + (U-u+r)E[Y] \tag{59}$$

$$= (p+r-u)\int_0^x (x-y)dF(y) - (p-c+r)x + rE[Y] \tag{60}$$

Meteorologist parameters

Similarly, a weather forecaster might be judged by the cost Cx of recommended protection x against a level y of adverse weather (e.g., rainfall) added to any loss $L(y-x)_+$ resulting from underprediction, leading to the scoring function

$$s_m(x,y) = Cx + L(y-x)_+, \tag{61}$$

which rearranges to

$$s_m(x,y) = C(x-y)_+ + (L-C)(y-x)_+ + Cy. \tag{62}$$

The minimizer of $\overline{s}_{m,F}(x) = E_F[s_m(x,Y)]$ given the forecaster's belief $Y \sim F$ is now

$$Q = F^{-1}\left(\frac{L-C}{C+L-C}\right) = F^{-1}\left(1 - \frac{C}{L}\right). \tag{63}$$

In particular, faced with the classical binary decision problem of whether to recommend an additional unit of protection given a current level of protection x, the forecaster's optimal decision rule under s_m is to recommend adding protection if $x < F^{-1}\left(1 - \frac{C}{L}\right)$, that is, if

$$1 - F(x) = \mathbb{P}_F\{y > x\} > \frac{C}{L}, \tag{64}$$

the *cost-loss ratio* of the problem.

There is considerable opportunity here for confusion as to the role of x if one is following (as we mostly are) the notation of (Ehm et al. 2016).

Multipoint Quantile Scoring Rules

Suppose we ask a forecaster to provide a set of point forecasts $\{x_1, x_2\}$ to be rewarded according to the aggregate scoring function

$$s(x_1, x_2, y) = s_{O_1,U_1}(x_1, y) + s_{O_2,U_2}(x_2, y). \tag{65}$$

Allocation Scoring Rules

We now develop *allocation scoring functions* that elicit forecasts x_i for outcomes $y_i, i = 1, \dots, N$, each with it's own incremental cost O_i and loss U_i as for a quantile scoring rule, but with the additional constraint

$$\sum_{i=1}^{N} w_i x_i = \mathbf{w}^T \mathbf{x} \le K \tag{66}$$

on the total provision available with $w_i > 0$. We assume $K > 0$, that only non-negative forecasts x_i, i.e., recommended allocations, are accepted, and that

$$g_i'(x_i) > 0 \text{ for } x_i \ge 0, \tag{67}$$

though it may be interesting to introduce regions where $g_i(x_i)$ is constant, such as for the wind speed power curve mentioned above.

According to (2) and (24) the objective function is now

$$\overline{s}_F(\mathbf{x}) = \sum_{i=1}^{N} \overline{s}_{F_i}(x_i) = \sum_{i=1}^{N} \left\{ O_i \int_{-\infty}^{x_i} (g_i(x_i) - g_i(y)) f_i(y) dy + U_i \int_{x_i}^{\infty} (g_i(y) - g_i(x_i)) f_i(y) dy \right\} \tag{68}$$

$$= \sum_{i=1}^{N} \kappa_i \left[g_i(x_i)(F_i(x_i) - \alpha_i) - \overline{g}_{i F_i}(x_i) + \alpha_i \overline{g}_{i F_i} \right] \tag{69}$$

and the elicited forecasts solve the allocation problem (AP)

$$\underset{\mathbf{x} \in \mathbb{R}^N, 0 \le \mathbf{x}}{\text{minimize}} \ \overline{s}_F(\mathbf{x}) \text{ subject to } \mathbf{w}^T \mathbf{x} \le K. \tag{70}$$

As a separable problem, this is amenable to being solved via its dual (see (Ruszczynski 2011), section 4.4). With the Lagrangian

$$L(\mathbf{x}, \lambda; K) = \sum_{i=1}^{N} \overline{s}_{F_i}(x_i) + \lambda \left(\sum_{i=1}^{N} w_i x_i - K \right) \tag{71}$$

we have the dual problem of maximizing over $\lambda \ge 0$ the objective function

$$L_D(\lambda; K) = \min_{0 \le \mathbf{x}} L(\mathbf{x}, \lambda) = -\lambda K + \sum_{i=1}^{N} \min_{0 \le x_i} \left\{ \lambda w_i x_i + \overline{s}_{F,i}(x_i) \right\} \tag{72}$$

which gives lower bounds on $\overline{s}_F(\mathbf{x}^\star)$ at a solution $\mathbf{x}^\star$ to the primal AP. In particular, when $O_i > 0$ for all i, $L_D(0; K)$ gives the total score $\overline{s}_F(\mathbf{q}_{\boldsymbol{\alpha}, F})$ when the constraint does not prevent the forecaster from giving the quantiles in each component. When the constraint is active, the forecaster will generally not be able to attain a total score this low and the greatest lower bound on their score will be $L_D(\lambda^\star; K)$ where $\lambda^\star = \lambda^\star(K)$ is the maximizer of (72).

The advantage of this dual formulation under separability is that when considering a candidate λ for $\lambda^\star$, we can calculate each summand $L_{D,i}(\lambda) := \min_{0 \le x} \left\{ \lambda w_i x + \overline{s}_{F,i}(x) \right\}$ independently. To this end, let $Z_{\lambda,i}$ be the set of $x \in \mathbb{R}$ where the weighted subgradient of $-\overline{s}_{F_i}(x)$ contains λ (c.f. (6)):

$$Z_{\lambda,i} := \left\{ x \mid \lambda \in -w_i^{-1} \partial \overline{s}_{F_i}(x) = w_i^{-1} \kappa_i g_i'(x)(\alpha_i - [F_i(x-), F_i(x)]) \right\}. \tag{73}$$

Then for $\lambda > 0$ we have

$$L_{D,i}(\lambda) = \min \left\{ \lambda w_i x + \overline{s}_{F_i}(x) \mid x \in \{0\} \cup \{Z_{\lambda,i} \cap (0, \infty)\} \right\} \tag{74}$$

Note that

- $Z_{\lambda,i}$ will contain any (necessarily closed) interval where $\overline{s}_{F_i}$ is linear with slope $-\lambda w_i$,
- unless $\lambda = 0 = O_i$, we have $\sup Z_{\lambda,i} < \infty$ since in this case $\lambda w_i x + \overline{s}_{F,i}(x)$ is eventually monotonically increasing in x,
- $Z_{\lambda,i} \cap [0, \infty) = \emptyset$ for $\lambda > \max_{x \geq 0}\{w_i^{-1}\kappa_i g_i'(x)(\alpha_i - F_i(x))\}$.

In the case that we will usually restrict to where g_i' is non-vanishing and non-increasing, $Z_{\lambda,i}$ will consist at most of a single interval (usually of width 0), and in the typical "pinball loss" case of $g_i'(x) \equiv 1$,

$$Z_{\lambda,i} = \begin{cases} Q_i(\alpha_i - \kappa_i^{-1} w_i \lambda) & \text{if } \alpha_i - \kappa_i^{-1} w_i \lambda > 0 \\ \emptyset & \text{otherwise.} \end{cases} \tag{75}$$

Let $Z_\lambda = \{\mathbf{x} \mid x_i \in Z_{\lambda,i}\}$ and $\Delta_{\overline{s}_F,K,\pm}(\lambda) = \mathbf{w}^T\mathbf{x}_\pm(\lambda) - K$ where

$$x_{-,i}(\lambda) = \min Z_{\lambda,i} \cap [0, \infty) \text{ and } x_{+,i}(\lambda) = \max Z_{\lambda,i} \cap [0, \infty) \tag{76}$$

and write $\Delta_{\overline{s}_F,K,\pm}(\lambda) = \Delta_{\overline{s}_F,K}(\lambda)$ when $\#Z_\lambda = 1$. If $\Delta_{\overline{s}_F,K,\pm}(\lambda) > 0$, the constraint is violated everywhere in $Z_\lambda \cap [0, \infty)^N$, and from (71) we see that $0 > \sup(\partial L_D(\lambda))$ so that $\lambda^\star > \lambda$, whereas if $\Delta_{\overline{s}_F,K,\pm}(\lambda) < 0$ there is under-utilization everywhere in $Z_\lambda \cap [0, \infty)^N$, $0 < \inf(\partial L_D(\lambda))$, and $\lambda^\star < \lambda$. But if $\Delta_{\overline{s}_F,K,-}(\lambda) \leq 0 \leq \Delta_{\overline{s}_F,K,+}(\lambda)$, we have $\lambda^\star = \lambda$ and a solution $\mathbf{x}^\star$ to the AP is obtained by solving $\mathbf{w}^T\mathbf{x} = K$ on $Z_\lambda \cap [0, \infty)^N$. For example, take $N = 2$, $F_1 = \text{Unif}[a, b]$, $F_2 = N(a, a/2)$ with $a > 0$, $\alpha_1 = .5, \alpha_2 = .6$, $\kappa_i = g_i' = w_i = 1$, and let $q = Q_{N(a,a/2)}(.1)$. Then the allocation $(a, q) \in Z_{.5}$ violates a constraint of $K = a$, but the allocation with $(0, q) \in Z_{.5}$ leaves slack. Thus the solution $(a - q, q)$ lies in $Z_{.5}$.

As such, we can solve, at least approximately, the AP as follows. We first determine, from representations of the F_i, the finite set

$$I_{\overline{s}_F} := \{\lambda \mid \#Z_\lambda > 1\} \tag{77}$$

and check whether for any $\lambda \in I_{\overline{s}_F}$ we can solve $\mathbf{w}^T\mathbf{x} = K$ with $\mathbf{x} \in Z_\lambda$. If possible, we have found the maximizer $\lambda^\star$ as well as a solution $\mathbf{x}$ to the AP.

If $I_{\overline{s}_F}$ does not yield an immediate solution, we proceed to a binary search for $\lambda^\star$ working inward from the interval $[\lambda_L, \lambda_U]$ where

$$\lambda_L = \max\left\{\lambda \in \bar{I}_{\overline{s}_F} \mid \Delta_{\overline{s}_F,K,-}(\lambda_\tau) > 0\right\} \tag{78}$$

$$\lambda_U = \min\left\{\lambda \in \bar{I}_{\overline{s}_F} \mid \Delta_{\overline{s}_F,K,+}(\lambda_\tau) < 0\right\} \tag{79}$$

$$\bar{I}_{\overline{s}_F} = \{I_{\overline{s}_F}, 0, \max_{i,x}\{w_i^{-1}\kappa_i g_i'(x)(\alpha_i - F_i(x))\}\}. \tag{80}$$

At each step τ of the search, we take $\lambda_\tau = 1/2(\lambda_{U,\tau} + \lambda_{L,\tau})$ and either discard the upper half of the current search interval if $\Delta_{\overline{s}_F,K}(\lambda_\tau) < 0$, setting $\lambda_{U,\tau+1} = \lambda_\tau$, or discard the bottom half if not, setting $\lambda_{L,\tau+1} = \lambda_\tau$. We continue until $\Delta_{\overline{s}_F,K}(\lambda_\tau)$ is sufficiently small.

If we also impose convexity on $\overline{s}_F(\mathbf{x})$ by requiring that all g_i have non-positive second derivatives, then since the constraint $\mathbf{w}^T\mathbf{x} \leq K$ is affine and the domain $\{\mathbf{x} \geq 0\}$ is convex, we have strong duality so that the optimal value $L_D(\lambda^\star)$ of the dual problem is also the attained minimum of the primal AP. An optimal allocation is then given by taking

$$q_{i,K} \in \operatorname{argmin}\left\{\lambda^\star w_i x + \overline{s}_{F_i}(x) \mid x \in Z_{\lambda^\star,i}\right\}. \tag{81}$$

Note that in the "easy" case of $g_i(x) = x$ for all i and K not too small, this becomes

$$q_{i,K} = Q_i(\alpha_i - \kappa_i^{-1} w_i \lambda^\star) \tag{82}$$

where $\lambda^\star$ is found as a root of

$$\sum_{i=1}^{N} w_i Q_i(\alpha_i - \kappa_i^{-1} w_i \lambda^\star) = K. \tag{83}$$

This partial solution to the AP seems to have first appeared in (Hadley and Whitin 1963).

See the appendix for a derivation and discussion of this solution method via Karush-Kuhn-Tucker equations rather than duality theory.

Marginal benefit interpretation of the dual variable

Equation (73) can be viewed as defining functions

$$\lambda_i(x_i) := -\frac{1}{w_i}\frac{\partial \overline{s}_F}{\partial x_i}(x_i) = w_i^{-1}\kappa_i g_i'(x_i)(\alpha_i - F_i(x_i)) \tag{84}$$

which can each be interpeted as a *marginal expected unit benefit* of x_i. That is, if $F_i(x_i) < \alpha_i$ we receive the positive "benefit" of reducing $\overline{s}_F$ at a rate of $\lambda_i(x_i)$ per additional unit of capacity allocated to the i'th target at "price" w_i. Note that with $\mathcal{S}_i$ the support of F_i we have

$$\lambda_i(x_i) = \begin{cases} \alpha_i w_i^{-1}\kappa_i g_i'(x_i) & x_i \leq \inf(\mathcal{S}_i) \\ -(1-\alpha_i) w_i^{-1}\kappa_i g_i'(x_i) & x_i \geq \sup(\mathcal{S}_i), \end{cases} \tag{85}$$

as in the [uniform][1) $Y \sim U[a,b]$] and [exponential][4) $g(x) = x$ and $Y \sim \text{Exp}(1/\sigma)$] examples.

We can demonstrate the KKT condition (A.21) directly in terms of the λ_i:

Proposition 1. *For a solution* $\mathbf{Q} = (Q_1, \ldots, Q_N)$ *of the AP we have*

$$\lambda_i(Q_i) = \lambda_j(Q_j) \textit{ whenever } Q_i, Q_j > 0. \tag{86}$$

Proof. Assume instead we had $\lambda_i(Q_i) > \lambda_j(Q_j)$ and $Q_i, Q_j > 0$. Moving away from $\mathbf{Q}$ in the direction $\mathbf{d}_{ij} = w_i^{-1}\mathbf{e}_i - w_j^{-1}\mathbf{e}_j$ we remain in $\mathbb{R}_+^N$ while respecting the constraint ($\langle \nabla \mathbf{w}^T\mathbf{x}, \mathbf{d}_{ij}\rangle = 0$) but reducing $\overline{s}_F$:

$$\langle \nabla \overline{s}_F(\mathbf{Q}), \mathbf{d}_{ij}\rangle = w_i^{-1}\frac{\partial \overline{s}_F}{\partial x_i}(Q_i) - w_j^{-1}\frac{\partial \overline{s}_F}{\partial x_j}(Q_j) = \lambda_j(Q_j) - \lambda_i(Q_i) < 0. \tag{87}$$

Therefore $\mathbf{Q}$ cannot be a constrained minimum of $\overline{s}_F$. □

But in line with KKT complementary slack conditions, the argument does not apply when $Q_j = 0$ and $\mathbf{d}_{ij}$ takes us immediately out of $\mathbb{R}_+^N$. This shows again how the set of indices $\{i | Q_i = 0\}$ arises as another variable of the problem.

Allocation scoring rule definition

Collect the AP parameters into $\boldsymbol{\beta} := \{\mathbf{w}, K, \boldsymbol{\alpha}, \boldsymbol{\kappa}\}$. The resulting $Q_i^{\boldsymbol{\beta}} = Q_i(\boldsymbol{\beta})$ form the Bayes act for the the forecast $\mathbf{F}$ and the realized losses and costs defines the scoring rule evaluated on $\mathbf{F}$:

$$\begin{aligned} S_{\boldsymbol{\beta}}(\mathbf{F}, \mathbf{y}) = s_{\boldsymbol{\beta}}(\mathbf{Q}^{\boldsymbol{\beta}}, \mathbf{y}) &= \sum O_i(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i))_+ + U_i(g_i(y_i) - g_i(Q_i^{\boldsymbol{\beta}}))_+ && (88) \\ &= \sum ((U_i + O_i)\mathbb{1}\{Q_i^{\boldsymbol{\beta}} > y_i\} - U_i)(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i)) && (89) \\ &= \sum \kappa_i(\mathbb{1}\{Q_i^{\boldsymbol{\beta}} > y_i\} - \alpha_i)(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i)) && (90) \end{aligned}$$

Remarks:

- Quantile rules allow you to vary cost ratios while allocation rules add capacity as a parameter.

- While quantile rules are undefined for zero overprediction costs, allocation rules (with $K < \infty$) are defined even when some $\alpha_i = 1$. This covers one of the problems originally motivating this project: find the Bayes act for allocation $\mathbf{x}$ of hospital supplies to locations $l_i, i = 1, \dots, N$ given distributions F_i of need Y_i, a total available stock K of supplies, and a loss function

$$l(\mathbf{x}, \mathbf{y}) = U \sum_{i=1}^{N} (y_i - x_i)_+. \tag{91}$$

- $s_{\boldsymbol{\beta}}$ is consistent by definition for the functional $\mathbf{Q}_{\boldsymbol{\beta}}$ and $S_{\boldsymbol{\beta}}$ is proper since

$$S_{\boldsymbol{\beta}}(\mathbf{F}, \mathbf{F}) = s_{\boldsymbol{\beta}}(\mathbf{q}_{\boldsymbol{\beta},\mathbf{F}}, \mathbf{F}) \leq s_{\boldsymbol{\beta}}(\mathbf{q}_{\boldsymbol{\beta},\mathbf{F}'}, \mathbf{F}) = S_{\boldsymbol{\beta}}(\mathbf{F}', \mathbf{F}) \tag{92}$$

  but is not strictly proper

An Oracle Adjustment

Differences of $s_{\boldsymbol{\beta}}$ and $S_{\boldsymbol{\beta}}$ between forecasts may become difficult to compare as K becomes small. To remedy this we can adjust scores by that of an "oracle" who sees into the future and allocates according to the point mass forecast $\delta_{\mathbf{y}}$. That is, the oracle solves the deterministic version of the AP with $\overline{s}_F(\mathbf{x})$ replaced by $s(\mathbf{x}, \mathbf{y})$. Since the oracle knows there is no benefit to allocating more than y_i to the i'th component, we can pose this problem as

$$\underset{\mathbf{x} \in \mathbb{R}^N, 0 \leq \mathbf{x} \leq \mathbf{y}}{\text{minimize}} \; s(\mathbf{x}, \mathbf{y})|_{\mathbf{x} \leq \mathbf{y}} = \sum_{i=1}^{N} -U_i(g_i(x_i) - g_i(y_i)) \text{ subject to } \sum_{i=1}^{N} w_i x_i \leq K. \tag{93}$$

which we can again solve via its dual. The Lagrangian is now

$$L(\mathbf{x}, \lambda) = \sum_{i=1}^{N} -U_i(g_i(x_i) - g_i(y_i)) + \lambda \left(\sum_{i=1}^{N} w_i x_i - K \right) \tag{94}$$

giving the dual problem of maximizing

$$L_D(\lambda) = \min_{0 \leq \mathbf{x} \leq \mathbf{y}} L(\mathbf{x}, \lambda) \tag{95}$$

$$= -\lambda K + \sum_{i=1}^{N} \min_{0 \leq x_i \leq y_i} (\lambda w_i x_i - U_i g_i(x_i) + U_i g_i(y_i)) \tag{96}$$

subject to $\lambda \geq 0$. We have

$$L_{D,i}(\lambda) := \min_{0 \leq x_i \leq y_i} (\lambda w_i x_i - U_i g_i(x_i) + U_i g_i(y_i)) \tag{97}$$

$$= \min \left\{ \lambda w_i x - U_i g_i(x) + U_i g_i(y_i) \mid x \in Z_{o,\lambda,i} \right\} \tag{98}$$

where

$$Z_{o,\lambda,i} = \{0, y_i\} \cup \{Z'_{o,\lambda,i} \cap (0, y_i)\} \tag{99}$$

and $Z'_{o,\lambda,i}$ is the possibly empty set of critical points solving

$$\lambda = w_i^{-1} U_i g_i'(x) = w_i^{-1} \kappa_i \alpha_i g_i'(x). \tag{100}$$

Note that $L_D(0) = L(\mathbf{y}, 0) = 0$ is the oracle's score when the outcome $\mathbf{y}$ does not activate the constraint.

We can again find a best lower bound on the oracle's score by approximating a maximizer $\lambda^\star$ with a binary search starting at $\lambda_L = 0$ and $\lambda_U > \max_{i,x}\{w_i^{-1} \kappa_i \alpha_i g_i'(x)\}$. And with our standard assumption that $g_i''(x) \leq$

0 for all i, under which the primal objective is convex and $\max_{i,x}\{w_i^{-1}\kappa_i\alpha_i g_i'(x)\} = \max_i\{w_i^{-1}\kappa_i\alpha_i g_i'(0)\}$, strong duality gives the oracle's score as $L_D(\lambda^\star)$, corresponding to an allocation

$$q_{o,i,K} \in \operatorname{argmin}\left\{\lambda^\star w_i x - U_i g_i(x_i) \mid x \in Z_{o,\lambda^\star,i}\right\}. \tag{101}$$

In the pinball loss case of $g_i(x) = x$ for all i, oracle allocation reduces to a linear program with piecewise linear dual

$$\begin{aligned} L_D(\lambda) &= -\lambda K + \sum_{i=1}^{N} \min_{0\le x_i \le y_i} \left(\lambda w_i x_i - U_i x_i + U_i y_i\right) && (102)\\ &= -\lambda K + \sum_{i=1}^{N} \min\left\{U_i y_i, \lambda w_i y_i\right\} && (103)\\ &= -\lambda K + \sum_{i=1}^{N} \left(U_i \mathbb{1}\{\lambda \ge U_i/w_i\} + \lambda w_i \mathbb{1}\{\lambda < U_i/w_i\}\right) y_i. && (104)\end{aligned}$$

The maximizer $\lambda^\star$ of $L_D(\lambda)$ for $\lambda \ge 0$ can now be found by evaluating at the $N{+}1$ points $\{\lambda_0 = 0, \lambda_i = U_i/w_i\}$. Again, $\lambda^\star = 0$ corresponds to an inactive constraint under which the oracle can allocate $q_{o,i,K} = y_i$. When $\lambda^\star = U_{i^\star}/w_{i^\star}$, we get

$$\begin{aligned} q_{o,i,K} &= \operatorname*{argmin}_{0\le x\le y_i}\left\{\frac{U_{i^\star}}{w_{i^\star}} w_i x - U_i x + U_i y_i\right\} && (105)\\ &= \operatorname*{argmin}_{0\le x\le y_i}\left\{\left(\frac{U_{i^\star}}{w_{i^\star}} w_i - U_i\right) x\right\} && (106)\\ &= \begin{cases} 0 & \text{if } \frac{U_{i^\star}}{w_{i^\star}} > \frac{U_i}{w_i} \\ y_i & \text{if } \frac{U_{i^\star}}{w_{i^\star}} < \frac{U_i}{w_i}. \end{cases} && (107)\end{aligned}$$

If there is a single $i^\star$ for which $\lambda^\star = U_i/w_i$, the oracle's allocation in this coordinate is determined by the (now active) constraint as

$$q_{o,i^\star,K} = w_{i^\star}^{-1}\left(K - \sum_{i \mid \frac{U_{i^\star}}{w_{i^\star}} < \frac{U_i}{w_i}} w_i y_i\right). \tag{108}$$

If $I_K = \{i \mid \lambda^\star = U_i/w_i\}$ includes $n_K > 1$ indices, then the solutions $q_{o,i,K} \in [0, y_i]$ for $i \in I_K$ are not unique because the oracle's score will be constant on

$$\left\{\mathbf{x} \mid \sum_{i\in I_K} w_i x_i = K - \sum_{\frac{U_{i^\star}}{w_{i^\star}} < \frac{U_i}{w_i}} w_i y_i,\ x_i = q_{o,i,K} \text{ for } i \notin I_K\right\}. \tag{109}$$

(Note that restricted to the coordinates in I_K, the score level sets are parallel to the affine constraint hyperplane.) If a specific solution is desired, additional criteria could be invoked. A fairness criterion, for example, might require the remaining resources $K - \sum_{\frac{U_{i^\star}}{w_{i^\star}} < \frac{U_i}{w_i}} w_i y_i$ to be divided equally among the n_K recipients.

Special Case I: F_i from same location-scale family and $w_i = 1$, $O_i = O, U_i = 1$.

Also assume the constraint is active, i.e.,

$$K < \sum F_i^{-1}(\alpha) = \sum \mu_i + \sigma_i \Phi^{-1}(\alpha) \tag{110}$$

where Φ is the standard CDF for the family and $\alpha = (1+O)^{-1}$. Then setting

$$\sum F_i^{-1}(\alpha - \lambda) = \sum \mu_i + \sigma_i \Phi^{-1}(\alpha - \lambda) = K \tag{111}$$

we get

$$\alpha - \lambda = \Phi\left(\frac{K - \sum \mu_j}{\sum \sigma_j}\right) := F(K) \tag{112}$$

and the Bayes act when the constraint is set to K is

$$\begin{aligned} Q_i(K) = F_i^{-1}(F(K)) &= \mu_i + \sigma_i \Phi^{-1}(F(K)) \\ &= \mu_i + \sigma_i \left(\frac{K - \sum \mu_j}{\sum \sigma_j}\right) \\ &= \mu_i + \tilde{\sigma}_i (K - \sum \mu_j) \end{aligned} \tag{113}$$

where $\tilde{\sigma_i}$ is the proportion of F's scale (e.g. SD) “due” to F_i.

Remarks:

- $Q_i(K)$ here does not depend on O unlike when the constraint is inactive or the O_i differ.
- An “excess or shortage in mean” $K - \sum \mu_j$ is divided among the components in proportion to their scale factors as adjustmments up or down from their locations μ_i. This suggests an interpretation of under-dispersion of a forecast in one component as leading to that component not receiving as much additional recources as it should when there is an excess or other components not recieving as much of scarce resources as they should when there is a shortage.

The score for the forecast $\{F_i\}$ is

$$s_K(\mathbf{Q}, \mathbf{y}) = \sum (1 - (1+O)\mathbb{1}\left\{\frac{K - \sum \mu_i}{\sum \sigma_i} > \frac{y_i - \mu_i}{\sigma_i}\right\})(g_i(y_i) - g_i(Q_i(K))) \tag{114}$$

In the motivating case that $O = 0$, this becomes

$$s_K(\mathbf{Q}, \mathbf{y}) = \sum \mathbb{1}\left\{\frac{K - \sum \mu_i}{\sum \sigma_i} \leq \frac{y_i - \mu_i}{\sigma_i}\right\}(g_i(y_i) - g_i(Q_i(K))) \tag{115}$$

That is, a Q_i is only penalized when the standardized observed excess demand in component i exceeds the standardized excess in resources or the standardized observed shortfall in demand is not as large as the standardized shortage of resources.

Special Case II: F_i uniform on $[a_i, b_i]$

Problem: This does not seem to work. Take $F_1 = F_2 = F_{[0,1]}, \alpha_1 = \alpha_2 = .5, w_1 = 1, w_2 = 2$. Then

$$\lambda = \frac{3/2 - K}{1/2 + 4/2} = \frac{3 - 2K}{5} > \frac{1}{2}$$

for $K < 1/4$, so that $Q_2(\alpha_2(1 - w_2\lambda))$ is undefined. The minimizer here is $Q_1 = K, Q_2 = 0$, which does satisfy the KKT equations (A.21) and (A.22). $Q_1 = 0, Q_2 = K/2$ does not.

Following (Abdel-Malek, Montanari, and Morales 2004), we can directly evaluate (86) to solve for λ in the case of the uniform location-scale family ($\mu_i = a_i, \sigma = b_i - a_i$):

$$\begin{aligned}
K &= \sum w_i F^{-1}_{[a_i,b_i]}\left(\alpha_i(1 - w_i\lambda)\right) \tag{116}\\
&= \sum w_i Q_{[a_i,b_i]}\left(\alpha_i(1 - w_i\lambda)\right) \tag{117}\\
&= \sum w_i\left[a_i + (b_i - a_i)(\alpha_i - w_i\alpha_i\lambda)\right] \tag{118}\\
&= \sum w_i Q_{[a_i,b_i]}(\alpha_i) - (b_i - a_i)w_i^2\alpha_i\lambda \tag{119}
\end{aligned}$$

Writing $D = \sum w_i Q_{[a_i,b_i]}(\alpha_i) - K$ for the "deficit" of the unconstrained solution, we have

$$\lambda = \frac{D}{\sum (b_j - a_j)w_j^2\alpha_j} \tag{120}$$

$$\tag{121}$$

and from ??,

$$\begin{aligned}
Q_i &= F^{-1}_{[a_i,b_i]}\left(\alpha_i\left(1 - \frac{w_i D}{\sum_j (b_j - a_j)w_j^2\alpha_j}\right)\right) \tag{122}\\
&= Q_{[a_i,b_i]}(\alpha_i) - \frac{(b_i - a_i)w_i\alpha_i D}{\sum (b_j - a_j)w_j^2\alpha_j}. \tag{123}
\end{aligned}$$

That is, each constrained Q_i is obtained by reducing the α_i quantile by an amount of the deficit proportional to the product of spread, resource weighting, and cost/loss ratio (what should this be called?) of the corresponding prediction target.

General solution method

Is there anything we can salvage here?

Again following (Abdel-Malek, Montanari, and Morales 2004), we can use this solution for uniform F_i as the basis of an iterative procedure for approximating the Q_i under a constraint for general predictive distributions. Assuming that the unconstrained solution $\{Q_i(\alpha_i)\}$ is not feasible and that we can evaluate F_i, F_i^{-1} and f_i at least approximately at arbitrary x_i, we first linearize the equations (86) at $\{(\alpha_i, Q_i(\alpha_i))\}$ by replacing the F_i^{-1} with the tangential approximations

$$\left(F_i^{(1)}\right)^{-1}(p) := Q_i(\alpha_i) + \frac{p - \alpha_i}{f_i(Q_i(\alpha_i))}, \tag{124}$$

that is, the quantile functions of the uniform cdf's

$$F_i^{(1)} := \min(0, \max(1, \alpha_i + f_i(Q_i(\alpha_i))(x_i - Q_i(\alpha_i)))). \tag{125}$$

As before, these equations

$$\begin{aligned}
K &= \left(F_i^{(1)}\right)^{-1}\left(\alpha_i(1 - w_i\lambda)\right) \tag{126}\\
&= \sum w_i Q_i(\alpha_i) - \frac{w_i^2\alpha_i\lambda}{f_i(Q_i(\alpha_i))} \tag{127}
\end{aligned}$$

are linear in λ and have the solution

$$\lambda^{(1)} = \frac{D}{\sum \frac{w_j^2\alpha_j}{f_j(Q_j(\alpha_j))}}. \tag{128}$$

This provides the first iterate

$$Q_i^{(1)} = F_i^{-1}(\alpha_i(1 - w_i\lambda^{(1)})) \tag{129}$$

and a new deficit (or negative surplus?)

$$D^{(1)} = \sum w_i Q_i^{(1)} - K. \tag{130}$$

Now we can repeat the process with new tengential approximations

$$\left(F_i^{(2)}\right)^{-1}(p) := Q_i^{(1)} + \frac{p - \alpha_i^{(1)}}{f_i\left(Q_i^{(1)}\right)} \tag{131}$$

where $\alpha_i^{(1)} := \alpha_i(1 - w_i\lambda^{(1)})$ to get the next λ iterate

$$\lambda^{(2)} = \frac{D^{(1)}}{\sum \frac{w_j^2 \alpha_j^{(1)}}{f_j\left(Q_j^{(1)}\right)}}, \tag{132}$$

continuing with $\alpha_i^{(\tau)} := \alpha_i^{(\tau-1)}(1 - w_i\lambda^{(\tau)})$ until the relative error at the τ'th iteration

$$\varepsilon^{(\tau)} = D^{(\tau)}/K \tag{133}$$

is sufficiently small.

References

Abdel-Malek, Layek, Roberto Montanari, and Libia Cristina Morales. 2004. "Exact, Approximate, and Generic Iterative Models for the Multi-Product Newsboy Problem with Budget Constraint." *International Journal of Production Economics* 91 (2): 189–98.

Ehm, Werner, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. 2016. "Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 505–62.

Gneiting, Tilmann. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106 (494): 746–62. https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78. https://doi.org/10.1198/016214506000001437.

Hadley, G., and Thomson M. Whitin. 1963. *Analysis of Inventory Systems.* Prentice-Hall International Series in Management. Prentice-Hall.

Jose, Victor Richmond R., and Robert L. Winkler. 2009. "Evaluating Quantile Assessments." *Operations Research* 57 (5): 1287–97. https://doi.org/10.1287/opre.1080.0665.

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical Optimization.* Springer.

Ruszczynski, Andrzej. 2011. *Nonlinear Optimization.* Princeton: Princeton University Press.

Zhang, Bin, Xiaoyan Xu, and Zhongsheng Hua. 2009. "A Binary Solution Method for the Multi-Product Newsboy Problem with Budget Constraint." *International Journal of Production Economics* 117 (1): 136–41.

‘

Appendix A. Derivation Details

Differentiation of $\overline{s}_F(x)$ in terms of a density f:

$$\frac{d}{dx}\overline{s}_F(x) = \frac{d}{dx}\left[O\int_{-\infty}^{x}(g(x)-g(y))f(y)dy + U\int_{x}^{\infty}(g(y)-g(x))f(y)dy\right] \tag{A.1}$$

$$= O(g(x)-g(x))f(x) + O\int_{-\infty}^{x}\frac{d}{dx}\left[(g(x)-g(y))f(y)\right]dy \tag{A.2}$$

$$-U(g(x)-g(x))f(x) + U\int_{x}^{\infty}\frac{d}{dx}\left[(g(y)-g(x))f(y)\right]dy \tag{A.3}$$

$$= O\int_{-\infty}^{x}g'(x)f(y)dy - U\int_{x}^{\infty}g'(x)f(y)dy \tag{A.4}$$

$$= g'(x)\left(OF(x)-U(1-F(x))\right) \tag{A.5}$$

$$= g'(x)(O+U)\left(F(x)-\frac{U}{O+U}\right) \tag{A.6}$$

$$= \kappa g'(x)\left(F(x)-\alpha\right) \tag{A.7}$$

Comments on the dual problem formulation:

In the main text, we noted that after defining

$$L(\mathbf{x},\lambda) = \sum_{i=1}^{N}\overline{s}_{F_i}(x_i) + \lambda\left(\sum_{i=1}^{N}w_i x_i - K\right), \tag{A.8}$$

we have the dual problem of maximizing over $\lambda \geq 0$ the objective function

$$L_D(\lambda) = \min_{0\leq\mathbf{x}} L(\mathbf{x},\lambda) = -\lambda K + \sum_{i=1}^{N}\min_{0\leq x_i}\left\{\lambda w_i x_i + \overline{s}_{F,i}(x_i)\right\}. \tag{A.9}$$

The solution to this dual problem gives lower bounds on a solution to the primal AP; we verify this in more detail here.

First, note that for any $\mathbf{x}$ that is a feasible value for the AP, the constraint $\mathbf{w}^T\mathbf{x} \leq K$ holds, so that $\mathbf{w}^T\mathbf{x} - K \leq 0$ and for any $\lambda \geq 0$, $\lambda\left(\sum_{i=1}^{N}w_i x_i - K\right) \leq 0$. Therefore, for any feasible $\mathbf{x}$ and $\lambda \geq 0$,

$$L(\mathbf{x},\lambda) = \sum_{i=1}^{N}\overline{s}_{F_i}(x_i) + \lambda\left(\sum_{i=1}^{N}w_i x_i - K\right) \leq \overline{s}_{\mathbf{F}}(\mathbf{x}). \tag{A.10}$$

Thus, for a fixed $\lambda \geq 0$,

$$\min_{0\leq\mathbf{x},\mathbf{w}^T\mathbf{x}\leq K} L(\mathbf{x},\lambda) \leq \min_{0\leq\mathbf{x},\mathbf{w}^T\mathbf{x}\leq K}\overline{s}_{\mathbf{F}}(\mathbf{x}) \tag{A.11}$$

Now since $\{\mathbf{x} : 0 \leq \mathbf{x}, \mathbf{w}^T\mathbf{x} \leq K\} \subseteq \{\mathbf{x} : 0 \leq \mathbf{x}\}$ and using the fact that if $B \subseteq A$ then $\min_{x\in A} f(x) \leq \min_{x\in B} f(x)$, we have

$$L_D(\lambda) = \min_{0\leq\mathbf{x}} L(\mathbf{x},\lambda) \tag{A.12}$$

$$\leq \min_{0\leq\mathbf{x},\mathbf{w}^T\mathbf{x}\leq K} L(\mathbf{x},\lambda) \tag{A.13}$$

$$\leq \min_{0\leq\mathbf{x},\mathbf{w}^T\mathbf{x}\leq K}\overline{s}_{\mathbf{F}}(\mathbf{x}), \tag{A.14}$$

where the right hand side of Equation (A.14) is the optimal value of the AP. Since this holds for all λ and in particular for the maximizer $\lambda^\star$, we have $L_D(\lambda^\star) \leq \overline{s}_{\mathbf{F}}(\mathbf{x}^\star)$, where $\mathbf{x}^\star$ denotes the AP solution. This inequality goes by the name of weak duality.

Solution of the AP via KKT equations:

To describe forecasts solving the AP (70) without the direct use of duality theory, we begin by stating the necessary Karush-Kuhn–Tucker (KKT) conditions for a point $\mathbf{Q} \in C_K$ to be a solution of the AP. Since the feasible set C_K is a polyhedron on which $\overline{s}_F$ is smooth, these are: there exists a vector of multipliers $\boldsymbol{\mu} = (\lambda, \mu_1, \dots, \mu_N) \in \mathbb{R}^{1+N}$ such that

$$\lambda, \mu_1, \dots, \mu_N \geq 0 \tag{A.15}$$

$$-\nabla \overline{s}_F(\mathbf{Q}) = D^T \boldsymbol{\mu} \tag{A.16}$$

$$\lambda(\mathbf{w}^T \mathbf{Q} - K) = 0 \tag{A.17}$$

$$\mu_i Q_i = 0, \quad i = 1, \dots, N \tag{A.18}$$

(see for example, (Nocedal and Wright 2006), Theorem 12.1). These equations express the fact that at a minimizer, either the gradient $\nabla \overline{s}_F$ must vanish (so that $\boldsymbol{\mu} = \mathbf{0}$), or the direction of greatest descent $-\nabla \overline{s}_F$ must lie in the normal (outwardly pointing) cone of C_K, which is the non-negative span of some subset (generically a single one) of the columns of D^T. This subset corresponds to which constraints (i.e. "sides" of C_K) are active, allowing the corresponding multipliers in the "complementary slackness" equations (A.17) and (A.18) to be positive.

From (6), using $\mathbf{e}_i$ to denote the ith coordinate vector, the "Lagrange multiplier" equation (A.16) becomes

$$-\sum_{i=1}^{N} \kappa_i g_i'(Q_i)(F_i(Q_i) - \alpha_i)\mathbf{e}_i = [\mathbf{w}| - \mathrm{Id}_N]\boldsymbol{\mu} = \sum_{i=1}^{N} (\lambda w_i - \mu_i)\mathbf{e}_i \tag{A.19}$$

which decomposes to

$$\lambda = w_i^{-1}(\kappa_i g_i'(Q_i)(\alpha_i - F_i(Q_i)) + \mu_i), \quad i = 1, \dots, N. \tag{A.20}$$

Using (A.18), we can write this as

$$\lambda = \begin{cases} w_i^{-1}\kappa_i g_i'(Q_i)(\alpha_i - F_i(Q_i)) & \text{for } Q_i > 0 \quad \text{(A.21)} \\ w_i^{-1}(\kappa_i g_i'(0)(\alpha_i - F_i(0)) + \mu_i) & \text{for } Q_i = 0. \quad \text{(A.22)} \end{cases}$$

This leaves us with two alternatives.

(I) $\lambda = 0$, i.e., the constraint K on the total resources allocated is not active. Then (A.21) forces

$$F_i(Q_i) = \alpha_i \text{ for } Q_i > 0, \tag{A.23}$$

and (A.22) forces

$$F_i(0) = \alpha_i + \mu_i / \kappa_i g_i'(0) \geq \alpha_i \text{ for } Q_i = 0. \tag{A.24}$$

That is, $\mathbf{Q}$ is a vector with entries $Q_i = \max(Q_i(\alpha_i), 0)$.

(II) $\lambda > 0$, so that the constraint K is active. In this case (A.21) forces

$$F_i(Q_i) < \alpha_i \text{ for } Q_i > 0, \tag{A.25}$$

and for any i with $Q_i = 0$, (A.22) along with $\mu_i \geq 0$ gives a lower bound on λ,

$$\lambda \geq w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0)). \tag{A.26}$$

And from (A.17), the constraint must be active, i.e.,

$$\mathbf{w}^T \mathbf{Q} = K. \tag{A.27}$$

Now since $\overline{s}_F$ is a continuous function on the compact set C_K, some solution $\mathbf{Q} \in C_K$ of the AP exists. Uniqueness, however, requires some additional conditions on the g_i over the simplex $\{\mathbf{x} \in C_K \mid \mathbf{w}^T\mathbf{x} = K\}$. One such set of conditions – that for convenience we'll require over all of C_K – is that the g_i have non-positive second derivatives,

$$g_i''(x_i) \leq 0 \text{ whenever } F_i(x_i) < \alpha_i \tag{A.28}$$

so that from (7), the second partial derivatives $\overline{s}_{F,x_ix_i}$ are all non-negative on C_K. Because $\overline{s}_{F,x_ix_j} = 0, i \neq j$, the Hessian $\nabla^2\overline{s}_F$ is then positive semi-definite on C_K so that $\overline{s}_F$ in convex on C_K and any local minimizer for the AP is actually global. (Note that we might have to consider minimizing sets if any of the densities f_i vanish at a minimizer, but we'll set aside this possibility for now.) And along with (A.20), requiring (A.28) gives the upper bound

$$\lambda \leq \min_i w_i^{-1}(\kappa_i g_i'(0)(\alpha_i - F_i(0)) + \mu_i). \tag{A.29}$$

Finding this unique $\mathbf{Q}$ can sometimes be straightforward. Suppose, for example, there are no point masses, $F_i : [0, \varepsilon_i] \to [0, \delta_i]$ are bijective for some $\varepsilon_i, \delta_i > 0$ and all i, and g_i' are all identically 1, and that the constraint is active with $\mathbf{w}^T\mathbf{Q}_\mathbf{F}(\boldsymbol{\alpha}) = K_1 > K$. Then we can write the equations (A.21) as

$$Q_i(\lambda) = F_i^{-1}(\alpha_i - \lambda w_i/\kappa_i) = F_i^{-1}(\alpha_i(1 - \lambda w_i/U_i)) \tag{A.30}$$

which are all defined for λ in the interval $I = [0, \lambda_0 = \min(U_i/w_i)]$. By our assumptions, the left hand side of the constraint equation (A.27)

$$\sum_{i=1}^{N} w_i Q_i(\lambda) = K. \tag{A.31}$$

is a continuous function mapping I to $[K_1, \sum_{i=1}^N w_i Q_i(\lambda_0)]$. If K is sufficently close to K_1, we will have a root $0 < \lambda^\star < \lambda_0$ which can be easily found using a formal or numerical search in I. And since our assumptions imply that $Q_i(\lambda^\star) > 0$, the KKT conditions (A.21) and (A.22) are met.

But if the constraint is too tight or the lower bounds of the supports of the F_i differ, we will potentially need to repeat the root search for each possible set $\{i \mid Q_i = 0\}$, facing intractability for large N. Our convexity assumption (A.28), however, besides ensuring existence and uniqueness, makes available a binary search method of (Zhang, Xu, and Hua 2009) which is only of polynomial complexity in N. In particular (A.28) guarantees that the functions (84) are decreasing in x_i (cf. (7)) from

$$\lambda_i(0) = w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0)) \quad \text{to} \quad \lambda_i(Q_i(\alpha_i)) = 0. \tag{A.32}$$

They can therefore can be inverted to give decreasing functions $x_i(\lambda_i)$ with

$$x_i(0) = Q_i(\alpha_i) \quad \text{and} \quad x_i(w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0))) = 0. \tag{A.33}$$

ZXH binary search algorithm

```
Input: $F_i, Q_i, \kappa_i, \alpha_i, w_i, g_i, i = 1, \ldots, N,$
$K, \varepsilon_K, \varepsilon_\lambda, \mathrm{root}(f, I)$    ▷ root a function returning a root of function $f$ on interval $I$

for $i = 1$ to $N$ do
    $x_i \leftarrow \max(0, \min(Q_i(\alpha_i), w_i^{-1}2K))$    ▷ ensure constraint is violated if any $\alpha_i = 1$
if $\mathbf{w}^T\mathbf{x} \leq K$ then return $\mathbf{Q} = \mathbf{x}$    ▷ return quantiles if they satisfy constraint
$\lambda_1 \leftarrow \lambda_L \leftarrow 0$, $\lambda_U \leftarrow \max_i \lambda_i(0)$, $\tau \leftarrow 2$
if $\lambda_U \leq 0$ then
    return $\mathbf{Q} = \mathbf{0}$    ▷ handle case where all expected component scores increasing at 0

while $|(\mathbf{w}^T\mathbf{x} - K)/K| > \varepsilon_K$ or $\lambda_U - \lambda_L > \varepsilon_\lambda$ do
    $\lambda_\tau \leftarrow (\lambda_U + \lambda_L)/2$
    for $i = 1$ to $N$ do
        if $\lambda_\tau < \lambda_i(0)$ then    ▷ that is, if, from (A.22) and (A.28), $\lambda_\tau$ cannot be a multiplier for $x_i = 0$

            if $\lambda_\tau < \lambda_{\tau-1}$ then
                $I_{i,\tau} \leftarrow [x_i, Q_i(\alpha_i)]$
            else
                $I_{i,\tau} \leftarrow [0, x_i]$
            $x_i \leftarrow \mathrm{root}(\lambda_i(x) - \lambda_\tau, I_{i,\tau})$    ▷ well-defined since $\lambda_i$, defined by (84), is decreasing

        else
            $x_i \leftarrow 0$    ▷ prevent decrease past 0 if $\lambda_\tau > \lambda_{\tau-1}$ and maintain at 0 if $\lambda_i(0) < \lambda_\tau < \lambda_{\tau-1}$

    if $\mathbf{w}^T\mathbf{x} < K$ then
        $\lambda_U \leftarrow \lambda_\tau$    ▷ unused capacity so $\lambda_{\tau+1} < \lambda_\tau$ and any $x_i$ could be increased but none will decrease
    else
        $\lambda_L \leftarrow \lambda_\tau$    ▷ capacity exceeded so $\lambda_{\tau+1} > \lambda_\tau$ and no $x_i$ will increase so any set to 0 will remain so

    $\tau \leftarrow \tau + 1$
return $\mathbf{Q} = \mathbf{x}$
```