Forecast Similarity Using Cramer Distance Approximation

Johannes Bracher, Evan Ray, Nick Reich, Nutcha Wattanachit

09/22/2021

Cramer Distance

Consider two predictive distributions F and G. Their *Cramer distance* or *integrated quadratic distance* is defined as

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$$

where $F(x)$ and $G(x)$ denote the cumulative distribution functions. It can also be written as

$$\text{CD}(F, G) = \mathbb{E}_{F,G}|x - y| - 0.5 \left[\mathbb{E}_F|x - x'| + \mathbb{E}_G|y - y'|\right], \tag{1}$$

where x, x' are independent random variables following F and y, y' are independent random variables following G. This formulation illustrates that the Cramer distance depends on the shift between F and G (first term) and the variability of both F and G (of which the two last expectations in above equation are a measure).

The Cramer distance is the divergence associated with the continuous ranked probability score (Thorarinsdottir 2013, Gneiting and Raftery 2007), which is defined by

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx =) \tag{2}$$

$$= 2 \int_0^1 ((\mathbf{1}(y \leq q_k^F) - \tau_k)(q_k^F - y) d\tau_k \tag{3}$$

where y denotes the observed value. Indeed, it is a generalization of the CRPS as it simplifies to the CRPS if one out of F and G is a one-point distribution. Indeed, it is a generalization of the CRPS as it simplifies to the CRPS if one out of F and G is a one-point distribution. The Cramer distance is commonly used to measure the similarity of forecast distributions (see Richardson et al 2020 for a recent application).

Cramer Distance Approximation for Equally-Spaced Intervals

Now assume that for each of the distributions F and G we only know K quantiles at equally spaced levels $\tau_1 = 1/(K+1), \tau_2 = 2/(K+1), \ldots, \tau_K = K/(K+1)$. Denote these quantiles by $q_1^F \leq q_1^F \leq \cdots \leq q_K^F$ and $q_1^G \leq q_2^G \leq \cdots \leq q_K^G$, respectively. It is well known that the CRPS can be approximated by an average of linear quantile scores (Laio and Tamea 2007, Gneiting and Raftery 2007):

$$\text{CRPS}(F, y) \approx \frac{1}{K} \times \sum_{k=1}^{K} 2\{\mathbf{1}(y \leq q_k^F) - k/(K+1)\} \times (q_k^F - y). \tag{4}$$

This approximation is equivalent to the weighted interval score (WIS) which is in use for evaluation of quantile forecasts at the Forecast Hub, see Section 2.2 of Bracher et al (2021). This approximation can be generalized to the Cramer distance as

$$\mathrm{CD}(F,G) \approx \frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times \mathbf{1}\{(i \leq j \wedge q_i^F > q_j^G) \vee (i \geq j \wedge q_i^F < q_j^G)\} \times \left|q_i^F - q_j^G\right|. \tag{5}$$

This can be seen as a sum of penalties for *incompatibility* of predictive quantiles. Whenever the predictive quantiles q_i^F and q_j^G are incompatible in the sense that they imply F and G are different distributions (because $q_F^i > q_G^j$ despite $i \leq j$ or vice versa), a penalty $\left|q_i^F - q_j^G\right|$ is added.

Cramer Distance Approximation for Unequally-Spaced Intervals

Suppose we have quantiles $q_1^F, ..., q_K^F$ and $q_1^G, ..., q_K^G$ at K probability levels $\tau_1, ..., \tau_K$ (with $\tau_1 = 0$) from two distributions F and G. Define the combined vector of quantiles $q_1, ..., q_{2K}$ by combining the vectors $q_1^F, ..., q_K^F$ and $q_1^G, ..., q_K^G$ and sorting them in an ascending order. The CRPS can be approximated as follows

$$\mathrm{CRPS}(F,y) \approx \frac{1}{K} \sum_{k=1}^{K} 2\{\mathbf{1}(y \leq q_k^F) - \tau_k\} \times (q_k^F - y). \tag{6}$$

This approximation can be generalized to the Cramer distance as

$$\mathrm{CD}(F,G) \approx \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times w_{ij} \times \mathbf{1}\{(i \leq j \wedge q_i^F > q_j^G) \vee (i \geq j \wedge q_i^F < q_j^G)\} \times \left|q_i^F - q_j^G\right|, \tag{7}$$

where $w_{ij} = |\tau_i - \tau_j|$ (the difference of the probability levels). The details on how to go from (7) to the Riemann sums are still being worked out. Essentially, we can approximate the Cramer distance by eliminating the tails of the integral to the left of q_1 and the right of q_{2K}, and approximating the center via a Riemann sum:

$$\begin{aligned}
\mathrm{CD}(F,G) &= \int_{-\infty}^{\infty} F(x) - G(x)^2 dx && (8)\\
&\approx \int_{q_1}^{q_{2K}} F(x) - G(x)^2 dx && (9)\\
&= \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 dx && (10)
\end{aligned}$$

There are a variety of options that can be used for each term in this sum, for instance:

Left-sided Riemann sum approximation

$$\begin{aligned}
\mathrm{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 && (11)\\
&\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) && (12)\\
& && (13)
\end{aligned}$$

Since $q_j \in \{q_1, ..., q_{2K}\}$ belongs to either $q_1^F, ..., q_K^F$ or $q_1^G, ..., q_K^G$, we can rewrite the above approximation using $\tau_1, ..., \tau_K$ as follows

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \tag{14}$$

$$= \sum_{j=1}^{2K-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) \tag{15}$$

where $\tau_j^F \in \tau_F$ and $\tau_j^G \in \tau_G$. τ_F and τ_G are vectors of length $2K-1$ with elements

$$\tau_j^F = \begin{cases} I(q_1 = q_1^F) \times \tau_{q_1}^F & \text{for } j = 1 \\ I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{q_j}^F + I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{j-1}^F & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^F$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from F, and τ_{j-1}^F is the $(j-1)^{th}$ probability level in τ_F.

$$\tau_j^G = \begin{cases} I(q_1 = q_1^G) \times \tau_{q_1}^G & \text{for } j = 1 \\ I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{q_j}^G + I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{j-1}^G & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^G$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from G, and τ_{j-1}^G is the $(j-1)^{th}$ probability level in τ_G.

Trapezoidal rule

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{16}$$

$$\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \tag{17}$$

$$\tag{18}$$

Similarly, we can rewrite the above approximation using $\tau_1, ..., \tau_K$ as defined in the left-sided Riemann sum approximation as follows

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \tag{19}$$

$$= \sum_{j=1}^{2K-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2} (q_{j+1} - q_j). \tag{20}$$

Cramer Distance Approximation for Unequally-Spaced Intervals and Different Probability Levels

We (probably) can further modify the formula of the Cramer distance approximation for unequally-spaced intervals to accommodate different probability levels from F and G. Suppose we have quantiles $q_1^F, ..., q_N^F$ at

K probability levels $\tau_1^F, ..., \tau_N^F$ from the distribution F, and $q_1^G, ..., q_M^G$ at M probability levels $\tau_1^G, ..., \tau_M^G$ from the distribution G. Define the combined vector of quantiles $q_1, ..., q_{N+M}$ by combining the vectors $q_1^F, ..., q_N^F$ and $q_1^G, ..., q_M^G$ and again sorting them in an ascending order. Using the same definitions as previously defined, we can approximate the Cramer distance via a Riemann sum as follows:

Left-sided Riemann sum approximation

$$\text{CD}(F, G) \approx \sum_{j=1}^{N+M-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{21}$$

$$\approx \sum_{j=1}^{N+M-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j), \tag{22}$$

$$\tag{23}$$

which we can rewrite using $\tau_1^F, ..., \tau_N^F$ and $\tau_1^G, ..., \tau_M^G$ as follows

$$\text{CD}(F, G) \approx \sum_{j=1}^{N+M-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \tag{24}$$

$$= \sum_{j=1}^{N+M-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) \tag{25}$$

where $\tau_j^F \in \tau_F$ and $\tau_j^G \in \tau_G$. τ_F and τ_G are vectors of length $N + M - 1$ with elements

$$\tau_j^F = \begin{cases} \tau_{q_j}^F & \text{if } q_j \in \{q_1^F, ..., q_N^F\} \\ \tau_{q_{j-1}}^F & \text{if } q_j \notin \{q_1^F, ..., q_N^F\} \end{cases}$$

where $\tau_{q_j}^F$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from F.

$$\tau_j^G = \begin{cases} \tau_{q_j}^G & \text{if } q_j \in \{q_1^G, ..., q_M^G\} \\ \tau_{q_{j-1}}^G & \text{if } q_j \notin \{q_1^G, ..., q_M^G\} \end{cases}$$

where $\tau_{q_j}^G$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from G.

Trapezoidal rule

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{26}$$

$$\approx \sum_{j=1}^{N+M-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j), \tag{27}$$

$$\tag{28}$$

which we can rewrite as follows

$$\mathrm{CD}(F,G) \approx \sum_{j=1}^{N+M-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2}(q_{j+1} - q_j) \tag{29}$$

$$= \sum_{j=1}^{N+M-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2}(q_{j+1} - q_j). \tag{30}$$

Decomposition of Approximated Cramer Distance

The Cramer distance is commonly used to measure the similarity of forecast distributions (see Richardson et al 2020 for a recent application). Now assume that for each of the distributions F and G we only know K quantiles at equally spaced levels $1/(K+1), 2/(K+1), \ldots, K/(K+1)$. Denote these quantiles by $q_1^F, \ldots, q_K^F$ and $q_1^G, \ldots, q_K^G$, respectively. This CRPS approximation given by (**??**) is equivalent to the weighted interval score (WIS) which is in use for evaluation of quantile forecasts at the Forecast Hub, see Section 2.2 of Bracher et al (2021). This approximation can be generalized to the Cramer distance as

$$\mathrm{CD}(F,G) \approx \frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{1}\{(i-j) \times (q_i^F - q_j^G) \leq 0\} \times \left| q_i^F - q_j^G \right|, \tag{31}$$

This can be seen as a sum of penalties for *incompatibility* of predictive quantiles. Whenever the predictive quantiles q_i^F and q_j^G are incompatible in the sense that they imply F and G are different distributions (e.g. because $q_F^i > q_G^j$ despite $i < j$ or $q_F^i \neq q_G^j$ despite $i = j$), a penalty $\left| q_i^F - q_j^G \right|$ is added to the sum. This corresponds to the shift which would be necessary to make q_F^i and q_G^j compatible.

A divergence measure for central prediction intervals with potentially different nominal coverages

Consider two central prediction intervals $[l^F, u^F]$ and $[l^G, u^G]$ with nominal levels α^F and α^G, respectively (meaning that l^F is the $(1-\alpha^F)/2$ quantile of F etc). We can define an *interval divergence* measure by comparing the two pairs of predictive quantiles and summing up the respective incompatibility penalties as in (31). Adapting notation to the interval formulation and structuring the sum slightly differently, this can be written as:

$$\begin{aligned}
\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = {} & \mathbf{1}(\alpha^F \leq \alpha^G) \times \left\{ \max(l^G - l^F, 0) + \max(u^F - u^G, 0) \right\} + \\
& \mathbf{1}(\alpha^F \geq \alpha^G) \times \left\{ \max(l^F - l^G, 0) + \max(u^G - u^F, 0) \right\} + \\
& \max(l^F - u^G, 0) + \\
& \max(l^G - u^F, 0)
\end{aligned}$$

The first row adds penalties for the case where $[l^F, u^F]$ should be nested in $[l^G, u^G]$, but at least one of its ends is more extreme than the respective end of $[l^G, u^G]$. The second row covers the converse case. The last two rows add penalties if the lower end of one interval exceeds the upper end of the other, i.e. the intervals do not overlap.

This can be seen as a (scaled version of a) generalization of the interval score, but writing out the exact relationship is a bit tedious.

We now define four auxiliary terms with an intuitive interpretation which add up to the interval divergence:

- The term
$$D_F = \mathbf{1}(\alpha^F \leq \alpha^G) \times \max\{(u^F - l^F) - (u^G - l^G), 0\}$$
is the sum of penalties resulting from F being more dispersed than G. It is positive whenever the interval $[l^F, u^F]$ is longer than $[l^G, u^G]$, even though it should be nested in the latter. D_F then tells us by how much we would need to shorten $[l^F, u^F]$ so it could fit into $[l^G, u^G]$.

- The term
$$D_G = \mathbf{1}(\alpha^G \leq \alpha^G) \times \max\{(u^G - l^G) - (u^F - l^F), 0\}$$
measures the converse, i.e. overdispersion of G relative to F.

- The term
$$S^F = \max\{\mathbf{1}(\alpha^G \leq \alpha^F) \times \max(l^F - l^G, 0) + \mathbf{1}(\alpha^F \leq \alpha^G) \times \max(u^F - u^G, 0) + \max(l^F - u^G, 0) - D_F - D_G, 0\}$$
sums over penalties for values in $\{l^F, u^F\}$ exceeding those from $\{l^G, u^G\}$ where they should not (only counting penalties not already covered in D_F or D_G). It thus represents an *upward shift* of F relative to G.

- The term
$$S^G = \max\{\mathbf{1}(\alpha^F \leq \alpha^G) \times \max(l^G - l^F, 0) + \mathbf{1}(\alpha^G \leq \alpha^F) \times \max(u^G - u^F, 0) + \max(l^G - u^F, 0) - D_G - D_F, 0\}$$
accordingly represents an *upward shift* of G relative to F.

It can be shown that
$$\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = D_F + D_G + S^F + S^G$$
Intuitively the interval divergence measures by how much we need to move the quantiles of the interval with lower nominal coverage so it fits into the one with larger nominal coverage.

Approximating the Cramer distance using interval divergences

Assuming K is even, the K equally spaced predictive quantiles of each distribution can seen as $L = K/2$ central prediction intervals with coverage levels $\alpha_i = 2i/(L+1), i = 1, \ldots, L$. Similarly to the definition of the WIS, the approximation (31) can also be expressed in terms of these intervals as

$$\mathrm{CD}(F, G) \approx \frac{1}{2L(2L+1)} \sum_{k=1}^{L} \sum_{m=1}^{L} \mathrm{ID}([l_k^F, u_k^F], [l_m^G, u_m^G], \alpha_k^F, \alpha_m^G).$$

This implies a decomposition of the Cramer distance into the four interpretable components defined for the interval divergence in the previous section. If G is a one-point distribution, the CD reduces to the WIS and the proposed decomposition reduces to the well-known decomposition of the WIS into dispersion, overprediction and underprediction components.

Note that in practice we usually have an uneven rather than even number K of predictive quantiles. In this case the median needs to be treated separately (comparisons of the "0% prediction interval'' need to be weighted down with a factor of 2; this is the same little quirk as the one identified by Ryan and Evan for the WIS a few months ago). The decomposition has the following properties:

- Additive shifts of the two distributions only affect the shift components, not the dispersion components.
- Consequently, if G and G are identical up to an additive shift, both dispersion components will be 0.
- If F and G are both symmetric and have the same median, the both shift components will be 0.
- I think that in general it is possible that both shift components or both dispersion components are greater than 0, which leads to a somewhat strange interpretation. But this should only concern constructed examples.

Decomposition Preliminaries

Motivation of the approximation

We start by splitting up the sum from (31) into

$$\begin{aligned}\mathrm{CD}(F,G) \approx & \frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times \mathbf{1}\{i \leq j \wedge q_i^F > q_j^G\} \times \left(q_i^F - q_j^G\right) \\ + & \frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times \mathbf{1}\{i \geq j \wedge q_i^F < q_j^G\} \times \left(q_j^G - q_i^F\right)\end{aligned} \tag{32}$$

We now denote by $q_1 \leq q_2 \leq \cdots \leq q_{2n}$ the pooled set of quantiles (across F and G). Further, we denote by $r_1^F, \ldots r_n^F$ and $r_1^G, \ldots r_n^G$ the ranks of the members of $q_1^F, \ldots, q_n^F$ and $q_1^G, \ldots, q_n^G$, respectively, within the pooled set of quantiles, i.e.

$$q_i^F = q_{r_i^F}. \tag{33}$$

Note that ranks are oriented such that larger ranks correspond to larger values. We now focus on the first of the two double sums from equation (32), which using (33) becomes

$$\frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times \mathbf{1}\{i \leq j \wedge r_i^F > r_j^G\} \times \left(q_{r_i^F} - q_{r_j^G}\right). \tag{34}$$

Now denote by

$$\delta_l = q_{l+1} - q_l$$

the increments in the pooled set of quantiles. We can then use a telescope sum and continue (34) as

$$= \frac{1}{K(K+1)} \sum_{i=1}^{K} \sum_{j=1}^{K} 2 \times \mathbf{1}\{i \leq j \wedge r_i^F > r_j^G\} \times \sum_{l=r_j^G}^{r_{i-1}^F} \delta_l \tag{35}$$

$$= \frac{1}{K(K+1)} \sum_{l=1}^{K} \delta_l \times 2 \times \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{1}\{i \leq j \wedge r_j^G \leq l < r_i^F\}. \tag{36}$$

$$\tag{37}$$

The double sum over the indicator function counts how many pairs of (i, j) exist for a given l such that r_i^F, but not r_j^G exceeds l, despite $i \leq j$. To determine this number consider

$$a_l^F = \sum_{i=1}^{n} \mathbf{1}(r_i^F \leq l) \tag{38}$$

$$a_l^G = \sum_{i=1}^{n} \mathbf{1}(r_i^G \leq l), \tag{39}$$

i.e., the numbers of ranks falling below l among the quantiles of F and G. If

$$a_l^F - a_l^G = b_l \Leftrightarrow a_l^G = a_l^F - b_l$$

we have

$$r_1^F \leq \cdots \leq r_{a_l^F}^F \leq l < r_{a_l^F+1}^F \leq \cdots \leq r_n^F, \tag{40}$$

$$r_1^G \leq \cdots \leq r_{a_l^F-b_l}^G \leq l < r_{a_l^F-b_l+1}^F \leq \cdots \leq r_n^G. \tag{41}$$

The case $(i \leq j \wedge r_j^G \leq l < r_i^F)$ thus arises for

1. the tuples $(r^F_{a^F_l}, r^G_{a^F_l}), (r^F_{a^F_l}, r^G_{a^F_l-1}), \dots, (r^F_{a^F_l}, r^G_{a^F_l-(b_l-1)})$, i.e. b_l times for $r^F_{a^F_l}$.
2. the tuples $(r^F_{a^F_l-1}, r^G_{a^F_l-1}), (r^F_{a^F_l}, r^G_{a^F_l-1}), \dots, (r^F_{a^F_l}, r^G_{a^F_l-(b_l-1)})$, i.e. $b_l - 1$ times for $r^F_{a^F_l-1}$.

$\vdots$

b_l. the tuple $(r^F_{a^F_l-b_l}, r^G_{a^F_l-b_l})$, i.e. once for $r^F_{a^F_l-b_l}$.

This results in a total of $b_l + (b_l - 1) + \cdots + 1 = b_l(b_l+1)/2$ tuples, and we can re-write expression (36) as

$$= \frac{1}{K(K+1)} \sum_{l=1}^{K} \delta_l \times b_l(b_l+1) \times \mathbf{1}(b_l > 0).$$

The same argument can be made for the second double sum in (32), and bringing the two back together again the overall approximation from (32) simplifies to

$$\mathrm{CD}(F,G) \approx \frac{1}{K(K+1)} \sum_{l=1}^{K} \delta_l \times b_l(b_l+1). \tag{42}$$

For large K we obviously have

$$F(q_l) - G(q_l) = \underbrace{\frac{a^F_l}{K} - \frac{a^G_l}{K}}_{=b_l/K} \approx \underbrace{\frac{a^F_l+1}{K+1} - \frac{a^G_l+1}{K+1}}_{(b_l+1)/(K+1)}, \tag{43}$$

meaning that (42) is a simple (left-sided) Riemann sum approximation of the Cramer divergence.

There are more direct ways of approximating the Cramer divergence using quantiles, e.g., using b_l^2 rather than $b_l \times (b_l+1)$ in (42)). The motivation for expression (31) is that if G is a point mass, the approximated Cramer divergence simplifies to the approximation (4) already in use for the CRPS in the context of forecast evaluation. To see this consider equation (32). With $q^G_1 = \cdots = q^G_K = y$ it becomes

$$\begin{aligned}
\mathrm{CD}(F,G) \approx{}& \frac{1}{K(K+1)} \sum_{i=1}^{K}\sum_{j=1}^{K} 2 \times \mathbf{1}(i \le j \wedge q^F_i > y) \times (q^F_i - y) \\
&+ \quad \frac{1}{K(K+1)} \sum_{i=1}^{K}\sum_{j=1}^{K} 2 \times \mathbf{1}(i \ge j \wedge q^F_i < y) \times (y - q^F_i) \\
={}& \frac{1}{K(K+1)} \left\{ \sum_{i=1}^{K} 2 \times \mathbf{1}(q^F_i > y) \times i \times (q^F_i - y) \quad + \quad \sum_{i=1}^{K} 2 \times \mathbf{1}(q^F_i < y) \times (K+1-i) \times (y - q^F_i) \right\} \\
={}& \frac{1}{K(K+1)} \sum_{i=1}^{K} 2 \times \{\mathbf{1}(q^F_i > y) \times (K+1) - i\} \times (q^F_i - y) \\
={}& \frac{1}{K} \sum_{i=1}^{K} 2 \times \{\mathbf{1}(q^F_i > y) - i/(K+1)\} \times (q^F_i - y).
\end{aligned}$$

This is precisely the approximation of the CRPS from equation (4).

Establishing a decomposition for the approximated Cramer distance

We now introduce a decomposition of the approximated Cramer distance into the four following components:

- larger dispersion of F relative to G,
- larger dispersion of G relative to F,
- upward shift of F relative to G,

- upward shift of G relative to F.

This decomposition is inspired by the decomposition of the interval score which translates to the weighted interval score (WIS). To extend it to the approximated Cramer distance, we need to express it in terms of symmetric prediction intervals, similar to the definition of the WIS via the interval score. In the following we introduce such a representation and decomposition.

A divergence measure for central prediction intervals with potentially different nominal coverages

Consider two central prediction intervals $[l^F, u^F]$ and $[l^G, u^G]$ with nominal levels $\alpha^F, \alpha^G \in [0, 1)$, respectively; l^F is thus the $(1 - \alpha^F)/2$ quantile of F, u^F is the $(1 + \alpha^F)/2$ quantile of F etc. Note that we include the boundary case $\alpha = 0$, even though it has somewhat peculiar behaviour. We can define an *interval divergence* measure by comparing the two pairs of predictive quantiles and summing up the four resulting incompatibility penalties as in (31). Writing this out completely gives the somewhat unwieldy expression

$$\begin{aligned}
\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = \quad & \mathbf{1}\big\{\{(1-\alpha^F)/2 \leq (1-\alpha^G)/2 \wedge l^F > l^G\} \\
& \qquad \vee \{(1-\alpha^F)/2 \geq (1-\alpha^G)/2 \wedge l^F < l^G\}\big\} \times |l^F - l^G| \\
+ \ & \mathbf{1}\big\{\{(1-\alpha^F)/2 \leq (1+\alpha^G)/2 \wedge l^F > u^G\} \\
& \qquad \vee \{(1-\alpha^F)/2 \geq (1+\alpha^G)/2 \wedge l^F < u^G\}\big\} \times |l^F - u^G| \\
+ \ & \mathbf{1}\big\{\{(1+\alpha^F)/2 \leq (1-\alpha^G)/2 \wedge u^F > l^G\} \\
& \qquad \vee \{(1+\alpha^F)/2 \geq (1-\alpha^G)/2 \wedge u^F < l^G\}\big\} \times |u^F - l^G| \\
+ \ & \mathbf{1}\big\{\{(1+\alpha^F)/2 \leq (1+\alpha^G)/2 \wedge u^F > u^G\} \\
& \qquad \vee \{(1+\alpha^F)/2 \geq (1+\alpha^G)/2 \wedge u^F < u^G\}\big\} \times |u^F - u^G|.
\end{aligned}$$

By construction we know that if $\alpha^F > 0$ or $\alpha^G > 0$ we have $1 - \alpha^F < 1 + \alpha^G$ and $1 - \alpha^G < 1 + \alpha^F$ while $(1 - \alpha^F)/2 \leq (1 - \alpha^G)/2 \Leftrightarrow \alpha_F \geq \alpha^G \Leftrightarrow (1 + \alpha^F)/2 \geq (1 + \alpha^G)/2$ etc. In this case the above can thus be simplified considerably to

$$\begin{aligned}
\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = \ & \mathbf{1}(\alpha^F \leq \alpha^G) \times \big\{\max(l^G - l^F, 0) + \max(u^F - u^G, 0)\big\} \\
+ \ & \mathbf{1}(\alpha^F \geq \alpha^G) \times \big\{\max(l^F - l^G, 0) + \max(u^G - u^F, 0)\big\} \\
& + \max(l^F - u^G, 0) \ + \\
& + \max(l^G - u^F, 0).
\end{aligned}$$

Here, the first row adds penalties for the case where $[l^F, u^F]$ should be nested in $[l^G, u^G]$, but at least one of its ends is more extreme than the respective end of $[l^G, u^G]$. The second row covers the converse case. The last two rows add penalties if the lower end of one interval exceeds the upper end of the other, i.e. the intervals do not overlap. This can be seen as a (scaled version of a) generalization of the interval score, but writing out the exact relationship is a bit tedious.

If $\alpha^F = \alpha^G = 0$ we have

$$\mathrm{ID}([l^F, u^F], [l^G, u^G], 0, 0) = 4 \times |m^F - m^G|$$

where $m^F = l^F = u^F$ and $m^G = l^G = u^G$ are the predictive medians of F and G, respectively.

We now define four auxiliary terms with an intuitive interpretation which add up to the interval divergence:

- The term

  $$D_F = \mathbf{1}(\alpha^F \leq \alpha^G) \times \max\{(u^F - l^F) - (u^G - l^G), 0\}$$

  is the sum of penalties resulting from F being more dispersed than G. It is positive whenever the interval $[l^F, u^F]$ is longer than $[l^G, u^G]$, even though it should be nested in the latter. D_F then tells us by how much we would need to shorten $[l^F, u^F]$ so it could fit into $[l^G, u^G]$.

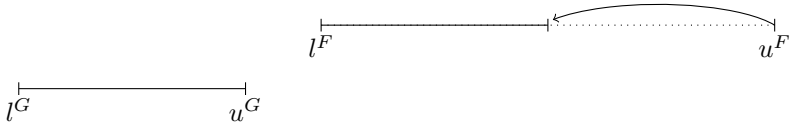
- The term
$$D_G = \mathbf{1}(\alpha^G \leq \alpha^G) \times \max\{(u^G - l^G) - (u^F - l^F), 0\}$$
measures the converse, i.e. overdispersion of G relative to F. Note that at most one of D_F and D_G can be positive. If $\alpha^F = \alpha^G = 0$ we always have $D^F = D^G = 0$.

- The term
$$S^F = \mathbf{1}\left\{l^F + u^F > l^G + u^G\right\} \times \left\{\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) - D^F - D^G\right\}$$
represents an *upward shift* of F relative to G. It is zero unless the center of $[l^F + u^F]$ exceeds that of $[l^G + u^G]$, in which case it absorbs the remaining penalties after accounting for differences in dispersion via D^F and D^G.

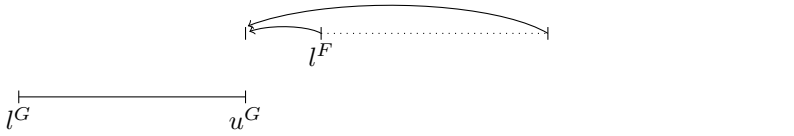
- The term
$$S^G = \mathbf{1}\left\{l^F + u^F < l^G + u^G\right\} \times \left\{\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) - D^F - D^G\right\}$$
accordingly represents an *upward shift* of G relative to F. Again note that at most one out of S^F and S^G can be positive.

It is easy to see that
$$\mathrm{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = D^F + D^G + S^F + S^G.$$
Intuitively the interval divergence measures by how much we need to move the quantiles of the interval with lower nominal coverage so it fits into the one with larger nominal coverage. The different components correspond to different types of moves we can make to achieve this We illustrate this using an example: Assume $[l^F, u^F]$ has lower nominal coverage than $[l^G, u^G]$, but is wider while $l^F > u^G$ (i.e., the intervals are non-overlapping):

To fit $[l^F, u^F]$ into $[l^G, u^G]$ we first need to shorten it by $D^F = (u^F - l^F) - (u^G - l^G)$. We perform this shortening at the end which is furthest from $[l^G, u^G]$:



Then we shift both ends of the resulting interval just onto the upper end u^G of the interval with larger nominal coverage:



The sum of the two necessary shifts (of the upper and lower end), which in this case simplifies to $S^F = 2l^F - u^G - l^G$ can be interpreted as the upward shift of $[l^F, u^F]$ with respect to $l^G, u^G]$.

Approximating the Cramer distance using interval divergences

Assuming K is even, the K equally spaced predictive quantiles of each distribution can seen as $L = K/2$ central prediction intervals with coverage levels $\alpha_i = 2i/(L+1), i = 1, \ldots, L$. Similarly to the definition of

the WIS, the approximation (31) can also be expressed in terms of these intervals as

$$\mathrm{CD}(F,G) \approx \frac{1}{2L(2L+1)} \sum_{k=1}^{L} \sum_{m=1}^{L} 2 \times \mathrm{ID}([l_k^F, u_k^F], [l_m^G, u_m^G], \alpha_k^F, \alpha_m^G).$$

This is easily seen as the involved double sum runs over the same discrepancy penalties as the one in equation (31), with four of them covered in each of the computed interval divergences.

If K is uneven, we get $L = (2K+1)/2$ central prediction intervals for each distribution, with coverage levels $\alpha_i = 2(i-1)/(L+1), i = 1, \ldots, L$. This means that the innermost prediction interval is just a single point at the predictive median. To avoid penalizing incompatibilities involving one of the medians more than once we then need to adjust the above to

$$\mathrm{CD}(F,G) \approx \frac{1}{(2L-0.5)(2L+0.5)} \sum_{k=1}^{L} \sum_{m=1}^{L} w_{km} \mathrm{ID}([l_k^F, u_k^F], [l_m^G, u_m^G], \alpha_k^F, \alpha_m^G). \tag{44}$$

with

$$w_{km} = \begin{cases} 1/4 & \text{if } k = 0 = m \\ 1/2 & \text{if } k = 0 \neq m \text{ or } k = 0 \neq m \\ 1 & \text{else} \end{cases}.$$

This representation implies a decomposition of the Cramer distance into the four interpretable components defined for the interval divergence in the previous section. Each component is just defined as the (appropriately weighted) average of the respective components at the different coverage levels. If G is a one-point distribution, the CD reduces to the WIS and the proposed decomposition reduces to the well-known decomposition of the WIS into dispersion, overprediction and underprediction components. The decomposition has the following further properties:

- Additive shifts of the two distributions affect the shift components, but not the dispersion components.
- Consequently, if G and G are identical up to an additive shift, both dispersion components will be 0.
- If F and G are both symmetric and have the same median, both shift components will be 0.
- It is possible that both shift components or both dispersion components are greater than 0, which leads to a somewhat strange interpretation. This corresponds to CDFs which cross more than once.