Prediction Interval Scoring Rules

Reference guide distilled from Greenberg (2018), "Calibration Scoring Rules for Practical Prediction Training"

1. Overview

This document summarizes the scoring system for **prediction interval** questions, where a user provides a confidence interval [L, U] at a fixed confidence level beta (e.g. beta = 0.8 for 80% confidence) and receives a score based on where the true answer x falls relative to that interval. The paper introduces two scoring rules: the **Distance** rule and the **Order of Magnitude** rule. Both share the same formula structure but differ in how they measure distance.

Source: Sections 7-9 of the original paper (pages 12-22).

2. Choosing a Scoring Rule

Distance rule — Use when the scale of the answer is roughly known to forecasters ahead of time. Examples: predicting a historical year, a percentage, an Oscar count. The score depends on raw numerical distance between x and the interval.

Order of Magnitude rule — Use when the answer could span many orders of magnitude and the main challenge is getting the rough scale right. Examples: "How many trees in the US?" or "How many babies born per day worldwide?" The score depends on the logarithmic ratio between x and the interval boundaries. This rule is also **unit-invariant** (changing inches to miles doesn't change the score).

Source: Sections 9.1.1 and 9.1.2 (pages 18-20). Guidance on choice: page 14-15.

3. Intermediate Variables: r, s, t

Both rules use three variables **r**, **s**, and **t**. They represent the same conceptual quantities (distance below, interval width, distance above) but are computed differently depending on the rule.

Variable	Meaning	Distance Rule	Order of Magnitude Rule
s	Width of interval	`(U - L) / c`	`log(U / L) / c`
r	How far x is below L (when x < L)	`(L - x) / c`	`log(L / x) / c`
t	How far x is above U (when x > U)	`(x - U) / c`	`log(x / U) / c`

The constant **c** is a scale parameter. The paper uses c = 100 for the Distance rule and c = ln(100) ~ 4.605 for the Order of Magnitude rule.

When x is **inside** the interval (L <= x <= U), r and t are redefined to measure position within the interval: for Distance, r = (x - L)/c and t = (U - x)/c; for Order of Magnitude, r = log(x/L)/c and t = log(U/x)/c. Note that s = r + t in both cases when x is inside the interval.

Source: Section 9.1.1 (page 19) and Section 9.1.2 (page 20).

4. The Core Scoring Formula

Both rules use the **same piecewise formula** (Section 9.2, page 20). Let beta be the confidence level (e.g. 0.8). Then:

Case 1: x is inside the interval (L <= x <= U)

```
Score = 4 * smax * (r * t / s^2) * (1 - s / (1 + s))
```

The term `4 * r * t / s^2` is a parabolic shape that equals 0 at the interval edges (x = L or x = U) and peaks at 1 when x is at the midpoint (arithmetic mean for Distance, geometric mean for Order of Magnitude). The term `(1 - s/(1+s))` = `1/(1+s)` penalizes wider intervals, approaching 0 as s grows to infinity. The maximum possible score is **smax**, achieved when the interval is infinitely narrow and perfectly centered on x.

Case 2: x is below the interval (x < L)

```
Score = (-2 / (1 - beta)) * r - (r / (1 + r)) * s
```

With beta = 0.8, the coefficient -2/(1-beta) = -10. The first term penalizes proportionally to how far x is below L. The second term adds a penalty that scales with interval width s, but is dampened by the factor r/(1+r) which approaches 1 for large misses.

Case 3: x is above the interval (x > U)

```
Score = (-2 / (1 - beta)) * t - (t / (1 + t)) * s
```

Symmetric to Case 2, using t instead of r.

Source: Unified formula in Section 9.2, page 20. Derivation/explanation on pages 20-21.

5. Final Adjustments (Section 9.3, page 22)

The raw formula above (called S') is modified in two ways to produce the final rules:

Adjustment 1: Interval expansion. Before scoring, the interval is slightly expanded by a factor delta (paper uses delta = 0.4). For the Distance rule, use [L - delta, U + delta] instead of [L, U]. For the Order of Magnitude rule, use [L * (1 - delta), U * (1 + delta)]. This ensures that if x lands right on the original boundary, the user still receives a small positive score rather than exactly zero.

Adjustment 2: Floor the score at smin. If the computed score falls below smin, clamp it to smin. The paper uses smin = -57.269 (derived as -(10 * log(99/50)) / log(50)). This prevents one catastrophic prediction from destroying a user's cumulative score.

Source: Section 9.3 (page 22) and Section 9.4 (pages 22-23).

6. Recommended Parameter Values

Parameter	Value	Meaning

smax	10	Max points per prediction (best case)
smin	-57.269	Min points per prediction (floor)
pmax	0.99	Max allowed probability (for choice predictions)
beta	0.8	Confidence level (80% interval)
delta	0.4	Interval expansion factor
c (Distance)	100	Scale: a miss of 100 units is 'moderately large'
c (OoM)	ln(100) ~ 4.605	Scale: a miss of 2 orders of magnitude is 'moderately large'

Source: Section 9.4 (pages 22-23).

7. Implementation Pseudocode

Below is pseudocode for the **Distance** rule. For the Order of Magnitude rule, replace the variable definitions as shown in Section 3 above.

```
function score_distance(x, L, U, beta=0.8, smax=10,
                        smin=-57.269, delta=0.4, c=100):
    # Step 1: Expand the interval
    L_exp = L - delta
    U_exp = U + delta

    # Step 2: Compute s (interval width)
    s = (U_exp - L_exp) / c

    # Step 3: Score based on where x falls
    if x < L_exp:
        r = (L_exp - x) / c
        raw = (-2/(1-beta)) * r - (r/(1+r)) * s
    elif x > U_exp:
        t = (x - U_exp) / c
        raw = (-2/(1-beta)) * t - (t/(1+t)) * s
    else:  # L_exp <= x <= U_exp
        r = (x - L_exp) / c
        t = (U_exp - x) / c
        raw = 4 * smax * (r * t) / (s * s) * (1 - s/(1+s))

    # Step 4: Clamp to floor
    return max(raw, smin)

function score_order_of_magnitude(x, L, U, beta=0.8, smax=10,
                        smin=-57.269, delta=0.4, c=4.605):
    # Step 1: Expand the interval
```

```
    L_exp = L * (1 - delta)
    U_exp = U * (1 + delta)

    # Step 2: Compute s
    s = log(U_exp / L_exp) / c

    # Step 3: Score based on where x falls
    if x < L_exp:
        r = log(L_exp / x) / c
        raw = (-2/(1-beta)) * r - (r/(1+r)) * s
    elif x > U_exp:
        t = log(x / U_exp) / c
        raw = (-2/(1-beta)) * t - (t/(1+t)) * s
    else:  # L_exp <= x <= U_exp
        r = log(x / L_exp) / c
        t = log(U_exp / x) / c
        raw = 4 * smax * (r * t) / (s * s) * (1 - s/(1+s))

    # Step 4: Clamp to floor
    return max(raw, smin)
```

8. Score Behavior Summary

Scenario	Score
x exactly at midpoint of a tiny interval	Approaches smax (best: +10)
x inside interval, near center	Positive, scales with precision
x inside interval, near edge	Small positive
x right at the boundary (x = L or x = U)	Small positive (due to delta expansion)
x just outside the interval	Small negative
x far outside the interval	Large negative (clamped at smin = -57.27)
Interval is infinitely wide	Approaches 0 (no information = no reward)

9. Key Design Properties (Why This Scoring Rule)

The paper identifies several properties that make these rules practical for training (Section 3, pages 2-3, and Section 9.2, pages 20-21):

Positive = correct, Negative = incorrect: The sign of the score immediately tells the user if they got it right.

Bounded scores: Scores range from smin (-57.27) to smax (+10). One bad prediction can't destroy everything.

Zero at boundary: Scoring ~0 when x lands on the interval edge provides a natural 'break-even' point.

Precision rewarded: Narrower intervals that still contain x earn more points.

Centering rewarded: x near the midpoint of the interval scores higher than x near the edge.

Infinite interval = zero: An infinitely wide interval (no information) earns zero points.

Continuous: Small changes in L, U, or x produce small changes in score. No jumps.

10. Important Caveat: Not Proper

The Distance and Order of Magnitude rules are **not proper scoring rules**. A proper scoring rule incentivizes honest reporting of beliefs (i.e., the user maximizes expected score by reporting their true confidence interval). The paper acknowledges this trade-off explicitly (Section 9.2, page 20): properness was sacrificed to gain the intuitive properties listed above. The standard proper scoring rules for prediction intervals (linear and log, Sections 7.1-7.2) lack bounded scores, the zero-at-boundary property, and within-interval sensitivity to centering.

11. Quick Reference: Where to Find Things in the Original Paper

Topic	Section	Pages
Desirable properties of scoring rules	3	2-3
Intuitive properties (7 criteria)	4	4-5
Formal definitions (S(p,e), S(x,L,U))	5	6
Proper scoring rules definition	5.2	6-7
Quadratic scoring rule	6.1	7-8
Brier scoring rule	6.2	8
Logarithmic scoring rule	6.3	9-11
General proper rule for intervals (theorem)	7	12
Linear interval scoring rule	7.1	13-14
Log interval scoring rule	7.2	14
Drawbacks of standard interval rules	7.3	15
Choice Predictions explained	8.1	16
Practical scoring rule transform	8.2	16-18
Distance scoring rule formula	9.1.1	18-19
Order of Magnitude scoring rule formula	9.1.2	19-20
Unified formula + explanation	9.2	20-21
Final adjustments (delta, smin)	9.3	22
Parameter choices	9.4	22-23