Gumbel-Box Volume

Motivation

The puzzle: Hard boxes have crisp, deterministic boundaries. Their volume is simply length × width × height (in higher dimensions, the product of interval lengths). But Gumbel boxes have **fuzzy** boundaries—they wiggle randomly. How do we compute the "average" volume when the boundaries themselves are random?

This is like asking: "If I draw a box on a piece of paper, but my hand shakes while drawing, what's the average size of the box I'll produce?" The answer isn't obvious, because we need to average over all possible "shaky" realizations.

The key insight—and here's the beautiful part—is that this expectation reduces to a special function that mathematicians have studied for centuries: the modified Bessel function of the second kind, order zero, denoted K_0. This connection emerges naturally from the structure of Gumbel distributions and provides both theoretical elegance and computational tractability. It's as if nature itself designed Gumbel distributions to produce this elegant result.

Historical context: The evolution from hard boxes to Gumbel boxes (Dasgupta et al., 2020) was motivated by the local identifiability problem: hard boxes produce zero gradients when boxes are disjoint or contained, preventing effective learning. Gumbel boxes solve this by introducing probabilistic boundaries, but this requires computing expected volumes over the joint distribution of random boundaries. The Bessel function formula provides an analytical solution, avoiding expensive numerical integration or Monte Carlo methods. This computational tractability is essential for practical applications.

Why Gumbel distributions? Gumbel distributions appear naturally in extreme value theory as the limiting distribution of maxima (or minima) of independent, identically distributed random variables. The Fisher-Tippett-Gnedenko theorem (1928) establishes that there are only three possible limiting distributions for normalized maxima: Gumbel (Type I), Fréchet (Type II), and Weibull (Type III). The Gumbel distribution is the only one of these that is max-stable in the sense that the maximum of independent Gumbel random variables remains Gumbel-distributed (see the Gumbel Max-Stability document). This makes Gumbel distributions the "natural" choice for modeling extreme events—in our case, the extreme coordinates (minimum and maximum) that define box boundaries. The max-stability property ensures that operations on Gumbel boxes remain within the Gumbel family, preserving mathematical structure throughout computations.

Definition

A **Gumbel box** models each coordinate interval $[X, Y]$ as:

- $X \sim \text{MinGumbel}(\mu_x, \beta)$ (minimum coordinate)
- $Y \sim \text{MaxGumbel}\big(\mu_y, \beta\big)$ (maximum coordinate)

where $\beta > 0$ is the scale parameter (constant across dimensions) and μ_x, μ_y are learnable location parameters.

The expected interval length $E[\max(Y - X, 0)]$ represents the average "size" of the box along this dimension, accounting for the probabilistic nature of the boundaries.

Note on dimensionality: For a d-dimensional Gumbel box, the expected volume is the product of expected interval lengths across all dimensions: $E[\text{Vol}(B)] = \prod_{i=1}^{d} E[\max(Y_i - X_i, 0)]$. This follows from the independence of coordinates across dimensions. The theorem below gives the formula for a single dimension.

Statement

Theorem (Gumbel-Box Volume). For a Gumbel box with $X \sim \text{MinGumbel}(\mu_x, \beta)$ and $Y \sim \text{MaxGumbel}(\mu_y, \beta)$, the expected interval length in one dimension is:

$$E[\max(Y - X, 0)] = 2\beta K_0\left(2e^{-\frac{\mu_y - \mu_x}{2\beta}}\right)$$

where K_0 is the modified Bessel function of the second kind, order zero.

For a d-dimensional Gumbel box, the expected volume is the product of expected interval lengths across dimensions (by independence of coordinates):

$$E[\text{Vol}(B)] = \prod_{i=1}^{d} E[\max(Y_i - X_i, 0)] = \prod_{i=1}^{d} 2\beta K_0\left(2e^{-\frac{\mu_{y,i} - \mu_{x,i}}{2\beta}}\right)$$

Proof

We compute the expectation by integrating over the joint distribution of X and Y:

$$E[\max(Y - X, 0)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max(y - x, 0) f_{X(x)} f_{Y(y)} \, \mathrm{d}x \, \mathrm{d}y$$

The Gumbel probability density functions are:

- $f_{X(x)} = \frac{1}{\beta} e^{\frac{x-\mu_x}{\beta} - e^{\frac{x-\mu_x}{\beta}}}$ (MinGumbel)
- $f_{Y(y)} = \frac{1}{\beta} e^{-\frac{y-\mu_y}{\beta} - e^{-\frac{y-\mu_y}{\beta}}}$ (MaxGumbel)

Step 1: Standardize the variables. We substitute $u = \frac{x-\mu_x}{\beta}$ and $v = \frac{y-\mu_y}{\beta}$, so $\mathrm{d}x = \beta \, \mathrm{d}u$ and $\mathrm{d}y = \beta \, \mathrm{d}v$. The region where $y > x$ (so $\max(y - x, 0) = y - x$) becomes $v > u - \delta$ where $\delta = \frac{\mu_y - \mu_x}{\beta}$ measures the separation between location parameters.

Step 2: Simplify the integrand. After changing variables in Step 1, the integrand has the double exponential structure $e^{u - e^u} e^{-v - e^{-v}}$. We change the order of integration to integrate over u first for fixed v.

For the region $v > u - \delta$, we make the substitution $w = u - v - \delta$ (so $u = w + v + \delta$). This transformation simplifies the integration domain. The Jacobian is 1, so $\mathrm{d}u = \mathrm{d}w$.

After this substitution, the integrand becomes:

$$e^{(w+v+\delta) - e^{w+v+\delta}} e^{-v - e^{-v}} = e^{w + \delta - e^{w+v+\delta} - e^{-v}}$$

Simplifying and recognizing the structure, we make a further substitution $s = e^w$, which transforms the integration domain. After algebraic manipulation, the double integral reduces to a single integral over a new variable t (related to s via $t = \text{arcsinh}\left(\frac{s}{2}\right)$ or similar transformation):

$$\int_0^{\infty} e^{-2e^{-\frac{\delta}{2}} \cosh t} \, \mathrm{d}t$$

This form reveals the underlying structure: the argument $2e^{-\frac{\delta}{2}}$ controls how the exponential decay interacts with the hyperbolic cosine. The appearance of $\cosh t$ is characteristic of integrals arising from Gumbel distributions and signals the connection to Bessel functions. The transformation to hyperbolic coordinates ($\cosh t$) reveals the underlying symmetry that makes the Bessel function appear naturally.

Step 3: Recognize the Bessel function. The integral representation of the modified Bessel function of the second kind, order zero, is:

$$K_0(z) = \int_0^\infty e^{-z\cosh t}\,\mathrm{d}t$$

Setting $z = 2e^{-\frac{\delta}{2}} = 2e^{-\frac{\mu_y-\mu_x}{2\beta}}$ and accounting for the 2β factor from the change of variables, we obtain the stated result.

Why Bessel functions appear: The appearance of K_0 is not coincidental—it emerges from the fundamental structure of Gumbel distributions and their relationship to extreme value theory. Bessel functions commonly arise in probability and statistics when dealing with products or ratios of random variables, sums of exponential random variables, and problems involving circular or cylindrical symmetry. In our case, the connection is deeper: Gumbel distributions are intimately related to exponential distributions through the transformation $X = -\ln(-\ln U)$ where U is uniform, and Bessel functions naturally appear when computing expectations involving exponential random variables.

The double exponential structure $e^{u-e^u}e^{-v-e^{-v}}$ in the Gumbel PDFs, when integrated over the region where $y > x$, produces a convolution-like integral. The transformation to hyperbolic coordinates ($\cosh t$) reveals the underlying symmetry, and the resulting integral matches the standard representation of the modified Bessel function $K_0(z) = \int_0^\infty e^{-z\cosh t}\,\mathrm{d}t$. This connection is fundamental: the modified Bessel function K_0 appears in the probability density function of the product of two normally distributed random variables, and more generally, Bessel functions are solutions to differential equations that arise naturally in problems involving exponential families and extreme value distributions. The appearance of $\cosh t$ in the integral representation is characteristic of problems involving exponential decay with hyperbolic geometry, which is why Bessel functions provide the natural analytical solution.

Numerical Approximation

Direct computation of K_0 can be numerically unstable for small arguments. For $z \to 0$, the asymptotic behavior is $K_0(z) \sim -\ln(\frac{z}{2}) - \gamma$, where $\gamma \approx 0.5772$ is Euler's constant. This logarithmic singularity reflects the behavior of the Bessel function near the origin.

To avoid numerical issues while maintaining smooth gradients, we use the stable approximation:

$$2\beta K_0\left(2e^{-\frac{x}{2\beta}}\right) \approx \beta \log\left(1 + \exp\left(\frac{x}{\beta} - 2\gamma\right)\right)$$

where $x = \mu_y - \mu_x$.

Why the softplus form works: For small arguments $z \to 0$, we have $K_0(z) \sim -\ln(\frac{z}{2}) - \gamma$. Substituting $z = 2e^{-\frac{x}{2\beta}}$ gives $K_0\left(2e^{-\frac{x}{2\beta}}\right) \sim -\ln\left(e^{-\frac{x}{2\beta}}\right) - \gamma = \frac{x}{2\beta} - \gamma$. Multiplying by 2β yields $x - 2\beta\gamma$. The softplus form $\beta\log\left(1 + \exp\left(\frac{x}{\beta} - 2\gamma\right)\right)$ approximates $\max(x - 2\beta\gamma, 0) +$ small correction, which matches this asymptotic behavior. For large x, the exponential dominates and we recover the linear behavior; for negative x, the correction term ensures smoothness.

The softplus form provides:

- **Smooth gradients**: Unlike the hard maximum, this approximation is differentiable everywhere
- **Numerical stability**: The form $\max(x, 0) + \log(1 + \exp(-|x|))$ avoids overflow. This is analogous to the log-sum-exp trick used in machine learning: by working in log-space and shifting by the maximum, we prevent numerical overflow when exponentiating large values. The softplus

function is the one-dimensional case of log-sum-exp, providing the same numerical stability guarantees.
- **Correct asymptotics**: It matches the Bessel function behavior in both small and large argument regimes

Error analysis: The relative error of this approximation is approximately $O(z^2)$ for small $z = 2e^{-\frac{x}{2\beta}}$. When $z < 0.1$ (i.e., when $\frac{x}{2\beta} > \ln(20) \approx 3$), the relative error is less than 1%. For $z < 0.01$, the relative error is less than 0.1%.

When the approximation breaks down: For very large β (relative to x), specifically when $\beta > \frac{x}{3}$, the approximation becomes less accurate. In practice, when $\beta > \frac{x}{10}$, direct computation of K_0 may be preferable, though the softplus form remains stable.

Numerical stability edge cases:

1. **Very small volumes**: When $\mu_y - \mu_x$ is very negative (boxes with expected negative length), the Bessel function argument $z = 2e^{-\frac{x}{2\beta}}$ becomes very large, and $K_0(z)$ decays exponentially. The softplus approximation $\beta \log\left(1 + \exp\left(\frac{x}{\beta} - 2\gamma\right)\right)$ correctly captures this exponential decay, but care must be taken to avoid underflow when $\frac{x}{\beta}$ is very negative. In practice, when $\frac{x}{\beta} < -20$, the expected volume is effectively zero (below machine precision), and the computation can be short-circuited.

2. **Very large volumes**: When $\mu_y - \mu_x$ is very large (boxes spanning most of the space), the Bessel function argument becomes very small, and $K_0(z) \sim -\ln(\frac{z}{2}) - \gamma$ dominates. The softplus approximation remains stable, but for extremely large $\frac{x}{\beta} > 50$, the linear term $x - 2\beta\gamma$ dominates, and the expected volume approaches x (the separation between expected boundaries). This is the correct asymptotic behavior: as the separation grows, the expected interval length approaches the separation itself.

3. **High-dimensional underflow**: For d-dimensional boxes, the volume is the product of d expected interval lengths. In log-space, this becomes a sum: $\log E[\text{Vol}] = \sum_{i=1}^{d} \log E[\max(Y_i - X_i, 0)]$. When any dimension has very small expected length (approaching machine epsilon), the log-volume becomes very negative. Care must be taken to handle dimensions where $E[\max(Y_i - X_i, 0)] < \varepsilon$ (machine epsilon), as these contribute $\log \varepsilon$ to the sum, potentially causing numerical issues. The library implementation clamps very small expected lengths to a minimum threshold (typically 10^-10) to prevent underflow while maintaining gradient flow.

4. **Boundary cases in intersection**: When computing intersection volumes, the formula $E[\text{Vol}(A \cap B)] = \prod_i E\left[\max\left(\min\left(Z_i^A, Z_i^B\right) - \max\left(z_i^A, z_i^B\right), 0\right)\right]$ can produce zero expected volume when boxes are far apart. The exponential decay bound $E[\text{Vol}(A \cap B)] \geq Ce^{-\frac{d}{\beta}}$ ensures the volume is always positive, but for very large separation d, the volume may be below machine precision. In practice, when $\frac{d}{\beta} > 20$, the intersection volume is effectively zero, and the computation can be short-circuited to avoid numerical issues.

5. **Temperature extremes**: When $\beta \to 0$ (hard boxes), the Bessel function argument $z = 2e^{-\frac{x}{2\beta}} \to 0$ for finite x, and $K_0(z) \sim -\ln(\frac{z}{2}) - \gamma$ diverges. However, the product $2\beta K_0(z)$ remains finite and approaches $\max(x, 0)$ as $\beta \to 0$, recovering the hard box volume. When $\beta \to \infty$ (very soft boxes), the Bessel function argument $z \to 2$, and $K_0(2) \approx 0.1139$ is finite. The expected volume approaches $2\beta * 0.1139$, which grows linearly with β. This behavior is correct: very soft boxes have large expected volumes due to high boundary variance.

Example

A worked example with numbers: Let's compute the expected volume of a Gumbel box with concrete numbers to see the machinery in action.

Consider a Gumbel box with $\mu_x = 0.0$, $\mu_y = 1.0$, and $\beta = 0.1$. Think of this as: the expected minimum is at 0, the expected maximum is at 1, and the "fuzziness" (scale) is 0.1—so the boundaries are relatively tight around their expected positions, but still random.

Step 1: Compute the Bessel function argument.
- $z = 2e^{-\frac{1.0-0.0}{2*0.1}} = 2e^{-5} \approx 0.0135$
- This is a very small number! The exponential decay e^{-5} makes it tiny.

Step 2: Evaluate the Bessel function.
- For small arguments, we use the asymptotic expansion: $K_0(z) \sim -\ln\left(\frac{z}{2}\right) - \gamma$
- $K_0(0.0135) \approx -\ln\left(\frac{0.0135}{2}\right) - 0.5772 \approx 4.27$
- Notice how the logarithm "undoes" the exponential, giving us a reasonable number.

Step 3: Compute the expected volume.
- $E[\max(Y - X, 0)] = 2 * 0.1 * 4.27 \approx 0.854$
- So the expected interval length is about 0.854, which is close to the separation of 1.0, but slightly less due to the probabilistic boundaries.

Step 4: Verify with the softplus approximation.
- $\beta \log\left(1 + \exp\left(\frac{1.0}{0.1} - 2 * 0.5772\right)\right) = 0.1 * \log(1 + \exp(10 - 1.1544)) \approx 0.854$
- The close agreement (both give 0.854) demonstrates that the approximation captures the essential behavior while remaining numerically stable. The softplus form avoids the logarithmic singularity that would cause numerical issues.

Notes

Why Bessel functions? The appearance of K_0 is not coincidental. Bessel functions arise naturally when computing expectations involving exponential random variables, and Gumbel distributions are intimately connected to exponentials through the transformation $X = -\ln(-\ln U)$ where U is uniform. This connection runs deep: the modified Bessel function K_0 appears in the probability density of products of normal random variables, suggesting a fundamental relationship between extreme value theory and geometric probability.

Mathematical structure: The double exponential structure $e^{u-e^u}e^{-v-e^{-v}}$ in the Gumbel PDFs, when integrated over the region $y > x$, produces an integral of the form $\int_0^\infty e^{-z\cosh t}\,\mathrm{d}t$ where $z = 2e^{-\frac{\delta}{2}}$ and $\delta = \frac{\mu_y - \mu_x}{\beta}$. This integral representation is the defining property of the modified Bessel function $K_0(z)$. The appearance of $\cosh t$ (hyperbolic cosine) reflects the underlying hyperbolic geometry: the transformation to hyperbolic coordinates reveals the symmetry that makes the Bessel function appear naturally. This is not just a computational convenience—the Bessel function provides the **exact** analytical solution, avoiding the need for numerical integration or Monte Carlo methods.

Computational considerations: For high-dimensional boxes, volume computation in log-space is essential to avoid numerical overflow. The library implementation computes $\log E[\mathrm{Vol}] = \sum_i \log E[\max(Y_i - X_i, 0)]$, then exponentiates only when necessary. This log-space computation is numerically stable and is the default in production implementations.

Complexity analysis: Computing the expected volume of a single d-dimensional Gumbel box requires d evaluations of the Bessel function (or its softplus approximation), giving time complexity $O(d)$. For N boxes, the total cost is $O(N * d)$. The Bessel function evaluation itself is typically $O(1)$ using standard library implementations (e.g., `scipy.special.k0` in Python, or optimized approximations). The softplus approximation $\beta \log\left(1 + \exp\left(\frac{x}{\beta} - 2\gamma\right)\right)$ is also $O(1)$ per dimension, making volume computation highly efficient. This linear complexity in dimension makes box embeddings scalable to high-dimensional spaces, unlike methods requiring numerical integration which would be $O\left(d^k\right)$ for k-th order integration methods.

Beyond Gumbel boxes: The Bessel function formula applies specifically to Gumbel-distributed boundaries. Other probabilistic box models (Gaussian-smoothed boxes, uniform boxes) require different volume calculations, typically involving numerical integration or Monte Carlo methods. The analytical tractability of Gumbel boxes is a key advantage that enables efficient training and inference.