STAT 33612 Autumn 2025 Scribe: Beining Wu
October 6, 2025 Lecturer: Frederic Koehler
These notes have not received the scrutiny of publication. They could be missing important references, etc.

Gumbel Trick, Random Energy Model (cont'd)

1 Gumbel Distribution

Recap. Suppose that $g_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, last time we concluded that

$$\mathbb{E}[\max_{i\le N} g_i]/\sqrt{2\log N} \to 1 \text{ as } N \to \infty. \tag{1}$$

The calculation consists of an upper bound based on the Chernoff bound and a lower bound based the concentration.

Convergence of maximal. For a class of light-tailed distributions including the Gaussian, the maximal of n iid samples, after proper centering and scaling, converges to the same limit distribution, which is the Gumbel distribution.

Here, we consider the extreme values of exponential distribution random variables as an illustration. Suppose that $Z_i \overset{\text{i.i.d.}}{\sim} \mathsf{EXP}(1)$, i.e., $\mathbb{P}(Z_i \ge t) = e^{-t}$ for $t \ge 0$. The heuristic for this is that the probability for each Z_i to be near $\log N$ is about $1/N$, and $(1-1/N)^N \approx e^{-1}$. We conjecture that $\max_i Z_i$ concentrates near $\log N$, similar to (1). For $t > 0$, we can calculate the CDF of $\max_i Z_i$ at $\log N + t$ as

$$\mathbb{P}(\max Z_i \le \log N + t) \overset{\text{ind}}{=} \left(1 - \frac{e^{-t}}{N}\right)^N \approx e^{-e^{-t}}. \tag{2}$$

Gumbel distribution. The gumbel distribution $\mathsf{Gumbel}(\mu, 1)$ is defined as the distribution with CDF/PDF

$$\mathbb{P}(X \le x) = e^{-e^{-(x-\mu)}}; \quad p(x) = e^{-(x-\mu)} e^{-e^{-(x-\mu)}}.$$

This distribution is heavily asymmetric. For $x > \mu$, $p(x)$ looks like $e^{-(x-\mu)}$ and decays exponentially. However for $x < \mu$, $p(x) \sim e^{-e^{-(x-\mu)}}$ decays double-exponentially. The heuristic is that for the maximal of exponential RV to be smaller than $\log N$ requires all of them to be smaller than $\log N$, which happens with probability $\sim e^{-e^{-t}}$.

Maximal of Gumbels. Since Gumbel distribution is the limiting distribution for the maximal of light-tailed distributions, it is natural to conjecture that maximals of Gumbel RVs are still Gumbel. Supppose that $X \sim \mathsf{Gumbel}(\alpha, 1)$ and $Y \sim \mathsf{Gumbel}(\beta, 1)$ are independent. We wonder how is $W = \max(X, Y)$ distributed.

Sketch for large α, β. We suppose that $\alpha = \log N$ and $\beta = \log M$ for $N, M \in \mathbb{Z}_{>0}$. Since α, β are both location parameters, we can write W as

$$\begin{aligned} W = \max\{X, Y\} &= \max\{ \underbrace{X - \alpha}_{\mathsf{Gumbel}(0,1)} + \alpha, \underbrace{Y - \beta}_{\mathsf{Gumbel}(0,1)} + \beta\} \\ &\overset{d.}{\approx} \max_i \{\max_{i\le N} Z_i, \max_{N<i\le M+N} Z_i\} \\ &= \max_{i \le N+M} Z_i. \end{aligned}$$

Here $Z_i \overset{\text{i.i.d.}}{\sim} \mathsf{Gumbel}(0,1)$ and the second line holds from (2). From this decomposition, we see that W follows $\mathsf{Gumbel}(\log(N+M),1)$ for large N, M, approximately. Roughly, we conclude that $W \sim \mathsf{Gumbel}(\log(e^\alpha + e^\beta), 1)$. This argument can only be used when M, N are large log-integers.

Exercise 1. Prove that $\max(X, Y) \sim \mathsf{Gumbel}(\log(e^\alpha + e^\beta), 1)$ *exactly* for all $\alpha, \beta \in \mathbb{R}$.

Given that the statement above is true, we can prove that $\max_{i \le N} X_i \sim \mathsf{Gumbel}(\log(\sum_i e^{\alpha_i}), 1)$, where $X_i \sim \mathsf{Gumbel}(\alpha_i, 1)$ are independent. We can also write it in a different way. Suppose that $\xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{Gumbel}(0,1)$. It holds that that

$$\begin{aligned}\mathbb{E}[\max_i\{\alpha_i + \xi_i\}] &= \mathbb{E}[\max_i X_i] \\ &= \log(\sum_i e^{\alpha_i}) + \gamma,\end{aligned}$$

where $\gamma = \mathbb{E}[\xi_1] = \lim_N(-\log N + \sum_{k \le N} k^{-1})$ is the Euler-Mascheroni constant. We can also note that the index for which the maximum is attained follows the softmax distribution with parameter α_i.

This identity is called the Gumbel trick.

Gumbel trick in REM. We illustrate the use of the Gumbel trick in the random energy model. For $g_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, we pair it with independent $\xi_i \sim \mathsf{Gumbel}(0,1)$. Then we have that

$$\begin{aligned}\mathbb{E}[\log \sum_i e^{\beta g_i}] + \gamma &= \mathbb{E}\big[\mathbb{E}\big[\max_i[\beta g_i + \xi_i] \mid g_1, \dots g_N\big]\big] \\ &= \mathbb{E}[\max_i[\beta g_i + \xi_i]] \\ &\ge \mathbb{E}[\max_i \beta g_i + \mathbb{E}[\xi_i]] = \beta\,\mathbb{E}[\max_i g_i] + \gamma.\end{aligned}$$

The last inequality comes from interchanging expectation and maximum.

2 Backgrounds in Statistical Physics

In statistical physics[1], we study the phase transition using random energy models. A typical example is the phase transition of water, which exhibit different phases (solid, liquid, gas) at different temperatures.

Suppose that we have a system of particles in a thermal reservoir at fixed temperature T, and the inverse temperature is $\beta = 1/T$. [2] The system is encoded with a state in the state space $x \in \mathcal{X}$, which for example can be the spins $\{+1, -1\}^N$, or 3D locations $\mathbb{R}^{3N}$, where N is the number of particles. For each state x, we associate the system with an energy $H(x)$, which is a function from $\mathcal{X}$ to $\mathbb{R}$. Given these componentns, the Boltzmann distribution over $\mathcal{X}$ is used to describe the chance of the system being in state x at equilibrium:

$$p(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)}, \quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta H(x)}. \tag{3}$$

For example in the water model, the energy is described by two parts, we can roughly specify the energy function as

$$H(x) = [\text{Polar bonds}] + [\text{Gravitational potential energy}].$$

The polar bonds describe the interaction between particles that are caused by molecular dipoles. This model is useful to describe the phase transition of water, as

[1] See Mezard and Montanari's textbook for more details.

[2] In physics literature, $\beta = 1/kT$ where k is the Boltzmann constant is more used.

- In the high temperature limit $\beta \to 0$, $p_\beta(x) \approx \mathrm{Unif}(\mathcal{X})$. And this says that the particle system appears to be non-interacting and non-structured.
- In the moderate temperature regime, the interaction between particles matters, and $p_\beta(x) \approx e^{-\beta H(x)}$ is not uniform.
- In the low temperature limit $\beta \to \infty$, $p_\beta(x)$ is concentrated on the minimizers of $H(x)$, i.e., $p_\beta(x) \sim \mathsf{Unif}(\arg\min_x H(x))$, where $\arg\min_x H(x) = \{x : H(x) = \min_y H(y)\}$. This means that the system is frozen in the ground states (ice).

We can formalize the low temperature limit as follows. For each $x \in \mathcal{X} \setminus \arg\min_x H(x)$, we have

$$\begin{aligned} p_\beta(x) &= \frac{e^{-\beta H(x)}}{\sum_{x\in\mathcal{X}} e^{-\beta H(x)}} \\ &\le \frac{e^{-\beta \min_x H(x)}}{\max_{x\in\mathcal{X}} e^{-\beta H(x)}} \\ &= \exp(-\beta(H(x) - \min_x H(x))) \to 0, \quad \text{as } \beta \to \infty. \end{aligned}$$

For $x \in \arg\min_x H(x)$, we have that

$$\begin{aligned} p_\beta(x) &= \frac{e^{-\beta \min_x H(x)}}{\sum_{x\in\mathcal{X}} e^{-\beta H(x)}} \\ &= \big(|\mathcal{X}| + \sum_{x\in\mathcal{X}\setminus\arg\min_x H(x)} e^{-\beta H(x) + \beta \min_x H(x)}\big)^{-1} \to \frac{1}{|\arg\min_x H(x)|}, \quad \text{as } \beta \to \infty. \end{aligned}$$

The model in (3) is called the canonical ensemble in statistical physics.

Free energy, REM. Some times we see formulation like $p_\beta(x) = \exp\{-\beta(F - E(x))\}$, where $F = \beta^{-1} \log Z_\beta$ is called the free energy, and $E(x) = H(x)$ is called the internal energy.

Sometimes we want to study the canonical ensemble with random energy function $H(x)$. For example, in the random energy model (REM), we suppose that $H(x)$ are independent Gaussian random variables for different x, i.e., $H(x) \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, N/2)$ for $x \in \mathcal{X} = \{+1, -1\}^N$.

In this case the p_β is a random measure on the state space. And we are interested in $\mathbb{E}[\log Z_\beta] = \mathbb{E}[\beta F]$ and the expectation is taken with respect to the randomness of $H(x)$.

Trailer. We will compute $\mathbb{E}[\log Z_\beta]$ in REM.

- Fact I. It holds that $\log Z_\beta \approx \mathbb{E}[\log Z_\beta]$. In various models this is easy to justify.
- Fact II. $Z_\beta \overset{?}{\approx} \mathbb{E}[Z_\beta]$. This might not be always correct. The heuristic is that Z_β is a exponential of other things. And the chance for Z_β to be large is considerable due to the heavy-tailedness.

In summary, $\log Z_\beta$ is tangible and we will compute the expectation in the next lecture.