Occam's Razor, Forced

The geometry behind the maximum-entropy, Boltzmann, and softmax form

Japjot Singh

This is the worked derivation behind the essay *A Very Good Loaded Die*. It rebuilds, from first principles, why "assume the least" has exactly one computable answer, the exponential (Boltzmann / softmax) form, and why that same shape runs a gas, a classifier, and an aligned language model. Every step is forced by the one before it; nothing is taken on faith.

1 The problem

A six-sided die has face values $E_i \in \{1, \dots, 6\}$. You are told exactly one thing: its mean is μ (say 4.5). Among all distributions consistent with that, pick the one that assumes the least, i.e. the one with the most entropy:

$$\max_p \; H(p) = -\sum_i p_i \ln p_i \qquad \text{s.t.} \qquad \sum_i p_i E_i = \mu, \quad \sum_i p_i = 1, \quad p_i > 0.$$

Write the two constraint gradients as

$$u_1 = \nabla g_{\text{mean}} = (E_1, \dots, E_n), \qquad u_2 = \nabla g_{\text{norm}} = (1, \dots, 1),$$

and set $N = \operatorname{span}(u_1, u_2)$ (the *normal space*) and $T = N^\perp$ (the *feasible directions*: the steps d that keep both constraints to first order, $\langle d,\, u_1\rangle = \langle d,\, u_2\rangle = 0$). Since $\mathbb{R}^n = N \oplus T$, every vector splits uniquely into an N-part and a T-part.

2 The geometry: the one fact everything rests on

Theorem 1. *At the constrained maximum $p^\star$, the gradient $\nabla H(p^\star)$ lies entirely in N. It has no component along any feasible direction: $\nabla H(p^\star) = \lambda u_1 + \nu u_2$ for unique scalars λ, ν.*

Proof. Split $\nabla H(p^\star) = s + t$ with $s \in N$, $t \in T$; we show $t = 0$.

(1) Competitors stay feasible. Fix $d \in T$. For $k = 1, 2$, $\langle p^\star + \varepsilon d,\, u_k\rangle = \langle p^\star,\, u_k\rangle + \varepsilon\langle d,\, u_k\rangle = \langle p^\star,\, u_k\rangle$ because $d \perp N \ni u_k$, so both constraints still read μ and 1. Positivity holds for small steps: with $\delta := \min_i p_i^\star/(1 + \max_i |d_i|)$, every $p^\star + \varepsilon d$ with $|\varepsilon| < \delta$ stays strictly positive, hence feasible.

(2) Restrict and apply Fermat. Let $\varphi(\varepsilon) := H(p^\star + \varepsilon d)$ on $(-\delta, \delta)$, the altitude along the straight feasible path through $p^\star$. Since $p^\star$ maximizes H over the feasible set, $\varphi(\varepsilon) \le \varphi(0)$ throughout, so $\varphi'(0) = 0$ (an interior one-variable maximum has zero derivative).

(3) Compute the rate. With $a_i(\varepsilon) = p_i^\star + \varepsilon d_i$, the product and chain rules give $\frac{d}{d\varepsilon}\big[-a_i \ln a_i\big] = -d_i \ln a_i - d_i$, so

$$\varphi'(0) = \sum_i d_i \,(-\ln p_i^\star - 1) = \langle \nabla H(p^\star),\, d\rangle.$$

Combined with (2): $\langle \nabla H(p^\star),\, d\rangle = 0$ for every $d \in T$.

(4) Force $t = 0$. Take $d = t \in T$:

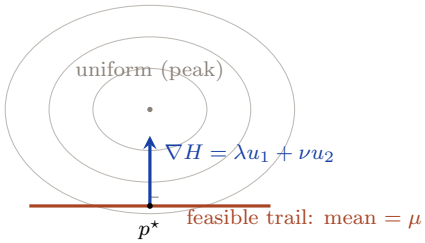
$$0 = \langle \nabla H(p^\star),\, t\rangle = \langle s + t,\, t\rangle = \underbrace{\langle s,\, t\rangle}_{0\ (N \perp T)} + \|t\|^2 = \|t\|^2 \implies t = 0.$$

So $\nabla H(p^\star) = s \in N$. Since N is two-dimensional with basis u_1, u_2 and coordinates in a basis are unique, $\nabla H(p^\star) = \lambda u_1 + \nu u_2$ with λ, ν forced, not chosen. □

Remark 1 (The double duty is Lagrange)**.** The same two vectors u_1, u_2 do two jobs at once. As *normals* they fence you in, defining the feasible slice you cannot leave. As a *basis* they build the trapped gradient $\lambda u_1 + \nu u_2$, the uphill that now points entirely out of that slice, which is exactly why $p^\star$ is stuck. That double role is the whole content of Lagrange multipliers.

3 Reading the constraints geometrically

Each u_k is a fixed measuring ruler; its constraint is the perpendicular slice through the ruler at one reading (μ on the energy ruler u_1, 1 on the mass ruler u_2). The feasible set is the intersection of the two slices: an unbounded affine flat $p_0 + T$, where p_0 is any one feasible die used as an anchor.



At the top of the feasible trail the gradient points straight off it, along the constraint normal.

Remark 2 (Dimension ledger)**.** Each independent equality constraint kills one dimension. A die with the mean and normalization constraints has $6 - 2 = 4$ free dimensions, so there is real room to choose and max-entropy has work to do. A coin with the same two constraints has $2 - 2 = 0$: a single distribution, nothing left to choose. Max-entropy only does something when $n >$ #constraints.

Remark 3 (Inner products are coordinates only against an orthonormal set)**.** Here $\langle u_1,\, u_2\rangle = \sum_i E_i = 21 \neq 0$, so the readings $\langle p,\, u_k\rangle$ are *not* the u-basis coordinates of p. In general $\langle p,\, u\rangle = \|u\| \times$ (length of p's shadow on u); "the mean is p's coordinate along the energy axis" is that statement for a *unit* ruler only. The constraints pin the whole N-part of p (its shadow), and leave the T-part free, which is the freedom entropy spends.

4 The Lagrangian, reverse-engineered

The optimality condition is $\nabla H(p^\star) - \lambda\, \nabla g_{\text{mean}} - \nu\, \nabla g_{\text{norm}} = 0$. We want a single function whose p-gradient *is* that expression, so we read it straight off:

$$L(p, \lambda, \nu) = H(p) - \lambda\big(g_{\text{mean}}(p) - \mu\big) - \nu\big(g_{\text{norm}}(p) - 1\big).$$

Setting every partial derivative to zero reproduces the entire problem: $\partial L/\partial p = 0$ is the geometric optimality condition; $\partial L/\partial \lambda = 0$ hands back $g_{\text{mean}} = \mu$; $\partial L/\partial \nu = 0$ hands back $g_{\text{norm}} = 1$. The subtracted constants $\mu, 1$ are exactly what make the multiplier-derivatives return the *right* level sets, and they make each penalty vanish on the feasible set, so $L(p^\star) = H(p^\star)$. A constrained maximization has become plain unconstrained stationarity of one function.

5 Turning the crank

Differentiate L in a single coordinate and set it to zero:

$$\frac{\partial L}{\partial p_i} = -\ln p_i - 1 - \lambda E_i - \nu = 0 \implies p_i = e^{-\lambda E_i}\, e^{-(1+\nu)}.$$

The factor $e^{-(1+\nu)}$ is the same for every face, so impose $\sum_i p_i = 1$ to *spend* ν:

$$e^{-(1+\nu)} \sum_i e^{-\lambda E_i} = 1 \implies e^{-(1+\nu)} = \frac{1}{Z(\lambda)}, \qquad Z(\lambda) := \sum_j e^{-\lambda E_j}.$$

The normalization multiplier is gone, folded into the partition function Z. What remains is the exact shape:

$$\boxed{p_i = \frac{e^{-\lambda E_i}}{Z(\lambda)}, \qquad Z(\lambda) = \sum_j e^{-\lambda E_j}.}$$

Remark 4 (Spend ν means solve it away)**.** "Spend ν" does not mean increase it; it means use the one equation not yet touched, $\sum_i p_i = 1$, to solve for ν and eliminate it ($\nu = \ln Z(\lambda) - 1$). Only λ survives into the answer, pinned by the mean through $\sum_i p_i E_i = \mu$.

Remark 5 (No positivity constraint was needed)**.** The boxed form is strictly positive automatically, an exponential is never zero, so $p_i > 0$ never binds. More deeply, $\partial H/\partial p_i = -\ln p_i - 1 \to +\infty$ as $p_i \to 0^+$: entropy repels the boundary and forces the maximum strictly interior, which is also the hypothesis $p_i^\star > 0$ the proof used.

6 The dial

Because λ is the slope of the best-achievable entropy against the demanded mean, it is a price:

$$\lambda = \frac{dH^\star}{d\mu},$$

the entropy you give up per unit you drag the mean (the envelope theorem). The curve $H^\star(\mu)$ is a concave hump peaked at the free mean 3.5, where $\lambda = 0$ (the fair die). Dragging the mean up to 4.5 forces $\lambda < 0$ (more weight on the high faces); as μ sweeps $1 \to 6$, λ sweeps $+\infty \to -\infty$. Numerically $\lambda \approx -0.371$ at $\mu = 4.5$, giving

$$p^\star \approx (0.054,\ 0.079,\ 0.114,\ 0.165,\ 0.240,\ 0.347),$$

each face a constant factor $e^{-\lambda} \approx 1.45$ larger than the one below it.

Remark 6 (The solution checks itself)**.** The optimality condition forces $\nabla H(p^\star)_i = \lambda E_i + \nu$, so the gradient's six components must lie on a single straight line in i, slope λ, intercept ν. Two faces pin λ and ν; the other four are redundant, and their agreeing (all six landing on the line) is a built-in verification of the solve. For $\mu = 4.5$ the gradient comes out as (1.90, 1.54, 1.17, 0.80, 0.43, 0.06), with every consecutive difference $\approx -0.37 = \lambda$.

7 The same machine

The boxed form $p_i = e^{-\lambda E_i}/Z$ is three familiar objects wearing one skeleton.

Softmax. With E_i the (negative) scores, $p = \mathrm{softmax}(-\lambda E)$. Among all distributions with a given expected score, this is the maximum-entropy one, which is why it sits at the end of essentially every classifier and language model.

Boltzmann. A gas molecule's "faces" are its energy levels; constraining the average energy gives probability $e^{-E/kT}/Z$ per level. Set $\lambda = 1/kT$ and it is the die formula character-for-character.

DPO / RLHF. Maximize reward while staying close to a reference policy π_{ref}, i.e. minimize KL divergence from π_{ref} subject to a reward target. The same crank gives the optimal policy

$$\pi(y) \;\propto\; \pi_{\text{ref}}(y)\, e^{\,\text{reward}(y)/\beta}.$$

Maximizing entropy was the special case of a *uniform* base measure; the general principle minimizes divergence from any prior, and here the prior is π_{ref}.

	max-entropy (the die)	KL-leash (DPO/RLHF)
objective	entropy $H(p)$	$-\mathrm{KL}(\pi \,\|\, \pi_{\text{ref}})$
base measure	uniform	π_{ref}
tilt	$e^{-\lambda E_i}$	$e^{\,\text{reward}/\beta}$
dial	λ (inverse temp.)	β (temperature)
spent multiplier	$\nu \to \ln Z(\lambda) - 1$	$\ln Z_\beta$ (the partition function)

The multipliers that survive are prices; the one that gets spent (ν, normalization) only enforces a definition. That is the whole arc: a loaded die, a gas, and a trillion-parameter model running the identical piece of mathematics.