Entropy, cross-entropy and KL divergence

Ten problems on entropy, cross-entropy and the Kullback–Leibler divergence: the Bernoulli entropy and its maximum, why the uniform distribution maximizes entropy, bits, nats and perplexity, cross-entropy as entropy plus KL, label smoothing, a proof that KL is non-negative, forward and reverse KL, the KL between two Gaussians and the VAE term with its gradients, with worked solutions and the mistakes that swap the arguments or drop a factor.

BEFORE YOU START

Entropy, cross-entropy and the KL divergence are three views of one quantity: the average number of nats it costs to describe outcomes from one distribution using another. Cross-entropy is the loss a classifier is trained on, the KL divergence is the term a variational autoencoder adds to it, and entropy is the floor neither can go below. These ten problems compute each by hand, prove the one inequality everything rests on, show that the two orders of KL prefer different fits, and derive the closed forms for Gaussians that a VAE uses. The five mistakes at the end are the ones that produce a plausible number: KL with its arguments swapped, a zero probability treated as harmless, a Gaussian log term without its $\frac{1}{2}$, a VAE gradient taken in the wrong variable, and a label-smoothed loss expected to reach zero.

- p and q are distributions over K outcomes $i = 1, \dots, K$: every $p_i \geq 0$ and $\sum_i p_i = 1$, and the same for q. $\log$ is the natural log, and quantities measured with it are in nats; $\log_2$ gives bits.
- The entropy is $H(p) = -\sum_i p_i \log p_i$, the cross-entropy is $H(p, q) = -\sum_i p_i \log q_i$, and the Kullback–Leibler divergence is $\mathrm{KL}(p \,\|\, q) = \sum_i p_i \log(p_i / q_i)$.
- Zero probabilities follow the usual conventions: a term with $p_i = 0$ is 0 in all three sums ($0 \log 0 = 0$, the limit of $t \log t$ as $t \to 0^+$), and a term with $p_i > 0$ and $q_i = 0$ makes $H(p, q)$ and $\mathrm{KL}(p \,\|\, q)$ equal to $+\infty$.
- u is the uniform distribution, $u_i = 1/K$. y is a one-hot target, 1 at the true class c and 0 elsewhere, and $\mathbf{1}$ is the all-ones vector.
- $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian with mean μ, variance $\sigma^2 > 0$ and density $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$. For densities the sums become integrals: $\mathrm{KL}(p \,\|\, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx$. $\mathbb{E}_p[f]$ is the expectation of f under p.
- The perplexity of a distribution or a model is e^H with H in nats, equivalently 2^H with H in bits: the number of equally likely outcomes that would have the same entropy.
- The softmax page derives the gradient of cross-entropy with respect to the logits, $s - y$, including soft targets; this page works with the distributions themselves.

Builds on: Maximum likelihood estimation: derivations by hand

Problems

1. The entropy of a coin with probability p of heads is $H(p) = -p\log p - (1-p)\log(1-p)$ for $0 < p < 1$. Compute $H'(p)$, find the p that maximizes H, and give the maximum in nats and in bits. difficulty ·

2. Over K outcomes, use a Lagrange multiplier to find the distribution p that maximizes $H(p)$. Then show that $\mathrm{KL}(p \,\|\, u) = \log K - H(p)$, and use it to confirm that the stationary point is the global maximum. difficulty ··

3. Let $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. Compute $H(p)$ in bits and in nats, and its perplexity. Then: a language model's mean negative log-likelihood on a test set is 2.3 nats per token. What is its perplexity? difficulty ·

4. Show that $H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q)$. Then take label smoothing with K classes and $0 < \varepsilon < 1$: the target is $p = (1-\varepsilon)y + \frac{\varepsilon}{K}\mathbf{1}$. Write $H(p, q)$ in terms of $\log q_c$ and $\sum_i \log q_i$, and give its minimum over q. difficulty ··

5. Use the inequality $\log t \le t - 1$ for $t > 0$ to prove that $\mathrm{KL}(p \,\|\, q) \ge 0$, with equality only when $q = p$. difficulty ···

6. Let $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\frac{9}{10}, \frac{1}{10})$. Compute $\mathrm{KL}(p \,\|\, q)$ and $\mathrm{KL}(q \,\|\, p)$. difficulty ·

7. Let $p = (\frac{1}{2}, 0, \frac{1}{2})$, a distribution with two modes, and consider two approximations: $q_A = (1, 0, 0)$, which keeps one mode, and $q_B = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, which covers everything. Compute $\mathrm{KL}(p \,\|\, q)$ and $\mathrm{KL}(q \,\|\, p)$ for both, and say which approximation each direction prefers. difficulty ··

8. Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$. Derive difficulty ···

$$\mathrm{KL}(p \,\|\, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

9. A VAE encoder outputs $\mu \in \mathbb{R}^d$ and $s \in \mathbb{R}^d$ with $s_j = \log \sigma_j^2$, defining $q = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$; the prior is $\mathcal{N}(0, I)$. Show that difficulty ···

$$\mathrm{KL}\left(q \,\|\, \mathcal{N}(0, I)\right) = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j^2 + e^{s_j} - s_j - 1\right),$$

and compute its gradients with respect to μ and s.

10. Data $x_1, \dots, x_N$ take values in $\{1, \dots, K\}$; outcome i occurs n_i times, difficulty ·· and $\hat{p}_i = n_i / N$ is the empirical distribution. A model assigns probabilities q. Show that the mean negative log-likelihood equals $H(\hat{p}, q)$, and use it to find the q that maximizes the likelihood and the smallest mean negative log-likelihood.

Answers

1. $H'(p) = \log \dfrac{1-p}{p}$; the maximum is at $p = \frac{1}{2}$, where $H = \ln 2 \approx 0.693$ nats $= 1$ bit
2. $p = u$, $p_i = 1/K$, with maximum $H(u) = \log K$; for every p, $H(p) = \log K - \mathrm{KL}(p \,\|\, u) \le \log K$
3. $H(p) = \frac{7}{4}$ bits $= \frac{7}{4} \ln 2 \approx 1.213$ nats; perplexity $2^{7/4} \approx 3.364$; a mean NLL of 2.3 nats gives perplexity $e^{2.3} \approx 9.97$
4. $H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q)$; with label smoothing $H(p, q) = -(1-\varepsilon)\log q_c - \frac{\varepsilon}{K}\sum_i \log q_i$, minimized at $q = p$ with value $H(p) > 0$
5. $\mathrm{KL}(p \,\|\, q) \ge 0$ for all p, q, with equality if and only if $q = p$
6. $\mathrm{KL}(p \,\|\, q) = \ln \frac{5}{3} \approx 0.511$; $\mathrm{KL}(q \,\|\, p) = 0.9 \ln 1.8 - 0.1 \ln 5 \approx 0.368$
7. $\mathrm{KL}(p \,\|\, q_A) = \infty$, $\mathrm{KL}(p \,\|\, q_B) = \ln \frac{3}{2}$; $\mathrm{KL}(q_A \,\|\, p) = \ln 2$, $\mathrm{KL}(q_B \,\|\, p) = \infty$: the forward KL prefers the covering q_B, the reverse KL the single-mode q_A
8. $\mathrm{KL}(p \,\|\, q) = \ln \dfrac{\sigma_2}{\sigma_1} + \dfrac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \dfrac{1}{2}$
9. $\mathrm{KL} = \frac{1}{2}\sum_j (\mu_j^2 + e^{s_j} - s_j - 1)$; $\nabla_\mu \mathrm{KL} = \mu$ and $\partial\,\mathrm{KL}/\partial s_j = \frac{1}{2}(e^{s_j} - 1) = \frac{1}{2}(\sigma_j^2 - 1)$
10. Mean NLL $= H(\hat{p}, q) = H(\hat{p}) + \mathrm{KL}(\hat{p} \,\|\, q)$; it is minimized at $q_i = n_i/N$, where it equals $H(\hat{p})$

Worked solutions

Problem 1

> The entropy of a coin with probability p of heads is $H(p) = -p\log p - (1-p)\log(1-p)$ for $0 < p < 1$. Compute $H'(p)$, find the p that maximizes H, and give the maximum in nats and in bits.

$\frac{d}{dp}\big[-p\log p\big] = -\log p - 1.$	Product rule: p differentiates to 1, and $p \cdot \frac{1}{p} = 1$.
$\frac{d}{dp}\big[-(1-p)\log(1-p)\big] = \log(1-p) + 1.$	The same derivative at $1-p$, times the inner derivative -1, which flips both signs.
$H'(p) = \log(1-p) - \log p = \log\dfrac{1-p}{p}.$	Add steps 1 and 2; the -1 and $+1$ cancel.
$H'(p) > 0$ for $p < \frac{1}{2}$, $H'(p) = 0$ at $p = \frac{1}{2}$, and $H'(p) < 0$ for $p > \frac{1}{2}$.	$\log t > 0$ exactly when $t > 1$, and $(1-p)/p > 1$ exactly when $p < \frac{1}{2}$. So H rises and then falls, and its one stationary point is the maximum.
$H(\frac{1}{2}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2.$	Both terms are equal.
$H'(p) = \log\dfrac{1-p}{p}$; **the maximum is at $p = \frac{1}{2}$, where $H = \ln 2 \approx 0.693$ nats $= 1$ bit**	Dividing by $\ln 2$ converts nats to bits, because $\log_2 t = \ln t / \ln 2$. A fair coin is the most uncertain coin, and one flip carries exactly one bit.

Problem 2

> Over K outcomes, use a Lagrange multiplier to find the distribution p that maximizes $H(p)$. Then show that $\mathrm{KL}(p \,\|\, u) = \log K - H(p)$, and use it to confirm that the stationary point is the global maximum.

$\mathcal{L}(p, \lambda) = -\sum_i p_i \log p_i + \lambda\big(\sum_i p_i - 1\big)$.	The constraint $\sum_i p_i = 1$ enters with a multiplier. The constraints $p_i \geq 0$ can be left out: $\partial H / \partial p_i = -\log p_i - 1 \to +\infty$ as $p_i \to 0^+$, so a maximizer never sits on that boundary.
$\partial \mathcal{L} / \partial p_i = -\log p_i - 1 + \lambda = 0$, so $p_i = e^{\lambda - 1}$.	Only the i-th term of each sum contains p_i. The right-hand side is the same for every i, so all the p_i are equal.
$\sum_i p_i = K e^{\lambda - 1} = 1$, so $p_i = 1/K$, and $H(u) = -\sum_i \frac{1}{K} \log \frac{1}{K} = \log K$.	The constraint fixes the common value.
$\mathrm{KL}(p \,\|\, u) = \sum_i p_i \log(K p_i) =$ $\log K \sum_i p_i + \sum_i p_i \log p_i = \log K - H(p)$.	$p_i / u_i = K p_i$, the log of a product is a sum of logs, and $\sum_i p_i = 1$.
$p = u$, $p_i = 1/K$, **with maximum** $H(u) = \log K$; **for every** p, $H(p) =$ $\log K - \mathrm{KL}(p \,\|\, u) \leq \log K$	$\mathrm{KL} \geq 0$ with equality only when the two distributions are equal (Problem 5), so step 4 bounds every entropy by $\log K$, with equality only at $p = u$. No second-derivative test is needed.

Problem 3

> Let $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. Compute $H(p)$ in bits and in nats, and its perplexity. Then: a language model's mean negative log-likelihood on a test set is 2.3 nats per token. What is its perplexity?

$H(p) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}$ bits.	Each probability is a power of 2, and $-\log_2 2^{-k} = k$.
$H(p) = \frac{7}{4} \ln 2 \approx 1.213$ nats.	$\ln t = \ln 2 \cdot \log_2 t$, so every term, and the sum, scales by $\ln 2$.

Perplexity $= 2^{7/4} = e^{(7/4)\ln 2} \approx 3.364.$	The base of the exponential must match the unit of H: 2 for bits, e for nats. Both give the same number. It lies between 2 and 4: the distribution is as uncertain as a fair choice among about 3.4 outcomes.
The mean negative log-likelihood is the cross-entropy between the test data and the model, in nats, so the perplexity is $e^{2.3}$.	Problem 10 shows the mean negative log-likelihood is a cross-entropy; the unit is nats because the loss used ln.
$H(p) = \frac{7}{4}$ **bits** $= \frac{7}{4}\ln 2 \approx 1.213$ **nats; perplexity** $2^{7/4} \approx 3.364$; **a mean NLL of** 2.3 **nats gives perplexity** $e^{2.3} \approx 9.97$	The model is, on average, as unsure of the next token as a uniform choice among about ten.

Problem 4

> Show that $H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q)$. Then take label smoothing with K classes and $0 < \varepsilon < 1$: the target is $p = (1-\varepsilon)y + \frac{\varepsilon}{K}\mathbf{1}$. Write $H(p, q)$ in terms of $\log q_c$ and $\sum_i \log q_i$, and give its minimum over q.

$\mathrm{KL}(p \,\|\, q) = \sum_i p_i \log p_i - \sum_i p_i \log q_i = -H(p) + H(p, q).$	$\log(p_i/q_i) = \log p_i - \log q_i$, and each sum is one of the definitions.
$H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q).$	Rearrange step 1.
$H(p)$ does not depend on q, so minimizing $H(p, q)$ over q is minimizing $\mathrm{KL}(p \,\|\, q)$, and the minimum is $H(p)$, at $q = p$.	$\mathrm{KL}(p \,\|\, q) \geq 0$, with equality only at $q = p$ (Problem 5).
$H(p, q) = -\sum_i \left[(1-\varepsilon)y_i + \frac{\varepsilon}{K}\right] \log q_i = -(1-\varepsilon)\log q_c - \frac{\varepsilon}{K}\sum_i \log q_i.$	Cross-entropy is linear in its first argument, and the one-hot y picks out the term $i = c$.

$H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q)$; **with label smoothing** $H(p, q) = -(1-\varepsilon)\log q_c - \frac{\varepsilon}{K}\sum_i \log q_i$, **minimized at** $q = p$ **with value** $H(p) > 0$	The second term penalizes any class whose probability goes to 0, which is how smoothing stops the logits growing without bound. Because p is not one-hot, $H(p) > 0$: the loss has a floor above zero (the last mistake below).

Problem 5

> Use the inequality $\log t \le t - 1$ for $t > 0$ to prove that $\mathrm{KL}(p \,\|\, q) \ge 0$, with equality only when $q = p$.

$\log t \le t - 1$ for every $t > 0$, with equality only at $t = 1$.	$f(t) = t - 1 - \log t$ has $f'(t) = 1 - 1/t$, negative for $t < 1$ and positive for $t > 1$, so its minimum is $f(1) = 0$.
Let $S = \{i : p_i > 0\}$. If $q_i = 0$ for some $i \in S$, then $\mathrm{KL}(p \,\|\, q) = +\infty > 0$. Otherwise $-\mathrm{KL}(p \,\|\, q) = \sum_{i \in S} p_i \log \dfrac{q_i}{p_i}$.	Terms with $p_i = 0$ are 0 by convention, and $-\log(p_i/q_i) = \log(q_i/p_i)$.
$\sum_{i \in S} p_i \log \dfrac{q_i}{p_i} \le \sum_{i \in S} p_i \left(\dfrac{q_i}{p_i} - 1 \right) = \sum_{i \in S} q_i - 1$.	Step 1 with $t = q_i/p_i > 0$, multiplied by the positive weight p_i; then $\sum_{i \in S} p_i = 1$.
$\sum_{i \in S} q_i \le \sum_i q_i = 1$, so $-\mathrm{KL}(p \,\|\, q) \le 0$.	Leaving out terms of a sum of non-negative numbers can only make it smaller.
Equality needs $q_i/p_i = 1$ for every $i \in S$ (step 3) and $\sum_{i \in S} q_i = 1$ (step 4).	The equality case of step 1 applies term by term, because every weight p_i in step 3 is positive.
So $q_i = p_i$ on S, and $q_i = 0 = p_i$ off it.	The first condition is $q_i = p_i$; and the q_i on S already sum to 1, which leaves nothing for the entries outside S, where $p_i = 0$ by the definition of S.

$\mathrm{KL}(p \,\|\, q) \geq 0$ **for all** p, q**, with equality if and only if** $q = p$	This is Gibbs' inequality. It is what makes $H(p)$ the floor of the cross-entropy (Problem 4), $\log K$ the ceiling of the entropy (Problem 2) and the empirical distribution the maximum-likelihood fit (Problem 10).

Problem 6

> Let $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\frac{9}{10}, \frac{1}{10})$. Compute $\mathrm{KL}(p \,\|\, q)$ and $\mathrm{KL}(q \,\|\, p)$.

$\mathrm{KL}(p \,\|\, q) = \frac{1}{2}\log\frac{1/2}{9/10} + \frac{1}{2}\log\frac{1/2}{1/10} = \frac{1}{2}\log\frac{5}{9} + \frac{1}{2}\log 5 = \frac{1}{2}\log\frac{25}{9} = \log\frac{5}{3}.$	The definition, then $\frac{1}{2}(\log a + \log b) = \frac{1}{2}\log ab$ and $\frac{1}{2}\log(25/9) = \log(5/3)$.
$\mathrm{KL}(q \,\|\, p) = \frac{9}{10}\log\frac{9/10}{1/2} + \frac{1}{10}\log\frac{1/10}{1/2} = 0.9\log 1.8 - 0.1\log 5.$	The weights now come from q, and $\log(1/5) = -\log 5$.
$\mathrm{KL}(p \,\|\, q) = \ln\frac{5}{3} \approx 0.511$; $\mathrm{KL}(q \,\|\, p) = 0.9\ln 1.8 - 0.1\ln 5 \approx 0.368$	The two differ, so KL is not symmetric and not a distance. The larger one weights by p, which puts half its mass on an outcome q gives only $\frac{1}{10}$: KL is large when the first distribution is often surprised by the second.

Problem 7

> Let $p = (\frac{1}{2}, 0, \frac{1}{2})$, a distribution with two modes, and consider two approximations: $q_A = (1, 0, 0)$, which keeps one mode, and $q_B = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, which covers everything. Compute $\mathrm{KL}(p \,\|\, q)$ and $\mathrm{KL}(q \,\|\, p)$ for both, and say which approximation each direction prefers.

$\mathrm{KL}(p \,\|\, q_A) = +\infty.$	Outcome 3 has $p_3 = \frac{1}{2} > 0$ and $q_{A,3} = 0$, so its term is $\frac{1}{2}\log(\frac{1}{2}/0) = +\infty$.
$\mathrm{KL}(p \,\|\, q_B) = \frac{1}{2}\log\frac{1/2}{1/3} + 0 + \frac{1}{2}\log\frac{1/2}{1/3} = \log\frac{3}{2} \approx 0.405.$	The middle term has $p_2 = 0$ and is 0 by convention.
$\mathrm{KL}(q_A \,\|\, p) = 1 \cdot \log\frac{1}{1/2} = \log 2 \approx 0.693.$	Now the weights are q_A's; the terms with $q_{A,i} = 0$ vanish.

$\mathrm{KL}(q_B \,\|\, p) = +\infty.$	Outcome 2 has $q_{B,2} = \frac{1}{3} > 0$ and $p_2 = 0$.
$\mathrm{KL}(p \,\|\, q_A) = \infty$, $\mathrm{KL}(p \,\|\, q_B) = \ln \frac{3}{2}$; $\mathrm{KL}(q_A \,\|\, p) = \ln 2$, $\mathrm{KL}(q_B \,\|\, p) = \infty$: **the forward KL prefers the covering q_B, the reverse KL the single-mode q_A**	$\mathrm{KL}(p \,\|\, q)$ is infinite when q misses mass that p has, so minimizing it over q spreads q out (mass-covering). $\mathrm{KL}(q \,\|\, p)$ is infinite when q puts mass where p has none, so minimizing it shrinks q onto one mode (mode-seeking). Maximum likelihood minimizes the forward direction (Problem 10); variational inference, including the VAE, minimizes the reverse.

Problem 8

Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$. Derive

$$\mathrm{KL}(p \,\|\, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

$\log p(x) = -\frac{1}{2}\log(2\pi\sigma_1^2) - \dfrac{(x-\mu_1)^2}{2\sigma_1^2}$, and likewise for q.	The log of the density in the definitions.
$\log p(x) - \log q(x) = \log \dfrac{\sigma_2}{\sigma_1} - \dfrac{(x-\mu_1)^2}{2\sigma_1^2} + \dfrac{(x-\mu_2)^2}{2\sigma_2^2}.$	The 2π cancels, and $-\frac{1}{2}\log\sigma_1^2 + \frac{1}{2}\log\sigma_2^2 = \log(\sigma_2/\sigma_1)$.
$\mathbb{E}_p\big[(x-\mu_1)^2\big] = \sigma_1^2.$	The definition of the variance of p.
$\mathbb{E}_p\big[(x-\mu_2)^2\big] = \sigma_1^2 + (\mu_1-\mu_2)^2.$	Write $x - \mu_2 = (x-\mu_1) + (\mu_1 - \mu_2)$ and expand: the cross term is $2(\mu_1 - \mu_2)\,\mathbb{E}_p[x-\mu_1] = 0$.
$\mathrm{KL}(p \,\|\, q) = \mathbb{E}_p[\log p(x) - \log q(x)] = \log \dfrac{\sigma_2}{\sigma_1} - \dfrac{\sigma_1^2}{2\sigma_1^2} + \dfrac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2}.$	KL is the expectation under p of the log-ratio; expectation is linear, so steps 3 and 4 apply term by term.

$\mathrm{KL}(p \,\|\, q) = \ln \dfrac{\sigma_2}{\sigma_1} + \dfrac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \dfrac{1}{2}$	Sanity check: with $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ it is $0 + \frac{1}{2} - \frac{1}{2} = 0$. The mean difference is measured in units of q's spread, σ_2, which is again not symmetric.

Problem 9

A VAE encoder outputs $\mu \in \mathbb{R}^d$ and $s \in \mathbb{R}^d$ with $s_j = \log \sigma_j^2$, defining $q = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$; the prior is $\mathcal{N}(0, I)$. Show that

$$\begin{aligned} &\mathrm{KL}\left(q \,\|\, \mathcal{N}(0, I)\right) \\ &\quad = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j^2 + e^{s_j} - s_j - 1\right), \end{aligned}$$

and compute its gradients with respect to μ and s.

Both densities are products over coordinates, so $\log q(x) - \log p_0(x) = \sum_j \left[\log q_j(x_j) - \log p_0(x_j)\right]$, and the KL is $\sum_j \mathrm{KL}\left(\mathcal{N}(\mu_j, \sigma_j^2) \,\|\, \mathcal{N}(0, 1)\right)$.	A diagonal covariance makes the coordinates independent, the log of a product is a sum, and the expectation of each term involves only x_j, whose distribution under q is $\mathcal{N}(\mu_j, \sigma_j^2)$.
Each term is $-\log \sigma_j + \frac{1}{2}(\sigma_j^2 + \mu_j^2) - \frac{1}{2}$.	Problem 8 with $\mu_1 = \mu_j$, $\sigma_1 = \sigma_j$, $\mu_2 = 0$, $\sigma_2 = 1$.
$-\log \sigma_j = -\frac{1}{2} \log \sigma_j^2 = -\frac{1}{2} s_j$, and $\sigma_j^2 = e^{s_j}$.	Rewrite in the variable the encoder outputs.
$\mathrm{KL} = \frac{1}{2} \sum_j \left(\mu_j^2 + e^{s_j} - s_j - 1\right)$.	Steps 2 and 3, summed over j.
$\partial \,\mathrm{KL} / \partial \mu_j = \mu_j$ and $\partial \,\mathrm{KL} / \partial s_j = \frac{1}{2}(e^{s_j} - 1)$.	Only the j-th term contains μ_j or s_j; $\frac{1}{2}\mu_j^2$ differentiates to μ_j, and $\frac{1}{2}(e^{s_j} - s_j)$ to $\frac{1}{2}(e^{s_j} - 1)$.
$\mathrm{KL} = \frac{1}{2} \sum_j (\mu_j^2 + e^{s_j} - s_j - 1)$; $\nabla_\mu \,\mathrm{KL} = \mu$ **and** $\partial \,\mathrm{KL} / \partial s_j = \frac{1}{2}(e^{s_j} - 1) = \frac{1}{2}(\sigma_j^2 - 1)$	Both gradients vanish exactly at $\mu = 0$, $\sigma^2 = \mathbf{1}$, the prior. Outputting $s = \log \sigma^2$ rather than σ^2 keeps the variance positive with no constraint on the network.

Problem 10

> Data $x_1, \dots, x_N$ take values in $\{1, \dots, K\}$; outcome i occurs n_i times, and $\hat{p}_i = n_i/N$ is the empirical distribution. A model assigns probabilities q. Show that the mean negative log-likelihood equals $H(\hat{p}, q)$, and use it to find the q that maximizes the likelihood and the smallest mean negative log-likelihood.

$\frac{1}{N}\sum_{n=1}^{N} -\log q_{x_n} = \sum_{i=1}^{K} \frac{n_i}{N}\,(-\log q_i).$	Group the terms by outcome: the term $-\log q_i$ appears once for each of the n_i data points equal to i.
$= H(\hat{p}, q) = H(\hat{p}) + \mathrm{KL}(\hat{p} \,\|\, q).$	The definition of cross-entropy with $\hat{p}_i = n_i/N$, then Problem 4.
$H(\hat{p})$ does not depend on q, and $\mathrm{KL}(\hat{p} \,\|\, q) \geq 0$ with equality only at $q = \hat{p}$.	Problem 5. Maximizing the likelihood is minimizing the mean negative log-likelihood, because log is increasing and $1/N > 0$.
Mean NLL $= H(\hat{p}, q) = H(\hat{p}) + \mathrm{KL}(\hat{p} \,\|\, q)$; **it is minimized at** $q_i = n_i/N$, **where it equals** $H(\hat{p})$	The maximum-likelihood page reaches the same estimate with a Lagrange multiplier. For a model family q_θ that cannot reach $\hat{p}$, step 2 still holds: maximum likelihood picks the θ minimizing the forward divergence $\mathrm{KL}(\hat{p} \,\|\, q_\theta)$.

Where this goes wrong

1. Maximum likelihood written as KL(q_θ ‖ p̂)

KL is often described as the divergence between the model and the data, and the order of its arguments looks like a choice of notation.

Mean NLL $= H(\hat{p}) + \mathrm{KL}(\hat{p} \,\|\, q_\theta)$	Right so far: Problem 10, step 2.
"Fitting a model to data minimizes how far the model is from the data, $\mathrm{KL}(\text{model} \,\|\, \text{data})$."	The phrasing that causes the mistake: it names the model first, but the order is fixed by which distribution the log-loss averages over, and the loss averages over the data.

Maximum likelihood minimizes $\mathrm{KL}(q_\theta \,\|\, \hat{p})$ ✗	$\hat{p}$ is 0 on every outcome that is absent from the data, so $\mathrm{KL}(q_\theta \,\|\, \hat{p}) = \infty$ for any model that gives an unseen outcome positive probability, including every softmax model. Maximum likelihood minimizes $\mathrm{KL}(\hat{p} \,\|\, q_\theta)$, the mass-covering direction (Problem 7); the reverse is the one variational inference minimizes.

2. KL term with q_i = 0 dropped as if it were 0 log 0

The convention $0 \log 0 = 0$ makes zero probabilities harmless in an entropy, and it is tempting to apply it to every zero in a KL.

$\mathrm{KL}(p \,\|\, q_A) = \sum_i p_i \log(p_i / q_{A,i})$ with $p = (\frac{1}{2}, 0, \frac{1}{2})$ and $q_A = (1, 0, 0)$	Right so far: the set-up of Problem 7.
"Terms with a zero probability contribute nothing, so keep only the outcomes where both are positive."	The shortcut that causes the mistake: the convention covers a zero weight p_i outside the log, not a zero q_i inside it.
$\mathrm{KL}(p \,\|\, q_A) = \frac{1}{2} \log \frac{1/2}{1} = -\frac{1}{2} \ln 2$ ✗	A negative KL is already a contradiction of Problem 5. The third term has $p_3 = \frac{1}{2}$ and $q_{A,3} = 0$, so it is $+\infty$, and so is the KL (Problem 7, step 1). In code the same slip appears as a mask that skips $q_i = 0$, or a clamp $q_i \geq 10^{-12}$ that turns ∞ into a large finite number.

3. Gaussian KL with log(σ₂²/σ₁²) and no ½

Gaussians are usually parametrized by their variances, and log-variance is what a network outputs, so the log term gets written with variances.

$\log p(x) = -\frac{1}{2} \log(2\pi\sigma_1^2) - (x - \mu_1)^2 / (2\sigma_1^2)$	Right so far: Problem 8, step 1.
"The normalizing constants contribute the log of the ratio of the variances."	The shortcut that causes the mistake: dropping the $\frac{1}{2}$ that comes from the square root in $1/\sqrt{2\pi\sigma^2}$.

$\mathrm{KL} = \log(\sigma_2^2/\sigma_1^2) + \dfrac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \dfrac{1}{2}$ ✗	The log term is twice what it should be: $\log(\sigma_2/\sigma_1) = \frac{1}{2}\log(\sigma_2^2/\sigma_1^2)$ (Problem 8). With $\mu_1 = \mu_2$, $\sigma_1 = 1$, $\sigma_2 = 2$ the correct value is $\ln 2 + \frac{1}{8} - \frac{1}{2} \approx 0.318$ and this gives $\ln 4 - \frac{3}{8} \approx 1.011$. The test $p = q$ still gives 0, so only a check with unequal variances catches it.

4. VAE gradient for log σ² computed as the derivative in σ

Code calls the encoder's second output “the variance”, and the KL is easiest to differentiate in the form it was first written in, with σ_j.

$\mathrm{KL}_j = -\log \sigma_j + \frac{1}{2}(\sigma_j^2 + \mu_j^2) - \frac{1}{2}$	Right so far: Problem 9, step 2.
“Differentiate with respect to the variance parameter, σ_j, and pass that to the encoder.”	The habit that causes the mistake: the encoder outputs $s_j = \log \sigma_j^2$, and the gradient backprop needs is with respect to that output.
$\partial\,\mathrm{KL}/\partial s_j = \sigma_j - 1/\sigma_j$ ✗	That is $\partial\,\mathrm{KL}/\partial\sigma_j$. With $\sigma_j = e^{s_j/2}$, $d\sigma_j/ds_j = \sigma_j/2$, so $\partial\,\mathrm{KL}/\partial s_j = (\sigma_j - 1/\sigma_j)\,\sigma_j/2 = \frac{1}{2}(\sigma_j^2 - 1)$ (Problem 9). Both vanish at $\sigma_j = 1$, which hides the slip, but at $\sigma_j = 3$ the wrong value is $\frac{8}{3}$ against the correct 4.

5. Reading a label-smoothed loss of 0.51 as far from optimal

A one-hot cross-entropy can be driven toward 0, and a loss curve trained with label smoothing gets read on the same scale.

$H(p, q) = H(p) + \mathrm{KL}(p \,\|\, q)$ with $p = (1 - \varepsilon)y + \frac{\varepsilon}{K}\mathbf{1}$	Right so far: Problem 4.
“A perfect model has zero loss.”	The assumption that causes the mistake: it holds only for a one-hot target, whose entropy is 0.

With $\varepsilon = 0.1$ and $K = 10$, a training loss that levels off at 0.51 nats means the model is still far from its targets ✗

The floor is $H(p)$, with $p_c = 0.91$ and the other nine entries 0.01: $-0.91 \ln 0.91 - 9(0.01) \ln 0.01 \approx 0.500$. A loss of 0.51 is a KL of about 0.01, almost at the optimum. Subtract $H(p)$ before reading the curve.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 25 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/entropy-cross-entropy-and-kl-divergence