

Jensen's inequality and convexity

Ten problems on convex functions and Jensen's inequality: the second-derivative test, the tangent-line proof of Jensen, the finite version by induction, AM–GM, the Jensen gaps behind the variance and the Gaussian moment generating function, why the ELBO is a lower bound and why averaging importance weights tightens it, the log-sum-exp bounds and its PSD Hessian, convexity-preserving operations applied to the logistic and hinge losses, the entropy ceiling and Gibbs' inequality, and the ambiguity decomposition for ensembles, with worked solutions and the mistakes that compose convex functions, apply Jensen to a plain sum, swap $E[1/X]$ for $1/E[X]$, call log-sum-exp strictly convex, or test convexity at one point.

BEFORE YOU START

One inequality sits under more of machine learning than any other: for a convex function, the function of an average is at most the average of the function. It is why the variance is nonnegative, why the KL divergence is nonnegative, why the evidence lower bound is a lower bound, why log-sum-exp is a smooth maximum, and why an ensemble's average prediction loses no more than its members do on average. These ten problems build it from the second-derivative test and the tangent line, prove the finite version from the definition alone, and then spend it: on the AM–GM inequality, on the moment generating function of a Gaussian, on the ELBO and the importance-weighted bound that tightens it, on the softmax Hessian, on the convexity of the logistic and hinge losses, on the entropy ceiling, and on the ambiguity decomposition. The five mistakes at the end are the ones that produce a true-looking inequality that is false: composing convex functions, applying Jensen to a sum instead of an average, replacing the mean of a reciprocal by the reciprocal of the mean, calling a flat direction strictly convex, and reading convexity off a single point.

- A function f on an interval I is convex if $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for all $x, y \in I$ and $\lambda \in [0, 1]$: the chord between two points of the graph lies on or above the graph. It is strictly convex if the inequality is strict whenever $x \neq y$ and $0 < \lambda < 1$, and concave if $-f$ is convex. The same definition applies to f on $\mathbb{R}^n$ with vectors x, y .
- Jensen's inequality: for a convex f and a random variable X with finite mean, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$; for concave f the inequality reverses. For a finite distribution with weights $p_i \geq 0$, $\sum_i p_i = 1$, it reads $f(\sum_i p_i x_i) \leq \sum_i p_i f(x_i)$. The difference $\mathbb{E}[f(X)] - f(\mathbb{E}[X])$ is the Jensen gap.
- log is the natural logarithm. $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid with $\sigma' = \sigma(1 - \sigma)$, and the softplus is $\text{sp}(t) = \log(1 + e^t)$, with $\text{sp}' = \sigma$ (the activation-functions page).

- For $z \in \mathbb{R}^n$, $\text{LSE}(z) = \log \sum_i e^{z_i}$ is the log-sum-exp and $s = \text{softmax}(z)$ has $s_i = e^{z_i} / \sum_j e^{z_j}$; $\mathbf{1}$ is the all-ones vector and $\text{diag}(s)$ the diagonal matrix with s on its diagonal. A symmetric matrix H is positive semidefinite (PSD) if $v^\top H v \geq 0$ for every v , and a twice-differentiable f on $\mathbb{R}^n$ is convex exactly when its Hessian $\nabla^2 f$ is PSD everywhere (the gradients-and-Hessians page).
- p and q are distributions over K outcomes with $p_i, q_i > 0$; the entropy is $H(p) = -\sum_i p_i \log p_i$ and $\text{KL}(p \parallel q) = \sum_i p_i \log(p_i/q_i)$ (the entropy page). For a latent-variable model, $p(x, z)$ is the joint, $p(x) = \sum_z p(x, z)$ the evidence, $q(z)$ a distribution the model chooses, and $w(z) = p(x, z)/q(z)$ the importance weight.

Builds on: Entropy, cross-entropy and KL divergence, Variance, covariance and correlation

Problems

1. Using the second derivative, decide which of e^x , $-\log x$, $x \log x$ and $\sqrt{x}$ are convex and which are concave on $x > 0$. difficulty ..

2. Let f be twice differentiable with $f'' \geq 0$ on an interval. Show that the tangent line lies below the graph, $f(y) \geq f(x) + f'(x)(y - x)$, and deduce Jensen's inequality $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ for any random variable X taking values in the interval with finite mean. difficulty ..

3. Without derivatives, show from the definition of convexity that $f(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i f(x_i)$ for weights $p_i \geq 0$ with $\sum_i p_i = 1$, by induction on n . Why does this version apply to $f(x) = |x|$, which has no second derivative? difficulty ..

4. From Jensen's inequality for log, prove the weighted AM–GM inequality $\prod_i a_i^{p_i} \leq \sum_i p_i a_i$ for $a_i > 0$ and weights $p_i \geq 0$, $\sum_i p_i = 1$. Evaluate both sides for $a = (1, 4)$ with equal weights and for $a = (1, 2, 4)$ with equal weights. difficulty ..

5. Show that each of the following is a Jensen gap and compute it: (a) $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$ for any X with finite variance; (b) $\mathbb{E}[e^X] - e^{\mathbb{E}[X]}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$, by first showing $\mathbb{E}[e^X] = e^{\mu + \sigma^2/2}$; (c) $\mathbb{E}[1/X] - 1/\mathbb{E}[X]$ for X equal to 1 or 4 with probability $\frac{1}{2}$ each. difficulty ..

6. For a latent-variable model with $z \sim q$ and $w(z) = p(x, z)/q(z)$, difficulty ... show that $\log p(x) \geq \mathbb{E}_q[\log w(z)]$, the evidence lower bound. Then, for K independent draws $z_1, \dots, z_K \sim q$, show that $\log p(x) \geq \mathbb{E}\left[\log \frac{1}{K} \sum_{k=1}^K w(z_k)\right] \geq \mathbb{E}_q[\log w(z)]$: averaging the weights inside the log tightens the bound.

7. For $\text{LSE}(z) = \log \sum_{i=1}^n e^{z_i}$, show that $\max_i z_i \leq \text{LSE}(z) \leq \max_i z_i + \log n$, that $\nabla \text{LSE}(z) = s = \text{softmax}(z)$, and that $\nabla^2 \text{LSE}(z) = \text{diag}(s) - ss^\top$ is PSD, so LSE is convex.

8. Show from the definition that (a) if f is convex then $g(x) = f(Ax + b)$ is convex, and (b) if $f_1, \dots, f_m$ are convex then $g(x) = \max_j f_j(x)$ is convex. Use (a) to show that the logistic loss $\ell(w) = \log(1 + e^{-y w^\top x})$ is convex in w for fixed $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, and compute its Hessian. Use (b) for the hinge loss $h(w) = \max(0, 1 - y w^\top x)$.

9. Use Jensen's inequality to show that (a) $H(p) \leq \log K$ for every distribution p over K outcomes, with equality for the uniform distribution, and (b) $\text{KL}(p \parallel q) \geq 0$. difficulty ..

10. An ensemble of M models makes predictions $\hat{y}_1, \dots, \hat{y}_M$ for a target y , and the ensemble predicts the average $\bar{y} = \frac{1}{M} \sum_m \hat{y}_m$. (a) Show that for any loss $\ell(\hat{y}, y)$ that is convex in $\hat{y}$, $\ell(\bar{y}, y) \leq \frac{1}{M} \sum_m \ell(\hat{y}_m, y)$. (b) For the squared loss, show the exact identity $(\bar{y} - y)^2 = \frac{1}{M} \sum_m (\hat{y}_m - y)^2 - \frac{1}{M} \sum_m (\hat{y}_m - \bar{y})^2$. difficulty ...

Answers

1. e^x , $-\log x$ and $x \log x$ are convex on $x > 0$ (second derivatives e^x , $1/x^2$, $1/x$); $\sqrt{x}$ and $\log x$ are concave (second derivatives $-\frac{1}{4}x^{-3/2}$ and $-1/x^2$)
2. $f(y) \geq f(x) + f'(x)(y - x)$, and $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$
3. $f(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i f(x_i)$
4. $\prod_i a_i^{p_i} \leq \sum_i p_i a_i$; for $(1, 4)$: $2 \leq 2.5$; for $(1, 2, 4)$: $2 \leq 7/3$
5. (a) the gap is $\text{Var}(X)$; (b) $\mathbb{E}[e^X] = e^{\mu + \sigma^2/2}$ and the gap is $e^\mu(e^{\sigma^2/2} - 1)$; (c) $\mathbb{E}[1/X] = 0.625 > 0.4 = 1/\mathbb{E}[X]$, gap 0.225
6. $\log p(x) \geq \mathbb{E}[\log \frac{1}{K} \sum_k w(z_k)] \geq \mathbb{E}_q[\log w(z)]$, the right-hand side being the ELBO
7. $\max_i z_i \leq \text{LSE}(z) \leq \max_i z_i + \log n$; $\nabla \text{LSE} = \text{softmax}(z)$; $\nabla^2 \text{LSE} = \text{diag}(s) - ss^\top$ is PSD, so LSE is convex
8. $f(Ax + b)$ and $\max_j f_j(x)$ are convex when f, f_j are; $\ell(w) = \text{sp}(-y w^\top x)$ is convex with $\nabla^2 \ell = \sigma(1 - \sigma)xx^\top$, $\sigma = \sigma(-y w^\top x)$; the hinge loss is a maximum of two affine functions, hence convex
9. $H(p) \leq \log K$ with equality for the uniform distribution; $\text{KL}(p \parallel q) \geq 0$
10. $\ell(\bar{y}, y) \leq \frac{1}{M} \sum_m \ell(\hat{y}_m, y)$ for convex ℓ ; for the squared loss, $(\bar{y} - y)^2 = \frac{1}{M} \sum_m (\hat{y}_m - y)^2 - \frac{1}{M} \sum_m (\hat{y}_m - \bar{y})^2$

Worked solutions

Problem 1

Using the second derivative, decide which of e^x , $-\log x$, $x \log x$ and $\sqrt{x}$ are convex and which are concave on $x > 0$.

$(e^x)'' = e^x > 0.$	Differentiate twice; the exponential is its own derivative and is positive everywhere.
$(-\log x)'' = (-1/x)' = 1/x^2 > 0.$	The derivative of $-\log x$ is $-1/x$, and the derivative of $-x^{-1}$ is x^{-2} .
$(x \log x)'' = (\log x + 1)' = 1/x > 0.$	Product rule for the first derivative: $1 \cdot \log x + x \cdot \frac{1}{x} = \log x + 1$.
$(\sqrt{x})'' = (\frac{1}{2}x^{-1/2})' = -\frac{1}{4}x^{-3/2} < 0.$	Power rule twice; the second derivative is negative for every $x > 0$.
e^x, $-\log x$ and $x \log x$ are convex on $x > 0$ (second derivatives e^x, $1/x^2$, $1/x$); $\sqrt{x}$ and $\log x$ are concave (second derivatives $-\frac{1}{4}x^{-3/2}$ and $-1/x^2$)	A nonnegative second derivative on an interval means the slope never decreases, so the graph bends upwards and every chord lies above it; Problem 2 makes that precise through the tangent line. The same test in several variables asks whether the Hessian is PSD (Problem 7). Concavity of $\log$ is the fact behind Problems 4, 6 and 9, and convexity of $x \log x$ is the fact behind the entropy being concave.

Problem 2

Let f be twice differentiable with $f'' \geq 0$ on an interval. Show that the tangent line lies below the graph, $f(y) \geq f(x) + f'(x)(y - x)$, and deduce Jensen's inequality $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ for any random variable X taking values in the interval with finite mean.

$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(\xi)(y - x)^2$ for some ξ between x and y .	Taylor's theorem with the Lagrange remainder at order one (the Taylor-series page).
--	--

$\frac{1}{2}f''(\xi)(y-x)^2 \geq 0$, so $f(y) \geq f(x) + f'(x)(y-x)$.	$f'' \geq 0$ on the interval and a square is nonnegative.
Put $x = \mu = \mathbb{E}[X]$ and $y = X$: $f(X) \geq f(\mu) + f'(\mu)(X - \mu)$ for every value of X .	Step 2 holds for every pair of points in the interval, and μ lies in the interval because it is an average of values in it.
$\mathbb{E}[f(X)] \geq f(\mu) + f'(\mu) \mathbb{E}[X - \mu] = f(\mu) + f'(\mu) \cdot 0$.	Expectation preserves inequalities and is linear; $f(\mu)$ and $f'(\mu)$ are constants, and $\mathbb{E}[X - \mu] = \mu - \mu = 0$.
$f(y) \geq f(x) + f'(x)(y-x)$, and $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$	The whole of Jensen is the supporting line at the mean: the graph sits above its tangent there, and the tangent's expectation is exactly $f(\mu)$ because the linear term averages to zero. The argument never used what the distribution of X is, so it covers sums, integrals and anything in between. If $f'' > 0$ strictly, step 1's remainder is zero only when $y = x$, so the gap is zero only when $X = \mu$ with probability one: equality in Jensen for a strictly convex f means X is constant, which is what pins the ELBO's equality case to the posterior in Problem 6.

Problem 3

Without derivatives, show from the definition of convexity that $f(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i f(x_i)$ for weights $p_i \geq 0$ with $\sum_i p_i = 1$, by induction on n . Why does this version apply to $f(x) = |x|$, which has no second derivative?

$n = 1$: $p_1 = 1$ and both sides are $f(x_1)$. $n = 2$: $f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)$ with $p_2 = 1 - p_1$.	The two-point case is the definition of convexity with $\lambda = p_1$.
---	--

Suppose the claim holds for $n - 1$ points, and take n points with $p_n < 1$. Write $\sum_{i=1}^n p_i x_i = p_n x_n + (1 - p_n) \bar{x} \text{ with } \bar{x} = \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} x_i.$	Factor $1 - p_n$ out of the first $n - 1$ terms; the new weights $p_i/(1 - p_n)$ are nonnegative and sum to $(1 - p_n)/(1 - p_n) = 1$, so $\bar{x}$ is a convex combination of $x_1, \dots, x_{n-1}$. If $p_n = 1$ there is nothing to prove.
$f\left(\sum_{i=1}^n p_i x_i\right) = f(p_n x_n + (1 - p_n) \bar{x}) \leq p_n f(x_n) + (1 - p_n) f(\bar{x}).$	The two-point case with $\lambda = p_n$ applied to the points x_n and $\bar{x}$.
$f(\bar{x}) \leq \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} f(x_i), \text{ so } (1 - p_n) f(\bar{x}) \leq \sum_{i=1}^{n-1} p_i f(x_i).$	The induction hypothesis on the $n - 1$ points with the new weights; multiply through by $1 - p_n > 0$.
$f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i)$	Steps 3 and 4 together. Only the chord definition was used, so the result holds for every convex function, differentiable or not. x is convex because $\lambda x + (1 - \lambda)y \leq \lambda x + (1 - \lambda) y$ is the triangle inequality, so $\mathbb{E}[X] \leq \mathbb{E} X$ for any finite distribution, and the same chord argument gives $\ \mathbb{E}[X]\ \leq \mathbb{E}\ X\$ for vectors. The induction fails if the weights do not sum to one, which is Mistake 2.

Problem 4

From Jensen's inequality for log, prove the weighted AM–GM inequality $\prod_i a_i^{p_i} \leq \sum_i p_i a_i$ for $a_i > 0$ and weights $p_i \geq 0$, $\sum_i p_i = 1$. Evaluate both sides for $a = (1, 4)$ with equal weights and for $a = (1, 2, 4)$ with equal weights.

log is concave on $x > 0$, so $\sum_i p_i \log a_i \leq \log\left(\sum_i p_i a_i\right)$.	Problem 1 for the concavity; Problem 3's finite Jensen with the inequality reversed for a concave function.
$\sum_i p_i \log a_i = \log \prod_i a_i^{p_i}.$	$p \log a = \log a^p$ and a sum of logs is the log of the product.

$\log \prod_i a_i^{p_i} \leq \log \sum_i p_i a_i$, so $\prod_i a_i^{p_i} \leq \sum_i p_i a_i$.	exp is increasing, so it preserves the inequality.
$a = (1, 4), p = (\frac{1}{2}, \frac{1}{2})$: $\sqrt{1 \cdot 4} = 2 \leq \frac{1}{2}(1 + 4) = 2.5$.	The geometric mean is $4^{1/2}$ and the arithmetic mean is $5/2$.
$a = (1, 2, 4), p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$: $(1 \cdot 2 \cdot 4)^{1/3} = 8^{1/3} = 2 \leq \frac{1}{3}(1 + 2 + 4) = \frac{7}{3}$.	$8 = 2^3$.
$\prod_i a_i^{p_i} \leq \sum_i p_i a_i$; for $(1, 4)$: $2 \leq 2.5$; for $(1, 2, 4)$: $2 \leq 7/3$	Equality needs the a_i all equal, because log is strictly concave (Problem 2's equality case). With $n = 2$ and equal weights the inequality is $\sqrt{ab} \leq \frac{1}{2}(a + b)$, which is $(\sqrt{a} - \sqrt{b})^2 \geq 0$ rearranged. The same step from the mean of logs to the log of the mean is the one taken, in the other direction, in Problem 6.

Problem 5

Show that each of the following is a Jensen gap and compute it: (a) $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$ for any X with finite variance; (b) $\mathbb{E}[e^X] - e^{\mathbb{E}[X]}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$, by first showing $\mathbb{E}[e^X] = e^{\mu + \sigma^2/2}$; (c) $\mathbb{E}[1/X] - 1/\mathbb{E}[X]$ for X equal to 1 or 4 with probability $\frac{1}{2}$ each.

(a) x^2 is convex, so $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, and the gap is $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{Var}(X) \geq 0$.	$(x^2)'' = 2 > 0$; the variance page's $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
(b) Write $X = \mu + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$: $\mathbb{E}[e^X] = e^\mu \mathbb{E}[e^{\sigma\varepsilon}]$.	A Gaussian is a shifted and scaled standard normal; e^μ is a constant.
$\mathbb{E}[e^{\sigma\varepsilon}] = \frac{1}{\sqrt{2\pi}} \int e^{\sigma\varepsilon - \varepsilon^2/2} d\varepsilon = e^{\sigma^2/2} \frac{1}{\sqrt{2\pi}} \int e^{-(\varepsilon - \sigma)^2/2} d\varepsilon = e^{\sigma^2/2}$.	Complete the square: $\sigma\varepsilon - \frac{1}{2}\varepsilon^2 = -\frac{1}{2}(\varepsilon - \sigma)^2 + \frac{1}{2}\sigma^2$; the remaining integral is the density of $\mathcal{N}(\sigma, 1)$, which integrates to 1.

$\mathbb{E}[e^X] = e^{\mu+\sigma^2/2}$, and the gap is $e^{\mu+\sigma^2/2} - e^\mu = e^\mu(e^{\sigma^2/2} - 1) > 0$.

e^x is convex (Problem 1) and $\mathbb{E}[X] = \mu$; $e^{\sigma^2/2} > 1$ for $\sigma > 0$.

(c) $1/x$ is convex on $x > 0$ since $(1/x)'' = 2/x^3 > 0$; $\mathbb{E}[1/X] = \frac{1}{2}(1 + \frac{1}{4}) = \frac{5}{8} = 0.625$ and $1/\mathbb{E}[X] = 1/\frac{5}{2} = 0.4$.

Each value has probability $\frac{1}{2}$; $\mathbb{E}[X] = \frac{1}{2}(1 + 4) = \frac{5}{2}$.

(a) the gap is $\text{Var}(X)$; (b) $\mathbb{E}[e^X] = e^{\mu+\sigma^2/2}$ and the gap is $e^\mu(e^{\sigma^2/2} - 1)$; (c) $\mathbb{E}[1/X] = 0.625 > 0.4 = 1/\mathbb{E}[X]$, gap 0.225

Three faces of one inequality. (a) says the variance is a Jensen gap, and Problem 7 reuses it to prove a Hessian is PSD. (b) is the moment generating function of a Gaussian; it is why the mean of a log-normal variable is $e^{\mu+\sigma^2/2}$, not e^μ , and why exponentiating a predicted log-price under-estimates the expected price by the factor $e^{\sigma^2/2}$. (c) is the harmonic-mean gap: the expected reciprocal exceeds the reciprocal of the expectation, which is Mistake 3 when it is forgotten.

Problem 6

For a latent-variable model with $z \sim q$ and $w(z) = p(x, z)/q(z)$, show that $\log p(x) \geq \mathbb{E}_q[\log w(z)]$, the evidence lower bound. Then, for K independent draws $z_1, \dots, z_K \sim q$, show that $\log p(x) \geq \mathbb{E}\left[\log \frac{1}{K} \sum_{k=1}^K w(z_k)\right] \geq \mathbb{E}_q[\log w(z)]$: averaging the weights inside the log tightens the bound.

$$\mathbb{E}_q[w(z)] = \sum_z q(z) \frac{p(x, z)}{q(z)} = \sum_z p(x, z) = p(x).$$

The $q(z)$ cancels; summing the joint over z gives the marginal.

$$\log p(x) = \log \mathbb{E}_q[w] \geq \mathbb{E}_q[\log w].$$

Jensen for the concave log (Problem 1), with the inequality reversed as for every concave function.

Let $\bar{w}_K = \frac{1}{K} \sum_k w(z_k)$. Then $\mathbb{E}[\bar{w}_K] = p(x)$, so $\log p(x) = \log \mathbb{E}[\bar{w}_K] \geq \mathbb{E}[\log \bar{w}_K]$.	Each $w(z_k)$ has mean $p(x)$ by step 1 and the mean of an average is the average of the means; then Jensen again on the random variable $\bar{w}_K$.
$\log \bar{w}_K = \log \left(\frac{1}{K} \sum_k w(z_k) \right) \geq \frac{1}{K} \sum_k \log w(z_k)$ for every draw.	Problem 3's finite Jensen for the concave log with equal weights $1/K$: the log of an average is at least the average of the logs.
Taking expectations in step 4: $\mathbb{E}[\log \bar{w}_K] \geq \frac{1}{K} \sum_k \mathbb{E}[\log w(z_k)] = \mathbb{E}_q[\log w(z)]$.	Expectation preserves inequalities; the z_k are identically distributed, so each term has the same mean.
$\log p(x) \geq \mathbb{E}[\log \frac{1}{K} \sum_k w(z_k)] \geq \mathbb{E}_q[\log w(z)]$, the right-hand side being the ELBO	The ELBO is the $K = 1$ case. Equality in step 2 needs w constant under q , which by Problem 2's equality case means $q(z) = p(z x)$; the ELBO page's Problem 2 computes the gap as $\text{KL}(q \ p(z x))$. The K -sample bound is tighter because $\bar{w}_K$ is less spread than a single w (its variance is $\text{Var}_q(w)/K$), so its Jensen gap is smaller; as $K \rightarrow \infty$, $\bar{w}_K \rightarrow p(x)$ and the bound closes, whatever q is. This is the importance-weighted bound, and the check verifies both inequalities on a model small enough to enumerate every pair of draws.

Problem 7

For $\text{LSE}(z) = \log \sum_{i=1}^n e^{z_i}$, show that $\max_i z_i \leq \text{LSE}(z) \leq \max_i z_i + \log n$, that $\nabla \text{LSE}(z) = s = \text{softmax}(z)$, and that $\nabla^2 \text{LSE}(z) = \text{diag}(s) - ss^\top$ is PSD, so LSE is convex.

Let $m = \max_i z_i$. Then $e^m \leq \sum_i e^{z_i} \leq n e^m$.	The sum contains the term e^m and all terms are positive; each of the n terms is at most e^m .
--	--

$m \leq \text{LSE}(z) \leq m + \log n.$	Take logs of step 1, which preserves the order; $\log(ne^m) = \log n + m.$
$\frac{\partial \text{LSE}}{\partial z_i} = \frac{e^{z_i}}{\sum_j e^{z_j}} = s_i.$	Chain rule: the derivative of $\log u$ is u'/u , and only the i -th term of the sum depends on z_i .
$\frac{\partial s_i}{\partial z_j} = s_i(\delta_{ij} - s_j), \text{ so } \nabla^2 \text{LSE} = \text{diag}(s) - ss^\top.$	The softmax Jacobian from the softmax page; the Hessian is the Jacobian of the gradient.
$v^\top (\text{diag}(s) - ss^\top) v = \sum_i s_i v_i^2 - \left(\sum_i s_i v_i \right)^2 = \mathbb{E}_s[V^2] - (\mathbb{E}_s[V])^2 = \text{Var}_s(V) \geq 0,$ where V takes the value v_i with probability s_i.	Expand the quadratic form; s is a probability vector, so the two sums are the second moment and the squared mean of V under s , and their difference is a variance, which is Problem 5(a)'s Jensen gap.
$\max_i z_i \leq \text{LSE}(z) \leq \max_i z_i + \log n;$ $\nabla \text{LSE} = \text{softmax}(z); \nabla^2 \text{LSE} = \text{diag}(s) - ss^\top \text{ is PSD, so LSE is convex}$	Log-sum-exp is a smooth maximum, never below the true maximum and never more than $\log n$ above it, which is why cross-entropy with softmax, $\text{LSE}(z) - z_y$, is a convex function of the logits: a convex function minus a linear one. The PSD proof is Jensen's gap in disguise, and the bound of step 2 is tight at both ends: $z = (c, -\infty, \dots)$ gives m and $z = c\mathbf{1}$ gives $m + \log n$. The Hessian is only semidefinite: $v = \mathbf{1}$ gives $\text{Var}_s(V) = 0$, which Mistake 4 overlooks.

Problem 8

Show from the definition that (a) if f is convex then $g(x) = f(Ax + b)$ is convex, and (b) if $f_1, \dots, f_m$ are convex then $g(x) = \max_j f_j(x)$ is convex. Use (a) to show that the logistic loss

$\ell(w) = \log(1 + e^{-y w^\top x})$ is convex in w for fixed $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, and compute its Hessian. Use (b) for the hinge loss $h(w) = \max(0, 1 - y w^\top x)$.

(a) $g(\lambda u + (1 - \lambda)v) = f(A(\lambda u + (1 - \lambda)v) + b) = f(\lambda(Au + b) + (1 - \lambda)(Av + b))$.

An affine map sends a convex combination of points to the same convex combination of their images: $\lambda b + (1 - \lambda)b = b$.

$\leq \lambda f(Au + b) + (1 - \lambda)f(Av + b) = \lambda g(u) + (1 - \lambda)g(v)$.

Convexity of f at the points $Au + b$ and $Av + b$.

(b) $f_j(\lambda u + (1 - \lambda)v) \leq \lambda f_j(u) + (1 - \lambda)f_j(v) \leq \lambda g(u) + (1 - \lambda)g(v)$ for every j , so the maximum over j of the left side, $g(\lambda u + (1 - \lambda)v)$, obeys the same bound.

Convexity of each f_j , then $f_j \leq g$ pointwise; a bound that holds for every j holds for the largest.

$\ell(w) = \text{sp}(t)$ with $t = -y w^\top x$, and $\text{sp}''(t) = \sigma'(t) = \sigma(t)(1 - \sigma(t)) > 0$, so sp is convex and ℓ is sp composed with the affine map $w \mapsto -y w^\top x$: convex by (a).

$\text{sp}' = \sigma$ (the activation-functions page); $\sigma \in (0, 1)$ makes the product positive; (a) with $A = -y x^\top$ ($1 \times d$) and $b = 0$.

$\nabla_w \ell = \sigma(t)(-y x)$ and $\nabla_w^2 \ell = \sigma(t)(1 - \sigma(t)) x x^\top$.

Chain rule through t : $\partial t / \partial w = -y x$, and $y^2 = 1$ in the second derivative. $x x^\top$ is PSD since $v^\top x x^\top v = (x^\top v)^2 \geq 0$, in agreement with (a).

$h(w) = \max(f_1, f_2)$ with $f_1(w) = 0$ and $f_2(w) = 1 - y w^\top x$, both affine, hence convex; so h is convex by (b).

An affine function satisfies the chord inequality with equality.

$f(Ax + b)$ and $\max_j f_j(x)$ are convex when f, f_j are; $\ell(w) = \text{sp}(-y w^\top x)$ is convex with $\nabla^2 \ell = \sigma(1 - \sigma) x x^\top$, $\sigma = \sigma(-y w^\top x)$; the hinge loss is a maximum of two affine functions, hence convex

These two rules, with nonnegative sums (a sum of convex functions is convex because the chord inequalities add), build almost every convex loss in machine learning from a few one-dimensional pieces: the logistic loss is softplus after an affine map, the hinge loss and ReLU are maxima of affine maps, the squared loss is u^2 after an affine map, and a mean over a dataset is a nonnegative sum. What the rules do not include is composition of two convex functions (Mistake 1) or products.

Problem 9

Use Jensen's inequality to show that (a) $H(p) \leq \log K$ for every distribution p over K outcomes, with equality for the uniform distribution, and (b) $\text{KL}(p \parallel q) \geq 0$.

(a) $H(p) = \sum_i p_i \log \frac{1}{p_i} = \mathbb{E}_p \left[\log \frac{1}{p_I} \right]$,
where I is an outcome drawn from p .

$-\log p_i = \log(1/p_i)$; a p -weighted sum is an expectation under p .

$\mathbb{E}_p \left[\log \frac{1}{p_I} \right] \leq \log \mathbb{E}_p \left[\frac{1}{p_I} \right] = \log \sum_i p_i \cdot \frac{1}{p_i} = \log K$.

Jensen for the concave log; the expectation of $1/p_I$ is a sum of K ones.

For $p_i = 1/K$: $H(p) = \sum_i \frac{1}{K} \log K = \log K$.

Direct evaluation; $1/p_I$ is the constant K , so there is no Jensen gap.

(b) $\text{KL}(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i} = -\mathbb{E}_p \left[\log \frac{q_I}{p_I} \right]$.

Flip the fraction and pull out the sign: $\log(p/q) = -\log(q/p)$.

$-\mathbb{E}_p \left[\log \frac{q_I}{p_I} \right] \geq -\log \mathbb{E}_p \left[\frac{q_I}{p_I} \right] = -\log \sum_i p_i \frac{q_i}{p_i} = -\log \sum_i q_i = -\log 1 = 0$.

Jensen for the concave log, multiplied by -1 , which flips $\leq$ to $\geq$; the p_i cancel and q sums to one.

$H(p) \leq \log K$ **with equality for the uniform distribution**; $\text{KL}(p \parallel q) \geq 0$

Both are Jensen on log with a cleverly chosen random variable, $1/p_I$ and q_I/p_I , whose expectations under p are known constants, K and 1. Equality in (b) needs q_I/p_I constant, and a constant ratio between two distributions is 1, so $q = p$: Gibbs' inequality, which the entropy page proves instead from $\log t \leq t - 1$, the tangent line of log at $t = 1$, which is Problem 2's supporting line at the point where $\mathbb{E}[q_I/p_I]$ sits.

Problem 10

An ensemble of M models makes predictions $\hat{y}_1, \dots, \hat{y}_M$ for a target y , and the ensemble predicts the average $\bar{y} = \frac{1}{M} \sum_m \hat{y}_m$. (a) Show that for any loss $\ell(\hat{y}, y)$ that is convex in $\hat{y}$, $\ell(\bar{y}, y) \leq \frac{1}{M} \sum_m \ell(\hat{y}_m, y)$. (b) For the squared loss, show the exact identity $(\bar{y} - y)^2 = \frac{1}{M} \sum_m (\hat{y}_m - y)^2 - \frac{1}{M} \sum_m (\hat{y}_m - \bar{y})^2$.

$$(a) \ell(\bar{y}, y) = \ell\left(\sum_m \frac{1}{M} \hat{y}_m, y\right) \leq \sum_m \frac{1}{M} \ell(\hat{y}_m, y).$$

Problem 3's finite Jensen with equal weights $1/M$, applied to $\ell(\cdot, y)$ with y held fixed.

$$(b) \hat{y}_m - y = (\hat{y}_m - \bar{y}) + (\bar{y} - y), \text{ so } (\hat{y}_m - y)^2 = (\hat{y}_m - \bar{y})^2 + 2(\hat{y}_m - \bar{y})(\bar{y} - y) + (\bar{y} - y)^2.$$

Add and subtract $\bar{y}$, then expand the square.

$$\frac{1}{M} \sum_m (\hat{y}_m - y)^2 = \frac{1}{M} \sum_m (\hat{y}_m - \bar{y})^2 + 2(\bar{y} - y) \cdot \frac{1}{M} \sum_m (\hat{y}_m - \bar{y}) + (\bar{y} - y)^2.$$

Average step 2 over m ; $(\bar{y} - y)$ does not depend on m .

$$\frac{1}{M} \sum_m (\hat{y}_m - \bar{y}) = \bar{y} - \bar{y} = 0, \text{ so the cross term vanishes.}$$

The deviations from a mean sum to zero.

$$\ell(\bar{y}, y) \leq \frac{1}{M} \sum_m \ell(\hat{y}_m, y) \text{ for convex } \ell;$$

for the squared loss, $(\bar{y} - y)^2 =$

$$\frac{1}{M} \sum_m (\hat{y}_m - y)^2 - \frac{1}{M} \sum_m (\hat{y}_m - \bar{y})^2$$

The ensemble's error is the members' average error minus the members' spread, the ambiguity decomposition: averaging can only help, and it helps most when the members disagree. (b) is Problem 5(a) applied across models, with the spread term the variance of the predictions, so the Jensen gap for the squared loss is exactly that variance. For the log loss $-\log \hat{p}_y$ with probabilities, (a) says the averaged probability scores at least as well on average as its members; the inequality is strict unless all members agree. This is the inequality behind bagging and model averaging, and it says nothing about any single member, which may be better than the ensemble; the bias–variance page quantifies how much averaging reduces variance.

Where this goes wrong

1. Composing two convex functions

Convexity survives affine maps inside and nonnegative sums and maxima outside, and the natural next guess is that it survives any composition.

$g(x) = x^2 - 1$ and $f(u) = u^2$ are both convex

Right so far: both have positive second derivatives.

“A convex function of a convex function is convex.”

The rule that causes the mistake: Problem 8(a) allows an affine inner function, not a convex one.

$f(g(x)) = (x^2 - 1)^2$ is convex ✗

Its second derivative is $12x^2 - 4$, negative on $|x| < 1/\sqrt{3}$, and it has two minima at $x = \pm 1$ with a local maximum at 0: the chord from $x = -1$ to $x = 1$ runs along 0 while the graph rises to 1 between them. Composition preserves convexity only when the outer function is also nondecreasing, which u^2 is not on $u < 0$; that is why $\text{sp}(\text{affine})$ in Problem 8 is safe (the inner map is affine) and why $\sigma(w^\top x)$ is not convex (sigmoid is neither convex nor monotone-convex), so the squared error of a logistic unit is not a convex function of w .

2. Jensen applied to a sum instead of an average

The finite inequality says f of a combination is at most the combination of f , and the weights are easy to forget.

$$f\left(\sum_i p_i x_i\right) \leq \sum_i p_i f(x_i) \text{ for } p_i \geq 0, \\ \sum_i p_i = 1$$

Right so far: Problem 3.

“The p_i are just positive weights, so any positive weights will do.”

The habit that causes the mistake: Problem 3's induction needs $\sum_i p_i = 1$ at every stage; with p_n and the rescaled $p_i/(1 - p_n)$, the normalisation is what makes $\bar{x}$ a point the chord inequality can be applied to.

$$f(x + y) \leq f(x) + f(y) \text{ for convex } f \text{ } \times$$

With $f(x) = x^2$ and $x = y = 1$: $f(2) = 4 > 2 = f(1) + f(1)$. For $f(x) = x^2$ the true statement is $f(x + y) = f(2 \cdot \frac{x+y}{2}) = 4f(\frac{x+y}{2}) \leq 2f(x) + 2f(y)$: the inequality holds only after the sum is rewritten as a multiple of an average. The fix is the same in every case: pull out the total weight, apply Jensen to the normalised combination, and keep the factor. LSE in Problem 7 is the one place a sum appears without weights, and there the inequality that survives is the bound $\text{LSE}(z) \leq \max_i z_i + \log n$, not a Jensen inequality.

3. The expected reciprocal taken to be the reciprocal of the expectation

The mean of a ratio comes up constantly: a mean rate, a mean waiting time, a mean of importance weights, and the shortcut takes expectations inside.

$\mathbb{E}[X] = \frac{5}{2}$ for $X \in \{1, 4\}$ with equal probability

Right so far: Problem 5(c).

“Expectation is linear, so it passes through $1/X$.”

The habit that causes the mistake: linearity passes expectation through $aX + b$, and $1/x$ is not linear.

$$\mathbb{E}[1/X] = 1/\mathbb{E}[X] = 0.4 \text{ } \times$$

$\mathbb{E}[1/X] = 0.625$ (Problem 5), and the direction is forced by Jensen: $1/x$ is convex on $x > 0$, so $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ always, with equality only for a constant X . The same error in the other direction treats $\mathbb{E}[\log X]$ as $\log \mathbb{E}[X]$, which would make the ELBO equal to $\log p(x)$ for every q (Problem 6), and treats $\mathbb{E}[e^X]$ as $e^{\mathbb{E}[X]}$, which drops the $e^{\sigma^2/2}$ of Problem 5(b). The estimators page on importance sampling meets the first form again in the self-normalised estimator, whose denominator is a mean of weights.

4. Calling log-sum-exp strictly convex

The Hessian $\text{diag}(s) - ss^\top$ has positive diagonal entries, and the quadratic form is a variance, which feels positive.

$v^\top \nabla^2 \text{LSE}(z) v = \text{Var}_s(V) \geq 0$	Right so far: Problem 7, step 5.
“A variance is positive, so the Hessian is positive definite and LSE is strictly convex.”	The shortcut that causes the mistake: a variance is zero when the variable is constant.
$\nabla^2 \text{LSE}(z) \succ 0$, so LSE is strictly convex and cross-entropy with softmax has a unique minimising logit vector $\mathbf{x}$	$v = \mathbf{1}$ gives V constant and $\mathbf{1}^\top (\text{diag}(s) - ss^\top) \mathbf{1} = 1 - 1 = 0$, so $\mathbf{1}$ is in the Hessian's null space at every z : along $\mathbf{1}$ the function is linear, $\text{LSE}(z + c\mathbf{1}) = \text{LSE}(z) + c$, and the softmax does not change at all (the softmax page's shift invariance). Cross-entropy $\text{LSE}(z) - z_y$ is constant along $\mathbf{1}$, so every minimiser comes with a whole line of them, and in a linear model $z = Wx$ the Hessian in W is singular for the same reason. The positive-definite-matrices page has the general test: PSD with a null vector is not PD, however positive the diagonal looks.

5. Checking the second derivative at one point

The second-derivative test for a minimum is local, and the same test is reached for when convexity is wanted.

$f(x) = x^4 - x^2$ has $f''(x) = 12x^2 - 2$, and $f''(1) = 10 > 0$	Right so far: a correct second derivative.
“ $f'' > 0$, so f is convex.”	The habit that causes the mistake: Problem 1's test asks for $f'' \geq 0$ on the whole interval, and Problem 2's tangent-line argument needs $f''(\xi) \geq 0$ at an unknown point ξ between x and y .

f is convex, so $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ for every X ✗

$f''(0) = -2 < 0$: the graph is a double well with minima at $x = \pm 1/\sqrt{2}$ and a local maximum at 0. Take $X = \pm 1/\sqrt{2}$ with probability $\frac{1}{2}$ each: $\mathbb{E}[X] = 0$, $f(0) = 0$, but $\mathbb{E}[f(X)] = -\frac{1}{4}$, so the claimed inequality reads $0 \leq -\frac{1}{4}$. One positive second derivative certifies a local minimum at a critical point, not convexity; convexity is a statement about every chord, and it fails as soon as f'' is negative anywhere between two points. The same error in several variables accepts a PSD Hessian at the current iterate as proof that the loss is convex, which for any neural network with a hidden layer it is not.