Maximum likelihood estimation: derivations by hand

Ten problems on maximum likelihood estimation: the Bernoulli, categorical, Poisson and Gaussian MLEs by hand, the curvature of the log-likelihood, why the variance MLE divides by N and is biased, least squares and cross-entropy as negative log-likelihoods, MAP with a Gaussian or Laplace prior as ridge and lasso with the exact λ, and the uniform MLE that no derivative finds, with worked solutions and the mistakes that maximize the wrong thing.

BEFORE YOU START

Almost every loss in machine learning is a negative log-likelihood in disguise. Least squares is the likelihood of Gaussian noise, binary and softmax cross-entropy are the likelihoods of Bernoulli and categorical labels, and weight decay is a Gaussian prior on the weights. These ten problems derive the standard maximum likelihood estimates by hand, check that each stationary point really is a maximum, work out what the estimates get wrong on average, and then turn the regression losses of the previous pages back into the probability models they came from, priors included. The five mistakes at the end are the ones that give a plausible number: the unbiased variance passed off as the maximum likelihood one, the exponential distribution's answer used for the Poisson, a constraint forgotten, a prior strength off by a factor of N, and a moment match used where the derivative has no zero.

- The data $x_1, \dots, x_N$ are independent draws from a distribution with density or probability mass $p(x; \theta)$, where θ is the parameter to estimate.
- The likelihood is $L(\theta) = \prod_n p(x_n; \theta)$, the probability of the data as a function of θ. The log-likelihood is $\ell(\theta) = \log L(\theta) = \sum_n \log p(x_n; \theta)$, with the natural log. Because log is strictly increasing, L and ℓ have the same maximizer, and the sum is easier to differentiate than the product.
- The maximum likelihood estimate (MLE) is $\hat{\theta} = \arg\max_\theta \ell(\theta)$. A zero derivative only finds a stationary point; a negative second derivative (or a concave ℓ) is what makes it a maximum.
- $\bar{x} = \frac{1}{N} \sum_n x_n$ is the sample mean.
- $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian with mean μ and variance σ^2, density $(2\pi\sigma^2)^{-1/2} \exp\big(-(x-\mu)^2/(2\sigma^2)\big)$.
- $\mathbb{E}$ is the expectation over a random sample. An estimator $\hat{\theta}$ is unbiased if $\mathbb{E}[\hat{\theta}] = \theta$, and its bias is $\mathbb{E}[\hat{\theta}] - \theta$. For independent variables the variance of a sum is the sum of the variances, and $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$.

- Regression uses the notation of the regression page: $X \in \mathbb{R}^{N \times d}$ has row n equal to $x^{(n)\top}$, $y \in \mathbb{R}^N$ holds the targets, $w \in \mathbb{R}^d$ the weights, $\|v\|^2 = \sum_j v_j^2$ and $\|v\|_1 = \sum_j |v_j|$.
- Bayes' rule for the weights: $p(w \mid y) = p(y \mid w)\, p(w) / p(y)$, where $p(w)$ is the prior. The maximum a posteriori (MAP) estimate maximizes $\log p(y \mid w) + \log p(w)$; $p(y)$ does not depend on w.

Builds on: Regression gradients: linear, logistic and softmax

Problems

1. A coin lands heads with probability θ. In N independent flips, $x_n = 1$ for heads and 0 for tails, and $k = \sum_n x_n$ is the number of heads, with $0 < k < N$. Find the MLE $\hat{\theta}$, show it is a maximum, and compute $\ell''(\hat{\theta})$. difficulty ··

2. A categorical variable takes one of K values, value i with probability π_i, where $\pi_i > 0$ and $\sum_i \pi_i = 1$. In N independent draws, value i appears $n_i > 0$ times. Find the MLE $\hat{\pi}$ with a Lagrange multiplier. difficulty ··

3. Counts $x_n \in \{0, 1, 2, \dots\}$ are independent Poisson with rate $\lambda > 0$: $P(x; \lambda) = \lambda^x e^{-\lambda} / x!$. Find the MLE $\hat{\lambda}$, assuming not every count is 0. difficulty ·

4. Let $x_1, \dots, x_N$ be independent draws from $\mathcal{N}(\mu, \sigma^2)$, with $N \geq 2$ and not all x_n equal. Find the joint MLE $(\hat{\mu}, \hat{\sigma}^2)$ and show it is a maximum. difficulty ··

5. The x_n are independent with mean μ and variance σ^2 (no Gaussian assumption needed). Show that the variance MLE of Problem 4 has $\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$. difficulty ···

6. Linear regression with Gaussian noise: $y_n = x^{(n)\top} w + \varepsilon_n$ with the ε_n independent $\mathcal{N}(0, \sigma^2)$. Write the negative log-likelihood, and find the MLE of w and σ^2 when the columns of X are independent. difficulty ··

7. Logistic regression models $P(y_n = 1 \mid x^{(n)}) = \sigma(x^{(n)\top} w)$, with labels $y_n \in \{0, 1\}$ independent given the inputs and $\sigma(t) = 1/(1 + e^{-t})$. Show that $-\ell(w)/N$ is the mean binary cross-entropy $L(w)$ of the regression page. difficulty ·

8. Keep the model of Problem 6 with σ^2 known, and give the weights independent Gaussian priors $w_j \sim \mathcal{N}(0, \tau^2)$. Show that the MAP estimate minimizes the mean ridge loss $\frac{1}{N}\|Xw - y\|^2 + \lambda\|w\|^2$, find λ, and write w_{MAP}. difficulty ···

9. Same model, but with independent Laplace priors $p(w_j) = \frac{1}{2b}e^{-|w_j|/b}$, $b > 0$. Show that the MAP estimate minimizes $\frac{1}{N}\|Xw - y\|^2 + \lambda\|w\|_1$ and find λ. difficulty ··

10. The x_n are independent and uniform on $[0, \theta]$, with $\theta > 0$ unknown and $p(x;\theta) = 1/\theta$ for $0 \le x \le \theta$, 0 otherwise. Find the MLE $\hat{\theta}$, explain why setting a derivative to zero cannot find it, and compute $\mathbb{E}[\hat{\theta}]$. difficulty ···

Answers

1. $\hat{\theta} = k/N$, and $\ell''(\hat{\theta}) = -\dfrac{N}{\hat{\theta}(1-\hat{\theta})}$
2. $\hat{\pi}_i = n_i / N$
3. $\hat{\lambda} = \bar{x}$
4. $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{N}\sum_n (x_n - \bar{x})^2$
5. $\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$
6. $\hat{w} = (X^\top X)^{-1} X^\top y$ and $\hat{\sigma}^2 = \frac{1}{N}\|X\hat{w} - y\|^2$
7. $-\ell(w)/N = -\frac{1}{N}\sum_n \left[y_n \log p_n + (1-y_n)\log(1-p_n) \right] = L(w)$
8. $\lambda = \dfrac{\sigma^2}{N\tau^2}$ and $w_{\text{MAP}} = \left(X^\top X + (\sigma^2/\tau^2)\, I\right)^{-1} X^\top y$
9. $\lambda = \dfrac{2\sigma^2}{Nb}$: the MAP estimate under a Laplace prior is the lasso
10. $\hat{\theta} = \max_n x_n$, with $\mathbb{E}[\hat{\theta}] = \dfrac{N\theta}{N+1}$

Worked solutions

Problem 1

> A coin lands heads with probability θ. In N independent flips, $x_n = 1$ for heads and 0 for tails, and $k = \sum_n x_n$ is the number of heads, with $0 < k < N$. Find the MLE $\hat\theta$, show it is a maximum, and compute $\ell''(\hat\theta)$.

$p(x;\theta) = \theta^x (1-\theta)^{1-x}$, so $\ell(\theta) = \sum_n \big[x_n \log\theta + (1-x_n)\log(1-\theta)\big] = k\log\theta + (N-k)\log(1-\theta)$.	One formula covers both outcomes: $x = 1$ leaves θ and $x = 0$ leaves $1-\theta$. The log of the product is the sum of the logs, and $\sum_n x_n = k$, $\sum_n (1-x_n) = N-k$.
$\ell'(\theta) = \dfrac{k}{\theta} - \dfrac{N-k}{1-\theta}.$	$\log u$ differentiates to $1/u$; the chain rule on $\log(1-\theta)$ supplies the minus sign.
$\ell'(\theta) = 0 \iff k(1-\theta) = (N-k)\,\theta \iff \theta = k/N.$	Multiply by $\theta(1-\theta) > 0$, which does not change where the expression is zero; the $k\theta$ terms cancel.
$\ell''(\theta) = -\dfrac{k}{\theta^2} - \dfrac{N-k}{(1-\theta)^2} < 0$ for every $\theta \in (0,1)$.	Both terms are negative because $0 < k < N$. So ℓ is strictly concave on $(0,1)$ and its one stationary point is the global maximum.
At $\hat\theta = k/N$: $\dfrac{k}{\hat\theta^2} = \dfrac{N}{\hat\theta}$ and $\dfrac{N-k}{(1-\hat\theta)^2} = \dfrac{N}{1-\hat\theta}$, whose sum is $\dfrac{N}{\hat\theta(1-\hat\theta)}$.	Substitute $k = N\hat\theta$ and $N-k = N(1-\hat\theta)$; over the common denominator $\hat\theta(1-\hat\theta)$ the numerator is $N(1-\hat\theta) + N\hat\theta = N$.
$\hat\theta = k/N$, **and** $\ell''(\hat\theta) = -\dfrac{N}{\hat\theta(1-\hat\theta)}$	The curvature grows in proportion to N: more flips make a sharper peak. Its reciprocal with the sign flipped, $\hat\theta(1-\hat\theta)/N$, is the usual estimate of the variance of $\hat\theta$, and here it is exactly $\mathrm{Var}(k/N) = \theta(1-\theta)/N$ with θ replaced by $\hat\theta$. If $k = 0$ or $k = N$, ℓ is monotonic and the maximum sits at the boundary 0 or 1, which is still k/N but is not a stationary point.

Problem 2

> A categorical variable takes one of K values, value i with probability π_i, where $\pi_i > 0$ and $\sum_i \pi_i = 1$. In N independent draws, value i appears $n_i > 0$ times. Find the MLE $\hat{\pi}$ with a Lagrange multiplier.

$\ell(\pi) = \sum_{i=1}^{K} n_i \log \pi_i$.	Each draw of value i contributes a factor π_i to the likelihood, and there are n_i of them.
$\mathcal{L}(\pi, \nu) = \sum_i n_i \log \pi_i + \nu\big(1 - \sum_i \pi_i\big)$, with multiplier ν.	The π_i are not free: on its own ℓ increases in every π_i and has no stationary point (Mistake 3). The multiplier term enforces $\sum_i \pi_i = 1$.
$\partial\mathcal{L}/\partial\pi_i = n_i/\pi_i - \nu = 0$, so $\pi_i = n_i/\nu$.	Only the i-th terms of the two sums contain π_i.
$\sum_i n_i/\nu = 1$, so $\nu = \sum_i n_i = N$.	Substitute step 3 into the constraint.
ℓ is a sum of concave functions $n_i \log \pi_i$, and the set $\{\pi : \sum_i \pi_i = 1,\ \pi_i > 0\}$ is convex.	A stationary point of the Lagrangian of a concave function under a linear equality constraint is the constrained global maximum.
$\hat{\pi}_i = n_i/N$	The estimate is the observed frequency. At $\hat{\pi}$ every partial derivative $n_i/\hat{\pi}_i$ equals N, so $\nabla\ell$ is parallel to $(1, \dots, 1)$: no move that keeps the sum at 1 can increase ℓ to first order. With $K = 2$ this is Problem 1.

Problem 3

> Counts $x_n \in \{0, 1, 2, \dots\}$ are independent Poisson with rate $\lambda > 0$: $P(x; \lambda) = \lambda^x e^{-\lambda}/x!$. Find the MLE $\hat{\lambda}$, assuming not every count is 0.

$\log P(x; \lambda) = x \log \lambda - \lambda - \log x!$, so $\ell(\lambda) = \big(\sum_n x_n\big) \log \lambda - N\lambda - \sum_n \log x_n!$.	The log turns the power into a product and the quotient into a difference; the last sum does not contain λ.

$\ell'(\lambda) = \dfrac{\sum_n x_n}{\lambda} - N = 0 \iff \lambda = \frac{1}{N}\sum_n x_n.$	The constant $\sum_n \log x_n!$ differentiates to 0.
$\ell''(\lambda) = -\dfrac{\sum_n x_n}{\lambda^2} < 0.$	$\sum_n x_n > 0$ by assumption, so ℓ is strictly concave and the stationary point is the maximum.
$\hat{\lambda} = \bar{x}$	The mean of a Poisson variable is λ, so the MLE is the sample mean. The exponential distribution, which models the waiting times between Poisson events, also has a parameter called a rate, but its mean is $1/\lambda$ and its MLE is $1/\bar{x}$ (Mistake 2).

Problem 4

> Let $x_1, \ldots, x_N$ be independent draws from $\mathcal{N}(\mu, \sigma^2)$, with $N \geq 2$ and not all x_n equal. Find the joint MLE $(\hat{\mu}, \hat{\sigma}^2)$ and show it is a maximum.

$\ell(\mu, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \dfrac{1}{2\sigma^2}\sum_n (x_n - \mu)^2.$	The log of the Gaussian density, summed over the data.
$\dfrac{\partial \ell}{\partial \mu} = \dfrac{1}{\sigma^2}\sum_n (x_n - \mu) = 0 \iff \mu = \bar{x},$ whatever σ^2 is.	$(x_n - \mu)^2$ differentiates to $-2(x_n - \mu)$, and the -2 cancels the $-\frac{1}{2}$. Since $\sum_n (x_n - \mu) = N\bar{x} - N\mu$, the zero does not depend on σ^2.
With $v = \sigma^2$ as the variable: $\dfrac{\partial \ell}{\partial v} = -\dfrac{N}{2v} + \dfrac{1}{2v^2}\sum_n (x_n - \mu)^2 = 0 \iff v = \frac{1}{N}\sum_n (x_n - \mu)^2.$	$\log(2\pi v)$ differentiates to $1/v$ and $1/v$ to $-1/v^2$. Multiply by $2v^2 > 0$ to solve.
Put $\mu = \bar{x}$ from step 2 into step 3.	The joint stationary point has to satisfy both equations, and step 2 fixes μ independently of v.

For each v, ℓ is a concave quadratic in μ with its peak at $\bar{x}$. With $S = \sum_n (x_n - \bar{x})^2 > 0$, the profile $\ell(\bar{x}, v) = -\frac{N}{2}\log(2\pi v) - S/(2v)$ has derivative $(S - Nv)/(2v^2)$, positive for $v < S/N$ and negative beyond.	So $\ell(\mu, v) \le \ell(\bar{x}, v) \le \ell(\bar{x}, S/N)$ for every (μ, v): the stationary point is the global maximum. $S > 0$ because the x_n are not all equal.
$\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{N}\sum_n (x_n - \bar{x})^2$	The variance estimate divides by N, not $N - 1$; Problem 5 shows what that costs. Differentiating with respect to σ instead of σ^2 gives the same answer, because the maximizer does not depend on how the parameter is written.

Problem 5

> The x_n are independent with mean μ and variance σ^2 (no Gaussian assumption needed). Show that the variance MLE of Problem 4 has $\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$.

$\sum_n (x_n - \bar{x})^2 = \sum_n (x_n - \mu)^2 - N(\bar{x} - \mu)^2$.	Write $x_n - \bar{x} = (x_n - \mu) - (\bar{x} - \mu)$ and expand. The cross term is $-2(\bar{x} - \mu)\sum_n (x_n - \mu) = -2N(\bar{x} - \mu)^2$, because $\sum_n (x_n - \mu) = N(\bar{x} - \mu)$; adding the $N(\bar{x} - \mu)^2$ from the squares leaves $-N(\bar{x} - \mu)^2$.
$\mathbb{E}\big[\sum_n (x_n - \mu)^2\big] = N\sigma^2$.	Each term has expectation $\mathrm{Var}(x_n) = \sigma^2$, and expectation is linear.
$\mathbb{E}\big[(\bar{x} - \mu)^2\big] = \mathrm{Var}(\bar{x}) = \sigma^2/N$.	$\mathbb{E}[\bar{x}] = \mu$, so this is the variance of $\bar{x}$. By independence $\mathrm{Var}\left(\sum_n x_n\right) = N\sigma^2$, and the factor $1/N$ divides that by N^2.
$\mathbb{E}\big[\sum_n (x_n - \bar{x})^2\big] = N\sigma^2 - N \cdot \sigma^2/N = (N-1)\sigma^2$.	Steps 1 to 3, by linearity of expectation.

$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$	Divide step 4 by N. The MLE is biased low: $\bar{x}$ is fitted to the same data, and it minimizes $\sum_n (x_n - c)^2$ over c, so the residuals about $\bar{x}$ are smaller than those about the true μ. Dividing by $N-1$ instead gives the unbiased sample variance; the two agree as $N \to \infty$.

Problem 6

> Linear regression with Gaussian noise: $y_n = x^{(n)\top} w + \varepsilon_n$ with the ε_n independent $\mathcal{N}(0, \sigma^2)$. Write the negative log-likelihood, and find the MLE of w and σ^2 when the columns of X are independent.

$y_n \sim \mathcal{N}\big(x^{(n)\top} w, \sigma^2\big).$	y_n is the constant $x^{(n)\top} w$ plus Gaussian noise, and adding a constant shifts the mean without changing the variance.
$-\ell(w, \sigma^2) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\lVert Xw - y\rVert^2.$	Problem 4, step 1, with μ replaced by $x^{(n)\top} w$ for example n; $\sum_n (y_n - x^{(n)\top} w)^2 = \lVert Xw - y\rVert^2$ because entry n of Xw is $x^{(n)\top} w$.
For every fixed σ^2, w enters only through $\lVert Xw - y\rVert^2$ with the positive coefficient $1/(2\sigma^2)$, so $\hat{w}$ minimizes $\lVert Xw - y\rVert^2$: $\hat{w} = (X^\top X)^{-1} X^\top y$.	Minimizing a positive multiple of a function has the same minimizer; the least-squares solution is the regression page, Problem 2. The noise level does not affect $\hat{w}$.
With $w = \hat{w}$, setting $\partial(-\ell)/\partial\sigma^2 = 0$ gives $\sigma^2 = \frac{1}{N}\lVert X\hat{w} - y\rVert^2$.	This is Problem 4, step 3, with the residuals $y_n - x^{(n)\top}\hat{w}$ in place of $x_n - \mu$.
$\hat{w} = (X^\top X)^{-1} X^\top y$ **and** $\hat{\sigma}^2 = \frac{1}{N}\lVert X\hat{w} - y\rVert^2$	Least squares is maximum likelihood under Gaussian noise, and the mean squared error at the optimum is the noise-variance MLE. Like Problem 5, it is biased low, now because d parameters were fitted: dividing by $N - d$ removes the bias.

Problem 7

Logistic regression models $P(y_n = 1 \mid x^{(n)}) = \sigma(x^{(n)\top} w)$, with labels $y_n \in \{0, 1\}$ independent given the inputs and $\sigma(t) = 1/(1 + e^{-t})$. Show that $-\ell(w)/N$ is the mean binary cross-entropy $L(w)$ of the regression page.

$P(y_n \mid x^{(n)}; w) = p_n^{y_n}(1 - p_n)^{1 - y_n}$ with $p_n = \sigma(x^{(n)\top} w)$.	The Bernoulli of Problem 1, with its probability now different for each example.
$\ell(w) = \sum_n \big[y_n \log p_n + (1 - y_n) \log(1 - p_n) \big]$.	The log of the product of the step 1 factors, as in Problem 1, step 1.
$-\ell(w)/N = -\frac{1}{N} \sum_n \big[y_n \log p_n + (1 - y_n) \log(1 - p_n) \big] = L(w)$	So minimizing the mean binary cross-entropy is maximum likelihood, and its gradient $\frac{1}{N} X^\top (\sigma(Xw) - y)$ (regression page, Problem 6) is the negative score divided by N. The $1/N$ scales the loss without moving the minimizer. In the same way, softmax cross-entropy is the negative log-likelihood of Problem 2's categorical distribution with per-example probabilities.

Problem 8

Keep the model of Problem 6 with σ^2 known, and give the weights independent Gaussian priors $w_j \sim \mathcal{N}(0, \tau^2)$. Show that the MAP estimate minimizes the mean ridge loss $\frac{1}{N}\|Xw - y\|^2 + \lambda \|w\|^2$, find λ, and write w_{MAP}.

$\log p(w \mid y) = \log p(y \mid w) + \log p(w) - \log p(y)$.	Bayes' rule; $\log p(y)$ does not depend on w.
$\log p(w) = \sum_j \big[-\frac{1}{2} \log(2\pi\tau^2) - w_j^2/(2\tau^2) \big] = \text{const} - \|w\|^2/(2\tau^2)$.	The prior is a product over j, so its log is a sum of Gaussian log-densities with mean 0.
$-\log p(w \mid y) = \dfrac{1}{2\sigma^2}\|Xw - y\|^2 + \dfrac{1}{2\tau^2}\|w\|^2 + c$, with c free of w.	Problem 6, step 2, for $-\log p(y \mid w)$; the $\log(2\pi\sigma^2)$ and $\log(2\pi\tau^2)$ terms and $\log p(y)$ go into c.

Multiplying by $2\sigma^2/N > 0$: $\frac{1}{N}\|Xw - y\|^2 + \frac{\sigma^2}{N\tau^2}\|w\|^2 + c'$.	A positive multiple has the same minimizer, and this one makes the data term a mean.
The mean ridge loss with λ has minimizer $(X^\top X + N\lambda I)^{-1}X^\top y$, and $N\lambda = \sigma^2/\tau^2$.	Regression page, Problem 3, which exists for every X.
$\lambda = \dfrac{\sigma^2}{N\tau^2}$ **and** $w_{\text{MAP}} = \left(X^\top X + (\sigma^2/\tau^2)\, I\right)^{-1} X^\top y$	Weight decay is a Gaussian prior. A wide prior ($\tau \to \infty$) gives $\lambda \to 0$ and the MLE of Problem 6. For the mean loss, λ shrinks as N grows: the prior's pull is fixed while the evidence accumulates.

Problem 9

> Same model, but with independent Laplace priors $p(w_j) = \frac{1}{2b}e^{-|w_j|/b}$, $b > 0$. Show that the MAP estimate minimizes $\frac{1}{N}\|Xw - y\|^2 + \lambda\|w\|_1$ and find λ.

$\log p(w) = -d\log(2b) - \|w\|_1/b.$	The log of a product of d Laplace densities: each contributes $-\log(2b) - \|w_j\|/b$.
$-\log p(w \mid y) = \dfrac{1}{2\sigma^2}\|Xw - y\|^2 + \dfrac{1}{b}\|w\|_1 + c.$	Problem 8, steps 1 and 3, with this prior in place of the Gaussian one; constants go into c.
Multiplying by $2\sigma^2/N$: $\frac{1}{N}\|Xw - y\|^2 + \frac{2\sigma^2}{Nb}\|w\|_1 + c'$.	The same positive rescaling as Problem 8, step 4.
$\lambda = \dfrac{2\sigma^2}{Nb}$**: the MAP estimate under a Laplace prior is the lasso**	The factor is $2\sigma^2$ here and σ^2 in Problem 8 because the Gaussian prior's exponent carries a $\frac{1}{2}$ and the Laplace prior's does not. $\|w_j\|$ has a corner at 0, so there is no closed form in general; the corner in the prior is also what lets the lasso set weights exactly to zero.

Problem 10

> The x_n are independent and uniform on $[0, \theta]$, with $\theta > 0$ unknown and $p(x; \theta) = 1/\theta$ for $0 \le x \le \theta$, 0 otherwise. Find the MLE $\hat{\theta}$, explain why setting a derivative to zero cannot find it, and compute $\mathbb{E}[\hat{\theta}]$.

$L(\theta) = \prod_n p(x_n; \theta) = \theta^{-N}$ if $\theta \ge m$, and 0 if $\theta < m$, where $m = \max_n x_n$.	One factor is 0 as soon as any x_n lies above θ, and every x_n lies in $[0, \theta]$ exactly when the largest one does.
$L'(\theta) = -N\theta^{-N-1} < 0$ for every $\theta > m$, and $L = 0$ below m.	Above m the likelihood strictly decreases, so no θ has a zero derivative; below m it is zero, the smallest possible value.
$\hat{\theta} = m$.	L jumps from 0 to m^{-N} at m and decreases afterwards, so its maximum is the left end of the region where it is positive. The maximum is at a discontinuity, where no derivative exists: the calculus recipe assumes a maximum in the interior of a smooth region.
$P(\hat{\theta} \le t) = P(\text{every } x_n \le t) = (t/\theta)^N$ for $0 \le t \le \theta$.	The maximum is at most t exactly when every point is, and by independence the N probabilities t/θ multiply.
$\mathbb{E}[\hat{\theta}] = \int_0^\theta t \cdot \dfrac{Nt^{N-1}}{\theta^N}\, dt = \dfrac{N}{\theta^N} \cdot \dfrac{\theta^{N+1}}{N+1} = \dfrac{N\theta}{N+1}$.	Differentiating step 4 in t gives the density Nt^{N-1}/θ^N; then integrate t^N.
$\hat{\theta} = \max_n x_n$, **with** $\mathbb{E}[\hat{\theta}] = \dfrac{N\theta}{N+1}$	The MLE can never exceed θ, so it is biased low, by $\theta/(N+1)$; $\frac{N+1}{N} \max_n x_n$ is unbiased. Problem 1 noted a boundary maximum at $k = 0$; here the maximum is always at a boundary.

Where this goes wrong

1. Variance MLE divided by N − 1

`pandas`, spreadsheet `VAR` functions and most statistics courses report the sample variance with $N - 1$ in the denominator, and that version is the one that sticks.

$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n (x_n - \bar{x})^2$	Right so far: Problem 4, step 3, at $\mu = \bar{x}$.
"The correct estimate of a variance divides by $N - 1$, so the maximum likelihood one must too."	The shortcut that causes the mistake: "correct" there means unbiased (Problem 5), which is a different criterion from maximizing the likelihood.
$\hat{\sigma}^2 = \frac{1}{N-1}\sum_n (x_n - \bar{x})^2$ ✗	At that value the derivative in step 1 is $-1/(2\sigma^2)$, not 0, so it is not the maximizer: the MLE divides by N (Problem 4). NumPy's `np.var` divides by N by default and `pandas` by $N - 1$, so the same data give two answers in one notebook.

2. Poisson rate estimated as one over the mean

The exponential and the Poisson distributions both have a parameter called the rate λ, and the exponential's MLE is $1/\bar{x}$.

$\ell(\lambda) = \left(\sum_n x_n\right) \log \lambda - N\lambda + \text{const}$	Right so far: Problem 3, step 1.
"A rate is events per unit time, the reciprocal of the mean, so $\hat{\lambda} = 1/\bar{x}$."	The analogy that causes the mistake: an exponential variable is a waiting time with mean $1/\lambda$; a Poisson variable is a count with mean λ.
$\hat{\lambda} = 1/\bar{x}$ ✗	At that value $\ell'(\lambda) = N\bar{x}^2 - N$, which is zero only when $\bar{x} = 1$. The MLE is $\bar{x}$ (Problem 3): counts averaging 4 per hour give $\hat{\lambda} = 4$ per hour, not 0.25.

3. Categorical log-likelihood made stationary without the constraint

Problem 1 had a single free parameter, and with K of them it is natural to set each partial derivative to zero in turn.

$\ell(\pi) = \sum_i n_i \log \pi_i$	Right so far: Problem 2, step 1.
"Set $\partial \ell / \partial \pi_i = 0$ for every i, as for any maximum of a function of several variables."	The shortcut that causes the mistake: the unconstrained condition applied to parameters that must sum to 1.
$\partial \ell / \partial \pi_i = n_i / \pi_i = 0$ ✗	This has no solution: $n_i / \pi_i > 0$, so ℓ increases in every π_i and, with nothing to stop it, every probability heads to 1. The constraint is what makes the classes compete for probability; with a multiplier the answer is $\hat{\pi}_i = n_i / N$ (Problem 2).

4. Ridge strength set to σ²/τ² in a mean loss

For the summed loss the textbook correspondence is $\lambda = \sigma^2 / \tau^2$, and libraries document their loss as a mean.

$-\log p(w \mid y) = \frac{1}{2\sigma^2} \lVert Xw - y \rVert^2 + \frac{1}{2\tau^2} \lVert w \rVert^2 + c$	Right so far: Problem 8, step 3.
"Multiply by $2\sigma^2$: the penalty weight is σ^2 / τ^2."	The shortcut that causes the mistake: that is right for $\lVert Xw - y \rVert^2 + \lambda \lVert w \rVert^2$, and the λ is then carried to a loss whose data term has a $1/N$.
$\frac{1}{N} \lVert Xw - y \rVert^2 + \frac{\sigma^2}{\tau^2} \lVert w \rVert^2$ has the MAP estimate as its minimizer ✗	Multiplying step 1 by $2\sigma^2 / N$ gives $\lambda = \sigma^2 / (N\tau^2)$ (Problem 8). With σ^2 / τ^2 the prior is N times too strong, and the more data there are, the further the estimate is from the posterior mode.

5. Uniform MLE taken as twice the mean

The derivative of θ^{-N} never vanishes, and matching the mean $\theta/2$ to $\bar{x}$ looks like a reasonable way to get an answer anyway.

$L(\theta) = \theta^{-N}$ for $\theta \geq \max_n x_n$, and 0 otherwise	Right so far: Problem 10, step 1.
"There is no stationary point, so match the mean instead: $\theta/2 = \bar{x}$."	The shortcut that causes the mistake: this is the method of moments, a different estimator with no claim to maximize L.
$\hat{\theta} = 2\bar{x}$ ✗	For the data $0.2, 0.3, 1.9$, $2\bar{x} = 1.6$, below the observation 1.9, so $L(1.6) = 0$: the estimate says the data were impossible. The MLE is $\max_n x_n = 1.9$ (Problem 10), and even when $2\bar{x}$ exceeds the maximum it has a smaller likelihood, because L decreases there.

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 17 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/maximum-likelihood-estimation