Bayesian inference: conjugate priors, posterior means and MAP estimates

Ten problems on Bayesian inference for a continuous parameter: a Beta posterior from a uniform prior and its MAP estimate, the posterior mean as a weighted average of the prior mean and the data, how the posterior narrows, updating one batch at a time, a Gamma prior on a Poisson error rate, the mean of a normal with known variance and its 95% credible interval, MAP regression with a Gaussian prior as ridge regression, the chance of the next two successes, a Beta prior built from a mean and a standard deviation, and add-one smoothing as a Dirichlet posterior mean, with worked solutions and the mistakes of squaring the posterior mean, forgetting the N in the data's precision, swapping the Gamma updates, inverting the ridge penalty and counting the first batch twice.

BEFORE YOU START

Maximum likelihood gives one number for a parameter. Bayesian inference gives a distribution: start with a prior density for the parameter, multiply by the likelihood of the data, and normalize. For a few prior and likelihood pairs, called conjugate, the posterior is in the same family as the prior, and the update is just arithmetic on the parameters. These ten problems work through the three conjugate pairs that matter most in practice, the Beta for a probability, the Gamma for a Poisson rate and the normal for a mean, then connect the posterior to the estimates that machine learning actually uses: the MAP estimate, which turns a Gaussian prior into ridge regression, and the posterior mean, which turns a uniform prior into add-one smoothing.

- **Bayes' rule for a parameter.** $p(\theta \mid \text{data}) \propto p(\text{data} \mid \theta)\, p(\theta)$: posterior is proportional to likelihood times prior. The missing constant makes the posterior integrate to 1, and it does not depend on θ.
- **Beta.** $\text{Beta}(\alpha, \beta)$ on $0 < \theta < 1$ has density proportional to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. Mean $\dfrac{\alpha}{\alpha+\beta}$, mode $\dfrac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$, and variance $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. $\text{Beta}(1,1)$ is the uniform prior.
- **Beta and binomial.** With prior $\text{Beta}(a, b)$ and k successes in N trials, the posterior is $\text{Beta}(a+k, b+N-k)$.
- **Gamma and Poisson.** $\text{Gamma}(\alpha, \beta)$, with shape α and **rate** β, has density proportional to $\lambda^{\alpha-1}e^{-\beta\lambda}$ on $\lambda > 0$, mean α/β, mode $(\alpha-1)/\beta$ for $\alpha \geq 1$ and variance α/β^2. With N Poisson counts adding up to S, the posterior is $\text{Gamma}(\alpha+S, \beta+N)$.
- **Normal mean, known variance.** With prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ and N observations from $\mathcal{N}(\mu, \sigma^2)$ with mean $\bar{x}$, precisions add: the posterior precision is $1/\tau_0^2 + N/\sigma^2$, and the posterior mean is the precision-weighted average of μ_0 and $\bar{x}$.

- **Estimates.** The MAP estimate is the posterior mode, the maximizer of $\log p(\text{data} \mid \theta) + \log p(\theta)$. With a uniform prior it is the maximum-likelihood estimate. The posterior mean minimizes the expected squared error.
- **Dirichlet.** $\text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ is the Beta for K categories, with mean $\alpha_j / \sum_i \alpha_i$. With category counts n_j the posterior is $\text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.

Builds on: Bayes' rule and conditional probability, Maximum likelihood estimation: derivations by hand

Problems

1. A new feature is shown to 12 users and 3 of them use it. With a uniform prior on the adoption rate θ, find the posterior, its mean and its MAP estimate, and compare them with the maximum-likelihood estimate. difficulty ·

2. A click-through rate has prior $\text{Beta}(6, 14)$. In 30 impressions there are 18 clicks. Find the posterior mean, and show it is a weighted average of the prior mean and the observed rate. What is the weight on the prior? difficulty ·

3. Find the standard deviation of the prior $\text{Beta}(6, 14)$ and of the posterior $\text{Beta}(24, 26)$ from Problem 2. difficulty ··

4. A conversion rate has prior $\text{Beta}(2, 2)$. On day 1, 4 of 10 visitors convert; on day 2, 7 of 10. Update one day at a time, then all at once, and compare. Give the final posterior mean and MAP estimate. difficulty ··

5. A service's errors per hour are Poisson with rate λ, and the prior is $\text{Gamma}(3, 2)$ (shape 3, rate 2). Over 5 hours the counts are $3, 1, 4, 0, 2$. Find the posterior, its mean, mode and standard deviation, and compare with the MLE. difficulty ··

6. A sensor's true offset μ has prior $\mathcal{N}(50, 16)$. Four readings, each $\mathcal{N}(\mu, 36)$, have mean $\bar{x} = 56$. Find the posterior mean and standard deviation of μ and a 95% credible interval. difficulty ··

7. Take $y = Xw + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and prior $w \sim \mathcal{N}(0, \tau^2 I)$. Show that the MAP estimate is ridge regression with $\lambda = \sigma^2/\tau^2$. Then compute it for $\sigma^2 = 1$, $\tau^2 = \frac{1}{2}$ and difficulty ···

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad y = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

8. With the posterior $\mathrm{Beta}(4, 10)$ of Problem 1, find the probability that the next user adopts the feature, and the probability that the next two both do. difficulty ··

9. An expert says a defect rate is "about 0.2, give or take 0.1". Find the Beta prior with mean 0.2 and standard deviation 0.1, then update it with 5 defects in 10 items and give the posterior mean. difficulty ··

10. A tiny language model sees the words *cat*, *dog* and *eel* 5, 0 and 3 times. Find the posterior mean of the word probabilities under a uniform $\mathrm{Dirichlet}(1, 1, 1)$ prior, and compare it with the maximum-likelihood estimate. difficulty ···

Answers

1. Posterior $\mathrm{Beta}(4, 10)$, mean $\frac{2}{7} \approx 0.286$, MAP 0.25, which equals the MLE $k/N = 0.25$
2. Posterior mean $0.48 = 0.4 \cdot 0.3 + 0.6 \cdot 0.6$, with weight $w = \frac{20}{50} = 0.4$ on the prior
3. Prior sd 0.1, posterior sd ≈ 0.070
4. Both routes give $\mathrm{Beta}(13, 11)$, with mean ≈ 0.542 and MAP ≈ 0.545
5. Posterior $\mathrm{Gamma}(13, 7)$: mean ≈ 1.857, mode ≈ 1.714, sd $\sqrt{13}/7 \approx 0.515$; the MLE is $S/N = 2$
6. Posterior $\mathcal{N}(53.84, 5.76)$, sd 2.4; 95% credible interval $53.84 \pm 1.96 \cdot 2.4 \approx (49.14, 58.54)$
7. $w_{\mathrm{MAP}} = (\frac{14}{15}, \frac{19}{15}) \approx (0.933, 1.267)$, ridge with $\lambda = \sigma^2/\tau^2 = 2$
8. $P(\text{next}) = \frac{2}{7} \approx 0.286$, $P(\text{next two}) = \frac{2}{21} \approx 0.095$
9. Prior $\mathrm{Beta}(3, 12)$; posterior $\mathrm{Beta}(8, 17)$ with mean $\frac{8}{25} = 0.32$
10. Posterior mean $(\frac{6}{11}, \frac{1}{11}, \frac{4}{11}) \approx (0.545, 0.091, 0.364)$; the MLE gives *dog* probability 0

Worked solutions

Problem 1

> A new feature is shown to 12 users and 3 of them use it. With a uniform prior on the adoption rate θ, find the posterior, its mean and its MAP estimate, and compare them with the maximum-likelihood estimate.

Prior $\mathrm{Beta}(1,1)$, $k=3$, $N=12$	The uniform density on $(0,1)$ is $\theta^0(1-\theta)^0$.
Posterior $\mathrm{Beta}(1+3, 1+9) = \mathrm{Beta}(4,10)$	Add the successes to α and the failures to β.
Mean $\frac{4}{14} = \frac{2}{7} \approx 0.286$	$\alpha/(\alpha+\beta)$.
MAP $\frac{4-1}{14-2} = \frac{3}{12} = 0.25$	The mode $(\alpha-1)/(\alpha+\beta-2)$.
Posterior $\mathrm{Beta}(4,10)$, **mean** $\frac{2}{7} \approx 0.286$, **MAP** 0.25, **which equals the MLE** $k/N = 0.25$	A flat prior adds nothing to the log-likelihood, so the posterior peaks where the likelihood does. The mean sits higher because the posterior is skewed to the right.

Problem 2

> A click-through rate has prior $\mathrm{Beta}(6,14)$. In 30 impressions there are 18 clicks. Find the posterior mean, and show it is a weighted average of the prior mean and the observed rate. What is the weight on the prior?

Posterior $\mathrm{Beta}(6+18, 14+12) = \mathrm{Beta}(24,26)$	12 impressions were not clicked.
Posterior mean $\frac{24}{50} = 0.48$	$\alpha/(\alpha+\beta)$.
Prior mean $\frac{6}{20} = 0.3$; observed rate $\frac{18}{30} = 0.6$	The two things being averaged.

$\dfrac{a+k}{a+b+N} = w\,\dfrac{a}{a+b} + (1-w)\,\dfrac{k}{N}$	Holds with $w = \dfrac{a+b}{a+b+N}$: the two terms are $\dfrac{a}{a+b+N}$ and $\dfrac{k}{a+b+N}$.
Posterior mean $0.48 = 0.4 \cdot 0.3 + 0.6 \cdot 0.6$, **with weight** $w = \frac{20}{50} = 0.4$ **on the prior**	The prior counts as $a + b = 20$ pseudo-observations against 30 real ones. As N grows, $w \to 0$ and the data take over.

Problem 3

> Find the standard deviation of the prior $\mathrm{Beta}(6, 14)$ and of the posterior $\mathrm{Beta}(24, 26)$ from Problem 2.

Prior variance $\dfrac{6 \cdot 14}{20^2 \cdot 21} = \dfrac{84}{8400} = 0.01$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.
Prior standard deviation 0.1	The square root.
Posterior variance $\dfrac{24 \cdot 26}{50^2 \cdot 51} = \dfrac{624}{127{,}500}$	The same formula with $\alpha + \beta = 50$.
≈ 0.004894	Less than half the prior variance.
Prior sd 0.1, **posterior sd** ≈ 0.070	The variance shrinks roughly like $1/(\alpha+\beta)$, the prior's 20 pseudo-observations plus the 30 real ones.

Problem 4

> A conversion rate has prior $\mathrm{Beta}(2, 2)$. On day 1, 4 of 10 visitors convert; on day 2, 7 of 10. Update one day at a time, then all at once, and compare. Give the final posterior mean and MAP estimate.

After day 1: $\mathrm{Beta}(2+4, 2+6) = \mathrm{Beta}(6, 8)$	6 visitors did not convert.

After day 2: $\text{Beta}(6+7, 8+3) = \text{Beta}(13, 11)$	Day 1's posterior is day 2's prior.
All at once: 11 of 20, so $\text{Beta}(2+11, 2+9)$	The same $\text{Beta}(13, 11)$.
Mean $\frac{13}{24} \approx 0.542$; MAP $\frac{12}{22} \approx 0.545$	$\alpha/(\alpha+\beta)$ and $(\alpha-1)/(\alpha+\beta-2)$.
Both routes give $\text{Beta}(13, 11)$, **with mean** ≈ 0.542 **and MAP** ≈ 0.545	Multiplying the likelihoods of two independent batches is the same as multiplying by one and then the other. Each observation must be used exactly once (Mistake 5).

Problem 5

> A service's errors per hour are Poisson with rate λ, and the prior is $\text{Gamma}(3, 2)$ (shape 3, rate 2). Over 5 hours the counts are $3, 1, 4, 0, 2$. Find the posterior, its mean, mode and standard deviation, and compare with the MLE.

$S = 3+1+4+0+2 = 10$ errors in $N = 5$ hours	Only the total and the number of hours enter the likelihood $\prod e^{-\lambda}\lambda^{x_i} \propto \lambda^S e^{-N\lambda}$.
Posterior $\propto \lambda^{10} e^{-5\lambda} \cdot \lambda^{2} e^{-2\lambda}$	The likelihood times the prior density.
$= \lambda^{12} e^{-7\lambda}$, so $\text{Gamma}(13, 7)$	Shape $\alpha + S$, rate $\beta + N$.
Mean $\frac{13}{7} \approx 1.857$; mode $\frac{12}{7} \approx 1.714$	α/β and $(\alpha-1)/\beta$.
Posterior $\text{Gamma}(13, 7)$: **mean** ≈ 1.857, **mode** ≈ 1.714, **sd** $\sqrt{13}/7 \approx 0.515$; **the MLE is** $S/N = 2$	The prior mean is $3/2 = 1.5$, worth 2 hours of data, so the posterior mean sits between 1.5 and 2, closer to the data's 5 hours.

Problem 6

A sensor's true offset μ has prior $\mathcal{N}(50, 16)$. Four readings, each $\mathcal{N}(\mu, 36)$, have mean $\bar{x} = 56$. Find the posterior mean and standard deviation of μ and a 95% credible interval.

Prior precision $\frac{1}{16} = 0.0625$	One over the prior variance.
Data precision $\frac{N}{\sigma^2} = \frac{4}{36} \approx 0.1111$	The mean of 4 readings has variance $36/4 = 9$.
Posterior precision 0.1736, variance 5.76	Precisions add; $1/0.1736 = 5.76$.
Mean $0.36 \cdot 50 + 0.64 \cdot 56 = 53.84$	Weights are the precisions over their sum: $0.0625/0.1736 = 0.36$.
Posterior $\mathcal{N}(53.84, 5.76)$, **sd** 2.4; 95% **credible interval** $53.84 \pm 1.96 \cdot 2.4 \approx (49.14, 58.54)$	The posterior sd is smaller than both the prior's 4 and the data's 3. The interval is a probability statement about μ given the data and the prior.

Problem 7

Take $y = Xw + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and prior $w \sim \mathcal{N}(0, \tau^2 I)$. Show that the MAP estimate is ridge regression with $\lambda = \sigma^2/\tau^2$. Then compute it for $\sigma^2 = 1$, $\tau^2 = \frac{1}{2}$ and

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad y = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

$-\log p(w \mid y) = \frac{1}{2\sigma^2}\lVert y - Xw\rVert^2 + \frac{1}{2\tau^2}\lVert w\rVert^2 + c$	The log-likelihood and the log-prior, with the terms that do not depend on w collected in c.
Multiply by $2\sigma^2$: minimize $\lVert y - Xw\rVert^2 + \lambda\lVert w\rVert^2$	A positive factor does not move the minimizer; $\lambda = \sigma^2/\tau^2$.
$w_{\text{MAP}} = (X^\top X + \lambda I)^{-1} X^\top y$	Set the gradient $-2X^\top(y - Xw) + 2\lambda w$ to 0.

$\lambda = 2$; $X^\top X + 2I = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$ and $X^\top y = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$	$X^\top X$ is $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.
$\frac{1}{15}\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} 5 \\ 6 \end{bmatrix} = \frac{1}{15}\begin{bmatrix} 14 \\ 19 \end{bmatrix}$	The 2×2 inverse, with determinant 15.
$w_{\text{MAP}} = (\frac{14}{15}, \frac{19}{15}) \approx (0.933, 1.267)$, **ridge with** $\lambda = \sigma^2/\tau^2 = 2$	Least squares gives $(\frac{4}{3}, \frac{7}{3})$; the prior pulls both weights toward 0. A tighter prior (smaller τ^2) means a larger λ.

Problem 8

> With the posterior $\text{Beta}(4, 10)$ of Problem 1, find the probability that the next user adopts the feature, and the probability that the next two both do.

$P(\text{next}) = E[\theta] = \frac{4}{14} = \frac{2}{7}$	Average $P(\text{success} \mid \theta) = \theta$ over the posterior.
$P(\text{next two}) = E[\theta^2]$	Given θ the two are independent with probability θ^2; average that over the posterior.
$E[\theta^2] = \dfrac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$	The second moment of a Beta.
$= \dfrac{4 \cdot 5}{14 \cdot 15} = \dfrac{20}{210}$	With $\alpha = 4$, $\beta = 10$.
$P(\text{next}) = \frac{2}{7} \approx 0.286$, $P(\text{next two}) = \frac{2}{21} \approx 0.095$	Larger than $(\frac{2}{7})^2 \approx 0.082$, because $E[\theta^2] = (E\theta)^2 + \text{Var}(\theta)$ (Mistake 1).

Problem 9

An expert says a defect rate is "about 0.2, give or take 0.1". Find the Beta prior with mean 0.2 and standard deviation 0.1, then update it with 5 defects in 10 items and give the posterior mean.

With $m = \frac{\alpha}{\alpha+\beta}$ and $n = \alpha + \beta$: $\text{Var} = \frac{m(1-m)}{n+1}$	The Beta variance rewritten in terms of the mean and the total n.
$\frac{0.2 \cdot 0.8}{n+1} = 0.01$, so $n + 1 = 16$	$0.1^2 = 0.01$.
$n = 15$: $\alpha = 0.2 \cdot 15 = 3$, $\beta = 12$	$\alpha = mn$, $\beta = (1-m)n$.
Posterior $\text{Beta}(3 + 5, 12 + 5) = \text{Beta}(8, 17)$	5 defects and 5 good items.
Prior $\text{Beta}(3, 12)$; **posterior** $\text{Beta}(8, 17)$ **with mean** $\frac{8}{25} = 0.32$	The expert's opinion is worth 15 items, so 10 items at a rate of 0.5 move the mean from 0.2 only to 0.32.

Problem 10

A tiny language model sees the words *cat*, *dog* and *eel* 5, 0 and 3 times. Find the posterior mean of the word probabilities under a uniform $\text{Dirichlet}(1, 1, 1)$ prior, and compare it with the maximum-likelihood estimate.

Posterior $\text{Dirichlet}(1 + 5, 1 + 0, 1 + 3) = \text{Dirichlet}(6, 1, 4)$	Add each count to its parameter.
Posterior mean $\frac{1}{11}(6, 1, 4)$	Each parameter over their sum 11.
MLE $\frac{1}{8}(5, 0, 3) = (0.625, 0, 0.375)$	Counts over the total 8.
$\frac{n_j + 1}{N + K}$ with $N = 8$, $K = 3$	The posterior mean adds one to every count and K to the total.

Posterior mean $(\frac{6}{11}, \frac{1}{11}, \frac{4}{11}) \approx (0.545, 0.091, 0.364)$; **the MLE gives *dog* probability** 0	This is add-one (Laplace) smoothing. Under the MLE an unseen word has probability 0, so any sentence containing it has likelihood 0 and log-likelihood $-\infty$.

Where this goes wrong

1. Squaring the posterior mean for two future successes

Problem 8 asks for two adoptions in a row, and two independent events multiply.

$P(\text{next}) = E[\theta] = \frac{2}{7}$	Right so far.
"The next two are independent, so square it."	They are independent only given θ. Unconditionally, a first adoption is evidence that θ is high, which raises the chance of the second.
$P(\text{next two}) = (\frac{2}{7})^2 = \frac{4}{49} \approx 0.082$ ✗	The right value is $E[\theta^2] = \frac{2}{21} \approx 0.095$. The gap is $\text{Var}(\theta) = \frac{2}{21} - \frac{4}{49} \approx 0.014$, the posterior's own uncertainty.

2. Forgetting the N in the data's precision

Problem 6 has a prior variance of 16 and a reading variance of 36, and the two variances are the numbers on the page.

Prior precision $\frac{1}{16}$	Right so far.
"The data precision is $\frac{1}{36}$."	That is the precision of a single reading. The mean of 4 readings has variance $36/4 = 9$, so the data precision is $4/36$.
Posterior variance $1/(\frac{1}{16} + \frac{1}{36}) \approx 11.08$, mean ≈ 51.85 ✗	The right answer is variance 5.76 and mean 53.84. Dropping N treats four readings as one, so the posterior stays too close to the prior and too wide (sd ≈ 3.33 instead of 2.4).

3. Swapping the Gamma updates

The Poisson update adds two numbers to the two Gamma parameters, and it is easy to pair them the wrong way.

Prior $\text{Gamma}(3, 2)$, $S = 10$ errors in $N = 5$ hours	Right so far: Problem 5.
"Add the hours to the shape and the errors to the rate."	The likelihood is $\lambda^S e^{-N\lambda}$: the count S is a power of λ, which belongs with the shape, and the exposure N multiplies λ in the exponent, which belongs with the rate.
Posterior $\text{Gamma}(3 + 5, 2 + 10) = \text{Gamma}(8, 12)$, mean ≈ 0.667 ✗	The right posterior is $\text{Gamma}(13, 7)$ with mean ≈ 1.857. A rate of 0.667 errors an hour is below both the prior mean 1.5 and the observed 2, which a posterior mean can never be.

4. Inverting the ridge penalty

Problem 7's prior variance is $\tau^2 = \frac{1}{2}$ and the noise variance is $\sigma^2 = 1$, and the ratio is easy to write upside down.

Minimize $\frac{1}{2\sigma^2}\|y - Xw\|^2 + \frac{1}{2\tau^2}\|w\|^2$	Right so far.
"$\lambda = \tau^2/\sigma^2$."	Multiplying through by $2\sigma^2$ leaves σ^2/τ^2 in front of $\|w\|^2$. A tight prior, small τ^2, has to mean a strong penalty.
$\lambda = \frac{1}{2}$, so $w \approx (1.238, 1.905)$ ✗	With $\lambda = 2$ the answer is $(0.933, 1.267)$. The inverted ratio makes a confident prior shrink the weights less than a vague one would.

5. Counting the first batch twice

In Problem 4, day 1's posterior becomes day 2's prior, and then all the data so far are used for the update.

After day 1: $\text{Beta}(6, 8)$	Right so far.

“Update with everything seen: 11 of 20.”	Day 1's 4 of 10 is already inside $\text{Beta}(6, 8)$. The update should use only the new data, 7 of 10.
Posterior $\text{Beta}(6 + 11, 8 + 9) = \text{Beta}(17, 17)$, mean 0.5 ✗	The right posterior is $\text{Beta}(13, 11)$, mean ≈ 0.542. Counting data twice also makes the posterior too confident: sd ≈ 0.085 instead of ≈ 0.100.

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

eulerpad.app/practice/bayesian-inference-and-conjugate-priors