

Gaussian process conditioning

Ten problems on Gaussian process regression: the predictive mean and covariance from the joint Gaussian, two training points worked by hand, the posterior mean as a sum of kernels and why the variance only falls, the RBF kernel's derivatives, the gradient of the log marginal likelihood with respect to the lengthscale, signal variance and noise, the noise gradient through the eigenvalues, conditioning one observation at a time, the linear kernel as ridge regression, the lengthscale limits, and the derivative of the posterior mean with respect to the lengthscale, with worked solutions and the mistakes that leave out the noise, the log-determinant, or the second term of a derivative through an inverse.

BEFORE YOU START

A Gaussian process is a prior over functions under which any finite set of function values is jointly Gaussian with a covariance given by a kernel, and Gaussian process regression is nothing more than the conditioning formula of the multivariate-Gaussian page applied to that joint: condition the unseen values on the observed ones. Every formula in the subject, the predictive mean and variance, the marginal likelihood, its hyperparameter gradients, comes from that one step plus the inverse-and-determinant calculus of the previous page. These ten problems do the step, run it on two data points by hand, read what the result says (a weighted sum of kernels, a variance that can only fall), differentiate the kernel and then the marginal likelihood with respect to the lengthscale, the signal variance and the noise, condition one observation at a time, show that a linear kernel recovers ridge regression, and take the lengthscale to its two limits. The five mistakes are each a missing term: the noise left out of the matrix that is inverted, noise added where none belongs, a marginal likelihood gradient without its log-determinant, a lengthscale derivative short a factor of 2, and a derivative of the posterior mean that holds the inverse fixed.

- f is a function with a Gaussian process prior with zero mean and kernel k : for any inputs $x_1, \dots, x_n$, the vector $(f(x_1), \dots, f(x_n))$ is Gaussian with mean 0 and covariance matrix $K_{ij} = k(x_i, x_j)$. Kernels here are positive definite, so every such matrix is symmetric positive semidefinite.
- Training inputs $X = (x_1, \dots, x_n)$ with function values $f = (f(x_1), \dots, f(x_n))^T$ and noisy observations $y = f + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ independent of f . Test inputs $X_* = (x_1^*, \dots, x_m^*)$ with values $f_* \in \mathbb{R}^m$.
- Kernel blocks: $K = k(X, X)$ is $n \times n$, $K_* = k(X, X_*)$ is $n \times m$ with $(K_*)_{ij} = k(x_i, x_j^*)$, and $K_{**} = k(X_*, X_*)$ is $m \times m$. For a single test point, $k_* = k(X, x_*) \in \mathbb{R}^n$ and $k_{**} = k(x_*, x_*)$. $K_y = K + \sigma^2 I$ and $\alpha = K_y^{-1} y$.

- Gaussian conditioning (the multivariate-Gaussian page, Problem 9): if (x_a, x_b) is jointly Gaussian with mean (μ_a, μ_b) and covariance blocks $\Sigma_{aa}, \Sigma_{ab}, \Sigma_{bb}$, then $x_a | x_b \sim \mathcal{N}(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$.
- The RBF (squared-exponential) kernel on scalar or vector inputs is $k(x, x') = s^2 \exp(-\|x - x'\|^2/(2\ell^2))$ with signal variance s^2 and lengthscale $\ell > 0$. Its hyperparameters, with σ^2 , are written θ . Numbers on this page use scalar inputs.
- From the previous page: $d(A^{-1}) = -A^{-1}(dA)A^{-1}$, $d \log \det A = \text{tr}(A^{-1}dA)$, and tr is cyclic. $\text{tr}(A)$ of a symmetric matrix with eigenvalues λ_i is $\sum_i \lambda_i$.

Builds on: [The multivariate Gaussian: gradients and identities](#), [Matrix inverse and log-determinant derivatives](#)

Problems

1. Write the joint distribution of (f_*, y) and condition on y to obtain the predictive distribution $f_* | y \sim \mathcal{N}(\mu_*, \Sigma_*)$ with $\mu_* = K_*^\top K_y^{-1}y$ and $\Sigma_* = K_{**} - K_*^\top K_y^{-1}K_*$. difficulty ..

2. Two training points $x_1 = 0, x_2 = 1$ with $y = (1, 2)^\top$, RBF kernel with $s^2 = 1, \ell = 1$, noise $\sigma^2 = 0.1$. Compute the predictive mean and variance at $x_* = 0.5$ to four decimal places. difficulty ..

3. Show that the posterior mean at any x_* is $\mu(x_*) = \sum_{i=1}^n \alpha_i k(x_i, x_*)$ with $\alpha = K_y^{-1}y$ not depending on x_* ; that the posterior variance never exceeds the prior variance; and that with $\sigma^2 = 0$ the posterior mean passes through every observation with zero variance there. difficulty ..

4. For the RBF kernel with $r = \|x - x'\|$, compute $\partial k / \partial \ell, \partial k / \partial s^2$ and, for a scalar test input, $\partial k(x_i, x_*) / \partial x_*$. Use the last to write the slope $\mu'(x_*)$ of the posterior mean. difficulty .

5. The log marginal likelihood is $\log p(y | X, \theta) = -\frac{1}{2}y^\top K_y^{-1}y - \frac{1}{2} \log \det K_y - \frac{n}{2} \log 2\pi$. For a hyperparameter θ with $\partial K_y / \partial \theta$ known, show that $\frac{\partial}{\partial \theta} \log p(y | X, \theta) = \frac{1}{2} \text{tr} \left((\alpha \alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta} \right)$, and write the two terms for $\theta = \ell$. difficulty ...

6. Let K have eigenvalues $\lambda_1, \dots, \lambda_n \geq 0$. Show that $\log \det K_y =$ difficulty $\dots$
 $\sum_i \log(\lambda_i + \sigma^2)$ and $\text{tr}(K_y^{-1}) = \sum_i 1/(\lambda_i + \sigma^2)$, hence that $\frac{\partial \log p}{\partial \sigma^2} = \frac{1}{2}(\|\alpha\|^2 - \text{tr } K_y^{-1})$, and
state the condition at a maximum over σ^2 .

7. After conditioning on n observations, the posterior over difficulty $\dots$
 $(f_*, f(x_{n+1}))$ is Gaussian with mean (μ_*, μ_{n+1}) , covariance Σ_* for f_* , cross-covariance $c =$
 $\text{Cov}(f_*, f(x_{n+1}))$ and variance $v = \text{Var}(f(x_{n+1}))$. A new observation $y_{n+1} = f(x_{n+1}) + \varepsilon_{n+1}$
arrives. Show that the updated posterior is $\mu'_* = \mu_* + \frac{c(y_{n+1} - \mu_{n+1})}{v + \sigma^2}$ and $\Sigma'_* = \Sigma_* -$
 $\frac{cc^\top}{v + \sigma^2}$, and say why this equals the batch posterior on all $n + 1$ points.

8. Take the linear kernel $k(x, x') = x^\top x'$ with inputs as rows of $X \in$ difficulty $\dots$
 $\mathbb{R}^{n \times d}$ and $X_* \in \mathbb{R}^{m \times d}$. Show that the posterior mean is $X_* w$ with $w = (X^\top X + \sigma^2 I)^{-1} X^\top y$,
the ridge solution with penalty σ^2 , and that the posterior covariance is $\sigma^2 X_* (X^\top X +$
 $\sigma^2 I)^{-1} X_*^\top$.

9. With the RBF kernel, find the posterior mean and variance at a test difficulty $\dots$
point as $\ell \rightarrow \infty$ and as $\ell \rightarrow 0$ (for x_* not equal to any training input).

10. For a single test point, $\mu_* = k_*^\top K_y^{-1} y$ depends on ℓ through both k_* difficulty $\dots$
and K . Compute $\frac{d\mu_*}{d\ell}$ in terms of α , $\partial k_*/\partial \ell$ and $\partial K/\partial \ell$.

Answers

1. $f_* \mid y \sim \mathcal{N}(K_*^\top K_y^{-1} y, K_{**} - K_*^\top K_y^{-1} K_*)$
2. $\mu_* \approx 1.5514$ and $\Sigma_* = 1 - 0.9128 \approx 0.0873$
3. $\mu(x_*) = \sum_i \alpha_i k(x_i, x_*)$; $\text{Var}(f(x_*) \mid y) \leq k_{**}$ everywhere; with $\sigma^2 = 0$, $\mu(x_j) = y_j$ and the variance at x_j is 0
4. $\partial k / \partial \ell = k r^2 / \ell^3$, $\partial k / \partial s^2 = k / s^2$, $\partial k(x_i, x_*) / \partial x_* = -k(x_i, x_*)(x_* - x_i) / \ell^2$, and $\mu'(x_*) = -\frac{1}{\ell^2} \sum_i \alpha_i (x_* - x_i) k(x_i, x_*)$
5. $\frac{\partial \log p}{\partial \theta} = \frac{1}{2} \text{tr} \left((\alpha \alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta} \right) = \frac{1}{2} \alpha^\top \frac{\partial K_y}{\partial \theta} \alpha - \frac{1}{2} \text{tr} \left(K_y^{-1} \frac{\partial K_y}{\partial \theta} \right)$
6. $\log \det K_y = \sum_i \log(\lambda_i + \sigma^2)$, $\text{tr} K_y^{-1} = \sum_i (\lambda_i + \sigma^2)^{-1}$, $\partial \log p / \partial \sigma^2 = \frac{1}{2} (\|\alpha\|^2 - \text{tr} K_y^{-1})$, and at the optimum $\|\alpha\|^2 = \text{tr} K_y^{-1}$
7. $\mu'_* = \mu_* + \frac{c(y_{n+1} - \mu_{n+1})}{v + \sigma^2}$, $\Sigma'_* = \Sigma_* - \frac{cc^\top}{v + \sigma^2}$, identical to the batch posterior on $n + 1$ points
8. $\mu_* = X_*(X^\top X + \sigma^2 I)^{-1} X^\top y$ and $\Sigma_* = \sigma^2 X_*(X^\top X + \sigma^2 I)^{-1} X_*^\top$
9. $\ell \rightarrow \infty$: $\mu_* \rightarrow \frac{ns^2}{ns^2 + \sigma^2} \bar{y}$ and $\Sigma_* \rightarrow \frac{s^2 \sigma^2}{ns^2 + \sigma^2}$ everywhere; $\ell \rightarrow 0$: $\mu_* \rightarrow 0$ and $\Sigma_* \rightarrow s^2$ away from the data
10. $\frac{d\mu_*}{d\ell} = \left(\frac{\partial k_*}{\partial \ell} \right)^\top \alpha - k_*^\top K_y^{-1} \frac{\partial K}{\partial \ell} \alpha$

Worked solutions

Problem 1

Write the joint distribution of (f_*, y) and condition on y to obtain the predictive distribution $f_* | y \sim \mathcal{N}(\mu_*, \Sigma_*)$ with $\mu_* = K_*^\top K_y^{-1} y$ and $\Sigma_* = K_{**} - K_*^\top K_y^{-1} K_*$.

(f, f_*) is jointly Gaussian with mean 0 and covariance $\begin{pmatrix} K & K_* \\ K_*^\top & K_{**} \end{pmatrix}$.	The prior applied to the $n + m$ inputs X and X_* together; the blocks are the kernel evaluated between the two sets.
$y = f + \varepsilon$ is Gaussian with $\text{Cov}(y) = K + \sigma^2 I = K_y$.	A sum of independent Gaussians is Gaussian and their covariances add; $\text{Cov}(\varepsilon) = \sigma^2 I$.
$\text{Cov}(f_*, y) = \text{Cov}(f_*, f) + \text{Cov}(f_*, \varepsilon) = K_*^\top + 0$.	Covariance is bilinear, and ε is independent of everything in f_*. The noise touches only the y block.
$(f_*, y) \sim \mathcal{N}\left(0, \begin{pmatrix} K_{**} & K_*^\top \\ K_* & K_y \end{pmatrix}\right)$.	Steps 1 to 3 assembled with $x_a = f_*$ and $x_b = y$; (f_*, y) is jointly Gaussian because it is a linear function of the Gaussian vector (f, f_*, ε) (the multivariate-Gaussian page, Problem 7).
$\mu_* = 0 + K_*^\top K_y^{-1}(y - 0)$ and $\Sigma_* = K_{**} - K_*^\top K_y^{-1} K_*$.	The conditioning formula with $\Sigma_{ab} = K_*^\top$, $\Sigma_{bb} = K_y$, $\Sigma_{aa} = K_{**}$ and both means 0.
$f_* y \sim \mathcal{N}(K_*^\top K_y^{-1} y, K_{**} - K_*^\top K_y^{-1} K_*)$	μ_* is $m \times 1$ and Σ_* is $m \times m$. The noise appears once, in the matrix that is inverted, because the observations are noisy and the quantity predicted, f_*, is not; to predict a noisy y_* instead, add $\sigma^2 I$ to Σ_* afterwards. The inverse of K_y is the whole cost, $O(n^3)$ once, after which each test point needs $O(n)$ for the mean and $O(n^2)$ for the variance.

Problem 2

Two training points $x_1 = 0, x_2 = 1$ with $y = (1, 2)^\top$, RBF kernel with $s^2 = 1, \ell = 1$, noise $\sigma^2 = 0.1$. Compute the predictive mean and variance at $x_* = 0.5$ to four decimal places.

$$K = \begin{pmatrix} 1 & e^{-1/2} \\ e^{-1/2} & 1 \end{pmatrix}, \text{ and } e^{-1/2} \approx 0.6065, \quad k(0, 1) = \exp(-1/2); k(x, x) = 1; \text{ add } \sigma^2 = 0.1 \text{ to the diagonal.}$$

so $K_y = \begin{pmatrix} 1.1 & 0.6065 \\ 0.6065 & 1.1 \end{pmatrix}$.

$$k_* = (e^{-1/8}, e^{-1/8})^\top \approx (0.8825, 0.8825)^\top$$

and $k_{**} = 1$. Both training points are at distance 0.5 from x_* : $\exp(-0.25/2) = \exp(-1/8)$.

$$K_y^{-1} = \frac{1}{a^2 - b^2} \begin{pmatrix} a & -b \\ -b & a \end{pmatrix} \text{ with } a = 1.1, \quad \text{The } 2 \times 2 \text{ inverse of a symmetric matrix with equal diagonal: swap nothing, negate the off-diagonal, divide by the determinant.}$$

$b = 0.6065, a^2 - b^2 \approx 1.21 - 0.3679 = 0.8421$.

$$\alpha = K_y^{-1}y = \frac{1}{0.8421} (1.1 - 2(0.6065), -0.6065 + 2(1.1))^\top \approx (-0.1343, 1.8922)^\top.$$

Multiply out: $1.1 - 1.2131 = -0.1131$ and $-0.6065 + 2.2 = 1.5935$, each divided by 0.8421.

$$\mu_* = k_*^\top \alpha = 0.8825 (-0.1343 + 1.8922) = 0.8825 \times 1.7580 \approx 1.5514.$$

Both entries of k_* are equal, so the dot product is 0.8825 times the sum of α .

$$k_*^\top K_y^{-1} k_* = 0.8825^2 \times (\text{sum of all entries of } K_y^{-1}) = 0.7788 \times \frac{2(a - b)}{a^2 - b^2} = 0.7788 \times \frac{0.9870}{0.8421} \approx 0.7788 \times 1.1720 \approx 0.9128.$$

With $k_* = c\mathbf{1}$, $k_*^\top M k_* = c^2 \mathbf{1}^\top M \mathbf{1}$, the sum of the entries; for step 3's inverse that is $2(a - b)/(a^2 - b^2)$.

$$\mu_* \approx 1.5514 \text{ and } \Sigma_* = 1 - 0.9128 \approx 0.0873$$

A little above the midpoint 1.5 of the two observations, because the second observation has the larger weight α_2 and the first a small negative one: α is not a set of averaging weights but the solution of $K_y \alpha = y$. The variance has fallen from the prior 1 to 0.0873, near the noise level 0.1, as a point half a lengthscale from two observations should.

Problem 3

Show that the posterior mean at any x_* is $\mu(x_*) = \sum_{i=1}^n \alpha_i k(x_i, x_*)$ with $\alpha = K_y^{-1} y$ not depending on x_* ; that the posterior variance never exceeds the prior variance; and that with $\sigma^2 = 0$ the posterior mean passes through every observation with zero variance there.

$$\mu(x_*) = k_*^\top K_y^{-1} y = k_*^\top \alpha = \sum_i (k_*)_i \alpha_i = \sum_i \alpha_i k(x_i, x_*).$$

Problem 1 for one test point; α depends on the training data only, so it is computed once.

$$K_*^\top K_y^{-1} K_* = (K_y^{-1/2} K_*)^\top (K_y^{-1/2} K_*) \text{ is positive semidefinite.}$$

K_y is positive definite (a positive semidefinite K plus $\sigma^2 I$), so it has a symmetric square root and $K_y^{-1} = K_y^{-1/2} K_y^{-1/2}$; a matrix of the form $B^\top B$ is positive semidefinite.

$$\text{So } \Sigma_* = K_{**} - K_*^\top K_y^{-1} K_* \preceq K_{**}, \text{ and in particular } \text{Var}(f(x_*) | y) = k_{**} - k_*^\top K_y^{-1} k_* \leq k_{**}.$$

Subtracting a positive semidefinite matrix can only lower each diagonal entry. The data cannot increase uncertainty about f anywhere.

$$\text{With } \sigma^2 = 0 \text{ and } x_* = x_j: k_* = K e_j \text{ (column } j \text{ of } K) \text{ and } k_{**} = K_{jj}.$$

$k(x_i, x_j)$ is entry (i, j) of K .

$$\mu(x_j) = e_j^\top K^\top K^{-1} y = e_j^\top y = y_j \text{ and } \text{Var} = K_{jj} - e_j^\top K K^{-1} K e_j = K_{jj} - K_{jj} = 0.$$

$K^\top = K$ and $K K^{-1} = I$, assuming K is invertible.

$\mu(x_*) = \sum_i \alpha_i k(x_i, x_*)$; $\text{Var}(f(x_*) \mid y) \leq k_{**}$ **everywhere; with $\sigma^2 = 0$, $\mu(x_j) = y_j$ and the variance at x_j is 0**

The posterior mean is a fixed linear combination of n copies of the kernel, one centred on each training input, so it inherits the kernel's shape: smooth for the RBF, returning to the prior mean 0 far from the data (Problem 9). The variance depends on the inputs X and x_* but not on y : where the data are is what reduces uncertainty, not what they say. Noise-free conditioning interpolates; with $\sigma^2 > 0$ the mean passes near, not through, the observations, and the variance at a training input is positive.

Problem 4

For the RBF kernel with $r = \|x - x'\|$, compute $\partial k / \partial \ell$, $\partial k / \partial s^2$ and, for a scalar test input, $\partial k(x_i, x_*) / \partial x_*$. Use the last to write the slope $\mu'(x_*)$ of the posterior mean.

$$\frac{\partial}{\partial \ell} \left(-\frac{r^2}{2\ell^2} \right) = -\frac{r^2}{2} \cdot (-2)\ell^{-3} = \frac{r^2}{\ell^3}.$$

$$\frac{d}{d\ell} \ell^{-2} = -2\ell^{-3}.$$

$$\frac{\partial k}{\partial \ell} = k \cdot \frac{r^2}{\ell^3}.$$

Chain rule through the exponential:
 $\frac{d}{d\ell} e^{u(\ell)} = e^u u'$, and $s^2 e^u = k$.

$$\frac{\partial k}{\partial s^2} = \exp \left(-\frac{r^2}{2\ell^2} \right) = \frac{k}{s^2}.$$

k is linear in s^2 .

$$\frac{\partial k(x_i, x_*)}{\partial x_*} = k(x_i, x_*) \cdot \frac{\partial}{\partial x_*} \left(-\frac{(x_* - x_i)^2}{2\ell^2} \right) = -k(x_i, x_*) \frac{x_* - x_i}{\ell^2}.$$

Chain rule; $\frac{d}{dx_*} (x_* - x_i)^2 = 2(x_* - x_i)$ and the 2 cancels.

$$\mu'(x_*) = \sum_i \alpha_i \frac{\partial k(x_i, x_*)}{\partial x_*} = -\frac{1}{\ell^2} \sum_i \alpha_i (x_* - x_i) k(x_i, x_*).$$

Differentiate Problem 3's representer form term by term; α does not depend on x_* .

$$\begin{aligned}\partial k / \partial \ell &= k r^2 / \ell^3, \partial k / \partial s^2 = k / s^2, \\ \partial k(x_i, x_*) / \partial x_* &= -k(x_i, x_*)(x_* - x_i) / \ell^2, \\ \text{and } \mu'(x_*) &= -\frac{1}{\ell^2} \sum_i \alpha_i (x_* - x_i) k(x_i, x_*)\end{aligned}$$

The lengthscale derivative is largest at $r \approx \sqrt{2} \ell$ and vanishes at $r = 0$ (the diagonal of K never depends on ℓ) and at large r ; the signal-variance derivative is the kernel with s^2 stripped out. The slope formula is how a Gaussian process gives derivatives of its mean function for free, and the same $\partial k / \partial x_*$ is the cross-covariance between f and f' .

Problem 5

The log marginal likelihood is $\log p(y | X, \theta) = -\frac{1}{2} y^\top K_y^{-1} y - \frac{1}{2} \log \det K_y - \frac{n}{2} \log 2\pi$. For a hyperparameter θ with $\partial K_y / \partial \theta$ known, show that $\frac{\partial}{\partial \theta} \log p(y | X, \theta) = \frac{1}{2} \text{tr} \left((\alpha \alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta} \right)$, and write the two terms for $\theta = \ell$.

$$\begin{aligned}\frac{\partial}{\partial \theta} (y^\top K_y^{-1} y) &= y^\top \frac{\partial K_y^{-1}}{\partial \theta} y = \\ &= -y^\top K_y^{-1} \frac{\partial K_y}{\partial \theta} K_y^{-1} y.\end{aligned}$$

y is data, so only the inverse moves; the previous page's Problem 4 gives $\partial(A^{-1})/\partial\theta = -A^{-1}(\partial A/\partial\theta)A^{-1}$.

$$= -\alpha^\top \frac{\partial K_y}{\partial \theta} \alpha = -\text{tr} \left(\alpha \alpha^\top \frac{\partial K_y}{\partial \theta} \right).$$

$K_y^{-1} y = \alpha$ and $y^\top K_y^{-1} = \alpha^\top$ since K_y is symmetric; a scalar is its own trace and the cyclic property moves α to the back.

$$\frac{\partial}{\partial \theta} \log \det K_y = \text{tr} \left(K_y^{-1} \frac{\partial K_y}{\partial \theta} \right).$$

The previous page's Problem 4 again, for $\log \det$.

$$\begin{aligned}\frac{\partial}{\partial \theta} \log p &= \frac{1}{2} \text{tr} \left(\alpha \alpha^\top \frac{\partial K_y}{\partial \theta} \right) - \\ &= \frac{1}{2} \text{tr} \left(K_y^{-1} \frac{\partial K_y}{\partial \theta} \right) = \frac{1}{2} \text{tr} \left((\alpha \alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta} \right).\end{aligned}$$

$-\frac{1}{2}$ times step 2 and $-\frac{1}{2}$ times step 3; the constant term has no θ ; the trace is linear.

For $\theta = \ell$: $\frac{\partial K_y}{\partial \ell} = \frac{\partial K}{\partial \ell}$ with entries $K_{ij} (x_i - x_j)^2 / \ell^3$, and the two terms are $\frac{1}{2} \alpha^\top \frac{\partial K}{\partial \ell} \alpha$ and $-\frac{1}{2} \text{tr} \left(K_y^{-1} \frac{\partial K}{\partial \ell} \right)$.

$\sigma^2 I$ does not depend on ℓ ; Problem 4 entry by entry.

$$\frac{\partial \log p}{\partial \theta} = \frac{1}{2} \text{tr} \left((\alpha \alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta} \right) = \frac{1}{2} \alpha^\top \frac{\partial K_y}{\partial \theta} \alpha - \frac{1}{2} \text{tr} \left(K_y^{-1} \frac{\partial K_y}{\partial \theta} \right)$$

The first term is the derivative of the data fit $-\frac{1}{2} y^\top K_y^{-1} y$ and the second of the complexity penalty $-\frac{1}{2} \log \det K_y$; the marginal likelihood trades them off automatically, which is why its maximum is a usable choice of hyperparameters without a validation set. The same $\partial K_y / \partial \theta$ with K/s^2 (Problem 4) gives the signal-variance gradient, and with I the noise gradient (Problem 6). Each gradient costs one $n \times n$ product once K_y^{-1} and α are known.

Problem 6

Let K have eigenvalues $\lambda_1, \dots, \lambda_n \geq 0$. Show that $\log \det K_y = \sum_i \log(\lambda_i + \sigma^2)$ and $\text{tr}(K_y^{-1}) = \sum_i 1/(\lambda_i + \sigma^2)$, hence that $\frac{\partial \log p}{\partial \sigma^2} = \frac{1}{2} (\|\alpha\|^2 - \text{tr} K_y^{-1})$, and state the condition at a maximum over σ^2 .

$K = Q \Lambda Q^\top$ with Q orthogonal, so $K_y = Q(\Lambda + \sigma^2 I)Q^\top$.

K is symmetric; $I = Q Q^\top$ puts the noise into the same basis, as on the previous page's Problem 4.

$\det K_y = \det(\Lambda + \sigma^2 I) = \prod_i (\lambda_i + \sigma^2)$, so $\log \det K_y = \sum_i \log(\lambda_i + \sigma^2)$.

$\det(Q D Q^\top) = \det Q \det D \det Q^\top = \det D$ since $\det Q = \pm 1$; a diagonal determinant is the product of the diagonal.

$K_y^{-1} = Q(\Lambda + \sigma^2 I)^{-1} Q^\top$ and $\text{tr} K_y^{-1} = \sum_i \frac{1}{\lambda_i + \sigma^2}$.

Inverse of $Q D Q^\top$ is $Q D^{-1} Q^\top$; the trace is cyclic, $\text{tr}(Q D^{-1} Q^\top) = \text{tr}(D^{-1})$.

$\frac{\partial K_y}{\partial \sigma^2} = I$, so Problem 5 gives $\frac{\partial \log p}{\partial \sigma^2} = \frac{1}{2} \alpha^\top \alpha - \frac{1}{2} \text{tr} K_y^{-1}$.

Substitute $\partial K_y / \partial \theta = I$ into both terms.

At a maximum over σ^2 (an interior one),

$$\|\alpha\|^2 = \text{tr} K_y^{-1} = \sum_i \frac{1}{\lambda_i + \sigma^2}.$$

Set the derivative to zero.

$\log \det K_y = \sum_i \log(\lambda_i + \sigma^2)$, $\text{tr } K_y^{-1} = \sum_i (\lambda_i + \sigma^2)^{-1}$, $\partial \log p / \partial \sigma^2 = \frac{1}{2} (\|\alpha\|^2 - \text{tr } K_y^{-1})$, **and at the optimum** $\|\alpha\|^2 = \text{tr } K_y^{-1}$

$\|\alpha\|^2 = y^\top K_y^{-2} y$ measures how hard the model is working to explain y and $\text{tr } K_y^{-1}$ is what that quantity would be on average for data drawn from the model itself ($\mathbb{E}[y^\top K_y^{-2} y] = \text{tr}(K_y^{-2} K_y)$ for $y \sim \mathcal{N}(0, K_y)$); the noise is tuned until the two agree. The sum form shows the log-determinant is dominated by the small eigenvalues of K , which σ^2 lifts: the noise is also the regulariser that keeps K_y invertible, and the "jitter" added to a noise-free kernel matrix is exactly a small σ^2 .

Problem 7

After conditioning on n observations, the posterior over $(f_*, f(x_{n+1}))$ is Gaussian with mean (μ_*, μ_{n+1}) , covariance Σ_* for f_* , cross-covariance $c = \text{Cov}(f_*, f(x_{n+1}))$ and variance $v = \text{Var}(f(x_{n+1}))$. A new observation $y_{n+1} = f(x_{n+1}) + \varepsilon_{n+1}$ arrives. Show that the updated posterior is $\mu'_* = \mu_* + \frac{c(y_{n+1} - \mu_{n+1})}{v + \sigma^2}$ and $\Sigma'_* = \Sigma_* - \frac{cc^\top}{v + \sigma^2}$, and say why this equals the batch posterior on all $n + 1$ points.

Given the first n observations, (f_*, y_{n+1}) is jointly Gaussian with mean (μ_*, μ_{n+1}) , $\text{Cov}(f_*) = \Sigma_*$, $\text{Cov}(f_*, y_{n+1}) = c$ and $\text{Var}(y_{n+1}) = v + \sigma^2$.

The posterior after n points is a Gaussian (Problem 1), and y_{n+1} adds independent noise of variance σ^2 to $f(x_{n+1})$, which changes its variance but not its covariance with f_* (as in Problem 1, step 3).

Condition on y_{n+1} : $\mu'_* = \mu_* + c(v + \sigma^2)^{-1}(y_{n+1} - \mu_{n+1})$.

The conditioning formula with $\Sigma_{ab} = c$ ($m \times 1$) and $\Sigma_{bb} = v + \sigma^2$, a scalar.

$\Sigma'_* = \Sigma_* - c(v + \sigma^2)^{-1}c^\top$.

The same formula's covariance; $cc^\top$ is $m \times m$ of rank one.

Conditioning on $y_1, \dots, y_n$ and then on y_{n+1} is conditioning on all $n + 1$ at once.

$p(f_* \mid y_{1:n+1}) \propto p(f_*, y_{n+1} \mid y_{1:n})$ by the definition of conditional probability, applied to the posterior after n points as the new "prior"; the Gaussian family is closed under conditioning, so the sequential answer is the batch answer.

$$\mu'_* = \mu_* + \frac{c(y_{n+1} - \mu_{n+1})}{v + \sigma^2}, \Sigma'_* = \Sigma_* - \frac{cc^\top}{v + \sigma^2}, \text{ **identical to the batch posterior on } n + 1 \text{ points}**}$$

The update is the residual $y_{n+1} - \mu_{n+1}$ (what the new observation says beyond what was predicted) spread over the test points in proportion to their posterior covariance with the new input, divided by the predicted variance of the observation; the variance drops by a rank-one amount that again does not involve y_{n+1} . In the K_y picture this is a rank-one update of the inverse, the previous page's Sherman–Morrison formula, and it is the step a Kalman filter takes for one scalar measurement.

Problem 8

Take the linear kernel $k(x, x') = x^\top x'$ with inputs as rows of $X \in \mathbb{R}^{n \times d}$ and $X_* \in \mathbb{R}^{m \times d}$. Show that the posterior mean is $X_* w$ with $w = (X^\top X + \sigma^2 I)^{-1} X^\top y$, the ridge solution with penalty σ^2 , and that the posterior covariance is $\sigma^2 X_* (X^\top X + \sigma^2 I)^{-1} X_*^\top$.

$$K = XX^\top, K_* = XX_*^\top, K_{**} = X_* X_*^\top.$$

Entry (i, j) of each is the dot product of the corresponding rows.

$$X^\top (XX^\top + \sigma^2 I) = (X^\top X + \sigma^2 I) X^\top.$$

Multiply out: both sides are $X^\top XX^\top + \sigma^2 X^\top$.

$$(X^\top X + \sigma^2 I)^{-1} X^\top = X^\top (XX^\top + \sigma^2 I)^{-1}.$$

Multiply step 2 by $(X^\top X + \sigma^2 I)^{-1}$ on the left and $(XX^\top + \sigma^2 I)^{-1}$ on the right; both matrices are positive definite, hence invertible. This is the push-through identity: a $d \times d$ inverse on one side, an $n \times n$ inverse on the other.

$$\mu_* = K_*^\top K_y^{-1} y = X_* X^\top (X X^\top + \sigma^2 I)^{-1} y = X_* (X^\top X + \sigma^2 I)^{-1} X^\top y = X_* w.$$

Problem 1, then step 3.

$$\Sigma_* = X_* X_*^\top - X_* X^\top (X X^\top + \sigma^2 I)^{-1} X X^\top = X_* (I - (X^\top X + \sigma^2 I)^{-1} X^\top X) X_*^\top.$$

Step 3 applied inside the second term, then factor X_* and $X_*^\top$ out.

$$I - (X^\top X + \sigma^2 I)^{-1} X^\top X = (X^\top X + \sigma^2 I)^{-1} (X^\top X + \sigma^2 I - X^\top X) = \sigma^2 (X^\top X + \sigma^2 I)^{-1}.$$

Write I as $(X^\top X + \sigma^2 I)^{-1} (X^\top X + \sigma^2 I)$ and subtract.

$$\mu_* = X_* (X^\top X + \sigma^2 I)^{-1} X^\top y \text{ and } \Sigma_* = \sigma^2 X_* (X^\top X + \sigma^2 I)^{-1} X_*^\top$$

Gaussian process regression with a linear kernel is Bayesian linear regression with prior $w \sim \mathcal{N}(0, I)$: the mean is the ridge fit with $\lambda = \sigma^2$ (the regression page, in the units of $\frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$), and the covariance is X_* times the posterior covariance of w , which is $\sigma^2 (X^\top X + \sigma^2 I)^{-1}$. The two sides of step 3 are the two ways to compute it: in weight space ($d \times d$) when $d < n$, in function space ($n \times n$) when the kernel has no finite d , which is the point of kernels.

Problem 9

With the RBF kernel, find the posterior mean and variance at a test point as $\ell \rightarrow \infty$ and as $\ell \rightarrow 0$ (for x_* not equal to any training input).

$$\text{As } \ell \rightarrow \infty, k(x, x') \rightarrow s^2 \text{ for every pair, so } K \rightarrow s^2 \mathbf{1}\mathbf{1}^\top, k_* \rightarrow s^2 \mathbf{1} \text{ and } k_{**} = s^2.$$

$$\exp(-r^2/2\ell^2) \rightarrow 1.$$

$(s^2 \mathbf{1} \mathbf{1}^\top + \sigma^2 I) \mathbf{1} = (ns^2 + \sigma^2) \mathbf{1}$, so $K_y^{-1} \mathbf{1} = \frac{\mathbf{1}}{ns^2 + \sigma^2}$.	$\mathbf{1}^\top \mathbf{1} = n$: $\mathbf{1}$ is an eigenvector of K_y with eigenvalue $ns^2 + \sigma^2$, and the inverse has the reciprocal eigenvalue on the same vector.
$\mu_* = s^2 \mathbf{1}^\top K_y^{-1} y = \frac{s^2 \mathbf{1}^\top y}{ns^2 + \sigma^2} = \frac{ns^2}{ns^2 + \sigma^2} \bar{y}$.	$\mathbf{1}^\top K_y^{-1} = (K_y^{-1} \mathbf{1})^\top$ by symmetry, and $\mathbf{1}^\top y = n\bar{y}$.
$\Sigma_* = s^2 - s^4 \mathbf{1}^\top K_y^{-1} \mathbf{1} = s^2 - \frac{ns^4}{ns^2 + \sigma^2} = \frac{s^2 \sigma^2}{ns^2 + \sigma^2}$.	Step 2 again; put over the common denominator.
As $\ell \rightarrow 0$, $k(x, x') \rightarrow 0$ for $x \neq x'$, so $K \rightarrow s^2 I$, $k_* \rightarrow 0$ and $k_{**} = s^2$.	$\exp(-r^2/2\ell^2) \rightarrow 0$ for $r > 0$, and is 1 at $r = 0$.
$\mu_* \rightarrow 0$ and $\Sigma_* \rightarrow s^2$.	$k_* = 0$ kills both the mean and the correction to the variance.
$\ell \rightarrow \infty$: $\mu_* \rightarrow \frac{ns^2}{ns^2 + \sigma^2} \bar{y}$ and $\Sigma_* \rightarrow \frac{s^2 \sigma^2}{ns^2 + \sigma^2}$ everywhere ; $\ell \rightarrow 0$: $\mu_* \rightarrow 0$ and $\Sigma_* \rightarrow s^2$ away from the data	A very long lengthscale says the function is constant, so the model fits one number, the sample mean shrunk towards the prior mean 0 by the factor $ns^2/(ns^2 + \sigma^2)$, with the variance of a mean of n noisy observations; a very short one says the observations tell nothing about any other point, so the posterior is the prior except on the data. The marginal likelihood of Problem 5 picks the ℓ between these at which the data look most like a draw from the prior.

Problem 10

For a single test point, $\mu_* = k_*^\top K_y^{-1} y$ depends on ℓ through both k_* and K . Compute $\frac{d\mu_*}{d\ell}$ in terms of α , $\partial k_*/\partial \ell$ and $\partial K/\partial \ell$.

$$\frac{d\mu_*}{d\ell} = \left(\frac{\partial k_*}{\partial \ell} \right)^\top K_y^{-1} y + k_*^\top \frac{\partial K_y^{-1}}{\partial \ell} y.$$

Product rule on the three factors; y is constant.

$\frac{\partial K_y^{-1}}{\partial \ell} = -K_y^{-1} \frac{\partial K}{\partial \ell} K_y^{-1}.$	The previous page's Problem 4; $\partial K_y / \partial \ell = \partial K / \partial \ell$ because the noise term has no ℓ .
$\frac{d\mu_*}{d\ell} = \left(\frac{\partial k_*}{\partial \ell} \right)^\top \alpha - k_*^\top K_y^{-1} \frac{\partial K}{\partial \ell} \alpha.$	$K_y^{-1} y = \alpha$ in both terms.
The entries are $(\partial k_* / \partial \ell)_i = k(x_i, x_*)(x_i - x_*)^2 / \ell^3$ and $(\partial K / \partial \ell)_{ij} = K_{ij}(x_i - x_j)^2 / \ell^3$.	Problem 4.
$\frac{d\mu_*}{d\ell} = \left(\frac{\partial k_*}{\partial \ell} \right)^\top \alpha - k_*^\top K_y^{-1} \frac{\partial K}{\partial \ell} \alpha$	The first term is how the prediction changes because the test point's correlations with the data change; the second, through the inverse, is how it changes because the data's correlations with each other change, which reweights α . The second term needs one more solve with K_y (of $\frac{\partial K}{\partial \ell} \alpha$, or of k_* by symmetry), and dropping it is the last mistake below. The same two-term structure gives the derivative of any prediction with respect to any hyperparameter, which is what makes the predictions themselves differentiable for downstream use.

Where this goes wrong

1. Inverting K instead of $K + \sigma^2 I$

The conditioning formula is written with the prior covariance, and K is the prior covariance.

(f, f_*) has prior covariance with blocks K, K_*, K_{**}	Right so far: Problem 1, step 1.
“Condition f_* on the observed values using the prior covariance of what was observed.”	The slip that causes the mistake: what was observed is $y = f + \varepsilon$, whose covariance is $K + \sigma^2 I$, not K .

$$\mu_* = K_*^\top K^{-1} y \quad \times$$

This is the $\sigma^2 = 0$ posterior (Problem 3), which interpolates every noisy observation exactly, so the mean wiggles through the noise and the variance is 0 at the data, when the data are known to be noisy. It also needs K invertible, which a kernel matrix with two nearby inputs is not, numerically; $\sigma^2 I$ is what makes K_y well conditioned (Problem 6). In Problem 2 the wrong mean is 1.6480 in place of 1.5514 and the variance 0.0305 in place of 0.0873.

2. Noise added to the cross-covariance

A convenient implementation folds the noise into the kernel as $k_y(x, x') = k(x, x') + \sigma^2[x = x']$ and uses k_y for every block.

$K_y = K + \sigma^2 I$ is correct for the training block

Right so far: the diagonal of K is where the observation noise lives.

“Use the same noisy kernel k_y to build K_* , so a test point equal to a training input gets the same treatment.”

The shortcut that causes the mistake: the noise is a property of the observations, not of the function, and f_* is a function value; $\text{Cov}(f(x_*), y_i) = k(x_*, x_i)$ with no σ^2 even when $x_* = x_i$ (Problem 1, step 3).

$$(K_*)_{ij} = k(x_i, x_j^*) + \sigma^2 \text{ when } x_j^* = x_i \quad \times$$

At a test point that coincides with a training input the cross-covariance is too large by σ^2 , so the posterior mean leans towards that observation's noisy value and the posterior variance comes out too small (it can go negative). The bug is invisible at test points away from the data and appears exactly where predictions are compared with the training targets. The noise belongs in K_y only, and in $\Sigma_* + \sigma^2 I$ if a noisy y_* is being predicted.

3. Marginal likelihood gradient without the log-determinant

The data-fit term is the one with y in it, and its derivative is the one that looks like a gradient of a loss.

$$\frac{\partial}{\partial \theta} \left(-\frac{1}{2} y^\top K_y^{-1} y \right) = \frac{1}{2} \alpha^\top \frac{\partial K_y}{\partial \theta} \alpha$$

Right so far: Problem 5, steps 1 and 2.

“The log det term is a normalising constant.”

The assumption that causes the mistake: it is constant in y , not in θ ; it is how the marginal likelihood charges for a kernel that could explain anything.

$$\frac{\partial \log p}{\partial \theta} = \frac{1}{2} \alpha^\top \frac{\partial K_y}{\partial \theta} \alpha \quad \times$$

The term $-\frac{1}{2} \text{tr}(K_y^{-1} \partial K_y / \partial \theta)$ is missing (Problem 5). For $\theta = \sigma^2$ the surviving term is $\frac{1}{2} \|\alpha\|^2 > 0$, so the "gradient" always says to increase the noise; for s^2 it is $\frac{1}{2} \alpha^\top K \alpha / s^2 > 0$, so it always says to increase the signal variance; the optimiser drives the hyperparameters off to infinity, where the data fit is best because everything is explained as a draw from a huge prior. Problem 6's balance $\|\alpha\|^2 = \text{tr} K_y^{-1}$ exists only because the second term is there.

4. Lengthscale derivative short a factor of 2

The exponent $-r^2/(2\ell^2)$ has a 2 in it, and it is tempting to let it cancel something.

$$k = s^2 \exp \left(-r^2 / (2\ell^2) \right)$$

Right so far.

“Differentiate the exponent: $-\frac{r^2}{2} \cdot$
 $\frac{d}{d\ell} \ell^{-2} = -\frac{r^2}{2} \cdot (-\ell^{-3}).$ ”

The slip that causes the mistake:
 $\frac{d}{d\ell} \ell^{-2} = -2\ell^{-3}$, and the 2 is exactly what the $\frac{1}{2}$ in the exponent cancels.

$$\frac{\partial k}{\partial \ell} = k \frac{r^2}{2\ell^3} \quad \times$$

The derivative is $k r^2 / \ell^3$ (Problem 4): every entry of $\partial K / \partial \ell$ is half its true value, so the lengthscale gradient of the marginal likelihood (Problem 5) is halved. The stationary point is unchanged, because a zero gradient is still zero, so a gradient-based optimiser still finds the right ℓ , only more slowly; a finite-difference check of $\partial \log p / \partial \ell$ catches it at once, and a derivative-based test of $\mu'(x_*)$ does not, since that uses $\partial k / \partial x_*$.

5. Derivative of the posterior mean with the inverse held fixed

$\alpha = K_y^{-1}y$ is computed once and stored, and it is easy to forget that it moves with the kernel.

$$\mu_* = k_*^\top \alpha$$

Right so far: Problem 3.

“ α is the fitted weight vector; differentiate the kernel features k_* and leave the weights alone.”

The analogy that causes the mistake: in a linear model the weights are parameters, independent of the features; here $\alpha = K_y^{-1}y$ is a function of ℓ through K .

$$\frac{d\mu_*}{d\ell} = \left(\frac{\partial k_*}{\partial \ell} \right)^\top \alpha \quad \times$$

The second term $-k_*^\top K_y^{-1}(\partial K / \partial \ell)\alpha$ (Problem 10) is missing: it is the derivative through the inverse, $d(A^{-1}) = -A^{-1}(dA)A^{-1}$, and it is of the same order as the first. A gradient check against finite differences in ℓ fails. In the marginal likelihood (Problem 5) every dependence on ℓ runs through K_y , so there the derivative through the inverse is not a correction but the whole data-fit term.