The multivariate Gaussian: gradients and identities

Ten problems on the multivariate Gaussian: the log-density and its normalizing constant, the score $-\Sigma^{-1}(x-\mu)$, the maximum-likelihood mean and covariance through the gradients of $\log\det\Sigma$ and $\operatorname{tr}(\Sigma^{-1}A)$, affine maps, the reparameterization trick, conditionals by the Schur complement and the product of two Gaussian densities, with worked solutions and the mistakes that put Σ where Σ^{-1} or its square root belongs.

BEFORE YOU START

The multivariate Gaussian is the distribution behind least squares, the VAE's latent space, Gaussian processes, Kalman filters and the noise in diffusion models, and nearly every manipulation of it comes down to a handful of identities: the gradient of a log-determinant, the gradient of an inverse, the covariance of a linear map and a completed square. These ten problems derive each one, use them to find the maximum-likelihood mean and covariance, and then work through sampling with the reparameterization trick, conditioning on part of the vector, and multiplying two densities together. The five mistakes at the end are the places where Σ turns up in the wrong form: a normalizing constant with $\det\Sigma$ where its square root belongs, a score with Σ where its inverse belongs, a matrix inverse differentiated as if it were a number, samples scaled by the covariance instead of its square root, and a conditional variance read off the marginal block.

- $x, \mu \in \mathbb{R}^d$ are column vectors, and $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric positive definite: $\Sigma^\top = \Sigma$ and $v^\top \Sigma v > 0$ for every $v \neq 0$. Then Σ is invertible, $\det\Sigma > 0$, and Σ^{-1} is symmetric positive definite too. $\Lambda = \Sigma^{-1}$ is called the precision matrix.
- The density is $\mathcal{N}(x;\mu,\Sigma) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\big(-\tfrac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\big)$, and $x \sim \mathcal{N}(\mu,\Sigma)$ means x has this density. Its mean is $\mathbb{E}[x] = \mu$ and its covariance is $\operatorname{Cov}[x] = \mathbb{E}[(x-\mu)(x-\mu)^\top] = \Sigma$. $\log$ is the natural log.
- $q(x) = (x-\mu)^\top\Sigma^{-1}(x-\mu)$ is the squared Mahalanobis distance from x to μ; the density is constant on the ellipsoids $q(x) = \text{const}$.
- The conventions are those of the matrix-calculus page: a gradient has the shape of the variable, and for a scalar f of a matrix X, $\nabla_X f$ is the matrix of the $\partial f/\partial X_{ij}$. If a small change dX changes f by $df = \operatorname{tr}(M\,dX)$ to first order, then $\nabla_X f = M^\top$, because $\operatorname{tr}(M\,dX) = \sum_{i,j} M_{ji}\,dX_{ij}$.
- Gradients with respect to Σ treat all d^2 entries as free variables and are then evaluated at a symmetric Σ. The other convention, which counts $\Sigma_{ij} = \Sigma_{ji}$ as one variable, turns a symmetric gradient G into $2G - \operatorname{diag}(G)$, where $\operatorname{diag}(G)$ keeps only the diagonal of G. The two vanish together, so they give the same maximum-likelihood estimates.

- tr is the trace. It is cyclic, $\operatorname{tr}(ABC) = \operatorname{tr}(CAB)$, and for a vector v, $v^\top M v = \operatorname{tr}(M v v^\top)$.
- Data: N independent examples $x^{(1)}, \dots, x^{(N)}$ from $\mathcal{N}(\mu, \Sigma)$, the log-likelihood $\ell(\mu, \Sigma) = \sum_n \log \mathcal{N}(x^{(n)}; \mu, \Sigma)$, and the sample mean $\bar{x} = \frac{1}{N} \sum_n x^{(n)}$. The maximum-likelihood page does the one-dimensional version.
- Every symmetric positive definite Σ has a Cholesky factor: $\Sigma = L L^\top$ with L lower-triangular and its diagonal positive.
- For Problem 9, x is split into blocks x_a and x_b, with $\mu = (\mu_a; \mu_b)$ and $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, so that $\Sigma_{ba} = \Sigma_{ab}^\top$.

Builds on: Matrix calculus conventions, Maximum likelihood estimation: derivations by hand

Problems

1. Write $\log \mathcal{N}(x; \mu, \Sigma)$ term by term. Then evaluate it for $d = 2$, $\mu = (1, -1)^\top$, $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and $x = (2, 0)^\top$. These numbers return in Problems 2, 7 and 9. difficulty ·

2. Compute the score $\nabla_x \log \mathcal{N}(x; \mu, \Sigma)$. Evaluate it at Problem 1's numbers, and write out entry i when $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$. difficulty ·

3. Compute $\nabla_\mu \ell(\mu, \Sigma)$ for N examples and show that the maximum-likelihood mean is $\bar{x}$, whatever Σ is. difficulty ·

4. Show that $\nabla_\Sigma \log\det \Sigma = \Sigma^{-1}$. Work first with a general invertible X with $\det X > 0$, then specialize to a symmetric Σ. difficulty ··

5. Let $A \in \mathbb{R}^{d \times d}$ be symmetric. Show that $\nabla_\Sigma \operatorname{tr}(\Sigma^{-1} A) = -\Sigma^{-1} A \Sigma^{-1}$, and deduce $\nabla_\Sigma (v^\top \Sigma^{-1} v)$ for a fixed vector v. difficulty ··

6. Put $\mu = \bar{x}$ (Problem 3) and let $S = \sum_n (x^{(n)} - \bar{x})(x^{(n)} - \bar{x})^\top$, the scatter matrix, which is invertible when there are enough examples. Compute $\nabla_\Sigma \ell$ and solve $\nabla_\Sigma \ell = 0$ for the maximum-likelihood covariance. difficulty ···

7. Let $x \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ and $y = Ax + b$. Take as known that an affine image of a Gaussian vector is Gaussian. Find $\mathbb{E}[y]$ and $\mathrm{Cov}[y]$, then give the distribution of $x_1 + x_2$ for Problem 1's μ and Σ. difficulty ·

8. The reparameterization trick. Let $\Sigma = LL^\top$ be the Cholesky factorization, $\varepsilon \sim \mathcal{N}(0, I)$ and $x = \mu + L\varepsilon$. Show that $x \sim \mathcal{N}(\mu, \Sigma)$. Then, for a symmetric B and $f(x) = x^\top B x$, compute $F(\mu, L) = \mathbb{E}[f(x)]$ in closed form, find $\nabla_\mu F$ and $\nabla_L F$, and show that they equal $\mathbb{E}[\nabla f(x)]$ and $\mathbb{E}[\nabla f(x)\, \varepsilon^\top]$. difficulty ··

9. Split $x \sim \mathcal{N}(\mu, \Sigma)$ into blocks x_a and x_b, and let $K = \Sigma_{ab}\Sigma_{bb}^{-1}$. Using $z = x_a - K x_b$, find the distribution of x_a given x_b. Then find the distribution of x_1 given $x_2 = 0$ for Problem 1's μ and Σ. difficulty ···

10. Show that, as functions of x, $\mathcal{N}(x; a, A)\,\mathcal{N}(x; b, B)$ is proportional to $\mathcal{N}(x; c, C)$ with $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$. In one dimension, find c and C for $a = 0$, $A = 4$, $b = 3$ and $B = 2$. difficulty ··

Answers

1. $\log \mathcal{N}(x;\mu,\Sigma) = -\log(2\pi) - \tfrac{1}{2}\log 3 - \tfrac{1}{3} \approx -2.721$
2. $\nabla_x \log \mathcal{N}(x;\mu,\Sigma) = -\Sigma^{-1}(x-\mu)$, which is $(-\tfrac{1}{3}, -\tfrac{1}{3})^\top$ at Problem 1's numbers
3. $\nabla_\mu \ell = N\Sigma^{-1}(\bar{x} - \mu)$, so $\hat{\mu} = \bar{x}$ for every Σ
4. $\nabla_X \log\det X = X^{-\top}$, so $\nabla_\Sigma \log\det \Sigma = \Sigma^{-1}$
5. $\nabla_\Sigma \operatorname{tr}(\Sigma^{-1}A) = -\Sigma^{-1}A\Sigma^{-1}$ and $\nabla_\Sigma(v^\top\Sigma^{-1}v) = -\Sigma^{-1}vv^\top\Sigma^{-1}$
6. $\nabla_\Sigma \ell = -\tfrac{N}{2}\Sigma^{-1} + \tfrac{1}{2}\Sigma^{-1}S\Sigma^{-1}$, so $\hat{\Sigma} = \tfrac{1}{N}\sum_n (x^{(n)} - \bar{x})(x^{(n)} - \bar{x})^\top$
7. $y \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$; for Problem 1, $x_1 + x_2 \sim \mathcal{N}(0, 6)$
8. $\nabla_\mu F = 2B\mu = \mathbb{E}[\nabla f(x)]$ and $\nabla_L F = 2BL = \mathbb{E}[\nabla f(x)\,\varepsilon^\top]$
9. $x_a \mid x_b \sim \mathcal{N}\big(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b),\ \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\big)$; for Problem 1, $x_1 \mid x_2 = 0 \sim \mathcal{N}(\tfrac{3}{2}, \tfrac{3}{2})$
10. $\mathcal{N}(x;a,A)\,\mathcal{N}(x;b,B) \propto \mathcal{N}(x;c,C)$ with $C = (A^{-1} + B^{-1})^{-1}$, $c = C(A^{-1}a + B^{-1}b)$; for the 1-D numbers, $c = 2$ and $C = \tfrac{4}{3}$

Worked solutions

Problem 1

> Write $\log \mathcal{N}(x; \mu, \Sigma)$ term by term. Then evaluate it for $d = 2$, $\mu = (1, -1)^\top$, $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and $x = (2, 0)^\top$. These numbers return in Problems 2, 7 and 9.

$\log \mathcal{N}(x; \mu, \Sigma) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$	The log of a product is the sum of the logs, $\log c^{-1/2} = -\frac{1}{2}\log c$, and $\log$ undoes $\exp$. The first two terms do not depend on x: they are what make the density integrate to 1.
$-\frac{d}{2}\log(2\pi) = -\log(2\pi)$	$d = 2$.
$\det\Sigma = 2 \cdot 2 - 1 \cdot 1 = 3$	The determinant of a 2×2 matrix is $ad - bc$.
$\Sigma^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$	For a 2×2 matrix, swap the diagonal entries, negate the off-diagonal ones and divide by the determinant; multiplying by Σ gives I.
$x - \mu = (1, 1)^\top$, $\Sigma^{-1}(x-\mu) = \frac{1}{3}(1, 1)^\top$ and $q(x) = \frac{1}{3}(1 + 1) = \frac{2}{3}$	Each row of the matrix in step 4 sums to $2 - 1 = 1$, and q is the dot product of $x - \mu$ with that vector.
$\log \mathcal{N}(x; \mu, \Sigma) = -\log(2\pi) - \frac{1}{2}\log 3 - \frac{1}{3} \approx -2.721$	Steps 2 to 5 in step 1: $-1.838 - 0.549 - 0.333$. Only the last term depends on x, and it depends on x only through the Mahalanobis distance.

Problem 2

> Compute the score $\nabla_x \log \mathcal{N}(x; \mu, \Sigma)$. Evaluate it at Problem 1's numbers, and write out entry i when $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$.

Only $-\frac{1}{2}q(x)$ depends on x.	The first two terms of Problem 1, step 1 are constants, and a constant has zero gradient.

With $v = x - \mu$, $\nabla_v(v^\top \Sigma^{-1} v) = 2\Sigma^{-1} v$.	$\nabla_v(v^\top A v) = (A + A^\top)v$ (the matrix-calculus page, Problem 4), and Σ^{-1} is symmetric because $(\Sigma^{-1})^\top = (\Sigma^\top)^{-1} = \Sigma^{-1}$.
$\nabla_x q = 2\Sigma^{-1}(x - \mu)$	$\partial v / \partial x = I$: subtracting a constant does not change a derivative.
With $\Sigma = \mathrm{diag}(\sigma_i^2)$, $\Sigma^{-1} = \mathrm{diag}(1/\sigma_i^2)$ and entry i of the score is $-(x_i - \mu_i)/\sigma_i^2$.	The inverse of a diagonal matrix inverts each diagonal entry. Each coordinate gets the one-dimensional score, because with a diagonal Σ the density is a product of d one-dimensional densities and its log a sum.
$\nabla_x \log \mathcal{N}(x; \mu, \Sigma) = -\Sigma^{-1}(x - \mu)$, **which is $(-\frac{1}{3}, -\frac{1}{3})^\top$ at Problem 1's numbers**	Step 3 times $-\frac{1}{2}$; Problem 1, step 5 gives $\Sigma^{-1}(x - \mu)$. The score points back toward μ, weighting the directions of small variance most. Here $x - \mu = (1, 1)^\top$ is an eigenvector of Σ with eigenvalue 3, so the score points straight back at μ, shrunk by a factor 3. Score-based diffusion models train a network to approximate this vector field.

Problem 3

> Compute $\nabla_\mu \ell(\mu, \Sigma)$ for N examples and show that the maximum-likelihood mean is $\bar{x}$, whatever Σ is.

$\nabla_\mu \log \mathcal{N}(x^{(n)}; \mu, \Sigma) = \Sigma^{-1}(x^{(n)} - \mu)$	Problem 2's computation with $v = x^{(n)} - \mu$, whose Jacobian with respect to μ is $-I$; that flips the sign.
$\nabla_\mu \ell = \Sigma^{-1} \sum_n (x^{(n)} - \mu) = N\Sigma^{-1}(\bar{x} - \mu)$	The gradient of a sum is the sum of the gradients, Σ^{-1} factors out of the sum, and $\sum_n x^{(n)} = N\bar{x}$.
$\nabla_\mu \ell = 0 \iff \mu = \bar{x}$	Σ^{-1} is invertible, so it sends only the zero vector to zero.

$\nabla_\mu^2 \ell = -N\Sigma^{-1}$, which is negative definite.	Step 2 is affine in μ, so its Jacobian is the matrix that multiplies μ. Σ^{-1} is positive definite, so ℓ is strictly concave in μ and its stationary point is the global maximum.
$\nabla_\mu \ell = N\Sigma^{-1}(\bar{x} - \mu)$, **so** $\hat{\mu} = \bar{x}$ **for every** Σ	Σ dropped out in step 3, so the mean can be estimated before the covariance is known. That is what lets Problem 6 put $\bar{x}$ in place of μ.

Problem 4

> Show that $\nabla_\Sigma \log\det\Sigma = \Sigma^{-1}$. Work first with a general invertible X with $\det X > 0$, then specialize to a symmetric Σ.

$\det X = \sum_j X_{ij} C_{ij}$ for any fixed row i, where C_{ij} is $(-1)^{i+j}$ times the determinant of X with row i and column j deleted.	The cofactor (Laplace) expansion along row i.
$\partial \det X / \partial X_{ij} = C_{ij}$	Every cofactor in the row-i expansion deletes row i, so X_{ij} appears only as the explicit factor of the j-th term.
$X^{-1} = C^\top / \det X$	The inverse is the adjugate divided by the determinant, and the adjugate is the transpose of the matrix of cofactors.
$\partial \log\det X / \partial X_{ij} = C_{ij} / \det X = (X^{-1})_{ji}$	Chain rule through the log, whose derivative is $1/\det X$; then step 3 read at entry (j, i).
$\nabla_X \log\det X = X^{-\top}$, **so** $\nabla_\Sigma \log\det\Sigma = \Sigma^{-1}$	Step 4 puts entry (j, i) of X^{-1} at position (i, j), which is the transpose. For a symmetric Σ, $\Sigma^{-\top} = \Sigma^{-1}$. In differential form, $d \log\det\Sigma = \operatorname{tr}(\Sigma^{-1}\, d\Sigma)$, which Problem 6 uses.

Problem 5

Let $A \in \mathbb{R}^{d \times d}$ be symmetric. Show that $\nabla_\Sigma \operatorname{tr}(\Sigma^{-1} A) = -\Sigma^{-1} A \Sigma^{-1}$, and deduce $\nabla_\Sigma (v^\top \Sigma^{-1} v)$ for a fixed vector v.

$d(\Sigma^{-1}) = -\Sigma^{-1}\, d\Sigma\, \Sigma^{-1}$	Differentiate $\Sigma\Sigma^{-1} = I$ with the product rule, $d\Sigma\,\Sigma^{-1} + \Sigma\, d(\Sigma^{-1}) = 0$, and multiply on the left by Σ^{-1}. The factors keep their order because matrices do not commute.
$d \operatorname{tr}(\Sigma^{-1} A) = -\operatorname{tr}(\Sigma^{-1}\, d\Sigma\, \Sigma^{-1} A) = \operatorname{tr}\big(-\Sigma^{-1} A \Sigma^{-1}\, d\Sigma\big)$	The trace is linear, so the differential passes inside; cycling moves $\Sigma^{-1} A$ from the end to the front.
$\nabla_\Sigma \operatorname{tr}(\Sigma^{-1} A) = \big(-\Sigma^{-1} A \Sigma^{-1}\big)^\top = -\Sigma^{-1} A \Sigma^{-1}$	$df = \operatorname{tr}(M\, d\Sigma)$ means the gradient is $M^\top$ (Before you start). Transposing reverses the order of a product, and Σ^{-1} and A are symmetric.
$v^\top \Sigma^{-1} v = \operatorname{tr}(\Sigma^{-1} v v^\top)$	A scalar is its own trace, and the trace is cyclic; $vv^\top$ is symmetric.
$\nabla_\Sigma \operatorname{tr}(\Sigma^{-1} A) = -\Sigma^{-1} A \Sigma^{-1}$ **and** $\nabla_\Sigma (v^\top \Sigma^{-1} v) = -\Sigma^{-1} v v^\top \Sigma^{-1}$	Step 3, then step 4 with $A = vv^\top$. With $d = 1$ it is the scalar rule $\frac{d}{ds}(a/s) = -a/s^2$. The minus sign says that a larger covariance makes every point closer in Mahalanobis distance.

Problem 6

Put $\mu = \bar{x}$ (Problem 3) and let $S = \sum_n (x^{(n)} - \bar{x})(x^{(n)} - \bar{x})^\top$, the scatter matrix, which is invertible when there are enough examples. Compute $\nabla_\Sigma \ell$ and solve $\nabla_\Sigma \ell = 0$ for the maximum-likelihood covariance.

$\ell(\Sigma) = -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log\det \Sigma - \frac{1}{2} \operatorname{tr}(\Sigma^{-1} S)$	Sum Problem 1, step 1 over the N examples. Each quadratic term is $\operatorname{tr}(\Sigma^{-1} v v^\top)$ (Problem 5, step 4), and a sum of traces is the trace of the sum.

$\nabla_\Sigma \ell = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1} S \Sigma^{-1}$	Problem 4 for the log-determinant and Problem 5 with $A = S$, which is symmetric as a sum of outer products $vv^\top$.
$\nabla_\Sigma \ell = 0 \iff -N\Sigma + S = 0$	Multiply on the left and on the right by Σ, then by 2. Σ is invertible, so this can be undone and the two equations have the same solutions.
With $\Lambda = \Sigma^{-1}$, $\ell = \frac{N}{2}\log\det\Lambda - \frac{1}{2}\operatorname{tr}(\Lambda S) + \text{const}$, which is concave in Λ.	$\log\det\Lambda = -\log\det\Sigma$. $\log\det$ is concave on positive definite matrices (a standard fact) and $\operatorname{tr}(\Lambda S)$ is linear, so the one stationary point is the global maximum. S/N is positive definite because S is a positive semidefinite sum of outer products and invertible.
$\nabla_\Sigma \ell = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1} S \Sigma^{-1}$, **so** $\hat\Sigma = \frac{1}{N}\sum_n (x^{(n)} - \bar{x})(x^{(n)} - \bar{x})^\top$	It divides by N, not $N-1$: it is NumPy's `np.cov(X.T, bias=True)`, while the default `np.cov` divides by $N-1$. The centered examples span at most $N-1$ dimensions, so S is invertible only if $N > d$; otherwise ℓ has no maximum, because Σ can collapse onto the data's subspace and $\log\det\Sigma \to -\infty$.

Problem 7

> Let $x \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$ and $y = Ax + b$. Take as known that an affine image of a Gaussian vector is Gaussian. Find $\mathbb{E}[y]$ and $\operatorname{Cov}[y]$, then give the distribution of $x_1 + x_2$ for Problem 1's μ and Σ.

$\mathbb{E}[y] = A\,\mathbb{E}[x] + b = A\mu + b$	Expectation is linear, and A and b are constants.
$y - \mathbb{E}[y] = A(x - \mu)$	b cancels.
$\operatorname{Cov}[y] = \mathbb{E}\big[A(x-\mu)(x-\mu)^\top A^\top\big] = A\Sigma A^\top$	$(Av)^\top = v^\top A^\top$, and the constants A and $A^\top$ come out of the expectation on either side.

For $x_1 + x_2$: $A = (1, 1)$ and $b = 0$, so $A\mu = 1 - 1 = 0$ and $A\Sigma A^\top = 2 + 1 + 1 + 2 = 6$.	$(1,1)\,\Sigma\,(1,1)^\top$ adds up every entry of Σ.
$y \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$; **for Problem 1,** $x_1 + x_2 \sim \mathcal{N}(0, 6)$	A Gaussian is fixed by its mean and covariance, so steps 1 and 3 give the whole distribution. The variance 6 exceeds $2 + 2$ by twice the covariance $\Sigma_{12} = 1$. With $A = (I \;\; 0)$ the same result says that a block of a Gaussian vector is Gaussian, with the matching blocks of μ and Σ; Problem 9 uses that.

Problem 8

> The reparameterization trick. Let $\Sigma = LL^\top$ be the Cholesky factorization, $\varepsilon \sim \mathcal{N}(0, I)$ and $x = \mu + L\varepsilon$. Show that $x \sim \mathcal{N}(\mu, \Sigma)$. Then, for a symmetric B and $f(x) = x^\top B x$, compute $F(\mu, L) = \mathbb{E}[f(x)]$ in closed form, find $\nabla_\mu F$ and $\nabla_L F$, and show that they equal $\mathbb{E}[\nabla f(x)]$ and $\mathbb{E}[\nabla f(x)\,\varepsilon^\top]$.

$x \sim \mathcal{N}(\mu, LIL^\top) = \mathcal{N}(\mu, \Sigma)$	Problem 7 with $A = L$, $b = \mu$, and ε of mean 0 and covariance I.
$f(x) = \mu^\top B\mu + 2\mu^\top BL\varepsilon + \varepsilon^\top L^\top BL\varepsilon$	Expand $(\mu + L\varepsilon)^\top B(\mu + L\varepsilon)$. The two cross terms are equal scalars because B is symmetric: $(L\varepsilon)^\top B\mu = \mu^\top BL\varepsilon$.
$F = \mu^\top B\mu + \operatorname{tr}(L^\top BL)$	$\mathbb{E}[\varepsilon] = 0$ removes the middle term, and $\mathbb{E}[\varepsilon^\top M\varepsilon] = \mathbb{E}[\operatorname{tr}(M\varepsilon\varepsilon^\top)] = \operatorname{tr}(M\,\mathbb{E}[\varepsilon\varepsilon^\top]) = \operatorname{tr}(M)$.
$\nabla_\mu F = 2B\mu$ and $\nabla_L F = 2BL$	$\nabla_x(x^\top Bx) = (B + B^\top)x$ and $\nabla_X \operatorname{tr}(X^\top BX) = (B + B^\top)X$ (the matrix-calculus page, Problems 4 and 8), with $B = B^\top$.
$\nabla f(x) = 2Bx$, so $\mathbb{E}[\nabla f(x)] = 2B\mu$ and $\mathbb{E}[\nabla f(x)\,\varepsilon^\top] = 2B\big(\mu\,\mathbb{E}[\varepsilon]^\top + L\,\mathbb{E}[\varepsilon\varepsilon^\top]\big) = 2BL$	Substitute $x = \mu + L\varepsilon$ and use linearity, $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^\top] = I$.

$\nabla_\mu F = 2B\mu = \mathbb{E}[\nabla f(x)]$ **and** $\nabla_L F = 2BL = \mathbb{E}[\nabla f(x)\,\varepsilon^\top]$	This holds for any smooth f: $F = \mathbb{E}_\varepsilon[f(\mu + L\varepsilon)]$, the distribution of ε does not involve the parameters, so the gradient passes inside the expectation, and $\partial f(\mu + L\varepsilon)/\partial L_{ij} = (\nabla f)_i\,\varepsilon_j$. That is why a VAE samples ε rather than x: the average of $\nabla f(x)\,\varepsilon^\top$ over samples is an unbiased estimate of the gradient. Only the lower triangle of L is free, so its gradient is the lower triangle of $2BL$.

Problem 9

> Split $x \sim \mathcal{N}(\mu, \Sigma)$ into blocks x_a and x_b, and let $K = \Sigma_{ab}\Sigma_{bb}^{-1}$. Using $z = x_a - Kx_b$, find the distribution of x_a given x_b. Then find the distribution of x_1 given $x_2 = 0$ for Problem 1's μ and Σ.

(z, x_b) is jointly Gaussian.	It is the affine image of x under $\begin{pmatrix} I & -K \\ 0 & I \end{pmatrix}$, so Problem 7 applies.
$\mathrm{Cov}[z, x_b] = \Sigma_{ab} - K\Sigma_{bb} = \Sigma_{ab} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{bb} = 0$	The cross-covariance $\mathbb{E}[(z - \mathbb{E}z)(x_b - \mu_b)^\top]$ is linear in z, and $\mathrm{Cov}[x_a, x_b] = \Sigma_{ab}$, $\mathrm{Cov}[x_b, x_b] = \Sigma_{bb}$.
z is independent of x_b.	For jointly Gaussian vectors, zero cross-covariance makes the joint covariance block-diagonal, so the quadratic in the exponent and the determinant both split, and the density factorizes.
$\mathbb{E}[z] = \mu_a - K\mu_b$ and $\mathrm{Cov}[z] = \Sigma_{aa} - K\Sigma_{ba} - \Sigma_{ab}K^\top + K\Sigma_{bb}K^\top = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$	Problem 7 with $A = (I \;\; -K)$. Each of the last three terms equals $\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$, because Σ_{bb}^{-1} is symmetric and $\Sigma_{ab}^\top = \Sigma_{ba}$; two carry a minus sign and one a plus.
Given x_b, $x_a = z + Kx_b$ with z still distributed as in step 4.	By step 3, conditioning on x_b does not change the distribution of z, and Kx_b becomes a constant shift: the mean moves by Kx_b and the covariance stays.

For Problem 1, $K = \Sigma_{12}/\Sigma_{22} = \frac{1}{2}$, the mean is $1 + \frac{1}{2}\big(0 - (-1)\big) = \frac{3}{2}$ and the variance is $2 - \frac{1}{2} \cdot 1 = \frac{3}{2}$.	Every block is 1×1: $\mu_a = 1$, $\mu_b = -1$, $\Sigma_{aa} = \Sigma_{bb} = 2$ and $\Sigma_{ab} = 1$.
$x_a \mid x_b \sim \mathcal{N}\big(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b),\ \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\big)$; **for Problem 1,** $x_1 \mid x_2 = 0 \sim \mathcal{N}(\frac{3}{2}, \frac{3}{2})$	The covariance is the Schur complement of Σ_{bb} in Σ, which by the block-inverse formula is $(\Lambda_{aa})^{-1}$, the inverse of the precision's top-left block. It does not depend on the observed x_b and is never larger than Σ_{aa}: observing a correlated variable can only narrow the distribution. Gaussian-process prediction is this formula.

Problem 10

> Show that, as functions of x, $\mathcal{N}(x; a, A)\,\mathcal{N}(x; b, B)$ is proportional to $\mathcal{N}(x; c, C)$ with $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$. In one dimension, find c and C for $a = 0$, $A = 4$, $b = 3$ and $B = 2$.

$\log\big[\mathcal{N}(x; a, A)\,\mathcal{N}(x; b, B)\big] = -\frac{1}{2}(x - a)^\top A^{-1}(x - a) - \frac{1}{2}(x - b)^\top B^{-1}(x - b) + \text{const}$	The log of a product is the sum of the logs, and the normalizing constants do not depend on x.
$= -\frac{1}{2}x^\top(A^{-1} + B^{-1})x + x^\top(A^{-1}a + B^{-1}b) + \text{const}$	$(x - a)^\top A^{-1}(x - a) = x^\top A^{-1}x - 2x^\top A^{-1}a + a^\top A^{-1}a$, the two cross terms being equal because A^{-1} is symmetric; the same for b.
$-\frac{1}{2}(x - c)^\top C^{-1}(x - c) = -\frac{1}{2}x^\top C^{-1}x + x^\top C^{-1}c + \text{const}$	The same expansion.
$C^{-1} = A^{-1} + B^{-1}$ and $C^{-1}c = A^{-1}a + B^{-1}b$	Two quadratics in x with the same quadratic and linear parts differ by a constant, so the densities differ by a constant factor. $A^{-1} + B^{-1}$ is positive definite as a sum of positive definite matrices, so C exists.

One dimension: $C = (\frac{1}{4} + \frac{1}{2})^{-1} = \frac{4}{3}$ and $c = \frac{4}{3}\left(\frac{0}{4} + \frac{3}{2}\right) = 2$	Step 4 with 1×1 matrices.
$\mathcal{N}(x; a, A)\,\mathcal{N}(x; b, B) \propto \mathcal{N}(x; c, C)$ **with** $C = (A^{-1} + B^{-1})^{-1}$, $c = C(A^{-1}a + B^{-1}b)$; **for the 1-D numbers,** $c = 2$ **and** $C = \frac{4}{3}$	Precisions add, and the new mean is a precision-weighted average: 2 lies nearer 3 than 0 because $B = 2$ is the more certain of the two. The constant of proportionality is $\mathcal{N}(a; b, A + B)$. This is Bayes' rule for a Gaussian prior on a mean and one Gaussian measurement of it, the update step of a Kalman filter.

Where this goes wrong

1. Normalizing constant with $\det \Sigma$ where its square root belongs

In one dimension the constant is $1/\sqrt{2\pi\sigma^2}$, and σ^2 is the variance, so the matrix version seems to need only the covariance in its place.

$p(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\!\big(-(x-\mu)^2/(2\sigma^2)\big)$ for $d = 1$	Right so far: the scalar density.
"Replace 2π by $(2\pi)^d$ under the root and σ^2 by $\det \Sigma$, the variance of the whole vector."	The shortcut that causes the mistake: the square root is carried onto 2π and left off the determinant.
$\log \mathcal{N}(x; \mu, \Sigma) = -\frac{d}{2}\log(2\pi) - \log\det\Sigma - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$ ✗	The coefficient is $-\frac{1}{2}$ (Problem 1); this density integrates to $1/\sqrt{\det \Sigma}$, and at Problem 1's point it is low by $\frac{1}{2}\log 3 \approx 0.549$. The mean estimate is unaffected, which hides the slip, but setting the Σ-gradient to zero now gives $S/(2N)$, half the maximum-likelihood covariance of Problem 6.

2. Score with Σ in place of Σ⁻¹

In one dimension the score is $-(x-\mu)/\sigma^2$, a division, and a division by a matrix is easy to write as a multiplication.

$\log \mathcal{N}(x;\mu,\Sigma) = \text{const} - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$	Right so far: Problem 1, step 1.
"The score pulls x back toward μ, scaled by the spread of the distribution."	The shortcut that causes the mistake: reading the covariance as the scale of the pull, when it is the inverse that sits in the exponent.
$\nabla_x \log \mathcal{N}(x;\mu,\Sigma) = -\Sigma(x-\mu)$ ✗	The score is $-\Sigma^{-1}(x-\mu)$ (Problem 2). At Problem 1's numbers this gives $(-3,-3)^\top$ instead of $(-\frac{1}{3},-\frac{1}{3})^\top$. The directions of large variance should get the weakest pull and here get the strongest; the units are wrong too, since with x in meters the score is per meter and $\Sigma(x-\mu)$ is in cubic meters.

3. Derivative of the inverse taken as −Σ⁻² dΣ

For a number, $d(1/s) = -ds/s^2$, and Σ^{-2} looks like the matrix version.

$d\,\mathrm{tr}(\Sigma^{-1}A) = \mathrm{tr}\left(d(\Sigma^{-1})\,A\right)$	Right so far: the trace is linear.
"$d(\Sigma^{-1}) = -\Sigma^{-2}\,d\Sigma$, as for numbers."	The analogy that causes the mistake: the scalar rule assumes $d\Sigma$ commutes with Σ, and a general perturbation does not.
$\nabla_\Sigma \mathrm{tr}(\Sigma^{-1}A) = -\Sigma^{-2}A$ ✗	The gradient is $-\Sigma^{-1}A\Sigma^{-1}$ (Problem 5), from $d(\Sigma^{-1}) = -\Sigma^{-1}\,d\Sigma\,\Sigma^{-1}$. The wrong form is generally not even symmetric, while the true gradient is. The two agree when A commutes with Σ, for example $A = I$, so a test with an identity matrix will not catch it.

4. Sampling with x = μ + Σε

A one-dimensional sample is $\mu + \sigma\varepsilon$, and Σ is the matrix that comes to hand.

With $\varepsilon \sim \mathcal{N}(0, I)$, $x = \mu + M\varepsilon$ has covariance $MM^\top$.	Right so far: Problem 7 with $A = M$.
"The matrix that scales the noise is the covariance."	The analogy that causes the mistake: in one dimension the multiplier is the standard deviation σ, and Σ plays the part of σ^2, not σ.
$x = \mu + \Sigma\varepsilon$ ✗	Its covariance is $\Sigma\Sigma^\top = \Sigma^2$: for Problem 1, $\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ instead of $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. The multiplier must satisfy $MM^\top = \Sigma$, which the Cholesky factor L does (Problem 8).

5. Conditional covariance read off as the block Σ_aa

The conditional mean needs a formula, but the variance of x_a seems to be sitting in Σ already, in its top-left block.

For Problem 1, $\mathrm{Var}[x_1] = \Sigma_{11} = 2$	Right so far: that is the marginal variance (Problem 7 with $A = (1, 0)$).
"Knowing x_2 moves the center of x_1 but not its spread, so the variance stays Σ_{11}."	The shortcut that causes the mistake: conditioning on a correlated variable removes the part of x_1 that x_2 predicts, and that part carries variance.
$x_1 \mid x_2 \sim \mathcal{N}\big(1 + \tfrac{1}{2}(x_2 + 1),\ 2\big)$ ✗	The mean is right and the variance is $\Sigma_{11} - \Sigma_{12}^2/\Sigma_{22} = \tfrac{3}{2}$ (Problem 9). Using the marginal block gives intervals that are too wide; it is correct only when $\Sigma_{ab} = 0$, when there was nothing to learn from x_b.

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

eulerpad.app/practice/multivariate-gaussian