27

Laplace's Method

The idea behind the Laplace approximation is simple. We assume that an unnormalized probability density $P^*(x)$, whose normalizing constant

$$Z_P \equiv \int P^*(x)\,\mathrm{d}x \tag{27.1}$$

is of interest, has a peak at a point x_0. We Taylor-expand the logarithm of $P^*(x)$ around this peak:

$$\ln P^*(x) \simeq \ln P^*(x_0) - \frac{c}{2}(x - x_0)^2 + \cdots, \tag{27.2}$$

where

$$c = -\left.\frac{\partial^2}{\partial x^2} \ln P^*(x)\right|_{x=x_0}. \tag{27.3}$$

We then approximate $P^*(x)$ by an unnormalized Gaussian,

$$Q^*(x) \equiv P^*(x_0) \exp\left[-\frac{c}{2}(x - x_0)^2\right], \tag{27.4}$$

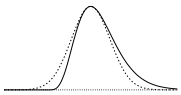
and we approximate the normalizing constant Z_P by the normalizing constant of this Gaussian,

$$Z_Q = P^*(x_0)\sqrt{\frac{2\pi}{c}}. \tag{27.5}$$

$P^*(x)$

$\ln P^*(x)$

$\ln P^*(x)$
& $\ln Q^*(x)$



$P^*(x)$
& $Q^*(x)$

We can generalize this integral to approximate Z_P for a density $P^*(\mathbf{x})$ over a K-dimensional space $\mathbf{x}$. If the matrix of second derivatives of $-\ln P^*(\mathbf{x})$ at the maximum $\mathbf{x}_0$ is $\mathbf{A}$, defined by:

$$A_{ij} = -\left.\frac{\partial^2}{\partial x_i \partial x_j} \ln P^*(\mathbf{x})\right|_{\mathbf{x}=\mathbf{x}_0}, \tag{27.6}$$

so that the expansion (27.2) is generalized to

$$\ln P^*(\mathbf{x}) \simeq \ln P^*(\mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathsf{T}}\mathbf{A}(\mathbf{x} - \mathbf{x}_0) + \cdots, \tag{27.7}$$

then the normalizing constant can be approximated by:

$$Z_P \simeq Z_Q = P^*(\mathbf{x}_0)\frac{1}{\sqrt{\det \frac{1}{2\pi}\mathbf{A}}} = P^*(\mathbf{x}_0)\sqrt{\frac{(2\pi)^K}{\det \mathbf{A}}}. \tag{27.8}$$

Predictions can be made using the approximation Q. Physicists also call this widely-used approximation the *saddle-point approximation*.

The fact that the normalizing constant of a Gaussian is given by

$$\int \mathrm{d}^K\mathbf{x}\, \exp\left[-\frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}\right] = \sqrt{\frac{(2\pi)^K}{\det \mathbf{A}}} \tag{27.9}$$

can be proved by making an orthogonal transformation into the basis $\mathbf{u}$ in which $\mathbf{A}$ is transformed into a diagonal matrix. The integral then separates into a product of one-dimensional integrals, each of the form

$$\int \mathrm{d}u_i \exp\left[-\frac{1}{2}\lambda_i u_i^2\right] = \sqrt{\frac{2\pi}{\lambda_i}}. \tag{27.10}$$

The product of the eigenvalues λ_i is the determinant of $\mathbf{A}$.

The Laplace approximation is basis-dependent: if x is transformed to a nonlinear function $u(x)$ and the density is transformed to $P(u) = P(x)\,|\mathrm{d}x/\mathrm{d}u|$ then in general the approximate normalizing constants Z_Q will be different. This can be viewed as a defect – since the true value Z_P is basis-independent – or an opportunity – because we can hunt for a choice of basis in which the Laplace approximation is most accurate.

▶ 27.1 Exercises

Exercise 27.1.[2] (See also exercise 22.8 (p.309).) A photon counter is pointed at a remote star for one minute, in order to infer the rate of photons arriving at the counter per minute, λ. Assuming the number of photons collected r has a Poisson distribution with mean λ,

$$P(r\,|\,\lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}, \tag{27.11}$$

and assuming the improper prior $P(\lambda) = 1/\lambda$, make Laplace approximations to the posterior distribution

(a) over λ

(b) over $\log\lambda$. [Note the improper prior transforms to $P(\log\lambda) =$ constant.]

▷ Exercise 27.2.[2] Use Laplace's method to approximate the integral

$$Z(u_1, u_2) = \int_{-\infty}^{\infty} \mathrm{d}a\, f(a)^{u_1}(1 - f(a))^{u_2}, \tag{27.12}$$

where $f(a) = 1/(1 + e^{-a})$ and u_1, u_2 are positive. Check the accuracy of the approximation against the exact answer (23.29, p.318) for $(u_1, u_2) = (1/2, 1/2)$ and $(u_1, u_2) = (1, 1)$. Measure the accuracy $(\log Z_P - \log Z_Q)$ in bits.

▷ Exercise 27.3.[3] Linear regression. N datapoints $\{(x^{(n)}, t^{(n)})\}$ are generated by the experimenter choosing each $x^{(n)}$, then the world delivering a noisy version of the linear function

$$y(x) = w_0 + w_1 x, \tag{27.13}$$

$$t^{(n)} \sim \mathrm{Normal}(y(x^{(n)}), \sigma_\nu^2). \tag{27.14}$$

Assuming Gaussian priors on w_0 and w_1, make the Laplace approximation to the posterior distribution of w_0 and w_1 (which is exact, in fact) and obtain the predictive distribution for the next datapoint $t^{(N+1)}$, given $x^{(N+1)}$.

(See MacKay (1992a) for further reading.)