Gaussian Process Kernels for Pattern Discovery and Extrapolation Supplementary Material

Andrew Gordon Wilson and Ryan Prescott Adams

1 Code

Please check `http://mlg.eng.cam.ac.uk/andrew/` for updates on code release.

2 Detailed Derivations for Spectral Mixture Kernels

A stationary kernel $k(x, x')$ is the inverse Fourier transform of its spectral density $S(s)$,

$$k(\tau) = \int S(s) e^{2\pi i s^\top \tau} ds\,, \tag{1}$$

where $\tau = x - x'$.

First suppose

$$S(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(s-\mu)^2\}\,, \tag{2}$$

where s, μ, σ and $\tau = x - x'$ are scalars. Substituting (2) into (1),

$$k(x, x') = \int \exp(2\pi i s(x - x')) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(s-\mu)^2) ds \tag{3}$$

let $\tau = x - x'$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp[2\pi i\tau - \frac{1}{2\sigma^2}(s^2 - 2\mu s + \mu^2)] ds \tag{4}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp[-\frac{1}{2\sigma^2} s^2 + (2\pi i\tau + \frac{\mu}{\sigma^2})s - \frac{\mu^2}{\sigma^2}] ds \tag{5}$$

let $a = \frac{1}{2\sigma^2}\,, b = 2\pi i\tau + \frac{\mu}{\sigma^2}\,, c = -\frac{\mu^2}{2\sigma^2}$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp(-a(s - \frac{b}{2a})^2) \exp(\frac{b^2}{4a} + c) ds \tag{6}$$

$$= \exp[(2\pi i\tau + \frac{\mu}{\sigma^2})^2 \frac{\sigma^2}{2} - \frac{\mu^2}{2\sigma^2}] \tag{7}$$

$$= \exp[(-4\pi^2\tau^2 + 4\pi i\tau \frac{\mu}{\sigma^2} + \frac{\mu^2}{\sigma^4})\frac{\sigma^2}{2} - \frac{\mu^2}{2\sigma^2}] \tag{8}$$

$$= \exp[-2\pi^2(x - x')^2\sigma^2][\cos(2\pi(x - x')\mu) + i\sin(2\pi(x - x')\mu))]\,. \tag{9}$$

Noting that the spectral density $S(s)$ must be symmetric (Rasmussen and Williams, 2006), we let

$$\phi(s\,;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(s-\mu)^2\}, \quad \text{and} \tag{10}$$

$$S(s) = [\phi(s) + \phi(-s)]/2\,. \tag{11}$$

Closely following the above derivation, substituting (11) into (1) gives

$$k(\tau) = \exp\{-2\pi^2\tau^2\sigma^2\}\cos(2\pi\tau\mu)\,. \tag{12}$$

If $\phi(s)$ is instead a mixture of Q Gaussians on $\mathbb{R}^P$, where the q^{th} component has mean vector $\boldsymbol{\mu}_q = (\mu_q^{(1)}, \ldots, \mu_q^{(P)})$ and covariance matrix $\mathbf{M}_q = \text{diag}(v_q^{(1)}, \ldots, v_q^{(P)})$, and τ_p is the p^{th} component of the P dimensional vector $\tau = x - x'$, then the integral in (1) becomes a sum of a product of the one dimensional integrals we encountered to derive (12), from which it follows that

$$k(\tau) = \sum_{q=1}^{Q} w_q \prod_{p=1}^{P} \exp\{-2\pi^2\tau_p^2 v_q^{(p)}\}\cos(2\pi\tau_p\mu_q^{(p)}). \tag{13}$$

3 Comment on Training Hyperparameters

Generally, we have had success naively training kernel hyperparameters using conjugate gradients (we use Carl Rasmussen's 2010 version of `minimize.m`) to maximize the marginal likelihood $p(\boldsymbol{y}|\theta)$ of the data $\boldsymbol{y}$ given hypers θ, having analytically integrated away a zero mean Gaussian process. We have found subtracting an empirical mean from the data prior to training hyperparameters (with conjugate gradients) undesirable, sometimes leading to local optima with lower marginal likelihoods, particularly on small datasets with a rising trend.

References

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for Machine Learning.* The MIT Press.