C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

Symbols and Notation

Matrices are capitalized and vectors are in bold type. We do not generally distinguish between probabilities and probability densities. A subscript asterisk, such as in X_*, indicates reference to a *test set* quantity. A superscript asterisk denotes complex conjugate.

Symbol	Meaning
$\backslash$	left matrix divide: $A\backslash\mathbf{b}$ is the vector $\mathbf{x}$ which solves $A\mathbf{x} = \mathbf{b}$
$\triangleq$	an equality which acts as a definition
$\overset{c}{=}$	equality up to an additive constant
$\lvert K\rvert$	determinant of K matrix
$\lvert\mathbf{y}\rvert$	Euclidean length of vector $\mathbf{y}$, i.e. $\big(\sum_i y_i^2\big)^{1/2}$
$\langle f, g\rangle_{\mathcal{H}}$	RKHS inner product
$\lVert f\rVert_{\mathcal{H}}$	RKHS norm
$\mathbf{y}^\top$	the transpose of vector $\mathbf{y}$
$\propto$	proportional to; e.g. $p(x\lvert y) \propto f(x, y)$ means that $p(x\lvert y)$ is equal to $f(x, y)$ times a factor which is independent of x
$\sim$	distributed according to; example: $x \sim \mathcal{N}(\mu, \sigma^2)$
∇ or $\nabla_{\mathbf{f}}$	partial derivatives (w.r.t. $\mathbf{f}$)
$\nabla\nabla$	the (Hessian) matrix of second derivatives
$\mathbf{0}$ or $\mathbf{0}_n$	vector of all 0's (of length n)
$\mathbf{1}$ or $\mathbf{1}_n$	vector of all 1's (of length n)
C	number of classes in a classification problem
$\mathrm{cholesky}(A)$	Cholesky decomposition: L is a lower triangular matrix such that $LL^\top = A$
$\mathrm{cov}(\mathbf{f}_*)$	Gaussian process posterior covariance
D	dimension of input space $\mathcal{X}$
$\mathcal{D}$	data set: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\lvert i = 1, \ldots, n\}$
$\mathrm{diag}(\mathbf{w})$	(vector argument) a diagonal matrix containing the elements of vector $\mathbf{w}$
$\mathrm{diag}(W)$	(matrix argument) a vector containing the diagonal elements of matrix W
δ_{pq}	Kronecker delta, $\delta_{pq} = 1$ iff $p = q$ and 0 otherwise
$\mathbb{E}$ or $\mathbb{E}_{q(x)}[z(x)]$	expectation; expectation of $z(x)$ when $x \sim q(x)$
$f(\mathbf{x})$ or $\mathbf{f}$	Gaussian process (or vector of) latent function values, $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top$
$\mathbf{f}_*$	Gaussian process (posterior) prediction (random variable)
$\bar{\mathbf{f}}_*$	Gaussian process posterior mean
$\mathcal{GP}$	Gaussian process: $f \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$, the function f is distributed as a Gaussian process with mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$
$h(\mathbf{x})$ or $\mathbf{h}(\mathbf{x})$	*either* fixed basis function (or set of basis functions) *or* weight function
H or $H(X)$	set of basis functions evaluated at all training points
I or I_n	the identity matrix (of size n)
$J_\nu(z)$	Bessel function of the first kind
$k(\mathbf{x}, \mathbf{x}')$	covariance (or kernel) function evaluated at $\mathbf{x}$ and $\mathbf{x}'$
K or $K(X, X)$	$n \times n$ covariance (or Gram) matrix
K_*	$n \times n_*$ matrix $K(X, X_*)$, the covariance between training and test cases
$\mathbf{k}(\mathbf{x}_*)$ or $\mathbf{k}_*$	vector, short for $K(X, \mathbf{x}_*)$, when there is only a single test case
K_f or K	covariance matrix for the (noise free) $\mathbf{f}$ values

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

Symbol	Meaning
K_y	covariance matrix for the (noisy) $\mathbf{y}$ values; for independent homoscedastic noise, $K_y = K_f + \sigma_n^2 I$
$K_\nu(z)$	modified Bessel function
$\mathcal{L}(a, b)$	loss function, the loss of predicting b, when a is true; note argument order
$\log(z)$	natural logarithm (base e)
$\log_2(z)$	logarithm to the base 2
ℓ or ℓ_d	characteristic length-scale (for input dimension d)
$\lambda(z)$	logistic function, $\lambda(z) = 1/\big(1 + \exp(-z)\big)$
$m(\mathbf{x})$	the mean function of a Gaussian process
μ	a measure (see section A.7)
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ or $\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \Sigma)$	(the variable $\mathbf{x}$ has a) Gaussian (Normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
$\mathcal{N}(\mathbf{x})$	short for unit Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$
n and n_*	number of training (and test) cases
N	dimension of feature space
N_H	number of hidden units in a neural network
$\mathbb{N}$	the natural numbers, the positive integers
$\mathcal{O}(\cdot)$	big Oh; for functions f and g on $\mathbb{N}$, we write $f(n) = \mathcal{O}(g(n))$ if the ratio $f(n)/g(n)$ remains bounded as $n \to \infty$
O	*either* matrix of all zeros *or* differential operator
$y\|x$ and $p(y\|x)$	conditional random variable y given x and its probability (density)
$\mathbb{P}_N$	the regular n-polygon
$\boldsymbol{\phi}(\mathbf{x}_i)$ or $\Phi(X)$	feature map of input $\mathbf{x}_i$ (or input set X)
$\Phi(z)$	cumulative unit Gaussian: $\Phi(z) = (2\pi)^{-1/2}\int_{-\infty}^{z} \exp(-t^2/2)dt$
$\pi(\mathbf{x})$	the sigmoid of the latent value: $\pi(\mathbf{x}) = \sigma(f(\mathbf{x}))$ (stochastic if $f(\mathbf{x})$ is stochastic)
$\hat{\pi}(\mathbf{x}_*)$	MAP prediction: π evaluated at $\bar{f}(\mathbf{x}_*)$.
$\bar{\pi}(\mathbf{x}_*)$	mean prediction: expected value of $\pi(\mathbf{x}_*)$. Note, in general that $\hat{\pi}(\mathbf{x}_*) \neq \bar{\pi}(\mathbf{x}_*)$
$\mathbb{R}$	the real numbers
$R_{\mathcal{L}}(f)$ or $R_{\mathcal{L}}(c)$	the risk or expected loss for f, or classifier c (averaged w.r.t. inputs and outputs)
$\tilde{R}_{\mathcal{L}}(l\|\mathbf{x}_*)$	expected loss for predicting l, averaged w.r.t. the model's pred. distr. at $\mathbf{x}_*$
$\mathcal{R}_c$	decision region for class c
$S(\mathbf{s})$	power spectrum
$\sigma(z)$	any sigmoid function, e.g. logistic $\lambda(z)$, cumulative Gaussian $\Phi(z)$, etc.
σ_f^2	variance of the (noise free) signal
σ_n^2	noise variance
$\boldsymbol{\theta}$	vector of hyperparameters (parameters of the covariance function)
$\mathrm{tr}(A)$	trace of (square) matrix A
$\mathbb{T}_l$	the circle with circumference l
$\mathbb{V}$ or $\mathbb{V}_{q(x)}[z(x)]$	variance; variance of $z(x)$ when $x \sim q(x)$
$\mathcal{X}$	input space and also the index set for the stochastic process
X	$D \times n$ matrix of the training inputs $\{\mathbf{x}_i\}_{i=1}^n$: the design matrix
X_*	matrix of test inputs
$\mathbf{x}_i$	the ith training input
x_{di}	the dth coordinate of the ith training input $\mathbf{x}_i$
$\mathbb{Z}$	the integers $\ldots, -2, -1, 0, 1, 2, \ldots$