Notation

“ · ”	event (in probability)
$\{\cdots\}$	set
$\lvert \cdot \rvert$	absolute value of a number, or cardinality (number of elements) of a set, or determinant of a matrix
$\lVert \cdot \rVert^2$	square of the norm; sum of the squared components of a vector
$\lfloor \cdot \rfloor$	floor; largest integer which is not larger than the argument
$[a, b]$	the interval of real numbers from a to b
$[\![\cdot]\!]$	evaluates to 1 if argument is true, and to 0 if it is false
∇	gradient operator, e.g., ∇E_{in} (gradient of $E_{\text{in}}(\mathbf{w})$ with respect to $\mathbf{w}$)
$(\cdot)^{-1}$	inverse
$(\cdot)^{\dagger}$	pseudo-inverse
$(\cdot)^{\mathrm{T}}$	transpose (columns become rows and vice versa)
$\binom{N}{k}$	number of ways to choose k objects from N distinct objects (equals $\frac{N!}{(N-k)!k!}$ where ‘!’ is the factorial)
$A \setminus B$	the set A with the elements from set B removed
$\mathbf{0}$	zero vector; a column vector whose components are all zeros
$\{1\} \times \mathbb{R}^d$	d-dimensional Euclidean space with an added ‘zeroth coordinate’ fixed to 1
ϵ	tolerance in approximating a target
δ	bound on the probability of exceeding ϵ (the approximation tolerance)
η	learning rate (step size in iterative learning, e.g., in stochastic gradient descent)
λ	regularization parameter
λ_C	regularization parameter corresponding to weight budget C
Ω	penalty for model complexity; either a bound on generalization error, or a regularization term
θ	logistic function $\theta(s) = e^s/(1 + e^s)$
Φ	feature transform, $\mathbf{z} = \Phi(\mathbf{x})$
Φ_{Q}	Qth-order polynomial transform

ϕ	a coordinate in the feature transform Φ, $z_i = \phi_i(\mathbf{x})$
μ	probability of a binary outcome
ν	fraction of a binary outcome in a sample
σ^2	variance of noise
$\mathcal{A}$	learning algorithm
$\mathrm{argmin}_a(\cdot)$	the value of a at which the minimum of the argument is achieved
$\mathcal{B}$	an event (in probability), usually 'bad' event
b	the bias term in a linear combination of inputs, also called w_0
bias	the bias term in bias-variance decomposition
$B(N, k)$	maximum number of dichotomies on N points with a break point k
C	bound on the size of weights in the soft order constraint
d	dimensionality of the input space $\mathcal{X} = \mathbb{R}^d$ or $\mathcal{X} = \{1\} \times \mathbb{R}^d$
$\tilde{d}$	dimensionality of the transformed space $\mathcal{Z}$
d_{VC},$d_{\text{VC}}(\mathcal{H})$	VC dimension of hypothesis set $\mathcal{H}$
$\mathcal{D}$	data set $\mathcal{D} = (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$; technically not a set, but a vector of elements $(\mathbf{x}_n, y_n)$. $\mathcal{D}$ is often the training set, but sometimes split into training and validation/test sets.
$\mathcal{D}_{\text{train}}$	subset of $\mathcal{D}$ used for training when a validation or test set is used.
$\mathcal{D}_{\text{val}}$	validation set; subset of $\mathcal{D}$ used for validation.
$E(h, f)$	error measure between hypothesis h and target function f
e^x	exponent of x in the natural base $e = 2.71828\cdots$
$\mathsf{e}(h(\mathbf{x}), f(\mathbf{x}))$	pointwise version of $E(h, f)$, e.g., $(h(\mathbf{x}) - f(\mathbf{x}))^2$
e_n	leave-one-out error on example n when this nth example is excluded in training [cross validation]
$\mathbb{E}[\cdot]$	expected value of argument
$\mathbb{E}_{\mathbf{x}}[\cdot]$	expected value with respect to $\mathbf{x}$
$\mathbb{E}[y\|\mathbf{x}]$	expected value of y given $\mathbf{x}$
E_{aug}	augmented error (in-sample error plus regularization term)
E_{in}, $E_{\text{in}}(h)$	in-sample error (training error) for hypothesis h
E_{cv}	cross validation error
E_{out}, $E_{\text{out}}(h)$	out-of-sample error for hypothesis h
$E_{\text{out}}^{\mathcal{D}}$	out-of-sample error when $\mathcal{D}$ is used for training
$\bar{E}_{\text{out}}$	expected out-of-sample error
E_{val}	validation error
E_{test}	test error
f	target function, $f\colon \mathcal{X} \to \mathcal{Y}$
g	final hypothesis $g \in \mathcal{H}$ selected by the learning algorithm; $g\colon \mathcal{X} \to \mathcal{Y}$
$g^{(\mathcal{D})}$	final hypothesis when the training set is $\mathcal{D}$
$\bar{g}$	average final hypothesis [bias-variance analysis]

g^-	final hypothesis when trained using $\mathcal{D}$ *minus* some points
$\mathbf{g}$	gradient, e.g., $\mathbf{g} = \nabla E_{\text{in}}$
h	a hypothesis $h \in \mathcal{H}$; $h\colon \mathcal{X} \to \mathcal{Y}$
$\tilde{h}$	a hypothesis in transformed space $\mathcal{Z}$
$\mathcal{H}$	hypothesis set
$\mathcal{H}_\Phi$	hypothesis set that corresponds to perceptrons in Φ-transformed space
$\mathcal{H}(C)$	restricted hypothesis set by weight budget C [soft order constraint]
$\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$	dichotomies (patterns of ± 1) generated by $\mathcal{H}$ on the points $\mathbf{x}_1, \cdots, \mathbf{x}_N$
H	The hat matrix [linear regression]
I	identity matrix; square matrix whose diagonal elements are 1 and off-diagonal elements are 0
K	size of validation set
L_q	qth-order Legendre polynomial
$\ln$	logarithm in base e
$\log_2$	logarithm in base 2
M	number of hypotheses
$m_{\mathcal{H}}(N)$	the growth function; maximum number of dichotomies generated by $\mathcal{H}$ on any N points
$\max(\cdot, \cdot)$	maximum of the two arguments
N	number of examples (size of $\mathcal{D}$)
$o(\cdot)$	absolute value of this term is asymptotically negligible compared to the argument
$O(\cdot)$	absolute value of this term is asymptotically smaller than a constant multiple of the argument
$P(\mathbf{x})$	(marginal) probability or probability density of $\mathbf{x}$
$P(y \mid \mathbf{x})$	conditional probability or probability density of y given $\mathbf{x}$
$P(\mathbf{x}, y)$	joint probability or probability density of $\mathbf{x}$ and y
$\mathbb{P}[\cdot]$	probability of an event
Q	order of polynomial transform
Q_f	complexity of f (order of polynomial defining f)
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^d$	d-dimensional Euclidean space
s	signal $s = \mathbf{w}^{\mathrm{T}}\mathbf{x} = \sum_i w_i x_i$ (i goes from 0 to d or 1 to d depending on whether $\mathbf{x}$ has the $x_0 = 1$ coordinate or not)
$\mathrm{sign}(\cdot)$	sign function, returning $+1$ for positive and -1 for negative
$\sup_a(.)$	supremum; smallest value that is $\geq$ the argument for all a
T	number of iterations, number of epochs
t	iteration number or epoch number
$\tanh(\cdot)$	hyperbolic tangent function; $\tanh(s) = (e^s - e^{-s})/(e^s + e^{-s})$
$\mathrm{trace}(\cdot)$	trace of square matrix (sum of diagonal elements)
V	number of subsets in V-fold cross validation ($V \times K = N$)
$\mathbf{v}$	direction in gradient descent (not necessarily a unit vector)

$\hat{\mathbf{v}}$	unit vector version of $\mathbf{v}$ [gradient descent]
var	the variance term in bias-variance decomposition
$\mathbf{w}$	weight vector (column vector)
$\tilde{\mathbf{w}}$	weight vector in transformed space $\mathcal{Z}$
$\hat{\mathbf{w}}$	selected weight vector [pocket algorithm]
$\mathbf{w}^*$	weight vector that separates the data
$\mathbf{w}_{\text{lin}}$	solution weight vector to linear regression
$\mathbf{w}_{\text{reg}}$	regularized solution to linear regression with weight decay
$\mathbf{w}_{\text{PLA}}$	solution weight vector of perceptron learning algorithm
w_0	added coordinate in weight vector $\mathbf{w}$ to represent bias b
$\mathbf{x}$	the input $\mathbf{x} \in \mathcal{X}$. Often a column vector $\mathbf{x} \in \mathbb{R}^d$ or $\mathbf{x} \in \{1\} \times \mathbb{R}^d$. x is used if input is scalar.
x_0	added coordinate to $\mathbf{x}$, fixed at $x_0 = 1$ to absorb the bias term in linear expressions
$\mathcal{X}$	input space whose elements are $\mathbf{x} \in \mathcal{X}$
X	matrix whose rows are the data inputs $\mathbf{x}_n$ [linear regression]
XOR	exclusive OR function (returns 1 if the number of 1's in its input is odd)
y	the output $y \in \mathcal{Y}$
$\mathbf{y}$	column vector whose components are the data set outputs y_n [linear regression]
$\hat{\mathbf{y}}$	estimate of $\mathbf{y}$ [linear regression]
$\mathcal{Y}$	output space whose elements are $y \in \mathcal{Y}$
$\mathcal{Z}$	transformed input space whose elements are $\mathbf{z} = \Phi(\mathbf{x})$
Z	matrix whose rows are the transformed inputs $\mathbf{z}_n = \Phi(\mathbf{x}_n)$ [linear regression]