Table of notation

ℓ	Loss function
$\mathcal{H}$, $\mathcal{F}$ or $\mathcal{B}$	Classes of functions
$R(h)$	Risk $\mathbb{E}(\ell(h(X), Y))$ for *fixed* h
$\hat{R}(h)$	Empirical risk $\frac{1}{n}\sum_{i=1}^n \ell(h(X_i), Y_i)$
h_0	Typically used to denote a minimiser of the risk over all functions $h : \mathcal{X} \to \mathcal{Y}$ e.g. a Bayes classifier in the classification setting or $\mathbb{E}(Y \mid X = \cdot)$ in a regression setting
h^*	Minimiser of the risk R over $\mathcal{H}$
$\hat{h}$	Minimiser of the empirical risk $\hat{R}$ over $\mathcal{H}$; this is random
$R(\hat{h})$	Risk $\mathbb{E}(\ell(\hat{h}(X), Y) \mid (X_i, Y_i)_{i=1}^n)$ for *random* $\hat{h}$; this is a random quantity depending on the data $(X_i, Y_i)_{i=1}^n$ used to train $\hat{h}$
$U \overset{d}{=} V$	U has the same distribution as V
ε_i	Typically a Rademacher random variable which takes values $1, -1$ each with probability $1/2$
$U \perp\!\!\!\perp V$	U is independent of V
$z_{1:n}$	Shorthand for $(z_1, \ldots, z_n)$
$\mathcal{F}(z_{1:n})$	Set of 'behaviours' of $\mathcal{F}$ on $z_{1:n}$ i.e. $\{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$
$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$	Empirical Rademacher complexity $\mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i) \mid Z_{1:n} = z_{1:n}\right) = \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(z_i)\right)$
$\mathcal{R}_n(\mathcal{F})$	Rademacher complexity $\mathbb{E}\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))$
$s(\mathcal{H}, n)$	Shattering coefficient $\max_{x_{1:n}\in\mathcal{X}^n} \lvert\mathcal{H}(x_{1:n})\rvert$
$\mathrm{VC}(\mathcal{H})$	VC dimension $\sup\{n \in \mathbb{N} : s(\mathcal{H}, n) = 2^n\}$
$R_\phi(h)$	ϕ-risk $\mathbb{E}\phi(Yh(X))$ of h
$\hat{R}_\phi(h)$	Empirical ϕ-risk $\frac{1}{n}\sum_{i=1}^n \phi(Y_i h(X_i))$ of h
$\operatorname{conv} S$	Convex hull of set S
$\pi_C(x)$	Projection of x onto closed convex set C
$\partial f(x)$	Subdifferential of f at x

Review of least squares regression

Given $\Phi \in \mathbb{R}^{n\times d}$ of full column rank with ith row $\varphi_i \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$, we have that

$$\operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \varphi_i^\top \beta)^2 = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \|y - \Phi\beta\|_2^2 = (\Phi^\top\Phi)^{-1}\Phi^\top y.$$

Indeed, we have that $\|\Phi z\|_2^2 = z^\top\Phi^\top\Phi z = 0$ if and only if $z = 0$, so $\Phi^\top\Phi$ is invertible. Then $P := \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$ is an orthogonal projection matrix onto the columns space of Φ (one can easily check that is is symmetric and $P^2 = P$). Thus

$$\begin{aligned}\|y - \Phi\beta\|_2^2 &= \|y - PY + Py - \Phi\beta\|_2^2 \\ &= \|(I-P)y\|_2^2 + \|Py - \Phi\beta\|_2^2\end{aligned}$$

since the cross term

$$\{(I-P)y\}^\top(Py - \Phi\beta) = y^\top(I-P)(Py - \Phi\beta) = 0$$

as $I - P$ sends everything in the column space of Φ to 0. [Note that this is a generalisation of the decomposition

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \mu)^2 + n(\mu - \bar{y})^2$$

where $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$; see for example the first display in the proof of Theorem 2, or the alternative expression for Q_m on page 9. To see the correspondence, take Φ to be a $n \times 1$ matrix of ones.] Thus to minimise the least squares term we require

$$\Phi(\Phi^\top\Phi)^{-1}\Phi^\top y = \Phi\beta$$

which multiplying on the left by $(\Phi^\top\Phi)^{-1}\Phi^\top$ gives that the minimiser is $\hat{\beta} = (\Phi^\top\Phi)^{-1}\Phi^\top y$.

An alternative derivation involves differentiating the least squares objective with respect to β to give gradient vector

$$\frac{\partial}{\partial\beta}\|y - \Phi\beta\|_2^2 = -2\Phi^\top(y - \Phi\beta)$$

Setting this to 0 once more gives the minimiser $\hat{\beta} = (\Phi^\top\Phi)^{-1}\Phi^\top y$. Indeed, the objective is strictly convex as the Hessian matrix $2\Phi^\top\Phi$ is positive definite, so this must be the unique minimiser of the objective.