1 Moore-Penrose Pseudoinverse

Throughout, we use the defining property of the inverse of a nonsingular square matrix $\mathbf{A}$: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

Problem 1.1. Group the last two factors and use the property above with $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$:

$$\begin{aligned}\mathbf{X}^+\mathbf{X} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}) \\ &= \mathbf{I}.\end{aligned}$$

Problem 1.2. $\mathbf{I}$ has the same dimension as $\mathbf{X}^\top\mathbf{X}$. Since $\mathbf{X}^\top \in \mathbb{R}^{M\times N}$ and $\mathbf{X} \in \mathbb{R}^{N\times M}$, we have $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{M\times M}$, so $\mathbf{I}$ is the $M \times M$ identity matrix.

Problem 2.1. Multiplying $\mathbf{X}$ on the *left* by $\mathbf{X}^+$ gives the identity, $\mathbf{X}^+\mathbf{X} = \mathbf{I}$ (Problem 1.1), so $\mathbf{X}^+$ acts as an inverse from the left. Multiplying on the right gives the $N \times N$ matrix

$$\mathbf{X}\mathbf{X}^+ = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

which in general is not the identity. Its rank is at most M, so when $N > M$ (more data points than weights) it cannot equal the $N \times N$ identity. In linear regression it maps the targets to the least-squares fit: with $\mathbf{w} = \mathbf{X}^+\mathbf{t}$, the fitted values are $\mathbf{X}\mathbf{w} = \mathbf{X}\mathbf{X}^+\mathbf{t}$.

Problem 2.2. A right-inverse $\mathbf{X}^R$ should satisfy $\mathbf{X}\mathbf{X}^R = \mathbf{I}$. By analogy with $\mathbf{X}^+$, take

$$\mathbf{X}^R = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}.$$

Check: $\mathbf{X}\mathbf{X}^R = (\mathbf{X}\mathbf{X}^\top)(\mathbf{X}\mathbf{X}^\top)^{-1} = \mathbf{I}$, the $N \times N$ identity.

Problem 2.3. We need the $N \times N$ matrix $\mathbf{X}\mathbf{X}^\top$ to be nonsingular. This requires $\mathbf{X}$ to have rank N, which is only possible when $N \le M$ (at least as many columns as rows). When $N < M$, the formula $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is undefined, because the $M \times M$ matrix $\mathbf{X}^\top\mathbf{X}$ has rank at most $N < M$ and is therefore singular; in fact $\mathbf{X}$ then has no left-inverse at all.

Problem 3.1. Insert $\mathbf{I} = \mathbf{X}\mathbf{X}^{-1}$ on the right and regroup:

$$\begin{aligned}\mathbf{X}\mathbf{X}^+ &= \mathbf{X}\mathbf{X}^+\mathbf{I} \\ &= \mathbf{X}\mathbf{X}^+\mathbf{X}\mathbf{X}^{-1} && \text{since } \mathbf{X}\mathbf{X}^{-1} = \mathbf{I} \\ &= \mathbf{X}(\mathbf{X}^+\mathbf{X})\mathbf{X}^{-1} \\ &= \mathbf{X}\mathbf{I}\mathbf{X}^{-1} && \text{by Problem 1.1} \\ &= \mathbf{X}\mathbf{X}^{-1} \\ &= \mathbf{I}.\end{aligned}$$

Problem 3.2. Insert $\mathbf{I} = \mathbf{X}^{-1}\mathbf{X}$ on the left and regroup:

$$\begin{aligned}
\mathbf{X}^{+} &= \mathbf{I}\mathbf{X}^{+} \\
&= \mathbf{X}^{-1}\mathbf{X}\mathbf{X}^{+} && \text{since } \mathbf{X}^{-1}\mathbf{X} = \mathbf{I} \\
&= \mathbf{X}^{-1}(\mathbf{X}\mathbf{X}^{+}) \\
&= \mathbf{X}^{-1}\mathbf{I} && \text{by Problem 3.1} \\
&= \mathbf{X}^{-1}.
\end{aligned}$$

Alternatively, for square nonsingular $\mathbf{X}$ we have $(\mathbf{X}^{\top}\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}^{\top})^{-1}$, so $\mathbf{X}^{+} = \mathbf{X}^{-1}(\mathbf{X}^{\top})^{-1}\mathbf{X}^{\top} = \mathbf{X}^{-1}$.

When $\mathbf{X}$ is square and nonsingular, the pseudoinverse is the ordinary inverse. The least-squares solution is then $\mathbf{w} = \mathbf{X}^{-1}\mathbf{t}$, which fits every training target exactly.

2 Vector Calculus

Problem 1.1. Write the inner product as a sum,

$$f(\mathbf{w}) = \mathbf{w}^\top \mathbf{x} = \sum_{k=1}^{N} w_k x_k.$$

Only the term with $k = i$ depends on w_i, so $\dfrac{\partial f}{\partial w_i} = x_i$.

Problem 1.2. Stack the partial derivatives from Problem 1.1:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \begin{bmatrix} \frac{\partial f}{\partial w_1} & \cdots & \frac{\partial f}{\partial w_N} \end{bmatrix}^\top = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}^\top = \mathbf{x}.$$

Problem 2.1. $f(\mathbf{w}) = \mathbf{w}^\top \mathbf{w} = \sum_{k=1}^{N} w_k^2$. Only the term with $k = i$ depends on w_i, so $\dfrac{\partial f}{\partial w_i} = 2w_i$.

Problem 2.2.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \begin{bmatrix} 2w_1 & \cdots & 2w_N \end{bmatrix}^\top = 2\mathbf{w}.$$

Problem 3.1. The partial derivative of a sum is the sum of the partial derivatives:

$$\frac{\partial f}{\partial w_i} = \frac{\partial g}{\partial w_i} + \frac{\partial h}{\partial w_i}.$$

Problem 3.2. Stack the partial derivatives and split the vector into two:

$$\begin{aligned} \nabla_{\mathbf{w}} f(\mathbf{w}) &= \begin{bmatrix} \frac{\partial g}{\partial w_1} + \frac{\partial h}{\partial w_1} & \cdots & \frac{\partial g}{\partial w_N} + \frac{\partial h}{\partial w_N} \end{bmatrix}^\top \\ &= \begin{bmatrix} \frac{\partial g}{\partial w_1} & \cdots & \frac{\partial g}{\partial w_N} \end{bmatrix}^\top + \begin{bmatrix} \frac{\partial h}{\partial w_1} & \cdots & \frac{\partial h}{\partial w_N} \end{bmatrix}^\top \\ &= \nabla_{\mathbf{w}} g(\mathbf{w}) + \nabla_{\mathbf{w}} h(\mathbf{w}). \end{aligned}$$

Problem 4.1. Let $\mathbf{A} = (a_{pq})$. Since $\mathbf{A}^\top = \mathbf{A}$, we have $a_{pq} = a_{qp}$. Write the quadratic form as a double sum:

$$f(\mathbf{w}) = \mathbf{w}^\top \mathbf{A} \mathbf{w} = \sum_{p=1}^{N} \sum_{q=1}^{N} w_p \, a_{pq} \, w_q.$$

By the product rule, each term $w_p a_{pq} w_q$ contributes $a_{pq} w_q$ when $p = i$ and $w_p a_{pq}$ when $q = i$ (and $2a_{ii} w_i$ when both hold, which is the sum of the two). Therefore

$$\begin{aligned} \frac{\partial f}{\partial w_i} &= \sum_{q=1}^{N} a_{iq} w_q + \sum_{p=1}^{N} w_p a_{pi} \\ &= \sum_{q=1}^{N} a_{iq} w_q + \sum_{p=1}^{N} a_{ip} w_p && \text{since } a_{pi} = a_{ip} \\ &= 2 \sum_{q=1}^{N} a_{iq} w_q. \end{aligned}$$

The last sum is the i-th entry of $\mathbf{Aw}$, so $\dfrac{\partial f}{\partial w_i} = 2(\mathbf{Aw})_i$.

Problem 4.2. Stack the partial derivatives:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2\begin{bmatrix}\sum_q a_{1q}w_q & \cdots & \sum_q a_{Nq}w_q\end{bmatrix}^\top = 2\begin{bmatrix}a_{11} & \cdots & a_{1N}\\ \vdots & \ddots & \vdots\\ a_{N1} & \cdots & a_{NN}\end{bmatrix}\begin{bmatrix}w_1\\ \vdots\\ w_N\end{bmatrix} = 2\mathbf{Aw}.$$

Without the symmetry assumption, the first line of the computation of $\partial f/\partial w_i$ in Problem 4.1 gives $(\mathbf{Aw})_i + (\mathbf{A}^\top\mathbf{w})_i$, so $\nabla_{\mathbf{w}} f(\mathbf{w}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{w}$.

Problem 5. Write the regulariser as $\frac{\lambda}{2}\|\mathbf{w}\|_2^2 = \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}$ and expand the squared error. The two cross terms are equal scalars, $\mathbf{t}^\top\mathbf{Xw} = \mathbf{w}^\top\mathbf{X}^\top\mathbf{t}$, so

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{t}^\top\mathbf{t} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{t} + \frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{Xw} + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}.$$

Differentiate term by term, using Problem 3 to add the gradients (and $\nabla_{\mathbf{w}}(c\,g) = c\,\nabla_{\mathbf{w}} g$ for a constant c, which follows in the same way, entry by entry):

- $\frac{1}{2}\mathbf{t}^\top\mathbf{t}$ does not depend on $\mathbf{w}$, so its gradient is $\mathbf{0}$;
- $-\mathbf{w}^\top\mathbf{X}^\top\mathbf{t}$ has gradient $-\mathbf{X}^\top\mathbf{t}$ (Problem 1 with $\mathbf{x} = \mathbf{X}^\top\mathbf{t}$);
- $\frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{Xw}$ has gradient $\mathbf{X}^\top\mathbf{Xw}$ (Problem 4 with $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$, which is symmetric);
- $\frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}$ has gradient $\lambda\mathbf{w}$ (Problem 2).

Hence

$$\boxed{\nabla_{\mathbf{w}} E(\mathbf{w}) = \mathbf{X}^\top\mathbf{Xw} - \mathbf{X}^\top\mathbf{t} + \lambda\mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{w} - \mathbf{X}^\top\mathbf{t}.}$$

Setting the gradient to zero gives the ridge regression solution from the lectures, $\mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{t}$; for $\lambda > 0$ the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible.

3 Probabilistic Inference

Problem 1. *Log-likelihood.* Given q, the flips are independent, each head has probability q and each tail has probability $1-q$. The sequence $\mathcal{D}$ has 6 heads and 4 tails, so

$$p(\mathcal{D}|q) = q^6(1-q)^4, \qquad \log p(\mathcal{D}|q) = 6\log q + 4\log(1-q).$$

(We observe a particular sequence, so there is no binomial coefficient. Including it would add a constant and would not change the maximiser.)

Maximiser. Set the derivative to zero:

$$\frac{d}{dq}\log p(\mathcal{D}|q) = \frac{6}{q} - \frac{4}{1-q} = 0 \implies 6(1-q) = 4q \implies \boxed{\hat{q} = 0.6.}$$

This stationary point is the maximum on $(0,1)$: the second derivative $-6/q^2 - 4/(1-q)^2$ is negative everywhere, and the log-likelihood tends to $-\infty$ as $q \to 0$ or $q \to 1$. In general, with h heads in N flips and $0 < h < N$, the same steps give $\hat{q} = h/N$, the fraction of heads. (If $h = 0$ or $h = N$, the likelihood is maximised at the endpoint $q = 0$ or $q = 1$, which is again h/N.)

Problem 2. Now q can take only three values, so there is no derivative to set to zero. Instead, evaluate the log-likelihood from Problem 1 at each candidate and take the largest:

$$\log p(\mathcal{D}|q) = \begin{cases} 6\log 0.3 + 4\log 0.7 \approx -8.65 & \text{for } q = 0.3, \\ 10\log 0.5 \approx -6.93 & \text{for } q = 0.5, \\ 6\log 0.7 + 4\log 0.3 \approx -6.96 & \text{for } q = 0.7. \end{cases}$$

So the maximum likelihood estimate is $\boxed{\hat{q} = 0.5.}$ The values for 0.5 and 0.7 are close: the likelihood ratio is $p(\mathcal{D}|0.5)/p(\mathcal{D}|0.7) \approx 1.02$.

Problem 3. Drawing one of the five coins uniformly at random gives the prior

$$p(q) = \begin{cases} 1/5 & \text{for } q = 0.3, \\ 1/5 & \text{for } q = 0.5, \\ 3/5 & \text{for } q = 0.7. \end{cases}$$

By Bayes' rule, $p(q|\mathcal{D}) = p(\mathcal{D}|q)\,p(q)/p(\mathcal{D})$, and $p(\mathcal{D})$ does not depend on q. The MAP estimate is therefore

$$\hat{q}_{\text{MAP}} = \operatorname*{argmax}_{q \in \mathcal{H}} \big[\log p(\mathcal{D}|q) + \log p(q)\big].$$

Adding $\log p(q)$ to the values from Problem 2:

$$\log p(\mathcal{D}|q) + \log p(q) \approx \begin{cases} -8.65 - 1.61 = -10.26 & \text{for } q = 0.3, \\ -6.93 - 1.61 = -8.54 & \text{for } q = 0.5, \\ -6.96 - 0.51 = -7.47 & \text{for } q = 0.7. \end{cases}$$

So the MAP estimate is $\boxed{\hat{q} = 0.7.}$ Normalising gives the posterior probabilities $p(q|\mathcal{D}) \approx 0.044,\ 0.244,\ 0.713$ for $q = 0.3,\ 0.5,\ 0.7$ (rounded).

The data and the likelihood are the same as in Problem 2. The answer changes because the prior favours $q = 0.7$ over $q = 0.5$ by a factor of 3, while the data favour $q = 0.5$ over $q = 0.7$ by a factor of only about 1.02.