# AI Camp — Math Competition

Time limit: 180 minutes Total score: 100 points

---

1. From likelihood to the logistic (cross-entropy) loss

   Consider a sample of feature vectors $X = \{x_1, \ldots, x_\ell\}$ and binary labels $Y = \{y_1, \ldots, y_\ell\}$, where $y_i \in \{0, 1\}$. We want to train a linear model with score

   $$z_i = \langle w, x_i \rangle,$$

   by minimizing an empirical risk of the form

   $$Q(w; X, Y) = \sum_{i=1}^{\ell} L(y_i, \langle w, x_i \rangle) \;\to\; \min_w,$$

   where $w$ is the weight vector and $L(y, z)$ is a smooth loss function.

   In logistic regression we model the probability of class $1$ as

   $$\tilde{y}_i := p(y_i = 1 \mid x_i, w).$$

   To measure the quality of such a probabilistic classifier, we use the likelihood $P(Y \mid X, w)$. Assume the pairs $(x_i, y_i)$ are independent across $i$.

   (a) (4 points) If $y_i \in \{0, 1\}$ and the model predicts $\tilde{y}_i \in [0, 1]$, write the probability of observing $y_i$ as a single expression $p(y_i \mid x_i)$.

   You know that for a Bernoulli label:

   - if $y_i = 1$, then $p(y_i \mid x_i) = \tilde{y}_i$,
   - if $y_i = 0$, then $p(y_i \mid x_i) = 1 - \tilde{y}_i$.

   Use the exponent trick:

   $$p(y_i \mid x_i) = \tilde{y}_i^{\square} \, (1 - \tilde{y}_i)^{\square},$$

   where the boxes are expressions involving $y_i$.

   (b) (2 points) Write the joint likelihood $P(Y \mid X, w)$ in terms of the per-example probabilities $p(y_i \mid x_i)$.

   Hint: what is the probability of independent events happening simultaneously?

   (c) (3 points) Take the logarithm of your expression and simplify it. Then write the negative log-likelihood (NLL).

   (d) (3 points) Recall the sigmoid function:

   $$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

   In logistic regression we set

   $$z_i = \langle w, x_i \rangle, \qquad \tilde{y}_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}}.$$

   Rewrite the per-sample negative log-likelihood $\ell_i$ in terms of $y_i$ and $z_i$.

   (e) (3 points) Show that $\ell_i$ can be written in the form

   $$\ell_i = \log(1 + e^{-z_i}) + (\text{something involving } (1 - y_i) z_i).$$

Name: _____

(f) (1 point) What simple transformation maps $y_i \in \{0, 1\}$ to a label $t_i \in \{-1, 1\}$?

(g) (2 points) Find a single expression in $t_i$ and $z_i$ that equals $\log(1+e^{-z_i})$ when $t_i = +1$ and equals $\log(1 + e^{z_i})$ when $t_i = -1$.

(h) (2 points) Write the final simplified expression for $\ell_i$ in terms of $t_i$ and $z_i$.

2. Logistic regression: separability, MAP, and gradients

Now assume labels are given as $t_i \in \{-1, 1\}$. If $p(y = 1 \mid x, w) = \sigma(w^\top x)$, we are interested in the probability the model assigns to the correct label.

(a) (2 points) Express the probability of the correct label as a single sigmoid expression in terms of $t_i$:

$$p(y_i \mid x_i, w) = \sigma(\dots).$$

(b) (1 point) Using the identity above, write the (joint) log-likelihood $\log P(Y \mid X, w)$ and the NLL.

(c) (1 point) Definition: the data are linearly separable if there exists a vector $u$ such that

$$t_i \, u^\top x_i > 0 \quad \text{for all } i.$$

If $t_i u^\top x_i > 0$ for all $i$, what happens to $t_i(\alpha u)^\top x_i$ as $\alpha \to +\infty$? (Does it go to $+\infty$, $-\infty$, or stay bounded?)

(d) (4 points) Assuming the data are linearly separable, find an upper bound for the log-likelihood. Is this bound achieved by any finite $w$? Conclude whether the maximum-likelihood estimate exists.

(e) (1 point) Switch from MLE to MAP (maximum a posteriori). Using Bayes' rule, write $P(w \mid X, Y)$ in terms of $P(Y \mid X, w)$ and a prior $P(w)$ (up to a proportionality constant).

(f) (2 points) Introduce L2 regularization and consider the regularized objective

$$J(w) = \text{NLL}(w) + \frac{\lambda}{2}\|w\|^2, \qquad \lambda > 0.$$

Describe what happens to $J(\alpha w)$ as $\alpha \to +\infty$.

(g) (2 points) Compare $J(w)$ to the negative log-posterior. What prior $P(w)$ corresponds to the penalty $\frac{\lambda}{2}\|w\|^2$?

(h) (2 points) Now return to the common $y_i \in \{0, 1\}$ form with

$$z_i = w^\top x_i, \qquad \tilde{y}_i = \sigma(z_i).$$

The per-sample negative log-likelihood is

$$\ell_i(w) = -\Big(y_i \log \tilde{y}_i + (1 - y_i) \log(1 - \tilde{y}_i)\Big).$$

Compute $\frac{d\ell_i}{dz_i}$ and simplify it to an expression involving $\tilde{y}_i$ and $y_i$.

Hint: you may use

$$\frac{d}{dz} \log \sigma(z) = 1 - \sigma(z), \qquad \frac{d}{dz} \log(1 - \sigma(z)) = -\sigma(z).$$

(i) (1 point) Using $z_i = w^\top x_i$, what is $\frac{\partial z_i}{\partial w}$?

(j) (2 points) Using the chain rule, show that

$$\nabla_w \mathrm{NLL}(w) = \sum_{i=1}^{\ell} \big(\sigma(w^\top x_i) - y_i\big)\, x_i.$$

(k) (1 point) What is $\nabla_w\big(\frac{\lambda}{2}\|w\|^2\big)$?

(l) (1 point) Combine the previous results to show that

$$\nabla_w J(w) = \sum_{i=1}^{\ell} \big(\sigma(w^\top x_i) - y_i\big)\, x_i + \lambda w.$$

3. Softmax regression and non-identifiability

In multiclass classification, logistic regression generalizes as follows: for each class $k \in \{1, \dots, K\}$ we have a weight vector $w_k$. The predicted probability is

$$P(y = k \mid x, W) = \frac{e^{\langle w_k, x\rangle}}{\sum_{j=1}^{K} e^{\langle w_j, x\rangle}}.$$

The (softmax) negative log-likelihood for a dataset of size $N$ is

$$L_{\mathrm{sm}}(W) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}[y_i = k]\, \log P(y_i = k \mid x_i, W),$$

where $\mathbb{I}[y_i = k] = 1$ if $y_i = k$ and $0$ otherwise.

Let $K = 2$. For simplicity, assume the data are linearly inseparable.

(a) (6 points) Let $a$ be any vector of the same dimension as $w_1$ and $w_2$. Define

$$w_1' = w_1 + a, \qquad w_2' = w_2 + a.$$

What happens to

$$P(y = 1 \mid x, W) = \frac{e^{\langle w_1, x\rangle}}{e^{\langle w_1, x\rangle} + e^{\langle w_2, x\rangle}}$$

when you replace $w_1, w_2$ by $w_1', w_2'$?

(b) (4 points) What does this invariance imply about whether an optimal solution $W^* = (w_1^*, w_2^*)$ can be unique?

(c) (4 points) Let $v := w_1 - w_2$. Rewrite $P(y = 1 \mid x, W)$ in terms of $v$ only.

(d) (6 points) Show that for $K = 2$ the softmax model reduces to a sigmoid:

$$P(y = 1 \mid x, W) = \sigma\big((w_1 - w_2)^\top x\big).$$

4. Decision trees: optimal constant prediction in a leaf

When building a decision tree, suppose a leaf contains $N$ objects $x_1, \dots, x_N$ with labels $y_1, \dots, y_N$. The prediction in this leaf is a constant $\tilde{y}$. Find the value of $\tilde{y}$ that minimizes each loss:

(a) (6 points) Mean Squared Error (regression):

$$Q = \frac{1}{N}\sum_{i=1}^{N} (y_i - \tilde{y})^2.$$

(b) (7 points) Mean Absolute Error (regression):

$$Q = \frac{1}{N}\sum_{i=1}^{N} |y_i - \tilde{y}|.$$

(c) (7 points) LogLoss (binary classification), with $\tilde{y} \in [0,1]$ and $y_i \in \{0,1\}$:

$$Q = -\frac{1}{N}\sum_{i=1}^{N} \Big(y_i \log \tilde{y} + (1-y_i)\log(1-\tilde{y})\Big).$$

5. Softmax: gradients and convexity

Consider multiclass logistic regression with $K$ classes and weight vectors $w_1, \dots, w_K \in \mathbb{R}^d$. For an object $x_i \in \mathbb{R}^d$ define the class scores

$$s_{ik} = w_k^\top x_i,$$

and the softmax probabilities

$$p_{ik} = P(y_i = k \mid x_i, W) = \frac{e^{s_{ik}}}{\sum_{j=1}^{K} e^{s_{ij}}}.$$

Let $y_i \in \{1, \dots, K\}$ be the true class and define one-hot targets $y_{ik} = [y_i = k]$.

(a) (3 points) Write the per-sample negative log-likelihood (cross-entropy) $\ell_i(W)$ using the one-hot targets $y_{ik}$ and the probabilities $p_{ik}$.

(b) (4 points) Let $\mathrm{LSE}(s_i) = \log\left(\sum_{j=1}^{K} e^{s_{ij}}\right)$. Show that

$$\frac{\partial}{\partial s_{ik}} \mathrm{LSE}(s_i) = p_{ik}.$$

(c) (4 points) Using the previous result, compute $\frac{\partial \ell_i}{\partial s_{ik}}$ and simplify your answer to an expression involving only $p_{ik}$ and $y_{ik}$.

(d) (3 points) Using $s_{ik} = w_k^\top x_i$, compute the gradient $\nabla_{w_k} \ell_i(W)$.

(e) (2 points) Write $\nabla_{w_k} L(W)$ for the full dataset loss

$$L(W) = \sum_{i=1}^{N} \ell_i(W).$$

(f) (4 points) Let $p_i = (p_{i1}, \dots, p_{iK})^\top$. Show that the Hessian of $\ell_i$ with respect to the score vector $s_i = (s_{i1}, \dots, s_{iK})^\top$ is

$$H_i = \nabla_{s_i}^2 \ell_i = \mathrm{diag}(p_i) - p_i p_i^\top,$$

and argue that $H_i$ is positive semidefinite.

# Answer Sheet

## AI Camp – Math Competition: Official Solutions

1. From likelihood to the logistic (cross-entropy) loss

(a) (4 points) Solution:

$$p(y_i \mid x_i) = \tilde{y}_i^{\,y_i}\big(1-\tilde{y}_i\big)^{\,1-y_i}.$$

So the boxes are $y_i$ and $1-y_i$.

(b) (2 points) Solution: Independence gives

$$P(Y \mid X, w) = \prod_{i=1}^{\ell} p(y_i \mid x_i) = \prod_{i=1}^{\ell} \tilde{y}_i^{\,y_i}\big(1-\tilde{y}_i\big)^{\,1-y_i}.$$

(c) (3 points) Solution: Taking logs,

$$\log P(Y \mid X, w) = \sum_{i=1}^{\ell} \Big( y_i \log \tilde{y}_i + (1-y_i)\log(1-\tilde{y}_i) \Big).$$

Hence the negative log-likelihood (NLL) is

$$\mathrm{NLL}(w) = -\log P(Y \mid X, w) = -\sum_{i=1}^{\ell} \Big( y_i \log \tilde{y}_i + (1-y_i)\log(1-\tilde{y}_i) \Big).$$

(d) (3 points) Solution: With $\tilde{y}_i = \sigma(z_i)$, $z_i = \langle w, x_i \rangle$,

$$\ell_i(w) = -\Big( y_i \log \sigma(z_i) + (1-y_i)\log(1-\sigma(z_i)) \Big).$$

(e) (3 points) Solution: Use

$$\sigma(z) = \frac{1}{1+e^{-z}}, \quad 1-\sigma(z) = \frac{1}{1+e^{z}},$$

so

$$\log \sigma(z) = -\log(1+e^{-z}), \qquad \log(1-\sigma(z)) = -\log(1+e^{z}).$$

Then

$$\begin{aligned} \ell_i &= y_i \log(1+e^{-z_i}) + (1-y_i)\log(1+e^{z_i}) \\ &= y_i \log(1+e^{-z_i}) + (1-y_i)\big(z_i + \log(1+e^{-z_i})\big) \\ &= \log(1+e^{-z_i}) + (1-y_i)z_i. \end{aligned}$$

So the ``something" is $(1-y_i)z_i$.

(f) (1 point) Solution: A standard mapping is

$$t_i = 2y_i - 1,$$

so $y_i = 1 \mapsto t_i = +1$ and $y_i = 0 \mapsto t_i = -1$.

(g) (2 points) Solution:

$$\log(1+e^{-t_i z_i}) = \begin{cases} \log(1+e^{-z_i}), & t_i = +1, \\ \log(1+e^{z_i}), & t_i = -1. \end{cases}$$

(h) (2 points) Solution: The final compact form is

$$\ell_i = \log\big(1 + e^{-t_i z_i}\big) = \log\big(1 + \exp(-t_i\, w^\top x_i)\big).$$

2. Logistic regression: separability, MAP, and gradients

(a) (2 points) Solution: Let $z_i = w^\top x_i$. If $t_i = +1$ (correct class is 1), the probability is $\sigma(z_i)$. If $t_i = -1$ (correct class is 0), the probability is $1-\sigma(z_i) = \sigma(-z_i)$. Thus, in one expression:

$$p(y_i \mid x_i, w) = \sigma(t_i z_i) = \sigma\big(t_i\, w^\top x_i\big).$$

(b) (1 point) Solution: The joint log-likelihood is

$$\log P(Y \mid X, w) = \sum_{i=1}^{\ell} \log \sigma\big(t_i w^\top x_i\big),$$

and the NLL is

$$\mathrm{NLL}(w) = -\sum_{i=1}^{\ell} \log \sigma\big(t_i w^\top x_i\big) = \sum_{i=1}^{\ell} \log\big(1 + e^{-t_i w^\top x_i}\big).$$

(c) (1 point) Solution: Since $t_i u^\top x_i > 0$, we have

$$t_i(\alpha u)^\top x_i = \alpha\, t_i u^\top x_i \xrightarrow[\alpha \to +\infty]{} +\infty.$$

(d) (4 points) Solution: For any real $a$, $\sigma(a) \in (0, 1)$ so $\log \sigma(a) \le 0$. Hence

$$\log P(Y \mid X, w) = \sum_{i=1}^{\ell} \log \sigma(t_i w^\top x_i) \le 0,$$

so an upper bound is $0$.

If the data are separable, pick $u$ with $t_i u^\top x_i > 0$ for all $i$ and let $w = \alpha u$. Then $t_i w^\top x_i \to +\infty$ so $\sigma(t_i w^\top x_i) \to 1$ and $\log \sigma(\cdot) \to 0$. Therefore the log-likelihood approaches $0$ as $\alpha \to +\infty$.

However, $\log \sigma(a) = 0$ would require $\sigma(a) = 1$, which is impossible for finite $a$. So the supremum $0$ is not attained by any finite $w$ and the MLE does not exist (the maximizing sequence has $\|w\| \to \infty$).

(e) (1 point) Solution: Bayes' rule gives

$$P(w \mid X, Y) = \frac{P(Y \mid X, w)\, P(w)}{P(Y \mid X)} \;\propto\; P(Y \mid X, w)\, P(w),$$

where $P(Y \mid X)$ does not depend on $w$.

(f) (2 points) Solution:

$$J(\alpha w) = \mathrm{NLL}(\alpha w) + \frac{\lambda}{2}\|\alpha w\|^2 = \mathrm{NLL}(\alpha w) + \frac{\lambda}{2}\alpha^2 \|w\|^2.$$

As $\alpha \to +\infty$, the quadratic term $\frac{\lambda}{2}\alpha^2\|w\|^2 \to +\infty$, so $J(\alpha w) \to +\infty$ (even if $\mathrm{NLL}(\alpha w)$ decreases).

(g) (2 points) Solution: MAP minimizes

$$-\log P(w \mid X, Y) = -\log P(Y \mid X, w) - \log P(w) + \text{const.}$$

Comparing to $J(w) = \text{NLL}(w) + \frac{\lambda}{2}\|w\|^2$, we need

$$-\log P(w) = \frac{\lambda}{2}\|w\|^2 + \text{const} \quad \Longleftrightarrow \quad P(w) \propto \exp\left(-\frac{\lambda}{2}\|w\|^2\right),$$

i.e. an isotropic Gaussian prior $w \sim \mathcal{N}(0, \lambda^{-1} I)$ (up to normalization).

(h) (2 points) Solution: Let $\tilde{y}_i = \sigma(z_i)$. Using the provided derivatives:

$$\begin{aligned}
\frac{d\ell_i}{dz_i} &= -\left(y_i \frac{d}{dz_i}\log \sigma(z_i) + (1 - y_i)\frac{d}{dz_i}\log(1 - \sigma(z_i))\right) \\
&= -\left(y_i(1 - \sigma(z_i)) + (1 - y_i)(-\sigma(z_i))\right) \\
&= -\left(y_i - \sigma(z_i)\right) = \sigma(z_i) - y_i \\
&= \tilde{y}_i - y_i.
\end{aligned}$$

(i) (1 point) Solution: Since $z_i = w^\top x_i$, we have

$$\frac{\partial z_i}{\partial w} = x_i.$$

(j) (2 points) Solution: By the chain rule,

$$\nabla_w \ell_i(w) = \frac{d\ell_i}{dz_i}\frac{\partial z_i}{\partial w} = (\sigma(z_i) - y_i)x_i.$$

Summing over $i$ gives

$$\nabla_w \text{NLL}(w) = \sum_{i=1}^{\ell} (\sigma(w^\top x_i) - y_i)x_i.$$

(k) (1 point) Solution:

$$\nabla_w\left(\frac{\lambda}{2}\|w\|^2\right) = \nabla_w\left(\frac{\lambda}{2} w^\top w\right) = \lambda w.$$

(l) (1 point) Solution: Combine the two gradients:

$$\nabla_w J(w) = \nabla_w \text{NLL}(w) + \nabla_w\left(\frac{\lambda}{2}\|w\|^2\right) = \sum_{i=1}^{\ell} (\sigma(w^\top x_i) - y_i)x_i + \lambda w.$$

3. Softmax regression and non-identifiability

(a) (6 points) Solution: With $w_1' = w_1 + a$ and $w_2' = w_2 + a$,

$$P'(y = 1 \mid x) = \frac{e^{\langle w_1 + a, x\rangle}}{e^{\langle w_1 + a, x\rangle} + e^{\langle w_2 + a, x\rangle}} = \frac{e^{\langle a, x\rangle} e^{\langle w_1, x\rangle}}{e^{\langle a, x\rangle} e^{\langle w_1, x\rangle} + e^{\langle a, x\rangle} e^{\langle w_2, x\rangle}} = P(y = 1 \mid x).$$

So the probability is invariant to adding the same vector $a$ to both $w_1$ and $w_2$.

(b) (4 points) Solution: If $W^* = (w_1^*, w_2^*)$ is optimal, then for any vector $a$,

$$W'^* = (w_1^* + a,\ w_2^* + a)$$

gives exactly the same predicted probabilities for all $x$, hence the same loss. Therefore the minimizer is not unique (parameters are not identifiable without an extra constraint).

(c) (4 points) Solution: Let $v = w_1 - w_2$. Then

$$P(y = 1 \mid x, W) = \frac{e^{w_1^\top x}}{e^{w_1^\top x} + e^{w_2^\top x}} = \frac{1}{1 + e^{w_2^\top x - w_1^\top x}} = \frac{1}{1 + e^{-v^\top x}}.$$

So it depends only on $v$.

(d) (6 points) Solution: Using the previous identity,

$$P(y = 1 \mid x, W) = \frac{1}{1 + e^{-v^\top x}} = \sigma(v^\top x) = \sigma\big((w_1 - w_2)^\top x\big).$$

4. Decision trees: optimal constant prediction in a leaf

(a) (6 points) Solution:

$$Q(\tilde{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y})^2 \quad \Rightarrow \quad \frac{dQ}{d\tilde{y}} = \frac{2}{N}\sum_{i=1}^{N}(\tilde{y} - y_i).$$

Set $\frac{dQ}{d\tilde{y}} = 0$:

$$\sum_{i=1}^{N}(\tilde{y} - y_i) = 0 \ \Rightarrow\ N\tilde{y} = \sum_{i=1}^{N} y_i \ \Rightarrow\ \tilde{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

So MSE is minimized by the mean.

(b) (7 points) Solution: The objective $\sum_{i=1}^{N} |y_i - \tilde{y}|$ is minimized by any median of $\{y_i\}_{i=1}^{N}$. Equivalently, any $\tilde{y}$ such that at least half the points satisfy $y_i \le \tilde{y}$ and at least half satisfy $y_i \ge \tilde{y}$.

(c) (7 points) Solution: Let $m = \sum_{i=1}^{N} y_i$ be the number of ones, so $\sum_{i=1}^{N}(1 - y_i) = N - m$. Differentiate:

$$Q(\tilde{y}) = -\frac{1}{N}\left(m \log \tilde{y} + (N - m)\log(1 - \tilde{y})\right),$$

$$\frac{dQ}{d\tilde{y}} = -\frac{1}{N}\left(\frac{m}{\tilde{y}} - \frac{N - m}{1 - \tilde{y}}\right).$$

Set to zero:

$$\frac{m}{\tilde{y}} = \frac{N - m}{1 - \tilde{y}} \ \Rightarrow\ m(1 - \tilde{y}) = (N - m)\tilde{y} \ \Rightarrow\ m = N\tilde{y} \ \Rightarrow\ \tilde{y} = \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

So LogLoss is minimized by the empirical fraction of class $1$ in the leaf.

5. Softmax: gradients and convexity

(a) (3 points) Solution: The per-sample cross-entropy is

$$\ell_i(W) = -\sum_{k=1}^{K} y_{ik} \log p_{ik} \quad (= -\log p_{i,y_i}).$$

(b) (4 points) Solution: Let $S_i = \sum_{j=1}^{K} e^{s_{ij}}$. Then

$$\mathsf{LSE}(s_i) = \log S_i \quad \Rightarrow \quad \frac{\partial}{\partial s_{ik}} \mathsf{LSE}(s_i) = \frac{1}{S_i} \frac{\partial S_i}{\partial s_{ik}} = \frac{1}{S_i} e^{s_{ik}} = \frac{e^{s_{ik}}}{\sum_{j=1}^{K} e^{s_{ij}}} = p_{ik}.$$

(c) (4 points) Solution: Using $\log p_{ik} = s_{ik} - \mathsf{LSE}(s_i)$ and $\sum_k y_{ik} = 1$,

$$\ell_i(W) = -\sum_k y_{ik} s_{ik} + \mathsf{LSE}(s_i).$$

Therefore

$$\frac{\partial \ell_i}{\partial s_{ik}} = -y_{ik} + \frac{\partial}{\partial s_{ik}} \mathsf{LSE}(s_i) = -y_{ik} + p_{ik} = p_{ik} - y_{ik}.$$

(d) (3 points) Solution: Since $s_{ik} = w_k^\top x_i$, we have $\frac{\partial s_{ik}}{\partial w_k} = x_i$, hence

$$\nabla_{w_k} \ell_i(W) = \frac{\partial \ell_i}{\partial s_{ik}} \frac{\partial s_{ik}}{\partial w_k} = (p_{ik} - y_{ik}) x_i.$$

(e) (2 points) Solution: Summing over $i$,

$$\nabla_{w_k} L(W) = \sum_{i=1}^{N} \nabla_{w_k} \ell_i(W) = \sum_{i=1}^{N} (p_{ik} - y_{ik}) x_i.$$

(f) (4 points) Solution: From the previous part, $\nabla_{s_i} \ell_i = p_i - y_i$, so

$$H_i = \nabla_{s_i}^2 \ell_i = \nabla_{s_i} p_i.$$

The softmax Jacobian is

$$\frac{\partial p_{ik}}{\partial s_{im}} = p_{ik}(\delta_{km} - p_{im}),$$

so in matrix form

$$H_i = \mathsf{diag}(p_i) - p_i p_i^\top.$$

To show positive semidefinite, take any vector $a \in \mathbb{R}^K$:

$$a^\top H_i a = \sum_k p_{ik} a_k^2 - \left( \sum_k p_{ik} a_k \right)^2 = \mathsf{Var}_{p_i}(a) \geq 0.$$

Hence $H_i \succeq 0$.