Practice problems to prepare for Exam 2

COMS 4771 Fall 2016

Problem 1 (Convexity)**.** Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is a twice-differentiable convex function.

(a) Let $\boldsymbol{w} \in \mathbb{R}^d$ be a fixed vector, and define $g\colon \mathbb{R}^d \to \mathbb{R}$ by $g(\boldsymbol{x}) := f(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$. Write a formula for the Hessian of g at a given point $\boldsymbol{x} \in \mathbb{R}^d$ solely in terms of $\boldsymbol{x}$, $\boldsymbol{w}$, and the second derivative of f at $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$.

(b) Continuing from (a), define $h\colon \mathbb{R}^d \to \mathbb{R}$ by

$$h(\boldsymbol{x}) \;:=\; \begin{cases} g(\boldsymbol{x}) & \text{if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq 0\,, \\ f'(0)\langle \boldsymbol{w}, \boldsymbol{x} \rangle & \text{if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle > 0\,. \end{cases}$$

Here, $f'(0)$ denotes the first derivative of f at zero. Is h a convex function? Answer with either "yes" or "no", and briefly explain your answer.

Problem 2 (Poisson linear regression)**.** Consider the statistical model $\mathcal{P} = \{P_{\boldsymbol{w}} : \boldsymbol{w} \in \mathbb{R}^d\}$ for data in $\mathbb{R}^d \times \mathbb{Z}_+$, where $(\boldsymbol{X}, Y) \sim P_{\boldsymbol{w}}$ means that

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \;\sim\; \mathrm{Poi}(e^{\langle \boldsymbol{w}, \boldsymbol{x} \rangle})\,.$$

Recall that $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ denotes the non-negative integers, and that $T \sim \mathrm{Poi}(\lambda)$ for $\lambda > 0$ means

$$\mathbb{P}(T = t) \;=\; \frac{\lambda^t e^{-\lambda}}{t!} \quad \text{for each } t \in \mathbb{Z}_+\,.$$

(a) Write a simple convex optimization problem (in the standard form from lecture) whose solution (which you may assume is unique) is the maximum likelihood estimators of $\boldsymbol{w}$ given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{Z}_+$.

(b) Derive a gradient descent algorithm for computing the maximum likelihood estimator of $\boldsymbol{w}$ given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{Z}_+$. Give concise and unambiguous pseudocode for your algorithm, and be explicit about how to compute gradients. You may use vector addition, scaling, and inner products as primitive operations (in addition to usual arithmetic operations); and natural logarithm (ln) and exponential (exp) functions as subroutines. Furthermore, assume the initial solution, step sizes, and number of iterations are provided as inputs.

(c) Now consider the statistical model $\mathcal{P}' := \{P_{\boldsymbol{w}} : \boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\|_2 \leq 1\}$, which is the subset of $\mathcal{P}$ in which the parameter vectors have norm at most one. Derive a *projected gradient descent algorithm* for computing the maximum likelihood estimator of $\boldsymbol{w}$ *in this new model* $\mathcal{P}'$ given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{Z}_+$. Give concise and unambiguous pseudocode for your algorithm, and be explicit about how to compute gradients *and how to compute the projection step*. The rest of the instructions are as in part (b).

Problem 3 (Neural networks and AdaBoost)**.**

(a) Let D_t denote the distribution over the training data $\{(x_i, y_i)\}_{i=1}^n$ used in the t-th iteration of AdaBoost, and let f_t be the classifier returned by the weak learning algorithm in the t-th iteration. Suppose that in every iteration $t = 1, 2, \dots$, the *weighted* training error rate of f_t with respect to D_t is 0.6—that is, always *worse* than random guessing. Will AdaBoost be able to reduce the (*unweighted*) training error rate below 0.3?

(b) Suppose AdaBoost is run on a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \{-1, 1\}$ with a weak learning algorithm that outputs linear classifiers. Is the final classifier returned by AdaBoost also a linear classifier in $\mathbb{R}^d$? Answer with either "yes" or "no", and briefly explain your answer.

(c) Suppose you construct a neural network function $f_{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}} \colon \mathbb{R}^d \to \mathbb{R}$ with a single hidden layer of k hidden units, where the link function used is simply the identity function $g(z) = z$. (Here, $\boldsymbol{W}^{(1)} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{W}^{(2)} \in \mathbb{R}^{1 \times k}$.) Let $F \colon \mathbb{R}^d \to \{-1, +1\}$ be the binary classifier given by $F(\boldsymbol{x}) = \operatorname{sign}(f(\boldsymbol{x}))$. Is F a linear classifier in $\mathbb{R}^d$? Answer with either "yes" or "no", and briefly explain your answer.

Problem 4 (Conditional probabilities)**.** Suppose you are faced with a binary classification problem, but you would like to learn a predictor of conditional probabilities rather than a classifier. You have access to an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from the distribution P over $\mathcal{X} \times \{-1, +1\}$ that you care about. You use a learning algorithm that finds an approximate solution $\hat{f}$ to the optimization problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i f(x_i)))] \,.$$

where $\mathcal{F}$ is a particular class of real-valued functions on $\mathcal{X}$. You form the function $\hat{\eta}$ using $\hat{\eta}(x) := 1/(1 + \exp(-\hat{f}(x)))$ for each $x \in \mathcal{X}$. Give at least three reasons why $\hat{\eta}$ may not give the exact conditional probabilities $\mathbb{P}(Y = +1 \mid X = x)$. (Each reason should be valid even if all of the other reasons you give do not hold.)

Problem 5 (Variant of square loss)**.** Let Y be a random variable taking values in $\{-1, +1\}$, with $\mathbb{P}(Y = +1) = \mu$ and $\mathbb{P}(Y = -1) = 1 - \mu$; here, μ is some number in $(0, 1)$. Define the loss function $\ell_{\text{msq}} \colon \mathbb{R} \to \mathbb{R}$ by

$$\ell_{\text{msq}}(z) \;:=\; \big(\max\{1 - z,\, 0\}\big)^2 .$$

What is the minimizer of the function $f \colon \mathbb{R} \to \mathbb{R}$ given by $f(s) := \mathbb{E}[\ell_{\text{msq}}(Ys)]$?

Problem 6 (One-against-all)**.** Suppose using the one-against-all reduction for 3-class classification (with conditional probability estimators), you find conditional probability estimators $f_1', f_2', f_3' \colon \mathcal{X} \to [0, 1]$ such that for each $k \in \{1, 2, 3\}$,

$$|f_k'(x) - \mathbb{P}(Y = k \mid X = x)| \;=\; 0.1 \quad \text{for each } x \in \mathcal{X} \,. \tag{1}$$

Here, (X, Y) is the $\mathcal{X} \times \{1, 2, 3\}$-valued random pair whose distribution is the one you care about. The final classifier $f \colon \mathcal{X} \to \{1, 2, 3\}$ is given by $f(x) := \arg\max_{k \in \{1,2,3\}} f_k'(x))$, with ties broken arbitrarily.

Assume X only takes on two possible values $\mathcal{X} = \{x_1, x_2\}$, with $\mathbb{P}(X = x_1) = \mathbb{P}(X = x_2) = 1/2$, and also that

$$\begin{bmatrix} \mathbb{P}(Y = 1 \mid X = x_1) \\ \mathbb{P}(Y = 2 \mid X = x_1) \\ \mathbb{P}(Y = 3 \mid X = x_1) \end{bmatrix} \;=\; \begin{bmatrix} 0.6 \\ 0.2 \\ 0.2 \end{bmatrix} , \qquad \begin{bmatrix} \mathbb{P}(Y = 1 \mid X = x_2) \\ \mathbb{P}(Y = 2 \mid X = x_2) \\ \mathbb{P}(Y = 3 \mid X = x_2) \end{bmatrix} \;=\; \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix} .$$

(a) What is the error rate of the Bayes classifier?

(b) What is the smallest that the error rate of f can be?

Hint: It depends on f_1', f_2', f_3' subject to the constraint in (1).

(c) What is the largest that the error rate of f can be?

(Same hint as above applies.)

Problem 7 (Generalized least squares linear regression)**.** Consider the statistical model $\mathcal{P} = \{P_{\boldsymbol{w}} : \boldsymbol{w} \in \mathbb{R}^d\}$ for data in $\mathbb{R}^d \times \mathbb{R}$, where $(\boldsymbol{X}, Y) \sim P_{\boldsymbol{w}}$ means that

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \;\sim\; \mathrm{N}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, \|\boldsymbol{x}\|_2^2)\,.$$

Give a formula for the maximum likelihood estimator of $\boldsymbol{w}$ given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{R}$ (regarded as an iid sample). You may assume that the design matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ (whose i-th row is $\boldsymbol{x}_i^\top$, for each $i = 1, 2, \ldots, n$) has rank d, and that none of the $\boldsymbol{x}_i$ is the zero vector.

Hint: you may use $\mathrm{diag}(c_1, c_2, \ldots, c_n)$ to denote the $n \times n$ diagonal matrix whose (i,i)-th entry is c_i, for each $i = 1, 2, \ldots, n$.

Problem 8 (Ridge regression)**.** Recall that if $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then the ridge regression objective function $\boldsymbol{w} \mapsto \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2$ can be written as

$$\boldsymbol{w} \;\mapsto\; \|\boldsymbol{A}(\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathrm{ols}})\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 + (\text{stuff not depending on } \boldsymbol{w})\,.$$

Write a formula for this "stuff not depending on $\boldsymbol{w}$" just in terms of $\boldsymbol{A}$ and $\boldsymbol{b}$.

Problem 9 (Eigendecomposition)**.** Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times d}$ be matrices where, for each $i = 1, 2, \ldots, n$, $\boldsymbol{a}_i^\top$ is the i-th row of $\boldsymbol{A}$, and $\boldsymbol{b}_i^\top$ is the i-th row of $\boldsymbol{B}$. Recall that the trace of a square matrix $\boldsymbol{M}$, denoted $\mathrm{tr}(\boldsymbol{M})$, is the sum of the diagonal entries of $\boldsymbol{M}$.

(a) Prove that

$$\mathrm{tr}(\boldsymbol{a}_i \boldsymbol{b}_i^\top) \;=\; \langle \boldsymbol{b}_i, \boldsymbol{a}_i \rangle\,.$$

(b) Use the fact that tr is a linear function to prove that $\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A}^\top)$.

(c) Use the fact from part (b) to prove that if $\boldsymbol{X}$ is a mean-zero random vector in $\mathbb{R}^d$, then

$$\lambda_1 + \lambda_2 + \cdots + \lambda_d \;=\; \mathbb{E}\,\|\boldsymbol{X}\|_2^2\,, \tag{2}$$

where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of $\mathrm{cov}(\boldsymbol{X})$.

(d) Continuing from (c), if $\mathbb{E}(\boldsymbol{X}) \neq \boldsymbol{0}$, then $\mathrm{tr}(\mathrm{cov}(\boldsymbol{X})) \neq \mathbb{E}\,\|\boldsymbol{X}\|_2^2$. Write down a modified version of the equation in (2) that is correct even when $\mathbb{E}(\boldsymbol{X}) \neq \boldsymbol{0}$. You should just have to make a very small change.

Problem 10 (Least squares and SVD)**.** Consider the least squares problem for $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{y} \in \mathbb{R}^n$:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\|_2^2\,.$$

Suppose $r := \mathrm{rank}(\boldsymbol{A}) < d$, so the least squares problem does not have a unique solution.

(a) Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ denote the SVD of $\boldsymbol{A}$, so $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ satisfy $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{S} \in \mathbb{R}^{r \times r}$ is diagonal and positive definite. Write the *normal equations* that every solution $\boldsymbol{w}$ to the least squares problem must satisfy, solely in terms of $\boldsymbol{w}$, $\boldsymbol{U}$, $\boldsymbol{S}$, $\boldsymbol{V}$, and $\boldsymbol{y}$.

(b) Continuing from (a), let $\boldsymbol{\Pi}$ denote the orthogonal projection to the range of $\boldsymbol{V}$. Explain, in one or two short sentences, why every solution $\boldsymbol{w}$ to the least squares problem must satisfy

$$\boldsymbol{\Pi}\boldsymbol{w} \;=\; \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\top \boldsymbol{y}\,.$$

(c) Continuing from (a) and (b), the matrix $\boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\top$ is called the *Moore-Penrose pseudoinverse of* $\boldsymbol{A}$, which we denote by $\boldsymbol{A}^\dagger \in \mathbb{R}^{d\times n}$. Explain, in one or two short sentences, why $\boldsymbol{A}^\dagger\boldsymbol{y}$ is the solution to the least squares problem smallest Euclidean norm.

Problem 11 (Matrix completion)**.** Consider the optimization problem

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} \quad & \sum_{(i,j)\in\Omega} \big(a_{i,j} - X_{i,j}\big)^2 \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{X}) \;\leq\; k \end{aligned}$$

where $\Omega \subseteq \{1,2,\ldots,m\}\times\{1,2,\ldots,n\}$ and $a_{i,j}\in\mathbb{R}$ for each $(i,j)\in\Omega$.

(a) Explain why the set $\{\boldsymbol{X}\in\mathbb{R}^{m\times n} : \operatorname{rank}(\boldsymbol{X})\leq k\}$ is generally not convex.

(b) Specify values of k such that the above optimization problem can be efficiently solved for any Ω and $\{a_{i,j} : (i,j)\in\Omega\}$.

(c) Specify a particular non-empty set Ω such that the above optimization problem can be efficiently solved for every k and $\{a_{i,j} : (i,j)\in\Omega\}$.

Problem 12 (Coordindate descent)**.** The idea of alternating optimization can be generalized to a scheme called *coordinate descent*. In this problem, you will derive a coordinate descent algorithm for linear regression. Let $\boldsymbol{A} = [\boldsymbol{a}_1|\boldsymbol{a}_2|\cdots|\boldsymbol{a}_d]\in\mathbb{R}^{n\times d}$ and $\boldsymbol{b}\in\mathbb{R}^n$. The objective function $f\colon\mathbb{R}^d\to\mathbb{R}$ is the least squares criterion

$$f(\boldsymbol{w}) \;:=\; \|\boldsymbol{A}\boldsymbol{w}-\boldsymbol{b}\|_2^2\,.$$

Suppose you have a current solution $\hat{\boldsymbol{w}} = (\hat{w}_1,\hat{w}_2,\ldots,\hat{w}_d)\in\mathbb{R}^d$. Fix a coordinate $i\in\{1,2,\ldots,d\}$, and consider objective function $g_i\colon\mathbb{R}\to\mathbb{R}$

$$g_i(w_i) \;:=\; f(\hat{w}_1,\ldots,\hat{w}_{i-1},w_i,\hat{w}_{i+1},\ldots,\hat{w}_d)\,.$$

Give a formula for the minimizer of g_i in terms of $\hat{\boldsymbol{w}}$, $\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_d$, and $\boldsymbol{b}$. You may assume that $\boldsymbol{a}_i\neq\boldsymbol{0}$ for each $i=1,2,\ldots,d$. What is the time complexity of computing the minimizer?

Postscript: The (randomized) coordinate descent algorithm is the following.

1: Initialize $\hat{\boldsymbol{w}}\in\mathbb{R}^d$ somehow.
2: **loop**
3: Pick $i\in\{1,2,\ldots,d\}$ uniformly at random.
4: Replace $\hat{w}_i$ with the minimizer of g_i as defined above.
5: **end loop**

Do you see how some of the computation required for a single iteration can be amortized?

Problem 13 (k-means objective)**.** The ESL text describes the k-means objective in terms of the partitioning of $\{\boldsymbol{x}_i\}_{i=1}^n$ into k clusters $C_1,C_2,\ldots,C_k$ (where $C_i\cap C_j=\emptyset$ and $\bigcup_{i=1}^k C_i = \{\boldsymbol{x}_i\}_{i=1}^n$):

$$\frac{1}{2}\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}\sum_{\boldsymbol{x}'\in C_i}\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2\,. \tag{3}$$

Assume each C_i is non-empty, let

$$\boldsymbol{c}_i \;:=\; \frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}\,,$$

and assume each $\boldsymbol{x} \in C_i$ is closer to $\boldsymbol{c}_i$ than any other $\boldsymbol{c}_j$, $j \neq i$. Is the expression in (3) the same as the k-means objective from lecture:

$$\sum_{i=1}^{n} \min_{j \in \{1,2,\ldots,k\}} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2^2 \, ?$$

Answer with either "yes" or "no", and briefly explain your answer.

Problem 14 (Mixture of two logistic regressions)**.** Consider the following latent variable model for labeled data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \{0,1\}$ (regarded as an iid sample).

- **Observed**: random labeled example $(\boldsymbol{X}, Y)$ in $\mathbb{R}^d \times \{0,1\}$.
- **Hidden**: hidden bit Z in $\{0,1\}$.
- **Model**: $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$. Here, $(\boldsymbol{X}, Y, Z) \sim P_{\boldsymbol{\theta}}$ means

$$\begin{aligned} Z \mid \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Bern}\big(\mathrm{logistic}(\langle \boldsymbol{\alpha}, \boldsymbol{x} \rangle)\big)\,, \\ Y \mid \boldsymbol{X} = \boldsymbol{x},\, Z = z &\sim \mathrm{Bern}\big(\mathrm{logistic}(\langle (1-z)\boldsymbol{\beta}_0 + z\boldsymbol{\beta}_1, \boldsymbol{x} \rangle)\big)\,. \end{aligned}$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$ be some initial parameters for this model. Regard $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n$ as an iid sample from $P_{\hat{\boldsymbol{\theta}}}$.

(a) Write an explicit formula for

$$P_{\hat{\boldsymbol{\theta}}}(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i,\, Y_i = y_i)$$

in terms of $\hat{\boldsymbol{\theta}}$ and $(\boldsymbol{x}_i, y_i)$.

(b) Derive a gradient descent (or "gradient ascent") algorithm for maximizing the expected complete log-likelihood function

$$\boldsymbol{\theta} \;\mapsto\; \mathbb{E}_{\hat{\boldsymbol{\theta}}}\big[\mathcal{L}_c(\boldsymbol{\theta}) \mid \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\big]\,.$$

Give concise and unambiguous pseudocode for your algorithm, and be explicit about how to compute gradients. You may use vector addition, scaling, and inner products as primitive operations (in addition to usual arithmetic operations); and natural logarithm (ln) and exponential (exp) functions as subroutines. Furthermore, assume the initial solution, step sizes, and number of iterations are provided as inputs.

Problem 15 (Variant of Mechanical Turk model)**.** Consider the following variant of the Mechanical Turk model for binary response data $\{X_{i,j}\}_{i \in [m], j \in [n]}$.

- **Observed**: predicted $\{0,1\}$-valued labels on m items from n workers $\{x_{i,j}\}_{i \in [m], j \in [n]}$.
- **Hidden**: correct $\{0,1\}$-valued labels $\{z_i\}_{i=1}^m$ for all m items.
- **Model**: $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_m, p_1, p_2, \ldots, p_n, r_1, r_2, \ldots, r_n) \in [0,1]^{m+2n}$.
  - Data for items $\{(Z_i, X_{i,1}, X_{i,2}, \ldots, X_{i,n})\}_{i=1}^m$ are independent.

– Nature determines correct label for item i:

$$P_{\boldsymbol{\theta}}(Z_i = 1) \;=\; 1 - P_{\boldsymbol{\theta}}(Z_i = 0) \;=\; \pi_i \,.$$

– Conditioned on Z_i, predicted labels $\{X_{i,j}\}_{j=1}^{n}$ from workers on item i are independent.

Conditioned on $Z_i = 0$, worker j is correct with probability p_j:

$$P_{\boldsymbol{\theta}}(X_{i,j} = 0 \mid Z_i = 0) \;=\; 1 - P_{\boldsymbol{\theta}}(X_{i,j} = 1 \mid Z_i = 0) \;=\; p_j \,.$$

Conditioned on $Z_i = 1$, worker j is correct with probability r_j:

$$P_{\boldsymbol{\theta}}(X_{i,j} = 1 \mid Z_i = 1) \;=\; 1 - P_{\boldsymbol{\theta}}(X_{i,j} = 0 \mid Z_i = 1) \;=\; r_j \,.$$

Give an E-M algorithm for this model. No need to specify the initial parameters or the number of iterations.