UNIVERSITÄT
BIELEFELD
Faculty of Technology

Introduction to Data Mining
07 - Item response theory

Benjamin Paaßen

WS 2023/2024, Bielefeld University

1. 1-parameter IRT model
2. Likelihood and posterior
3. Optimization procedure
4. 2-parameter IRT model
5. 3-parameter IRT model

Motivation

- Which latent **student abilities** and **item difficulties** would best explain the following data?

	Task 1	Task 2	Task 3	ability	difficulty		
					1	2	3
Student 1	1	1	0	2	1	0	-1
Student 2	1	0	0	1	0	-1	-2
Student 3	1	1	0	2	1	0	-1
Student 4	1	1	1	3	2	1	0
Student 5	1	1	0	2	1	0	-1
Student 6	0	0	0	0	-1	-2	-3
Student 7	1	1	1	3	2	1	0
Student 8	1	0	0	1	0	-1	-2
Student 9	1	1	1	3	2	1	0
Student 10	1	0	0	1	0	-1	-2

Problem: Noise

UNIVERSITÄT BIELEFELD
Faculty of Technology

- But what do we do with the following data?

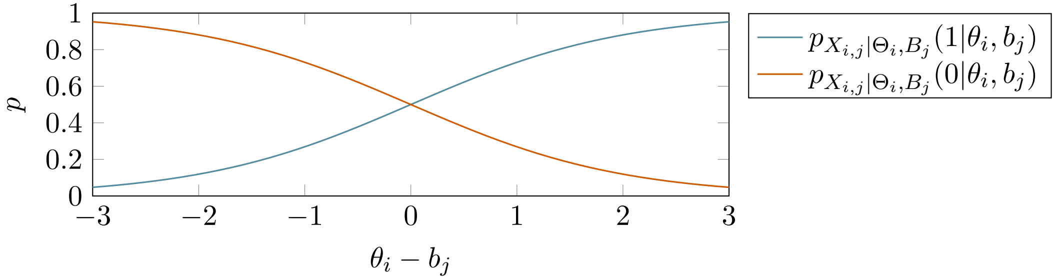
	Task 1	Task 2	Task 3	ability	difficulty 1	difficulty 2	difficulty 3
Student 1	1	1	0	2	1	0	-1
Student 2	1	0	0	1	0	-1	-2
Student 3	0	1	0	1	0	-1	-2
Student 4	1	0	1	2	1	0	-1
Student 5	1	1	0	2	1	0	-1
Student 6	0	0	0	0	-1	-2	-3
Student 7	1	1	1	3	2	1	0
Student 8	1	0	0	1	0	-1	-2
Student 9	0	1	1	2	1	0	-1
Student 10	1	0	0	1	0	-1	-2

- This is where IRT comes in (and to explain survey responses and many other things)

1-parameter IRT model

Probabilistic model

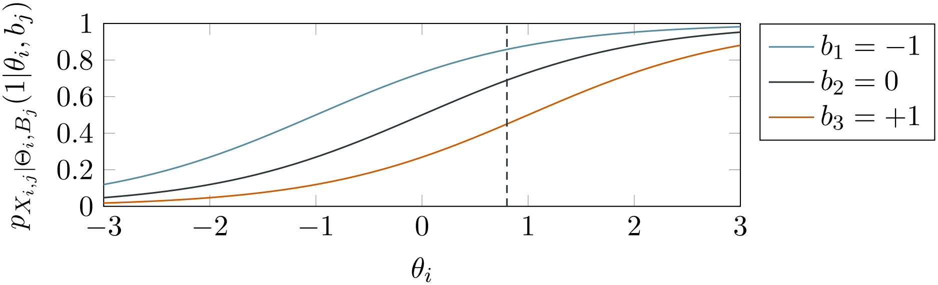
- For each student i, sample ability θ_i from a standard normal Gaussian
- For each item j, sample difficulty b_j from a standard normal Gaussian
- Probability of student i successfully completing item j:
  $p_{X_{i,j}|\Theta_i,B_j}(1|\theta_i, b_j) = \frac{1}{1+\exp[-(\theta_i - b_j)]}$



Likelihood and posterior

Optimization procedure: example

- Assume we would know the difficulties of our three items
- What is the most likely θ_i if $X_{i,1} = 1$, $X_{i,2} = 1$, $X_{i,3} = 0$?



Likelihood

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Assume observed passes/fails $x_{1,1}, \ldots, x_{N,M} \in \{0, 1\}$ are given.

$$p(\boldsymbol{X}|\vec{\theta}, \vec{b}) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(X_{i,j} = x_{i,j}|\vec{\theta}, \vec{b}) \qquad \text{(cond. independence)}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{M} p_{X_{i,j}|\Theta_i, B_j}(x_{i,j}|\theta_i, b_j) \qquad \text{(independence)}$$

$$\Rightarrow -\log\left[p(\boldsymbol{X}|\vec{\theta}, \vec{b})\right] = -\sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j} \cdot \log[p_{X_{i,j}|\Theta_i, B_j}(1|\theta_i, b_j)] + (1 - x_{i,j}) \cdot \log[p_{X_{i,j}|\Theta_i, B_j}(0|\theta_i, b_j)]$$

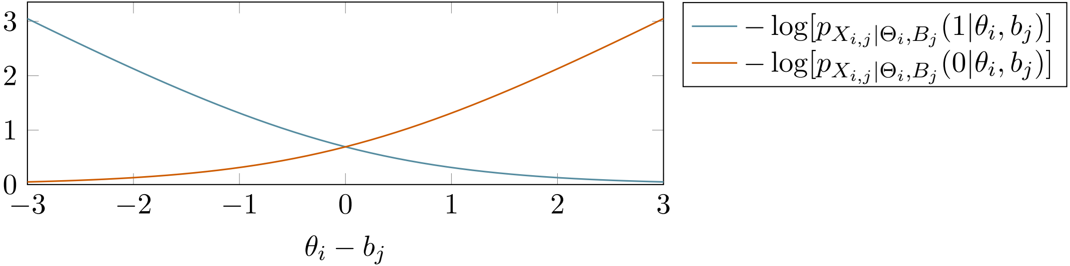
Likelihood (continued)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Note:

$$
\begin{aligned}
-\log[p_{X_{i,j}|\Theta_i,B_j}(1|\theta_i, b_j)] &= -\log\left(\frac{1}{1+\exp[-(\theta_i - b_j)]}\right) \\
&= \log\Big(1+\exp[-(\theta_i - b_j)]\Big) \\
-\log[p_{X_{i,j}|\Theta_i,B_j}(0|\theta_i, b_j)] &= -\log\left(1 - \frac{1}{1+\exp[-(\theta_i - b_j)]}\right) \\
&= -\log\left(\frac{1+\exp[-(\theta_i - b_j)] - 1}{1+\exp[-(\theta_i - b_j)]}\right) \\
&= -\log\left(\frac{1}{1+\exp[\theta_i - b_j]}\right) = \log\Big(1+\exp[\theta_i - b_j]\Big)
\end{aligned}
$$

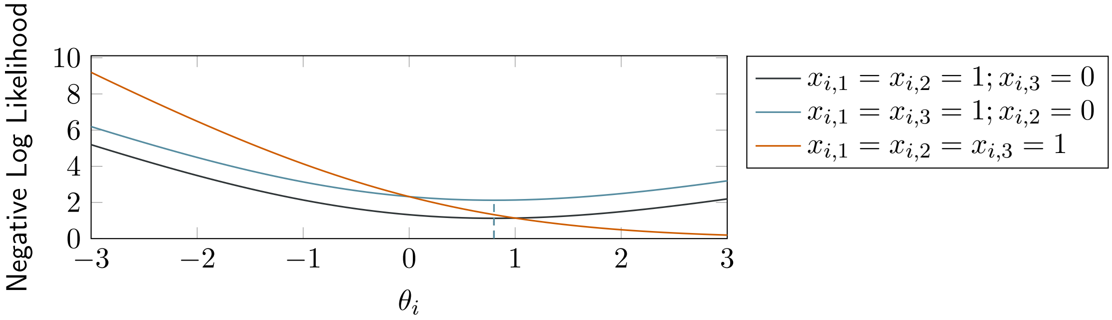
Likelihood (continued)



$\Rightarrow$ Behavior is similar to $\text{ReLU}(-[\theta_i - b_j])$ and $\text{ReLU}(\theta_i - b_j)$

$\Rightarrow$ if $x_{i,j} = 1$, $\theta_i > b_j$ is fine, but we punish if θ_i gets smaller b_j (and vice versa if $x_{i,j} = 0$)

Likelihood: Example



Problems with maximum likelihood

UNIVERSITÄT BIELEFELD
Faculty of Technology

- unbounded: if students get all items wrong, optimum θ_i is at $-\infty$; if students get all items right, optimum is at $+\infty$
- ambigous: adding any constant c to θ_i and b_j yields the same difference $\theta_i - b_j$

$\Rightarrow$ Utilize the priors/marginal densities p_{Θ_i} and p_{B_j} (Bayesian view)

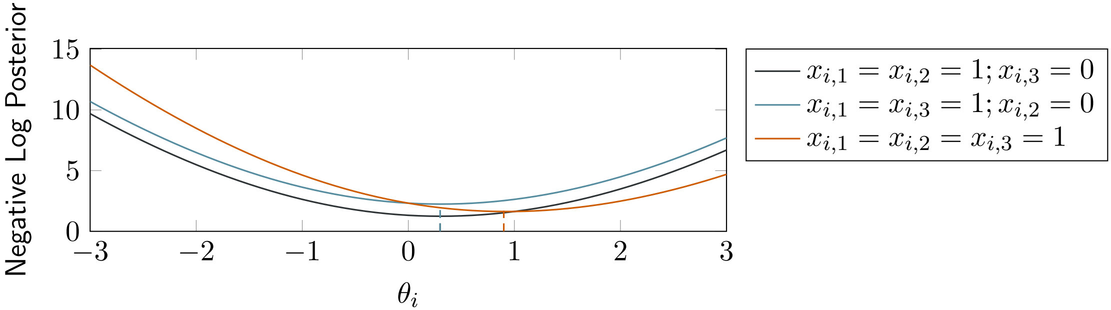
Maximum a-posteriori

- Instead of the likelihood $p(\boldsymbol{X}|\vec{\theta}, \vec{b})$ we wish to maximize the **posterior** $p(\vec{\theta}, \vec{b}|\boldsymbol{X})$

$$
\begin{aligned}
p(\vec{\theta}, \vec{b}|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\vec{\theta}, \vec{b}) \cdot p(\vec{\theta}, \vec{b})}{p(\boldsymbol{X})} \\
\Rightarrow -\log[p(\vec{\theta}, \vec{b}|\boldsymbol{X})] &= -\log[p(\boldsymbol{X}|\vec{\theta}, \vec{b})] - \log[p(\vec{\theta}, \vec{b})] + const. \\
&= -\log[p(\boldsymbol{X}|\vec{\theta}, \vec{b})] + \frac{1}{2}\sum_{i=1}^{N} \theta_i^2 + \frac{1}{2}\sum_{j=1}^{M} b_j^2 + const.
\end{aligned}
$$

Posterior: Example

UNIVERSITÄT BIELEFELD
Faculty of Technology



Optimization procedure

Optimization procedure: Overview

1. Calculate gradient of negative log likelihood and posterior
2. Notice that there is no closed-form solution for gradient $= 0$:(
3. Use logistic regression solvers, instead

Gradient

- Let $z_{i,j} = \theta_i - b_j$, let $p_{i,j} = p_{X_{i,j}|\Theta_i,B_j}(1|\theta_i, b_j) = 1/(1 + \exp(-z_{i,j}))$

$$\frac{\partial}{\partial z_{i,j}} - \log[p_{X_{i,j}|\Theta_i,B_j}(1|\theta_i, b_j)] = \frac{\partial}{\partial z_{i,j}} \log\Big(1 + \exp(-z_{i,j})\Big)$$

$$= \frac{1}{1 + \exp(-z_{i,j})} \cdot \exp(-z_{i,j}) \cdot (-1)$$

$$= -\frac{1}{1 + \exp(z_{i,j})} = p_{i,j} - 1$$

$$\frac{\partial}{\partial z_{i,j}} - \log[p_{X_{i,j}|\Theta_i,B_j}(0|\theta_i, b_j)] = \frac{\partial}{\partial z_{i,j}} \log\Big(1 + \exp(z_{i,j})\Big)$$

$$= \frac{1}{1 + \exp(z_{i,j})} \cdot \exp(z_{i,j})$$

$$= \frac{1}{1 + \exp(-z_{i,j})} = p_{i,j}$$

Gradient (part II)

$$\begin{aligned}
\Rightarrow \frac{\partial}{\partial z_{i,j}} - \log\left[p(\boldsymbol{X}|\vec{\theta}, \vec{b})\right] &= -\sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j} \cdot (p_{i,j} - 1) + (1 - x_{i,j}) \cdot p_{i,j} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j} \cdot (1 - p_{i,j}) + (x_{i,j} - 1) \cdot p_{i,j} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j} - x_{i,j} \cdot p_{i,j} + x_{i,j} \cdot p_{i,j} - p_{i,j} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j} - p_{i,j}
\end{aligned}$$

Gradient (part III)

UNIVERSITÄT BIELEFELD
Faculty of Technology

$$\Rightarrow \frac{\partial}{\partial \theta_i} - \log\left[p(\theta, \vec{b}|\boldsymbol{X})\right] = \frac{\partial}{\partial z_{i,j}} - \log[p(\boldsymbol{X}|\theta, \vec{b})] \cdot \frac{\partial z_{i,j}}{\partial \theta_i} + \frac{\partial}{\partial \theta_i} \frac{1}{2} \sum_{i=1}^{N} \theta_i^2$$

$$= \sum_{j=1}^{M} x_{i,j} - p_{i,j} + \theta_i$$

$$\Rightarrow \frac{\partial}{\partial b_j} - \log\left[p(\theta, \vec{b}|\boldsymbol{X})\right] = \frac{\partial}{\partial z_{i,j}} - \log[p(\boldsymbol{X}|\theta, \vec{b})] \cdot \frac{\partial z_{i,j}}{\partial b_j} + \frac{\partial}{\partial b_j} \frac{1}{2} \sum_{j=1}^{M} b_j$$

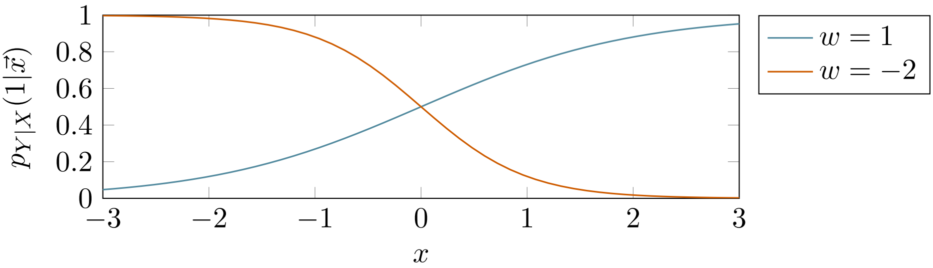
$$= \sum_{i=1}^{N} p_{i,j} - x_{i,j} + b_j$$

$\Rightarrow$ Convex :)

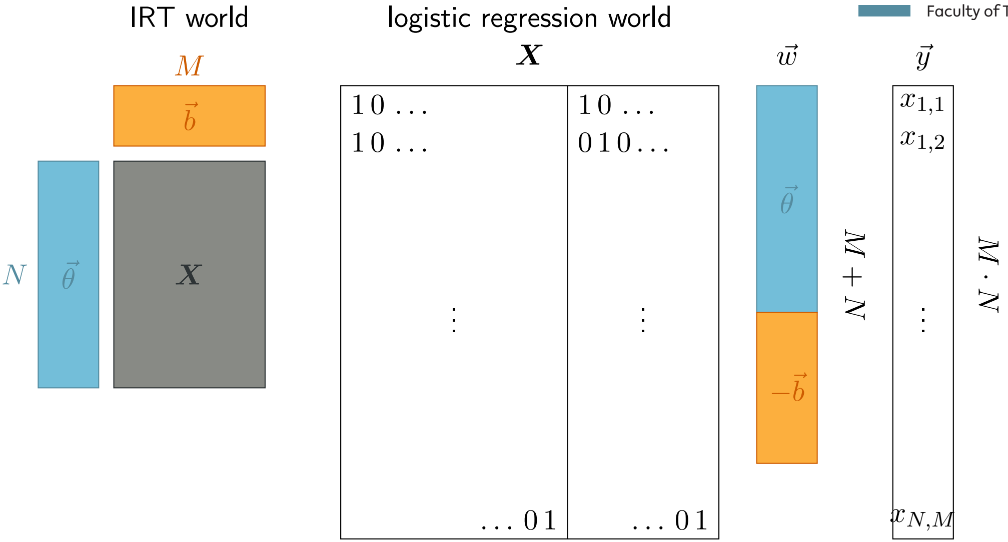
$\Rightarrow$ But no analytical/closed-form solution for gradient $= 0$:(

Logistic Regression

- Key idea: re-phrase item response theory as a logistic regression, then use logistic regression algorithms (e.g. `sklearn.linear_model.LogisticRegression`)
- Logistic regression: binary classifier, i.e. try to predict whether a feature vector $\vec{x}$ should have label 0 or 1
- Model: $p_{Y|X}(1|\vec{x}) = 1/(1 + \exp[-\vec{w}^T \cdot \vec{x}])$ for learned weights $\vec{w}$



Logistic Regression IRT



Logistic Regression IRT (continued)

- Idea: Translate each entry of the IRT data matrix $\boldsymbol{X}$ into one data point with label (!) $x_{i,j}$
- **careful with notation!** In logreg world, $\boldsymbol{X}$ is the feature matrix, $\vec{y}$ is the label vector (filled with entries of the IRT-world version of $\boldsymbol{X}$)
- Weight vector $\vec{w}$: concatenation of $\vec{\theta}$ and $-\vec{b}$
- Feature vector $\vec{x}_{i,j}$: concatentation of ith unit vector and jth unit vector

$\Rightarrow$ $\vec{w}^T \cdot \vec{x}_{i,j} = \theta_i - b_j$

$\Rightarrow$ Logistic regression becomes an IRT model

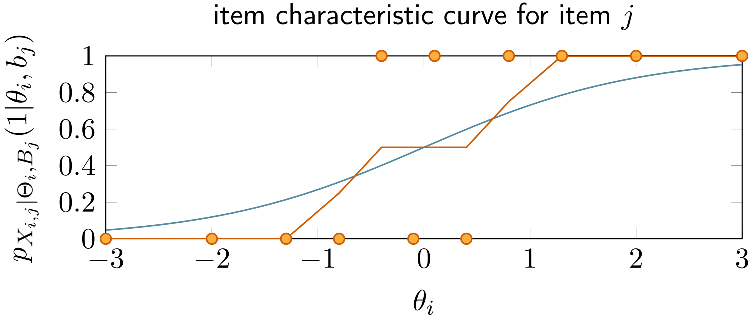
UNIVERSITÄT BIELEFELD

Faculty of Technology

2-parameter IRT model

Motivation: Item characteristic curve

- Is the probability $p_{X_{i,j}|\Theta_i,B_j}(1|\theta_i, b_j)$ accurate? (**calibration**)
- Let's look at one **item characteristic curve**



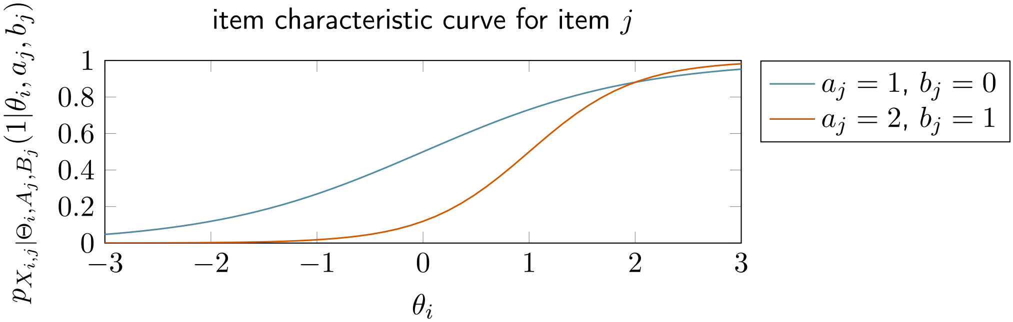
- The empiric probability is **steeper** than the predicted one

2-parameter model

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Idea: add **slope** or **discrimination** parameter a_j for each item

$$p_{X_{i,j}|\Theta_i,A_j,B_j}(1|\theta_i,a_j,b_j) = \frac{1}{1+\exp[-a_j\cdot(\theta_i-b_j)]}$$



Optimiation procedure: overview

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Argument taken from Cai and Thissen (2014)
- Idea: If we would know every student's ability, optimizing item parameters would be easy (just logistic regression for the item-characteristic curve)

$\Rightarrow$ EM algorithm

  - **expectation step:** compute posterior for abilities θ
  - **maximization step:** solve one logistic regression per item j to identify a_j and b_j

Item posterior

UNIVERSITÄT BIELEFELD
Faculty of Technology

$$p(\vec{a}, \vec{b} | \boldsymbol{X}) = \frac{p(\boldsymbol{X} | \vec{a}, \vec{b}) \cdot p(\vec{a}, \vec{b})}{p(\boldsymbol{X})}$$

$$p(\boldsymbol{X} | \vec{a}, \vec{b}) = \prod_{i=1}^{N} p(\vec{x}_i | \vec{a}, \vec{b}) = \prod_{i=1}^{N} \int p(\vec{x}_i, \theta | \vec{a}, \vec{b}) d\theta$$

$$= \prod_{i=1}^{N} \int \prod_{j=1}^{M} p_{X_{i,j} | \Theta_i, A_j, B_j}(x_{i,j} | \theta, a_j, b_j) \cdot p_{\Theta_i}(\theta) d\theta$$

- Nasty, non-convex structure with integral of product :(

$\Rightarrow$ Trick 1: replace integral by sum over sampled values (Bock-Aitkin approach)

$\Rightarrow$ Trick 2: optimize expected neg. log likelihood instead (like in GMMs, EM approach)

Expectation step: ability posterior

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Assume that abilities can only take one of the values $\theta_1, \ldots, \theta_K$ (e.g. -3, -2.9, ..., 2.9, 3)

$$
\begin{aligned}
\gamma_{i,k} &= p(\Theta_i = \theta_k | \vec{a}, \vec{b}, \vec{x}_i) \\
&= \frac{p(\vec{x}_i | \vec{a}, \vec{b}, \Theta_i = \theta_k) \cdot p_{\Theta_i}(\theta_k)}{p(\vec{x}_i)} \\
&= \frac{\prod_{j=1}^M p_{X_{i,j}|\Theta_i, A_j, B_j}(x_{i,j}|\theta_k, a_j, b_j) \cdot p_{\Theta_i}(\theta_k)}{\sum_{l=1}^K \prod_{j=1}^M p_{X_{i,j}|\Theta_i, A_j, B_j}(x_{i,j}|\theta_l, a_j, b_j) \cdot p_{\Theta_i}(\theta_l)}
\end{aligned}
$$

- Can be effectively computed :)

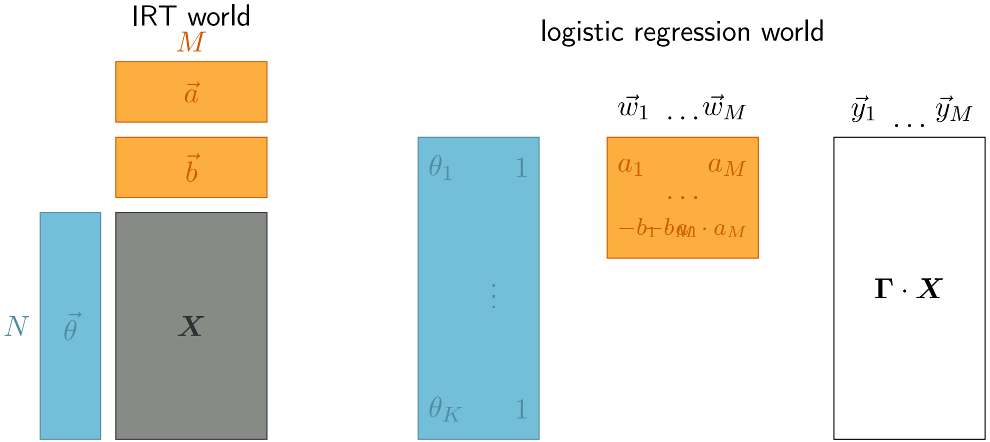
Maximization step: Exp. Neg. Log Likelihood

$$
\begin{aligned}
Q = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{M} \gamma_{i,k} \cdot \Big(&- x_{i,j} \cdot \log[p_{X_{i,j}|\Theta_i,A_j,B_j}(1|\theta_k, a_j, b_j)] \\
&- (1 - x_{i,j}) \cdot \log[p_{X_{i,j}|\Theta_i,A_j,B_j}(0|\theta_k, a_j, b_j)] \Big) \\
= \sum_{j=1}^{M} \sum_{k=1}^{K} &- \log[p_{X_{i,j}|\Theta_i,A_j,B_j}(1|\theta_k, a_j, b_j)] \cdot \Big(\sum_{i=1}^{N} \gamma_{i,k} \cdot x_{i,j} \Big) \\
&- \log[p_{X_{i,j}|\Theta_i,A_j,B_j}(0|\theta_k, a_j, b_j)] \cdot \Big(\sum_{i=1}^{N} \gamma_{i,k} \cdot (1 - x_{i,j}) \Big)
\end{aligned}
$$

$\Rightarrow$ item-wise logistic regression – but with continuous label $\sum_{i=1}^{N} \gamma_{i,k} \cdot x_{i,j}$

M-Step: Logistic Regression

UNIVERSITÄT BIELEFELD
Faculty of Technology



Maximization Step (continued)

- Idea: Each item j has its separate logistic regression with K data points and 2 parameters
- Weight vector $\vec{w}$ for the jth problem: $(a_j, -b_j \cdot a_j)$
- Label $y_k = \sum_{i=1}^{N} \gamma_{i,k} \cdot x_{i,j}$
- Feature vector $\vec{x}_k$: $(\theta_k, 1)^T$

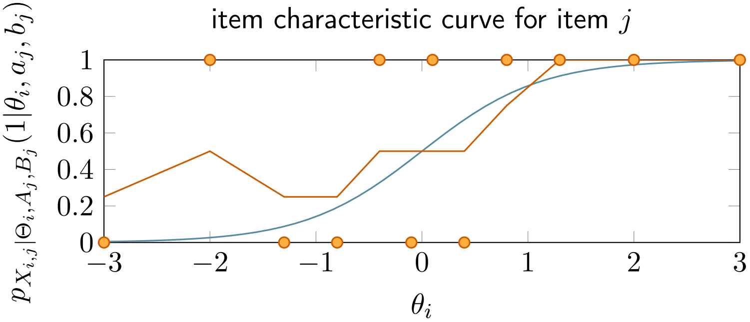
$\Rightarrow$ $\vec{w}^T \cdot \vec{x}_k = a_j \cdot \theta_k - b_j \cdot a_j = a_j \cdot (\theta_k - b_j)$

$\Rightarrow$ Logistic regression becomes 2-parameter IRT model

3-parameter IRT model

Motivation

- What if students can just guess the right answer? (e.g. multiple choice)?

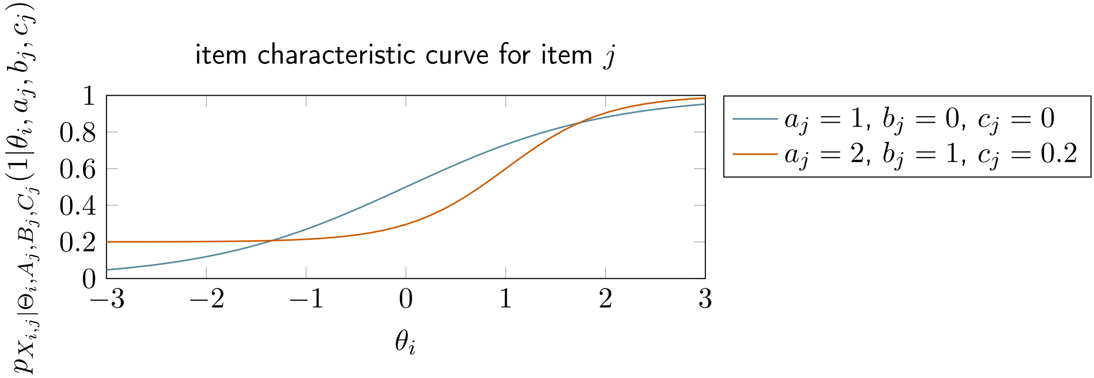


3-parameter model

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Idea: add **guessing** or **base rate** parameter c_j for each item

$$p_{X_{i,j}|\Theta_i,A_j,B_j,C_j}(1|\theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-a_j \cdot (\theta_i - b_j)]}$$



Summary

UNIVERSITÄT BIELEFELD
Faculty of Technology

- IRT tries to model the chance of each student i to pass item j
- 1-parameter model: student ability θ_i and item difficulty b_j; easy to optimize via logistic regression
- 2-parameter model: $+$ item discrimination a_j; tough to optimize, e.g. Box-Aitkin approach or Markov chain monte carlo
- 3-parameter model: $+$ item base rate c_j; even tougher to optimize – not discussed here
- Overall: one of the most interpretable algorithms out there (and very successful for such a simple model)

Challenges and outlook

- How to handle intermediate values between pass/fail?
- How to handle multiple skills? (e.g. VAEs, next session)
- How to generalize to new students? (e.g. VAEs, next session)
- How does ability develop over time? (dynamic models, future sessions)

Literature I

Baker, Frank (2001). **The Basics of Item Response Theory**. 2nd ed. ERIC. URL: https://files.eric.ed.gov/fulltext/ED458219.pdf.

Cai, Li and David Thissen (2014). "Modern approaches to parameter estimation in item response theory". In: **Handbook of item response theory modeling**. Ed. by Steven Reise and Dennis Revicki, pp. 41–59. URL: https://books.google.de/books?hl=en&lr=&id=yDiLBQAAQBAJ&oi=fnd&pg=PA41.

Paaßen, Benjamin, Andreas Bertsch, et al. (2021). "Analyzing Student Success and Mistakes in Virtual Microscope Structure Search Tasks". English. In: **Proceedings of the 15th International Conference on Educational Data Mining (EDM 2021)** (virtual). Ed. by François Bouchet et al. International Educational Datamining Society. URL: https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_67.pdf.

Literature II

UNIVERSITÄT BIELEFELD
Faculty of Technology

Paaßen, Benjamin, Christina Göpfert, and Niels Pinkwart (July 24, 2022). "Faster Confidence Intervals for Item Response Theory via an Approximate Likelihood". In: **Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)** (Durham, UK). Ed. by Alexandra I. Cristea et al., pp. 555–559. DOI: 10.5281/zenodo.6852950.