1-1-2015

Optimistic and Parallel Ising Model Estimation

James Brofos
Dartmouth College

Rui Shu
Dartmouth College

Dartmouth College Computer Science Technical Report TR2015-766

OPTIMISTIC AND PARALLEL ISING MODEL ESTIMATION

By James Brofos and Rui Shu

Dartmouth College

We consider a new method for estimating the structure of Ising graphical models from data. We assume that the data is observed with error, so that it is, in a sense, unreliable. We propose and investigate an "optimistic" estimator; that is, an approach that seeks to correct the log-likelihood objective function when some amount of the data is known to be mismeasured. We derive an interior point algorithm that constructs our estimator efficiently, and demonstrate that it leads naturally to a parallel procedure for recovering the graphical structure of Ising models. We show that the optimistic estimator has performance comparable to, and exceeding, regularized logistic regression in the presence of noise.

1. Introduction. Let $X_1, \ldots, X_k$ be random variables taking values in the dichotomous set $\{-1, +1\}$. We define $G = (V, E)$ to be a simple, undirected graph such that $V = [k] = \{1, \ldots, k\}$. The Ising model associates to each vertex in G a random variable such that,

$$\mathbb{P}\left[X_1 = x_1, \ldots, X_k = x_k\right] = \frac{1}{Z\left(\theta\right)} \exp\left\{ \sum_{(u,v) \in E} \theta_{u,v} x_u x_v \right\}, \tag{1.1}$$

where $\theta_{u,v}$ is regarded as the edge weight in G between vertices u and v. The term $Z\left(\theta\right)$ serves the role of a normalizing constant and is typically called the partition function of the Ising graphical model.

The structure estimation procedure we consider is based on conditional probabilities for each vertex. Defining the neighborhood of a vertex v as

$$\mathrm{ne}\left(v\right) = \left\{u \in V : (u, v) \in E\right\}, \tag{1.2}$$

it can be shown that the conditional probability distributions in the Ising model satisfy the logistic relation

$$\log\left(\frac{\mathbb{P}\left[X_v = x_v | X_u = x_u, u \neq v\right]}{1 - \mathbb{P}\left[X_v = x_v | X_u = x_u, u \neq v\right]}\right) = \sum_{u \in \mathrm{ne}(v)} 2\theta_{u,v} x_u \tag{1.3}$$

$$= \sum_{u \in \mathrm{ne}(v)} \beta_{u,v} x_u. \tag{1.4}$$

Estimation of the underlying graphical model, and, in particular, the edge weight vector θ, can therefore be achieved by estimating the k logistic regression problems assuming the form of (1.3). In the next section we derive an algorithm for an estimator of the edge weight coefficients when X_v is observed in a faulty manner.

2. Main Results. Denote by $\mathbf{X} \in \{-1, +1\}^{n \times k}$ the data drawn in an i.i.d. fashion from an Ising model with individual rows of the matrix being denoted by $\mathbf{X}_i$ for $i \in [n]$. The purpose of this analysis is to construct a logistic regression model that optimistically corrects measurement errors. In the language of robust optimization, we define an uncertainty set $\mathcal{U}$ as

$$\mathcal{U} = \left\{ \Delta \in \{0,1\}^n : \sum_{i=1}^{n} \Delta_i \leq \Gamma \right\}, \tag{2.1}$$

where $\Gamma \in [n]$ is assumed to be a known parameter. The natural interpretation of Γ is that it is an upper bound on the maximum number of flipped bits that may be present in the response variable $\mathbf{y}$. A $\Gamma = 0$ corresponds to noiseless observations of y whereas $\Gamma = n$ corresponds to a potentially very noisy observation where every bit is flipped from its true value!

Given a set of n observations $(\mathbf{y}_i, \mathbf{X}_i)$ for $i \in [n]$ we assume that $\mathbf{y}$ has been corrupted so that the true response variable for the i^{th} observation is $|\mathbf{y}_i - \Delta_i|$. Therefore, Δ may be viewed as a binary vector that encodes a one in the i^{th} position if $\mathbf{y}_i$ is measured with error and zero otherwise. As above, we assume that there are at most Γ corruptions of the response variable.

Proposition 2.1. *We evaluate an optimistic approach for obtaining the coefficients in a logistic regression. In particular, the estimator we propose is defined to be the solution to the optimization problem,*

$$\max_{\beta, \beta_0} \max_{\Delta \in \mathcal{U}} \log \mathcal{L}\left(|\mathbf{y} - \Delta|, \mathbf{X}, \beta, \beta_0\right), \tag{2.2}$$

where $\log \mathcal{L}(\cdot)$ *is the log-likelihood function of the logistic regression model.*

Remark 2.1. The intuition behind the estimator in (2.2) is that we view $\mathbf{y}$ as a worst-case corruption from $|\mathbf{y} - \Delta|$ in terms of the likelihood. Therefore, in the same traditional line as maximum likelihood, we attempt to "correct" the response variable $\mathbf{y}$ so as to maximize the likelihood for a fixed pair (β, β_0). The form of this estimator is somewhat contrary to existing literature on robust optimization where instead $|\mathbf{y} - \Delta|$ is assumed to be the *least* likely configuration of the response variable.

2.1. *Dual Formulation of the Optimistic Estimator.*

THEOREM 2.1. *The inner optimization problem of the optimistic approach in (2.2) has the same objective value as the maximization problem,*

$$\max \quad -\Gamma p - \sum_{i=1}^{n} q_i + \mathbf{y}_i \left(\beta' \mathbf{X}_i + \beta_0\right) - \log\left(1 + e^{\beta' \mathbf{X}_i + \beta_0}\right) \tag{2.3}$$

$$\textit{Such that} \quad p + q_i \geq (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right) \tag{2.4}$$

$$p \geq 0 \quad \textit{and} \quad q_i \geq 0 \quad \forall \quad i \in [n]\,. \tag{2.5}$$

PROOF. Observe that the log-likelihood of the logistic regression model in (2.2) may be expressed,

$$\begin{aligned} &\sum_{i=1}^{n} (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right) \Delta_i + \sum_{i=1}^{n} \mathbf{y}_i \left(\beta' \mathbf{X}_i + \beta_0\right) \\ &\qquad - \sum_{i=1}^{n} \log\left(1 + \exp\left\{\beta' \mathbf{X}_i + \beta_0\right\}\right). \end{aligned} \tag{2.6}$$

Only the terms under the first sum possess a dependency on Δ. Therefore, the optimal solution to the inner maximization problem in (2.2) is identical to that obtained from the problem,

$$\max \quad \sum_{i=1}^{n} (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right) \Delta_i \tag{2.7}$$

$$\text{Such that} \quad \sum_{i=1}^{n} \Delta_i \leq \Gamma \quad \text{and} \quad 0 \leq \Delta_i \leq 1 \quad \forall \quad i \in [n]\,. \tag{2.8}$$

It is easy to see by strong duality that this maximization problem is equivalent to the linear program relaxation,

$$\max \quad -\Gamma p - \sum_{i=1}^{n} q_i \tag{2.9}$$

$$\text{Such that} \quad p + q_i \geq (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right) \tag{2.10}$$

$$p \geq 0, q_i \geq 0 \quad \forall \quad i \in [n]\,. \tag{2.11}$$

The equivalence of objective functions follows quickly from here. □

2.2. *Interior Point Derivation.* For ease of notation we begin straight away by defining a function that expresses the dual objective function of Theorem 2.1. In particular, let $\mathcal{Z}(p, q, \beta, \beta_0)$ be the objective function in (2.3). Apparently the optimistic optimization problem,

$$\begin{aligned} &\max && \mathcal{Z}(p, q, \beta, \beta_0) && (2.12)\\ &\text{Such that} && p + q_i - (-1)^{\mathbf{y}_i}\left(\beta'\mathbf{X}_i + \beta_0\right) \geq 0 && (2.13)\\ & && p \geq 0 \ \text{ and } \ q_i \geq 0 \ \ \forall \ \ i \in [n], && (2.14) \end{aligned}$$

may be solved using an interior point method. We devote the remainder of this section to the proper derivation of that interior point algorithm.

LEMMA 2.1. *The optimal solution to the optimization problem is equivalent to the solution of the unconstrained maximization problem as $\mu \to 0$ iteratively,*

$$\begin{aligned} \max \quad & \mathcal{Z}(p, q, \beta, \beta_0) + \mu \sum_{i=1}^{n} \log\left(p + q_i - (-1)^{\mathbf{y}_i}\left(\beta'\mathbf{X}_i + \beta_0\right)\right) \\ & + \mu \log p + \mu \sum_{i=1}^{n} \log q_i. \end{aligned} \tag{2.15}$$

In order to further simplify the notation, we denote the objective function in (2.15) by $\mathcal{H}(p, q, \beta, \beta_0)$. This $\mathcal{H}(\cdot)$ essentially serves the role of the barrier function and μ is the so-called barrier parameter.

In the style of Newtonian optimization algorithms, we first compute the derivatives of $\mathcal{H}(\cdot)$. This is conceptually simple but notationally quite cumbersome so we report only the ultimate results.

$$\frac{\partial \mathcal{H}}{\partial p} = -\Gamma + \mu \sum_{i=1}^{n} \frac{1}{p + q_i - (-1)^{\mathbf{y}_i}\left(\beta'\mathbf{X}_i + \beta_0\right)} + \frac{\mu}{p}, \tag{2.16}$$

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial \beta_0} = & -\sum_{i=1}^{n} \mathbf{y}_i - \frac{e^{\beta'\mathbf{X}_i + \beta_0}}{1 + e^{\beta'\mathbf{X}_i + \beta_0}} \\ & - \mu \sum_{i=1}^{n} \frac{(-1)^{\mathbf{y}_i}}{p + q_i - (-1)^{\mathbf{y}_i}\left(\beta'\mathbf{X}_i + \beta_0\right)}, \end{aligned} \tag{2.17}$$

$$\frac{\partial \mathcal{H}}{\partial q_i} = -1 + \frac{\mu}{p + q_i - (-1)^{\mathbf{y}_i}\left(\beta' \mathbf{X}_i + \beta_0\right)} + \frac{\mu}{q_i}, \tag{2.18}$$

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial \beta_j} = &-\sum_{i=1}^{n} \mathbf{y}_i \mathbf{X}_{i,j} - \frac{e^{\beta' \mathbf{X}_i + \beta_0}}{1 + e^{\beta' \mathbf{X}_i + \beta_0}} \mathbf{X}_{i,j} \\ &- \mu \sum_{i=1}^{n} \frac{(-1)^{\mathbf{y}_i}}{p + q_i - (-1)^{\mathbf{y}_i}\left(\beta' \mathbf{X}_i + \beta_0\right)} \mathbf{X}_{i,j}.\end{aligned} \tag{2.19}$$

Additionally, we require the matrix of second derivatives. We spare the reader the equations that are used to construct the Hessian and we refer the reader to the appendix for a treatment of these quantities.

2.3. *Explicit Interior Point Algorithm for the Optimistic Estimator.* We present here an explicit algorithm for constructing the coefficients of the optimistic estimator. It can be readily observed (by function composition rules) that $\mathcal{H}(p, q, \beta, \beta_0)$ is a concave function of its parameters. This, in conjunction with the twice differentiable objective function, makes the solution to (2.15) tractable for an interior point method.

The details of the algorithm are outlined as follows.

Data: The design matrix $\mathbf{X}$, the response vector $\mathbf{y}$, and a positive integer Γ.
Result: A vector $\hat{\theta}$ that estimates the edge weights of the Ising graph θ.
Initialize $(p, q, \beta, \beta_0) = (1, \mathbf{1}, \mathbf{0}, 0)$. Set a stopping criterion ϵ. Set the barrier parameter $\mu = 1$.
while *True* **do**
  Calculate the vector update direction $(\Delta p, \Delta q, \Delta \beta, \Delta \beta_0) = -\left(\nabla^2 \mathcal{H}\right)^{-1} \nabla \mathcal{H}$.
  if $||\nabla \mathcal{H}||_\infty < \epsilon$ **then**
    Update the barrier parameter $\mu = \mu \times \omega$ for some quantity $0 < \omega < 1$.
  end
  if $(2n + 1)(\mu) < \epsilon$ **then**
    Break.
  end
  For some $\alpha > 0$ update $(p, q, \beta, \beta_0) = (p, q, \beta, \beta_0) + \alpha(\Delta p, \Delta q, \Delta \beta, \Delta \beta_0)$. This α may be determined by the Armijo rule for instance.
end
Output: The edge weight coefficients $\hat{\theta} = \frac{\beta}{2}$.

Algorithm 1: Interior Point Algorithm for the Optimistic Estimator

REMARK 2.2. It is interesting to note that the algorithm derived here never explicitly requires that $\mathbf{X}$ be a dichotomous matrix. In fact, this same interior point algorithm holds for arbitrary design matrices. However, as the Ising graphical model was our proposed focus area, we devote the remainder

of this work to empirically demonstrating that the optimistic estimator is competitive with, and even exceeds, state-of-the-art methods for estimating the structure of such graphical models.

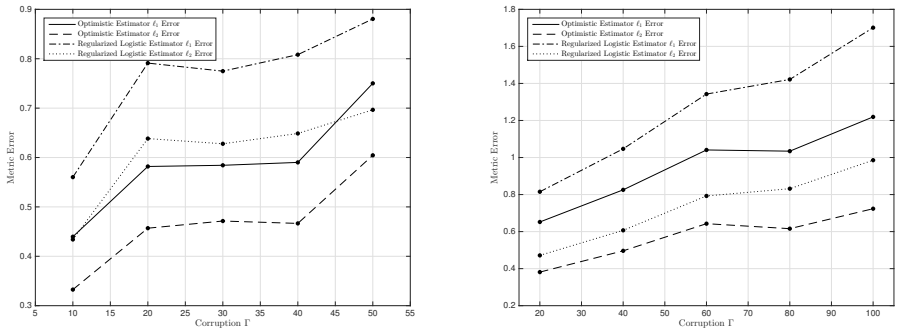
3. Software & Numerical Experiments. In order to evaluate the efficacy of our approach we implement several numerical experiments and illustrate the results in this section. The algorithms discussed here were implemented in the Python 2.7.9 programming language. Details of the implementation are publicly available on the software's homepage: https://github.com/JamesBrofos/Optimistic-Ising-Estimation.

3.1. *Parallel Estimation Paradigm.* As indicated previously, estimation of Ising graphical models via a logistic regression framework is advantageous because structure can be inferred in parallel. In particular, this is achieved by estimating k logistic regression models (one for each of the k vertices appearing in the graph). To exploit this embarrassingly parallel property with distributed computing, we take a master-worker-based approach to the estimation problem, assigning to the master process a queue of optimistic estimators to construct which in turn tasks these to worker processes as they become available.

Remark 3.1. We note that structure estimation with this distributed logistic regression framework does contain a degree of redundancy. It is easily seen that for vertices u and v and $u \in \text{ne}(v)$ in practice the edge weight $\theta_{u,v}$ is estimated twice: once for the regression of u on $\text{ne}(u)$ and again for the regression of v on $\text{ne}(v)$. In an ideal environment, these estimates would be equal, though this is in practice unlikely to be the case. In an effort to correct this likely issue, we simply average the coefficients obtained from either logistic regression model and take this value as our ultimate estimated edge weight between vertices u and v.

We make use of Message Passing Interface (MPI), a common mechanism for programming and coordinating distributed systems. MPI provides a standardized message passing interface that permits multiple processes in a cluster to transmit and receive information.

3.2. *Experimental Design.* We compare the optimistic estimator to ℓ_1-regularized logistic regression, a state-of-the-art method with appealing theoretical properties. To evaluate the performance of these two approaches we generate synthetic data from an Ising graphical model and task the algorithms with recovering the edge weights. In particular, we design two experiments as follows:



(a) Metric Errors for $n = 100$.

(b) Metric Errors for $n = 500$.

Fig 1: Measurements of the ℓ_1 and ℓ_2 errors for logistic regression models under uncertainty in the response variable. As expected, for fixed sample size, increasing the corruption of the dependent variable drives the error metric higher on average. We compare the optimistic estimator with ℓ_1-penalized logistic regression and find that the optimistic estimator outperforms the alternative in the presence of errors in the variables.

1. We drawn 100 samples from an Ising graphical model with 3 vertices. Edge weights are generated uniformly at random from a standard normal distribution. We corrupt the data such that the true likelihood function is minimized for $\Gamma \in \{10, 20, 30, 40\}$.
2. We drawn 500 samples from an Ising graphical model with 5 vertices. Edge weights are again generated uniformly at random from a standard normal distribution. We corrupt the data such that the true likelihood function is minimized for $\Gamma \in \{20, 40, 60, 80, 100\}$.

One vertex is selected as the dependent variable and the remaining vertices are then taken as the explanatory variables. We repeat these experiments fifty times for each value of Γ and report the results. As a performance metric, we provide the average ℓ_1 and ℓ_2 distance of the estimated edge weights from their true values.

3.3. *Results.* The results of these experiments are shown in Figure 1. We observe that the optimistic estimator, in the presence of unreliably reported dependent variables, is consistently superior in terms of both error metrics. We infer that the performance of either estimator decreases at approximately the same rate Γ increases.

4. Estimation of Ising Graphical Models. Having at this point empirically demonstrated a consistent performance increase related to the optimistic estimator we propose, we turn our attention now to evaluating the error of our approach on some common Ising models. We evaluate the performance of the optimistic estimator on the k logistic regression problems on three different graphical model classes. The first is a 15-vertex Ising chain. The second is 4×4 Ising grid. The third is a $3 \times 3 \times 3$ Ising chimera. For each case we select $n \in \{50, 100, 250, 500\}$ and in each case set $\Gamma = \frac{n}{50}$. This selection of Γ is low and was chosen to reflect a high confidence in the accuracy of the approximate distribution generated through Gibbs sampling.

The procedure for evaluating performance is the same for each variety of underlying graphical structure. The edge weight parameter $\theta_{u,v}$ is drawn uniformly at random from the interval $[-1, +1]$ for every $(u, v) \in E$. Samples are then drawn approximately from the Ising model using Gibbs sampling. For each vertex, an optimistic logistic regression model is fit and the ℓ_1 and ℓ_2 distances of the estimate from the truth are recorded.

Remark 4.1. We note that it is often perfectly reasonable to anticipate that data drawn from an Ising graphical model is obtained with some amount of error. This is because most mechanisms for sampling from the underlying distribution are only approximate and rely on Monte Carlo techniques to estimate the behavior of samples.

The results of this analysis are presented in Table 1. As expected the accuracy of the optimistic estimator improves, *ceteris paribus*, as a function of the number of samples. More complex classes of graphs are more difficult to estimate that simpler models, with the chimera achieving the highest error in both the ℓ_1 and ℓ_2 norms. In order to implement a distributed estimation scheme, we use a 2.9GHz MacBook Pro with four independent processes. Three of these processes are devoted to the construction of the optimistic estimators, whereas the remaining process handles the delegation of tasks. We show that adopting the distributed approach to estimation significantly decreases the wallclock time required to construct a single optimistic estimator.

5. Conclusion. We propose an estimator of Ising graphical models (and more generally of logistic relations) that seeks to correct inaccurately measured response variables so that the data makes as much "sense," in terms of the likelihood, as possible. We derive an interior point algorithm that can be used to efficiently construct our estimator. We demonstrate that our approach is competitive with, and exceeds, regularized logistic regression

TABLE 1

We measure the average performance of the optimistic estimator on three classes of graphs and report the edge weight reconstruction accuracy and time-to-completion. Parallel computing was performed on a 2.9GHz MacBook Pro with four processes. The reported time is wallclock time and is shown in minutes.

		Performance Metrics			
		ℓ_1-norm	ℓ_2-norm	Sequential Time	Parallel Time
$n = 50$	Ising Chain	19.0141 (12.8499)	9.0303 (5.8798)	3.029 (0.381)	1.457 (0.721)
	Ising Grid	31.7966 (20.8320)	12.7103 (9.0285)	8.887 (2.139)	3.474 (1.164)
	Ising Chimera	112.9065 (46.0315)	22.3757 (10.1972)	100.357 (10.795)	30.205 (5.367)
$n = 100$	Ising Chain	11.1084 (8.2285)	6.0591 (6.5071)	4.864 (2.011)	1.745 (1.153)
	Ising Grid	17.6933 (9.8632)	7.3620 (6.2498)	10.486 (2.049)	3.231 (1.345)
	Ising Chimera	76.0397 (28.5035)	18.4424 (10.7930)	110.634 (20.462)	38.901 (15.189)
$n = 250$	Ising Chain	7.2526 (3.9337)	2.9544 (1.4452)	6.375 (1.998)	2.012 (0.993)
	Ising Grid	15.8873 (17.1335)	5.7050 (7.9849)	12.553 (2.856)	4.032 (2.134)
	Ising Chimera	40.2116 (13.6180)	8.5715 (4.7329)	130.971 (30.735)	50.420 (22.787)
$n = 500$	Ising Chain	4.4213 (2.021)	1.2967 (1.560)	10.138 (2.078)	4.337 (2.624)
	Ising Grid	11.298 (10.069)	3.217 (5.054)	18.971 (4.638)	7.341 (2.584)
	Ising Chimera	30.698 (8.278)	4.837 (1.021)	160.083 (28.625)	70.012 (25.311)

Note: Standard deviations are shown in parentheses.

in structure estimation tasks in Ising models when variables are measured with error. We also demonstrate that significant speed increases in structure estimation may be achieved by leveraging an embarrassingly parallel computing architecture.

APPENDIX A: SECOND DERIVATIVE MATRIX

We present the second derivatives relevant to the construction of the Hessian matrix for $\mathcal{H}(\cdot)$, holding the barrier parameter constant. By the equality of mixed partials, we calculate only ten derivatives. We begin with the unmixed second derivatives.

$$\frac{\partial^2 \mathcal{H}}{\partial p^2} = -\mu \sum_{i=1}^{n} \frac{1}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2} - \frac{\mu}{p^2}, \tag{A.1}$$

$$\frac{\partial^2 \mathcal{H}}{\partial q_j q_i} = \begin{cases} -\frac{\mu}{\left(p+q_i-(-1)^{\mathbf{y}_i}(\beta' \mathbf{X}_i+\beta_0)\right)^2} - \frac{\mu}{q_i^2} & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \tag{A.2}$$

$$\begin{aligned} \frac{\partial^2 \mathcal{H}}{\partial \beta_0^2} =& \sum_{i=1}^{n} \frac{e^{\beta' \mathbf{X}_i + \beta_0}}{\left(1 + e^{\beta' \mathbf{X}_i + \beta_0}\right)^2} \\ &+ \mu \sum_{i=1}^{n} \frac{1}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}, \end{aligned} \tag{A.3}$$

$$\begin{aligned} \frac{\partial^2 \mathcal{H}}{\partial \beta_k \partial \beta_j} =& \sum_{i=1}^{n} \frac{e^{\beta' \mathbf{X}_i + \beta_0}}{\left(1 + e^{\beta' \mathbf{X}_i + \beta_0}\right)^2} \mathbf{X}_{i,j} \mathbf{X}_{i,k} \\ &+ \mu \sum_{i=1}^{n} \frac{\mathbf{X}_{i,j} \mathbf{X}_{i,k}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}. \end{aligned} \tag{A.4}$$

We continue on with the mixed partials as follows.

$$\frac{\partial^2 \mathcal{H}}{\partial q_i \partial p} = -\frac{\mu}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}, \tag{A.5}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \beta_0 \partial p} = \mu \sum_{i=1}^{n} \frac{(-1)^{\mathbf{y}_i}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}, \tag{A.6}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \beta_j \partial q_i} = \mu \frac{(-1)^{\mathbf{y}_i}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2} \mathbf{X}_{i,j}, \tag{A.7}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \beta_j \partial p} = \mu \sum_{i=1}^{n} \frac{(-1)^{\mathbf{y}_i}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2} \mathbf{X}_{i,j}, \tag{A.8}$$

$$\begin{aligned} \frac{\partial^2 \mathcal{H}}{\partial \beta_0 \partial \beta_j} =& \sum_{i=1}^{n} \frac{e^{\beta' \mathbf{X}_i + \beta_0}}{\left(1 + e^{\beta' \mathbf{X}_i + \beta_0}\right)^2} \mathbf{X}_{i,j} \\ &+ \mu \sum_{i=1}^{n} \frac{\mathbf{X}_{i,j}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}, \end{aligned} \tag{A.9}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \beta_0 \partial q_i} = \mu \frac{(-1)^{\mathbf{y}_i}}{\left(p + q_i - (-1)^{\mathbf{y}_i} \left(\beta' \mathbf{X}_i + \beta_0\right)\right)^2}. \tag{A.10}$$