Stat

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Misclassification Excess Risk Bounds for 1-Bit Matrix Completion

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence: The Tien Mai (the.t.mai@ntnu.no)

Received: 28 November 2023 | **Revised:** 16 July 2024 | **Accepted:** 13 August 2024

Keywords: binary classification | logistic model | matrix completion | misclassification excess risk | optimal rate

ABSTRACT

This study investigates the misclassification excess risk bound in the context of 1-bit matrix completion, a significant problem in machine learning involving the recovery of an unknown matrix from a limited subset of its entries. Matrix completion has garnered considerable attention in the last two decades due to its diverse applications across various fields. Unlike conventional approaches that deal with real-valued samples, 1-bit matrix completion is concerned with binary observations. While prior research has predominantly focused on the estimation error of proposed estimators, our study shifts attention to the prediction error. This paper offers theoretical analysis regarding the prediction errors of two previous works utilizing the logistic regression model: one employing a max-norm constrained minimization and the other employing nuclear-norm penalization. Significantly, our findings demonstrate that the latter achieves the minimax-optimal rate without the need for an additional logarithmic term. These novel results contribute to a deeper understanding of 1-bit matrix completion by shedding light on the predictive performance of specific methodologies.

1 | Introduction

The matrix completion problem is an extensively investigated issue within the realms of machine learning and statistics, garnering significant attention in recent years. This surge in interest is fuelled by contemporary applications such as recommendation systems (Bobadilla et al. 2013; Koren, Bell, and Volinsky 2009) and particularly the famous Netflix challenge (Bennett and Lanning 2007), image processing (Ji et al. 2010; Han et al. 2014), genotype imputation (Chi et al. 2013; Jiang et al. 2016), and quantum statistics (Gross 2011). The task of reconstructing a matrix without any supplementary information is inherently considered unattainable. Nevertheless, the feasibility of this task may be enhanced under specific assumptions regarding the inherent structure of the matrix awaiting recovery. This feasibility is demonstrated by works such as Candès and Tao (2010), Candes and Plan (2010) and Candès and Recht (2009), where the pivotal assumption is that the matrix has a low rank structure. This assumption is particularly natural in various practical scenarios, such as recommendation systems, where it signifies the presence of a small number of latent features elucidating user preferences. Different approaches for matrix completion have been proposed and studied from theoretical and computational points of views; see, for example, Alquier and Ridgway (2020), Chatterjee (2015), Chen et al. (2019), Lim and Teh (2007), Mai and Alquier (2015), Recht and Ré (2013), Salakhutdinov and Mnih (2008) and Tsybakov, Koltchinskii, and Lounici (2011).

The previously mentioned papers primarily concentrated on matrices with real-valued entries. Nevertheless, in numerous practical applications, the observed entries are subjected to significant quantization, frequently restricted to a single bit and belonging to the set $\{-1,1\}$. This quantization scheme is prevalent in various scenarios, such as voting or rating data and survey responses. In these contexts, the typical nature of responses involves binary distinctions, such as 'yes/no', 'like/

dislike', or 'true/false', highlighting the inherent discrete and binary characteristics of the observed data. The problem of recovering a matrix from partial binary (1-bit) observations is usually referred to as 1-bit matrix completion, which was first introduced and studied in Davenport et al. (2014). Since then, various works have been proposed and studied in the context of 1-bit matrix completion; see, for example, Alquier, Cottet, and Lecué (2019), Cai and Zhou (2013), Cottet and Alquier (2018), Hsieh, Natarajan, and Dhillon (2015), Herbster, Pasteris, and Pontil (2016) and Klopp et al. (2015) among others. The predominant emphasis in existing studies lies in addressing estimation errors, demonstrating the consistency of the proposed estimators. However, there is a notable scarcity of attention dedicated to ensuring a guarantee on the probability of encountering prediction errors.

From a machine learning point of view, dealing with binary output is called a classification problem, which is one of the most important setups in statistical learning. This problem has been studied in various contexts, see for examples in Devroye, Györfi, and Lugosi (1996) and Vapnik (2000), while the surveys of the state-of-the-art can be found in Boucheron, Bousquet, and Lugosi (2005), Hastie, Tibshirani, and Friedman (2009), Bühlmann and van de Geer (2011) and Giraud (2022). Let us consider a binary classification scenario involving a high-dimensional feature vector denoted as $x \in \mathbb{R}^d$ and the class label outcome $Y|x$ follows a binomial distribution with parameters $p(x)$, as in Abramovich and Grinshtein (2018). The accuracy of a classifier η is defined by a misclassification risk

$$R(\eta) = \mathbb{P}(Y \neq \eta(x)).$$

It is well-known that the Bayes classifier,

$$\eta^*(x) = \mathbb{1}_{[p(x)\geq 1/2]},$$

minimizes $R(\eta)$, that is, $\eta^* = \text{argmin}R(\eta)$. However, the probability function $p(x)$ is unknown, and the resulting classifier $\widehat{\eta}(x)$ should be designed from the data $\mathcal{S}$: a random sample of n independent observations $(x_1, Y_1), \ldots, (x_n, Y_n)$. The design points x_i may be considered as fixed or random. The corresponding (conditional) misclassification error of $\widehat{\eta}$ is

$$R(\widehat{\eta}) = \mathbb{P}(Y \neq \widehat{\eta}(x) \,|\, \mathcal{S})$$

and the goodness of $\widehat{\eta}$ with respect to η^* is measured by the misclassification excess risk, defined as

$$excess(\widehat{\eta}, \eta^*) = \mathbb{E}R(\widehat{\eta}) - R(\eta^*).$$

A standard approach to obtain $\widehat{\eta}$ is to assume some (parametric or nonparametric) model for $p(x)$. The most commonly used models is logistic regression, where it is assumed that $p(x) = e^{\beta^t x}/(1 + e^{\beta^t x})$ and $\beta \in \mathbb{R}^d$ is a vector of unknown regression coefficients. The corresponding Bayes classifier is a linear classifier $\eta^*(x) = \mathbb{1}_{[p(x)\geq 1/2]} = \mathbb{1}_{[\beta^t x \geq 0]}$. One then estimates β from the data by the maximum likelihood estimator $\widehat{\beta}$, plugs-in $\widehat{\beta}$ and the resulting (linear) classifier is $\widehat{\eta}(x) = \mathbb{1}_{[\widehat{p}(x)\geq 1/2]} = \mathbb{1}_{[\widehat{\beta}^t x \geq 0]}$. This approach has been adapted to the context of 1-bit matrix completion first in the work (Davenport et al. 2014) and then studied in Cai and Zhou (2013), Klopp et al. (2015), Alaya and Klopp (2019) and Hsieh, Natarajan, and Dhillon (2015).

An alternative (nonparametric) technique for deriving a classifier, $\widehat{\eta}$, from available data is empirical risk minimization. In this approach, the objective is to minimize the empirical risk, $\widehat{R}_n(\eta)$, which is a substitute for the true misclassification risk $R(\eta)$.

$$\widehat{R}_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i \neq \eta(x_i)]}$$

over a given class of classifiers, see for examples Boucheron, Bousquet, and Lugosi (2005) and Giraud (2022). However, applying this strategy directly in practical applications poses computational challenges due to its nonconvex and nonsmooth nature. Typically, this challenge is mitigated by adopting a related convex minimization surrogate, such as hinge loss (Bartlett, Jordan, and McAuliffe 2006; Zhang 2004). Notably, this approach has been less explored in the context of 1-bit matrix completion. To the best of our knowledge, the initial efforts to apply this method to 1-bit matrix completion are found in the works of Cottet and Alquier (2018) and Alquier, Cottet, and Lecué (2019).

1.1 | Related Works and Contributions

In the context of 1-bit matrix completion, the observation of n samples Y_{ij}, where $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$, follows a Bernoulli distribution with parameter $f(M_{ij})$. Here, f is a link function mapping from $\mathbb{R}$ to the interval [0,1], and M is a $m_1 \times m_2$ real matrix to be recovered, as detailed in Cai and Zhou (2013), Davenport et al. (2014) and Klopp et al. (2015).

While substantial progress has been made in addressing the estimation error for 1-bit matrix completion, as discussed in Section 2.2, there is a noticeable scarcity of studies focusing on the misclassification excess risk. The pioneering work in addressing misclassification excess risk, to the best of our knowledge, is presented in Cottet and Alquier (2018), which studies a variational Bayesian method with hinge loss. However, the study of misclassification excess risk is confined to a highly restrictive noiseless setting and utilizes PAC-Bayesian techniques. An alternative approach is proposed in Alquier, Cottet, and Lecué (2019), which introduces a nuclear-norm penalization method and achieves a similar result, using hinge loss, to Cottet and Alquier (2018) in a more general case. Specifically, they establish a misclassification excess error of order $r^*(m_1 + m_2)\log(m_1 + m_2)/n$, where r^* represents the rank of M^*, and M^* is the true model parameter.

For logistic models, Alquier, Cottet, and Lecué (2019) obtain a slower rate of $\sqrt{r^*(m_1 + m_2)\log(m_1 + m_2)/n}$, and similarly, Alaya and Klopp (2019) derive a misclassification excess error of order $\sqrt{r^*(m_1 + m_2)/n}$ without an additional logarithmic term.

In this work, we make two significant contributions: firstly, it establishes an upper bound for misclassification excess error for the method proposed by Cai and Zhou (2013), and secondly, it

demonstrates that for logistic models, a fast rate of misclassification excess error of order $r^*(m_1+m_2)/n$ can be achieved without an extra logarithmic term, which is minimax-optimal. These contributions are novel as of our current knowledge.

1.2 | Notation and Organization of the Paper

In this paper, we use $[d]$ to denote the set of integers $\{1, \ldots, d\}$. We use capital letter to denote a matrix (e.g., M) and standard text to denote its entries (e.g., $M_{i,j}$). We let $\|M\|$ denote the operator norm of M, $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ denote the Frobenius norm of M, $\|M\|_*$ denote the nuclear or Schatten-1 norm of M (the sum of the singular values), and $\|M\|_\infty = \max_{i,j}|M_{i,j}|$ denote the entry-wise infinity-norm of M and for $p, q \geq 1$ put $\|M\|_{p,q} = \left(\sum_j \left(\sum_i |M_{ij}|^p\right)^{q/p}\right)^{1/q}$. Finally, for an event $\mathcal{E}$, $\mathbb{1}_{[\mathcal{E}]}$ is the indicator function for that event, that is, $\mathbb{1}_{[\mathcal{E}]}$ is 1 if $\mathcal{E}$ occurs and 0 otherwise. The matrix max-norm of M is defined as $\|M\|_{max} = \min_{M=UV^\top} \|U\|_{2,\infty}\|V\|_{2,\infty}$.

The rest of the paper is organized as follows: In Section 2, we outline the 1-bit matrix completion problem, introduce various methods to address it, and provide a brief discussion on estimation error. Section 3 presents our primary results concerning misclassification excess error across different methods. All technical proofs are given in Section 4. We summarize our findings in the conclusion, covered in Section 5.

2 | Model and Method

2.1 | 1-Bit Matrix Completion

Assume that the observations follow a binomial distribution parametrized by a matrix $X^* \in \mathbb{R}^{m_1 \times m_2}$. Assume in addition that an $i.i.d$ sequence of coefficients (indexes) $(\omega_i)_{i=1}^n \in ([m_1]\times[m_2])^n, n < m_1 m_2$, is revealed uniformly, put $\Omega = (\omega_i)_{i=1}^n$. The observations associated to these coefficients are denoted by $(Y_i)_{i=1}^n \in \{-1,1\}^n$ and distributed as follows:

$$Y_i = \begin{cases} +1 & \text{with probability } f(X^*_{\omega_i}), \\ -1 & \text{with probability } 1-f(X^*_{\omega_i}). \end{cases} \tag{1}$$

where f is the logistic link function, which is common in statistics, with $f(x) = e^x/(1+e^x)$. For ease of notation, we often write X_i, X_i^* instead of $X_{\omega_i}, X^*_{\omega_i}$. Specifically, we assume that each entry of $[m_1]\times[m_2], \omega_i$, is uniformly observed with probability $n/(m_1 m_2)$, independently.

Define

$$d = m_1 + m_2,\ M = \max(m_1, m_2),\ m = \min(m_1, m_2).$$

In order to estimate either X^* or $f(X^*)$, a first strategy was propose in Davenport et al. (2014) that maximizes the log-likelihood function of the variable X given the observations subject to a set of convex constraints. In our case, the log-likelihood function is given by

$$\ell_{\Omega,Y}(X) := \sum_{(i,j)\in\Omega} \left(\mathbb{1}_{[Y_{i,j}=1]}\log(f(X_{i,j})) + \mathbb{1}_{[Y_{i,j}=-1]}\log(1-f(X_{i,j}))\right).$$

Put, for some $r, \gamma > 0$.

$$\begin{aligned} K_*(\gamma, r) &= \left\{X \in \mathbb{R}^{m_1\times m_2} : \|X\|_\infty \leq \gamma, \|X\|_* \leq \gamma\sqrt{r m_1 m_2}\right\}, \\ K_{max}(\gamma, r) &= \left\{X \in \mathbb{R}^{m_1\times m_2} : \|X\|_\infty \leq \gamma, \|X\|_{max} \leq \gamma\sqrt{r}\right\}. \end{aligned}$$

It is noted that for matrices X of rank at most r, we have that $K_{max}(\gamma, r) \subset K_*(\gamma, r)$, see Cai and Zhou (2013).

Davenport et al. (2014) employ the following convex programme:

$$\widehat{X}^{(1)} = \underset{X\in K_*(\gamma,r)}{argmax}\, \ell_{\Omega,Y}(X).$$

Additionally, Cai and Zhou (2013) propose a max-norm constrained method as

$$\widehat{X}^{(2)} = \underset{X\in K_{max}(\gamma,r)}{argmax}\, \ell_{\Omega,Y}(X).$$

Instead of adopting a constrained method, Klopp et al. (2015) suggest a nuclear-norm penalized approach as follows:

$$\widehat{X}^{nu-pen} = \underset{\|X\|_\infty \leq \gamma}{argmin}\, \Phi_Y^\lambda(X), \text{ where } \Phi_Y^\lambda(X) = \frac{1}{n}\ell_{\Omega,Y}(X) + \lambda\|X\|_* \tag{2}$$

with $\lambda > 0$ being a regularization parameter. This estimator undergoes further examination in Alquier, Cottet, and Lecué (2019) and is expanded to a broader context in Alaya and Klopp (2019).

2.2 | On the Estimation Error

Under the assumption that $\text{rank}(X^*) \leq r$ and $\|X\|_\infty \leq \gamma$, Davenport et al. (2014) show that

$$\frac{1}{m_1 m_2}\|\widehat{X}^{(1)} - X^*\|_F^2 \leq c_\gamma\sqrt{\frac{rd}{n}},$$

where c_γ is a constant depending only on γ, see Theorem 1 in Davenport et al. (2014). A similar result using max-norm minimization was also obtained in Cai and Zhou (2013). The paper of Klopp et al. (2015) proves a faster estimation error rate as

$$\frac{1}{m_1 m_2}\|\widehat{X}^{nu-pen} - X^*\|_F^2 \leq c_\gamma\frac{rd\,\log(d)}{n}.$$

A comparable result is established in Alquier, Cottet, and Lecué (2019), specifically in Theorem 4.2. Subsequently, this rate has been recently enhanced to rd/n, without the presence of a logarithmic term, as demonstrated in Alaya and Klopp (2019) (refer to Theorem 7). Moreover, Theorem 3 in Klopp et al. (2015) establishes a lower bound, indicating that the estimation error cannot surpass rd/n. Consequently, the work presented in Alaya and Klopp (2019) attains the precise minimax estimation rate of convergence for 1-bit matrix completion.

3 | Misclassification Excess Risk Bounds

Following the standard approach in classification theory (Vapnik 2000), we now study the misclassification excess risk for 1-bit matrix completion. The corresponding (ideal) Bayes classifier, in the logistic regression, is a linear classifier

$$\eta^*(X^*_\omega) = \begin{cases} +1 & \text{if } f(X^*_\omega) \geq 0.5, \\ -1 & \text{if } f(X^*_\omega) < 0.5, \end{cases} = \text{sign}(X^*_\omega).$$

The accuracy of a classifier η to predict a new entry of the matrix is assessed by a misclassification error

$$R(\eta) = \mathbb{P}\left[Y_1 \neq \eta(X_{\omega_1})\right].$$

One then estimates η from the data $\mathcal{S}$ consisting of a random sample of n independent observations $(Y_1, \omega_1), \ldots, (Y_n, \omega_n)$ by using one of the methods mentioned above to get an estimator of the underlying matrix $\widehat{X}$. The corresponding (conditional) misclassification error of $\widehat{\eta} = \text{sign}(\widehat{X})$ is

$$R(\widehat{\eta}) = \mathbb{P}(Y \neq \widehat{\eta} \,|\, \mathcal{S}),$$

and the goodness of $\widehat{\eta}$ w.r.t. η^* is measured by the misclassification excess risk

$$excess(\widehat{\eta}, \eta^*) = \mathbb{E}\, R(\widehat{\eta}) - \inf_{\eta} R(\eta) = \mathbb{E}\, R(\widehat{\eta}) - R(\eta^*).$$

3.1 | Misclassification Excess Risk Bound

We now present a first result on misclassification excess error for the method proposed in Cai and Zhou (2013). The proof is given in Section 4.

Theorem 1. We have for the estimator $\widehat{X}^{(2)}$ that

$$excess(\widehat{\eta}, \eta^*) \leq \sqrt{C_\gamma \sqrt{\frac{\text{rank}(X^*)d}{n}}},$$

where C_γ is a constant depending only on γ.

Up to our knowledge, the misclassification excess risk bound in Theorem 1 is novel. It brings more information on the behaviour of the methods proposed in Cai and Zhou (2013) and Davenport et al. (2014).

Remark 1. In the logistic regression model, the rate as established in Theorem 1 is comparatively slower than the one derived by Alaya and Klopp (2019, 13) (refer to Corollary 15), which is of the order $\sqrt{\text{rank}(X^*)d/n}$. However, it is crucial to acknowledge that the result in Alaya and Klopp (2019) is obtained under the additional assumption of the so-called 'low-noise' condition. Importantly, in the subsequent section, we demonstrate that the method presented in Cai and Zhou (2013) can achieve the same rate as in Alaya and Klopp (2019) when this type of assumption is satisfied.

3.2 | Improved Bounds With Low-Noise Assumption

The primary challenges for any classifier manifest in the vicinity of the boundary $\{x : p(x) = 1/2\}$, or equivalently, a hyperplane $\beta^t x = 0$ for the logistic regression model, where accurate prediction of the class label becomes particularly challenging. However, in regions where $p(x)$ is sufficiently away from 1/2 (referred to as the margin or low-noise condition), there exists potential for improving the bounds on misclassification excess risk, as established in the preceding section.

In alignment with the approach outlined in Barlett, Jordan, and McAuliffe (2006, 146) (see Lemma 5), we introduce the following low-noise assumption. This kind of assumption was introduced in Mammen and Tsybakov (1999) and Tsybakov (2004).

Assumption 1. Consider the logistic regression model (1) and assume that, for some $c > 0$,

$$\mathbb{P}\left(0 < \left|f(X) - \frac{1}{2}\right| < \frac{1}{2c}\right) = 0.$$

Remark 2. Assumption 1 essentially assumes the existence of the probability is not very close to 1/2. This implies that the decision boundary between classes is well-defined and that instances near the boundary are not highly ambiguous. Under Assumption 1, the rates we achieve in Theorem 2 are faster compared with those in Theorem 1 without Assumption 1. As highlighted in Bartlett, Jordan, and McAuliffe (2006), under Assumption 1, the misclassification excess errors for methods using convex surrogate losses are of the same order as those using the 0-1 loss.

Theorem 2. Under Assumption 1, we have for the estimator $\widehat{X}^{(2)}$ that

$$excess(\widehat{\eta}, \eta^*) \leq C_\gamma \sqrt{\frac{\text{rank}(X^*)d}{n}}$$

where C_γ is a constant depending only on γ.

Remark 3. The upper bound for misclassification excess risk, as presented in Theorem 2, shares similarities with the results reported in Alquier, Cottet, and Lecué (2019) (Theorem 4.2) and Alaya and Klopp (2019) (Corollary 15 and Remark 9). Notably, our technical approach is considerably simpler compared with the methodologies employed in these two references.

Remark 4. A faster rate of $\text{rank}(X^*)M\log(d)/n$ for classification excess error was initially achieved in Cottet and Alquier (2018), to the best of our knowledge. However, it is important to note that this result is obtained under a highly restrictive noiseless setting. Furthermore, the study in Cottet and Alquier (2018) employs a variational Bayesian method using hinge loss and utilizes the PAC-Bayesian bound technique.

A comparable rate, $\text{rank}(X^*)M\log(d)/n$, is also obtained by an alternative estimator based on the hinge loss, rather than the logistic loss, as demonstrated in Theorem 4.4 of Alquier, Cottet, and Lecué (2019, 2319). However, it is noteworthy that the analysis in Alquier, Cottet, and Lecué (2019) for the constraint nuclear-norm estimator with logistic loss yields a slower rate. Specifically, Theorem 4.2 in Alquier, Cottet, and Lecué (2019, 2137–2138) establishes that the rate for the excess is $\sqrt{\text{rank}(X^*)M\log(d)/n}$.

To the best of our knowledge, the misclassification excess risk bound presented in Theorem 2 is also novel. Despite its comparatively slower rate compared with Cottet and Alquier (2018) and Alquier, Cottet, and Lecué (2019), this result contributes valuable insights into the performance characteristics of the methods introduced in Cai and Zhou (2013) and Davenport et al. (2014).

3.3 | Sharp Rate

In the subsequent theorem, we demonstrate that a more precise misclassification excess error, without an additional logarithmic term, can be achieved for the nuclear-norm penalization method within the logistic model. This outcome signifies an enhancement over the findings presented in the reference (Alaya and Klopp 2019).

Theorem 3. Under the assumption 1, for the estimator $\widehat{X}^{nu-pen}$, we have that

$$excess(\widehat{\eta}, \eta^*) \leq C_\gamma \frac{\text{rank}(X^*)d}{n},$$

where C_γ is a constant depending only on γ.

The distinctive technical approach employed to derive the results in the aforementioned theorem lies in leveraging the low-noise assumption. By utilizing the outcomes from the paper (Bartlett, Jordan, and McAuliffe 2006), a faster rate is achieved, deviating from the methodology in Alaya and Klopp (2019) where the results from Zhang (2004) were employed. This divergence in approach enables the attainment of a fast rate.

Remark 5. In Theorem 4.5 of Alquier, Cottet, and Lecué (2019, 2140), a lower bound of the order $\text{rank}(X^*)d/n$ for the misclassification excess risk was established. Consequently, Theorem 3 demonstrates that the estimator $\widehat{X}^{nu-pen}$ attains the optimal-minimax misclassification excess rate without the presence of a logarithmic term. This constitutes a novel contribution that enhances our understanding of 1-bit matrix completion.

4 | Proofs

To begin, we first define some additional notation that we will need for the proofs. For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on a finite set A, let $KL(\mathcal{P}\,\|\,\mathcal{Q})$ denote the Kullback–Leibler (KL) divergence,

$$KL(\mathcal{P}, \mathcal{Q}) = \sum_{x \in A} \mathcal{P}(x) \log\left(\frac{\mathcal{P}(x)}{\mathcal{Q}(x)}\right),$$

where $\mathcal{P}(x)$ denotes the probability of the outcome x under the distribution $\mathcal{P}$. We will abuse this notation slightly by overloading it in two ways. First, for scalar inputs $a_1, a_2 \in [0,1]$, we will set

$$KL(a_1, a_2) = a_1 \log\left(\frac{a_1}{a_2}\right) + (1-a_1)\log\left(\frac{1-a_1}{1-a_2}\right).$$

Second, for two matrices $U, V \in [0,1]^{d_1 \times d_2}$, we define

$$KL(U, V) = \frac{1}{d_1 d_2} \sum_{i,j} KL(U_{i,j}, V_{i,j}).$$

The following lemmas states a general result that relates the misclassification excess risk and the Kullback–Leiber divergence. These remarkable results are established in Zhang (2004) and Bartlett, Jordan, and McAuliffe (2006).

Lemma 1. We have that

$$\mathbb{E}R(\widehat{\eta}) - R(\eta^*) \leq \sqrt{2\mathbb{E}\,KL(f(X^*), f(\widehat{X}))}.$$

Proof. This result is presented in Theorem 2.1 in Zhang (2004), see in particularly Section 3.5 in Zhang (2004). □

Lemma 2. Under the low-noise condition 1, there exists a constant $C > 0$ that

$$\mathbb{E}R(\widehat{\eta}) - R(\eta^*) \leq C\mathbb{E}\,KL(f(X^*), f(\widehat{X})).$$

Proof. The proof can be found in the proof of Theorem 3 in Abramovich and Grinshtein (2018, 3074), which make use of Theorem 3 in Bartlett, Jordan, and McAuliffe (2006). □

The following lemma is an important result from Cai and Zhou (2013). Put

$$\beta_\gamma = \frac{(1+e^\gamma)^2}{e^\gamma},\ U_\gamma = 2\log(e^{\gamma/2} + e^{-\gamma/2}).$$

Lemma 3. Assume that $\|X\|_\infty \leq \gamma$. Then, with probability at least $1-\delta$ for any $\delta > 0$, we have that

$$KL(f(X^*), f(\widehat{X}^{(2)})) \leq \beta_\gamma \sqrt{\frac{rd}{n}} + U_\gamma \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. The proof is an important result in the proof of Theorem 3 in Cai and Zhou (2013, 3641) (details in eq. (35)). □

Lemma 4. Let U be a nonnegative random variable. Assume that

$$\mathbb{P}(U > a_1 + a_2\epsilon) \le e^{-\epsilon^2},$$

then we have that

$$\mathbb{E}(U) \le a_1 + ca_2$$

where $c > 0$ is a numerical constant.

Proof. As U is a nonnegative random variable, we have that

$$\begin{aligned}\mathbb{E}(U) = \int_0^\infty \mathbb{P}(U>x)dx &= \int_0^{a_1} \mathbb{P}(U>x)dx + \int_{a_1}^\infty \mathbb{P}(U>x)dx \\ &\le \int_0^{a_1} 1dx + a_2\int_0^\infty \mathbb{P}(U > a_1 + a_2\epsilon)d\epsilon \\ &\le a_1 + a_2\int_0^\infty e^{-\epsilon^2}d\epsilon = a_1 + a_2\pi/2.\end{aligned}$$

The proof is completed. □ □

Proof of Theorem 1. Applying result from Lemma 3 and Lemma 4 we obtain that

$$\mathbb{E}\,KL(f(X^*), f(\hat{X}^{(2)})) \le \beta_\gamma\sqrt{\frac{rd}{n}} + c\frac{U_\gamma}{\sqrt{n}}.$$

The results in Theorem 1 is followed by using Lemma 1 that

$$\mathbb{E}R(\hat{\eta}) - R(\eta^*) \le \sqrt{2\mathbb{E}\,KL\left(f(X^*), f\left(\hat{X}\right)\right)} \le \sqrt{2\beta_\gamma\sqrt{\frac{rd}{n}} + 2c\frac{U_\gamma}{\sqrt{n}}},$$

where c is a numerical constant. The proof is completed.

Proof of Theorem 2. The proof follows a similar argument as in the proof of Theorem 1. However, under the low-noise assumption, we make use of Lemma 2. We obtain, for some positive constant $C > 0$, that

$$\mathbb{E}R(\hat{\eta}) - R(\eta^*) \le C\mathbb{E}\,KL\left(f(X^*), f\left(\hat{X}\right)\right) \le C\beta_\gamma\sqrt{\frac{rd}{n}} + Cc\frac{U_\gamma}{\sqrt{n}},$$

where c is a numerical constant. Thus, the result of the theorem is obtained.

Proof of Theorem 3. From Corollary 15, Alaya and Klopp (2019, 13), we have for the nuclear-norm penalization estimator in (2) that

$$\mathbb{E}\,KL\left(f(X^*), f\left(\hat{X}\right)\right) \le C_\gamma\frac{\mathrm{rank}(X^*)d}{n}.$$

Applying result from Lemma 2, we obtain the result of the theorem.

5 | Conclusion

In this work, we investigated the problem of matrix completion, which involves the recovery of a matrix from an incomplete set of sampled entries. Specifically, our focus was on binary observations, a well-suited context for example for voting data expressed as 'yes/no' or 'true/false' responses. Our analysis delves into the misclassification excess error associated with certain methods designed for 1-bit matrix completion—a topic that has received limited attention in existing literature.

Our study presents significant original contributions. Firstly, we establish an upper bound for the misclassification excess error concerning the method proposed in Cai and Zhou (2013). Secondly, we demonstrate that a method incorporating nuclear-norm penalization with a logistic regression model can achieve a fast rate of convergence. Furthermore, we establish that this rate is minimax-optimal without additional (multiplicative) logarithmic terms.

It is important to highlight that the current analysis, as well as in Davenport et al. (2014), Cai and Zhou (2013), Klopp et al. (2015), and Alaya and Klopp (2019), assumes the correctness of the model, specifically that the link function f is accurately specified. A potential avenue for future research could involve exploring scenarios where model misspecification occurs and investigating the associated prediction error. While some initial work in this direction has been undertaken without model specification in Cottet and Alquier (2018), to the best of our knowledge, such exploration has not been pursued for nuclear-norm penalization or constrained methods.

Author Contributions

I am the only author of this paper.

Acknowledgements

The author acknowledges the support of the Norwegian Research Council, grant number 309960, through the Centre for Geophysical Forecasting at NTNU. The author expresses gratitude to the associate editor and the anonymous reviewer for their valuable feedback, which significantly enhanced the quality of the manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

Abramovich, F., and V. Grinshtein. (2018). "High-Dimensional Classification by Sparse Logistic Regression." *IEEE Transactions on Information Theory* 65, no. 5: 3068–3079.

Alaya, M. Z., and O. Klopp. (2019). "Collective Matrix Completion." *Journal of Machine Learning Research* 20: 148.

Alquier, P., V. Cottet, and G. Lecué. (2019). "Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures With Lipschitz Loss Functions." *Annals of Statistics* 47, no. 4: 2117–2144.

Alquier, P., and J. Ridgway. (2020). "Concentration of Tempered Posteriors and of Their Variational Approximations." *Annals of Statistics* 48, no. 3: 1475–1497.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe. (2006). "Convexity, Classification, and Risk Bounds." *Journal of the American Statistical Association* 101, no. 473: 138–156.

Bennett, J., and S. Lanning. (2007). "The Netflix Prize." In *Proceedings of KDD Cup and Workshop*, Vol. 2007, 35. New York.

Bobadilla, J., F. Ortega, A. Hernando, and A. Gutiérrez. (2013). "Recommender Systems Survey." *Knowledge-Based Systems* 46: 109–132.

Boucheron, S., O. Bousquet, and G. Lugosi. (2005). "Theory of Classification: A Survey of Some Recent Advances." *ESAIM: Probability and Statistics* 9: 323–375.

Bühlmann, P., and S. van de Geer. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics. Springer, Heidelberg.

Cai, T., and W.-X. Zhou. (2013). "A Max-Norm Constrained Minimization Approach to 1-Bit Matrix Completion." *Journal of Machine Learning Research* 14, no. 1: 3619–3647.

Candes, E. J., and Y. Plan. (2010). "Matrix Completion With Noise." *Proceedings of the IEEE* 98, no. 6: 925–936.

Candès, E. J., and B. Recht. (2009). "Exact Matrix Completion Via Convex Optimization." *Foundations of Computational Mathematics* 9, no. 6: 717–772.

Candès, E. J., and T. Tao. (2010). "The Power of Convex Relaxation: Near-Optimal Matrix Completion." *IEEE Transactions on Information Theory* 56, no. 5: 2053–2080.

Chatterjee, S. (2015). "Matrix Estimation by Universal Singular Value Thresholding." *Annals of Statistics* 43, no. 1: 177–214.

Chen, Y., J. Fan, C. Ma, and Y. Yan. (2019). "Inference and Uncertainty Quantification for Noisy Matrix Completion." *Proceedings of the National Academy of Sciences* 116, no. 46: 22931–22937.

Chi, E. C., H. Zhou, G. K. Chen, D. O. Del Vecchyo, and K. Lange. (2013). "Genotype Imputation Via Matrix Completion." *Genome Research* 23, no. 3: 509–518.

Cottet, V., and P. Alquier. (2018). "1-Bit Matrix Completion: Pac-Bayesian Analysis of a Variational Approximation." *Machine Learning* 107, no. 3: 579–603.

Davenport, M. A., Y. Plan, E. Van Den Berg, and M. Wootters. (2014). "1-Bit Matrix Completion." *Information and Inference: A Journal of the IMA* 3, no. 3: 189–223.

Devroye, L., L. Györfi, and G. Lugosi. (1996). *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics (New York), Vol. 31. New York: Springer-Verlag.

Giraud, C. (2022). *Introduction to High-Dimensional Statistics* (2nd ed.), Monographs on Statistics and Applied Probability, Vol. 168. Boca Raton, FL: CRC Press.

Gross, D. (2011). "Recovering Low-Rank Matrices From Few Coefficients in any Basis." *IEEE Transactions on Information Theory* 57, no. 3: 1548–1566.

Han, X., J. Wu, L. Wang, Y. Chen, L. Senhadji, and H. Shu. (2014). "Linear Total Variation Approximate Regularized Nuclear Norm Optimization for Matrix Completion." *Abstract and Applied Analysis* 2014: 765782.

Hastie, T., R. Tibshirani, and J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer Series in Statistics. New York: Springer.

Herbster, M., S. Pasteris, and M. Pontil. (2016). "Mistake Bounds for Binary Matrix Completion." In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Vol. 29, 1–9. Curran Associates, Inc.

Hsieh, C.-J., N. Natarajan, and I. Dhillon. (2015). "PU Learning for Matrix Completion." In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach, and D. Blei, Proceedings of Machine Learning Research, Vol. 37, 2445–2453. Lille, France: PMLR.

Ji, H., C. Liu, Z. Shen, and Y. Xu. (2010). "Robust Video Denoising Using Low Rank Matrix Completion." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1791–1798.

Jiang, B., S. Ma, J. Causey, et al. (2016). "Sparrec: An Effective Matrix Completion Framework of Missing Data Imputation for GWAS." *Scientific Reports* 6, no. 1: 35534.

Klopp, O., J. Lafond, E. Moulines, and J. Salmon. (2015). "Adaptive Multinomial Matrix Completion." *Electronic Journal of Statistics* 9: 2950–2975.

Koren, Y., R. Bell, and C. Volinsky. (2009). "Matrix Factorization Techniques for Recommender Systems." *Computer* 42, no. 8: 30–37.

Lim, Y. J., and Y. W. Teh. (2007). "Variational Bayesian Approach to Movie Rating Prediction." *Proceedings of KDD Cup and Workshop* 7: 15–21.

Mai, T. T., and P. Alquier. (2015). "A Bayesian Approach for Noisy Matrix Completion: Optimal Rate Under General Sampling Distribution." *Electronic Journal of Statistics* 9, no. 1: 823–841.

Mammen, E., and A. B. Tsybakov. (1999). "Smooth Discrimination Analysis." *Annals of Statistics* 27, no. 6: 1808–1829.

Recht, B., and C. Ré. (2013). "Parallel Stochastic Gradient Algorithms for Large-Scale Matrix Completion." *Mathematical Programming Computation* 5, no. 2: 201–226.

Salakhutdinov, R., and A. Mnih. (2008). "Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo." In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 880–887. New York, NY, USA: Association for Computing Machinery.

Tsybakov, A. B. (2004). "Optimal Aggregation of Classifiers in Statistical Learning." *Annals of Statistics* 32, no. 1: 135–166.

Tsybakov, A. B., V. Koltchinskii, and K. Lounici. (2011). "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion." *Annals of Statistics* 39, no. 5: 2302–2329.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory* (2nd ed.), Statistics for Engineering and Information Science. New York: Springer-Verlag.

Zhang, T. (2004). "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *Annals of Statistics* 32, no. 1: 56–85.