Learning Sparse Potential Games in Polynomial Time and Sample Complexity

Asish Ghoshal and Jean Honorio
Department of Computer Science
Purdue University
West Lafayette, IN - 47906
{aghoshal, jhonorio}@purdue.edu

We consider the problem of learning sparse potential games — a non-parametric class of graphical games where the payoffs are given by potential functions — from observations of strategic interactions. We show that a polynomial time method based on \( l_1, l_2 \)-group regularized logistic regression recovers the \( \epsilon \)-Nash equilibria set of the true game in \( O(m^4d^4 \log(pd)) \) samples, where \( m \) is the maximum number of pure strategies of a player, \( p \) is the number of players and \( d \) is the maximum degree of the game graph. Under slightly more stringent conditions on the payoff functions of the true game, we show that our method recovers the pure-strategy Nash equilibria (PSNE) set of the true game exactly. We also show that \( \Omega (\log m + d \log p) \) samples are necessary for any method to recover the PSNE set of the true game.

1. Introduction and Related Work

Over the past several decades, non-cooperative game theory has emerged as a powerful mathematical framework for reasoning about strategic interactions between self-interested agents. Traditionally, research in game theory has focused on computing various solution concepts like Nash equilibria — which characterizes the stable outcome of the overall behavior of self-interested agents — assuming full knowledge of the payoffs involved. In many real-world settings, however, only the final outcome of the game is observed, in which case inferring the latent payoffs of the players from observations of behavioral data becomes imperative.

In recent times, there has been a surge of interest in learning games from behavioral data c.f. [Irfan and Ortiz, 2014; Honorio and Ortiz, 2015; Ghoshal and Honorio, 2016; Garg and Jaakkola, 2016; Ghoshal and Honorio, 2017]. For instance, Irfan and Ortiz (2014) identified the most influential senators in the U.S congress — a small coalition
of senators whose collective behavior forced every other senator to a unique choice of action — by learning a linear influence game from congressional voting records. Garg and Jaakkola (2016) showed that a tree-structured potential game learned from U.S. Supreme Court data was able to recover the known ideologies of the justices. However, many open problems remain in this area of active research. One such problem is whether there exists efficient (polynomial time) methods for learning potential games from noisy observations of strategic interactions. This is the focus of the current paper.

Various methods have been proposed for learning games from data. Honorio and Ortiz (2015) proposed a maximum-likelihood approach to learn “linear influence games” — a class of parametric graphical games with linear payoffs. However, in addition to being exponential time, the maximum-likelihood approach of Honorio and Ortiz (2015) also assumed a specific observation model for the strategy profiles. Ghoshal and Honorio (2016) proposed a polynomial time algorithm, based on $\ell_1$-regularized logistic regression, for learning linear influence games. They again assumed the specific observation model proposed by Honorio and Ortiz (2015) in which the strategy profiles (or joint actions) were drawn from a mixture of uniform distributions: one over the pure-strategy Nash equilibria (PSNE) set, and the other over the complement of the PSNE set. Ghoshal and Honorio (2017) obtained necessary and sufficient conditions for learning linear influence games under arbitrary observation model. Finally, Garg and Jaakkola (2016) use a discriminative, max-margin based approach, to learn tree-structured potential games. However, their method is exponential time and they show that learning potential games is NP-hard under this max-margin setting, even when the class of graphs is restricted to trees. Furthermore, all the aforementioned works, with the exception of Garg and Jaakkola (2016), consider binary strategies only. In this paper, we propose a polynomial time algorithm for learning potential games (or equivalently poly-matrix games (Janovskaja, 1968)), which are non-parametric graphical games where the pairwise payoffs between players are characterized by matrices (or pairwise potential functions). In this setting, each player has a finite number of pure-strategies.

Our Contributions. We propose an $\ell_{1,2}$ group-regularized logistic regression method to learn potential games, which has been considered by Garg and Jaakkola (2016) and is a generalization of linear influence games considered by Ghoshal and Honorio (2017). We make no assumptions on the latent payoff functions and show that our polynomial time algorithm recovers an $\varepsilon$-Nash equilibrium of the true game, with high probability, if the number of samples is $O(m^4d^4\log(pd))$, where $p$ is the number of players, $d$ is the maximum degree of the game graph and $m$ is the maximum number of pure-strategies of a player. Under slightly more stringent conditions on the payoff functions of the underlying game, we show that our method recovers the Nash equilibria set exactly. We further generalize the observation model from Ghoshal and Honorio (2017) in the sense that we allow strategy profiles in the non-Nash equilibria set to have zero measure. This should be compared with the results of Garg and Jaakkola (2016) who show that learning tree-structured potential games is NP-hard under a max-margin setting. We also obtain necessary conditions on learning potential games and show that $\Omega (\log m + d \log p)$. 

2
samples are required by any method for recovering the PSNE set of a potential game from observations of strategy profiles.

Finally, we conclude this section by referring the reader to the work of Jalali et al. (2011) who analyze $\ell_{1,2}$-regularized logistic regression for learning undirected graphical models. However, our setting differs from that of learning discrete graphical models in many ways. First, unlike discrete graphical models, where the underlying distribution over the variables is described by a potential function that factorizes over the cliques of the graph, we make no assumptions whatsoever on the generative distribution of data. Further, we are interested in recovering the PSNE set of a game, since the graph structure in generally unidentifiable from observational data, whereas Jalali et al. (2011) obtain guarantees on the graph structure of the discrete graphical model. As a result, our theoretical analysis and proofs differ significantly from those of Jalali et al. (2011).

2. Notation and Problem Formulation

In this section, we introduce our notation and formally define the problem of learning potential games from behavioral data. A $p$-player potential game is a graphical game where the set of nodes of the graph denote players and the edges correspond to two-player games. We will denote the graph by $G = ([p], E)$, where $[p] \define \{1, \ldots, p\}$ is the vertex set and $E \subseteq [p] \times [p]$ is set of directed edges. An edge $(i, j) \in E$ denotes the directed edge $i \to j$. Each player $i$ has a set of pure-strategies or actions $A_i$, and the set of pure-strategy profiles or joint actions of all the $p$ players is denoted by $A = \times_{i \in [p]} A_i$. We will denote $A_{-i} \define \times_{j \in \{-i\}} A_j$. With each edge $(i, j) \in E$ is associated a payoff matrix $u^{ij} : A_i \times A_j \to \mathbb{R}$, such that $u^{ij}(x_i, x_j)$ gives the finite payoff of the $i$-th player (with respect to the $j$-th player), when player $i$ plays $x_i \in A_i$ and player $j$ plays $x_j \in A_j$. We assume that $(i, j) \in E$, if and only if $u^{ij}(\cdot, \cdot) \neq 0$. Given a strategy profile $\mathbf{x} \in A$, the total payoff, or simply the payoff, of the $i$-th player is given by the following potential function:

$$u^i(x_i, \mathbf{x}_{-i}; G) = u^{ii}(x_i) + \sum_{j \in N_i} u^{ij}(x_i, x_j), \tag{1}$$

where $N_i(G) \define \{j \in [p]|(i, j) \in E\}$ is the set of neighbors of $i$ in the graph $G$, and $u^{ii} : A_i \to \mathbb{R}$ gives the (finite) individual payoff of $i$ for playing $x_i$. We will denote the number of neighbors of player $i$ by $d_i \define |N_i(G)|$, and the maximum degree of the graph $G$ by $d = \max\{d_1, \ldots, d_p\}$. A potential game $G = (G, \mathcal{U})$ is then completely defined by a graph $G = ([p], E)$ and a collection of potential functions $\mathcal{U}(G) = \{u^i : A_{-i} \to \mathbb{R}\}_{i \in [p]}$, where each of the payoff functions $u^i(\cdot; G)$ decomposes according to (1). Finally, we will also assume that the number of strategies of each player, $m_i \define |A_i|$, is non-zero and $O(1)$ with respect to $p$ and $d$, and that $m \define \max\{m_i\}$.

The pure-strategy Nash equilibria (PSNE) set for the game $G = (G, \mathcal{U})$ is given by the set of strategy profiles where no player has any incentive to unilaterally deviate from its
strategy given the strategy profiles of its neighbors, and is defined as follows:

\[ \mathcal{NE}(G) = \left\{ x \in A \mid x_i \in \arg\max_{a \in A_i} u^i(a, x_{-i}) \right\}. \]  

(2)

The set of \( \varepsilon \)-Nash equilibria of the game \( G \) are those strategy profiles where each player can gain at most \( \varepsilon \) payoff by deviating from its strategy, and is defined as follows:

\[ \varepsilon\mathcal{NE}(G) = \left\{ x \in A \mid u^i(x_i, x_{-i}) \geq u^i(a, x_{-i}) - \varepsilon, \forall a \in A_i, \forall i \in [p] \right\}. \]  

(3)

We define the problem of learning potential games from data as follows. Assume that we are given "noisy" observations of strategy profiles, or joint actions, \( D = \{x^{(l)} \in A \}_{l \in [n]} \) drawn from a game \( G = (G, \mathcal{U}) \). The noise process models our uncertainty over the individual actions of the players due to observation noise, for instance, when we observe the actions through a noisy channel, or due to the unobserved dynamics of gameplay during which equilibrium is reached. By "observations drawn from a game" we simply mean that there exists a distribution \( P \), from which the strategy profiles are drawn, satisfying the following condition:

\[ \forall x, x' \text{ such that } x \in \mathcal{NE}(G), x' \in A \setminus \mathcal{NE}(G), \mathcal{P}(x) > \mathcal{P}(x'). \]

The above condition ensures that the signal level is more than the noise level. This should be compared with the observation model of Ghoshal and Honorio (2017), who assume that \( \forall x' \in A \setminus \mathcal{NE}(G), \mathcal{P}(x') > 0 \). Our observation model thus encompasses specific observation models considered in prior literature (Honorio and Ortiz, 2015; Ghoshal and Honorio, 2016): the global and local noise model. The global noise model is parameterized by a constant \( q \in (\mathcal{NE}(G)/|A|, 1) \) such that the probability of observing a strategy profile \( x \in A \) is given by a mixture of two uniform distributions:

\[ \mathcal{P}_g(x; G) = \frac{1}{|\mathcal{NE}(G)|} \mathbb{1}[x \in \mathcal{NE}(G)] + (1 - q) \frac{1}{|A| - |\mathcal{NE}(G)|} \mathbb{1}[x \notin \mathcal{NE}(G)]. \]  

(4)

In the local noise model, we observe strategy profiles \( x \) from the PSNE set with each entry (strategy) corrupted independently. Therefore, in the local noise model we have the following distribution over strategy profiles:

\[ \mathcal{P}_l(x; G) = \frac{1}{|\mathcal{NE}(G)|} \sum_{y \in \mathcal{NE}(G)} \prod_{i=1}^p (q_i)^{1[x_i = y_i]} \left( \frac{1 - q_i}{m_i - 1} \right)^{1[x_i \neq y_i]}, \]  

(5)

with \( q_i > 0.5 \) for all \( i \in [p] \).

The learning problem then corresponds to recovering a game \( \hat{G} = (\hat{G}, \hat{\mathcal{U}}) \) from \( D \) such that \( \mathcal{NE}(\hat{G}) = \mathcal{NE}(G) \) with high probability. Note that in our definition of the learning problem, we do not impose any restriction on the “closeness” of the recovered graph \( \hat{G} \) to the true graph \( G \). This is because multiple graphs \( G \) can give rise to the same PSNE set under different payoff functions and thus be unidentifiable from observations of joint actions alone.
3. Method

In this section, we describe our method for learning potential games from observational data. The individual and pairwise payoffs can be equivalently written, in linear form, as follows:

\[ u^{i,j}(x_i) = (\theta^{i,0})^T f^{i,0}(x_i), \quad u^{i,j}(x_i, x_j) = (\theta^{i,j})^T f^{i,j}(x_i, x_j), \]

where for \( j \in \mathcal{N}_i \), \( f^{i,j}(x_i, x_j) = (1 [x_i = a, x_j = b])_{a \in A_i, b \in A_j} \) and \( \theta^{i,j} = (\theta^{i,j}_{a,b})_{a \in A_i, b \in A_j} \), \( f^{i,0}(x_i) = (1 [x_i = a])_{a \in A_i} \) and \( \theta^{i,0} = (\theta^{i,0}_a)_{a \in A_i} \). Note that \( f^{i,j} \in \{0, 1\}^{(m_i m_j)}, \theta^{i,j} \in \mathbb{R}^{(m_i m_j)} \neq 0, f^{i,0}(x_i) \in \{0, 1\}^{m_i}, \text{ and } \theta^{i,0} \in \mathbb{R}^{m_i} \). Let

\[ \theta^i \overset{\text{def}}{=} (\theta^{i,0}, \theta^{i,1}, \ldots, \theta^{i,i-1}, \theta^{i,i+1}, \ldots, \theta^{i,p}), \]

\[ f^i(x_i, x_{-i}) \overset{\text{def}}{=} (f^{i,0}(x_i), f^{i,1}(x_i, x_1), \ldots, f^{i,i-1}(x_i, x_{i-1}), f^{i,i+1}(x_i, x_{i+1}), \ldots, f^{i,p}(x_i, x_p)), \]

with \( \theta^{i,j} = 0 \) for \( j > 0 \land j \notin \mathcal{N}_i \) and \( \theta^i \in \mathbb{R}^{(m_i + \sum_{j \in \mathcal{N}_i} m_j)}, f^i(x_i, x_{-i}) \in \{0, 1\}^{(m_i + \sum_{j \in \mathcal{N}_i} m_j)} \). Thus the payoff for the \( i \)-th player can be written, in linear form, as:

\[ u^i(x_i, x_{-i}) = (\theta^i)^T f^i(x_i, x_{-i}). \]

The learning problem then corresponds to learning the parameters \( \theta^i \) for each player \( i \). The sparsity pattern of \( \theta^i \) identifies the neighbors of \( i \). The way this differs from the binary strategies considered by [Ghoshal and Honorio, 2017] is that the parameters \( \theta^i \) have a group-sparsity structure, i.e., for all \( j > 0 \land j \notin \mathcal{N}_i \) the entire group of parameters \( \theta^{i,j} \) is zero. In order to ensure that the payoffs are finite, we will assume that the parameters for the \( i \)-th player belong to the set \( \Theta^i \overset{\text{def}}{=} \{y \in \mathbb{R}^{(m_i + \sum_{j \in \mathcal{N}_i} m_j)} \mid ||y||_\infty < \infty \} \).

Our approach for estimating the parameters \( \theta^i \) is to perform one-versus-rest multinomial logistic regression with \( \ell_{1,2} \) group-sparse regularization. In more detail, we obtain estimators \( \hat{\theta}^i \) by solving the following optimization problem for each \( i \in [p] \):

\[ \hat{\theta}^i = \arg \min_{\theta \in \Theta^i} L^i(D; \theta) + \lambda ||\theta||_{1,2}, \]

\[ L^i(D; \theta) = \frac{1}{n} \sum_{l=1}^n \ell^l(x^{(l)}; \theta), \quad \ell^l(x; \theta) = -\log \left( \frac{\exp(\theta^T f^l(x_i, x_{-i}))}{\sum_{a \in A_i} \exp(\theta^T f^l(a, x_{-i}))} \right), \]

where \( ||\theta||_{1,2} = \sum_{j \in [p]} ||\theta_j||_2 \), with \( \theta_j \) being the \( j \)-th group of \( \theta \). When referring to a block of a matrix or vector we will use bold letters, e.g., \( \theta_j \) denotes the \( j \)-th group or block of \( \theta \), while \( \theta_{j} \) denotes the \( j \)-th element of \( \theta \). In general, we define the \( \ell_{a,b} \) group structured norm as follows: \( ||\theta||_{a,b} = ||(||\theta_1||_a, \ldots, ||\theta_p||_b)||_a \). Also, when using group structured norms, we will use the group structure as shown in [7], i.e., we will assume that there are \( p \) groups and, in the context of the \( i \)-th player, the sizes of the groups are: \{\( m_i, m_i m_1, \ldots, m_i m_{i-1}, m_i m_{i+1}, \ldots, m_i m_p \}\). Finally, we will define the support set of \( \theta^i \) as the set of all indices corresponding to the active groups, i.e.,
S_i = \{(j,k)|j \in \{0\} \cup \mathcal{N}_i \text{ and } k \in [m_i] \text{ for } j = 0, k \in [m_im_j] \text{ for } j > 0\}, \text{ where } j \text{ can be thought of as indexing the groups, while } k \text{ can be thought of as indexing the elements within the } j\text{-th group. Thus, } |S_i| = m_i + \sum_{j \in \mathcal{N}_i} m_i m_j.

After estimating the parameters $\hat{\Theta}_i$ for each $i \in [p]$, the payoff functions are simply estimated to be $\hat{u}^i(x_i, x_{-i}) = (\hat{\Theta}_i)^T f^i(x_i, x_{-i})$. Finally, the graph $\hat{G} = ([p], \hat{E})$ is given by the group-sparsity structure of $\hat{u}$’s, i.e., $\hat{u}^{i,j}(\cdot, \cdot) \neq 0 \implies (i,j) \in \hat{E}$.

4. Sufficient Conditions

First, we obtain sufficient conditions on the number of samples $n$ to ensure successful PSNE recovery. Since our theoretical results depend on certain properties of the Hessian of the loss function defined above, we introduce the Hessian matrix in this paragraph. Let $\mathbf{H}^i(x; \theta)$ denote the Hessian of $\ell^i(x; \theta)$. A little calculation shows that the $(j,k)$-th block of the Hessian matrix for the $i$-th player is given as:

$$
\mathbf{H}_{j,k}^i(x; \theta) = \sum_{a \in A_i} \sigma^i(a, x_{-i}; \theta) f^{i,j}(a, x_j) f^{i,k}(a, x_k)^T - (\sum_{a \in A_i} \sigma^i(a, x_{-i}; \theta) f^{i,j}(a, x_j)) (\sum_{a \in A_i} \sigma^i(a, x_{-i}; \theta) f^{i,k}(a, x_k))^T,
$$

where we have overloaded the notation $f^{i,j}(x_i, x_j)$ to also include $f^{i,0}(x_i)$, i.e., we let $f^{i,0}(x_i, x_0) \overset{\text{def}}{=} f^{i,0}(x_i)$. We will denote the $i$-th expected Hessian matrix at any parameter $\theta \in \Theta^i$ as $\mathbf{H}^i(\theta) = \mathbb{E}_x [\mathbf{H}^i(x; \theta)]$, and the $i$-th Hessian matrix at the true parameter $\theta^i$ as $\mathbf{H}^i(\theta^i)$. We will also drop the superscript $i$ from the $i$-th Hessian matrix, whenever clear from context. We will denote the finite sample version of $\mathbf{H}^i(\theta^i)$ by $\mathbf{H}^i(D, \theta^i)$, i.e., $\mathbf{H}^i(D, \theta^i) = \frac{1}{n} \sum_{i=1}^n \mathbf{H}^i(x^{(i)}, \theta^i)$. Finally, we will denote the Hessian matrix restricted to the true support set $S_i$ by: $\mathbf{H}^i(\cdot; \theta^i_{S_i}) \in \mathbb{R}^{|S_i| \times |S_i|}$. In order to prove our main result, we will present a series of technical lemmas slowly building towards our main result. Detailed proofs of the lemmas are given in Appendix A.

The following lemma states that the $i$-th population Hessian is positive definite. Specifically, the $i$-th population Hessian evaluated at the true parameter $\theta^i$, are positive definite with the minimum eigenvalue being $C_{\text{min}}$. We prove the following lemma by showing that the loss function given by $[11]$, when restricted to an arbitrary line, is strongly convex as long as the payoffs are finite.

**Lemma 1** (Minimum eigenvalue of population Hessian). For $\theta^i \in \Theta^i$, $\lambda_{\text{min}}(\mathbf{H}^i(\theta^i)) \overset{\text{def}}{=} C_{\text{min}} > 0$.

Given that population Hessian matrices are positive-definite, we then show that the finite sample Hessian matrices, evaluated at any parameter $\theta_{S_i}$, are positive definite with high probability. We use tools from random matrix theory developed by [Tropp 2012] to prove the following lemma.
Lemma 2 (Minimum eigenvalue of finite sample Hessian). Let $\theta \in \Theta^i$ be any arbitrary vector and let $\lambda_{\min}(H^i(\theta_{S_i})) \overset{\text{def}}{=} \lambda_{\min} > 0$. Then, if the number of samples satisfies the following condition:

$$
n \geq \frac{8(d_i + 1)}{\lambda_{\min}} \log \left( \frac{m_i(1 + d_i m)}{\delta} \right),$$

then $\lambda_{\min}(H^i(D; \theta_{S_i})) \geq \frac{\lambda_{\min}}{2}$ with probability at least $1 - \delta$ for some $\delta \in (0, 1)$.

Now that we have shown that the loss function given by (11) is strongly convex (Lemmas 1 and 2), we exploit strong convexity to control the difference between the true parameter and the estimator $\|\theta^i - \hat{\theta}^i\|_{1,2}$. However, before proceeding further, we need to bound the $\ell_{\infty,2}$ norm of the gradient, as done in the following lemma. We prove the lemma by using McDiarmid’s inequality to show that in each group the finite sample gradient concentrates around the expected gradient, and then use a union bound over all the groups to control the $\ell_{\infty,2}$ norm.

Lemma 3 (Gradient bound). Let $\|E_{x} \left[ \nabla \ell^i(x; \theta^i) \right] \|_{\infty,2} = \nu$, then we have that

$$
\|\nabla L^i(D; \theta^i)\|_{\infty,2} \leq \nu + \sqrt{\frac{2}{n} \log \left( \frac{2(d_i + 1)}{\delta} \right)},
$$

with probability at least $1 - \delta$.

Note that the expected gradient at the parameter $\theta^i$ does not vanish, i.e., $\|E_{x} \left[ \nabla \ell^i(x; \theta^i) \right] \|_{\infty,2} = \nu$. This is because of the mismatch between the generating distribution $P$ and the softmax distribution used for learning the parameters, as in (11). Indeed, if the data were drawn from a Markov random field, which induces a softmax distribution on the conditional distribution of node given the rest of the nodes, the parameter $\nu = 0$. However this is not the case for us. An unfortunate consequence of this is that, even with an infinite number of samples, our method will not be able to recover the parameters $\theta^i$ exactly. Thus, without additional assumptions on the payoffs, our method only recovers the $\varepsilon$-Nash equilibrium of the game.

With the required technical results in place, we are now ready to bound $\|\theta^i - \hat{\theta}^i\|_{1,2}$. Our analysis has two steps. First, we bound the norm in the true support set, i.e., $\|\theta^i_{S_i} - \hat{\theta}^i_{S_i}\|_{1,2}$. Then, we show that the norm of the difference between the true parameter and the estimator, outside the support set, is a constant factor (specifically 3) of the difference in the support set. For the first step with use a proof technique originally developed by Rothman et al. (2008) in a different context, while the second step follows from matrix algebra and optimality of the estimator $\hat{\theta}^i$ for the problem (10).

The following technical lemma, which will be used later on in our proof to bound $\|\hat{\theta}^i_{S_i} - \hat{\theta}^i_{S_i}\|_{1,2}$, lower bounds the minimum eigenvalue of the $i$-th population Hessian at an arbitrary parameter $\theta \in \Theta^i$, in terms of the minimum eigenvalue of the $i$-th population Hessian at the true parameter $\theta^i$. 


**Lemma 4** (Minimum population eigenvalue at arbitrary parameter). Let $\theta \in \Theta^i$ be any vector. Then the minimum eigenvalue of $i$-th population Hessian matrix evaluated at $\theta_{S_i}$ is lower bounded as follows:

$$\lambda_{\min}(H^i(\theta_{S_i})) \geq \lambda_{\min}(H^i(\theta_{S_i}^i)) - \frac{1}{4} (d_i + 1) m^2 \|\theta_{S_i} - \theta_{S_i}^i\|_{1,2}.$$ 

Now, we are ready to bound the difference between the true parameter $\theta^i$ and its estimator $\hat{\theta}^i$, in the true support set $S_i$.

**Lemma 5** (Error of the $i$-th estimator on the support set). If the regularization parameter and number of samples satisfy the following condition:

$$\lambda \geq 2 \left( \nu + \sqrt{\frac{2 \log \left( \frac{2(d_i + 1)}{\delta} \right)}{n}} \right), \quad n > \frac{2}{N(m, d)} \log \left( \frac{2(d_i + 1)}{\delta} \right),$$

where $N(m, d) = \left\{ \frac{C_{\text{min}}}{(36m^2(d+1)^2) - \nu} \right\}^2$, and $C_{\text{min}} \overset{\text{def}}{=} \lambda_{\min}(H^i(\theta_{S_i}^i))$; then with probability at least $1 - \delta$, for some $\delta \in (0, 1)$, we have:

$$\|\hat{\theta}_{S_i}^i - \theta_{S_i}^i\|_{1,2} \leq \frac{6(d_i + 1)}{C_{\text{min}}} \lambda.$$ (14)

Next, we bound the difference between the true parameter $\theta^i$ and its estimator $\hat{\theta}^i$.

**Lemma 6** (Error of the $i$-th parameter estimator). Under the same conditions on the regularization parameter and number of samples as in Lemma 5, we have, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$,

$$\|\hat{\theta}^i - \theta^i\|_{1,2} \leq \frac{24(d_i + 1)}{C_{\text{min}}} \lambda.$$

Now that we have control over $\|\theta^i - \hat{\theta}^i\|_{1,2}$ for all $i \in [p]$, we are ready to prove our main result concerning the sufficient number of samples needed by our method to guarantee PSNE recovery with high probability.

**Theorem 1.** Let $G = (G, \mathcal{U})$, with $\mathcal{U} = \{u^i : A_{-i} \to \mathbb{R}\}_{i \in [p]}$, be the true potential graphical game over $p$ players and maximum degree $d$, from which the data set $D$ is drawn. Let $\hat{G} = (\hat{G}, \hat{\mathcal{U}})$, with $\hat{\mathcal{U}} = \{\hat{u}^i : A_{-i} \to \mathbb{R}\}_{i \in [p]}$, be the game learned from the data set $D$ by solving the optimization problem (10) for each $i \in [p]$. Then if the regularization parameter and the number of samples satisfy the condition:

$$\lambda \geq 2 \left( \nu + \sqrt{\frac{2 \log \left( \frac{2p(d + 1)}{\delta} \right)}{n}} \right), \quad n > \max \left\{ \frac{2}{N(m, d)} \log \left( \frac{2p(d + 1)}{\delta} \right), \frac{8(d + 1)}{C_{\text{min}}} \log \left( \frac{m(1 + dm)}{\delta} \right) \right\},$$

where $N(m, d) = \left\{ \frac{C_{\text{min}}}{(36m^2(d+1)^2) - \nu} \right\}^2$, then we have that the following hold with probability at least $1 - \delta$, for some $\delta \in (0, 1)$:
(i) \( \mathcal{NE}(\hat{G}) = \varepsilon \cdot \mathcal{NE}(G) \), with \( \varepsilon = \frac{48(d_i + 1)}{C_{\min}} \).

(ii) Additionally, if the true game \( G \) satisfies the condition: \( \forall i \in [p], \forall (x_i, x_{-i}), (x'_i, x_{-i}) \in \mathcal{A} \) such that \((x_i, x_{-i}) \in \mathcal{NE}(G) \land (x'_i, x_{-i}) \notin \mathcal{NE}(G) \implies u^i(x_i, x_{-i}) > u^i(x'_i, x_{-i}) + \varepsilon \). Then, \( \mathcal{NE}(\hat{G}) = \mathcal{NE}(G) \).

Proof. Note that \( \|f^i(x_i, x_{-i})\|_{\infty, 2} = \max\{\|f^i,0(x_i)\|_2, \|f^i,1(x_i, x_1)\|_2, \ldots, \|f^i,p(x_i, x_p)\|_2\} = 1 \), for any \( x \in \mathcal{A} \), since each binary vector \( f^{i,j}(x_i, x_j) \) has a single “1” at exactly one location. Then, from the Cauchy-Schwartz inequality, Lemma 6, and a union bound over all players, we have that:

\[
(\forall x \in \mathcal{A}, \forall i \in [p]) \|\hat{u}^i(x_i, x_{-i}) - u^i(x_i, x_{-i})\| = \|\hat{\theta}^i - \theta^i\|T \| f^i(x_i, x_{-i})\|_{\infty, 2} \\
\leq \|\hat{\theta}^i - \theta^i\|_1 \| f^i(x_i, x_{-i})\|_{\infty, 2} \\
= \|\hat{\theta}^i - \theta^i\|_1 \leq \frac{24(d_i + 1)}{C_{\min}} \lambda = \frac{\varepsilon}{2},
\]

with probability at least \( 1 - \rho \delta \). Now consider any \( x \in \mathcal{NE}(\hat{G}) \) and any \( i \in [p] \). Since \( x \in \mathcal{NE}(\hat{G}) \), we have from (15):

\[
u^i(x_i, x_{-i}) + \varepsilon/2 \geq \hat{u}^i(x_i, x_{-i}) \geq \hat{u}^i(x'_i, x_{-i}) \quad (\forall x'_i \in \mathcal{A}_i) \]

\[
u^i(x_i, x_{-i}) \geq \hat{u}^i(x'_i, x_{-i}) - \varepsilon \quad (\forall x'_i \in \mathcal{A}_i) \]

where the last line again follows from (15). This proves that \( \mathcal{NE}(\hat{G}) \subseteq \varepsilon \cdot \mathcal{NE}(G) \). Using exactly the same arguments as above, we can also show that for any \( x \in \mathcal{NE}(\hat{G}) \):

\[
\hat{u}^i(x_i, x_{-i}) \geq \hat{u}^i(x_i, x_{-i}) - \varepsilon \quad (\forall x'_i \in \mathcal{A}_i),
\]

which proves that \( \mathcal{NE}(\hat{G}) \subseteq \varepsilon \cdot \mathcal{NE}(G) \). Thus we have that \( \mathcal{NE}(\hat{G}) = \varepsilon \cdot \mathcal{NE}(G) \), i.e., the set of joint strategy profiles \( x \in \mathcal{NE}(\hat{G}) \) form an \( \varepsilon \)-Nash equilibrium set of the true game \( G \). This proves our first claim. For our second claim, consider any \((x_i, x_{-i}) \in \mathcal{NE}(G) \) and \((x'_i, x_{-i}) \notin \mathcal{NE}(G) \). Then:

\[
u^i(x_i, x_{-i}) > u^i(x'_i, x_{-i}) + \varepsilon \implies \hat{u}^i(x_i, x_{-i}) > \hat{u}^i(x'_i, x_{-i}) \geq \hat{u}^i(x'_i, x_{-i}) - \varepsilon/2 + \varepsilon \]

\[
\implies \hat{u}^i(x_i, x_{-i}) > \hat{u}^i(x'_i, x_{-i}) \implies (x_i, x_{-i}) \in \mathcal{NE}(\hat{G}) \land (x'_i, x_{-i}) \notin \mathcal{NE}(\hat{G}),
\]

where the first line holds by assumption, and the second line again follows from (15). Thus we have that \( \mathcal{NE}(\hat{G}) = \mathcal{NE}(G) \). By setting the probability of error \( \rho \delta = \delta' \) for some \( \delta' \in (0, 1) \) we prove our claim. The second part of the lower bound on the number of samples is due to Lemma 2.

Remark 1. The sufficient number of samples needed by our method to guarantee PSNE recovery, with probability at least \( 1 - \delta \), scales as \( O \left( m^4 d^4 \log(p d^4) \right) \). This should be compared with the results of Jalali et al. (2011) for learning undirected graphical models. They show that \( O \left( m^2 d^2 \log(m^2 p) \right) \) are sufficient for learning \( m \)-ary discrete graphical
models. However, their sample complexity hides a constant $K$ that is related to the maximum eigenvalue of the scatter matrix, which we have upper bounded by $m^2d^2$ in our case, leading to a slightly higher sample complexity.

**Remark 2.** Note that as $n \to \infty$, the regularization parameter $\lambda \to 2\nu$, where $\nu$ is the maximum norm of the expected gradient at the true parameter $\theta^t$ across all $i \in [p]$. Thus, even with an infinite number of samples, our method recovers the $\epsilon$-Nash equilibria set of the true game with $\epsilon \to \frac{96(d+1)\nu}{C_{\min}}$ as $n \to \infty$.

5. Necessary Conditions

In this section, we obtain an information-theoretic lower bound on the number of samples needed to learn sparse potential games. Let $\mathfrak{G}_{p,d,m}$ be set of potential games over $p$ players, with degree at most $d$, and maximum number of strategies per player being $m$. Our approach for doing so is to treat the inference procedure as a communication channel, where nature picks a game $G^*$ from the set $\mathfrak{G}_{p,d,m}$ and then generates a data set $D$ of $n$ strategy profiles. A decoder $\psi : \mathcal{A}^n \to \mathfrak{G}_{p,d,m}$ then maps $D$ to a game $\hat{G} \in \mathfrak{G}_{p,d,m}$. We wish to obtain lower bounds on the number of samples required by any decoder $\psi$ to recover the true game consistently. In this setting, we define the minimax estimation error as follows:

$$p_{err} = \min_{\psi} \sup_{G^* \in \mathfrak{G}_{p,d,m}} \Pr \{ N\mathcal{E}(\psi(D)) \neq N\mathcal{E}(G^*) \},$$

where the probability is computed over the data distribution. For obtaining necessary conditions on the sample complexity, we assume that the data distribution follows the global noise model described in (4). The following theorem prescribes the number of samples needed for learning sparse potential games. Our proof of the theorem constitutes constructing restricted ensembles of “hard-to-learn” potential games, from which nature picks a game uniformly at random and generates data. We then use the Fano’s technique to lower bound the minimax error. The use of restricted ensembles is customary for obtaining information-theoretic lower bounds, c.f. (Santhanam and Wainwright 2012, Wang et al. 2010).

**Theorem 2.** If the number of samples $n \leq \frac{\log m + \log (d)}{2\log 2} - 1$, then estimation fails with $p_{err} \geq \frac{1}{2}$.

**Proof.** Consider the following restricted ensemble $\tilde{\mathfrak{G}} \subset \mathfrak{G}_{p,d,m}$ of $p$-player potential games with degree $d$, and the set of pure-strategies of each player being $\mathcal{A}_i = [m]$. Each $G = (G, U) \in \tilde{\mathfrak{G}}_{p,d,m}$ is characterized by a set $\mathcal{I}$ of influential players, and a set $\mathcal{I}^c \overset{\text{def}}{=} [p] \setminus \mathcal{I}$ of non-influential players, with $|\mathcal{I}| = d$. The graph $G$ is a complete (directed) bipartite graph from the set $\mathcal{I}$ to $\mathcal{I}^c$. Further with each game $G \in \tilde{\mathfrak{G}}_{p,d,m}$ is associated a strategy $a \in [m]$ that determines the individual and pairwise payoffs for any $i \in \mathcal{I}$ and $j \in \mathcal{I}^c$ as follows:

$$u_i^a(x_i) = 1[x_i = a], u_i^a(x_j) = 0(\forall x_j \in [m]), u_i^{a,j}(x_j, x_i) = 1[x_i = a, x_j = (a \mod m) + 1].$$
Thus, each $G \in \tilde{\mathcal{G}}$ game has a exactly one unique Nash equilibrium where the influential players play $a$ and the non-influential players play $(a \mod m) + 1$. Thus we have that $|\tilde{\mathcal{G}}| = m^d$. Nature picks a game $G$ uniformly at random from $\tilde{\mathcal{G}}$ by randomly selecting a set of $d$ players as “influential” and then randomly picking a strategy $a \in [m]$ for the influential players. Nature then generates a dataset $D$ using the global noise model with parameter $q \in (\frac{1}{m^p}, \frac{2}{(m^p+1)})$. Then from the Fano’s inequality we have that:

$$p_{err} \geq 1 - \frac{I(D; G) + \log 2}{H(G)}, \quad (16)$$

where $I(.)$ and $H(.)$ denote mutual information and entropy respectively. The mutual information $I(D; G)$ can be bounded, using a result by [Yu (1997)], as follows:

$$I(D; G) \leq \frac{1}{|\tilde{\mathcal{G}}|^2} \sum_{G_1 \in \tilde{\mathcal{G}}} \sum_{G_2 \in \tilde{\mathcal{G}}} \text{KL}(P_{D|G=\hat{G}_1} \| P_{D|G=\hat{G}_2}), \quad (17)$$

where $P_{D|G=\hat{G}_1}$ (respectively $P_{D|G=\hat{G}_2}$) denotes the data distribution under $G_1$ (respectively $G_2$). The KL divergence term from (17) can be bounded as follows:

$$\text{KL}(P_{D|G=\hat{G}_1} \| P_{D|G=\hat{G}_2}) = \frac{n}{\sum_{x \in \mathcal{A}} q \log \frac{q(m^p - 1)}{1 - q} + \sum_{x \in \mathcal{NE}(G_2)} (1 - q) \log \frac{1 - q}{q(m^p - 1)}} \leq n \log 2, \quad (18)$$

where the first line follows from the fact that the samples are i.i.d, and the second line follows from the fact the the distributions $P_{D|G=\hat{G}_1}$ and $P_{D|G=\hat{G}_2}$ assign the same probability to $x \in \mathcal{A} \setminus (\mathcal{NE}(G_1) \cup \mathcal{NE}(G_2))$. The last line follows from the fact that $q \in (\frac{1}{m^p}, \frac{2}{(m^p+1)})$, which implies $\frac{1 - q}{(m^p - 1)} < q \leq \frac{2(1 - q)}{m^p - 1}$. Putting together (16), (17) and (18), we have that if

$$n \leq \frac{\log m + \log \binom{p}{d}}{2 \log 2} - 1,$$

then $p_{err} \geq \frac{1}{2}$. Since, learning the ensemble $\mathcal{G}$ is at least as hard as learning a subset of $\mathcal{G}$, our claim follows.

**Remark 3.** From the above theorem we have that, the number of samples needed by any conceivable method grows as $\Omega(\log m + d \log p)$, assuming that $d = o(p)$. Therefore, the method based on $\ell_{1,2}$-regularized logistic regression is information-theoretically optimal in the number of players, for learning sparse potential games.

Synthetic experiments validating our theoretical results can be found in Appendix B.
Concluding Remarks. We conclude this exposition with a discussion of potential avenues for future work. In this paper we considered the problem of learning a very general, and widely used, class of graphical games called potential games, involving players with pure strategies. One can also consider mixed strategies, which would entail learning distributions, instead of “sets”, under the framework of non-cooperative maximization of utility. Further, one can also consider other solution concepts like correlated equilibria.

References

Boyd, S. and Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Garg, V. and Jaakkola, T. (2016). Learning Tree Structured Potential Games. In Advances in Neural Information Processing Systems 29, pages 1552–1560.

Ghoshal, A. and Honorio, J. (2016). From behavior to sparse graphical games: Efficient recovery of equilibria. In 54th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2016, Monticello, IL, USA, September 27-30, 2016, pages 1220–1227.

Ghoshal, A. and Honorio, J. (2017). Learning graphical games from behavioral data: Sufficient and necessary conditions. In Artificial Intelligence and Statistics, pages 1532–1540.

Honorio, J. and Ortiz, L. (2015). Learning the structure and parameters of large-population graphical games from behavioral data. Journal of Machine Learning Research, 16:1157–1210.

Irfan, M. T. and Ortiz, L. E. (2014). On influence, stable behavior, and the most influential individuals in networks: A game-theoretic approach. Artificial Intelligence, 215:79–119.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011). On Learning Discrete Graphical Models using Group-Sparse Regularization. In AISTATS, pages 378–387.

Janovskaja, E. (1968). Equilibrium situations in multi-matrix games. Litovskii Matematicheskii Sbornik, 8:381–384.

Johnson, C. R. and Nylen, P. (1991). Monotonicity properties of norms. Linear Algebra and its Applications, 148:43–58.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. Electronic Journal of Statistics, 2:494–515.

Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. Information Theory, IEEE Transactions on, 58(7):4117–4134.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. Foundations of computational mathematics, 12(May):389–434.
Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for Gaussian Markov random fields. In *ISIT*, pages 1373–1377. Citeseer.

Yu, B. (1997). *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, chapter Assouad, Fano, and Le Cam, pages 423–435. Springer New York, New York, NY.
Appendix A Detailed Proofs

Proof of Lemma 1 (Minimum eigenvalue of population Hessian).
Fix any \( \theta^0, \theta^1 \in \Theta^i \), with \( \theta^1 \neq 0 \). For any \( t \in (-\infty, \infty) \), let \( F(t; x_i) \overset{\text{def}}{=} (\theta^0 + t\theta^1)^T f(x_i, x_{-i}) \). Then for \( x \in A^i \),
\[
\ell(x; \theta^0 + t\theta^1) = -F(t; x_i) + \log(\sum_{a \in A^i} \exp(F(t; a))).
\] (19)

A little calculation shows that the double derivative of \( \ell(x; \theta^0 + t\theta^1) \) with respect to \( t \) is as follows:
\[
\frac{\partial^2 \ell(x; \theta^0 + t\theta^1)}{\partial t^2} = \sum_{a \in A^i} \sigma(t; a) F'(a)^2 - \left( \sum_{a \in A^i} \sigma(t; a) F'(a) \right)^2,
\] (20)
where \( F'(a) \) is the derivative of \( F(t; a) \) with respect to \( t \). Since \( F(t; a) \) is a linear function of \( t \), \( F'(a) \) is not a function of \( t \). Also note that \( \sum_{a \in A^i} \sigma(t; a) = 1 \). Since \( \theta^0, \theta^1 \) have bounded norm and \( t \in (-\infty, \infty) \), we have that \( \sigma(t; a) > 0 \), \( \forall a \in A^i \). Therefore, from (20), the strict convexity of \( \cdot^2 \) and Jensen’s inequality, we have:
\[
\frac{\partial^2 \ell(x; \theta^0 + t\theta^1)}{\partial t^2} > 0 \quad (\forall t \in (-\infty, \infty)).
\]
Thus we have that \( \ell(x, \theta) \) is strongly convex, i.e., \( \lambda_{\min}(H^i(x; \theta)) > 0 \), \( \forall \theta \in \Theta^i \). Finally, by concavity of \( \lambda_{\min}(\cdot) \) [Boyd and Vandenberghe 2004] and the Jensen’s inequality we have:
\[
\lambda_{\min}(H^i(\theta^i)) = \lambda_{\min}(\mathbb{E}_x [H^i(x; \theta^i)]) \geq \mathbb{E}_x [\lambda_{\min}(H^i(x; \theta^i))] > 0.
\]

Proof of Lemma 2 (Minimum eigenvalue of finite sample Hessian).
To simply notation in the proof we will denote \( S_i \) by \( S \). The \((j, k)\) block of \( H(D; \theta_S) \), where \( j, k \in \{0\} \cup N_i \), can be written as:
\[
H_{j,k}(D; \theta_S) = \sum_{a \in A^i} \sigma^i(a, x_{-i}; \theta_S) f^{i,j}(a, x_j)(f^{i,k}(a, x_k))^T - \sum_{a, b \in A^i} \sigma^i(a, x_{-i}; \theta_S) f^{i,j}(a, x_j)f^{i,k}(b, x_k)^T,
\]

Learning Sparse Potential Games in Polynomial Time and Sample Complexity
where the matrices $B$ and $R$ have been defined above (blockwise). Since the matrix $R$ is positive semi-definite $\lambda_{\max}(H(D; \theta_S)) \leq \lambda_{\max}(B(D; \theta_S))$. Further, since $B$ is positive semi-definite, we have from Lemma 7:

$$
\lambda_{\max}(B(D; \theta_S)) \leq \sum_{j \in \{0\} \cup N_i} \lambda_{\max}(B_{ij}(D; \theta_S))
$$

$$
\leq (d_i + 1) \max_{j \in \{0\} \cup N_i} \lambda_{\max}\left(\frac{1}{n} \sum_{l=1}^{n} \sigma^i(a, x_j^{(l)}; \theta_S) f^{i,j}(a, x_j^{(l)}) (f^{i,j}(a, x_j^{(l)}))^T\right)
$$

$$
\leq (d_i + 1) \frac{1}{n} \sum_{l=1}^{n} \sigma^i(a, x_j^{(l)}; \theta_S) \lambda_{\max}\left(f^{i,j}(a, x_j^{(l)}) (f^{i,j}(a, x_j^{(l)}))^T\right)
$$

$$
= d_i + 1.
$$

Thus we have that $\lambda_{\max}(H(D; \theta_S)) \leq \lambda_{\max}(B(D; \theta_S)) \leq d_i + 1 \overset{\text{def}}{=} R$. Also note that $H(D; \theta_S) \in \mathbb{R}^{S \times S}$, with $|S| \leq m_i(1 + d_i m)$. Then using the matrix Chernoff bounds by Tropp (2012), we have:

$$
\Pr \left\{ \lambda_{\min}(H(D; \theta_S)) \leq (1 - \delta) \lambda_{\min} \right\} \leq |S| \left(\frac{\exp(-\delta)}{(1 - \delta)^{1 - \delta}}\right)^{\lambda_{\min}/n}
$$

Setting $\delta = 1/2$ we get:

$$
\Pr \left\{ \lambda_{\min}(H(D; \theta_S)) \geq \frac{\lambda_{\min}}{2} \right\} \geq 1 - m_i(1 + d_i m) \exp\left(-\frac{n\lambda_{\min}}{8(d_i + 1)}\right)
$$

Controlling the probability of error to be at most $\delta$ we obtain the lower bound on the number of samples. \hfill \Box

**Proof of Lemma 3** (Gradient bound).

A simple calculation shows that

$$
\frac{\partial \ell_i(x; \theta^i)}{\partial \theta^i} = -f^{i,j}(x_i, x_j) + \sum_{a \in A_i} \sigma^i(a, x_{-i}; \theta^i) f^{i,j}(a, x_j), \quad (21)
$$

where $\sigma^i(\cdot)$ has been defined in [13]. Let $g_j(x^{(1)}, \ldots, x^{(n)}) = (g_j(x^{(1)}, \ldots, x^{(n)}))_{j \in \{0\} \cup N_i}$, where $g_j(\cdot) = \|\frac{1}{n} \sum_{l=1}^{n} \frac{\partial \ell_j(x_j; \theta^j)}{\partial \theta^j}\|_2$. Then $\|g(\cdot)\|_\infty = \|\nabla L^j(D; \theta^j)\|_\infty, 2$ and $\|E_x [g(\cdot)]\|_\infty = \|E_x [\nabla \ell^j(x; \theta)]\|_\infty, 2 = \nu$. Then, for any $x^{(l)} \neq x^{(l)}'$ we have that:

$$
|g_j(x^{(1)}, \ldots, x^{(n)}) - g_j(x^{(1)}, \ldots, x^{(l)}, \ldots, x^{(n)})| = \frac{1}{n} \left\| f^{i,j}(x_i, x'_j) - f^{i,j}(x_i, x_j) + \sum_{a \in A_i} \sigma^i(a, x_j; \theta^i) f^{i,j}(a, x_j) - \sigma^i(a, x_j; \theta^i) f^{i,j}(a, x'_j) \right\|_2
$$

$$
\leq \frac{1}{n} \left(2 + \sum_{a \in A_i} (\sigma^i(a, x_j; \theta^i))^2 + (\sigma^i(a, x_j; \theta^i))^2\right)^{1/2} \leq \frac{1}{n}(2 + 2)^{1/2} = 2/n,
$$
where in the last line we used the fact that \( \sum \sigma^i(a, \cdot) = 1 \) along with the Cauchy-Schwartz inequality. Then using the McDiarmid’s inequality we have:

\[
\Pr \{|g_j(\cdot) - \mathbb{E}_x[g_j(\cdot)]| \leq t\} \geq 1 - 2 \exp\left(-\frac{nt^2}{2}\right).
\]

Then using a union bound over all \( j \) we have:

\[
\Pr \left\{ \max_j |g_j(\cdot) - \mathbb{E}_x[g_j(\cdot)]| \leq t \right\} \geq 1 - 2(d_i + 1) \exp\left(-\frac{nt^2}{2}\right)
\]

\[
\implies \Pr \{|g(\cdot) - \mathbb{E}_x[g(\cdot)]|_\infty \leq t\} \geq 1 - 2(d_i + 1) \exp\left(-\frac{nt^2}{2}\right)
\]

\[
\implies \Pr \{|g(\cdot)|_\infty - |\mathbb{E}_x[g(\cdot)]|_\infty \leq t\} \geq 1 - 2(d_i + 1) \exp\left(-\frac{nt^2}{2}\right)
\]

\[
\implies \Pr \{|g(\cdot)|_\infty \leq \nu + t\} \geq 1 - 2(d_i + 1) \exp\left(-\frac{nt^2}{2}\right),
\]

where in the third line we used the reverse triangle inequality. Setting the probability of error to be \( \delta \) and solving for \( t \), we prove our claim. \( \square \)

**Proof of Lemma 3 (Minimum population eigenvalue at arbitrary parameter).**

To simplify notation in the proof we will denote \( S_i \) by \( S \). The population Hessian matrix at \( \mathbf{H}(\theta_S^\dagger) \) can also be written as \( \mathbf{H}(\theta_S^\dagger + \Delta S) \), where \( \Delta S = \theta_S - \theta_S^\dagger \). Using the variational characterization of the minimum eigenvalue of \( \mathbf{H}(\theta_S^\dagger + \Delta S) \) and the Taylor’s theorem, we have:

\[
\lambda_{\min}(\mathbf{H}(\theta_S^\dagger + \Delta S)) = \min_{\{y \in \mathbb{R}^{|S|}||y||_2 = 1\}} \sum_{i,j \in S} y_i \{H_{i,j}(\theta_S^\dagger) + (\nabla H_{i,j}(\bar{\theta}))^T \Delta_S\} y_j
\]

\[
\geq \lambda_{\min}(\mathbf{H}(\theta_S^\dagger)) - \max_{\{y \in \mathbb{R}^{|S|}||y||_2 = 1\}} \sum_{i,j \in S} y_i \{((\nabla H_{i,j}(\bar{\theta}))^T \Delta_S\} y_j
\]

\[
\geq \lambda_{\min}(\mathbf{H}(\theta_S^\dagger)) - \max_{\{y \in \mathbb{R}^{|S|}||y||_2 = 1\}} \sum_{i,j \in S} y_i \{|((\nabla H_{i,j}(\bar{\theta}))^T \Delta_S|\} y_j, \quad (22)
\]

where \( \bar{\theta} = t\theta_S^\dagger + (1 - t)\theta_S \) for some \( t \in [0,1] \), and the third line follows from the monotonicity property of the spectral norm \( ||\cdot||_2 \) [Johnson and Nylen, 1991]. For any vector \( \theta \in \Theta^t \), let \( \mathbf{A}(\theta_S) = (A_{i,j}(\theta_S)) \), where \( A_{i,j}(\theta_S) = \langle (\nabla H_{i,j}(\theta_S))^T \Delta_S \rangle \). Then,

\[
||\mathbf{A}(\theta_S)||_2 = ||\mathbb{E}_x[\mathbf{A}(x; \theta_S)]||_2 \leq \max_{x \in \mathcal{A}} ||\mathbf{A}(x; \theta_S)||_2. \quad (23)
\]

Now consider the \( (j, k) \) block of \( \mathbf{A}(x; \theta_S) \) for any \( x \in \mathcal{A} \), where \( j, k \in \{0\} \cup \mathcal{N}_i \). Then, from [12] we have that:

\[
\mathbf{A}_{j,k}(x; \theta) = \sum_{a \in \mathcal{A}_i} \langle (\nabla \sigma^i(a, x_{-i}; \theta))^T \Delta_S \rangle \mathbf{f}^{i,j}(a, x_j)(\mathbf{f}^{i,k}(a, x_k))^T
\]

\[
+ \sum_{b \in \mathcal{A}_i} \langle \sigma^i(b, x_{-i}; \theta) \nabla \sigma^i(a, x_{-i}; \theta) + \sigma^i(a, x_{-i}; \theta) \nabla \sigma^i(b, x_{-i}; \theta) \rangle^T \Delta_S \mathbf{f}^{i,j}(a, x_j)\mathbf{f}^{i,k}(a, x_k)^T.
\]

\[
R_{j,k}(x; \theta)
\]

16
Thus, \( A(x; \theta) = B(x; \theta) - R(x; \theta) \), where the matrices \( B \) and \( R \) have been defined above (block-wise). Observe that the matrix \( R \) is positive semi-definite. Therefore, \( \|A(x; \theta)\|_2 \leq \|B(x; \theta)\|_2 \). Finally, since \( B \) is positive semi-definite, the spectral norm of \( B \) is at most the sum of the spectral norms of the diagonal blocks (c.f. Lemma 7). Therefore, we have

\[
\|B(x; \theta)\|_2 \leq \sum_{j \in \{0\} \cup N_i} \|B_{j,j}(x; \theta)\|_2 \leq (d_i + 1) \left( \max_{j \in \{0\} \cup N_i} \|B_{j,j}(x; \theta)\|_2 \right).
\] (24)

A little calculation shows that

\[
\frac{\partial \sigma^i(a, x_{-i}; \theta)}{\partial \theta_j} = \sigma^i(a, x_{-i}; \theta) \left\{ f^{i,j}(a, x_j) - \sum_{a' \in A_i} \sigma^i(a', x_{-i}; \theta) f^{i,j}(a', x_j) \right\},
\]

and \( \|\partial \sigma^i(a, x_{-i}; \theta)/\partial \theta_j\|_\infty \leq 1/4 \). Further, since for any given \( a \in A_i \), at most \( m_j + 1 \) elements of the partial derivative vector above are non-zero, we have \( \|\partial \sigma^i(a, x_{-i}; \theta)/\partial \theta_j\|_2 \leq (m_j + 1)/4 \) and \( \|\nabla \sigma^i(a, x_{-i}; \theta)\|_\infty \leq (m_j + 1)/4 \leq (m + 1)/4 \). Then using the Cauchy-Schwartz inequality and the monotonicity property of spectral norm [Johnson and Nylen, 1991] we have:

\[
\|B_{j,j}(x; \theta)\|_2 \leq \left\| \sum_{a \in A_i} \|\nabla \sigma^i(a, x_{-i}; \theta)\|_\infty 2 \|\Delta_S\|_{1,2} f^{i,j}(a, x_j)(f^{i,j}(a, x_j))^T \right\|_2 \leq \frac{1}{4}m_i m \|\Delta_S\|_{1,2}.
\] (25)

Putting together (22), (23), (24) and (25) we get

\[
\lambda_{\min}(H(\theta^*_S + \Delta_S)) \geq \lambda_{\min}(H(\theta^*_S)) - \|A(\theta_S)\|_2
\]

\[
\geq \lambda_{\min}(H(\theta^*_S)) - (d_i + 1) \left( \max_{x \in A_i} \max_{j \in \{0\} \cup N_i} \|B_{j,j}(x; \theta)\|_2 \right)
\]

\[
\geq \lambda_{\min}(H(\theta^*_S)) - \frac{1}{4}(d_i + 1)m_i m \|\Delta_S\|_{1,2}.
\]

\( \square \)

**Proof of Lemma 3 (Error of the i-th estimator on the support set).**

To simplify notation in the proof, we will write \( S \) instead of \( S_i \). Recall that \( L^i(D; \theta) \) is the empirical loss for the i-th player for parameter \( \theta \). For the purpose of the proof we will often write \( L(\theta) \) instead of \( L^i(D; \theta) \). Let \( F(\theta) = L(\theta) + \lambda\|\theta\|_{1,2} \). For any \( \theta \in \Theta^i \), let \( \Delta_S = \theta_S - \theta^*_S \) denote the difference between \( \theta \) and the true parameter \( \theta^* \) on the true support set \( S \). We introduce the following shifted and reparameterized regularized loss function:

\[
\tilde{F}(\Delta_S) = L(\theta^*_S + \Delta_S) - L(\theta^*_S) + \lambda(\|\theta^*_S + \Delta_S\|_{1,2} - \|\theta^*_S\|_{1,2}).
\] (26)
which takes the value 0 at the true parameter $\theta^*$, i.e., $\tilde{F}(0) = 0$. Let $\hat{\Delta}_S = \hat{\theta}_S^i - \theta_S^i$, where $\hat{\theta}^i$ minimizes $F(\theta)$. Since $\hat{\theta}^i$ minimizes $F(\theta)$, we must have that $\tilde{F}(\hat{\Delta}_S) \leq 0$. Thus, in order to upper bound $\|\Delta_S\|_{1,2} = \|\theta_S - \theta_S^i\|_{1,2} \leq b$, we show that there exists an $\ell_{1,2}$ ball of radius $b$ such that function $\tilde{F}(\Delta_S)$ is strictly positive on the surface of the ball.

To see this, assume the contrary, i.e., $\forall \Delta \in \Theta^i \land \|\Delta_S\|_{1,2} = b$, $\tilde{F}(\Delta_S) > 0$, but $\Delta_S$ lies outside the ball, i.e., $\|\Delta_S\|_{1,2} > b$. Then, there exists a $t \in (0,1)$ such that $(1 - t)0 + t\hat{\Delta}_S$ lies on the surface of the ball, i.e., $\|(1 - t)0 + t\hat{\Delta}_S\|_{1,2} = b$. However, by convexity of $\tilde{F}$ we have that

$$0 < \tilde{F}((1 - t)0 + t\hat{\Delta}_S) \leq (1 - t)\tilde{F}(0) + t\tilde{F}(\hat{\Delta}_S) = t\tilde{F}(\hat{\Delta}_S),$$

which implies that $\tilde{F}(\Delta_S) > 0$ and therefore is a contradiction to the fact that $\tilde{F}(\Delta_S) \leq 0$. Going forward, our strategy would be to lower bound $\tilde{F}(\Delta_S)$ in terms of $\|\Delta_S\|_{1,2} = b$. We then set the lower bound to 0 and solve for $b$, to obtain the radius of the $\ell_{1,2}$ ball on which the function is non-negative. Towards that end we first lower bound the first term of (26).

Using the Taylor’s theorem and the Cauchy-Schwartz inequality, for some $t \in [0,1]$, we have:

$$L(\theta_S^i + \Delta_S) - L(\theta_S^i) = \nabla L(\theta_S^i)^T \Delta_S + \Delta_S^T \nabla^2 L(\theta_S^i + t\Delta_S) \Delta_S,$$

$$\geq -\|\nabla L(\theta_S^i)\|_{\infty,2}\|\Delta_S\|_{1,2} + \|\Delta_S\|^2 \lambda_{\min}(H(D; \theta_S^i + t\Delta_S))$$

$$\geq -\frac{b\lambda}{2} + \frac{\|\Delta_S\|^2_{1,2}}{d_i + 1} \lambda_{\min}(H(D; \theta_S^i + t\Delta_S))$$

$$\geq -\frac{b\lambda}{2} + \frac{b^2}{2(d_i + 1)} C_{\min} - \frac{m^2b(d_i + 1)}{4}$$

$$\geq -\frac{b\lambda}{2} + \frac{b^2C_{\min}}{4(d_i + 1)},$$

(27)

where the third follows from our assumption that $\|\nabla L(\theta^i)\|_{\infty,2} \leq \lambda/2$ and the fact for any vector $x$, $\|x\|_2 \geq (1/\sqrt{d_i})\|x\|_{1,2}$ where the $\ell_{1,2}$ norm is evaluated over $g$ groups. The fourth line follows from Lemma 4 with $t = 1$ and Lemma 2. Finally, in the last line we assumed that $b \leq 2C_{\min}/(m^2(d_i + 1))$ — an assumption that we will verify momentarily. The second term of (26) is easily lower bounded using the reverse triangle inequality as follows:

$$\lambda(\|\theta_S^i + \Delta_S\|_{1,2} - \|\theta_S^i\|_{1,2}) \geq -\lambda\|\Delta_S\|_{1,2} = -b\lambda$$

(28)

Putting together (26), (27) and (28) we get:

$$\tilde{F}(\Delta_S) \geq -\frac{b\lambda}{2} + \frac{b^2C_{\min}}{4(d_i + 1)} - b\lambda.$$

Setting the above to zero and solving for $b$ we get:

$$b = \frac{6\lambda(d_i + 1)}{C_{\min}}.$$
Finally, coming back to our assumption that $b \leq 2C_{\text{min}}/(m^2(d_i + 1))$, it is easy to show that the assumption holds if the regularization parameter $\lambda$ satisfies:

$$
\lambda \leq \frac{C_{\text{min}}^2}{3m^2(d_i + 1)^2}.
$$

The lower bound on the number of samples is obtained by ensuring that the lower bound on $\lambda$ is less than the upper bound. The final claim follows from using the high probability bound on $\|\nabla L(\theta^i)\|_{\infty,2}$ from Lemma 3.

Proof of Lemma 6 (Error of the i-th parameter estimator).

$\Delta \overset{\text{def}}{=} \hat{\theta}_i - \theta^i$. We will denote the true support of $\theta^i$ by $S$, and the complement of $S$ by $S^c$. We will also simply write $L(\theta)$ instead of $L(\mathcal{D}; \theta)$. For any vector $y$, let $y_S$ denote the vector $y$ with elements not in the support set $S$ zeroed out, i.e.,

$$(y_S)_j = \begin{cases} y_j & j \in S, \\ 0 & \text{otherwise} \end{cases}$$

Then by definition of $S$, we have that $\|\theta_S^i\|_{1,2} = \|\theta^i\|_{1,2}$.

$$
\|\theta^i\|_{1,2} = \|\theta^i + \Delta\|_{1,2} = \|\theta_S^i + \Delta_S + \Delta_{S^c}\|_{1,2}
= \|\theta_S^i + \Delta_S\|_{1,2} + \|\Delta_{S^c}\|_{1,2}
\geq \|\theta_S^i\|_{1,2} - \|\Delta_S\|_{1,2} + \|\Delta_{S^c}\|_{1,2},
$$

where in the second line follows from the fact that the index sets $S$ and $S^c$ have non-overlapping groups, and in the last line we used the reverse triangle inequality. Rearranging the terms of the previous equation, and from the fact that $\|\theta_S^i\|_{1,2} = \|\theta^i\|_{1,2}$, we get:

$$
\|\theta^i\|_{1,2} - \|\hat{\theta}^i\|_{1,2} \leq \|\Delta_S\|_{1,2} - \|\Delta_{S^c}\|_{1,2} \tag{29}
$$

Next, by optimality of $\hat{\theta}^i$ we have that $L(\theta^i) + \lambda\|\theta^i\|_{1,2} \geq L(\hat{\theta}^i) + \lambda\|\hat{\theta}^i\|_{1,2}$. Rearranging the terms and continuing, we get

$$
\lambda(\|\theta^i\|_{1,2} - \|\hat{\theta}^i\|_{1,2}) \geq L(\hat{\theta}^i) - L(\theta^i)
\geq (\nabla L(\hat{\theta}^i))^T(\hat{\theta}^i - \theta^i)
\geq -\|\nabla L(\hat{\theta}^i)\|_{\infty,2}\|\Delta\|_{1,2}
\geq -\frac{\lambda}{2}\|\Delta\|_{1,2}, \tag{30}
$$

where the third line follows from the convexity of $L(\cdot)$, the fourth line follows from the Cauchy-Schwartz inequality and the last line follows from our assumption that $\lambda \geq 2\|\nabla L(\theta^i)\|_{\infty,2}$. Thus, from (29) and (30) we have that

$$
\frac{1}{2}\|\Delta\|_{1,2} \geq \|\Delta_{S^c}\|_{1,2} - \|\Delta_S\|_{1,2}
\implies \frac{1}{2}\|\Delta_S\|_{1,2} + \frac{1}{2}\|\Delta_{S^c}\|_{1,2} \geq \|\Delta_{S^c}\|_{1,2} - \|\Delta_S\|_{1,2}
\implies 3\|\Delta_S\|_{1,2} \geq \|\Delta_{S^c}\|_{1,2}.
$$
Finally, from the above inequality, we have $\|\Delta\|_{1,2} = \|\Delta_S\|_{1,2} + \|\Delta_{\bar{S}}\|_{1,2} \leq 4\|\Delta_S\|_{1,2}$. The final result follows from the upper bound on $\|\Delta_S\|_{1,2}$ derived in Lemma 5.

**Lemma 7** (Maximum eigenvalue of block positive semi-definite matrix). Let $X \in \mathbb{R}^{n \times n} \succeq 0$ be any positive semi-definite matrix, with $X_{i,i}$ being the $i$-th diagonal block of $X$. Then

$$\lambda_{\max}(X) \leq \sum_i \lambda_{\max}(X_{i,i}).$$

**Proof.** We will prove the result by decomposing $X$ into two blocks as follows:

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \end{bmatrix},$$

where $X_{1,1} \in \mathbb{R}^{n_1 \times n_1}$, $X_{2,2} \in \mathbb{R}^{n_2 \times n_2}$ and $n_1 + n_2 = 1$. The general result for multiple diagonal blocks is obtained by recursively decomposing the blocks $X_{1,1}$ and $X_{2,2}$. Any unit vector $x$ can be written as $x = c_1 x_1(x) + c_2 x_2(x)$, with $x_1(x) = (x_1/\|x_1(x)\|_2, \ldots, x_{n_1}/\|x_1(x)\|_2, 0)$, $x_2(x) = (0, x_{n_2}/\|x_2(x)\|_2, \ldots, x_{n_2}/\|x_2(x)\|_2)$, and $c_1(x) = \|x_1(x)\|_2$ (similarly $c_2(x)$). For notational simplicity we will drop the $(x)$s. Note that $c_1^2 + c_2^2 = 1$, thus $c = (c_1, c_2)$ is also a unit vector. Further, for any unit vector $x$, we have $x^T X x = c^T Y c$, where

$$Y = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{bmatrix} \in \mathbb{R}^{2 \times 2}.$$ 

Note that $x_1^T x_1 \leq \lambda_{\max}(X_{1,1})$ and $x_2^T x_2 \leq \lambda_{\max}(X_{2,2})$ for all $x$. Thus, using the variational characterization of the maximum eigenvalue of $X$ we get:

$$\lambda_{\max}(X) = \max_{\|x\|_2 = 1} x^T X x = \max_{\|c\|_2 = 1} c^T Y c \leq \max_{\|c\|_2 = 1} c^T Y c = \lambda_{\max}(Y) \leq \text{Tr}(Y) \quad (\text{since } Y \text{ is positive semi-definite})$$

$$\leq \lambda_{\max}(X_{1,1}) + \lambda_{\max}(X_{2,2}),$$

where the third line follows from the fact that the maximization is over a superset of the set $\{c = (\|x_1(x)\|_2, \|x_2(x)\|_2) : \|x\|_2 = 1\}$. \qed

**Appendix B Experiments**

In order to validate our theoretical results, we performed various synthetic experiments which are described next. We generated random potential games $\mathcal{G}$ by first generating random graphs over $p$ players and degree $d$ and number of pure strategies $m = 2$ per player. For each edge in the graph, we sampled random $m \times m$ payoff matrices by randomly setting the entries to $\pm 1$ with equal probability. We then generated a data set $\mathcal{D}$ from the game using the local noise model (5), with the noise parameter $q_i = 0.6$. 

20
Figure 1: Probability of exact recovery of the PSNE set computed across 40 randomly sampled potential games with the number of samples set to $n = 10^c(d + 1)^2 \log (2p(d+1)/\delta)$, where $c$ is the control parameter shown in the x-axis, and $\delta = 0.01$. We observe that the scaling of the sample complexity prescribed by Theorem 1 indeed holds in practice. The results show a phase transition behavior, where if the number of samples is less than $c(d + 1)^2 \log (p(d+1)/\delta)$, for some constant $c$, then PSNE recovery fails with high probability, while if the number of samples is at least $C(d + 1)^2 \log (p(d+1)/\delta)$, for some constant $C$, then PSNE recovery succeeds with high probability.