Supplementary Material

The supplementary material is organized as follows. In Section A we recall all the notation needed for the proofs. In Section B we restate the main statistical results from the paper and in Section C we provide proofs of these results. In Section D we state the optimization results and give proofs in Section E.

A Notation and Preliminary Remarks

- $L_X(a_m)$: negative log likelihood of complete data X given mth row a_m.
- $L_X^{(q)}(a_m)$: degree-q Taylor series approximation of $L_X(a_m)$.
- $L_Z(a_m)$: negative log likelihood of missing data Z given mth row a_m. Loss function is unbiased in the sense that $\mathbb{E}[L_Z(a_m)|X] = L_X(a_m)$.
- $L_Z^{(q)}(a_m)$ degree-q Taylor series approximation of $L_Z(a_m)$
- $R^{(q)}(a_m) = L_X(a_m) - L_Z^{(q)}(a_m)$
- $\mathbb{B}_{1,\infty}(1) = \{A \in \mathbb{R}^{M\times M} : \|a_m\|_1 \leq 1 \text{ for all } m\}$
- p: fraction of data which is observed
- ρ: $\max_m \|a_m^*\|_0$

Finally, we introduce additional notation which will be helpful in the proofs of Lemmas B.4 and B.5. First let $\mathcal{U}_d$ denote the set of all monomials of degree d. We represent an element $U \in \mathcal{U}_d$ as a list containing d elements. An element in the list corresponds to the index of a term in the monomial (the list can potentially have repeated elements). For an example, the monomial $x_1^2x_3$ can be represented as the list $(1, 1, 3)$.

For a polynomial function h we let $c_{U,h}$ denote the coefficient of the monomial U in h. Finally we define the order of a list to denote the number of unique elements in the list, so $|(1, 2)| = 2$ whereas $|(1, 1)| = 1$.

Example *Consider the function $h(x_1, x_2) = x_1^2 + 4x_1x_2$. Using all the notation above, we can decompose h as*

$$h(x_1, x_2) = \sum_{U\in\mathcal{U}_2} c_{U,h} \prod_{u\in U} x_u$$

where $\mathcal{U}_2 = \{(1, 1), (1, 2), (2, 2)\}$ with corresponding coefficients $c_{(1,1),h} = 1$, $c_{(1,2),h} = 4$ and $c_{(2,2),h} = 0$.

Remark 1. We next make several observations by applying the notation above to functions which appear in the likelihoods L_X and L_Z. First, for a fixed t, m we decompose the following function as a sum of monomials:

$$\begin{aligned} h(a_{m,1}X_{t,1}, \ldots, a_{m,M}X_{t,M}) :=& (a_{m,1}X_{t,1} + \ldots + a_{t,M}X_{t,M})^d \\ =& \sum_{U\in\mathcal{U}_d} c_{U,h} \left(\prod_{u\in U} a_{m,u}\right)\left(\prod_{u\in U} X_{t,u}\right) \end{aligned}$$

and note that

$$\sum_{U\in U_d} c_{U,h} \prod_{u\in U} a_{m,u} = (a_{m,1} + \ldots + a_{m,M})^d \tag{A.1}$$

We also have

$$\begin{aligned} \nabla_{a_{m,j}} h =& d \cdot X_{t,j}(a_{m,1}X_{t,1} + \ldots + a_{m,M}X_{t,M})^{d-1} \\ =& \sum_{U\in\mathcal{U}_{d-1}} c_{U,\nabla_{a_{m,j}}h} X_{t,j} \left(\prod_{u\in U} a_{m,u}\right)\left(\prod_{u\in U} X_{t,u}\right) \end{aligned}$$

and we similarly note that

$$\sum_{U\in\mathcal{U}_{d-1}} c_{U,\nabla_{a_{m,j}}h}\prod_{u\in U} a_{m,u} = d\cdot(a_{m,1}+\ldots+a_{m,M})^{d-1} \tag{A.2}$$

Next consider the function g which appears in the missing data likelihood L_Z.

$$g(a_{m,1}Z_{t,1},\ldots,a_{m,M}Z_{t,M}) := \sum_{U\in\mathcal{U}_d} c_{U,g}\left(\prod_{u\in U} a_{m,u}\right)\left(\prod_{u\in U} Z_{t,u}\right)$$

where $c_{U,g} = \frac{c_{U,h}}{p^{|U|}}$. This observation will be important for our analysis because it allows us to leverage Equation A.1. Similarly we have

$$\nabla_{a_{m,j}} g = \sum_{U\in\mathcal{U}_{d-1}} c_{U,\nabla g_{a_{m,j}}} Z_{t,j}\left(\prod_{u\in U} a_{m,u}\right)\left(\prod_{u\in U} Z_{t,u}\right)$$

where $c_{U,\nabla_{a_{m,j}}g} = \frac{c_{U,\nabla_{a_{m,j}}h}}{p^{|U|}}$ allowing us to use Equation A.2.

Remark 2. Using an identical argument to Lemma B.4 one can show that for $a_m \in \mathbb{B}_1(1)$, $|L_{Z,p}(a_m) - L^{(q)}_{Z,p}(a_m)| \leq (p\pi)^{-q}$. This implies that $\lim_{q\to\infty} L^{(q)}_{Z,p}$ converges uniformly on $\mathbb{B}_1(1)$ so that $L_{Z,p}(a_m)$ is well defined on this ball. Moreover, it implies that $\lim_{q\to\infty}\mathbb{E}\left[|L^{(q)}_{Z,p}(a_m)|\middle|X\right]$ converges and so

$$\mathbb{E}[L_{Z,p}(a_m)|X] = \lim_{q\to\infty}\mathbb{E}[L^{(q)}_{Z,p}(a_m)|X] = \lim_{q\to\infty} L^{(q)}_X(a_m) = L_X(a_m)$$

and $L_{Z,p}(a_m)$ satisfies (3.1)

B Statistical Results

We assume $A^* \in \mathbb{B}_{1,\infty}(1)$ and $p \geq \frac{1}{\pi}$. We take $q \in \mathbb{N}\cup\{\infty\}$.

Theorem B.1 (Accuracy of $L^{(q)}_Z$)**.** *Suppose* $\widehat{A} \in \arg\min_{A\in\mathbb{B}_{1,\infty}(1)} L^{(q)}_Z(A) + \lambda\|A\|_1$ *where* $\lambda \asymp \frac{\log(MT)}{\sqrt{T}(p\pi-1)} + \frac{1}{(p\pi)^q}$. *Then*

$$\|\widehat{A} - A^*\|_F^2 \lesssim \frac{s\log^2(MT)}{T(\pi p-1)^2} + \frac{s}{(p\pi)^{2n}}$$

for $T \gtrsim \rho^2\log(MT)$ *with probability at least* $1-\frac{1}{T}$.

The proof of Theorem B.1 relies on the following supplementary lemmas.

Lemma B.2. *Let* $f(x) = \log(1+\exp(x))$. *Then* $|\frac{f^{(q)}(0)}{q!}| \lesssim \frac{1}{q\pi^q}$.

Lemma B.3 (Truncation error of ∇L_X)**.** *Suppose* $\|a_m\|_1 \leq 1$. *Then*

$$\|\nabla L_X(a_m) - \nabla L^{(q)}_X(a_m)\|_\infty \lesssim \pi^{-q}.$$

Lemma B.4 (Truncation Error of ∇L_Z)**.** *Suppose* $\|a_m\|_1 \leq 1$. *Then*

$$\|\nabla L_Z(a_m) - \nabla L^{(q)}_Z(a_m)\|_\infty \lesssim (p\pi)^{-q}.$$

Lemma B.5.

$$\sup_{\|a_m\|_1\leq 1}\left\|\nabla L^{\log(T)}_Z(a_m) - \nabla L^{\log(T)}_X(a_m)\right\|_\infty \lesssim \frac{\log(MT)}{\sqrt{T}(p\pi-1)}$$

with probability at least $1-\frac{\log(T)^2}{MT^2}$.

Lemma B.6.

$$\sup_{\|a_m\|_1 \le 1} \left\| \nabla L_Z^{(q)}(a_m) - \nabla L_X(a_m) \right\|_\infty \lesssim \frac{\log(MT)}{\sqrt{T}(p\pi - 1)} + \frac{1}{(p\pi)^q}$$

with probability at least $1 - \frac{\log(T)^2}{MT^2}$.

Lemma B.7. *Let* $\|v\|_T^2 = \frac{1}{T}\sum_t (v^\top X_t)^2$ *and* $\tilde{R} = \min(R_{\min}, 1 - R_{\max})$. *For any* $v \in \mathbb{R}^M$ *we have*

$$\|v\|_T^2 \ge \frac{\tilde{R}}{2}\|v\|_2^2 - \sqrt{\frac{3\log(M)}{T}}\|v\|_1^2$$

and

$$\|v\|_T^2 \le \frac{1}{4}\|v\|_2^2 + \sqrt{\frac{3\log(M)}{T}}\|v\|_1^2$$

with probability at least $1 - \frac{1}{M}$.

C Proofs of Statistical Results

C.1 Proof of Theorem B.1

Part 1: Controlling the Remainder We set $\lambda \asymp \frac{\log(MT)}{\sqrt{T}(p\pi - 1)} + \frac{1}{(p\pi)^q}$ and let $\triangle_m = \widehat{a}_m - a_m^*$. Note that the loss functions are decomposable, i.e., $L_Z(A) = \sum_m L_Z(a_m)$. Since $\widehat{A} \in \arg\min_{A \in \mathbb{B}_{1,\infty}(1)} L_Z^{(q)}(A)$ we have

$$L_Z^{(q)}(\widehat{a}_m) \le L_Z^{(q)}(a_m^*)$$

and so

$$L_X(\widehat{a}_m) \le L_X(a_m^*) + \left(R^{(q)}(\widehat{a}_m) - R^{(q)}(a_m^*) \right).$$

Define $\triangle_m := \widehat{a}_m - a_m^*$. By the mean value theorem, there exists some $v \in \mathbb{B}_1(1)$ such that

$$R^{(q)}(\widehat{a}_m) - R^{(q)}(a_m^*) = \langle \nabla R^{(q)}(v), \triangle_m \rangle$$

so by Lemma B.6

$$|R^{(q)}(\widehat{a}_m) - R^{(q)}(a_m^*)| \le \|\triangle_m\|_1 \|\nabla R^{(q)}(v)\|_\infty \le \frac{\lambda}{4}\|\triangle_m\|_1$$

with probability $1 - \frac{\log(T)^2}{MT^2}$. We condition on this event for the remainder of the proof.

Part 2: Setting Up the Standard Equations The next several steps follow standard techniques for ℓ_1 regularization in GLMs. Expanding L_X and using the substitution

$$X_{t+1,m} = \mathbb{E}[X_{t+1,m}|X_t] + \epsilon_{t,m} = f'(a_m^{*\top} X_t) + \epsilon_{t,m}$$

gives

$$\begin{aligned}\frac{1}{T}\sum_t f(\widehat{a}_m^\top X_t) - (\widehat{a}_m^\top X_t)(f'(a_m^{*\top} X_t) + \epsilon_{t,m}) + \lambda\|\widehat{a}_m\|_1 \\ \le \frac{1}{T}\sum_t f(a_m^{*\top} X_t) - (a_m^{*\top} X_t)(f'(a_m^{*\top} X_t) + \epsilon_{t,m}) + \lambda\|a_m^*\|_1 + \frac{\lambda}{4}\|\triangle_m\|_1\end{aligned}$$

Rearranging terms yields

$$\begin{aligned}\frac{1}{T}\sum_t f(\widehat{a}_m^\top X_t) - f(a_m^{*\top} X_t) - f'(a_m^{*\top} X_t)\triangle_m^\top X_t \\ \le |\frac{1}{T}\sum_t \epsilon_{t,m}\triangle_m^\top X_t| + \lambda\left(\|a_m^*\|_1 - \|\widehat{a}_m\|_1\right) + \frac{\lambda}{4}\|\triangle_m\|_1\end{aligned} \tag{C.1}$$

Since $f(x) = \log(1 + \exp(x))$ is σ-strongly convex on $[R_{\min}, R_{\max}]$ with $\sigma = \frac{R_{min}}{(1+R_{\min})^2}$ we can lower bound Equation (C.1) by $\frac{\sigma}{T}\sum_t(\triangle_m^\top X_t)^2$. Also note

$$\left|\frac{1}{T}\sum_t \epsilon_{t,m}\triangle_m^\top X_t\right| \leq \|\triangle_m\|_1 \left\|\frac{1}{T}\sum_t \epsilon_{t,m}X_t\right\|_\infty.$$

Using Theorem 2.2 in Hall et al. (2016), $\left\|\frac{1}{T}\sum_t \epsilon_{t,m}X_t\right\|_\infty \leq \frac{\lambda}{4}$ with probability at least $1 - \frac{1}{MT}$. Applying these observations to Equation (C.1) gives

$$\frac{1}{T}\sum_{t=1}^T(\triangle_m^\top X_t)^2 \leq \frac{\lambda}{2}\|\triangle_m\|_1 + \lambda\|a_m^*\|_1 - \lambda\|\widehat{a}_m\|_1.$$

Define $S := \{i : a^*_{m,i} \neq 0\}$ and $\rho_m := \|a_m^*\|_0$. Then

$$\frac{1}{T}\sum_{t=1}^T(\triangle_m^\top X_t)^2 \leq \frac{3\lambda}{2}\|\triangle_{m,S}\|_1 - \frac{\lambda}{2}\|\triangle_{m,S^c}\|_1 \tag{C.2}$$

and so

$$\frac{1}{T}\sum_{t=1}^T(\triangle_m^\top X_t)^2 \leq \frac{3\lambda}{2}\|\triangle_{m,S}\|_1 \leq \frac{3\lambda\sqrt{\rho_m}}{2}\|\triangle_{m,S}\|_2 \leq \frac{3\lambda\sqrt{\rho_m}}{2}\|\triangle_m\|_2. \tag{C.3}$$

Restriction of $\triangle_m$ to Cone It remains to lower bound $\frac{1}{T}\sum_t(\triangle_m^\top X_t)^2$ in terms of $\|\triangle_m\|_2^2$. In order to do this we will rely heavily on the fact that $\triangle_m$ is not an arbitrary vector. Instead we show that $\triangle_m$ must lie in a cone with important properties. In particular, returning to Equation (C.2) and observing that $0 \leq \frac{1}{T}\sum_t(\triangle_m^\top X_t)^2$ it follows that $\|\triangle_{m,S^c}\|_1 \leq 3\|\triangle_{m,S}\|_1$. Thus

$$\|\triangle_m\|_1 = \|\triangle_{m,S}\|_1 + \|\triangle_{m,S^c}\|_1 \leq 4\|\triangle_{m,S}\|_1. \tag{C.4}$$

Restricted Eigenvalue Condition As mentioned in the previous section, our goal is to lower bound $\frac{1}{T}\sum_{t=1}^T(\triangle_m^\top X_t)^2$. This is commonly referred to as a restricted eigenvalue condition in the literature, and it is closely related to the restricted strong convexity condition proved in Lemma D.2. In particular, Lemma B.7 guarantees that there exist universal constants c_1 and c_2 such that

$$\frac{1}{T}\sum_t(\triangle_m^\top X_t)^2 \geq \frac{c_1}{2}\|\triangle_m\|_2^2 - c_2\sqrt{\frac{\log(MT)}{T}}\|\triangle_m\|_1^2.$$

For arbitrary vectors this lower bound can be negative; however, by Equation (C.4)

$$\|\triangle_m\|_1^2 \leq 16\|\triangle_{m,S}\|_1^2 \leq 16\rho_m\|\triangle_{m,S}\|_2^2 \leq 16\rho_m\|\triangle_m\|_2^2$$

and thus

$$\frac{1}{T}\sum_t(\triangle_m^\top X_t)^2 \geq \frac{c_1}{2}\|\triangle_m\|_2^2 - 16c_2\rho_m\sqrt{\frac{\log(MT)}{T}}\|\triangle_m\|_2^2.$$

Hence if $T \gtrsim \rho_m^2\log(MT)$ it follows that

$$\frac{1}{T}\sum_t(\triangle_m^\top X_t)^2 \geq c\|\triangle_m\|_2^2$$

for a universal constant c. Plugging this in to Equation (C.3) gives that

$$\|\triangle_m\|_2 \lesssim \sqrt{\rho_m}\lambda$$

with probability $1 - \frac{1}{MT}$. Taking a union bound we conclude that

$$\|A^* - \widehat{A}\|_F^2 = \sum_{m=1}^M \|\triangle_m\|_2^2 \lesssim \sum_{m=1}^M \rho_m\lambda^2 = s\lambda^2$$

with probability $1 - \frac{1}{T}$. □

C.2 Proof of Lemma B.2

A computation shows that

$$f^{(q)}(0) = \frac{1}{2^q} \sum_{m=0}^{q-2} (-1)^m A(q-1, m)$$

where the $A(q, m)$ are the Eulerian numbers. The alternating sum of the Eulerian numbers for fixed q can be given as

$$\sum_{m=0}^{q-2} (-1)^q A(q-1, m) = \frac{2^q(2^q - 1)B_q}{q}$$

where B_q is the qth Bernoulli number (see the derivation of Equation 4.8 in Carlitz (1959)). Thus

$$\left| \frac{f^{(q)}(0)}{q!} \right| \leq \frac{2^q |B_q|}{q! q}.$$

Using Alzer (2000) we have the bound $|B_q| \lesssim \frac{q!}{(2\pi)^q}$. Thus

$$\left| \frac{f^{(q)}(0)}{q!} \right| \lesssim \frac{2^q q!}{q! q 2^q \pi^q} = \frac{1}{q\pi^q}.$$

□

C.3 Proof of Lemma B.3

For any j we have

$$\begin{aligned}
\left| (\nabla L_X(a_m) - \nabla L_X^{(q)}(a_m))_j \right| &= \left| \frac{1}{T} \sum_t \sum_{i=q}^{\infty} \frac{f^{(i)}(0)}{(i-1)!} (a_m^T X_t)^{i-1} X_{t,j} \right| \\
&\leq \frac{1}{T} \sum_t \sum_{i=q}^{\infty} \pi^{-i} \\
&\lesssim \frac{1}{T} \sum_t \pi^{-q} = \pi^{-q},
\end{aligned}$$

where the last two lines use Lemma B.2 along with the fact that $\|a_m\|_1 \leq 1$. □

C.4 Proof of Lemma B.4

Differentiating $L_Z(a_m)$ with respect to $a_{m,j}$ gives

$$\nabla_j L_Z(a_m) = \sum_{d=1}^{\infty} \left(\frac{f^{(d)}(0)}{d!T} \sum_t \sum_{U \in \mathcal{U}_{d-1}} \frac{c_{U, \nabla_j f}}{p^{|U|}} \left(\prod_{u \in U} a_{m,u} \right) \left(\prod_{u \in U} Z_{t,u} \right) \right)$$

We first bound

$$g_d(a_m) := \frac{f^{(d)}(0)}{d!T} \sum_t \sum_{U \in \mathcal{U}_{d-1}} \frac{c_{U, \nabla_j f}}{p^{|U|}} \left(\prod_{u \in U} a_{m,u} \right) \left(\prod_{u \in U} Z_{t,u} \right)$$

which is the degree $d-1$ term of $\nabla_j L_Z(a_m)$. All the terms other than the $a_{m,u}$ in $g_d(a_m)$ are always non-negative and $|U| \leq d$, so

$$|g_d(a_m)| \leq \frac{f^{(d)}(0)}{d! p^d} \frac{1}{T} \sum_t \sum_{U \in \mathcal{U}_{d-1}} c_{U, \nabla_j f} \prod_{u \in U} |a_{m,u}|.$$

By Equation A.2

$$\sum_{U \in \mathcal{U}_{d-1}} c_{U, \nabla_j f} \prod_{u \in U} |a_{m,u}| = d(|a_{m,1}| + \ldots + |a_{m,M}|)^{d-1} \leq d.$$

We conclude that

$$|g_d(a_m)| \leq \frac{1}{T} \sum_t \frac{f^{(d)}(0)}{(d-1)!p^d} = \frac{f^{(d)}(0)}{(d-1)!p^d}.$$

Using Lemma B.2

$$\frac{f^{(d)}(0)}{(d-1)!p^d} \lesssim \frac{1}{(p\pi)^d}.$$

Overall, we have concluded that $|g_d(a_m)| \leq \frac{1}{(p\pi)^d}$. We are ultimately interested in $|\nabla_j L_Z(a_m) - \nabla_j L_Z^{(q)}(a_m)|$. This is the sum of the degree d terms of $\nabla_j L_Z(a_m)$ for all $d \geq q$. In other words,

$$|\nabla_j L_Z(a_m) - \nabla_j L_Z^{(q)}(a_m)| \leq \sum_{d=q}^{\infty} |g_d(a_m)| \leq \sum_{d=q}^{\infty} \frac{1}{(p\pi)^d} \lesssim \frac{1}{(p\pi)^q}$$

as claimed. □

C.5 Proof of Lemma B.5

We begin by bounding individual monomials of $\left\| \nabla L_Z^{\log(T)}(a_m) - \nabla L_X^{\log(T)}(a_m) \right\|_\infty$. We then extend these individual bounds to bounds on the entire expression.

Bounding Individual Monomials Following the notation introduced in Section A a degree $d-1$ monomial of $\nabla_j L_X^{\log(T)}(a_m)$ with index U is of the form

$$m_{U,X} := \frac{f^d(0)}{d!T} \sum_t c_{U,\nabla_j f} \left(\prod_{u \in U} a_{m,u} \right) \left(\prod_{u \in U} X_{t,u} \right).$$

Meanwhile, the degree $d-1$ monomial of $\nabla_j L_Z^{\log(T)}(a_m)$ with index U is of the form

$$m_{U,Z} := \frac{f^d(0)}{d!Tp^{|U|}} \sum_t c_{U,\nabla_j f} \left(\prod_{u \in U} a_{m,u} \right) \left(\prod_{u \in U} Z_{t,u} \right).$$

The difference of these monomials is given by

$$\begin{aligned}
&|m_{U,X} - m_{U,Z}| \\
=&\left| \frac{f^d(0)}{Td!} \Big(\sum_t c_{U,\nabla_j f} \prod_{u \in U} a_{m,u} \prod_{u \in U} X_{t,u} - \frac{1}{p^{|U|}} \sum_t c_{U,\nabla_j f} \prod_{u \in U} a_{m,u} \prod_{u \in U} Z_{t,u} \Big) \right| \\
=&\frac{f^d(0) c_{U,\nabla_j h} \prod_{u \in U} |a_{m,u}|}{Td!} \left| \left(\sum_t \prod_{u \in U} X_{t,u} - \frac{1}{p^{|U|}} \sum_t \prod_{u \in U} Z_{t,u} \right) \right|.
\end{aligned}$$

Observe that

$$\frac{1}{p^{|U|}} \prod_{u \in U} Z_{t,u} \in \{0, \frac{1}{p^{|U|}}\}$$

and

$$\mathbb{E}\left[\frac{1}{p^{|U|}} \prod_{u \in U} Z_{t,u} \right] = \prod_{u \in U} X_{t,u}.$$

We apply Hoeffding's inequality to conclude that

$$\mathbb{P}\left(\left| \sum_t \frac{1}{p^{|U|}} \prod_{u \in U} Z_{t,u} - \sum_t \prod_{u \in U} X_{t,u} \right| \geq \frac{\log(MT)\sqrt{T}}{p^{|U|}} \right) \leq 2\exp(-2\log^2(MT)). \tag{C.5}$$

Thus

$$|m_{U,X} - m_{U,Z}| \leq \frac{f^d(0) \log(MT) c_{U,\nabla_j h} \prod_{u \in U} |a_{m,u}|}{d!\sqrt{T}p^{|U|}} \tag{C.6}$$

with probability at least $1 - 2\exp(-2\log^2(MT))$.

Extension to Entire Expression We need to take a union bound so that this holds for all monomials of degree at most $\log(T)$. However, since $Z_{t,u}$ and $X_{t,u}$ are binary random variables

$$\sum_t \prod_{u\in U} Z_{t,u} - \sum_t \prod_{u\in U} X_{t,u} = \sum_t Z_{t,v} \prod_{u\in U} Z_{t,u} - \sum_t X_{t,v} \prod_{u\in U} X_{t,u}$$

whenever $v \in U$.

Suppose we have shown that Equation (C.5) holds for all monomials of degree $< d$. Now to show it holds for all monomials of degree d we only need to show that (C.5) holds for all monomials of degree d that have d distinct terms. The remaining concentrations are already covered by the degree $d-1$ monomials. There are $\binom{M}{d} \leq M^d$ monomials of degree d with d distinct terms. Hence we need to take a union bound over at most

$$\sum_{d=1}^{\log(T)} M^d \leq \log(T) M^{\log(T)}$$

monomials of degree $\leq \log(T)$ and so

$$\begin{aligned}\mathbb{P}\Bigg(\Big|\sum_t \frac{1}{p^{|U|}} \prod_{u\in U} Z_{t,u} - \sum_t \prod_{u\in U} X_{t,u}\Big| \geq \frac{\log(MT)\sqrt{T}}{p^{|U|}} \text{ for } \geq 1 \text{ monomial of degree} \leq \log(T)\Bigg) \\ \leq 2\log(T)\exp\left(\log(M)\log(T) - 2\log^2(MT)\right) \leq \frac{1}{M^2T^2}.\end{aligned}$$

We condition on this event and recall that $\mathcal{U}_{d-1}$ denotes the set of all monomials of degree $d-1$. Using Equation (C.6), the difference between the degree $d-1$ terms of $\nabla_j L_X^{\log(T)}(a_m)$ and $\nabla_j L_Z^{\log(T)}(a_m)$ can be bounded by

$$\sum_{U\in\mathcal{U}_{d-1}} |m_{U,X} - m_{U,Z}| \leq \frac{f^d(0)\log(MT)}{d!\sqrt{T}} \sum_{U\in\mathcal{U}_{d-1}} \frac{c_{U,\nabla_j f}}{p^{|U|}} \prod_{u\in U} |a_{m,u}|.$$

Using Lemma B.2 and Equation A.2 along with the fact that $|U| \leq d$,

$$\begin{aligned}\frac{f^d(0)\log(MT)}{d!\sqrt{T}} \sum_{U\in\mathcal{U}_{d-1}} \frac{c_{U,\nabla_j f}}{p^{|U|}} \prod_{u\in U} |a_{m,u}| &\leq \frac{f^d(0)\log(MT)}{(d-1)!p^d\sqrt{T}} \\ &\leq \frac{\log(MT)}{\sqrt{T}(p\pi)^d}\end{aligned}$$

where the final inequality uses Lemma B.2. Thus we have the bound

$$\begin{aligned}\left|\nabla_j L_X^{\log(T)}(a_m) - \nabla_j L_Z^{\log(T)}(a_m)\right| &\leq \sum_{d=1}^{\log(T)} \frac{\log(MT)}{\sqrt{T}(p\pi)^d} \\ &\lesssim \frac{\log(MT)}{\sqrt{T}(p\pi - 1)}\end{aligned}$$

with probability at least $1 - \frac{\log(T)^2}{MT^2}$. □

C.6 Proof of Lemma B.6

By the triangle inequality we have

$$\begin{aligned}\left|\left(\nabla L_Z^{(q)}(a_m) - \nabla L_X(a_m)\right)_j\right| \leq \\ &\left|\left(\nabla L_Z^{(q)}(a_m) - \nabla L_Z^{\log(T)}(a_m)\right)_j\right| \\ &+ \left|\left(\nabla L_X^{\log(T)}(a_m) - \nabla L_X(a_m)\right)_j\right| \\ &+ \left|\left(\nabla L_Z^{\log(T)}(a_m) - \nabla L_X^{\log(T)}(a_m)\right)_j\right|.\end{aligned}$$

By Lemmas B.3 and B.4 the first two terms can be bounded by $\max\left((p\pi)^{-\log(T)}, (p\pi)^{-q}\right)$ while by Lemma B.5 the final term can be bounded is

$$\lesssim \frac{\log(MT)}{\sqrt{T}(p\pi - 1)}$$

with probability at least $1 - \frac{\log(T)^2}{MT^2}$. We conclude that

$$\|\nabla R^{(q)}(a_m)\|_\infty \lesssim \frac{\log(MT)}{\sqrt{T}(p\pi - 1)} + \frac{1}{(p\pi)^{-q}}$$

with probability at least $1 - \frac{\log(T)^2}{MT^2}$ as claimed. □

C.7 Proof of Lemma B.7

We have

$$\|v\|_T^2 = \frac{1}{T}\sum_t v^\top \left(\mathbb{E}[(X_t X_t^\top)|X_{t-1}]\right) v - \frac{1}{T}\sum_t v^\top \left(X_t X_t^\top - \mathbb{E}[X_t X_t^\top|X_{t-1}]\right) v \tag{C.7}$$

By Theorem 2.1 in Hall et al. (2016),

$$\frac{1}{T}\sum_t v^\top \left(\mathbb{E}[(X_t X_t^\top)|X_{t-1}]\right) v \geq \frac{\tilde{R}}{2}\|v\|_2^2 \tag{C.8}$$

Now we define the matrix $G \in \mathbb{R}^{M\times M}$ as follows:

$$G := \frac{1}{T}\sum_{t\in\mathcal{T}} (X_t X_t^\top - \mathbb{E}[X_t X_t^\top|X_{t-1}]).$$

Note that each entry of G is a martingale and

$$v^\top \left(\sum_t X_t X_t^\top - \mathbb{E}[X_t X_t^\top|X_{t-1}]\right) v \leq \|v\|_1^2 \max_{m,m'} |G_{m,m'}|.$$

Applying the Azuma-Hoeffding inequality we conclude that for any m, m'

$$\mathbb{P}\left(|G_{m,m'}| \geq \sqrt{3\log(M)T}\right) \leq \frac{1}{M^3}$$

and therefore

$$\mathbb{P}\left(\max_{m,m'} |G_{m,m'}| \geq \sqrt{3\log(M)T}\right) \leq \frac{1}{M}.$$

Overall we have concluded

$$\frac{1}{T}\sum_t v^\top \left(X_t X_t^\top - \mathbb{E}[X_t X_t^\top|X_{t-1}]\right) v \leq \frac{\sqrt{3\log(M)}}{\sqrt{T}}\|v\|_1^2 \tag{C.9}$$

with probability at least $1 - \frac{1}{M}$. Combining Equations (C.7),(C.8) and (C.9) give the desired result. □

D Optimization Results

The main result of this section is the following Theorem.

Theorem D.1. *Suppose $A^* \in \mathbb{B}_{1,\infty}(1)$ and $\|a_m^*\|_0 > 0$ for at least $\frac{M}{C}$ rows of A^* where C is a universal constant. Let $\tilde{A} \in \mathbb{B}_{1,\infty}(1)$ be any stationary point of $L_Z^{(q)}(A) + \lambda\|A\|_1$ where $\lambda \asymp \frac{\log(MT)}{\sqrt{T}(p\pi-1)} + \frac{1}{(p\pi)^q}$. Then*

$$\|\tilde{A} - A^*\|_F^2 \lesssim s\left(\sqrt{\frac{\log(MT)}{T}} + \frac{1}{(p\pi)^q}\right)$$

with probability at least $1 - \frac{\log(T)}{T^2}$.

In order to prove Theorem D.1 we need to introduce notions of Restricted Strong Convexity (RSC) and Restricted Smoothness (RSM) from Agarwal et al. (2012). To do this we first define the first order Taylor expansion to a function L:

$$T_L(v, w) = L(v) - L(w) - \langle \nabla L(w), v - w\rangle.$$

Definition 3 (Restricted Strong Convexity)**.** A loss function L satisfies the RSC condition with parameters α, τ if

$$T_L(v, w) \geq \frac{\alpha}{2}\|v - w\|_2^2 - \tau\|v - w\|_1^2$$

for all $v, w \in \mathbb{B}_1(1)$.

Definition 4 (Restricted Smoothness)**.** A loss function L satisfies the RSM condition with parameters α, τ if

$$T_L(v, w) \geq \frac{\alpha}{2}\|v - w\|_2^2 + \tau\|v - w\|_1^2$$

for all $v, w \in \mathbb{B}_1(1)$.

Lemma D.2. *The RSC and RSM conditions are satisfied for $L_Z^{(q)}$ with constants $\alpha = c_1$ and $\tau = c_2\left(\sqrt{\frac{\log(MT)}{T}} + \frac{1}{(p\pi)^q}\right)$ with probability at least $1 - \frac{\log(T)}{T^2}$ where c_1 and c_2 are universal constants.*

Combining Lemma D.2 with Theorem 2 in Agarwal et al. (2012) gives the following corollary.

Corollary D.3. *Under the conditions of Theorem B.1, let a_m^s denote the s'th iteration of projected gradient descent using the loss function $\phi(a_m) = L_Z^{(q)}(a_m) - \lambda\|a_m\|_1$. There exists some S such that for all $s > S$ we have*

$$\phi(a_m^s) - \phi(\widehat{a}_m) \lesssim \tau\rho_m.$$

E Proofs of Optimization Results

E.1 Proof of Lemma D.2

We begin by computing the first order Taylor approximation:

$$\begin{aligned} &T_{L_Z^{(q)}}(v, w) \\ &\quad = L_Z^{(q)}(v) - L_Z^{(q)}(w) - \langle \nabla L_Z^{(q)}(w), v - w \rangle \\ &\quad = \underbrace{L_X(v) - L_X(w) - \langle \nabla L_X(w), v - w \rangle}_{1} + \underbrace{R^{(q)}(v) - R^{(q)}(w) - \langle \nabla R^{(q)}(w), v - w \rangle}_{2}. \end{aligned}$$

We first handle term (1), which is the first order Taylor error for L_X. A computation shows that this is equal to

$$\frac{1}{T} \sum_T f(v^\top X_t) - f(w^\top X_t) - f'(w^\top X_t) \langle v - w, X_t \rangle$$

where again $f(x) = \log(1 + \exp(x))$. Since f is σ-strongly convex on $[R_{\min}, R_{\max}]$ with $\sigma = \frac{R_{\min}}{(1+R_{\min})^2}$ we can lower bound term (1):

$$\frac{\sigma}{T} \sum_T \langle v - w, X_t \rangle^2 \leq L_X(v) - L_X(w) - \langle \nabla L_X(w), v - w \rangle.$$

Using Lemma B.7, term (1) can be bounded below by

and can be bounded above by

$$\frac{\sigma}{4} \|v - w\|_2^2 + \sigma \sqrt{\frac{3 \log(M)}{T}} \|v - w\|_1^2 \tag{E.1}$$

and below by

$$\frac{\sigma \tilde{R}}{2} \|v - w\|_2^2 - \sigma \sqrt{\frac{3 \log(M)}{T}} \|v - w\|_1^2. \tag{E.2}$$

It remains to handle term (2) which is $T_{R^{(q)}}(v, w)$. By the mean value theorem there exists some $u \in \mathbb{B}_1(1)$ such that $R^{(q)}(v) - R^{(q)}(w) = \langle \nabla R^{(q)}(u), v - w \rangle$. Thus we can bound term (2) by

$$\left(\|\nabla R^{(q)}(u)\|_\infty + \|\nabla R^{(q)}(w)\|_\infty \right) \|v - w\|_1.$$

By Lemma B.6

$$\|\nabla R^{(q)}(u)\|_\infty + \|\nabla R^{(q)}(w)\|_\infty \lesssim \frac{\log(MT)}{\sqrt{T}} + \frac{1}{(p\pi)^q}$$

with probability $1 - \frac{2 \log(T)}{MT^2}$. Combining this with our bounds on term 1 in Equations (E.1) and (E.2) gives the final result. □

E.2 Proof of Theorem D.1

By Corollary D.3 we have

$$\left(L_Z^{(q)}(a_m^s) - L_Z^{(q)}(\widehat{a}_m \right) + (\lambda \|\widehat{a}_m\|_1 - \lambda \|a_m^s\|_1) \leq \tau \rho_m.$$

Since $\widehat{a}_m$ is a stationary point it satisfies

$$\langle \nabla L_Z^{(q)}(\widehat{a}_m) - \lambda \widehat{a}_m, a_m^s - \widehat{a}_m \rangle \geq 0.$$

Using this, along with the fact that $a_m^s, \widehat{a}_m \in \mathbb{B}_1(1)$ we get that

$$T_{L_Z^{(q)}}(a_m^s, \widehat{a}_m) \leq \tau \rho_m + 3\lambda.$$

Using the RSC condition from Lemma D.2 we conclude

$$\|\widehat{a}_m - a_m^s\|_2^2 \lesssim \tau \rho_m + \tau + \lambda \lesssim (\rho_m + 1) \left(\sqrt{\frac{\log(MT)}{T}} + \frac{1}{(p\pi)^q} \right).$$

Finally we apply the statistical error bound on $\|\widehat{a}_m - a_m^*\|_2^2$ from Theorem B.1 along with the triangle inequality to conclude that

$$\|a_m^s - a_m^*\|_2^2 \lesssim (\rho_m + 1)\left(\sqrt{\frac{\log(MT)}{T}} + \frac{1}{(p\pi)^q}\right).$$

Summing over all m and assuming $\|a_m^*\|_0 > 0$ for at least $\frac{M}{C}$ values of m allows us to conclude that

$$\|A^s - A^*\|_F^2 \lesssim s\left(\sqrt{\frac{\log(MT)}{T}} + \frac{1}{(p\pi)^q}\right).$$

To get the final form of the result, we recall that A^s is the sth iteration of projected gradient descent run with an arbitrary initialization within $\mathbb{B}_{1,\infty}(1)$. In particular, if we initialize A^0 at a stationary point then $A^s = A^0$ which gives the final form of the Theorem.