A Proof of Section 2

A.1 Proof of Theorem 1

Proof. We rewrite the spectral norm error in terms of the polynomial representations (8) and (9) as

$$\begin{aligned}
&\|\widehat{\mathcal{S}_m}(x) - \mathcal{S}_m(x)\|_{\mathrm{sp}} \\
\le\ & \sum_{\lambda\in\Lambda_m} c_m(\lambda)\, \|\mathrm{sym}(\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}}) - \mathrm{sym}(\bigotimes_{j\in\lambda} \mathcal{G}_j)\|_{\mathrm{sp}} \\
\le\ & \sum_{\lambda\in\Lambda_m} c_m(\lambda)\, \|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\mathrm{sp}}\ , \qquad (23)
\end{aligned}$$

where the last inequality comes from the fact that $\|\mathrm{sym}(\mathcal{T})\|_{\mathrm{sp}} \le \|\mathcal{T}\|_{\mathrm{sp}}$. Then we study each term in (23). For simplicity of notation, denote the estimation error $\mathcal{E}_j^{(p)} \triangleq \widehat{\mathcal{A}_j^{(p)}} - \mathcal{G}_j$, then we have

$$\begin{aligned}
&\|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\mathrm{sp}} \\
=\ & \|\bigotimes_{j\in\lambda} (\mathcal{E}_j^{(p)} + \mathcal{G}_j) - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\mathrm{sp}} \\
=\ & \|\sum_{\nu\subset\lambda} \left(\left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \right) - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\mathrm{sp}} \\
=\ & \|\sum_{\nu\subset\lambda,\nu\neq\emptyset} \left(\left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \right) \|_{\mathrm{sp}} \\
\le\ & \sum_{\nu\subset\lambda,\nu\neq\emptyset} \| \left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \|_{\mathrm{sp}} \\
\le\ & \sum_{\nu\subset\lambda,\nu\neq\emptyset} \left((\prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}}) \times (\prod_{j\in\lambda\setminus\nu} \|\mathcal{G}_j\|_{\mathrm{sp}}) \right) . \qquad (24)
\end{aligned}$$

Now we study the spectral norm of $\mathcal{E}_j^{(p)}$, which can be upper bounded by the Frobenius norm. Then by Lemma 2.1, we have,

$$\begin{aligned}
\|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}} \le \|\mathcal{E}_j^{(p)}\|_{\mathcal{F}} &= \sqrt{\sum_{i_1,\ldots,i_j} \left(\mathcal{E}_j^{(p)}\right)^2_{(i_1,\ldots,i_j)}} \\
&= O(d^{j/2} h^{p+1-j}) + O_p(d^{j/2}(nh^{d+2j})^{-1/2}). \qquad (25)
\end{aligned}$$

Since for any $j \le m$, we have $h^{p+1-j} \to 0$ and $nh^{d+2j} \to \infty$ as $n \to \infty$. So for sufficiently large n, we have $\sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}} \le 1$ with high probability. Then, plug it into (24), we get

$$\begin{aligned}
&\|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\mathrm{sp}} \\
\le\ & \sum_{\nu\subset\lambda,\nu\neq\emptyset} \left((\prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}}) \times \prod_{j\in\lambda\setminus\nu} C_j \right) \\
\le\ & C \sum_{\nu\subset\lambda,\nu\neq\emptyset} \prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}} \\
=\ & C \left(\prod_{j\in\lambda} (1 + \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}}) - 1 \right) \\
\le\ & C \left(\exp\{\sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}}\} - 1 \right) \\
\le\ & 2C \sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\mathrm{sp}} \\
=\ & O(d^{j_{\max}/2} h^{p+1-j_{\max}}) \\
& + O_p(d^{j_{\max}/2}(nh^{d+2j_{\max}})^{-1/2}) \qquad (26)
\end{aligned}$$

here constant $C = \max_\nu \prod_{j\in\lambda\setminus\nu} C_j$ and $j_{\max} = \max\{j : j \in \lambda\}$. The last inequality comes from the fact that $e^y - 1 \le 2y$ for any $y \le 1$. Since λ is a partition of integer m, we have $j_{\max} \le m$, and the equation holds if and only if $\lambda = \{m\}$. Therefore the only term in (23) that achieves $O(d^{m/2}h^{p+1-m}) + O_p(d^{m/2}(nh^{d+2m})^{-1/2})$ is $\|\widehat{\mathcal{A}}_m^{(p)} - \mathcal{G}_m\|_{\mathrm{sp}}$, with $c_m(\lambda) = 1$. Therefore, we complete the proof. □

B Proofs of Section 3

The key technical lemma behind our results is the Stein's lemma and its generalizations which we present below.

Lemma B.1 (Stein et al. (1972)). Let $x \sim \mathcal{N}(0, I_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ be such that both $\mathbb{E}[\nabla g(x)]$ and $\mathbb{E}[g(x)x]$ exist and are finite. Then

$$\mathbb{E}[g(x)x] = \mathbb{E}[\nabla_x g(x)]. \qquad (27)$$

The following lemma generalizes Stein's lemma to more general distributions and higher-order derivatives.

Lemma B.2 (Sedghi and Anandkumar (2014)). Let $m \ge 1$ and $\mathcal{S}_m(x)$ be defined as in (2). Then for any $g : \mathbb{R}^d \to \mathbb{R}$ satisfying some regularity conditions, we have

$$\mathbb{E}[g(x) \cdot \mathcal{S}_m(x)] = \mathbb{E}[\nabla_x^{(m)} g(x)]. \qquad (28)$$

The following theorem gives an alternate characterization of the loss function L and is the key step in the proof of Theorem 2.

Theorem 4. *The loss function $L(\cdot)$ defined in (19) satisfies that*

$$\begin{aligned} & L(A) \\ = & \sum_{i\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \sum_{j,k\in[d], j\neq k} \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2 \\ & -\mu \sum_{i,j\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^4 \\ & +\lambda \sum_{i\in[d]} (\|a_i\| - 1)^2 \,. \end{aligned} \tag{29}$$

Proof. Since η is zero-mean and independent of x, we have that

$$\mathbb{E}[y \cdot \mathcal{S}_m(x)] = \sum_{i\in[k]} w_i^* \mathbb{E}[g(\langle a_i^*, x\rangle \cdot \mathcal{S}_m(x))] \,, \tag{30}$$

□

Putting $m = 4$ in Lemma B.2, in view of (30), we obtain that

$$\mathbb{E}[y \cdot \mathcal{S}_4(x)] = \sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)](a_i^*)^{\otimes 4}. \tag{31}$$

Thus for any fixed a_j, a_k, we have

$$\begin{aligned} & \mathbb{E}[y \cdot \mathcal{S}_4(x)(a_j, a_j, a_k, a_k)] = \mathbb{E}[y \cdot t_1(x)] \\ = & \sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2, \end{aligned} \tag{32}$$

$$\begin{aligned} & \mathbb{E}[y \cdot \mathcal{S}_4(x)(a_j, a_j, a_j, a_j)] = \mathbb{E}[y \cdot t_2(x)] \\ = & \sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^4. \end{aligned} \tag{33}$$

Now summing over j, k finishes the proof.

B.1 Proof of Theorem 2

Proof. The proof directly follows from Theorem 2.3 of Ge et al. (2017) and Theorem 4. □

B.2 Proof of Theorem 3

We formally state our assumptions for the finite sample landscape analysis below.

Assumption 3. (a) $\|x\|$ has exponentially decaying tails, i.e.

$$\mathbb{P}\left[\|x\|^2 \geq t\right] \leq K_1 e^{-K_2 t^2}, \quad \forall t \geq 0, \tag{34}$$

for some constants $K_1, K_2 > 0$.

(b) Let $l(x, y, A)$ be such that $L(A) = \mathbb{E}[l(x, y, A)] + \lambda \sum_{i\in[k]} (\|a_i\|^2 - 1)^2$. Then there exists a constant $K > 0$ which is at most a polynomial in d and a constant $p \in \mathbb{N}$ such that

$$\begin{aligned} \|\nabla_A l(x, y, A)\| & \leq K \|x\|^p \,, \\ \left\|\nabla_A^2 l(x, y, A)\right\| & \leq K \|x\|^p \,, \end{aligned} \tag{35}$$

for all A such that $\|A_i\| \leq 2$.

In order to establish that the gradient and the Hessian of L are close to their finite sample counterparts, we first consider its truncated version L_T defined as

$$L_T \triangleq \mathbb{E}[l(x, y, A)\mathbb{1}_E], \quad E \triangleq \{\|x\| \leq R\}, \tag{36}$$

where $R = Cd\log(1/\varepsilon)$ for some $\varepsilon < 0$. It follows that L_T is well behaved and exhibits uniform convergence of empirical gradients/Hessians to its population version Ge et al. (2017) for A with bounded norm. Then Theorem 3 follows from showing that the gradient and the Hessian of L_T are close to that of L as well in this setting, which we prove in Lemma B.3. Next we combine this result with Lemma E.5 of Ge et al. (2017) which shows that A with large row norms must also have large gradients and hence cannot be local minima. First we define L_T

Lemma B.3. Let L_T be defined as in (36) and Assumption 3 hold. Then for a sufficiently large constant C and a sufficiently small $\varepsilon > 0$, we have that

$$\|\nabla L(A) - \nabla L_T(A)\|_2 \leq \varepsilon, \tag{37}$$

$$\left\|\nabla^2 L(A) - \nabla^2 L_T(A)\right\|_2 \leq \varepsilon. \tag{38}$$

for all A with row norm $\|A_i\| \leq 2$.

Proof. We have that

$$\begin{aligned} & \|\nabla L(A) - \nabla L_T(A)\|_2 \\ = & \|\mathbb{E}[\nabla l(x, y, A)(1 - \mathbb{1}_E)]\| \\ \overset{(a)}{\leq} & \mathbb{E}[\|\nabla l(x, y, A)\| \, \mathbb{1}\{\|x\| \geq R\}] \\ = & \sum_{i\geq 0} \mathbb{E}[\|\nabla l(x, y, A)\| \, \mathbb{1}\{\|x\| \in [2^i R, 2^{i+1} R]\}] \\ \overset{(b)}{\leq} & \sum_{i\geq 0} K(2^{i+1}R)^p \mathbb{P}\left[\|x\| \geq 2^i R\right] \\ \leq & \sum_{i\geq 0} K(2^{i+1}R)^p e^{-2^i R} \\ \overset{(c)}{\leq} & \sum_{i\geq 0} e^{-2^{i-1}R} \\ = & \sum_{i\geq 0} \varepsilon^{Cd2^{i-1}} \\ \overset{(d)}{\leq} & \sum_{i\geq 0} \varepsilon/2^{i+1} = \varepsilon, \end{aligned} \tag{39}$$

where (a) follows from the Jensen's inequality, (b) follows from Assumption 3, (c) follows from the fact that $K(2x)^p e^{-x} \leq e^{-x/2}$ for x sufficiently large, and (d) follows from choosing C sufficiently large. Similarly for $\left\|\nabla^2 L(A) - \nabla^2 L_T(A)\right\|_2$. □

We are now ready to prove Theorem 3.

Proof. Let A be such that norms of all the rows are less than 2. Then we have from Lemma B.3 that

$$\|\nabla L(A) - \nabla L_T(A)\|_2 \le \varepsilon/4, \tag{40}$$
$$\left\|\nabla^2 L(A) - \nabla^2 L_T(A)\right\|_2 \le \tau_0/4. \tag{41}$$

Notice that the gradient and Hessian of $l(x, y, A)\mathbb{1}_E$ are bounded $\tau = \text{poly}(d, 1/\varepsilon)$ for some fixed polynomial poly. Hence using the uniform convergence of the sample gradients/Hessians to their population counterparts (Ge et al., 2017, Theorem E.3), we have that

$$\left\|\nabla L_T(A) - \nabla \hat{L}_T(A)\right\|_2 \le \varepsilon/6, \tag{42}$$
$$\left\|\nabla^2 L_T(A) - \nabla^2 \hat{L}_T(A)\right\|_2 \le \tau_0/6, \tag{43}$$

whenever $N \ge \text{poly}(d, 1/\varepsilon)$, with high probability. Moreover, from standard concentration inequalities (such as multivariate Chebyshev) it follows that

$$\left\|\nabla \hat{L}(A) - \nabla \hat{L}_T(A) - (\nabla L(A) - \nabla L_T(A))\right\|_2 \le \varepsilon/6, \tag{44}$$
$$\left\|\nabla^2 \hat{L}(A) - \nabla^2 \hat{L}_T(A) - (\nabla^2 L(A) - \nabla^2 L_T(A))\right\|_2 \le \tau_0/6, \tag{45}$$

with high probability, whenever $N \ge \text{poly}(d, 1/\varepsilon)$. Hence, we obtain that

$$\left\|\nabla L(A) - \nabla \hat{L}(A)\right\|_2 \le \varepsilon/2, \tag{46}$$
$$\left\|\nabla^2 L(A) - \nabla^2 \hat{L}(A)\right\|_2 \le \tau_0/2. \tag{47}$$

If A is such that there exists a row A_i with $\|A_i\| \ge 2$, we have from (Ge et al., 2017, Lemma E.5) that $\langle \nabla \hat{L}(A), A_i, \ge \rangle c\lambda \|A_i\|^4$ for a small constant c and thus A cannot be a local minimum for $\hat{L}$. Hence all local minima of $\hat{L}$ must have $\|A_i\| \le 2$ and thus in view of (47) it follows that it also a ε-approximate local minima of L, or more concretely,

$$\|\nabla L(A)\| \le \varepsilon, \quad \nabla^2 L(A) \succcurlyeq -\tau_0 I_d. \tag{48}$$

□

B.3 Landscape design for $k < d$

In the setting where $k = d$ and the regressors $a_1^*, \ldots, a_d^*$ are linearly independent, our loss functions $L_4(\cdot)$ can modified in a straightforward manner to arrive at the loss function $F(\cdot)$ defined in Appendix C.2 of Ge et al. (2017). Hence we have the same landscape properties as that of Theorem B.1 of Ge et al. (2017). The proof is exactly similar to that of our Theorem 2.

In a more geneal scenario where $k < d$ and the regressors $a_1^*, \ldots, a_d^*$ are linearly independent, it turns out that our loss function $L_4(\cdot)$ can also be transformed to obtain the loss $\mathcal{F}(\cdot)$ in Appendix C.3 of Ge et al. (2017) to arrive at Theorem C.1 of Ge et al. (2017) in our setting. The proof is again similar.