A Positive-Definiteness of Θ

Lemma 1. *Let $\Theta = (\theta_{i,j}) \in \mathbb{R}^{m\times m}$ be a symmetric matrix such that $\theta_{i,i} = 1$ and $\sum_{j\neq i} |\theta_{i,j}| \leq 1$ for all $i \in [m]$. Then Θ is symmetric positive semidefinite.*

Proof. Define vectors $\mathbf{z}_1, \ldots, \mathbf{z}_m$ in $\mathbb{R}^{m^2}$ as follows. Let $z_{i,k,l}$ denote coordinate $(k-1)m + l$ of $\mathbf{z}_i$, and set

$$z_{i,k,l} \;=\; \begin{cases} \sqrt{1 - \sum_{j\neq i} |\theta_{i,j}|} & \text{if} \;\; k = i = l \\ \text{sign}(\theta_{i,l})\sqrt{\frac{1}{2}|\theta_{i,l}|} & \text{if} \;\; k = i \neq l \\ \sqrt{\frac{1}{2}|\theta_{i,k}|} & \text{if} \;\; k \neq i = l \\ 0 & \text{otherwise} \end{cases} .$$

Note that Θ is the Gram matrix of $\mathbf{z}_1, \ldots, \mathbf{z}_m$. □

B Technical Result for Sec. 3

In the derivation of our algorithm in Sec. 3, we used the assertion that if

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_t, \boldsymbol{\xi}_t) - \inf_{\mathbf{w}\in\mathcal{W}, \boldsymbol{\xi}\in\mathbb{R}^m} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}, \boldsymbol{\xi}) \leq R.$$

Then it also holds that

$$f(\bar{\mathbf{w}}_T) + h(\bar{\boldsymbol{\xi}}_T) \;\leq\; f(\mathbf{w}^\star) + h(\boldsymbol{\xi}^\star) + R.$$

We will now show why this is true. By definition of $(\mathbf{w}^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$, as a saddle point solution, we have

$$\begin{aligned} &\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_t, \boldsymbol{\xi}_t) \;\leq\; \inf_{\mathbf{w},\boldsymbol{\xi}} \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}, \boldsymbol{\xi}) + TR \;=\; \inf_{\mathbf{w},\boldsymbol{\xi}} \sum_{t=1}^{T} \mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1}) + TR \\ &\leq \inf_{\mathbf{w},\boldsymbol{\xi}} \sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{t=1}^{T} \mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + TR \;=\; T\mathcal{L}(\mathbf{w}^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) + TR \\ &=\; T(f(\mathbf{w}^\star) + h(\boldsymbol{\xi}^\star)) + TR, \end{aligned}$$

where in the last transition we use the fact that at the saddle point, the terms other than the primal function vanishes, due to the KKT conditions. On the other hand, by definition of $\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1}$ as maximizing the Lagrangian with respect to $\mathbf{w}_t, \boldsymbol{\xi}_t$, we have

$$\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_t, \boldsymbol{\xi}_t) = \sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t, \boldsymbol{\xi}_t, \boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1}) \geq \sum_{t=1}^{T} f(\mathbf{w}_t) + h(\boldsymbol{\xi}_t).$$

Let $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$ and $\bar{\boldsymbol{\xi}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t$. Combining the two inequalities above, and using Jensen's inequality, we get that

$$f(\bar{\mathbf{w}}_T) + h(\bar{\boldsymbol{\xi}}_T) \;\leq\; \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t) + h(\boldsymbol{\xi}_t) \;\leq\; f(\mathbf{w}^\star) + h(\boldsymbol{\xi}^\star) + R$$

as required.

C Proof of Thm. 1

We begin with a few definitions. First, note that we can write the function class $\mathcal{H} \otimes^k \mathcal{G}$ as

$$\{\mathbf{x} \mapsto \langle \mathbf{h}(\mathbf{x}), \mathbf{g}(\mathbf{x})\rangle, \mathbf{h} \in \mathcal{H}^k, \mathbf{g} \in \mathcal{G}^k\},$$

where

$$\mathcal{H}^k = \{\mathbf{x} \mapsto (h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})) : h_1, \ldots, h_k \in \mathcal{H}\}$$

and

$$\mathcal{G}^k = \{\mathbf{x} \mapsto (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})) : g_1, \ldots, g_k \in \mathcal{G} \text{ disjoint}\}.$$

For any fixed $\mathbf{g} \in \mathcal{G}^k$, define

$$\mathcal{H}^k_{\mathbf{g}} \;=\; \{\mathbf{x} \mapsto \langle \mathbf{h}(\mathbf{x}), \mathbf{g}(\mathbf{x})\rangle : \mathbf{h} \in \mathcal{H}^k\}.$$

Also, for any sequence of data points $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, define

$$\mathcal{G}^k_{\mathbf{x}} \;=\; \{\mathbf{g}(\mathbf{x}_1), \ldots, \mathbf{g}(\mathbf{x}_m) \;:\; \mathbf{g} \in \mathcal{G}^k\}.$$

Note that since $\mathcal{G}$ has VC-dimension $d_{\mathcal{G}}$, then by Sauer's lemma, we for any fixed $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ it holds that

$$|\{g(\mathbf{x}_1), \ldots, g(\mathbf{x}_m) \;:\; g \in \mathcal{G}\}| \leq \left(\frac{em}{d_{\mathcal{G}}}\right)^{d_{\mathcal{G}}} \leq m^{d_{\mathcal{G}}},$$

where we use the assumption that $d_{\mathcal{G}} > 2$ (note that the form of Sauer's lemma used here assumes $m \geq d_{\mathcal{G}}$, but our theorem holds vacuously otherwise). Therefore, we have

$$|\mathcal{G}^k_{\mathbf{x}}| \leq m^{kd_{\mathcal{G}}} \tag{7}$$

for any $\mathbf{x}$. Finally, let

$$\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) = \sup_{\mathbf{f} \in \mathcal{H}^k_{\mathbf{g}}} \frac{1}{m} \sum_{i=1}^m \sigma_i \mathbf{f}(\mathbf{x}_i)$$

be the empirical counterpart of $\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})$ (without expectation over σ_i).

Let $\lambda > 0$ be a parameter whose value will be determined later. We have the following:

$$\begin{aligned}
\exp\left(\lambda m \mathcal{R}_m(\mathcal{H} \otimes^k \mathcal{G})\right) &= \exp\left(\lambda \mathbb{E}\left[\sup_{\mathbf{g} \in \mathcal{G}^k, \mathbf{h} \in \mathcal{H}^k} \sum_{i=1}^m \sigma_i \langle \mathbf{g}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}_i)\rangle\right]\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda \sup_{\mathbf{g} \in \mathcal{G}^k, \mathbf{h} \in \mathcal{H}^k} \sum_{i=1}^m \sigma_i \langle \mathbf{g}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}_i)\rangle\right)\right] \\
&= \mathbb{E}\left[\sup_{\mathbf{g} \in \mathcal{G}^k} \lambda \exp\left(\sup_{\mathbf{h} \in \mathcal{H}^k} \sum_{i=1}^m \sigma_i \langle \mathbf{g}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}_i)\rangle\right)\right] \\
&\leq \sum_{\mathbf{g} \in \mathcal{G}^k_{\mathbf{x}}} \mathbb{E}\left[\exp\left(\lambda \sup_{\mathbf{h} \in \mathcal{H}^k} \sum_{i=1}^m \sigma_i \langle \mathbf{g}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}_i)\rangle\right)\right] \\
&= \sum_{\mathbf{g} \in \mathcal{G}^k_{\mathbf{x}}} \mathbb{E}\left[\exp\left(\lambda m \hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}})\right)\right] \\
&= \sum_{\mathbf{g} \in \mathcal{G}^k_{\mathbf{x}}} \exp\left(\lambda m \mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})\right) \mathbb{E}\left[\exp\left(\lambda\left(m\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) - m\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})\right)\right)\right] \qquad (8)
\end{aligned}$$

Let us now consider the random variable $m\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) - \mathbb{E}[m\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})]$. Clearly, it is zero-mean. Also, it satisfies the bounded difference property with parameter 2: namely, for any instantiation of $\sigma_1, \ldots, \sigma_m$, changing one of these values results in changing the value of $m\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) = \sup_{\mathbf{f} \in \mathcal{H}^k_{\mathbf{g}}} \sum_i \sigma_i \mathbf{f}(\mathbf{x}_i)$ by at most 2. Invoking McDiarmid's inequality, we get that

$$\Pr\left(\left|m\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) - \mathbb{E}[m\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})]\right| > x\right) \leq \exp\left(-\frac{x^2}{2m}\right).$$

for all $x > 0$.

Using this inequality, and invoking Theorem 1 in [12] (which is a variant of Azuma's inequality for Martingales with subgaussian tails) with respect to the random variable $m\hat{\mathcal{R}}_m(\mathcal{H}^k_{\mathbf{g}}) - \mathbb{E}[m\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}})]$, we get that Eq. (8) can be upper bounded by

$$\sum_{\mathbf{g} \in \mathcal{G}^k_{\mathbf{x}}} \exp\left(\lambda m \mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}}) + 14 m \lambda^2\right).$$

Using Eq. (7), which bounds the cardinality of $\mathcal{G}^k_{\mathbf{x}}$, we can upper bound the above by

$$\leq m^{kd_{\mathcal{G}}} \exp\left(\lambda m \sup_{\mathbf{g}\in\mathcal{G}^k} \mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}}) + 14m\lambda^2\right),$$

Now, recall that this is an upper bound on $\exp\left(\lambda m \mathcal{R}_m(\mathcal{H} \otimes^k \mathcal{G})\right)$, from which we started in Eq. (8). Taking logarithms and dividing by λm, we get that

$$\mathcal{R}_m(\mathcal{H} \otimes^k \mathcal{G}) \leq \frac{kd_{\mathcal{G}} \log(m)}{\lambda m} + \sup_{\mathbf{g}\in\mathcal{G}^k} \mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}}) + 14\lambda.$$

Choosing λ which minimizes the expression above, we get that

$$\mathcal{R}_m(\mathcal{H} \otimes^k \mathcal{G}) \leq \sup_{\mathbf{g}\in\mathcal{G}^k} \mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}}) + 8\sqrt{\frac{kd_{\mathcal{G}} \log(m)}{m}}.$$

It remains to upper bound the first term in the inequality above. For any fixed $\mathbf{g} \in \mathcal{G}^k$, let A_j be the region of the data space in which $g_j(\mathbf{x}) = 1$. Then we have

$$\begin{aligned}
\mathcal{R}_m(\mathcal{H}^k_{\mathbf{g}}) &= \mathbb{E} \sup_{h_1,\ldots,h_k} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \sum_{j=1}^{k} h_j(\mathbf{x}_i) g_j(\mathbf{x}_i) \\
&= \mathbb{E} \sup_{h_1,\ldots,h_k} \frac{1}{m} \sum_{j=1}^{k} \sum_{i:\mathbf{x}_i\in A_j} \sigma_i h_j(\mathbf{x}_i) \\
&\leq \sum_{j=1}^{k} \mathbb{E} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i:\mathbf{x}_i\in A_j} \sigma_i h(\mathbf{x}_i).
\end{aligned}$$

Now, if we assume that $\mathcal{R}_m(\mathcal{H}) \leq \sqrt{d_{\mathcal{H}}/m}$ for any sample of size m, then we can rewrite the above as

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{|i : \mathbf{x}_i \in A_j|}{m} \mathbb{E} \sup_{h\in\mathcal{H}} \frac{1}{|i : \mathbf{x}_i \in A_j|} \sum_{i:\mathbf{x}_i\in A_j} \sigma_i h(\mathbf{x}_i) \leq \sum_{j=1}^{k} \frac{|i : \mathbf{x}_i \in A_j|}{m} \sqrt{\frac{d_{\mathcal{H}}}{|i : \mathbf{x}_i \in A_j|}} \\
&= \frac{\sqrt{d_{\mathcal{H}}} \sum_{j=1}^{k} \sqrt{|i : \mathbf{x}_i \in A_j|}}{m},
\end{aligned}$$

from which the result follows.