7. Some additional discussions

We would like to make a few more comments on the possible future directions based on the current work.

Remark 7.1 (Better computation methods)**.** *As is mentioned above, both* (11) *and* (12) *are highly nontrivial constrained non-convex optimization problems. However, in view of the recent progress on the algorithmic convergence of non-convex phase retrieval, which is closely related to PCA, under a generative model (i.e. (Hand et al., 2018)), it seems reasonable to ask the question whether or not there also exists a relatively efficient algorithm (approximately) solving the non-convex PCA under a generative prior and in particular the ReLU generative prior, which has not been answered to the best of authors' knowledge. We leave this direction for future works.*

Remark 7.2 (Weaker moments on score functions)**.** *We would like to comment that the subgaussian assumption of the score function $S_p(\mathbf{x})$. Previously, such an assumption has been adopted, for example, when analyzing non-asymptotic convergence properties in inverse regression problems (e.g. (Babichev et al., 2018)) and parameter estimation of high dimensional distributions (e.g. (Han et al., 2018)). In our current scenario, it seems unlikely to drop this assumption for general $G(\cdot)$ function, because of the necessity of the chaining method. Though it is interesting whether or not one can obtain a competitive convergence result for specific $G(\cdot)$ functions under weaker moment assumptions. For example, for the multilayer ReLU function, the proving technique involves only a "one-step chaining" argument, which might be improvable to work under weaker moments on $S_p(\mathbf{x})$.*

8. Proof of results

8.1. Proof of Theorem 3.2

Proof. We start from (13) in the proof of Theorem 3.1. Using Lemma 5.1 with $t = G(\widehat{\theta}) - \lambda G(\theta^*)$ and (14), we have

$$\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2^2 \le \sigma_y \|S_p(\mathbf{x})\|_{\psi_2} \sqrt{\frac{1+\kappa}{\kappa}\frac{1}{m}} \cdot \|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 + \left|\langle \mathbb{E}[\widetilde{y} S_p(\mathbf{x})] - \mathbb{E}_m[\widetilde{y} S_p(\mathbf{x})], G(\widehat{\theta}) - G(\overline{\theta})\rangle\right|.$$

Using the fact that $\lambda G(\theta^*) = G(\lambda\theta^*)$ for ReLU generative function when $\lambda > 0$, we have $G(\overline{\theta}) = G(\lambda\theta^*)$. Dividing $\|G(\widehat{\theta}) - G(\lambda\theta^*)\|_2$ from both sides gives

$$\begin{aligned}\|G(\widehat{\theta}) - G(\lambda\theta^*)\|_2 \le& \sigma_y \|S_p(\mathbf{x})\|_{\psi_2} \sqrt{\frac{1+\kappa}{\kappa}\frac{1}{m}} + \frac{\left|\langle \mathbb{E}[\widetilde{y} S_p(\mathbf{x})] - \mathbb{E}_m[\widetilde{y} S_p(\mathbf{x})], G(\widehat{\theta}) - G(\lambda\theta^*)\rangle\right|}{\|G(\widehat{\theta}) - G(\lambda\theta^*)\|_2} \\ \le& \sigma_y \|S_p(\mathbf{x})\|_{\psi_2} \sqrt{\frac{1+\kappa}{\kappa}\frac{1}{m}} + \sup_{t,t'\in\mathbb{R}^k} \frac{|\langle \mathbb{E}[\widetilde{y} S_p(\mathbf{x})] - \mathbb{E}_m[\widetilde{y} S_p(\mathbf{x})], G(t) - G(t')\rangle|}{\|G(t) - G(t')\|_2}.\end{aligned}$$

To this point, our goal is to bound the supreme of the empirical process on the right hand side. First of all, by symmetrization inequality (Lemma 11.2), it is enough to bound

$$\sup_{t,t'\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{y}_i \langle S_p(\mathbf{x}_i), G(t) - G(t')\rangle\right|}{\|G(t) - G(t')\|_2}$$

We will then use the piecewise linear structure of the ReLU function. Note that the ReLU network has n layers with each layer having at most d nodes, where each layer of the network is a linear transformation followed by at most d pointwise nonlinearities. Consider any node in the first layer, which can be written as $\max\{\langle \mathbf{w}, \mathbf{x}\rangle, 0\}$ with a weight vector $\mathbf{w}$ and an input vector $\mathbf{x}$, splits the input space $\mathbb{R}^k$ into two disjoint pieces, namely $\mathcal{P}_1$ and $\mathcal{P}_2$, where for any input in $\mathcal{P}_1$, the node is a linear mapping $\langle \mathbf{w}, \mathbf{x}\rangle$ and for any input in $\mathcal{P}_2$ is the other linear mapping $\langle 0, \mathbf{x}\rangle$.

Thus, each node in the first layer corresponds to a splitting hyperplane in $\mathbb{R}^k$. An induction argument on the number of hyperplanes shows that d nodes in the first layer can split the input space into at most $d^k + 1$ pieces (See, for example, (Winder, 1966) for details). For the second layer, we can consider each piece after the first layer, which is a subset of $\mathbb{R}^k$ and will then be further split into at most $d^k + 1$ pieces. Thus, we will get at most $(d^k + 1)^2$ pieces after the second layer. Continuing this argument through all n layers and we have the input space $\mathbb{R}^k$ is split into at most $(d^k + 1)^n \le (2d)^{kn}$ pieces, where within each piece the function $G(\cdot)$ is simply a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^d$.

Now, we consider any two pieces, namely $\mathcal{P}_1,\ \mathcal{P}_2 \subseteq \mathbb{R}^k$, from the aforementioned collection of pieces, and aim at bounding the following quantity:

$$\sup_{t_1\in\mathcal{P}_1, t_2\in\mathcal{P}_2} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{y}_i \langle S_p(\mathbf{x}_i), G(t_1)-G(t_2)\rangle\right|}{\|G(t_1)-G(t_2)\|_2}.$$

By the previous argument, we know that within $\mathcal{P}_1$ and $\mathcal{P}_2$, the function $G(\cdot)$ can simply be represented by some fixed linear maps W_1 and W_2, respectively. As a consequence, it is enough to bound

$$\begin{aligned}
&\sup_{t_1\in\mathcal{P}_1, t_2\in\mathcal{P}_2} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{y}_i \langle S_p(\mathbf{x}_i), W_1t_1-W_2t_2\rangle\right|}{\|W_1t_1-W_2t_2\|_2}\\
\le &\sup_{t_1,\ t_2\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{y}_i \langle S_p(\mathbf{x}_i), W_1t_1-W_2t_2\rangle\right|}{\|W_1t_1-W_2t_2\|_2}\\
\le &\sup_{t\in\mathbb{R}^{2k}} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{y}_i \langle S_p(\mathbf{x}_i), W_0t\rangle\right|}{\|W_0t\|_2},
\end{aligned}$$

where $W_0 := [W_1,\ -W_2]$, and the last inequality follows from concatenating t_1 and t_2 to form a vector $t\in\mathbb{R}^{2k}$ and then expanding the set to take supremum over $t\in\mathbb{R}^{2k}$. Let $\mathcal{E}_{2k}$ be the subspace in $\mathbb{R}^d$ spanned by the $2k$ columns of W_0, then, the above supremum can be rewritten as

$$H_m := \sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b\rangle\right|.$$

To bound the supremum, we consider a $1/2$-covering net of the set $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$, namely, $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$. A simple volume argument shows that the cardinality $|\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)| \le 3^{2k}$. Using Lemma 9.1 with $\Delta = \eta+4k+2kn\log(2d)$, and taking a union bound over the covering net, we get with probability at least

$$1-3^{2k}\cdot e^{-(\eta+4k+2kn\log(2d))} \ge 1-e^{-\eta-2kn\log(2d)},$$

the following holds,[4]

$$\begin{aligned}
&\sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b\rangle\right| \le\\
&\qquad C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+kn\log(2d)}{m}}\left(\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}+\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right). \quad (20)
\end{aligned}$$

Let $P_{\mathcal{N}}(\cdot)$ be the projection of any point in $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$ onto $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$. we have

$$\begin{aligned}
H_m &\le \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b\rangle\right| + \sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b-P_{\mathcal{N}}(b)\rangle\right|\\
&\le \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b\rangle\right| + \frac{1}{2}\sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i \frac{\langle S_p(\mathbf{x}_i), b-P_{\mathcal{N}}(b)\rangle}{\|b-P_{\mathcal{N}}(b)\|_2}\right|\\
&\le \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i), b\rangle\right| + \frac{1}{2}H_m, \quad (21)
\end{aligned}$$

where the second equality follows from the definition of $\frac{1}{2}$-covering net. Combining the above bound with (20) gives with probability at least $1-e^{-\eta-2kn\log(2d)}$,

$$H_m \le C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+kn\log(2d)}{m}}\left(\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}+\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right),$$

[4] We consider only the case $\eta+4k+2kn\log(2d) \ge \log(em)$, and the case $\eta+4k+2kn\log(2d) < \log(em)$ is similar.

where $C>0$ is an absolute constant. This further implies with probability at least $1-e^{-\eta-2kn\log(2d)}$,

$$\sup_{t_1\in\mathcal{P}_1,t_2\in\mathcal{P}_2}\frac{\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i),G(t_1)-G(t_2)\rangle\right|}{\|G(t_1)-G(t_2)\|_2}$$
$$\leq C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+kn\log(2d)}{m}}\left(\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}+\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right)$$

Taking a further union bound over all combinations of different $\mathcal{P}_1,\ \mathcal{P}_2$ pieces, which have at most $(2d)^{2kn}$ possibilities, implies with probability at least $1-e^{-\eta}$

$$\sup_{t,t'\in\mathbb{R}^k}\frac{\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i),G(t)-G(t')\rangle\right|}{\|G(t)-G(t')\|_2}$$
$$\leq C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+kn\log(2d)}{m}}\left(\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}+\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right)$$

Finally, we apply Bernstein's inequality on the two terms $\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}$ and $\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}$, respectively:

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2-\mathbb{E}\big[|\tilde{y}|^2\big]\geq C\left(\sqrt{\frac{\mathbb{E}[|\tilde{y}|^4]\beta}{m}}+\frac{\tau^2\beta}{m}\right)\right)\leq e^{-\beta},\ \ \beta\geq 0,$$

where $C>1$ is an absolute constant. Substituting the bound $\mathbb{E}\big[|\tilde{y}|^4\big]\leq\mathbb{E}\big[y^4\big]\leq\|y\|_{L_q}^4$, $\mathbb{E}\big[|\tilde{y}|^2\big]\leq\|y\|_{L_q}^2$, and $\tau=m^{1/2(1+\kappa)}\sigma_y$ gives

$$\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}\leq C\sigma_y\left(1+\frac{\beta^{1/4}}{m^{1/4}}+\frac{\beta^{1/2}}{m^{\kappa/(1+\kappa)}}\right)\leq C\sqrt{\beta}\sigma_y,$$

with probability at least $1-e^{-\beta}$ for any $\beta\geq 1$. Similarly, we have,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}-\mathbb{E}\Big[|\tilde{y}|^{2(1+\kappa)}\Big]\geq C\left(\sqrt{\frac{\mathbb{E}\big[|\tilde{y}|^{4(1+\kappa)}\big]\beta}{m}}+\frac{\tau^{2(1+\kappa)}\beta}{m}\right)\right)\leq e^{-\beta}.$$

Recall that $\tau=m^{1/2(1+\kappa)}\sigma_y$. Thus,

$$\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{1/2(1+\kappa)}\leq C\sigma_y\left(1+\frac{\beta^{1/4(1+\kappa)}}{m^{1/4(1+\kappa)}}+\beta^{1/2(1+\kappa)}\right)\leq C'\sigma_y\beta^{1/2(1+\kappa)},$$

with probability at least $1-e^{-\beta}$. Overall, we get with probability at least $1-e^{-\beta}-e^{-\eta}$,

$$\sup_{t,t'\in\mathbb{R}^k}\frac{\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{y}_i\langle S_p(\mathbf{x}_i),G(t)-G(t')\rangle\right|}{\|G(t)-G(t')\|_2}\leq C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{\eta+kn\log(2d)}{m}},$$

which implies the claim. □

8.2. Proof of Theorem 4.2

Proof. We start from (17) in the previous proof and use Lemma 5.4 with $k\log(Lr)$ replaced by $kn\log(2d)$,

$$\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T-G(\theta^*)G(\theta^*)^T\|_F^2$$
$$\leq\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{\frac{4(1+\kappa)}{\kappa}\frac{kn\log(2d)}{m}}\|G(\hat{\theta})-G(\theta^*)\|_2+|\langle\mathbf{S}-\mathbf{S}_0,G(\hat{\theta})G(\hat{\theta})^T-G(\theta^*)G(\theta^*)^T\rangle|.$$

Dividing $\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F$ from both sides, and by (19), we have

$$\begin{aligned}
&\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F \\
\leq&\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{\frac{4(1+\kappa)}{\kappa}\frac{kn\log(2d)}{m}}\frac{\|G(\hat{\theta})-G(\theta^*)\|_2}{\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F} + \frac{|\langle \mathbf{S}-\mathbf{S}_0, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle|}{\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F} \\
\leq&\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{\frac{4(1+\kappa)}{\kappa}\frac{kn\log(2d)}{m}} + \sqrt{2}\cdot\frac{|\langle \mathbf{S}-\mathbf{S}_0, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle|}{\|G(\hat{\theta})-G(\theta^*)\|_2} \\
\leq&\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{\frac{4(1+\kappa)}{\kappa}\frac{kn\log(2d)}{m}} + \sqrt{2}\cdot\sup_{t_1,t_2\in\mathbb{R}^k,\ \|G(t_1)\|_2=1,\ \|G(t_2)\|_2=1}\frac{|\langle \mathbf{S}-\mathbf{S}_0, G(t_1)G(t_1)^T - G(t_2)G(t_2)^T\rangle|}{\|G(t_1)-G(t_2)\|_2},
\end{aligned}$$

where the second inequality follows from Lemma 11.7. In order to bound the supreme on the right hand side, we recall the definition (15), (16) of S and S_0, respectively, and apply the symmetrization inequality. Then, it is enough to bound

$$\sup_{t_1,t_2\in\mathbb{R}^k,\ \|G(t_1)\|_2=1,\ \|G(t_2)\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\left(|\langle G(t_1), S_p(\mathbf{x}_i)\rangle|^2 - |\langle G(t_2), S_p(\mathbf{x}_i)\rangle|^2\right)}{\|G(t_1)-G(t_2)\|_2}\right|.$$

By the same argument as that of Section 8.1 proving Theorem 3.2, we have the ReLU function G split the input space $\mathbb{R}^k$ into at most $(d^k+1)^n \leq (2d)^{kn}$ pieces, where within each piece the function $G(\cdot)$ is simply a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^d$.

Now, we consider any two pieces, namely $\mathcal{P}_1,\ \mathcal{P}_2 \subseteq \mathbb{R}^k$, from the aforementioned collection of pieces, and aim at bounding the following quantity:

$$\sup_{t_1\in\mathcal{P}_1,\ t_2\in\mathcal{P}_2,\ \|G(t_1)\|_2=1,\ \|G(t_2)\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\left(|\langle G(t_1), S_p(\mathbf{x}_i)\rangle|^2 - |\langle G(t_2), S_p(\mathbf{x}_i)\rangle|^2\right)}{\|G(t_1)-G(t_2)\|_2}\right|.$$

Since within $\mathcal{P}_1$ and $\mathcal{P}_2$, the function $G(\cdot)$ can simply be represented by some fixed linear maps W_1 and W_2, respectively. As a consequence, it is enough to bound

$$\begin{aligned}
&\sup_{t_1\in\mathcal{P}_1,\ t_2\in\mathcal{P}_2,\ \|W_1t_1\|_2=1,\ \|W_2t_2\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\left(|\langle W_1t_1, S_p(\mathbf{x}_i)\rangle|^2 - |\langle W_2t_2, S_p(\mathbf{x}_i)\rangle|^2\right)}{\|W_1t_1-W_2t_2\|_2}\right| \\
\leq&\sup_{t_1,\ t_2\in\mathbb{R}^k,\ \|W_1t_1\|_2=1,\ \|W_2t_2\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\langle W_1t_1+W_2t_2, S_p(\mathbf{x}_i)\rangle\langle W_1t_1-W_2t_2, S_p(\mathbf{x}_i)\rangle}{\|W_1t_1-W_2t_2\|_2}\right| \\
\leq&\sup_{b_1,\ b_2\in\mathcal{E}^{2k},\ \|b_1\|_2\leq 2,\ \|b_2\|_2\leq 2}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\langle b_1, S_p(\mathbf{x}_i)\rangle\langle b_2, S_p(\mathbf{x}_i)\rangle}{\|b_2\|_2}\right| \\
=&\sup_{b_1\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ b_2\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1, S_p(\mathbf{x}_i)\rangle\langle b_2, S_p(\mathbf{x}_i)\rangle\right| := H_m
\end{aligned}$$

where $\mathcal{E}^{2k}$ denotes the $2k$-subspace of $\mathbb{R}^d$ spanned by the columns of $[W_1,\ W_2]$, $\mathcal{S}^{d-1}(2)$ denotes the sphere of radius 2 in $\mathbb{R}^d$, and the second from the last inequality follows from replacing $W_1t_1+W_2t_2$ with b_1 and $W_1t_1-W_2t_2$ with b_2, both of which have length bounded by 2.

To this point, we consider a $1/4$-covering net of the set $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$ and a $1/2$ covering net of the set $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)$, namely, $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/4)$ and $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2), 1/2)$, respectively. A simple volume argument shows that the cardinalities $|\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/4)| \leq 5^{2k}$ and $|\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2), 1/2)| \leq 5^{2k}$. Using Lemma 10.1 with $\Delta = \eta + 16k + 2kn\log(2d)$, and taking a union bound over the two covering nets, we get with probability at least

$$1 - 5^{4k}\cdot e^{-(\eta+16k+2kn\log(2d))} \geq 1 - e^{-\eta-2kn\log(2d)},$$

the following holds,[5]

$$\sup_{b_1\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ 1/2),\ b_2\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/4)}\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1,S_p(\mathbf{x}_i)\rangle\langle b_2,S_p(\mathbf{x}_i)\rangle\right|$$
$$\le C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\eta+16k+2kn\log(2d)}\cdot\frac{1+\kappa}{\kappa}\cdot\Big(Q_m\big(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\big)^{1/2}+$$
$$\sqrt{m}\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q}+\|S_p(\mathbf{x})\|_{\psi_2}m^{\frac{\kappa}{2(1+\kappa)}}\Big(\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\Big)^{\frac{1}{2(1+\kappa)}}\Big). \quad (22)$$

Now, let $P_{\mathcal{N}_1}$ be the projection onto $\mathcal{N}\left(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ 1/2\right)$ and $P_{\mathcal{N}_2}$ be the projection onto $\mathcal{N}\left(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},\ 1/4\right)$. We then adopt a similar "one-step chaining" argument as (21).

$$\begin{aligned}H_m\le&\sup_{b_1\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ 1/2),\ b_2\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/4)}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1,S_p(\mathbf{x}_i)\rangle\langle b_2,S_p(\mathbf{x}_i)\rangle\right|\\&+\sup_{b_1\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ b_2\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1-P_{\mathcal{N}_1}(b_1),S_p(\mathbf{x}_i)\rangle\langle b_2-P_{\mathcal{N}_2}(b_2),S_p(\mathbf{x}_i)\rangle\right|\\&+\sup_{b_1\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ b_2\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle P_{\mathcal{N}_1}(b_1),S_p(\mathbf{x}_i)\rangle\langle b_2-P_{\mathcal{N}_2}(b_2),S_p(\mathbf{x}_i)\rangle\right|\\&+\sup_{b_1\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ b_2\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1-P_{\mathcal{N}_1}(b_1),S_p(\mathbf{x}_i)\rangle\langle P_{\mathcal{N}_2}(b_2),S_p(\mathbf{x}_i)\rangle\right|\\\le&\sup_{b_1\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2),\ 1/2),\ b_2\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/4)}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle b_1,S_p(\mathbf{x}_i)\rangle\langle b_2,S_p(\mathbf{x}_i)\rangle\right|+\frac{9}{16}H_m,\end{aligned}$$

where in the first inequality we use the fact that the second term is bounded by $H_m/16$, the third and the last terms are both bounded by $H_m/4$, which implies H_m satisfies the same bound as (22) (with a different constant C) dividing by m.

Taking a further union bound over all combinations of different $\mathcal{P}_1,\ \mathcal{P}_2$ pieces, which contains at most $(2d)^{2kn}$ possibilities, implies with probability at least $1-e^{-\eta}$,

$$\sup_{t_1,t_2\in\mathbb{R}^k,\ \|G(t_1)\|_2=1,\ \|G(t_2)\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m\frac{\varepsilon_i\tilde{y}_i\left(|\langle G(t_1),S_p(\mathbf{x}_i)\rangle|^2-|\langle G(t_2),S_p(\mathbf{x}_i)\rangle|^2\right)}{\|G(t_1)-G(t_2)\|_2}\right|$$
$$\le C\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{\frac{\eta+16k+2kn\log(2d)}{m}}\cdot\frac{1+\kappa}{\kappa}\cdot\Big(\frac{\max_{\mathcal{E}^{2k}}Q_m\big(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\big)^{1/2}}{\sqrt{m}}$$
$$+\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q}+\|S_p(\mathbf{x})\|_{\psi_2}\Big(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\Big)^{\frac{1}{2(1+\kappa)}}\Big), \quad (23)$$

where $\max_{\mathcal{E}^{2k}}Q_m\big(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\big)$ is taken over all possible $(2d)^{2kn}$ subspaces.

It remains to bound the two terms $\left(\frac{1}{m}\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}$ and $\max_{\mathcal{E}^{2k}}Q_m\big(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\big)^{1/2}$, respectively. First, by Bernstein's inequality, we have

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m|\tilde{y}_i|^{2(1+\kappa)}-\mathbb{E}\Big[|\tilde{y}|^{2(1+\kappa)}\Big]\ge C\left(\sqrt{\frac{\mathbb{E}\big[|\tilde{y}|^{4(1+\kappa)}\big]\beta}{m}}+\frac{\tau^{2(1+\kappa)}\beta}{m}\right)\right)\le e^{-\beta},\ \ \beta\ge0.$$

[5]We consider only the case $\eta+4k+2kn\log(2d)\ge\log(em)$, and the case $\eta+4k+2kn\log(2d)<\log(em)$ is similar.

Recall that $\tau = \left(\frac{m}{kn\log(2d)}\right)^{1/2(1+\kappa)}\sigma_y$. Thus,

$$\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^{2(1+\kappa)}\right)^{1/2(1+\kappa)} \leq C\sigma_y\left(1+\left(\frac{kn\log(2d)}{m}\right)^{1/4(1+\kappa)}\beta^{1/4(1+\kappa)}+\beta^{1/2(1+\kappa)}\right)$$
$$\leq 3C\sigma_y\beta^{1/2(1+\kappa)} \leq 3C\sigma_y\sqrt{\beta}, \quad (24)$$

with probability at least $1-e^{-\beta}$, where the last inequality follows from $m \geq kn\log(2d)$. Next, for the term $\max_{\mathcal{E}^{2k}} Q_m\big(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\big)^{1/2}$, we bound it in Lemma 10.4. Substituting (24) and the bound in Lemma 10.4 with $u=\sqrt{\beta}$ into (23) and rearranging terms give

$$\sup_{t_1,t_2\in\mathbb{R}^k,\ \|G(t_1)\|_2=1,\ \|G(t_2)\|_2=1}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{\varepsilon_i\tilde{y}_i\left(|\langle G(t_1),S_p(\mathbf{x}_i)\rangle|^2-|\langle G(t_2),S_p(\mathbf{x}_i)\rangle|^2\right)}{\|G(t_1)-G(t_2)\|_2}\right|$$
$$\leq C\frac{1+\kappa}{\kappa}(\|S_p(\mathbf{x})\|_{\psi_2}+\|S_p(\mathbf{x})\|_{\psi_2}^2)\sigma_y\sqrt{\beta}\sqrt{\frac{\eta+kn\log(2d)}{m}},$$

with probability at least $1-e^{-\beta}-e^{-\eta}$, finishing the proof. □

9. Proof of technical lemmas in Section 5.1

9.1. Proof of Lemma 5.1

Proof. First of all, note that $|y-\widetilde{y}| \leq |y|\cdot\mathbf{1}_{\{|y|>\tau\}}$, where $\mathbf{1}_{\{|y|>\tau\}}$ is an indicator function. Thus, it follows,

$$|\mathbb{E}[y\langle S_p(\mathbf{x}),t\rangle-\widetilde{y}\langle S_p(\mathbf{x}),t\rangle]| \leq \mathbb{E}[|y-\widetilde{y}|\cdot|\langle S_p(\mathbf{x}),t\rangle|] \leq \mathbb{E}\big[|y|\cdot\mathbf{1}_{\{|y|>\tau\}}\cdot|\langle S_p(\mathbf{x}),t\rangle|\big].$$

By Cauchy-Schwarz and then Holder's inequality, we have

$$\begin{aligned}|\mathbb{E}[y\langle S_p(\mathbf{x}),t\rangle-\widetilde{y}\langle S_p(\mathbf{x}),t\rangle]| \leq& \mathbb{E}\big[|y|^2\cdot|\langle S_p(\mathbf{x}),t\rangle|^2\big]^{1/2}\cdot Pr(|y|>\tau)^{1/2}\\ \leq& \mathbb{E}\Big[|y|^{2(1+\kappa)}\Big]^{1/(2(1+\kappa))}\mathbb{E}\Big[|\langle S_p(\mathbf{x}),t\rangle|^{\frac{2(1+\kappa)}{\kappa}}\Big]^{\frac{\kappa}{2(1+\kappa)}}\cdot Pr(|y|>\tau)^{1/2}\\ \leq& \|y\|_{L_q}\sqrt{\frac{2(1+\kappa)}{\kappa}}\|S_p(\mathbf{x})\|_{\psi_2}\|t\|_2\cdot Pr(|y|>\tau)^{1/2}.\end{aligned}$$

Thus, it follows,

$$Pr(|y|>\tau) \leq \frac{\mathbb{E}\big[|y|^{2(1+\kappa)}\big]}{\tau^{2(1+\kappa)}} \leq \frac{\|y\|_{L_q}^{q/2}}{\big(m^{1/2(1+\kappa)}\sigma_y\big)^{q/2}} \leq \frac{1}{m^{q/4(1+\kappa)}} \leq \frac{1}{m},$$

where the first inequality follows from Markov inequality, the third inequality follows from the fact that $\sigma_y \geq \|y\|_{L_q}$, and the last inequality follows from $q > 4(1+\kappa)$. This implies the claim. □

9.2. Subgaussian concentration of a multiplier process

The proof of Lemma 5.2 relies on the following key result which provides a subgaussian type concentration for a truncated heavy-tailed multiplier process.

Lemma 9.1. *Under Assumption 3.1 and $\tilde{y}_i = \mathrm{sign}(y_i)|y_i|\wedge\tau$, where $\tau = m^{1/2(1+\kappa)}\sigma_y$ with $\sigma_y \geq \|y\|_{L_q}$, $\kappa\in(0,\frac{q}{4}-1)$ being any chosen constants, we have for any fixed $t,t'\in\mathbb{R}^d$ and any fixed $\Delta\geq 1$, with probability at least $1-\exp(-\Delta)$,*

- *When $\Delta < \log(em)$,*

$$\left|\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t-t'\rangle\right| \leq C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\Delta m}.$$

- *When* $\log(em) \le \Delta < m$,

$$\left|\sum_{i=1}^m \varepsilon_i \tilde{y}_i \langle S_p(\mathbf{x}_i), t-t'\rangle\right| \le C\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\sqrt{\Delta}\sqrt{\frac{1+\kappa}{\kappa}}\left(\left(\sum_{i=1}^m |\tilde{y}_i|^2\right)^{1/2} + m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^m |\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right).$$

- *When* $\Delta \ge m$,

$$\left|\sum_{i=1}^m \varepsilon_i \tilde{y}_i \langle S_p(\mathbf{x}_i), t-t'\rangle\right| \le C\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\sqrt{\Delta}\left(\sum_{i=1}^m |\tilde{y}_i|^2\right)^{1/2}.$$

where, for each of the three cases, $C > 1$ *is an absolute constant.*

Remark 9.1. *Not that in Lemma 9.1 is not a uniform concentration result since it holds for any single pair* (t, t')*. However, the power of this lemma is that it then allows us to take union bound over all* (t, t') *before taking care of the heavy-tailed random variables* $\{\tilde{y}_i\}_{i=1}^m$*, thereby obtaining a bound of optimal order.*

Proof of Lemma 9.1. We separate the proof of three cases:

1. Proof of the first case: Since the quantity Δ is relatively small, we will just apply Bernstein's inequality (Lemma 11.1). First of all, for any integer $p \ge 2$, the following holds,

$$\begin{aligned}
&\mathbb{E}\left[|\varepsilon\tilde{y}\langle S_p(\mathbf{x}), t-t'\rangle|^p\right]\\
\le&\mathbb{E}\left[|\varepsilon\langle S_p(\mathbf{x}), t-t'\rangle|^p\, y^2\cdot|\tilde{y}|^{p-2}\right]\\
\le&\mathbb{E}\left[|\langle S_p(\mathbf{x}), t-t'\rangle|^p\, y^2\right]\cdot\tau^{p-2}\\
\le&\tau^{p-2}\mathbb{E}\left[|\langle S_p(\mathbf{x}), t-t'\rangle|^{2p}\right]^{1/2}\mathbb{E}\left[|y|^4\right]^{1/2}\\
\le&\|y\|_{L_q}^2\tau^{p-2}\|S_p(\mathbf{x})\|_{\psi_2}^p\,(2p)^{p/2}\,\|t-t'\|_2^p,
\end{aligned}$$

where the second inequality follows from the truncation, the third inequality from Hölder's inequality, and the last inequality follows from the following bound: For any $p \ge 4$, and any $\mathbf{v} \in \mathbb{R}^d$,

$$\left(\mathbb{E}\langle S_p(\mathbf{x}), \mathbf{v}\rangle^{2p}\right)^{1/2p} \le (2p)^{1/2}\,\|S_p(\mathbf{x})\|_{\psi_2}\|\mathbf{v}\|_2.$$

Next, by Stirling's approximation, $p! \ge \sqrt{2\pi}\sqrt{p}(p/e)^p$, thus there exist some absolute constants $C', C'' > 0$ such that

$$\mathbb{E}\left[|\varepsilon\tilde{y}\langle S_p(\mathbf{x}), t-t'\rangle|^p\right] \le p!\left(C'\|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right)^2\left(C''\tau\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right)^{p-2}.$$

Thus, substituting $\tau = cm^{1/2(1+\kappa)}\sigma_y$ and by Bernstein's inequality, we have with probability at least $1-e^{-u}$,

$$\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i), t-t'\rangle\right| \le C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\left(\sqrt{\frac{u}{m}} + \frac{u}{m^{1-\frac{1}{2(1+\kappa)}}}\right).$$

Now, we take $u = \Delta$, which gives

$$\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i), t-t'\rangle\right| \le C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\sqrt{\frac{\Delta}{m}}\left(1 + \frac{\sqrt{\Delta}}{m^{\frac{1}{2}-\frac{1}{2(1+\kappa)}}}\right),$$

where C is an absolute constant. Since $\Delta < \log(em)$, it follows

$$\frac{\sqrt{\Delta}}{m^{\frac{1}{2}-\frac{1}{2(1+\kappa)}}} \le \sqrt{\frac{\log(em)}{m^{\frac{\kappa}{1+\kappa}}}} \le \sqrt{\frac{1+\kappa}{\kappa}},$$

which implies the first case of the lemma.

2. Proof of the second case: This case is somewhat more involved. The main difficulty is that simply applying the truncation bound $|\tilde{y}| \leq \tau$ results a bound looser than subgaussian on probability and suboptimal in terms of number of samples m. The solution is to refrain from bounding $\tilde{y}$ altogether and consider working on the term $\langle S_p(\mathbf{x}), t - t' \rangle$ only, which is subgaussian anyway. In particular, we will use the Montgomery-Smith inequality (Lemma 11.3) to effectively separate $\tilde{y}$ from $\langle S_p(\mathbf{x}), t - t' \rangle$.

Let $\omega_i = \langle S_p(\mathbf{x}_i), t - t' \rangle$. For any sequence of scalars $\{X_i\}_{i=1}^m$, let $\{X_i^*\}_{i=1}^m$ is the *non-increasing* rearrangement of $\{|X_i|\}_{i=1}^m$. Then, in view of (48) in Lemma 11.3, we choose the index set I to be the union of the p largest entries of $\{\omega_i\}_{i=1}^m$ and p largest entries of $\{\tilde{y}_i\}_{i=1}^m$, where p is a number to be chosen later. Then, we obtain the following bound with probability at least $1 - \exp(-u^2/2)$

$$\begin{aligned}\left|\sum_{i=1}^m \varepsilon_i \tilde{y}_i \omega_i\right| \leq & \sum_{i\in I} |\tilde{y}_i \omega_i| + u \left(\sum_{i\notin I} |\tilde{y}_i \omega_i|^2\right)^{1/2} \\ \leq & \left(\sum_{i\in I} |\tilde{y}_i|^2\right)^{1/2} \left(\sum_{i\in I} |\omega_i|^2\right)^{1/2} + u \left(\sum_{i\notin I} |\tilde{y}_i|^{2r}\right)^{1/2r} \left(\sum_{i\notin I} |\omega_i|^{2r'}\right)^{1/2r'} \\ \leq & 2\left(\sum_{i\leq p} |\tilde{y}_i^*|^2\right)^{1/2} \left(\sum_{i\leq p} |\omega_i^*|^2\right)^{1/2} + u \left(\sum_{i>p} |\tilde{y}_i^*|^{2r}\right)^{1/2r} \left(\sum_{i>p} |\omega_i^*|^{2r'}\right)^{1/2r'}, \end{aligned} \tag{25}$$

where r, r' are conjugate exponents $\frac{1}{r} + \frac{1}{r'} = 1$, the second inequality follows from Holder's inequality and the last inequality follows from the definition of the set I that

$$\left(2\sum_{i\leq p} |\tilde{y}_i^*|^2\right)^{1/2} \geq \left(\sum_{i\in I} |\tilde{y}_i|^2\right)^{1/2}, \quad \left(2\sum_{i\leq p} |\omega_i^*|^2\right)^{1/2} \geq \left(\sum_{i\in I} |\omega_i|^2\right)^{1/2}$$

and

$$\left(\sum_{i>p} |\tilde{y}_i^*|^{2r}\right)^{1/2r} \geq \left(\sum_{i\notin I} |\tilde{y}_i|^{2r}\right)^{1/2r}, \quad \left(\sum_{i>p} |\omega_i^*|^{2r'}\right)^{1/2r'} \geq \left(\sum_{i\notin I} |\omega_i|^{2r'}\right)^{1/2r'}.$$

Next, we take $u = \sqrt{\Delta}$, so that the probability matches the statement in the lemma. Now we need to choose an appropriate cut-off index p. The intuition is that this cut off index should be determined by how large Δ is, so that the magnitude of the latter $m - p$ entries in $\{\omega_i^*\}_{i=1}^m$ does not depend on Δ, which is possible via the fact that they are in lower ordered positions thus "naturally" small. On the other hand, by the second part of Lemma 11.3, $p \approx u^2 = \Delta$ makes the decreased ordering bound close to the infimum. Thus, we will choose $p = \lfloor \Delta / \log(em/\Delta) \rfloor$, and choose $r = 1 + \kappa$, $r' = (1+\kappa)/\kappa$.

In the following, we will bound $\left(\sum_{i\leq p} |\omega_i^*|^2\right)^{1/2}$ and $\left(\sum_{i>p} |\omega_i^*|^{2r'}\right)^{1/2r'}$, respectively.

- **Bounding the term $\left(\sum_{i\leq p} |\omega_i^*|^2\right)^{1/2}$:** By assumption, we have $S_p(\mathbf{x}_i)$ is a subgaussian vector, thus, ω_i is a subgaussian random variable and ω_i^2 is subexponential. Furthermore,

$$\|\omega_i^2\|_{\psi_1} \leq 2\|\omega_i\|_{\psi_2}^2 \leq 2\|S_p(\mathbf{x}_i)\|_{\psi_2}^2 \|t - t'\|_2^2.$$

  It then follows from Bernstein's inequality (Lemma 11.1) that for any fixed set $J \subseteq \{1, 2, \cdots, m\}$ with $|J| = p$,

$$\mathbb{P}\left(\left|\frac{1}{p}\sum_{i\in J} \left(\omega_i^2 - \mathbb{E}[\omega_i^2]\right)\right| \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}^2 \|t - t'\|_2^2 \left(\sqrt{\frac{2u}{p}} + \frac{u}{p}\right)\right) \leq 2\exp(-u),$$

where $C > 0$ is an absolute constant. We choose $u = 4\Delta$. Since by assumption $\Delta \geq \lfloor \Delta / \log(em/\Delta) \rfloor = p \geq 1$, the factor Δ/p dominates the right hand side. Note that $\mathbb{E}\big[\omega_i^2\big] \leq \|S_p(\mathbf{x}_i)\|_{\psi_2}^2 \|t' - t\|_2^2$, we have following bound,

$$\mathbb{P}\left(\left(\sum_{i\in J}\omega_i^2\right)^{1/2} \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'-t\|_2\sqrt{\Delta}\right) \leq 2\exp(-4\Delta),$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\left(\sum_{i=1}^{k}|\omega_i^*|^2\right)^{1/2} \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'-t\|_2\sqrt{\Delta}\right)\\
=&\mathbb{P}\left(\exists J\subseteq\{1,\cdots,m\},\ |J|=p : \left(\sum_{i\in J}\omega_i^2\right)^{1/2} \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'-t\|_2\sqrt{\Delta}\right)\\
\leq&\binom{m}{p}\cdot\mathbb{P}\left(\left(\sum_{i\in J}\omega_i^2\right)^{1/2} \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'-t\|_2\sqrt{\Delta}\right)\\
\leq&2\binom{m}{p}\exp(-4\Delta) \leq 2\left(\frac{em}{p}\right)^p\exp(-4\Delta) \leq 2\exp(-\Delta),
\end{aligned} \tag{26}$$

where the last inequality follows from the setup that $p = \lfloor \Delta/\log(em/\Delta)\rfloor$ and thus $\left(\frac{em}{p}\right)^p \leq \exp(3\Delta)$, which obviously holds when $p = 1$ due to $\log(em) < \Delta$, and when $p \geq 2$, we have the following chain of inequalities,

$$\begin{aligned}
\left(\frac{em}{p}\right)^p \leq 2\exp\left(\frac{\Delta}{\log\frac{em}{\Delta}}\log\left(\frac{em}{\frac{\Delta}{\log\frac{em}{\Delta}}-1}\right)\right) &\leq 2\exp\left(\frac{\Delta}{\log\frac{em}{\Delta}}\log\left(\frac{em}{\Delta-\log\frac{em}{\Delta}}\log\frac{em}{\Delta}\right)\right)\\
&\leq 2\exp\left(\frac{\Delta}{\log\frac{em}{\Delta}}\log\left(\frac{2em}{\Delta}\log\frac{em}{\Delta}\right)\right) \leq \exp(3\Delta),
\end{aligned} \tag{27}$$

where the second from the last inequality follows from $\Delta \geq 2\log(em/\Delta)$ due to $p \geq 2$.

- **Bounding the term** $\left(\sum_{i>p}|\omega_i^*|^{2r'}\right)^{1/2r'}$, with $r' = (1+\kappa)/\kappa$: First of all, by assumption, we have $S_p(\mathbf{x}_i)$ is subgaussian, thus, for any index $j \in \{1,2,\cdots,m\}$,

$$\mathbb{P}\left(|\omega_j| - \mathbb{E}[|\omega_j|] \geq C\beta\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right) \leq e^{-\beta^2},$$

for any $\beta \geq 0$, where $C > 0$ is a positive constant. For the rest of the proof in this case, the index i will be dedicated for non-increasing ordering index, i.e. $\{\omega_i^*\}_{i=1}^m$. For each index i, we then choose

$$\beta = c_\kappa(m/i)^{\kappa/4(1+\kappa)},\ \ c_\kappa := \sqrt{\frac{5(1+\kappa)}{\kappa}}.$$

The reason for such choices will be clear as we compute the probabilistic bound for ω_i^*'s. Before doing that, we first substitute above choice of β into the sbugaussian bound and get for any arbitary index $j \in \{1,2,\cdots,m\}$,

$$\mathbb{P}\left(|\omega_j| - \mathbb{E}[|\omega_j|] \geq Cc_\kappa\left(\frac{m}{i}\right)^{\kappa/4(1+\kappa)}\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right) \leq \exp\left(-c_\kappa^2\left(\frac{m}{i}\right)^{\kappa/2(1+\kappa)}\right).$$

Since $\mathbb{E}[|\omega_j|] \leq \mathbb{E}\big[\omega_j^2\big]^{1/2} \leq \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2$, it follows

$$\mathbb{P}\left(|\omega_j| \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\kappa/4(1+\kappa)}\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right) \leq \exp\left(-c_\kappa^2\left(\frac{m}{i}\right)^{\kappa/2(1+\kappa)}\right).$$

Thus, it follows

$$\begin{aligned}
&\mathbb{P}\left(\omega_i^* \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right)\\
=&\mathbb{P}\left(\exists J\subseteq\{1,\cdots,m\},\ |J|=i:\ \omega_j \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2,\ \forall j\in J\right)\\
\leq&\binom{m}{i}\cdot\mathbb{P}\left(|\omega_j| \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right)^i\\
\leq&\binom{m}{i}\cdot\exp\left(-c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right) \leq \left(\frac{em}{i}\right)^i \exp\left(-c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right)\\
=&\exp\left(i\log(em/i) - c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right).
\end{aligned}$$

Taking a union bound over all indices i gives

$$\begin{aligned}
&\mathbb{P}\left(\exists i\in\{1,2,\cdots,m\},\ \omega_i^* \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right)\\
\leq&\sum_{i=p+1}^{m}\exp\left(i\log(em/i) - c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right)\\
\leq&m\cdot\exp\left(p\log(em/p) - c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} p^{\frac{2+\kappa}{2(1+\kappa)}}\right),
\end{aligned}$$

where the second inequality follows from the fact that $c_\kappa = \sqrt{\frac{5(1+\kappa)}{\kappa}}$ and thus, the term $i\log(em/i) - c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}$ is monotonically decreasing with respect to the index i. To bound the probability on the right hand size, we first use the same argument as (27) along with $\log(em)\leq\Delta$ and get

$$m\cdot\exp\left(p\log(em/p)\right) \leq \exp\left(\log(em)+p\log(em/p)\right) \leq \exp(4\Delta). \tag{28}$$

For the latter term, we claim that setting $c_\kappa = \sqrt{\frac{5(1+\kappa)}{\kappa}}$ gives

$$c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} p^{\frac{2+\kappa}{2(1+\kappa)}} \geq 5\Delta. \tag{29}$$

To see why this is true, we first substitute the definition $p=\Delta/\log(em/\Delta)$, $c_\kappa = \sqrt{\frac{5(1+\kappa)}{\kappa}}$ and it is enough to show

$$\frac{1+\kappa}{\kappa}\left(\frac{m}{\Delta}\right)^{\frac{\kappa}{2(1+\kappa)}} \geq \left(\log\frac{em}{\Delta}\right)^{\frac{2+\kappa}{2(1+\kappa)}},$$

which is equivalent to showing

$$\left(\frac{1+\kappa}{\kappa}\right)^{\frac{2(1+\kappa)}{\kappa}}\frac{m}{\Delta} - \left(\log\frac{em}{\Delta}\right)^{\frac{2+\kappa}{\kappa}} \geq 0.$$

Note that $m\geq\Delta$, this is further equivalent to saying the minimum of the function

$$f(x) = \left(\frac{1+\kappa}{\kappa}\right)^{\frac{2(1+\kappa)}{\kappa}} - \frac{(1+\log x)^{\frac{2+\kappa}{\kappa}}}{x}$$

is nonnegative over $x\geq 1$ region, which is easy to verify. Thus, combining (28) and (29) gives

$$\mathbb{P}\left(\exists i\in\{1,2,\cdots,m\},\ \omega_i^* \geq (1+Cc_\kappa)\left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\right) \leq \exp(-\Delta).$$

This implies with probability at least $1-\exp(-\Delta)$,

$$\begin{aligned}\sum_{i>p}|\omega_i^*|^{2r'} \leq& \sum_{i=p+1}^{m}(1+Cc_\kappa)^{\frac{2(1+\kappa)}{\kappa}}\left(\frac{m}{i}\right)^{1/2}\|S_p(\mathbf{x})\|_{\psi_2}^{\frac{2(1+\kappa)}{\kappa}}\|t-t'\|_2^{\frac{2(1+\kappa)}{\kappa}}\\ \leq&(1+Cc_\kappa)^{\frac{2(1+\kappa)}{\kappa}}\|S_p(\mathbf{x})\|_{\psi_2}^{\frac{2(1+\kappa)}{\kappa}}\|t-t'\|_2^{\frac{2(1+\kappa)}{\kappa}}\sum_{i=1}^{m}(m/i)^{1/2}\\ \leq&\left((C+1)\sqrt{\frac{1+\kappa}{\kappa}}\right)^{\frac{2(1+\kappa)}{\kappa}}\|S_p(\mathbf{x})\|_{\psi_2}^{\frac{2(1+\kappa)}{\kappa}}\|t-t'\|_2^{\frac{2(1+\kappa)}{\kappa}}m\end{aligned}$$

which implies with probability at least $1-\exp(-\Delta)$,

$$\left(\sum_{i>p}|\omega_i^*|^{2r'}\right)^{1/2r'} \leq C'\sqrt{\frac{1+\kappa}{\kappa}}\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2 m^{\kappa/2(1+\kappa)}, \tag{30}$$

for positive constant $C'=1+C$.

Finally, substituting (26) and (30) into (39) finishes the second part of the Lemma 9.1.

3. Proof of the third case: By Cauchy-Schwarz inequality, we have

$$\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t-t'\rangle\right| \leq \left(\frac{1}{m}\sum_{i=1}^{m}\omega_i^2\right)^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}\tilde{y}_i^2\right)^{1/2},$$

where $\omega_i=\langle S_p(\mathbf{x}),t-t'\rangle$. Thus, ω_i^2 is sub-exponential with

$$\|\omega_i^2\|_{\psi_1}\leq 2\|\omega_i\|_{\psi_2}^2\leq 2\|S_p(\mathbf{x}_i)\|_{\psi_2}^2\|t-t'\|_2^2.$$

we use Bernstein's inequality again,

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(\omega_i^2-\mathbb{E}\left[\omega_i^2\right]\right)\right|\geq C\|S_p(\mathbf{x})\|_{\psi_2}^2\|t-t'\|_2^2\left(\sqrt{\frac{2u}{m}}+\frac{u}{m}\right)\right)\leq 2\exp(-u).$$

We take $u=\Delta$. Using the fact that $\Delta>m$, we have the term $\frac{\Delta}{m}$ dominates the right hand side. Furthermore, $\mathbb{E}\left[\omega_j^2\right]\leq \|S_p(\mathbf{x})\|_{\psi_2}^2\|t-t'\|_2^2\leq\frac{\Delta}{m}\|S_p(\mathbf{x})\|_{\psi_2}^2\|t-t'\|_2^2$. Thus,

$$\mathbb{P}\left(\left(\frac{1}{m}\sum_{i=1}^{m}\omega_i^2\right)^{1/2}\geq C\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\sqrt{\frac{\Delta}{m}}\right)\leq 2\exp(-\Delta),$$

for some absolute constant $C>0$, which implies the third part of the lemma. □

9.3. Uniform concentration over a δ-covering net

Using the previous subgaussian concentration result, we can proof the following lemma which gives a uniform bound over a δ-covering net of the set $G(\mathcal{B}^k(r))$. For the rest of the proof, we use $\mathcal{B}$ to denote $\mathcal{B}^k(r)$ for simplicity.

Lemma 9.2. *For any $\delta\in(0,r)$, let $\mathcal{N}(\delta,G(\mathcal{B}))$ be a δ-covering net of the set $G(\mathcal{B})$ with respect to the norm $\|\cdot\|_2$, such that*

$$\log|\mathcal{N}(\delta,G(\mathcal{B}))|\leq 2k\log(4Lr/\delta),$$

where $\mathcal{B}$ is the ball in $\mathbb{R}^k$ with radius r and $c>0$ is an absolute constant. Under Assumption 3.1 and $\tilde{y}_i=\operatorname{sign}(y_i)|y_i|\wedge\tau$, where $\tau=cm^{1/2(1+\kappa)}\sigma_y$, for any $\eta,\beta\geq 1$, with probability at least $1-e^{-\beta}-e^{-\eta}$, for any $t,t'\in\mathcal{N}(\delta,G(\mathcal{B}))$,

$$\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t-t'\rangle\right|\leq C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|t-t'\|_2.$$

Proof of Lemma 9.2. First of all, we need to show that there does exist a net satisfying the proposed cardinality bound. In order to construct such a net, let $\mathcal{N}(\delta/L, \mathcal{B})$ be the δ/L-covering of the set $\mathcal{B} \subseteq \mathbb{R}^k$. It is known that there exists a net such that

$$\log|\mathcal{N}(\delta/L, \mathcal{B})| \leq k \cdot \log(4Lr/\delta).$$

Then, due to the L-Lipschitz property of G, the set $G(\mathcal{N}(\delta/L, \mathcal{B}))$ forms a δ-net of $G(\mathcal{B})$ with the same cardinality bound. Thus, we let $\mathcal{N}(\delta, G(\mathcal{B})) = G(\mathcal{N}(\delta/L, \mathcal{B}))$ and it follows

$$\log|\mathcal{N}(\delta, G(\mathcal{B}))| = \log|G(\mathcal{N}(\delta/L, \mathcal{B}))| \leq k \cdot \log(4Lr/\delta). \tag{31}$$

To this point, we let $\Delta = \eta + 4k \cdot \log(4Lr/\delta)$ and try to apply Lemma 9.1:

- When $\beta + 4k \cdot \log(4Lr/\delta) \leq \log(em)$, by taking a union bound over all $t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$, we have with probability $1 - e^{-\eta}$, $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \tilde{y}_i \langle S_p(\mathbf{x}_i), t - t'\rangle\right| \leq C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\|t - t'\|_2\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta + k \cdot \log(4Lr/\delta)}{m}}.$$

- When $\log(em) \leq \eta + 4k \cdot \log(4Lr/\delta) \leq m$, again, taking a union bound over all $t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$, we have with probability $1 - e^{-\eta}$, $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \tilde{y}_i \langle S_p(\mathbf{x}_i), t - t'\rangle\right| \leq C\|S_p(\mathbf{x})\|_{\psi_2}\|t - t'\|_2\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta + k \cdot \log(4Lr/\delta)}{m}} \cdot \left(\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^2\right)^{1/2} + \left(\frac{1}{m}\sum_{i=1}^{m}|\widetilde{y_i}|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right). \tag{32}$$

To this point, we apply Bernstein's inequality again on the two terms $\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^2\right)^{1/2}$ and $\left(\frac{1}{m}\sum_{i=1}^{m}|\widetilde{y_i}|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}$, respectively:

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^2 - \mathbb{E}\big[|\tilde{y}|^2\big] \geq C\left(\sqrt{\frac{\mathbb{E}[|\tilde{y}|^4]\beta}{m}} + \frac{\tau^2\beta}{m}\right)\right) \leq e^{-\beta}, \ \ \beta \geq 0,$$

where $C > 1$ is an absolute constant. Substituting the bound $\mathbb{E}\big[|\tilde{y}|^4\big] \leq \mathbb{E}\big[y^4\big] \leq \|y\|_{L_q}^4$, $\mathbb{E}\big[|\tilde{y}|^2\big] \leq \|y\|_{L_q}^2$, and $\tau = m^{1/2(1+\kappa)}\sigma_y$ gives

$$\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^2\right)^{1/2} \leq C\sigma_y\left(1 + \frac{\beta^{1/4}}{m^{1/4}} + \frac{\beta^{1/2}}{m^{\kappa/(1+\kappa)}}\right) \leq C\sqrt{\beta}\|y\|_{L_q}, \tag{33}$$

with probability at least $1 - e^{-\beta}$ for any $\beta \geq 1$. Similarly, we have,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^{2(1+\kappa)} - \mathbb{E}\Big[|\tilde{y}|^{2(1+\kappa)}\Big] \geq C\left(\sqrt{\frac{\mathbb{E}\big[|\tilde{y}|^{4(1+\kappa)}\big]\beta}{m}} + \frac{\tau^{2(1+\kappa)}\beta}{m}\right)\right) \leq e^{-\beta}, \ \ \beta \geq 0.$$

Recall that $\tau = m^{1/2(1+\kappa)}\sigma_y$. Thus,

$$\left(\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i|^{2(1+\kappa)}\right)^{1/2(1+\kappa)} \leq C\sigma_y\left(1 + \frac{\beta^{1/4(1+\kappa)}}{m^{1/4(1+\kappa)}} + \beta^{1/2(1+\kappa)}\right), \tag{34}$$

with probability at least $1 - e^{-\beta}$. Overall, substitute (34) and (33) into (32) and use the fact that $\|y\|_{L_q} \leq \sigma_y$, we have with probability at least $1 - e^{-\eta} - e^{-\beta}$, $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \tilde{y}_i \langle S_p(\mathbf{x}_i), t - t'\rangle\right| \leq C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}\|t - t'\|_2\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta + k \cdot \log(4Lr/\delta)}{m}}.$$

- When $\eta + 4k \cdot \log(4Lr/\delta) \geq m$, this case is similar to the last case and omitted for brevity.

Overall, we finished the proof. □

9.4. Proof of Lemma 5.2

Proof. To simplify the notations, for this proof, we define the following semi-norm: For any $t, t' \in G(\mathcal{B})$,

$$\|t - t'\|_{G_m} := \left| \frac{1}{m} \sum_{i=1}^{m} \tilde{y}_i \langle S_p(\mathbf{x}_i), t - t' \rangle - \mathbb{E}[\tilde{y} \langle S_p(\mathbf{x}), t - t' \rangle] \right|.$$

Also, for any subset $T \subseteq G(\mathcal{B})$, let $P_T(\cdot)$ be the projection on to this set.

The approach we will take is Dudley's chaining technique together with the fact that $G(\mathcal{B})$ is a Lipschitz map of an $\mathbb{R}^k$ ball. Specifically, for any $\delta > 0$, we consider a sequence of covering nets $\left\{\hat{T}_i\right\}_{i=0}^{\ell}$ of the set $G(\mathcal{B})$ with respect to $\|\cdot\|_2$, where $\hat{T}_i$ denotes the $2^{-i}\delta$-covering of the set $\hat{T}_{i+1},\ i = 0, 1, \cdots, \ell - 1$ and $\hat{T}_\ell = \mathcal{N}(2^{-\ell}\delta, G(\mathcal{B}))$ is the minimum $2^{-\ell}\delta$-covering of the set $G(\mathcal{B})$. We have $\widehat{T}_0 \subseteq \widehat{T}_1 \subseteq \cdots \subseteq \widehat{T}_\ell$. The way we construct these nets and bound the cardinalities are of similar flavor to that of last lemma. We consider a sequence of covering nets $\{T_i\}_{i=0}^{\ell}$ on the ball $\mathcal{B}$ with respect to $\|\cdot\|_2$, where $T_i := \mathcal{N}(2^{-i}\delta/L, T_{i+1})$ and $T_\ell := \mathcal{N}(2^{-\ell}\delta/L, \mathcal{B})$. It is known that there exists a net such that

$$\log |\mathcal{N}(2^{-\ell}\delta/L, \mathcal{B})| \leq k \cdot \log\left(4Lr/(2^{-\ell}\delta)\right) \leq k \cdot \log(4Lr/\delta) + k\ell,$$

and there exists nets such that

$$\log |\mathcal{N}(2^{-i}\delta/L, T_{i+1})| \leq \log |\mathcal{N}(2^{-i}\delta/L, \mathcal{B})| \leq k \cdot \log(4Lr/\delta) + ki,\ \forall i \in \{0, 1, \cdots, \ell - 1\}.$$

Furthermore, we have $T_0 \subseteq T_1 \subseteq T_2 \cdots \subseteq T_\ell$. Then, due to the Lipschitz property of the map $G(\cdot)$, the set $\hat{T}_i := G(T_i)$ will be a $2^{-i}\delta$ covering net of the set $\hat{T}_{i+1} = G(T_i), i = 0, 1, \cdots, \ell - 1$ and $\hat{T}_\ell := G(T_\ell)$ will be a $2^{-\ell}\delta$ covering net of $G(\mathcal{B})$, as a consequence,

$$\log \left|\hat{T}_i\right| \leq k \cdot \log(4Lr/\delta) + ki,\ \forall i \in \{0, 1, \cdots, \ell\}. \tag{35}$$

Now, for any $t, t' \in G(\mathcal{B})$, we have

$$\|t - t'\|_{G_m} \leq \left\|t - P_{\hat{T}_\ell}(t)\right\|_{G_m} + \left\|P_{\hat{T}_\ell}(t) - P_{\hat{T}_\ell}(t')\right\|_{G_m} + \left\|P_{\hat{T}_\ell}(t') - t'\right\|_{G_m} \tag{36}$$

Next, we will focus on bounding the term $\left\|P_{\hat{T}_\ell}(t) - P_{\hat{T}_\ell}(t')\right\|_{G_m}$ and then take limit $\ell \to \infty$ so that the first and third term vanish. Again, for simplicity of notations, starting from a point $t \in G(\mathcal{B})$ and $\hat{t}_\ell := P_{\hat{T}_\ell}(t)$, we sequentially define $\hat{t}_i = P_{\hat{T}_i}(\hat{t}_{i+1}),\ i = 0, 1, \cdots, \ell - 1$. Then, it follows for any $t, t' \in G(\mathcal{B})$,

$$\begin{aligned}
\|\hat{t}_\ell - \hat{t}'_\ell\|_{G_m} \leq & \sum_{i=1}^{\ell} \|\hat{t}_i - \hat{t}_{i-1}\|_{G_m} + \|\hat{t}_0 - \hat{t}'_0\|_{G_m} + \sum_{i=1}^{\ell} \|\hat{t}'_i - \hat{t}'_{i-1}\|_{G_m} \\
\leq & 2 \sup_{t,t' \in G(\mathcal{B})} \sum_{i=1}^{\ell} \|\hat{t}_i - \hat{t}_{i-1}\|_{G_m} + \|\hat{t}_0 - \hat{t}'_0\|_{G_m} \\
\leq & 2 \sum_{i=1}^{\ell} \sup_{t \in G(\mathcal{B})} \|\hat{t}_i - \hat{t}_{i-1}\|_{G_m} + \|\hat{t}_0 - \hat{t}'_0\|_{G_m} \\
\leq & 2 \sum_{i=1}^{\ell} \sup_{t \in \hat{T}_i} \left\|t - P_{\hat{T}_{i-1}}(t)\right\|_{G_m} + \|\hat{t}_0 - \hat{t}'_0\|_{G_m}
\end{aligned} \tag{37}$$

Note that by definition of covering nets, $\|\hat{t}_i - \hat{t}_{i-1}\|_2 \leq 2^{-i+1}\delta$. Next, we apply Lemma 9.1 to bound $\sup_{t \in G(\mathcal{B})} \|\hat{t}_i - \hat{t}_{i-1}\|_{G_m},\ \forall i$ by choosing $\Delta = \eta + 3k \cdot \log(4Lr/\delta) + 3ki$ for $\eta \geq 0$, which gives

- When $\eta + 3k \cdot \log(4Lr/\delta) + 3ki \leq \log(em)$, using the symmetrization inequality (Lemma 11.2) and taking a union bound over all $t \in \hat{T}_i$ with (35), we have for any i with probability at least $1 - \exp\left(-\eta - k \cdot \log(4Lr/\delta) - ki\right)$,

$$\sup_{t \in \hat{T}_i} \left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq C \|S_p(\mathbf{x})\|_{\psi_2} \sigma_y \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}}$$

  Taking a further union bound over all indices $i = 0, 1, 2 \cdots, \ell$ gives with probability at least

$$1 - \sum_{i=0}^{\ell} \exp\left(-\eta - k \cdot \log(4Lr/\delta) - ki\right) \geq 1 - c \cdot \exp(-\eta),$$

  for some absolute constant $c > 0$,

$$\sup_{t \in \hat{T}_i} \left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq C \|S_p(\mathbf{x})\|_{\psi_2} \sigma_y \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}},$$
$$\forall i \in \{0, 1, 2, \cdots, \ell\}, \ s.t. \ \eta + 3k \cdot \log(4Lr/\delta) + 3ki \leq \log(em)$$

- When $\eta + 3k \cdot \log(4Lr/\delta) + 3ki > \log(em)$, using the symmetrization inequality (Lemma 11.2) and taking a union bound over all $t \in \hat{T}_i$ with (35), we have with probability at least $1 - \exp\left(-\eta - k \cdot \log(4Lr/\delta) - ki\right)$, $\forall t \in \hat{T}_i$

$$\left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq C \|S_p(\mathbf{x})\|_{\psi_2} \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}}$$
$$\cdot \left(\left(\frac{1}{m} \sum_{i=1}^{m} |\tilde{y}_i|^2 \right)^{1/2} + \left(\frac{1}{m} \sum_{i=1}^{m} |\widetilde{y_i}|^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \right).$$

  Taking a further union bound over all indices $i = 0, 1, 2 \cdots, \ell$ gives with probability at least

$$1 - \sum_{i=0}^{\ell} \exp\left(-\eta - k \cdot \log(4Lr/\delta) - ki\right) \geq 1 - c \cdot \exp(-\eta),$$

  for some absolute constant $c > 0$, $\forall t \in \hat{T}_i$

$$\left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq C \|S_p(\mathbf{x})\|_{\psi_2} \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}}$$
$$\cdot \left(\left(\frac{1}{m} \sum_{i=1}^{m} |\tilde{y}_i|^2 \right)^{1/2} + \left(\frac{1}{m} \sum_{i=1}^{m} |\widetilde{y_i}|^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \right),$$
$$\forall i \in \{0, 1, 2, \cdots, \ell\}, \ s.t. \ \eta + 3k \cdot \log(4Lr/\delta) + 3ki > \log(em).$$

  By the same argument as that of last lemma using Bernstein's inequality, we reach at

$$\sup_{t \in \hat{T}_i} \left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq$$
$$C \|S_p(\mathbf{x})\|_{\psi_2} \sigma_y \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\beta} \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}},$$
$$\forall i \in \{0, 1, 2, \cdots, \ell\}, \ s.t. \ \eta + 3k \cdot \log(4Lr/\delta) + 3ki > \log(em),$$

  with probability at least $1 - e^{-\beta} - c \cdot e^{-\eta}$ for some constant $C > 0$.

Overall, we have with probability at least $1 - e^{-\beta} - e^{-\eta}$, $\beta, \eta \geq 1$,

$$\sup_{t \in \hat{T}_i} \left\| t - P_{\hat{T}_{i-1}}(t) \right\|_{G_m} \leq C \|S_p(\mathbf{x})\|_{\psi_2} \sigma_y \sqrt{\frac{1+\kappa}{\kappa}} \cdot 2^{-i+1} \delta \sqrt{\beta} \sqrt{\frac{\eta + 3k \cdot \log(4Lr/\delta) + 3ki}{m}}, \forall i \in \{0, 1, 2, \cdots, \ell\}.$$

Therefore, with the same probability,

$$\begin{aligned}\sum_{i=1}^{\ell}\sup_{t\in\hat{T}_i}\left\|t-P_{\hat{T}_{i-1}}(t)\right\|_{G_m} \leq & C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\frac{1+\kappa}{\kappa}}\cdot\delta\sqrt{\beta}\cdot\sum_{i=0}^{\ell}2^{-i+1}\sqrt{\frac{\eta+3k\cdot\log(4Lr/\delta)+3ki}{m}}\\ \leq & C'\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\frac{1+\kappa}{\kappa}}\cdot\delta\sqrt{\beta}\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}},\end{aligned}$$

for some absolute constant $C'>0$. Substituting this bound into (49) gives

$$\|\hat{t}_\ell-\hat{t}'_\ell\|_{G_m}\leq 2C'\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}\delta+\|\hat{t}_0-\hat{t}'_0\|_{G_m}.$$

To bound $\|\hat{t}_0-\hat{t}'_0\|_{G_m}$, recall that $\hat{T}_0$ is a δ-covering net of $\hat{T}_1\subseteq G(\mathcal{B})$ and by construction $\log|\hat{T}_0|\leq k\log(4Lr/\delta)$. Thus, if we choose the union of $\hat{T}_0$ and the covering chosen in Lemma 9.2, i.e. $\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B}))$, the resulting net is still a δ-covering net of $G(\mathcal{B})$ and satisfying the proposed cardinality bound. Thus, using the symmetrization inequality (Lemma 11.2) and then Lemma 9.2, we get with probability at least $1-e^{-\beta}-e^{-u}$, $\forall t_0,t'_0\in\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B}))$,

$$\sup_{t_0,t'_0\in\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B}))}\frac{\|t_0-t'_0\|_{G_m}}{\|t_0-t'_0\|_2}\leq C''\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}},$$

for some constant $C''>0$, and in particular, $\forall t,t'\in G(\mathcal{B})$, we have the corresponding $\hat{t}_0,\hat{t}'_0\in\hat{T}_0$ satisfies

$$\|\hat{t}_0-\hat{t}'_0\|_{G_m}\leq C''\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q}\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|\hat{t}_0-\hat{t}'_0\|_2,$$

Overall, we have with probability at least $1-e^{-\beta}-e^{-u}$, $\forall t,t'\in G(\mathcal{B})$,

$$\begin{aligned}\|\hat{t}_\ell-\hat{t}'_\ell\|_{G_m}\leq 2C'\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}\cdot\delta\\ +C''\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|\hat{t}_0-\hat{t}'_0\|_2.\end{aligned}$$

Since this bound simultaneously holds for any $\ell\in\mathbb{N}$, substituting it into (36) and taking the limit $\ell\to\infty$ give with probability at least $1-e^{-\beta}-e^{-u}$, the same bound holds, i.e. $\forall t,t'\in G(\mathcal{B})$,

$$\begin{aligned}\|t-t'\|_{G_m}\leq 2C'\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}\cdot\delta\\ +C''\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|\hat{t}_0-\hat{t}'_0\|_2.\end{aligned}$$

Note that by construction $\|t-\hat{t}_0\|_2\leq 2\delta$ and $\|t'-\hat{t}'_0\|_2\leq 2\delta$. Thus,

$$\|\hat{t}_0-\hat{t}'_0\|_2\leq 4\delta+\|t-t'\|_2.$$

Substituting this bound into the last probabilistic bound finishes the proof. □

10. Proof of technical lemmas in Section 5.2

10.1. Proof of Lemma 5.4

Proof. First of all, note that

$$\begin{aligned}&\left|\mathbb{E}\left[(\tilde{y}-y)\left(|\langle t,S_p(\mathbf{x})\rangle|^2-|\langle t',S_p(\mathbf{x})\rangle|^2\right)\right]\right|\\ =&\left|\mathbb{E}[(\tilde{y}-y)\langle t+t',S_p(\mathbf{x})\rangle\langle t-t',S_p(\mathbf{x})\rangle]\right|\\ \leq&\mathbb{E}\left[|y|\cdot 1_{\{|y|\geq\tau\}}\cdot|\langle t+t',S_p(\mathbf{x})\rangle\langle t-t',S_p(\mathbf{x})\rangle|\right]\end{aligned}$$

Applying Holder's inequality,

$$\begin{aligned}
&\left|\mathbb{E}\left[(\tilde{y}-y)\left(|\langle t, S_p(\mathbf{x})\rangle|^2-|\langle t', S_p(\mathbf{x})\rangle|^2\right)\right]\right| \\
\leq&\mathbb{E}\left[|y|^2\cdot|\langle t+t', S_p(\mathbf{x})\rangle\langle t-t', S_p(\mathbf{x})\rangle|^2\right]^{1/2}\cdot Pr(|y|\geq\tau)^{1/2} \\
\leq&\mathbb{E}\left[|y|^4\cdot|\langle t+t', S_p(\mathbf{x})\rangle|^4\right]^{1/4}\mathbb{E}\left[|\langle t-t', S_p(\mathbf{x})\rangle|^4\right]^{1/4}\cdot Pr(|y|\geq\tau)^{1/2} \\
\leq&\mathbb{E}\left[|y|^{4(1+\kappa)}\right]^{1/4(1+\kappa)}\mathbb{E}\left[|\langle t+t', S_p(\mathbf{x})\rangle|^{4(1+\kappa)/\kappa}\right]^{\kappa/4(1+\kappa)}\cdot\mathbb{E}\left[|\langle t-t', S_p(\mathbf{x})\rangle|^4\right]^{1/4}\cdot Pr(|y|\geq\tau)^{1/2} \\
\leq&\sqrt{\frac{4(1+\kappa)}{\kappa}}\|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}\|t-t'\|_2\cdot Pr(|y|\geq\tau)^{1/2}.
\end{aligned}$$

Furthermore,

$$Pr(|y|>\tau)\leq\frac{\mathbb{E}\left[|y|^{2(1+\kappa)}\right]}{\tau^{2(1+\kappa)}}\leq\frac{\|y\|_{L_q}^{q/2}}{\left(m^{1/2(1+\kappa)}\sigma_y\right)^{q/2}}\leq\frac{1}{m^{q/4(1+\kappa)}}\leq\frac{1}{m},$$

where the first inequality follows from Markov inequality, the third inequality follows from the fact that $\sigma_y\geq\|y\|_{L_q}$, and the last inequality follows from $q>4(1+\kappa)$. This implies the claim. □

10.2. Subgaussian concentration of a quadratic processes

The proof of Lemma 5.5 relies on the following key results which finds a "subgaussian" type concentration for a heavy-tailed quadratic proces.

Lemma 10.1. *Consider any fixed $\Delta\geq 1$. Under Assumption 3.1 and $\tilde{y}_i=\operatorname{sign}(y_i)|y_i|\wedge\tau$, where $\tau=\left(\frac{m}{B}\right)^{1/2(1+\kappa)}\sigma_y$ with $\sigma_y\geq\|y\|_{L_q}$, $\kappa\in(0,\frac{q}{4}-1)$, $B\in[1,\Delta]$ being any chosen constants, we have for any fixed $t\in T_1$, $t'\in T_2$, where $T_1,\ T_2\subseteq\mathbb{R}^d$ are bounded measurable sets, then, with probability at least $1-\exp(-\Delta)$,*

- *When $\Delta<\log(em)$,*

$$\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t\rangle\langle S_p(\mathbf{x}_i),t'\rangle\right|\leq C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2\frac{1+\kappa}{\kappa}\sqrt{\Delta m}.$$

- *When $\log(em)\leq\Delta<m$,*

$$\begin{aligned}
\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t\rangle\langle S_p(\mathbf{x}_i),t'\rangle\right|&\leq C\|S_p(\mathbf{x})\|_{\psi_2}\|t'\|_2\sqrt{\Delta}\frac{1+\kappa}{\kappa} \\
&\cdot\left(Q_m(T_1)^{1/2}+\sqrt{m}\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q}\|t\|_2+\|S_p(\mathbf{x})\|_{\psi_2}m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^m|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\|t\|_2\right).
\end{aligned}$$

- *When $\Delta\geq m$,*

$$\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t\rangle\langle S_p(\mathbf{x}_i),t'\rangle\right|\leq C\|S_p(\mathbf{x})\|_{\psi_2}\|t'\|_2\sqrt{\Delta}\left(Q_m(T_1)^{1/2}+\|S_p(\mathbf{x})\|_{\psi_2}\left(\sum_{i=1}^m|\widetilde{y}_i|^2\right)^{1/2}\|t\|_2\right).$$

where, for each of the three cases, $C>1$ is an absolute constant, and

$$Q_m(T_1):=\sup_{t\in T_1}\left|\sum_{i=1}^m\left(\tilde{y}_i|\langle S_p(\mathbf{x}_i),t\rangle|^2-\mathbb{E}\left[\tilde{y}|\langle S_p(\mathbf{x}),t\rangle|^2\right]\right)\right|. \tag{38}$$

Remark 10.1. *Note that this lemma bears some similarities with Lemma 9.1 except that we are now facing a potentially more complicated quadratic process instead of a multiplier process. The proof technique is similar.*

Proof of Lemma 10.1. We will prove the three cases, respectively.

1. Proof of the first case: Since the quantity Δ is relatively small, we will just apply Bernstein's inequality (Lemma 11.1). First of all, for any integer $p \geq 2$, the following holds,

$$\begin{aligned}
&\mathbb{E}\big[|\varepsilon\tilde{y}\langle S_p(\mathbf{x}),t\rangle\langle S_p(\mathbf{x}),t'\rangle|^p\big]\\
\leq&\mathbb{E}\Big[|\tilde{y}\langle S_p(\mathbf{x}),t\rangle|^{\frac{3}{2}p}\Big]^{2/3}\mathbb{E}\big[|\langle S_p(\mathbf{x}),t'\rangle|^{3p}\big]^{1/3}\\
\leq&\mathbb{E}\Big[|\langle S_p(\mathbf{x}),t\rangle|^{\frac{3}{2}p}\,y^3\cdot|\tilde{y}|^{\frac{3}{2}p-3}\Big]^{1/3}\mathbb{E}\big[|\langle S_p(\mathbf{x}),t'\rangle|^{3p}\big]^{1/3}\\
\leq&\mathbb{E}\Big[|\langle S_p(\mathbf{x}),t\rangle|^{\frac{3}{2}p}\,y^3\Big]^{2/3}\mathbb{E}\big[|\langle S_p(\mathbf{x}),t'\rangle|^{2p}\big]^{1/2}\cdot\tau^{p-2}\\
\leq&\tau^{p-2}\mathbb{E}\Big[|\langle S_p(\mathbf{x}),t\rangle|^{6p}\Big]^{1/6}\mathbb{E}\big[|y|^4\big]^{1/2}\mathbb{E}\big[|\langle S_p(\mathbf{x}),t'\rangle|^{2p}\big]^{1/2}\\
\leq&\|y\|_{L_q}^2\tau^{p-2}\|S_p(\mathbf{x})\|_{\psi_2}^{2p}\left(3\sqrt{2}p\right)^p\|t\|_2^p\|t'\|_2^p,
\end{aligned}$$

where the last inequality follows from the following bound: For any $p\geq 4,\ k>0$, and any $\mathbf{v}\in\mathbb{R}^d$,

$$\left(\mathbb{E}\langle S_p(\mathbf{x}),\mathbf{v}\rangle^{kp}\right)^{1/kp}\leq(kp)^{1/2}\,\|S_p(\mathbf{x})\|_{\psi_2}\|\mathbf{v}\|_2$$

Next, by Stirling's approximation, $p!\geq\sqrt{2\pi}\sqrt{p}(p/e)^p$, thus there exist some absolute constants $C',C''>0$ such that

$$\mathbb{E}\big[|\varepsilon\tilde{y}\langle S_p(\mathbf{x}),t-t'\rangle|^p\big]\leq p!\left(C'\|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2\right)^2\left(C''\tau\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2\right)^{p-2}.$$

Thus, substituting $\tau=\left(\frac{m}{B}\right)^{1/2(1+\kappa)}\sigma_y$ and by Bernstein's inequality, we have with probability at least $1-e^{-u}$,

$$\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t-t'\rangle\right|\leq C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2\left(\sqrt{\frac{u}{m}}+\frac{B^{1/2(1+\kappa)}u}{m^{1-\frac{1}{2(1+\kappa)}}}\right).$$

Now, we take $u=\Delta$, which gives

$$\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t-t'\rangle\right|\leq C\sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2\sqrt{\frac{\Delta}{m}}\left(1+\frac{B^{1/2(1+\kappa)}\sqrt{\Delta}}{m^{\frac{1}{2}-\frac{1}{2(1+\kappa)}}}\right),$$

where C is an absolute constant. Since $B\leq\Delta<\log(em)$, it follows

$$\frac{B^{1/2(1+\kappa)}\sqrt{\Delta}}{m^{\frac{1}{2}-\frac{1}{2(1+\kappa)}}}\leq\frac{\Delta}{m^{\frac{1}{2}-\frac{1}{2(1+\kappa)}}}\leq\frac{\log(em)}{m^{\frac{\kappa}{2(1+\kappa)}}}\leq\frac{1+\kappa}{\kappa},$$

which implies the first case of the lemma.

2. Proof of the second case: We start from the ordering bound (39) in the proof of Lemma 9.1, but with a slightly different decomposition. Let $\omega_i=\langle S_p(\mathbf{x}_i),t\rangle$, $\tilde{\omega}_i=\langle S_p(\mathbf{x}_i),t'\rangle$. Set the index I to be the union of the p largest entries of $\{\omega_i\}_{i=1}^m$, the p largest entries of $\{\tilde{\omega}_i\}_{i=1}^m$, and the p largest entries of $\{\tilde{y}_i\}_{i=1}^m$, where p is a number to be chosen later. Then, it follows

with probability at least $1-e^{-u^2/2}$,

$$
\begin{aligned}
&\left|\sum_{i=1}^m \varepsilon_i \tilde{y}_i \omega_i \tilde{\omega}_i\right| \\
\leq & \sum_{i\in I}|\tilde{y}_i\omega_i\tilde{\omega}_i| + u\left(\sum_{i\notin I}|\tilde{y}_i\omega_i\tilde{\omega}_i|^2\right)^{1/2} \\
\leq & \left(\sum_{i\in I}|\tilde{y}_i\omega_i|^2\right)^{1/2}\left(\sum_{i\in I}|\tilde{\omega}_i|^2\right)^{1/2} + u\left(\sum_{i\notin I}|\tilde{y}_i|^{2r}\right)^{1/2r}\left(\sum_{i\notin I}|\omega_i\tilde{\omega}_i|^{2r'}\right)^{1/2r'} \\
\leq & \left(\sum_{i\in I}|\tilde{y}_i\omega_i|^2\right)^{1/2}\left(\sum_{i\in I}|\tilde{\omega}_i|^2\right)^{1/2} + u\left(\sum_{i\notin I}|\tilde{y}_i|^{2r}\right)^{1/2r}\left(\sum_{i\notin I}|\omega_i|^{4r'}\right)^{1/4r'}\left(\sum_{i\notin I}|\tilde{\omega}_i|^{4r'}\right)^{1/4r'} \\
\leq & 3\left(\sum_{i\leq p}|(\tilde{y}_i\omega_i)^*|^2\right)^{1/2}\left(\sum_{i\leq p}|\tilde{\omega}_i^*|^2\right)^{1/2} + u\left(\sum_{i>p}|\tilde{y}_i^*|^{2r}\right)^{1/2r}\left(\sum_{i>p}|(\omega_i)^*|^{4r'}\right)^{1/4r'}\left(\sum_{i>p}|(\tilde{\omega}_i)^*|^{4r'}\right)^{1/4r'} \\
\leq & 3\left(Q_m(T_1)^{1/2} + \sqrt{m}\mathbb{E}\big[|\tilde{y}_i\omega_i|^2\big]^{1/2}\right)\left(\sum_{i\leq p}|\tilde{\omega}_i^*|^2\right)^{1/2} + u\left(\sum_{i=1}^m|\tilde{y}_i|^{2r}\right)^{1/2r}\left(\sum_{i>p}|(\omega_i)^*|^{4r'}\right)^{1/4r'}\left(\sum_{i>p}|(\tilde{\omega}_i)^*|^{4r'}\right)^{1/4r'},
\end{aligned}
\tag{39}
$$

where the first inequality follows from Lemma 11.3, the second and third inequality both follow from Holder's inequality, and the last inequality follows from the definition of $Q_m(T_1)$ in (38) that

$$
\begin{aligned}
\left(\sum_{i\leq p}|(\tilde{y}_i\omega_i)^*|^2\right)^{1/2} \leq \left(\sum_{i=1}^m|(\tilde{y}_i\omega_i)|^2\right)^{1/2} &\leq \left(\sum_{i=1}^m\big(|(\tilde{y}_i\omega_i)|^2 - \mathbb{E}\big[|(\tilde{y}_i\omega_i)|^2\big]\big)\right)^{1/2} + \big(m\mathbb{E}\big[|(\tilde{y}_i\omega_i)|^2\big]\big)^{1/2} \\
&\leq Q_m(T_1)^{1/2} + \sqrt{m}\mathbb{E}\big[|\tilde{y}_i\omega_i|^2\big]^{1/2}.
\end{aligned}
$$

Note that

$$
\mathbb{E}\big[|\tilde{y}_i\omega_i|^2\big]^{1/2} \leq \mathbb{E}\big[y_i^4\big]^{1/4}\mathbb{E}\big[\omega_i^4\big]^{1/4} \leq \|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}\|t\|_2,
$$

it remains to bound $\left(\sum_{i\leq p}|\tilde{\omega}_i^*|^2\right)^{1/2}$ and $\left(\sum_{i>p}|\omega_i^*|^{4r'}\right)^{1/4r'}\left(\sum_{i>p}|\tilde{\omega}_i^*|^{4r'}\right)^{1/4r'}$.

Again choosing $p = \lfloor \Delta/\log(em/\Delta)\rfloor$, and choose $r = 1+\kappa$, $r' = (1+\kappa)/\kappa$. Then, for the term $\left(\sum_{i\leq p}|\omega_i^*|^2\right)^{1/2}$ in (39), by the same argument leading to (26), one can easily show that

$$
\mathbb{P}\left(\left(\sum_{i=1}^k|\tilde{\omega}_i^*|^2\right)^{1/2} \geq C\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'\|_2\sqrt{\Delta}\right) \leq 2\exp(-\Delta) \tag{40}
$$

For the terms $\left(\sum_{i>p}|\omega_i^*|^{4r'}\right)^{1/4r'}$ and $\left(\sum_{i>p}|\tilde{\omega}_i^*|^{4r'}\right)^{1/4r'}$, by assumption, we have $S_p(\mathbf{x}_i)$ is subgaussian, thus, for any index $j\in\{1,2,\cdots,m\}$,

$$
\mathbb{P}\left(|\omega_j| - \mathbb{E}[|\omega_j|] \geq C\beta\|S_p(\mathbf{x})\|_{\psi_2}\|t\|_2\right) \leq e^{-\beta^2},
$$

for any $\beta\geq 0$, where $C>0$ is a positive constant. For the rest of the proof in this case, the index i will be dedicated for non-increasing ordering index, i.e. $\{\omega_i^*\}_{i=1}^m$. For each index i, we then choose

$$
\beta = c_\kappa(m/i)^{\kappa/8(1+\kappa)}, \;\; c_\kappa := \sqrt{\frac{10(1+\kappa)}{\kappa}},
$$

and following the same argument leading to (30) gives with probability $1-\exp(-\Delta)$

$$\left(\sum_{i>p}|\omega_i^*|^{4r'}\right)^{1/4r'} \leq C'\sqrt{\frac{1+\kappa}{\kappa}}\|S_p(\mathbf{x})\|_{\psi_2}\|t\|_2 m^{\kappa/4(1+\kappa)}.$$

Substituting the above two bounds along with (40) into (39) gives with probability $1-\exp(-\Delta)-\exp(-u^2)$,

$$\left|\sum_{i=1}^m \varepsilon_i\tilde{y}_i\omega_i\tilde{\omega}_i\right| \leq C\left(Q_m(T_1)^{1/2}+\sqrt{m}\|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}\|t\|_2\right)\|S_p(\mathbf{x}_i)\|_{\psi_2}\|t'\|_2\sqrt{\Delta}$$
$$+u\left(\sum_{i=1}^m|\tilde{y}_i|^{2r}\right)^{1/2r}\frac{1+\kappa}{\kappa}\|S_p(\mathbf{x})\|_{\psi_2}^2\|t\|_2\|t'\|_2 m^{\kappa/2(1+\kappa)}.$$

Taking $u=\sqrt{\Delta}$ finishes the second the part of the lemma.

3. Proof of the third case: By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left|\sum_{i=1}^m \varepsilon_i\tilde{y}_i\langle S_p(\mathbf{x}_i),t\rangle\langle S_p(\mathbf{x}_i),t'\rangle\right| \\
&\leq\left(\sum_{i=1}^m\tilde{y}_i^2\omega_i^2\right)^{1/2}\left(\sum_{i=1}^m\tilde{\omega}_i^2\right)^{1/2} \\
&\leq\left(\left|\sum_{i=1}^m\left(\tilde{y}_i^2\omega_i^2-\mathbb{E}\big[\tilde{y}_i^2\omega_i^2\big]\right)\right|^{1/2}+\sqrt{m}\mathbb{E}\big[\tilde{y}_i^2\omega_i^2\big]^{1/2}\right)\left(\sum_{i=1}^m\tilde{\omega}_i^2\right)^{1/2} \\
&\leq\left(Q_m(T_1)^{1/2}+\|y\|_{L_q}\|S_p(\mathbf{x})\|_{\psi_2}\right)\left(\sum_{i=1}^m\tilde{\omega}_i^2\right)^{1/2}
\end{aligned}$$

where $\omega_i=\langle S_p(\mathbf{x}),t\rangle,\ \tilde{\omega}_i=\langle S_p(\mathbf{x}),t'\rangle$. Thus, by the fact that $\tilde{\omega}_i^2$ is sub-exponential with

$$\|\tilde{\omega}_i^2\|_{\psi_1}\leq 2\|\tilde{\omega}_i\|_{\psi_2}^2\leq 2\|S_p(\mathbf{x}_i)\|_{\psi_2}^2\|t'\|_2^2,$$

we use Bernstein's inequality again,

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m\left(\omega_i^2-\mathbb{E}\big[\omega_i^2\big]\right)\right|\geq C\|S_p(\mathbf{x})\|_{\psi_2}^2\|t'\|_2^2\left(\sqrt{\frac{2u}{m}}+\frac{u}{m}\right)\right)\leq 2\exp(-u).$$

We take $u=\Delta$. Using the fact that $\Delta>m$, we have the term $\frac{\Delta}{m}$ dominates the right hand side. Furthermore, $\mathbb{E}\big[\tilde{\omega}_j^2\big]\leq\|S_p(\mathbf{x})\|_{\psi_2}^2\|t'\|_2^2\leq\frac{\Delta}{m}\|S_p(\mathbf{x})\|_{\psi_2}^2\|t'\|_2^2$. Thus,

$$\mathbb{P}\left(\left(\frac{1}{m}\sum_{i=1}^m\tilde{\omega}_i^2\right)^{1/2}\geq C\|S_p(\mathbf{x})\|_{\psi_2}\|t'\|_2\sqrt{\frac{\Delta}{m}}\right)\leq 2\exp(-\Delta),$$

for some absolute constant $C>0$, which implies the third part of the lemma. □

10.3. Uniform concentration over a δ-covering net

We will now use Lemma 10.1 to establish a uniform concentration result of the following quadratic process

$$\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{y}_i\left(|\langle t,S_p(\mathbf{x}_i)\rangle|^2-|\langle t',S_p(\mathbf{x}_i)\rangle|^2\right)\right|$$

over a δ-covering net of the set $G(\mathcal{B}^k(r))\cap\mathcal{S}^{d-1}$. Again, we will use $\mathcal{B}$ to denote $\mathcal{B}^k(r)$ in this section for simplicity.

Lemma 10.2. *For any $\delta \in (0,1)$, let $\mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$ be a δ-covering net of the set $G(\mathcal{B}) \cap \mathcal{S}^{d-1}$ with respect to the norm $\|\cdot\|_2$, such that*

$$\log |\mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})| \leq 2k \log(4Lr/\delta),$$

where $\mathcal{B}$ is the ball in $\mathbb{R}^k$ with radius r and $c > 0$ is an absolute constant. Suppose $m \geq k \log(Lr)$. Under Assumption 3.1 and $\tilde{y}_i = \operatorname{sign}(y_i)|y_i| \wedge \tau$, where $\tau = \left(\frac{m}{k \log(Lr)}\right)^{1/2(1+\kappa)} \sigma_y$, for any $\eta, \beta \geq 1$, with probability at least $1 - e^{-\beta} - e^{-\eta}$, for any $t, t' \in \mathcal{N}(\delta, G(\mathcal{B}))$,

$$\left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \tilde{y}_i \left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2 \right) \right| \leq C \left(\|S_p(\mathbf{x})\|_{\psi_2}^2 + \|S_p(\mathbf{x})\|_{\psi_2} \right) \sigma_y \sqrt{\beta} \frac{1+\kappa}{\kappa} \sqrt{\frac{\eta + k \log(4Lr/\delta)}{m}} \|t - t'\|_2.$$

Proof of Lemma 10.2. First of all, by (31), there exists a net such that the cardinality

$$\mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1}) \leq \mathcal{N}(\delta, G(\mathcal{B})) \leq k \log(Lr/\delta).$$

Next we set $\Delta = \eta + 4k \cdot \log(4Lr/\delta) \geq k \log(Lr)$ since $\delta < 1$, and by Lemma 10.1,

- When $\eta + 4k \cdot \log(4Lr/\delta) \leq \log(em)$, take a union bound over all $t, t' \in \mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$, which implies with probability at least $1 - e^{-\eta}$, $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$,

$$\begin{aligned} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \tilde{y}_i \left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2 \right) \right| &= \left| \sum_{i=1}^m \varepsilon_i \tilde{y}_i \langle t + t', S_p(\mathbf{x}_i)\rangle \langle t - t', S_p(\mathbf{x}_i)\rangle \right| \\ &\leq C \left(\|S_p(\mathbf{x})\|_{\psi_2}^2 + \|S_p(\mathbf{x})\|_{\psi_2} \right) \sigma_y \sqrt{\frac{\eta + k \log(4Lr/\delta)}{m}} \|t - t'\|_2. \end{aligned}$$

- When $\eta + 4k \cdot \log(4Lr/\delta) > \log(em)$, we have $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$, with probability at least $1 - \exp(-\Delta)$,

$$\begin{aligned} &\left| \sum_{i=1}^m \varepsilon_i \tilde{y}_i \left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2 \right) \right| = \left| \sum_{i=1}^m \varepsilon_i \tilde{y}_i \langle t + t', S_p(\mathbf{x}_i)\rangle \langle t - t', S_p(\mathbf{x}_i)\rangle \right| \\ \leq & C \|S_p(\mathbf{x})\|_{\psi_2} \|t - t'\|_2 \sqrt{\Delta} \frac{1+\kappa}{\kappa} \\ & \cdot \left(Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2} + \sqrt{m} \|S_p(\mathbf{x})\|_{\psi_2} \|y\|_{L_q} \|t + t'\|_2 + \|S_p(\mathbf{x})\|_{\psi_2} m^{\frac{\kappa}{2(1+\kappa)}} \left(\sum_{i=1}^m |\widetilde{y}_i|^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \|t + t'\|_2 \right) \\ \leq & \sqrt{2} C \|S_p(\mathbf{x})\|_{\psi_2} \|t - t'\|_2 \sqrt{\Delta} \frac{1+\kappa}{\kappa} \\ & \cdot \left(Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2} + \sqrt{m} \|S_p(\mathbf{x})\|_{\psi_2} \|y\|_{L_q} + \|S_p(\mathbf{x})\|_{\psi_2} m^{\frac{\kappa}{2(1+\kappa)}} \left(\sum_{i=1}^m |\widetilde{y}_i|^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \right) \end{aligned}$$

where

$$\begin{aligned} Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2} &= \sup_{t,t' \in \mathcal{N}(G(\mathcal{B}) \cap \mathcal{S}^{d-1}, \delta)} \left| \sum_{i=1}^m \left(\tilde{y}_i |\langle S_p(\mathbf{x}_i), t + t'\rangle|^2 - \mathbb{E}\left[\tilde{y} |\langle S_p(\mathbf{x}), t + t'\rangle|^2 \right] \right) \right|^{1/2} \\ &\leq 2 \sup_{t \in \mathcal{N}(G(\mathcal{B}) \cap \mathcal{S}^{d-1}, \delta)} \left| \sum_{i=1}^m \left(\tilde{y}_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 - \mathbb{E}\left[\tilde{y} |\langle S_p(\mathbf{x}), t\rangle|^2 \right] \right) \right|^{1/2} \\ &\leq 2 \sup_{t \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}} \left| \sum_{i=1}^m \left(\tilde{y}_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 - \mathbb{E}\left[\tilde{y} |\langle S_p(\mathbf{x}), t\rangle|^2 \right] \right) \right|^{1/2} \end{aligned} \tag{41}$$

Take a union bound over all $t, t' \in \mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$, and use $\Delta = \eta + 4k \cdot \log(4Lr/\delta)$, we have with probability at least $1 - \exp(-\eta)$, $\forall t, t' \in \mathcal{N}(\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$,

$$\left|\sum_{i=1}^m \varepsilon_i \tilde{y}_i \left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right| \leq \sqrt{2}C\|S_p(\mathbf{x})\|_{\psi_2}\|t - t'\|_2 \frac{1+\kappa}{\kappa}\sqrt{\eta + 4k \cdot \log(4Lr/\delta)}$$
$$\cdot \left(Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2} + \sqrt{m}\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q} + \|S_p(\mathbf{x})\|_{\psi_2} m^{\frac{\kappa}{2(1+\kappa)}} \left(\sum_{i=1}^m |\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right) \quad (42)$$

It remains to bound the two terms $\left(\sum_{i=1}^m |\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}$ and $Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2}$, respectively. First, by Bernstein's inequality, we have

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m |\tilde{y}_i|^{2(1+\kappa)} - \mathbb{E}\left[|\tilde{y}|^{2(1+\kappa)}\right] \geq C\left(\sqrt{\frac{\mathbb{E}\left[|\tilde{y}|^{4(1+\kappa)}\right]\beta}{m}} + \frac{\tau^{2(1+\kappa)}\beta}{m}\right)\right) \leq e^{-\beta}, \ \ \beta \geq 0.$$

Recall that $\tau = \left(\frac{m}{k\log(Lr)}\right)^{1/2(1+\kappa)}\sigma_y$. Thus,

$$\left(\frac{1}{m}\sum_{i=1}^m |\tilde{y}_i|^{2(1+\kappa)}\right)^{1/2(1+\kappa)} \leq C\sigma_y\left(1 + \left(\frac{k\log(Lr)}{m}\right)^{1/4(1+\kappa)}\beta^{1/4(1+\kappa)} + \beta^{1/2(1+\kappa)}\right) \leq 3C\sigma_y\beta^{1/2(1+\kappa)}, \quad (43)$$

with probability at least $1 - e^{-\beta}$, where the last inequality follows from $m \geq k\log(Lr)$. Next, for the term $Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2}$, recall (41) and substitute in the bound in Lemma 10.3 with $u = \sqrt{\beta}$, we have with probability at least $1 - \exp(-\beta)$,

$$Q_m(G(\mathcal{B}) \cap \mathcal{S}^{d-1})^{1/2} \leq C\sigma_y\left(\|S_p(\mathbf{x})\|_{\psi_2} + \|S_p(\mathbf{x})\|_{\psi_2}^{1/2}\right)\sqrt{m\beta}, \quad (44)$$

where we use the fact that $m \geq k\log(Lr)$. Finally, substitute (44) and (43) into (42) and using the fact that $\|y\|_{L_q} \leq \sigma_y$ finish the proof for this case.

Overall, we finish the proof. □

10.4. Proof of Lemma 5.5

Proof. For the rest of the proof, for any $t, t' \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}$, where $\mathcal{B} = \mathcal{B}^k(r)$, we will denote

$$H_m(t, t') := \left|\frac{1}{m}\sum_{i=1}^m \tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right) - \mathbb{E}\left[\tilde{y}\left(|\langle t, S_p(\mathbf{x})\rangle|^2 - |\langle t', S_p(\mathbf{x})\rangle|^2\right)\right]\right|.$$

For any $\delta > 0$, we consider a sequence of covering nets $\left\{\hat{T}_i\right\}_{i=0}^{\ell}$ of the set $G(\mathcal{B}) \cap \mathcal{S}^{d-1}$ with respect to $\|\cdot\|_2$, where $\hat{T}_i$ denotes the $2^{-i}\delta$-covering of the set $\hat{T}_{i+1}$, $i = 0, 1, \cdots, \ell - 1$ and $\hat{T}_\ell = \mathcal{N}(2^{-\ell}\delta, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$ is the minimum $2^{-\ell}\delta$-covering of the set $G(\mathcal{B}) \cap \mathcal{S}^{d-1}$. We have $\widehat{T}_0 \subseteq \widehat{T}_1 \subseteq \cdots \subseteq \widehat{T}_\ell$. By (35), we know that

$$\log\left|\hat{T}_i\right| \leq k \cdot \log(4Lr/\delta) + ki, \ \forall i \in \{0, 1, \cdots, \ell\}.$$

For any point $t \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}$, define $\hat{t}_\ell := P_{\hat{T}_\ell}(t)$, and $\hat{t}_i = P_{\hat{T}_i}(\hat{t}_{i+1})$, $i = 0, 1, \cdots, \ell - 1$, where for any set $T \in \mathbb{R}^d$, $P_T(\cdot)$ denotes the projection on to this set. Then, it follows

$$H_m(t, t') \leq H_m(t, \hat{t}_\ell) + H_m(\hat{t}_\ell, \hat{t}'_\ell) + H_m(t', \hat{t}'_\ell) \quad (45)$$

using the same argument leading to (49), it follows for any $t, t' \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}$,

$$H_m(\hat{t}_\ell, \hat{t}'_\ell) \le 2\sum_{i=1}^{\ell} \sup_{t\in\hat{T}_i} H_m\big(t,\ P_{\hat{T}_{i-1}}(t)\big) + H_m\big(\hat{t}_0,\ \hat{t}'_0\big) \tag{46}$$

Next, we aim to bound $\sup_{t\in\hat{T}_i} H_m\big(t,\ P_{\hat{T}_{i-1}}(t)\big),\ \forall i$. First, by symmetrization inequality, it is enough to bound

$$\sup_{t\in\hat{T}_i}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon\tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right|.$$

We apply Lemma 10.1 by choosing $\Delta = \eta + 3k\cdot\log(4Lr/\delta) + 3ki$ for $\eta \ge 0$,

- When $\eta + 3k\cdot\log(4Lr/\delta) + 3ki \le \log(em)$, take a union bound over all $t \in \hat{T}_i$, we have for any i with probability at least $1 - \exp\left(-\eta - k\cdot\log(4Lr/\delta) - ki\right)$,

$$\sup_{t\in\hat{T}_i}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon\tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right| = \sup_{t\in\hat{T}_i}\left|\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\langle t+t', S_p(\mathbf{x}_i)\rangle\langle t-t', S_p(\mathbf{x}_i)\rangle\right|$$
$$\le C\left(\|S_p(\mathbf{x})\|_{\psi_2}^2 + \|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\sqrt{\frac{\eta + 3k\cdot\log(4Lr/\delta) + 3ki}{m}}2^{-i+1}\delta.$$

  Taking a further union bound over all indices $i = 0, 1, 2\cdots, \ell$ gives with probability at least

$$1 - \sum_{i=0}^{\ell}\exp\left(-\eta - k\cdot\log(4Lr/\delta) - ki\right) \ge 1 - c\cdot\exp(-\eta),$$

  for some absolute constant $c > 0$,

$$\sup_{t\in\hat{T}_i}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon\tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right|$$
$$\le C\left(\|S_p(\mathbf{x})\|_{\psi_2}^2 + \|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\sqrt{\beta}\sqrt{\frac{\eta + 3k\cdot\log(4Lr/\delta) + 3ki}{m}}2^{-i+1}\delta,$$
$$\forall i \in \{0, 1, 2, \cdots, \ell\},\ s.t.\ \eta + 3k\cdot\log(4Lr/\delta) + 3ki \le \log(em).$$

- When $\eta + 3k\cdot\log(4Lr/\delta) + 3ki > \log(em)$, take a union bound over all $t \in \hat{T}_i$, we have for any i with probability at least $1 - \exp\left(-\eta - k\cdot\log(4Lr/\delta) - ki\right)$,

$$\sup_{t\in\hat{T}_i}\left|\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right|$$
$$= \sup_{t\in\hat{T}_i}\left|\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\langle t+t', S_p(\mathbf{x}_i)\rangle\langle t-t', S_p(\mathbf{x}_i)\rangle\right| \le \sqrt{2}C\|S_p(\mathbf{x})\|_{\psi_2}\frac{1+\kappa}{\kappa}\sqrt{\eta + 4k\cdot\log(4Lr/\delta)}$$
$$\cdot\left(Q_m(G(\mathcal{B})\cap\mathcal{S}^{d-1})^{1/2} + \sqrt{m}\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q} + \|S_p(\mathbf{x})\|_{\psi_2}m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^{m}|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right)\cdot 2^{-i+1}\delta.$$

  Taking a further union bound over all indices $i = 0, 1, 2\cdots, \ell$ gives with probability at least $1 - c\cdot\exp(-\eta)$, for some absolute constant $c > 0$,

$$\sup_{t\in\hat{T}_i}\left|\sum_{i=1}^{m}\varepsilon_i\tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right)\right| \le \sqrt{2}C\|S_p(\mathbf{x})\|_{\psi_2}\frac{1+\kappa}{\kappa}\sqrt{\eta + 4k\cdot\log(4Lr/\delta)}$$
$$\cdot\left(Q_m(G(\mathcal{B})\cap\mathcal{S}^{d-1})^{1/2} + \sqrt{m}\|S_p(\mathbf{x})\|_{\psi_2}\|y\|_{L_q} + \|S_p(\mathbf{x})\|_{\psi_2}m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^{m}|\widetilde{y}_i|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right)\cdot 2^{-i+1}\delta,$$
$$\forall i \in \{0, 1, 2, \cdots, \ell\},\ s.t.\ \eta + 3k\cdot\log(4Lr/\delta) + 3ki > \log(em).$$

Combining with bounds (43), (44) and rearranging terms delivers

$$\sup_{t\in\hat{T}_i}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon\tilde{y}_i\left(|\langle t,S_p(\mathbf{x}_i)\rangle|^2-|\langle t',S_p(\mathbf{x}_i)\rangle|^2\right)\right| \leq C\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\sqrt{\frac{\eta+3k\cdot\log(4Lr/\delta)+3ki}{m}}2^{-i+1}\delta, \quad \forall i\in\{0,1,2,\cdots,\ell\},\ s.t.\ \eta+3k\cdot\log(4Lr/\delta)+3ki\leq\log(em).$$

with probability at least $1-c\cdot\exp(-\eta)-\exp(-\beta)$.

Overall, with probability at least $1-\exp(-\eta)-\exp(-\beta)$,

$$\sum_{i=1}^{\ell}\sup_{t\in\hat{T}_i}H_m\big(t,\ P_{\hat{T}_{i-1}}(t)\big)\leq C\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\cdot\delta\cdot\sum_{i=0}^{\ell}2^{-i+1}\sqrt{\frac{\eta+3k\cdot\log(4Lr/\delta)+3ki}{m}}$$
$$\leq C'\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\cdot\delta\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}},$$

for some absolute constant $C'>0$. Substituting this bound into (46),

$$H_m(\hat{t}_\ell,\hat{t}'_\ell)\leq C'\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\cdot\delta\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}+H_m\big(\hat{t}_0,\ \hat{t}'_0\big), \tag{47}$$

with probability at least $1-\exp(-\eta)-\exp(-\beta)$, for some absolute constant $C>0$.

To bound $H_m\big(\hat{t}_0,\ \hat{t}'_0\big)$, recall that $\hat{T}_0$ is a δ-covering net of $\hat{T}_1\subseteq G(\mathcal{B})\cap\mathcal{S}^{d-1}$ and by construction $\log|\hat{T}_0|\leq k\log(4Lr/\delta)$. Thus, if we choose the union of $\hat{T}_0$ and the covering chosen in Lemma 10.2, i.e. $\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B})\cap\mathcal{S}^{d-1})$, the resulting net is still a δ-covering net of $G(\mathcal{B})\cap\mathcal{S}^{d-1}$ and satisfying the proposed cardinality bound. Thus, using the symmetrization inequality (Lemma 11.2), and then Lemma 10.2, we get with probability at least $1-e^{-\beta}-e^{-u}$, $\forall t_0,t'_0\in\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B})\cap\mathcal{S}^{d-1})$,

$$\sup_{t_0,t'_0\in\hat{T}_0\cup\mathcal{N}(\delta,G(\mathcal{B})\cap\mathcal{S}^{d-1})}\frac{H_m\big(\hat{t}_0,\ \hat{t}'_0\big)}{\|t_0-t'_0\|_2}\leq C''\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\sqrt{\beta}\frac{1+\kappa}{\kappa}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}},$$

for some constant $C''>0$. Substitute this bound into (47), we have with probability at least $1-e^{-\beta}-e^{-u}$, $\forall t,t'\in G(\mathcal{B})\cap\mathcal{S}^{d-1}$,

$$H_m(\hat{t}_\ell,\ \hat{t}'_\ell)\leq 2C'\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\cdot\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}\delta$$
$$+C''\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\sqrt{\beta}\frac{1+\kappa}{\kappa}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|\hat{t}_0-\hat{t}'_0\|_2.$$

Since this bound simultaneously holds for any $\ell\in\mathbb{N}$, substituting it into (45) and taking the limit $\ell\to\infty$ give with probability at least $1-e^{-\beta}-e^{-u}$, the same bound holds, i.e. $\forall t,t'\in G(\mathcal{B})\cap\mathcal{S}^{d-1}$,

$$H_m(\hat{t},\ \hat{t}')\leq 2C'\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\frac{1+\kappa}{\kappa}\sqrt{\beta}\cdot\sqrt{\frac{\eta+k\cdot(\log(4Lr/\delta)+1)}{m}}\delta$$
$$+C''\left(\|S_p(\mathbf{x})\|_{\psi_2}^2+\|S_p(\mathbf{x})\|_{\psi_2}\right)\sigma_y\sqrt{\beta}\frac{1+\kappa}{\kappa}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}\|\hat{t}_0-\hat{t}'_0\|_2.$$

Note that by construction $\|t-\hat{t}_0\|_2\leq 2\delta$ and $\|t'-\hat{t}'_0\|_2\leq 2\delta$. Thus,

$$\|\hat{t}_0-\hat{t}'_0\|_2\leq 4\delta+\|t-t'\|_2.$$

Substituting this bound into the last probabilistic bound finishes the proof. □

10.5. Bounds on quadratic processes

In this section, we prove two supporting lemmas on the bound of quadratic processes used in previous sections.

Lemma 10.3. *Suppose $m \geq k\log(Lr)$. Under Assumption 3.1 and $\tilde{y}_i = \mathrm{sign}(y_i)|y_i| \wedge \tau$, where $\tau = \left(\frac{m}{k\log(Lr)}\right)^{1/2(1+\kappa)}\sigma_y$, for any $u \geq 1$, with probability at least $1-e^{-u^2}$,*

$$\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\left(\tilde{y}_i|\langle S_p(\mathbf{x}_i),t\rangle|^2 - \mathbb{E}\left[\tilde{y}|\langle S_p(\mathbf{x}),t\rangle|^2\right]\right)\right|$$
$$\leq C\sigma_y^2\left(\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{mk\log(Lr)} + \|S_p(\mathbf{x})\|_{\psi_2}m + \|S_p(\mathbf{x})\|_{\psi_2}u\frac{m}{\sqrt{k\log(Lr)}} + \|S_p(\mathbf{x})\|_{\psi_2}^2u^2\sqrt{\frac{m}{k\log(Lr)}}\right),$$

where $C>0$ is an absolute constant.

Proof of Lemma 10.3. First of all, by symmetrization inequality (Lemma 11.2), it is enough to consider the following supremum:

$$\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i|\langle S_p(\mathbf{x}_i),t\rangle|^2\right|.$$

By contraction principle (Lemma 11.4), for any $v>0$,

$$Pr\left(\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i|\langle S_p(\mathbf{x}_i),t\rangle|^2\right| \geq \left(\frac{m}{k\log(Lr)}\right)^{1/2(1+\kappa)}\sigma_y v\right)$$
$$\leq 2Pr\left(\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\varepsilon_i|\langle S_p(\mathbf{x}_i),t\rangle|^2\right| \geq v\right).$$

Thus, it is enough to bound the right hand side of the above inequality. We apply symmetrization inequality again and Lemma 11.5, which gives with probability at least $1-e^{-u^2}$, for $u\geq 1$,

$$\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\varepsilon_i|\langle S_p(\mathbf{x}_i),t\rangle|^2\right| \leq C_1\left(\|S_p(\mathbf{x})\|_{\psi_2}\gamma_2(G(\mathcal{B})\cap\mathcal{S}^{d-1})\sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2\gamma_2(G(\mathcal{B})\cap\mathcal{S}^{d-1})^2\right)$$
$$+ C_2\left(\|S_p(\mathbf{x})\|_{\psi_2}u\sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2u^2\right),$$

where we use the fact that for any specific t, $|\langle S_p(\mathbf{x}_i),t\rangle|^2$ is a subexponential random variable with $\sigma \leq \|S_p(\mathbf{x})\|_{\psi_2}$ and $K \leq \|S_p(\mathbf{x})\|_{\psi_2}^2$, and the radius $\Delta(G(\mathcal{B})\cap\mathcal{S}^{d-1}) \leq 1$. Next, apply the bound in Lemma 11.6, which gives $\gamma_2(G(\mathcal{B})\cap\mathcal{S}^{d-1}) \leq C\sqrt{k\log(Lr)}$, and we have with proability at least $1-e^{-u^2}$, for $u\geq 1$, ,

$$\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}}\left|\sum_{i=1}^m\varepsilon_i\tilde{y}_i|\langle S_p(\mathbf{x}_i),t\rangle|^2\right|$$
$$\leq C'\left(\frac{m}{k\log(Lr)}\right)^{1/2(1+\kappa)}\sigma_y\left(\left(\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{mk\log(Lr)} + \|S_p(\mathbf{x})\|_{\psi_2}^2k\log(Lr)\right)\right.$$
$$\left.+\left(\|S_p(\mathbf{x})\|_{\psi_2}u\sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2u^2\right)\right).$$

Rearranging the terms yields the claim of the lemma. □

Lemma 10.4. *Suppose $m \geq kn\log(2d)$. Under Assumption 3.1 and $\tilde{y}_i = \mathrm{sign}(y_i)|y_i| \wedge \tau$, where $\tau = \left(\frac{m}{kn\log(2d)}\right)^{1/2(1+\kappa)}\sigma_y$, for any $u\geq 1$, with probability at least $1-e^{-u^2}$,*

$$\max_{\mathcal{E}^{2k}} Q_m\left(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}(2)\right)$$
$$\leq C\sigma_y^2\left(\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{mkn\log(2d)} + \|S_p(\mathbf{x})\|_{\psi_2}m + u\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{mkn\log(2d)} + \frac{u^2\|S_p(\mathbf{x})\|_{\psi_2}^2kn\log(2d)}{\sqrt{m}}\right),$$

where $C > 0$ is an absolute constant and the maximum is taken over all possible $(2d)^{2kn}$ different $2k$-subspaces $\mathcal{E}^{2k} \in \mathbb{R}^d$, $\mathcal{S}^{d-1}(2)$ is the sphere in $\mathbb{R}^d$ with radius 2, and $Q_m(\cdot)$ is defined in (38).

Proof of Lemma 10.4. Consider first any specific $2k$-subspaces $\mathcal{E}^{2k} \in \mathbb{R}^d$. Similar to the previous proof, by symmetrization inequality (Lemma 11.2), to bound $Q_m\big(\mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)\big)$, it is enough to consider the following supremum:

$$\sup_{t \in \mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)} \left| \sum_{i=1}^m \varepsilon_i \tilde{y}_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 \right|.$$

By contraction principle (Lemma 11.4), for any $v > 0$,

$$Pr\left(\sup_{t \in \mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)} \left| \sum_{i=1}^m \varepsilon_i \tilde{y}_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 \right| \geq \left(\frac{m}{kn \log(2d)} \right)^{1/2(1+\kappa)} \sigma_y v \right) \leq 2Pr\left(\sup_{t \in \mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)} \left| \sum_{i=1}^m \varepsilon_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 \right| \geq v \right).$$

We apply symmetrization inequality again and Lemma 11.5, which gives with probability at least $1 - e^{-\tilde{u}^2}$, for $\tilde{u} \geq 1$,

$$\sup_{t \in \mathcal{E}^{2k} \cap \mathcal{S}^{d-1}} \left| \sum_{i=1}^m \varepsilon_i |\langle S_p(\mathbf{x}_i), t\rangle|^2 \right| \leq C_1 \left(\|S_p(\mathbf{x})\|_{\psi_2} \gamma_2(\mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)) \sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2 \gamma_2(\mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2))^2 \right) + C_2 \left(\|S_p(\mathbf{x})\|_{\psi_2} \tilde{u} \sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2 \tilde{u}^2 \right),$$

with the bound $\gamma_2(\mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)) \leq 2\sqrt{2k}$. Next, take $\tilde{u} = u + \sqrt{2kn \log(2d)}$, and take a union bound over all $(2d)^{2kn}$ different $2k$-subspaces gives with probability at least $1 - e^{-u^2}$,

$$\max_{\mathcal{E}^{2k}} \; Q_m\big(\mathcal{E}^{2k} \cap \mathcal{S}^{d-1}(2)\big) \leq C_1 \left(\|S_p(\mathbf{x})\|_{\psi_2} \sqrt{mk} + \|S_p(\mathbf{x})\|_{\psi_2}^2 k \right) + C_2 \left(\|S_p(\mathbf{x})\|_{\psi_2} (u + \sqrt{2kn \log(2d)}) \sqrt{m} + \|S_p(\mathbf{x})\|_{\psi_2}^2 (u^2 + 2kn \log(2d)) \right),$$

which implies the claim. □

11. Some probability and linear algebra tools

11.1. Some probability bounds

We recall the following well-known concentration inequality,

Lemma 11.1 (Bernstein's inequality)**.** *Let $X_1, \cdots, X_m$ be a sequence of independent centered random variables. Assume that there exist positive constants f and D such that for all integers $p \geq 2$*

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}[|X_i|^p] \leq \frac{p!}{2} f^2 D^{p-2},$$

then

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^m X_i \right| \geq \frac{f}{\sqrt{m}} \sqrt{2u} + \frac{D}{m} u \right) \leq 2 \exp(-u).$$

In particular, if $X_1, \cdots, X_m$ are all sub-exponential random variables, then f and D can be chosen as $f = \frac{1}{m} \sum_{i=1}^m \|X_i\|_{\psi_1}$ and $D = \max_{i=1\ldots m} \|X_i\|_{\psi_1}$.

The following version of Symmetrization inequality can be found, for example, in (Van Der Vaart & Wellner, 1996).

Lemma 11.2 (Symmetrization inequality). *Let $\{Z_t(i)\}_{i=1}^m$ be i.i.d. copies of a mean 0 stochastic process $\{Z_t : t \in T\}$. For every $1 \le i \le m$, let $g_t(i) : T \to \mathbb{R}$ be an arbitrary function. Let $\{\varepsilon_i\}_{i=1}^m$ be a sequence of independent Rademacher random variables. Then, for every $x > 0$,*

$$\left(1 - \frac{4m}{x^2} \sup_{t\in T} var(Z_t)\right) \cdot Pr\left(\sup_{t\in T}\left|\sum_{i=1}^m Z_t(i)\right| > x\right) \le 2Pr\left(\sup_{t\in T}\left|\sum_{i=1}^m \varepsilon_i(Z_t(i) - g_t(i))\right| > \frac{x}{4}\right),$$

where $var(Z_t) = \mathbb{E}\big[(Z_t - \mathbb{E}[Z_t])^2\big]$.

The following is a classical bound on the Rademacher process:

Lemma 11.3 ((Montgomery-Smith, 1990)). *Let $\mathbf{X} = [X_1, \cdots, X_m]$ be a sequence of scalars. Define the following quantity:*

$$K_{1,2}(\mathbf{X}, u) := \inf\left\{\sum_{i\in I}|X_i| + u\left(\sum_{i\notin I}|X_i|^2\right)^{1/2}, \ I \subseteq \{1, 2, \cdots, m\}\right\}.$$

Then, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^m \varepsilon_i X_i\right| \ge K_{1,2}(\mathbf{X}, u)\right) \le 2\exp(-u^2/2). \tag{48}$$

Furthermore, there exists a universal constant $c > 0$ such that

$$c^{-1}K_{1,2}(\mathbf{X}, u) \le \sum_{i=1}^{\lfloor u^2\rfloor} X_i^* + u\left(\sum_{i=\lfloor u^2\rfloor+1}^m (X_i^*)^2\right)^{1/2} \le cK_{1,2}(\mathbf{X}, u)$$

where $\{X_i^*\}_{i=1}^m$ is the non-increasing rearrangement of $\{|X_i|\}_{i=1}^m$ and $\{\varepsilon_i\}_{i=1}^m$ is a sequence of i.i.d. Rademancher random variables independent of $\{X_i\}_{i=1}^m$.

The following version of contraction principle is a direct generalization of Theorem 4.4 of (Ledoux & Talagrand, 2013). The proof is the same replacing norm in Banach space by a semi-norm in $\mathbb{R}^d$.

Lemma 11.4 (Contraction principle). *For any sequence $\{\mathbf{x}_i\}_{i=1}^m$ in $\mathbb{R}^d$, any semi-norm $\|\cdot\|$ and any real numbers $\{\alpha_i\}_{i=1}^m$ such that $|\alpha_i| \le 1$,*

$$Pr\left(\left\|\sum_{i=1}^m \alpha_i\varepsilon_i\mathbf{x}_i\right\| > u\right) \le 2Pr\left(\left\|\sum_{i=1}^m \varepsilon_i\mathbf{x}_i\right\| > u\right),$$

where $u > 0$ is any constant and $\{\varepsilon_i\}_{i=1}^m$ is a sequence of independent Rademacher random variables.

The following lemma bounds the size of the quadratic process in terms of the Gaussian mean width of a set:

Lemma 11.5 ((Dirksen et al., 2015)). *Let T be any measurable set in $\mathbb{R}^d$ and define the Gaussian mean width of T as*

$$\gamma_2(T) := \mathbb{E}\left[\sup_{t\in T}\langle \mathbf{g}, t\rangle\right],$$

where $g \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$. Let

$$A_t := \frac{1}{m}\sum_{i=1}^m |\langle X_i, t\rangle|^2 - \mathbb{E}\big[|\langle X_i, t\rangle|^2\big],$$

where $\{X_i\}_{i=1}^m$ are i.i.d. subgaussian vectors. Let σ, K be constants satisfying $q \ge 2$,

$$\sup_{t\in T}\frac{1}{m}\sum_{i=1}^m \mathbb{E}\Big[\big||\langle X_i, t\rangle|^2 - \mathbb{E}\big[|\langle X_i, t\rangle|^2\big]\big|^q\Big] \le \frac{q!}{2}\sigma^2K^{q-2},$$

and let $\|X\|_{\psi_2} := \max_{i=1,2,\cdots,m} \|X_i\|_{\psi_2}$. *Then, with probability at least* $1 - e^{-u^2}$, *for* $u \geq 1$,

$$\sup_{t\in T} |A_t| \leq C_1 \left(\Delta(T)\|X\|_{\psi_2} \frac{\gamma_2(T)}{\sqrt{m}} + \|X\|_{\psi_2}^2 \frac{\gamma_2(T)^2}{m} \right) + C_2 \left(\frac{\sigma u}{\sqrt{m}} + \frac{K u^2}{m} \right),$$

where $C_1, C_2 \geq 1$ *are absolute constants and* $\Delta(T) := \sup_{t,s\in T} \|t - s\|_2$.

Finally, the following lemma bounds the Gaussian mean width of the set $G(\mathcal{B}^k(r))$:

Lemma 11.6. *Let* $\mathcal{B}^k(r)$ *be the ball of radius* r *in* $\mathbb{R}^k$, *and let* $G : \mathbb{R}^k \to \mathbb{R}^d$ *be an* L*-Lipschitz function, then,*

$$\gamma_2(G(\mathcal{B}^k(r)) \cap \mathcal{S}^{d-1}) \leq C\sqrt{k \log(4Lr)},$$

where $C > 0$ *is an absolute constant.*

Proof of Lemma 11.6. We will use $\mathcal{B}$ to denote $\mathcal{B}^k(r)$. The approach we will take is Dudley's chaining technique. Let $\varepsilon = 2^{-\ell}$ for some positive integer ℓ. Let $\widehat{T}$ be an ε-cover of $G(\mathcal{B}) \cap \mathcal{S}^{d-1}$ with Euclidean distance $\|\cdot\|_2$. Then, we have for any $t \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}$,

$$\langle g, t\rangle = \langle g, t - \widehat{t}\rangle + \langle g, \widehat{t}\rangle \leq 2 \sup_{\|t-\widehat{t}\|_2 \leq \varepsilon} \langle g, t - \widehat{t}\rangle + \sup_{\widehat{t}\in\widehat{T}} \langle g, \widehat{t}\rangle,$$

where in the first equality, we pick $\widehat{t} \in \widehat{T}$ such that $\|\widehat{t} - t\|_2 \leq \varepsilon$. Thus, it follows

$$\mathbb{E}\left[\sup_{t\in G(\mathcal{B})\cap\mathcal{S}^{d-1}} \langle g, t\rangle \right] \leq \mathbb{E}\left[\sup_{\|t-\widehat{t}\|_2 \leq \varepsilon} \langle g, t - \widehat{t}\rangle \right] + \mathbb{E}\left[\sup_{\widehat{t}\in\widehat{T}} \langle g, \widehat{t}\rangle \right] \tag{49}$$

To bound the second term in (49), consider a sequence of progressively better approximations of $\widehat{T}$ as follows: Let $\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_\ell > 0$ such that $\varepsilon_i = 2^{-i}$ and $\widehat{T}_\ell = \widehat{T}$. For any $i \geq 1$, let $\widehat{T}_{i-1}$ be the minimum ε_{i-1} cover of $\widehat{T}_i$. We have $\widehat{T}_0 \subseteq \widehat{T}_1 \subseteq \cdots \subseteq \widehat{T}_\ell$. Now for any $\widehat{t}, \widehat{t}' \in \widehat{T}$,

$$\langle g, \widehat{t}\rangle = \langle g, \widehat{t}_\ell - \widehat{t}'_\ell\rangle = \sum_{i=1}^{\ell} \langle g, \widehat{t}_i - \widehat{t}_{i-1}\rangle + \langle g, \widehat{t}_0\rangle,$$

where $\widehat{t}_i \in \widehat{T}_i$. For any $i \geq 1$, we choose $\widehat{t}_{i-1} = f_{i-1}(\widehat{t}_i)$, where f_{i-1} maps any point in $\widehat{T}_i$ to its nearest point in $\widehat{T}_{i-1}$. Then, it follows,

$$\begin{aligned}
\mathbb{E}\left[\sup_{\widehat{t},\,\widehat{t}'\in\widehat{T}} \langle g, \widehat{t} - \widehat{t}'\rangle \right] &\leq \sum_{i=1}^{\ell} \mathbb{E}\left[\sup_{\widehat{t}_i\in\widehat{T}_i} \langle g, \widehat{t}_i - f_{i-1}(\widehat{t}_i)\rangle \right] + \mathbb{E}\left[\sup_{\widehat{t}_0\in\widehat{T}_0} \langle g, \widehat{t}_0\rangle \right] \\
&\leq \sum_{i=1}^{\ell} 2^{-(i-1)} \sqrt{2\log\left|\widehat{T}_i\right|} + 2\sqrt{2\log\left|\widehat{T}_0\right|} \\
&\leq \sum_{i=1}^{\ell} 2^{-(i-1)} \sqrt{2\log \mathcal{N}(2^{-i}, G(\mathcal{B}))} + 2\sqrt{2\log \mathcal{N}(1, G(\mathcal{B}))}, \qquad (50)
\end{aligned}$$

where for any $\varepsilon > 0$, $\mathcal{N}(\varepsilon, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$ denotes the minimum ε-covering number of the set $G(\mathcal{B})$, and the second from the last inequality follows from the standard Gaussian maximum estimate that for a set of pairs H from $(T, \|\cdot\|_2)$ with $T \subseteq \mathbb{R}^d$,

$$\mathbb{E}\left[\max_{t\in H} \langle g, t\rangle \right] \leq \max_{t\in H} \|t\|_2 \sqrt{\log|H|}.$$

It remains to bound $\mathcal{N}(\varepsilon, G(\mathcal{B}) \cap \mathcal{S}^{d-1})$ for some $\varepsilon \in (0, 1)$. Let $\mathcal{N}(\varepsilon/L, \mathcal{B})$ be the ε/L-covering number of the set $\mathcal{B} \subseteq \mathbb{R}^k$. It is known that there exists such a net M_ε that

$$\log|M_\varepsilon| \leq k \cdot \log(4Lr/\varepsilon).$$

Then, due to the L-Lipschitz property of G, the set $G(M_\varepsilon)$ forms a ε-net of $G(\mathcal{B})$ with the same cardinality bound. As a consequence, for any i, we have

$$\mathcal{N}(2^{-i}, G(\mathcal{B}) \cap \mathcal{S}^{d-1}) \leq \mathcal{N}(2^{-i}, G(\mathcal{B})) \leq ik + k\log(4Lr).$$

Substituting this bound into (50) gives

$$\mathbb{E}\left[\sup_{\widehat{t}} \langle g, \widehat{t}\rangle\right] \leq \sum_{i=1}^{\ell} 2^{-(i-1)}\sqrt{2ik + 2k\log(4Lr)} + 2\sqrt{2k\log(4Lr)} \leq C\sqrt{k\log(4Lr)},$$

for some positive constant C. This finishes bounding the second term in (49). Finally, since the bound (49) holds for any ε, taking $\ell \to \infty$ gives $\varepsilon \to 0$ in (49) and the first term will go to 0. This implies

$$\mathbb{E}\left[\sup_{t \in G(\mathcal{B}) \cap \mathcal{S}^{d-1}} \langle g, t\rangle\right] \leq C\sqrt{k\log(4Lr)},$$

finishing the proof. □

11.2. Some linear algebra inequalities

Lemma 11.7 (Lemma A.1.2 of (Vu & Lei, 2012))**.** *Let* $\mathbf{x},\ \mathbf{y} \in \mathcal{S}^{d-1}$*, then*

$$\|\mathbf{x}\mathbf{x}^T - \mathbf{y}\mathbf{y}^T\|_F^2 \leq 2\|\mathbf{x} - \mathbf{y}\|_2^2.$$

If, in addition, $\|\mathbf{x} - \mathbf{y}\|_2 \leq \sqrt{2}$*, then,*

$$\|\mathbf{x}\mathbf{x}^T - \mathbf{y}\mathbf{y}^T\|_F^2 \geq \|\mathbf{x} - \mathbf{y}\|_2^2$$