Structure of This Supplementary Document

This supplementary document contains the technical proofs for the theoretical results of the AISTATS-2021 paper entitled "**Stability and Risk Bounds of Iterative Hard Thresholding**". The content of this document is organized as follows:

- In Appendix A we collect a number of auxiliary lemmas that will be used in our analysis.
- In Appendix B, we present the technical proofs of main results in Section 2.
- In Appendix C, we present the technical proofs of main results in Section 3.

A Some Auxiliary Lemmas

This section is devoted to presenting a set of preliminary results that are useful in the proof of our main results.

Generalization bounds for uniformly stable algorithms. To prove the stability implied risk bounds, we need the following lemma from Bousquet et al. (2020, Corollary 8) which gives a near-tight high probability generalization error bound for uniformly stable learning algorithms.

Lemma 1 (Generalization bound implied by uniform stability). *Let $A : \mathcal{X}^n \mapsto \mathcal{W}$ be a learning algorithm that has uniform stability γ with respect to a loss function $\ell(\cdot;\cdot) \leq M$. Then for any $\delta \in (0,1)$, the following generalization bound holds with probability at least $1-\delta$ over S:*

$$\left|\mathbb{E}_\xi\left[\ell(A(S);\xi)\right] - \frac{1}{n}\sum_{i=1}^n \ell(A(S),\xi_i)\right| \leq \mathcal{O}\left(\gamma \log(n)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

RIP-condition-free convergence rate of IHT. The rate of convergence and parameter estimation error of IHT have been extensively analyzed under RIP (or restricted strong condition number) bounding conditions (Bahmani et al., 2013; Yuan et al., 2014). The RIP-type conditions, however, are unrealistic in many applications. To remedy this deficiency, sparsity-level relaxation strategy was considered in Jain et al. (2014); Yuan et al. (2018) with which the high-dimensional estimation consistency of IHT can be established under arbitrary restricted strong condition number. In order to make our analysis more realistic for high-dimensional problems, we choose to work on the following RIP-condition-free convergence rate bound, which is essentially from Jain et al. (2014), for IHT invoking on the empirical risk F_S.

Lemma 2 (Convergence rate of IHT). *Assume that F_S is L_{3k}-smooth and μ_{3k}-strongly convex. Consider $\bar{k}$ such that $k \geq \frac{32L_{3k}^2}{\mu_{3k}^2}\bar{k}$. Let $\bar{w}_{S,k} = \arg\min_{\|w\|_0 \leq \bar{k}} F_S(w)$. Set $\eta = \frac{2}{3L_{3k}}$. Then for any $\epsilon > 0$, IHT outputs $w_{S,k}^{(t)}$ satisfying $F_S(w_{S,k}^{(t)}) \leq F_S(\bar{w}_{S,k}) + \epsilon$, after*

$$t \geq \mathcal{O}\left(\frac{L_{3k}}{\mu_{3k}}\log\left(\frac{F_S(w_{S,k}^{(0)})}{\epsilon}\right)\right)$$

rounds of iteration.

Localized Rademacher Complexities and data dependent risk bounds. Let us define $\|\ell(w;\cdot) - \ell(w';\cdot)\|_\infty := \max_{\xi\in\mathcal{X}} |\ell(w;\xi) - \ell(w';\xi)|$. We further introduce following concept of Localized Rademacher Complexity which plays an important role in deriving the fast rates of convergence:

$$R_S(r_n; w^*) := \mathbb{E}_\varepsilon\left[\sup_{\|\ell(w;\cdot)-\ell(w^*;\cdot)\|_\infty \leq r_n}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i[\ell(w;\xi_i) - \ell(w^*;\xi_i)]\right|\right],$$

where $r_n > 0$ and w^* are fixed and $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables, i.e., symmetric Bernoulli random variables taking values $+1$ and -1 with probability $1/2$ each. $R_S(\delta; w^*)$ can be used as a data dependent complexity measure of the target parametric class around w^* that allows one to estimate the accuracy of approximation of $F(w) - F(w^*)$ by $F_S(w) - F_S(w^*)$ based on the data. The following localized concentration bound is elementary and it can implied immediately by the symmetrization and McDiarmid's inequalities (see, e.g., Koltchinskii, 2006).

Lemma 3 (Data dependent local concentration bound)**.** *For any fixed w^* and all $\delta \in (0,1)$, the following bound holds with probability at least $1-\delta$:*

$$\sup_{\|\ell(w;\cdot)-\ell(w^*;\cdot)\|_\infty \le r_n} |F_S(w) - F_S(w^*) - (F(w) - F(w^*))| \le 2R_S(r_n; w^*) + 3r_n\sqrt{\frac{2\log(2/\delta)}{n}}.$$

The following elementary lemma (see, e.g., Yuan et al., 2018, Lemma 14) is useful in our analysis.

Lemma 4. *Assume that f is μ_s-strongly convex. Then for any w, w' such that $\|w-w'\|_0 \le s$ and $f(w) \le f(w')+\epsilon$ for some $\epsilon \ge 0$, the following bound holds*

$$\|w - w'\| \le \frac{2\sqrt{s}\|\nabla f(w')\|_\infty}{\mu_s} + \sqrt{\frac{2\epsilon}{\mu_s}},$$

where $I = supp(w)$ and $I' = supp(w')$.

B Proofs of the Results in Section 2

In this section, we present the technical proofs of the main results stated in Section 2.

B.1 Proof of Theorem 1

In this subsection, we present a detailed proof of Theorem 1.

A key lemma. For a given index set $J \subseteq [p]$, let us consider the following restrictive estimator over J:

$$w_{S|J} = \mathop{\arg\min}_{w\in\mathcal{W}, \mathrm{supp}(w)\subseteq J} F_S(w). \tag{A.1}$$

We present the following lemma about the uniform generalization gap of $w_{S|J}$ for all J with $|J| = k$ which is crucial to our proof.

Lemma 5. *Assume that the loss function ℓ is smooth and G-Lipschitz continuous with respect to its first argument and $\ell(\cdot;\xi) \le M$ for all ξ. Suppose that F_S is μ_k-strongly convex with probability at least $1-\delta'_n$ over the random draw of S. Let $\mathcal{J} = \{J \subseteq [p] : |J| = k\}$ be the set of index set of cardinality k. Then for any $\delta \in (0, 1-\delta'_n)$ and $\lambda > 0$, it holds with probability at least $1-\delta-\delta'_n$ over the random draw of S that*

$$\sup_{J\subseteq\mathcal{J}} \left|F(w_{S|J}) - F_S(w_{S|J})\right| \le \mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\left(\log\left(\frac{1}{\delta}\right) + k\log\left(\frac{ep}{k}\right)\right) + M\sqrt{\frac{\log(1/\delta)+k\log(ep/k)}{n}} + \frac{\lambda G\sqrt{M}}{\mu_k\sqrt{\mu_k}}\right).$$

Proof. Let us consider the following defined ℓ_2-regularized ℓ_0-ERM estimator for any given $\lambda > 0$:

$$w_{\lambda,S|J} := \mathop{\arg\min}_{w\in\mathcal{W}, \mathrm{supp}(w)\subseteq J} \left\{F_{\lambda,S}(w) := F_S(w) + \frac{\lambda}{2}\|w\|^2\right\}.$$

The reason for introducing the additional ℓ_2-regularization term is to guarantee uniform stability of the hypothetical estimator $w_{\lambda,S|J}$. Based on the standard proof arguments (see, e.g., Shalev-Shwartz et al., 2009) we can show that the optimal model $w_{\lambda,S|J}$ has uniform stability $\gamma = \frac{4G^2}{\lambda n}$. Indeed, let $S^{(i)}$ be a sample set that is identical to S except that one of the ξ_i is replaced by another random sample ξ'_i. Then we can derive that

$$\begin{aligned}
&F_{\lambda,S}(w_{\lambda,S^{(i)}|J}) - F_{\lambda,S}(w_{\lambda,S|J})\\
=&\frac{1}{n}\sum_{j\neq i}\left(\ell(w_{\lambda,S^{(i)}|J};\xi_j) - \ell(w_{\lambda,S|J};\xi_j)\right) + \frac{1}{n}\left(\ell(w_{\lambda,S^{(i)}|J};\xi_i) - \ell(w_{\lambda,S|J};\xi_i)\right) + \frac{\lambda}{2}\|w_{\lambda,S^{(i)}|J}\|^2 - \frac{\lambda}{2}\|w_{\lambda,S|J}\|^2\\
=&F_{\lambda,S^{(i)}}(w_{\lambda,S^{(i)}|J}) - F_{\lambda,S^{(i)}}(w_{\lambda,S|J}) + \frac{1}{n}\left(\ell(w_{\lambda,S^{(i)}|J};\xi_i) - \ell(w_{\lambda,S|J};\xi_i)\right) - \frac{1}{n}\left(\ell(w_{\lambda,S^{(i)}|J};\xi'_i) - \ell(w_{\lambda,S|J};\xi'_i)\right)\\
\le&\frac{1}{n}\left|\ell(w_{\lambda,S^{(i)}|J};\xi_i) - \ell(w_{\lambda,S|J};\xi_i)\right| + \frac{1}{n}\left|\ell(w_{\lambda,S^{(i)}|J};\xi'_i) - \ell(w_{\lambda,S|J};\xi'_i)\right|\\
\le&\frac{2G}{n}\|w_{\lambda,S^{(i)}|J} - w_{\lambda,S|J}\|,
\end{aligned}$$

where we have used the optimality of $w_{\lambda,S^{(i)}|J}$ with respect to $F_{\lambda,S^{(i)}}(w)$ and the Lipschitz continuity of the loss function $\ell(w;\xi)$. Since $F_{\lambda,S}$ is λ-strongly convex and $w_{\lambda,S|J}$ is optimal for $F_{\lambda,S}(w)$ over the supporting set J, we have

$$F_{\lambda,S}(w_{\lambda,S^{(i)}|J}) \geq F_{\lambda,S}(w_{\lambda,S|J}) + \frac{\lambda}{2}\|w_{\lambda,S^{(i)}|J} - w_{\lambda,S|J}\|^2.$$

By combing the preceding two inequalities we arrive at $\|w_{\lambda,S^{(i)}|J} - w_{\lambda,S|J}\| \leq \frac{4G}{\lambda n}$. Consequently from the Lipschitz continuity of ℓ we have that for any sample ξ

$$|\ell(w_{\lambda,S^{(i)}|J};\xi) - \ell(w_{\lambda,S|J};\xi)| \leq G\|w^{(i)}_{\lambda,S|J} - w_{\lambda,S|J}\| \leq \frac{4G^2}{\lambda n}.$$

This confirms that the optimal model $w_{\lambda,S|J}$ has uniform stability $\gamma = \frac{4G^2}{\lambda n}$. By invoking Lemma 1 we obtain that with probability at least $1-\delta$ over random draw of S,

$$\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| \leq \mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{A.2}$$

Let $\mathcal{J} = \{J \subseteq [p] : |J| = k\}$ be the set of index set of cardinality k. It is standard to verify $|\mathcal{J}| = \binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$ (Rigollet, 2015, Lemma 2.7). Then for each $J \in \mathcal{J}$, based on (A.2) we must have that with probability at least $1 - \frac{\delta}{|\mathcal{J}|}$ over S, the following generalization gap is valid for any $\lambda > 0$:

$$\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| \leq \mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\log\left(\frac{|\mathcal{J}|}{\delta}\right) + M\sqrt{\frac{\log(|\mathcal{J}|/\delta)}{n}}\right).$$

Then by union probability we obtain that the following bound holds with probability at least $1-\delta$,

$$\sup_{J\subseteq\mathcal{J}}\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| \leq \mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\left(\log\left(\frac{1}{\delta}\right) + k\log\left(\frac{ep}{k}\right)\right) + M\sqrt{\frac{\log(1/\delta) + k\log(ep/k)}{n}}\right). \tag{A.3}$$

Next, we show how to bound the estimator difference $\sup_{J\subseteq\mathcal{J}}\|w_{S|J} - w_{\lambda,S|J}\|$. The strong convexity assumption of F_S implies that the following bound holds with probability at least $1-\delta'_n$ over S for all $J \subseteq \mathcal{J}$:

$$\lambda\|w_{S|J}\| = \|\nabla_J F_{\lambda,S}(w_{S|J}) - \nabla_J F_{\lambda,S}(w_{\lambda,S|J})\| \geq (\mu_k + \lambda)\|w_{S|J} - w_{\lambda,S|J}\|,$$

where the notation $\nabla_J g$ denotes the restriction of gradient ∇g over J and we have used the optimality of $w_{\lambda,S|J}$ and $w_{S|J}$ over J which implies that

$$\nabla_J F_{\lambda,S}(w_{\lambda,S|J}) = 0, \quad \nabla_J F_{\lambda,S}(w_{S|J}) = \nabla_J F_S(w_{S|J}) + \lambda w_{S|J} = \lambda w_{S|J}.$$

In the meanwhile, since $\ell(0;\cdot) \in (0, M)$, we must have the following bound holds with probability at least $1-\delta'_n$ over S for all $J \subseteq \mathcal{J}$:

$$M \geq F_S(0) \geq F_S(0) - F_S(w_{S|J}) \geq \frac{\mu_k}{2}\|w_{S|J}\|^2,$$

which leads to $\|w_{S|J}\| \leq \sqrt{2M/\mu_k}$. Then it follows readily from the previous two inequalities that

$$\|w_{S|J} - w_{\lambda,S|J}\| \leq \frac{\lambda}{\mu_k + \lambda}\|w_{S|J}\| \leq \frac{\lambda\sqrt{2M}}{\sqrt{\mu_k}(\mu_k + \lambda)} \leq \frac{\lambda\sqrt{2M}}{\mu_k\sqrt{\mu_k}}.$$

Since the loss function is G-Lipschitz continuous, the following is then valid with probability at least $1-\delta'_n$ over the random draw of S for all $J \subseteq \mathcal{J}$:

$$\begin{aligned}
&\left|F(w_{S|J}) - F_S(w_{S|J})\right| \\
\leq &\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| + |F_S(w_{S|J}) - F_S(w_{\lambda,S|J})| + |F(w_{S|J}) - F(w_{\lambda,S|J})| \\
\leq &\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| + 2G\|w_{S|J} - w_{\lambda,S|J}\| \\
\leq &\left|F(w_{\lambda,S|J}) - F_S(w_{\lambda,S|J})\right| + \frac{2\lambda G\sqrt{2M}}{\mu_k\sqrt{\mu_k}}.
\end{aligned}$$

In view of the above bound and the bound in (A.3), with probability at least $1-\delta-\delta'_n$ over S we have

$$\sup_{J\subseteq\mathcal{J}}\left|F(w_{S|J})-F_S(w_{S|J})\right|\leq\mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\left(\log\left(\frac{1}{\delta}\right)+k\log\left(\frac{ep}{k}\right)\right)+M\sqrt{\frac{\log(1/\delta)+k\log(ep/k)}{n}}+\frac{\lambda G\sqrt{M}}{\mu_k\sqrt{\mu_k}}\right).$$

The proof is concluded. □

Now we are in the position to prove Theorem 1.

Proof of Theorem 1. Let $\mathcal{J}=\{J\subseteq[p]:|J|=k\}$ be the set of index set of cardinality k. For any random sample set S, by the definition of $\tilde{w}^{(t)}_{S,k}$ we always have $\tilde{w}^{(t)}_{S,k}\in\{w_{S|J}:J\in\mathcal{J}\}$. Applying Lemma 5 yields that with probability at least $1-\delta-\delta'_n$,

$$\left|F(\tilde{w}^{(t)}_{S,k})-F_S(\tilde{w}^{(t)}_{S,k})\right|\leq\mathcal{O}\left(\frac{G^2}{\lambda n}\log(n)\left(\log\left(\frac{1}{\delta}\right)+k\log\left(\frac{ep}{k}\right)\right)+M\sqrt{\frac{\log(1/\delta)+k\log(ep/k)}{n}}+\frac{\lambda G\sqrt{M}}{\mu_k\sqrt{\mu_k}}\right).$$

Setting $\lambda=\sqrt{\frac{G\mu_k^{1.5}\log(n)(\log(1/\delta)+k\log(ep/k))}{nM^{0.5}}}$ in the above and preserving leading terms yields

$$\left|F(\tilde{w}^{(t)}_{S,k})-F_S(\tilde{w}^{(t)}_{S,k})\right|\leq\mathcal{O}\left(\frac{G^{3/2}M^{1/4}}{\mu_k^{3/4}}\sqrt{\frac{\log(n)(\log(1/\delta)+k\log(ep/k))}{n}}\right). \tag{A.4}$$

For any $\epsilon>0$, given that $t=\mathcal{O}\left(\frac{L_{3k}}{\mu_{3k}}\log\left(\frac{F_S(w^{(0)}_{S,k})}{\epsilon}\right)\right)=\mathcal{O}\left(\frac{L_{3k}}{\mu_{3k}}\log\left(\frac{M}{\epsilon}\right)\right)$ is sufficiently large, we can bound the sparse excess risk $F(\tilde{w}^{(t)}_{S,k})-F(\bar{w})$ as

$$\begin{aligned}F(\tilde{w}^{(t)}_{S,k})-F(\bar{w})=&F(\tilde{w}^{(t)}_{S,k})-F_S(\tilde{w}^{(t)}_{S,k})+F_S(\tilde{w}^{(t)}_{S,k})-F_S(\bar{w})+F_S(\bar{w})-F(\bar{w})\\ \leq&\left|F(\tilde{w}^{(t)}_{S,k})-F_S(\tilde{w}^{(t)}_{S,k})\right|+|F_S(\bar{w})-F(\bar{w})|+\epsilon,\end{aligned}$$

where in the last inequality we have used the bound $F_S(\tilde{w}^{(t)}_{S,k})\leq F_S(w^{(t)}_{S,k})\leq F_S(\bar{w})+\epsilon$ which is implied by the definition of $\tilde{w}^{(t)}_{S,k}$ and Lemma 2. Since $\ell(\bar{w};\xi)\leq M$, from Hoeffding's inequality we know that with probability at least $1-\delta/2$,

$$|F_S(\bar{w})-F(\bar{w})|\leq\mathcal{O}\left(M\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Based on the generalization gap bound (A.4) and by union probability we get with probability at least $1-\delta$

$$\begin{aligned}&F(\tilde{w}^{(t)}_{S,k})-F(\bar{w})\\ \leq&\left|F(\tilde{w}^{(t)}_{S,k})-F_S(\tilde{w}^{(t)}_{S,k})\right|+|F_S(\bar{w})-F(\bar{w})|+\epsilon\\ \leq&\mathcal{O}\left(\frac{G^{3/2}M^{1/4}}{\mu_k^{3/4}}\sqrt{\frac{\log(n)(\log(1/\delta)+k\log(ep/k))}{n}}+M\sqrt{\frac{\log(1/\delta)}{n}}+\epsilon\right).\end{aligned}$$

Setting $\epsilon=\mathcal{O}(\sqrt{k\log(n)\log(ep/k)/n})$ yields the desired bound (keep in mind the monotonicity of restricted smoothness and strong convexity). This completes the proof. □

B.2 Proof of Corollary 1

In this subsection we prove Corollary 1 which is an application of Theorem 1 to sparse logistic regression models. We first present the following lemma, which follows immediately from Agarwal et al. (2012, Lemma 6), to be used for proving the main result.

Lemma 6. *Suppose x_i are drawn i.i.d. from a zero-mean sub-Gaussian distribution with covariance matrix $\Sigma \succ 0$. Let $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$. Assume that $\Sigma_{jj} \leq \sigma^2$. Then there exist universal positive constants c_0 and c_1 such that for all $w \in \mathbb{R}^p$*

$$\frac{\|X^\top w\|^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}w\|^2 - c_1\frac{\sigma^2\log(p)}{n}\|w\|_1^2$$

holds with probability at least $1-\exp\{-c_0 n\}$.

Proof of Corollary 1. Given that $\|x_i\| \leq 1$, we have $\ell(w;\xi_i)$ is L-smooth with $L \leq 4s(2y_i w^\top x_i)(1-s(2y_i w^\top x_i)) \leq 1$. Since $\|w\| \leq R$, we must have $|y_i w^\top x_i| \leq R$ and thus the logistic loss $\ell(w;\xi_i) = \log(1+\exp(-2y_i w^\top x_i))$ satisfies $\ell(w;\xi_i) \leq \mathcal{O}(R)$ and $[\Lambda(w)]_{ii} = 4s(2y_i w^\top x_i)(1-s(2y_i w^\top x_i)) \geq \frac{4}{(1+\exp(2R))^2} \geq \frac{1}{\exp(4R)}$. It follows that

$$\nabla^2 F_S(w) = \frac{1}{n}X\Lambda(w)X^\top \succeq \frac{1}{n\exp(4R)}XX^\top = \frac{1}{\exp(4R)}\Sigma.$$

In view of Lemma 6 and the fact $\|w\|_1 \leq \sqrt{k}\|w\|$ when $\|w\|_0 \leq k$ we can verify that with probability at least $1-\exp\{-c_0 n\}$, $F_S(w)$ is μ_{4k}-strongly convex with

$$\mu_{4k} = \frac{1}{\exp(4R)}\left(\frac{1}{2}\lambda_{\min}(\Sigma) - \frac{kc_1\log(p)}{n}\right).$$

Provided that $n \geq \frac{4kc_1\log(p)}{\lambda_{\min}(\Sigma)}$, we have $\mu_{4k} \geq \frac{\lambda_{\min}(\Sigma)}{4\exp(4R)}$ holds with probability at least $1-\exp\{-c_0 n\}$. By invoking Theorem 1, after sufficiently large $T \geq \mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}(\Sigma)}\log\left(\frac{nR}{k\log(n)\log(p/k)}\right)\right)$ rounds of IHT iteration, with probability at least $1-\delta-\exp\{-c_0 n\}$ the sparse excess risk of IHT converges at the rate of

$$\mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}^{3/4}(\Sigma)}\sqrt{\frac{\log(n)(\log(1/\delta)+k\log(p/k))}{n}} + R\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

This completes the proof. □

C Proofs of the Results in Section 3

In this section, we present the technical proofs of the main results stated in Section 3.

C.1 Proof of Theorem 2

In this subsection, we prove Theorem 2. For any fixed $J \subseteq [p]$ with $|J| = k$, let

$$w_J^* = \operatorname*{arg\,min}_{\mathrm{supp}(w)\subseteq J} F(w).$$

Before proving the main result, we first establish a key lemma which shows a uniform fast rate of $w_{S|J}$ (recall the definition in (A.1)) towards w_J^* for all J if the population risk is restricted strongly convex and the loss is Lipschitz continuous. To ease notation, we define an abbreviation of loss function as $\ell_w(\cdot) := \ell(w;\cdot)$. Particularly, we write the localized Rademacher complexity restricted over J at w_J^* as:

$$R_{S|J}(r_n; w_J^*) := \mathbb{E}_\varepsilon\left[\sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\|_\infty \leq r_n}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i[\ell_w(\xi_i) - \ell_{w_J^*}(\xi_i)]\right|\right],$$

where $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables, i.e., symmetric Bernoulli random variables taking values $+1$ and -1 with probability $1/2$ each. The following preliminary result is standard yet useful in our analysis. We provide its proof for the sake of completeness.

Lemma 7. *Under Assumption 4, there exists some absolute constant $C > 0$ such that*

$$R_{S|J}(Gr_n; w_J^*) \leq CGr_n\sqrt{\frac{k}{n}}\log^{3/2}\left(\frac{1}{r_n}\sqrt{\frac{n}{k}}\right).$$

Proof. Let us restrict the analysis over $\mathcal{W}_J$ as a restriction of $\mathcal{W}$ over J. Since $\mathcal{W}$ is assumed to be a subset of unit ℓ_2-sphere, it is standard (see, for instance, Böröczky and Wintsche, 2003) to bound the covering number of $\mathcal{W}_J$ at scale ϵ with respect to the ℓ_2-distance as $\log \mathcal{N}(\epsilon, \mathcal{W}_J, \ell_2) \leq \mathcal{O}\left(k \log\left(1/\epsilon\right)\right)$. Since the loss function $\ell(w;\xi)$ is G-Lipschitz continuous with respect to w, it can be verified that the covering number of the class of functions $\mathcal{L}_J = \{\xi \mapsto \ell_w(\xi) \mid w \in \mathcal{W}_J\}$ with respect to ℓ_∞-distance $\|\ell_{w_1} - \ell_{w_2}\|_\infty$ is given by

$$\log \mathcal{N}(\epsilon, \mathcal{L}_J, \ell_\infty) \leq \log \mathcal{N}(\epsilon/G, \mathcal{W}_J, \ell_2) \leq \mathcal{O}\left(k \log(G/\epsilon)\right).$$

Based on the result from Srebro et al. (2010, Lemma A.3) on the connection between Rademacher complexity and covering number we can show that

$$\begin{aligned}
&R_{S|J}(Gr_n; w_J^*) \\
\leq &\inf_{\alpha>0}\left\{4\alpha + 10\int_\alpha^{Gr_n}\sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{L}_J, \ell_\infty)}{n}}d\epsilon\right\} \leq \mathcal{O}\left(4Gr_n\sqrt{\frac{k}{n}} + 10\int_{Gr_n\sqrt{\frac{k}{n}}}^{Gr_n}\sqrt{\frac{k\log(G/\epsilon)}{n}}d\epsilon\right) \\
\leq &\mathcal{O}\left(4Gr_n\sqrt{\frac{k}{n}} + 10Gr_n\sqrt{\frac{k}{n}}\int_{Gr_n\sqrt{\frac{k}{n}}}^{Gr_n}\frac{\sqrt{\log(G/\epsilon)}}{\epsilon}d\epsilon\right) \\
\overset{\zeta_1}{\leq} &\mathcal{O}\left(4Gr_n\sqrt{\frac{k}{n}} + 6.67Gr_n\sqrt{\frac{k}{n}}\log^{3/2}\left(\frac{\sqrt{n}}{r_n\sqrt{k}}\right)\right) \\
\leq &\mathcal{O}\left(Gr_n\sqrt{\frac{k}{n}}\log^{3/2}\left(\frac{\sqrt{n}}{r_n\sqrt{k}}\right)\right),
\end{aligned}$$

where in "ζ_1" we have used the following fact for $c > b > a > 0$:

$$\int_a^b x^{-1}\sqrt{\log\left(\frac{c}{x}\right)}dx = \frac{2}{3}\left(\log^{3/2}\left(\frac{c}{a}\right) - \log^{3/2}\left(\frac{c}{b}\right)\right) \leq \frac{2}{3}\log^{3/2}\left(\frac{c}{a}\right).$$

This proves the desired bound. □

The following lemma presents a uniform fast rate of $w_{S|J}$ for all J.

Lemma 8. *Suppose that Assumptions 1, 4 are valid. Then for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\sup_{J\subseteq[p],|J|=k} F(w_{S|J}) - F(w_J^*) \leq \mathcal{O}\left(\frac{G^2k(\log^3(\rho n) + \log(ep/k)) + \log(1/\delta)}{\rho n}\right).$$

Proof. Fix a subset J with $|J| = k$. Let $r_n > 0$ be an arbitrary scalar that satisfies

$$\frac{\rho r_n^2}{2} \geq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}}. \tag{A.5}$$

Our first step is to show that

$$\mathbb{P}\left(F(w_{S|J}) - F(w_J^*) \leq \frac{\rho r_n^2}{2}\right) \geq 1-\delta. \tag{A.6}$$

To this end, suppose the event $\|w_{S|J} - w_J^*\| > r_n$ occurs. We can verify that the following event occurs consequently:

$$\sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\|_\infty \leq Gr_n} |F_S(w) - F_S(w_J^*) - (F(w) - F(w_J^*))| \geq 2R_S(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}}.$$

Indeed, let us consider

$$\tilde{w}_J = (1-\eta_n)w_J^* + \eta_n w_{S|J},$$

where $\eta_n = \frac{r_n}{\|w_{S|J} - w_J^*\|} < 1$. It is direct to verify that $\|\tilde{w}_J - w_J^*\| = r_n$. Since F_S is convex, we must have

$$F_S(\tilde{w}_J) \leq (1-\eta_n)F_S(w_J^*) + \eta_n F_S(w_{S|J}) \leq F_S(w_J^*).$$

Note that $\|\ell_{\tilde{w}_J} - \ell_{w_J^*}\|_\infty \leq G\|\tilde{w}_J - w_J^*\| = Gr_n$. Therefore, we have

$$\begin{aligned}
&\sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\| \leq Gr_n} |F_S(w) - F_S(w_J^*) - (F(w) - F(w_J^*))| \\
&\geq |F_S(\tilde{w}_J) - F_S(w_J^*) - (F(\tilde{w}_J) - F(w_J^*))| \\
&\geq |F(\tilde{w}_J) - F(w_J^*)| \\
&\overset{\zeta_1}{\geq} \frac{\rho}{2}\|\tilde{w}_J - w_J^*\|^2 = \frac{\rho r_n^2}{2} \geq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}},
\end{aligned}$$

where in "ζ_1" we have used Assumption 4 and the last inequality follows from (A.5). Then, invoking Lemma 3 over the supporting set J yields

$$\begin{aligned}
&\mathbb{P}\left(\|w_{S|J} - w_J^*\| > r_n\right) \\
\leq &\mathbb{P}\left(\sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\| \leq Gr_n} |F_S(w) - F_S(w_J^*) - (F(w) - F(w_J^*))| \geq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}}\right) \\
\leq &\frac{\delta}{2}.
\end{aligned}$$

Now let us consider the following three events:

$$\begin{aligned}
\mathcal{E}_1 &: \left\{F(w_{S|J}) - F(w_J^*) \leq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}}\right\}, \\
\mathcal{E}_2 &: \left\{\|w_{S|J} - w_J^*\| \leq r_n\right\}, \\
\mathcal{E}_3 &: \left\{\sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\| \leq Gr_n} |F_S(w) - F_S(w_J^*) - (F(w) - F(w_J^*))| \leq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}}\right\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\|w_{S|J} - w_J^*\| \leq r_n \\
\Rightarrow &\|\ell_{w_{S|J}} - \ell_{w_J^*}\| \leq Gr_n \\
\Rightarrow &F(w_{S|J}) - F(w_J^*) \leq \sup_{\mathrm{supp}(w)\subseteq J, \|\ell_w - \ell_{w_J^*}\| \leq Gr_n} |F_S(w) - F_S(w_J^*) - (F(w) - F(w_J^*))|.
\end{aligned}$$

Therefore, we must have

$$\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_2) \geq 1 - \mathbb{P}\left(\overline{\mathcal{E}_2}\right) - \mathbb{P}\left(\overline{\mathcal{E}_3}\right) \geq 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta,$$

which together with (A.5) implies the desired bound in (A.6).

The next step is to properly choose r_n so as to fulfill the key condition of (A.5). Based on the bound on $R_{S|J}(Gr_n; w_J^*)$ as summarized in Lemma 7, there exists some $C > 0$ such that

$$\begin{aligned}
&\frac{\rho r_n^2}{2} \geq 2R_{S|J}(Gr_n; w_J^*) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}} \\
\Leftarrow &\frac{\rho r_n^2}{2} \geq 2CGr_n\sqrt{\frac{k}{n}}\log^{3/2}\left(\frac{\sqrt{n}}{r_n\sqrt{k}}\right) + \frac{3Gr_n\sqrt{2\log(4/\delta)}}{\sqrt{n}} \\
\Leftarrow &r_n \geq \frac{4CG\sqrt{k}\log^{3/2}\left(\frac{\sqrt{n}}{r_n\sqrt{k}}\right) + 6G\sqrt{2\log(4/\delta)}}{\rho\sqrt{n}}.
\end{aligned}$$

Therefore, it suffices to choose

$$r_n = \mathcal{O}\left(\frac{G\sqrt{k}\log^{3/2}(\rho n) + G\sqrt{\log(1/\delta)}}{\rho\sqrt{n}}\right) \leq \mathcal{O}\left(\frac{G}{\rho}\sqrt{\frac{k\log^3(\rho n) + \log(1/\delta)}{n}}\right).$$

Substituting the above choice of r_n to (A.6) yields

$$F(w_{S|J}) - F(w_J^*) \leq \mathcal{O}\left(\frac{G^2 k \log^3(\rho n) + \log(1/\delta)}{\rho n}\right).$$

As the final step, since there are at most $\binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$ different J, by union probability we get

$$\sup_{J\subseteq[p],|J|=k} F(w_{S|J}) - F(w_J^*) \leq \mathcal{O}\left(\frac{G^2 k(\log^3(\rho n) + \log(ep/k)) + \log(1/\delta)}{\rho n}\right).$$

This completes the proof. □

To prove the main result, we also need to prove the following lemma which basically provides a sufficient condition to guarantee the support recovery performance of IHT.

Lemma 9. *Suppose that F_S is μ_{2k}-strongly convex with probability at least $1-\delta_n'$. Assume that the loss function ℓ is G-Lipschitz. Suppose that there exists a $\bar{k}$-sparse vector $\bar{w}$ such that*

$$\bar{w}_{\min} > \frac{2\sqrt{2k}\|\nabla F(\bar{w})\|_\infty}{\mu_{2k}} + \frac{3G}{\mu_{2k}}\sqrt{\frac{k\log(p/\delta)}{n}}$$

for some $\delta \in (0, 1-\delta_n')$. Then for sufficiently large $T \geq \mathcal{O}\left(\frac{L_{2k}}{\mu_{2k}}\log\left(\frac{n\mu_{2k}}{kG\log(n)\log(p/k)}\right)\right)$ rounds of IHT iteration, the support recovery $supp(\bar{w}) \subseteq supp(w_{S,k}^{(T)})$ holds with probability at least $1-\delta-\delta_n'$.

Proof. Let us consider a fixed $\bar{w}$. Since the G-Lipschitz condition implies $\|\nabla\ell(\bar{w};\cdot)\| \leq G$, from the Hoeffding concentration bound we know that with probability at least $1-\delta$ over S,

$$\|\nabla F_S(\bar{w}) - \nabla F(\bar{w})\| \leq G\sqrt{\frac{\log(p/\delta)}{2n}}.$$

Then with probability at least $1-\delta$,

$$\begin{aligned}\|\nabla F_S(\bar{w})\|_\infty &\leq \|\nabla F(\bar{w})\|_\infty + \|\nabla F_S(\bar{w}) - \nabla F(\bar{w})\|_\infty \\ &\leq \|\nabla F(\bar{w})\|_\infty + \|\nabla F_S(\bar{w}) - \nabla F(\bar{w})\| \leq \|\nabla F(\bar{w})\|_\infty + G\sqrt{\frac{\log(p/\delta)}{2n}}.\end{aligned} \tag{A.7}$$

Since with probability at least $1-\delta_n'$ the empirical risk F_S is μ_{2k}-strongly convex, the bound in Lemma 2 implies that the following holds for sufficiently large $T \geq \mathcal{O}\left(\frac{L_{2k}}{\mu_{2k}}\log\left(\frac{n\mu_{2k}}{kG\log(n)\log(p/k)}\right)\right)$ with probability at least $1-\delta_n'$:

$$F_S(\tilde{w}_{S,k}^{(T)}) \leq F_S(\bar{w}) + \frac{G^2 k\log(p/\delta)}{2\mu_{2k}n}.$$

Invoking Lemma 4 to the above with $w = \tilde{w}_{S,k}^{(T)}, w' = \bar{w}$ and $\epsilon = \frac{kG^2\log(1/\delta)}{2\mu_{2k}n}$ yields that with probability at least $1-\delta_n'$,

$$\|\tilde{w}_{S,k}^{(T)} - \bar{w}\| \leq \frac{2\sqrt{2k}\|\nabla F_S(\bar{w})\|_\infty}{\mu_{2k}} + \sqrt{\frac{2\epsilon}{\mu_{2k}}} = \frac{2\sqrt{2k}\|\nabla F_S(\bar{w})\|_\infty}{\mu_{2k}} + \frac{G}{\mu_{2k}}\sqrt{\frac{k\log(p/\delta)}{n}}.$$

Using (A.7) and union probability argument we obtain that with probability at least $1-\delta-\delta_n'$,

$$\|\tilde{w}_{S,k}^{(T)} - \bar{w}\| \leq \frac{2\sqrt{2k}\|\nabla F(\bar{w})\|_\infty}{\mu_{2k}} + \frac{3G}{\mu_{2k}}\sqrt{\frac{k\log(p/\delta)}{n}}.$$

Consequently from the condition on $\bar{w}_{\min}$ we must have $\mathrm{supp}(\tilde{w}_{S,k}^{(T)}) \supseteq \mathrm{supp}(\bar{w})$ holds with probability at least $1-\delta-\delta_n'$. □

We are now ready to prove the main result of Theorem 2.

Proof of Theorem 2. In what follows, we denote $\tilde{J} = \mathrm{supp}(\tilde{w}_{S,k}^{(T)})$ and $w_{\tilde{J}}^* = \arg\min_{w\in\mathcal{W},\mathrm{supp}(w)\subseteq\tilde{J}} F(w)$. Let us define the following three events associated with the sample set S:

$$\mathcal{E}_1 : \left\{ F(\tilde{w}_{S,k}^{(T)}) - F(\bar{w}) \leq \mathcal{O}\left(\frac{G^2 k(\log^3(\rho n) + \log(ep/k)) + \log(1/\delta)}{\rho n} \right) \right\},$$
$$\mathcal{E}_2 : \left\{ F(\tilde{w}_{S,k}^{(T)}) - F(w_{\tilde{J}}^*) \leq \mathcal{O}\left(\frac{G^2 k(\log^3(\rho n) + \log(ep/k)) + \log(1/\delta)}{\rho n} \right) \right\},$$
$$\mathcal{E}_3 := \left\{ \mathrm{supp}(\bar{w}) \subseteq \tilde{J} \right\}.$$

We claim that $\mathcal{E}_1 \cap \mathcal{E}_3 \supseteq \mathcal{E}_2 \cap \mathcal{E}_3$. Indeed, for any $S \in \mathcal{E}_2 \cap \mathcal{E}_3$, we have

$$\begin{aligned} &\mathrm{supp}(\bar{w}) \subseteq \tilde{J} \\ \Rightarrow &F(\tilde{w}_{S,k}^{(T)}) - F(\bar{w}) \leq F(\tilde{w}_{S,k}^{(T)}) - F(w_{\tilde{J}}^*) \leq \mathcal{O}\left(\frac{G^2 k(\log^3(\rho n) + \log(ep/k)) + \log(1/\delta)}{\rho n} \right), \end{aligned}$$

which implies $S \in \mathcal{E}_1$ and thus $S \in \mathcal{E}_1 \cap \mathcal{E}_3$.

Given the condition on $\bar{w}$ and $\delta_n' \leq \frac{\delta}{4}$, it follows from Lemma 9 that $\mathrm{supp}(\bar{w}) \subseteq \tilde{J}$ holds with probability at least $1 - \frac{\delta}{2}$, i.e.,

$$\mathbb{P}\left(\mathcal{E}_3\right) \geq 1 - \frac{\delta}{2}.$$

In the meanwhile, noting $\tilde{w}_{S,k}^{(T)} = w_{S|\tilde{J}}$ and invoking Lemma 8 yields

$$\mathbb{P}\left(\mathcal{E}_2\right) \geq 1 - \frac{\delta}{2}.$$

Combining the above leads to

$$\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_3) \geq \mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \mathbb{P}(\overline{\mathcal{E}}_2) - \mathbb{P}(\overline{\mathcal{E}}_3) \geq 1 - \delta.$$

This proves the desired bound (keep in mind the monotonicity of restricted smoothness and strong convexity). □

C.2 Proof of Theorem 3

We need the following lemma which can be derived based on the concentration bound of sub-Gaussian random variables.

Lemma 10. *Under Assumption 5, for any $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*

$$\|\nabla F_S(\bar{w})\|_\infty \leq \sigma\sqrt{\frac{2\log(p/\delta)}{n}}.$$

Proof. Consider a fixed index $j \in [p]$. Since $\nabla_j \ell(\bar{w};\xi)$ are assumed to be σ^2-sub-Gaussian and $\nabla F(\bar{w}) = \mathbb{E}_\xi\left[\nabla\ell(\bar{w};\xi)\right] = 0$, we must have $\nabla_j \ell(\bar{w};\xi)$ are zero-mean σ^2-sub-Gaussian. Thus it is known from the Hoeffding inequality that for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\nabla_j F_S(\bar{w})| > \varepsilon\right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{\xi_i \in S} \nabla_j \ell(\bar{w};\xi_i) \right| > \varepsilon \right) \leq \exp\left\{ -\frac{n\varepsilon^2}{2\sigma^2} \right\}.$$

By the union bound we have

$$\mathbb{P}(\|\nabla F_S(\bar{w})\|_\infty > \varepsilon) \leq p \exp\left\{ -\frac{n\varepsilon^2}{2\sigma^2} \right\}.$$

By choosing $\varepsilon = \sqrt{\frac{2\sigma^2\log(p/\delta)}{n}}$ in the above inequality we obtain that with probability at least $1-\delta$,

$$\|\nabla F_S(\bar{w})\|_\infty \leq \sqrt{\frac{2\sigma^2\log(p/\delta)}{n}}.$$

This completes the proof. □

We are now ready to prove the main result of Theorem 3.

Proof of Theorem 3. Since by assumption $F_S(w)$ is L_{4k}-smooth and μ_{4k}-strongly convex with probability at least $1-\delta_n'$, Lemma 2 shows that $F_S(w_{S,k}^{(T)}) - F_S(\bar{w}) \leq \epsilon$ with probability at least $1-\delta_n'$ provided that $t \geq \mathcal{O}\left(\frac{L_{4k}}{\mu_{4k}} \log\left(\frac{1}{\epsilon}\right)\right)$. Then by invoking Lemma 4 we obtain that with probability at least $1-\delta_n'$,

$$\|w_{S,k}^{(T)} - \bar{w}\|^2 \leq \frac{16k\|\nabla F_S(\bar{w})\|_\infty^2}{\mu_{2k}^2} + \frac{4\epsilon}{\mu_{2k}} \leq \frac{16k\|\nabla F_S(\bar{w})\|_\infty^2}{\mu_{4k}^2} + \frac{4\epsilon}{\mu_{4k}}.$$

From Lemma 10 we know that with probability at least $1-\delta$,

$$\|\nabla F_S(\bar{w})\|_\infty \leq \sigma\sqrt{\frac{2\log(p/\delta)}{n}}.$$

Then by union probability the following holds with probability at least $1-\delta-\delta_n'$:

$$\|w_{S,k}^{(T)} - \bar{w}\|^2 \leq \frac{32}{\mu_{4k}^2}\left(\frac{k\sigma^2\log(p/\delta)}{n}\right) + \frac{4\epsilon}{\mu_{4k}}.$$

Based on the Lipschitz smoothness of F we can show

$$F(w_{S,k}^{(T)}) - F(\bar{w}) \leq \frac{L}{2}\|w_{S,k}^{(t)} - \bar{w}\|^2 \leq \frac{16L}{\mu_{4k}^2}\left(\frac{k\sigma^2\log(p/\delta)}{n}\right) + \frac{2L\epsilon}{\mu_{4k}}.$$

Setting $\epsilon = \frac{1}{\mu_{4k}}\left(\frac{k\sigma^2\log(p/\delta)}{n}\right)$ yields the desired high probability bound of sparse excess risk. □

C.3 Proofs of Corollary 2 and Corollary 3

We first prove Corollary 2 which is an application of Theorem 3 to sparse linear regression models.

Proof of Corollary 2. Let $\xi = \{x, \varepsilon\}$ in which x is zero-mean sub-Gaussian with covariance matrix $\Sigma \succ 0$ and ε is zero-mean σ^2-sub-Gaussian. Since x and ε are independent, it can be directly verified that $\nabla F(\bar{w}) = \mathbb{E}_\xi\left[\nabla \ell(\bar{w};\xi)\right] = \mathbb{E}_{\varepsilon,x}\left[-\varepsilon x\right] = 0$. Given that $\Sigma_{jj} \leq 1$, it can be shown that $\nabla_j \ell(\bar{w};\xi_i) = -\varepsilon_i [x_i]_j$ are zero-mean σ^2-sub-Gaussian variables, which indicates that Assumption 5 holds. Clearly, F is L-smooth with $L = \lambda_{\max}(\Sigma)$.

Based on Lemma 6 and the fact $\|w\|_1 \leq \sqrt{k}\|w\|$ when $\|w\|_0 \leq k$, it holds with probability at least $1-\exp\{-c_0 n\}$ that $F_S(w)$ is μ_{4k}-strongly convex with

$$\mu_{4k} = \frac{1}{2}\lambda_{\min}(\Sigma) - \frac{kc_1\log(p)}{n}.$$

Provided that $n \geq \frac{4kc_1\log(p)}{\lambda_{\min}(\Sigma)}$, we have $\mu_{4k} \geq \frac{1}{4}\lambda_{\min}(\Sigma)$ holds with probability at least $1-\exp\{-c_0 n\}$. Similarly, we can show that $F_S(w)$ is L_{4k}-smooth with $L_{4k} = \mathcal{O}(\lambda_{\max}(\Sigma))$. Provided that

$$T \geq \mathcal{O}\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\log\left(\frac{n\lambda_{\min}(\Sigma)}{k\sigma^2\log(p/\delta)}\right)\right)$$

is sufficiently large, by applying the high probability bound in Theorem 3 we obtain that with probability at least $1-\delta-\exp\{-c_0 n\}$,

$$F(w_{S,k}^{(T)}) - F(\bar{w}) \leq \mathcal{O}\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}^2(\Sigma)}\left(\frac{k\sigma^2\log(p/\delta)}{n}\right)\right).$$

This proves the desired bounds. □

In what follows we prove Corollary 3 as an application of Theorem 3 to sparse logistic regression models.

Proof of Corollary 3. Let $\xi = \{x, y\}$ in which x is zero-mean sub-Gaussian with covariance matrix $\Sigma \succ 0$ and $y \in \{-1, 1\}$ is generated by $\mathbb{P}(y|x; \bar{w}) = \frac{\exp(2y\bar{w}^\top x)}{1+\exp(2y\bar{w}^\top x)}$. The logistic loss function at ξ_i is given by $\ell(w; \xi_i) = \log(1 + \exp(-2y_i w^\top x_i))$. We first show that $\nabla F(\bar{w}) = \mathbb{E}_\xi [\nabla \ell(\bar{w}; \xi)] = 0$. Indeed,

$$
\begin{aligned}
&\mathbb{E}_\xi [\nabla \ell(\bar{w}; \xi)] \\
=&\mathbb{E}_{x,y} \left[\nabla \log(1 + \exp(-2y\bar{w}^\top x))\right] = \mathbb{E}_x \left[\mathbb{E}_{y|x} \left[\nabla \log(1 + \exp(-2y\bar{w}^\top x)) \mid x\right]\right] \\
=&\mathbb{E}_x \left[\mathbb{P}(y = 1 \mid x)\nabla \log(1 + \exp(-2\bar{w}^\top x)) + \mathbb{P}(y = -1 \mid x)\nabla \log(1 + \exp(2\bar{w}^\top x))\right] \\
=&\mathbb{E}_x \left[\frac{\exp(2\bar{w}^\top x)}{1 + \exp(2\bar{w}^\top x)} \frac{-2x \exp(-2\bar{w}^\top x)}{1 + \exp(-2\bar{w}^\top x)} + \frac{1}{1 + \exp(2\bar{w}^\top x)} \frac{2x \exp(2\bar{w}^\top x)}{1 + \exp(2\bar{w}^\top x)}\right] = 0.
\end{aligned}
$$

Next we show that $\nabla_j \ell(\bar{w}; \xi) = \frac{-2y[x]_j \exp(-2y\bar{w}^\top x)}{1+\exp(-2y\bar{w}^\top x)}$ is a zero-mean sub-Gaussian random variable. Clearly, $\mathbb{E}[\nabla_j \ell(\bar{w}; \xi)] = 0$. Since $y \in \{-1, 1\}$ and $[x]_j$ is $\frac{\sigma^2}{32}$-sub-Gaussian, we can show the following

$$
\mathbb{P}\left(|\nabla_j \ell(\bar{w}; \xi)| \geq t\right) = \mathbb{P}\left(\frac{2|[x]_j| \exp(-2y\bar{w}^\top x)}{1 + \exp(-2y\bar{w}^\top x)} \geq t\right) \leq \mathbb{P}\left(|[x]_j| \geq \frac{t}{2}\right) \leq 2 \exp\left(-\frac{4t^2}{\sigma^2}\right).
$$

Then based on the result of Rigollet (2015, Lemma 1.5) we know that for any $\lambda > 0$,

$$
\mathbb{E}_\xi \left[\exp(\lambda \nabla_j \ell(\bar{w}; \xi))\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right),
$$

which shows that $\nabla_j \ell(\bar{w}; \xi)$ is σ^2-sub-Gaussian. This verifies the validness of Assumption 5.

By invoking Lemma 6 we obtain that if $n \geq \frac{4\sigma^2 k c_1 \log(p)}{\lambda_{\min}(\Sigma)}$, then it holds with probability at least $1 - \exp\{-c_0 n\}$ that $F_S(w)$ is μ_{4k}-strongly convex with $\mu_{4k} \geq \frac{\lambda_{\min}(\Sigma)}{\exp(4R)}$. It is standard to verify that F and F_S are $\mathcal{O}(1)$-smooth almost surely. Therefore, provided that

$$
T \geq \mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}(\Sigma)} \log\left(\frac{n\lambda_{\min}(\Sigma)}{k \exp(R)\sigma^2 \log(p/\delta)}\right)\right)
$$

is sufficiently large, by applying the bound in Theorem 3 we obtain that the following bound holds with probability at least $1 - \delta - \exp\{-c_0 n\}$:

$$
F(w_{S,k}^{(T)}) - F(\bar{w}) \leq \mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}^2(\Sigma)} \left(\frac{k\sigma^2 \log(p/\delta)}{n}\right)\right).
$$

This concludes the proof. □