A IRLS and the Scaled Huber Loss - Supplementary Details

We recapitulate below the definitions of the Huber loss, the scaled (and translated) Huber loss and, given a model $\mathbf{w}^0$ and data $(\mathbf{x}_i, y_i)_{i=1}^n$, other allied functions.

$$h_\epsilon(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le \epsilon \\ \epsilon |x| - \frac{1}{2}\epsilon^2 & |x| > \epsilon \end{cases}$$

$$f_\epsilon(x) = \begin{cases} \frac{1}{2}\left(\frac{x^2}{\epsilon} + \epsilon\right) & |x| \le \epsilon \\ |x| & |x| > \epsilon \end{cases}$$

$$g_\epsilon(x; a) := \frac{1}{2}\left(\frac{x^2}{\max\{|a|, \epsilon\}} + \max\{|a|, \epsilon\}\right)$$

$$\ell_\epsilon(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n f_\epsilon\left(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i\right)$$

$$\wp_\epsilon(\mathbf{w}; \mathbf{w}^0) := \sum_{i=1}^n g_\epsilon\left(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i; \langle \mathbf{w}^0, \mathbf{x}_i\rangle - y_i\right)$$

The claim that M-truncated IRLS minimizes $\wp_{\frac{1}{M}}(\mathbf{w}; \mathbf{w}^0)$ to obtain the next model can be easily verified using the equivalence between the truncation and regularization techniques explained in Footnote 1 (see §5 for the footnote). In the following, we establish that $g_\epsilon(\cdot;\cdot)$ is a valid majorizer for f_ϵ for any $\epsilon > 0$.

Claim 3. *For any $a, x \in \mathbb{R}, \epsilon > 0$, we have $g_\epsilon(a; a) = f_\epsilon(a)$ as well as $g_\epsilon(x; a) \ge f_\epsilon(x)$.*

Proof. We have, for the first claim,

$$g_\epsilon(a; a) = \frac{1}{2}\left(\frac{a^2}{\max\{|a|, \epsilon\}} + \max\{|a|, \epsilon\}\right) = \begin{cases} \frac{1}{2}\left(\frac{a^2}{\epsilon} + \epsilon\right) & |a| \le \epsilon \\ |a| & |a| > \epsilon \end{cases} = f_\epsilon(a).$$

For the second claim, we consider two simple cases

Case 1 $|x| > \epsilon$: In this case we have $f_\epsilon(x) = |x|$ and we always have $\frac{1}{2}\left(\frac{x^2}{\max\{|a|,\epsilon\}} + \max\{|a|, \epsilon\}\right) \ge |x|$.

Case 2 $|x| \le \epsilon$: In this case denote $b = \max\{|a|, \epsilon\}$. Then we have $b \ge \epsilon \ge |x|$ which gives us $x^2 \le b\epsilon$. Thus, we have $g_\epsilon(x; a) - f_\epsilon(x) = \frac{1}{2}\left(\frac{x^2}{b} + b\right) - \frac{1}{2}\left(\frac{x^2}{\epsilon} + \epsilon\right) = \frac{(b-\epsilon)(b\epsilon - x^2)}{2b\epsilon} \ge 0$. □

The following claim shows that we have $f'_\epsilon(x)|_{x=a} = g'_\epsilon(x; a)|_{x=a}$ for any ϵ, a. This immediately establishes that $\nabla\wp_\epsilon(\mathbf{w}^0; \mathbf{w}^0) = \nabla\ell_\epsilon(\mathbf{w}^0)$ for any model $\mathbf{w}^0$.

Claim 4. *For any $a, x \in \mathbb{R}, \epsilon > 0$, we have $f'_\epsilon(x)|_{x=a} = g'_\epsilon(x; a)|_{x=a}$.*

Proof. We have $g'_\epsilon(x; a) = \frac{x}{\max\{|a|,\epsilon\}}$ which gives us

$$g'_\epsilon(x; a)|_{x=a} = \begin{cases} \frac{a}{\epsilon} & |a| \le \epsilon \\ \text{sign}(a) & |a| > \epsilon, \end{cases}$$

whereas we have

$$f'_\epsilon(x) = \begin{cases} \frac{x}{\epsilon} & |x| \le \epsilon \\ \text{sign}(x) & |x| > \epsilon \end{cases},$$

which establishes the claim. □

B Supporting Results

In this section we prove a few results used in the convergence analysis of STIR.

Lemma 5. *Suppose we have data covariates $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ generated from an isotropic but otherwise arbitrary sub-Gaussian distribution. Then for any fixed set $S \subset [n]$ and $n = \Omega\left(d + \log \frac{1}{\delta}\right)$, with probability at least $1 - \delta$,*

$$0.99\,|S| \le \lambda_{\min}(X_S X_S^\top) \le \lambda_{\max}(X_S X_S^\top) \le 1.01\,|S|\,,$$

where the constant inside $\Omega(\cdot)$ depends only on the sub-Gaussian distribution and universal constants.

Proof. This is a special case of [6, Lemma 16] for isotropic distributions. Note that since our adversary is partially adaptive, the sets of good and bad points G, B are fixed and this lemma applies to both G and B. □

Lemma 6. *Suppose our data covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$ are generated from a sub-Gaussian distribution with sub-Gaussian norm R. Then with probability at least $1-\delta$, we have $R_X := \max_{i\in[n]} \|\mathbf{x}_i\|_2 \le \|\boldsymbol{\mu}\|_2 + \mathcal{O}\left(R\sqrt{d + \log \frac{n}{\delta}}\right)$.*

Proof. If $\mathbf{x}$ is R-sub-Gaussian with mean $\boldsymbol{\mu}$, then for any unit vector $\mathbf{v} \in S^{d-1}$, $\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu}\rangle$ is centered as well as $2R$-sub-Gaussian which gives us

$$\mathbb{P}\left[|\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu}\rangle| \ge t\right] \le 2\exp\left[-t^2/2R^2\right]$$

If $\mathbf{v}^1, \mathbf{v}^2 \in S^{d-1}$, such that $\left\|\mathbf{v}^1 - \mathbf{v}^2\right\|_2 \le \frac{1}{2}$, then we have $\left|\left\langle \mathbf{v}^1 - \mathbf{v}^2, \mathbf{x} - \boldsymbol{\mu}\right\rangle\right| \le \frac{1}{2} \cdot \|\mathbf{x} - \boldsymbol{\mu}\|_2$. Thus, taking a union bound over a 1/2-net over S^{d-1} gives us

$$\mathbb{P}\left[\max_{\mathbf{v}\in S^{d-1}} |\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu}\rangle| \ge \frac{1}{2} \cdot \|\mathbf{x} - \boldsymbol{\mu}\|_2 + t\right] = \mathbb{P}\left[\|\mathbf{x}\|_2 \ge \|\boldsymbol{\mu}\|_2 + 2t\right] \le 2 \cdot 5^d \exp\left[-t^2/2R^2\right]$$

Taking $t^2 = 2R^2(d\log 5 + \log\frac{n}{\delta} + \log 2)$ proves the result.

$$\mathbb{P}\left[\max_{i\in[n]} \|\mathbf{x}_i\|_2 > \|\boldsymbol{\mu}\|_2 + R\sqrt{2\left(d\log 5 + \log\frac{n}{\delta} + \log 2\right)}\right] \le \delta$$ □

In the following, we establish that the scaled Huber loss is Lipschitz. This will be helpful in transferring our convergence guarantees to those with respect to the Huber and absolute loss functions.

Lemma 7. *For any $\epsilon > 0$, we have $|\ell_\epsilon(\mathbf{w}) - \ell_\epsilon(\mathbf{w}')| \le \|\mathbf{w} - \mathbf{w}'\|_2 \cdot \sqrt{1.01}$.*

Proof. The function $f_\epsilon(\cdot)$ is clearly 1-Lipschitz for any $\epsilon > 0$. This means that we have

$$\begin{aligned}|\ell_\epsilon(\mathbf{w}) - \ell_\epsilon(\mathbf{w}')| &\le \frac{1}{n}\sum_{i=1}^n |\langle \mathbf{w}, \mathbf{x}_i\rangle - \langle \mathbf{w}', \mathbf{x}_i\rangle| = \frac{1}{n}\left\|X^\top(\mathbf{w} - \mathbf{w}')\right\|_1 \le \frac{1}{\sqrt{n}}\left\|X^\top(\mathbf{w} - \mathbf{w}')\right\|_2 \\ &\le \frac{1}{\sqrt{n}}\|X\|_2\|\mathbf{w} - \mathbf{w}'\|_2 \le \|\mathbf{w} - \mathbf{w}'\|_2 \cdot \sqrt{1.01},\end{aligned}$$

where the last step follows due to Lemma 5. □

C Convergence Analysis - Supplementary Details

We begin by restating Theorem 1, the main result that we will prove in this section.

Theorem 1. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \le \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \le \frac{c}{2.88\eta + c}$, where $c > 0$ is a constant that depends only on $\mathcal{D}$, then for any $\epsilon > 0$, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n) + \log\frac{1}{M_1\epsilon}\right)\right)$, after $K = \mathcal{O}\left(\log\frac{1}{M_1\epsilon}\right)$ stages, we must have $\left\|\mathbf{w}^K - \mathbf{w}^*\right\|_2 \le \epsilon$. Moreover, each stage consists of only $\mathcal{O}(1)$ iterations.*

Proof. As mentioned before, notice that this is indeed a global convergence guarantee since it places no restrictions on the initial model $\mathbf{w}^0$. The only requirement is that the accompanying initial truncation parameter M_1 complement the model initialization by satisfying $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$. In particular, if initialized at the origin, as Algorithms 1 and 2 do, we need only ensure $M_1 \leq \frac{1}{R_W}$ where $R_W = \|\mathbf{w}^*\|_2$. This can be done using a simple binary search to identify an appropriate value of M_1. Recall that both STIR and STIR-GD operate in stages. We introduce a notion of a *well-initialized* stage below.

Definition 2 (Well-initialized Stage)**.** *A stage in the execution of STIR or STIR-GD is said to be well-initialized if, given the truncation parameter M_T which will be used during that stage, at the beginning of that stage T, we are in possession of a model $\mathbf{w}^{T,1}$ that satisfies $\left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M_T}$.*

Note that the initialization of STIR and STIR-GD with respect to the setting of M_1 ensure $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$ which implies that the very first stage is always well-initialized. Now, Lemmata 8 and 9 show that, if the preconditions of this theorem are satisfied, then a stage T, started off with a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ (see Algorithm 1, line 3) and a truncation parameter M_T that satisfy the well-initialized condition i.e. $\left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M_T}$, will ensure with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, that there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M_T}$ for all $\tau \geq t_0$.

An application of the triangle inequality shows that we will have $\left\|\mathbf{w}^{T,t_0} - \mathbf{w}^{T,t_0+1}\right\|_2 \leq \frac{2}{\eta M_T}$ which implies (see Algorithm 1, line 5) that we will exit this stage at the $(t_0+1)^{\text{th}}$ inner iteration. However, notice that at this point we are endowed with $\left\|\mathbf{w}^{T+1,1} - \mathbf{w}^*\right\|_2 = \left\|\mathbf{w}^{T+1} - \mathbf{w}^*\right\|_2 = \left\|\mathbf{w}^{T,t_0+1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M_T} = \frac{1}{M_{T+1}}$. Note that this means that stage $(T+1)$ is well-initialized too.

Thus, whenever a stage T is well-initialized, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, we have $\left\|\mathbf{w}^{T+1,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta}\left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2$. Since we always set $\eta > 1$, there exists an upper bound $T_0 = \mathcal{O}\left(\log \frac{1}{M_1 \epsilon}\right)$ on the number of stages. Thus, an application of union bound shows that we must have $\left\|\mathbf{w}^{T_0+1,1} - \mathbf{w}^*\right\|_2 \leq \epsilon$ with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right) + \log \frac{1}{M_1 \epsilon}\right) = 1 - \exp(-\tilde{\Omega}(n))$ for all $\epsilon = \frac{1}{n^{\mathcal{O}(1)}}$. □

Lemma 8. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. Suppose we initialize a stage T within an execution of STIR with truncation level M, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ such that $\alpha \leq \frac{c}{2.88\eta + c}$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$. Here c is the constant of the WSC property and depends only on the distribution $\mathcal{D}$ (see Lemma 12).*

Proof. Let $\mathbf{w}^{T,\tau}$ be a model encountered by STIR within this stage and let $\mathbf{r} = X^\top \mathbf{w}^{T,\tau} - \mathbf{y}$ denote the residuals due to $\mathbf{w}^{T,\tau}$ and $S = \text{diag}(\mathbf{s})$ denote the diagonal matrix of weights where $\mathbf{s}_i = \min\left\{\frac{1}{|\mathbf{r}_i|}, M\right\}$. Then STIR will choose as the next model $\mathbf{w}^{T,\tau+1} = (XSX^\top)^{-1}XS\mathbf{y} = \mathbf{w}^* + (XSX^\top)^{-1}XS\mathbf{b}$ which gives us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{\|XS\mathbf{b}\|_2}{\lambda_{\min}(XSX^\top)}$$

Now by Lemma 5, with probability at least $1 - \exp(-\Omega(n-d))$, we have $\|X_B\|_2 = \sqrt{\lambda_{\max}(X_B X_B^\top)} \leq \sqrt{1.01B}$. By Lemma 10 we have, again with probability at least $1 - \exp(-\Omega(n-d))$

$$\|S\mathbf{b}\|_2 \leq \sqrt{4B(1 + 1.01M^2\|\mathbf{w} - \mathbf{w}^*\|_2^2)} \leq 2\sqrt{2.01B}$$

It should be noted that Lemma 10 relies precisely on Lemma 5 to derive its confidence assurance. Since the nature of Lemma 5 is such that it need be established only once, and not repeatedly for every iteration, we have, with probability at least $1 - \exp(-\Omega(n-d))$, for *all iterations* within this stage (actually all iterations across all stages), both Lemma 10 and Lemma 5 hold simultaneously.

Using Lemma 12, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, we have $\lambda_{\min}(XSX^\top) \geq \lambda_{\min}(X_G S_G X_G^\top) \geq 0.99c \cdot GM$. Note that since all models $\mathbf{w}^{T,\tau}, \tau \geq 1$ in this stage will at least satisfy

$\left\|\mathbf{w}^{T,\tau}-\mathbf{w}^*\right\|_2 \leq \frac{1}{M}$ (since the initial model $\mathbf{w}^{T,1}$ satisfies this by assumption and STIR offers monotonic convergence), the result of Lemma 12 applies uniformly to all these models and need not be applied separately to each model in this stage. Using these results to upper bound $\|XS\mathbf{b}\|_2$ and lower bound $\lambda_{\min}(XSX^\top)$ shows that at either we must have

$$\left\|\mathbf{w}^{T,\tau+1}-\mathbf{w}^*\right\|_2 \leq \frac{2B\sqrt{2.0301}}{0.99c\cdot GM}$$

or else if the above is not true, then we must instead have

$$\left\|\mathbf{w}^{T,\tau+1}-\mathbf{w}^*\right\|_2 \leq 0.99\cdot\|\mathbf{w}-\mathbf{w}^*\|_2$$

Note that since we have $\alpha \leq \frac{c}{2.88\eta+c}$, we get $\frac{2B\sqrt{2.0301}}{0.99c\cdot GM} \leq \frac{1}{\eta M}$. Thus, it is assured that after $t_0=\mathcal{O}(\log\eta)=\mathcal{O}(1)$ iterations, iterates $\mathbf{w}^{T,\tau}$ of STIR will satisfy $\left\|\mathbf{w}^{T,\tau}-\mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$ □

Lemma 9. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. Suppose we initialize a stage T within an execution of STIR-GD with truncation level M, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ such that $\alpha \leq \frac{c}{2.88\eta+c}$ and $\|\mathbf{w}-\mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1-\exp(-\Omega(n-d\log(d+n)))$, there exists an upper bound of $t_0=\mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau}-\mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$.*

Proof. As observed before, all models $\mathbf{w}^{T,\tau}, \tau \geq 1$ in this stage at least satisfy $\left\|\mathbf{w}^{T,\tau}-\mathbf{w}^*\right\|_2 \leq \frac{1}{M}$ since the initial model $\mathbf{w}^{T,1}$ satisfies this by assumption and we will see below that STIR-GD offers monotonic convergence. Thus, Lemma 12 applies uniformly to all these models and thus, with probability at least $1-\exp(-\Omega(n-d\log(d+n)))$, for all $\tau \geq 1$, the function $\wp_{\frac{1}{M}}(\cdot,\mathbf{w}^{T,\tau})$ (refer to §6 for notation) is γ-strongly convex for $\gamma \geq 0.99c\cdot GM$.

Similarly, Lemma 5 tells us that, again with probability at least $1-\exp(-\Omega(n-d\log(d+n)))$, for all $\tau \geq 1$,the function $\wp_{\frac{1}{M}}(\cdot,\mathbf{w}^{T,\tau})$ is δ-strongly smooth for $\delta \leq 1.01Mn$. From now on, we will be using the shorthand $\wp(\cdot) := \wp_{\frac{1}{M}}(\cdot,\mathbf{w}^{T,\tau})$ to avoid notational clutter.

If we denote $\mathbf{g}^t := \nabla\wp(\mathbf{w}^{T,\tau}) = \wp_{\frac{1}{M}}(\mathbf{w}^{T,\tau},\mathbf{w}^{T,\tau})$, then it is clear that STIR-GD will choose as the next model as $\mathbf{w}^{T,\tau+1} := \mathbf{w}^{T,\tau} - \frac{C}{Mn}\cdot\mathbf{g}^t$. For sake of notational simplicity, we will abbreviate $\mathbf{w} := \mathbf{w}^{T,\tau}, \mathbf{w}^+ := \mathbf{w}^{T,\tau+1}, \mathbf{g} := \mathbf{g}^t$. Then, applying strong smoothness tells us that

$$\begin{aligned}
\wp(\mathbf{w}^+)-\wp(\mathbf{w}) &\leq \left\langle \mathbf{g}, \mathbf{w}^+-\mathbf{w}\right\rangle + \frac{\delta}{2}\left\|\mathbf{w}^+-\mathbf{w}\right\|_2^2 \\
&= \left\langle \mathbf{g}, \mathbf{w}^+-\mathbf{w}^*\right\rangle + \left\langle \mathbf{g}, \mathbf{w}^*-\mathbf{w}\right\rangle + \frac{\delta}{2}\left\|\mathbf{w}^+-\mathbf{w}\right\|_2^2 \\
&= \frac{Mn}{C}\cdot\left\langle \mathbf{w}-\mathbf{w}^+, \mathbf{w}^+-\mathbf{w}^*\right\rangle + \left\langle \mathbf{g}, \mathbf{w}^*-\mathbf{w}\right\rangle + \frac{\delta}{2}\left\|\mathbf{w}^+-\mathbf{w}\right\|_2^2 \\
&= \frac{Mn}{2C}\left(\|\mathbf{w}-\mathbf{w}^*\|_2^2 - \left\|\mathbf{w}^+-\mathbf{w}^*\right\|_2^2\right) + \left\langle \mathbf{g}, \mathbf{w}^*-\mathbf{w}\right\rangle + \left(\frac{\delta}{2}-\frac{Mn}{2C}\right)\left\|\mathbf{w}^+-\mathbf{w}\right\|_2^2 \\
&\leq \frac{Mn}{2C}\left(\|\mathbf{w}-\mathbf{w}^*\|_2^2 - \left\|\mathbf{w}^+-\mathbf{w}^*\right\|_2^2\right) + \left\langle \mathbf{g}, \mathbf{w}^*-\mathbf{w}\right\rangle,
\end{aligned}$$

where the fifth step holds for any $C \leq \frac{Mn}{\delta} \leq 0.99$. Strong smoothness on the other hand tells us that

$$\left\langle \mathbf{g}, \mathbf{w}^*-\mathbf{w}\right\rangle \leq \wp(\mathbf{w}^*)-\wp(\mathbf{w}) - \frac{\gamma}{2}\|\mathbf{w}-\mathbf{w}^*\|_2^2$$

Combining the above two results gives us

$$\wp(\mathbf{w}^+)-\wp(\mathbf{w}^*) \leq \frac{Mn}{2C}\left(\|\mathbf{w}-\mathbf{w}^*\|_2^2 - \left\|\mathbf{w}^+-\mathbf{w}^*\right\|_2^2\right) - \frac{\gamma}{2}\|\mathbf{w}-\mathbf{w}^*\|_2^2$$

Now, we can either have $\wp(\mathbf{w}^+)-\wp(\mathbf{w}^*) \geq 0$ in which case we get $\|\mathbf{w}^+-\mathbf{w}^*\|_2 \leq \sqrt{1-\frac{C\gamma}{Mn}}\|\mathbf{w}-\mathbf{w}^*\|_2 \leq \sqrt{1-\frac{0.99cCG}{n}}\|\mathbf{w}-\mathbf{w}^*\|_2$ or else $\wp(\mathbf{w}^+)-\wp(\mathbf{w}^*) < 0$ in which case applying strong convexity once again yields

$$\frac{\gamma}{2}\left\|\mathbf{w}^+-\mathbf{w}^*\right\|_2^2 \leq \wp(\mathbf{w}^+)-\wp(\mathbf{w}^*) + \left\langle \nabla\wp(\mathbf{w}^*), \mathbf{w}^*-\mathbf{w}^+\right\rangle \leq \left\langle \nabla\wp(\mathbf{w}^*), \mathbf{w}^*-\mathbf{w}^+\right\rangle$$

Now notice that $\nabla_{\wp}(\mathbf{w}^*) = XS\mathbf{b}$ and Lemmata 10 and 5 tell us that $\|XS\mathbf{b}\|_2 \leq 2B\sqrt{5.05}$ which give us $\|\mathbf{w}^+ - \mathbf{w}^*\|_2 \leq \frac{2B\sqrt{2.0301}}{\gamma} \leq \frac{2B\sqrt{2.0301}}{0.99cGM} < \frac{1}{\eta M}$ whenever $\frac{B}{G} \leq \frac{0.99c}{2\eta\sqrt{2.0301}}$. This completes the proof of the result upon making similar arguments as those made in the proof of Lemma 9. □

C.1 Bounding the Weights on Bad Points

The following lemma establishes that neither STIR nor STIR-GD put too much weight on bad points.

Lemma 10. *Suppose during the execution of STIR or STIR-GD, we encounter a model* $\mathbf{w}$ *while the truncation parameter is* M*. Denote* $\|\mathbf{w} - \mathbf{w}^*\|_2 = \epsilon$ *and let* $S = diag(\mathbf{s})$ *be the diagonal matrix of* M*-truncated weights assigned due to residuals induced by* $\mathbf{w}$*. Then, with probability at least* $1 - \exp(-\Omega(n-d))$*, we must have*

$$\|S\mathbf{b}\|_2^2 \leq 4B(1 + 1.01M^2\epsilon^2),$$

where we recall that $\mathbf{b}$ *denotes the vector of corruptions.*

Proof. Let $\boldsymbol{\Delta} := \mathbf{w} - \mathbf{w}^*$ and let b_i denote the corruption on the data point $\mathbf{x}_i$. The proof proceeds via a simple case analysis

Case 1: $|b_i| \leq 2|\boldsymbol{\Delta} \cdot \mathbf{x}_i|$ In this case we simply bound $(s_i b_i)^2 \leq M^2 b_i^2 \leq 4M^2(\boldsymbol{\Delta} \cdot \mathbf{x}_i)^2$.

Case 2: $|b_i| > 2|\boldsymbol{\Delta} \cdot \mathbf{x}_i|$ In this case we have $|r_i| = |\boldsymbol{\Delta} \cdot \mathbf{x}_i - b_i| \geq |b_i| - |\boldsymbol{\Delta} \cdot \mathbf{x}_i| \geq \frac{|b_i|}{2}$ and thus we must have $s_i \leq \frac{2}{|b_i|}$ (due to possible truncation) and thus $(s_i b_i)^2 \leq 4$.

Thus, we get

$$\|S\mathbf{b}\|_2^2 = \sum_{i \in B}(s_i b_i)^2 \leq 4 \cdot \sum_{i \in B} \max\left\{1, M^2(\boldsymbol{\Delta} \cdot \mathbf{x}_i)^2\right\} \leq 4(B + M^2\epsilon^2\lambda_{\max}(X_B X_B^\top)) \leq 4(B + 1.01M^2\epsilon^2 B),$$

where the last step follows due to Lemma 5 which holds with probability at least $1 - \exp(-\Omega(n-d))$ and finishes the proof. □

C.2 Convergence with respect to Huber and Absolute Loss

A relatively straightforward application of Theorem 1 alongwith some Lipschitzness properties allows us to show that STIR and STIR-GD also ensure convergence to the optimal objective value with respect to the Huber and absolute loss functions. These are widely used in robust regression applications.

Theorem 11. *Under the same preconditions as those in Theorem 1, we are assured with probability at least* $1 - \exp(-\tilde{\Omega}(n))$*, that after* $K = \mathcal{O}\left(\log \frac{1}{M_1 \epsilon}\right)$ *stages, both STIR and STIR-GD must produce a model* $\mathbf{w}^K$ *so that*

1. $\ell_\epsilon(\mathbf{w}^K) \leq \ell_\epsilon(\mathbf{w}^*) + \sqrt{1.01}\epsilon$
2. $\frac{1}{n}\left\|X^\top \mathbf{w}^K - \mathbf{y}\right\|_1 \leq \frac{1}{n}\left\|X^\top \mathbf{w}^* - \mathbf{y}\right\|_1 + \frac{3\sqrt{1.01}}{2}\epsilon$.

Proof. The first part follows directly from Lemma 7 and Theorem 1. The second part follows due to the fact that $|x| \leq f_\epsilon(x) \leq |x| + \frac{\epsilon}{2}$ for any $\epsilon > 0$ and thus,

$$\frac{1}{n}\left\|X^\top \mathbf{w}^K - \mathbf{y}\right\|_1 \leq \ell_\epsilon(\mathbf{w}^K) \leq \ell_\epsilon(\mathbf{w}^*) + \sqrt{1.01}\epsilon \leq \frac{1}{n}\left\|X^\top \mathbf{w}^* - \mathbf{y}\right\|_1 + \frac{3\sqrt{1.01}}{2}\epsilon,$$

where the second inequality in the above chain follows from part 1 of this claim. □

D Establishing WSC/WSS - Supplementary Details

Recall that for any $r > 0$ and $M > 0$, $\mathcal{S}_M(r)$ denotes the set of all diagonal M-truncated weight matrices STIR could possibly generate with respect to models residing in the radius R ball centered at $\mathbf{w}^*$ i.e.

$$\mathcal{S}_M(r) := \left\{ S = \operatorname{diag}(\mathbf{s}), \mathbf{s}_i = \min\left\{ \frac{1}{|\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i|}, M \right\}, \mathbf{w} \in \mathcal{B}_2(\mathbf{w}^*, r) \right\},$$

then we have the following result.

Lemma 12. *Suppose the data covariates $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ are generated from an isotropic R-sub-Gaussian distribution $\mathcal{D}$, and G denotes the set of uncorrupted points (as well as the size of that set) then there exists a constant c that depends only on the distribution $\mathcal{D}$ such that for any fixed value of $M > 0$,*

$$\left.\begin{array}{c} \mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99c \cdot GM\right] \\ \mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\max}(X_G S_G X_G^\top) > 1.01 \cdot GM\right] \end{array}\right\} \leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right),$$

where the constants inside $\Omega(\cdot)$ are clarified in the proof. In particular, if $\mathcal{D}$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, I_d)$, then we can take $c = 0.96$.

Proof. The bound for the largest eigenvalue follows directly due to the fact that all weights are upper bounded by M and hence $X_G S_G X_G^\top \preceq M \cdot X_G X_G^\top$ and applying Lemma 5. For the bound on the smallest eigenvalue, notice that Lemma 14 shows us that for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$, i.e. a set of M-truncated weights that correspond to some fixed model $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$, we have

$$\mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995c \cdot GM\right] \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right]$$

Recall that we let $R_X := \max_{i\in[n]} \|\mathbf{x}_i\|_2$ denote the maximum Euclidean length of any covariate. However, Lemma 15 shows us that if $\mathbf{w}^1, \mathbf{w}^2 \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ are two models such that $\left\|\mathbf{w}^1 - \mathbf{w}^2\right\|_2 \leq \tau$ then, conditioned on the value of R_X, the following holds *almost surely*.

$$\left|\lambda_{\min}(X_G S_G^1 X_G^\top) - \lambda_{\min}(X_G S_G^2 X_G^\top)\right| \leq 2G\tau M^2 R_X^3$$

This prompts us to initiate a uniform convergence argument by setting up a τ-net over $\mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ for $\tau = \frac{c}{400R_X^3 M}$. Note that such a net has at most $\left(\frac{800R_X^3}{c}\right)^d$ elements by applying standard covering number bounds for the Euclidean ball [28, Corollary 4.2.13]. Taking a union bound over this net gives us

$$\begin{aligned} \mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99c \cdot GM\right] &\leq 2 \cdot \left(\frac{7200R_X^3}{c}\right)^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right] \\ &\leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right), \end{aligned}$$

where in the last step we used Lemma 6 to bound $R_X = \mathcal{O}\left(R\sqrt{d+n}\right)$ with probability at least $1 - \exp(-\Omega(n))$. For the specific bound on the constant c for various distributions, including the Gaussian distribution, we refer the reader to Section D.1. □

The proof of the above result relies on several intermediate results which we prove in succession below. In the first result Lemma 13, we establish expected bounds on the extremal singular values of the matrix $X_G S_G X_G^\top$ corresponding to a fixed model $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$. In the next result Lemma 14, we establish the same result, but this time with high probability instead of in expectation. The next result Lemma 15 establishes that extremal singular values corresponding to two models close to each other must be (deterministically) close.

Lemma 13 (Pointwise Expectation)**.** *With the same preconditions as in Lemma 12, there must exist a constant $c > 0$ that depends only on $\mathcal{D}$ such that for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$, and fixed vector unit $\mathbf{v} \in S^{d-1}$, we have*

$$c \cdot GM \leq \mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] \leq GM.$$

In particular, if $\mathcal{D}$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, I_d)$, then we can take $c = 0.96$.

Proof. Let $\mathbf{x} \sim \mathcal{D}$ and let $y = \langle \mathbf{w}^*, \mathbf{x} \rangle$. Then if we let $\mathbf{\Delta} := \frac{\mathbf{w}-\mathbf{w}^*}{\|\mathbf{w}-\mathbf{w}^*\|_2}$ (note that $\|\mathbf{w}-\mathbf{w}^*\| \leq \frac{1}{M}$), then we have $s = \min\left\{\frac{1}{|\langle \mathbf{w},\mathbf{x}\rangle - y|}, M\right\} \geq M \cdot \min\left\{\frac{1}{|\langle \mathbf{\Delta},\mathbf{x}\rangle|}, 1\right\}$ as well as $s \leq M$. Then by linearity of expectation we have

$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] = \mathbb{E}\left[\sum_{i\in G} \mathbf{s}_i \langle \mathbf{x}_i, \mathbf{v}\rangle^2\right] = G \cdot \mathbb{E}\left[s \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right] \leq GM \cdot \mathbb{E}\left[\langle \mathbf{x}, \mathbf{v}\rangle^2\right] = GM,$$

since $\mathcal{D}$ is isotropic. We also get

$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] = G \cdot \mathbb{E}\left[s \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right] \geq GM \cdot \mathbb{E}\left[\min\left\{\frac{1}{|\langle \mathbf{\Delta}, \mathbf{x}\rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right] \geq c \cdot GM,$$

where, for any distribution $\mathcal{D}$ over $\mathbb{R}^d$, we define the constant c as

$$c := \inf_{\mathbf{u},\mathbf{v}\in S^{d-1}} \left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\min\left\{\frac{1}{|\langle \mathbf{u}, \mathbf{x}\rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right]\right\}.$$

This concludes the proof. For the specific bound on the constant c for various distributions, including the Gaussian distribution, we refer the reader to Section D.1. □

Lemma 14 (Pointwise Convergence)**.** *With the same preconditions as in Lemma 12, for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$,*

$$\left.\begin{matrix}\mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995c \cdot GM\right] \\ \mathbb{P}\left[\lambda_{\max}(X_G S_G X_G^\top) > 1.005 \cdot GM\right]\end{matrix}\right\} \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right]$$

Proof. Note that for any square symmetric matrix $A \in \mathbb{R}^{d\times d}$, we have $c - \delta \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq c + \delta$ for some $\delta > 0$ iff $\left|\mathbf{v}^\top A \mathbf{v} - c\right| \leq \delta$ for all $\mathbf{v} \in S^{d-1}$ which itself happens iff $\|A - c \cdot I\|_2 \leq \delta$. Now, if $\mathcal{N}_\epsilon$ denotes an ϵ-net over S^{d-1}, then for any square symmetric matrix $B \in \mathbb{R}^{d\times d}$, we have $\|B\|_2 \leq (1-2\epsilon)^{-1} \sup_{\mathbf{v}\in\mathcal{N}_\epsilon} \left|\mathbf{v}^\top B \mathbf{v}\right|$. Thus, setting $B = A - c \cdot I$ and $\epsilon = 1/4$, we have $\|A - c \cdot I\|_2 \leq 2 \sup_{\mathbf{v}\in\mathcal{N}_{1/4}} \left|\mathbf{v}^\top A \mathbf{v} - c\right|$.

Let $\mathbf{x} \sim \mathcal{D}$ and $t = \sqrt{\min\left\{\frac{1}{|\langle \mathbf{w}-\mathbf{w}^*, \mathbf{x}\rangle|}, M\right\}} \leq \sqrt{M}$ and for any fixed $\mathbf{v} \in S^{d-1}$, let $Z := t \cdot \langle \mathbf{x}, \mathbf{v}\rangle$. Then we have

$$\|Z\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}\left[|Z|^p\right]\right)^{1/p} \leq \sqrt{M} \cdot \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}\left[|\langle \mathbf{x}, \mathbf{v}\rangle|^p\right]\right)^{1/p} = R\sqrt{M},$$

where the last step follows by observing that since $\mathcal{D}$ is R-sub-Gaussian, $\|\langle \mathbf{x}_1, \mathbf{v}\rangle\|_{\Psi_2} \leq R$. Thus, Z is $R\sqrt{M}$-sub-Gaussian. This implies Z^2 is MR^2-subexponential (see [28, Lemma 2.7.6]), as well as $Z^2 - \mathbb{E}Z^2$ is $2MR^2$-subexponential by centering and applying the triangle inequality. Note that Lemma 13 implicitly establishes that $\mu := \mathbb{E}Z^2 \in [cM, M]$. Let $Z_1, Z_2, \ldots, Z_G$ be independent realizations of Z with respect to a fixed vector $\mathbf{v}$. Then we have

$$\begin{aligned}\mathbb{P}\left[\left|\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v} - G\mu\right| \geq \varepsilon \cdot GM\right] &= \mathbb{P}\left[\left|\sum_{i\in G}(Z_i^2 - \mu)\right| \geq \varepsilon \cdot GM\right] \\ &\leq 2\exp\left[-m \cdot \min\left\{\frac{(\varepsilon \cdot GM)^2}{4M^2R^4G}, \frac{\varepsilon \cdot GM}{2MR^2}\right\}\right] \\ &\leq 2\exp\left[-\frac{mn\varepsilon^2}{8R^4}\right]\end{aligned}$$

where $m > 0$ is a universal constant and in the last step we used $G \geq n/2$ and w.l.o.g. we assumed that $\varepsilon \leq 2R^2$. Taking a union bound over all 9^d elements of $\mathcal{N}_{1/4}$, we get

$$\begin{aligned}\mathbb{P}\left[\left\|X_G S_G X_G^\top - G\mu \cdot I\right\|_2 \geq \varepsilon \cdot GM\right] &\leq \mathbb{P}\left[\max_{\mathbf{v}\in\mathcal{N}_{1/4}} \left|\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v} - G\mu\right| \geq \frac{\varepsilon}{2} \cdot GM\right] \\ &\leq 2 \cdot 9^d \exp\left[-\frac{mn\varepsilon^2}{8R^4}\right]\end{aligned}$$

Setting $\varepsilon = 0.005c$ and noticing that $\mu \in [cM, M]$ by Lemma 13 finishes the proof. □

Lemma 15 (Approximation Bound)**.** *Consider two models* $\mathbf{w}^1, \mathbf{w}^2 \in \mathbb{R}^d$ *such that* $\left\|\mathbf{w}^1 - \mathbf{w}^2\right\|_2 \leq \tau$ *and let* $\mathbf{s}^1, \mathbf{s}^2$ *denote the* M*-truncated weight vectors they induce i.e.* $s_i^j = \min\left\{M, \frac{1}{|\langle \mathbf{w}^j, \mathbf{x}_i\rangle - y_i|}\right\}, j = 1, 2$. *Also let* $S^1 = diag(\mathbf{s}^1)$ *and* $S^2 = diag(\mathbf{s}^2)$. *Then for any* $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ *such that* $\|\mathbf{x}_i\|_2 \leq R_X$ *for all* i*,*

$$\left|\lambda_{\min}(XS^1X^\top) - \lambda_{\min}(XS^2X^\top)\right| \leq 2n\tau M^2 R_X^3$$

Proof. We have the following four cases with respect to the weights $s_i^j = \min\left\{M, \frac{1}{|\langle \mathbf{w}^j, \mathbf{x}_i\rangle - y_i|}\right\}, j = 1, 2$ these two models generate on any data point $\mathbf{x}_i \in \mathcal{B}_2(R_X)$. Note that we do not assume that these data points are generated from $\mathcal{D}$, just that they are bounded inside the ball $\mathcal{B}_2(R_X)$. Also note that although $\left|s_i^1 - s_i^2\right| \leq M$ trivially holds by virtue of truncation, such a result is not sufficient for us since our later analyses would like to be able to show $\left|s_i^1 - s_i^2\right| \leq \frac{M}{1000}$ by setting τ to be really small.

Case 1 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$ and $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$. Here $s_i^1 = s_i^2 = M$ i.e. $\left|s_i^1 - s_i^2\right| = 0$.

Case 2 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$ but $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$. In this case $s_i^2 = M > s_i^1$. Thus,

$$\left|s_i^1 - s_i^2\right| = M - \frac{1}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i|} \leq M - \frac{1}{|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i| + \tau R_X} \leq M - \frac{M}{1 + \tau M R_X} < 2\tau M^2 R_X$$

Case 3 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$ but $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$. This is similar to Case 2 above.

Case 4 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$ and $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$. In this case we have

$$\left|\frac{1}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i|} - \frac{1}{|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i|}\right| \leq \frac{\left|\langle \mathbf{w}^1 - \mathbf{w}^2, \mathbf{x}_i\rangle\right|}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i| \cdot |\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i|} \leq 2\tau M^2 R_X$$

This tells us that $\left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \leq 2n\tau M^2 R_X$. Now, if we let $S^1 = \mathrm{diag}(\mathbf{s}^1)$ and $S^2 = \mathrm{diag}(\mathbf{s}^2)$, then for any unit vector $\mathbf{v} \in S^{d-1}$, denoting $R_X := \max_{i\in[n]} \|\mathbf{x}_i\|_2$ we have

$$\left|\mathbf{v}^\top XS^1X^\top\mathbf{v} - \mathbf{v}^\top XS^2X^\top\mathbf{v}\right| = \left|\sum_{i=1}^n \left(\mathbf{s}_i^1 - \mathbf{s}_i^2\right)\langle \mathbf{x}_i, \mathbf{v}\rangle^2\right| \leq \left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \cdot \max_{i\in[n]} \langle \mathbf{x}_i, \mathbf{v}\rangle^2 \leq \left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \cdot R_X^2 \leq 2n\tau M^2 R_X^3.$$

This proves that $\left\|XS^1X^\top - XS^2X^\top\right\|_2 \leq 2n\tau M^2 R_X^3$ and concludes the proof. □

D.1 Calculation of Distribution-specific Constants

The WSC/WSS bounds from Lemma 12 are parametrized by a constant c that lower bounds on the singular values of the matrix $X_G S_G X_G^\top$. Recall that for any covariate distribution $\mathcal{D}$, the constant is defined as

$$c := \inf_{\mathbf{u},\mathbf{v}\in S^{d-1}} \left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\min\left\{\frac{1}{|\langle \mathbf{u}, \mathbf{x}\rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right]\right\}.$$

Below we present some interesting cases where this constant is lower bounded.

Centered Isotropic Gaussian For the special case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, I_d)$, notice that by rotational symmetry, we can, without loss of generality, take $\mathbf{u} = (1, 0, 0, \ldots, 0)$ and $\mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ where $v_1^2 + v_2^2 = 1$. Thus, if we consider $x_1, x_2 \sim \mathcal{N}(0, 1)$ i.i.d. then $c \geq \inf_{(v_1,v_2)\in S^1} f(v_1, v_2)$ where

$$\begin{aligned}
f(v_1, v_2) &= \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2x_1^2 + v_2^2x_2^2 + 2v_1v_2x_1x_2)\right] \\
&= \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2x_1^2 + v_2^2x_2^2)\right] \\
&= \mathop{\mathbb{E}}_{x_1\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2x_1^2 + v_2^2)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{2}{\pi}} \left(\int_0^1 (v_1^2 t^2 + v_2^2) e^{-t^2/2} dt + \int_1^\infty \left(v_1^2 t + \frac{v_2^2}{t} \right) e^{-t^2/2} dt \right) \\
&\geq 0.6827 \cdot v_1^2 + 0.9060 \cdot v_2^2
\end{aligned}
$$

where in the second step we used the independence of x_1, x_2 and $\mathbb{E}[x_2] = 0$, in the third step we used independence once more and $\mathbb{E}\left[x_2^2\right] = 1$, and in the last step we used standard bounds on the error function and the exponential integral. This gives us $c \geq \inf_{(v_1,v_2)\in S^1} \left\{0.6827 \cdot v_1^2 + 0.9060 \cdot v_2^2\right\} \geq 0.68$.

Centered Non-isotropic Gaussian For the case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, \Sigma)$, we have $\mathbf{x} \sim \mathcal{D} \equiv \Sigma^{1/2} \cdot \mathcal{N}(\mathbf{0}, I_d)$. Thus, for any fixed unit vector $\mathbf{v}$, we have $\langle \mathbf{v}, \mathbf{x} \rangle \sim \langle \tilde{\mathbf{v}}, \mathbf{z} \rangle$ where $\tilde{\mathbf{v}} = \Sigma^{-1/2}\mathbf{v}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$. We also have $\|\tilde{\mathbf{v}}\|_2 \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$ where $\lambda = \lambda_{\min}(\Sigma)$ and $\Lambda = \lambda_{\max}(\Sigma)$. Note that we must insist on having $\lambda = \lambda_{\min}(\Sigma) > 0$ failing which, as the calculations show below, there is no hope of expecting c to be bounded away from 0. Now for any fixed vectors $\mathbf{u}, \mathbf{v}$ we first perform rotations so that we have $\tilde{\mathbf{u}} = (u, 0, 0, \ldots, 0)$ and $\tilde{\mathbf{v}} = (v_1, v_2, 0, 0, \ldots, 0)$ where we can assume w.l.o.g. that $u \geq 0$. Note that since $\{\|\tilde{\mathbf{u}}\|_2, \|\tilde{\mathbf{v}}\|_2\} \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$, we have $(v_1, v_2) \in S^r$ and $r, u \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$. This gives us $c \geq \inf_{(v_1,v_2)\in S^r} f(v_1, v_2)$ where

$$
\begin{aligned}
f(v_1, v_2) &= \mathop{\mathbb{E}}_{x_1, x_2 \sim \mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{u \cdot |x_1|}, 1 \right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2 + 2 v_1 v_2 x_1 x_2) \right] \\
&= \mathop{\mathbb{E}}_{x_1, x_2 \sim \mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{u \cdot |x_1|}, 1 \right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2) \right] \\
&= \mathop{\mathbb{E}}_{x_1 \sim \mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{u \cdot |x_1|}, 1 \right\} \cdot (v_1^2 x_1^2 + v_2^2) \right] \\
&= \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(\int_0^{\frac{1}{u}} u(v_1^2 t^2 + v_2^2) e^{-t^2/2} dt + \int_{\frac{1}{u}}^\infty \left(v_1^2 t + \frac{v_2^2}{t} \right) e^{-t^2/2} dt \right) \\
&\geq \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(\int_0^{\frac{1}{u}} u(v_1^2 t^2 + v_2^2) e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} dt + v_1^2 e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} + \frac{v_2^2}{2} \int_{\frac{1}{2}\left(\frac{1}{u}\right)^2}^\infty \frac{1}{z} e^{-z} dz \right) \\
&\geq \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} \left(\frac{v_1^2}{3u^2} + v_2^2 \right) + v_1^2 e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} + \frac{v_2^2}{4} e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} \log\left(1 + 4u^2\right) \right) \\
&\geq \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} \left(v_1^2 \left(1 + \frac{\lambda}{3} \right) + v_2^2 \left(1 + \frac{1}{4} \log\left(1 + \frac{4}{\Lambda} \right) \right) \right) \\
&\geq \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} (v_1^2 + v_2^2) \\
&= r^2 \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} \\
&\geq \frac{1}{\Lambda}\sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}}
\end{aligned}
$$

where in the second and third steps we used independence of x_1, x_2, $\mathbb{E}[x_2] = 0$ and $\mathbb{E}\left[x_2^2\right] = 1$ as before, and in the sixth step we used lower bounds on the exponential integral.

Non-centered Isotropic Gaussian We discuss two techniques to handle the case of non-centered covariates.

- **Pairing Trick** This technique requires changes to the data points and relies on the fact that the difference of two i.i.d. non-centered Gaussian random variables is a centered Gaussian random variable with double the variance. Thus, given n covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, I_d)$ and corresponding responses $y_1, \ldots, y_n$, create $n/2$ data points (assume without loss of generality that n is even) $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mathbf{x}_{i+n/2}}{\sqrt{2}}$ and $\tilde{y}_i = \frac{y_i - y_{i+n/2}}{\sqrt{2}}$. Clearly $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, 2 \cdot I_d)$. However, this method has drawbacks since it is likely to increase the proportion of corrupted data points. If α fraction of the original points were corrupted, at most 2α fraction of the new points would be corrupted.

- **Direct Centering** Suppose we have data from a distribution $\mathcal{D} = \mathcal{N}(\boldsymbol{\mu}, I_d)$. As earlier, by rotational symmetry, we can take $\mathbf{u} = (1, 0, 0, \ldots, 0), \mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, 0, 0, \ldots, 0)$. Assume $\|\boldsymbol{\mu}\|_2 = \rho$ and, without loss of generality, $\rho \geq 2$. Letting $\langle \boldsymbol{\mu}, \mathbf{v} \rangle =: p \leq \rho$ and $x_1, x_2, x_3 \sim \mathcal{N}(0,1)$ i.i.d. gives $c \geq \inf_{(v_1,v_2)\in S^1} f(v_1, v_2)$ where, as before, independence of x_1, x_2, x_3 and the fact that $\mathbb{E}[x_2] = 0$ and $\mathbb{E}[x_2^2] = 1$, gives us

$$f(v_1, v_2) = \mathop{\mathbb{E}}_{x_1\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{|x_1 + \mu_1|}, 1 \right\} \cdot ((p + v_1 x_1)^2 + v_2^2) \right]$$

Now, since $(v_1, v_2) \in S^1$ we get two cases (recall that we have assumed w.l.o.g. $\rho \geq 2$)

Case 1: $v_2^2 \geq \frac{1}{2}$ In this case $f(v_1, v_2) \geq \frac{1}{2} \mathop{\mathbb{E}}_{x_1\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{|x_1+\mu_1|}, 1 \right\} \right] \geq \Omega\left(\exp^{-\rho^2/2} \log\left(1 + \frac{1}{\rho^2}\right) \right)$.

Case 2: $v_1^2 \geq \frac{1}{2}$ In this case, if $x_1 \geq 2\sqrt{2}\rho$, then $|v_1 x_1 + p| \geq \frac{v_1 x_1}{2}$, as well as $|x_1 + \mu_1| \leq 2x_1$.

$$\begin{aligned} f(v_1, v_2) &\geq \mathop{\mathbb{E}}_{x_1\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{|x_1 + \mu_1|}, 1 \right\} (p + v_1 x_1)^2 \cdot \mathbb{I}\left\{ x_1 \geq 2\sqrt{2}\rho \right\} \right] \\ &\geq \mathop{\mathbb{E}}_{x_1\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{2x_1}, 1 \right\} \frac{x_1^2}{8} \cdot \mathbb{I}\left\{ x_1 \geq \max 2\sqrt{2}\rho \right\} \right] \geq \frac{1}{16} e^{-4\rho^2} \end{aligned}$$

Since the value ρ influences the final bound on c very heavily, it is advisable to avoid a large ρ value. One way to ensure this is to algorithmically center the covariates i.e. use $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ where $\hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$. This would (approximately) center the covariates and ensure an effective value of $\rho \approx \mathcal{O}\left(\sqrt{\frac{d}{n}}\right)$

Bounded Sub-Gaussian Suppose our covariate distribution has bounded support i.e. $\mathrm{supp}(\mathcal{D}) \subset \mathcal{B}_2(\rho)$ for some $\rho > 0$. Assume $\rho \geq 1$ w.l.o.g. Also, using the centering trick above, assume that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}] = \mathbf{0}$. Then we have $|\langle \mathbf{u}, \mathbf{x} \rangle| \leq \rho$ which implies $\min\left\{ \frac{1}{|\langle \mathbf{u}, \mathbf{x} \rangle|}, 1 \right\} \geq \frac{1}{\rho}$. Let Σ denote the covariance of the distribution $\mathcal{D}$ and let $\lambda := \lambda_{\min}(\Sigma)$ denote its smallest eigenvalue. This gives us $c \geq \frac{1}{\rho} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} \left[\langle \mathbf{x}, \mathbf{v} \rangle^2 \right] \geq \frac{\lambda}{\rho}$.

E Corruptions and Dense Noise - Supplementary Details

In this section, we will provide details of the convergence analysis of STIR and STIR-GD in the setting where even the "good" points experience sub-Gaussian noise. Thus, we will assume that our data is generated as $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b} + \boldsymbol{\epsilon}$ where, as before $\|\mathbf{b}\|_0 \leq \alpha \cdot n$ and $\boldsymbol{\epsilon} \sim \mathcal{D}_\varepsilon$ where $\mathcal{D}_\varepsilon$ is a σ-sub-Gaussian distribution with zero mean and real support. As mentioned before, we can tolerate noise with non-zero mean as well, by using the same pairing trick we used to center the covariates in Appendix D.1. This would have a side effect of at most doubling the corruption rate α. We will denote, as before $B := \mathrm{supp}(\mathbf{b})$ and $G := [n] \setminus B$. Our covariates will continue to be sampled from an R sub-Gaussian distribution $\mathcal{D}$ with support over $\mathbb{R}^d$. We (re)state the main result of this section below.

Theorem 2. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted with the rest subjected to sub-Gaussian noise sampled from a distribution $\mathcal{D}_\varepsilon$ with sub-Gaussian norm σ. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$, where $c_\varepsilon > 0$ is a constant that depends only on the distributions $\mathcal{D}$ and $\mathcal{D}_\varepsilon$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n) + \log\frac{1}{M_1\sigma}\right)\right)$, after $K = \mathcal{O}\left(\log\frac{1}{M_1\sigma}\right)$ stages, each of which has only $\mathcal{O}(1)$ iterations, we must have $\left\|\mathbf{w}^K - \mathbf{w}^*\right\|_2 \leq \mathcal{O}(\sigma)$.*

Proof. The overall proof of this result follows exactly the same way as the result in Theorem 1. We will still utilize the notion of a *well-initialized stage* and establish (see Lemma 16 below) a convergence guarantee for each well-initialized stage. However, Lemma 16 will itself require a few new results to be proved.

However, note that Lemma 8, a similar result for well-initialized stages in the setting without dense noise, required two results, namely Lemmata 12 and 10 that established the WSC/WSS properties and bounded the weight put

on bad points. Those results implicitly assumed that good points incur absolutely no modification to their response value which is no longer true here since in the setting being considered here, even good points do incur sub-Gaussian noise in their responses. Thus, we will establish below Lemmata 17 and 18 which will establish those results in the dense noise setting. We note that a similar convergence guarantee may be established for STIR-GD in the dense noise setting as well.

However, note that this result only guarantees a convergence to $\left\|\mathbf{w}^{K,1} - \mathbf{w}^*\right\|_2 \leq \mathcal{O}(\sigma)$ and thus, does not ensure a consistent solution. A technical reason for this is because Lemma 17 holds true only for values of $M \leq \mathcal{O}\left(\frac{1}{\sigma}\right)$ which restricts the application of this result to offer errors much smaller than σ. It would be interesting to show, as [5] do, that STIR, or a variant, does offer consistent estimates.

For sake of notational simplicity, we will assume that $\boldsymbol{\epsilon}_B = \mathbf{0}$ by shifting any sub-Gaussian noise a bad point, say $j \in B$ does incur, into the corruption value corresponding to that point i.e. $\mathbf{b}_j$. This is without loss of generality since we impose no constraints on the corruptions other than that they be sparse, in particular the corruptions need not be bounded and can thus, absorb sub-Gaussian noise values into them. □

Lemma 16. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted with the rest experiencing noise generated i.i.d. from a distribution $\mathcal{D}_\varepsilon$ with sub-Gaussian norm σ. Suppose we initialize a stage T within an execution of STIR with truncation level $M \leq \frac{c_\varepsilon}{8\eta\sigma}$, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^T T, 1$ such that $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$. Here c_ε is the constant of the WSC property and depends only on the distributions $\mathcal{D}$ and $\mathcal{D}_\varepsilon$ (see Lemma 17).*

Proof. Let $\mathbf{w}^{T,\tau}$ be a model encountered by STIR within this stage and let $\mathbf{r} = X^\top \mathbf{w}^{T,\tau} - \mathbf{y}$ denote the residuals due to $\mathbf{w}^{T,\tau}$ and $S = \text{diag}(\mathbf{s})$ denote the diagonal matrix of weights where $\mathbf{s}_i = \min\left\{\frac{1}{|\mathbf{r}_i|}, M\right\}$. Then STIR will choose as the next model $\mathbf{w}^{T,\tau+1} = (XSX^\top)^{-1}XS\mathbf{y} = \mathbf{w}^* + (XSX^\top)^{-1}XS(\mathbf{b} + \boldsymbol{\epsilon})$ which gives us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{\|XS(\mathbf{b}+\boldsymbol{\epsilon})\|_2}{\lambda_{\min}(XSX^\top)}$$

Now by Lemma 5, with probability at least $1 - \exp(-\Omega(n-d))$, we have $\|X_B\|_2 = \sqrt{\lambda_{\max}(X_B X_B^\top)} \leq \sqrt{1.01B}$. By Lemma 10, with the same probability, we have

$$\|S\mathbf{b}\|_2 \leq \sqrt{4B(1 + 1.01M^2\|\mathbf{w} - \mathbf{w}^*\|_2^2)} \leq 2\sqrt{2.01B},$$

whereas by Lemma 18, as we have restricted $M \leq \frac{1}{8\sigma}$, we have, yet again with the same probability,

$$\|XS\boldsymbol{\epsilon}\|_2 = \|X_G S_G \boldsymbol{\epsilon}_G\| \leq 4MG\sigma\sqrt{1.01} \leq \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G,$$

where the first equality follows due to our convention that $\text{supp}(\boldsymbol{\epsilon}) = G$ since for bad points in the set B, we clubbed any sub-Gaussian noise into the corruption itself, thus leaving $\boldsymbol{\epsilon}_B = \mathbf{0}$. Now, by Lemma 17, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, we have $\lambda_{\min}(XSX^\top) \geq \lambda_{\min}(X_G S_G X_G^\top) \geq 0.99c_\varepsilon \cdot GM$. This give us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{2B\sqrt{2.0301} + \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G}{0.99c_\varepsilon \cdot GM} \leq \frac{2B\sqrt{2.0301}}{0.99c_\varepsilon \cdot GM} + \frac{\sqrt{1.01}}{1.98\eta \cdot M}$$

Now, since we have $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$, we also have $\frac{2B\sqrt{2.0301}}{0.99c_\varepsilon \cdot GM} \leq \left(1 - \frac{\sqrt{1.01}}{1.98}\right)\frac{1}{\eta M}$ and thus, $\frac{2B\sqrt{2.0301} + \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G}{0.99c_\varepsilon \cdot GM} \leq \frac{1}{\eta M}$. Arguing as we did in the proof of Lemma 8, we must either have $\|\mathbf{w}^+ - \mathbf{w}^*\|_2 \leq \frac{2B\sqrt{2.0301}}{0.9801c_\varepsilon \cdot GM} + \frac{\sqrt{1.01}}{1.9602\eta \cdot M}$ and if that does not happen, we must instead have

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq 0.99 \cdot \left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2$$

This proves the claimed result. □

E.1 Establishing WSC/WSS in Presence of Dense Noise

We will rework a counterpart to Lemma 12 in this section.

Lemma 17. *Given the problem setting above, then there exists a constant $c_\varepsilon > 0$ that depends only on the distributions $\mathcal{D}, \mathcal{D}_\varepsilon$ such that for any $M \in \left[0, \frac{1}{\sigma}\right]$, we have*

$$\mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99 c_\varepsilon \cdot GM\right] \leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$$

In particular, for standard Gaussian covariates and Gaussian noise with variance σ^2, we can take $c_\varepsilon \geq 0.52$.

Proof. Let $\mathbf{x} \sim \mathcal{D}, \epsilon \sim \mathcal{D}_\varepsilon$ and let $y = \langle \mathbf{w}^*, \mathbf{x}\rangle + \epsilon$ be the response of an uncorrupted data point and $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ be any fixed model. Then if we let $\mathbf{\Delta} := \mathbf{w} - \mathbf{w}^*$, the weight s that the model $\mathbf{w}$ would cause STIR to put on this (clean) data point must satisfy $s \geq \min\left\{\frac{1}{|\langle \mathbf{\Delta}, \mathbf{x}\rangle - \epsilon|}, M\right\}$. This gives us, for any fixed $\mathbf{v} \in S^{d-1}$,

$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] \geq c_\varepsilon \cdot GM,$$

where we define,

$$c_\varepsilon := \inf_{\substack{0 \leq r \leq \frac{1}{M} \\ \mathbf{u}, \mathbf{v} \in S^{d-1}}} \left\{ \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}, \epsilon\sim\mathcal{D}_\varepsilon}\left[\min\left\{\frac{1}{|Mr\langle \mathbf{u}, \mathbf{x}\rangle - M\epsilon|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right]\right\}$$

We analyze the constant c for the Gaussian case at the end of the proof. For now, we proceed as in Lemma 14 and realize that the sub-Gaussian norm calculations continue to hold in this case since they simply upper bound the weights by M, and get

$$\mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995 c_\varepsilon \cdot GM\right] \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005 c_\varepsilon)^2}{8R^4}\right]$$

After this we notice that the proof of Lemma 15 pays no heed to corruptions or additional noise and hence, continues to hold in this setting too. Proceeding as in the proof of Lemma 12 to set up a τ-net over $\mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ and taking a union bound over this net finishes the proof.

For the special case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, I_d)$ and $\mathcal{D}_\varepsilon = \mathcal{N}(0, \sigma^2)$, by rotational symmetry, we can, without loss of generality, take $\mathbf{u} = (1, 0, 0, \ldots, 0)$ and $\mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ where $v_1^2 + v_2^2 = 1$. Thus, if $x_1, x_2, \epsilon \sim \mathcal{N}(0,1)$ i.i.d. then $c \geq \inf_{(v_1,v_2)\in S^1, r\in\left[0,\frac{1}{M}\right]} f(v_1, v_2, r)$ where

$$\begin{aligned}
f(v_1, v_2, r) &= \mathop{\mathbb{E}}_{x_1,x_2,\epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2 + 2v_1 v_2 x_1 x_2)\right] \\
&= \mathop{\mathbb{E}}_{x_1,x_2,\epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2)\right] \\
&= \mathop{\mathbb{E}}_{x_1,\epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2)\right] \\
&= v_1^2 \cdot \underbrace{\mathop{\mathbb{E}}_{x_1,\epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} x_1^2\right]}_{(A)} + v_2^2 \cdot \underbrace{\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{M\sqrt{r^2+\sigma^2}\,|z|}, 1\right\}\right]}_{(B)}
\end{aligned}$$

where in the second step we used the independence of x_1, x_2 and $\mathbb{E}[x_2] = 0$, in the third step we used independence once more and $\mathbb{E}\left[x_2^2\right] = 1$. In the fourth step, we substituted $\sqrt{r^2+\sigma^2}z = rx_1 - \sigma\epsilon$ and noticed that $rx_1 - \sigma\epsilon \sim \mathcal{N}(0, (r^2+\sigma^2))$ i.e. $z \sim \mathcal{N}(0,1)$. To bound (B) we notice $r \leq \frac{1}{M}$ and $M \leq \frac{1}{\sigma}$ and use standard bounds on Gaussian and exponential integrals to get

$$(B) \geq \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{\sqrt{2}\,|z|}, 1\right\}\right] \geq 0.815$$

To bound (A), we use the fact that pairwise orthogonal projections of a standard Gaussian vector yield independent variables. Thus, if we denote $a = Mr, b = M\sigma$ and $z = \frac{ax_1 - b\epsilon}{\sqrt{a^2+b^2}}, w = \frac{bx_1 + a\epsilon}{\sqrt{a^2+b^2}}$, then $z, w \sim \mathcal{N}(0,1)$ as well as $z \perp w$. Thus, we have

$$\begin{aligned}
(A) &= \mathop{\mathbb{E}}_{z,w\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{M\sqrt{r^2+\sigma^2}\,|z|}, 1 \right\} \cdot \left(\frac{r^2z^2 + \sigma^2w^2 + 2r\sigma zw}{r^2+\sigma^2} \right) \right] \\
&\geq \mathop{\mathbb{E}}_{z,w\sim\mathcal{N}(0,1)} \left[\min\left\{ \frac{1}{\sqrt{2}\,|z|}, 1 \right\} \cdot \left(\frac{r^2z^2 + \sigma^2w^2}{r^2+\sigma^2} \right) \right] \\
&\geq \frac{0.52r^2}{r^2+\sigma^2} + \frac{0.815\sigma^2}{r^2+\sigma^2} = 0.52 + \frac{0.295\sigma^2}{r^2+\sigma^2}
\end{aligned}$$

where in the second step we used $M \leq \frac{1}{\sigma}$ and $r \leq \frac{1}{M}$, independence of z and w and the fact that $\mathbb{E}[w] = 0, \mathbb{E}\left[w^2\right] = 1$ and the last step uses standard bounds on Gaussian and exponential integrals. □

E.2 Bounding the Weights on Good Points

Although Lemma 10 continues to hold in this case, since good points also incur modifications to their response values, albeit modifications that are stochastic and not adversarial, we need an analogous result for the good points in this case as well.

Lemma 18. *Suppose σ is the sub-Gaussian norm of the noise distribution $\mathcal{D}_\varepsilon$ and the identity of the good points G is chosen independently of the covariates. Then for any $M > 0$, if S is the diagonal matrix of M-truncated weights assigned to the data points by a model $\mathbf{w}$, then with probability at least $1 - \exp(-\Omega(n-d))$,*

$$\|X_G S_G \boldsymbol{\epsilon}_G\|_2 \leq 4MG\sigma\sqrt{1.01}$$

Proof. We have, by applying Lemma 5, with probability at least $1 - \exp(-\Omega(n-d))$,

$$\|X_G S_G \boldsymbol{\epsilon}_G\|_2 \leq \sqrt{\lambda_{\max}(X_G X_G^\top)} \cdot \|S_G \boldsymbol{\epsilon}_G\| \leq \sqrt{1.01G} \cdot \|S\|_2 \|\boldsymbol{\epsilon}_G\|_2 \leq \sqrt{1.01G}M \cdot \|\boldsymbol{\epsilon}_G\|_2,$$

where the last inequality follows since S is a diagonal matrix and by M-truncation, the maximum value of any weight is M. Now, since our noise is σ sub-Gaussian and unbiased, we have, for any fixed $\mathbf{u} \in S^{G-1}$, $\mathbb{E}[\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle] = 0$, as well as, by applying the Hoeffding's inequality,

$$\mathbb{P}[|\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle| \geq t] \leq 3\exp\left(-\frac{t^2}{2\sigma^2} \right)$$

Now, if $\mathbf{u}^1, \mathbf{u}^2 \in S^{G-1}$, such that $\left\|\mathbf{u}^1 - \mathbf{u}^2\right\|_2 \leq \frac{1}{2}$, then we have $\left|\left\langle \mathbf{u}^1 - \mathbf{u}^2, \boldsymbol{\epsilon} \right\rangle\right| \leq \frac{1}{2} \cdot \|\boldsymbol{\epsilon}\|_2$. Thus, taking a union bound over a 1/2-net over S^{G-1} gives us

$$\mathbb{P}\left[\|\boldsymbol{\epsilon}\|_2 = \max_{\mathbf{u}\in S^{G-1}} \langle \mathbf{u}, \boldsymbol{\epsilon} \rangle \geq \frac{1}{2} \cdot \|\boldsymbol{\epsilon}\|_2 + t \right] = \mathbb{P}[\|\boldsymbol{\epsilon}\|_2 \geq 2t] \leq 3 \cdot 5^G \exp\left[-t^2/2\sigma^2\right]$$

Setting $t = \sigma\sqrt{4G}$ establishes the result. □

F Robust Linear Bandits

In this section, we briefly discuss the linear contextual bandit problem with corrupted arm pulls. We refer the reader to [19] for a more relaxed introduction to the problem as well as formal regret bounds. Indeed, the discussion here is adapted from the discussion in [19].

F.1 Problem Setting

The stochastic linear contextual bandit framework [1, 20] considers a (possibly infinite) set of *arms*. Arms correspond to various actions that can be performed by the algorithm. For instance, in a recommendation setting, arms may correspond to various products that are available for sale, for instance, at an e-commerce website, or in a quantitative trading setting, arms may correspond to stocks that are available for sale/purchase.

Problem Setting 1 Adversarial Linear Bandits

for $t = 1, 2, 3..$ **do**
  Player receives a set of contexts $A_t = \left\{\mathbf{x}^{t,1}, \ldots, \mathbf{x}^{t,n_t}\right\} \subset \mathbb{R}^d$
  Player plays an arm, $\hat{\mathbf{x}}^t \in A_t$
  Clean reward is generated $r_t^* = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$ conditioned on $\mathcal{H}^t$
  Adversary inspects $\hat{\mathbf{x}}^t, r_t^*, \mathcal{H}^t$ and chooses b_t //`while making sure` $|\tau \leq t : b_\tau \neq 0| \leq \eta \cdot (t+1)$
  Player receives reward, $r_t = r_t^* + b_t$
end for

Algorithm 4 WUCB-Lin: Weighted UCB for Linear Contextual Bandits

Input: Upper bounds σ_0 (on sub-Gaussian norm of noise distribution), B (on magnitude of corruption), α_0 (on fraction of corrupted points), initial truncation M_1, increment rate η

for $t = 1, 2, \ldots, T$ **do**
  Receive set of arms A_t
  Play arm $\hat{\mathbf{x}}^t = \underset{\mathbf{x} \in A_t, \mathbf{w} \in C_{t-1}}{\arg\max} \langle \mathbf{x}, \mathbf{w} \rangle$
  Receive reward r_t
  $(\hat{\mathbf{w}}^t, S^t) \leftarrow \mathsf{STIR}\left(\{\hat{\mathbf{x}}^\tau, r_\tau\}_{\tau=1}^t, M_1, \eta\right)$ //`Denote` $S^t = \mathtt{diag}(s_1^t, s_2^t, \ldots, s_t^t)$
  $V^t \leftarrow \sum_{\tau \leq t} s_\tau^t \hat{\mathbf{x}}^\tau (\hat{\mathbf{x}}^\tau)^\top$, $X^t \leftarrow \left[\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots, \hat{\mathbf{x}}^t\right]$
  $\bar{\mathbf{w}}^t \leftarrow (V^t)^{-1} X^t S^t \mathbf{y}$
  $C_t \leftarrow \{\mathbf{w} : \left\|\mathbf{w} - \bar{\mathbf{w}}^t\right\|_{V^t} \leq \sigma_0 \sqrt{d \log T} + \alpha_0 BT\}$
end for

Every arm $\mathbf{a}$ is parametrized by a vector $\mathbf{a} \in \mathbb{R}^d$ (we abuse notation to denote the arm and its corresponding parametrization using the same notation). Recall that the set of all arms is potentially infinite. However, not all arms may be available at every time step. For instance, an e-commerce website would not like to recommend products not currently in stock. Similarly, stocks not currently in one's possession cannot be sold.

At each time step t, the algorithm receives a set of n_t arms (also called *contexts*) $A_t = \left\{\mathbf{x}^{t,1}, \ldots, \mathbf{x}^{t,n_t}\right\} \subset \mathbb{R}^d$ that can be played or *pulled* in this round. Pulling an arm is akin o performing the action associated with that arm, for example, recommending an item or selling a stock unit. The context set A_t, as well as the number n_t of contexts available can vary across time steps. The algorithm selects and pulls an arm $\hat{\mathbf{x}}^t \in A_t$ as per its arm selection policy. In response, a reward r_t is generated. Let $\mathcal{H}^t = \left\{A_1, \hat{\mathbf{x}}^1, r_1, \ldots, A_{t-1}, \hat{\mathbf{x}}^{t-1}, r_{t-1}, A_t, \hat{\mathbf{x}}^t\right\}$.

F.2 Adversary Model

In the stochastic linear bandit setting, as has been studied in prior work [1, 20] , at every time step, the reward r_t is generated using a *model vector* $\mathbf{w}^* \in \mathbb{R}^d$ (that is not known to the algorithm) as follows: $r_t = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$, where ϵ_t is a *noise* value that is typically assumed to be (conditionally) centered and σ-sub-Gaussian, i.e., $\mathbb{E}\left[\epsilon_t \,|\, \mathcal{H}^t\right] = 0$, as well as for some $\sigma > 0$, we have $\mathbb{E}\left[\exp(\lambda \epsilon_t) \,|\, \mathcal{H}^t\right] \leq \exp(\lambda^2 \sigma^2 / 2)$ for any $\lambda > 0$.

However, recent works [19, 22] have considered settings where the rewards may suffer not only sub-Gaussian noise, but also adversarial corruptions that are introduced by an *adaptive adversary* that is able to view the on-goings of the online process and at any time instant t, *after* observing the history $\mathcal{H}^t$ and the "clean" reward value, i.e., $\langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$, is able to add a corruption value b_t to the reward. For notational uniformity, we will assume that for time instants where the adversary chooses not to do anything, $b_t = 0$. Thus, the final reward to the player at every time step is $r_t = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t + b_t$. This model is described in Problem Setting 1.

For sake of simplicity we will assume that, for some $B > 0$, the final (possibly corrupted) reward presented to the player satisfies $r_t \in [-B, B]$ almost surely. The only constraint the adversary need observe while introducing the corruptions is that at no point in the online process, should the adversary have corrupted more than an η fraction of the observed rewards. Formally, let $G_t = \{\tau < t : b_\tau = 0\}$ and $B_t = \{\tau < t : b_\tau \neq 0\}$ denote the set of "good" and "bad" time instances till time t. We insist that $|B_t| \leq \eta \cdot t$ for all t.

F.3 Notion of Regret

The goal of the algorithm is to maximize the cumulative reward it receives over the time steps $\sum_{t=1}^{T} r_t$. However, a more popular technique of casting this objective is in the form of *cumulative pseudo regret*. At time t, let $\mathbf{x}^{t,*} = \arg\max_{\mathbf{x} \in A_t} \langle \mathbf{w}^*, \mathbf{x} \rangle$ be the arm among those available that yields the highest expected (uncorrupted) reward. The cumulative pseudo regret of a policy π is defined as follows

$$\bar{R}_T(\pi) = \sum_{t=1}^{T} \left\langle \mathbf{w}^*, \mathbf{x}^{t,*} \right\rangle - \mathbb{E}\left[r_t\right].$$

Note that the best arm here may change across time-steps.

F.4 WUCB-Lin: An Algorithm for Robust Linear Bandits

We use the notation $\|\mathbf{x}\|_M = \sqrt{\mathbf{x}^\top M \mathbf{x}}$ for a vector $\mathbf{x} \in \mathbb{R}^d$ and a matrix $M \in \mathbb{R}^{d \times d}$. We reproduce, for convenience, the WUCB-Lin algorithm in Algorithm 4. WUCB-Lin builds upon the OFUL principle [1] for linear contextual bandits. At every step, WUCB-Lin uses rewards obtained from previous arm pulls to obtain an estimate $\hat{\mathbf{w}}^t$ of the true model vector $\mathbf{w}^*$.

Whereas classical algorithms utilize ordinary least squares to solve this problem, WUCB-Lin utilizes STIR (actually STIR-GD for sake of speed) to obtain this estimate. This lends resilience to the algorithm against the (possibly several) past arm pulls whose rewards got corrupted by the adversary. The previous work of [19] used the TORRENT algorithm for the same purpose.

The next step in executing the OFUL principle is the construction of a *confidence set*. It is common to use an ellipsoidal confidence set with the ellipsoid induced by the covariance matrix of the arm vectors pulled so far. The work of [19] modifies this to only consider arms considered as clean by the TORRENT algorithm while constructing the confidence ellipsoid.

Since STIR, instead of selecting a specific subset of arms like TORRENT, instead would assign weights to all previously pulled arms, with a small weight indicating a high likelihood of the arm pull being a corrupted one and a large weight indicating a high likelihood of the arm pull being a clean one. Thus, STIR utilizes these weights to construct a *weighted covariance matrix* which is then used to define the confidence ellipsoid and carry out the arm selection step.