Supplementary to "Empirical Risk Minimization for Losses without Variance"

This supplementary is organized as follows. Section A gives two Catoni-type influence functions that satisfy the Hölder's continuity assumption. Section B provides three examples to explain the upper bounds obtained in the main theorems. Additional simulation results are given in Section C. All technical proofs are collected from Sections D - H. Additional discussions on algorithms are given in Section I.

A Examples satisfying Hölder continuity

The following two special functions have Hölder's continuity.

1. (unbounded case):

$$\phi_1(x) = \begin{cases} \log(1 + x + C_\varepsilon |x|^{1+\varepsilon}) & x \geq 0 \\ -\log(1 - x + C_\varepsilon |x|^{1+\varepsilon}) & x < 0. \end{cases} \tag{S1}$$

2. (bounded case):

$$\phi_2(x) = \begin{cases} -\log(1 + A_2 + C_\varepsilon A_2^{1+\varepsilon}) & \text{if } x \leq -A_2 \\ -\log(1 - x + C_\varepsilon |x|^{1+\varepsilon}) & \text{if } -A_2 \leq x \leq 0, \\ \log(1 + x + C_\varepsilon |x|^{1+\varepsilon}) & \text{if } 0 < x \leq A_1, \\ \log(1 + A_1 + C_\varepsilon A_1^{1+\varepsilon}) & \text{if } x \geq A_1 \end{cases} \tag{S2}$$

with both $|A_1|, |A_2| > ((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}$.

B Illustrative Examples

The messages from our main theorems are

a Theorem 3 applies to the settings when losses could be unbounded without variance, but the differences of two loss functions (i.e., $|f(X) - f'(X)|$) are bounded.

b Theorem 4 can be applied to the settings even if the differences of two loss functions are unbounded.

In this section, we provide several examples to help readers to understand our theoretical results.

L_1 **regression.** In this setting, we let $\mathcal{F} = \{f_g(z, y) = |g(z) - y| : g \in \mathcal{G}\}$ and assume that $E\big|g(Z) - Y|^{1+\varepsilon} \le v$ for every $g \in \mathcal{G}$. For the maximum distance, since

$$D(f_g, f_{g'}) = \sup_{z,y} ||g(z) - y| - |g'(z) - y|| \le d_\infty(g, g'),$$

the covering number of $\mathcal{F}$ under the distance D is bounded by the covering number of $\mathcal{G}$ under the sup norm. Similarly, for the norm d_p with $p = 1 + \varepsilon$, we have

$$d_p(f_g, f_{g'}) = (\mathbb{E}|f_g(X) - f_{g'}(X)|^p)^{1/p} \le (\mathbb{E}|g(z) - g'(z)|^p)^{1/p} = d_p(g, g').$$

Hence the covering number of $\mathcal{F}$ under distance the d_p is bounded by the covering number of $\mathcal{G}$ under the same distance. Applying Theorem **??**, we obtain the following result.

Proposition 1. *In the L_1-regression problem,*

$$m_{\hat{f}} - m^* \le 6L_\varepsilon \left(2^{2+\varepsilon} C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n}\right) + C' \log(2/\delta)\Big(\frac{2\alpha^{(\varepsilon-1)/2}}{3n}\gamma_{1,\varepsilon}(\mathcal{G}, d_\infty) + \sqrt{\frac{\alpha^{\varepsilon-1}}{n}}\gamma_{2,\varepsilon}(\mathcal{G}, d_p)\Big) \tag{S3}$$

with probability $1 - \delta$ for any n that satisfies the (α, δ) condition and the η-condition for

$$\eta = 2L_\varepsilon A'_\alpha(\delta) + C' \log(2/\delta)\Big(\frac{2\alpha^{(\varepsilon-1)/2}}{3n}\gamma_{1,\varepsilon}(\mathcal{G}, d_\infty) + \sqrt{\frac{\alpha^{\varepsilon-1}}{n}}\gamma_{2,\varepsilon}(\mathcal{G}, d_p)\Big)$$

with $C' = 384 C_{3\varepsilon} \log 2$ and $p = 1 + \varepsilon$.

L_2 **regression.** In this setting, we let $\mathcal{F} = \{f_g(z, y) = (g(z) - y)^2 : g \in \mathcal{G}\}$ with $d_\infty(g, g')$ being bounded and apply Theorem **??** with some straightforward calculations to get the next result.

Proposition 2. *In the L_2-regression problem, it holds that with $\mathbb{E}[|f_g|^{1+\varepsilon}] \le \infty$ for any*

$f_g \in \mathcal{F}$.

$$
\begin{aligned}
& m_{\hat{f}} - m^* \qquad \text{(S4)} \\
\leq\ & 6L_\varepsilon(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n}) \\
& + KC_{3\varepsilon}\sqrt{\frac{\log(8/\delta)}{n\alpha^{1-\varepsilon}}}(\Delta^{1+\varepsilon} + \mathbb{E}[|Y|^{1+\varepsilon}] + \sqrt{8v/n\delta})^{1/(1+\varepsilon)}\gamma_{2,\varepsilon}(\mathcal{G}, d_\infty)
\end{aligned}
$$

with probability $1-2\delta$ for any $n \geq N_0$ and a universal constant K. (N_0 is still a positive constant satisfying (α,δ) condition and η-condition and Δ is a positive constant larger than $diam_{d_p}(\mathcal{F})^{(1+\varepsilon)/2}$.)

Remark S1. *The above result applies to the special linear model $Y = \beta^T X + \epsilon$ with $\mathbb{E}[|\epsilon|^{2(1+\varepsilon)}] < \infty$. Compared to the state of art result (Hsu & Sabato, 2016), our result is established under an even weaker moment condition, that is, the fourth moment of error term ϵ does not exist.*

Kernel Learning. Consider the following optimization problem,

$$
\hat{f} = \arg \min_{f = L \circ h \in \mathcal{F}} \left\{ \hat{\mu}_f + \lambda_n \|h\|_{\mathcal{H}}^2 \right\}, \qquad \text{(S5)}
$$

where $\mathcal{F} = L \circ \mathcal{H}$, where L is a deterministic loss function and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) associated with kernel $K(x,y)$. In this section, we assume $L \circ \mathcal{H}$ is $L_{1+\varepsilon}$-integrable and takes the form that $L(Y - h(X))$, loss function L satisfies that $|L(Y - h_1(X)) - L(Y - h_2(X))| \leq C(Y)|h_1(X) - h_2(X)|$ for any $h_1, h_2 \in \mathcal{H}$ where $C(Y)$ is an square integrable function. Kernel K is assumed to be a Mercer kernel. Moreover, without loss of generality, we can always assume the true underlying h^* has a bounded norm, and particularly we assume $\|h^*\|_{\mathcal{H}}^2 \leq 1$.

Proposition 3. *In the kernel regression problem described as above, it holds that*

$$m_{\hat{f}} - m^* \leq 6L_\varepsilon(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n} + \lambda_n) \\ + KC_{3\varepsilon}\log(2/\delta)(\frac{2\alpha^{(\varepsilon-1)/2}}{3n}\gamma_{1,\varepsilon}(L\circ\mathcal{H}, D) + \sqrt{\frac{\alpha^{\varepsilon-1}}{n}}\gamma_{2,\varepsilon}(L\circ\mathcal{H}, d_p))$$

with probability $1-2\delta$ *for any* $n \geq N_0$ *and a universal constant* K*. (*N_0 *is a large constant satisfying* (α, δ) *and* η*-condition.)*

Remark S2. *By taking* $\mathcal{H} = \{h(Z) \mid \beta^T Z, , \beta \in \mathbb{R}^d\}$ *with kernel* $K(f_1, f_2) = \beta_1 \cdot \beta_2$*,* $\mathcal{F} = \{f(X)|(Y - h(Z))^2; h \in \mathcal{H}\}$ *and influence function* $\phi(x) = x$*, then* (S5) *is reduced to the standard ridge regression*

$$\arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T Z_i)^2 + \lambda_n \|\beta\|_2^2.$$

Remark S3. *In deep learning, the RKHS can be taken as the space spanned by ReLU functions.*

C Additional Simulations

C.1 Regression with Contamination

We next consider a regression problem with contamination, where we in particular assume that the clean data follows $Y_i = X_i^T w_* + \xi_i$, where ξ_i's are standard normal random variables. The data are contaminated in the following fashion. $\tilde{Y}_i = Y_i$ with probability $1-\eta$ and $\tilde{Y}_i = (2u_i - 1)\tilde{\xi}_i$ with probability η. Here $\eta \in (0, 1)$ is the contamination rate and $\tilde{\xi}_i$, u_i are the same as in the previous setting. In this scenario, we fix $d = 8$ and choose contamination probability $\eta \in \{5\%, 10\%, 20\%, 30\%, 40\%\}$. The choices of tail parameter, sample size and w^* remain the same as in the simulation section of the main paper. The results of estimation errors are shown in Figure S1.

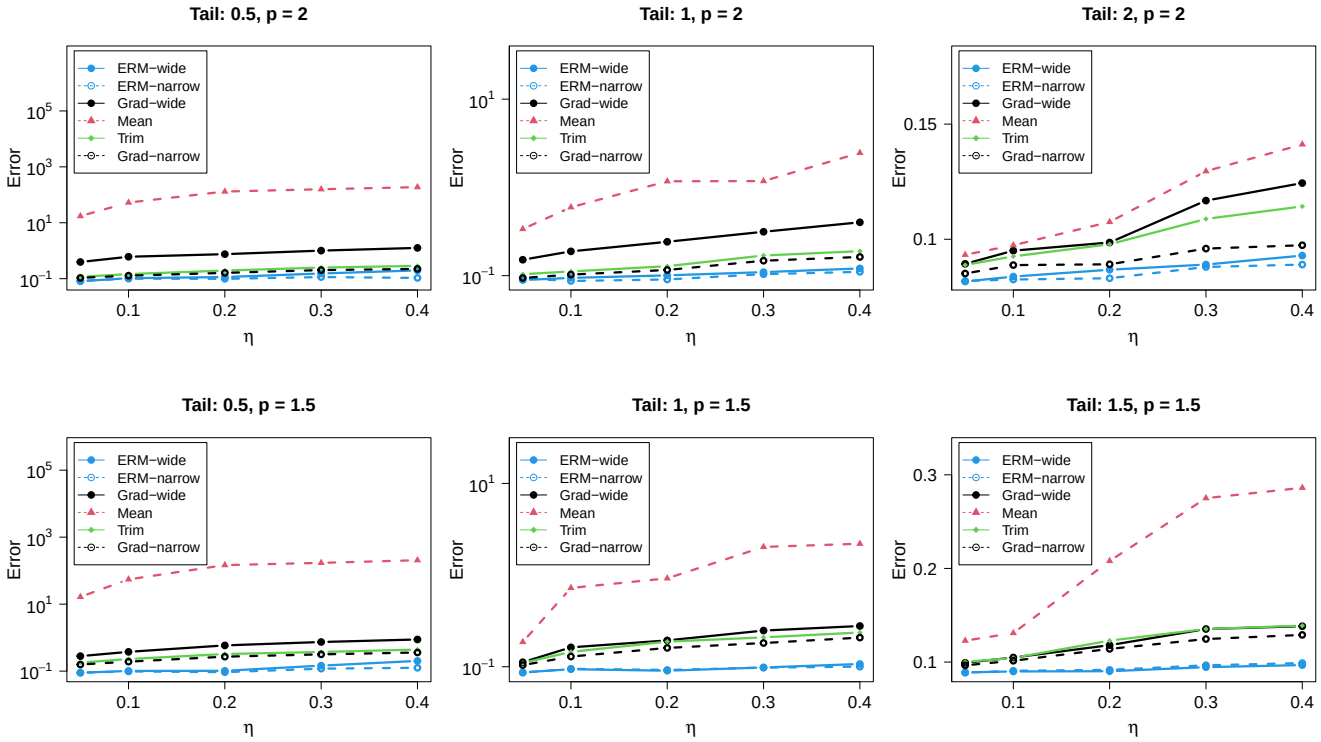

Figure S1: Comparison between six methods in regression problems under different contamination rates and shape parameters.

C.2 K-means Clustering

We next consider a K-means clustering problem in this section. The data generation is described as follows,

$$Y_i = W^*_{c_i} + \boldsymbol{\xi}_i \text{ with } c_i \sim \text{Multinom}(1, 1, \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ such that $0 < \pi_k < 1$ and $\sum_{k=1}^K \pi_k = 1$. $W^* = (W^*_k)$ is a d by k matrix. $\boldsymbol{\xi}_i \in \mathbb{R}^d$ and each of its coordinates follows a symmetrized Pareto random variables as described before.

For optimization, we postulate the following formulation.

$$\min_W \sum_i \min_{k\in[K]} l(Y_i, W_k), \tag{S6}$$

where $W = (W_k)$ is a d by k matrix with W_k being its k-th column. We then perform the following estimation scheme for each of the six algorithms until the convergence.

- For each cluster k, update $W_k^{(t+1)} = W_k^{(t)} - \gamma_t g_k^{(t)}$ where $g_k^{(t)}$ is obtained via using

ERM-wide (ERM-narrow, Grad-wide, Grad-narrow, Mean or Grad-trim) algorithm.

- For each i, we assign class label $c_i := \arg\min_k l(Y_i, W_k^{(t+1)})$.

In this clustering task, we consider three different settings. (i) We fix $d = 2$, $a = 1$ and let $K \in \{2, 3, 4, 5, 6\}$. (ii) We fix $d = 4$, $K = 2$ and let $a \in \{0.5, 1, 1.5, 2.5, 3.5\}$. (iii) We fix $d = 2$, $K = 3$ and let $\eta \in \{5\%, 10\%, 20\%, 30\%, 40\%\}$. The sample size is fixed at 1000 and centers W^*'s are randomly generated from the normal distribution with zero mean and standard deviation equal to 4. The average of estimation errors ($\|\hat{W} - W^*\|_2$) are provided in Figure S2.

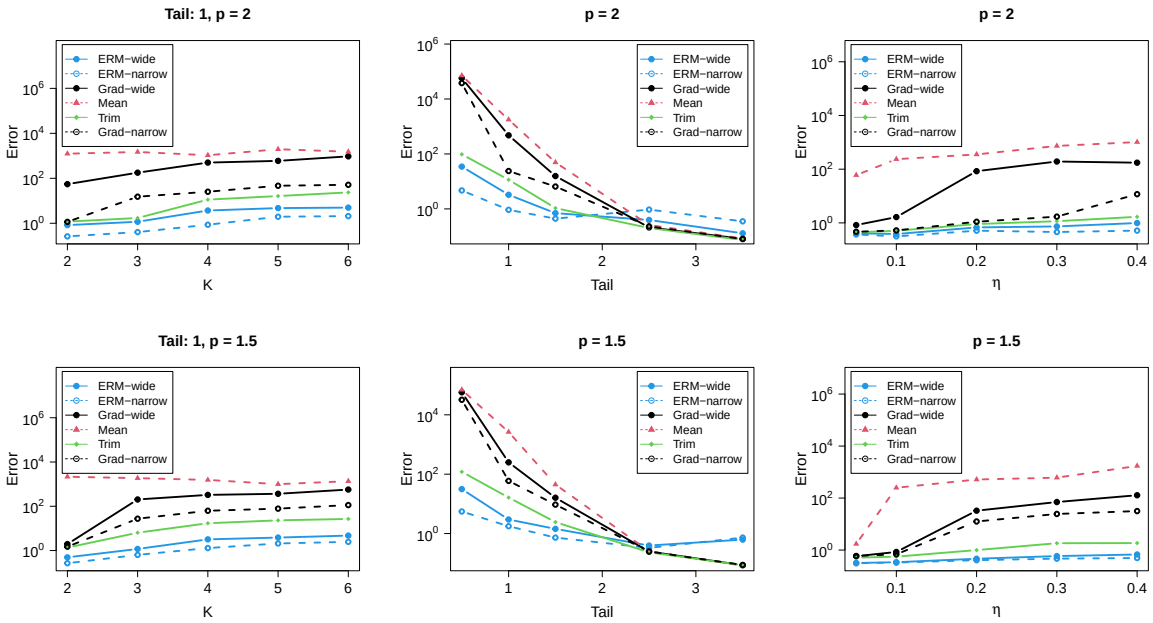

Figure S2: Comparison between six methods in k-means clustering problems.

From Figures S1 - S2, we can see the ERM based algorithm always achieves lower estimation errors.

C.3 Comparison with Additional Algorithms

There exist quite a few robust estimation methods in the literature. In this study, we consider comparing the proposed algorithm with the loss truncation-based algorithm (Y. Xu et al., 2020) (denoted as alg_{Xu}), robust coordinate gradient descent method (Merad & Gaïffas, 2023) (denoted s alg_{CD}), and adaptive Huber estimator (Sun, Zhou, & Fan, 2020) (denoted as alg_{ada}).

The regression setting is specified as follows,

$$Y_i = X_i^T w_* + \xi_i,$$

where ξ_i's are noise terms. Three types of noises are considered.

1 (Symmetrized Pareto) $\xi_i = (2u_i - 1)\tilde{\xi}_i$ with $\tilde{\xi}_i \sim_{i.i.d.} F_{pareto}(x)$ and $u_i = \text{Bernoulli}(0.5)$, $F_{pareto}(x) = 1 - \frac{1}{x^{1+a}}$ and p is the shape parameter.

2 (Mixture of Pareto) $\xi_i = \tilde{\xi}_i^{(1)} - b \cdot \tilde{\xi}_i^{(2)}$ with $\tilde{\xi}_i^{(1)}$ being the Pareto random variable with shape parameter p and $\tilde{\xi}_i^{(2)}$ being the other Pareto random variable with shape parameter $p + u$ with $u \sim Unif[0, 1]$. The constant b is chosen to satisfy $\mathbb{E}[\xi_i] = 0$.

3 (Pareto and Log-normal) $\xi_i = \tilde{\xi}_i^{(1)} - b \cdot \tilde{\xi}_i^{(2)}$ with $\tilde{\xi}_i^{(1)}$ being the Pareto random variable with shape parameter p and $\tilde{\xi}_i^{(2)}$ being the log-normal random variable with $\mu = 2, \sigma = 2$. The constant b is also chosen to satisfy $\mathbb{E}[\xi_i] = 0$.

In Algorithm alg_{Xu}, we choose the truncation loss as $\phi(x) = \log(1 + x + x^2/2)$. In Algorithm alg_{CD}, we choose the robust gradient estimator as the median of mean. In Algorithm alg_{ada}, we choose $\lambda = 0$ since we here consider a non-sparse regression problem. The sample size is chosen to be 5000. The true parameter w^* is d-dimensional ($d \in \{5, 10, 20, 40, 80\}$) and its entries are randomly chosen from $\{-2, 2\}$. Each case is replicated for 50 times and the average results are given in Figure S3.

By Figure S3, we observe that our method achieves the smallest estimation error in all six cases, while Algorithm alg_{ada}, which is originally designed for sparse regression problems, has the largest estimation error. Different types of noises do not have too much impact on the estimation results.

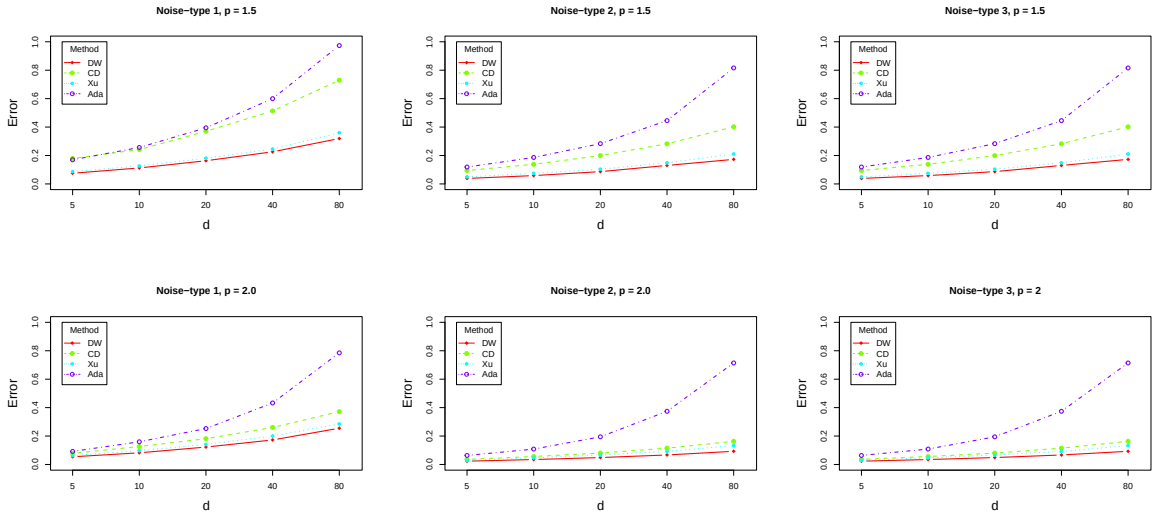

Figure S3: Comparisons with three robust estimation methods under regression problems with non-symmetric noises.

C.4 Time Comparison with alg_{CD}

Moreover, we report the computational time comparisons between the proposed double-

weighted algorithm and coordinate gradient descent method alg_{CD} for the regression task described in Section C.3 with the first noise type (i.e. symmetric Pareto noise) in Figure S4. For both methods, the termination criterion is $|w^{(t+1)} - w^{(t)}|_\infty \leq 10^{-4}$. (For the other two noise types, the results are similar and hence we omit them here.)

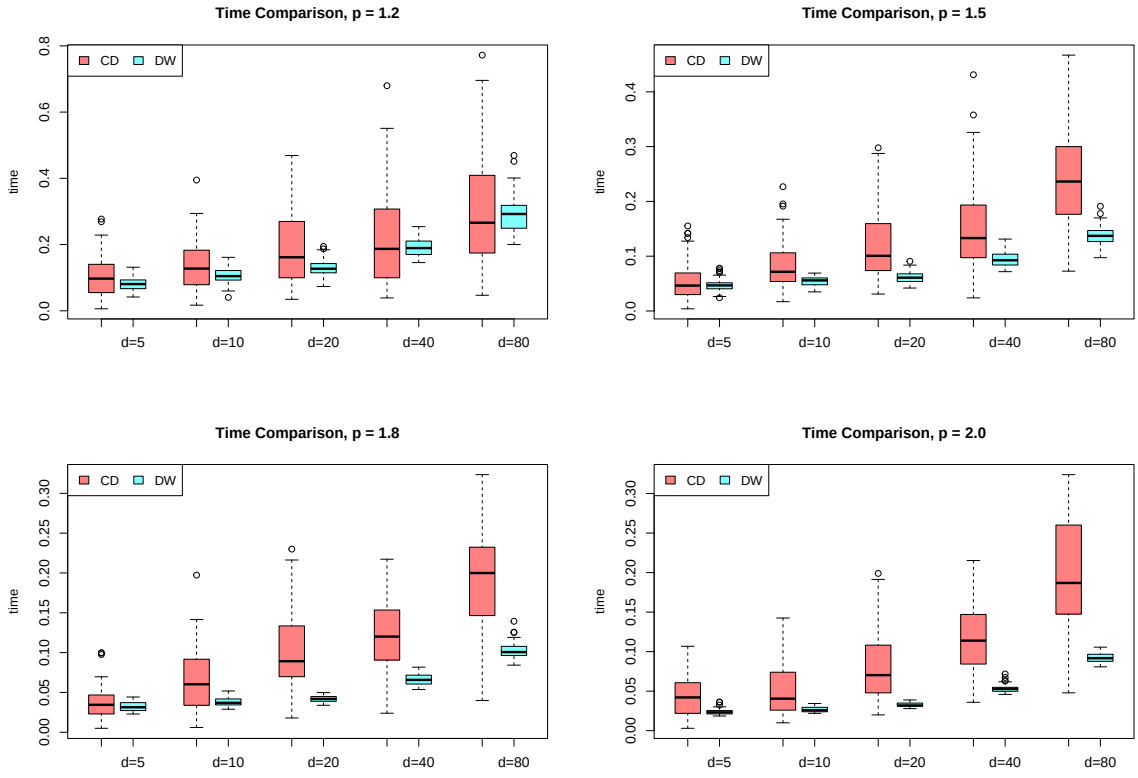

Figure S4: Box-plots of computational times (unit: second) of two methods.

From Figure S4, we can see that the computational time of the proposed double-weighted algorithm is smaller and more consistent than the coordinate gradient descent method as d gets larger. This suggests that our method could be more useful in a large-scale optimization problem.

C.5 Regression with Multi-dimensional Complex Function

In this study, we consider a regression problem for a more complex function by using Pytorch platform. The setting is the same as that in Section **??** of the main paper except that the underlying function is six-dimensional, that is,

$$f(x) = \frac{3}{2}\exp\{x_1/2 + x_2 - \sqrt{x_3 + 5}\} - \cos\{0.01 + |x_4 - 2x_5 + 3x_6|\},$$

with $x = (x_1, ..., x_6)$. We fit the data with a two-layer ReLU network with 512 hidden units. The results are plotted in Figure S5.

As we can see from Figure S5, the proposed method and the trimmed method give the similar prediction errors, which are larger than e^{-1} and are quite way from zero. This phenomenon suggests that the prediction error is largely caused by approximation bias instead of stochastic variability. In other words, the curse of dimensionality is a more severe problem than the heavy-tailness of the data.

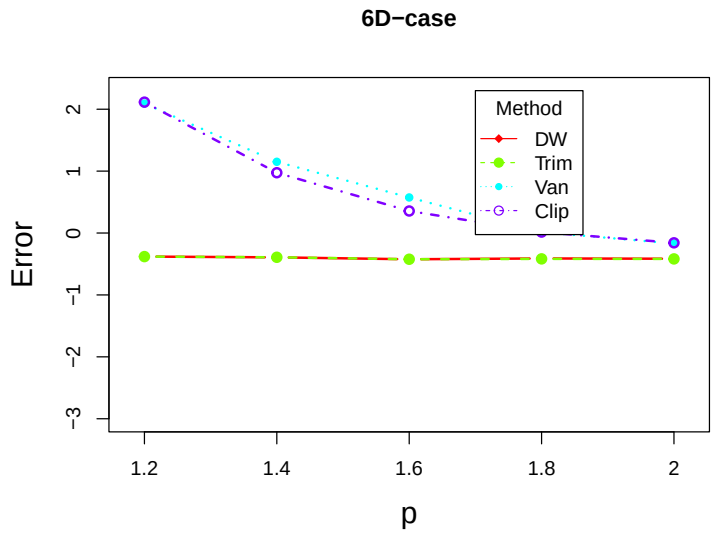

Figure S5: Prediction Error. The values are reported under log-scale.

D Proof of Results in Section ??

The following lemma gives the range of C_ε in the influence function ϕ.

Lemma S1. *A function ϕ satisfying (??) exists if and only if $C_\varepsilon \geq \left(\frac{\varepsilon}{1+\varepsilon}\right)^{\frac{1+\varepsilon}{2}} \left(\frac{1-\varepsilon}{\varepsilon}\right)^{\frac{1-\varepsilon}{2}}$.*

Remark S4. *Empirically, we find that smaller C_ε leads to more robust results. Therefore, throughout the paper, we can always treat $C_\varepsilon = \left(\frac{\varepsilon}{1+\varepsilon}\right)^{\frac{1+\varepsilon}{2}} \left(\frac{1-\varepsilon}{\varepsilon}\right)^{\frac{1-\varepsilon}{2}}$. When $\varepsilon = 1$, we recover the coefficient in Catoni (2012), namely $C_1 = 1/2$. (Here the standard convention $0^0 := 1$ applies.)*

Proof of Lemma S1. A necessary and sufficient condition for the existence of a function satisfying (??) is given by

$$(1 - x + C_\varepsilon x^{1+\varepsilon})(1 + x + C_\varepsilon x^{1+\varepsilon}) \geq 1, \ \forall \ x \geq 0.$$

After rearrangement, this reduces to

$$2C_\varepsilon x^{1+\varepsilon} + C_\varepsilon^2 x^{2(1+\varepsilon)} \geq x^2, \ \forall \ x \geq 0,$$

which is equivalent to the condition

$$C_\varepsilon^2 x^{2\varepsilon} + 2C_\varepsilon x^{\varepsilon-1} > 1, \ \forall \ x > 0. \tag{S7}$$

The minimum of the expression in the left hand side over $x > 0$ is achieved at

$$x_* = \left(\frac{1-\varepsilon}{C_\varepsilon \varepsilon}\right)^{\frac{1}{1+\varepsilon}},$$

and substituting this value in (S7) and solving for C_ε produces the desired result. ■

To prove Theorem **??**, if suffices to prove the following Theorem S1 which is the extended version of Theorem **??**.

We introduce the $(h,\ \alpha,\ \delta)$-condition,

$$C_p\alpha^{\varepsilon}(1-h)^{-\varepsilon} < 1/2; \tag{S8}$$

$$h^{-\varepsilon}\alpha^{1+\varepsilon}C_p v + \frac{\log(2/\delta)}{n} \leq \frac{\varepsilon}{1+\varepsilon}(1-h)\Big(\frac{1}{(1+\varepsilon)C_p}\Big)^{1/\varepsilon}; \tag{S9}$$

$$h^{-\varepsilon}C_p\alpha^{\varepsilon}v + C_p\alpha^{\varepsilon}(1-h)^{-\varepsilon} + \frac{\log(2/\delta)}{\alpha n} < 1 \tag{S10}$$

hold. In fact, the condition is very mild since that (S8) - (S10) are easy to be satisfied when n is large and α is small with any fixed h and δ. Here h is a tuning parameter in $(0,1)$ and it appears since "$a+bx+c|x|^p=0$"-type equation does not admit a closed form solution and we need to find an approximation to it. In the main paper, we simply treat $h=\frac{1}{2}$ for reader convenience.

Theorem S1. *Let $\{X_i\}_{i=1}^n$ be i.i.d random variables with mean μ and $\mathbb{E}|X_1-\mu|^{1+\varepsilon} \leq v$. Let $\delta \in (0,1)$, $\varepsilon \in (0,1)$ and $h \in (0,1)$. Assume that $(h,\ \alpha,\ \delta)$-condition holds, then we have the Catoni's M-estimator $\widetilde{\mu}_c$ satisfies*

$$|\widetilde{\mu}_c - \mu| \leq 2\Big(h^{-\varepsilon}C_\varepsilon\alpha^{\varepsilon}v + \frac{\log(2/\delta)}{\alpha n}\Big) \tag{S11}$$

with probability $1-\delta$. Especially, we take $\alpha = \Big(\frac{\log(2/\delta)}{nC_\varepsilon v}\Big)^{\frac{1}{1+\varepsilon}} h^{\frac{\varepsilon}{1+\varepsilon}}$, it holds

$$|\widetilde{\mu}_c - \mu| \leq 4(C_\varepsilon v)^{\frac{1}{1+\varepsilon}} h^{\frac{-\varepsilon}{1+\varepsilon}} \Big(\frac{\log(2/\delta)}{n}\Big)^{\frac{\varepsilon}{1+\varepsilon}}. \tag{S12}$$

Proof of Theorem S1. As in Catoni (2012), define

$$r_n(\theta) = \sum_{i=1}^{n} \phi\Big(\alpha(X_i-\theta)\Big),$$

and note that $r_n(\theta)$ is non-increasing in $\theta \in \mathbb{R}$. Using the upper bound on the influence

function in (**??**),

$$\begin{aligned}\mathbb{E}\Big[\exp(r_n(\theta))\Big] &= \Big(\mathbb{E}\Big[\exp\Big(\phi\Big(\alpha(X_1-\theta)\Big)\Big]\Big)^n \\ &\le \Big(\mathbb{E}\Big[1+\alpha(X_1-\theta)+C_\varepsilon\alpha^{1+\varepsilon}|X_1-\theta|^{1+\varepsilon}\Big]\Big)^n \\ &= \Big(1+\alpha(\mu-\theta)+C_\varepsilon\alpha^{1+\varepsilon}\mathbb{E}|X_1-\theta|^{1+\varepsilon}\Big)^n.\end{aligned}$$

We will use a convexity upper bound as follows. For any $a, b \ge 0$ and $0 < h < 1$,

$$\begin{aligned}(a+b)^{1+\varepsilon} &= \Big(h\frac{a}{h}+(1-h)\frac{b}{1-h}\Big)^{1+\varepsilon} \\ &\le h\Big(\frac{a}{h}\Big)^{1+\varepsilon}+(1-h)\Big(\frac{b}{1-h}\Big)^{1+\varepsilon} = \frac{a^{1+\varepsilon}}{h^\varepsilon}+\frac{b^{1+\varepsilon}}{(1-h)^\varepsilon}. \qquad \text{(CB)}\end{aligned}$$

Therefore, for any $0 < h < 1$,

$$\mathbb{E}|X_1-\theta|^{1+\varepsilon} \le h^{-\varepsilon}\mathbb{E}|X_1-\mu|^{1+\varepsilon}+(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}. \qquad \text{(S13)}$$

This leads to worse constants than in Catoni (2012), and is the price to pay for the generalization. Using the above bound, we obtain

$$\begin{aligned}\mathbb{E}\Big[\exp(r_n(\theta))\Big] &\le \Big(1+\alpha(\mu-\theta)+h^{-\varepsilon}C_\varepsilon\alpha^{1+\varepsilon}v+C_\varepsilon\alpha^{1+\varepsilon}(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}\Big)^n \\ &\le \exp\Big(\alpha n(\mu-\theta)+nh^{-\varepsilon}C_\varepsilon\alpha^{1+\varepsilon}v+nC_\varepsilon\alpha^{1+\varepsilon}(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}\Big).\end{aligned}$$

Similarly, using the lower bound on the influence function in (**??**), we obtain by symmetric arguments

$$\mathbb{E}\Big[\exp(-r_n(\theta))\Big] \le \exp\Big(-\alpha n(\mu-\theta)+nh^{-\varepsilon}C_\varepsilon\alpha^{1+\varepsilon}v+nC_\varepsilon\alpha^{1+\varepsilon}(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}\Big).$$

Let $\delta \in (0,1)$. As in Catoni (2012), we define

$$\begin{aligned}B_+(\theta) &= (\mu-\theta)+h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v+C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}+\frac{\log(2/\delta)}{\alpha n}, \\ B_-(\theta) &= (\mu-\theta)-h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v-C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}|\mu-\theta|^{1+\varepsilon}-\frac{\log(2/\delta)}{\alpha n}.\end{aligned}$$

By the exponential Markov inequality, we have

$$\mathbb{P}\Big\{r_n(\theta) \geq n\alpha B_+(\theta)\Big\} \leq \frac{\mathbb{E}\Big[\exp(r_n(\theta))\Big]}{\exp(\alpha n B_+(\theta))} \leq \delta/2,$$
$$\mathbb{P}\Big\{r_n(\theta) \leq n\alpha B_-(\theta)\Big\} \leq \frac{\mathbb{E}\Big[\exp(-r_n(\theta))\Big]}{\exp(-\alpha n B_-(\theta))} \leq \delta/2. \tag{S14}$$

Note that the function B_+ is a strictly convex function of θ and $B_+(\theta) \to \infty$ as $|\theta| \to \infty$. Therefore, B_+ has a unique minimum on $\mathbb{R}$, which is achieved at

$$\theta_* = \mu + \frac{1-h}{\alpha}\Big(\frac{1}{(1+\varepsilon)C_\varepsilon}\Big)^{\frac{1}{\varepsilon}},$$

so that

$$\min_{\theta\in\mathbb{R}} B_+(\theta) = B_+(\theta_*) = h^{-\varepsilon}\alpha^\varepsilon C_\varepsilon v - \frac{\varepsilon}{1+\varepsilon}\frac{1-h}{\alpha}\Big(\frac{1}{(1+\varepsilon)C_\varepsilon}\Big)^{\frac{1}{\varepsilon}} + \frac{\log(2/\delta)}{\alpha n}.$$

Suppose that this minimum is non-positive, i.e.

$$h^{-\varepsilon}\alpha^{1+\varepsilon}C_\varepsilon v + \frac{\log(2/\delta)}{n} \leq \frac{\varepsilon}{1+\varepsilon}(1-h)\Big(\frac{1}{(1+\varepsilon)C_\varepsilon}\Big)^{\frac{1}{\varepsilon}}. \tag{S15}$$

Then the equation

$$B_+(\theta) = 0$$

has a real root, and, if the inequality is strict, it has two real roots. Since $B_+(\mu) > 0$ and $\theta_* > \mu$, the roots are larger than μ. Letting $z = \mu - \theta$, the equation

$$\widehat{B}_+(z) = z + h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v + C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}|z|^{1+\varepsilon} + \frac{\log(2/\delta)}{\alpha n},$$

has a root $z_+ = \mu - \theta_+(\alpha)$ in $[-1, 0)$ using (S10). This is because $\widehat{B}_+(0) > 0$ and $\widehat{B}_+(-1) = h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v + C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon} + \frac{\log(2/\delta)}{\alpha n} - 1 < 0$. Additionally, since $|z|^{1+\varepsilon} < -z$ for $z \in (-1, 0)$, we have that

$$\widehat{B}_+(z) \leq z + h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v - C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}z + \frac{\log(2/\delta)}{\alpha n}. \tag{S16}$$

We let $z_{+,0} := -\frac{h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v+\frac{\log(2/\delta)}{\alpha n}}{1-C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}}$ and get $\widehat{B}_+(z_{+,0}) < 0$ from (S16). Therefore, it holds $\mu-\theta_+(\alpha) > z_{+,0} = -\frac{h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v+\frac{\log(2/\delta)}{\alpha n}}{1-C_\varepsilon\alpha^\varepsilon(1-h)^{-\varepsilon}}$. Further using (S8), we have

$$\mu-\theta_+(\alpha) := z_+ \geq -2(h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v+\frac{\log(2/\delta)}{\alpha n}).$$

By the monotonicity of the root, we know $\widetilde{\mu}_c \leq \theta_+(\alpha)$. Thus, it holds

$$\mu-\widetilde{\mu}_c := z_+ \geq -2(h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v+\frac{\log(2/\delta)}{\alpha n}).$$

Symmetric arguments establish the bounds in the other direction. Finally, by the special choice that $\alpha = \left(\frac{\log(2/\delta)}{nC_\varepsilon v}\right)^{\frac{1}{1+\varepsilon}} h^{\frac{\varepsilon}{1+\varepsilon}}$, it is straightforward to compute that $|\widetilde{\mu}_c-\mu_c| \leq 4(C_\varepsilon v)^{\frac{1}{1+\varepsilon}} h^{\frac{-\varepsilon}{1+\varepsilon}}\left(\frac{\log(2/\delta)}{n}\right)^{\frac{\varepsilon}{1+\varepsilon}}$. ■

E Proof of Results in Section ??

Proof of Theorem ??. According to Theorem **??**, we know

$$\mathbb{P}\bigg(|\hat{\mu}_f-m_f| \geq 2(2^\varepsilon C_\varepsilon\alpha^\varepsilon v+\frac{\log(2|\mathcal{F}|/\delta)}{\alpha n})\bigg) \leq \frac{\delta}{|\mathcal{F}|} \tag{S17}$$

for any fixed f. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}|\hat{\mu}_f-m_f| \geq 2(2^\varepsilon C_\varepsilon\alpha^\varepsilon v+\frac{\log(2|\mathcal{F}|/\delta)}{\alpha n})\bigg)\\
\leq\ &\sum_{f\in\mathcal{F}}\mathbb{P}\bigg(|\hat{\mu}_f-m_f| \geq 2(2^\varepsilon C_\varepsilon\alpha^\varepsilon v+\frac{\log(2|\mathcal{F}|/\delta)}{\alpha n})\bigg)\\
\leq\ &|\mathcal{F}|\frac{\delta}{|\mathcal{F}|}=\delta.
\end{aligned} \tag{S18}$$

Finally, by (**??**), we arrive at that

$$m_{\hat{f}}-m^* \leq 4(2^\varepsilon C_\varepsilon\alpha^\varepsilon v+\frac{\log(2|\mathcal{F}|/\delta)}{\alpha n})$$

holds with probability at least $1-\delta$. ■

The following lemma shows the existence of the Lipschitz constant for the default choice of influence function ϕ. Therefore, the requirement of ϕ being Lipschitz continuous is not stringent.

Lemma S2. *Consider influence function* $\phi(x) = sign(x)\log(1+|x|+C_\varepsilon|x|^{1+\varepsilon})$. *Then it is a Lipschitz function with a Lipschitz constant* L_ε *not exceeding* $\max\{(1+(1+\varepsilon)C_\varepsilon), 1+\varepsilon\}$.

Proof of Lemma S2. We can easily compute the derivative of $\phi(x)$ for $x \neq 0$. That is,

$$\phi'(x) = \frac{1+(1+\varepsilon)C_\varepsilon|x|^\varepsilon}{1+|x|+C_\varepsilon|x|^{1+\varepsilon}}, \tag{S19}$$

so $|\phi'(x)| \leq 1+\varepsilon$ if $|x| \geq 1$ and $|\phi'(x)| \leq 1+(1+\varepsilon)C_\varepsilon$ if $0 < |x| < 1$, showing the claimed Lipschitz property. ■

Before proving Lemma **??** and Lemma **??**, we need to introduce two approximate functions,

$$\begin{aligned} B_f^+(\mu,\eta) &= (m_f-\mu) + C_p\alpha^\varepsilon(1-h)^{-\varepsilon}|m_f-\mu|^p + h^{-\varepsilon}C_p\alpha^\varepsilon v + \eta, \\ B_f^-(\mu,\eta) &= (m_f-\mu) - C_p\alpha^\varepsilon(1-h)^{-\varepsilon}|m_f-\mu|^p - h^{-\varepsilon}C_p\alpha^\varepsilon v - \eta, \end{aligned}$$

and let

$$\mu_f^+(\eta) = m_f + 2h^{-\varepsilon}C_p\alpha^\varepsilon v + 2\eta, \quad \mu_f^-(\eta) = m_f - 2h^{-\varepsilon}C_p\alpha^\varepsilon v - 2\eta.$$

We additionally introduce the extended η-condition,

$$C_p\alpha^\varepsilon(1-h)^{-\varepsilon}2^p(h^{-\varepsilon}C_p\alpha^\varepsilon v + 2\eta)^{p-1} < 1, \tag{S20}$$

where $h \in (0,1)$. It reduces to the η-condition given in the main paper by taking $h = 1/2$. Under (**??**), it is easy to check that both $B_f^+(\mu,\eta) = 0$ and $B_f^-(\mu,\eta) = 0$ have at least one solution. Furthermore, it can be seen that $\mu_f^+(\eta)$ is the upper bound of the smallest root of $B_f^+(\mu,\eta)$ and $\mu_f^-(\eta)$ is the lower bound of the largest root of $B_f^-(\mu,\eta)$.

Proof of Lemma ??. To prove Lemma **??**, we need the following Lemma S3 - Lemma S4.

Lemma S3. *For any fixed $f \in \mathcal{F}$ and $\mu \in \mathbb{R}$, it holds*

$$B_f^-(\mu, 0) \le \bar{r}_f(\mu) \le B_f^+(\mu, 0), \tag{S21}$$

and, therefore, $m_f - 2h^{-\epsilon}C_p\alpha^\epsilon v \le \bar{\mu}_f \le m_f + 2h^{-\epsilon}C_p\alpha^\epsilon v$. In particular,

$$B_{\hat{f}}^-(\mu, 0) \le \bar{r}_{\hat{f}}(\mu) \le B_{\hat{f}}^+(\mu, 0).$$

For any μ and η such that $\bar{r}_{\hat{f}}(\mu) < \eta$, if extended η-condition (S20) *holds, then*

$$m_{\hat{f}} \le \mu + 2h^{-\epsilon}C_p\alpha^\epsilon v + 2\eta. \tag{S22}$$

Lemma S4. *Let $\mu_0 = m_{f^*} + A_\alpha(\delta)$. Then on the event,*

$$\Omega_{f^*}(\delta) := \{\omega : |\hat{\mu}_{f^*} - m_{f^*}| \le A_\alpha(\delta)\},$$

the following inequalities hold:

(i.) $\hat{r}_{\hat{f}}(\mu_0) \le 0$; *(ii.)* $\bar{r}_{f^*}(\mu_0) \le 0$; *(iii.)* $-\hat{r}_{f^*}(\mu_0) \le 2L_\varepsilon A_\alpha(\delta)$.

Thanks to above lemmas, we can see that, with probability at least $1 - 2\delta$, it holds

$$\begin{aligned}
\bar{r}_{\hat{f}}(\mu_0) &\le \hat{r}_{\hat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \hat{r}_{f^*}(\mu_0) + |\bar{r}_{\hat{f}}(\mu_0) - \hat{r}_{\hat{f}}(\mu_0) - \bar{r}_{f^*}(\mu_0) + \hat{r}_{f^*}(\mu_0)| \\
&\le \hat{r}_{\hat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \hat{r}_{f^*}(\mu_0) + \sup_{f\in\mathcal{F}} |\bar{r}_{\hat{f}}(\mu_0) - \hat{r}_{\hat{f}}(\mu_0) - \bar{r}_{f^*}(\mu_0) + \hat{r}_{f^*}(\mu_0)| \\
&\le \hat{r}_{\hat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \hat{r}_{f^*}(\mu_0) + Q(\mu_0, \delta) \\
&\le 0 + 0 + 2L_\varepsilon A_\alpha(\delta) + Q(\mu_0, \delta) \\
&= 2L_\varepsilon A_\alpha(\delta) + Q(\mu_0, \delta),
\end{aligned} \tag{S23}$$

where the first inequality in (S23) follows from the triangle inequality, the third inequality in (S23) follows from the definition of quantile function Q where we define the $1-\delta$ quantile of $\sup_{f\in\mathcal{F}}|X_f(\mu)-X_{f^*}(\mu)|$ by $Q(\mu,\delta)$, i.e., the minimum possible q satisfying that

$$\mathbb{P}(\sup_{f\in\mathcal{F}}|X_f(\mu)-X_{f^*}(\mu)|\le q)\ge 1-\delta.$$

The fourth inequality in (S23) is according to Lemma S4 that $\hat{r}_{\hat{f}}(\mu_0)\le 0$, $\bar{r}_{f^*}(\mu_0)\le 0$ and $-\hat{r}_{f^*}(\mu_0)\le 2L_\varepsilon A_\alpha(\delta)$. By choosing $h=1/2$, this completes the proof of lemma. ■

Proof of Lemma S3. We write $Y=\alpha(f(X)-\mu)$ and use the fact that $\phi(x)\le\log(1+x+C_\varepsilon|x|^{1+\varepsilon})$. Then

$$\begin{aligned}
\exp\{\alpha\bar{r}_f(\mu)\} &\le \exp\{\mathbb{E}[\log(1+Y+C_\varepsilon|Y|^{1+\varepsilon})]\}\\
&\le \mathbb{E}[1+Y+C_\varepsilon|Y|^{1+\varepsilon}],\\
&= 1+\alpha(m_f-\mu)+C_\varepsilon\mathbb{E}[|\alpha(f(X)-m_f+m_f-\mu)|^{1+\varepsilon}]\\
&\le 1+\alpha(m_f-\mu)+h^{-\varepsilon}C_\varepsilon\alpha^{1+\epsilon}v+C_\varepsilon\alpha^{1+\varepsilon}(1-h)^{-\varepsilon}|m_f-\mu|^{1+\varepsilon}\\
&\le \exp\{\alpha B_f^+(\mu,0)\}, \qquad \text{(S24)}
\end{aligned}$$

where we use the convexity upper bound as follows,

$$(a+b)^{1+\varepsilon}\le\frac{a^{1+\varepsilon}}{h^\varepsilon}+\frac{b^{1+\varepsilon}}{(1-h)^\varepsilon}.$$

Therefore, we have $\bar{r}_f(\mu)\le B_f^+(\mu,0)$ held for any $f\in\mathcal{F}$. Recall that $\bar{\mu}_f$ satisfies $\bar{r}_f(\mu)=0$, therefore $\bar{\mu}_f\le\mu_f^+(0)\le m_f+2h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v$. The other side of inequality is similar.

If $\bar{r}_{\hat{f}}(\mu)\le\eta$, then $B_{\hat{f}}^-(\mu,0)\le\eta$ which is equivalent to $B_{\hat{f}}^-(\mu,\eta)\le 0$. Note that $\bar{r}_{\hat{f}}(\mu)$ is a non-increasing function, μ is then above the largest solution to $B_{\hat{f}}^-(\mu,\eta)=0$. Thus, $\mu_{\hat{f}}^-(\eta)\le\mu$ which implies $m_{\hat{f}}\le\mu+2h^{-\varepsilon}C_p\alpha^\varepsilon v+2\eta$. This concludes the proof. ■

Proof of Lemma S4. For (i.), on $\Omega_{f^*}(\delta)$ and by the definition of $\hat{f}$, we have

$$\hat{\mu}_{\hat{f}}\le\hat{\mu}_{f^*}\le m_{f^*}+A_\alpha(\delta)=\mu_0.$$

Since $\hat{r}_{\hat{f}}$ is a non-increasing function of μ, $\hat{r}_{\hat{f}}(\mu_0)\le\hat{r}_{\hat{f}}(\mu_{\hat{f}})=0$.

For (ii.), by Lemma S3, $\bar{\mu}_{f^*}\le m_{f^*}+2h^{-\varepsilon}C_\varepsilon\alpha^\varepsilon v\le m_{f^*}+A_\alpha(\delta)=\mu_0$. Again by the fact that $\bar{r}_{f^*}$ is a non-increasing function, we have $\bar{r}_{f^*}(\mu_0)\le\bar{r}_{f^*}(\bar{\mu}_{f^*})=0$.

For (iii.), by Lemma S2, we can get

$$\begin{aligned} |\hat{r}_{f^*}(\mu_0)| &= |\hat{r}_{f^*}(\hat{\mu}_{f^*}) - \hat{r}_{f^*}(\mu_0)| \leq L_\varepsilon |\hat{\mu}_{f^*} - \mu_0| \\ &\leq L_\varepsilon (|\hat{\mu}_{f^*} - m_{f^*}| + |m_{f^*} - \mu_0|) \\ &\leq 2L_\varepsilon A_\alpha(\delta). \end{aligned} \tag{S25}$$

This implies $-\hat{r}_{f^*}(\mu_0) \leq 2L_\varepsilon A_\alpha(\delta)$. ■

Proof of Lemma ??. The result is the special case of Lemma S3 by taking $\mu = \mu_0$ and $h = 1/2$. ■

F Proof of Results in Section ??

Proof of Lemma ??. We let $e_n(T) := \inf\{\epsilon : N(T, d, \epsilon/2) \leq N_n\}$ with $N_n = 2^{2^n}$. We can construct a partition $\mathcal{A}_n^*$ such that $|\mathcal{A}_n^*| \leq 2^{2^n}$ and $\Delta(A) \leq e_n(T)$ for any $A \in \mathcal{A}_n^*$.

By the definition of $e_n(T)$, we know that $e_{n+1}(T) \leq e_n(T)$ and for any $\epsilon < e_n(T)$, it holds $N(T, d, \epsilon/2) > N_n$, i.e., $N(T, d, \epsilon/2) \geq 1 + N_n$. So we have

$$\begin{aligned} & (\log(1+N_n))^{1/\beta}((e_n(T))^{(1+\varepsilon)/2} - (e_{n+1}(T))^{(1+\varepsilon)/2}) \\ = \; & (\log(1+N_{n-1}))^{1/\beta} \int_{e_{n+1}(T)}^{e_n(T)} \frac{(1+\varepsilon)}{2} \epsilon^{(\varepsilon-1)/2} d\epsilon \\ \leq \; & \frac{(1+\varepsilon)}{2} \int_{e_{n+1}(T)}^{e_n(T)} \epsilon^{(\varepsilon-1)/2} (\log(N(T, d, \epsilon/2)))^{1/\beta} d\epsilon. \end{aligned} \tag{S26}$$

Note that $\log(1 + N_n) \geq 2^n \log 2$ for any $n \geq 0$, we sum over n and get

$$(\log 2)^{1/\beta} \sum_n 2^{(n)/\beta}((e_n(T))^{(1+\varepsilon)/2} - (e_{n+1}(T))^{(1+\varepsilon)/2}) \leq \frac{(1+\varepsilon)}{2} \int_0^{e_0(T)} \epsilon^{(\varepsilon-1)/2} (\log(N(T, d, \epsilon/2)))^{1/\beta} d\epsilon.$$

Furthermore,

$$\begin{aligned} & \sum_n 2^{(n)/\beta}((e_n(T))^{(1+\varepsilon)/2} - (e_{n+1}(T))^{(1+\varepsilon)/2}) \\ = \; & \sum_{n \geq 0} 2^{(n)/\beta}(e_n(T))^{(1+\varepsilon)/2} - \sum_{n \geq 1} 2^{(n-1)/\beta}(e_n(T))^{(1+\varepsilon)/2} \\ \geq \; & (1 - 2^{-1/\beta}) \sum_{n \geq 0} 2^{n/\beta}(e_n(T))^{(1+\varepsilon)/2}. \end{aligned} \tag{S27}$$

Therefore, we have that

$$\begin{aligned}\sum_{n\geq 0} 2^{n/\beta}(e_n(T))^{(1+\varepsilon)/2} &\leq \frac{1}{(\log 2)^{1/\beta}(1-2^{-1/\beta})}\frac{1+\varepsilon}{2}\int_0^{e_0(T)} \epsilon^{(\varepsilon-1)/2}(\log(N(T,d,\epsilon/2)))^{1/\beta}d\epsilon \\ &\leq C_{\beta,\varepsilon}\int_0^{\infty} \epsilon^{(\varepsilon-1)/2}(\log(N(T,d,\epsilon/2)))^{1/\beta}d\epsilon.\end{aligned}$$

Finally, by the definition of $\gamma_{\beta,\varepsilon}(T,d)$, we have

$$\begin{aligned}&\gamma_{\beta,\varepsilon}(T,d) \\ \leq\ & \sup_{t\in T}\sum_{n\geq 0} 2^{n/\beta}(\Delta(A_n^*(t)))^{(1+\varepsilon)/2} \\ \leq\ & \sum_{n\geq 0} 2^{n/\beta}\sup_{t\in T}(\Delta(A_n^*(t)))^{(1+\varepsilon)/2} \\ \leq\ & \sum_{n\geq 0} 2^{n/\beta}(e_n(T))^{(1+\varepsilon)/2} \\ \leq\ & C_{\beta,\varepsilon}\int_0^{\infty} \epsilon^{(\varepsilon-1)/2}(\log(N(T,d,\epsilon/2)))^{1/\beta}d\epsilon. \qquad \text{(S28)}\end{aligned}$$

This completes the proof. ■

Proof of Theorem ??.

By recalling

$$X_f(\mu) = \frac{1}{n}\sum_{i=1}^{n}[\frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X_i)-\mu))]], \qquad \text{(S29)}$$

we then know

$$\begin{aligned}&n(X_f(\mu) - X_{f'}(\mu)) \\ =\ & \sum_{i=1}^{n}\big[\frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X)-\mu))] - (\frac{1}{\alpha}\phi(\alpha(f'(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f'(X)-\mu))])\big].\end{aligned}$$

Using Hölder property of ϕ, we have

$$\begin{aligned}
& \mathrm{Var}[\frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X)-\mu))] - (\frac{1}{\alpha}\phi(\alpha(f'(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f'(X)-\mu))])] \\
= & \frac{1}{\alpha^2}\mathrm{Var}[\phi(\alpha(f(X_i)-\mu)) - \mathbb{E}[\phi(\alpha(f(X)-\mu))] - (\phi(\alpha(f'(X_i)-\mu)) - \mathbb{E}[\phi(\alpha(f'(X)-\mu))])] \\
\leq & \frac{1}{\alpha^2}\mathbb{E}[(\phi(\alpha(f(X_i)-\mu)) - \phi(\alpha(f'(X_i)-\mu)))^2] \\
\leq & \frac{C_{3\varepsilon}^2}{\alpha^2}\mathbb{E}[|\alpha(f(X_i)-f'(X_i))|^p] \\
= & \frac{C_{3\varepsilon}^2\alpha^p}{\alpha^2}(d_p(f,f'))^p, \qquad \text{(S30)}
\end{aligned}$$

where $d_p(f,f') := (\mathbb{E}[|f(X)-f'(X)|^p])^{1/p}$ with $p = 1+\varepsilon$. Additionally, we have

$$\begin{aligned}
& \left|\mathbb{E}[\phi(\alpha(f(X)-\mu))] - \mathbb{E}[\phi(\alpha(f'(X)-\mu))]\right| \\
\leq & \mathbb{E}[\left|\phi(\alpha(f(X)-\mu)) - \phi(\alpha(f'(X)-\mu))\right|] \\
\leq & C_{3\varepsilon}\mathbb{E}[|\alpha(f(X_i)-f'(X_i))|^{p/2}] \\
\leq & C_{3\varepsilon}\alpha^{p/2}(D(f,f'))^{p/2}. \qquad \text{(S31)}
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
& \left|\frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X)-\mu))] - (\frac{1}{\alpha}\phi(\alpha(f'(X_i)-\mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f'(X)-\mu))])\right| \\
\leq & \frac{2}{\alpha}C_{3\varepsilon}\alpha^{p/2}(D(f,f'))^{p/2}. \qquad \text{(S32)}
\end{aligned}$$

Then we can apply Bernstein inequality to get

$$\begin{aligned}
& \mathbb{P}(n|X_f(\mu) - X_{f'}(\mu)| > nt) \\
\leq & 2\exp\{-\frac{n^2t^2}{2(nC_{3\varepsilon}^2\alpha^p d_p^p(f,f')/\alpha^2 + 2C_{3\varepsilon}\alpha^{p/2-1}(D(f,f'))^{p/2}nt/3)}\} \\
= & 2\exp\{-\frac{nt^2}{2(C_{3\varepsilon}^2\alpha^{p-2}d_p^p(f,f') + 2C_{3\varepsilon}\alpha^{p/2-1}(D(f,f'))^{p/2}t/3)}\}. \qquad \text{(S33)}
\end{aligned}$$

We then recall the following lemma, which is Lemma 2.2.10 from Van Der Vaart and Wellner (1996).

Lemma S5. *Let $a, b > 0$, assume that the random variables satisfy,*

$$\mathbb{P}(|X_i| > x) \leq 2\exp\{-\frac{1}{2}\frac{x^2}{b+ax}\}$$

for any $x > 0$. Then

$$\|\max_{1\leq i\leq m} X_i\|_{\phi_1} \leq 48(a\log(1+m) + \sqrt{b}\sqrt{\log(1+m)}).$$

We write $\tilde{X}_i = |X_{f_i}(\mu) - X_{f_i'}(\mu)|$. Then Lemma S5 gives us that

$$\|\max_{1\leq i\leq m} \tilde{X}_i\|_{\phi_1} \leq 48C_{3p}\left(\frac{2\alpha^{p/2-1}}{3n}(D_m)^{p/2}\log(1+m) + \sqrt{\frac{\alpha^{p-2}}{n}}d_{p,m}^{p/2}\sqrt{\log(1+m)}\right), \quad \text{(S34)}$$

where $D_m := \max_i D(f_i, f_i')$ and $d_{p,m} := \max_i d_p(f_i, f_i')$.

We now derive a bound on $Q(\mu, \delta)$. Consider an admissible sequence $(\mathcal{B}_n)$ such that for all $f \in \mathcal{F}$,

$$\sum_{n\geq 0} 2^n(\Delta_D(B_n(f)))^{(1+\varepsilon)/2} \leq 2\gamma_{1,\varepsilon}(\mathcal{F}, D)$$

and an admissible sequence $(\mathcal{C}_n)$ such that for all $f \in \mathcal{F}$,

$$\sum_{n\geq 0} 2^{n/2}(\Delta_{d_p}(C_n(f)))^{(1+\varepsilon)/2} \leq 2\gamma_{2,\varepsilon}(\mathcal{F}, d_p).$$

Now we define an admissible sequence by intersecting the elements of $(\mathcal{B}_{n-1})$ and $(\mathcal{C}_{n-1})$: set $\mathcal{A}_0 = \{\mathcal{F}\}$ and set

$$\mathcal{A}_n = \{B \cap C : B \in \mathcal{B}_{n-1} \text{ and } C \in \mathcal{C}_{n-1}\}.$$

Define a sequence of finite sets $\mathcal{F}_0 = \{f\} \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ such that $\mathcal{F}_n$ contains a single point in each set of $\mathcal{A}_n$. For any $f' \in \mathcal{F}$, denote by $\pi_n(f')$ the unique elements of $\mathcal{F}_n$ in $A_n(f')$. Then by continuity of ϕ,

$$X_{f'}(\mu) - X_f(\mu) = \sum_{k=0}^{\infty}(X_{\pi_{k+1}(f')}(\mu) - X_{\pi_k(f')}(\mu)) \quad \text{(S35)}$$

a.s. Using the fact that $\|\cdot\|_{\phi_1}$ is a norm and (S34) we have

$$\begin{aligned}
&\| \sup_{f' \in \mathcal{F}} |X_f(\mu) - X_{f'}(\mu)| \|_{\phi_1} \\
\leq\ & \sum_{k=0}^{\infty} \| \max_{f' \in \mathcal{F}_{k+1}} |X_{\pi_{k+1}(f')(\mu)} - X_{\pi_k(f')}(\mu)| \|_{\phi_1} \\
\leq\ & 48 C_{3\varepsilon} \sum_k \left(\frac{2\alpha^{p/2-1}}{3n} (\Delta_D(B_k(f')))^{p/2} \log(1 + 2^{2^{k+1}}) + \sqrt{\frac{\alpha^{p-2}}{n}} (\Delta_{d_p}(C_k(f')))^{p/2} \sqrt{\log(1 + 2^{2^{k+1}})} \right) \\
\leq\ & 192 \log(2) C_{3\varepsilon} \sum_k \left(\frac{2\alpha^{p/2-1}}{3n} (\Delta_D(B_k(f')))^{p/2} 2^k + \sqrt{\frac{\alpha^{p-2}}{n}} (\Delta_{d_p}(C_k(f')))^{p/2} 2^{k/2} \right) \\
\leq\ & 384 \log(2) C_{3\varepsilon} \left(\frac{2\alpha^{p/2-1}}{3n} \gamma_{1,\varepsilon}(\mathcal{F}, D) + \sqrt{\frac{\alpha^{p-2}}{n}} \gamma_{2,\varepsilon}(\mathcal{F}, d_p) \right), \qquad \text{(S36)}
\end{aligned}$$

where we have use the fact that $\log(1 + 2^{2^{k+1}}) \leq 4 \log(2) 2^k$.

Since

$$X \leq \|X\|_{\phi_1} \log(2/\delta)$$

with probability at least $1 - \delta$ for any sub-exponential random variable X, we conclude that

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| \leq 384 \log(2) C_{3\varepsilon} \left(\frac{2\alpha^{p/2-1}}{3n} \gamma_{1,\varepsilon}(\mathcal{F}, D) + \sqrt{\frac{\alpha^{p-2}}{n}} \gamma_{2,\varepsilon}(\mathcal{F}, d_p)) \log(2/\delta) \right) \right) \geq 1 - \delta.$$

In particular,

$$Q(\mu, \delta) \leq 384 \log(2) C_{3\varepsilon} \log(2/\delta) \left(\frac{2\alpha^{p/2-1}}{3n} \gamma_{1,\varepsilon}(\mathcal{F}, D) + \sqrt{\frac{\alpha^{p-2}}{n}} \gamma_{2,\varepsilon}(\mathcal{F}, d_p) \right) \qquad \text{(S37)}$$

for every μ.

We put together (**??**), (S37) and the obvious observation

$$\mathbb{E}|W - \mathbb{E}W|^p \leq 2^p \mathbb{E}|W|^p$$

valid for any random variable W with a finite pth moment with $p = 1 + \varepsilon$. This gives us the desired result. ■

Proof of Theorem ??.

To prove Theorem **??**, we further define the following distance $d_{X,X'}(f,g) = (\sum_{i=1}^n (Z_i(f) - Z_i(g))^2)^{1/2}$ to quantify the difference between any two functions f and g, where $Z_i(f) := \frac{1}{n\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{n\alpha}\phi(\alpha(f(X_i')-\mu))$ for any fixed f.

By calculations, we can derive that

$$\begin{aligned}
& d_{X,X'}(f,g) \\
= & \left(\frac{1}{n^2\alpha^2}\sum_{i=1}^n (\phi(\alpha(f(X_i)-\mu)) - \phi(\alpha(f(X_i')-\mu)) - \phi(\alpha(g(X_i)-\mu)) + \phi(\alpha(g(X_i')-\mu)))^2\right)^{1/2} \\
\leq & \frac{1}{n\alpha}\Bigg((\sum_{i=1}^n (\phi(\alpha(f(X_i)-\mu)) - \phi(\alpha(g(X_i)-\mu)))^2)^{1/2} \\
& + (\sum_{i=1}^n (\phi(\alpha(f(X_i')-\mu)) - \phi(\alpha(g(X_i')-\mu)))^2)^{1/2}\Bigg) \\
\leq & \frac{C_{3\varepsilon}}{n\alpha^{1-p/2}}\left((\sum_{i=1}^n |f(X_i)-g(X_i)|^p)^{1/2} + (\sum_{i=1}^n |f(X_i')-g(X_i')|^p)^{1/2}\right) \\
= & \frac{C_{3\varepsilon}}{n^{1/2}\alpha^{1-p/2}}\left((\frac{1}{n}\sum_{i=1}^n |f(X_i)-g(X_i)|^p)^{1/2} + (\frac{1}{n}\sum_{i=1}^n |f(X_i')-g(X_i')|^p)^{1/2}\right) \qquad \text{(S38)}
\end{aligned}$$

with $p = 1+\varepsilon$.

Next, we provide a lemma that characterizes the relationship between distances and γ-functionals.

Lemma S6. *For any distances d and d_1, d_2 satisfying that $d(t,t') \leq a(d_1(t,t')^{(1+\varepsilon)/2} + d_2(t,t')^{(1+\varepsilon)/2})$, we have*

$$\gamma_2(T,d) \leq a2^{3/2}(\gamma_{2,\varepsilon}(T,d_1) + \gamma_{2,\varepsilon}(T,d_2)).$$

Proof of Lemma S6. By the definition of γ-functional, we can find an admissible sequence $(\mathcal{B}_n)$ such that for all $t \in T$,

$$\sum_{n\geq 0} 2^{n/2}(\Delta_{d_1}(B_n(t)))^{(1+\varepsilon)/2} \leq 2\gamma_{2,\varepsilon}(T,d_1)$$

and an admissible sequence $(\mathcal{C}_n)$ such that for all $t \in T$,

$$\sum_{n\geq 0} 2^{n/2}(\Delta_{d_2}(C_n(t)))^{(1+\varepsilon)/2} \leq 2\gamma_{2,\varepsilon}(T, d_2).$$

Similarly, we could construct an admissible sequence by intersecting the elements in $(\mathcal{B}_{n-1})$ and $(\mathcal{C}_{n-1})$: set $\mathcal{A}_0 = T$ and set

$$\mathcal{A}_n = \{B \cap C : B \in \mathcal{B}_{n-1} \text{ and } C \in \mathcal{C}_{n-1}\}.$$

Again $\mathcal{A}_n$ is increasing and has at most 2^{2^n} sets.

By definition of $\gamma_2(T, d)$, we have

$$\begin{aligned}
\gamma_2(T,d) &\leq \sup_{t\in T}\sum_{n\geq 0} 2^{n/2}\Delta(A_n(t)) \\
&\leq \sup_{t\in T}\sum_{n\geq 0} 2^{n/2}a((\Delta_{d_1}(A_n(t)))^{(1+\varepsilon)/2} + (\Delta_{d_1}(A_n(t)))^{(1+\varepsilon)/2}) \\
&\qquad \text{(by the relationship between } d, d_1, d_2.) \\
&\leq \sup_{t\in T}\sum_{n\geq 0} 2^{n/2}a(\Delta_{d_1}(B_{n-1}(t)))^{(1+\varepsilon)/2} + (\Delta_{d_2}(C_{n-1}(t)))^{(1+\varepsilon)/2}) \\
&\leq a2^{1/2}\sup_{t\in T}\sum_{n\geq 0} 2^{(n-1)/2}(\Delta_{d_1}(b_{n-1}(t)))^{(1+\varepsilon)/2} + (\Delta_{d_2}(C_{n-1}(t)))^{(1+\varepsilon)/2}) \\
&\leq a2^{1/2}\big(\sup_{t\in T}\sum_{n\geq 0} 2^{(n-1)/2}(\Delta_{d_1}(B_{n-1}(t)))^{(1+\varepsilon)/2} + \sup_{t\in T}\sum_{n\geq 0} 2^{(n-1)/2}(\Delta_{d_2}(C_{n-1}(t)))^{(1+\varepsilon)/2}\big) \\
&\leq a2^{3/2}(\gamma_{2,\varepsilon}(T, d_1) + \gamma_{2,\varepsilon}(T, d_2)). \qquad \text{(S39)}
\end{aligned}$$

■

Therefore, according to inequality (S38) and Lemma S6, we arrive at

$$\gamma_2(\mathcal{F}, d_{X,X'}) \leq \frac{C_{3p}2^{3/2}}{\sqrt{n}\alpha^{1-p/2}}(\gamma_{2,p}(\mathcal{F}, d_{X,p}) + \gamma_{2,p}(\mathcal{F}, d_{X',p})), \qquad \text{(S40)}$$

where $d_{X,p}(f, g) := (\frac{1}{n}\sum_i |f(X_i) - g(X_i)|^p)^{1/p}$ with $p = 1 + \varepsilon$.

Moreover, for any fixed f, we also introduce a symmetrized random variable $Z(f) := \sum_{i=1}^n \epsilon_i Z_i(f)$, where ϵ_i's are independent Rademacher random variables. By Hoeffding's

inequality, we have

$$\mathbb{P}_{\epsilon_1,\ldots,\epsilon_n}(|Z(f)-Z(g)|>t)\leq 2\exp\{-\frac{t^2}{2d_{X,X'}(f,g)^2}\}, \tag{S41}$$

where $\mathbb{P}_{(\epsilon_1,\ldots,\epsilon_n)}$ denotes the probability with respect to the Rademacher variables only.

According to equation (11) in Brownlees, Joly, and Lugosi (2015), we know that

$$\begin{aligned}&\mathbb{E}_{(\epsilon_1,\ldots,\epsilon_n)}\left[\exp\{\lambda\sup_{f\in\mathcal{F}}|\sum_{i=1}^n\epsilon_i[Z_i(f)-Z_i(f^*)]|\}\right]\\ \leq\ & 2\exp\{\lambda^2L_\varepsilon^2\gamma_2(\mathcal{F},d_{X,X'})^2/4\}.\end{aligned} \tag{S42}$$

Next we can compute the high probability bound of $\sup_{f\in\mathcal{F}}|Z(F)-Z(f^*)|$. Specifically, it holds

$$\begin{aligned}&\mathbb{P}(\sup_{f\in\mathcal{F}}|Z(F)-Z(f^*)|>t)\\ \leq\ & \mathbb{P}(\sup_{f\in\mathcal{F}}|Z(f)-Z(f^*)|>t\mid\gamma_{2,\varepsilon}(\mathcal{F},d_{X,p})\leq\Gamma_\delta,\gamma_{2,\varepsilon}(\mathcal{F},d_{X',p})\leq\Gamma_\delta)+2\mathbb{P}(\gamma_{2,\varepsilon}(\mathcal{F},d_{X,\varepsilon})>\Gamma_\delta)\\ \leq\ & \mathbb{E}_{X,X'}[\mathbb{E}_{(\epsilon_1,\ldots,\epsilon_n)}[\exp\{\lambda\sup_{f\in\mathcal{F}}|\sum_{i=1}^n\epsilon_i(Z_i(f)-Z_i(f^*))|\}]|\gamma_{2,\varepsilon}(\mathcal{F},d_{X,p})\leq\Gamma_\delta,\gamma_{2,\varepsilon}(\mathcal{F},d_{X',p})\leq\Gamma_\delta]\\ &\cdot\exp\{-\lambda t\}+\delta/4\\ \leq\ & 2\exp\{\frac{8C_{3\varepsilon}^2\lambda^2L_\varepsilon^2}{n\alpha^{2-p}}\Gamma_\delta^2-\lambda t\}+\frac{\delta}{4}.\end{aligned} \tag{S43}$$

We optimize over λ and it gives $\lambda=\frac{tn\alpha^{2-p}}{16C_{3\varepsilon}^2L_\varepsilon^2\Gamma_\delta^2}$. Then the right hand side of (S43) becomes $\exp\{-\frac{t^2n\alpha^{2-p}}{32L_\varepsilon^2C_{3\varepsilon}^2\Gamma_\delta^2}\}+\delta/4$. By letting $t=\sqrt{32}C_{3\varepsilon}L_\varepsilon\Gamma_\delta\sqrt{\log(8/\delta)}n^{-1/2}\alpha^{-(1-p/2)}$, we obtain that

$$\mathbb{P}(\sup_{f\in\mathcal{F}}|Z(f)-Z(f^*)|>t)\leq\delta/2.$$

A standard symmetrization inequality of tail probabilities of empirical process guarantees that

$$\mathbb{P}(\sup_{f\in\mathcal{F}}|X_f(\mu)-X_{f^*}(\mu)|>2t)\leq 2\mathbb{P}(\sup_{f\in\mathcal{F}}|Z(f)-Z(f^*)|>t)$$

as long as, for any $f\in\mathcal{F}$, it holds $\mathbb{P}(|X_f(\mu)-X_{f^*}(\mu)|>t)<\frac{1}{2}$.

Recall that $X_f(\mu)-X_{f^*}(\mu)$ is a mean-zero random variable. Then by Chebyshev's

inequality, we know that

$$\frac{\mathrm{Var}(X_f(\mu) - X_{f^*}(\mu))}{t^2} \leq \frac{(d_p(f, f^*))^p}{n\alpha^{2-p}t^2} < \frac{1}{2}. \tag{S44}$$

Therefore, $\mathbb{P}(|X_f(\mu) - X_{f^*}(\mu)| > t) < \frac{1}{2}$ holds for any f when $t > \sqrt{2}(\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2} n^{-1/2}\alpha^{-(1-p/2)}$.

Furthermore, without loss of generality, we can assume $C_{3\varepsilon}L_\varepsilon > 1$. Note that $0 < \delta < 1$. Thus $\sqrt{\log(8/\delta)} > 1/4$ and $\sqrt{32}L_\varepsilon\Gamma_\delta\sqrt{\log(8/\delta)} \geq \sqrt{2} \cdot \mathrm{diam}_{d_p}(\mathcal{F})$ provided that $\Gamma_\delta > (\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2}$. Therefore, we have that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| > \sqrt{32}C_{3\varepsilon}L_\varepsilon\Gamma_\delta\sqrt{\frac{\log(8/\delta)}{n\alpha^{2-p}}} \right) \leq \delta$$

when $\Gamma_\delta > (\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2}$. Similarly, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| > \sqrt{32}C_{3\varepsilon}L_\varepsilon(\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2}\sqrt{\frac{\log(8/\delta)}{n\alpha^{2-p}}} \right) \leq \delta$$

when $\Gamma_\delta \leq (\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2}$. To sum up, we have

$$Q(\mu_0, \delta) \leq \sqrt{32}C_{3\varepsilon}L_\varepsilon \max\{\Gamma_\delta, (\mathrm{diam}_{d_p}(\mathcal{F}))^{p/2}\}\sqrt{\frac{\log(8/\delta)}{n\alpha^{2-p}}}. \tag{S45}$$

By above inequality, (**??**) and $p = 1 + \varepsilon$, it concludes the proof of Theorem **??**. ■

G Proof of Results in Section B

Proof of Proposition 2.

It is straightforward to see that

$$|(g(Z_i) - Y_i)^2 - (g'(Z_i) - Y_i)^2 \;\; \leq \;\; d_\infty(g, g')\big(|Y_i - g(Z_i)| + |Y_i - g'(Z_i)|\big). \qquad \text{(S46)}$$

Thus

$$d_{X,p}(f_g, f_{g'}) \leq d_\infty(g, g') \left[\left(\frac{1}{n} \sum_{i=1}^n |Y_i - g(Z_i)|^p \right)^{1/p} + \left(\frac{1}{n} \sum_{i=1}^n |Y_i - g'(Z_i)|^p \right)^{1/p} \right],$$

with $p = 1 + \varepsilon$. By Chebyshev's inequality, it holds that

$$\frac{1}{n} \sum_{i=1}^n |Y_i|^p \leq \mathbb{E}[|Y|^p] + \sqrt{8v/n\delta}$$

with probability at most $\delta/8$. Choosing Δ to be upper bound of $d_\infty(g, g')$ for any $g, g' \in \mathcal{G}$, we then have

$$d_X(f, f') \leq 2^{1+(2-p)/p} d_\infty(g, g') \Big(\Delta^p + \mathbb{E}[|Y|^p] + \sqrt{8v/n\delta} \Big)^{1/p}$$

holds with probability at least $1 - \delta/8$. By definition, it is easy to see that $\gamma_{2,\varepsilon}(\mathcal{G}, d_1) \leq c\gamma_{2,\varepsilon}(\mathcal{G}, d_2)$ for any distances d_1, d_2 satisfying $d_1 \leq c d_2$. Then, we know that

$$\Gamma_\delta \leq \Gamma_\delta(\Delta) := 2^{1+(2-p)/p} (\Delta^p + \mathbb{E}[|Y|^p] + \sqrt{8\sigma^2/n\delta})^{1/p} \cdot \gamma_{2,\varepsilon}(\mathcal{G}, d_\infty).$$

By choosing Δ large enough, it holds $\Gamma_\delta(\Delta) \geq \Delta \geq \mathrm{diam}_{d_p}(\mathcal{F})^{p/2}$. ■

Proof of Proposition 3.

By representer theorem, the optimizer of (S5) has the form, $\hat{h}(x) = \sum_{i=1}^n c_i K(x_i, x)$. Therefore, solving (S5) requires handling with an n by n matrix, which is computationally expensive in most cases. In the following, we prove a stronger version. We consider a smaller Hilbert space $\mathcal{H}_s$ instead of $\mathcal{H}$.

Note that the kernel K is a Mercer kernel which admits the approximation

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \varphi(v_j, x)\varphi(v_j, y).$$

We define a smaller RKHS space

$$\mathcal{H}_s = \{h(x) : h(x) = \sum_{i=s}^{S} c_i \varphi(v_i, x), c_i \in \mathbb{R}\},$$

where $v_i, i = 1, \ldots, S$ are S features with $S \ll n$. Then we practically solve the following estimator,

$$\hat{f}_{H_s} = \arg \min_{f=L\circ h \in L\circ \mathcal{H}_s} \{\hat{\mu}_f + \lambda_n \|h\|^2_{\mathcal{H}_s}\}. \tag{S47}$$

Let $\tilde{m}^* := \min_{f\in L\circ\mathcal{H}_s} m_f$. Given $v_1, \ldots, v_S$ and recalling the fact that $\gamma_{\beta,p}(L \circ \mathcal{H}_s, d) \leq \gamma_{\beta,p}(\mathcal{F}, d)$ for any β, distance d and sub-space $\mathcal{H}_s$, we apply Theorem **??** or Theorem **??** (simply modifying the proof by setting $A_\alpha(\delta) = 2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n} + \lambda_n + \text{err}$) and obtain **stronger result**,

$$m_{\hat{f}_{H_s}} - \tilde{m}^* \leq 6L_\varepsilon(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n} + \lambda_n + \text{err}) + 2Q_{1,\mathcal{H}_s}(\delta) \tag{S48}$$

holds with probability $1 - \delta$. Here err is an approximation error which will be explain later in this section and

$$Q_{1,\mathcal{H}_s}(\delta) \;\; = \;\; K \cdot C_{3\varepsilon} \log(2/\delta)(\frac{2\alpha^{(\varepsilon-1)/2}}{3n}\gamma_{1,\varepsilon}(L \circ \mathcal{H}_s, D) + \sqrt{\frac{\alpha^{\varepsilon-1}}{n}}\gamma_{2,\varepsilon}(L \circ \mathcal{H}_s, d_p)).$$

Remark S5. *We can obtain the upper bounds of $\gamma_{1,\varepsilon}(L \circ \mathcal{H}_s, D), \gamma_{2,\varepsilon}(L \circ \mathcal{H}_s, d_p))$ by computing the covering number $N(L \circ \mathcal{H}_s, D, \epsilon/2)$ and $N(L \circ \mathcal{H}_s, d_p, \epsilon/2)$ in specific cases. For example, suppose loss function L is c_1-Lipschitz continuous with respect to argument $h(x)$. Then $N(L \circ \mathcal{H}_s, D, \epsilon/2) \leq N(\mathcal{H}_s, D, \epsilon/2c_1)$ and $N(L \circ \mathcal{H}_s, d_p, \epsilon/2) \leq N(\mathcal{H}_s, d_p, \epsilon/2c_1)$. Write*

$\mathbb{C} = \{(c_1, \ldots, c_s) : c_i \in [-b, b]\}$ *and assume eigen-functions satisfy* $\max_i \sup_x \varphi(v_i, x) \leq B$ *and* $\max_i \mathbb{E}[|\varphi(v_i, X)|^p]^{1/p} \leq B$. *We know* $N(\mathcal{H}_s, D, \epsilon/2c_1) \leq N(\mathbb{C}, \ell_1, \epsilon/2c_1B)$ *and* $N(\mathcal{H}_s, d_p, \epsilon/2c_1) \leq N(\mathbb{C}, \ell_p, \epsilon/2c_1BS^{(p-1)/p})$. *Finally, it is know that* $N(\mathbb{C}, \ell_1, \epsilon/2c_1B) = O\big((\frac{S}{\epsilon})^S\big)$ *and* $N(\mathbb{C}, \ell_p, \epsilon/2c_1BS^{(p-1)/p}) = O\big((\frac{\sqrt{S}}{\epsilon})^S\big)$. *Upper bounds of* $\gamma_{1,\varepsilon}(L \circ \mathcal{H}_s, D), \gamma_{2,\varepsilon}(L \circ \mathcal{H}_s, d_p)$ *is then obtained from Lemma* **??**.

Then problem is reduced to understanding the difference $\tilde{m}^* - m^*$. By the definition, we know

$$\tilde{m}^* - m^* = \tilde{m}^* - m_{f_0} + m_{f_0} - m^* \leq m_{f_0} - m^* =: \text{err} \tag{S49}$$

for any $f_0 \in L \circ \mathcal{H}_s$. We need to find a suitable $f_0 = L \circ h_0$ such that $m_{f_0} - m^*$ is as small as possible (i.e., approximating $L \circ h^*$ as close as possible). By definition of m_f, we have

$$\begin{aligned} m_{f_0} - m^* &= \mathbb{E}[L(Y - h_0(X))] - \mathbb{E}[L(Y - h^*(X))] \\ &\leq \mathbb{E}|C(Y)(h_0(X) - h^*(X))| \\ &\leq \sqrt{\mathbb{E}[C^2(Y)]}\sqrt{\mathbb{E}[(h_0(X) - h^*(X))^2]} \\ &\leq C\sqrt{\mathbb{E}[(h_0(X) - h^*(X))^2]}. \end{aligned} \tag{S50}$$

by adjusting constant C. The last inequality uses the assumption that $C(Y)$ is square integrable.

Since K is a Mercer kernel which satisfies $K(x, y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\varphi_i(y)$ with $\{\varphi_i(\cdot)\}$ are orthonormal bases in $L_2(X)$, λ_i are non-increasing. Then we know $\hat{K}(x, y) = \sum_{j=1}^{S} \lambda_{i_j} \varphi_{i_j}(x)\varphi_{i_j}(y)$. In addition, h^* can be decomposed as $h^* = \sum_{i=1}^{\infty} a_i^* \varphi_i(x)$ satisfying that $\sum_{i=1}^{\infty} (a_i^*)^2/\lambda_i \leq 1$. To this end, we deliberately choose $h_0(x) = \sum_{j=1}^{S} a_{i_j}^* \varphi_{i_j}(x)$.

$$\begin{aligned} \|h_0(x) - h^*(x)\|_{L_2} &= \sqrt{\int |h_0(x) - h^*(x)|^2 dx} \\ &\leq \sqrt{\sum_{i=1}^{\infty} (a_i^*)^2 - \sum_{j=1}^{S} (a_{i_j}^*)^2} \\ &\leq \sqrt{\lambda_{i_0}}, \end{aligned} \tag{S51}$$

where $i_0 = \arg\min\{i : \ i \text{ is not in } i_1, \ldots, i_S\}$. The last inequality (S51) uses the fact that

$$\sum_{i=1}^{\infty}(a_i^*)^2 - \sum_{j=1}^{S}(a_{i_j}^*)^2 = \sum_{i=i_0}^{\infty}(a_i^*)^2 \leq \lambda_{i_0}(\sum_{i=i_0}^{\infty}(a_i^*)^2/\lambda_i) \leq \lambda_{i_0}\Big(\sum_{j=1}^{\infty}(a_j^*)^2/\lambda_j\Big) \leq \lambda_{i_0}.$$

Therefore, with (S50), we have

$$\text{err} = m_{f_0} - m^* \leq C\sqrt{\lambda_{i_0}}$$

and plug this back into (S48) to conclude the analysis of excess risk, $m_{\tilde{f}_{H_s}} - m^*$. Finally, note that err $= 0$ when $\mathcal{H}_s = \mathcal{H}$. This completes the proof. ■

H Proof of Results in Section ??

Proof of Theorem ??. The proof consists of two main steps.

Step 1. Our goal here is to obtain the uniform concentration bounds of differences between gradients $g^{(t)}$'s and their expectations.

To begin with, we consider the j-th coordinate of the gradient. We let $r_{n,w}(\theta) = \frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(\nabla f_w(X_i)[j]-\theta))$ and we let $\hat{\theta}_w$ be the solution to $r_{n,w}(\theta)=0$. We then consider the following two cases, where Case 1 is the special case of Case 2.

Case 1. It holds $|\nabla f_{w_1}(X)[j]-\nabla f_{w_2}(X)[j]| \le R\|w_1-w_2\|$ for any X. (That is, Assumption **A1** is replaced by bounded Lipschitz condition.)

Case 2. It holds $|\nabla f_{w_1}(X)[j]-\nabla f_{w_2}(X)[j]| \le R_B\|w_1-w_2\|$ for any $\|X\|\le B$. (Assumption **A1**.)

In the first case, we show that the smoothness of the loss function implies a Lipschitz property of the estimator (He & Shao, 1996; Holland & Ikeda, 2019).

To see this, by Lipschitz assumption, we observe that

$$\begin{aligned}
&\frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(\nabla f_{w_1}(X_i)[j]-R\|w_1-w_2\|-\theta))\\
\le\ & r_{n,w_2}(\theta)\\
\le\ & \frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(\nabla f_{w_1}(X_i)[j]+R\|w_1-w_2\|-\theta)).
\end{aligned} \tag{S52}$$

Since $\phi(\cdot-\theta)$ is non-increasing, then $\hat{\theta}_{w_2}$ is no greater than the solution to $\sum_{i=1}^{n}\phi(\alpha(\nabla f_{w_1}(X_i)[j]+R\|w_1-w_2\|-\theta))=0$ and is no smaller than $\sum_{i=1}^{n}\phi(\alpha(\nabla f_{w_1}(X_i)[j]-R\|w_1-w_2\|-\theta))=0$. It is also easy to see that $\hat{\theta}_{w_1}\pm R\|w_1-w_2\|$s are the solutions to $\sum_{i=1}^{n}\phi(\alpha(\nabla f_{w_1}(X_i)\pm R\|w_1-w_2\|-\theta))=0$, respectively. Therefore, we have

$$\hat{\theta}_{w_1}-R\|w_1-w_2\| \le \hat{\theta}_{w_2} \le \hat{\theta}_{w_1}+R\|w_1-w_2\|.$$

In other words,

$$|\hat{\theta}_{w_1}-\hat{\theta}_{w_2}| \le R\|w_1-w_2\|. \tag{S53}$$

This leads to the desired Lipschitz property.

For the second case, it might be hard to directly get the similar Lipschitz property like (S53). But fortunately, we can find a good proxy estimator that enjoys this property and the proxy is not far away from the true estimator.

We define $\tilde{X}_{i,\eta}$ be the truncation version of X_i at level η, that is,

$$\tilde{X}_{i,\eta} = X_i \mathbf{1}_{|X_i| \le B_\eta}, \tag{S54}$$

where B_η is defined in Remark **??**. According to Theorem 4 of Chung and Lu (2006), we know that

$$\mathbb{P}(\sharp|i : \tilde{X}_{i,\eta} \neq X_i| \ge n\eta + \lambda) \le \exp\{-\frac{\lambda^2}{2(n\eta + \lambda/3)}\}. \tag{S55}$$

By taking $\eta = 2n\eta$, we have

$$\mathbb{P}(\sharp|i : \tilde{X}_{i,\eta} \neq X_i| \ge 3n\eta) \le \delta, \tag{S56}$$

whenever $\eta \ge \frac{\log(1/\delta)}{n}$. In other words, with probability at least $1-\delta$, there are at most $3n\eta$ $\tilde{X}_{i,\eta}$'s differ from the original X_i's. We define the event $E_{good} := \{\sharp|i : \tilde{X}_{i,\eta} \neq X_i| \le 3n\eta\}$.

Moreover, we are able to define $\tilde{r}_{n,w,\eta}(\theta) = \frac{1}{n\alpha}\sum_{i=1}^n \phi(\alpha(\nabla f_w(\tilde{X}_{i,\eta}) - \theta))$ and let $\hat{\theta}_{w,\eta,\zeta}$ be the solution to $\tilde{r}_{n,w,\eta}(\theta) = \zeta$ for any fixed ζ. Next, we study the difference between $\hat{\theta}_{w,\eta,\zeta}$ and $\hat{\theta}_{w,\eta,0}$. Without loss of generality, we assume $\zeta > 0$ and it is easy to see that $\hat{\theta}_{w,\eta,\zeta} \le \hat{\theta}_{w,\eta,0}$ by the fact that ϕ is non-increasing.

By assumption A1 and the optimality of w^*, we can obtain that, for any w in the parameter space and $\|X\| \le B_{\frac{1}{2}}$, it holds

$$|\nabla f_w(X)| = |\nabla f_w(X)[j] - \nabla f_{w^*}(X)[j]| \le R_{B_{\frac{1}{2}}} \|w - w^*\| \le R_{B_{\frac{1}{2}}} D_w,$$

where D_w is the diameter of parameter space. Let event $E_{\text{small},x} = \{\#\{i : \|X_i\| \ge B_{\frac{1}{2}}\} \ge n/4\}$. By Hoeffding inequality, we know event $E_{\text{small},x}$ happens with probability at least $1 - \exp\{-n/4\} \ge 1 - \delta$ for sufficiently large $n \ge 4\log(1/\delta)$.

On event $E_{\text{small},x}$, it is not hard to get that

$$\begin{aligned}
\zeta &= -\int_{\hat{\theta}_{w,\eta,\zeta}}^{\hat{\theta}_{w,\eta,0}} \tilde{r}'_{n,w,\eta}(\theta)d\theta \\
&= -\frac{1}{n\alpha}\sum_{i=1}^{n}\int_{\hat{\theta}_{w,\eta,\zeta}}^{\hat{\theta}_{w,\eta,0}} \frac{\partial\phi(\alpha(\nabla f_w(\tilde{X}_{i,\eta})[j]-\theta))}{\partial\theta}d\theta \\
&\geq \frac{1}{4}\cdot\frac{1}{2}\int_{\hat{\theta}_{w,\eta,\zeta}}^{\hat{\theta}_{w,\eta,0}} 1 d\theta \qquad \text{(S57)} \\
&= \frac{1}{8}|\hat{\theta}_{w,\eta,0}-\hat{\theta}_{w,\eta,\zeta}|. \qquad \text{(S58)}
\end{aligned}$$

Here (S57) uses the fact that $\frac{\partial\phi(\alpha(\nabla f_w(\tilde{X}_{i,\eta})-\theta))}{\partial\theta} \geq \frac{1}{2}\alpha$ under the requirement that $\alpha(R_{B_{\frac{1}{2}}}+1)D_w \leq x_c$ and the assumption $\phi'(x) \geq \frac{1}{2}$ for $|x| \leq x_c$.

Therefore, by (S58), we have

$$\theta_{w,\eta,\zeta} \geq \hat{\theta}_{w,\eta,0} - 8\zeta \qquad \text{(S59)}$$

for any w and $\zeta > 0$ on event $E_{\text{small},x}$. Similarly, it holds

$$\theta_{w,\eta,-\zeta} \leq \hat{\theta}_{w,\eta,0} + 8\zeta. \qquad \text{(S60)}$$

By the assumptions on ϕ, on event E_{good}, we have

$$\begin{aligned}
&|r_{n,w}(\theta)-\tilde{r}_{n,w,\eta}(\theta)| \\
=\ &|\frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(\nabla f_w(X_i)[j]-\theta)) - \frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(\nabla f_w(\tilde{X}_{i,\eta})[j]-\theta))| \\
=\ &|\frac{1}{n\alpha}\sum_{i:X_i\neq\tilde{X}_{i,\eta}}\phi(\alpha(\nabla f_w(X_i)[j]-\theta)) - \phi(\alpha(\nabla f_w(\tilde{X}_{i,\eta})[j]-\theta))| \\
\leq\ &3\eta\cdot 2A/\alpha \qquad \text{(S61)}
\end{aligned}$$

held for any w and θ, where $A := \sup_x|\phi(x)|$ (A is finite since $\phi'(x) \equiv 0$ when $|x| \geq x_{cut}$). As a result,

$$\tilde{r}_{n,w,\eta}(\theta) - \frac{6A\eta}{\alpha} \leq r_{n,w}(\theta) \leq \tilde{r}_{n,w,\eta}(\theta) + \frac{6A\eta}{\alpha},$$

Again, by the fact that $r_{n,w}(\theta)$ is non-increasing, we have

$$\hat{\theta}_{w,\eta,\frac{6A\eta}{\alpha}} \leq \hat{\theta}_w \leq \hat{\theta}_{w,\eta,-\frac{6A\eta}{\alpha}}. \tag{S62}$$

Together with (S59) - (S60) by taking $\zeta = 6A\eta/\alpha$, we arrive at

$$\hat{\theta}_{w,\eta,0} - \frac{8 \cdot 6A\eta}{\alpha} \leq \hat{\theta}_w \leq \hat{\theta}_{w,\eta,0} + \frac{8 \cdot 6A\eta}{\alpha}. \tag{S63}$$

We view $\hat{\theta}_{w,\eta,0}$ as the **proxy** of $\hat{\theta}_w$ at level η. (S63) says that the difference between the proxy and $\hat{\theta}_w$ is no more than $\frac{48A\eta}{\alpha}$ on event $E_{good} \cap E_{small,x}$. As long as we could choose η sufficiently small, the proxy $\hat{\theta}_{w,\eta,0}$ is very close to $\hat{\theta}_w$.

Following the proof in Case 1, we can straightforwardly get

$$|\hat{\theta}_{w_1,\eta,0} - \hat{\theta}_{w_2,\eta,0}| \leq R_{B_\eta}\|w_1 - w_2\| \tag{S64}$$

for any w_1, w_2. In other words, Lipschitz continuity property (S53) holds for the proxy estimators.

Note that our goal is to study $\sup_w |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]|$, it is sufficient to study $\sup_{w\in\mathcal{N}_\epsilon} |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]|$, where $\mathcal{N}_\epsilon$ is an ϵ-net over the parameter space. To see this, we take any w in the parameter space and let $w' \in \mathcal{N}_\epsilon$ with $\|w - w'\| \leq \epsilon$. Therefore, it holds

$$\begin{aligned}
& |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]| \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| \\
& + |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]] - (\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]])| \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)]| + |\hat{\theta}_w - \hat{\theta}_{w'}| + L_f\|w - w'\| \quad \text{(by Assumption A2)} \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| + |\hat{\theta}_w - \hat{\theta}_{w,\eta,0}| + |\hat{\theta}_{w,\eta,0} - \hat{\theta}_{w',\eta,0}| + |\hat{\theta}_{w',\eta,0} - \hat{\theta}_{w'}| + L_f\|w - w'\| \quad \text{(use proxy)} \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| + |\hat{\theta}_{w,\eta,0} - \hat{\theta}_{w',\eta,0}| + 2 \cdot \frac{48A\eta}{\alpha} + L_f\|w - w'\| \quad \text{(by (S63))} \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| + 2 \cdot \frac{48A\eta}{\alpha} + (R_{B_\eta} + L_f)\|w - w'\| \quad \text{(by (S64))} \\
\leq\ & |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| + 2 \cdot \frac{48A\eta}{\alpha} + (R_{B_\eta} + L_f)\epsilon \quad \text{(by property of } \mathcal{N}_\epsilon). \qquad \text{(S65)}
\end{aligned}$$

As a result, we have that

$$\sup_w |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]| \leq \sup_{w\in\mathcal{N}_\epsilon} |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]| + \frac{96A\eta}{\alpha} + +(R_{B_\eta} + L_f)\epsilon. \tag{S66}$$

In particular, we can take $\epsilon = \epsilon_0 := \frac{96A\eta}{(R_{B_\eta}+L_f)\alpha}$ and the above inequality becomes

$$\sup_w |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]| \leq \sup_{w\in\mathcal{N}_{\epsilon_0}} |\hat{\theta}_w - \mathbb{E}[\nabla f_w(X)[j]]| + \frac{192A\eta}{\alpha}. \tag{S67}$$

By the basic property of covering number for compact subsets of Eculidean space, we know

$$|\mathcal{N}_{\epsilon_0}| \leq (\frac{3D_w}{2\epsilon_0})^d = (\frac{D_w(R_{B_\eta}+L_f)\alpha}{64A\eta})^d.$$

By (2.10), we know

$$\mathbb{P}(|\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| \geq A_\alpha(\delta')) \leq \delta'$$

for any δ'. By taking $\delta' = \delta/\Big(d(\frac{D_w(R_{B_\eta}+L_f)\alpha}{64A\eta})^d\Big)$, we have

$$\mathbb{P}(\sup_{w\in\mathcal{N}_{\epsilon_0}} |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| \geq A_\alpha(\delta')) \leq \delta/d$$

by the union bound. Together with (S66), we have that

$$\mathbb{P}(\{\sup_w |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| \leq A_\alpha(\delta') + \frac{192A\eta}{\alpha}\} \cap E_{good} \cap E_{small,x}) \geq 1 - 2\delta - \delta/d. \tag{S68}$$

Note the requirement that $\frac{192A\eta}{\alpha} \leq A_\alpha(\delta')$, the term $\frac{192A\eta}{\alpha}$ can be absorbed into $A_\alpha(\delta')$ by multiplying a constant 2. Thus we obtain

$$\mathbb{P}(\sup_w |\hat{\theta}_{w'} - \mathbb{E}[\nabla f_{w'}(X)[j]]| \leq 2A_\alpha(\delta')) \geq 1 - 2\delta - \delta/d. \tag{S69}$$

Therefore, we complete the first step.

Step 2. We prove the convergence of Algorithm **??** under two scenarios, (1) F is differentiable and L_f-Lipschitz continuous, (2) F is κ strongly-convex and L_f-Lipschitz continuous.

For the first scenario, by the Lipschitz continuity, we have

$$\begin{aligned} & F(w^{(t+1)}) - F(w^{(t)}) \leq \nabla F(w^{(t)})(w^{(t+1)} - w^{(t)}) + \frac{L_f}{2}\|w^{(t+1)} - w^{(t)}\|^2 \\ = \; & -\gamma_t \nabla F(w^{(t)}) g^{(t)} + \frac{L_f \gamma_t^2}{2}\|g^{(t)}\|^2 \\ = \; & -\gamma_t \nabla F(w^{(t)})(\nabla F(w^{(t)}) + \zeta^{(t)}) + \frac{L_f \gamma_t^2}{2}\|\nabla F(w^{(t)}) + \zeta^{(t)}\|^2 \\ \leq \; & -\gamma_t \|\nabla F(w^{(t)})\|^2 + \gamma_t \sqrt{d}\bar{A}_\alpha(\delta)\|\nabla F(w^{(t)})\| + L_f \gamma_t^2 (\|\nabla F(w^{(t)})\|^2 + d(\bar{A}_\alpha(\delta))^2) \end{aligned} \tag{S70}$$

where $\zeta^{(t)} := g^{(t)} - \nabla F(w^{(t)})$ and, by applying (S69) for each $j \in [d]$, it is easy to see that $\|\zeta^{(t)}\| \leq \sqrt{d}\bar{A}_\alpha(\delta)$ holds with probability at least $1 - 3\delta$ for any t with $\bar{A}_\alpha(\delta) := 2A_\alpha(\delta')$.

By direct calculation, we can find that

$$\gamma_t \sqrt{d}\bar{A}_\alpha(\delta)\|\nabla F(w^{(t)})\| + L\gamma_t^2 d(\bar{A}_\alpha(\delta))^2 \leq \frac{\gamma_t - \gamma_t^2 L_f}{2}\|\nabla F(w^{(t)})\|^2$$

when $\gamma_t \leq \frac{4}{9L_f}$ and $\|\nabla F(w^{(t)})\| \geq \sqrt{d}\bar{A}_\alpha(\delta)$. Together with (S70), we know

$$F(w^{(t+1)}) - F(w^{(t)}) \leq -\frac{\gamma_t - \gamma_t^2 L_f}{2}\|\nabla F(w^{(t)})\|^2 \leq -\frac{5}{18}\gamma_t \|\nabla F(w^{(t)})\|^2 \tag{S71}$$

holds with probability at least $1-3\delta$. Define T_{stop} to be the smallest t such that $\|\nabla F(w^{(t)})\| \leq \sqrt{d}\bar{A}_\alpha(\delta)$. Therefore,

$$\begin{aligned} & F(w^{(T_{stop})}) - F(w^{(0)}) \\ = \; & \sum_{t=0}^{T_{stop}-1} F(w^{(t+1)}) - F(w^{(t)}) \\ \leq \; & \sum_{t=0}^{T_{stop}-1} -\frac{5}{18}\gamma_t \|\nabla F(w^{(t)})\|^2 \;\; \text{(by (S71))} \\ \leq \; & -\frac{5}{18}\gamma_t d(\bar{A}_\alpha(\delta))^2 \;\; \text{(by the definition of } T_{stop}) \end{aligned} \tag{S72}$$

By re-organizing (S72) and recalling the definition that $F(w) = m_{f_w}$ and the optimality

of f^* that $m_{f^*} \leq m_{f_{w^{(T_{stop})}}}$, then we know

$$\sum_{t=1}^{T_{stop}} \gamma_t \leq \frac{18(m_{f_{w^{(0)}}} - m_{f^*})}{5d(\bar{A}_\alpha(\delta))^2} \tag{S73}$$

holds with probability at least $1-3\delta$. This complete the proof by using the assumption that $\gamma_t \equiv \gamma$.

In the second scenario, we can compute that

$$\begin{aligned} & \|w^{(t+1)} - w^*\| = \|w^{(t)} - \gamma_t g^{(t)} - w^*\| \\ \leq \; & \|\|w^{(t)} - \gamma_t \nabla F(w^{(t)}) - w^*\|\| + \gamma_t \|\nabla F(w^{(t)}) - g^{(t)}\|. \end{aligned} \tag{S74}$$

The first term of (S74) can be handled via standard method in Nesterov (2003). It then follows that

$$\|w^{(t)} - \gamma_t \nabla F(w^{(t)}) - w^*\|^2 \leq (1 - \frac{2\gamma_t \kappa L_f}{\kappa + L_f}) \|w^{(t)} - w^*\|^2.$$

For the second term, by the same logic, it is bounded via $\gamma_t \zeta$ with statistical error $\zeta := \sqrt{d}\bar{A}_\alpha(\delta)$.

Therefore, we have

$$\|w^{(t+1)} - w^*\| \leq (\prod_{s=0}^{t} a_t) \|w^0 - w^*\| + \zeta (\sum_{s=0}^{t} \gamma_s \prod_{s'=s+1}^{t} a_s), \tag{S75}$$

where $a_t = \sqrt{1 - \frac{2\gamma_t \kappa L_f}{\kappa + L_f}}$ and $\prod_{s'=t+1}^{t} \equiv 1$. Especially, if $\gamma_t \equiv \gamma$ and we let $a \equiv \sqrt{1 - \frac{2\gamma \kappa L_f}{\kappa + L_f}}$, we have

$$\|w^{(t+1)} - w^*\| \leq a^{t+1} \|w^0 - w^*\| + \zeta \gamma \frac{1 - a^{t+1}}{1 - a}. \tag{S76}$$

Finally, it is bounded that $\frac{1-a^{t+1}}{1-a} \leq \gamma / (1 - \sqrt{1 - \frac{2\gamma \kappa L_f}{\kappa + L_f}})$ which concludes the proof. ■

Proof of Theorem ??.

We do Taylor expansion for

$$\sum_{i=1}^{n} \phi(\alpha(f_{w^{(t+1)}}(X_i) - \mu)) \tag{S77}$$

and (S77) becomes

$$\begin{aligned}
&\sum_{i=1}^{n}\phi(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)})+\alpha(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)+(\hat{\mu}^{(t)}-\mu)))\\
=\;&\sum_{i=1}^{n}\phi(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))+\sum_{i=1}^{n}\phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))\alpha(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)+(\hat{\mu}^{(t)}-\mu))+\\
&+O(\sum_{i=1}^{n}\alpha^2(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)+(\hat{\mu}^{(t)}-\mu))^2). \qquad \text{(S78)}
\end{aligned}$$

Therefore, let $\mu^{(t+1)}$ be the solution to (S77) $= 0$. It satisfies

$$\begin{aligned}
\mu^{(t+1)} \;=\;& \frac{\sum_i \phi(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))}{\alpha\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))}+\hat{\mu}^{(t)}+\sum_i \nu_i^{(t)}(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i))\\
&+O(\frac{\sum_i \alpha^2(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)+(\hat{\mu}^{(t)}-\mu^{(t+1)}))^2}{\alpha\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))}). \qquad \text{(S79)}
\end{aligned}$$

By recalling the update that

$$\hat{\mu}^{(t+1)}=\hat{\mu}^{(t)}+\sum_i \nu_i^{(t)}(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)),$$

we could compute the difference, $\mathrm{err}_\mu^{(t+1)} := |\mu^{(t+1)}-\hat{\mu}^{(t+1)}|$,

$$\mu^{(t+1)}-\hat{\mu}^{(t+1)}=\underbrace{\frac{\sum_i \phi(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))}{\alpha\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))}}_{\text{term}_1}+\underbrace{O(\frac{\sum_i \alpha^2(f_{w^{(t+1)}}(X_i)-f_{w^{(t)}}(X_i)+(\hat{\mu}^{(t)}-\mu^{(t+1)}))^2}{\alpha\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))})}_{\text{term}_2}. \qquad \text{(S80)}$$

We next prove that (i) term_1 is $O_p(\mathrm{err}_\mu^{(t)}+\mathrm{err}_\mu^{(t)2})$; (ii) term_2 is $O_p(\alpha+\alpha^{p-1})$, where $p=1+\varepsilon$.

We first control the denominator term $\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)}))$ in term_1. By the definition of the Catoni's influence function, it is easy to see that $\phi'(0)=1$. Moreover, by (**??**) - (**??**), we know that $\phi'(x)\geq 1/2$ when $|x|\leq x_c$. We then define the index set $\mathcal{I}_{large}^{(t)} := \{i : |\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)})|\geq x_c\}$. It can be computed that

$$\mathbb{P}(|\alpha(f_{w^{(t)}}(X_i)-\hat{\mu}^{(t)})|\geq x_c)\leq C\alpha^p \qquad \text{(S81)}$$

for a universal constant C. By Hoeffding inequality for binary variables, we have

$$|\mathcal{I}_{large}^{(t)}| \leq \alpha^p n \log n \tag{S82}$$

with probability going to 1 as n goes to infinity. Therefore, by the property that $\phi'(x) \leq 1$, we have

$$n \geq \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) \geq \sum_{i \in \mathcal{I}_{large}^{(t)}} \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) \geq \frac{1}{2}(n - n\alpha^p \log n). \tag{S83}$$

For the numerator term $\sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))$ in term_1, we know

$$\begin{aligned}
\sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) &= \sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) - 0 \\
&= \sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) - \sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})) \\
&\qquad \text{(since } \mu^{(t)} \text{ is the solution to } \sum_i \phi(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})) = 0) \\
&\leq \sum_i \alpha\phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))|\hat{\mu}^{(t)} - \mu^{(t)}| + O(n\alpha^2|\hat{\mu}^{(t)} - \mu^{(t)}|^2) \\
&= \alpha \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))\text{err}_\mu^{(t)} + O(n\alpha^2 \text{err}_\mu^{(t)2}).
\end{aligned} \tag{S84}$$

Therefore, the first term can be bounded by,

$$\begin{aligned}
\text{term}_1 &\leq \frac{\alpha \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))\text{err}_\mu^{(t)} + O(n\alpha^2 \text{err}_\mu^{(t)2})}{\alpha \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))} \\
&\leq \text{err}_\mu^{(t)} + O(\alpha \text{err}_\mu^{(t)2}).
\end{aligned} \tag{S85}$$

For term_2, we define index set $\mathcal{I}_x^{(t)} := \{i : |f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)})| \geq x\}$ given level x. Following the idea of deriving (S69), we know $\sup_w |\hat{\mu}_f - m_f| = o_p(1)$. Thus both $\hat{\mu}^{(t)}$ and $\mu^{(t+1)}$ are bounded by some constant R with probability going to 1. Furthermore, $\mathbb{E}[|f_w^{(t+1)}(X_i) - f_w^{(t)}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)})|^p] \leq p(\mathbb{E}[|f_w^{(t+1)}(X_i)|^p] + \mathbb{E}[|f_w^{(t)}(X_i)|^p] + R^p) \leq \tilde{C}$ for some universal constant $\tilde{C}$.

Again, by adjusting the constant $\tilde{C}$, we can get that

$$\mathbb{P}(|f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)})| \geq x) \leq \tilde{C}/x^p \tag{S86}$$

and $|\mathcal{I}_x^{(t)}| \leq n \log n / x^p$ holds for all fixed x and t with probability going to 1 as n goes to infinity by using Hoeffding's inequality.

Then with probability going to 1, the numerator in term_2 can be bounded by

$$\begin{aligned}
& \sum_{i \notin \mathcal{I}_1^{(t)}} \alpha^2 (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)}))^2 \\
& + \sum_{i \in \mathcal{I}_1^{(t)}} \alpha^2 (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)}))^2 \\
\leq \; & \alpha^2 |\mathcal{I}_1^{(t)c}| + \sum_{i \in \mathcal{I}_1^{(t)}} \alpha^2 (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu))^2 \qquad \text{(S87)} \\
\leq \; & \alpha^2 n + \sum_{k=2}^{x_{cut}/\alpha} \sum_{i \in \mathcal{I}_{k-1}^{(t)} - \mathcal{I}_k^{(t)}} \alpha^2 (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i) + (\hat{\mu}^{(t)} - \mu^{(t+1)}))^2 \\
\leq \; & \alpha^2 n + \sum_{k=2}^{x_{cut}/\alpha} \sum_{i \in \mathcal{I}_{k-1}^{(t)} - \mathcal{I}_k^{(t)}} \alpha^2 k^2 \\
\leq \; & \alpha^2 n + 2\alpha^2 \int_1^{x_{cut}/\alpha} x n \log n / x^p dx \quad \text{(Fubini theorem)} \\
\leq \; & \alpha^2 n + 2\alpha^2 n \log n (x_{cut}/\alpha)^{2-p}, \qquad \text{(S88)}
\end{aligned}$$

where, by the definition, x_{cut} is the threshold that $\phi'(x) \equiv 0$ if $|x| \geq x_{cut}$.

Therefore,

$$\begin{aligned}
\text{term}_2 \; & \leq \; (\alpha^2 n + 2\alpha^2 n \log n (x_{cut}/\alpha)^{2-p}) / \big(\frac{\alpha}{2}(n - n\alpha^p \log n)\big) \\
& \leq \; 5(\log n)(\alpha + \alpha^{p-1}). \qquad \text{(S89)}
\end{aligned}$$

Putting everything together, we have

$$\text{err}_\mu^{(t+1)} \leq \text{err}_\mu^{(t)} + O(\alpha \text{err}_\mu^{(t)2}) + 5(\log n)(\alpha + \alpha^{p-1}). \qquad \text{(S90)}$$

By the choice of $\alpha = O(n^{-c_0})$ for some constant $c_0 \in (0,1)$, $\alpha \text{err}_\mu^{(t)2}$ is smaller than $(\log n)(\alpha + \alpha^{p-1})$ for any $t \leq T_{end}(\varrho)$. Thus we have

$$\text{err}_\mu^{(t+1)} \leq \text{err}_\mu^{(t)} + 6(\log n)(\alpha + \alpha^{p-1}) \qquad \text{(S91)}$$

and it arrives at

$$\mathrm{err}_\mu^{(t)} \le 6t(\log n)(\alpha + \alpha^{(p-1)}) \tag{S92}$$

for any $t \le T_{end}(\varrho)$. This also implies that $\hat{\mu}^{(t+1)} = \mu^{(t+1)} + O(\mathrm{err}_\mu^{(t)}) = m_{f_{w^{(t+1)}}} + o_p(1) + O(\mathrm{err}_\mu^{(t)})$ is bounded from above.

Moreover, we compare the difference between weights $\tilde{\nu}_i := \frac{\phi'(\alpha(f_{w^{(t)}}(X_i)-\mu^{(t)}))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\mu^{(t)}))}$ and $\nu_i^{(t)}$. That is,

$$\begin{aligned}
& |\tilde{\nu}_i - \nu_i^{(t)}| \\
= \; & |\frac{\phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)}))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)}))} - \frac{\phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))}| \\
\le \; & |\frac{|(\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) - \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})))\phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)}))|}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})) \cdot \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))}| \\
& +|\frac{|(\phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) - \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)}))) \cdot \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)}))|}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})) \cdot \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))}| \\
\le \; & C_{\phi''} n\alpha \mathrm{err}_\mu^{(t)} / (\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \mu^{(t)})) \cdot \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)}))) \\
& + C_{\phi''} \alpha \mathrm{err}_\mu^{(t)} / \sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) \\
\le \; & 3C_{\phi''} \alpha \mathrm{err}_\mu^{(t)} / n,
\end{aligned} \tag{S93}$$

where $C_{\phi''} := \max_{x:|x| \le x_{cut}} \phi''(x)$ and the last inequality uses the fact that $\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}^{(t)})) \ge \frac{1}{2}(n - n\alpha^p \log n)$ by (S83).

As a result, by recalling the definition of gradient $g^{(t)}$, we could get that

$$\begin{aligned}
|\zeta^{(t)}| \; &:= \; |\nabla_w \hat{\mu}_{f_{w^{(t)}}} - g^{(t)}| \\
&\le \; \sum_i |\tilde{\nu}_i - \nu_i^{(t)}| \cdot |\nabla_w f_{w^{(t)}}(X_i)| \\
&\le \; \sum_i \frac{3C_{\phi''} \alpha \mathrm{err}_\mu^{(t)}}{n} |\nabla_w f_{w^{(t)}}(X_i)| \\
&\le \; 3\tilde{C} \alpha \mathrm{err}_\mu^{(t)},
\end{aligned} \tag{S94}$$

elementwisely with probability going to 1, by adjusting the constant $\tilde{C}$. The last inequality uses the fact that, with probability going to 1, $\sum_i |\nabla_w f_w(X_i)|/n$ is uniformly upper bounded from above.

Finally, we analyze the difference $\hat{\mu}_f$ of at each time step,

$$\begin{aligned}
&\hat{\mu}_{f_{w^{(t+1)}}} - \hat{\mu}_{f_{w^{(t)}}} \leq \nabla_w \hat{\mu}_{f_{w^{(t)}}}(w^{(t+1)} - w^{(t)}) + \frac{L}{2}\|w^{(t+1)} - w^{(t)}\|^2 \\
= \; & -\gamma_t \nabla_w \hat{\mu}_{f_{w^{(t)}}} g^{(t)} + \frac{L\gamma_t^2}{2}\|g^{(t)}\|^2 \\
= \; & -\gamma_t \nabla_w \hat{\mu}_{f_{w^{(t)}}}(\nabla_w \hat{\mu}_{f_{w^{(t)}}} + \zeta^{(t)}) + \frac{L\gamma_t^2}{2}\|\nabla_w \hat{\mu}_{f_{w^{(t)}}} + \zeta^{(t)}\|^2 \\
\leq \; & -\gamma_t \|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\|^2 + \gamma_t \|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\|\|\zeta^{(t)}\| + L\gamma_t^2(\|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\|^2 + \|\zeta^{(t)}\|^2), \qquad \text{(S95)}
\end{aligned}$$

When $\|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\| \geq 2\|\zeta^{(t)}\|$ and $\gamma_t \leq 1/5L$, we have

$$\hat{\mu}_{f_{w^{(t+1)}}} - \hat{\mu}_{f_{w^{(t)}}} \leq -\frac{1}{4}\gamma_t \|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\|^2. \qquad \text{(S96)}$$

Moreover, from (S92) and (S94), we know $\|\zeta^{(t)}\| \leq \sqrt{d}\tilde{C}\alpha^p 36t(\log n)$. Since $T_{end}(\varrho) := \min\{t : \|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\| \leq \varrho\}$, therefore we have

$$T_{end}(\varrho) \leq \frac{\hat{\mu}_{f_{w^{(0)}}}}{\gamma \varrho^2} \qquad \text{(S97)}$$

as long as

$$\varrho \geq \sqrt{d}\tilde{C}\alpha^p 36t(\log n) \qquad \text{(S98)}$$

holds for any $t \leq T_{end}(\varrho)$. In other words, it suffices to have $\varrho \geq \left(36\tilde{C}\alpha^p\sqrt{d}(\log n)\frac{\hat{\mu}_{f_{w^{(0)}}}}{\gamma}\right)^{1/3}$ to make (S98) held. This concludes the proof. ■

I More Discussions on Algorithm ??

Remark S6 (**Comparison with Truncated Loss-based Methods**). *By straightforward calculations, we can find that the truncated loss based method (L. Xu, Yao, Yao, & Zhang, 2023; Y. Xu et al., 2020) is equivalent to assigning weight* $\nu_i^{(t),trunc} = \frac{\phi'(\alpha(f_{w^{(t)}}(X_i)))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)))}$ *to*

sample i in the t-th step. Then the following two observations explain why our algorithm is preferable. (i) Since $\mathbb{E}[f_w(X)] = m_{f_w}$ which is usually non-zero, therefore we should assign larger weights to those samples X_i's with $f_w(X_i)$ closer to m_{f_w} rather than those with $f_w(X_i)$ closer to 0. Therefore, truncated loss based method can lead to a larger bias than ours. (ii) Consider the weight formula $\nu_i^{(t)}(\mu) = \frac{\phi'(\alpha(f_{w^{(t)}}(X_i)-\mu))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i)-\mu))}$ by treating μ as an additional tuning parameter. Therefore, the truncated loss based method always fixes $\mu \equiv 0$ while our method allows μ to be updated adaptively.

Remark S7 (**Comparison with Coordinate Descent Methods**). *In the optimization literature, coordinate gradient descent (CD) is another popular approach. However, for the model with a large number of parameters, CD is less computationally efficient since each forward pass (i.e. the computation of loss value) is only used for one parameter update. Moreover, CD is very uncommon to be implemented in deep learning framework, since it requires sampling a single parameter coordinate throughout different neural layers in each update.*

References

Brownlees, C., Joly, E., & Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, *43*(6), 2507–2536.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'ihp probabilités et statistiques* (Vol. 48, pp. 1148–1185).

Chung, F., & Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, *3*(1), 79–127.

He, X., & Shao, Q.-M. (1996). A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, *24*(6), 2608–2630.

Holland, M., & Ikeda, K. (2019). Better generalization with less data using robust gradient descent. In *International conference on machine learning* (pp. 2761–2770).

Hsu, D., & Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, *17*(1), 543–582.

Merad, I., & Gaïffas, S. (2023). Robust supervised learning with coordinate gradient descent. *Statistics and Computing*, *33*(5), 116.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course* (Vol. 87). Springer Science & Business Media.

Sun, Q., Zhou, W.-X., & Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, *115*(529), 254–265.

Van Der Vaart, A. W., & Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Xu, L., Yao, F., Yao, Q., & Zhang, H. (2023). Non-asymptotic guarantees for robust statistical learning under infinite variance assumption. *Journal of Machine Learning Research*, *24*(92), 1–46.

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R., & Yang, T. (2020). Learning with non-convex truncated losses by sgd. In *Uncertainty in artificial intelligence* (pp. 701–711).