A. Proofs for Structure of the Influence Function

To prove that the influence function $\sigma(\mathcal{S})$ is a coverage function, the key is that non-negative combinations of coverage functions are still coverage functions. We state and prove the property for the case of combining two coverage functions, while for the general case we can simply repeat the argument.

Lemma 4. *Suppose $c^{(1)}$ and $c^{(2)}$ are two coverage functions mapping from $2^{\mathcal{V}}$ to $\mathbb{R}_+$. If $\alpha^{(1)} \geq 0$ and $\alpha^{(2)} \geq 0$, then $\sum_{\ell=1}^{2} \alpha^{(\ell)} c^{(\ell)}$ is also a coverage function mapping from $2^{\mathcal{V}}$ to $\mathbb{R}_+$.*

Proof. By definition, for $\ell = 1, 2$, there exists a universe $\mathcal{U}^{(\ell)}$, a set of weights $\left\{a_u^{(\ell)}\right\}_{u \in \mathcal{U}^{(\ell)}}$, and a family of subsets $\left\{\mathcal{A}_v^{(\ell)} : \mathcal{A}_v^{(\ell)} \subseteq \mathcal{U}^{(\ell)}\right\}_{v \in \mathcal{V}}$ such that for any $\mathcal{S} \subseteq \mathcal{V}$,

$$c^{(\ell)}(\mathcal{S}) = \sum_{u \in \bigcup_{v \in \mathcal{S}} \mathcal{A}_v^{(\ell)}} a_u^{(\ell)}.$$

Define a new universe $\mathcal{U} = \bigcup_{\ell=1}^{2} \mathcal{U}^{(\ell)}$, where elements in $\mathcal{U}^{(\ell)} (\ell = 1, 2)$ are treated as different elements. Define the corresponding weights a_u for $u \in \mathcal{U}$ as follows: if $u \in \mathcal{U}^{(\ell)}$, then $a_u = \alpha^{(\ell)} a_u^{(\ell)}$. Define a family of subsets $\{\mathcal{A}_v : \mathcal{A}_v \subseteq \mathcal{U}\}_{v \in \mathcal{V}}$ where $\mathcal{A}_v = \bigcup_{\ell=1}^{2} \mathcal{A}_v^{(\ell)}$. Then the corresponding coverage function is

$$\begin{aligned} c(\mathcal{S}) &= \sum_{u \in \bigcup_{v \in \mathcal{S}} \mathcal{A}_v} a_u = \sum_{u \in \bigcup_{\ell=1}^{2} \bigcup_{v \in \mathcal{S}} \mathcal{A}_v^{(\ell)}} a_u = \sum_{\ell=1}^{2} \sum_{u \in \bigcup_{v \in \mathcal{S}} \mathcal{A}_v^{(\ell)}} a_u = \sum_{\ell=1}^{2} \sum_{u \in \bigcup_{v \in \mathcal{S}} \mathcal{A}_v^{(\ell)}} \alpha^{(\ell)} a_u^{(\ell)} \\ &= \sum_{\ell=1}^{2} \alpha^{(\ell)} \sum_{u \in \bigcup_{v \in \mathcal{S}} \mathcal{A}_v^{(\ell)}} a_u^{(\ell)} = \sum_{\ell=1}^{2} \alpha^{(\ell)} c^{(\ell)}(\mathcal{S}). \end{aligned}$$

Therefore, $c(\mathcal{S}) = \sum_{\ell=1}^{2} \alpha^{(\ell)} c^{(\ell)}$ is a coverage function. □

Since $\Phi(\mathcal{S}|\boldsymbol{R})$ is a coverage function for any fixed $\boldsymbol{R}$, and the influence

$$\sigma(\mathcal{S}) = \mathbb{E}_{\boldsymbol{R} \sim p_{\boldsymbol{R}}} \left[\Phi(\mathcal{S}|\boldsymbol{R})\right]$$

is a convex combination of $\Phi(\mathcal{S}|\boldsymbol{R})$, we have the following corollary.

Corollary 5. *The influence function $\sigma(\mathcal{S})$ is a coverage function.*

Note If we naively construct the universe for the influence function as in the proof of Lemma 4, this will lead to a universe of size $2^d d$, which is exponential in d. It seems to imply that the function is difficult to learn. However, as shown in (Badanidiyuru et al., 2012), there exists a coverage function that is a $(1+\epsilon)$ multiplicative approximation to σ, and is defined on a universe of size $O\left(\frac{d^2}{\epsilon^2}\right)$. This suggests that there are structures in a coverage function that make learning tractable, even if it is defined on an exponentially large universe. On the other hand, the proof in (Badanidiyuru et al., 2012) does not immediately lead to an efficient learning algorithm, since the construction explicitly makes use of the weights of the elements in the universe defining σ.

B. Proofs for Random Basis Function Approximation

In this section, we fix a node j and $f_j(\chi_{\mathcal{S}}) = \mathbb{E}_{r \sim p_j(r)} \left[\phi(\chi_{\mathcal{S}}^\top r)\right]$. Suppose a set of K random features $\{r_{j1}, \ldots, r_{jK}\}$ is drawn from the distribution $q_j(r)$ over $\{0,1\}^n$. We show that given sufficiently many random features, there exists a convex combination of the random basis functions that approximates the truth f_j.

The number of random features needed depends on how close the sample distribution q_j is to the true distribution p_j. The "distance" between the two is formalized in the following definition.

Definition 6. *Let C be the minimum value such that*

$$p_j(r) \leq C q_j(r) \text{ for all } j \in [d], r \in \{0,1\}^n.$$

We first introduce an intermediate class $\widetilde{\mathcal{F}}^w$ that depends on C, and show that there exists a function in $\widetilde{\mathcal{F}}^w$ that is close to f_j. We then utilize the structure of our problem to show that the same is true for a class $\widehat{\mathcal{F}}^w$ that does not depend on C. In

particular, define

$$\widetilde{\mathcal{F}}^w := \left\{ f^w(\chi_{\mathcal{S}}) = \sum_{k=1}^K w_k \phi(\chi_{\mathcal{S}}^\top r_{jk}) \middle| 0 \le w_k \le \frac{C}{K} \right\}, \tag{19}$$

$$\widehat{\mathcal{F}}^w := \left\{ f^w(\chi_{\mathcal{S}}) = \sum_{k=1}^K w_k \phi(\chi_{\mathcal{S}}^\top r_{jk}) \middle| w_k \ge 0, \sum_{k=1}^K w_k \le 1 \right\}. \tag{20}$$

Lemma 7. *Let p_χ be any distribution of $\chi_{\mathcal{S}}$. If $r_{j1}, \dots, r_{jK}$ are drawn i.i.d. from $q_j(r)$, then with probability at least $1-\delta$ over $r_{j1}, \dots, r_{jK}$, there exists $\widetilde{f} \in \widetilde{\mathcal{F}}^w$ such that*

$$\Pr_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left| \widetilde{f}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}) \right| \ge \epsilon \right] \ge \epsilon^2 / C$$

when $K = O(\frac{C^2}{\epsilon^2} \log \frac{C}{\epsilon\delta})$. Consequently,

$$\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}}) \right)^2 \right] \le 3\epsilon^2.$$

Proof. Here we prove the first statement, which is stronger and implies the second one. Let $f^k(\chi_{\mathcal{S}}) = \frac{p(r_{jk})}{q(r_{jk})} \phi(\chi_{\mathcal{S}}^\top r_{jk})$ for $k = 1, \dots, K$. Then $\mathbb{E}_{r_{jk} \sim q_j(r)}[f^k] = f_j$. Let $\widetilde{f}(\chi_{\mathcal{S}}) = \frac{1}{K} \sum_{i=1}^K \frac{p(r_{jk})}{q(r_{jk})} \phi(\chi_{\mathcal{S}}^\top r_{jk})$ be the sample average of these functions. Then $\widetilde{f} \in \widetilde{\mathcal{F}}^w$ since $0 \le \frac{1}{K} \frac{p(r_{jk})}{q(r_{jk})} \le \frac{C}{K}$.

By Hoeffding's inequality, when $K = O(\frac{C^2}{\epsilon^2} \log \frac{C}{\delta\epsilon})$, for any fixed $\mathcal{S}$ we have

$$\Pr_r \left[\left| \widetilde{f}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}) \right| \ge \epsilon \right] \le \delta\epsilon^2 / C$$

where $\Pr_r$ is over the random sample of $r_{j1}, \dots, r_{jK}$. This leads to

$$\Pr_{\chi_{\mathcal{S}} \sim p_\chi} \Pr_r \left[\left| \widetilde{f}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}) \right| \ge \epsilon \right] \le \delta\epsilon^2 / C.$$

Exchanging $\Pr_{\chi_{\mathcal{S}} \sim p_\chi}$ and $\Pr_r$ by Fubini's theorem, and then by Markov's inequality, we have

$$\Pr_r \left\{ \Pr_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left| \widetilde{f}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}) \right| \ge \epsilon \right] \ge \epsilon^2 / C \right\} \le \delta.$$

This means with probability at least $1-\delta$ over the random sample of $r_{j1}, \dots, r_{jK}$, on at least $1 - \epsilon^2/C$ probability mass of the distribution of $\mathcal{S}$, $[\widetilde{f}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}})]^2 \le \epsilon^2$. Since $|\widetilde{f}(\chi_{\mathcal{S}})| \le C$ and $|f_j(\chi_{\mathcal{S}})| \le 1$,

$$\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}}) \right)^2 \right] \le \epsilon^2(1-\epsilon^2) + (C+1)\epsilon^2/C < 3\epsilon^2.$$

□

Note that for learning over $\widetilde{\mathcal{F}}^w$, the parameter C needs to be determined. However, there are additional structures in our problem that can be utilized to further restrict $\widetilde{\mathcal{F}}^w$ and get rid of the dependence on C.

Lemma 1. *Let p_χ be any distribution of $\chi_{\mathcal{S}}$. If $K = O(\frac{C^2}{\epsilon^2} \log \frac{C}{\epsilon\delta})$ and $r_{j1}, \dots, r_{jK}$ are drawn iid from $q_j(r)$, then with probability at least $1-\delta$ over $r_{j1}, \dots, r_{jK}$, there exists $\widehat{f} \in \widehat{\mathcal{F}}^w$ such that*

$$\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}(\chi_{\mathcal{S}}) \right)^2 \right] \le \epsilon^2.$$

Proof. Construct a distribution $\Delta_{\mathbf{1}}$ that assigns probability 1 to $\chi_{\mathcal{S}} = \mathbf{1}$ and probability 0 to all other source sets. Note that the definition of $\widetilde{f}$ is independent of the distribution of $\chi_{\mathcal{S}}$, so that we can apply Lemma 7 for $\widetilde{f}$ on both $\Delta_{\mathbf{1}}$ and p_χ.

Without loss of generality, assume $r_{jk} \ne \mathbf{0}$ for any k, since otherwise we can remove r_{jk} without changing $\widetilde{f}$. Then $\mathbf{1}^\top r_{jk} > 0$ and thus $\widetilde{f}(\mathbf{1}) = \sum_{k=1}^K w_k$. By Lemma 7 on $\Delta_{\mathbf{1}}$, with probability $1 - \delta/2$ we have

$$\sqrt{\mathbb{E}_{\chi_{\mathcal{S}} \sim \Delta_{\mathbf{1}}} \left[\left(f_j(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}}) \right)^2 \right]} = |f_j(\mathbf{1}) - \widetilde{f}(\mathbf{1})| = \left| f_j(\mathbf{1}) - \sum_{k=1}^K w_k \right| \le \frac{\epsilon}{2}.$$

Then $\sum_{k=1}^{K} w_k \le f_j(\mathbf{1}) + \frac{\epsilon}{2} \le 1 + \frac{\epsilon}{2}$. Define $\widehat{f} = \widetilde{f}/(1+\epsilon/2)$. Then $\widehat{f} \in \widehat{\mathcal{F}}^w$ and

$$|\widehat{f}(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}})| = \frac{\epsilon/2}{1+\epsilon/2}\widetilde{f}(\chi_{\mathcal{S}}) \le \frac{\epsilon/2}{1+\epsilon/2}\sum_{k=1}^{K} w_k \le \frac{\epsilon}{2}.$$

By Lemma 7 on p_χ, with probability $1-\delta/2$ we have

$$\mathbb{E}_{\chi_{\mathcal{S}}\sim p_\chi}\left[\left(f_j(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}})\right)^2\right] \le \frac{\epsilon^2}{4}.$$

Then we have

$$\mathbb{E}_{\chi_{\mathcal{S}}\sim p_\chi}\left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}(\chi_{\mathcal{S}})\right)^2\right] \le 2\mathbb{E}_{\chi_{\mathcal{S}}\sim p_\chi}\left[\left(f_j(\chi_{\mathcal{S}}) - \widetilde{f}(\chi_{\mathcal{S}})\right)^2 + \left(\widetilde{f}(\chi_{\mathcal{S}}) - \widehat{f}(\chi_{\mathcal{S}})\right)^2\right] \le 2\left(\frac{\epsilon}{2}\right)^2 + \frac{2\epsilon^2}{4} = \epsilon^2$$

which completes the proof. □

C. Proofs for Sample Complexity

In this section, we provide the complete proof for the sample complexity of learning the weights of the random basis functions by maximum likelihood estimation (MLE). We are not aware of any previous work providing the analysis of MLE for the hypothesis class in our problem (the weighted sum of the random basis functions). Therefore, we adopt the general framework in (Birgé & Massart, 1998), which analyzes the sample complexity based on a particular dimension notion for the hypothesis class. Then we bound the dimension of our hypothesis class, which then leads to our sample bound. The techniques used in bounding the dimension can be extended to other hypothesis classes, and thus may be of independent interest.

In the following, we first review the framework and paraphrase their result for distributions over a discrete domain, since this suffices for our purpose. We then apply the result to learning the conditional probability f_j for an individual node j, and finally prove the bound for the entire influence function.

C.1. Review of MLE for probability estimation

The MLE estimator is defined as follows. Suppose we observe m data points $Z_1, \ldots, Z_m$ independent identically distributed according to the true probability function p^* over a discrete domain $\mathcal{Z}$. The hypothesis class $\mathcal{H}$ is a set of functions, each of which is the square root[1] of a probability function. That is, for each $h \in \mathcal{H}$, $h = \sqrt{p_h}$ where p_h is a probability over $\mathcal{Z}$. The MLE estimator is $\widehat{h} = \operatorname{argmax}_{h\in\mathcal{H}} \sum_{i=1}^{m} \log\left[h(Z_i)\right]$. More generally, an approximate MLE estimator is $\widehat{h}$ such that

$$\sum_{i=1}^{m} \log\left[\widehat{h}(Z_i)\right] + 1 \ge \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \log\left[h(Z_i)\right]. \tag{21}$$

The goal is to analyze how the difference between $\hat{h}$ and the truth $h^* = \sqrt{p^*}$ decreases with the sample size m.

Complexity of the hypothesis class To analyze the sample complexity, we need to introduce some metric over the hypotheses and some notion bounding the complexity of the hypothesis class based on the metric. Given $h, \widetilde{h}$ that are the square roots of two probabilities, the ℓ_2 distance is

$$d(h,\widetilde{h}) := \|h - \widetilde{h}\| = \sqrt{\sum_{Z\in\mathcal{Z}} [h(Z) - \widetilde{h}(Z)]^2}. \tag{22}$$

Note that $d(h,\widetilde{h})/\sqrt{2}$ is just the Hellinger distance. Similar to the ℓ_2 distance, we can define ℓ_∞ distance:

$$d_\infty(h,\widetilde{h}) := \|h - \widetilde{h}\|_\infty = \max_{Z\in\mathcal{Z}} |h(Z) - \widetilde{h}(Z)|. \tag{23}$$

Both the ℓ_2 and ℓ_∞ distances are bounded over all square roots of probabilities, so a hypothesis class with such metrics is always a bounded metric space. To measure the complexity of such a metric space, a common notion is the following:

Definition 8. *Given a set $\mathcal{B}$ equipped with metric d, and a real number $\epsilon > 0$, $\mathcal{T} \subseteq \mathcal{B}$ is an ϵ-covering of $\mathcal{B}$ if the following*

[1] We will always talk about the square root of the probabilities. This is because the ℓ_2 distance over such hypotheses correspond to the Hellinger distance, which plays a key role in the analysis of MLE and appears in the final bound.

holds: for every $h \in \mathcal{B}$ there exists $\widetilde{h} \in \mathcal{T}$ such that $d(h, \widetilde{h}) < \epsilon$.

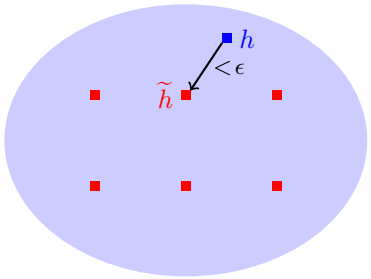


Figure 3. Illustration of ϵ-covering.

Intuitively, if we construct balls aroud points in $\mathcal{T}$ with radius ϵ, then these balls can cover all points in $\mathcal{B}$. Note that the dimension depends on the metric d. The result in (Birgé & Massart, 1998) actually depends on both the ℓ_2 and ℓ_∞ metrics on $\mathcal{H}$. More precisely, we introduce the following $\ell_{2,\infty}$ dimension[2].

Definition 9 ((Birgé & Massart, 1998)). *The $\ell_{2,\infty}$ dimension of $\mathcal{H}$ is the minimum $D \geq 1$ such that there exist constants $c_0 \geq 1$ and $c_1 \geq 1$ satisfying the following. For each $\epsilon > 0$ and each ball $\mathcal{B} \subseteq \mathcal{H}$ with radius $R \geq 5\epsilon$, one can find $\mathcal{T}$ with*

$$|\mathcal{T}| \leq (c_0 R/\epsilon)^D$$

that is an ϵ-covering of $\mathcal{B}$ for the ℓ_2 metric and a $c_1\epsilon$-covering for the ℓ_∞ metric.

The condition says that for any given distance threshold ϵ and any sufficiently large ball in $\mathcal{H}$, we can find a finite $O(\epsilon)$-covering $\mathcal{T}$ that is simultaneously with respect to both the ℓ_2 metric and the ℓ_∞ metric, and the size of the covering depends exponentially on the dimension D.

Sample complexity based on $\ell_{2,\infty}$ dimension The following result bounds the expected squred ℓ_2 distance between the MLE estimator and the truth, by a constant times the best Kullback-Leibler divergence from the truth to any hypothesis, plus a penalty term roughly $\tilde{O}(D/m)$ where D is the dimension of $\mathcal{H}$ and m is the number of data points. The Kullback-Leibler divergence between $h, \widetilde{h} \in \mathcal{H}$ is defined as

$$\mathrm{KL}(h, \widetilde{h}) := \mathbb{E}_{Z \sim p_h(Z)} \left[\log \frac{h^2(Z)}{\widetilde{h}^2(Z)} \right]. \tag{24}$$

Theorem 10 (Theorem 3 in (Birgé & Massart, 1998)). *Assume $\mathcal{H}$ has $\ell_{2,\infty}$ dimension $D \in [1, m]$. Let $\widehat{h}$ be an approximate MLE estimator,* i.e.*, it satisfies (21). Then there is a constant $c > 0$ such that*

$$\mathbb{E}_{\mathcal{D}_m}[d^2(h^*, \widehat{h})] \leq c \inf_{h \in \mathcal{H}} \mathrm{KL}(h^*, h) + \frac{cD}{m}(1 + \log[c_0(1 + c_1)])$$

where $\mathbb{E}_{\mathcal{D}_m}$ is with respect to the randomness in the data $Z_1, \ldots, Z_m$ generated from the true distribution (h^*)[2].

On the right hand side of the bound is the Kullback-Leibler divergence, instead of the the squared distance as on the left. The following lemma is useful for connecting the two.

Lemma 11 (Eqn. (7.5) and (7.6) in Lemma 5 in (Birgé & Massart, 1998)). *If h and $\widetilde{h}$ are the square roots of two probabilities and $\|h/\widetilde{h}\|_\infty < +\infty$, then*

$$d^2(h, \widetilde{h}) \leq \mathrm{KL}(h, \widetilde{h}) \leq 2[1 + \log \|h/\widetilde{h}\|_\infty] d^2(h, \widetilde{h}).$$

[2]The result (Birgé & Massart, 1998) actually depends on a covering property, which basically says that the $\ell_{2,\infty}$ dimension of $\mathcal{H}$ is bounded by D. For our purpose, it is more convenient to introduce a definition of the dimension. Also note that in (Birgé & Massart, 1998), the covering property actually requires that $\mathcal{T}$ is simultaneously an ϵ-net of $\mathcal{B}$ for the ℓ_2 metric and a $c_1\epsilon$-net for the ℓ_∞ metric. But in fact, this requirement can be relaxed to that $\mathcal{T}$ is a covering (instead of a net) as in our definition. See Assumption $\boldsymbol{M}'_{2,\infty}$ and Theorem 10 in their subsequent work (**?**).

C.2. Estimation for individual node

Here we consider learning f_j for a fixed node j. Assume that the event stated in Lemma 1 happens, and fix the set of random features $r_{j1}, \ldots, r_{jK}$. We first formalize our hypothesis class for learning f_j, and then analyze the sample complexity.

Hypotheis class Recall that for learning f_j, we get training data in the form $Z_i = (\chi_{\mathcal{S}_i}, y_{ij})$, where $\chi_{\mathcal{S}_i} \in \{0,1\}^d$ is the indicator vector of $\mathcal{S}_i$ and $y_{ij} \in \{0,1\}$ indicates whether node j gets influenced by $\mathcal{S}_i$. Let p^* denote the true distribution

$$p^*(\chi_{\mathcal{S}}, y) = p_\chi(\chi_{\mathcal{S}}) p(y|\chi = \chi_{\mathcal{S}})$$

where p_χ is the distribution of $\chi_{\mathcal{S}}$, and $p(y|\chi = \chi_{\mathcal{S}})$ is the conditional probability

$$p(y|\chi = \chi_{\mathcal{S}}) = [f_j(\chi_{\mathcal{S}})]^y [1 - f_j(\chi_{\mathcal{S}})]^{1-y}.$$

Similarly, given a function f, define the distribution induced as

$$p(\chi_{\mathcal{S}}, y|f) = p_\chi(\chi_{\mathcal{S}}) p(y|\chi = \chi_{\mathcal{S}}, f) \quad \text{where} \quad p(y|\chi = \chi_{\mathcal{S}}, f) = [f(\chi_{\mathcal{S}})]^y [1 - f(\chi_{\mathcal{S}})]^{1-y}.$$

We could define our hypothesis class as the square roots of the probability distributions induced by functions in $\widehat{\mathcal{F}}^w$. Unfortunately, there is some subtle technical difficulty: $p(\chi_{\mathcal{S}}, y|f)$ can be arbitrarily close to 0, in which case our technique for bounding the dimension of our hypothesis class fails (in particular, we cannot construct coverings for our hypotheses based on coverings for the weights; see the proof of Lemma 15). Therefore, we add a small offset to functions in $\widehat{\mathcal{F}}^w$ and ensure that they are bounded away from 0. More precisely, define

$$\widehat{\mathcal{F}}^{w,\lambda} := \left\{ f^{w,\lambda} \;\middle|\; f^{w,\lambda} = f^w + \lambda, f^w \in (1-2\lambda)\widehat{\mathcal{F}}^w \right\} \tag{25}$$

where $\lambda \in (0,1)$ is a constant whose value will be determined later. For any $f^{w,\lambda} \in \widehat{\mathcal{F}}^{w,\lambda}$, we have $\lambda \le f^{w,\lambda}(\chi_{\mathcal{S}}) \le 1-\lambda$ for any $\chi_{\mathcal{S}}$. Then the probability $p(\chi_{\mathcal{S}}, y|f^{w,\lambda})$ introduced by $f^{w,\lambda}$ satisfies that $p(\chi_{\mathcal{S}}, y|f^{w,\lambda}) \ge \lambda$ for any $\chi_{\mathcal{S}}$ and y, which will allow us to use our technique.

Still, for $\widehat{\mathcal{F}}^{w,\lambda}$ to be meaningful, we need to show there exists a function in $\widehat{\mathcal{F}}^{w,\lambda}$ close to f_j. The following lemma shows that this is true as long as λ is small.

Lemma 12. *Assume that the statement in Lemma 1 happens. Then there exists $\widehat{f}^{w,\lambda} \in \widehat{\mathcal{F}}^{w,\lambda}$ such that*

$$\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}^{w,\lambda}(\chi_{\mathcal{S}}) \right)^2 \right] \le 2\epsilon^2 + 2\lambda^2.$$

Proof. Let $\widehat{f}^w \in \widehat{\mathcal{F}}^w$ be such that $\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}^w(\chi_{\mathcal{S}}) \right)^2 \right] \le \epsilon^2$. Define $\widehat{f}^{w,\lambda} = (1-2\lambda)\widehat{f}^w + \lambda$. Then $|\widehat{f}^w(\chi_{\mathcal{S}}) - \widehat{f}^{w,\lambda}(\chi_{\mathcal{S}})| = |\lambda - 2\lambda \widehat{f}^w(\chi_{\mathcal{S}})| \le \lambda$. The lemma then follows from

$$\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}^{w,\lambda}(\chi_{\mathcal{S}}) \right)^2 \right] \le 2\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(f_j(\chi_{\mathcal{S}}) - \widehat{f}^w(\chi_{\mathcal{S}}) \right)^2 \right] + 2\mathbb{E}_{\chi_{\mathcal{S}} \sim p_\chi} \left[\left(\widehat{f}^{w,\lambda}(\chi_{\mathcal{S}}) - \widehat{f}^w(\chi_{\mathcal{S}}) \right)^2 \right].$$

□

Therefore, our hypothesis class is defined as

$$\mathcal{H}_K := \left\{ \sqrt{p(\chi_{\mathcal{S}}, y|f^{w,\lambda})} \;\middle|\; f^{w,\lambda} \in \widehat{\mathcal{F}}^{w,\lambda} \right\}. \tag{26}$$

In other words, $\mathcal{H}_K$ is the square roots of the probabilities induced by $\widehat{\mathcal{F}}^{w,\lambda}$. Let $h^* = \sqrt{p^*(\chi_{\mathcal{S}}, y)}$ denote the element corresponding to the true distribution. Note that we do not assume h^* is in $\mathcal{H}_K$.

Sample complexity To bound the dimension of $\mathcal{H}_K$ and apply Theorem 10, the key is to construct coverings for $\mathcal{H}_K$ based on those for the weights, since the feasible set of weights is a subset of $\mathbb{R}^K$ which has nice structure. We first relate the topology of $\mathcal{H}_K$ to that of the weights w in Lemma 14, which makes the construction possible. We then bound on the dimension in Lemma 15, and subsequently bound the sample complexity in Lemma 2.

To begin with, let $\Delta := \{w | w \ge \mathbf{0}, \|w\|_1 \le 1 - 2\lambda\}$ denote the feasible set of the weights w of the functions in $\widehat{\mathcal{F}}^{w,\lambda}$, and consider a mapping $\pi : \Delta \to \mathcal{H}_K$ as follows:

$$\pi(w) := \sqrt{p(\cdot|f^{w,\lambda})}, \text{ where } f^{w,\lambda}(\chi_{\mathcal{S}}) = \sum_{k=1}^K w_k \phi(\chi_{\mathcal{S}}^\top r_k) + \lambda.$$

Lemma 14 shows that the ℓ_2 distance between $\pi(w)$ and $\pi(w')$ is approximately the ℓ_∞ distance between w and w', relating the topology of $\mathcal{H}_K$ to that of the weights w. The following quantity is useful in the process:

Definition 13. *Let $A^j = \Sigma\Phi^j$ where Σ is a $2^n \times 2^n$ diagonal matrix with entries $\Sigma_{\chi_\mathcal{S},\chi_\mathcal{S}} = \sqrt{p_\chi(\chi_\mathcal{S})}$, and Φ^j is a $2^n \times K$ matrix with entries $\Phi^j_{\chi_\mathcal{S},k} = \phi(\chi_\mathcal{S}^\top r_{jk})$. Define*

$$\Lambda^j := \min_{w \neq 0} \frac{\|A^j w\|}{\|w\|}, \quad \Lambda = \min_{j \in [d]} \Lambda^j.$$

Intuitively, Λ reflects how the change in w affects $A^j w$, which subsequently affects the corresponding hypothesis in $\mathcal{H}_K$. This quantity thus goes into the relation between the distance on the set of w and the distance on $\mathcal{H}_K$, as shown in Lemma 14.

Lemma 14. *For an $w, w' \in \Delta$,*

$$\frac{\Lambda}{2}\|w - w'\|_\infty \leq \|\pi(w) - \pi(w')\| \leq \frac{K}{\sqrt{2\lambda}}\|w - w'\|_\infty.$$

Proof. For simplicity, let f be a shorthand of $f^{w,\lambda}(\chi_\mathcal{S})$ and f' be a shorthand of $f^{w',\lambda}(\chi_\mathcal{S})$ in the proof.
(1) By definition of the norm in (22), we have

$$\begin{aligned}
\|\pi(w) - \pi(w')\|^2 &= \sum_{(\chi_\mathcal{S}, y)} (\sqrt{p(\chi_\mathcal{S}, y|f)} - \sqrt{p(\chi_\mathcal{S}, y|f')})^2 \\
&= \sum_{\chi_\mathcal{S}} p_\chi(\chi_\mathcal{S}) \sum_y (\sqrt{p(y|\chi = \chi_\mathcal{S}, f)} - \sqrt{p(y|\chi = \chi_\mathcal{S}, f')})^2 \\
&= \sum_{\chi_\mathcal{S}} p_\chi(\chi_\mathcal{S}) \left[(\sqrt{f} - \sqrt{f'})^2 + (\sqrt{1-f} - \sqrt{1-f'})^2 \right] \\
&= \mathbb{E}_{p_\chi} \left[(\sqrt{f} - \sqrt{f'})^2 + (\sqrt{1-f} - \sqrt{1-f'})^2 \right].
\end{aligned}$$

This leads to

$$\|\pi(w) - \pi(w')\|^2 \geq \mathbb{E}_{p_\chi} \left[(\sqrt{f} - \sqrt{f'})^2 \right] \geq \frac{1}{4}\mathbb{E}_{p_\chi} \left[(f - f')^2 \right] = \frac{1}{4} \sum_{\chi_\mathcal{S}} p_\chi(\chi_\mathcal{S})(f - f')^2$$

where the second inequality follows from Lemma 16. The right hand side expands to

$$\frac{1}{4} \sum_{\chi_\mathcal{S}} p_\chi(\chi_\mathcal{S})(f - f')^2 = \frac{1}{4} \sum_{\chi_\mathcal{S}} p_\chi(\chi_\mathcal{S}) \left[\sum_{k=1}^K \phi(\chi_\mathcal{S}^\top r_{jk})(w_k - w'_k) \right]^2 = \frac{1}{4}\|A^j w - A^j w'\|^2.$$

where the last step follows from the definition of A^j. So

$$\|\pi(w) - \pi(w')\|^2 \geq \frac{1}{4}\|A^j w - A^j w'\|^2 \geq \frac{\Lambda^2}{4}\|w - w'\|^2 \geq \frac{\Lambda^2}{4}\|w - w'\|_\infty^2$$

where the second inequality follows from the definition of Λ.
(2) By definition we have

$$|f(\chi_\mathcal{S}) - f'(\chi_\mathcal{S})| \leq \|w - w'\|_1 \leq K\|w - w'\|_\infty$$

for any $\chi_\mathcal{S}$. Then

$$\begin{aligned}
\|\pi(w) - \pi(w')\|^2 &= \mathbb{E}_{p_\chi} \left[(\sqrt{f} - \sqrt{f'})^2 + (\sqrt{1-f} - \sqrt{1-f'})^2 \right] \\
&\leq \mathbb{E}_{p_\chi} \left[\frac{(f - f')^2}{4\lambda} + \frac{((1-f) - (1-f'))^2}{4\lambda} \right] \leq \frac{K^2}{2\lambda}\|w - w'\|_\infty^2
\end{aligned}$$

where the first inequality follows from Lemma 16.(2) and the fact that $\lambda \leq f \leq 1 - \lambda$ and $\lambda \leq f' \leq 1 - \lambda$. □

Lemma 15. *The $\ell_{2,\infty}$ dimension of $\mathcal{H}_K$ is at most K.*

Proof. To bound the dimension, the key is to construct coverings of small sizes. By Lemma 14, the ℓ_2 metric on $\mathcal{H}_K$ approximately corresponds to the ℓ_∞ metric on the set of weights. So based on coverings for the weights with respect to

the ℓ_∞ metric, we can construct coverings for $\mathcal{H}_K$ with respect to the ℓ_2 metric. We then show that they are also coverings with respect to the ℓ_∞ metric. The bound on the dimension then follows from the sizes of these coverings.

More precisely, given $\epsilon > 0$ and a ball $\mathcal{B} \subseteq \mathcal{H}_K$ with radius $R > 5\epsilon$, we construct an ϵ-covering $\mathcal{T}$ as follows. Define $\mathcal{B}^w = \pi^{-1}(\mathcal{B})$. By Lemma 14, the radius of $\mathcal{B}^w$ is at most $R^w = \frac{2}{\Lambda}R$ (with respect to the ℓ_∞ metric). Now consider finding an ϵ^w-covering for $\mathcal{B}^w$ with respect to the ℓ_∞ metric, where $\epsilon^w = (\frac{K}{\sqrt{2\lambda}})^{-1}\epsilon$. Since $\mathcal{B}^w \subseteq \mathbb{R}^K$, by taking the grid with length $\epsilon^w/2$ on each dimension, we can get such a covering $\mathcal{T}^w$ with

$$|\mathcal{T}^w| \le \left(\frac{4R^w}{\epsilon^w}\right)^K \le \left(\frac{8K}{\sqrt{2\lambda}\Lambda}\frac{R}{\epsilon}\right)^K.$$

Let $\mathcal{T} = \pi(\mathcal{T}^w)$, and for any $h \in \mathcal{B}$ find $\widetilde{h}$ as follows. Suppose $w_h \in \mathcal{B}^w$ satisfies $\pi(w_h) = h$ and $w_{\widetilde{h}}$ is the nearest neighbor of w_h in $\mathcal{T}^w$, then we set $\widetilde{h} = \pi(w_{\widetilde{h}})$. See Figure 4 for an illustration.

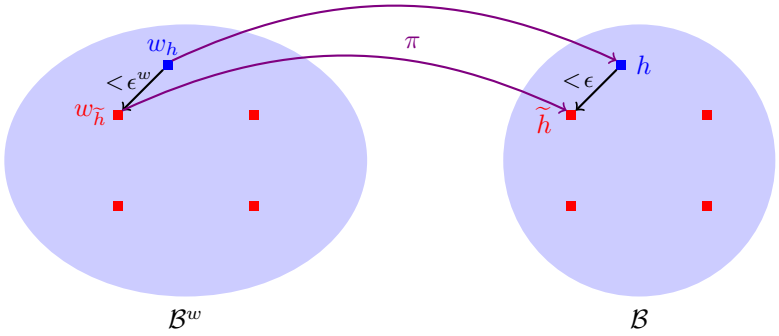


Figure 4. Illustration of the mapping.

First, we argue that $\mathcal{T}$ is an ϵ-covering w.r.t. the ℓ_2 metric, *i.e.*, $d(h, \widetilde{h}) < \epsilon$ for any $h \in \mathcal{B}$. It follows from Lemma 14:

$$d(h, \widetilde{h}) \le \frac{K}{\sqrt{2\lambda}}\|w_h - w_{\widetilde{h}}\|_\infty < \frac{K}{\sqrt{2\lambda}}\epsilon^w = \epsilon.$$

Second, we argue that $\mathcal{T}$ is also an $O(\epsilon)$-covering w.r.t. the ℓ_∞ metric, *i.e.*, $d_\infty(h, \widetilde{h}) = \|h - \widetilde{h}\|_\infty = O(\epsilon)$ for any $h \in \mathcal{B}$. We have $\|h - \widetilde{h}\| < \epsilon$, then $\|w_h - w_{\widetilde{h}}\|_\infty < \frac{2}{\Lambda}\epsilon$ by Lemma 14. Let $f_h := f^{w_h,\lambda}$ and $f_{\widetilde{h}} := f^{w_{\widetilde{h}},\lambda}$. Then

$$|f_h(\chi_\mathcal{S}) - f_{\widetilde{h}}(\chi_\mathcal{S})| \le \|w_s - w_t\|_1 \le K\|w_s - w_t\|_\infty < \frac{2K}{\Lambda}\epsilon$$

for any $\chi_\mathcal{S}$, and thus

$$\begin{aligned}\|\pi(w_s) - \pi(w_t)\|_\infty &= \max_{\chi_\mathcal{S}} \max\left\{|\sqrt{f_h} - \sqrt{f_{\widetilde{h}}}|, |\sqrt{1-f_h} - \sqrt{1-f_{\widetilde{h}}}|\right\} \\ &\le \max_{\chi_\mathcal{S}} |\sqrt{f_h} - \sqrt{f_{\widetilde{h}}}| \le \max_{\chi_\mathcal{S}} \frac{|f_h - f_{\widetilde{h}}|}{2\sqrt{\lambda}} < \frac{K}{\Lambda\sqrt{\lambda}}\epsilon\end{aligned}$$

where the second inequality follows from Lemma 16.(2).

So the conditions in the definition of the dimension are satisfied with $D = K$ and $c_0 = c_1 = O\left(\frac{K}{\Lambda\sqrt{\lambda}}\right)$, and thus the dimension of $\mathcal{H}_K$ is at most K. □

Lemma 2. *Assume the statement in Lemma 1 happens. Let $\widehat{h}$ be an approximate MLE estimator,* i.e.*, it satisfies (21). Let $\widehat{f}_j^{w,\lambda}$ be the corresponding function in $\widehat{\mathcal{F}}^{w,\lambda}$. Then when $m = O(\frac{K}{\epsilon}\log\frac{K}{\lambda\Lambda})$,*

$$\mathbb{E}_{\mathcal{D}^m}\left[\mathbb{E}_{p_\chi}[(\widehat{f}(\chi_\mathcal{S}) - f_j(\chi_\mathcal{S}))^2]\right] \le 8c\left(\epsilon + \frac{\epsilon^2 + \lambda^2}{\lambda}\left[1 + \log\frac{1}{\lambda}\right]\right)$$

where c is the constant in Theorem 10.

Proof. The lemma follows from Theorem 10 and Lemma 15. On the left hand side of that bound in Theorem 10, we have

$$d^2(h^*, \widehat{h}) = \mathbb{E}_{p_\chi}\left[\left(\sqrt{\widehat{f}_j^{w,\lambda}} - \sqrt{f_j}\right)^2 + \left(\sqrt{1-\widehat{f}_j^{w,\lambda}} - \sqrt{1-f_j}\right)^2\right]$$
$$\geq \mathbb{E}_{p_\chi}\left[\left(\sqrt{\widehat{f}_j^{w,\lambda}} - \sqrt{f_j}\right)^2\right] \geq \frac{1}{4}\mathbb{E}_{p_\chi}\left[\left(\widehat{f}_j^{w,\lambda} - f_j\right)^2\right]$$

where the last inequality follows from Lemma 16.(1).

On the right hand side of the bound, the number of points m is sufficiently large so that the penalty term is at most ϵ. So it suffices to show that $\inf_{h\in\mathcal{H}_K} \mathrm{KL}(h^*, h) \leq 2[1+\log\frac{1}{\lambda}]\frac{\epsilon^2+\lambda^2}{\lambda}$. By Lemma 12, there exists $\widehat{f}^{w,\lambda} \in \widehat{\mathcal{F}}^{w,\lambda}$ such that $E_{p_\chi}[(f_j - \widehat{f}^{w,\lambda})^2] \leq 2\epsilon^2 + 2\lambda^2$. Let $\widehat{h}^{w,\lambda} = \sqrt{p(\cdot|\widehat{f}^{w,\lambda})}$ denote the element in $\mathcal{H}_K$ corresponding to $\widehat{f}^{w,\lambda}$. Then

$$\mathrm{KL}(h^*, \widehat{h}^{w,\lambda}) \leq 2\left[1+\log\left\|\frac{h^*}{\widehat{h}^{w,\lambda}}\right\|_\infty\right] d^2(h^*, \widehat{h}^{w,\lambda}) \leq 2\left[1+\log\frac{1}{\lambda}\right] d^2(h^*, \widehat{h}^{w,\lambda})$$

where the first inequality follows from Lemma 11, and the second inequality follows from the definition of h^* and $\widehat{h}^{w,\lambda}$, and the fact that $p(\chi_\mathcal{S}, y|\widehat{f}^{w,\lambda}) \geq \lambda$ for any $\chi_\mathcal{S}$ and y. The proof is completed by noting

$$d^2(h^*, \widehat{h}^{w,\lambda}) = \mathbb{E}_{p_\chi}\left[\left(\sqrt{f_j} - \sqrt{\widehat{f}^{w,\lambda}}\right)^2 + \left(\sqrt{1-f_j} - \sqrt{1-\widehat{f}^{w,\lambda}}\right)^2\right]$$
$$\leq \frac{\mathbb{E}_{p_\chi}[(f_j - \widehat{f}^{w,\lambda})^2]}{4\lambda} + \frac{\mathbb{E}_{p_\chi}[((1-f_j) - (1-\widehat{f}^{w,\lambda}))^2]}{4\lambda} \leq \frac{\epsilon^2+\lambda^2}{\lambda}$$

where the first inequality follows from Lemma 16.(2), and the last inequality follows from the choice of $\widehat{f}^{w,\lambda}$ as in Lemma 12. □

Below are some technical facts that are used in the analysis.

Lemma 16. *(1) If $f_1, f_2 \in [0,1]$, then $4(\sqrt{f_1} - \sqrt{f_2})^2 \geq (f_1 - f_2)^2$.*
(2) If $f_1 \geq \lambda > 0$ and $f_2 \geq \lambda$, then $|\sqrt{f_1} - \sqrt{f_2}| \leq \frac{|f_1 - f_2|}{2\sqrt{\lambda}}$.

Proof. Both claims follow from the fact that $f_1 - f_2 = (\sqrt{f_1} - \sqrt{f_2})(\sqrt{f_1} + \sqrt{f_2})$. □

C.3. Estimation of the entire influence function

We now combine the bounds for individual nodes to get the sample complexity for learning the entire influence function.

Theorem 3. *Let $\epsilon \in (0, 1/4)$ and $\lambda = \frac{\epsilon}{c' d \log\frac{d}{\epsilon}}$ where $c' > 0$ is a sufficiently large constant. If $K = O(\frac{C^2 d^2}{\epsilon^2}\log^2\frac{d}{\epsilon}[\log\frac{Cd}{\delta} + \log\frac{d}{\epsilon}])$,*

$$m = O\left(\frac{C^2 d^3}{\epsilon^3}\log^3\frac{d}{\epsilon}\left[\log\frac{1}{\Lambda} + \log\frac{Cd}{\epsilon} + \log\frac{d}{\delta}\right]\right)$$

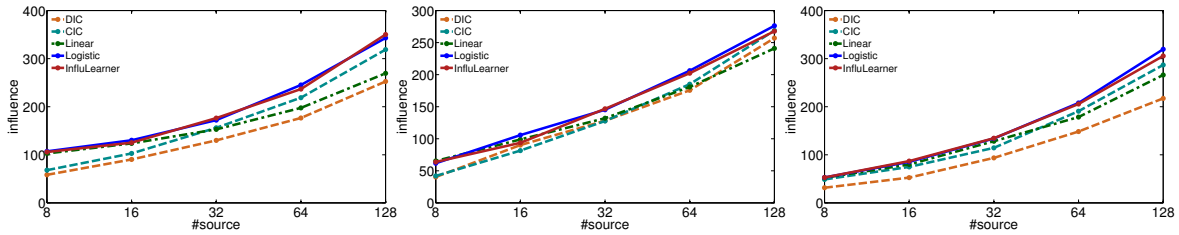
then with probability $1-\delta$ over the drawing of the random features,

$$\mathbb{E}_{\mathcal{D}_m}\left[\mathbb{E}_{p_\chi}\left[\left(\sum_{j=1}^d \widehat{f}_j^{w,\lambda}(\chi_\mathcal{S}) - \sigma(\mathcal{S})\right)^2\right]\right] \leq \epsilon$$

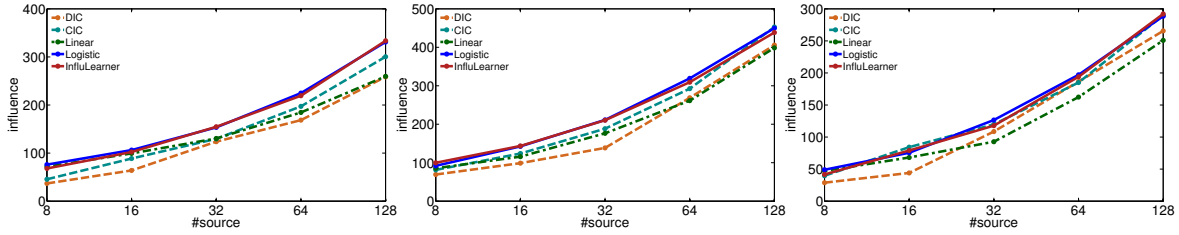
where $\mathbb{E}_{\mathcal{D}_m}$ is with respect to the randomness of $\{(\chi_{\mathcal{S}_i}, \boldsymbol{y}_i)\}_{i=1}^m$. The running time of the algorithm is $O(dmK)$.

Proof. Let $\lambda = \epsilon_0$ and $\epsilon_0 = \frac{\epsilon}{c' d \log\frac{d}{\epsilon}}$ where $c' > 0$ is a sufficiently large constant, so that $8c\left(\epsilon_0 + \frac{\epsilon_0^2+\lambda^2}{\lambda}\left[1+\log\frac{1}{\lambda}\right]\right) \leq \frac{\epsilon}{2d}$ where c is the constant in Lemma 2.

Apply Lemma 1 with error rate ϵ_0 and confidence parameter δ/d. Then when $K = O(\frac{C^2}{\epsilon_0^2}\log\frac{Cd}{\epsilon\delta})$, with probability at least



(a) Group 2 (b) Group 3 (c) Group 4



(a) Group 5 (b) Group 6 (c) Group 7

Figure 5. Expected influence vs. #sources on the real hold-out testing data.

$1-\delta$, for each node $j \in [d]$ there exists $\widehat{f}'_j \in \widehat{\mathcal{F}}^w$ satisfying

$$\mathbb{E}_{p_\chi}\left[(\widehat{f}'_j(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}))^2\right] \leq \epsilon_0^2.$$

Then by Lemma 2, when $m = O\left(\frac{K}{\epsilon_0}\log\frac{K}{\lambda\Lambda}\right)$, for each node $j \in [d]$ we find $\widehat{f}_j^{w,\lambda}$ satisfying

$$\mathbb{E}_{\mathcal{D}_m}\left[\mathbb{E}_{p_\chi}\left[(\widehat{f}_j^{w,\lambda}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}))^2\right]\right] \leq \frac{\epsilon}{2d}.$$

The theorem follows from the fact that $\mathbb{E}_{p_\chi}[(\sum_{j=1}^d \widehat{f}_j^{w,\lambda}(\chi_{\mathcal{S}}) - \sigma(\mathcal{S}))^2] \leq 2\sum_{j=1}^d \mathbb{E}_{p_\chi}[(\widehat{f}_j^{w,\lambda}(\chi_{\mathcal{S}}) - f_j(\chi_{\mathcal{S}}))^2]$.

Runtime. The maximum likelihood estimation only needs to be solved approximately. In particular, it suffices to get $\widehat{h}$ such that

$$\sum_{i=1}^m \log[\widehat{h}(Z_i)] + 1 \geq \sup_{h\in\mathcal{H}_K}\sum_{i=1}^m \log[h(Z_i)].$$

By the convergence rate of EG (see Section 4.4 in (Kivinen & Warmuth, 1997)), we only need $O(1/\eta)$ iterations, where the learning rate η can be viewed as a constant. Each iteration takes time $O(mK)$, and we need to use EG for each of the d nodes. Hence, the total time is $O(dmK)$. □

C.4. Additional experimental results

We report the additional experimental evaluations on the application of the learnt influence functions to the continuous-time influence maximization problem on the rest six groups of hold-out real testing cascades datasets. Compared to DIC and Linear regression, Figure 5 verifies that the performance of INFLULEARNER, Modified Logistic and CIC are better and more consistent with each other.