On the Efficiency of ERM in Feature Learning

Ayoub El Hanchi, Chris J. Maddison, Murat A. Erdogdu

University of Toronto & Vector Institute

Motivation

- Classical ML: learn a linear predictor on top of a feature map.
- Modern ML: jointly learn a feature map and a linear predictor.
- By putting the burden of picking a feature map on the model and data, we should expect that we need more samples to learn.

But just how many more samples?

Setup

- We evaluate the quality of a predictor $f : \mathcal{X} \to \mathbb{R}$ through its risk

$$R(f) := \mathrm{E}[\ell(f(X), Y)], \quad R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Setup

- We evaluate the quality of a predictor $f : \mathcal{X} \to \mathbb{R}$ through its risk

$$R(f) := \mathrm{E}[\ell(f(X), Y)], \quad R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

- We consider classes of predictors induced by arbitrary collections of feature maps $(\phi_t)_{t\in\mathcal{T}}$, $\phi_t : \mathcal{X} \to \mathbb{R}^d$,

$$\mathcal{F} := \bigcup_{t\in\mathcal{T}} \mathcal{F}_t, \qquad \mathcal{F}_t := \left\{x \mapsto \langle w, \phi_t(x)\rangle \mid w \in \mathbb{R}^d\right\}.$$

Setup

- We evaluate the quality of a predictor $f : \mathcal{X} \to \mathbb{R}$ through its risk

$$R(f) := \mathrm{E}[\ell(f(X), Y)], \quad R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

- We consider classes of predictors induced by arbitrary collections of feature maps $(\phi_t)_{t\in\mathcal{T}}$, $\phi_t : \mathcal{X} \to \mathbb{R}^d$,

$$\mathcal{F} := \bigcup_{t\in\mathcal{T}} \mathcal{F}_t, \qquad \mathcal{F}_t := \left\{x \mapsto \langle w, \phi_t(x)\rangle \mid w \in \mathbb{R}^d\right\}.$$

- Goal is to compare the excess risk of the following procedures

ERM procedure	Oracle procedure
$\hat{f}_{n,ERM} \in \operatorname*{argmin}_{f\in\mathcal{F}} R_n(f)$,	$\hat{f}_{n,oracle} \in \operatorname*{argmin}_{f\in\mathcal{F}_{t_*}} R_n(f)$.

$$\mathcal{E}(\hat{f}) := R(\hat{f}) - \min_{f\in\mathcal{F}} R(f), \quad t_* := \operatorname*{argmin}_{t\in\mathcal{T}} \min_{f\in\mathcal{F}_t} R(f).$$

Background

- Conventional wisdom:

  excess risk of ERM $\propto$ size of model class.

- Since
  1. $\hat{f}_{ERM}$ corresponds to ERM on the large class $\mathcal{F}$,
  2. $\hat{f}_{oracle}$ corresponds to ERM on the small class $\mathcal{F}_{t_*}$,

  this suggests that

$$\frac{\mathcal{E}(\hat{f}_{n,ERM})}{\mathcal{E}(\hat{f}_{n,oracle})} \gg 1.$$

Asymptotic result

- Under mild assumptions, we prove that

$$\lim_{n\to\infty} \mathrm{P}\left(1 \leq \frac{\mathcal{E}(\hat{f}_{n,ERM})}{\mathcal{E}(\hat{f}_{n,oracle})} \leq \_\right) = 1.$$

Asymptotic result

- Under mild assumptions, we prove that

$$\lim_{n\to\infty} \mathrm{P}\left(1 \leq \frac{\mathcal{E}(\hat{f}_{n,ERM})}{\mathcal{E}(\hat{f}_{n,oracle})} \leq 2\right) = 1.$$

Asymptotic result

- Under mild assumptions, we prove that

$$\lim_{n\to\infty} \mathrm{P}\left(1 \leq \frac{\mathcal{E}(\hat{f}_{n,ERM})}{\mathcal{E}(\hat{f}_{n,oracle})} \leq 2\right) = 1.$$

- Asymptotically, the difficulty of learning with ERM over the large class of predictors

$$\mathcal{F} := \bigcup_{t\in\mathcal{T}} \mathcal{F}_t,$$

is, up to a factor of two, the same as that of learning with ERM over the linear class of predictors

$$\mathcal{F}_{t_*} := \left\{x \mapsto \langle w, \phi_{t_*}(x)\rangle \mid w \in \mathbb{R}^d\right\}$$

!

Asymptotic result: on the assumption

How mild is the assumption?

- Weak Law of large numbers (WLLN):

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}[X]\right| > \varepsilon\right) = 0.$$

- A collection of random variables $(X_t)_{t\in\mathcal{T}}$ satisfies the UWLLN if

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathrm{P}\left(\sup_{t\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n} X_{t,i} - \mathrm{E}[X_t]\right| > \varepsilon\right) = 0.$$

- The assumption in our result is that certain collections of random variables arising from the problem satisfy the UWLLN. This always holds when $\mathcal{T}$ is finite. In general, this is an assumption on the size of $\mathcal{T}$, appropriately measured.

Nonasymptotic result

What happens non-asymptotically?

Nonasymptotic result

What happens non-asymptotically?

- There is a sequence of subsets $(S_n)_{n=1}^{\infty}$ of $\mathcal{T}$ such that
  1. $S_1 \supset S_2 \supset S_3 \ldots,$
  2. $\bigcap_{n=1}^{\infty} S_n = \{t_*\},$

$$\mathcal{E}(\hat{f}_{n,ERM}) \lesssim (\text{size of } S_n) \cdot \left(\sup_{s \in S_n} \mathcal{E}(\hat{f}_s) \right)$$

where $\hat{f}_s$ is an ERM over the class $\mathcal{F}_s$. Note that as $n \to \infty$, we recover the asymptotic result, up to an absolute constant.

Nonasymptotic result

What happens non-asymptotically?

- There is a sequence of subsets $(S_n)_{n=1}^{\infty}$ of $\mathcal{T}$ such that
  1. $S_1 \supset S_2 \supset S_3 \ldots,$
  2. $\bigcap_{n=1}^{\infty} S_n = \{t_*\},$

  $$\mathcal{E}(\hat{f}_{n,ERM}) \lesssim (\text{size of } S_n) \cdot \left(\sup_{s \in S_n} \mathcal{E}(\hat{f}_s) \right)$$

  where $\hat{f}_s$ is an ERM over the class $\mathcal{F}_s$. Note that as $n \to \infty$, we recover the asymptotic result, up to an absolute constant.
- The subsets S_n correspond to the sublevel sets, for $\varepsilon_n = O(1/n)$,

  $$S_n = \{t \in \mathcal{T} \mid \Delta(t) \leq \varepsilon_n\},$$

  of the function

  $$\Delta(t) := \min_{f \in \mathcal{F}_t} R(f) - \min_{f \in \mathcal{F}_{t_*}} R(f),$$

  that measures the suboptimality of the feature map indexed by t.

Main Takeaways

- Asymptotically and under mild assumptions, learning a feature map in addition to learning a linear predictor with ERM induces a negligible sample complexity overhead.

Main Takeaways

- Asymptotically and under mild assumptions, learning a feature map in addition to learning a linear predictor with ERM induces a negligible sample complexity overhead.
- Non-asymptotically, this overhead is controlled by the size of the set of feature maps that are ε_n as good as the best feature map, for $\varepsilon_n = O(1/n)$.

Main Takeaways

- Asymptotically and under mild assumptions, learning a feature map in addition to learning a linear predictor with ERM induces a negligible sample complexity overhead.
- Non-asymptotically, this overhead is controlled by the size of the set of feature maps that are ε_n as good as the best feature map, for $\varepsilon_n = O(1/n)$.
- Future directions: can we verify that the assumptions hold for model classes and distributions of practical interest?

Main Takeaways

- Asymptotically and under mild assumptions, learning a feature map in addition to learning a linear predictor with ERM induces a negligible sample complexity overhead.
- Non-asymptotically, this overhead is controlled by the size of the set of feature maps that are ε_n as good as the best feature map, for $\varepsilon_n = O(1/n)$.
- Future directions: can we verify that the assumptions hold for model classes and distributions of practical interest?

Thank you for your attention!