# FROM POLARIZATION OF BELIEF TO ACTIVE LEARNING THEORY: A DIAMETER APPROACH

**Gauthier Guinet**
MIT
gguinet@mit.edu

December 9, 2020

**ABSTRACT**

In [Haghtalab et al., 2019], polarization of belief is studied through the lens of statistical learning theory. Aside from the innovative ideas, the main theoretical contribution is the introduction of diameter inequalities on an hypothesis class, leveraging only the structure induced by the pseudo metric related to the 0-1 loss. Such diameter is mapped to the maximal disagreement between agents and thus the potential polarization. More precisely, they establish some PAC style bounds on the maximal distance between two penalized ERM hypothesis and study the impact of small modification of the distribution on this distance.

With this in mind, this work leverages their framework to further study diameter inequalities under the existence of penalization, without making any assumptions on the structure of the hypothesis space nor on the form of such penalization. Particular attention is given to asymptotic diameter and convergence of empirical and expected approximation sets, called Rashomon Sets. Roughly speaking, we wonder to what extent polarization is robust w.r.t. the penalization? In others words, we analyse the impact of modifications of the penalization associated with hypothesis (i.e.education) on polarization.

The second part of the work lays the groundwork of an algorithm whose goal is to introduce bias in the initial distribution in order to reduce maximal diameter, studying an open question of [Haghtalab et al., 2019]. In particular, some links are established with a line of work in Active Learning community tackling related questions [Tosh and Dasgupta, 2017, Hanneke, 2011].

## 1 A Statistical Learning approach to Polarization

The context of this work a the study on the polarization of belief phenomena through the lens of Statistical and Active Learning Theories. Classical modeling of polarization mainly studies the differences induced by the exposure to different sources of information. Let's cite for instance the literature on Social Learning (in Bayesian and non-Bayesian framework) or on Stochastic models of opinion dynamics, widely used in Statistical Physics or Complex Systems for instance. Following the ideas of [Haghtalab et al., 2019], we shall explore a more pernicious form of polarization that may appears even when people are exposed to similar type of information. Statistical Learning tools provide a powerful modeling to such type of phenomena.

### 1.1 First definitions and notations

Throughout this work, we shall consider a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a given input space and $\mathcal{Y} = \{-1, +1\}$. We also consider the total variation distance $\mathcal{TV}$ between two distributions $\mathcal{D}$ and $\tilde{\mathcal{D}}$ defined on their marginal distribution on $\mathcal{X}$ ($\mathcal{D} \downarrow \mathcal{X}$). This restriction is in a sense justified as all the distributions we shall consider have the same conditional label distribution.

Next, for a given hypothesis class $\mathcal{F}$, we introduce the error defined on $\mathcal{D}$ by the 0-1 loss:

$$\mathrm{err}_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(f(x) \neq y)] = \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$$

We make the assumption that $\mathcal{D}$ is *realizable* i.e. that there exists $f^\star \in \mathcal{F}$ such that $\mathrm{err}_{\mathcal{D}}(f) = 0$. We also introduce the empirical error associated with a training set $\mathcal{S} = \{(x_i, y_i)\}_{i\in[m]}$:

$$\mathrm{err}_{\mathcal{S}}(f) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(f\left(x_i\right) \neq y_i\right)$$

Finally, a key notion we shall consider throughout this work is the disagreement between two hypothesis $f, \tilde{f} \in \mathcal{F}$, defined by:

$$\Delta_{\mathcal{D}}\left(f, f'\right) := \mathbb{P}_{x\sim\mathcal{D}\downarrow\mathcal{X}}\left[f(x) \neq f'(x)\right]$$

This defines a pseudo-metric on $\mathcal{F}$, which induce a natural although limited structure on this set. Thanks to that, we can define the diameter of any given hypothesis set $\mathcal{H}$ as:

$$\mathrm{diam}_{\mathcal{D}}(\mathcal{H}) := \sup_{f,f'\in\mathcal{H}} \Delta_{\mathcal{D}}\left(f, f'\right)$$

More precisely, this corresponds to the largest disagreement between two hypotheses in this class and bear great sense in our polarization approach.

### 1.2 The objective cost model

As we know, classical results on Learning Theory shows us that thanks the *realizable* assumption and under mild conditions, the learning process of Empirical Risk Minimizer Agents will converge towards an hypothesis in agreement with $f^\star$ if given sufficient data points. PAC theory also gives us some results on the speed of this convergence in function of some key parameters. In this case, we shall say that there is no polarization of belief as two learning agents seeing different realization of the distribution $\mathcal{D}$ shall reach with sufficient time hypothesis almost in agreement.

One of the great interest of [Haghtalab et al., 2019] is the introduction of a notion of complexity of learning: indeed, agents might not be totally rationals and have difficulties learning very *complex* hypothesis. Thus, they rather seek a balance between accuracy and complexity. Under this very simple assumption, one can then show that two agents receiving different samples from the same distribution $\mathcal{D}$ might reach a large disagreement.

More precisely, we suppose the knowledge of a non-negative function $\phi$ such that for a given $f \in \mathcal{F}$, $\phi(f)$ traduces the complexity of using hypothesis $f$. $\phi$ is very related to the concept of penalization or regularization, but the goal here is slightly different. For instance, one could see $\phi$ as the number of features in a boolean function, or the depth of a decision list. Our goal is to remain as general as possible, as such generality allows to tackle more complex cases, $\mathcal{F}$ could be a meta-hypothesis space (neural networks and linear functions for instance). One could also consider $\phi$ as a prior of learning agents for certain hypothesis or model complex bounded rationality phenomena. Finally, as we shall introduce no other structure on $\mathcal{F}$ than the one induce by the pseudo-metric associated with the 0-1 loss, introduction of classical properties of $\phi$ is not possible. On what follows, we shall refer to $\phi$ as a complexity function to state clearly the difference. This leads us to introduce for a given distribution $\mathcal{D}$ and a training set $\mathcal{S}$ the cost of an hypothesis as:

$$\mathrm{cost}_{\mathcal{D}}(f) := \mathrm{err}_{\mathcal{D}}(f) + \phi(f) \text{ and } \mathrm{cost}_{S}(f) := \mathrm{err}_{S}(f) + \phi(f)$$

More generally, for $\lambda \geq 0$, we shall generalize this to:

$$\mathrm{cost}^{\lambda}_{\mathcal{D}}(f) := \mathrm{err}_{\mathcal{D}}(f) + \lambda\phi(f) \text{ and } \mathrm{cost}^{\lambda}_{\mathcal{S}}(f) := \mathrm{err}_{\mathcal{S}}(f) + \lambda\phi(f)$$

The following simple example from [Haghtalab et al., 2019] showcase the polarization phenomena:

**Example 1.1** (A simple polarization case)**.** *Let's consider* $\mathcal{F} := \left\{f_{\mathbf{w}}(\mathbf{x}) = \mathrm{sign}(<\mathbf{w}, \mathbf{x}>) \mid \mathbf{w} \in \mathbb{R}^d\right\}$ *associated with* $\mathcal{X} := \mathbb{R}^d$*. Let* $\mathcal{D}$ *be the distribution with weight* $\frac{1}{2d}$ *on each* $(e_i, +1)$ *and* $(-e_i, -1)$ *for all* $i \in [d]$ *where* $e_i$ *is the* $i^{th}$ *unit vector. Finally, let's define* $\phi(\mathbf{w}) = \frac{1}{2d}\|\mathbf{w}\|_0$*. We can show that for any* $m$ *and two sets* $\mathcal{S}_1, \mathcal{S}_2$ *of* $m$ *i.i.d. samples from* $\mathcal{D}$*, with probability* $\frac{1}{4}$*, there exists* $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ *such that* $\Delta_{\mathcal{D}}\left(f_1, f_2\right) > \frac{1}{6}$*.*

To tackle to this polarization problem, the authors suggest the following theoretical solution:

**Theorem 1.2** (Diameter reduction through Bias introduction)**.** *For a hypothesis class* $\mathcal{F}$ *(with finite VC dimension), a realizable distribution* $\mathcal{D}$ *on* $\mathcal{X} \times \mathcal{Y}$*, a parameter* $\alpha \in [0, 1]$ *and a maximum level of disagreement* $\gamma > 0$*, there exists*

$$m \in O\left(\gamma^{-4}\alpha^{-2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

*and realizable distribution* $\tilde{\mathcal{D}}$*, with* $\mathcal{TV}(\mathcal{D}, \tilde{\mathcal{D}}) \leq \frac{\alpha}{2}$*, such that if two sets* $S_1$ *and* $S_2$ *of size at least* $m$ *are sampled from* $\widetilde{\mathcal{D}}$*, then with probability at least* $1 - \delta$ *any two cost-minimizing hypotheses* $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ *for* $i \in \{1, 2\}$

1. *have at most $\gamma$ disagreement over $\mathcal{D}$, i.e., $\Delta_{\mathcal{D}}\left(\widetilde{f}_1, \widetilde{f}_2\right) \leq \gamma$, and*
2. *have a cost that is optimal up to $3\alpha$ on $\mathcal{D}$, i.e.*

$$\mathrm{cost}_{\mathcal{D}}\left(\widetilde{f}_i\right) \leq \underset{f\in\mathcal{F}}{\mathrm{argmin}}\, \mathrm{cost}_{\mathcal{D}}(f) + 3\alpha$$

Less formally, one can biased the distribution $\mathcal{D}$ towards a close and still realizable distribution $\tilde{\mathcal{D}}$ such that the hypothesis learned with data from $\tilde{\mathcal{D}}$ will be considerably closer w.r.t $\Delta_{\mathcal{D}}$ while staying close to optimal.

### 1.3 Limitations and motivations

The starting points of our work are the 3 following observations:

- Although relevant theoretically, the way this bias is constructed is of no use in practice as it involves the construction of a complex set, the Rashomon set that we shall introduce bellow. An open question is then if there exists a tractable or alternative way to reduce this potential disagreement.
- Then, [Haghtalab et al., 2019] also present a lower bound on the number of samples needed to reduce polarization through this bias technique. Yet, this a considerable gap between the upper bound on our theoreme and the lower bound provided. This might be due to the use of uniform convergence of hypothesis on this set, where weaker notions seems to be sufficient. This calls for a better understanding of the dynamics of learning and convergence over this set.
- Finally, through example 1.1., we witnessed the existence of a necessary polarization. As seen above, a first way to tackle it is to shift the distribution, which is a very general idea. We advocate that another potential interesting solution is *education* of agents i.e. working on the complexity function rather than on the distribution. Indeed, an interesting insight is that the polarization of this example vanish if we induce small modifications of the complexity function.

Thus, in this work, we shall explore the two following avenues:

- **Robustness, Complexity and Learning:** To what extent polarization is robust w.r.t. the complexity functions ? In others words, what is the impact of modifications of the complexity function associated with hypothesis (i.e. *education*) on polarization ? Would this polarization disappear for a complexity tending to 0 ? Is there an opportunity of doing global or local complexity modifications in order to reduce this ? This line is also of independent interest as it involves a deeper study of the disagreement phenomena under the existence of complexity.
- **Active Learning and Polarization:** This idea of biasing or shifting the distribution $\mathcal{D}$ towards a distribution $\tilde{\mathcal{D}}$ is widely used in the field of **Active Learning Theory**. Although not exactly stated this way, we believe that some interesting links exists. Indeed, one branch of Active Learning studies specifically how to query a minimal number of points in order to reduce the diameter of a subset of the hypothesis space, the version space (all hypothesis consistent with current sample $\mathcal{S}$) [Hanneke, 2011, Tosh and Dasgupta, 2017]. Moreover, as stated above, one of our goal is to get a better understanding of the convergence dynamics w.r.t. the number of needed data points to lower disagreement, in order to lower such number. Such goal is at the core of Active Learning, which brings both theoretical and practical tools we would like to leverage.

## 2 Robustness, Complexity and Learning

For a given distribution $\mathcal{D}$, and $\epsilon, \lambda > 0$, we define the Rashomon set [Fisher et al., 2019, Semenova et al., 2020]:

$$\mathcal{F}_{\epsilon}^{\mathcal{D}}(\lambda) := \left\{ f \in \mathcal{F} \mid \mathrm{cost}_{\mathcal{D}}^{\lambda}(f) \leq \min_{f'\in\mathcal{F}} \mathrm{cost}_{\mathcal{D}}^{\lambda}\left(f'\right) + \epsilon \right\}$$

We remark that for a sequence $\mathcal{S}$ of points, we can define the empirical set associated with the discrete distribution allocating uniform mass on the drawn points. We also define the ball centered in $f^\star$ of diameter $\epsilon$:

$$\mathcal{B}(f^\star, \epsilon) := \left\{ f \in \mathcal{F} \mid \Delta_{\mathcal{D}}(f^\star, f) = \mathrm{err}_{\mathcal{D}}(f) - \underbrace{\mathrm{err}\,\mathcal{D}(f^\star)}_{=0} \leq \epsilon \right\} = \mathcal{F}_{\epsilon}^{\mathcal{D}}(0)$$

As $\mathrm{err}_{\mathcal{D}}(f^\star) = 0$, we see that the definition of this ball is independent of the choice of the optimal hypothesis $f^\star$. Similarly, for a sequence $\mathcal{S}$, we can define an empirical ball centered in $f^\star$ of diameter $\epsilon$, associated to the pseudo metric $\Delta_{\mathcal{S}}$. Interested by a measure of distance between those two set, we next introduce the Hausdorff associated with the pseudo metric $\Delta$:

$$d_{\mathcal{H}}(\epsilon, \lambda) = d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) := \max( \sup_{f \in \mathcal{B}(f^\star, \epsilon)} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda), \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda))$$

We also define $d_{\mathcal{H}}^{\mathcal{S}}(\epsilon, \lambda) = d(\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda), \mathcal{F}_\epsilon^{\mathcal{S}}(0))$ the distance between the sets defined by an empirical distribution $\mathcal{S}$. All the proof are differed to the appendix.

The following lemma justify our interest in this metric:

**Lemma 2.1** (Hausdorff Diameter inequality)**.** *For two sets $A$ and $B$ such that $\mathrm{diam}(A), \mathrm{diam}(B) < \infty$, $d_{\mathcal{H}}(A, B)$ is finite and we have:*

$$|\mathrm{diam}(B) - \mathrm{diam}(A)| \leq 2 d_{\mathcal{H}}(A, B)$$

In particular, we have:

$$\forall \lambda, \epsilon > 0, \quad \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) - 2 d_{\mathcal{H}}(\epsilon, \lambda) \leq \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) + 2 d_{\mathcal{H}}(\epsilon, \lambda)$$

The diameter is 2-lipschitz w.r.t. the Hausdorff distance and thus we shift the problem of studying the convergence of the diameter to the convergence of the distance. Indeed, this distance involves a uniform notion of convergence, which is very suited for our needs. As stated above, we are interested in the limit of those two distance when $\lambda \to 0^+$ and when $\epsilon \to 0^+$. Yet, so far, we haven't defined a structure on $\mathcal{F}$ aside from the one induced by the pseudo-metric $\Delta$, which is very limited. Thus, one has to be cautious when considering limits. As seen in the lecture, similar results exists when the error and the complexity of the hypothesis have interesting forms (ridge or lasso regression under least square loss for instance). Yet, here we interested in general type of complexity functions associated with the hypothesis whereas previous results involve a very specific form. Moreover, the evolution (in term of inclusion) of the set $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ in function of $\lambda$ is not obvious, as we shall see bellow. Figure 1 summarize the different relations we are interested in.

$$\mathrm{diam}_{\mathcal{S}}\, \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda) \xrightarrow[\lambda \to 0^+]{??} \mathrm{diam}_{\mathcal{S}}\, \mathcal{B}^{\mathcal{S}}(f^\star, \epsilon) \xrightarrow[\epsilon \to 0^+]{??} \mathrm{diam}_{\mathcal{S}}\{f \in \mathcal{F} \mid \mathrm{err}_{\mathcal{S}}(f) = 0\} = 0$$

$$\mathrm{diam}_{\mathcal{D}}\, \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) \xrightarrow[\lambda \to 0^+]{??} \mathrm{diam}_{\mathcal{D}}\, \mathcal{B}(f^\star, \epsilon) \xrightarrow[\epsilon \to 0^+]{??} \mathrm{diam}_{\mathcal{D}}\{f \in \mathcal{F} \mid \mathrm{err}_{\mathcal{D}}(f) = 0\} = 0$$

$$\mathrm{diam}_{\mathcal{D}}\, \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda) \xrightarrow[\lambda \to 0^+]{??} \mathrm{diam}_{\mathcal{D}}\, \mathcal{B}^{\mathcal{S}}(f^\star, \epsilon) \xrightarrow[\epsilon \to 0^+]{??} \mathrm{diam}_{\mathcal{D}}\{f \in \mathcal{F} \mid \mathrm{err}_{\mathcal{S}}(f) = 0\}$$

Figure 1: Global goals and motivations

To do so, we need to study the two minimax quantities introduced in the distance, which is the goal of the following sections.

### 2.1 Distance of $\mathcal{F}_\epsilon^{\mathcal{D}}$ to $\mathcal{B}(f^\star, \epsilon)$

The following plot describes the evolution of $\mathcal{F}_\epsilon^{\mathcal{D}}$ in function of $\lambda$:

First, we note that $\mathcal{F}_\epsilon^{\mathcal{D}}$ is clearly decreasing in function of $\epsilon$ and that $\lambda \mapsto \min \mathrm{cost}_{\mathcal{D}}^{\lambda}(f' t$ is concave (as a pointwise inf). Yet, the evolution of the set in function of $\lambda$ is rather complex. At a given "level of complexity" $\lambda$, the behavior of the hypothesis in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ is one of the following type.

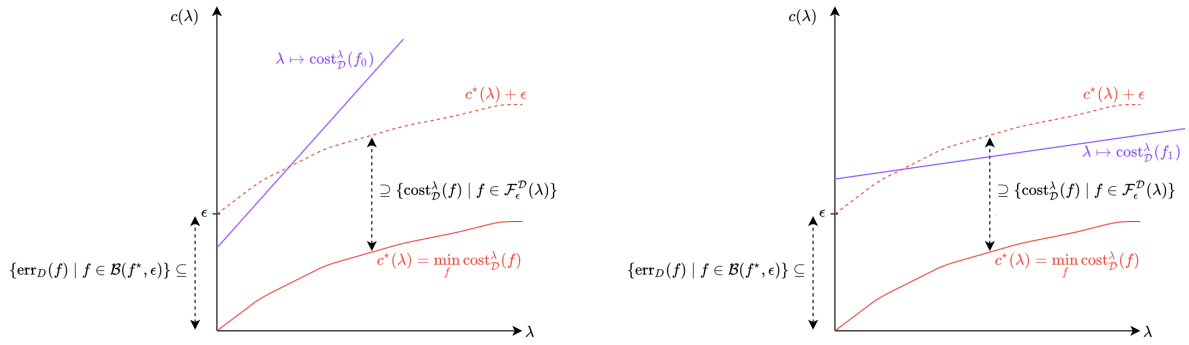


Figure 2: Evolution of Rashomon Set Values in function of $\lambda$

**Lemma 2.2** ($\lambda$-evolution of $\mathcal{F}_\epsilon^{\mathcal{D}}$)**.** *For $f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$, we have the two following behaviors:*

- *If $\exists \lambda_0 \leq \lambda$ with $\lambda_0 > 0$, $f_\lambda \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$, then $\forall \lambda_1 \leq \lambda_0$, $f_\lambda \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_1)$. Moreover, $f_\lambda \notin \mathcal{B}(f^\star, \epsilon)$.*
- *Otherwise, $\forall \lambda_0 \leq \lambda$ such that $\lambda_0 > 0$, $f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$, and we have $f_\lambda \in \mathcal{B}(f^\star, \epsilon)$.*

Yet, the difficulty is double to tackle the uniform continuity:

- First, the set $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ might continue to grow when $\lambda \to 0^+$. Hypothesis $f_0$ in the figure 2 and Example 2.3 show such a phenomena. Thus, considering it at a given time step $\lambda$ doesn't take into account the fact that it can still increase afterwards.
- Secondly, we need a uniform parameter $\lambda_0$ associated with the removal of an hypothesis of the class and not a per hypothesis version. Lemma 2.5 gives the existence of such $\lambda_0$ under certain conditions

**Example 2.3** (A complex evolution of $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$)**.** *With $\phi = \|.\|_1$, and the distribution $\mathcal{D}_1$ described in previous example, we have for $\epsilon, \lambda > 0$:*

$$\mathcal{B}(f^\star, \epsilon) = \left\{ \mathbf{w} \in \left(\mathbb{R}_+^\star\right)^d \right\}$$
$$\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) = \left\{ \mathbf{w} \in \left(\mathbb{R}_+^\star\right)^d \mid \|\mathbf{w}\|_1 \leq \frac{\epsilon}{\lambda} \right\}$$

Hopefully, we can have the following result:

**Theorem 2.4** (Convergence equality for finite image hypothesis set)**.** *If $\{\text{err}_{\mathcal{D}}(f) \mid f \in \mathcal{F}\}$ is finite, then there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) = 0$$

Regarding upper bounds, we can show that:

**Lemma 2.5** (An $\epsilon$-upper bound)**.** *For a given distribution $\mathcal{D}$, we have:*

$$\forall \lambda > 0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) \leq \min(1, \epsilon + c^\star(\lambda) - \Gamma) \leq \min(1, \epsilon + c^\star(\lambda))$$

*where $\Gamma \geq 0$ is the minimal gradient of an affine function tangent to $\lambda_0 \mapsto c^\star(\lambda_0)$ and going through $c^\star(\lambda) + \epsilon$. Moreover, there exists for all $\lambda > 0$, a distribution $\mathcal{D}$, an $\epsilon > 0$ and an hypothesis space $\mathcal{F}$ where the equality is reached (for a fixed lambda !).*

Note that as $c^\star$ tends to 0, we can pick an upper bound an $\lambda_0$ such that this quantity is bellow every $\tilde{\epsilon}$.

Finally, the following example shows that there exists scenarios where this distance is lower bounded for all $\lambda > 0$, which make the uniform convergence impossible (hence the need of stronger hypothesis).

**Example 2.6** (No uniform convergence in general case)**.** *Let's $\mathcal{F} = \{f^\star\} \cup \{(f^k)_{k \in \mathbb{N}}\}$, where for a given $k$, $\phi(f^k) = 0$ and $\text{err}_{\mathcal{D}}(f^k) = \epsilon + \frac{1}{k}$. Finally, let's assume that $\phi(f^\star) = 0$ and $\text{err}_{\mathcal{D}}(f^\star) = 0$. We have $\mathcal{B}(f^\star, \epsilon) = \{f^\star\}$, For a*

*given $\lambda > 0$ (small enough to ensure that $f^\star$ is the minimal cost hypothesis), $f^{k_\lambda} \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$, where $k_\lambda = 1 + \lfloor 1/\lambda \rfloor$. This involves that $\inf_{f \in \mathcal{B}(f^\star,\epsilon)} \Delta(f, f^{k_\lambda}) = \Delta(f^\star, f^{k_\lambda}) = \epsilon + \frac{1}{k}$. Thus, we have:*

$$\forall \lambda > 0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star,\epsilon)} \Delta(f, f_\lambda) > \epsilon \tag{1}$$

### 2.2 Distance of $\mathcal{B}(f^\star, \epsilon)$ to $\mathcal{F}_\epsilon^{\mathcal{D}}$

Regarding the set $\mathcal{B}(f^\star, \epsilon)$, example 2.7 presents a lower bound in the general case scenario.

**Example 2.7** (A lower bound in general case)**.** *Let's take $\mathcal{F} = \{f_0, f^\star\}$ and $\phi$ such that $\phi(f) = 1$ if $f = f_0$ and 0 otherwise. Finally, let's assume that $\mathrm{err}_{\mathcal{D}}(f^\star) = 0$ and $\mathrm{err}_{\mathcal{D}}(f_0) = \epsilon$. Thus, in particular, we have $\Delta(f^\star, f_0) = \epsilon$. Then, we have for $\lambda > 0$:*

$$\mathcal{B}(f^\star, \epsilon) = \{f_0, f^\star\}$$
$$\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) = \{f^\star\}$$

*We can deduce from this that:*

$$\forall \lambda > 0, \quad \sup_{f \in \mathcal{B}(f^\star,\epsilon)} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) = \epsilon$$

Yet, with this in mind, we can deduce the two following properties, by making distinction between its interior $\mathcal{B}(f^\star, \epsilon)^\circ$and frontier $\partial\mathcal{B}(f^\star, \epsilon)$:

The following lemma generalizes the idea of example 2.7:

**Lemma 2.8** (No convergence on the frontier)**.** *If $f \in \partial\mathcal{B}(f^\star, \epsilon)$ and $\phi(f^\star) > \phi(f)$, then $\forall \lambda > 0, f \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$.*

Then, we show a pointwise result for $f \in \mathcal{B}(f^\star, \epsilon)^\circ$.

**Lemma 2.9** (Pointwise interior convergence)**.** *If $f \in \mathcal{B}(f^\star, \epsilon)^\circ$, then there exists $\lambda_0$ such that $\forall \lambda \leq \lambda_0$, $f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$.*

Finally, an simpler upper bound can be provided for $\lambda$ small enough.

**Lemma 2.10** (An $\epsilon$-type upper bound)**.** *There exists $\lambda_0 > 0$ such that*

$$\forall \lambda \leq \lambda_0, \quad \sup_{f_0 \in \mathcal{B}(f^\star,\epsilon)} \inf_{f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f_0, f) \leq \epsilon$$

During the proof, we note that if $\phi(f^\star) > 0$, we have $\lambda_0 \geq \frac{\epsilon}{\phi(f^\star)}$. This still true in a sense if $\phi(f^\star) = 0$ as any $\lambda$ works.

Under some assumptions, we can show that the distance is equal to 0 for $\lambda$ small enough.

**Lemma 2.11** (First result on convergence)**.** *If there exists $\eta > 0$ such that $\forall f \in \mathcal{B}(f^\star, \epsilon)^\circ, \mathrm{err}_{\mathcal{D}}(f) \leq (1 - \eta)\epsilon$ and $\phi$ is bounded on $\mathcal{B}(f^\star, \epsilon)^\circ$, then there exists $\lambda_0$ such that $\forall \lambda \leq \lambda_0, \mathcal{B}(f^\star, \epsilon)^\circ \subset \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. In particular, we can deduce that:*

$$\forall \lambda \leq \lambda_0, \quad \sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) = 0$$

We do not need the hypothesis stating that $\phi$ must be bounded on $\mathcal{B}(f^\star, \epsilon)^\circ$ if we make stronger assumptions on the distribution.

**Lemma 2.12** (Second result on convergence)**.** *If $\mathcal{F}$ has a finite number of patterns on $\mathcal{D}$, then, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad \sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) = 0$$

In particular, for distribution with finite support, the hypothesis of this theorem are verified.

Finally, if we make stronger assumptions on the set $\mathcal{F}$ i.e. that it has finite VC dimension, we can show the following, which we consider as one of our main theorem.

**Theorem 2.13** (Main result on convergence)**.** *If $\mathcal{F}$ has finite VC dimension, then:*

$$\sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) \underset{\lambda \to 0^+}{\to} 0$$

This is strong result, at it involves a uniform notion of convergence (vs a pointwise), which is not granted at all. The key object allowing this is the existence of a finite $\eta$-net covering of $\mathcal{F}$ for all $\eta > 0$ which is possible as $\mathcal{F}$ has finite VC dimension.

### 2.3 General convergence of diameter

The goal of this section is to summarize the results from the two previous subsections on $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ and $\mathcal{B}(f^\star, \epsilon)$ in order to reach global results on the diameter of these sets. In what follow, we shall consider a given distribution $\mathcal{D}$ and hypothesis space $\mathcal{F}$, with an associated complexity function $\phi$. In a first part, we shall tackle approximation result and then switch to a more learning theory approach.

**Theorem 2.14** (An $\epsilon$-approximation)**.** *For all $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) \leq 2\epsilon$$

*Thus, we have in particular:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}(\mathcal{B}(f^\star, \epsilon)) - 2\epsilon \leq \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq \operatorname{diam}(\mathcal{B}(f^\star, \epsilon)) + 2\epsilon$$

From this, we can deduce that:

**Corollary 2.14.1** ($\epsilon$-decrease)**.** *For any $\mathcal{D}$, there exists $\lambda_0$ such that:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 6\epsilon$$

This shows that we do have a hope of reducing the disagreement by lowering the complexity of the agents. Yet, we have seen in our proofs that we could only show the upper bound $\lambda_0 \geq \frac{\epsilon}{\phi(f^\star)}$. Thus, as we progressively decrease $\epsilon$, this involves also working more on the complexity (although at a linear rate).

For the following result, let's assume that we are in a Tsybakov's low-noise scenario, a condition widely studied in active and passive learning literature. More precisely, following the modeling of [Hanneke, 2011], let's assume that there exist finite constants $\mu > 0$ and $\kappa \geq 1$ such that:

$$\forall \varepsilon > 0, \operatorname{diam}_{\mathcal{D}}(\mathcal{B}(f^\star, \epsilon)) \leq \mu \varepsilon^{1/\kappa}$$

For instance, this condition is satisfied when there exists $\mu' > 0$ and $\kappa \geq 1$ such that:

$$\exists h \in \mathbb{F} : \forall h' \in \mathbb{C} \quad \operatorname{err}_{\mathcal{D}}(h') \geq \mu' \mathbb{P}\{h(X) \neq h'(X)\}^\kappa$$

Moreover, as in our case, we assumed the Bayes optimal classifier is in $\mathcal{F}$ (realizable assumption), this is true if there exists $\mu'' >$ and $\alpha > 0$ such that:

$$\forall \varepsilon > 0, \quad \mathbb{P}\{|\eta(X) - 1/2| \leq \varepsilon\} \leq \mu'' \varepsilon^\alpha$$

where $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$. Under those assumptions, we have:

**Theorem 2.15** (Stronger $\epsilon$-approximation under low-noise scenario)**.** *Under the Tsybakov's low-noise assumption, there exists $\lambda_0$ such that, for all $\epsilon > 0$ and $\tilde{\epsilon}$ arbitrarily close to 0:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 2\epsilon + \mu \epsilon^{1/\kappa} + \tilde{\epsilon}$$

By making an extra assumption, we can get considerably stronger results, as the theorem bellow shows:

**Theorem 2.16** (Equality of diameter for finite patterns)**.** *If there is a finite number of patterns of $\mathcal{F}$ on $\mathcal{D}$, for all except a finite number of $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) = 0$$

*From this, we can deduce in particular that:*

$$\forall \lambda \leq \lambda_0, \quad \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) = \operatorname{diam}(\mathcal{B}(f^\star, \epsilon))$$

This is true in particular is $\mathcal{F}$ is finite (but most results are obvious in this case). Please note that as we are using a pseudo-metric, we can't state that those two sets are equals . In particular, we can deduce the corollary, using the pseudo-metric associated with the distribution of a sample $\mathcal{S}$:

**Corollary 2.16.1** (Diameter under empirical measure)**.** *For any sample $\mathcal{S}$ drawn form $\mathcal{D}$, for all except of finite number of $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad \operatorname{diam}_{\mathcal{S}}(\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)) = \operatorname{diam}_{\mathcal{S}}(\mathcal{B}^{\mathcal{S}}(f^\star, \epsilon))$$

In other words, this means that we can lower the complexity such that the empirical Rashomon set and the empirical hypothesis ball have same diameter **under the pseudo-metric induced by** $\mathcal{S}$.

Then, using classical learning theory results, we reach the following result:

**Lemma 2.17** (PAC approximation with blowup)**.** *If $\mathcal{F}$ has finite VC dimension, for $\epsilon, \delta > 0$, there exists $N(\epsilon, \delta)$ such that for all sample $\mathcal{S}$ of size at least $N(\epsilon, \delta)$, we have with probability as least $1 - \delta$:*

$$\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) \subset \mathcal{F}_{2\epsilon}^{\mathcal{S}}(\lambda) \quad \textit{and} \quad \mathcal{B}(f^\star, \epsilon) \subset \mathcal{B}^{\mathcal{S}}(f^\star, 2\epsilon)$$

*Moreover, we also have:*

$$\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda) \subset \mathcal{F}_{2\epsilon}^{\mathcal{D}}(\lambda) \quad \textit{and} \quad \mathcal{B}^{\mathcal{S}}(f^\star, \epsilon) \subset \mathcal{B}(f^\star, 2\epsilon)$$

One key remark is the fact that a "blowup" of the sets w.r.t. $\epsilon$ was needed to ensure to have all hypothesis included. The choice of $2\epsilon$ to ensure this blow-up is pretty arbitrary. Another key remark is that we used rather strong tools of uniform convergence of hypothesis when we are in really only interested by properties of hypothesis on the Rashomon sets, which can be weaker. This will mainly have an impact on the size of $N(\epsilon, \delta)$. One should also remark that $N$ doesn't depend on $\lambda$. This allows to show the following result:

**Theorem 2.18** (Diameter of empirical Rashomon set)**.** *For any $\epsilon, \delta > 0$, there exists $N(\epsilon, \delta)$ and $\lambda_0(\epsilon)$ such that for all set $\mathcal{S}$ for size at least $N(\epsilon, \delta)$, with probability $1 - \delta$:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}_{\mathcal{D}}(\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)) \leq \operatorname{diam}_{\mathcal{D}}(\mathcal{F}_{2\epsilon}^{\mathcal{D}}(\lambda)) \leq 12\epsilon$$

This is of particular interest as this set contain all $\epsilon$-optimal agents w.r.t. the empirical distribution. Thus, it tells us about the real effective diameter we are interest in and on the possibility of "educating" agents in order to lower disagreement. Note that here the parameter $\lambda_0$ doesn't depend on $\mathcal{S}$. As a summary, Figure 3 shows the different results we were able to show and conclude this section. To establish the missing junction between $\operatorname{diam}_{\mathcal{S}}$ and $\operatorname{diam}_{\mathcal{D}}$, similarly one would have to leverage uniform results of convergence between empirical and expected distribution. The notion of Glivenko–Cantelli class could be of use for this (w.r.t. the sets $\{f(x) \neq \tilde{f}(x) | x \in \mathcal{D}\}_{f, \tilde{f} \in \mathcal{F}}$) and the challenge is to combine the approximation of $\operatorname{diam}_{\mathcal{S}}$ and $\operatorname{diam}_{\mathcal{D}}$ and of $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ and $\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)$ is order to obtain powerful results.

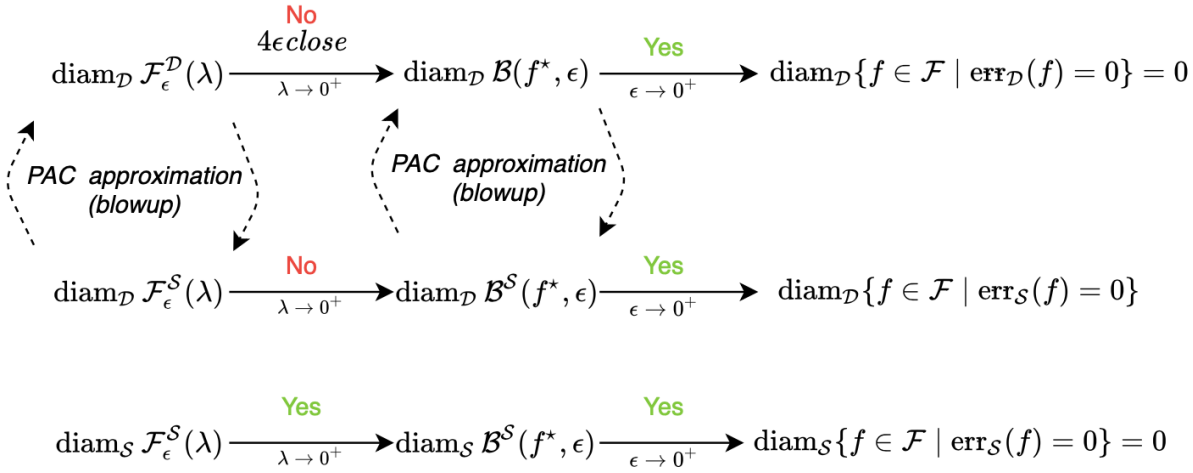


Figure 3: A global summary of results

## 3 Active Learning and Polarization

### 3.1 Towards a dynamic reduction of disagreement

As stated above, a line of work in Active Learning theory is focusing on a diameter based approach of learning, which we find we relevant in regard to the results stated above [Tosh and Dasgupta, 2017, Hanneke, 2007, Hanneke, 2011]. Indeed, by adding a bias in the distribution towards certain subspace of the hypothesis space, you allow faster rate

of convergence (w.r.t. classical supervised learning). More precisely, we would like to study in our case whether an empirical construction of the distribution $\tilde{\mathcal{D}}$ or an approximation of it is possible in order to lower disagreement. Thus, the goal of this section is to discuss the potential implementation of an learning algorithm to introduce this bias.

Looking a bit more precisely in the proof used in [Haghtalab et al., 2019], the key argument is to introduce a bias toward an hypothesis $f$ in $\mathcal{F}_\epsilon^{\mathcal{D}}$ (we suppose the $\lambda$ fixed here). With a distance $\alpha$ allowed between $\mathcal{D}$ and $\tilde{\mathcal{D}}$, this allowed to get guarantees of the form:

$$\operatorname{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\tilde{\mathcal{D}}}\right) \in O\left(\frac{\epsilon}{\alpha \operatorname{err}_{\mathcal{D}}(f)}\right)$$

This leads usually to pick the maximal error hypothesis to get the best bounds. Yet, the introduction of this bias involves the knowledge of $\mathcal{F}_\epsilon^{\mathcal{D}}$, $\operatorname{err}_{\mathcal{D}}$ and $f^\star$, in order to construct $\tilde{\mathcal{D}}$. Moreover, this construction allows to remain within a range $3\alpha$ of optimal solution (the interpretation of this last result is less obvious with the polarization spirit in mind).

Let's detail more precisely the construction:

- **Step 0**: First, we fix a parameter $\gamma$ (maximum disagreement allowed) and $\alpha$ (maximum loss of cost and distance to $\mathcal{D}$). This leads to a choice of $\epsilon = \frac{\alpha\gamma^2}{16}$ to trade of.
- **Step 1**: Then, we determine $\mathcal{F}_\epsilon^{\mathcal{D}}$ which involves finding both the cost minimizing hypothesis and the set of hypothesis whose cost is at a range of $\epsilon$ of it. Note that we have no results on the disagreement between hypothesis in $\mathcal{F}_\epsilon^{\mathcal{D}}$ so far.
- **Step 2**: We then pick an hypothesis $\tilde{f}$ in $\mathcal{F}_\epsilon^{\mathcal{D}}$ with ideally maximal $\operatorname{err}_{\mathcal{D}}$.
- **Step 3**: Finally, we build $\tilde{\mathcal{D}} := (1-\alpha)\mathcal{D} + \alpha\mathcal{P}$ where:

  $$\mathcal{P} := \mathcal{D} \mid \left\{x \mid \tilde{f}(x) = f^*(x)\right\}$$

  By doing this, we add bias towards less accurate but less complex solution. One key observation is that we are trying to give more importance in $\mathcal{D}$ to points where $\tilde{f}$ will agree with $f^\star$.

With a learning perspective in mind, one may now wonder how to implement all those steps without having access to the distribution $\mathcal{D}$ (and hence to $\mathcal{F}_\epsilon^{\mathcal{D}}$, $f^\star$ and $\operatorname{err}_{\mathcal{D}}$) but only to sample $\mathcal{S}$ coming from one or several learning agents. The algorithm would act as a coordinator agent, observing the realization of $\mathcal{S}$ associated with agents and adding potentially new points to the distribution. Let's study the different steps to get an overview of a potential algorithm:

- **Step 1**: We want to be able to say if an hypothesis belongs to the set $\mathcal{F}_\epsilon^{\mathcal{D}}$. This involves being able to compute both $c_{\mathcal{D}}^\star := \min_{f\in\mathcal{F}} \operatorname{cost}_{\mathcal{D}}(f)$ and $\operatorname{err}_{\mathcal{D}}(f)$ (or equivalently $\operatorname{cost}_{\mathcal{D}}(f)$ if we suppose $\phi$ known). Using Learning Theory results, we know that for enough points, $\operatorname{cost}_{\mathcal{S}}(f)$ will be close uniformly to $\operatorname{cost}_{\mathcal{D}}(f)$. Thus, we can use $c_{\mathcal{D}}^\star$ as an approximation of $c_{\mathcal{S}}^\star$.
- **Step 2**: Then, we want to sample hypothesis, evaluate their empirical cost as an approximation of the expected cost. If the sample process of hypothesis is sufficiently big, that would bring guarantees on the approximation of the max error hypothesis in $\mathcal{F}_\epsilon^{\mathcal{D}}$.
- **Step 3**: Finally, to find points from $\mathcal{P}$ to add bias in the distribution, we need to have a candidate for $f^\star$ and an approximation of $\mathcal{D}$. One option is to use a Version Space (the set of hypothesis consistent with the sampled set $\mathcal{S}$). The diameter of such set is converging to zero (at a pace dictated by a well know coefficient called disagreement coefficient [Hanneke, 2011]). Thus, for sufficient points, we expect to make few mistake in the choice of a candidate for $f^\star$.

By nature, those ideas are dynamic and the distribution is biased at each step, by giving new data points to the agents. Although the process might not be exact (and add unwanted bias that we can't correct), we expect to be able to derive guarantees showing a decrease of maximum disagreement for a sample set $S$ of sufficient size.

### 3.2 Approximation of expected cost minimizer, sampling from Rashomon Set and bias creation

First, we need an approximation $\hat{\mathcal{D}}$ of $\mathcal{D}$, which can be done using non-parametric density estimation for instance. For a given set $\mathcal{S}$, let's define the version space:

$$V := \{f \in \mathcal{F} : f(x) = y, \forall (x,y) \in S\}$$

For a subset of hypothesis $\mathcal{H}$, let's also introduce it subset of hypothesis coherent the labeling of $x$ as $y$:

$$V_x^y(\mathcal{H}) := \{f \in \mathcal{F} : f(x) = y, h \in \mathcal{H}\}$$

Then, to sample an hypothesis from $\mathcal{F}_\epsilon^{\mathcal{D}}$, let's assume that we have access to a distribution $\rho$ on $\mathcal{F}$, or a way of sampling hypothesis. Thus, by sampling sufficient hypothesis in $\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)$ (where sufficient is defined w.r.t. relative volume of this region in total hypothesis space), we can pick a subset of hypothesis with maximal empirical error $\mathcal{E}$ (Importance sampling could help to fasten the convergence if we have a good prior for this region).

We know that those empirical errors might evolve between 0 and $2\epsilon$ with high probability for $n$ large enough. Even if they disagree in some points with the error free optimal solution, we know that the empirical error of those hypothesis might converge to zero, in some scenarios (with infinite support distribution for instance). Yet, if the expected error of those hypothesis goes to 0, the bias we are going to introduce tends to $\mathcal{D}$ and thus doesn't modify the initial distribution.

A first natural idea is to enforce the modified distribution to say realizable while introducing the bias by picking points $(x, y)$ verifying:

$$\max_{(x,y)\in S} \rho(V_x^y(\mathcal{E}))$$

This quantity might be equal to zero, if all maximal error hypothesis selected disagree with the true labeling on all the sampled set so far. Yet, if this happens for long time, it means that the disagreement in $\mathcal{F}_\epsilon^{\mathcal{S}}$ is zero (for $\epsilon$ small enough) as all hypothesis will completely disagree with the true labeling and thus, our problem will not exist. The introduction of the bias towards this point is done using the density approximation in order to match the mass of distribution $\mathcal{D}$ on this point. Yet, we believe that the mass added initially to the new points, modeled with a parameter $\alpha_n$ should progressively increase as we gain in confidence on our approximation.

A more interesting track could be to look for $(x, y)$ verifying:

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \rho(V_x^y(\mathcal{E}) \cup V_x^y(V))$$

In this case, we would trade off between the maximal agreement between the majority vote of the version space and points where the maximal empirical error hypothesis is reached. This voting procedure can be estimated and implemented thanks to our distribution on $\mathcal{F}$. Again, those quantities can be estimated using sampling from $\mathcal{F}$ (or importance sampling ideas again). Our intuition is that the maximization on $\mathcal{D}$ allows to get faster convergence rates although it might make the new distribution non-realizable.

## 4 Conclusion and further directions

In this work, we have studied the impact of general form penalized on the potential disagreement of learning agent i.e. the diameter of the Rashomon Set induced by pseudo metric associated with 0-1 loss. We have shown that a global lowering of the amplitude of the penalization can lead to bounds on the diameter but that the asymptotic behavior, which exhibit both a dependence in $\epsilon$ and $\lambda$ limits the range of results in the general case and call for more assumptions (such as the finiteness of $\{\mathrm{err}_D(f) \mid f \in \mathcal{F}\}$) ). Then, we studied the different steps needed in order to tackle an open question in [Haghtalab et al., 2019]: is it possible to bias in an learning fashion the initial distribution in order to lower the potential disagreement between agents ? By making link with some Active Learning ideas, we've suggested an algorithm and prepared the ground for a deeper theoretical analysis. Finally, several others points appears to be of great interest with this framework in mind:

- The disagreement remains a prediction type of notion and call to be implemented in an action taking context where agents would use those predictions to interact with an environment. Thus, disagreement would lead to differences in reward. The contextual bandits framework seems to be very suited to answer this question and may be a very promising avenue for this work. In particular, the recent work of [Foster et al., 2020] bridge some interesting gaps.
- One a different topic, one could also be interested in modeling similar framework with different type of agents (with distinct hypothesis space and/or complexity functions). This lead to interesting modeling scenarios and leverage more realistic description of polarization. Rashomon sets have also been studied extensively with results under the assumption of growing hypothesis space [Semenova et al., 2020].

## References

[Fisher et al., 2019] Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.

[Foster et al., 2020] Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. (2020). Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective.

[Haghtalab et al., 2019] Haghtalab, N., Jackson, M., and Procaccia, A. (2019). Polarization through the lens of learning theory.

[Hanneke, 2007] Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 353–360, New York, NY, USA. Association for Computing Machinery.

[Hanneke, 2011] Hanneke, S. (2011). Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361.

[Semenova et al., 2020] Semenova, L., Rudin, C., and Parr, R. (2020). A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning.

[Tosh and Dasgupta, 2017] Tosh, C. and Dasgupta, S. (2017). Diameter-based active learning.

# 5 Appendix

**Lemma 5.1.** *For two sets $A$ and $B$ such that $\mathrm{diam}(A), \mathrm{diam}(B) < \infty$, $d_{\mathcal{H}}(A, B)$ is finite and we have:*

$$\mathrm{diam}(A) - 2d_{\mathcal{H}}(A, B) \leq \mathrm{diam}(B) \leq \mathrm{diam}(A) + 2d_{\mathcal{H}}(A, B)$$

*Proof.* Use the fact that $d_{\mathcal{H}}(A, B)$ is the is the infimum of all $d \in \mathbb{R}$ such that A is in the $d$-neighborhood of B and B is in the $d$-neighborhood of A. Let's note $r = d_{\mathcal{H}}(A, B)$. For every $\epsilon > 0$, each set $A$ and $B$ are in the $r + \epsilon$ neighborhood of the other. We take any two points $x, y \in B$. Each has a corresponding point $x', y'$ in $A$ at distance at most $r + \epsilon$, i.e.

$$d(x, x') < r + \epsilon$$
$$d(y, y') < r + \epsilon$$

since $x', y' \in A$, we have

$$d(x', y') \leq \mathrm{diam}(A)$$

Apply the triangle inequality to get

$$d(x, y) \leq \mathrm{diam}(A) + 2r + 2\epsilon$$

since $x$ and $y$ were arbitrary points in $B$, and $\epsilon > 0$ was arbitrary, we get

$$\mathrm{diam}(B) \leq \mathrm{diam}(A) + 2r$$

The other inequality follows the same way. □

**Lemma 5.2.** *For $f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$, we have the two following behaviors:*

- *If $\exists \lambda_0 \leq \lambda$ with $\lambda_0 > 0$, $f_\lambda \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$, then $\forall \lambda_1 \leq \lambda_0$, $f_\lambda \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_1)$. Moreover, $f_\lambda \notin \mathcal{B}(f^\star, \epsilon)$.*
- *Otherwise, $\forall \lambda_0 \leq \lambda$ such that $\lambda_0 > 0$, $f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$, and we have $f_\lambda \in \mathcal{B}(f^\star, \epsilon)$.*

*Proof.* First, let's suppose that there exists $\lambda_0 > 0$ such that $f_\lambda \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$. Let's note $c^\star(\lambda) = \mathrm{argmin}_{\tilde{f} \in \mathcal{F}} \mathrm{cost}_{\mathcal{D}}^{\lambda}(\tilde{f})$. Thus, we have $\mathrm{cost}_{\mathcal{D}}^{\lambda_0}(f) > \epsilon + c^\star(\lambda_0)$ and $\mathrm{cost}_{\mathcal{D}}^{\lambda}(f) \leq \epsilon + c^\star(\lambda)$. We know that $\lambda \mapsto \epsilon + c^\star(\lambda)$ is concave, thus the curve will be bellow the affine function going through the image of $\lambda$ and $\lambda_0$. On the other hand, $\lambda \mapsto \mathrm{cost}_{\mathcal{D}}^{\lambda}(f)$ will be strictly above this curve starting under $\lambda_0$. If $f$ belongs to $\mathcal{B}(f^\star, \epsilon)$, then the expected error of $f$ in under $\epsilon$ and thus, if $f$ belongs to a certain $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$, it will remain in the set. Then, if $f_\lambda \notin \mathcal{B}(f^\star, \epsilon)$, then the expected error of $f$ is strictly above $\epsilon$ and thus, as $c^\star$ tends to 0 for $\lambda \to 0^+$ (consider the cost of $f^\star$ to show this), we can pick $\lambda$ such that $c^\star(\lambda) + \epsilon$ is bellow the expected error of $f$ and thus, $f$ leave $\mathcal{F}_\epsilon^{\mathcal{D}}$, which contradicts our initial hypothesis.

□

**Theorem 5.3** (Convergence equality for finite image hypothesis set)**.** *If $\{\mathrm{err}_{\mathcal{D}}(f) \mid f \in \mathcal{F}\}$ is finite, then there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) = 0$$

*Proof.* If $\{\mathrm{err}_{\mathcal{D}}(f) \mid f \in \mathcal{F}\}$ is finite, then there exists $\eta > 0$ such that if $\mathrm{err}_{\mathcal{D}}(f) \leq \epsilon + \eta$, then $\mathrm{err}_{\mathcal{D}}(f) \leq \epsilon$. In other words, there is not accumulation points in $\epsilon$. As $c^\star$ tends to 0 for $\lambda \to 0^+$, we can pick $\lambda_0$ such that $c^\star(\lambda_0) \leq \eta$ (and this will still be true for $\lambda \leq \lambda_0$ as $c^\star$ is decreasing). Thus, let's take $f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda_0)$ (same reasoning apply for $\lambda \leq \lambda_0$). As $\phi$ is non-negative, it means that $\{\mathrm{err}_{\mathcal{D}}(f) \leq \epsilon + \eta$ and thus $\{\mathrm{err}_{\mathcal{D}}(f) \leq \epsilon$. Thus, $f \in \mathcal{B}(f^\star, \epsilon)$ and $\inf_{\tilde{f} \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, \tilde{f}) = 0$. □

**Lemma 5.4.** *For a given distribution $\mathcal{D}$, we have:*

$$\forall \lambda > 0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) \leq \min(1, \epsilon + c^\star(\lambda) - \Gamma) \leq \min(1, \epsilon + c^\star(\lambda))$$

*where $\Gamma \geq 0$ is the minimal gradient of an affine function tangent to $\lambda_0 \mapsto c^\star(\lambda_0)$ and going through $c^\star(\lambda) + \epsilon$. Moreover, there exists for all $\lambda > 0$, a distribution $\mathcal{D}$, an $\epsilon > 0$ and an hypothesis space $\mathcal{F}$ where the equality is reached (for a fixed lambda !).*

*Proof.* The bound of 1 is quite straightforward. To show the other bound, one just has to take a point $f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ and to bound its expected error by the maximal point in the $y$-axis it can reach. It's easy to show that this depends on the maximal value of the cost in $\lambda$ ($c^\star(\lambda)$) and minimal possible direction coefficient $\Gamma >\geq 0$. Then, this point is at distance $\epsilon + c^\star(\lambda) - \Gamma$ of $f^\star$. This geometrical interpretation then helps to show that the bound is sharp at a given $\lambda$. □

**Lemma 5.5.** *If $f \in \mathcal{B}(f^\star, \epsilon)^\circ$, then there exists $\lambda_0$ such that $\forall \lambda \leq \lambda_0$, $f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$.*

*Proof.* One just has to pick $\lambda$ small enough to ensure that $\mathrm{err}_{\mathcal{D}}(f) + \lambda|\phi| \leq \epsilon$, which is possible as the expected error of $f$ is strictly bellow $\epsilon$. □

**Lemma 5.6.** *If $f \in \partial\mathcal{B}(f^\star, \epsilon)$ and $\phi(f^\star) > \phi(f)$, then $\forall \lambda > 0, f \notin \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$.*

*Proof.* We can show that for all $\lambda$, $\mathrm{cost}_{\mathcal{D}}^{\lambda}(f) > \epsilon + \mathrm{cost}_{\mathcal{D}}^{\lambda}(f^\star) \geq c^\star(\lambda) + \epsilon$ □

**Lemma 5.7.** *There exists $\lambda_0 > 0$ such that*

$$\forall \lambda \leq \lambda_0, \sup_{\mathcal{B}(f^\star,\epsilon)} \inf_{f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f_0, f) \leq \epsilon$$

*Proof.* One can for instance pick $\lambda_0 = \frac{\epsilon}{|\phi(f^\star)|}$ if $|\phi(f^\star)| > 0$ and $\lambda > 0$ otherwise. □

**Lemma 5.8.** *If there exists $\eta > 0$ such that $\forall f \in \mathcal{B}(f^\star, \epsilon)^\circ, \mathrm{err}_{\mathcal{D}}(f) \leq (1-\eta)\epsilon$ and $\phi$ is bounded on $\mathcal{B}(f^\star, \epsilon)^\circ$, then there exists $\lambda_0$ such that $\forall \lambda \leq \lambda_0, \mathcal{B}(f^\star, \epsilon)^\circ \subset \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. In particular, we can deduce that:*

$$\forall \lambda \leq \lambda_0, \sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) = 0$$

*Proof.* We use similar ideas: for a given $f \in \mathcal{B}(f^\star, \epsilon)^\circ$, we define:

$$\lambda_f = \frac{\epsilon - \mathrm{err}_{\mathcal{D}}(f)}{|\phi(f)|}$$

if $|\phi(f)| > 0$ and $\lambda = 1$ otherwise. We have for all $\lambda \leq \lambda_f$, $f \in \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)$. The conditions $\phi$ bounded and $\mathrm{err}_{\mathcal{D}}(f) \leq (1-\eta)\epsilon$ ensure that $\lambda_f$ is lower-bounded by a strictly positive constant $\lambda_0$. □

**Theorem 5.9.** *If $\mathcal{F}$ has finite VC dimension:*

$$\sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) \to_{\lambda \to 0^+} 0$$

*Proof.* Key ideas: Although we have pointwise convergence, the uniformity of it w.r.t. $\lambda$ is not obvious. If we assume that this is not true, if involves that for a sequence of $\lambda$ that tends to $0^+$, we are able to find $f_\lambda \in \mathcal{B}(f^\star, \epsilon)^\circ$ such that the distance between $f_\lambda$ and $\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)$ is strictly above a certain threshold $\eta$. Yet, the trick is to add sufficient functions in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ with pointwise convergence to build a finite $\frac{\eta}{2}$ covering of $\mathcal{F}$ that would be included (and remain) in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. This would contradict our non-uniform convergence as all functions in $\mathcal{B}(f^\star, \epsilon)^\circ$ would then be at a distance $\frac{\eta}{2}$ of a function in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. To do so, we take a given finite $\frac{\eta}{2}$ covering of $\mathcal{F}$ that exists thanks to the fact that $\mathcal{F}$ has finite VC dimension. Then, we progressively add the functions $f_\lambda$ in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ (which is possible by decreasing $\lambda$ and they will remain in such set as shown above). Our hypothesis ensure that we can successively add a potential infinite number of functions that would be at a distance of $\eta$ of all the functions already in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. Yet, to do so, two such functions $f_\lambda$ can't be associated with the same function in the $\frac{\eta}{2}$ covering of $\mathcal{F}$, or they would be at a distance of $\eta$ and this would contradict the non-uniform hypothesis. But, as the covering is finite, we can't add an infinite number of new functions and thus we reach a contradiction. □

**Lemma 5.10.** *If $\mathcal{D}$ has a finite support, then, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \sup_{f \in \mathcal{B}(f^\star,\epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) = 0$$

*Proof.* For all possible pattern, if there exist an hypothesis in $\mathcal{B}(f^\star, \epsilon)^\circ$ realizing this pattern, we pick $\lambda$ small enough to ensure that such hypothesis in included in $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$. As there is only a finite number of such hypothesis, picking the minimizer of the associated $\lambda$ ensure that all hypothesis in $\mathcal{B}(f^\star, \epsilon)^\circ$ are in full agreement with an hypothesis in $\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)$ and thus that the maximal distance is 0. □

**Theorem 5.11.** *For all $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) \leq 2\epsilon$$

$$\forall \lambda \leq \lambda_0, \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) - 2\epsilon \leq \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) + 2\epsilon$$

*Proof.* We take as $\lambda_0$ the minima of the $\lambda_0$ provided by lemma 2.10 and of a $\lambda$ such that $c^\star(\lambda) \leq \epsilon$. Thanks to lemma 2.10, we have:

$$\sup_{\mathcal{B}(f^\star,\epsilon)} \inf_{f \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f_0, f) \leq \epsilon$$

and thanks to lemma 2.5, we have:

$$\sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star,\epsilon)} \Delta(f, f_\lambda) \leq 2\epsilon$$

Thus, for $\lambda \leq \lambda_0$, we have:

$$d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) \leq 2\epsilon$$

The second inequality is immediate to get using the property of the Hausdorff distance of lemma 2.1. □

**Corollary 5.11.1.** *For any $\mathcal{D}$, there exists $\lambda_0$ such that:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 4\epsilon$$

*Proof.* We using theorem 2.14 and the fact that $\operatorname{diam}(\mathcal{B}(f^\star, \epsilon)) \leq 2\epsilon$. □

**Theorem 5.12.** *Under the Tsybakov's low-noise assumption, there exists $\lambda_0$ such that, for all $\epsilon > 0$ and $\tilde{\epsilon}$ arbitrarily close to 0:*

$$\forall \lambda \leq \lambda_0, \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 2\epsilon + \mu\epsilon^{1/\kappa} + \tilde{\epsilon}$$

*Proof.* Using previous results, we can actually showed that the Hausdorff distance is upper bounded by $\epsilon + \tilde{\epsilon}$, where the second quantity can be taken arbitrarily close to 0, as $c^\lambda$ tends to 0 in $0^+$. We combine this with the Tsybakov upper bound on $\operatorname{diam}(\mathcal{B}(f^\star, \epsilon))$ to get the main result. □

**Theorem 5.13.** *If there is a finite number of patterns of $\mathcal{F}$ on $\mathcal{D}$, for all except a finite number of $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) = 0$$

$$\forall \lambda \leq \lambda_0, \quad \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) = \operatorname{diam}(\mathcal{B}(f^\star, \epsilon))$$

*Proof.* The condition on $\mathcal{F}$ and $\mathcal{D}$ involves in particular that $\{\operatorname{err}_{\mathcal{D}}(f) \mid f \in \mathcal{F}\}$ is finite. By removing from the possible $\epsilon$ all values in $\{\operatorname{err}_{\mathcal{D}}(f) \mid f \in \mathcal{F}\}$, we can also use lemma 2.13 and the sets $\mathcal{B}(f^\star, \epsilon)$ and $\mathcal{B}(f^\star, \epsilon)^\circ$ are equals, thus we can conclude. □

**Corollary 5.13.1.** *For any sample $\mathcal{S}$ drawn form $\mathcal{D}$, for all except of finite number of $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:*

$$\forall \lambda \leq \lambda_0, \quad \operatorname{diam}_{\mathcal{S}}(\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)) = \operatorname{diam}_{\mathcal{S}}(\mathcal{B}^{\mathcal{S}}(f^\star, \epsilon))$$

*Proof.* One just have to note that the distribution associated with $\mathcal{S}$ has finite support, so we can apply theorem 2.15. □

**Lemma 5.14.** *If $\mathcal{F}$ has finite VC dimension, for $\epsilon, \delta > 0$, there exists $N(\epsilon, \delta)$ such that for all sample $\mathcal{S}$ of size at least $N(\epsilon, \delta)$, we have with probability as least $1 - \delta$:*

$$\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) \subset \mathcal{F}_{2\epsilon}^{\mathcal{S}}(\lambda) \quad \textit{and} \quad \mathcal{B}(f^\star, \epsilon) \subset \mathcal{B}^{\mathcal{S}}(f^\star, 2\epsilon)$$

*Moreover, we also have:*

$$\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda) \subset \mathcal{F}_{2\epsilon}^{\mathcal{D}}(\lambda) \quad \textit{and} \quad \mathcal{B}^{\mathcal{S}}(f^\star, \epsilon) \subset \mathcal{B}(f^\star, 2\epsilon)$$

*Proof.* Using classical results from learning theory, there exists $N(\epsilon, \delta)$ such that with probability at least $1 - \delta$, for $S$ of size at least $N(\epsilon, \delta)$, we have:

$$\forall f \in \mathcal{F}, |\operatorname{err}_{\mathcal{D}}(f) - \operatorname{err}_{\mathcal{S}}(f)| \leq \tilde{\epsilon}$$

Let's start by showing that $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) \subset \mathcal{F}_{2\epsilon}^{\mathcal{S}}(\lambda)$. Using the previous result, we have for a given $\in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$:

$$\begin{aligned}\mathrm{cost}_{\mathcal{S}}(f) &\leq \mathrm{cost}_{\mathcal{D}}(f) + \tilde{\epsilon} \\ &\leq \mathrm{cost}_{\mathcal{D}}(f^\star) + \tilde{\epsilon} + \epsilon \\ &\leq \mathrm{cost}_{\mathcal{D}}(f_{\mathcal{S}}^\star) + \tilde{\epsilon} + \epsilon \\ &\leq \mathrm{cost}_{\mathcal{S}}(f_{\mathcal{S}}^\star) + 2\tilde{\epsilon} + \epsilon\end{aligned}$$

Thus, by taking $\tilde{\epsilon} \leq \frac{\epsilon}{2}$, we can ensure the result. To show the result for the unit ball, similar reasoning give the results (using the fact that $\mathrm{err}_{\mathcal{D}}(f^\star) = 0$ can remove one $\tilde{\epsilon}$). □

**Theorem 5.15.** *For any $\epsilon, \delta > 0$, there exists $N(\epsilon, \delta)$ and $\lambda_0(\epsilon)$ such that for all set $\mathcal{S}$ for size at least $N(\epsilon, \delta)$, with probability $1 - \delta$:*

$$\forall \lambda \leq \lambda_0, \mathrm{diam}_{\mathcal{S}}(\mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)) \leq \mathrm{diam}_{\mathcal{D}}(\mathcal{F}_{2\epsilon}^{\mathcal{D}}(\lambda)) \leq 12\epsilon$$

*Proof.* We combine the previous lemma and corrolary 2.14.1 □