Concentration: uniform convergence, VC dimension and the fundamental theorem of PAC learning

Master 2 Mathematics and Computer Science

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval
2020-2021
1. PAC learning

2. Uniform convergence for infinite classes: VC dimension and Sauer’s lemma

3. Finite VC dimension implies Uniform Convergence

4. Finite VC-dimension implies learnability
PAC learning
Learning framework

• Domain $\mathcal{X}$, label space $\mathcal{Y} = \{0, 1\}$
• Unknown distribution $D$ on $\mathcal{X} \times \mathcal{Y}$
• Sample $S = (X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} D$
• $h : \mathcal{X} \rightarrow \mathcal{Y}, h \in \mathcal{H}$ hypothesis class
• loss function $\ell(y, y') = \mathbb{1}\{y \neq y'\}$
• generalization error (loss)
  $L_D(h) = \mathbb{E}_D[\ell(h(X), Y)] = \mathbb{E}_D[h(X) \neq Y]$
• training error $L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}$
• agnostic learning $\neq$ realizable assumption (when there exists $h^*$ such that $L_S(h^*) = 0$)
• learning algorithm: $S \mapsto \hat{h}_n$ such that $L_D(\hat{h}_n) - \inf_{h \in \mathcal{H}} L_D(h)$ small
**Definition**

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a function $n_\mathcal{H} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_n$ such that for every $\epsilon, \delta \in (0, 1)$, for every distribution $D$ on $X \times Y$ when $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{iid}{\sim} D$,

$$\mathbb{P}\left( L_D(\hat{h}_n) \geq \inf_{h' \in \mathcal{H}} L_D(h') + \epsilon \right) \leq \delta$$

for all $n \geq n_\mathcal{H}(\epsilon, \delta)$.

The smallest possible function $n_\mathcal{H}$ is called the *sample complexity* of learning $\mathcal{H}$. 

---

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A hypothesis class $\mathcal{H}$ is <em>agnostic PAC learnable</em> if there exists a function $n_\mathcal{H} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}<em>n$ such that for every $\epsilon, \delta \in (0, 1)$, for every distribution $D$ on $X \times Y$ when $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{iid}{\sim} D$, [ \mathbb{P}\left( L_D(\hat{h}<em>n) \geq \inf</em>{h' \in \mathcal{H}} L_D(h') + \epsilon \right) \leq \delta ] for all $n \geq n</em>\mathcal{H}(\epsilon, \delta)$.</td>
</tr>
</tbody>
</table>
Learning via uniform convergence

**Definition**
A training set $S$ is called $\epsilon$-representative (wrt domain $\mathcal{X} \times \mathcal{Y}$, hypothesis class $\mathcal{H}$, loss function $\ell$ and distribution $D$) if

$$\forall h \in \mathcal{H}, \left| L_S(h) - L_D(h) \right| \leq \epsilon.$$ 

**Lemma**
If $S$ is $\epsilon/2$-representative, then any ERM $\hat{h}_n$ defined by $\hat{h}_n \in \arg \min_{h \in \mathcal{H}} L_S(h)$ satisfies:

$$L_D(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

Proof: for every $h \in \mathcal{H}$,

$$L_D(\hat{h}_n) \leq L_S(\hat{h}_n) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$
Uniform Convergence Property

**Definition**
A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (wrt $\mathcal{X} \times \mathcal{Y}$ and $\ell$) if there exists a function $n_{\mathcal{H}}^{UC} : (0,1)^2 \rightarrow \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every distribution $D$ over $\mathcal{X} \times \mathcal{Y}$, a sample $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{\text{iid}}{\sim} D$ of size $n \geq n_{\mathcal{H}}^{UC}(\epsilon, \delta)$ has probability at least $1 - \delta$ to be $\epsilon$-representative.

**Corollary**
If $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostically PAC learnable with a sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq n_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$. Furthermore, the ERM is a successful PAC learner for $\mathcal{H}$. 
Finite classes areagnostically PAC-learnable

**Theorem**

Let $\mathcal{H}$ be a finite hypothesis class. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2} \right\rceil .$$

Moreover, $\mathcal{H}$ is agnostically PAC learnable using an ERM algorithm with sample complexity

$$n_{\mathcal{H}}(\epsilon, \delta) \leq 2n_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2 \log \frac{2|\mathcal{H}|}{\delta}}{\epsilon^2} \right\rceil .$$

Proof: Hoeffding’s inequality and the union bound.
Uniform convergence for infinite classes: VC dimension and Sauer’s lemma
**Definition**

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0, 1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The **restriction** of $\mathcal{H}$ to $C$ is the set of functions $C \to \{0, 1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \left\{ (c_1, \ldots, c_m) \to (h(c_1), \ldots, h(c_m)) : h \in \mathcal{H} \right\}.$$

**Shattering**

A hypothesis class $\mathcal{H}$ **shatters** a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0, 1\}^C$.

**Example:**

- $\mathcal{H} = \{1_{]-\infty, a]} : a \in \mathbb{R}\}$.
- $\mathcal{H}_{\text{rec}}^2 = \{h(a_1, b_1, a_2, b_2) : a_1 \leq b_1 \text{ and } a_2 \leq b_2\}$ where
  $$h(a_1, b_1, a_2, b_2)(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2; \\ 0 & \text{otherwise}. \end{cases}$$
VC dimension

**Definition**

The *Vapnik Chervonenkis dimension* VCdim(\(\mathcal{H}\)) of a hypothesis class \(\mathcal{H}\) is the maximal size of a set \(C \subset \mathcal{X}\) that can be shattered by \(\mathcal{H}\). If \(\mathcal{H}\) can shatter sets of arbitrarily large size we say that VCdim(\(\mathcal{H}\)) = \(\infty\).

Example:

- \(\mathcal{H} = \{ [1]_{-\infty, a] : a \in \mathbb{R}\}\}.
- \(\mathcal{H}^2_{\text{rec}} = \{ h(a_1, b_1, a_2, b_2) : a_1 \leq b_1 \text{ and } a_2 \leq b_2 \}\) where

\[
h(a_1, b_1, a_2, b_2)(x_1, x_2) = \begin{cases} 
1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2; \\
0 & \text{otherwise}.
\end{cases}
\]
Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of 0–1 loss. Then the following propositions are equivalent:

1. $\mathcal{H}$ has the uniform convergence property,
2. any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$,
3. $\mathcal{H}$ is agnostic PAC learnable,
4. $\mathcal{H}$ has finite VC-dimension.
Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Assume that $d := \text{VCdim}(\mathcal{H}) < \infty$. Then there exist constants $C_1, C_2$ such that:

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n^{\text{UC}}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2},$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2},$$
## Definition

Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as the maximal number of different functions that can be obtained by restricting $\mathcal{H}$ to a set of size $m$:

$$
\tau_{\mathcal{H}}(m) = \max_{C \subseteq X : |C| = m} |\mathcal{H}_C|.
$$

Note: if $\text{VCdim}(\mathcal{H}) = d$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$.

## Sauer’s lemma

Let $\mathcal{H}$ be a hypothesis class with $d = \text{VCdim}(\mathcal{H}) < \infty$. Then, for all $m \geq d$,

$$
\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d.
$$

Think of example: $\mathcal{H} = \{1_{(-\infty, a]} : a \in \mathbb{R}\}$ with $d = \text{VCdim}(\mathcal{H}) = 1$. 


Proof of Sauer’s lemma 1/2

In fact we prove the stronger claim:

\[ |\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}. \]

where the last inequality holds since no set of size larger than \(d\) is shattered by \(\mathcal{H}\). The proof is by induction.

\(m=1\): The empty set is always considered to be shattered by \(\mathcal{H}\). Hence, either \(|\mathcal{H}_C| = 1\) and \(d = 0\), inequality 1 \(\leq 1\), or \(d \geq 1\) and the inequality is \(2 \leq 2\).

**Induction:** Let \(C = \{c_1, \ldots, c_m\}\), and let \(C' = \{c_2, \ldots, c_m\}\). We note functions like vectors, and we define

\[ Y_0 = \left\{ (y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ or } (1, y_2, \ldots, y_m) \in \mathcal{H}_C \right\}, \]

\[ Y_1 = \left\{ (y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ and } (1, y_2, \ldots, y_m) \in \mathcal{H}_C \right\}. \]

Then \( |\mathcal{H}_C| = |Y_0| + |Y_1| \). Moreover, \(Y_0 = \mathcal{H}_C'\) and hence by the induction hypothesis:

\[ |Y_0| = |\mathcal{H}_C'| \leq |\{B' \subset C' : \mathcal{H} \text{ shatters } B'\}| = |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| \]

Next, define

\[ \mathcal{H}' = \left\{ h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } \forall 1 \leq i \leq n, h'(c_i) = \begin{cases} 1 - h(c_1) & \text{if } i = 1 \\ h(c_i) & \text{otherwise} \end{cases} \right\} \]

Note that \(\mathcal{H}'\) shatters \(B' \subset C'\) iff \(\mathcal{H}'\) shatters \(B' \cup \{c_1\}\), and that \(Y_1 = \mathcal{H}_C'\). Hence, by the induction hypothesis,

\[ |Y_1| = |\mathcal{H}_C'| \leq |\{B' \subset C' : \mathcal{H}' \text{ shatters } B'\}| = |\{B' \subset C' : \mathcal{H}' \text{ shatters } B' \cup \{c_1\}\}| \]

\[ = |\{B \subset C : c_1 \in B \text{ and } \mathcal{H}' \text{ shatters } B\}| \leq |\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}|. \]

Overall,

\[ |\mathcal{H}_C| = |Y_0| + |Y_1| \leq |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| + \{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| = |\{B \subset C : \mathcal{H} \text{ shatters } B\}|. \]
Proof of Sauer’s lemma 2/2

For the last inequality, one may observe that if \( m \geq 2d \), defining \( N \sim B(m, 1/2) \), Chernoff’s inequality and inequality \( \log(u) \geq (u - 1)/u \) yield

\[
-\log \mathbb{P}(N \leq d) \geq m \text{kl} \left( \frac{d}{m}, \frac{1}{2} \right) \geq d \log \frac{2d}{m} + (m - d) \log \frac{2(m - d)}{m}
\]

\[
\geq m \log(2) + d \log \frac{d}{m} + (m - d) \frac{-d/m}{(m - d)/m}
\]

\[
= m \log(2) + d \log \frac{d}{em},
\]

and hence

\[
\sum_{i=0}^{d} \binom{m}{i} = 2^m \mathbb{P}(N \leq d) \leq \exp \left( -d \log \frac{d}{em} \right) = \left( \frac{em}{d} \right)^d.
\]

Besides, for the case \( d \leq m \leq 2d \), the inequality is obvious since \( (em/d)^d \geq 2^m \): indeed, function \( f : x \mapsto -x \log(x/e) \) is increasing on \([0, 1]\), and hence for all \( d \leq m \leq 2d \):

\[
\frac{d}{m} \log \frac{em}{d} = f(d/m) \geq f(1/2) = \frac{1}{2} \log(2e) \geq \log(2),
\]

which implies

\[
\left( \frac{em}{d} \right)^d = \exp \left( d \log \frac{em}{d} \right) \geq \exp(m \log(2)) = 2^m.
\]

Alternatively, you may simply observe that for all \( m \geq d \),

\[
\left( \frac{d}{m} \right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left( \frac{d}{m} \right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left( \frac{d}{m} \right)^i \binom{m}{i} = \left( 1 + \frac{d}{m} \right)^m \leq e^d.
\]
Finite VC dimension implies Uniform Convergence
Finite VC dimension implies Uniform Convergence

**Theorem**

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every distribution $D$ and for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the choice of the sample $S \sim D^\otimes n$ we have

$$
\sup_{h \in \mathcal{H}} \left| L_D(h) - L_S(h) \right| \leq \frac{1 + \sqrt{\log (\tau_{\mathcal{H}}(2n))}}{\delta \sqrt{n/2}}.
$$

Note: this result is sufficient to prove that finite VC-dim $\iff$ learnable, but the dependency in $\delta$ is not correct at all: roughly speaking, the factor $1/\delta$ can be replaced by $\log(1/\delta)$. 
We consider the 0-1 loss $\ell(h, (x, y)) = 1 \{ h(x) \neq y \}$, or any $[0, 1]$-valued loss $\ell$. We denote $Z_i = (X_i, Y_i)$, and observe that $L_D(h) = \mathbb{E}_{Z_i}[\ell(h, Z_i)] = \mathbb{E}_{S'}[L_{S'}(h)]$ if $S' = Z_1', \ldots, Z_n'$ denotes another iid sample of $D$. Hence,

$$\mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \right] = \mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'}[L_{S'}(h)] - L_S(h) \right| \right] = \mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'}[L_{S'}(h) - L_S(h)] \right| \right]$$

$$\leq \mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} \mathbb{E}_{S'} \left[ |L_{S'}(h) - L_S(h)| \right] \right] \leq \mathbb{E}_S \left[ \mathbb{E}_{S'} \left[ \sup_{h \in \mathcal{H}} |L_{S'}(h) - L_S(h)| \right] \right]$$

$$= \mathbb{E}_{S, S'} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(h, Z_i') - \ell(h, Z_i) \right]$$

$$= \mathbb{E}_{S, S'} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\ell(h, Z_i') - \ell(h, Z_i)) \right] \quad \text{for all } \sigma \in \{-1, 1\}^n$$

$$= \mathbb{E}_\Sigma \mathbb{E}_{S, S'} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \Sigma_i (\ell(h, Z_i') - \ell(h, Z_i)) \right] \quad \text{if } \Sigma \sim \mathcal{U}(\{-1, 1\}^n)$$

$$= \mathbb{E}_{S, S'} \mathbb{E}_\Sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \Sigma_i (\ell(h, Z_i') - \ell(h, Z_i)) \right].$$

Now, for every $S, S'$, let $C = C_{S, S'}$ be the instances appearing in $S$ and $S'$. Then $\forall \sigma \in \{-1, 1\}^n$,

$$\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, Z_i') - \ell(h, Z_i)) \right| = \max_{h \in \mathcal{H}_C} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, Z_i') - \ell(h, Z_i)) \right|.$$
Proof: symmetrization and Rademacher complexity (2/2)

Moreover, for every \( h \in \mathcal{H}_C \) let \( Z_h = \frac{1}{n} \sum_{i=1}^{n} \Sigma_i (\ell(h, Z'_i) - \ell(h, Z_i)) \). Then \( \mathbb{E}_\Sigma [Z_h] = 0 \), each summand belongs to \([-1, 1]\) and by Hoeffding’s inequality, for every \( \epsilon > 0 \):

\[
\mathbb{P}_\Sigma [\left| Z_h \right| \geq \epsilon] \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right).
\]

Hence, by the union bound,

\[
\mathbb{P}_\Sigma [\max_{h \in \mathcal{H}_C} |Z_h| \geq \epsilon] \leq 2 |\mathcal{H}_C| \exp\left(-\frac{n\epsilon^2}{2}\right).
\]

The following lemma permits to deduce that

\[
\mathbb{E}_\Sigma \left[ \max_{h \in \mathcal{H}_C} |Z_h| \right] \leq \frac{1 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{n/2}} \leq \frac{1 + \sqrt{\log(\tau_\mathcal{H}(2n))}}{\sqrt{n/2}}.
\]

Hence,

\[
\mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \right] \leq \mathbb{E}_{S, S'} \mathbb{E}_\Sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \Sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right] \leq \frac{1 + \sqrt{\log(\tau_\mathcal{H}(2n))}}{\sqrt{n/2}},
\]

and we conclude by using Markov’s inequality (poor idea! Better: McDiarmid’s inequality).
Lemma

Let \( a > 0, \) \( b > 1, \) and let \( Z \) be a real-valued random variable such that for all \( t \geq 0, \) \( \mathbb{P}(Z \geq t) \leq 2b \exp \left( -\frac{t^2}{a^2} \right). \) Then

\[
\mathbb{E}[Z] \leq a \left( \sqrt{\log(b)} + \frac{1}{\sqrt{\log(b)}} \right).
\]

Proof:

\[
\mathbb{E}[Z] \leq \int_0^\infty \mathbb{P}(Z \geq t) dt \leq a \sqrt{\log(b)} + \int_{a\sqrt{\log(b)}}^\infty 2b \exp \left( -\frac{t^2}{a^2} \right) dt
\]

\[
\leq a \sqrt{\log(b)} + 2b \int_{a\sqrt{\log(b)}}^\infty \frac{t}{a \sqrt{\log(b)}} \exp \left( -\frac{t^2}{a^2} \right) dt
\]

\[
= a \sqrt{\log(b)} + \frac{2b}{a \sqrt{\log(b)}} \times \frac{a^2}{2} \exp \left( -\frac{(a\sqrt{\log(b)})^2}{a^2} \right)
\]

\[
= a \sqrt{\log(b)} + \frac{a}{\sqrt{\log(b)}}.
\]

NB: cutting at \( a\sqrt{\log(2b)} \) gives a better but less nice inequality for our use.
Finite VC-dimension implies learnability
It suffices to prove that finite VC-dim implies the uniform convergence property. From Sauer’s lemma, for all $m \geq d/2$ we have $\tau_{\mathcal{H}}(2n) \leq (2en/d)^d$. With the previous theorem, this yields that with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \leq \frac{1 + \sqrt{d \log (2en/d)}}{\delta \sqrt{n/2}} \leq \frac{1}{\delta} \sqrt{\frac{8d \log (2en/d)}{n}}$$

as soon as $\sqrt{d \log (2en/d)} \geq 1$. To ensure that this is at most $\epsilon$, one may choose

$$n \geq \frac{8d \log(n)}{(\delta \epsilon)^2} + \frac{8d \log(2e/d)}{(\delta \epsilon)^2}.$$

By the following lemma, it is sufficient that

$$n \geq \frac{32d \log \left( \frac{4d}{(\delta \epsilon)^2} \right) + 16d \log \left( \frac{2e}{d} \right)}{(\delta \epsilon)^2}.$$
Technical Lemma

Lemma
Let $a > 0$. Then

$$x \geq 2a \log(a) \implies x \geq a \log(x).$$

Proof: For $a \leq e$, true for every $x > 0$. Otherwise, for $a \geq \sqrt{e}$ we have $2a \log(a) \geq a$ and thus for every $t \geq 2a \log(a)$, as $f : t \mapsto t - a \log(t)$ is increasing on $[a, \infty)$, $f(t) \geq f(2a \log(a)) = a \log(a) - a \log(2 \log(a)) \geq 0$, since for every $a > 0$ it holds that $a \geq 2 \log(a)$.

Lemma
Let $a \geq 1$, $b > 0$. Then

$$x \geq 4a \log(2a) + 2b \implies x \geq a \log(x) + b.$$ 

Proof: It suffices to check that $x \geq 2a \log(x)$ (given by the above lemma) and that $x \geq 2b$ (obvious since $4a \log(2a) \geq 0$).