CSC 665: Vapnik-Chervonenkis (VC) Theory

Chicheng Zhang

September 12, 2019

1 Infinite hypothesis classes can be PAC learnable

In the last lecture, we have seen that the size of a hypothesis class $\mathcal{H}$ can be a important factor of sample complexity of learning from that $\mathcal{H}$. Specifically, if $\mathcal{H}$ is finite, then it has a PAC sample complexity upper bound of $O(\frac{1}{\epsilon}(\ln|\mathcal{H}|+\ln\frac{1}{\delta}))$, and an agnostic PAC sample complexity upper bound of $O(\frac{1}{\epsilon^2}(\ln|\mathcal{H}|+\ln\frac{1}{\delta}))$. Does that mean that if $\mathcal{H}$ is infinite, then $\mathcal{H}$ is not PAC learnable?

In this section, we give a counterexample, showing that for the hypothesis class of threshold functions on the $[0,1]$ interval, $\mathcal{H}$ is PAC learnable. To formalize the statement, we need some notation setup.

1. The instance domain $\mathcal{X}$ be the $[0,1]$ interval,
2. The label space $\mathcal{Y}$ be $\{-1,+1\}$.
3. The hypothesis class $\mathcal{H} = \left\{ h_t \triangleq 2\mathbf{1}(x > t) - 1 : t \in [0,1] \right\}$ is the set of threshold functions over $[0,1]$. Given classifier h_t, it will classify all examples x on the left of t as label -1, and classify all examples x on the right of t (including t) as label $+1$. Note that $\mathcal{H}$ is (uncountably) infinite.

Recall that the consistency algorithm is one that returns a classifier $\hat{h}$ in $\mathcal{H}$ that agrees with all training examples. We have the following theorem on the sample complexity of the consistency algorithm.

Theorem 1. *Suppose D is a distribution over $[0,1]$ that is realizable with respect to $\mathcal{H}$. Then, for any $\epsilon \in (0, \frac{1}{2})$, $\delta \in (0,1)$, given $m \geq \frac{1}{\epsilon}\ln\frac{2}{\delta}$ training examples drawn iid from D, the consistency algorithm returns a classifier $h_{\hat{t}}$ such that with probability $1-\delta$,*

$$\mathrm{err}(h_{\hat{t}}, D) \leq \epsilon.$$

Proof. We will only consider the setting where D is a continuous probability distribution that has density on $[0,1]$. (For a rigorous proof for general D, see Appendix A for details.)

Consider two points t_L and t_R, which are defined such that

$$\mathbb{P}(x \in [t^\star, t_R]) = \epsilon,$$

$$\mathbb{P}(x \in [t_L, t^\star]) = \epsilon.$$

We consider the setting where such t_L and t_R exists. See Appendix A for the proof for the general setting where t_L and t_R may not exist.

We are going to show that given a sample size $m \geq \frac{1}{\epsilon}\ln\frac{2}{\delta}$, there exists an event $\bar{E}$, such that at least one sample is in $[t^\star, t_R]$, and at least one sample is in $[t_L, t^\star]$. Note that if this happens, then the returned threshold $\hat{t}$ will be inside $[t_L, t_R]$.

If $\hat{t}$ is in $[t_L, t^\star]$, then

$$\mathrm{err}(h_{\hat{t}}, D) = \mathbb{P}(x \in [\hat{t}, t^\star] \leq \mathbb{P}(x \in [t_L, t^\star]) = \epsilon.$$

Similarly, if $\hat{t}$ is in $[t^\star, t_R]$, then

$$\text{err}(h_{\hat{t}}, D) = \mathbb{P}(x \in [t^\star, \hat{t}] \leq \mathbb{P}(x \in [t^\star, t_R]) = \epsilon.$$

Now define event E_L (resp. E_R) be such that no sample is in $[t_L, t^\star]$ (resp. $[t^\star, t_R]$), and define $E = E_L \cup E_R$. It suffices to show $\mathbb{P}(E) \leq \delta$. Indeed,

$$\mathbb{P}(E_L) = (1 - \mathbb{P}(x \in [t_L, t^\star]))^m \leq e^{-\epsilon m} = \delta/2$$

and similarly $\mathbb{P}(E_R) \leq \delta/2$. This implies that $\mathbb{P}(E) \leq \mathbb{P}(E_L) + \mathbb{P}(E_R) \leq \delta$. $\square$

2 VC dimension

We provide a more refined characterization of the complexity of a hypothesis class. Generally, if a hypothesis class is more expressive, then we may need more samples to learn from them. But how can we measure the expressiveness of a hypothesis class?

Definition 1. *Given a hypothesis class $\mathcal{H}$ and a set of unlabeled examples $S = \{x_1, \dots, x_n\}$, define the projection of $\mathcal{H}$ to S as:*

$$\Pi_{\mathcal{H}}(S) = \left\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\right\}.$$

Intuitively, if $\mathcal{H}$ is more expressive, then $|\Pi_{\mathcal{H}}(S)|$ is larger. The largest possible value of $|\Pi_{\mathcal{H}}(S)|$ is 2^n, where $\mathcal{H}$ achieves all possible $+1/-1$ labelings on S. In this case, we call that S is *shattered* by $\mathcal{H}$.

Definition 2. *The VC dimension of $\mathcal{H}$ (abbrev. $\mathrm{VC}(\mathcal{H})$), is the largest nonnegative integer d such that there exists S of size d that is shattered by $\mathcal{H}$. If no such d exists, we $\mathrm{VC}(\mathcal{H})$ is defined to be infinity.*

We have the following more checkable definition of VC dimension:

Lemma 1. *Suppose we are given a hypothesis class $\mathcal{H}$ and an integer d. Then $\mathrm{VC}(\mathcal{H}) = d$ is equivalent to the following two holding simultaneously:*

1. *There exists a set of examples of size d that is shattered by $\mathcal{H}$.*
2. *Any set of examples of size $d+1$ are not shattered by $\mathcal{H}$.*

Examples of VC dimension:

1. Thresholds in $\mathbb{R}$. $\mathcal{H} = \left\{h_t(x) \triangleq 2\mathbf{1}(x > t) - 1 : t \in [0,1]\right\}$. It can be seen that $\{0.5\}$ is shattered by $\mathcal{H}$. However, consider any set $S = \{x_1, x_2\}$. Suppose $x_1 \leq x_2$. Then it is impossible to find h_t in $\mathcal{H}$ such that $h_t(x_1) = +1$ and $h_t(x_2) = -1$. Therefore, $\mathrm{VC}(\mathcal{H}) = 1$.

2. Intervals in $\mathbb{R}$. $\mathcal{H} = \left\{h_{a,b}(x) \triangleq 2\mathbf{1}(a \leq x \leq b) - 1 : t \in [0,1]\right\}$. It can be seen that $\{0.2, 0.5\}$ is shattered by $\mathcal{H}$. However, consider any set $S = \{x_1, x_2, x_3\}$. Suppose $x_1 \leq x_2 \leq x_3$. Then it is impossible to find $h_{a,b}$ in $\mathcal{H}$ such that $h_{a,b}(x_1) = +1$, $h_{a,b}(x_2) = -1$, $h_{a,b}(x_3) = +1$. The reason is that: $h_{a,b}(x_1) = +1$ implies that $a \leq x_1$; $h_{a,b}(x_3) = +1$ implies that $x_3 \leq b$. However, this would imply that $x_2 \in [a,b]$, therefore $h_{a,b}(x_2) = +1$, and $h_{a,b}(x_2) = -1$ is impossible. Therefore, $\mathrm{VC}(\mathcal{H}) = 2$.

3. Homogeneous linear classifiers in $\mathbb{R}^d$. $\mathcal{H} = \left\{h_w(x) \triangleq 2\mathbf{1}(w \cdot x > 0) - 1 : w \in \mathbb{R}^d\right\}$. It can be seen that the canonical basis vectors $\{e_1, \dots, e_d\}$ (or more generally, any set of linearly independent examples) is shattered by $\mathcal{H}$. To see this, note that given a set of linearly independent examples $x_1, \dots, x_m$, consider the matrix $M \in \mathbb{R}^{m \times d}$ whose rows are the x_i's. Note that M has rank m, therefore its columns also spans the whole $\mathbb{R}^m$. Hence, for any vector l in $\mathbb{R}^m$, there is a vector w in $\mathbb{R}^d$, such that

$$Mw = \begin{bmatrix} \langle w, x_1 \rangle \\ \dots \\ \langle w, x_d \rangle \end{bmatrix} = l.$$

This immediately implies that for any labeling in $\{-1,+1\}^m$, there is a linear classifier in $\mathbb{R}^m$ that achieves that labeling. Hence, $\mathrm{VC}(\mathcal{H}) \geq d$.

However, consider any set $S = \{x_1, \ldots, x_{d+1}\}$. We now show that S is not shatterable.

First, $x_1, \ldots, x_{d+1}$ are $d+1$ vectors in $\mathbb{R}^d$, therefore they must be linearly dependent. Thus, there exists $\alpha_1, \ldots, \alpha_{d+1}$ not all zero, such that

$$\sum_{i=1}^{d+1} \alpha_i x_i = 0. \tag{1}$$

Furthermore, there exists $\alpha_1, \ldots, \alpha_{d+1}$, such that there exists $i^\star$ in $\{1, \ldots, d+1\}$, $\alpha_i^\star > 0$,

$$\sum_{i=1}^{d+1} \alpha_i x_i = 0.$$

The reason is as follows: if there already exists a positive α_i in Equation 1, then we are done; otherwise, we can flip the sign of the α_i's and ensuring at least one positive α_i.

Now consider the following labeling $(l_1, \ldots, l_{d+1})$, where $l_i = \begin{cases} +1, \alpha_i > 0 \\ -1, \alpha_i \leq 0 \end{cases}$. Can a linear classifier achieve such labeling? Suppose there is a w that achieves so. Then, for all i in $\{1, \ldots, d+1\}$,

$$\begin{cases} w \cdot x_i > 0, \alpha_i > 0 \\ w \cdot x_i \leq 0, \alpha_i \leq 0 \end{cases}$$

Thus, for all i, $\alpha_i \langle w, x_i \rangle \geq 0$. Specifically, for index $i^\star$, $\alpha_{i^\star} \langle w, x_{i^\star} \rangle > 0$. Summing over all i's, this implies that

$$\sum_{i=1}^{d+1} \alpha_i \langle w, x_i \rangle > 0.$$

This contradicts Equation 1, which would imply that

$$\sum_{i=1}^{d+1} \alpha_i \langle w, x_i \rangle = 0.$$

4. Non-homogeneous linear classifiers in $\mathbb{R}^d$. $\mathcal{H} = \left\{ h_w(x) \triangleq 2\mathbf{1}(w \cdot x + w_0 > 0) - 1 : (w, w_0) \in \mathbb{R}^{d+1} \right\}$. Using the same reasoning as above, it can be shown that the VC dimension of $\mathcal{H}$ is $d+1$. We leave the proof to you as an exercise.

Finally, we define the notion of growth function, which measures the largest possible number of for $\mathcal{H}$ to datasets of fixed size n.

Definition 3. *Define the growth function $\mathcal{S}(\mathcal{H}, n)$ as the maximum number of labelings one can generate on a dataset of size n, formally,*

$$\mathcal{S}(\mathcal{H}, n) = \max_{S : |S| = n} |\Pi_{\mathcal{H}}(S)|.$$

We also have the simple observation for finite classes.

Lemma 2. *If $\mathcal{H}$ is finite, then $\mathcal{S}(\mathcal{H}, n) \leq |\mathcal{H}|$ and $\mathrm{VC}(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.*

Proof. The first statement is trivial as $|\Pi_{\mathcal{H}}(S)| \leq |\mathcal{H}|$. For the second statement, suppose $\mathcal{H}$ shatters S. Then, $2^{|S|} \leq |\Pi_{\mathcal{H}}(S)| \leq |\mathcal{H}|$, implying that $|S| \leq \log_2 |\mathcal{H}|$. Therefore, $\mathrm{VC}(\mathcal{H})$, the maximum sizes of a dataset shatterable by $\mathcal{H}$ is at most $\log_2 |\mathcal{H}|$. $\square$

3 Sauer's Lemma: bounding the growth function

Suppose we have a hypothesis class $\mathcal{H}$ of VC dimension d, and a set of m examples $\{x_1, \ldots, x_m\}$. We already know that when $m \le d$, $\mathcal{S}(\mathcal{H}, n)$ can be as large as 2^m. Can we give a good characterization of $\mathcal{S}(\mathcal{H}, n)$ when $m > d$ (other than the trivial upper bound of $2^m - 1$)? We have the following important combinatorial lemma, discovered independently by several authors (including Sauer, Shelah, Perles, Vapnik and Chervonenkis) in the 70s.

Theorem 2 (Sauer's Lemma)**.** *Suppose $\mathcal{H}$ is a nonempty hypothesis class, and $S = \{x_1, \ldots, x_n\}$ is a set of m unlabeled examples. Then,*

$$|\Pi_{\mathcal{H}}(S)| \le |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}|.$$

Consequently, if $\mathrm{VC}(\mathcal{H}) = d$*, then*

$$\mathcal{S}(\mathcal{H}, m) \le \sum_{i=0}^{d} \binom{m}{i}.$$

(The right hand side is often abbreviated as $\binom{m}{\le d}$.) Here we use the convention that $\mathcal{H}$ always shatters an empty set.

Remark. We see that the growth function, as a function of m, has the following behavior on its upper bound: when $m \le d$, the upper bound grows exponentially with m; however, when $m > d$, the upper bound grows as a polynomial of m, which is substantially slower. We will see in the next section why this type of growth is useful for establishing uniform convergence guarantees.

Proof. We will show the first claim by induction on the size of sample m.

- Base case. If $m = 1$, then there are two subcases to consider: if $\mathcal{H}$ classifies x_1 in both $+1$ and -1 labels, then the left hand size is 2, and the right hand side is also 2. Otherwise, $\mathcal{H}$ classifies x_1 in only one label, then both sides are equal to 1.

- Inductive case. Before proceeding, we need the following important definition. Define a modification of the original hypothesis class $\mathcal{H}$: for every labeling $(l_1, \ldots, l_m)$ in $\Pi_{\mathcal{H}}(S)$, we select one representative classifier h in $\mathcal{H}$ that achieves the labeling; we call the collection of the classifers selected $\mathcal{H}_S$. Note that $|\,\mathcal{H}_S\,| = \Pi_{\mathcal{H}}(S)$. In addition, define $S' = \{x_1, \ldots, x_{m-1}\}$.

  Now, given $\mathcal{H}_S$, let us decompose it to two hypothesis classes, $\mathcal{H}_1$ and $\mathcal{H}_2$, in the following manner. Consider a labeling $(l_1, \ldots, l_{m-1})$ achieved by $\mathcal{H}_2$ on examples S'.

  - If both $(l_1, \ldots, l_{m-1}, +1)$ and $(l_1, \ldots, l_{m-1}, -1)$ are achievable by $\mathcal{H}_S$, then we allocate the pair of classifiers such that one of them goes to $\mathcal{H}_1$, and the other goes to $\mathcal{H}_2$.
  - Otherwise, only one of $(l_1, \ldots, l_{m-1}, +1)$ and $(l_1, \ldots, l_{m-1}, -1)$ is achievable by $\mathcal{H}_S$, then we send the classifier that achieves that labeling to $\mathcal{H}_1$.

  See Tables 1, 2 and 3 for an example.

Classifier	x_1	x_2	x_3	x_4
h_1	−	−	−	−
h_2	−	−	−	+
h_3	−	+	−	+
h_4	+	−	−	−
h_5	+	−	−	+

Table 1: An example with $m = 4$ and $|\,\mathcal{H}_S\,| = 5$. The matrix shows $\mathcal{H}_S$'s labelings on $\{x_1, x_2, x_3, x_4\}$.

Classifier	x_1	x_2	x_3	x_4
h_1	$-$	$-$	$-$	$-$
h_3	$-$	$+$	$-$	$+$
h_5	$+$	$-$	$-$	$+$

Table 2: $\mathcal{H}_2$'s labelings on $\{x_1, x_2, x_3, x_4\}$.

Classifier	x_1	x_2	x_3	x_4
h_2	$-$	$-$	$-$	$+$
h_4	$+$	$-$	$-$	$-$

Table 3: $\mathcal{H}_1$'s labelings on $\{x_1, x_2, x_3, x_4\}$.

By construction, we have the following three simple but important observations:

Claim 1. *1.* $|\mathcal{H}_1| = |\Pi_{\mathcal{H}_1}(S')|$, $|\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')|$.

2. If $T \subset S'$ and $\mathcal{H}_1$ shatters T, then it is also achieved by $\mathcal{H}_S$.

3. If $T \subset S'$ and $\mathcal{H}_2$ shatters T, then $\mathcal{H}_S$ shatters $T \cup \{x_m\}$.

Now, let us upper bound the size of $\mathcal{H}_S$:

$$\begin{aligned}
|\mathcal{H}_S| &= |\mathcal{H}_1| + |\mathcal{H}_2| \\
&= |\Pi_{\mathcal{H}_1}(S')| + |\Pi_{\mathcal{H}_2}(S')| \\
&\leq |\left\{T \subseteq S' : T \text{ shattered by } \mathcal{H}_1\right\}| + |\left\{T \subset S' : T \text{ shattered by } \mathcal{H}_2\right\}| \\
&\leq |\left\{T \subseteq S' : T \text{ shattered by } \mathcal{H}\right\}| + |\left\{T \subseteq S' : T \cup x_m \text{ shattered by } \mathcal{H}\right\}| \\
&= |\left\{T \subseteq S : x_m \notin T, T \text{ shattered by } \mathcal{H}\right\}| + |\{T \subseteq S : x_m \in T, T \text{ shattered by } \mathcal{H}\}| \\
&= |\{T \subseteq S : T \text{ shattered by } \mathcal{H}\}|
\end{aligned}$$

For the second statement, observe that all subsets T shatterable by $\mathcal{H}$ is of size at most d. The right hand size of exactly the number of subsets of size at most d. □

Proof of Claim 3. We show the three items respectively.

1. The first statement is trivial, as by construction, for every labeling $(l_1, \ldots, l_{m-1})$, there is at most one classifier in $\mathcal{H}_1$ (resp. $\mathcal{H}_2$).

2. The second statement is also trivial, as $\mathcal{H}_1$ is a subset of $\mathcal{H}_S$.

3. Suppose some classifier h in $\mathcal{H}_2$ achieves certain labeling $(b_1, \ldots, b_{|T|})$ on T. Suppose h's full labeling on S' is $(l_1, \ldots, l_{m-1})$ (which is consistent with $(b_1, \ldots, b_{|T|})$). Then by construction, both $(l_1, \ldots, l_{m-1}, +1)$ and $(l_1, \ldots, l_{m-1}, -1)$ are achieved by $\mathcal{H}_S$. This implies that $\mathcal{H}_S$ achieves labelings $(b_1, \ldots, b_{|T|}, +1)$ and $(b_1, \ldots, b_{|T|}, +1)$ on $T \cup \{x_m\}$. Therefore, if $\mathcal{H}_2$ achieves all $2^{|T|}$ labelings on T, then $\mathcal{H}_S$ achieves all $2^{|T|+1}$ labelings on $T \cup \{x_m\}$.

Example. Consider the example in Tables 1 and 3. Observe that $\mathcal{H}_2$ shatters $T = \{x_1\}$ with classifiers h_2 and h_4. It can be seen that $\mathcal{H}_S$ also shatters $T \cup \{x_4\} = \{x_1, x_4\}$ with classifiers h_1, h_2, h_4 and h_5.) □

Remark. The growth function bound $\binom{m}{\leq d}$ can further be upper bounded by $(m+1)^d$ or $(\frac{em}{d})^d$.

A A rigorous proof of Theorem 1

Proof. As D is realizable wrt $\mathcal{H}$, there exists a classifier $h_{t^\star}$ that has zero error on D.

Let us consider two critical thresholds t_L and t_R, defined as follows:

$$t_L = \sup\left\{t \in [0,1] : \mathbb{P}(t \leq x \leq t^\star) \geq \epsilon\right\}$$

If $\mathbb{P}(0 \le x \le t^\star) < \epsilon$, then t_L is defined as 0.

$$t_R = \inf\left\{t \in [0,1] : \mathbb{P}(t^\star < x \le t) \ge \epsilon\right\}$$

If $\mathbb{P}(t^\star < x \le 1) < \epsilon$, then t_R is defined as 1.

Suppose for the moment that both t_L and t_R are in $(0,1)$. Our plan is to show the following:

1. With probability $1-\delta$, the returned threshold $\hat{t}$ lies in $[t_L, t_R)$.
2. Wherever $\hat{t}$ lies in $[t_L, t_R)$, $h_{\hat{t}}$ has error at most ϵ.

We show the two items respectively:

1. By Lemma 3, we have that

$$\mathbb{P}(t^\star < x \le t_R) \ge \epsilon, \quad \mathbb{P}(t_L \le x \le t^\star) \ge \epsilon.$$

Now, consider event E_L (resp. E_R) as the one that for all i, none of x_i are in $[t_L, t^\star]$ (resp. $(t^\star, t_R]$). In addition, define $E = E_L \cup E_R$.

Observe that

$$\mathbb{P}(E_L) = \mathbb{P}(\text{for all } i, x_i \notin [t_L, t^\star]) \le (1-\epsilon)^m \le e^{-m\epsilon} \le \delta/2.$$

Similarly, $\mathbb{P}(E_R) \le \delta/2$. By union bound, $\mathbb{P}(E) \le \mathbb{P}(E_L) + \mathbb{P}(E_R) \le \delta$. Therefore, in the event $\bar{E}$ (which happens with probability $1-\delta$), there is an x_i (resp. x_j) in $[t_L, t^\star]$ (resp. $(t^\star, t_R]$). Note that x_i has label -1 and x_j has label $+1$. Thus, the consistency algorithm will return a threshold $\hat{t}$ between $[t_L, t_R)$ (Note that $\hat{t}$ cannot be t_R, as this would misclassify x_j).

2. Suppose $\bar{E}$ happens. We show that the generalization error of the returned threshold classifier $h_{\hat{t}}$ is at most ϵ.

   (a) Suppose $\hat{t} < t^\star$. As argued above, $\hat{t} \ge t_L$. Therefore,

   $$\mathrm{err}(h_{\hat{t}}, D) = \mathbb{P}(\hat{t} < x \le t^\star) \le \mathbb{P}(t_L < x \le t^\star) \le \epsilon,$$

   where the inequality is from item 1 of Lemma 3.

   (b) Suppose $\hat{t} \ge t^\star$. As argued above, $\hat{t} < t_R$. Therefore,

   $$\mathrm{err}(h_{\hat{t}}, D) = \mathbb{P}(t^\star < x \le \hat{t}) \le \mathbb{P}(t^\star < x < t_R) \le \epsilon,$$

   where the inequality is from item 2 of Lemma 3.

Now for the general case, where t_L can be 0 or t_R can be 1. Note that both cannot happen at the same time. Suppose $t_L = 0$, then by the exact same reasoning, we can show that with probability $1-\delta$, $\hat{t}$ is in $[t^\star, t_R)$ or $[0, t^\star]$. In the former case, as have been argued before,

$$\mathrm{err}(h_{\hat{t}}, D) \le \mathbb{P}(t^\star \le x \le \hat{t}) \le \mathbb{P}(t^\star \le x < t_R) \le \epsilon.$$

In the latter case,

$$\mathrm{err}(h_{\hat{t}}, D) = \mathbb{P}(\hat{t} < x \le t^\star) \le \mathbb{P}(0 \le x \le t^\star) \le \epsilon.$$

In summary, with probability $1-\delta$, $\mathrm{err}(h_{\hat{t}}, D) \le \epsilon$. The case of $t_R = 1$ is symmetric and is left as exercise. $\square$

The following lemma crucially uses the continuity property of probability measure, that is, If $A_1 \subset \ldots \subset A_n \subset \ldots$, and $A = \cup_{n=1}^{\infty} A_n$ (abbrev. $A_n \uparrow A$), then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

Lemma 3. *1. Suppose* $\mathbb{P}(0 \le x \le t^\star) \ge \epsilon$*. Consider*

$$t_L \triangleq \sup\left\{t \in [0,1] : \mathbb{P}(t \le x \le t^\star) \ge \epsilon\right\}.$$

Then,

$$\mathbb{P}(t_L \le x \le t^\star) \ge \epsilon,$$
$$\mathbb{P}(t_L < x \le t^\star) \le \epsilon.$$

2. Suppose $\mathbb{P}(t^\star < x \le 1) \ge \epsilon$*. Consider*

$$t_R \triangleq \inf\left\{t \in [0,1] : \mathbb{P}(t^\star < x \le t) \ge \epsilon\right\}.$$

Then,

$$\mathbb{P}(t^\star < x \le t_R) \ge \epsilon,$$
$$\mathbb{P}(t^\star < x < t_R) \le \epsilon.$$

Proof. We only show the first item. The second item is left as an exercise.

First, by the definition of t_L, for all $t < t_L$, $\mathbb{P}(t \le x \le t^\star) \ge \epsilon$. As events $\left\{t_L - \frac{1}{n} \le x \le t^\star\right\} \downarrow \{t_L \le x \le t^\star\}$ as $n \to \infty$, this implies that

$$\mathbb{P}(t_L \le x \le t^\star) = \lim_{n\to\infty} \mathbb{P}(t_L - \frac{1}{n} \le x \le t^\star) \ge \epsilon.$$

Second, by the definition of t_L, for all $t > t_L$, $\mathbb{P}(t \le x \le t^\star) < \epsilon$. As events $\left\{t_L + \frac{1}{n} \le x \le t^\star\right\} \downarrow \{t_L < x \le t^\star\}$ as $n \to \infty$, this implies that

$$\mathbb{P}(t_L < x \le t^\star) = \lim_{n\to\infty} \mathbb{P}(t_L + \frac{1}{n} \le x \le t^\star) \le \epsilon. \qquad \square$$