Machine Learning Theory 2026 Lecture 2

Tim van Erven

Download these slides from elo.mastermath.nl!

- ▶ Review
- ▶ (Agnostic) PAC learning
- ▶ Agnostic PAC-learnability for finite classes
- ▶ Uniform convergence
- ▶ No-Free-Lunch Theorem

Formal Setup Review

$$S = \begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_m \\ \boldsymbol{X}_m \end{pmatrix} \quad \sim \quad \mathcal{D}$$

Risk: $\quad L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h, \boldsymbol{X}, Y)] \qquad$ for $(\boldsymbol{X}, Y) \sim \mathcal{D}$

Empirical Risk: $\quad L_S(h) = \dfrac{1}{m} \sum_{i=1}^{m} \ell(h, \boldsymbol{X}_i, Y_i) \quad$ for $(\boldsymbol{X}_i, Y_i)$ in S

Classification (0/1-loss counts mistakes):

$$\ell(h, \boldsymbol{X}, Y) = \mathbf{1}\{h(\boldsymbol{X}) \neq Y\} = \begin{cases} 0 & \text{if } h(\boldsymbol{X}) = Y \\ 1 & \text{if } h(\boldsymbol{X}) \neq Y \end{cases}$$

Regression (Squared Error):

$$\ell(h, \boldsymbol{X}, Y) = (Y - h(\boldsymbol{X}))^2$$

No Overfitting for (Multiclass) Classification

Realizability assumption: Exists perfect predictor $h^* \in \mathcal{H}$, i.e. $\Pr(h^*(\boldsymbol{X}) = Y) = 1$.

Theorem (First Example of PAC-Learning)

*Assume $\mathcal{H}$ is **finite**, **realizability** holds. Choose any $\delta \in (0, 1)$, $\epsilon > 0$. Then, for all $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$, ERM over $\mathcal{H}$ guarantees*

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta.$$

NB Lower bound on m does not depend on $\mathcal{D}$ or on h^*!

PAC learning: probably approximately correct

(Agnostic) PAC Learning

- ▶ PAC learning (always for binary classification)
- ▶ Agnostic PAC learning for binary classification
- ▶ Agnostic PAC learning in general
- ▶ Improper Agnostic PAC learning in general

Definition: PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm**[1] that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

---

[1] The algorithm's choice h_S is allowed to depend on ϵ and δ as well.

Definition: PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm**[1] that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Sample complexity:
The function $m_{\mathcal{H}}$ such that $m_{\mathcal{H}}(\epsilon, \delta)$ is smallest possible for all ϵ, δ

---

[1]The algorithm's choice h_S is allowed to depend on ϵ and δ as well.

No Overfitting for (Multiclass) Classification

Theorem (First Example of PAC-Learning)

*Assume $\mathcal{H}$ is **finite**, **realizability** holds. Choose any $\delta \in (0, 1)$, $\epsilon > 0$. Then, for all $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$, ERM over $\mathcal{H}$ guarantees*

$$L_{\mathcal{D}}(h_S) \leq \epsilon$$

with probability at least $1 - \delta$.

For binary classification this is equivalent to:

Theorem

*Every **finite** hypothesis class $\mathcal{H}$ is PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

and learning algorithm ERM.

Definition: PAC Learning
(Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0, 1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Agnostic PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **Agnostic PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ ~~for which realizability holds w.r.t. $\mathcal{H}$~~
- and all $\epsilon, \delta \in (0, 1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Agnostic PAC Learning ~~(Binary Classification)~~ (In General)

A hypothesis class $\mathcal{H}$ is **Agnostic PAC-learnable** if there exist

▶ a **function** $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$
▶ and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

▶ distributions $\mathcal{D}$ ~~for which realizability holds w.r.t. $\mathcal{H}$~~
▶ and all $\epsilon, \delta \in (0, 1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Agnostic PAC Learning (In General)

A hypothesis class $\mathcal{H}$ is **Agnostic PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Improper Agnostic PAC Learning (In General)

A hypothesis class $\mathcal{H}$ is **Improperly** **Agnostic PAC-learnable** if there exist

▶ a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
▶ and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

▶ distributions $\mathcal{D}$
▶ and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Agnostic PAC-Learnability for Finite Classes via Uniform Convergence

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

*Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Then every **finite** hypothesis class $\mathcal{H}$ is **agnostically** PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$. Then every finite hypothesis class $\mathcal{H}$ is agnostically PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

▶ Worse dependence on ϵ compared to $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$ for PAC-learnability

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

*Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$. Then every **finite** hypothesis class $\mathcal{H}$ is **agnostically** PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

- Worse dependence on ϵ compared to $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$ for PAC-learnability
- Losses with different range $[a, b]$ can be reduced to $[0, 1]$ range by subtracting a and dividing by $(b - a)$.

Technical Tool: Uniform Convergence

A hypothesis class $\mathcal{H}$ has the **uniform convergence** property if there exists

▶ a **function** $m_{\mathcal{H}}^{\mathbf{UC}} : (0,1)^2 \to \mathbb{N}$

such that for all

▶ distributions $\mathcal{D}$
▶ and all $\epsilon, \delta \in (0,1)$

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta).$$

Uniform Convergence → Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

Lemma

If $\mathcal{H}$ has the **uniform convergence property***, then it is* **agnostic PAC-learnable** *with*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\tfrac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

Uniform Convergence $\rightarrow$ Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

Lemma

*If $\mathcal{H}$ has the **uniform convergence property**, then it is **agnostic PAC-learnable** with*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\tfrac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

▶ We will prove uniform convergence for finite $\mathcal{H}$ and loss range $[0, 1]$
▶ Then the desired agnostic PAC-learnability follows

Proof (Handwritten)

To show, for h_S ERM hypothesis:

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}^{\mathrm{UC}}\left(\tfrac{\epsilon}{2}, \delta\right).$$

Assuming uniform convergence, applied for $\epsilon/2$:

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \tfrac{\epsilon}{2} \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}^{\mathrm{UC}}\left(\tfrac{\epsilon}{2}, \delta\right).$$

Proof: On the event that $|L_{\mathcal{D}}(h) - L_S(h)| \leq \tfrac{\epsilon}{2}$ for all $h \in \mathcal{H}$, we have for all $h' \in \mathcal{H}$

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h') + \frac{\epsilon}{2} \leq L_D(h') + \epsilon.$$

Then take the infimum over h'.

Uniform Convergence for Finite Classes

Lemma (Bounded Loss, Finite Class)

*Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$. Then every **finite** hypothesis class $\mathcal{H}$ has the **uniform convergence property** with*

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

To show:

$$\Pr\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \right) \geq 1 - \delta$$

$$\text{whenever} \quad m \geq \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

Proof (Handwritten)

$$\Pr\big(\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon\big) \overset{?}{\geq} 1 - \delta$$

$$\Pr\big(\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\big) \overset{?}{\leq} \delta$$

$$\Pr\big(\text{exists } h \in \mathcal{H} : |L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\big) \overset{?}{\leq} \delta$$

Part I (union bound):

$$\Pr\big(\text{exists } h \in \mathcal{H} : |L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\big) \leq \sum_{h\in\mathcal{H}} \Pr\big(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\big)$$

Part II (Hoeffding's inequality): Let $Z_i = \ell(h, \boldsymbol{X}_i, Y_i) \in [0, 1]$.

$$\Pr\big(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\big) = \Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z]\right| > \epsilon\right) \overset{\text{Hoeffding}}{\leq} 2e^{-2m\epsilon^2}$$

Proof Continued (Handwritten)

Part I+II:

$$\Pr\left(\text{exists } h \in \mathcal{H} : |L_\mathcal{D}(h) - L_S(h)| > \epsilon\right) \leq \sum_{h \in \mathcal{H}} \Pr\left(|L_\mathcal{D}(h) - L_S(h)| > \epsilon\right)$$

$$\leq |\mathcal{H}| 2 e^{-2m\epsilon^2} \overset{?}{\leq} \delta$$

Yes, for $m \geq \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2}$

Putting Everything Together

Theorem (Bounded Loss, Finite Class)

*Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$. Then every **finite** hypothesis class $\mathcal{H}$ has the uniform convergence property with*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil,$$

*and is therefore **agnostically** PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\tfrac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

No-Free-Lunch Theorem

No-Free-Lunch Theorem
(Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

Let A be **any learning algorithm** *for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

*Let A be **any learning algorithm** for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_\mathcal{D}(f) = 0$.*
2. $\Pr\left(L_\mathcal{D}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

Interpretation:

- $\mathcal{H}_{\text{all}} = $ all functions from $\mathcal{X}$ to $\{-1, +1\}$
- $m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$ for any $\epsilon < 1/8$, $\delta < 1/7$

No-Free-Lunch Theorem
(Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

*Let A be **any learning algorithm** for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

Interpretation:

▶ $\mathcal{H}_{\mathsf{all}}$ = all functions from $\mathcal{X}$ to $\{-1, +1\}$
▶ $m_{\mathcal{H}_{\mathsf{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$ for any $\epsilon < 1/8$, $\delta < 1/7$

Corollary

Suppose $|\mathcal{X}| = \infty$. Then $\mathcal{H}_{all}$ is not PAC-learnable.

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

*Let A be **any learning algorithm** for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

Proof Intuition:

▶ Suppose $\mathcal{D}$ is uniform on $2m$ points in $\mathcal{X}$,
   and $Y = f(X)$ for some unknown function f.

▶ From S we only know $f(X)$ for m observed points.

▶ Without any assumptions about f, learner cannot do better than
   random guessing on m unobserved points.

Pick $C \subseteq \mathcal{X}$ of size $|C| = 2m$.

$T = 2^{2m}$ possible binary classifiers over C:

$$f_1, \ldots, f_T$$

For each $i \in \{1, \ldots, T\}$, let D_i be dist over $C \times \{0,1\}$ such that X is uniformly chosen from C and f_i is a perfect labeling function:

$$D_i(X = x) = \frac{1}{|C|} \quad \text{for } x \in C$$

$$D_i(y = f_i(x) \mid x) = 1$$

Will show: for every alg A

$$\max_i \, \mathbb{E}_{s \sim D_i} \left[L_{D_i}(A(s)) \right] \geq \frac{1}{4} \qquad (1)$$

Let $D = D_i$ for the i that achieves the max.

Will also show that (1) implies

$$Pr\left(L_D(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}. \qquad (2)$$

Proof of (1):

There are $k = (2m)^m$ possible sequences

$$x_1, \ldots, x_m \quad \text{with values in } C$$

Let $S_j = \left(\binom{f_i(x_1)}{x_1}, \ldots, \binom{f_i(x_m)}{x_m}\right)$ be S,

the j-th sequence S_j of $x_1, \ldots, x_m$

labeled according to f_i.

$$\mathbb{E}_{S \sim D_i^m} \left[\sum L_{D_i}(A(S)) \right] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

Instead of worst possible labeling i, it is hard enough to choose i at random:

$$\max_i \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

$$= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

$$\geq \min_j \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

$$L_{D_i}(h) = \frac{1}{2m} \sum_{x \in C} \mathbb{1}\left[h(x) \neq f_i(x) \right]$$

$$\geq \frac{1}{2p} \sum_{\substack{x \in C \\ x \notin S_j}} \mathbb{1}\left[h(x) \neq f_i(x) \right]$$

where $p = |C \setminus S_j| \geq m$

Hence

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{\substack{x \in C \\ x \notin S_j}} \mathbb{1}\left[A(S_j^i)(x) \neq f_i(x) \right]$$

$$= \frac{1}{2p} \sum_{\substack{x \in C \\ x \notin S_j}} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\left[A(S_j^i)(x) \neq f_i(x) \right]$$

$$\geq \frac{1}{2} \min_{\substack{x \in C \\ x \notin S_j}} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\left[A(S_j^i)(x) \neq f_i(x) \right] \qquad (*)$$

Let x^* achieve the minimum.

Partition $f_1, \ldots, f_T$ into pairs $(f_i, f_{i'})$
s.t. $f_i(x) = f_{i'}(x)$ if $x \neq x^*$
$f_i(x) \neq f_{i'}(x)$ if $x = x^*$

Then

$$\mathbb{1}\left[A(S_j^i)(x^*) \neq f_i(x^*)\right] + \mathbb{1}\left[A(S_j^i)(x^*) \neq f_{i'}(x^*)\right] = 1,$$

so

$$\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}\left[A(S_j^i)(x^*) \neq f_i(x^*)\right] = \frac{1}{2}$$

Plugging into (*) gives

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{4}$$

Tracing back chain of inequalities proves (1).

For (2) we will show that,

for $z = L_{D_i}(A(S)) \in [0,1]$.

$\mathbb{E}[z] \geq \frac{1}{4}$ implies $\Pr(z \geq \frac{1}{8}) \geq \frac{1}{7}$

Lemma B.1 in book: for z taking values in $[0,1]$

$$\Pr(z \geq a) \geq \frac{\mathbb{E}[z] - a}{1-a}$$

$\mathbb{E}[z] \geq \frac{1}{4}$, $a = \frac{1}{8}$ $\Rightarrow$ $\Pr(z \geq \frac{1}{8}) \geq \frac{\frac{1}{4} - \frac{1}{8}}{7/8} = \frac{1}{7}$.