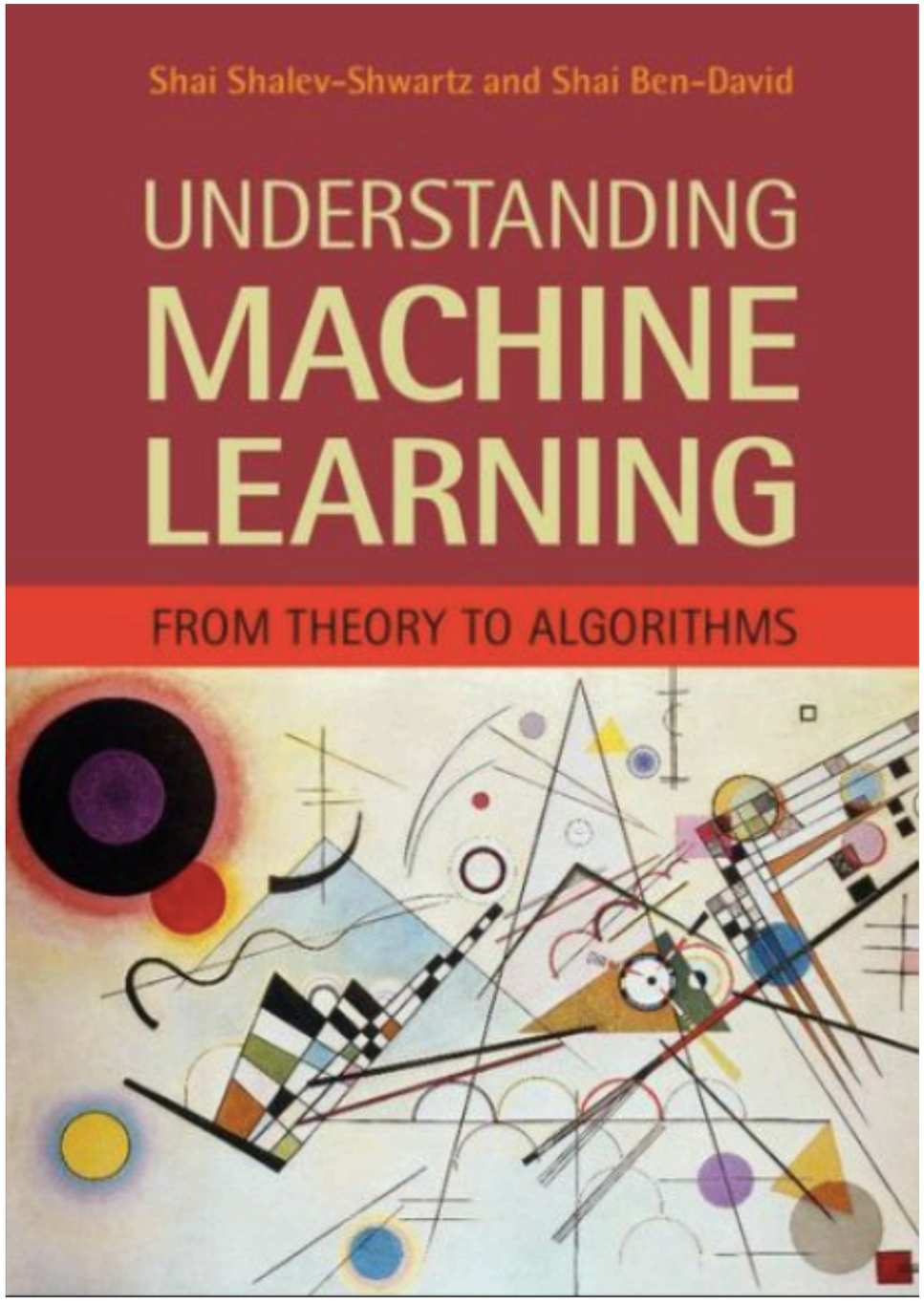
بسم الله الرحمن الرحیم

یادگیری برخط

جلسه ششم: نظریه یادگیری ـ تعریف‌ها

منبع

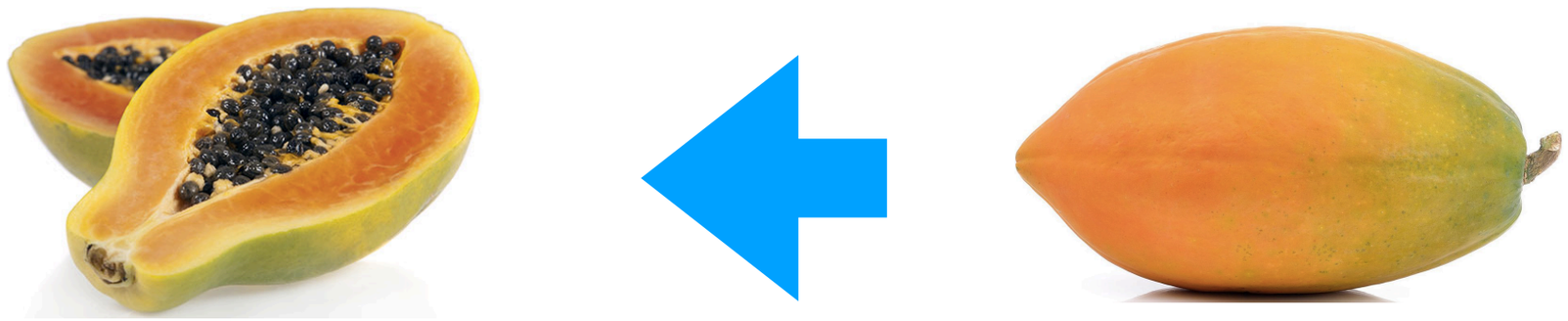

Shai Shalev-Shwartz and Shai Ben-David
UNDERSTANDING
MACHINE
LEARNING
FROM THEORY TO ALGORITHMS

یادگیری «پاپای شیرین»

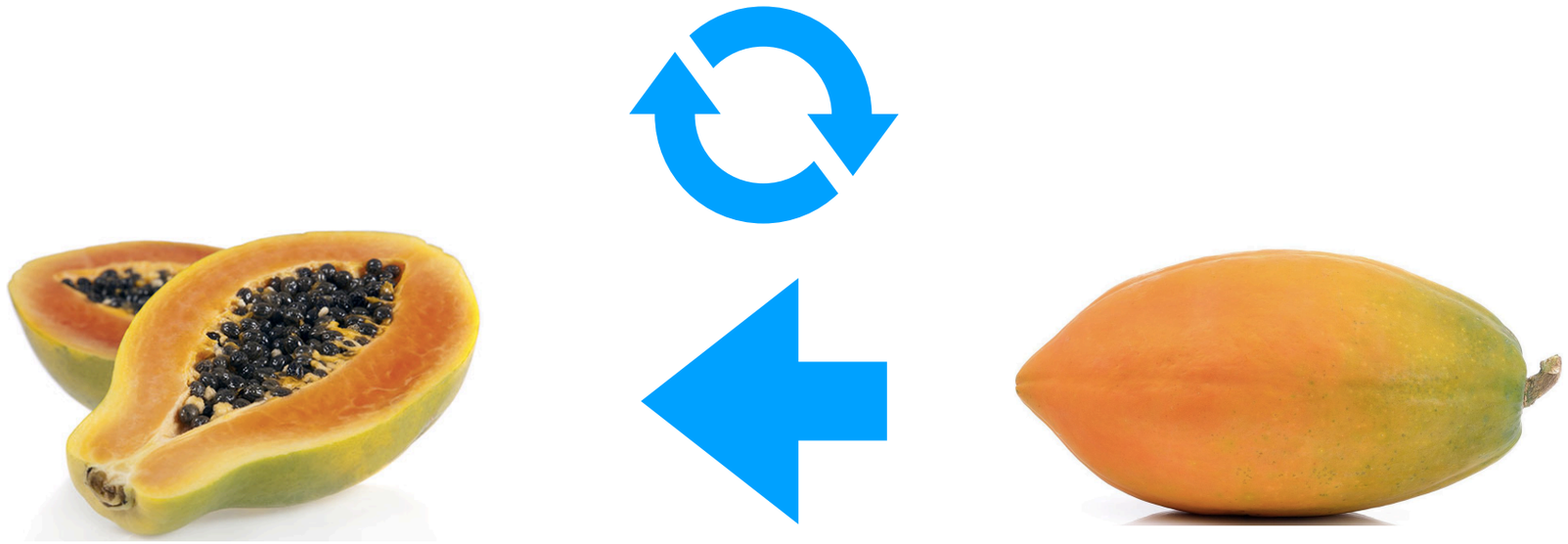
یادگیری «پاپایای شیرین»

یادگیری «پاپایای شیرین»

یادگیری «پاپایای شیرین»



یادگیری «پاپایای شیرین»



پیش‌فرض: پاپایاهای شیرین، محدوده حجم و محدوده وزن خاصی دارند

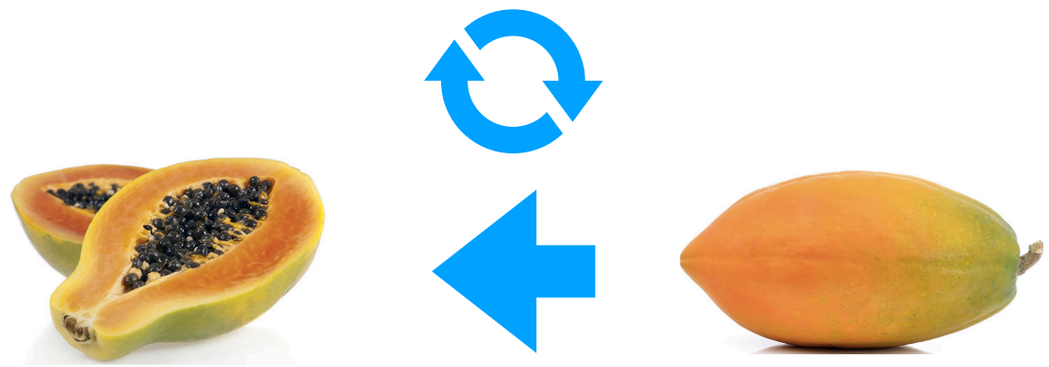
پیش‌فرض: پاپایاهای شیرین، محدوده حجم و محدوده وزن خاصی دارند

اما کدام محدوده حجم و وزن؟

پیش‌فرض: پاپایاهای شیرین، محدوده حجم و محدوده وزن خاصی دارند

اما کدام محدوده حجم و وزن؟

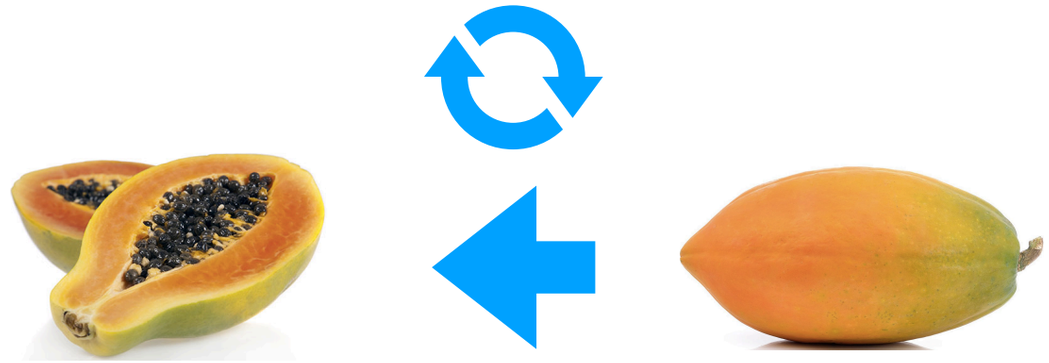
تجربیات



پیش‌فرض: پاپایاهای شیرین، محدوده حجم و محدوده وزن خاصی دارند

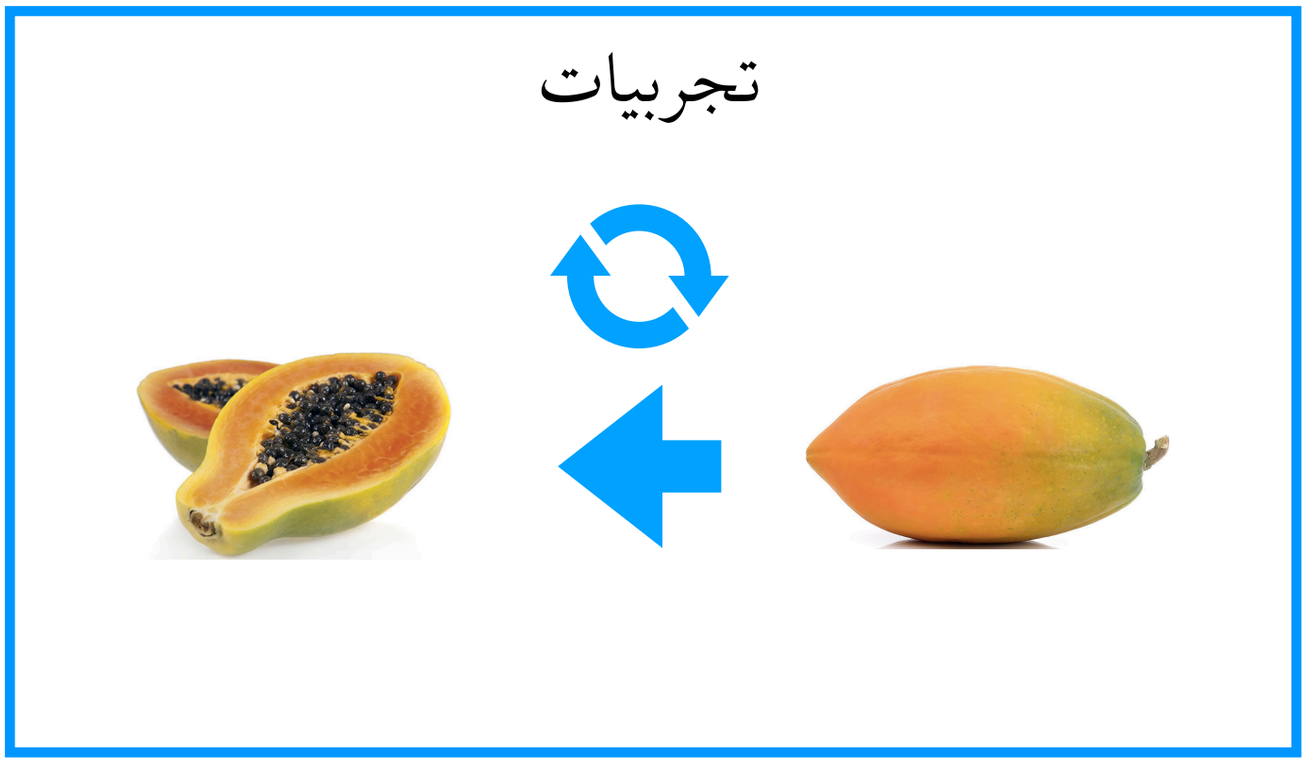
اما کدام محدوده حجم و وزن؟

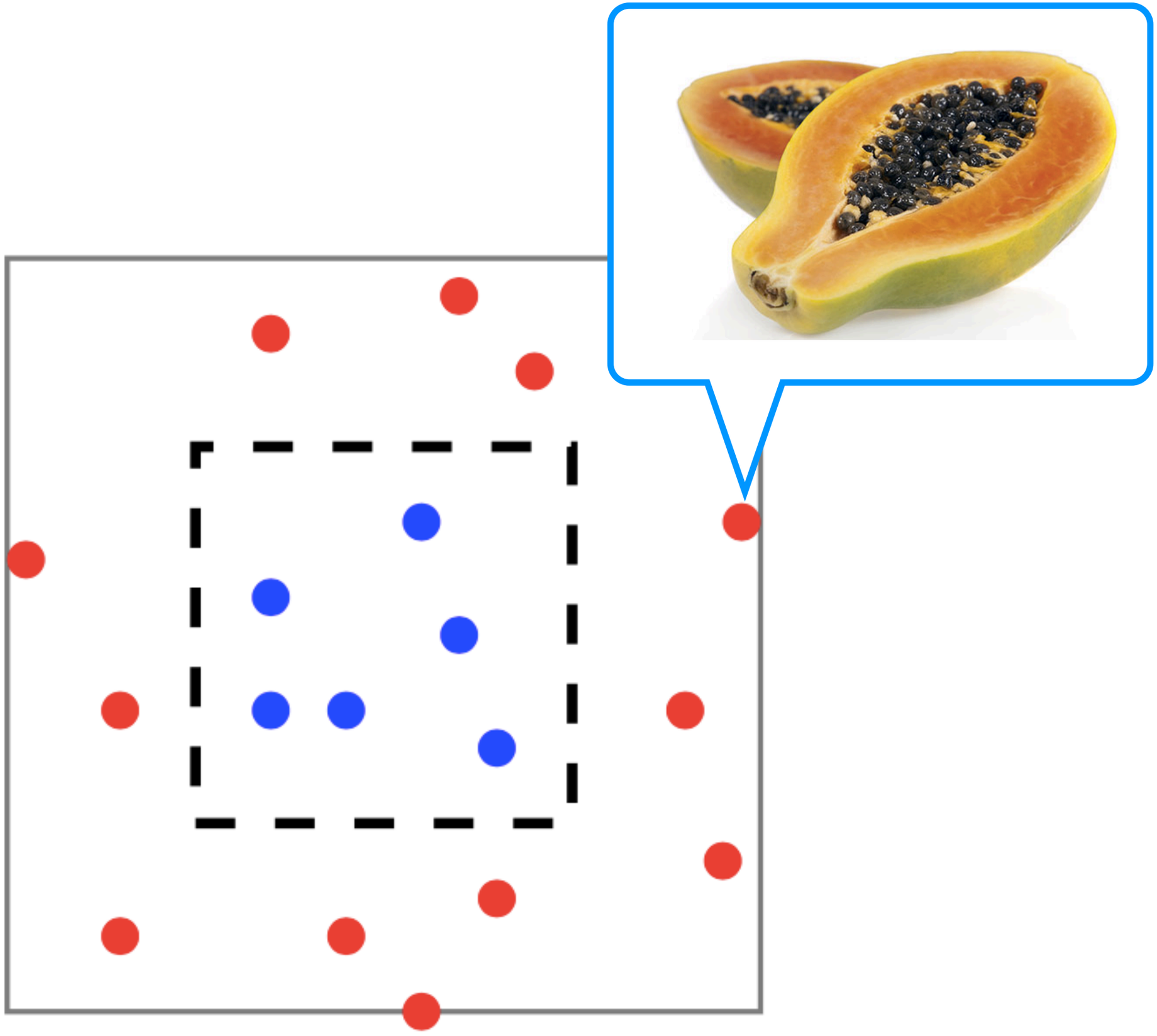
تجربیات

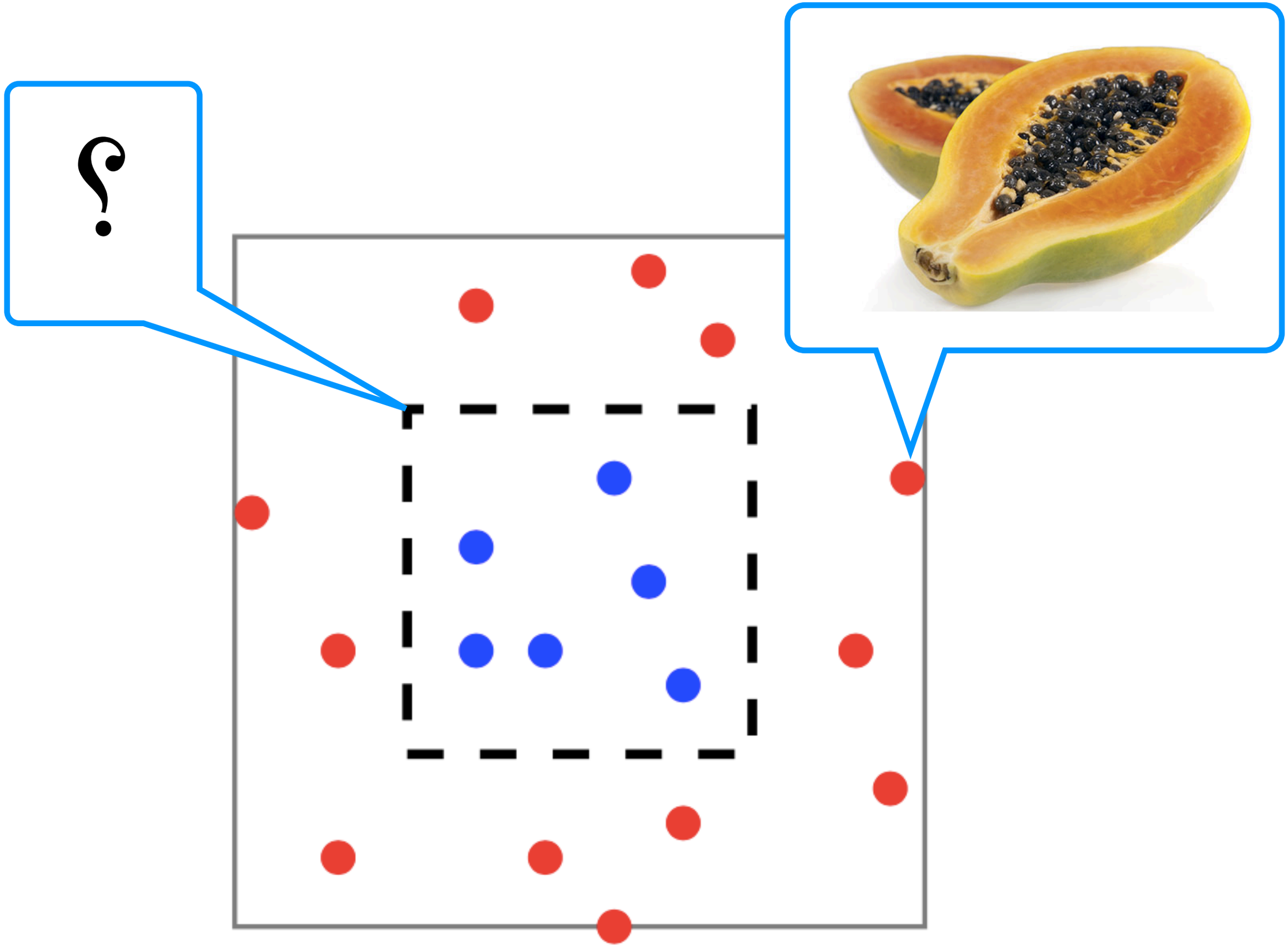


پیش‌فرض: پاپایاهای شیرین، محدوده حجم و محدوده وزن خاصی دارند

اما کدام محدوده حجم و وزن؟







نمادگذاری

- دامنه ($\mathcal{X}$)،
- برچسب ($\mathcal{Y}$)
- توزیع ($\mathcal{D}$) روی دامنه
- تابع حقیقت:
$f : \mathcal{X} \to \mathcal{Y}$

نمادگذاری

- دامنه ($\mathcal{X}$)،
- برچسب ($\mathcal{Y}$)
- توزیع ($\mathcal{D}$) روی دامنه
- تابع حقیقت:

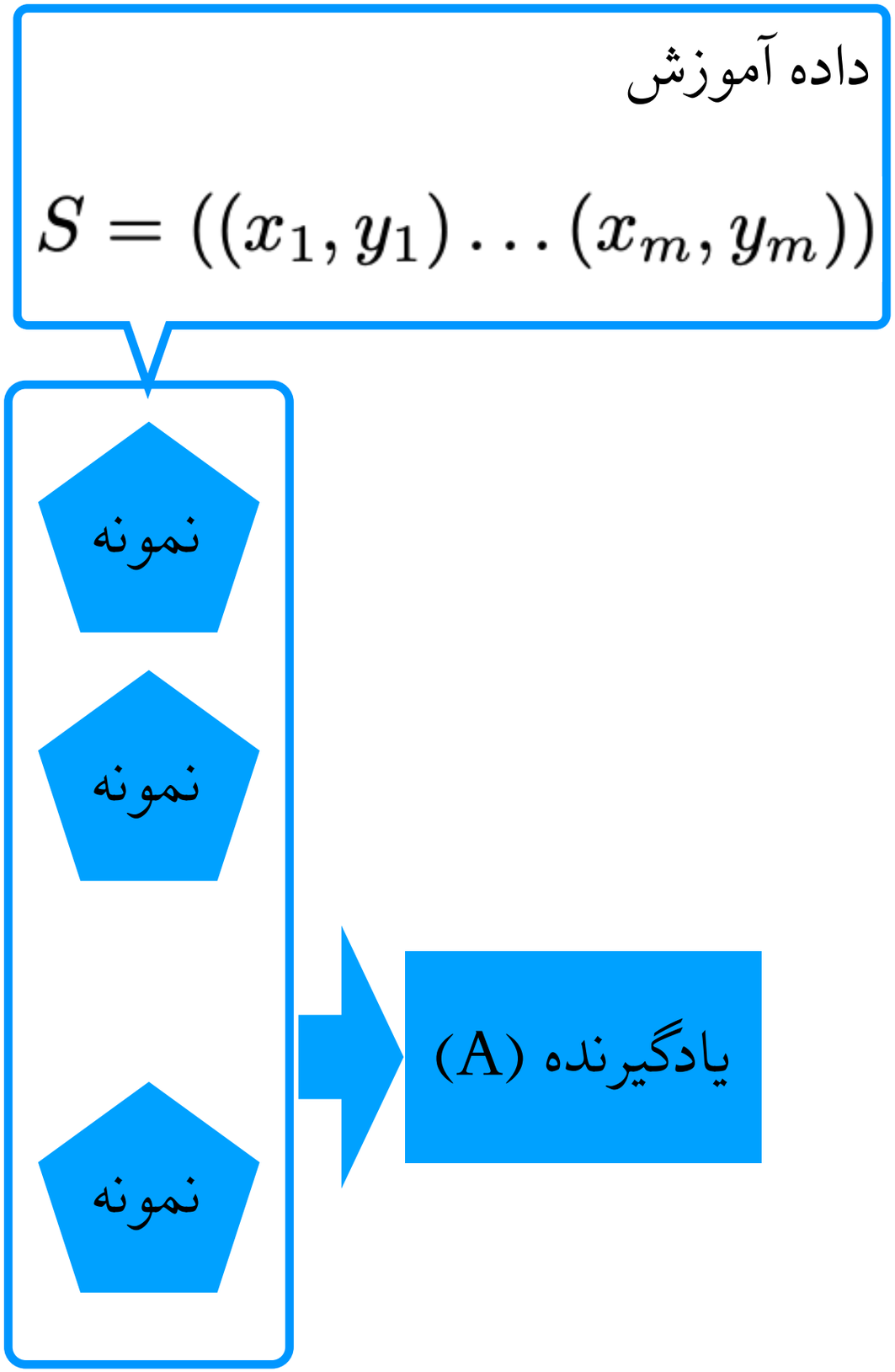
$$f : \mathcal{X} \to \mathcal{Y}$$

یادگیرنده (A)

نمادگذاری

- دامنه ($\mathcal{X}$)،
- برچسب ($\mathcal{Y}$)
- توزیع ($\mathcal{D}$) روی دامنه
- تابع حقیقت:

$$f : \mathcal{X} \to \mathcal{Y}$$



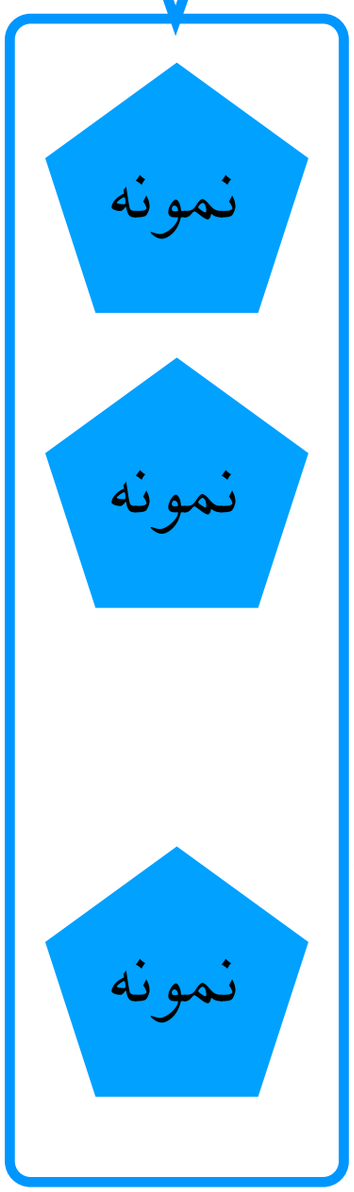
نمادگذاری

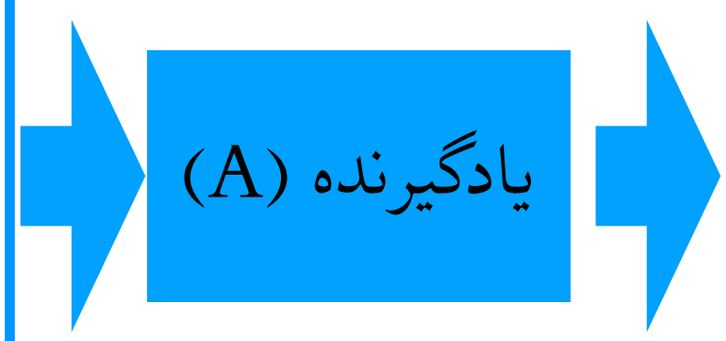
- دامنه ($\mathcal{X}$)،
- برچسب ($\mathcal{Y}$)
- توزیع ($\mathcal{D}$) روی دامنه
- تابع حقیقت:

$$f : \mathcal{X} \to \mathcal{Y}$$

داده آموزش

$$S = ((x_1, y_1) \dots (x_m, y_m))$$





$$h : \mathcal{X} \to \mathcal{Y}$$

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)]$$

تعریف: یادگیری = تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$

$$\Pr_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری = تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$

$$\Pr_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

وابستگی به S؟

تعریف: یادگیری = تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری = تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

احتمالا

تعریف: یادگیری = تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

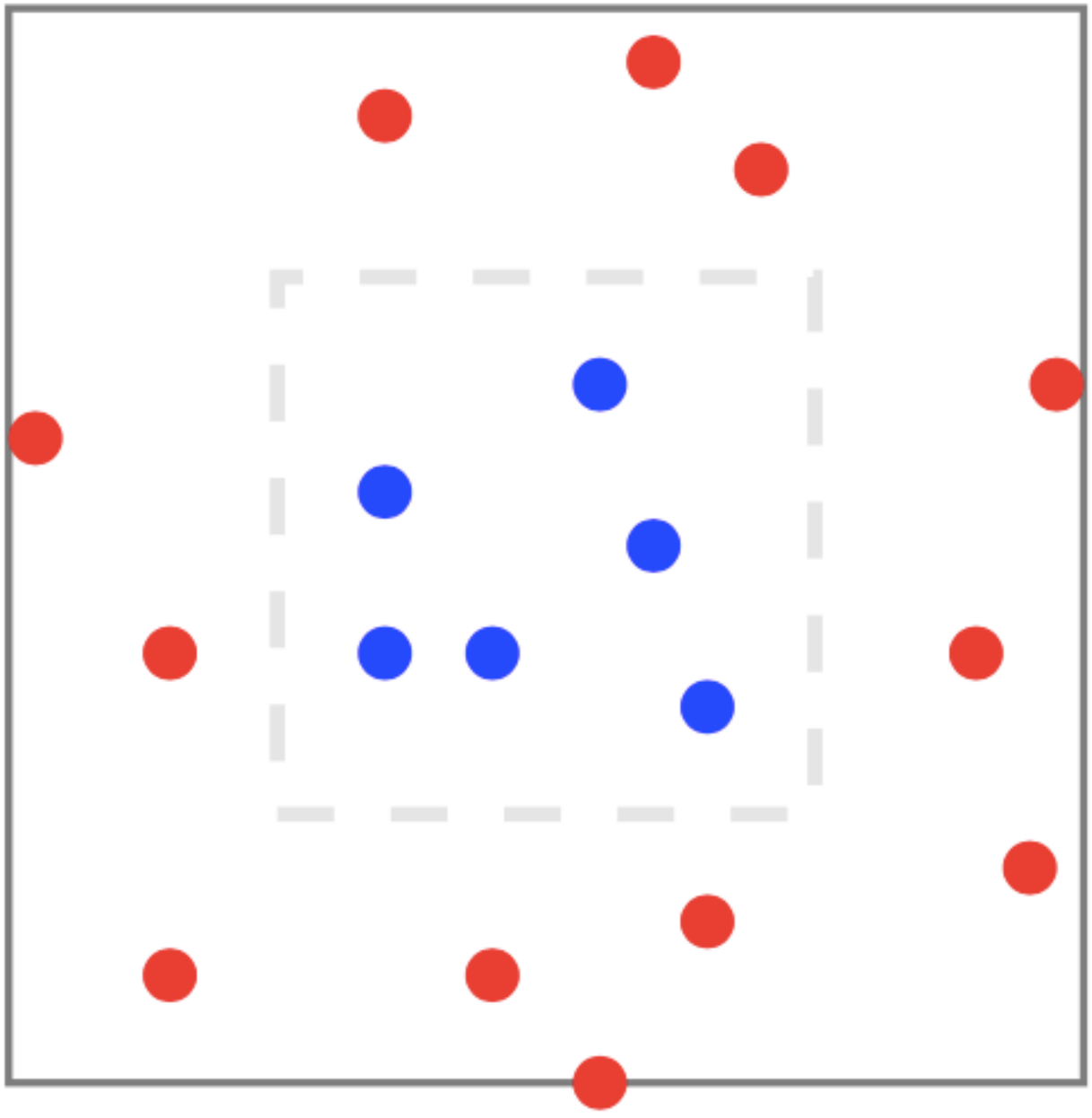
تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

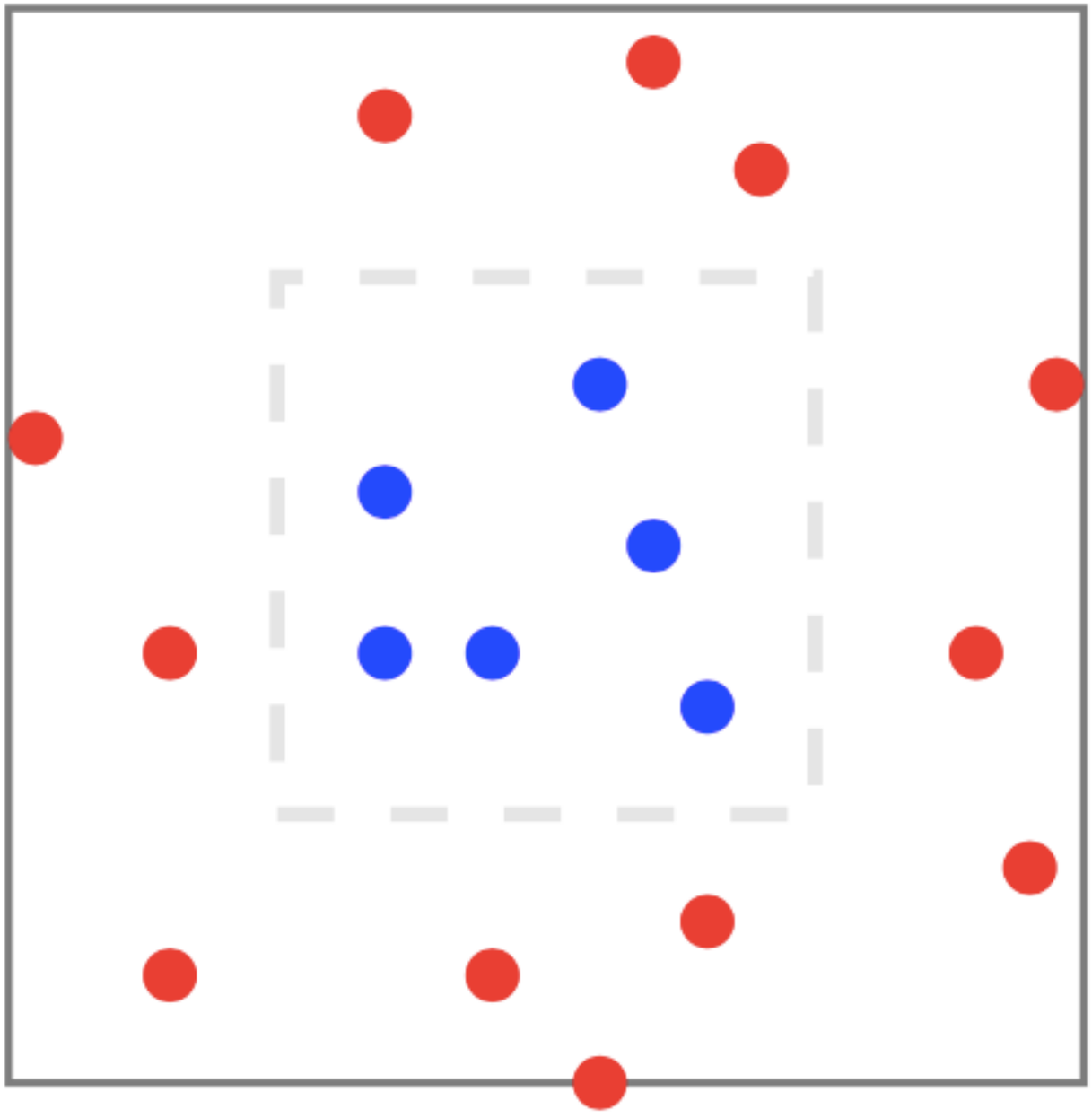
$$\mathbb{P}\mathrm{r}_{S\sim\mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

خروجی A: h؟

اهمیت دانش پیشین:



اهمیت دانش پیشین:



$$h \in \mathscr{H}$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ با توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ $(f \in \mathscr{H})$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ با توزیع $\mathscr{D}$ روی $\mathscr{X}$ با تابع $f : \mathscr{X} \to \mathscr{Y}$ $(f \in \mathscr{H})$ را یاد می‌گیرد، اگر به ازای هر $\epsilon > 0$ و هر $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

الگوریتم یادگیری $\mathscr{H}$ و $\mathscr{D}$ و f

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X}$ و تابع $f : \mathscr{X} \to \mathscr{Y}$ $(f \in \mathscr{H})$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

الگوریتم یادگیری $\mathscr{H}$ و $\mathscr{D}$ و f

تعریف: یادگیری احتمالا تقریبا درست

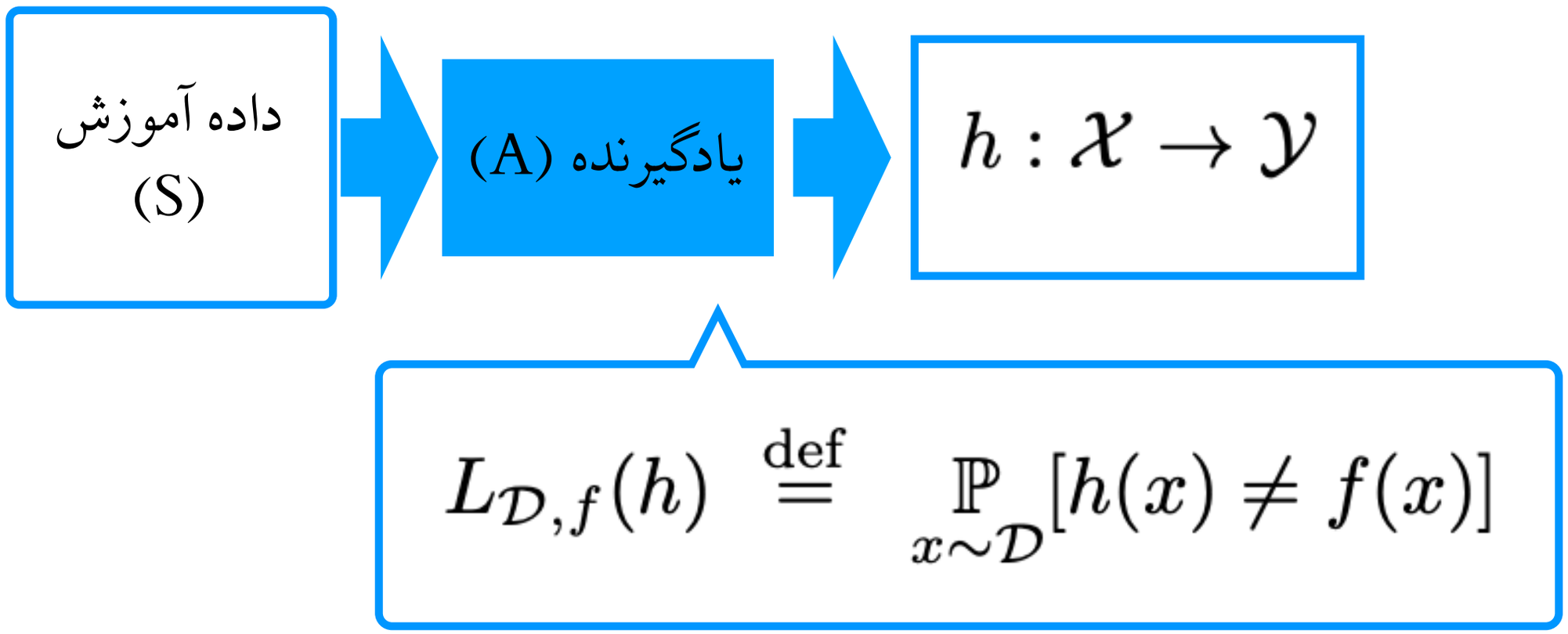
الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X}$ و تابع $f : \mathscr{X} \to \mathscr{Y}$ $(f \in \mathscr{H})$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

الگوریتم یادگیری $\mathscr{H}$ و $\mathscr{D}$ و f

الگوریتم یادگیری $\mathscr{H}$

یادگیری توابع پیوسته



یادگیری توابع پیوسته

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$$

$$L_{D,f,l}(h) := \mathbb{E}_{x\sim\mathcal{D}}[l(h,f,x)]$$

یادگیری توابع پیوسته

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$$

$$l(h, f, x) := 1[h(x) = f(x)]$$

$$L_{D,f,l}(h) := \mathbb{E}_{x\sim\mathcal{D}}[l(h, f, x)]$$

یادگیری توابع پیوسته

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$l(h, f, x) := 1[h(x) = f(x)]$$

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)]$$

$$l(h, f, x) := (h(x) - f(x))^2$$

$$L_{D,f,l}(h) := \mathbb{E}_{x \sim \mathcal{D}}[l(h, f, x)]$$

حذف تابع حقیقت f

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{D,f,l}(h) := \mathbb{E}_{x \sim \mathcal{D}}[l(h, f, x)]$$

حذف تابع حقیقت f

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{D,f,l}(h) := \mathbb{E}_{x \sim \mathcal{D}}[l(h, f, x)]$$

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[l(h, (x, y))]$$

حذف تابع حقیقت f

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{D,f,l}(h) := \mathbb{E}_{x\sim\mathcal{D}}[l(h,f,x)]$$

فرض: D سازگار با H است

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h,(x,y))]$$

حذف تابع حقیقت f

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{D,f,l}(h) := \mathbb{E}_{x\sim\mathcal{D}}[l(h,f,x)]$$

فرض: D سازگار با H است

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h,(x,y))]$$

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$$

حذف تابع حقیقت f

داده آموزش (S) → یادگیرنده (A) → $h : \mathcal{X} \to \mathcal{Y}$

$$L_{D,f,l}(h) := \mathbb{E}_{x\sim\mathcal{D}}[l(h,f,x)]$$

فرض: D سازگار با H است

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h,(x,y))]$$

$$L_{\mathcal{D},l}(h) := \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$$

$$L_{\mathcal{D}}(h) := \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},l}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

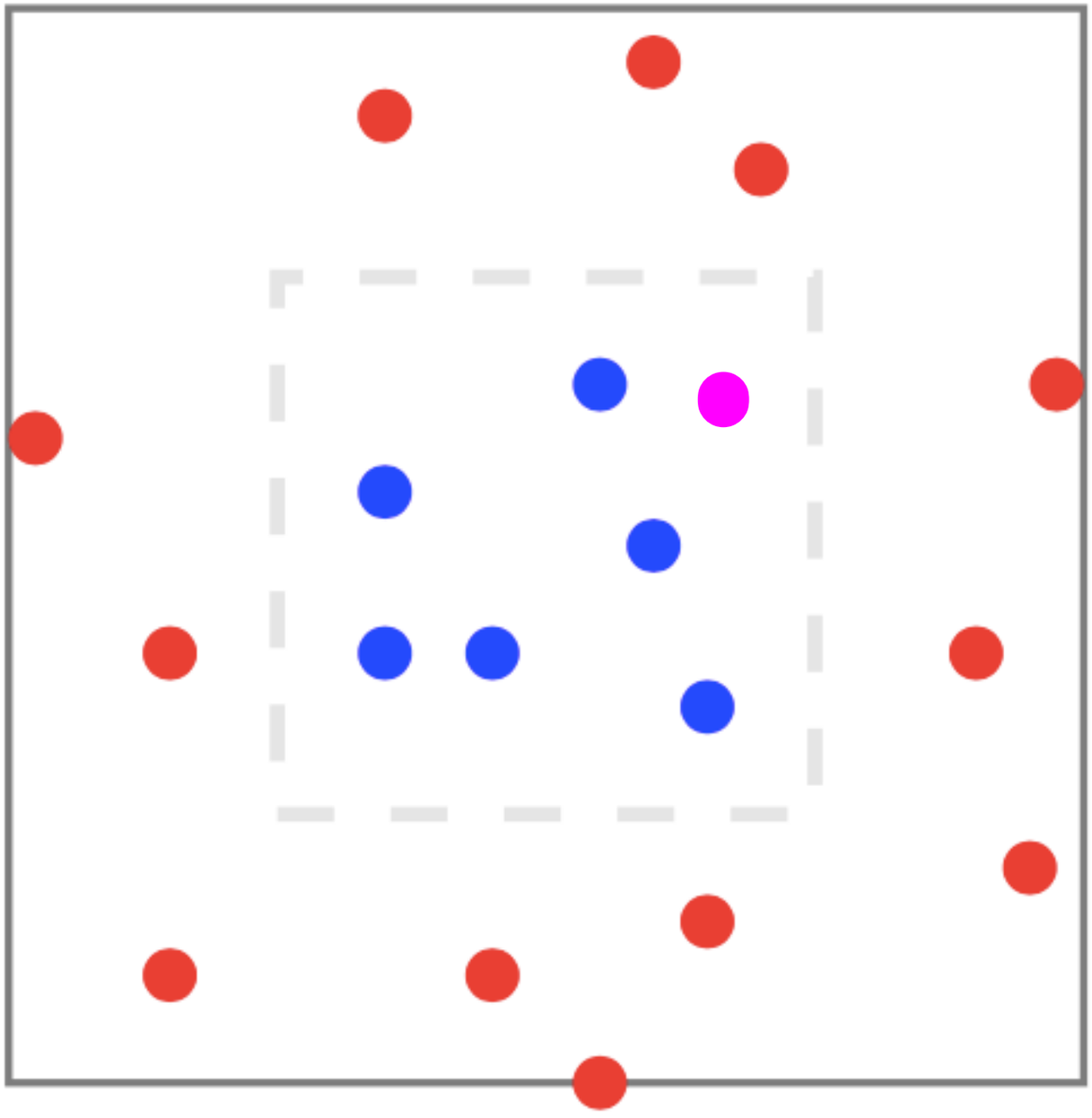
$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},l}(A(S)) \leq \epsilon] \geq 1 - \delta$$

ندانم‌گرایی:

۱) عدم هم‌خوانی

۲) عدم هم‌خوانی با خودش!

مثال: ندانم‌گرایی عدم هم‌خوانی با خودش



تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

یک الگوریتم یادگیری

الگوریتم ERM:

- ورودی: $S = (z_1, \ldots, z_m)$
- خروجی: $h \in \mathscr{H}$

الگوریتم ERM:

- ورودی: $S = (z_1, \ldots, z_m)$
- خروجی: $h \in \mathscr{H}$
- ایده: کدام h روی نمونه‌های آموزش خوب کار می‌کند؟

الگوریتم ERM:

- ورودی: $S = (z_1, \ldots, z_m)$
- خروجی: $h \in \mathscr{H}$
- ایده: کدام h روی نمونه‌های آموزش خوب کار می‌کند؟

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

الگوریتم ERM:

- ورودی: $S = (z_1, \ldots, z_m)$
- خروجی: $h \in \mathscr{H}$
- ایده: کدام h روی نمونه‌های آموزش خوب کار می‌کند؟

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

الگوریتم (ERM):

$$h_S \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- اگر $L_D(h)$ها را درست تخمین بزنیم:

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- اگر $L_D(h)$ها را درست تخمین بزنیم:
- اگر $L_D(h)$ها را با خطای $\epsilon/2$ تخمین بزنیم:
- فرضیه خروجی h و فرضیه بهینه h':

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- اگر $L_D(h)$ها را درست تخمین بزنیم:
- اگر $L_D(h)$ها را با خطای $\epsilon/2$ تخمین بزنیم:
- فرضیه خروجی h و فرضیه بهینه h':

$$L_D(h) - \epsilon/2 \leq L_S(h) \leq L_S(h') \leq L_D(h') + \epsilon/2$$

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- اگر $L_D(h)$ها را درست تخمین بزنیم:
- اگر $L_D(h)$ها را با خطای $\epsilon/2$ تخمین بزنیم:
- فرضیه خروجی h و فرضیه بهینه h':

$$L_D(h) \leq L_S(h) + \epsilon/2$$

$$L_D(h) - \epsilon/2 \leq L_S(h) \leq L_S(h') \leq L_D(h') + \epsilon/2$$

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- اگر $L_D(h)$ها را درست تخمین بزنیم:
- اگر $L_D(h)$ها را با خطای $\epsilon/2$ تخمین بزنیم:
- فرضیه خروجی h و فرضیه بهینه h':

$$L_D(h) \leq L_S(h) + \epsilon/2$$

$$L_S(h') \leq L_D(h') + \epsilon/2$$

$$L_D(h) - \epsilon/2 \leq L_S(h) \leq L_S(h') \leq L_D(h') + \epsilon/2$$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
- کافی است $\delta \geq |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
- کافی است $\delta \geq |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
- $\ln\delta \geq \ln|\mathscr{H}| + \ln 2 - 2m(\epsilon/2)^2$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
- کافی است $\delta \geq |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
- $\ln\delta \geq \ln|\mathscr{H}| + \ln 2 - 2m(\epsilon/2)^2$
- $2m(\epsilon/2)^2 > \ln 1/\delta + \log|\mathscr{H}| + 1$

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}\left[\overline{X}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|2e^{-2m(\epsilon/2)^2}$
- کافی است $\delta \geq |\mathscr{H}|2e^{-2m(\epsilon/2)^2}$
- $\ln\delta \geq \ln|\mathscr{H}| + \ln 2 - 2m(\epsilon/2)^2$
- $2m(\epsilon/2)^2 > \ln 1/\delta + \log|\mathscr{H}| + 1$
- $m = \Omega(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$

چرا منظم‌سازی؟

بسم الله الرحمن الرحیم

یادگیری برخط

جلسه هفتم: نظریه یادگیری ـ تعریف‌ها (۲)

مروری بر تعریف یادگیری
احتمالا تقریبا درست

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m}[L_{\mathscr{D}}(A(S)) \leq \epsilon] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m}[L_{\mathscr{D}}(A(S)) \leq \epsilon] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

در صورت وجود الگوریتم A، می‌گوییم مجموعه فرضیه‌های $\mathscr{H}$ PAC_یادگیری‌پذیر است.

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$
و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$
و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

در صورت وجود الگوریتم A، می‌گوییم مجموعه
فرضیه‌های $\mathscr{H}$ PAC ندانم‌گرا_ یادگیری‌پذیر است.

الگوریتم ERM:

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$$

الگوریتم (ERM):

$$h_S \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h)$$

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی $\mathscr{H}$ را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $< |\mathscr{H}|\,2e^{-2m(\epsilon/2)^2}$
  - کافی است $m = \Omega(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$
- فرضیه خروجی h و فرضیه بهینه h':

$$L_D(h) - \epsilon/2 \leq L_S(h) \leq L_S(h') \leq L_D(h') + \epsilon/2$$

$$L_D(h) \leq L_D(h') + \epsilon$$

؟

قضیه: الگوریتم ERM مجموعه فرضیه‌های متناهی $\mathscr{H}$ را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

- به ازای هر فرضیه h:

$$P[\,|L_S(h) - L_D(h)| \geq \epsilon/2] < 2e^{-2m(\epsilon/2)^2}$$

- احتمال اینکه حداقل یکی از فرضیه‌ها بد باشد $|\mathscr{H}|\,2e^{-2m(\epsilon/2)^2} >$
  - کافی است $m = \Omega(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$
- فرضیه خروجی h و فرضیه بهینه h':

$$L_D(h) - \epsilon/2 \leq L_S(h) \leq L_S(h') \leq L_D(h') + \epsilon/2$$

$$L_D(h) \leq L_D(h') + \epsilon$$

تعریف: مجموعه فرضیه‌های $\mathscr{H}$ خاصیت **هم‌گرایی یکنواخت** را دارد، اگر به ازای هر توزیع $\mathscr{D}$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که

$$\mathbb{P}_{S\sim\mathscr{D}^m}\left[\forall h \in \mathscr{H} : |L_{\mathscr{D}}(h) - L_S(h)| \leq \epsilon\right] \geq 1 - \delta$$

تعریف: مجموعه فرضیه‌های $\mathscr{H}$ خاصیت **هم‌گرایی یکنواخت** را دارد، اگر به ازای هر توزیع $\mathscr{D}$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که

$$\mathbb{P}_{S\sim\mathscr{D}^m}\left[\forall h \in \mathscr{H} : |L_{\mathscr{D}}(h) - L_S(h)| \leq \epsilon\right] \geq 1 - \delta$$

قضیه: الگوریتم ERM مجموعه فرضیه‌های $\mathscr{H}$ با خاصیت هم‌گرایی یکنواخت را با تابع زیان $l(h, z) \in [0,1]$ یاد می‌گیرد.

مثال: مجموعه فرضیه‌های نامتناهی، اما یادگیری‌پذیر

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ که $h_a(x) = \mathbb{1}_{[x<a]}$ یادگیری‌پذیر است .

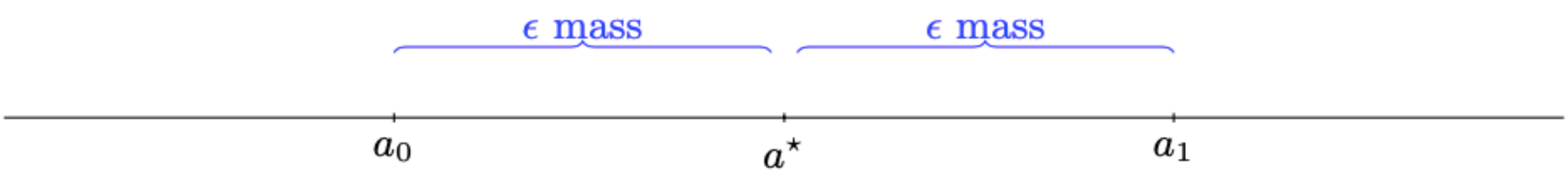
مثال: مجموعه فرضیه‌های نامتناهی، اما یادگیری‌پذیر

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ که $h_a(x) = \mathbb{1}_{[x<a]}$ یادگیری‌پذیر است .
- اثبات:
  - a^* جواب واقعی و h_S جواب ما.

مثال: مجموعه فرضیه‌های نامتناهی، اما یادگیری‌پذیر

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ که $h_a(x) = \mathbb{1}_{[x<a]}$ یادگیری‌پذیر است .
- اثبات:
  - a^* جواب واقعی و h_S جواب ما.

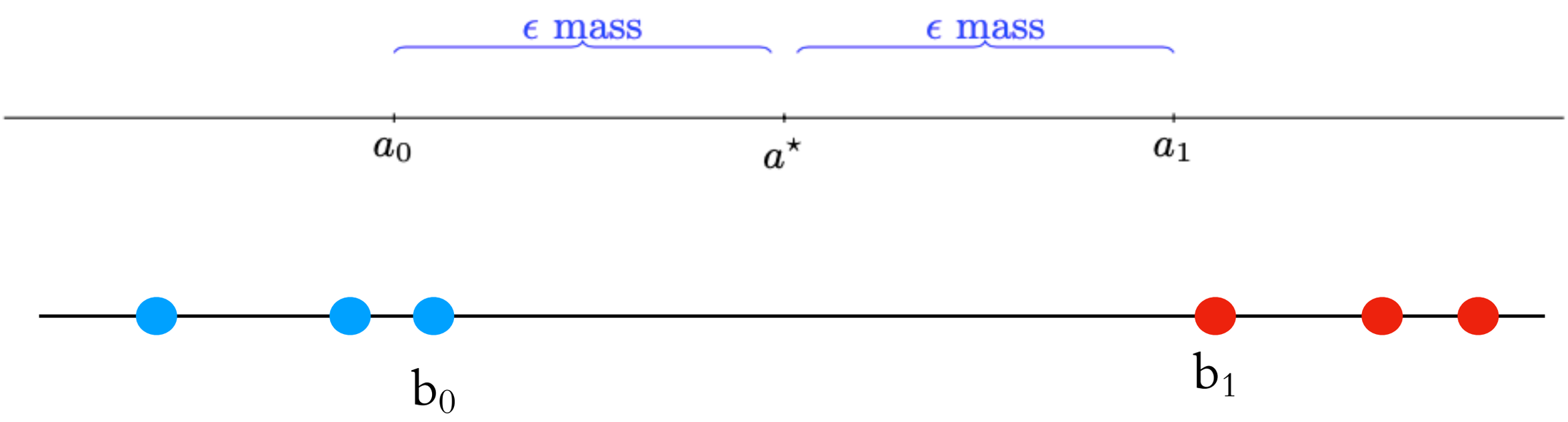
$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



مثال: مجموعه فرضیه‌های نامتناهی، اما یادگیری‌پذیر

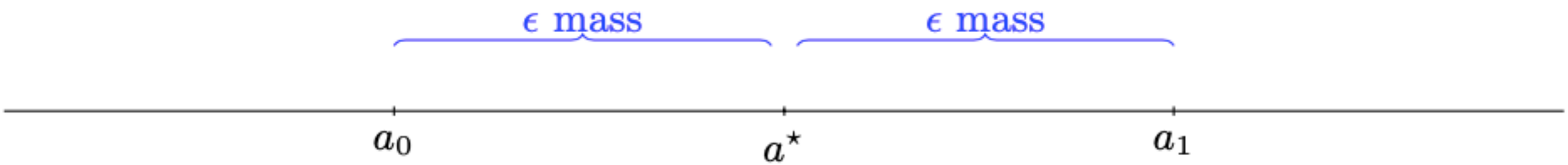
- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ که $h_a(x) = \mathbb{1}_{[x<a]}$ یادگیری‌پذیر است .
- اثبات:
  - a^* جواب واقعی و h_S جواب ما.

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



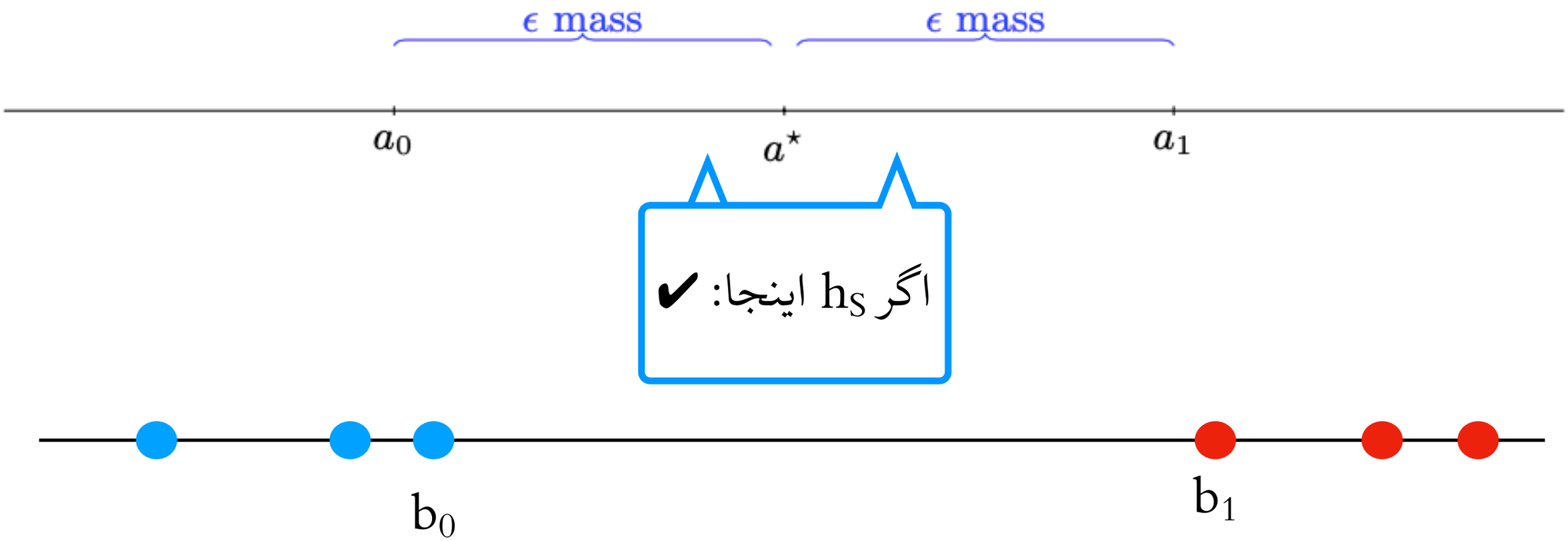
- a^* جواب واقعی و h_S جواب ما.

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



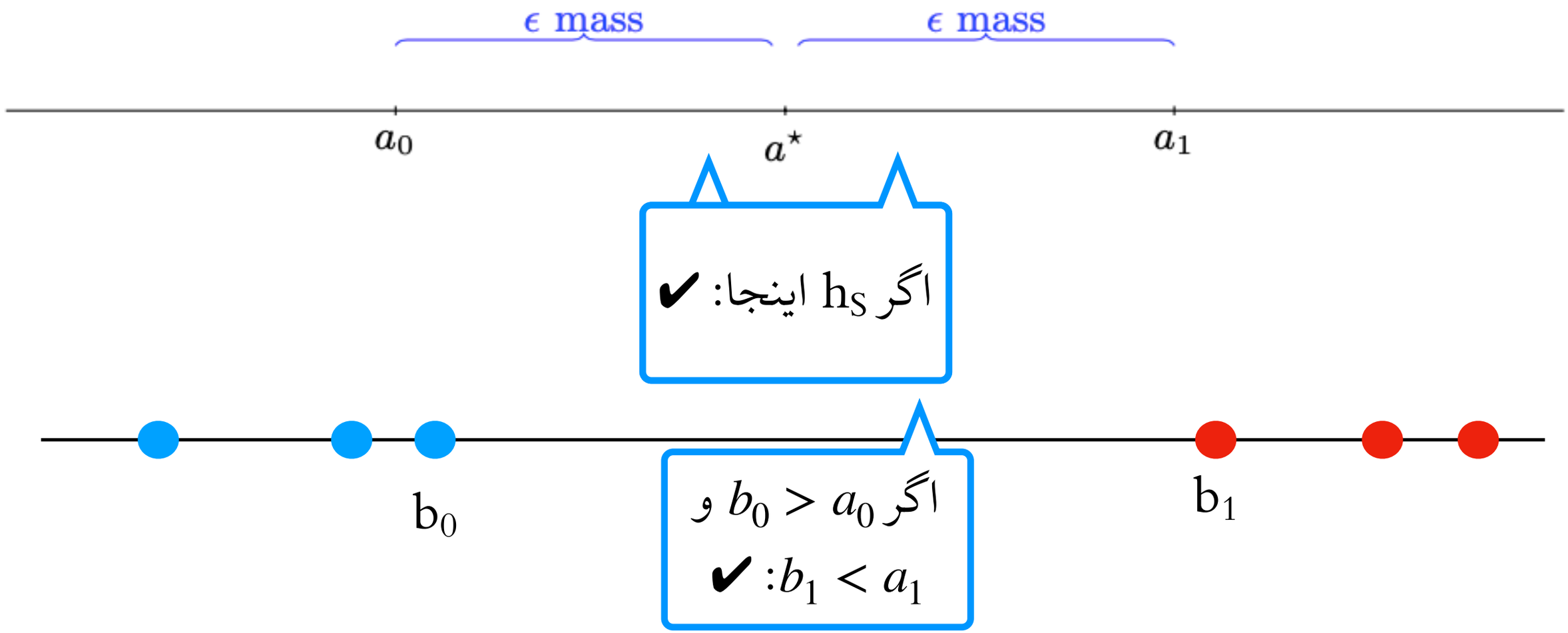
- a^* جواب واقعی و h_S جواب ما.

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$

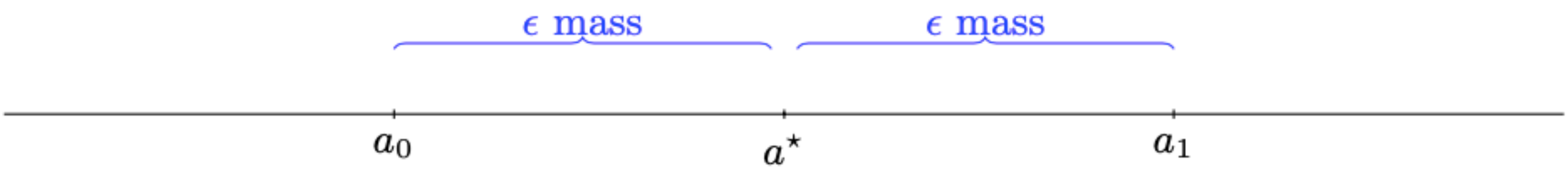


- a^* جواب واقعی و h_S جواب ما.

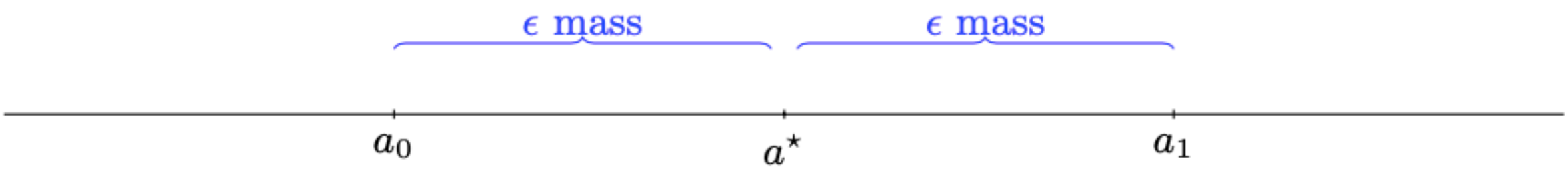
$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



$$\mathop{\mathbb{P}}_{x \sim \mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x \sim \mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$

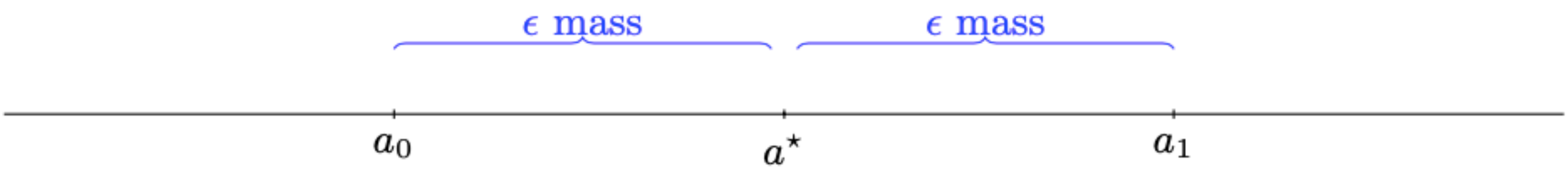


$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



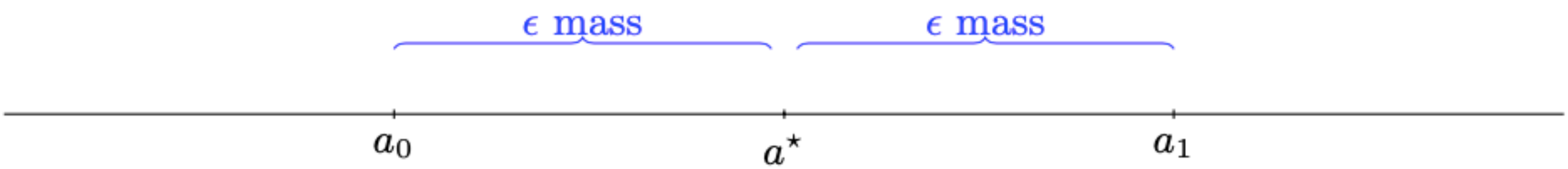
$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0]$$

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



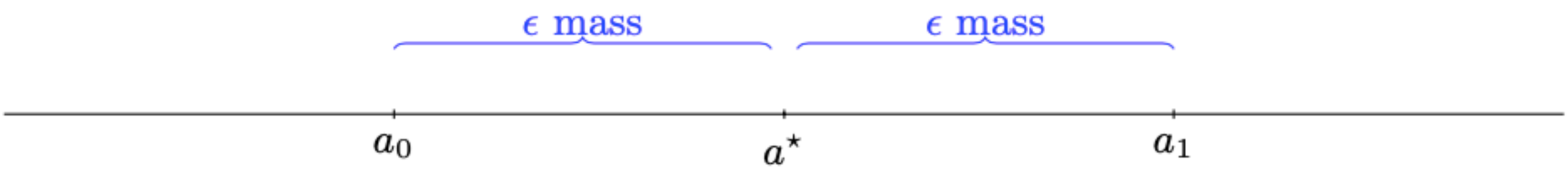
$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] \;=\; \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\; x \notin (a_0, a^\star)]$$

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



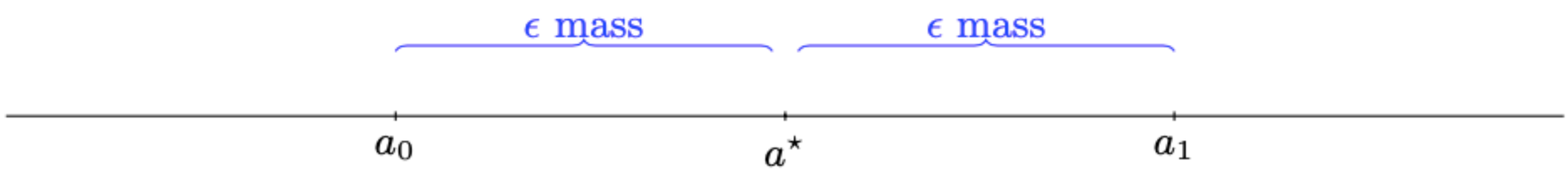
$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] \;=\; \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\; x \notin (a_0, a^\star)] \;= (1-\epsilon)^m$$

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] \;=\; \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\; x \notin (a_0, a^\star)] \;= (1-\epsilon)^m \;\leq e^{-\epsilon\, m}$$

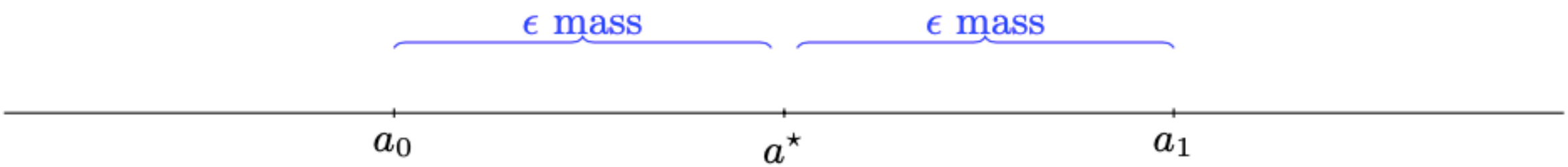
$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\ x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m}$$

$$m > \log(2/\delta)/\epsilon$$

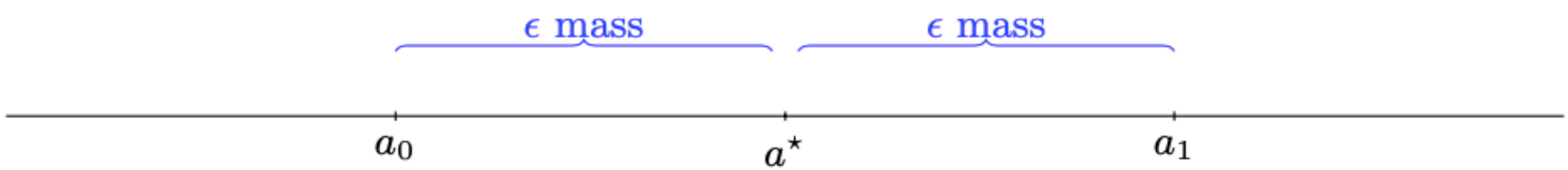
$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\ x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta/2$$

$$m > \log(2/\delta)/\epsilon$$

$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$

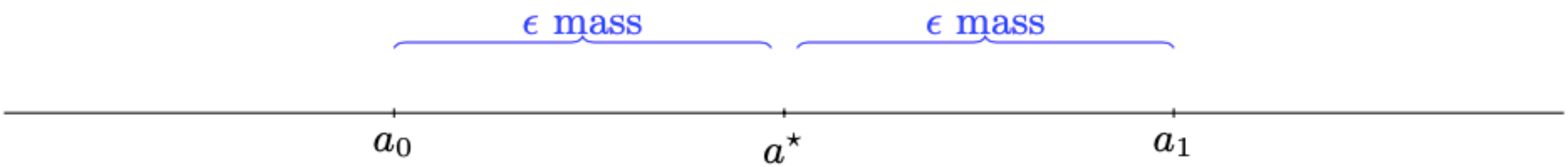


$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\ x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta/2$$

$$m > \log(2/\delta)/\epsilon$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2$$

$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S,\ x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta/2$$

$$m > \log(2/\delta)/\epsilon$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2$$

احتمال اینکه هیچ کدام از اتفاق‌ها رخ ندهد $1 - \delta$ =<

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Then, the following 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. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

بزرگ‌ترین $C \subset \mathscr{X}$ که با $h \in \mathscr{H}$ می‌توان همه ترکیب برچسب‌ها را روی آن گذاشت.

مثال‌ها:

- رگرسیون خطی:
- $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$
- $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$
- $l(h_w, (x, y)) = (h_w(x) - y)^2 = (\langle w, x \rangle - y)^2$

مثال‌ها:

- رگرسیون خطی:
  - $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$
  - $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$
  - $l(h_w, (x, y)) = (h_w(x) - y)^2 = (\langle w, x \rangle - y)^2$
- طبقه‌بندی دودویی خطی:
  - $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$
  - $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$
  - $l(h_w, (x, y)) = \max(1 - y\langle w, x \rangle, 0)$

بسم الله الرحمن الرحیم

یادگیری برخط

جلسه هشتم: نظریه یادگیری –
یادگیری برخط =>> یادگیری PAC

مرور

یادگیری یعنی

- تعمیم از L_S به L_D

- مثال: بچه‌ای که یاد می‌گیرد

یادگیری یعنی

- تعمیم از L_S به L_D

- مثال: بچه‌ای که یاد می‌گیرد
- مثال: علم فیزیک

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m}[L_{\mathscr{D},l}(A(S)) \leq \epsilon] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ و $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}\mathrm{r}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

الگوریتم ERM:

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$$

الگوریتم (ERM):

$$h_S \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h)$$

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Then, the following 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. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

بزرگ‌ترین $C \subset \mathscr{X}$ که با $h \in \mathscr{H}$ می‌توان همه ترکیب برچسب‌ها را روی آن گذاشت.

یادگیری برخط، ابزاری برای بهینه‌سازی در کمک به یادگیری

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

$$\mathbb{E}[l_t(x_t)] =$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

$\mathbb{E}[l_t(x_t)] =$ $\quad \mathbb{E}[\alpha_t l(x_t, z_t)] =$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t l(x_t, z_t)] = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(x_t, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] =$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t l(x_t, z_t)] = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(x_t, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \mathbb{E}[\alpha_t L_D(x_t)]$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t L_D(x_t)]$$

$$= \qquad\qquad\qquad\qquad >$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t L_D(x_t)]$$

$$\mathbb{E}[\sum_t l_t(x_t)] = \mathbb{E}[\sum_t \alpha_t L_D(x_t)]$$

$$= \qquad\qquad\qquad\qquad >$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t L_D(x_t)]$$

$$\mathbb{E}[\sum_t l_t(x_t)] = \ \mathbb{E}[\sum_t \alpha_t L_D(x_t)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \ = \ \mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t \alpha_t L_D(x_t)] \ \geq$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t L_D(x_t)]$$

$$\mathbb{E}[\sum_t l_t(x_t)] = \ \mathbb{E}[\sum_t \alpha_t L_D(x_t)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \ = \ \mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t \alpha_t L_D(x_t)] \ \geq \ \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\mathbb{E}[l_t(x_t)] = \quad \mathbb{E}[\alpha_t L_D(x_t)]$$

$$\mathbb{E}[\sum_t l_t(x_t)] = \mathbb{E}[\sum_t \alpha_t L_D(x_t)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] = \mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t \alpha_t L_D(x_t)] \geq \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\boxed{\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \geq \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]}$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u, T)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u, T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u, T)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \;\; \geq \;\; \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u,T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[l_t(u)] = \quad \mathbb{E}[\alpha_t l(u, z_t)] = \quad \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(u, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \quad \mathbb{E}[\alpha_t L_D(u)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u,T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[l_t(u)] = \quad \mathbb{E}[\alpha_t l(u, z_t)] = \quad \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(u, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \quad \mathbb{E}[\alpha_t L_D(u)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \alpha_t L_D(u) + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u,T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[l_t(u)] = \quad \mathbb{E}[\alpha_t l(u, z_t)] = \quad \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(u, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \quad \mathbb{E}[\alpha_t L_D(u)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \alpha_t L_D(u) + \mathbb{E}[R(u,T)]$$

$$\frac{1}{\sum \alpha_t} \sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq L_D(u) + \frac{1}{\sum \alpha_t} \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_{t} l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_{t} \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u,T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[l_t(u)] = \mathbb{E}[\alpha_t l(u, z_t)] = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(u, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \mathbb{E}[\alpha_t L_D(u)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \alpha_t L_D(u) + \mathbb{E}[R(u,T)]$$

$$\frac{1}{\sum \alpha_t} \sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq L_D(u) + \frac{1}{\sum \alpha_t} \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[\frac{1}{\sum \alpha_t} \sum_{t} l_t(x_t)] \geq \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_{t} \alpha_t x_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] - \sum_{t=1}^{T} \mathbb{E}[l_t(u)] \leq \mathbb{E}[R(u,T)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \mathbb{E}[l_t(u)] + \mathbb{E}[R(u,T)]$$

$$\mathbb{E}[l_t(u)] = \mathbb{E}[\alpha_t l(u, z_t)] = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\alpha_t \mathbb{E}_{z_t}[l(u, z_t) \,|\, z_1, \ldots, z_{t-1}]\right] = \mathbb{E}[\alpha_t L_D(u)]$$

$$\sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq \sum_{t=1}^{T} \alpha_t L_D(u) + \mathbb{E}[R(u,T)]$$

$$\frac{1}{\sum \alpha_t} \sum_{t=1}^{T} \mathbb{E}[l_t(x_t)] \leq L_D(u) + \frac{1}{\sum \alpha_t} \mathbb{E}[R(u,T)]$$

$$L_D(u) + \frac{1}{\sum \alpha_t} \mathbb{E}[R(u,T)] \geq \mathbb{E}[\frac{1}{\sum \alpha_t} \sum_t l_t(x_t)] \geq \mathbb{E}[L_D(\frac{1}{\sum \alpha_t} \sum_t \alpha_t x_t)]$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

$$L_D(u) + \frac{1}{\sum \alpha_t}\mathbb{E}[R(u,T)] \geq \quad \mathbb{E}[\frac{1}{\sum \alpha_t}\sum_t l_t(x_t)] \quad \geq \quad \mathbb{E}[L_D(\frac{1}{\sum \alpha_t}\sum_t \alpha_t x_t)]$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x} \rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x, z)]$ و $l(x, z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u, T)]}{\sum_{t=1}^{T}\alpha_t}$$

دسته‌بندی دوتایی

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$y_t \in \{-1, +1\}$

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T} \alpha_t} \sum_{t=1}^{T} \alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

$$\boxed{\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

$$\boxed{\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}}$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

فرض: η ثابت

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

$$x_1 = 0$$

الگوریتم GD:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

فرض: η ثابت

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

$$\sum_t f(x_t) - f(x^*) \leq \sum_t \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2} T G^2$$

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

$$\sum_t f(x_t) - f(x^*) \leq \sum_t \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2} T G^2$$

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

$$\sum_t f(x_t) - f(x^*) \leq \sum_t \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2} T G^2$$

$$\leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2} T G^2$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$$

$$f(x_t) - f(x^*) \leq \frac{1}{\eta}(x_t - x_{t+1})^\top (x_t - x^*)$$

$$\|a - b\|^2 - \|a\|^2 - \|b\|^2 = -2a^\top b$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2)$$

$$= \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

$$\sum_t f(x_t) - f(x^*) \leq \sum_t \frac{1}{2\eta}(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2) + \frac{\eta}{2}TG^2$$

$$\leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$ $\quad z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2}TG^2}{\sum_{t=1}^{T}\alpha_t}$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان:

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان: $= \nabla_x \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x} \rangle, 0)$ $\begin{cases} -yz & \text{در غیر این صورت} \\ 0 & 1 - y\langle \boldsymbol{z}, \boldsymbol{x} \rangle \leq 0 \end{cases}$

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان: $\nabla_x \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0) = \begin{cases} -yz & \text{در غیر این صورت} \\ 0 & 1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle \leq 0 \end{cases}$

$$G = R$$

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان: $\nabla_x \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0) = \begin{cases} -yz & \text{در غیر این صورت} \\ 0 & 1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle \leq 0 \end{cases}$

$$G = R$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان: $\nabla_x \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0) = \begin{cases} -yz & \text{در غیر این صورت} \\ 0 & 1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle \leq 0 \end{cases}$

$$G = R$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T} \leq L_D(u) + \frac{\frac{R}{2}\|x_1 - u\|^2 + \frac{1}{2}R}{\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{\frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2} T G^2}{\sum_{t=1}^{T} \alpha_t}$$

G: بیشینه گرادیان: $\nabla_x \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0) = \begin{cases} -yz & \text{در غیر این صورت} \\ 0 & 1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle \leq 0 \end{cases}$

$$G = R$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\leq L_D(u) + \frac{\frac{R\sqrt{T}}{2}\|x_1 - u\|^2 + \frac{1}{2R\sqrt{T}} T G^2}{T} \leq L_D(u) + \frac{\frac{R}{2}\|x_1 - u\|^2 + \frac{1}{2}R}{\sqrt{T}}$$

$$\leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$$y_t \in \{-1, +1\}$$

$$z_t \in \mathbb{R}^d$$

$$\|z_i\| \leq R$$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$\alpha_t = 1$ $\qquad$ $\eta = \dfrac{1}{R\sqrt{T}}$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$\alpha_t = 1$ $\qquad$ $\eta = \dfrac{1}{R\sqrt{T}}$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}} = L_D(u) + \frac{R}{2}\frac{\|u\|^2}{\sqrt{T}} + \frac{R}{2\sqrt{T}}$$

دسته‌بندی دوتایی

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle, 0)$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$\alpha_t = 1$ $\qquad$ $\eta = \frac{1}{R\sqrt{T}}$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}} = L_D(u) + \frac{R}{2}\frac{\|u\|^2}{\sqrt{T}} + \frac{R}{2\sqrt{T}}$$

منظم‌سازی

دسته‌بندی دوتایی با تابع زیان لجستیک

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x} \rangle)) \qquad y_t \in \{-1, +1\} \qquad z_t \in \mathbb{R}^d$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1, +1\} \qquad z_t \in \mathbb{R}^d$$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \ln(1 + \exp(-y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle))$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1, +1\} \qquad z_t \in \mathbb{R}^d$$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \ln(1 + \exp(-y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle))$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + ???$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x} \rangle))$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)}$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle)$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle) \; (-yz)$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle) \; (-yz)$$

$$G = R$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1, +1\} \qquad z_t \in \mathbb{R}^d$$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \ln(1 + \exp(-y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle))$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\alpha_t = 1$$

$$\eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$f(\boldsymbol{x}, (\boldsymbol{z}, y)) = \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1, +1\} \qquad z_t \in \mathbb{R}^d$$

$$\min_x L_D(x) := \frac{1}{N}\sum_{i=1}^{N} \ln(1 + \exp(-y_i\langle \boldsymbol{z}_i, \boldsymbol{x}\rangle))$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\alpha_t = 1$$

$$\eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}} = L_D(u) + \frac{R}{2}\frac{\|u\|^2}{\sqrt{T}} + \frac{R}{2\sqrt{T}}$$

بسم الله الرحمن الرحیم

یادگیری برخط

جلسه نهم: نظریه یادگیری –
یادگیری برخط =>> یادگیری PAC

مرور

تعریف: یادگیری احتمالا تقریبا درست

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ که با فرضیه‌ای $h \in \mathscr{H}$ هم‌خوانی داشته باشد و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m}[L_{\mathscr{D}}(A(S)) \leq \epsilon] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

در صورت وجود الگوریتم A، می‌گوییم مجموعه فرضیه‌های $\mathscr{H}$ PAC_یادگیری‌پذیر است.

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$
و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

$$L_{\mathscr{D}}(h) := \mathbb{E}_{z \sim \mathscr{D}}[l(h, z)]$$

در صورت وجود الگوریتم A، می‌گوییم مجموعه
فرضیه‌های $\mathscr{H}$ PAC ندانم‌گرا_ یادگیری‌پذیر است.

الگوریتم ERM:

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$$

الگوریتم (ERM):

$$h_S \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h)$$

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Then, the following 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. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

بزرگ‌ترین $C \subset \mathscr{X}$ که با $h \in \mathscr{H}$ می‌توان همه ترکیب برچسب‌ها را روی آن گذاشت.

یادگیری برخط، ابزاری برای بهینه‌سازی در کمک به یادگیری

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

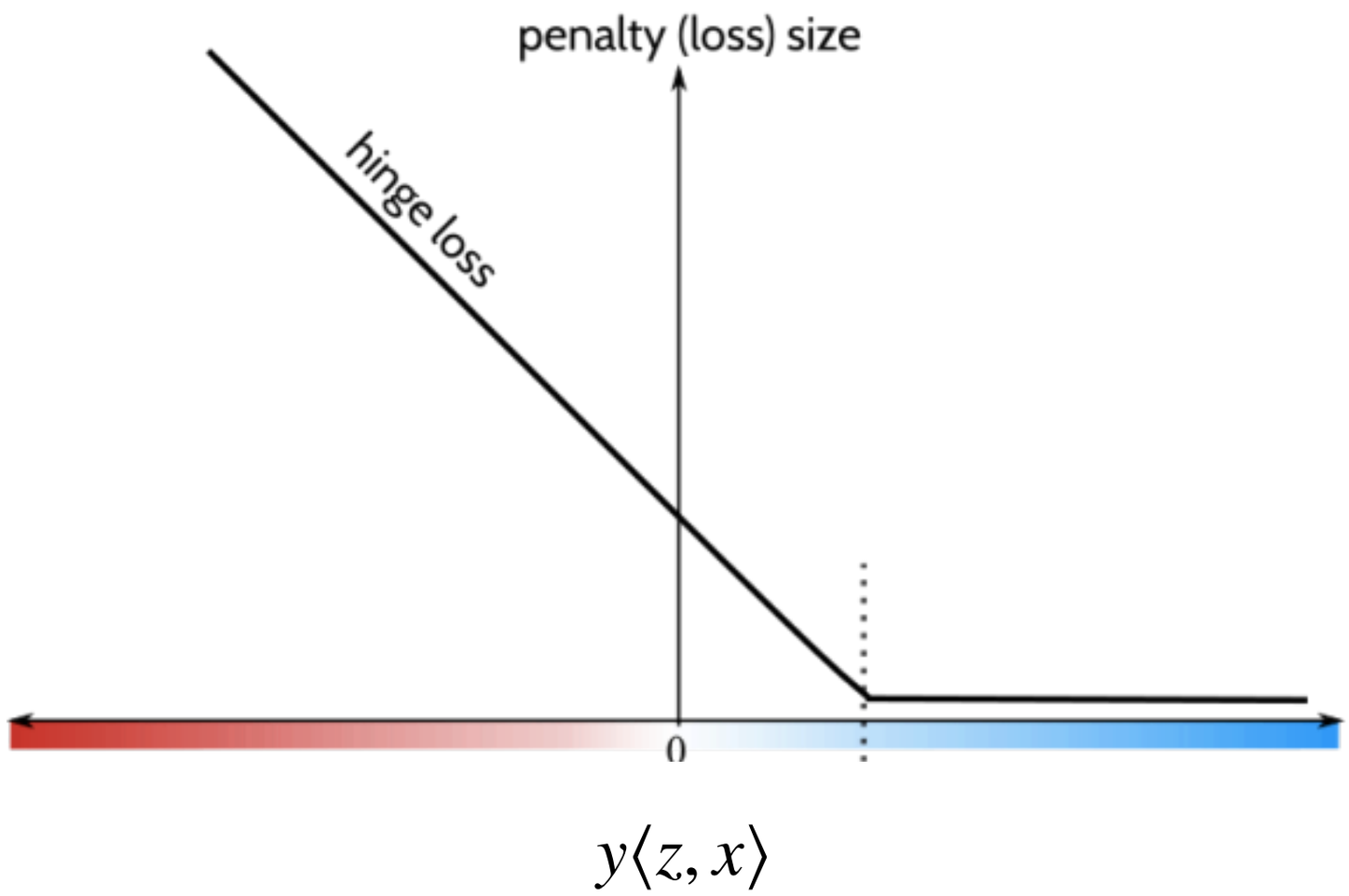
دسته‌بندی دوتایی

$$l(x,(z,y)) := \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$



دسته‌بندی دوتایی

$z_t \in \mathbb{R}^d$ $\quad$ $y_t \in \{-1, +1\}$ $\quad$ $l(x, (z, y)) := \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$

$\|z_i\| \leq R$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x, z)]$ و $l(x, z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u, T)]}{\sum_{t=1}^{T}\alpha_t}$$

دسته‌بندی دوتایی

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$y_t \in \{-1, +1\}$

$l(x, (z, y)) := \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x} \rangle, 0).$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u, T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z \sim D}[l(x, z)]$ و $l(x, z)$ محدب روی ورودی اول

● در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T} \alpha_t} \sum_{t=1}^{T} \alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u, T)]}{\sum_{t=1}^{T} \alpha_t}$$

دسته‌بندی دوتایی

$$l(x,(z,y)) := \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$$

$y_t \in \{-1, +1\}$

$z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

دسته‌بندی دوتایی

$l(x,(z,y)) := \max(1 - y\langle \boldsymbol{z}, \boldsymbol{x}\rangle, 0).$ $\quad y_t \in \{-1,+1\}$ $\quad z_t \in \mathbb{R}^d$

$\|z_i\| \leq R$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2+1}{\sqrt{T}}$$

LD یا LS؟

قضیه: در الگوریتم GD با η ثابت داریم

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

دسته‌بندی دوتایی ـ نسخه یادگیری

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی ـ نسخه یادگیری

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

$L_D(\bar{x}) - L_D(u)$

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

دسته‌بندی دوتایی ـ نسخه یادگیری

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

$L_D(\bar{x}) - L_D(u)$

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a} = \delta$$

دسته‌بندی دوتایی ـ نسخه یادگیری

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

$L_D(\bar{x}) - L_D(u)$

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a} = \delta$$

$$a = \frac{R}{2\delta}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی ـ نسخه یادگیری

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t} \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

$L_D(\bar{x}) - L_D(u)$

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a} = \delta$$

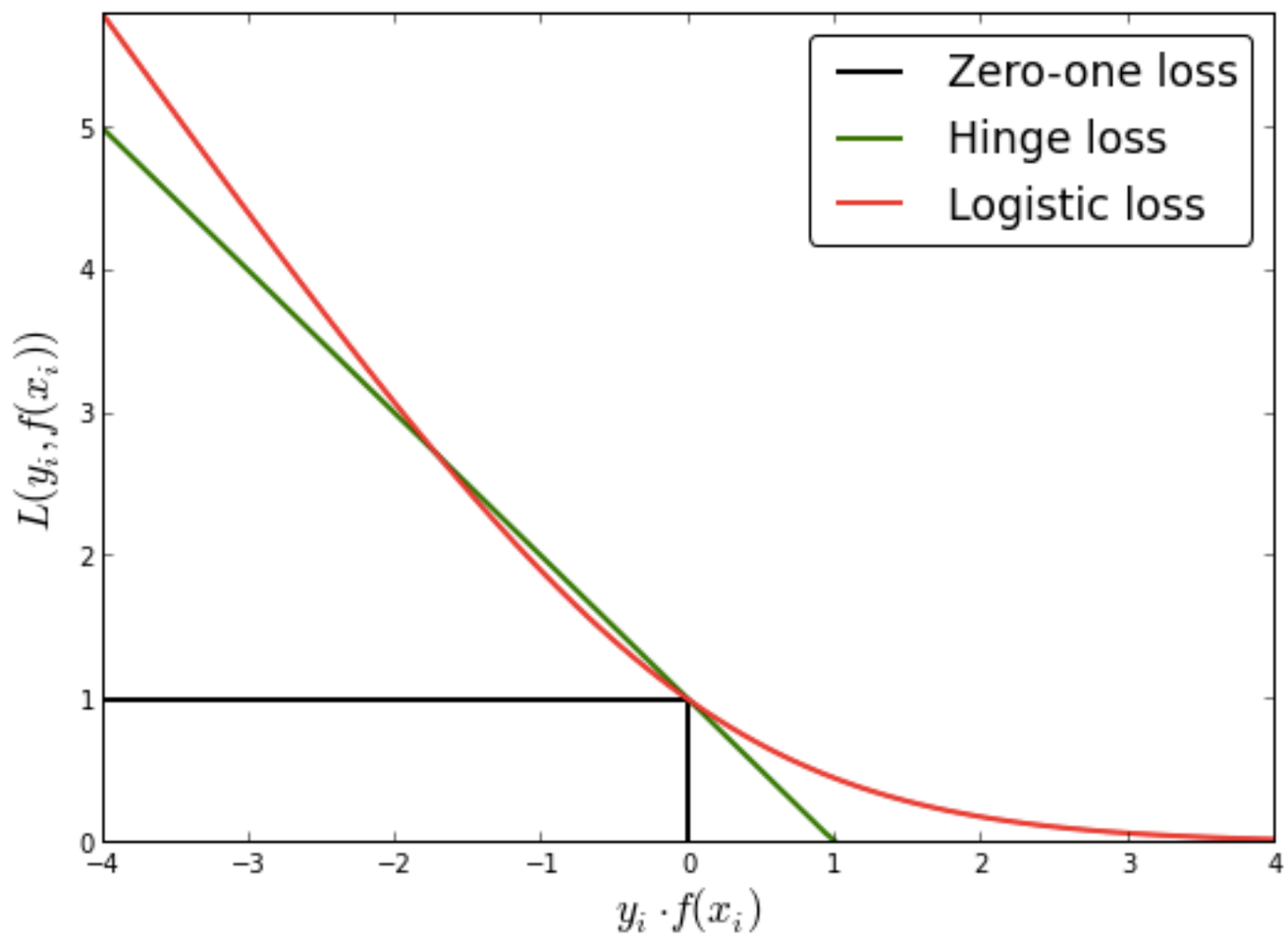
$$a = \frac{R}{2\delta}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

با احتمال $1 - \delta$ داریم:

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2\delta}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$



دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2} T G^2$$

$$\alpha_t = 1 \qquad \eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + ???$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x} \rangle))$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)}$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle)$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle) \; (-yz)$$

G: بیشینه گرادیان: $\nabla_x \ln(1 + \exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle))$

$$\frac{1}{1 + \exp(-y\langle z, x\rangle)} \exp(-y\langle z, x\rangle) \; (-yz)$$

$$G = R$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2} T G^2$$

$$\alpha_t = 1$$

$$\eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2 + 1}{\sqrt{T}}$$

دسته‌بندی دوتایی با تابع زیان لجستیک

$$l(x,(z,y)) := \ln(1+\exp(-y\langle \boldsymbol{z}, \boldsymbol{x}\rangle)) \qquad y_t \in \{-1,+1\} \qquad z_t \in \mathbb{R}^d$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T} \alpha_t}$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2} T G^2$$

$$\alpha_t = 1$$

$$\eta = \frac{1}{R\sqrt{T}}$$

$$\mathbb{E}[L_D(\bar{x})] \leq L_D(u) + \frac{R}{2}\frac{\|u\|^2+1}{\sqrt{T}} = L_D(u) + \frac{R}{2}\frac{\|u\|^2}{\sqrt{T}} + \frac{R}{2\sqrt{T}}$$

یادگیری برخط =>> یادگیری PAC

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\mathrm{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

قضیه:

- $x \in \mathbb{R}^d$
- $L_D := \mathbb{E}_{z\sim D}[l(x,z)]$ و $l(x,z)$ محدب روی ورودی اول

- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(x) := \alpha_t l(x, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $x_1, \ldots, x_T$ را تولید کند.

$$\mathbb{E}\left[L_D\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t x_t\right)\right] \leq L_D(u) + \frac{\mathbb{E}[R(u,T)]}{\sum_{t=1}^{T}\alpha_t}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

مارتینگل

$$\mathbb{E}[|Z_t|] < \infty,$$

$$\mathbb{E}[Z_{t+1}|Z_1, \ldots, Z_t] = Z_t$$

مارتینگل

$$\mathbb{E}[|Z_t|] < \infty,$$

$$\mathbb{E}[Z_{t+1}|Z_1, \ldots, Z_t] = Z_t$$

$$\Leftarrow |Z_t - Z_{t+1}| \leq B,$$

$$\mathbb{P}[Z_T - Z_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2B^2T}\right)$$

$$\mathbb{P}[Z_0 - Z_T \geq \epsilon]. \leq \exp\left(-\frac{\epsilon^2}{2B^2T}\right)$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$$l_t(h) := l(h, z_t)$$

$$\mathbb{E}_{z_t}[l_t(h_t) \,|\, z_1, \ldots, z_{t-1}] = L_D(h_t)$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$$l_t(h) := l(h, z_t)$$

$$\mathbb{E}_{z_t}[l_t(h_t) \,|\, z_1, \ldots, z_{t-1}] = L_D(h_t)$$

$$\mathbb{E}_{z_t}[L_D(h_t) - l_t(h_t) \,|\, z_1, \ldots, z_{t-1}] = 0$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$$l_t(h) := l(h, z_t)$$

$$\mathbb{E}_{z_t}[l_t(h_t) \,|\, z_1, \ldots, z_{t-1}] = L_D(h_t)$$

$$\mathbb{E}_{z_t}[L_D(h_t) - l_t(h_t) \,|\, z_1, \ldots, z_{t-1}] = 0$$

$$\sum_t L_D(h_t) - l_t(h_t)$$ مارتینگل است

مارتینگل

$$\mathbb{E}[|Z_t|] < \infty,$$

$$\mathbb{E}[Z_{t+1} | Z_1, \ldots, Z_t] = Z_t$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$\sum_t L_D(h_t) - l_t(h_t)$ مارتینگل است

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$\sum_t L_D(h_t) - l_t(h_t)$ مارتینگل است

$$\mathbb{P}[Z_T - Z_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2B^2T}\right)$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$\sum_t L_D(h_t) - l_t(h_t)$ مارتینگل است

$$\mathbb{P}[Z_T - Z_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2B^2T}\right)$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$\sum_t L_D(h_t) - l_t(h_t)$ مارتینگل است

$$\mathbb{P}[Z_T - Z_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2B^2T}\right)$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T}$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T} \qquad \sqrt{2T \ln 2/\delta} \leq \epsilon$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T} \qquad \sqrt{2T \ln 2/\delta} \leq \epsilon$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t L_D(h_t) - l_t(h_t) \leq \sqrt{2T \ln 2/\delta}$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T} \qquad \sqrt{2T \ln 2/\delta} \leq \epsilon$$

با احتمال حداقل

$1 - \delta/2$

$$\sum_t L_D(h_t) - l_t(h_t) \leq \sqrt{2T \ln 2/\delta}$$

$$\sum_t l_t(h_t) - \sum_t l_t(u) = \mathrm{Regret}(u) \leq R(u, T)$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T} \qquad \sqrt{2T \ln 2/\delta} \leq \epsilon$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t L_D(h_t) - l_t(h_t) \leq \sqrt{2T \ln 2/\delta}$$

$$\sum_t l_t(h_t) - \sum_t l_t(u) = \mathrm{Regret}(u) \leq R(u, T)$$

$$\sum_t L_D(h_t) \leq \sum_t l_t(u) + R(u, T) + \sqrt{2T \ln 2/\delta}$$

$$\mathbb{P}[\sum_t L_D(h_t) - l_t(h_t) \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2T}} \overset{?}{\leq} \delta/2$$

$$-\frac{\epsilon^2}{2T} \leq \ln \delta/2 \qquad \ln 2/\delta \leq \frac{\epsilon^2}{2T} \qquad \sqrt{2T \ln 2/\delta} \leq \epsilon$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t L_D(h_t) - l_t(h_t) \leq \sqrt{2T \ln 2/\delta}$$

$$\sum_t l_t(h_t) - \sum_t l_t(u) = \mathrm{Regret}(u) \leq R(u, T)$$

$$\sum_t L_D(h_t) \leq \sum_t l_t(u) + R(u, T) + \sqrt{2T \ln 2/\delta}$$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u, T) + \sqrt{\frac{2 \ln 2/\delta}{T}}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

$$\sum_t l_t(u) - L_D(u) \leq \sqrt{2T \ln 2/\delta}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

$$\sum_t l_t(u) - L_D(u) \leq \sqrt{2T \ln 2/\delta}$$

$$\sum_t l_t(u) \leq TL_D(u) + \sqrt{2T \ln 2/\delta}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t l_t(u) - L_D(u) \leq \sqrt{2T\ln 2/\delta}$$

$$\sum_t l_t(u) \leq TL_D(u) + \sqrt{2T\ln 2/\delta}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t l_t(u) - L_D(u) \leq \sqrt{2T\ln 2/\delta}$$

$$\sum_t l_t(u) \leq TL_D(u) + \sqrt{2T\ln 2/\delta}$$

$$\frac{1}{T}\sum_t L_D(h_t) \leq L_D(u) + \frac{1}{T}R(u,T) + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

با احتمال حداقل
$1 - \delta/2$

$$\frac{1}{T}\sum_t L_D(h_t) \leq \frac{1}{T}\sum_t l_t(u) + \frac{1}{T}R(u,T) + \sqrt{\frac{2\ln 2/\delta}{T}}$$

$$\mathbb{E}_{z_t}[l_t(u) \,|\, z_1, \ldots, z_{t-1}] = L_D(u)$$

با احتمال حداقل
$1 - \delta/2$

$$\sum_t l_t(u) - L_D(u) \leq \sqrt{2T\ln 2/\delta}$$

$$\sum_t l_t(u) \leq TL_D(u) + \sqrt{2T\ln 2/\delta}$$

با احتمال حداقل
$1 - \delta$

$$\frac{1}{T}\sum_t L_D(h_t) \leq L_D(u) + \frac{1}{T}R(u,T) + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

Theorem 3.3. *For α-strongly convex loss functions, online gradient descent with step sizes $\eta_t = \frac{1}{\alpha t}$ achieves the following guarantee for all $T \geq 1$*

$$\text{regret}_T \leq \frac{G^2}{2\alpha}(1 + \log T).$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\mathrm{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

مثال (رگرسیون خطی):

$\mathscr{X} \times \mathscr{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathscr{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x,y)) = (\langle w, x \rangle - y)^2$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T} \sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u, T)}{T} + 2\sqrt{\frac{2 \ln 2/\delta}{T}}$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

$$\nabla_w(\langle w, x \rangle - y)^2 = 2(\langle w, x \rangle - y)x$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

$$\nabla_w(\langle w, x \rangle - y)^2 = 2(\langle w, x \rangle - y)x \qquad ||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.1. *Online gradient descent with step sizes* $\{\eta_t = \frac{D}{G\sqrt{t}},\ t \in [T]\}$ *guarantees the following for all* $T \geq 1$*:*

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^\star \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \frac{3}{2} GD\sqrt{T}$$

$$\nabla_w(\langle w, x \rangle - y)^2 = 2(\langle w, x \rangle - y)x \qquad ||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$R(u,T) \leq \frac{3}{2} 2(A+1)B^2 B\sqrt{T} = O(AB^3\sqrt{T})$$

مثال (رگرسیون خطی):

$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{O(AB^3\sqrt{T})}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

مثال (رگرسیون خطی):

$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{O(AB^3\sqrt{T})}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

$$\frac{1}{\sqrt{T}}(O(AB^3) + 2\sqrt{2\ln 2/\delta}) \leq^? \epsilon$$

مثال (رگرسیون خطی):

$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{O(AB^3\sqrt{T})}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

$$\frac{1}{\sqrt{T}}(O(AB^3) + 2\sqrt{2\ln 2/\delta}) \leq^? \epsilon$$

$$\Omega(\frac{1}{\epsilon^2}A^2B^6 + \frac{1}{\epsilon^2}\ln 2/\delta) \leq T$$

مثال (رگرسیون خطی):

$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{O(AB^3\sqrt{T})}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

$$\frac{1}{\sqrt{T}}(O(AB^3) + 2\sqrt{2\ln 2/\delta}) \leq^? \epsilon$$

$$\Omega(\frac{1}{\epsilon^2}A^2B^6 + \frac{1}{\epsilon^2}\ln 2/\delta) \leq T$$

رگرسیون خطی با محدودیت $||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$

را می‌توان با $\Omega(\frac{1}{\epsilon^2}A^2B^6 + \frac{1}{\epsilon^2}\ln 2/\delta)$ نمونه یاد گرفت!

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$\leq \frac{1}{2\eta}A^2 + \frac{\eta}{2}O(A^2B^4)T = O(A^2B^2\sqrt{T})$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

$$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$\leq \frac{1}{2\eta}A^2 + \frac{\eta}{2}O(A^2B^4)T = O(A^2B^2\sqrt{T})$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$\leq \frac{1}{2\eta}A^2 + \frac{\eta}{2}O(A^2B^4)T = O(A^2B^2\sqrt{T})$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\nabla_w(\langle w, x \rangle - y)^2 = 2(\langle w, x \rangle - y)x \qquad ||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$\leq \frac{1}{2\eta}A^2 + \frac{\eta}{2}O(A^2B^4)T = O(A^2B^2\sqrt{T})$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

قضیه: در الگوریتم GD با η ثابت داریم

$$\sum_t f(x_t) - f(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}TG^2$$

$$\nabla_w(\langle w, x \rangle - y)^2 = 2(\langle w, x \rangle - y)x \qquad ||x|| \leq B \quad |y| \leq B \quad ||w|| \leq A$$

$$R(u,T) \leq \frac{1}{2\eta}\|u\|^2 + \frac{\eta}{2}(2(A+1)B^2)^2 T \leq \frac{1}{2\eta}A^2 + \frac{\eta}{2}O(A^2B^4)T = O(A^2B^2\sqrt{T})$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

$$O(A^2B^2\sqrt{T})$$

$$O\left(\frac{A^2B^2\sqrt{T}}{T}\right) + O\left(\sqrt{\frac{\ln 2/\delta}{T}}\right) \leq^? \epsilon$$

$$O\left(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta\right) \leq^? T$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

$$O(A^2B^2\sqrt{T})$$

$$O\left(\frac{A^2B^2\sqrt{T}}{T}\right) + O\left(\sqrt{\frac{\ln 2/\delta}{T}}\right) \leq^? \epsilon$$

$$O\left(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta\right) \leq^? T$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^{?} \epsilon$$

$$R(u,T) \leq O(A^2B^2\sqrt{T})$$

$$O(\frac{A^2B^2\sqrt{T}}{T}) + O(\sqrt{\frac{\ln 2/\delta}{T}}) \leq^{?} \epsilon$$

$$O(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta) \leq^{?} T$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u\in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^{?} \epsilon$$

$$R(u,T) \leq O(A^2B^2\sqrt{T})$$

$$O(\frac{A^2B^2\sqrt{T}}{T}) + O(\sqrt{\frac{\ln 2/\delta}{T}}) \leq^{?} \epsilon$$

$$O(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta) \leq^{?} T$$

رگرسیون خطی با محدودیت $||w|| \leq A$ $|y| \leq B$ $||x|| \leq B$

را می‌توان با $\Omega(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta)$ نمونه یاد گرفت!

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u\in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \leq^? \epsilon$$

$$R(u,T) \leq O(A^2B^2\sqrt{T})$$

$$O(\frac{A^2B^2\sqrt{T}}{T}) + O(\sqrt{\frac{\ln 2/\delta}{T}}) \leq^? \epsilon \qquad O(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta) \leq^? T$$

رگرسیون خطی با محدودیت $||w|| \leq A \quad |y| \leq B \quad ||x|| \leq B$

را می‌توان با $\Omega(\frac{1}{\epsilon^2}A^4B^4 + \frac{1}{\epsilon^2}\ln 2/\delta)$ نمونه یاد گرفت!

قدیمی

رگرسیون خطی با محدودیت $||w|| \leq A \quad |y| \leq B \quad ||x|| \leq B$

را می‌توان با $\Omega(\frac{1}{\epsilon^2}A^2B^6 + \frac{1}{\epsilon^2}\ln 2/\delta)$ نمونه یاد گرفت!

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.3. *For α-strongly convex loss functions, online gradient descent with step sizes $\eta_t = \frac{1}{\alpha t}$ achieves the following guarantee for all $T \geq 1$*

$$\text{regret}_T \leq \frac{G^2}{2\alpha}(1 + \log T).$$

مثال (رگرسیون خطی):

$\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ و $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ که $h_w(x) = \langle w, x \rangle$ و

$$l(h_w, (x, y)) = (\langle w, x \rangle - y)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}} \overset{?}{\leq} \epsilon$$

Theorem 3.3. *For α-strongly convex loss functions, online gradient descent with step sizes $\eta_t = \frac{1}{\alpha t}$ achieves the following guarantee for all $T \geq 1$*

$$\text{regret}_T \leq \frac{G^2}{2\alpha}(1 + \log T).$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

قضیه:

- الگوریتم یادگیری برخط روی $\mathscr{H}$ با $\text{Regret}(u) \leq R(u,T)$ به ازای هر $u \in \mathscr{H}$ و تابع زیان $l(h,z) \in [0,1]$.
- در مرحله t: با $z_t \sim D$ تابع زیان برخط $l_t(h) := l(h, z_t)$ را بساز و به الگوریتم یادگیری برخط بده تا $h_1, \ldots, h_T$ را تولید کند.
- با احتمال حداقل $1 - \delta$ داریم:

$$\frac{1}{T}\sum_{t=1}^{T} L_D(h_t) \leq \min_{u \in H} L_D(u) + \frac{R(u,T)}{T} + 2\sqrt{\frac{2\ln 2/\delta}{T}}$$

این یادگیری است؟

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

$\epsilon/2$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

$\epsilon/2$

$$L_S(\hat{h}) \leq L_S(h^*)$$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

$\epsilon/2$

$$L_S(\hat{h}) \leq L_S(h^*)$$

$$L_D(\hat{h}) - \epsilon/2 \leq L_S(\hat{h})$$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

$\epsilon/2$

$$L_S(\hat{h}) \leq L_S(h^*)$$

$$L_D(\hat{h}) - \epsilon/2 \leq L_S(\hat{h})$$

$$L_S(h^*) \leq L_D(h^*) + \epsilon/2$$

قضیه:

مجموعه $C = \{h_1, \ldots, h_{|C|}\}$ داده شده، و $\hat{h} = \arg\min_{h \in C} L_S(h)$ که C شامل T نمونه مستقل از D است. با احتمال حداقل $1 - \delta$ داریم:

$$L_D(\hat{h}) \leq \min_{h \in C} L_D(h) + 4\sqrt{\frac{2\ln(2|C|/\delta)}{T}}$$

با احتمال حداقل $1 - \delta$ برای همه hها داریم:

برگه بعدی

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{8\ln(2|C|/\delta)}{T}}$$

$\epsilon/2$

$$L_S(\hat{h}) \leq L_S(h^*)$$

$$L_D(\hat{h}) - \epsilon/2 \leq L_S(\hat{h})$$

$$L_S(h^*) \leq L_D(h^*) + \epsilon/2$$

$$L_D(\hat{h}) \leq L_D(h^*) + \epsilon$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}}$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}} \leq \delta$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}} \leq \delta$$

$$-\frac{\epsilon^2 T}{8} < \ln \frac{\delta}{2|C|}$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}} \leq \delta$$

$$-\frac{\epsilon^2 T}{8} \leq \ln\frac{\delta}{2|C|} \qquad \frac{\epsilon^2 T}{8} \geq \ln\frac{2|C|}{\delta}$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}} \leq \delta$$

$$e^{-\frac{\epsilon^2 T}{8}} \leq \frac{\delta}{2|C|} \qquad -\frac{\epsilon^2 T}{8} \leq \ln\frac{\delta}{2|C|} \qquad \frac{\epsilon^2 T}{8} \geq \ln\frac{2|C|}{\delta}$$

برای یک h:

$$\mathbb{P}[\,|L_D(h) - L_S(h)| > \epsilon/2] \leq 2e^{-\frac{\epsilon^2 T}{8}}$$

$$\mathbb{P}[\,\exists h \in C : |L_D(h) - L_S(h)| > \epsilon/2] \leq 2|C|e^{-\frac{\epsilon^2 T}{8}} \leq \delta$$

$$e^{-\frac{\epsilon^2 T}{8}} \leq \frac{\delta}{2|C|} \qquad -\frac{\epsilon^2 T}{8} \leq \ln\frac{\delta}{2|C|} \qquad \frac{\epsilon^2 T}{8} \geq \ln\frac{2|C|}{\delta}$$

$$\epsilon \geq \sqrt{\frac{8\ln\frac{2|C|}{\delta}}{T}}$$

یادگیری ناپذیری

غذای مجانی در کار نیست!

THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

COROLLARY 5.2 *Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable.*

یادگیری غیر یکنواخت

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را به صورت **غیر یک‌نواخت** یاد می‌گیرد، اگر تابع $m(\epsilon, \delta, h) \to \mathbb{N}$ وجود داشته باشد که به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ و به ازای هر $\epsilon > 0$ و $\delta > 0$ داشته باشیم:

$$\mathbb{P}_{S \sim \mathscr{D}^{m(\epsilon,\delta,h')}} \left[L_{\mathscr{D}}(A(S)) \leq L_D(h') + \epsilon \right] \geq 1 - \delta$$

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را به صورت **غیر یک‌نواخت** یاد می‌گیرد، اگر تابع $m(\epsilon, \delta, h) \to \mathbb{N}$ وجود داشته باشد که به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ و به ازای هر $\epsilon > 0$ و $\delta > 0$ داشته باشیم:

$$\mathbb{P}_{S \sim \mathscr{D}^{m(\epsilon,\delta,h')}} \left[L_{\mathscr{D}}(A(S)) \leq L_D(h') + \epsilon \right] \geq 1 - \delta$$

تعریف: یادگیری احتمالا تقریبا درست ندانم‌گرا

الگوریتم A مجموعه فرضیه‌های $\mathscr{H}$ را یاد می‌گیرد، اگر به ازای هر توزیع $\mathscr{D}$ روی $\mathscr{X} \times \mathscr{Y}$ و به ازای هر $\epsilon > 0$ و $\delta > 0$ یک $m(\epsilon, \delta)$ وجود داشته باشد که:

$$\mathbb{P}_{S \sim \mathscr{D}^m} \left[L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_D(h') + \epsilon \right] \geq 1 - \delta$$

THEOREM 7.2 *A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.*

`Structural Risk Minimization (SRM)`

prior knowledge:
$\mathcal{H} = \bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m^{\mathrm{UC}}_{\mathcal{H}_n}$
$w : \mathbb{N} \to [0,1]$ where $\sum_n w(n) \leq 1$
define: ϵ_n as in Equation (7.1) ; $n(h)$ as in Equation (7.4)
input: training set $S \sim \mathcal{D}^m$, confidence δ
output: $h \in \operatorname{argmin}_{h \in \mathcal{H}} \left[L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta) \right]$

جمع‌بندی

- یادگیری برخط ==> یادگیری