Margin-based generalization error bounds for classification

Chicheng Zhang

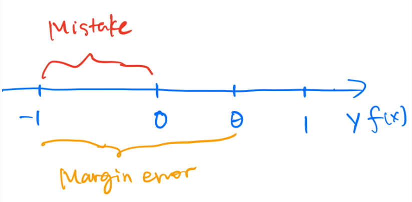
CSC 588, University of Arizona

Introduction

In the boosting lecture, we saw:

Theorem
Suppose base class $\mathcal{B}$ is finite, $\mathcal{C}(\mathcal{B}) = \{\sum_{h\in\mathcal{B}} \alpha_h h(x) : \sum_{h\in\mathcal{B}} |\alpha_h| \leq 1\}$ is the set of voting classifiers over $\mathcal{B}$. Fix margin $\theta \in [0,1]$. Then, for any distribution D, with probability $1-\delta$, for all $f \in \mathcal{C}(\mathcal{B})$,

$$\mathbb{P}_D(yf(x) \leq 0) \leq \underbrace{\mathbb{P}_S(yf(x) \leq \theta)}_{\text{“Margin error” of } f} + O\left(\frac{1}{\theta}\sqrt{\frac{\ln\frac{|\mathcal{B}|}{\delta}}{m}}\right)$$



Overview of this lecture

Questions:

- Can we develop some geometric intuition on this result?
- Can we generalize this result to analyze other large-margin classifiers?
- How can we prove this result?
- Can we use the insights obtained to design practical algorithms?

A geometric interpretation of boosting's margin bound

- Let $\mathcal{B} = \{h_1, \ldots, h_d\}$
- For every x, define its corresponding $z = (h_1(x), \ldots, h_d(x))$
- Any element in $\mathcal{C}(\mathcal{B})$, $f_\alpha(x) = \sum_{i=1}^{d} \alpha_i h_i(x)$ can be alternatively written as $g_\alpha(z) = \langle \alpha, z \rangle$

Theorem (Restated version)
Fix margin $\theta \in [0, 1]$. Then, for any distribution D, with probability $1 - \delta$, for all α such that $\|\alpha\|_1 \leq 1$,

$$\mathbb{P}_D(y \langle \alpha, z \rangle \leq 0) \leq \underbrace{\mathbb{P}_S(y \langle \alpha, z \rangle \leq \theta)}_{\text{“Margin error” of } g_\alpha(z) = \langle \alpha, z \rangle} + O\left(\frac{1}{\theta} \sqrt{\frac{\ln \frac{d}{\delta}}{m}} \right)$$

Margin bounds for linear classifiers: general ℓ_1/ℓ_∞ version

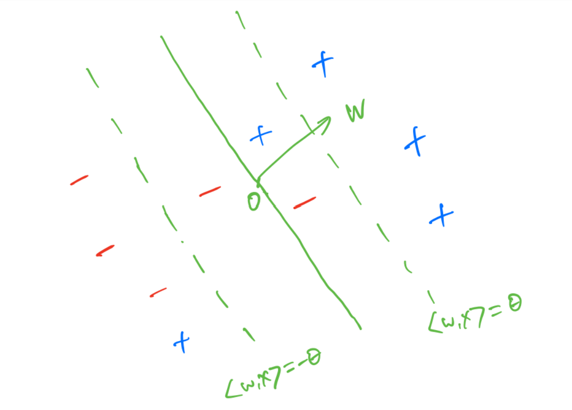
Theorem (general ℓ_1/ℓ_∞ margin bound)
Fix $B_1, R_\infty > 0$, and margin $\theta \in (0, B_1 R_\infty]$. Suppose D is a distribution over $\left\{x \in \mathbb{R}^d : \|x\|_\infty \le R_\infty\right\} \times \{\pm 1\}$. Then, with probability $1-\delta$, for all $w \in \mathbb{R}^d$ such that $\|w\|_1 \le B_1$,

$$\mathbb{P}_D(y\langle w, x\rangle \le 0) \le \mathbb{P}_S(y\langle w, x\rangle \le \theta) + O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln\frac{d}{\delta}}{m}}\right)$$

Remarks:

- Larger $\theta \implies$ smaller "generalization gap" term
- The bound is almost-dimension free, cf. VC theory ($O(\sqrt{\frac{d}{m}})$ term)
- Scale-invariance: scaling w and θ by the same factor (e.g. 10) results in the same bound

Margin error in linear classification: an illustration

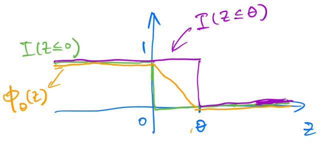


- $\mathbb{P}_S(y\langle w, x\rangle \leq 0) = 2/10$
- $\mathbb{P}_S(y\langle w, x\rangle \leq \theta) = 4/10$

Proof of general ℓ_1/ℓ_∞ margin bound

Step 1: Bridging 0-1 error and margin error using the "ramp-loss" $\ell_\theta(w,(x,y)) = \phi_\theta(y\langle w,x\rangle)$, where

$$\phi_\theta(z) = \begin{cases} 1, & z \le 0 \\ 1 - \frac{z}{\theta}, & 0 \le z \le \theta \\ 0, & z \ge \theta, \end{cases}$$



observe:

1. ϕ_θ is $\frac{1}{\theta}$-Lipschitz
2. $I(z \le 0) \le \phi_\theta(z) \le I(z \le \theta)$, therefore:

$$L_\theta(w,D) = \mathbb{E}_D\left[\ell_\theta(w,(x,y))\right] \ge \mathbb{P}_D(y\langle w,x\rangle \le 0),$$

$$L_\theta(w,S) = \mathbb{E}_S\left[\ell_\theta(w,(x,y))\right] \le \mathbb{P}_S(y\langle w,x\rangle \le \theta).$$

Are $L_\theta(w,S)$ and $L_\theta(w,D)$ close?

Proof of general ℓ_1/ℓ_∞ margin bound (cont'd)

Step 2: Uniform concentration between $L_\theta(w, S)$ and $L_\theta(w, D)$

1. Last lecture $\implies$ With probability $1 - \delta$, for all w such that $\|w\|_1 \leq B_1$:

$$|L_\theta(w, S) - L_\theta(w, D)| \leq 4\sqrt{\frac{\ln \frac{4}{\delta}}{2m}} + 4\,\mathrm{Rad}_m(\mathcal{F}),$$

   where $\mathcal{F} = \{\ell_\theta(w, (x, y)) : \|w\|_1 \leq B_1\}$

2. Bounding $\mathrm{Rad}_m(\mathcal{F})$:

$$\begin{aligned} \mathrm{Rad}_m(\mathcal{F}) &= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_1 \leq B_1} \sum_{i=1}^{m} \sigma_i \phi_\theta(y_i \langle w, x_i\rangle)\right] \\ &\leq \frac{1}{\theta} \cdot \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_1 \leq B_1} \sum_{i=1}^{m} \sigma_i y_i \langle w, x_i\rangle\right] \text{(Contraction ineq.)} \\ &= \frac{1}{\theta}\,\mathrm{Rad}_m(\mathcal{H}), \end{aligned}$$

   where $\mathcal{H} = \{g_w(x) := \langle w, x\rangle : \|w\|_1 \leq B_1\}$.

Bounding $\mathrm{Rad}_m(\mathcal{H})$

Theorem
If $\mathcal{H} = \{g_w(x) : \|w\|_1 \leq B_1\}$, and S is a set of examples that lie in $\left\{x \in \mathbb{R}^d : \|x\|_\infty \leq R_\infty\right\}$. Then $\mathrm{Rad}_S(\mathcal{H}) \leq B_1 R_\infty \sqrt{\frac{2\ln(2d)}{m}}$.

Proof.

$$
\begin{aligned}
\mathrm{Rad}_S(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_{w:\|w\|_1 \leq B_1} \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle\right] \\
&= B_1 \cdot \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|_\infty\right] \\
&\leq B_1 \cdot \mathbb{E}_\sigma\left[\underbrace{\max\left(\max_{j=1}^d \sum_{i=1}^m \sigma_i x_{i,j}, \max_{j=1}^d \sum_{i=1}^m \sigma_i(-x_{i,j})\right)}_{\text{max over } 2d \text{ r.v.'s, each } mR_\infty^2\text{-subgaussian}}\right] \\
&\leq B_1 R_\infty \sqrt{\frac{2\ln(2d)}{m}} \qquad \text{(Massart's finite lemma)}
\end{aligned}
$$

Dual norms

Definition
Given a norm $\|\cdot\|$, and vector $u \in \mathbb{R}^d$, define

$$\|u\|_\star = \sup_{v:\|v\|\leq 1} \langle u, v\rangle$$

to be the dual norm ($\|\cdot\|_\star$) of u.

Example of dual norms:

- $\|\cdot\|_1$ has dual norm $\|\cdot\|_\infty$
- $\|\cdot\|_2$ has dual norm $\|\cdot\|_2$
- More generally, for $p \in [1, \infty]$, $\|\cdot\|_p$ ($\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$) has norm $\|\cdot\|_q$, where q is the conjugate exponent of p ($\frac{1}{p} + \frac{1}{q} = 1$).

Proof of general ℓ_1/ℓ_∞ margin bound (cont'd)

Step 3: putting everything together

- Step 2 $\implies$ With probability $1-\delta$, for all w such that $\|w\|_1 \leq B_1$:

$$L_\theta(w, D) - L_\theta(w, S) \leq 4\sqrt{\frac{\ln \frac{4}{\delta}}{2m}} + 4\frac{B_1 R_\infty}{\theta}\sqrt{\frac{2\ln(2d)}{m}} = O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln \frac{d}{\delta}}{m}}\right)$$

- Step 1 $\implies$
  $L_\theta(w, D) \geq \mathbb{P}_D(y\langle w, x\rangle \leq 0)$, and $L_\theta(w, S) \leq \mathbb{P}_S(y\langle w, x\rangle \leq \theta)$
- Combining,

$$\mathbb{P}_D(y\langle w, x\rangle \leq 0) \leq \mathbb{P}_S(y\langle w, x\rangle \leq \theta) + O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln \frac{d}{\delta}}{m}}\right). \quad \square$$

Margin bounds for linear classifiers: ℓ_2/ℓ_2 version

What if our data satisfy other geometric constraints (intead of lying in ℓ_∞ balls)?

Theorem (general ℓ_2/ℓ_2 margin bound)
Fix $B_2, R_2 > 0$, and margin $\theta \in (0, B_2 R_2]$. Suppose D is a distribution over $\left\{x \in \mathbb{R}^d : \|x\|_2 \leq R_2\right\} \times \{\pm 1\}$. Then, with probability $1-\delta$, for all $w \in \mathbb{R}^d$ such that $\|w\|_2 \leq B_2$,

$$\mathbb{P}_D(y\langle w, x\rangle \leq 0) \leq \mathbb{P}_S(y\langle w, x\rangle \leq \theta) + O\left(\frac{B_2 R_2}{\theta}\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right)$$

Proof sketch.
Same as the proof of ℓ_1/ℓ_∞ bound, except that we now bound $\mathrm{Rad}_S(\mathcal{H})$ by $B_2 R_2 \sqrt{\frac{1}{m}}$ (last lecture). $\square$

ℓ_1/ℓ_∞ vs. ℓ_2/ℓ_2 bounds

Bound type	Constraint on x	Constraint on w	Bound
ℓ_1/ℓ_∞	$\|x\|_\infty \leq R_\infty$	$\|w\|_1 \leq B_1$	$\tilde{O}(B_1 R_\infty \sqrt{\frac{1}{m\theta^2}})$
ℓ_2/ℓ_2	$\|x\|_2 \leq R_2$	$\|w\|_2 \leq B_2$	$\tilde{O}(B_2 R_2 \sqrt{\frac{1}{m\theta^2}})$

Incomparable in general:

- Suppose D is supported on $\{x : \|x\|_\infty \leq X_\infty\}$, and we investigate the generalization error bound of some w with $\|w\|_1 \leq W_1$
  - Idea 1: applying ℓ_1/ℓ_∞ bound directly $\implies \tilde{O}(W_1 X_\infty \sqrt{\frac{1}{m\theta^2}})$
  - Idea 2: applying ℓ_2/ℓ_2 bound
    - $B_2 = W_1$
    - $R_2 = \sqrt{d} X_\infty$
    - Bound: $\tilde{O}(\sqrt{d} W_1 X_\infty \sqrt{\frac{1}{m\theta^2}})$
  - ℓ_1/ℓ_∞ bound is is a factor of $\sqrt{d}$ better in this case
- Exercise: construct a setting when ℓ_2/ℓ_2 bound is a factor of $\sqrt{d}$ better than ℓ_1/ℓ_∞ bound

Margin bounds for neural nets (Bartlett, Foster, Telgarsky, 2017)

Theorem 1.1. *Let nonlinearities* $(\sigma_1, \ldots, \sigma_L)$ *and reference matrices* $(M_1, \ldots, M_L)$ *be given as above (i.e.,* σ_i *is* ρ_i*-Lipschitz and* $\sigma_i(0) = 0$*). Then for* $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ *drawn iid from any probability distribution over* $\mathbb{R}^d \times \{1, \ldots, k\}$*, with probability at least* $1 - \delta$ *over* $((x_i, y_i))_{i=1}^n$*, every margin* $\gamma > 0$ *and network* $F_\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^k$ *with weight matrices* $\mathcal{A} = (A_1, \ldots, A_L)$ *satisfy*

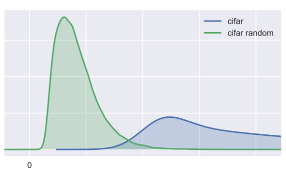
$$\Pr\left[\arg\max_j F_\mathcal{A}(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(F_\mathcal{A}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_2 R_\mathcal{A}}{\gamma n}\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right),$$

where $\widehat{\mathcal{R}}_\gamma(f) \leq n^{-1}\sum_i \mathbb{1}\left[f(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} f(x_i)_j\right]$ *and* $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

Informally:

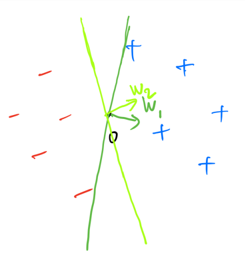
$$F\text{'s generalization error} \leq F\text{'s margin error at } \gamma + O\left(\frac{1}{\gamma}\sqrt{\frac{1}{m}}\right)$$

Normalized margin distribution is a reasonable indicator of generalization performance for neural networks:



Support vector machines: From bounds to algorithms

- Suppose D is realizable wrt $\mathcal{H} = \{h_w(x) := \mathsf{sign}(\langle w, x\rangle) : w \in \mathbb{R}^d\}$
- Given S a set of iid m training examples from D, how to best pick a $w \in \mathbb{R}^d$ such that $\mathbb{P}_D(y\langle w, x\rangle \leq 0)$ is small?



- Idea: Fix $\theta = 1$, pick w such that $\mathbb{P}_S(y\langle w, x\rangle \leq 1) = 0$, and $\|w\|_2$ is as small as possible
- Direction w_2 is "better" than w_1, as it requires a smaller scaling factor $\alpha > 0$ to ensure $\mathbb{P}_S(y\langle \alpha w, x\rangle \leq 1) = 0$

This motivates the optimization problem:

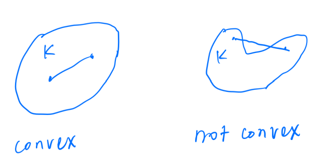
$$\min_{w \in \mathbb{R}^d} \|w\|_2$$
$$\text{subject to: } y_i \langle w, x_i \rangle \geq 1, \forall i \in \{1, \ldots, m\},$$

called the Support Vector Machine (SVM) problem. Remarks:

1. This is a convex optimization problem: convex objective function, convex constraint set
2. Equivalently, the objective function can be replaced with $\frac{1}{2}\|w\|_2^2$
3. If we minimize $\|w\|_1$ instead, this is called ℓ_1-SVM problem

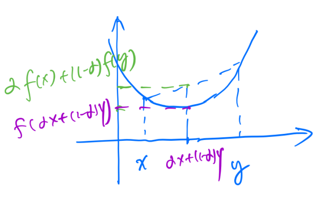
Convex optimization basics

- K is said to be a convex set, if for every $x, y \in K$ and $\alpha \in (0, 1)$, $\alpha x + (1 - \alpha)y \in K$



- f is said to be a convex function with domain C, if for all $x, y \in C$, and $\alpha \in (0, 1)$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$



Convex optimization basics (cont'd)

Optimization problems of the form:

$$\min_{x \in \mathbb{R}^d} f(w)$$
$$\text{subject to: } \; x \in C$$

is said to be a convex optimization problem, if f and C are convex. Convex optimization problems is a class of "easy" optimization problems, which admits efficient solvers (e.g. CVXPY)

SVM: generalization properties

Corollary
Fix $R_2 > 0$, and margin $\gamma \in (0, R_2]$. D is a distribution, such that

1. *it is supported on $\left\{x \in \mathbb{R}^d : \|x\|_2 \leq R_2\right\}$;*
2. *there exists a unit vector w^* that satisfies $\mathbb{P}_D(y \langle w^*, x\rangle \leq \gamma) = 0$.*

Then, with probability $1 - \delta$ over the draw of training examples S, the (ℓ_2-)SVM solution $\hat{w}$ satisfies that:

$$\mathbb{P}_D(y \langle \hat{w}, x\rangle \leq 0) \leq O\left(\frac{R_2}{\gamma}\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$$

Proof sketch.

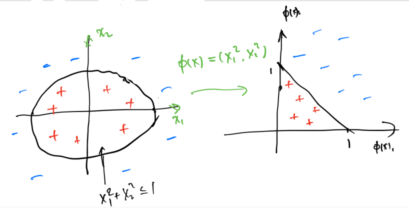
- $\frac{w^*}{\gamma}$ is a feasible solution of the SVM optimization problem $\implies$ $\|\hat{w}\|_2 \leq \|\frac{w^*}{\gamma}\|_2 = \frac{1}{\gamma}$
- Use ℓ_2/ℓ_2 margin bound on $\hat{w}$ and $\theta = 1$. $\square$

SVM: practical considerations

- In practice, data is rarely linearly separable
- Two general ways to cope with linear non-separability:
    - Introducing nonlinear feature maps (basis functions)
    - Modifying the SVM optimization problem by allowing some examples to be incorrectly classified

SVM with nonlinear feature maps

- Define $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$, $(x_i, y_i) \to (\phi(x_i), y_i)$



- $\hat{w} \in \mathbb{R}^{d'} \leftarrow$ Solve SVM on $(\phi(x_i), y_i)_{i=1}^{d'}$
- Final predictor: on x, predict $\hat{h}(x) = \text{sign}(\langle \hat{w}, \phi(x) \rangle)$
- There are SVM solvers that has time complexity independent of d' and outputs a implicit representation of $\hat{h}$, using the so-called "kernel trick"

SVM with soft margins

- Introducing a "slack variable" ξ_i for each example i:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{m} \xi_i$$

$$\text{subject to: } y_i \langle w, x_i \rangle \geq 1 - \xi_i, \forall i \in \{1, \ldots, m\},$$

$$\xi_i \geq 0, \forall i \in \{1, \ldots, m\}$$

- $\lambda \downarrow \implies$ penalizes ξ_i harder
- Try eliminating variable ξ_i: for any fixed w, the optimal ξ_i is such that

$$\min_{\xi_i} \xi_i, \text{ s.t. } \xi_i \geq 0 \wedge \xi_i \geq 1 - y_i \langle w, x_i \rangle,$$

  i.e. $\xi_i = \max(0, 1 - y_i \langle w, x_i \rangle) =: (1 - y_i \langle w, x_i \rangle)_+$; so soft-margin SVM problem is equivalent to

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{\text{complexity regularizer}} + \underbrace{\sum_{i=1}^{m} (1 - y_i \langle w, x_i \rangle)_+}_{\text{empirical risk}}$$

Regularized loss minimization: general formulations

$$\min_{w \in \mathbb{R}^d} \underbrace{\lambda \cdot R(w)}_{\text{complexity regularizer}} + \underbrace{\sum_{i=1}^{m} \ell(f_w(x_i), y_i)}_{\text{empirical risk}}$$

Popular choices of:

- $R(w)$: $\|w\|_1$, $\|w\|_2^2$, $\sum_{i=1}^d w_i \ln w_i$ (negative entropy)
- $f_w(x)$: $\langle w, x \rangle$ (linear), $\langle w_2, \sigma(W_1 x) \rangle$ (one-hidden-layer network)
- $\ell(\hat{y}, y)$:
  - for regression: $|\hat{y} - y|^p$,
  - for classification: $\phi(y \cdot \hat{y})$, where $\phi(z)$ can take e^{-z} (boosting), $(1-z)_+$ (SVM), $\ln(1+e^{-z})$ (logistic regression), etc

What have we learned?

- Margin-based generalization error bounds for linear classifiers
- ℓ_1/ℓ_∞ vs. ℓ_2/ℓ_2 bounds
- Using margin theory to guide the design of practical algorithms: SVMs and regularized loss minimization

Why are large-margin distributions easy to learn? An alternative perspective

- Let $\mathcal{X} \subset \left\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\right\}$
- D is linearly separable over $\mathcal{X} \times \{\pm 1\}$ with margin $\gamma > 0$, i.e. there exists w^*, such that $\|w^*\|_2 \leq 1$ and with probability 1,

$$y \langle w^*, x \rangle \geq \gamma.$$

- SVM over m iid training examples $\implies$ $\hat{w}$, such that with high probability,

$$\mathbb{P}_D(y \langle \hat{w}, x \rangle \leq 0) \leq \cancelto{0}{\mathbb{P}_S(y \langle \hat{w}, x \rangle \leq 1)} + O\left(\frac{1}{\gamma}\sqrt{\frac{1}{m}}\right)$$

- Recall: if $\frac{1}{\gamma^2} \ll d$, this is much lower than VC-based generalization error bound

- A matrix $A \in \mathbb{R}^{l \times d}$ can be viewed as a transformation ϕ_A,

$$\phi_A : \mathbb{R}^d \to \mathbb{R}^l, \quad x \mapsto Ax$$

  $l \ll d \implies$ dimensionality reduction
- Given D, let D_A be the joint distribution of (Ax, y), where $(x, y) \sim D$

Lemma (Johnson-Lindenstrauss)
Given the setting as above; let $l = O(\frac{\log |\mathcal{X}|}{\gamma^2})$. *With high probability over the draw of a random matrix* A *(from some fixed distribution), the distribution* D_A *is still linearly separable.*

We call such matrix A a *J-L transform*.

Large-margin linear learner using J-L transform

Algorithm:

- Input: training examples $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$
- Generate a random J-L transform $A \in \mathbb{R}^{l \times d}$
- Transformed training data $S_A = \langle (Ax_1, y_1), \ldots, (Ax_m, y_m) \rangle$
- Use the consistency algorithm to find a classifier $\hat{w} \in \mathbb{R}^l$ consistent with S_A
- Return classifier $\hat{h}(x) = \mathrm{sign}(\langle \hat{w}, Ax \rangle)$

Analysis:

- By J-L, D_A is linearly separable and $\hat{w}$ has zero training error on S_A
- VC inequality $\implies \mathbb{P}_{(z,y)\sim D_A}(y \langle \hat{w}, z \rangle \leq 0) \leq O(\sqrt{\frac{l}{m}}) = O\left(\frac{1}{\gamma}\sqrt{\frac{1}{m}}\right)$
- Finally, we recognize that

$$\mathbb{P}_{(z,y)\sim D_A}(y \langle \hat{w}, z \rangle \leq 0) = \mathbb{P}_{(x,y)\sim D}(\hat{h}(x) \neq y)$$