# Computational and Statistical Learning Theory

TTIC 31120

**Prof. Nati Srebro**

## Lecture 12:

Weak Learnability and the $\ell_1$ margin

Converse to Scale-Sensitive Learning

Stability

Convex-Lipschitz-Bounded Problems

# Prediction Margin

- For a predictor $h: \mathcal{X} \to \mathbb{R}$ and binary labels $y = \pm 1$:
  - Margin on a single example: $yh(x)$
  - Margin on a training set: $margin(h) = \min_{(x_i, y_i) \in S} y_i h(x_i)$

- Most classification loss functions are a function of the margin:
  - $loss^{mrg}(h(x); y) = \left[\left[margin < 1\right]\right]$
  - $loss^{hinge}(h(x); y) = \left[1 - margin\right]_+$
  - $loss^{logistic}(h(x); y) = \log\left(1 + e^{-margin}\right)$
  - $loss^{exp}(h(x); y) = e^{-margin}$
  - $loss^{sq}(h(x); y) = \left(y - h(x)\right)^2 = (1 - margin)^2$

# Complexity and Margin

- Recall: $\forall^{\delta}_{S\sim\mathcal{D}^m}\forall_{h\in\mathcal{H}}\ L^{01}_{\mathcal{D}}(h) \le L^{mrg}_{S}(h) + \mathcal{R}_S(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{m}}$
- $margin_S(h) = \gamma$ ➔ $L^{mrg}_{S}\left(\frac{h}{\gamma}\right) = 0$

$$L^{01}_{\mathcal{D}}(h) \le L^{mrg}_{S}\left(\frac{h}{\gamma}\right) + \mathcal{R}_S\left(\frac{1}{\gamma}\mathcal{H}\right) + \sqrt{\frac{\log 1/\delta}{m}}$$

$$\le \frac{1}{\gamma}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{m}} \le \sqrt{\left(\frac{B\|\phi(x)\|_2}{\gamma}\right)^2 \cdot \frac{1}{m}} + \sqrt{\frac{\log 1/\delta}{m}}$$

$\mathcal{H} = \{h_w(x) = \langle w, \phi(x)\rangle \mid \|w\|_2 \le B\}$

- $\ell_2$-margin: $\sup_w \frac{margin(w)}{\|w\|_2} = \sup_{\|w\|_2\le 1} margin(w)$

  Even better to consider relative $\ell_2$-margin: $\sup_w \frac{margin(w)}{\|w\|_2 \cdot \sup\|\phi(x)\|_2}$

- $\ell_1$-margin: $\sup_w \frac{margin(w)}{\|w\|_1}$

  Relative $\ell_1$-margin: $\sup_w \frac{margin(w)}{\|w\|_1 \cdot \sup\|\phi(x)\|_\infty}$

# Boosting and the $\ell_1$ Margin

Weak learning: can always find $f$ with $L^{01}(f) \leq \frac{1}{2} - \frac{\gamma}{2}$

$\Downarrow$

After $T = \frac{48 \log 2m}{\gamma^4}$ iteration, AdaBoost finds predictor with $\frac{margin(w)}{\|w\|_1} \geq \frac{\gamma}{2}$

over $\phi(x)[f] = f(x) \in \pm 1$, i.e. $\|\phi(x)\|_\infty = 1$

(and as $T \to \infty$, $\lim \frac{margin(w)}{\|w\|_1} \geq \gamma$)

$\Downarrow$

$\mathcal{B} = \{\text{weak predictors } f\}$

Need $m = \left( \frac{\left(\frac{\gamma}{2}\right)^2 VCdim(\mathcal{B})}{\epsilon^2} \right)$ samples to ensure $L^{01}(w) \leq \epsilon$

- Can we understand AdaBoost purely in terms of $\ell_1$-margin?
  Can we get a guarantee for AdaBoost that replies on existence of large $\ell_1$-margin predictor, instead of on "weak learnability"?
- The AdaBoost analysis shows weak learning ➔ $\ell_1$-margin. Converse?

# Weak Learning and the $\ell_1$ Margin

- Consider a base class $\mathcal{B} = \{f: \mathcal{X} \to \pm 1\}$ and the corresponding feature map $\phi: \mathcal{X} \to \mathbb{R}^{\mathcal{B}}$ defined as $\phi(x)[f] = f(x)$.

- Goal: relate weak learnability using predictors in $\mathcal{B}$ to $\ell_1$-margin using $\phi(x)$

- Weak learnability:

  $h: \mathcal{X} \to \pm 1$ is $\gamma$-weakly learnable using $\mathcal{B}$ if for any distribution $\mathcal{D}(\mathcal{X})$, there exists $f \in \mathcal{B}$ s.t. $\Pr_{x \sim \mathcal{D}}[f(x) = h(x)] \geq \frac{1}{2} + \frac{\gamma}{2}$

- Assume that $\mathcal{B}$ is symmetric, i.e. for any $f \in \mathcal{B}$, also $-f \in \mathcal{B}$
  - This allows us to consider only $w \geq 0$, and so $\|w\|_1 = \sum w[f]$
  - If $w[f] < 0$, instead use $w[-f] > 0$

  (without assuming $\mathcal{B}$ is symmetric, we will need to talk about margin attainable only with $w \geq 0$)

# Weak Learning and the $\ell_1$ Margin

- Best possible $\ell_1$ margin for a labeling $h$:

$$\gamma_1 = \sup_{\|w\|_1 \le 1} \min_x h(x)\langle \phi(x), w \rangle$$

- For finite domain $\mathcal{X} = \{x_1, \ldots, x_n\}$ and finite base class $\mathcal{B}$ (i.e. $\phi(x) \in \mathbb{R}^d$ is finite dimensional) consider matrix $A \in \pm 1^{n \times d}$ with rows

$$A_i = h(x_i)\phi(x_i)$$

- Can write the $\ell_1$ margin as:

$$\gamma_1 = \max_{\|w\|_1 \le 1} \min_i A_i w = \max_{\|w\|_1 \le 1} \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} p' A w$$

  and since $\mathcal{B}$ is symmetric:

$$\gamma_1 = \max_{\substack{w \in \mathbb{R}^d_+ \\ \|w\|_1 = 1}} \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} p' A w$$

# Weak Learning and the $\ell_1$ Margin

- Best possible weak-learnability "edge" for $h: \mathcal{X} \to \pm 1$:

$$\gamma = \min_{\mathcal{D}} \max_{f \in \mathcal{B}} \left( 2 \Pr_{x \sim \mathcal{D}} [h(x) = f(x)] - 1 \right) = \min_{\mathcal{D}} \max_{f \in \mathcal{B}} \sum_{x} \mathcal{D}(x) h(x) f(x)$$

- For a finite domain, and in terms of the matrix with rows $A_i = h(x_i)\phi(x_i)$ and columns $A^j$:

$$\gamma = \min_{\substack{p \in \mathbb{R}_+^n \\ \|p\|_1 = 1}} \max_{j \in 1..d} \sum_i p_i h(x_i) \phi(x_i) = \min_{\substack{p \in \mathbb{R}_+^n \\ \|p\|_1 = 1}} \max_{j \in 1..d} p' A^j$$

AdaBoost

$$\leq = \min_{\substack{p \in \mathbb{R}_+^n \\ \|p\|_1 = 1}} \max_{\substack{w \in \mathbb{R}_+^d \\ \|w\|_1 = 1}} p' A w \geq \max_{\substack{w \in \mathbb{R}_+^d \\ \|w\|_1 = 1}} \min_{\substack{p \in \mathbb{R}_+^n \\ \|p\|_1 = 1}} p' A w = \gamma_1$$

# Weak Learning and the $\ell_1$ Margin

- Best possible weak-learnability "edge" for $h: \mathcal{X} \to \pm 1$:

$$\gamma = \min_{\mathcal{D}} \max_{f \in \mathcal{B}} \left( 2 \Pr_{x \sim \mathcal{D}} [h(x) = f(x)] - 1 \right) = \min_{\mathcal{D}} \max_{f \in \mathcal{B}} \sum_{x} \mathcal{D}(x) h(x) f(x)$$

- For a finite domain, and in terms of the matrix with rows $A_i = h(x_i)\phi(x_i)$ and columns $A^j$:

$$\gamma = \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} \max_{j \in 1..d} \sum_i p_i h(x_i) \phi(x_i) = \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} \max_{j \in 1..d} p' A^j$$

$$= \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} \max_{\substack{w \in \mathbb{R}^d_+ \\ \|w\|_1 = 1}} p' A w = \max_{\substack{w \in \mathbb{R}^d_+ \\ \|w\|_1 = 1}} \min_{\substack{p \in \mathbb{R}^n_+ \\ \|p\|_1 = 1}} p' A w = \gamma_1$$

**Strong duality**

# Weak Learning and the $\ell_1$ Margin

- Conclusion:

  $\gamma$-weakly learnable using predictors from base class $\mathcal{B}$

  (i.e. for any distribution, can get error $\leq \frac{1}{2} - \frac{\gamma}{2}$ using predictor for $\mathcal{B}$)

  if and only if

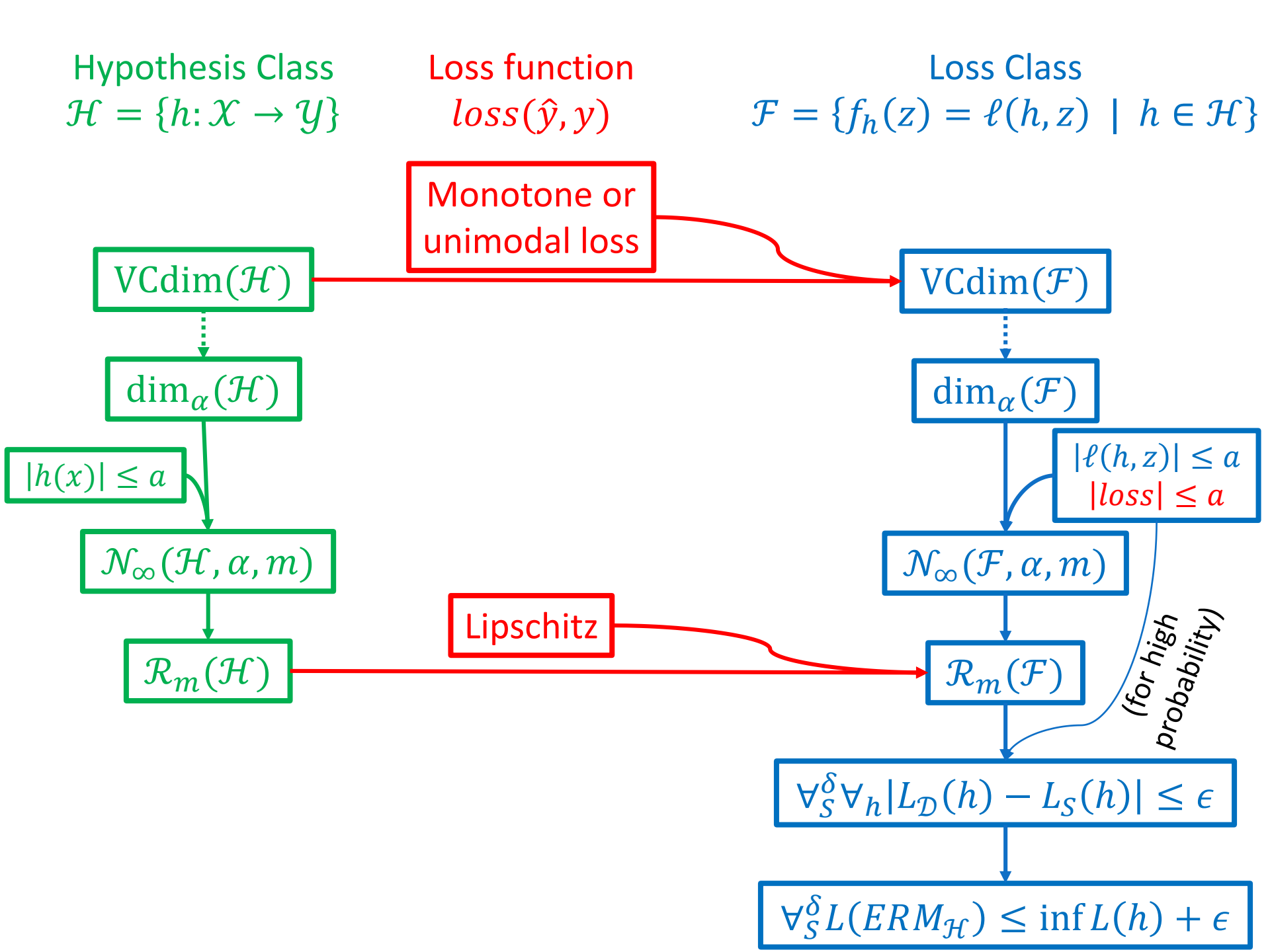
  realizable with $\ell_1$ margin $\gamma$ using $\phi(x) = \left(h(x)\right)_{h \in \mathcal{B}}$

  (i.e. there exists $\|w\|_1 = {}^1/_\gamma$ with $L^{mrg}(x \mapsto \langle w, \phi(x)\rangle) = 0$)

- AdaBoost can be viewed as an algorithm for maximizing the $\ell_1$ margin:
  $\exists_w \frac{margin_S(w)}{\|w\|_1} \geq \gamma$ ➔ AdaBoost finds $\frac{margin_S(w)}{\|w\|_1} \geq \frac{\gamma}{2}$ in $O\left(\frac{\log(m)}{\gamma^4}\right)$ steps,
  and eventually converges to the maximal $\ell_1$ margin solution.

# Loss, Regularizer and Efficient Representation

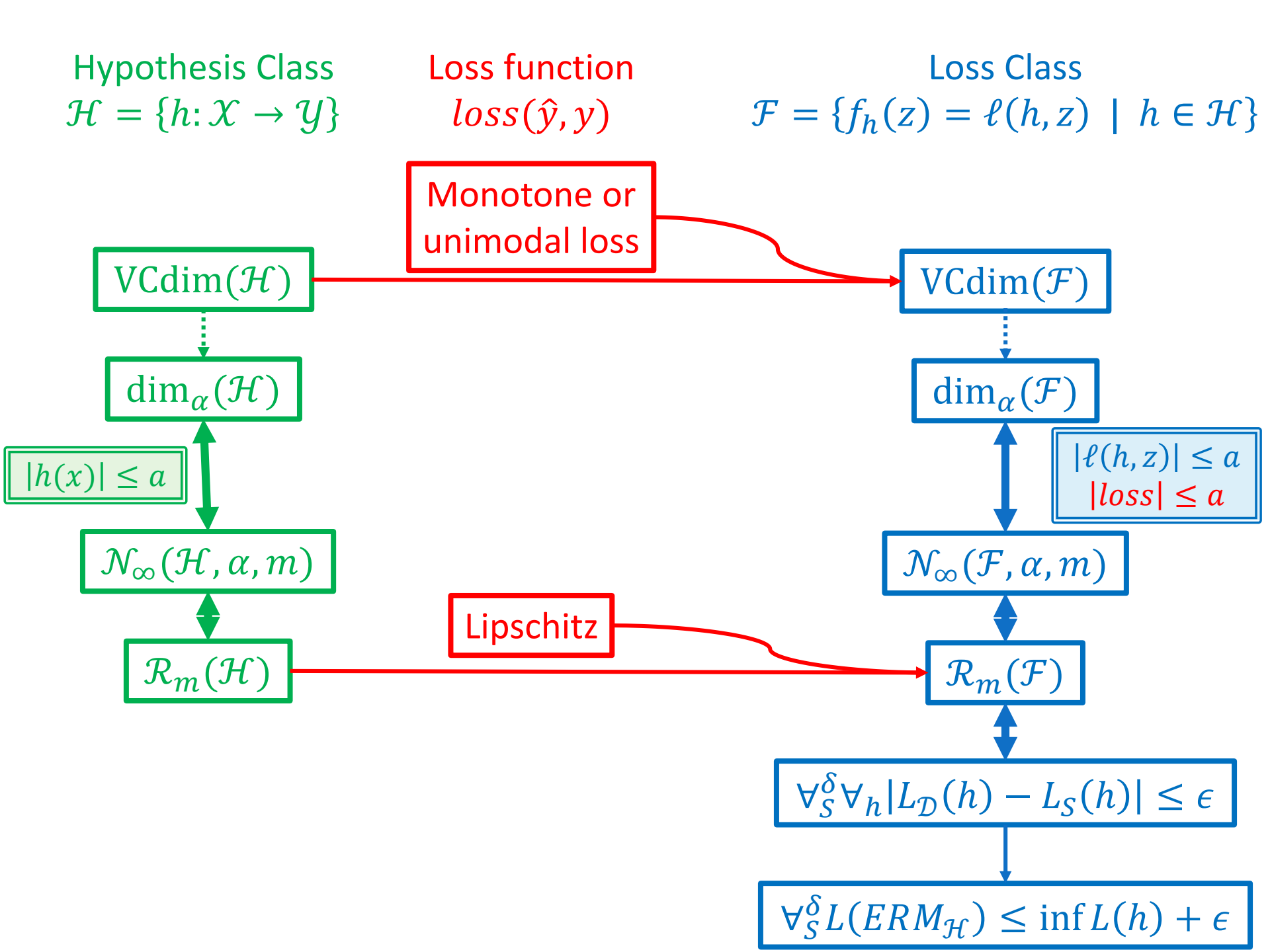
- SVM:
  - $\ell_2$ regularization ➔ dimension independent generalization
  - Hine loss
  - Represent infinite dimensional space via kernels
- Boosting:
  - $\ell_1$ regularization
    ➔ sample complexity depends on $\log(d)$ or $VCdim(\text{features})$
  - Exp-loss / hard margin
  - Represent infinite dimensional space via "weak learning oracle", i.e. oracle for finding high-derivative feature


Hypothesis Class
$\mathcal{H} = \{h: \mathcal{X} \to \mathcal{Y}\}$
Loss function
$loss(\hat{y}, y)$
Loss Class
$\mathcal{F} = \{f_h(z) = \ell(h, z) \mid h \in \mathcal{H}\}$
Monotone or unimodal loss
$\mathrm{VCdim}(\mathcal{H})$
$\mathrm{VCdim}(\mathcal{F})$
$\dim_\alpha(\mathcal{H})$
$\dim_\alpha(\mathcal{F})$
$|h(x)| \leq a$
$|\ell(h, z)| \leq a$
$|loss| \leq a$
$\mathcal{N}_\infty(\mathcal{H}, \alpha, m)$
$\mathcal{N}_\infty(\mathcal{F}, \alpha, m)$
Lipschitz
$\mathcal{R}_m(\mathcal{H})$
$\mathcal{R}_m(\mathcal{F})$
(for high probability)
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \leq \epsilon$
$\forall_S^\delta L(ERM_\mathcal{H}) \leq \inf L(h) + \epsilon$

# Converse: ULLN

- For bounded loss, the following are equivalent:
  - Finite fat-shattering dimension at every scale $\alpha > 0$
  - Finite covering number at every scale $\alpha > 0$
  - Radamacher complexity $\mathcal{R}_m \to 0$ as $m \to \infty$
  - $\sup_f |\mathbb{E}f - \mathbb{E}_S f| \to 0$ as $m \to \infty$

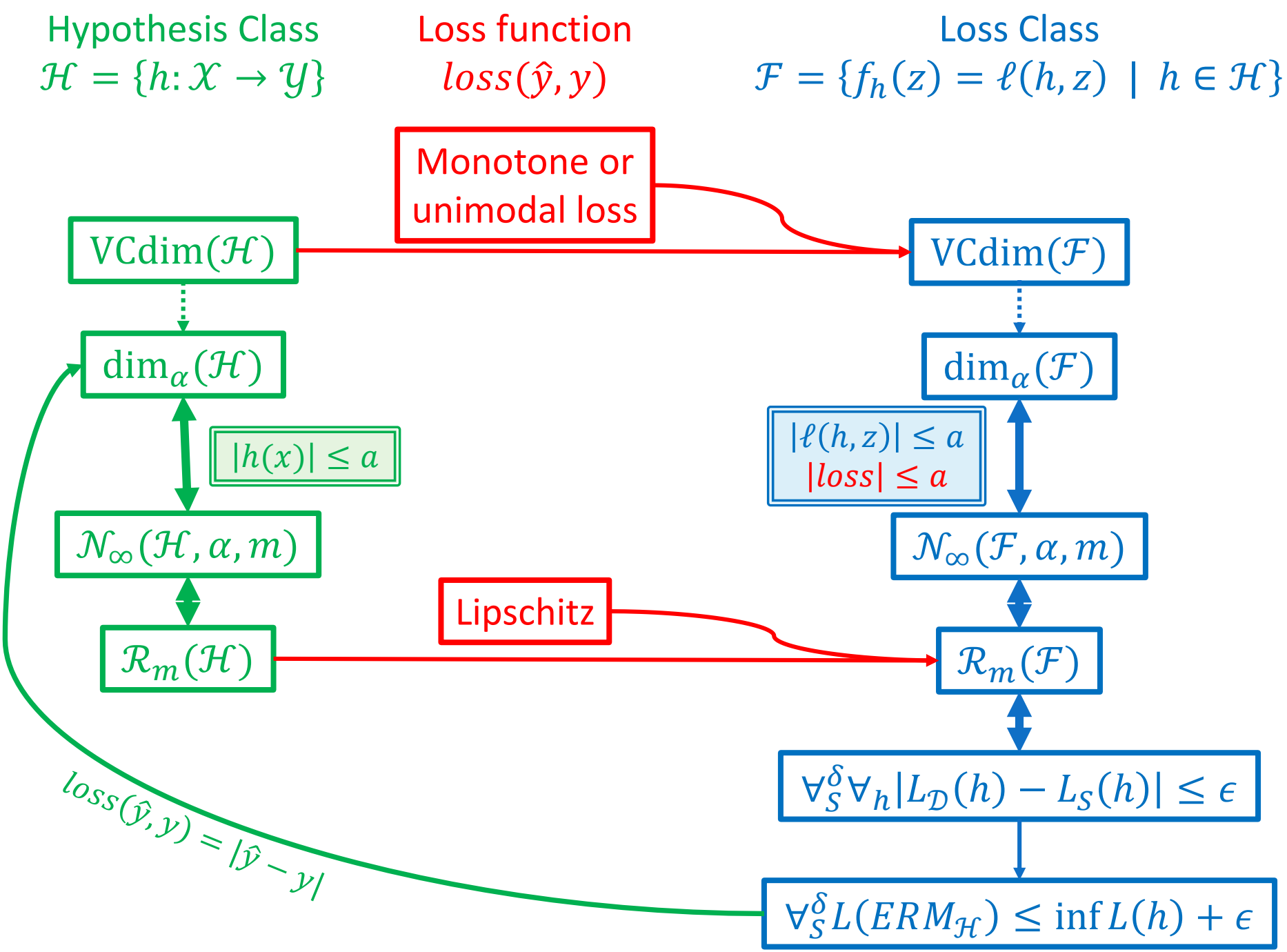
  (and equivalent quantitatively, up to log-factors)


Hypothesis Class
$\mathcal{H} = \{h: \mathcal{X} \to \mathcal{Y}\}$
Loss function
$loss(\hat{y}, y)$
Loss Class
$\mathcal{F} = \{f_h(z) = \ell(h, z) \mid h \in \mathcal{H}\}$
Monotone or unimodal loss
$\mathrm{VCdim}(\mathcal{H})$
$\mathrm{VCdim}(\mathcal{F})$
$\dim_\alpha(\mathcal{H})$
$\dim_\alpha(\mathcal{F})$
$|h(x)| \le a$
$|\ell(h, z)| \le a$
$|loss| \le a$
$\mathcal{N}_\infty(\mathcal{H}, \alpha, m)$
$\mathcal{N}_\infty(\mathcal{F}, \alpha, m)$
Lipschitz
$\mathcal{R}_m(\mathcal{H})$
$\mathcal{R}_m(\mathcal{F})$
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \le \epsilon$
$\forall_S^\delta L(ERM_\mathcal{H}) \le \inf L(h) + \epsilon$

# Fundamental Theorem of (Real Valued) Learning

- Finite fat-shattering dimension $\dim_\alpha(\mathcal{H})$
  ➔ learnable, with sample complexity $\propto \dim_\alpha(\mathcal{H})$
- Can't expect converse for any loss function
  - E.g. trivial loss $loss(\hat{y}; y) = 0$
  - Or partially trivial: ramp loss, $\dim_\alpha(\mathcal{H}) = \infty$, but $\forall_{h\in\mathcal{H},x\in\mathcal{X}} h(x) > 5$
- Focus on $loss(\hat{y}, y) = |\hat{y} - y|$
- **Theorem**: With $loss(\hat{y}, y) = |\hat{y} - y|$, for any $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, any learning rule $A$ and any $\alpha > 0$, there exists $\mathcal{D}$ and $h \in \mathcal{H}$, $L_{\mathcal{D}}(h) = 0$, but with $m < \frac{1}{4}\dim_\alpha(\mathcal{H})$ samples, $\mathbb{E}\left[L\left(A(S)\right)\right] > \frac{\alpha}{4}$

  i.e. sample complexity to get error $\frac{\alpha}{4}$ is at least $\propto \dim_\alpha(\mathcal{H})$
- **Conclusion**:
  - Fat-shattering dimension tightly characterizes learnability
  - If learnable, learnable using ERM with near-optimal sample complexity.


Hypothesis Class
$\mathcal{H} = \{h: \mathcal{X} \to \mathcal{Y}\}$
Loss function
$loss(\hat{y}, y)$
Loss Class
$\mathcal{F} = \{f_h(z) = \ell(h, z) \mid h \in \mathcal{H}\}$
Monotone or unimodal loss
$\mathrm{VCdim}(\mathcal{H})$
$\mathrm{VCdim}(\mathcal{F})$
$\dim_\alpha(\mathcal{H})$
$\dim_\alpha(\mathcal{F})$
$|h(x)| \le a$
$|\ell(h, z)| \le a$
$|loss| \le a$
$\mathcal{N}_\infty(\mathcal{H}, \alpha, m)$
$\mathcal{N}_\infty(\mathcal{F}, \alpha, m)$
Lipschitz
$\mathcal{R}_m(\mathcal{H})$
$\mathcal{R}_m(\mathcal{F})$
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \le \epsilon$
$loss(\hat{y}, y) = |\hat{y} - y|$
$\forall_S^\delta L(ERM_\mathcal{H}) \le \inf L(h) + \epsilon$

# General Learning Setting

$$\min_{h \in \overline{\mathcal{H}}} L(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)]$$

- Is learnability equivalent to finite fat-shattering dimension?

- Consider $\mathcal{Z} = \mathbb{R}, \overline{\mathcal{H}} = \{h: \mathbb{R} \to \mathbb{R}\}, \ell(h, z) = h(z)$
  $$\mathcal{H} = \{h(z) = 0\} \cup \{h: \mathbb{R} \to \mathbb{R} \mid 1 \leq h \leq 2\}$$
- $\dim_\alpha(\mathcal{H}) = \infty$ for $\alpha < {}^1/_2$
- But: $\mathrm{ERM}(S)(z) = 0$ "learns" with excess error 0!

… If learnable, can we always learn with ERM?

# A Different Approach: Stability

- **Definition**: A learning rule $A: S \mapsto h$ is (uniformly replacement) **stable** with rate $\beta(m)$ if, for all $z_1, \dots, z_m$ and $z_i'$:

$$|\ell(A(z_1, \dots, z_m), z_i) - \ell(A(\underbrace{z_1, \dots, z_i', \dots, z_m}_{S_i'}), z_i)| \leq \beta(m)$$

- **Theorem**: If A is stable with rate $\beta(m)$ then $\forall_{\mathcal{D}}$:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A(S)\right)\right] \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S\left(A(S)\right)\right] + \beta(m)$$

**Proof**:

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A(S)\right)\right] &= \mathbb{E}[\ell(A(z_1, \dots, z_i', \dots, z_m), z_i)] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\ell(A(z_1, \dots, z_i', \dots, z_m), z_i)] \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\ell(A(z_1, \dots, z_m), z_i)] + \beta(m)\right) \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \ell(A(S), z_i)\right] + \beta(m) = \mathbb{E}\left[L_S\left(A(S)\right)\right] + \beta(m)
\end{aligned}$$

# Stability of Linear Predictors?

supervised learning: $z = (x, y)$, $\ell\big(h, (x, y)\big) = loss(h(x), y)$

$$\mathcal{X} = \{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\}, \mathcal{Y} = [-1,1], loss(\hat{y}, y) = |\hat{y} - y|$$
$$\mathcal{H} = \{x \mapsto \langle w, x \rangle \mid \|w\|_2 \leq 2\}$$

- Is $A(S) = ERM_{\mathcal{H}}(S)$ stable?
- For any $m$, consider:
  - $x_1 = x_2 = \cdots = x_{m-1} = (1,0),\ y_1 = y_2 = \cdots = y_{m-1} = 1$
  - $x_m = (0,1)$, $y_m = 1$, which is replaced with $x'_m = (0,1)$, $y'_m = -1$

  ➔ $A(S) = (1,1)$ and $\ell(A(S), z_m) = 1$, but $A(S'_m) = (1,-1)$ and $\ell(A(S'_m), z_m) = 2$
- $ERM_{\mathcal{H}}$ does not have stability better than 2 (worst possible), even as $m \to \infty$

# Stability and Regularization

- Consider instead

$$RERM_\lambda(S) = \arg\min_w L_S(w) + \lambda\|w\|_2^2$$

over $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_2 \le R\}$ with $loss(\hat{y}, y) = |\hat{y} - y|$

- Claim: $RERM_\lambda$ is $\beta(m) = \frac{2R^2}{\lambda m}$ stable
- How can we use this to learn $\mathcal{H}_B = \{w \mid \|w\|_2 \le B\}$ ?

$$\mathbb{E}\left[L_{\mathcal{D}}\left(RERM_\lambda(S)\right)\right] \le \mathbb{E}\left[L_S\left(RERM_\lambda(S)\right)\right] + \beta(m)$$

$$\le \mathbb{E}\left[L_S\left(RERM_\lambda(S)\right) + \lambda\|RERM_\lambda(S)\|_2^2\right] + \beta(m)$$

$$\le \mathbb{E}[L_S(w) + \lambda\|w\|_2^2] + \beta(m) = L_{\mathcal{D}}(w) + \lambda\|w\|_2^2 + \frac{2R^2}{\lambda m}$$

$$\le \inf_{\|w\|_2 \le B} L(w) + \lambda B^2 + \frac{2R^2}{\lambda m} = \inf_{\|w\|_2 \le B} L(w) + \sqrt{\frac{8B^2R^2}{m}}$$

$$\lambda = \sqrt{2R^2/B^2m}$$

# Two Views of Regularization

## Uniform Convergence

- Limiting to $\|w\| \leq B$ ensure uniform convergence of $L_s(w)$ to $L(w)$
- Motivates $ERM_B(S) = \arg\min_{\|w\| \leq B} L_S(w)$
- SRM variant, balancing complexity and approximation, is $RERM_\lambda(S)$

## Stability

- Adding a regularizer ensures stability, and thus generalization
- Motivates $RERM_\lambda(S) = \arg\min L_S(w) + \lambda\|w\|^2$
- To learn $\|w\| \leq B$, use $\lambda \propto \frac{1}{B\sqrt{m}}$

- We still need to prove stability!
- We will consider broader class of generalized learning problems with Lipschitz objective

# Convex-Bounded-Lipschitz Problems

- For a generalized learning problem
$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{z \sim \mathcal{D}}[\ell(w, z)]$$
with domain z $\in \mathcal{Z}$ and a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^d$, we say:
- The problem is **convex** if for every $z$, $\ell(w, z)$ is convex in $w$
- The problem is ***G*-Lipschitz** if for every $z$, $\ell(w, z)$ is $G$-Lipschitz in $w$:
$$\forall_{z \in \mathcal{Z}} \forall_{w, w' \in \mathcal{H}} |\ell(w, z) - \ell(w', z)| \leq G \cdot \|w - w'\|_2$$
- Or ***G*-Lipschitz with respect to a norm $\|\boldsymbol{w}\|$**:
$$\forall_{z \in \mathcal{Z}} \forall_{w, w' \in \mathcal{H}} |\ell(w, z) - \ell(w', z)| \leq G \cdot \|w - w'\|$$
- The problem is ***B*-bounded w.r.t norm $\|\boldsymbol{w}\|$** if $\forall_{w \in \mathcal{H}} \|w\| \leq B$

For simplicity we write $w \in \mathbb{R}^d$. Actually, we can consider $w \in \mathcal{W}$ for some Banach space (normed vector space) $\mathcal{W}$ with norm $\|w\|$

# Linear Prediction as a Generalized Lipschitz Problem

$$z = (x, y) \in \mathcal{X} \times \mathcal{Y},\ \phi: \mathcal{X} \to \mathbb{R}^d,\ loss: \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$$
$$\ell\big(w, (x, y)\big) = loss(\langle w, \phi(x)\rangle; y)$$

- If $loss(\hat{y}; y)$ is convex in $\hat{y}$, the problem is convex

- If $loss(\hat{y}; y)$ is $g$-Lipschitz in $\hat{y}$ (as a scalar function):

$$\begin{aligned}\left|\ell\big(w, (x, y)\big) - \ell\big(w', (x, y)\big)\right| &= |loss(\langle w, \phi(x)\rangle; y) - loss(\langle w', \phi(x)\rangle; y)| \\ &\leq g \cdot |\langle w, \phi(x)\rangle - \langle w', \phi(x)\rangle| = g \cdot |\langle w - w', \phi(x)\rangle| \\ &\leq g\|\phi(x)\|_2 \cdot \|w - w'\|_2\end{aligned}$$

➔ If $\|\phi\|_2 \leq R$, then the problem is $G = gR$ Lipschitz (w.r.t $\|\cdot\|_2$)

- For any norm $\|w\|$: $\left|\ell\big(w, (x, y)\big) - \ell\big(w', (x, y)\big)\right| \leq g\|\phi(x)\|_* \cdot \|w - w'\|$

➔ If $\|\phi\|_* \leq R$ for the dual norm, then the problem is $G = gR$ Lipschitz

# Stability for Convex Lipschitz Problems

- For a convex $G$-Lipschitz (w.r.t $\|w\|_2$) generalized learning problem, consider

$$RERM_\lambda(S) = \arg\min_w L_S(w) + \lambda\|w\|_2^2$$

- Claim: $RERM_\lambda$ is $\beta(m) = \frac{2G^2}{\lambda m}$ stable
- Proof: homework

- Conclusion: using $\lambda = \sqrt{\frac{2G^2}{B^2 m}}$, can learn any convex $G$-Lipschitz, $B$-bounded Generalized Learning Problem (w.r.t $\|w\|_2$) with sample complexity $O\left(\frac{B^2G^2}{\epsilon^2}\right)$

# Back to Converse of Fundamental Theorem of Learning

- For (bounded) supervised learning problems (with abs loss):
  - Learnable if and only if fat shattering dim at every scale is finite
  - Fat shattering dimension exactly characterizes sample complexity
  - If learnable, we always have ULLN, and always learnable with ERM, with “optimal” sample complexity

- For generalized linear problems:
  - Finite fat shattering dimension ➔ Learnable with ERM
  - No strict converse because of “silly” problems, with complex irrelevant parts
  - Converse for “non trivial” problems?
  - If learnable, always learnable with ERM?

# Center of Mass with Missing Data

- Center of mass (mean estimation) problem:

$$\mathcal{Z} = \{z \in \mathbb{R}^\infty \mid \|z\|_2 \leq 1\}, \mathcal{H} = \{w \in \mathbb{R}^\infty \mid \|w\|_2 \leq 1\}$$

$$\ell(h, z) = \|h - z\|_2 = \sum_i (h[i] - z[i])^2$$

- Center of mass with missing data:

$$\mathcal{Z} = \{(I, z_I) = (I, \{z[i]\}_{i \in I}) \mid \|z\|_2 \leq 1, I \subseteq \text{coordinates}\}$$

$$\ell\big(h, (I, z_I)\big) = \sum_{i \in I} (h[i] - z[i])^2$$

- 4-Lipschitz and 1-Bounded convex problem wrt $\|w\|_2$, hence learnable with $RERM_\lambda$
- But: consider distribution $(I, z_I) \sim \mathcal{D}$ with $\Pr[i \in I] = 1/2$ independently for all $i$, and $z_i = 0$ almost surely.
  - $L_{\mathcal{D}}(0) = 0$
  - For any finite training set, there is (with probability one) some never-observed coordinate $j$. Consider the standard basis vector $e_j$
  - $L_S(e_j) = 0$, hence its an ERM, but $L_{\mathcal{D}}(e_j) = 1/2$

➔ No ULLN, and not learnable with ERM