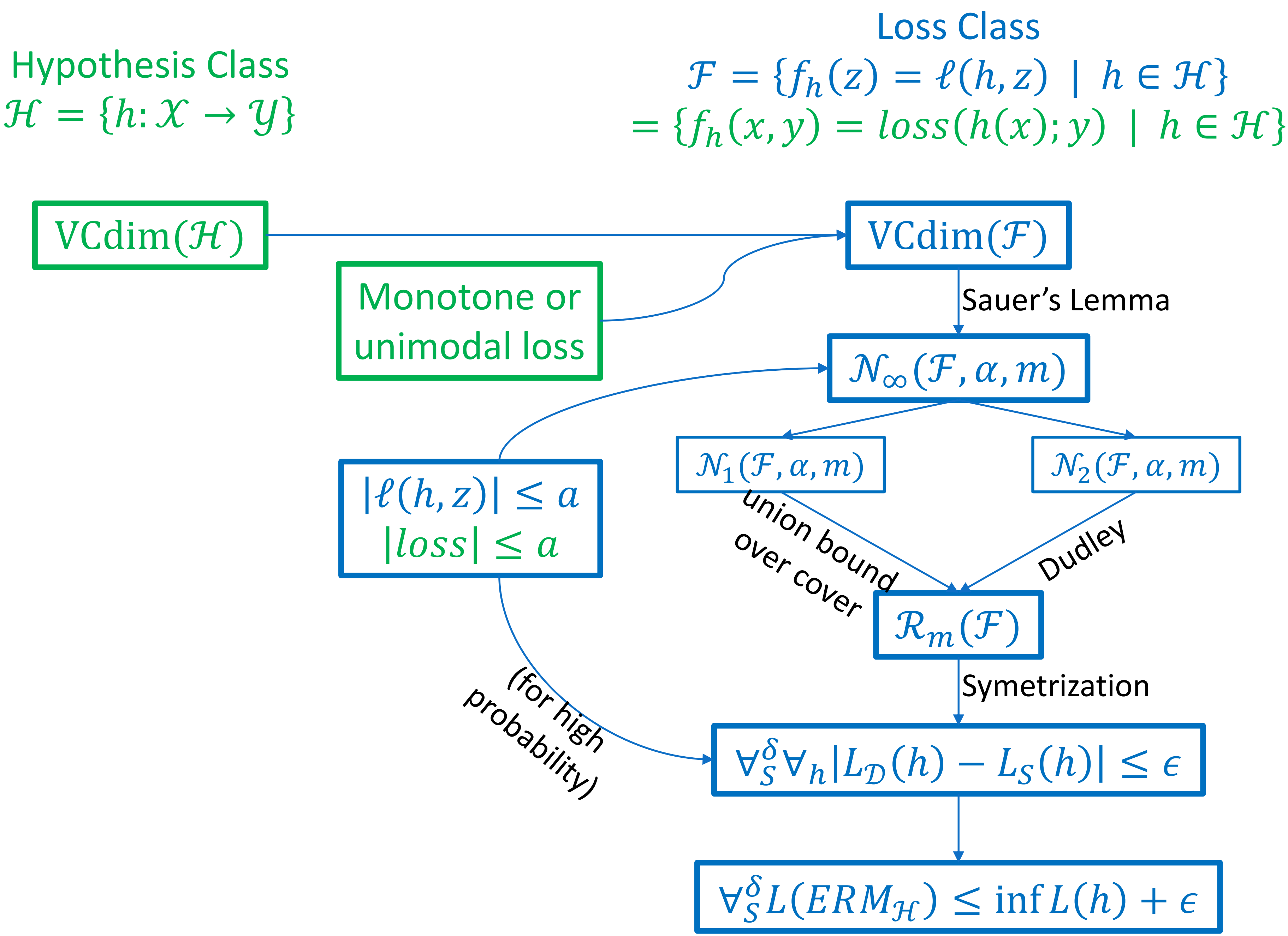
Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 10:

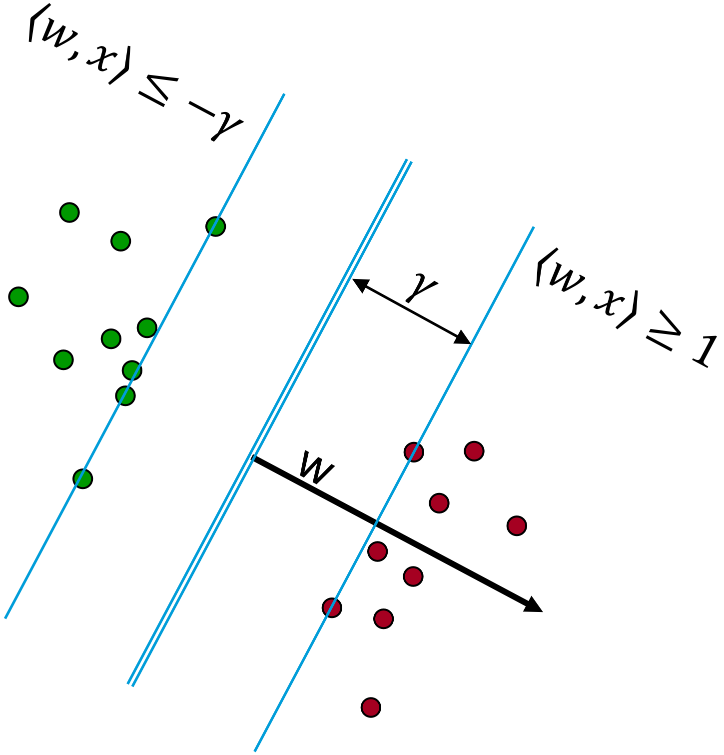
Scale-Sensitive Classes


Hypothesis Class
$\mathcal{H} = \{h: \mathcal{X} \to \mathcal{Y}\}$
Loss Class
$\mathcal{F} = \{f_h(z) = \ell(h,z) \mid h \in \mathcal{H}\}$
$= \{f_h(x,y) = loss(h(x);y) \mid h \in \mathcal{H}\}$
$\mathrm{VCdim}(\mathcal{H})$
Monotone or unimodal loss
$\mathrm{VCdim}(\mathcal{F})$
Sauer's Lemma
$\mathcal{N}_\infty(\mathcal{F},\alpha,m)$
$\mathcal{N}_1(\mathcal{F},\alpha,m)$
$\mathcal{N}_2(\mathcal{F},\alpha,m)$
$|\ell(h,z)| \le a$
$|loss| \le a$
union bound over cover
Dudley
$\mathcal{R}_m(\mathcal{F})$
Symetrization
(for high probability)
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \le \epsilon$
$\forall_S^\delta L(ERM_\mathcal{H}) \le \inf L(h) + \epsilon$

Beyond the VC Dimension

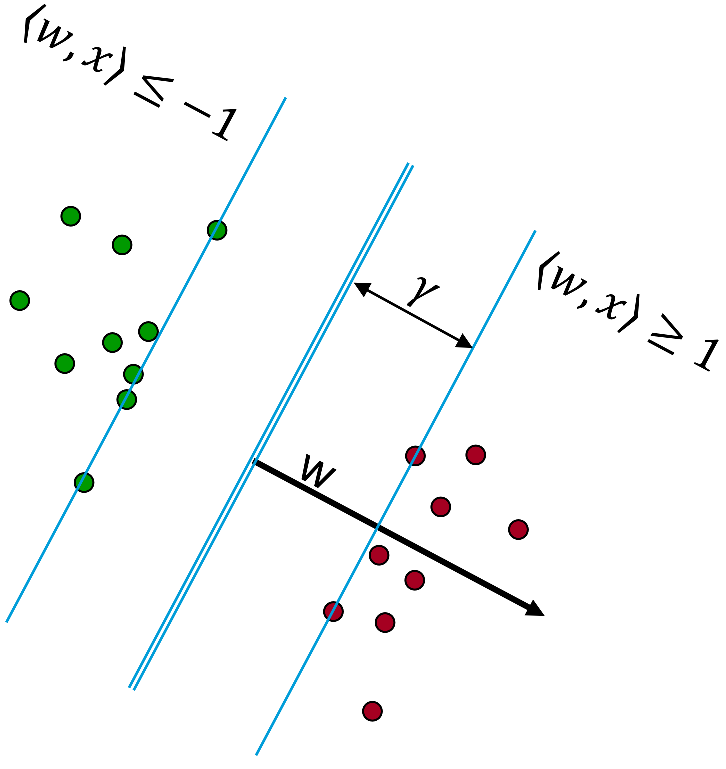
- So far: complexity control only via (VC subgraph) dimension $\approx$ number of parameters
- What is the role of the margin?
- Or of norm regularization, as in SVMs, LASSO, etc?

Reminder: Support Vector Machines



$\|w\| = 1$

Reminder: Support Vector Machines

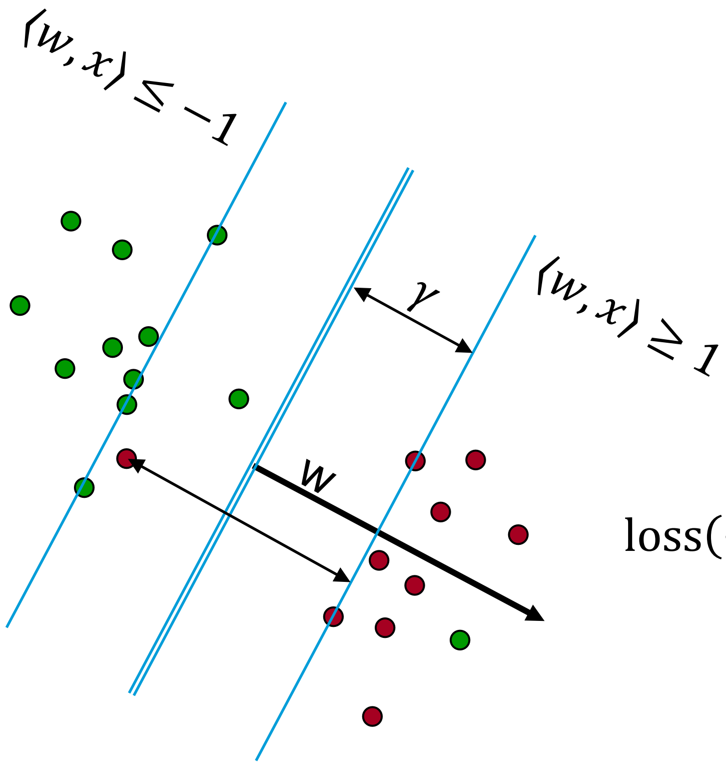


Margin: $\gamma = 1/\|w\|$

Reminder: Support Vector Machines

$$\min \quad \|w\|, \quad L_S(w)$$

Possibly serliazed as: $\min \frac{\lambda}{2}\|w\|^2 + L_S(w)$



Margin: $\gamma = 1/\|w\|$

$\text{loss}(\langle w, x\rangle; y) = [[y\langle w, x\rangle < 1]]$

Or: $[1 - y\langle w, x\rangle]_+$

Norm Constrained Linear Predictors

$$\mathcal{H}_B = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B \}$$

- What is the VC-subgraph dimension of $\mathcal{H}_B$?
  - Can shatter the d standard basis vectors $e_1, e_2, \dots, e_d$ with thresholds $\theta_1 = \theta_2 = \cdots = 0$ and arbitrarly small norm
  - For labels $y_1, \dots, y_d$, set $w = \frac{B}{\sqrt{d}}(y_1, y_2, \dots, y_d)$
  - $\mathrm{VCdim}(\mathcal{H}_B) = d$ (for any $B > 0$)
- VC-subgraph dimension, and Pollard's notion of shattering not relevant.
- Covering numbers still relevant and can depend on B
- How can we bound the covering number in this case?

Fat-Shattering Dimension

- **Definition**: $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ **$\boldsymbol{\alpha}$-shatters** $S = \{z_1, \dots, z_m\}$ if $\exists_{\theta_1,\theta_2,\dots,\theta_m \in \mathbb{R}}$ s.t. $\forall_{y_1,y_2,\dots,y_m \in \pm 1} \exists_{f \in \mathcal{F}}$ s.t. $\forall_i$:
$$y_i = +1 \Rightarrow f(z_i) > \theta_i + \alpha$$
$$y_i = -1 \Rightarrow f(z_i) < \theta_i - \alpha$$

- **Definition**: The **fat shattering dimension** $\dim_\alpha(\mathcal{F})$ of $\mathcal{F}$ is the largest m, s.t. there exists $S \in \mathcal{Z}^m$ that it α-shattered by $\mathcal{F}$

- **Theorem**: For $\mathcal{F} = \{f: \mathcal{Z} \to [-a, a]\}$ with $\dim_\alpha(\mathcal{F}) \leq D(\alpha)$:

$$\mathcal{N}_p(\mathcal{F}, \alpha, m) \leq \mathcal{N}_\infty(\mathcal{F}, \alpha, m) \leq \sum_{k=1}^{D(\alpha)} \binom{m}{k} \left(\frac{a}{\alpha}\right)^k \leq \left(\frac{em}{D(\alpha)} \frac{a}{\alpha}\right)^{D(\alpha)}$$

Fat-Shattering of Linear Predictors

$$\mathcal{H}_B = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\}$$

- For $\mathcal{X} = \mathbb{R}^d$
  (i.e. $\mathcal{H}_B = \{f: \mathbb{R}^d \to \mathbb{R} \mid f(x) = \langle w, x \rangle, w \in \mathbb{R}^d, \|w\|_2 \le B\}$)
  - $\dim_\alpha(\mathcal{H}_B) = d$

- For $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \le R\}$
  - $\dim_0(\mathcal{H}_B) = VCdim(\mathcal{H}_B) = d$
  - $\dim_\alpha(\mathcal{H}_B) \le d$, but maybe smaller?

Fat-Shattering Linear Predictors

$$\mathcal{X}_R = \{x \in \mathbb{R}^d \mid \|x\| \le R\} \qquad \mathcal{H}_B = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\}$$

Claim: $\dim_\alpha(\mathcal{H}_B) < \left(\frac{BR}{\alpha}\right)^2$ (as a predictors over $\mathcal{X}_R$)

Proof: Consider $x_1, .., x_m$ that can be α-shattered with thresholds $\theta_1, \ldots, \theta_m$. For every sign pattern $y \in \pm 1^m$ $\exists w(y)$ s.t. $\forall_i y_i(\langle w(y), x_i\rangle - \theta_i) > \alpha$

And so:

$$m\alpha < \sum_i y_i\left(\langle w(y), x_i\rangle - \theta_i\right) = \langle w(y), \sum_i y_i x_i\rangle - \sum_i y_i\theta_i \le \|w\| \|\sum_i y_i x_i\| - \sum_i y_i \theta_i$$

Considering y_i as independent random signs and taking an expectation over them:

$$m\alpha < B \cdot \mathbb{E}_y[\|\sum_i y_i x_i\|] - \cancel{\mathbb{E}_y[\sum_i y_i\theta_i]} \le B\sqrt{\mathbb{E}_y\left[\|\sum_i y_i x_i\|^2\right]}$$

$$= B\sqrt{\mathbb{E}\left[\sum_i \|y_i x_i\|^2 + \sum_{i \ne j}\langle y_i x_i, y_j x_j\rangle\right]} = B\sqrt{\sum_i \mathbb{E}[y_i^2]\|x_i\|^2 + \sum_{i\ne j} \cancel{\mathbb{E}[y_i y_j]}\langle x_i, x_j\rangle} \le BR\sqrt{m}$$

$\Rightarrow m\alpha < BR\sqrt{m} \Rightarrow m < \left(\frac{BR}{\alpha}\right)^2$

Norm-Regularized Linear Predictors

$$\mathcal{X}_R = \{x \in \mathbb{R}^d \mid \|x\| \le R\} \qquad \mathcal{H}_B = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\}$$

$$\boxed{\dim_\alpha(\mathcal{H}_B) \le \left(\frac{BR}{\alpha}\right)^2}$$

$$\boxed{\langle w, x\rangle \le BR}$$

$$\boxed{\log \mathcal{N}_\infty(\mathcal{H}_B, \alpha, m) \le \left(\frac{BR}{\alpha}\right)^2 \log\left(\frac{em\alpha}{BR}\right)}$$

$$\mathcal{R}_m(\mathcal{H}_B) \le \alpha + BR\sqrt{\frac{\left(\frac{BR}{\alpha}\right)^2 \log\left(\frac{em\alpha}{BR}\right)}{2m}}$$

$$\mathcal{R}_m(\mathcal{H}_B) \le 4\alpha_0 + 10\int_{\alpha_0}^{BR}\sqrt{\frac{\left(\frac{BR}{\alpha}\right)^2 \log\left(\frac{em\alpha}{BR}\right)}{m}}\,d\alpha$$

$\alpha = BR/\sqrt[4]{m}$

$\alpha_0 = BR/\sqrt{m}$

$$\boxed{\mathcal{R}_m(\mathcal{H}_B) \le \frac{3BR\sqrt{\log m}}{\sqrt[4]{m}}}$$

$$\boxed{\mathcal{R}_m(\mathcal{H}_B) \le 14\sqrt{\frac{B^2R^2\log^3(m)}{m}}}$$

Directly Bounding the Rademacher Complexity

$$\mathcal{R}_S(\mathcal{H}) = \mathbb{E}_\xi\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \xi_i h(x_i)\right]$$

- $\mathcal{R}_S(\mathcal{H}_B) = \mathbb{E}_\xi\left[\sup_{\|w\|\leq B} \frac{1}{m}\sum_i \xi_i\langle w, x_i\rangle\right] = \frac{1}{m}\mathbb{E}_\xi\left[\sup_{\|w\|\leq B}\langle w, \sum_i \xi_i x_i\rangle\right]$

$$= \frac{1}{m}\mathbb{E}_\xi[B\|\textstyle\sum_i \xi_i x_i\|] \leq \frac{B}{m}\sqrt{\mathbb{E}\left[\left\|\sum_i \xi_i x_i\right\|^2\right]}$$

$$= \frac{B}{m}\sqrt{\textstyle\sum_i \mathbb{E}\left[\xi_i^2\right]\|x_i\|^2 + \sum_{i\neq j}\mathbb{E}[\xi_i\xi_j]\langle x_i, x_j\rangle} = \sqrt{\frac{B^2\left(\frac{1}{m}\sum_i\|x_i\|^2\right)}{m}}$$

- Simpler and tighter (avoids log-factors) than going via fat-shattering
- $\mathcal{R}_S(\mathcal{H}_B)$ only depends on average $\|x_i\|^2$ inside S.
  - Fat-shattering dimension depends on maximum norm in $\mathcal{X}_B$

  ➔ $\mathcal{R}_{\mathcal{D}^m}(\mathcal{H}_B) \leq \sqrt{\frac{B^2\mathbb{E}[\|x\|^2]}{m}}$ (distribution-dependent bound)
- Actually, dependence on $\mathcal{R}_S$ enough:

$$\forall_S^\delta \forall_{f\in\mathcal{F}} |\mathbb{E}_\mathcal{D} f - \mathbb{E}_S f| \leq 2\mathcal{R}_S(\mathcal{F}) + 4a\sqrt{\frac{\log\frac{2}{\delta}}{m}}$$

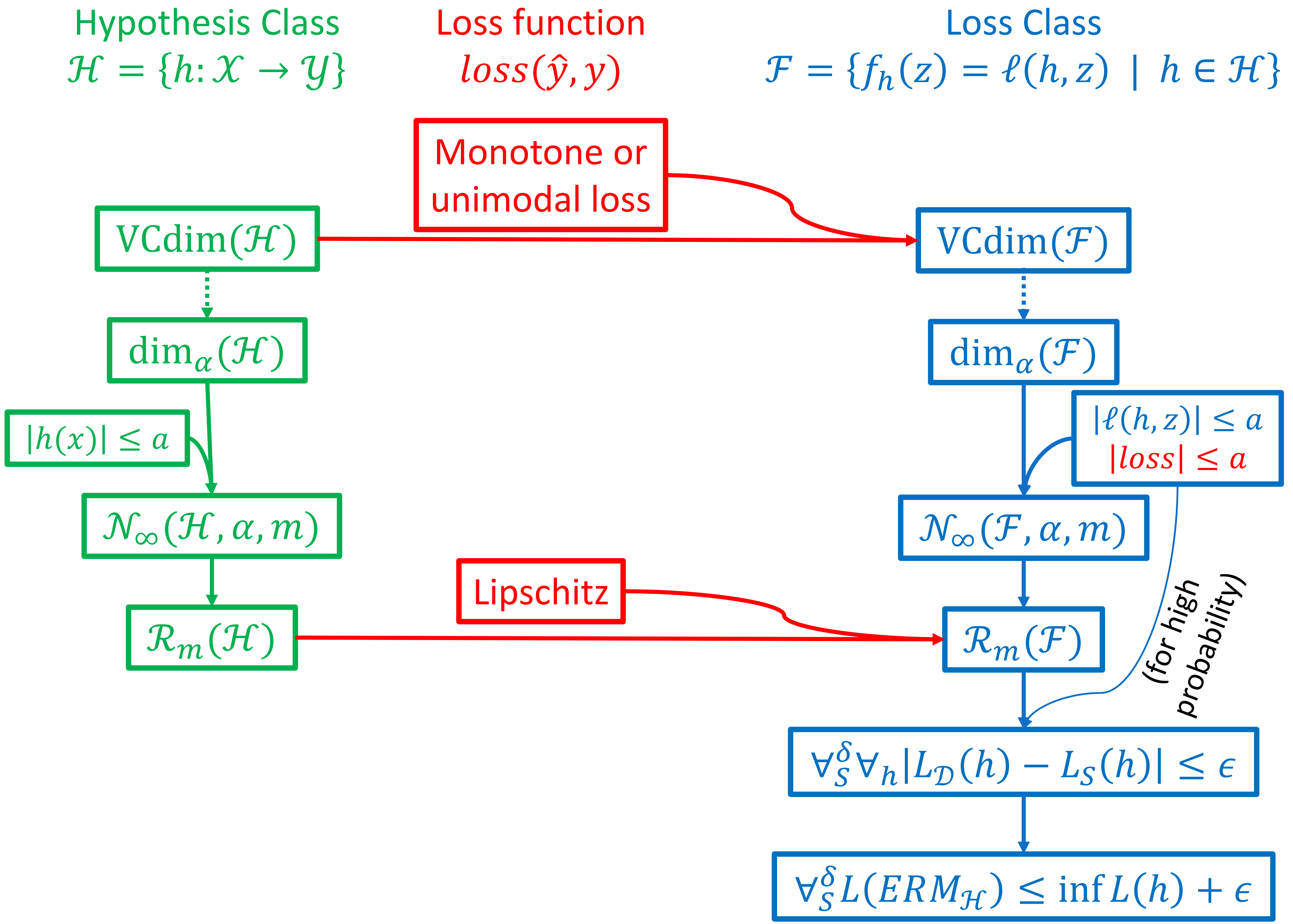
From Hypothesis to Loss Class

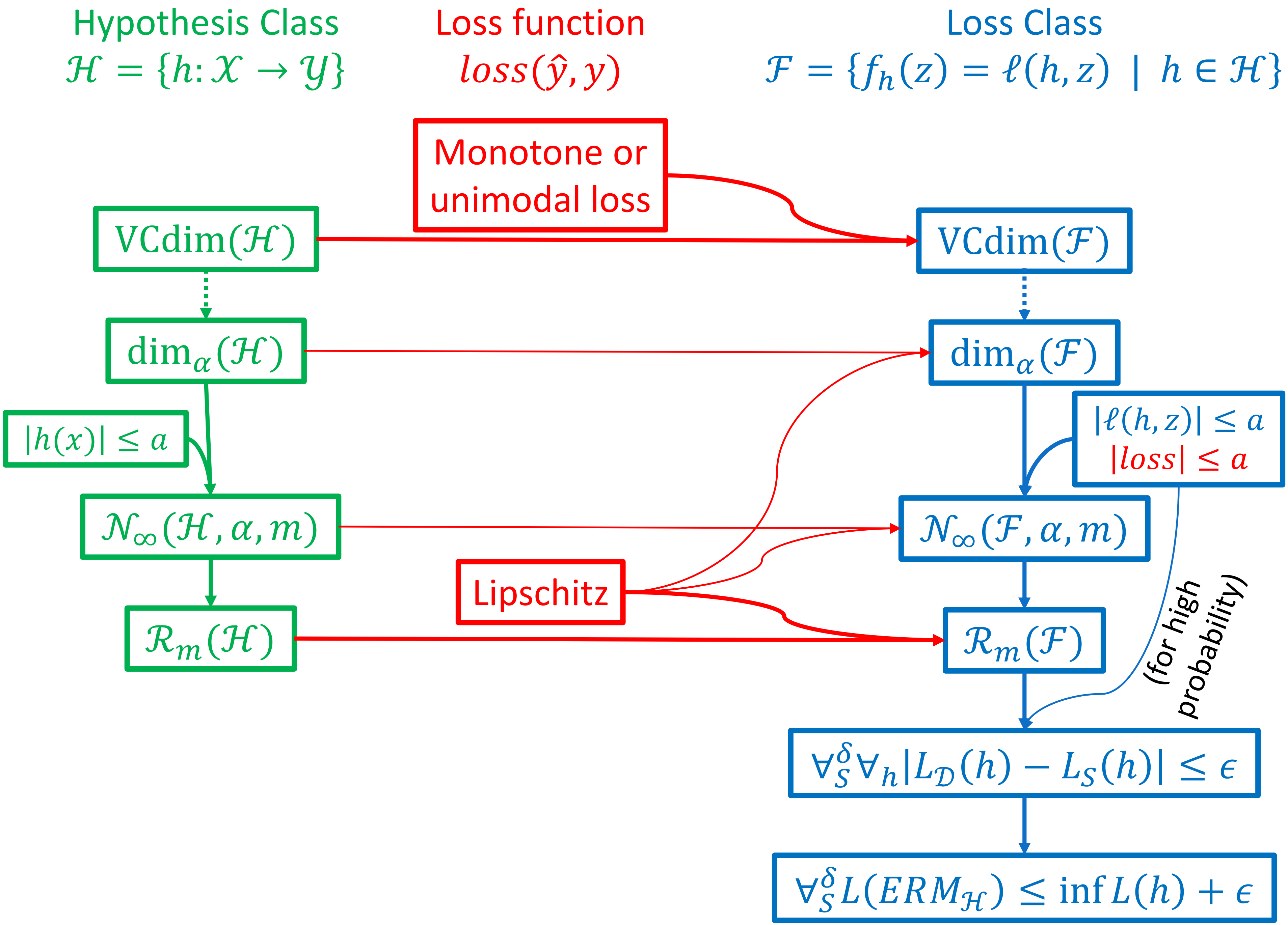
- **Definition**: $loss: \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ (i.e. with $\hat{\mathcal{Y}} = \mathbb{R}$) is ***G*-Lipschitz** (with respect to $\hat{y}$) if $\forall y, \hat{y}_1, \hat{y}_2$, $|loss(\hat{y}_1; y) - loss(\hat{y}_2; y)| \leq G \cdot |\hat{y}_1 - \hat{y}_2|$
(if differentiable, equivalent to $|loss'(\hat{y}; y)| \leq G$)

  - $loss(\hat{y}; y) = [[sign(\hat{y}) \neq y]]$ No!
  - $loss(\hat{y}; y) = [1 - \hat{y}y]_+$ G=1
  - $loss(\hat{y}; y) = \log(1 + e^{-\hat{y}y})$ G=1
  - $loss(\hat{y}; y) = |\hat{y} - y|$ G=1
  - $loss(\hat{y}; y) = (\hat{y} - y)^2$ Not over $\mathbb{R}$. $G = 4$a if $|y|, |\hat{y}| \leq a$

- **Lipschitz Contraction Lemma**: For $\mathcal{F} = \{(x, y) \mapsto loss(h(x); y) \mid h \in \mathcal{H}\}$, if the loss is G-Lipschitz, then

$$\mathcal{R}_S(\mathcal{F}) \leq G \cdot \mathcal{R}_S(\mathcal{H})$$


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
$\mathcal{R}_m(\mathcal{H})$
$\mathcal{R}_m(\mathcal{F})$
(for high probability)
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \le \epsilon$
$\forall_S^\delta L(ERM_\mathcal{H}) \le \inf L(h) + \epsilon$


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
$\mathcal{R}_m(\mathcal{H})$
$\mathcal{R}_m(\mathcal{F})$
(for high probability)
$\forall_S^\delta \forall_h |L_\mathcal{D}(h) - L_S(h)| \le \epsilon$
$\forall_S^\delta L(ERM_\mathcal{H}) \le \inf L(h) + \epsilon$

Parametric vs Scale Sensitive

- Parametric Complexity Control
  - Finite VC (subgraph) dimension
  - Only depend on structure of loss (monotone, unimodal), not on continuity
  - $\log \mathcal{N}_\infty(\mathcal{F}, \alpha, m)$ only depends logarithmically on α

    ➔ no need for Dudley (up to log factors)

- Scale-Sensitive Control
  - VC subgraph dimension might be infinite
  - Scale-sensitive hypothesis class: fat shattering dim decreases with α (typical scaling is $1/\alpha^2$)
  - Scale-sensitive loss: loss must be Lipschitz continuous
  - $\log \mathcal{N}_\infty(\mathcal{F}, \alpha, m)$ depends on α (typically as $1/\alpha^2$)

    ➔ Need Dudley in order to get correct dependence

Regularized Linear Prediction

- For a G-Lipschitz loss function:

$$\forall^{\delta}_{S\sim\mathcal{D}^m} \forall_{\|w\|\leq B} \; |L_S(w) - L_{\mathcal{D}}(w)| \leq 2G\sqrt{\frac{B^2 \mathbb{E}[\|x\|^2] \log 2/\delta}{m}}$$

➔ sample complexity $m = O\left(\frac{B^2\mathbb{E}[\|x\|^2]}{\epsilon^2}\right)$

- No dependence on the dimensionality!
- Valid even for linear prediction in very high, even infinite, dimensions—as long as data is bounded (or at least $\mathbb{E}[\|x\|^2]$ is bounded) and there is a good low-norm predictor, we can learn with sample complexity $\propto \|w^*\|^2$.

Margin-Based Learning

- Back to the geometrical margin:
  - Can we learn in high (infinite) dimensions is if we have a margin?
  - How does the sample complexity depend on the margin?

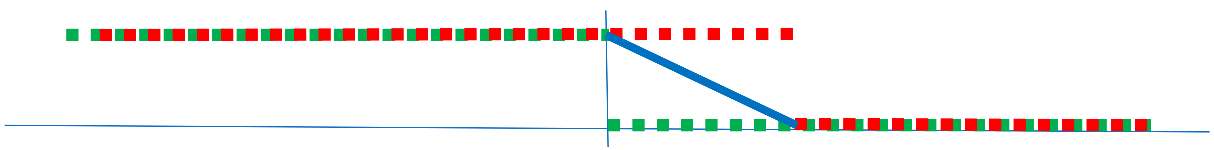
- Geometric margin: $y\langle w, x\rangle \geq \gamma$ for $\|w\| = 1$
  - How can we learn if $\exists_{\|w\|=1} \Pr[y\langle w, x\rangle \geq \gamma] = 1$ (or close to 1), with large γ?
- We'll re-normalize to: $y\langle w, x\rangle \geq 1$ with $\|w\| = 1/\gamma$
  - $loss^{mrg}(\hat{y}; y) = [[\hat{y}y < 1]]$
  - How can we learn if $L_{\mathcal{D}}^{m}(w)$ is small for some low-norm w?

- What can we say about:

$$ERM_B^{mrg}(S) = \arg\min_{\|w\|\leq B} L_S^{mrg}(w)$$

Margin and Ramp Loss

- We want to rely on: $loss^{mrg}(\hat{y};y) = [[\hat{y}y < 1]]$
- Use the 1-Lipschitz *ramp loss*: $loss^{ramp}(\hat{y};y) = \begin{cases} 0 & \hat{y}y \geq 1 \\ 1-\hat{y}y & 0 < \hat{y}y < 1 \\ 1 & \hat{y}y < 0 \end{cases}$



$$loss^{01} \leq loss^{ramp} \leq loss^{mrg}$$

- For any $\|w\| \leq B$, with probability $\geq 1-\delta$:

$$L_{\mathcal{D}}^{01}\left(ERM_B^{mrg}(S)\right) \leq L_{\mathcal{D}}^{ramp}\left(ERM_B^{mrg}(S)\right) \leq L_{\boldsymbol{S}}^{ramp}\left(ERM_B^{mrg}(S)\right) + 2\sqrt{\frac{B^2R^2+\log^4/_\delta}{m}}$$

$$\leq L_{\boldsymbol{S}}^{mrg}\left(ERM_B^{mrg}(S)\right) + 2\sqrt{\frac{B^2R^2+\log^4/_\delta}{m}} \leq L_{\boldsymbol{S}}^{mrg}(w) + 2\sqrt{\frac{B^2R^2+\log^4/_\delta}{m}}$$

$$\leq L_{\mathcal{D}}^{mrg}(w) + 2\sqrt{\frac{B^2R^2+\log^4/_\delta}{m}}$$

Single Hoefding bound (no need for union bound)

Margin-Based Learning Guarantee

- W.p. $\geq 1-\delta$, $L_{\mathcal{D}}^{01}\left(ERM_B^{mrg}(S)\right) \leq \inf_{\|w\|\leq B} L_{\mathcal{D}}^{mrg}(w) + 3\sqrt{\frac{B^2\mathbb{E}\|x\|^2 + \log\frac{4}{\delta}}{m}}$

- Is this a PAC-learning guarantee?

- In terms of margin: if the data is separable by margin γ except for L^* fraction of the points, we can find a predictor with 0/1 error $L^* + \epsilon$ using $O\left(\frac{R^2}{\gamma^2\epsilon^2}\right)$ samples.

- Also for hinge loss:

$$\forall^{\delta}_{S\sim\mathcal{D}^m} L_{\mathcal{D}}^{01}\left(ERM_B^{hinge}(S)\right) \leq \inf_{\|w\|\leq B} L_{\mathcal{D}}^{hinge}(w) + 3\sqrt{\frac{B^2R^2\log\frac{4}{\delta}}{m}}$$

Surrogate Losses

- Minimizing 0/1 error is problematic
  - Computationally intractable
  - Not scale-sensitive—can't learn in high dim even with norm regularization
- Instead, minimize upper bound on 0/1 error
- Minimizing margin loss or ramp loss
  - Upper bound on 0/1 error
  - Scale sensitive—can generalize even in infinite dim
  - But still not tractable
- Minimize hinge loss
  - Upper bound on 0/1 error
  - Scale sensitive (Lipschitz continuous) ➔ generalization
  - Convex ➔ tractability
  - But: to ensure success, need low $L_{\mathcal{D}}^{hinge}(w)$, not enough low $L_{\mathcal{D}}^{01}(w)$ or $L_{\mathcal{D}}^{mrg}(w)$

Other Regularized Classes

- "Geometric (Euclidean) Margin" corresponds to the Euclidean norm $\|w\|_2$
- Separating to a scale-sensitive loss (e.g. hinge loss, logistic loss, exp-loss, the intractable margin loss) and a scale-sensitive class, allows us to consider other "margins"
- E.g. ℓ_1 margin, corresponding to $y\langle w, x\rangle > 1$ with low $\|w\|_1$

  ➔ $\mathcal{H}_B = \{x \mapsto \langle w, x\rangle \mid \|w\|_1 \leq B\}$
- More generally, can define such a class hierarchy for any regularizer on w

- To ensure tractability, we will focus on linear prediction, with a convex regularizer $r(w)$, and a convex loss function:

$$\mathcal{H}_B = \{x \mapsto \langle w, \phi(x)\rangle \mid r(w) \leq B\}$$

  This ensures that the ERM/SRM problem is convex:

$$\min_{r(w)\leq B} L_S(w) \qquad \text{or} \qquad \min L_S(w) + \lambda r(w)$$

Convex Learning ➔ Linear Learning

- Consider supervised learning with a "non-degenerate" $loss(\hat{y}; y)$
- **Claim**: $\ell(h_w, (x, y))$ will be convex in a parametrization w only if $h_w(x)$ is affine in w. I.e.:
$$h_w(x) = \langle w, \phi(x) \rangle + \phi_0(x)$$
- Proof sketch: if the loss is non-degenerate, it must sometimes (for some value of y) be increasing in $\hat{y}$ and sometimes decreasing. If its increasing, for $loss(h_w(x); y)$ to be convex in w, we must have $h_w(x)$ convex in w. But if its decreasing, it must be concave in w.

- Conclusion: the only form of tractable *supervised* learning is linear learning with a convex loss and convex regularizer or constraint on w.

Generalized Linear Learning

- Different loss functions
  - Hinge, logistic, exp-loss, multi-class, structured, etc
- Different regularizers
  - ℓ_2, ℓ_1 (LASSO), group-regularizers, matrix-regularizers, etc
- Different feature spaces and different computationally efficient ways of representing them
  - Kernels
  - Boosting (implicit through weak learning oracle)
  - Indirectly

- Statistical Complexity of such classes?
- Computational efficiency?
- Relationships and interpertations