

# Empirical Risk Minimization

Vojtěch Franc

February 24, 2020

Prediction task and its solution by learning from data

Empirical risk minimization

Statistical consistency

Uniform Law of Large Numbers

**XEP33SML – Structured Model Learning, Summer 2020**

# Structured Output Prediction: the statistical model

## The setting

- ◆  $\mathcal{X}$  set of input observations
- ◆  $\mathcal{Y}$  finite set of hidden states, e.g.
  - Flat classification:  $\mathcal{Y} = \{1, \dots, K\}$
  - Structured classif.:  $\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_{|\mathcal{V}|}$  is a labeling of parts  $\mathcal{V}$ .
- ◆  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  randomly drawn from r.v. with p.d.f.  $p(x, y)$
- ◆  $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$  loss function

# Structured Output Prediction: the statistical model

## The setting

- ◆  $\mathcal{X}$  set of input observations
- ◆  $\mathcal{Y}$  finite set of hidden states, e.g.
  - Flat classification:  $\mathcal{Y} = \{1, \dots, K\}$
  - Structured classif.:  $\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_{|\mathcal{V}|}$  is a labeling of parts  $\mathcal{V}$ .
- ◆  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  randomly drawn from r.v. with p.d.f.  $p(x, y)$
- ◆  $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$  loss function

**The task:** find a strategy  $h: \mathcal{X} \rightarrow \mathcal{Y}$  with the minimal expected risk

$$R^* = \min_{h: \mathcal{X} \rightarrow \mathcal{Y}} R(h) \quad \text{where} \quad R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$$

## Solving the prediction problem from examples

- ◆ **Assumption:** we have an access to examples

$$\{(x^1, y^1), (x^2, y^2), \dots\}$$

drawn from i.i.d. r.v. distributed according to unknown  $p(x, y)$ .

## Solving the prediction problem from examples

- ◆ **Assumption:** we have an access to examples

$$\{(x^1, y^1), (x^2, y^2), \dots\}$$

drawn from i.i.d. r.v. distributed according to unknown  $p(x, y)$ .

- ◆ a) **Testing:** Estimate  $R(h)$  of a given  $h: \mathcal{X} \rightarrow \mathcal{Y}$  using **test set**

$$\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l\}$$

drawn i.i.d. from  $p(x, y)$ .

## Solving the prediction problem from examples

- ◆ **Assumption:** we have an access to examples

$$\{(x^1, y^1), (x^2, y^2), \dots\}$$

drawn from i.i.d. r.v. distributed according to unknown  $p(x, y)$ .

- ◆ a) **Testing:** Estimate  $R(h)$  of a given  $h: \mathcal{X} \rightarrow \mathcal{Y}$  using **test set**

$$\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l\}$$

drawn i.i.d. from  $p(x, y)$ .

- ◆ b) **Learning:** find  $h: \mathcal{X} \rightarrow \mathcal{Y}$  with small  $R(h)$  using **training set**

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m\}$$

drawn i.i.d. from  $p(x, y)$ .

## Estimation of the expected risk from examples

- ◆ Given a predictor  $h: \mathcal{X} \rightarrow \mathcal{Y}$ , compute the **empirical risk**

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$

and use it as a proxy for  $R(h) = \mathbb{E}_{(x,y) \sim p}(\ell(y, h(x)))$ .

- ◆ The value of the empirical risk  $R_{\mathcal{S}^l}(h)$  is a **random number**.
- ◆ Application of **Hoeffding inequality**: for any  $\varepsilon > 0$  the probability of seeing a “bad test set” can be bound by

$$\mathbb{P}_{\mathcal{S}^l \sim p} \left( \left| R_{\mathcal{S}^l}(h) - R(h) \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\min} - \ell_{\max})^2}}$$

## Learning algorithm

- ◆ **Learning:** find a strategy  $h: \mathcal{X} \rightarrow \mathcal{Y}$  with a small  $R(h)$  using the training set of examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m\}$$

drawn from i.i.d. according to unknown  $p(x, y)$ .

- ◆ Use prior knowledge to select hypothesis space

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$$

- ◆ The learning algorithm

$$A: \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$$

selects strategy  $h_m = A(\mathcal{T}^m)$  based on the training set  $\mathcal{T}^m$ .



## Generative learning (to come later)

1. Use the training set  $\mathcal{T}^m = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i \in \{1, \dots, m\}\}$  to approximate  $p(x, y)$  by  $\hat{p}(x, y)$ .

For example, use the Maximum-Likelihood method:

- (a) Guess the shape of the distribution, e.g.

$$\hat{p}_{\mathbf{w}}(x, y) = \frac{1}{Z(\mathbf{w})} \exp\langle \mathbf{w}, \phi(x, y) \rangle, \quad \mathbf{w} \in \mathcal{W}$$

- (b) Find the ML estimate

$$\mathbf{w}_m \in \operatorname{argmax}_{\mathbf{w} \in \mathcal{W}} \sum_{i=1}^m \log \hat{p}_{\mathbf{w}}(x^i, y^i)$$

2. Construct a plug-in classifier

$$h_m(x) \in \operatorname{argmin}_{h: \mathcal{X} \rightarrow \mathcal{Y}} \mathbb{E}_{(x, y) \sim \hat{p}_{\mathbf{w}_m}} [\ell(y, h(x))]$$

## Discriminative learning by Empirical Risk Minimization

- ◆ Use the training set  $\mathcal{T}^m = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i \in \{1, \dots, m\}\}$  to approximate the expected risk  $R(h)$  by the empirical risk

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^m \ell(y^i, h(x^i))$$

- ◆ The ERM learning algorithm returns  $h_m$  such that

$$h_m \in \underset{h \in \mathcal{H}}{\text{Argmin}} R_{\mathcal{T}^m}(h) \quad (1)$$

- ◆ Depending on the choice of  $\mathcal{H}$ ,  $\ell$  and algorithm solving (1) we get individual instances, e.g.: Structured-Output Perceptron, Structured-Output Support Vector Machines, Logistic regression, Neural Networks learned by back-propagation, AdaBoost, . . . .

## Errors characterizing a learning algorithm

### The characters of the play:

- ◆  $R^* = \min_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$  best attainable (Bayes) risk
- ◆  $R(h_{\mathcal{H}})$  best risk in  $\mathcal{H}$ ;  $h_{\mathcal{H}} \in \text{Argmin}_{h \in \mathcal{H}} R(h)$
- ◆  $R(h_m)$  risk of  $h_m = A(\mathcal{T}_m)$  learned from  $\mathcal{T}^m$

## Errors characterizing a learning algorithm

### The characters of the play:

- ◆  $R^* = \min_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$  best attainable (Bayes) risk
- ◆  $R(h_{\mathcal{H}})$  best risk in  $\mathcal{H}$ ;  $h_{\mathcal{H}} \in \text{Argmin}_{h \in \mathcal{H}} R(h)$
- ◆  $R(h_m)$  risk of  $h_m = A(\mathcal{T}_m)$  learned from  $\mathcal{T}^m$

**Excess error:** the quantity we want to minimize

$$\underbrace{\left( R(h_m) - R^* \right)}_{\text{excess error}} = \underbrace{\left( R(h_m) - R(h_{\mathcal{H}}) \right)}_{\text{estimation error}} + \underbrace{\left( R(h_{\mathcal{H}}) - R^* \right)}_{\text{approximation error}}$$

## Errors characterizing a learning algorithm

### The characters of the play:

- ◆  $R^* = \min_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$  best attainable (Bayes) risk
- ◆  $R(h_{\mathcal{H}})$  best risk in  $\mathcal{H}$ ;  $h_{\mathcal{H}} \in \text{Argmin}_{h \in \mathcal{H}} R(h)$
- ◆  $R(h_m)$  risk of  $h_m = A(\mathcal{T}_m)$  learned from  $\mathcal{T}^m$

**Excess error:** the quantity we want to minimize

$$\underbrace{\left( R(h_m) - R^* \right)}_{\text{excess error}} = \underbrace{\left( R(h_m) - R(h_{\mathcal{H}}) \right)}_{\text{estimation error}} + \underbrace{\left( R(h_{\mathcal{H}}) - R^* \right)}_{\text{approximation error}}$$

- ◆ The excess and the estimation error are random variables
- ◆ The estimation error depends on  $m$  and  $\mathcal{H}$
- ◆ The approximation error depends on  $\mathcal{H}$  (so called inductive bias)

## Statistically consistent learning algorithm

**Definition 1.** The algorithm  $A: \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$  is statistically consistent in  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  w.r.t.  $p(x, y)$  if for every  $\varepsilon > 0$  and  $\delta \in (0, 1)$  there exist  $m_0 \in \mathcal{N}$  such that

$$\mathbb{P}_{\mathcal{T}^m \sim p} \left( R(A(\mathcal{T}^m)) - R(h_{\mathcal{H}}) \geq \varepsilon \right) \leq 1 - \delta$$

holds for every  $m \geq m_0$ .

If  $A$  is consistent for any  $p(x, y)$  then  $A$  is *universally consistent in  $\mathcal{H}$* .

## Statistically consistent learning algorithm

**Definition 1.** The algorithm  $A: \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$  is statistically consistent in  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  w.r.t.  $p(x, y)$  if for every  $\varepsilon > 0$  and  $\delta \in (0, 1)$  there exist  $m_0 \in \mathcal{N}$  such that

$$\mathbb{P}_{\mathcal{T}^m \sim p} \left( R(A(\mathcal{T}^m)) - R(h_{\mathcal{H}}) \geq \varepsilon \right) \leq 1 - \delta$$

holds for every  $m \geq m_0$ .

If  $A$  is consistent for any  $p(x, y)$  then  $A$  is *universally consistent in  $\mathcal{H}$* .

Question:

- ◆ Is the ERM learning algorithm statistically consistent ?

## Example: ERM is not consistent $\mathcal{H}$ is unconstrained

- ◆ Let  $\mathcal{X} = [a, b] \subset \mathbb{R}$ ,  $\mathcal{Y} = \{+1, -1\}$ ,  $\ell(y, y') = [y \neq y']$ ,  $p(x | y = +1)$  and  $p(x | y = -1)$  be uniform distributions on  $\mathcal{X}$  and  $p(y = +1) = 0.8$ .



## Example: ERM is not consistent $\mathcal{H}$ is unconstrained

- ◆ Let  $\mathcal{X} = [a, b] \subset \mathbb{R}$ ,  $\mathcal{Y} = \{+1, -1\}$ ,  $\ell(y, y') = [y \neq y']$ ,  $p(x | y = +1)$  and  $p(x | y = -1)$  be uniform distributions on  $\mathcal{X}$  and  $p(y = +1) = 0.8$ .
- ◆ The optimal strategy is  $h(x) = +1$  with the Bayes risk  $R^* = 0.2$ .

## Example: ERM is not consistent $\mathcal{H}$ is unconstrained

- ◆ Let  $\mathcal{X} = [a, b] \subset \mathbb{R}$ ,  $\mathcal{Y} = \{+1, -1\}$ ,  $\ell(y, y') = [y \neq y']$ ,  $p(x | y = +1)$  and  $p(x | y = -1)$  be uniform distributions on  $\mathcal{X}$  and  $p(y = +1) = 0.8$ .
- ◆ The optimal strategy is  $h(x) = +1$  with the Bayes risk  $R^* = 0.2$ .
- ◆ Consider learning algorithm which for a given training set  $\mathcal{T}^m = \{(x^1, y^1), \dots, (x^m, y^m)\}$  returns strategy

$$h_m(x) = \begin{cases} y^j & \text{if } x = x^j \text{ for some } j \in \{1, \dots, m\} \\ -1 & \text{otherwise} \end{cases}$$

## Example: ERM is not consistent $\mathcal{H}$ is unconstrained

- ◆ Let  $\mathcal{X} = [a, b] \subset \mathbb{R}$ ,  $\mathcal{Y} = \{+1, -1\}$ ,  $\ell(y, y') = [y \neq y']$ ,  $p(x | y = +1)$  and  $p(x | y = -1)$  be uniform distributions on  $\mathcal{X}$  and  $p(y = +1) = 0.8$ .
- ◆ The optimal strategy is  $h(x) = +1$  with the Bayes risk  $R^* = 0.2$ .
- ◆ Consider learning algorithm which for a given training set  $\mathcal{T}^m = \{(x^1, y^1), \dots, (x^m, y^m)\}$  returns strategy
 
$$h_m(x) = \begin{cases} y^j & \text{if } x = x^j \text{ for some } j \in \{1, \dots, m\} \\ -1 & \text{otherwise} \end{cases}$$
- ◆ The empirical risk is  $R_{\mathcal{T}^m}(h_m) = 0$  with probability 1 for any  $m$ .
- ◆ The expected risk is  $R(h_m) = 0.8$  for any  $m$ .

## Uniform Law of Large Numbers

**Definition 2.** *The hypothesis space  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  satisfies the uniform law of large numbers if for every distribution  $p(x, y)$ ,  $\varepsilon > 0$  and  $\delta \in (0, 1)$  there exists  $m_0 \in \mathcal{N}$  such that*

$$\mathbb{P}_{\mathcal{T}^m \sim p} \left( \sup_{h \in \mathcal{H}} |R(h) - R_{\mathcal{T}^m}(h)| \geq \varepsilon \right) \leq 1 - \delta$$

*holds for every  $m \geq m_0$ .*

## Uniform Law of Large Numbers

**Definition 2.** *The hypothesis space  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  satisfies the uniform law of large numbers if for every distribution  $p(x, y)$ ,  $\varepsilon > 0$  and  $\delta \in (0, 1)$  there exists  $m_0 \in \mathcal{N}$  such that*

$$\mathbb{P}_{\mathcal{T}^m \sim p} \left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 1 - \delta$$

*holds for every  $m \geq m_0$ .*

**Theorem 1.** *If  $\mathcal{H}$  satisfies ULLN then ERM is statistically consistent in  $\mathcal{H}$ .*

## Proof: ULLN implies consistency of ERM

For fixed  $\mathcal{T}^m$  and  $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$  we have:

$$\begin{aligned}
 R(h_m) - R(h_{\mathcal{H}}) &= \left( R(h_m) - R_{\mathcal{T}^m}(h_m) \right) + \left( R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}}) \right) \\
 &\leq \left( R(h_m) - R_{\mathcal{T}^m}(h_m) \right) + \left( R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}}) \right) \\
 &\leq 2 \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right|
 \end{aligned}$$

Therefore  $\varepsilon \leq R(h_m) - R(h_{\mathcal{H}})$  implies  $\frac{\varepsilon}{2} \leq \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right|$  and

$$\mathbb{P} \left( R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \right) \leq \mathbb{P} \left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \frac{\varepsilon}{2} \right)$$

## Two examples of $\mathcal{H}$ which satisfy ULLN

1.  $\mathcal{H}$  is a finite set and  $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [\ell_{min}, \ell_{max}]$ . Then,

$$\mathbb{P}_{\mathcal{T} \sim p^m} \left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 2|\mathcal{H}| \exp \left( \frac{-2m \varepsilon^2}{(\ell_{max} - \ell_{min})^2} \right)$$

holds for any  $\varepsilon > 0$  and  $m \in \mathcal{N}$ .

## Two examples of $\mathcal{H}$ which satisfy ULLN

1.  $\mathcal{H}$  is a finite set and  $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [\ell_{min}, \ell_{max}]$ . Then,

$$\mathbb{P}_{\mathcal{T} \sim p^m} \left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 2|\mathcal{H}| \exp \left( \frac{-2m \varepsilon^2}{(\ell_{max} - \ell_{min})^2} \right)$$

holds for any  $\varepsilon > 0$  and  $m \in \mathcal{N}$ .

2.  $\ell(y, y') = [y \neq y']$ ,  $\mathcal{Y} = \{+1, -1\}$  and VC-dimension of  $\mathcal{H}$  is finite. VC-dimension  $d$  of  $\mathcal{H}$  is the maximal number of inputs which can be classified by strategies from  $\mathcal{H}$  in all possible (that is  $2^d$ ) ways. Then,

$$\mathbb{P}_{\mathcal{T} \sim p^m} \left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 4 \left( \frac{2em}{d} \right)^d e^{-\frac{m\varepsilon^2}{8}}$$



## Rademacher Complexity



- ◆ Let  $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $p(z) = p(x, y)$  and  $g(z) = \ell(y, h(x))$ .

## Rademacher Complexity

- ◆ Let  $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $p(z) = p(x, y)$  and  $g(z) = \ell(y, h(x))$ .

**Definition 3.** Let  $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$  be a set of functions  $g: \mathcal{Z} \rightarrow [a, b]$  where  $a, b \in \mathbb{R}$  and  $a < b$ . Let  $\mathcal{U}^m = \{z^1, \dots, z^m\} \in \mathcal{Z}^m$  be drawn i.i.d. from  $p(z)$ .

## Rademacher Complexity

- ◆ Let  $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $p(z) = p(x, y)$  and  $g(z) = \ell(y, h(x))$ .

**Definition 3.** Let  $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$  be a set of functions  $g: \mathcal{Z} \rightarrow [a, b]$  where  $a, b \in \mathbb{R}$  and  $a < b$ . Let  $\mathcal{U}^m = \{z^1, \dots, z^m\} \in \mathcal{Z}^m$  be drawn i.i.d. from  $p(z)$ .

The *empirical Rademacher complexity* of  $\mathcal{G}$  w.r.t. to the sample  $\mathcal{U}^m$  is

$$\hat{\mathfrak{R}}_m(\mathcal{G}, \mathcal{U}^m) = \mathbb{E}_{\sigma \sim \text{Unif}\{-1, +1\}} \left[ \sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right]$$

## Rademacher Complexity

- ◆ Let  $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $p(z) = p(x, y)$  and  $g(z) = \ell(y, h(x))$ .

**Definition 3.** Let  $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$  be a set of functions  $g: \mathcal{Z} \rightarrow [a, b]$  where  $a, b \in \mathbb{R}$  and  $a < b$ . Let  $\mathcal{U}^m = \{z^1, \dots, z^m\} \in \mathcal{Z}^m$  be drawn i.i.d. from  $p(z)$ .

The *empirical Rademacher complexity* of  $\mathcal{G}$  w.r.t. to the sample  $\mathcal{U}^m$  is

$$\hat{\mathfrak{R}}_m(\mathcal{G}, \mathcal{U}^m) = \mathbb{E}_{\sigma \sim \text{Unif}\{-1, +1\}} \left[ \sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right]$$

The *Rademacher complexity* of  $\mathcal{G}$  w.r.t. distribution  $p(z)$  is

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{\mathcal{U}^m \sim p^m(z)} \left[ \hat{\mathfrak{R}}_m(\mathcal{G}, \mathcal{U}^m) \right]$$

## Rademacher-based uniform convergence

- ◆ Let  $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$  be a set of functions. Then, for every  $\delta \in (0, 1)$

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}_{z \sim p}(g(z)) - \frac{1}{m} \sum_{i=1}^m g(z_i) \right| \leq 2 \mathfrak{R}_m(\mathcal{G}) + (b - a) \sqrt{\frac{\log 2/\delta}{2m}}$$

holds with probability  $1 - \delta$  at least, w.r.t.  $\mathcal{U}^m = \{z^1, \dots, z^m\} \sim p^m(z)$ .

## Rademacher-based uniform convergence

- ◆ Let  $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$  be a set of functions. Then, for every  $\delta \in (0, 1)$

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}_{z \sim p}(g(z)) - \frac{1}{m} \sum_{i=1}^m g(z_i) \right| \leq 2 \mathfrak{R}_m(\mathcal{G}) + (b - a) \sqrt{\frac{\log 2/\delta}{2m}}$$

holds with probability  $1 - \delta$  at least, w.r.t.  $\mathcal{U}^m = \{z^1, \dots, z^m\} \sim p^m(z)$ .

- ◆ For every  $\delta \in (0, 1)$

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}_{z \sim p}(g(z)) - \frac{1}{m} \sum_{i=1}^m g(z_i) \right| \leq 3 \hat{\mathfrak{R}}_m(\mathcal{G}, \mathcal{U}^m) + (b - a) \sqrt{\frac{\log 4/\delta}{2m}}$$

holds with probability  $1 - \delta$  at least, w.r.t.  $\mathcal{U}^m = \{z^1, \dots, z^m\} \sim p^m(z)$ .

## Example: Rademacher complexity of linear functions

- ◆ Assume that  $\mathcal{X} \subseteq \mathbb{R}^n$  and  $p(\mathbf{x}, y)$  is such that  $\|\mathbf{x}\| \leq R$ .
- ◆ Assume that

$$\mathcal{G} = \left\{ \psi(\langle \mathbf{w}, \mathbf{x} \rangle, y) \mid \|\mathbf{w}\|_2 \leq B \right\}$$

where  $\psi: \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$  is such that  $f(t) = \psi(t, y)$  is  $\rho$ -Lipschitz continuous for all  $y \in \mathcal{Y}$ .

## Example: Rademacher complexity of linear functions

- ◆ Assume that  $\mathcal{X} \subseteq \mathbb{R}^n$  and  $p(\mathbf{x}, y)$  is such that  $\|\mathbf{x}\| \leq R$ .
- ◆ Assume that

$$\mathcal{G} = \left\{ \psi(\langle \mathbf{w}, \mathbf{x} \rangle, y) \mid \|\mathbf{w}\|_2 \leq B \right\}$$

where  $\psi: \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$  is such that  $f(t) = \psi(t, y)$  is  $\rho$ -Lipschitz continuous for all  $y \in \mathcal{Y}$ .

E.g.  $\psi(t, y) = \max\{0, 1 - t y\}$  and  $\psi(t) = |t - y|$  are 1-Lipschitz.



## Example: Rademacher complexity of linear functions

◆ Assume that  $\mathcal{X} \subseteq \mathbb{R}^n$  and  $p(\mathbf{x}, y)$  is such that  $\|\mathbf{x}\| \leq R$ .

◆ Assume that

$$\mathcal{G} = \left\{ \psi(\langle \mathbf{w}, \mathbf{x} \rangle, y) \mid \|\mathbf{w}\|_2 \leq B \right\}$$

where  $\psi: \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$  is such that  $f(t) = \psi(t, y)$  is  $\rho$ -Lipschitz continuous for all  $y \in \mathcal{Y}$ .

E.g.  $\psi(t, y) = \max\{0, 1 - t y\}$  and  $\psi(t) = |t - y|$  are 1-Lipschitz.

◆ Then,

$$\hat{\mathfrak{R}}_m(\mathcal{G}) \leq \frac{\rho B R}{\sqrt{m}}$$

◆ We can also compute

$$b = \max_{t \in [-BR, BR]} \psi(t, y) \quad \text{and} \quad a = \min_{t \in [-BR, BR]} \psi(t, y)$$