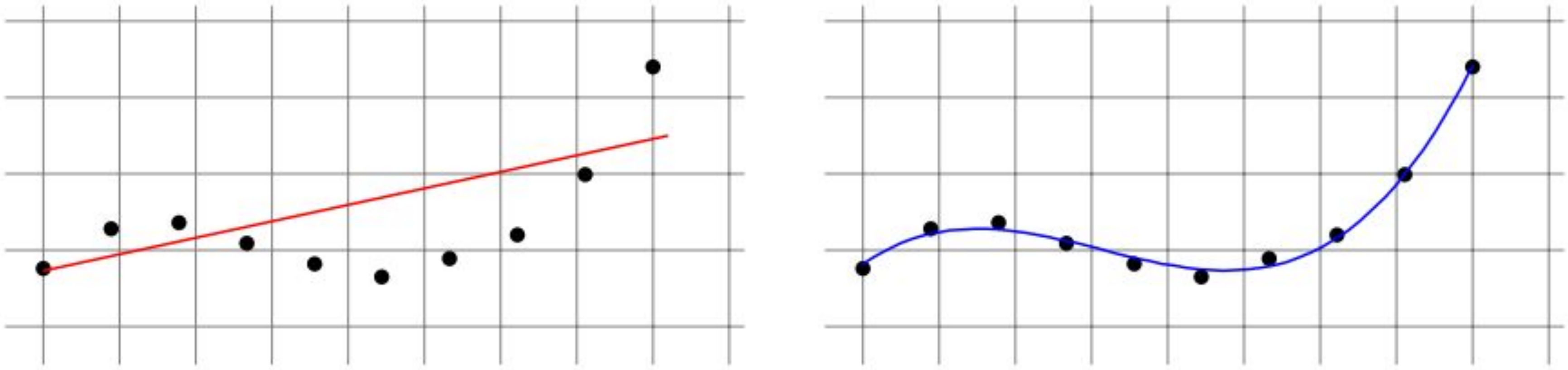
Regression Problem



Regression Problem

- Training data: sample drawn i.i.d. from set X according to some distribution D,

$$S=((x_1,y_1),\ldots,(x_m,y_m))\in X\times Y,$$

with $Y\subseteq\mathbb{R}$ is a measurable subset.

- Loss function: $L\colon Y\times Y\to\mathbb{R}_+$ a measure of closeness, typically $L(y,y')=(y'-y)^2$ or $L(y,y')=|y'-y|^p$ for some $p\geq 1$.

- Problem: find hypothesis $h\colon X\to\mathbb{R}$ in H with small generalization error with respect to target f

$$R_D(h)=\underset{x\sim D}{\mathrm{E}}\left[L\big(h(x),f(x)\big)\right].$$

Notes

- Empirical error:

$$\widehat{R}_D(h) = \frac{1}{m} \sum_{i=1}^{m} L\big(h(x_i), y_i\big).$$

- In much of what follows:
  - $Y = \mathbb{R}$ or $Y = [-M, M]$ for some $M > 0$.
  - $L(y, y') = (y' - y)^2 \longrightarrow$ mean squared error.

Generalization Bound - Finite H

- **Theorem**: let H be a finite hypothesis set, and assume that L is bounded by M. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall h \in H, R(h) \leq \widehat{R}(h) + M \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}.$$

- **Proof**: By the union bound,

$$\Pr\left[\sup_{h \in H} \left|R(h) - \widehat{R}(h)\right| > \epsilon\right] \leq \sum_{h \in H} \Pr\left[\left|R(h) - \widehat{R}(h)\right| > \epsilon\right].$$

By Hoeffding's bound, for a fixed h,

$$\Pr\left[\left|R(h) - \widehat{R}(h)\right| > \epsilon\right] \leq 2e^{-\frac{2m\epsilon^2}{M^2}}.$$

Rademacher Complexity of Lp Loss

- **Theorem**: Let $p \geq 1$, $H_p = \{x \mapsto |h(x) - f(x)|^p \colon h \in H\}$. Assume that $\sup_{x \in X, h \in H} |h(x) - f(x)| \leq M$. Then, for any sample S of size m,

$$\widehat{\mathfrak{R}}_S(H_p) \leq pM^{p-1} \widehat{\mathfrak{R}}_S(H).$$

Rad. Complexity Regression Bound

- **Theorem**: Let $p \geq 1$ and assume that $\|h - f\|_\infty \leq M$ for all $h \in H$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $h \in H$,

$$\mathrm{E}\left[\left|h(x) - f(x)\right|^p\right] \leq \frac{1}{m}\sum_{i=1}^{m}\left|h(x_i) - f(x_i)\right|^p + 2pM^{p-1}\mathfrak{R}_m(H) + M^p\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

$$\mathrm{E}\left[\left|h(x) - f(x)\right|^p\right] \leq \frac{1}{m}\sum_{i=1}^{m}\left|h(x_i) - f(x_i)\right|^p + 2pM^{p-1}\widehat{\mathfrak{R}}_S(H) + 3M^p\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

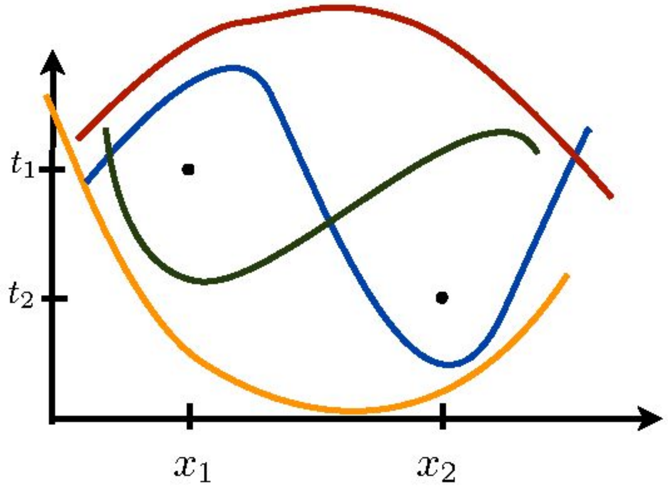
Notes

- As discussed for binary classification:
  - estimating the Rademacher complexity can be computationally hard for some Hs.
  - can we come up instead with a combinatorial measure that is easier to compute?

Shattering

- Definition: Let G be a family of functions mapping from X to $\mathbb{R}$. $A = \{x_1, \ldots, x_m\}$ is shattered by G if there exist $t_1, \ldots, t_m \in \mathbb{R}$ such that

$$\left| \left\{ \begin{bmatrix} \operatorname{sgn}\left(g(x_1) - t_1\right) \\ \vdots \\ \operatorname{sgn}\left(g(x_m) - t_m\right) \end{bmatrix} : g \in G \right\} \right| = 2^m.$$

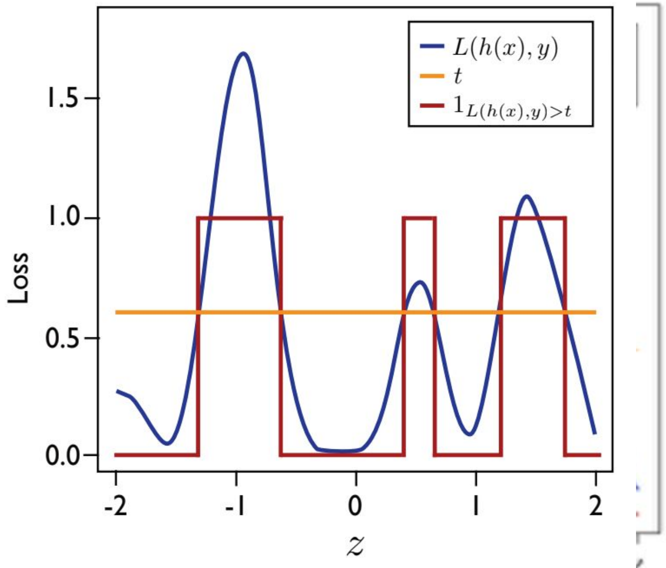


Pseudo-Dimension

(Pollard, 1984)

- **Definition**: Let G be a family of functions mapping from X to $\mathbb{R}$. The pseudo-dimension of G, $\mathrm{Pdim}(G)$, is the size of the largest set shattered by G.
- **Definition** (equivalent, see also (Vapnik, 1995)):

$$\mathrm{Pdim}(G) = \mathrm{VCdim}\Big(\big\{(x,t) \mapsto 1_{(g(x)-t)>0} \colon g \in G\big\}\Big).$$



Pseudo-Dimension - Properties

- **Theorem**: Pseudo-dimension of hyperplanes.

$$\text{Pdim}(\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b \colon \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}) = N + 1.$$

- **Theorem**: Pseudo-dimension of a vector space of real-valued functions H:

$$\text{Pdim}(H) = \dim(H).$$

Generalization Bound - Pdim

- Theorem: Let H be a family of real-valued functions. Assume that $\text{Pdim}(\{L(h, f) \colon h \in H\}) = d < \infty$ and that the loss L is bounded by M. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + M\sqrt{\frac{2d \log \frac{em}{d}}{m}} + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$