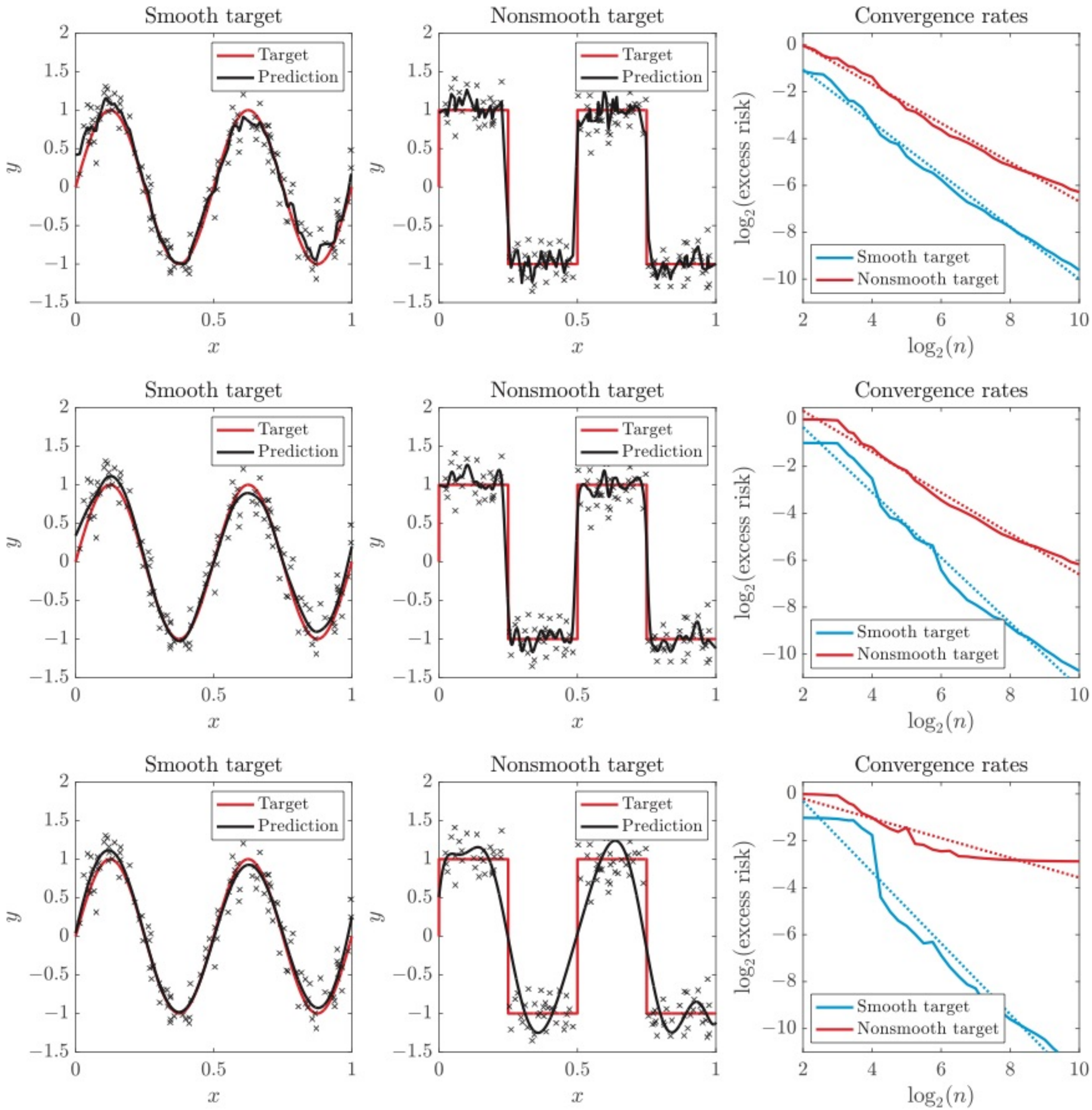


Smooth target
Nonsmooth target
Convergence rates
Target
Prediction
Smooth target
Nonsmooth target
x
y
$\log_2(n)$
$\log_2$(excess risk)

Summary: data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ — iid — $\subset \mathbb{R}$

loss function $\ell(y, z)$ + Expected risk

$$R(f) = \mathbb{E}[\ell(y, f(x))]$$

Regression: $\mathcal{Y} = \mathbb{R}$ $\ell(y,z) = (y-z)^2$

Classif. $\mathcal{Y} = \{-1, 1\}$ — $1_{y \neq z}$ $\longrightarrow$ logistic loss G-Lipschitz

$\ell : \{-1,1\} \times \mathbb{R}$

$\ell(y,z) = \log(1 + e^{-yz})$

Excess risk $f : \mathcal{X} \to \mathbb{R}$ if $f \in \mathcal{F}$

$$\underbrace{R(f) - R_*}_{\text{Bayes Risk}} = \underbrace{R(f) - \inf_{g \in \mathcal{F}} R(g)}_{\text{estimation error}} + \underbrace{\inf_{g \in \mathcal{F}} R(g) - R_*}_{\text{approx. error}}$$

Linear models: $f_\theta(x) = \theta^T \varphi(x)$ where $\varphi : \mathcal{X} \to \mathbb{R}^d$ $\|\varphi(x)\| \leq R$

$\mathcal{F} = \{f_\theta, \|\theta\| \leq D\}$

for Exact ERM $\mathbb{E}(\text{estimation error}) \leq 4 \frac{GRD}{\sqrt{n}}$

single pass SGD

Single pass SGD: $\theta_0 = 0$

$$\theta_k = \theta_{k-1} - \gamma g_k$$

$$\mathbb{E}[g_k | \theta_{k-1}] = \nabla_\theta R(f_\theta) \Big|_{\theta = \theta_{k-1}}$$

expected risk

with $g_k = \nabla_\theta \ell(y_k, f_\theta(x_k)) \Big|_{\theta = \theta_{k-1}}$

$k \in \{1, \dots, n\}$

single pass

(1) with projection on ball of radius D

$$\mathbb{E} R(f_{\bar\theta_n}) \le R(f_\theta) + \frac{GRD}{\sqrt{n}}$$

averaged iterate

if $\gamma \propto \frac{1}{\sqrt{n}}$

$\forall \|\theta\| \le$

(2) without projection

$$\mathbb{E} R(f_{\bar\theta_n}) \le R(f_\theta) + \frac{1}{2\gamma n}\|\theta\|^2 + \frac{\gamma G^2 R^2}{2} \qquad \forall \theta \in \mathbb{R}^d$$

TODAY:

$d \to +\infty$

- deal with approx error
- adaptivity to smoothness

1D

any D

inf-dim = $f_\theta(x) = \langle \theta, \varphi(x) \rangle$ with $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space

$\theta \in \mathcal{H}$

Representer theorem (1971) : $\min_{\theta \in \mathcal{H}} \hat{R}(f_\theta) + \frac{\lambda}{2} \|\theta\|^2$ $\left(\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \right)$

can be found as $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$ for $\alpha \in \mathbb{R}^n$

SGD : the k-th iterate of SGD is of the form $\sum_{i=1}^{k} \alpha_i \varphi(x_i)$
starting from 0.

proof : $\theta_k = \theta_{k-1} - \gamma \nabla_\theta \ell(y_k, \langle \theta, \varphi(x_k) \rangle) \Big|_{\theta = \theta_{k-1}}$

$= \theta_{k-1} - \gamma \underbrace{\ell'(y_k, \langle \theta_{k-1}, \varphi(x_k) \rangle)}_{\in \mathbb{R} \; = -\frac{\alpha_k}{\gamma}} \cdot \varphi(x_k)$

kernel trick

$\theta_k = \theta_{k-1} + \alpha_k \varphi(x_k)$

To run SGD need $\langle \theta_{k-1}, \varphi(x_k) \rangle = \sum_{i=1}^{k-1} \alpha_i \underbrace{\langle \varphi(x_i), \varphi(x_k) \rangle}_{\text{kernel fct } K(x_i, x_k)}$

"$O(n^2)$"

3 types of estimators:

$$\begin{cases} \hat{\theta} \leftarrow \arg\min \hat{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|^2 \\ \bar{\theta} \leftarrow \arg\min \hat{R}(f_\theta) \text{ s.t. } \|\theta\| \leq D \\ SGD: \theta_k = \theta_{k-1} - \gamma \,\square \end{cases}$$

$R(f_{\hat{\theta}})$

$$\boxed{\mathbb{E} R(f_{\bar{\theta}_n}) \leq R(f_\theta) + \frac{1}{2\gamma n}\|\theta\|^2 + \frac{\gamma G^2 R^2}{2} \quad \forall \theta \in \cancel{\mathbb{R}^d}\ \mathcal{H}}$$

Bounding approx error – $f_*: \mathcal{X} \to \mathbb{R}$ $\nexists$ $\langle \theta_*, \varphi(x) \rangle$ for some θ_*

$$\forall \theta,\ R(f_\theta) - R(f_*) = \mathbb{E}\left[\ell(y, f_\theta(x)) - \ell(y, f_*(x))\right]$$

$$\leq G\, \mathbb{E}\,|f_\theta(x) - f_*(x)| \leq G\, \|f_\theta - f_*\|_{L_2(p)}$$

↳ Jensen's inequality

$$\leq \frac{K}{2}\|f_\theta - f_*\|^2_{L^2(p)} + \frac{G^2}{2K} \qquad K = 1/\gamma R^2$$

$$\leq \frac{1}{2\gamma R^2}\|f_\theta - f_*\|^2_{L^2(p)} + \frac{\gamma G^2 R^2}{2}$$

trick: $|\varepsilon| \leq \frac{\varepsilon^2}{2\eta} + \frac{\eta}{2}$ with equality if $\eta = |\varepsilon|$.

$$\underbrace{\mathbb{E}\, R(f_{\hat\theta_n}) - R(f_*)}_{\text{excess risk}} \le \frac{1}{2\gamma R^2}\|f_\theta - f_*\|_{L^2}^2 + \frac{1}{2\gamma n}\|\theta\|^2 + \gamma G^2 R^2 \quad \forall \theta$$

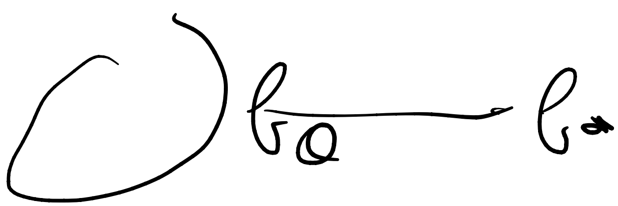
$$\le \frac{1}{2\gamma R^2} \underbrace{\inf_\theta \left\{ \|f_\theta - f_*\|^2 + \frac{R^2}{n}\|\theta\|^2 \right\}}_{A\left(\frac{R^2}{n}\right)} + \gamma G^2 R^2$$

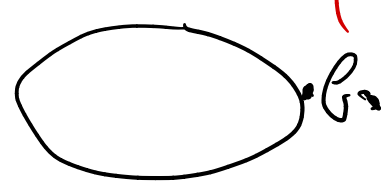
① simple case $f_*(x) = \langle \theta_*, \varphi(x) \rangle \Rightarrow A\left(\frac{R^2}{n}\right) \le \frac{R^2}{n}\|\theta_*\|^2$

excess risk $\le \frac{1}{2\gamma R^2}\frac{R^2}{n}\|\theta_*\|^2 + \gamma G^2 R^2 = \frac{1}{2\gamma n}\|\theta_*\|^2 + \gamma G^2 R^2$

$\le \frac{\sqrt{2}\|\theta_*\| G R}{\sqrt{n}}$

② general case: need to bound $\boxed{A(\lambda) = \inf_\theta \|f_\theta - f_*\|_{L^2}^2 + \lambda \|\theta\|^2}$





Simple case = $\mathcal{X} = [0,1)$, p uniform in $[0,1)$

Fourier series : $f(x) = \sum_{m \in \mathbb{Z}} \hat{f}_m \underline{e^{2i\pi m x}}$ orthonormal basis in L^2

Parseval $\|f\|^2_{L_2} = \sum_{m \in \mathbb{Z}} |\hat{f}_m|^2$

$\varphi(x)_{m \in \mathbb{Z}} = e^{2i\pi m x} \hat{c}_m^{1/2}$ where $\hat{c}_m \in \mathbb{R}_+$

$K(x,x') = \langle \varphi(x), \varphi(x') \rangle = \sum_{m \in \mathbb{Z}} \varphi(x)_m \varphi(x')_m^* = \sum_{m \in \mathbb{Z}} \hat{c}_m e^{2i\pi m(x-x')}$

$= c(x-y)$ for $c \in L^2$

Defining $\mathcal{H} := \{ f_\theta(x) = \langle \theta, \varphi(x) \rangle, \ \|\theta\|^2 < \infty \}$.

$\|\theta\|^2 = \sum_{m \in \mathbb{Z}} |\hat{\theta}_m|^2 = \sum_{m \in \mathbb{Z}} \left| \frac{\hat{f}_m}{\hat{c}_m^{1/2}} \right|^2 = \sum_{m \in \mathbb{Z}} \frac{1}{\hat{c}_m} |\hat{f}_m|^2$

Ex: $\hat{c}_m = \frac{1}{1 + |m|^{2s}}$ $\qquad 1 + m^{2s}$

$= \|f\|^2_{L^2} + \|f^{(s)}\|^2_{L^2}$

$$A(\lambda) = \inf_{\Theta} \|b_\Theta - b_\theta\|_{L^2}^2 + \lambda \|b_\Theta\|^2 = \inf_{\Theta} \underbrace{\sum_{m\in\mathbb{Z}} |\hat{b}_{\Theta m} - \hat{b}_{\theta m}|^2 + \frac{|\hat{b}_{\Theta m}|^2}{\hat{c}_m}}_{\text{depends only on } \hat{b}_{\Theta m}}$$

$$= \sum_{m\in\mathbb{Z}} \frac{\lambda c_m^{-1}}{1+\lambda c_m^{-1}} |\hat{b}_{\theta m}|^2$$

Assumptions:

$$\hat{c}_m = \frac{1}{1+m^{2s}}$$

s - Sobolev space

$$\sum_{m\in\mathbb{Z}} |\hat{b}_{\theta m}|^2 (1+m^{2t}) < \infty.$$

$b_\theta \in t$ - Sobolev space

$$\boxed{t \le s}$$

$$\le \sum_{m\in\mathbb{Z}} \boxed{\frac{\lambda c_m^{-1}}{1+\lambda c_m^{-1}} \frac{1}{1+m^{2t}}} (1+m^{2t}) |\hat{b}_{\theta m}|^2$$

$\le C(\lambda)$ $\forall m$

$$\boxed{C(\lambda) \le \lambda^{t/s}}$$

$$\underbrace{\mathbb{E}\, R(f_{\hat\theta_n}) - R(f_*)}_{\text{excess risk}} \le \frac{1}{2\gamma R^2}\|f_\theta - f_*\|_{L^2}^2 + \frac{1}{2\gamma n}\|\theta\|^2 + \gamma G^2 R^2 \quad \forall \theta$$

$$\le \underbrace{\frac{1}{2\gamma R^2} \inf_\theta \left\{ \|f_\theta - f_*\|^2 + \frac{R^2}{n}\|\theta\|^2 \right\}}_{A\left(\frac{R^2}{n}\right) \le \left(\frac{R^2}{n}\right)^{t/s}} + \gamma G^2 R^2$$

$$\le \frac{1}{\gamma}\frac{1}{n^{t/s}} + \gamma \le \frac{1}{n^{t/2s}} \quad \text{x cst}$$

$$\gamma \propto \frac{1}{n^{t/2s}}$$

$$\mathcal{G}_{\Gamma_{L}}^{\;t<s}$$

$f_* \in t$-Sobolev space

model f_θ – s-Sobolev space $s > t$

ⓒ only upper bounds

① rates are slower than $t = s$

② power $t/2s$ increases with $t \Rightarrow$ adaptivity in d-dimension

- Sobolev space of order $s > \frac{d}{2}$ model
- f_σ is t-Sobolev space if $t < s$

$\Rightarrow$ rate $= O\left(n^{-t/2s}\right)$

$t = 1 \quad s = d/2$

rate $O\left(n^{-1/d}\right)$

Curse of dimensionality