A statistical study of regularized boosting methods

Gábor Lugosi

Universitat Pompeu Fabra, Barcelona

lugosi@upf.es

Nicolas Vayatis

Université Paris VI

vayatis@ccr.jussieu.fr

Overview

- Binary classification: notations
- Combining classifiers
- Heuristics of boosting algorithms
- Main result on consistency
- Extensions of the main result
- Simulations results

Binary classification

Observation: $X \in \mathbb{R}^d$, distribution μ

Label/Class: $Y \in \{-1, +1\}$

Regression function:

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$$

Data sample:

$$D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\} \text{ i.i.d.}$$

Classifier:

makes a prediction $g_n(X) \in \{-1, +1\}$

A performance/error measure:

$$L(g_n) = \mathbb{P}\{g_n(X) \neq Y \mid D_n\} = \mathbb{E}\mathbb{I}_{\{Y \cdot g_n(X) < 0\}}$$

Bayes classifier and error:

$$g^* = \arg\min_{\text{all } g} L(g)$$

$$L^* = L(g^*) = \mathbb{E}\{\min(\eta(X), 1 - \eta(X))\}$$

Learning

Inputs:

- a model class $\mathcal{C}$ of indicator functions
- a sample-based criterion:

$$\min_{g \in \mathcal{C}} K(g, D_n)$$

Output: a classifier g_n

Goal of learning:

minimize generalization error $L(g_n) \geq L^*$

Examples of learning algorithms:
perceptron, neural networks, decision trees

Learning (2)

Complexity trade-off: estimation vs. approximation

$$L(g_n)-L^* = (L(g_n)-\inf_{g\in\mathcal{C}} L(g))+(\inf_{g\in\mathcal{C}} L(g)-L^*)$$

Empirical risk minimization:

$$g_n = \arg\min_{g\in\mathcal{C}} \left(\widehat{L}_n(g) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i g(X_i)<0\}} \right)$$

From Vapnik-Chervonenkis (VC) theory:

If the class $\mathcal{C}$ is not "too large" (finite VC dimension):

$$L(g_n) - \inf_{g\in\mathcal{C}} L(g) \to 0 \text{ a.s.}$$

However: a poor class often leads to poor performance...

Solutions: increase class complexity or combine!

Combination methods

$$\Pi_1 \text{ distribution over } D_n \to \text{select } h_1 \in \mathcal{C}$$

$$\ldots$$

$$\Pi_t \text{ distribution over } D_n \to \text{select } h_t \in \mathcal{C}$$

$$\ldots$$

Final prediction:

$$g_n(X) = \text{sign}\left(\sum_t w_t h_t(X)\right)$$

where

- uniform weights $\to$ Bagging

  Π_t distribution of a bootstrap subsample

- adaptive weights $\to$ Boosting

  $w_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

  Π_t updated iteratively from Π_{t-1}

Combination methods

Estimator: $f = \sum_i w_i h_i \in \mathcal{F}$,

where $h_i \in \mathcal{C}$, w_i are (convex) weights.

Model class: $\mathcal{F} = \mathcal{L}(\mathcal{C})$ or $\mathrm{conv}(\mathcal{C})$

Classifier: $g(X) = \mathrm{sign}\ f(X)$

Remarks:

- convex hulls of simple classes have infinite complexity
- derived algorithms are amazingly efficient

Open question:

Do boosting algorithms overfit?

Formally: Statistical consistency? Practical efficiency?

Leo Breiman: "understanding why boosting works is the main open problem in Machine Learning".

Previous work

Origins: Schapire (1990), Freund and Schapire (1995, 1996) for boosting, Breiman (1996) for bagging

Empirical studies: too many!!!

$\rightarrow$ visit www.boosting.org

Boosting as gradient descent: Breiman (1997) Friedman-Hastie-Tibshirani (1998), Collins-Schapire-Singer (2000), Mason-Bartlett-Baxter-Frean (1999).

Margin analysis:
Schapire-Freund-Bartlett-Lee (1998), Koltchinskii-Panchenko (2000), Blanchard (2001)

About consistency: only very recent work from Breiman (2000), Jiang (2000, 2001), Mannor-Meir-Mendelson (2001), Bühlmann-Yu (2001), Zhang (2001), Mannor-Meir-Zhang (2002).

Common belief

Observation:

$$L(g) = L(f) \leq A(f) = \mathbb{E}\exp(-Y \cdot f(X))$$

Empirical criterion:

$$A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \exp(-Y_i f(X_i))$$

Minimizers:

$$\widehat{f}_n = \arg\min_{\mathcal{F}} A_n(f)$$

$$f^* = \arg\min_{\text{all } f} A(f) = \frac{1}{2}\log\left(\frac{\eta}{1-\eta}\right)$$

Hopefully:

$$L(\widehat{f}_n) \to L(f^*) = L^*, \ \text{ almost surely}$$

Our challenge: prove it!!!

More notations

Let $\phi : \mathbb{R} \to \mathbb{R}_+$

- strictly increasing,
- strictly convex,
- $\phi(x) \geq \mathbb{I}_{[x \geq 0]}$ for all $x \in \mathbb{R}$,
- $\lim_{x \to -\infty} \phi(x) = 0$

Cost functional:

$$A^\lambda(f) = A(\lambda f) = \mathbb{E}\phi(-\lambda Y \cdot f(X)) \ .$$

where λ is a smoothing parameter.

Empirical cost functional:

$$A_n^\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(-\lambda Y_i \cdot f(X_i)) \ .$$

Typical example: $\phi = \exp$

$\to$ other choices? ...

Main result

Assume

- $\mathcal{F} = \mathrm{conv}(\mathcal{C})$ contains the indicators of all subrectangles of $\mathbb{R}^d$
- Let λ_n such that $\lambda_n \to \infty$ and

$$\lambda_n \phi'(\lambda_n) \sqrt{\frac{\ln n}{n}} \to 0 \quad \text{as } n \to \infty$$

- $f_n = \widehat{f}_n^{\lambda_n} = \arg\min_{f \in \mathcal{F}} A_n^{\lambda_n}(f) \in \mathcal{F}$

Then, we have

$$\lim_{n\to\infty} L(f_n) = L^* \quad \text{almost surely.}$$

Comments

Universal result: noise-resistant strategy

Denseness assumption: fulfilled by decision trees with $T > d$ terminal nodes

Key of the result: the smoothing parameter λ governs complexity trade-off

- accurate estimation $\rightarrow$ λ small
- reduce approximation error $\rightarrow$ λ large

First simple lemma

Let f_n such that

$$\lim_{n\to\infty} A(f_n) = A^*$$

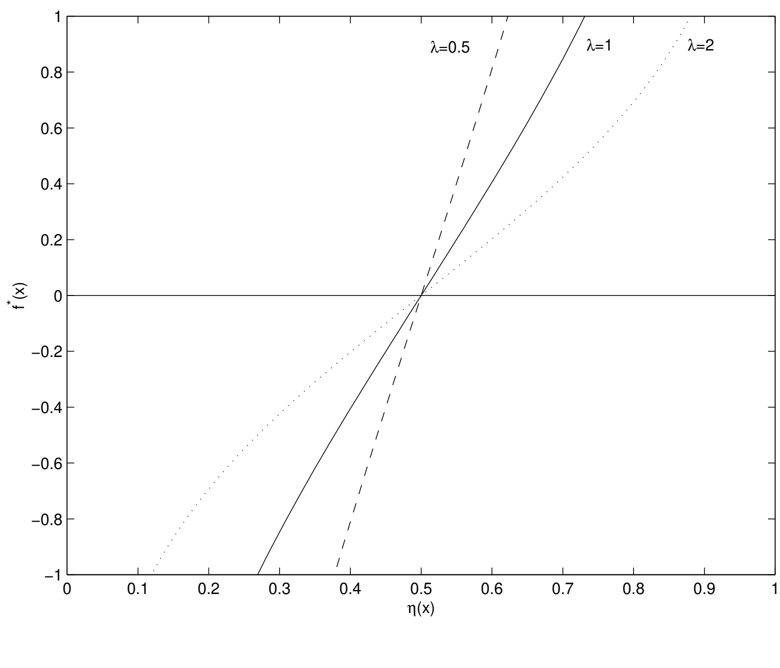
and

$$g_n(x) = \text{sign } f_n(X)\,.$$

Then

$$L(g_n) \to L^* \text{ a.s.}$$

Characterization of f^*_λ



Proof sketch

Estimation/approximation error decomposition with respect to the cost function ϕ:

$$
\begin{aligned}
&A(\lambda_n f_n) - A^* \\
&= \left(A^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A^{\lambda_n}(\overline{f}_{\lambda_n}) \right) + \left(\inf_{f \in \lambda_n \cdot \mathcal{F}} A(f) - A^* \right)
\end{aligned}
$$

where

$$
\overline{f}_\lambda = \arg\min_{f \in \mathcal{F}} A^\lambda(f)
$$

Estimation error:

We have

$$
A^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A^{\lambda_n}(\overline{f}_{\lambda_n}) \le 2 \sup_{f \in \mathcal{F}} \left| A^\lambda(f) - A_n^\lambda(f) \right|
$$

$$
\left.
\begin{array}{l}
A^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A_n^{\lambda_n}(\widehat{f}_n^{\lambda_n}) \\[3em]
A_n^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A^{\lambda_n}(\overline{f}_{\lambda_n})
\end{array}
\right\} \le \sup_{f \in \mathcal{F}} \left| A^\lambda(f) - A_n^\lambda(f) \right|
$$

Part I: Denseness

If $\mathcal{F}$ contains the indicators of all subrectangles of $\mathbb{R}^d$, then

$$\lim_{n\to\infty} \inf_{f\in\lambda\cdot\mathcal{F}} A(f) = A^*$$

where $A^* = \inf A(f)$ over all measurable f.

Part II: Concentration

For any $\delta > 0$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left|A^\lambda(f) - A_n^\lambda(f)\right|$$

$$\leq 4\lambda\phi'(\lambda)\sqrt{\frac{2V\ln(4n+2)}{n}} + \lambda\phi'(\lambda)\sqrt{\frac{\ln(1/\delta)}{2n}}\,.$$

(Koltchinskii & Panchenko (2000))

Second result

Strategy based on a penalized criterion is also Bayes-risk consistent.

Formally:

Consider positive $\lambda_k \to +\infty$ and

$$f_n = \arg\min_{k \geq 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) \ ,$$

where

$$\tilde{A}_n^{\lambda_k}(f) = A_n^{\lambda_k}(f) + 5\lambda_k \phi'(\lambda_k)\sqrt{\frac{V \ln n + \ln(nk)}{n}}$$

and

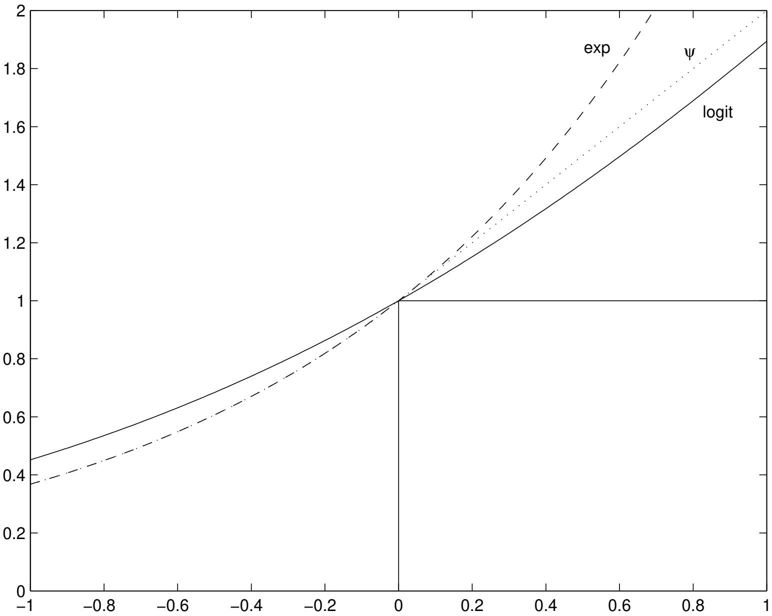
$$\widehat{f}_n^{\lambda_k} = \arg\min_{f \in \mathcal{F}} A_n^{\lambda_k}(f) \, .$$

Under the denseness assumption, we have

$$\lim_{n \to \infty} L(f_n) = L^* \quad \text{almost surely.}$$

Choices for the cost function

- $\phi(x) = \exp(x)$,
$$\phi'(\lambda) = \exp(\lambda)$$
- $\phi(x) = \text{logit}(x) := \log_2(1 + \exp(x))$,
$$\phi'(\lambda) = \frac{\exp(\lambda)}{1 + \exp(\lambda)}$$
- $\phi(x) = \psi(x) := \min\{\exp(x), |x| + 1\}$,
$$\phi'(\lambda) = 1$$



Simulations

Set-up

- Generate 6-dimensional synthetic data samples of size $n = 100, ..., 500$ from 'twonorm', 'threenorm', 'ringnorm' generators.
- For each λ, run the boosting algorithm to minimize $A_n^\lambda(f)$ over the convex hull of all decision stumps (CPU time from 10 to 50 seconds for 300 iterations).
- Estimate the expected cost $A^\lambda(\widehat{f_n^\lambda})$ and the generalization error $L(\widehat{f_n^\lambda})$ over a test set of size m.

Comments

- influence of the cost function
- comparison of the minimizers
- sensitivity to the level of noise

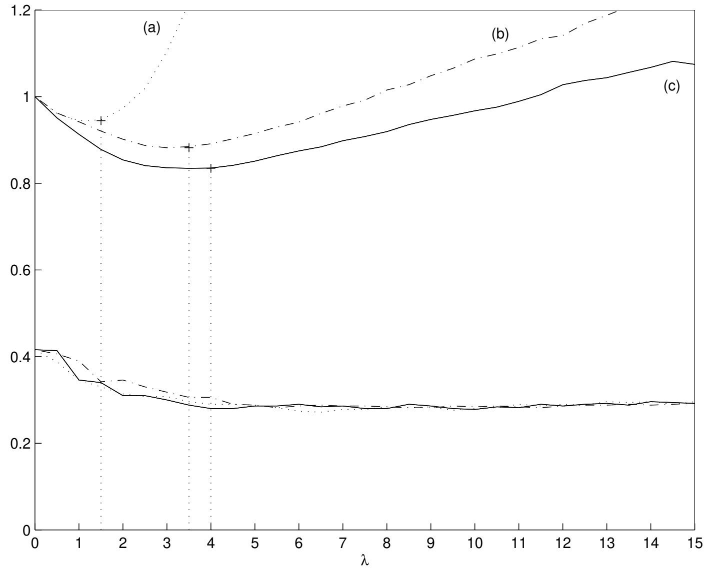


Figure 1: Threenorm. Cost $\phi = \psi$. $d = 6$. $\eta = 0.1$. $n = 100$. $m = 500$. Plot of the cost $A^{\lambda}(\widehat{f}_n^{\lambda})$ (upper curves) and test error (lower curves) for various cost functions (a) exp, (b) logit, (c) ψ.

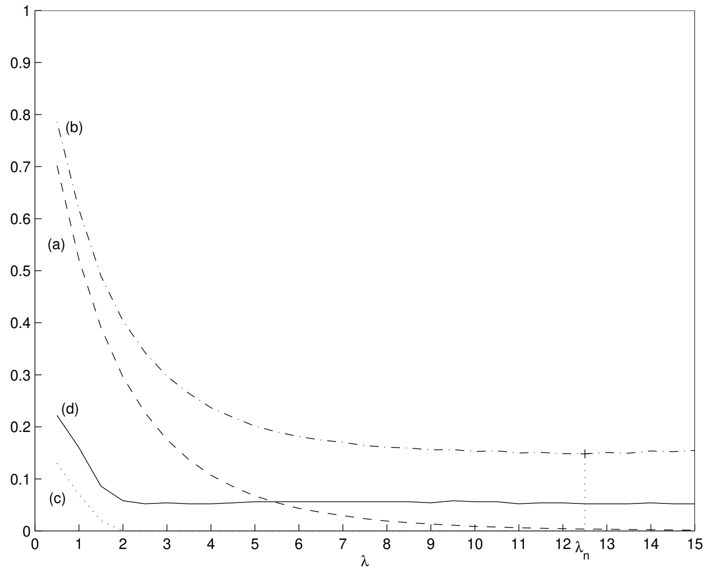


Figure 2: Twonorm. Cost $\phi = \psi$. $d = 6$. $n = 100$. $m = 500$. (a) $A_n^{\lambda}(\widehat{f}_n^{\lambda})$. (b) $A^{\lambda}(\widehat{f}_n^{\lambda})$. (c) training error. (d) test error.

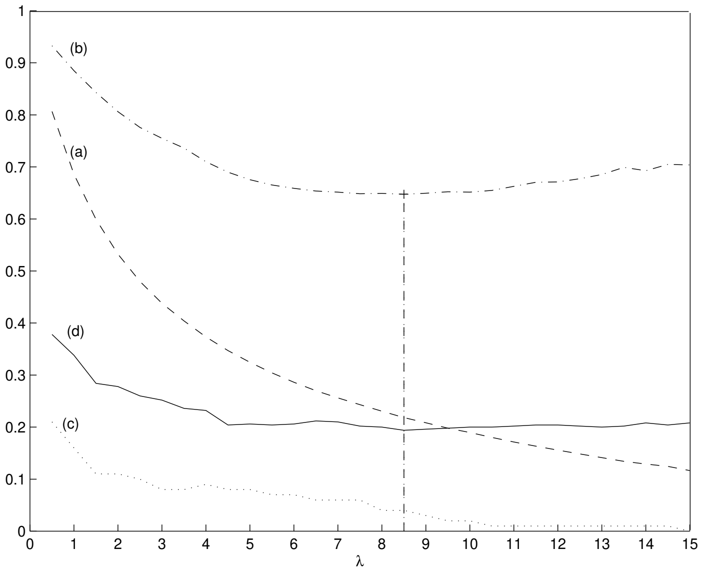


Figure 3: Threenorm. Cost $\phi = \psi$. $d = 6$. $n = 100$. $m = 500$. (a) $A_n^\lambda(\widehat{f}_n^\lambda)$. (b) $A^\lambda(\widehat{f}_n^\lambda)$. (c) training error. (d) test error.

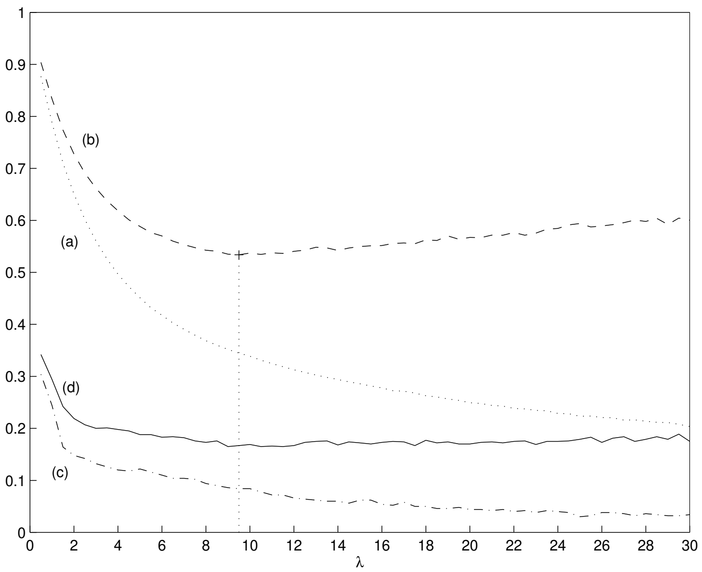


Figure 4: Threenorm. Cost $\phi = \phi$. $d = 6$. $n = 500$. $m = 1000$. (a) $A_n^\lambda(\widehat{f}_n^\lambda)$. (b) $A^\lambda(\widehat{f}_n^\lambda)$. (c) training error. (d) test error.

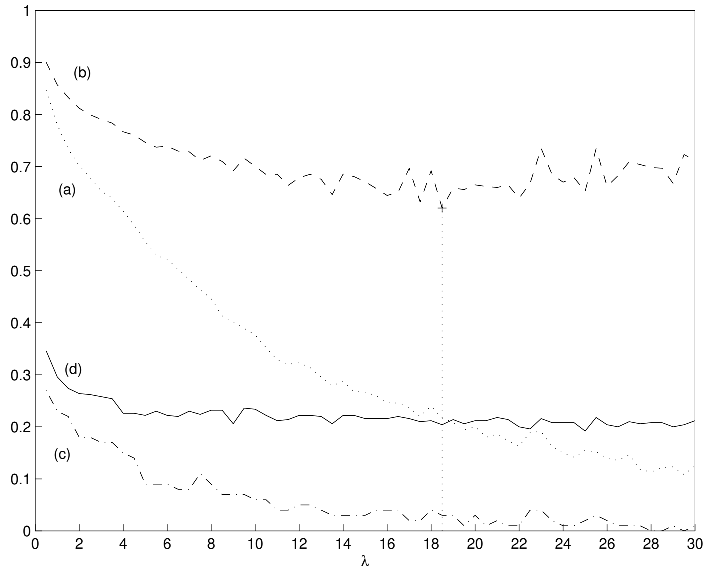


Figure 5: Ringnorm. Cost $\phi = \psi$. $d = 6$. $n = 100$. $m = 500$. (a) $A_n^{\lambda}(\widehat{f}_n^{\lambda})$. (b) $A^{\lambda}(\widehat{f}_n^{\lambda})$. (c) training error. (d) test error.

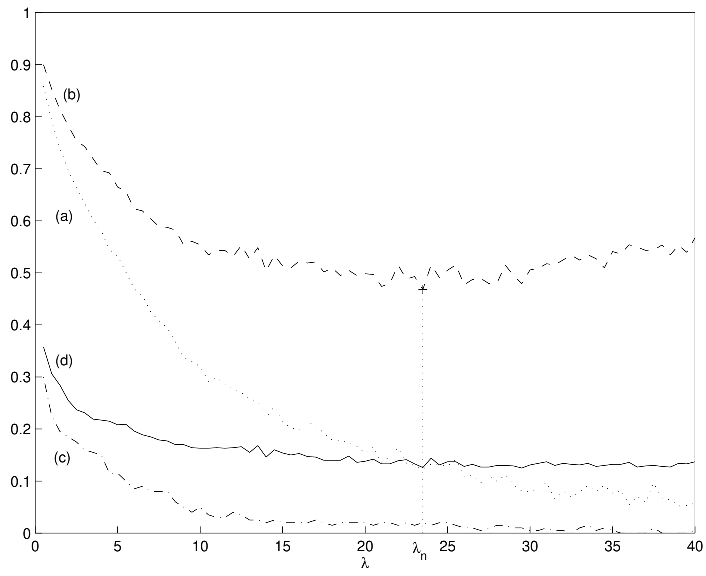


Figure 6: Ringnorm. Cost $\phi = \psi$. $d = 6$. $n = 200$. $m = 1000$. (a) $A_n^{\lambda}(\widehat{f}_n^{\lambda})$. (b) $A^{\lambda}(\widehat{f}_n^{\lambda})$. (c) training error. (d) test error.

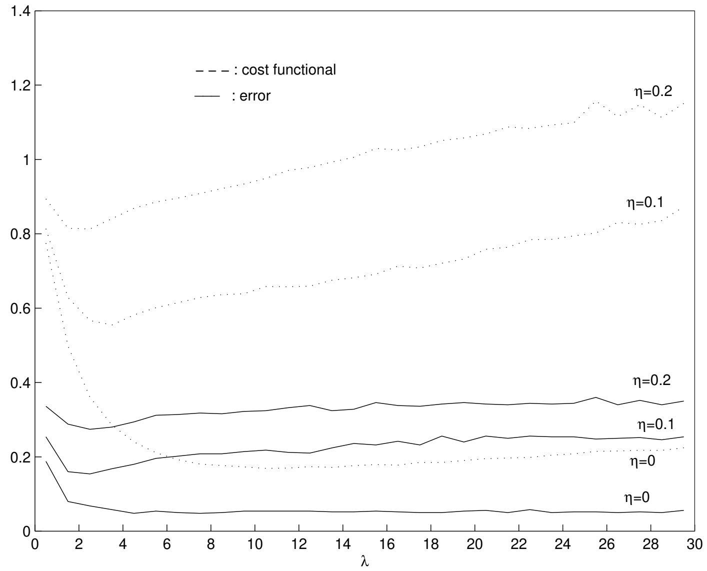


Figure 7: Twonorm. Cost $\phi = \psi$. $d = 6$. $n = 100$. $m = 500$. Plots of $A^{\lambda}(\widehat{f}_n^{\lambda})$ and of the test error for levels of noise $\eta = 0, 0.1, 0.2$.

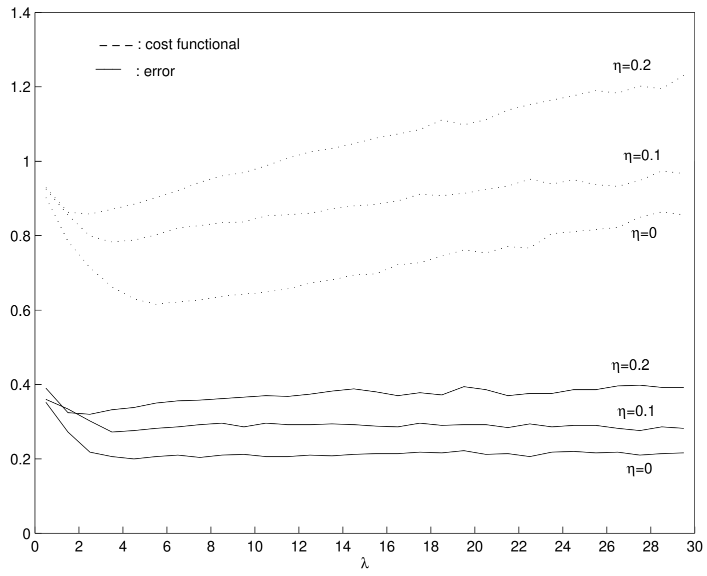


Figure 8: Threenorm. Cost $\phi = \psi$. $d = 6$. $n = 100$. $m = 500$. Plots of $A^{\lambda}(\widehat{f}_n^{\lambda})$ and of the test error for levels of noise $\eta = 0, 0.1, 0.2$.

Further work

- Non-asymptotic behavior has to be better understood.
- Rates of convergence for boosting. Use of approximation results.
- Distribution-dependent results.