Towards a better understanding of early stopping for boosting algorithms

Yuting Wei
Department of Statistics, Stanford University

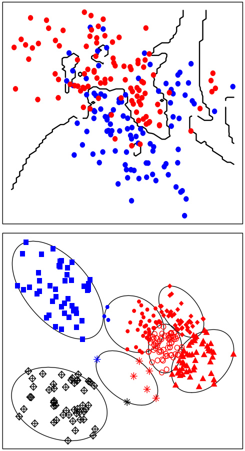
University of Cambridge
Nov 2nd, 2018

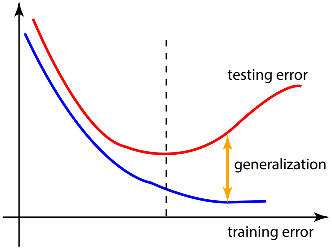
Fan Yang
ETH Zürich

Martin Wainwright
UC Berkeley

Overfitting and Generalization

Textbook examples

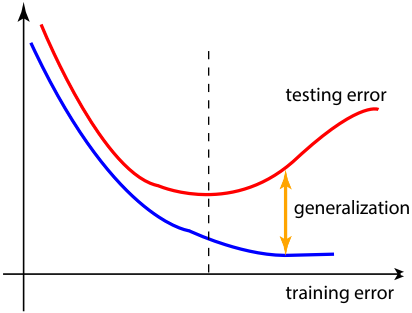




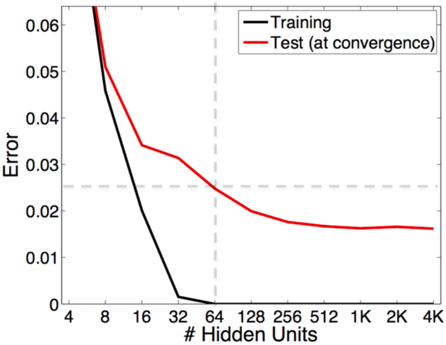
From *"The elements of statistical learning"* by Hastie, Tibshirani, Friedman

Lessons we learned...

- simpler models generalize better
- regularization is needed



Recent observed phenomenon



3-layer neural nets on MNIST (similar results on CIFAR)

Neyshabur, Tomioka, Srebro ICLR'15

Our reactions to technologies:

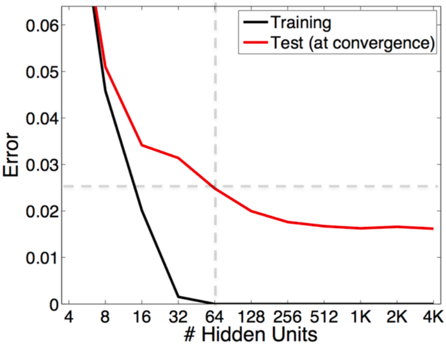
1. Anything that's in the world when you're born is normal and ordinary and is just a natural part of the way the world works.

2. Anything that's invented between when you're 15 and 35 is new and exciting and revolutionary and you can probably get a career in it.

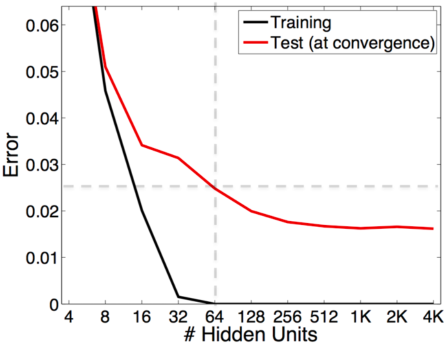
3. Anything invented after you're 35 is against the natural order of things.

—Douglas Adams, British author

Recent observed phenomenon(continued)

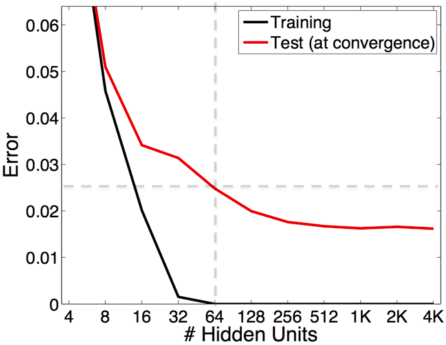


Recent observed phenomenon(continued)



- What is the right complexity measure?

Recent observed phenomenon(continued)



- What is the right complexity measure?
- What is this "error" here?

Does neural networks overfit the data?

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

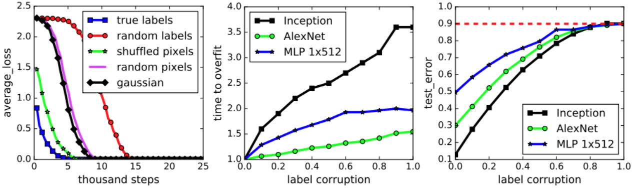
Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

- Can fit any training data, given enough time and large enough network



(a) learning curves (b) convergence slowdown (c) generalization error growth

Zhang et al. '17

We need some form of regularization!

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$
- f^* minimizes $\mathrm{Loss}(f; \mathbb{P})$ in function space $\mathcal{F}$

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$
- f^* minimizes $\text{Loss}(f; \mathbb{P})$ in function space $\mathcal{F}$
- Find a proper estimator $\widehat{f}$ in $\mathcal{F}$ based on D_n

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$
- f^* minimizes $\text{Loss}(f; \mathbb{P})$ in function space $\mathcal{F}$
- Find a proper estimator $\widehat{f}$ in $\mathcal{F}$ based on D_n
- Construct $\mathcal{L}_n$ and $\widehat{f} = \min_{f \in \mathcal{F}} \mathcal{L}_n(f)$

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$
- f^* minimizes $\text{Loss}(f; \mathbb{P})$ in function space $\mathcal{F}$
- Find a proper estimator $\widehat{f}$ in $\mathcal{F}$ based on D_n
- Construct $\mathcal{L}_n$ and $\widehat{f} = \min_{f \in \mathcal{F}} \mathcal{L}_n(f)$

For example:

- Squared loss $\mathcal{L}_n(f) = \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i))^2$

Empirical risk minimization

- Collect data $D_n = \{x_i, y_i\}_1^n \sim \mathbb{P}$
- f^* minimizes $\text{Loss}(f; \mathbb{P})$ in function space $\mathcal{F}$
- Find a proper estimator $\widehat{f}$ in $\mathcal{F}$ based on D_n
- Construct $\mathcal{L}_n$ and $\widehat{f} = \min_{f \in \mathcal{F}} \mathcal{L}_n(f)$

For example:

- Squared loss $\mathcal{L}_n(f) = \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i))^2$
- Function class with norm $\|f\|_2^2 = \int f^2(x) dx$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg\min_{\|f\|_{\mathcal{F}} \leq R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \ \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg \min_{\|f\|_{\mathcal{F}} \leq R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Bounds on population loss

$$\bar{\mathcal{L}}(f) = \mathbb{E}_{X,Y}\ \mathcal{L}_n(f; X_1^n, Y_1^n)$$

$$f^* := \operatorname{argmin}_{f \in \mathcal{F}} \bar{\mathcal{L}}(f)$$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg \min_{\|f\|_{\mathcal{F}} \leq R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Bounds on population loss

$$\bar{\mathcal{L}}(f) = \mathbb{E}_{X,Y}\ \mathcal{L}_n(f; X_1^n, Y_1^n)$$

$$f^* := \operatorname{argmin}_{f \in \mathcal{F}} \bar{\mathcal{L}}(f)$$

Excess loss: $\bar{\mathcal{L}}(\widehat{f}) - \bar{\mathcal{L}}(f^*)$

depends on complexity of $\mathcal{F}, f^*$ and radius R

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg\min_{\|f\|_{\mathcal{F}} \leq R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Bounds on population loss

$$\bar{\mathcal{L}}(f) = \mathbb{E}_{X,Y}\ \mathcal{L}_n(f; X_1^n, Y_1^n)$$
$$f^* := \operatorname*{argmin}_{f \in \mathcal{F}} \bar{\mathcal{L}}(f)$$

Excess loss: $\bar{\mathcal{L}}(\widehat{f}) - \bar{\mathcal{L}}(f^*)$

depends on complexity of $\mathcal{F}, f^*$ and radius R

Algorithmic regularization

Based on unconstrained problem

$$f \to \mathcal{L}_n(f; X_1^n, Y_1^n)$$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg \min_{\|f\|_{\mathcal{F}} \le R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Bounds on population loss

$$\bar{\mathcal{L}}(f) = \mathbb{E}_{X,Y}\, \mathcal{L}_n(f; X_1^n, Y_1^n)$$

$$f^* := \operatorname*{argmin}_{f \in \mathcal{F}} \bar{\mathcal{L}}(f)$$

Excess loss: $\bar{\mathcal{L}}(\widehat{f}) - \bar{\mathcal{L}}(f^*)$

depends on complexity of $\mathcal{F}, f^*$ and radius R

Algorithmic regularization

Based on unconstrained problem

$$f \to \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Generate a sequence of iterates $\{f^t\}_{t=1}^{\infty}$

$$f^{t+1} = f^t - \alpha^t g^t$$

From penalized to algorithmic regularization

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Penalized regularization

Risk minimization with constraints

$$\widehat{f} := \arg \min_{\|f\|_{\mathcal{F}} \leq R} \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Bounds on population loss

$$\bar{\mathcal{L}}(f) = \mathbb{E}_{X,Y} \, \mathcal{L}_n(f; X_1^n, Y_1^n)$$
$$f^* := \operatorname*{argmin}_{f \in \mathcal{F}} \bar{\mathcal{L}}(f)$$

Excess loss: $\bar{\mathcal{L}}(\widehat{f}) - \bar{\mathcal{L}}(f^*)$

depends on complexity of $\mathcal{F}, f^*$ and radius R

Algorithmic regularization

Based on unconstrained problem

$$f \to \mathcal{L}_n(f; X_1^n, Y_1^n)$$

Generate a sequence of iterates $\{f^t\}_{t=1}^{\infty}$

$$f^{t+1} = f^t - \alpha^t g^t$$

Regularization by "stopping early"

early stopped estimator depends on complexity of $\mathcal{F}, f^*$, step sizes and algorithm nature

Boosting via functional gradient descent

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n: \ \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Schapire'90, Freund & Schapire'95,'97, Breiman '95,'96, Mason et al.'99...

Boosting via functional gradient descent

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Given step size $\alpha^t > 0$,

$$f^{t+1} = f^t - \alpha^t g^t \qquad \text{where } g^t := \Pi_{\mathcal{F}}(\nabla \mathcal{L}_n(f^t))$$

Schapire'90, Freund & Schapire'95,'97, Breiman '95,'96, Mason et al.'99...

Boosting via functional gradient descent

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Given step size $\alpha^t > 0$,

$$f^{t+1} = f^t - \alpha^t g^t \qquad \text{where } g^t := \Pi_{\mathcal{F}}(\nabla \mathcal{L}_n(f^t))$$

e.g. $f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha X X^T (f^t(x_1^n) - y) \Rightarrow$ re-fitting the residual

Schapire'90, Freund & Schapire'95,'97, Breiman '95,'96, Mason et al.'99...

Boosting via functional gradient descent

Empirical loss function	Function class $\mathcal{F}$
$\mathcal{L}_n : \mathcal{F} \to \mathbb{R}$	Norm $\|\cdot\|_{\mathcal{F}}$

Given step size $\alpha^t > 0$,

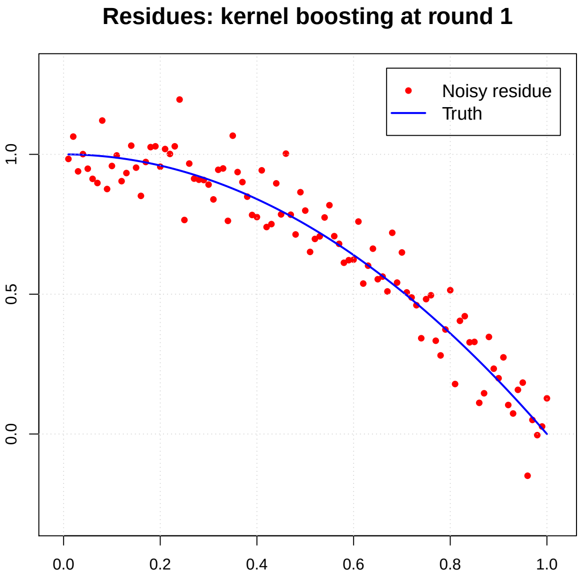
$$f^{t+1} = f^t - \alpha^t g^t \qquad \text{where } g^t := \Pi_{\mathcal{F}}(\nabla \mathcal{L}_n(f^t))$$

e.g. $f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha X X^T (f^t(x_1^n) - y) \Rightarrow$ re-fitting the residual

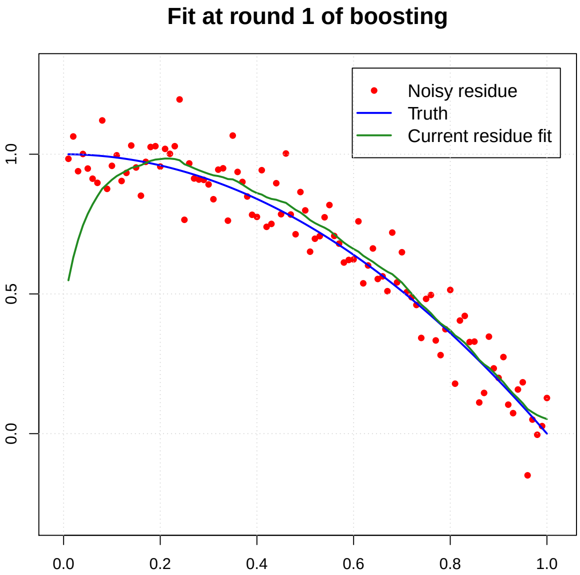
- ℓ_2-boosting: least-squares loss $\frac{1}{2}(y - f(x))^2$
- LogitBoost: logistic regression loss $\ln(1 + e^{-yf(x)})$
- AdaBoost: exponential loss $\exp(-yf(x))$

Schapire'90, Freund & Schapire'95,'97, Breiman '95,'96, Mason et al.'99...

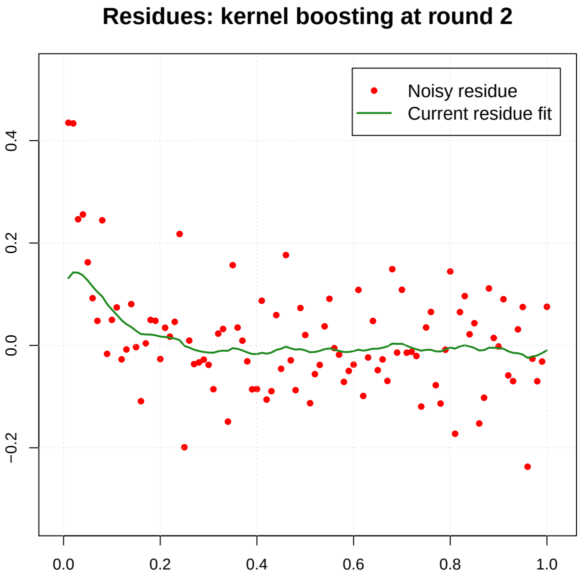
Boosting with a Laplacian kernel



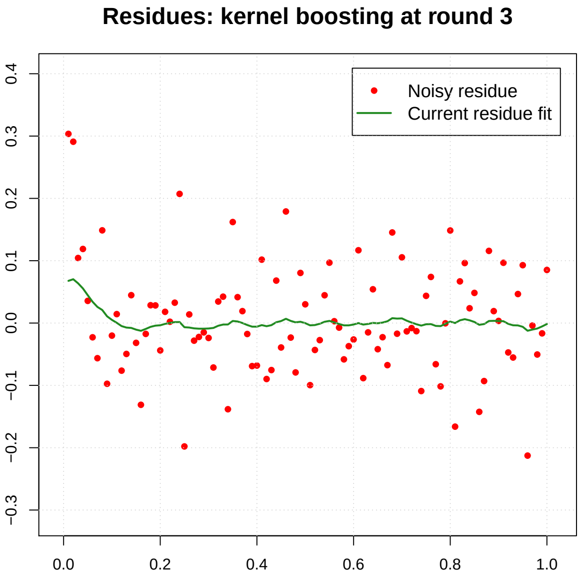
Boosting with a Laplacian kernel



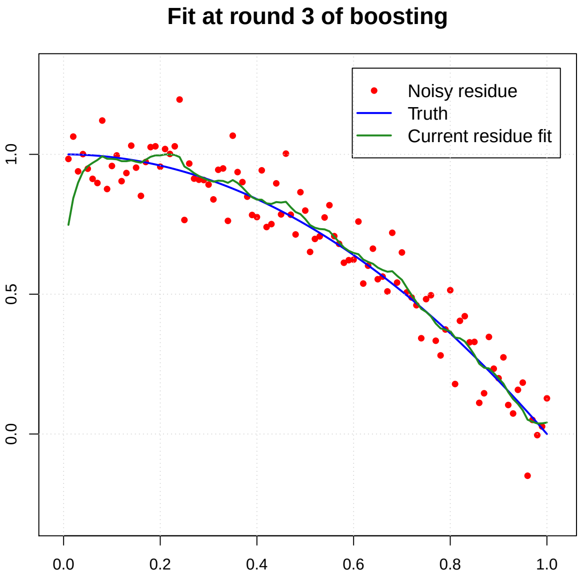
Boosting with a Laplacian kernel



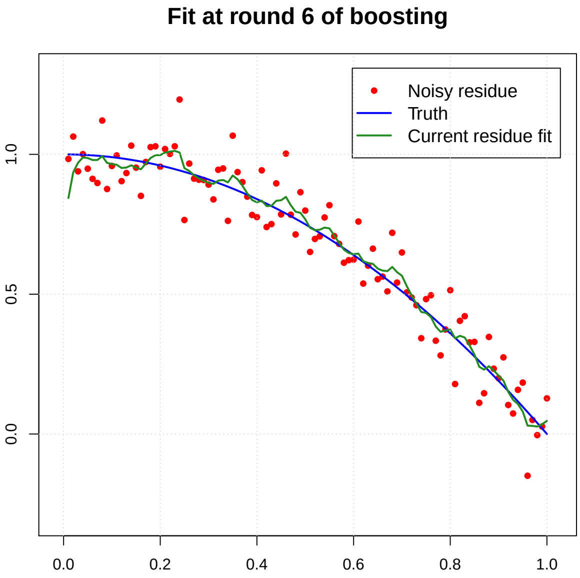
Boosting with a Laplacian kernel



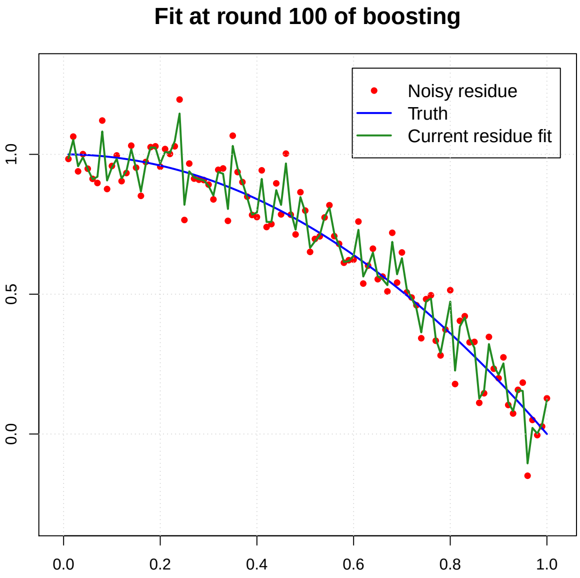
Boosting with a Laplacian kernel



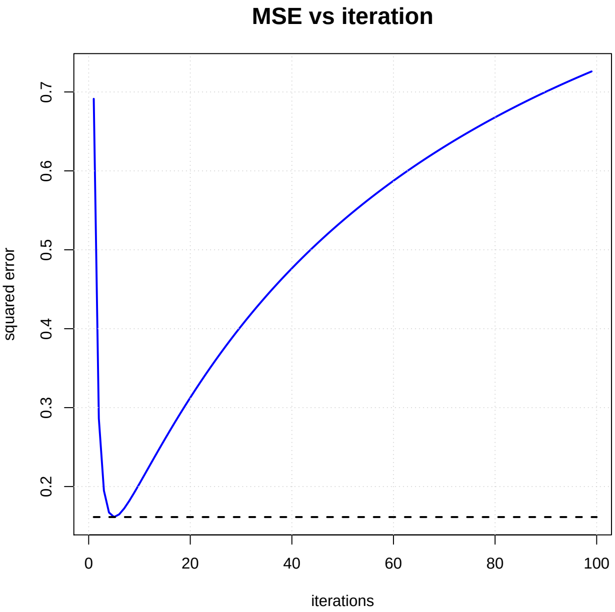
Boosting with a Laplacian kernel



Boosting with a Laplacian kernel



Mean-squared error $\|f^t - f^*\|_2^2$ versus iteration

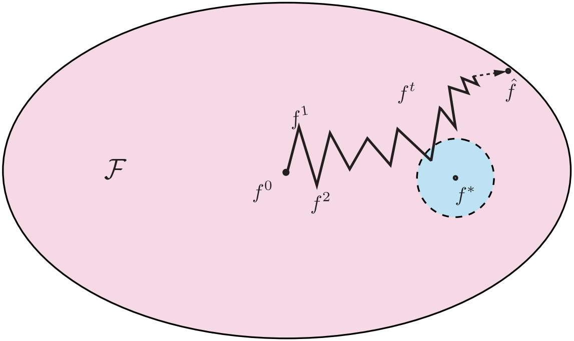


Early stopping for boosting

- Boosting algorithm:

$$f^{t+1} = f^t - \alpha^t \Pi_{\mathcal{F}}(\nabla \mathcal{L}_n(f^t))$$

Generate a sequence: $f^1, f^2, \cdots f^T \cdots f^\infty$.



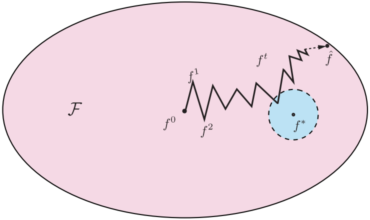
Early stopping for boosting

- Boosting algorithm:

  $$f^{t+1} = f^t - \alpha^t \Pi_{\mathcal{F}}(\nabla \mathcal{L}_n(f^t))$$

  Generate a sequence:

  $$f^1, f^2, \;\cdots\; f^T \;\cdots\; f^\infty.$$



What we would like:

Data-dependent stopping time T such that

$\mathcal{L}_n(f^T) \approx \mathcal{L}_n(f^*)$ where f^* is the population minimizer

$\|f^T - f^*\|_2 \to 0$ at the minimax-optimal rate as $n \to \infty$

Related results

- Consistency result of boosting algorithms
  [Zhang'04, Zhang and Yu'05, Bartlett and Traskin'06, Bickel et al.'06]

Related results

- Consistency result of boosting algorithms
  [Zhang'04, Zhang and Yu'05, Bartlett and Traskin'06, Bickel et al.'06]
- Optimal rate
  - Bühlmann and Yu'03 proves optimality for early stopping of ℓ_2-**boosting** for spline classes
  - Raskutti et al.'13 considers ℓ_2-**boosting** for kernel classes and establishes connection to the localized Rademacher complexity

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function and constant step size α, and any iterate $t = 1, 2, \ldots, \lfloor \frac{1}{\delta_n^2} \rfloor$,

$$\underbrace{\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*)}_{\text{Excess loss}} \lesssim \underbrace{\frac{1}{\alpha t}}_{\text{Opt error}} + \underbrace{\delta_n^2}_{\text{Stat error}},$$

with high probability over the randomized realization.

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function and constant step size α, and any iterate $t = 1, 2, \ldots, \lfloor \frac{1}{\delta_n^2} \rfloor$,

$$\underbrace{\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*)}_{\text{Excess loss}} \lesssim \underbrace{\frac{1}{\alpha t}}_{\text{Opt error}} + \underbrace{\delta_n^2}_{\text{Stat error}},$$

with high probability over the randomized realization.

Statistical error is determined by fixed point equation:

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \min\left\{1, \frac{\mu_i}{\delta^2}\right\}} = \frac{\delta}{\sigma},$$

where μ_i are the eigenvalues of the kernel operator, and σ is the noise level.

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function and constant step size α, and any iterate $t = 1, 2, \ldots, \lfloor \frac{1}{\delta_n^2} \rfloor$,

$$\underbrace{\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*)}_{\text{Excess loss}} \lesssim \underbrace{\frac{1}{\alpha t}}_{\text{Opt error}} + \underbrace{\delta_n^2}_{\text{Stat error}},$$

with high probability over the randomized realization.

Function space $\mathcal{F}$:

- Reproducing kernel Hilbert space (RKHS) Wahba'90, Gu' 02, Berlinet and Thomas-Agnan'04
- Examples: splines functions, polynomials, Lipschitz functions, Sobolev functions...

supp

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function and constant step size α, and any iterate $t = 1, 2, \ldots, \lfloor \frac{1}{\delta_n^2} \rfloor$,

$$\underbrace{\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*)}_{\text{Excess loss}} \lesssim \underbrace{\frac{1}{\alpha t}}_{\text{Opt error}} + \underbrace{\delta_n^2}_{\text{Stat error}},$$

with high probability over the randomized realization.

Loss functions:

- Regression (e.g. least squares)
- Classification (e.g. Logistic, Adaboost)

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function, constant step size α, and stopping criteria $T = \lfloor \frac{1}{\delta_n^2} \rfloor$, the excess loss

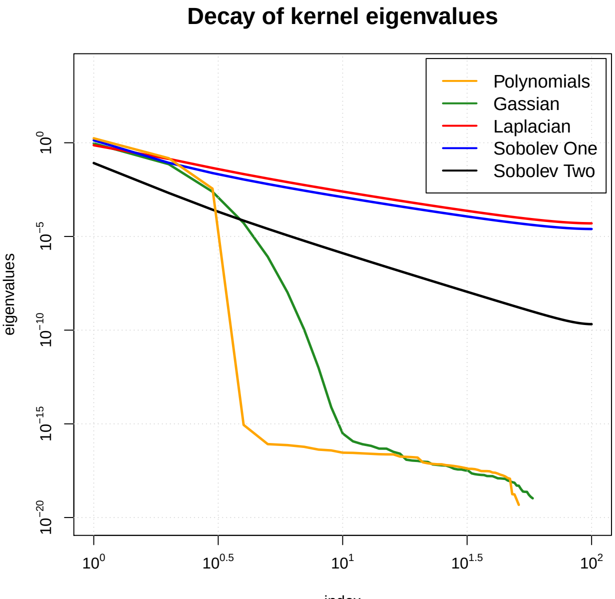
$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2.$$

Statistical error is determined by fixed point equation:

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \min\left\{1, \frac{\mu_i}{\delta^2}\right\}} = \frac{\delta}{\sigma},$$

where μ_i are the eigenvalues of the kernel operator, and σ is the noise level.

Main results



Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function, constant step size α, and stopping criteria $T = \lfloor \frac{1}{\delta_n^2} \rfloor$, the excess loss

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2.$$

- Examples:

function class $\mathcal{F}$	δ_n^2
Polynomial with degree D	$\frac{D}{n}$
Gaussian kernel space	$\frac{\sqrt{\log n}}{n}$
Lipchitz functions	$n^{-2/3}$
β-smooth kernel space, d-dim	$n^{-\frac{2\beta}{2\beta+d}}$

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function, constant step size α, and stopping criteria $T = \lfloor \frac{1}{\delta_n^2} \rfloor$, the excess loss

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2.$$

- Examples:

function class $\mathcal{F}$	δ_n^2	T
Polynomial with degree D	$\frac{D}{n}$	$\frac{n}{D}$
Gaussian kernel space	$\frac{\sqrt{\log n}}{n}$	$\frac{n}{\sqrt{\log n}}$
Lipchitz functions	$n^{-2/3}$	$n^{2/3}$
β-smooth kernel space, d-dim	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{\frac{2\beta}{2\beta+d}}$

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function, constant step size α, and stopping criteria $T = \lfloor \frac{1}{\delta_n^2} \rfloor$, the excess loss

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2.$$

- Examples:

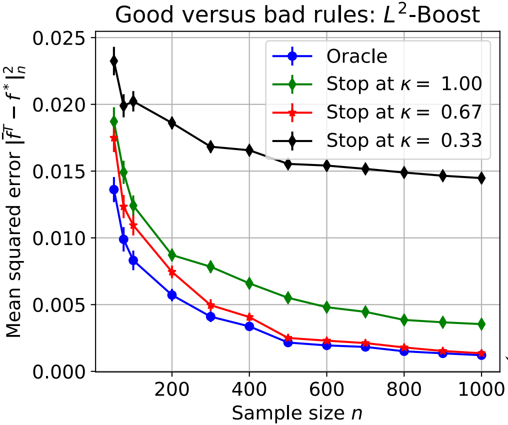
function class $\mathcal{F}$	δ_n^2	T
Polynomial with degree D	$\frac{D}{n}$	$\frac{n}{D}$
Gaussian kernel space	$\frac{\sqrt{\log n}}{n}$	$\frac{n}{\sqrt{\log n}}$
Lipchitz functions	$n^{-2/3}$	$n^{2/3}$
β-smooth kernel space, d-dim	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{\frac{2\beta}{2\beta+d}}$

Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps

Numerical results: Logit-Boost

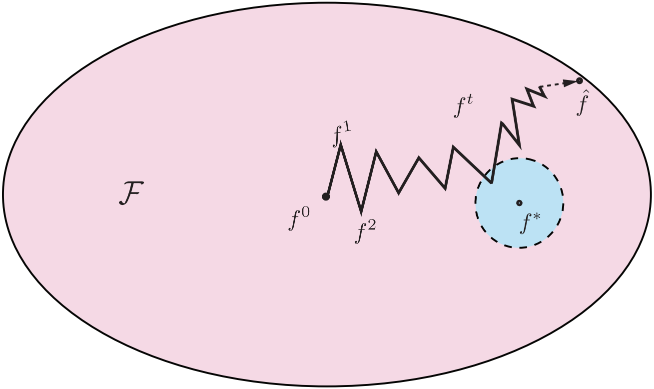
- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^\kappa$ steps



Oracle best!

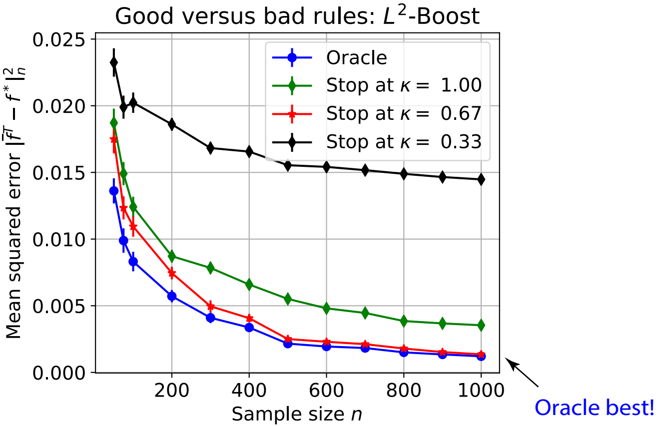
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps



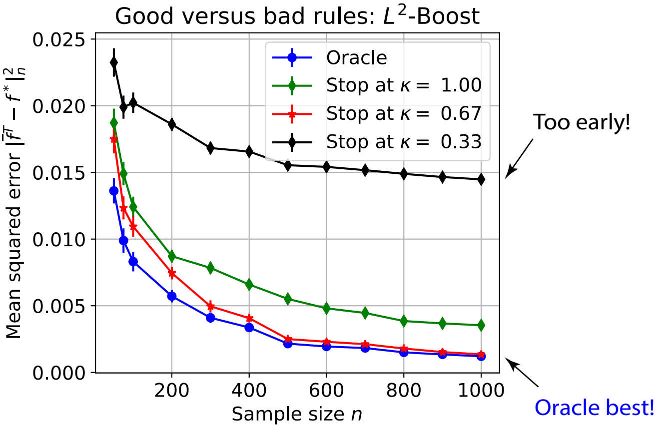
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps



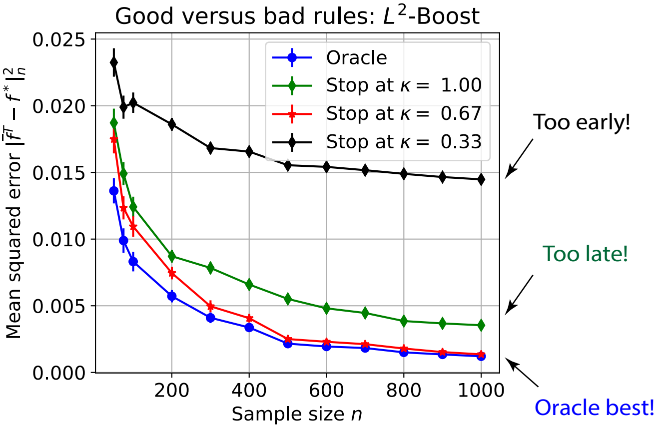
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps



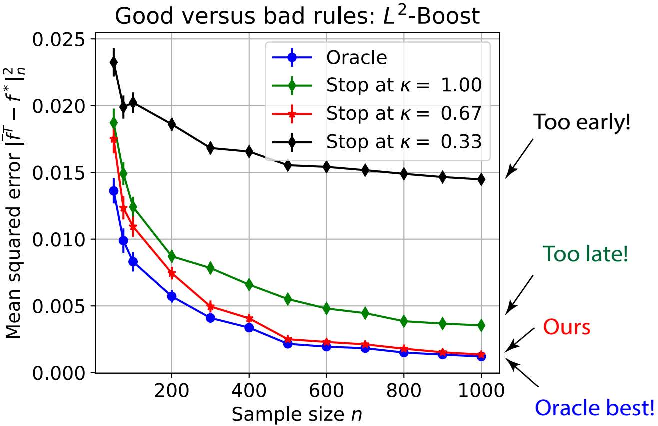
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps



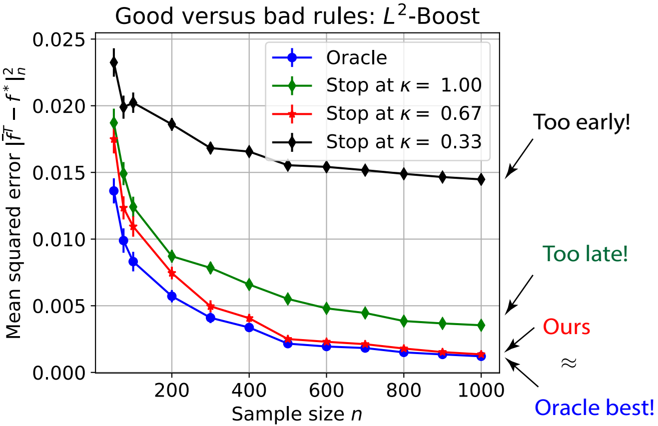
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^\kappa$ steps



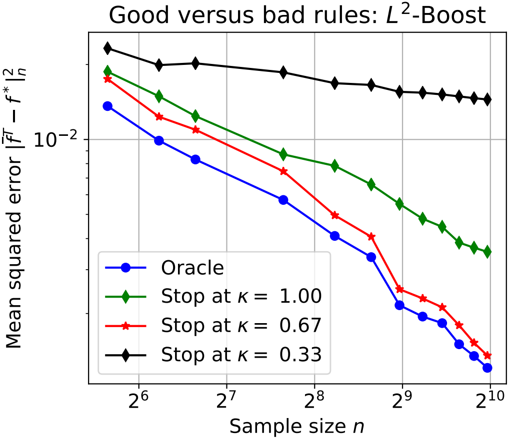
Numerical results: Logit-Boost

- Setting: $\mathbb{P}(y_i = 1) = \frac{\exp(2f^*(x_i))}{1+\exp(2f^*(x_i))}$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^{\kappa}$ steps

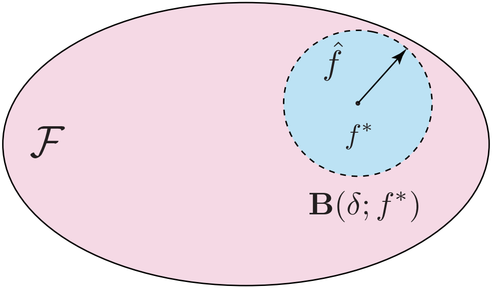


Numerical results: L^2-Boost (logscale)

- Setting: $y_i = f^*(x_i) + w_i$ where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$
- Stop after $\propto n^\kappa$ steps



Tools for sharp analysis



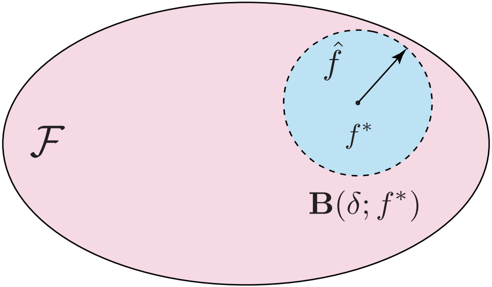
Gaussian complexity

How much you can align with i.i.d. noise sequence $\{w_i\}_1^n \sim N(0,1)$?

$$\mathcal{G}_n(\quad, \mathcal{F}) = \mathbb{E}_w \sup_{\boxed{f \in \mathcal{F}}} \boxed{\left| \frac{1}{n} \sum_{i=1}^{n} w_i (f(x_i) - f^*(x_i)) \right|}$$

(e.g., van de Geer'00, Bartlett et al.'05, Koltchinski '06)

Tools for sharp analysis



Localized Gaussian complexity

How much you can align with i.i.d. noise sequence $\{w_i\}_1^n \sim N(0,1)$?

$$\mathcal{G}_n(\boxed{\delta}, \mathcal{F}) = \mathbb{E}_w \sup_{\substack{f \in \mathcal{F} \\ \boxed{\|f - f^*\| \leq \delta}}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i (f(x_i) - f^*(x_i)) \right|$$

(e.g., van de Geer'00, Bartlett et al.'05, Koltchinski '06)

Main results

Theorem (W*, Yang* & Wainwright '17)

For any kernel class $\mathcal{F}$, any regular loss function, constant step size α, and stopping criteria $T = \lfloor \frac{1}{\delta_n^2} \rfloor$, the excess loss

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2.$$

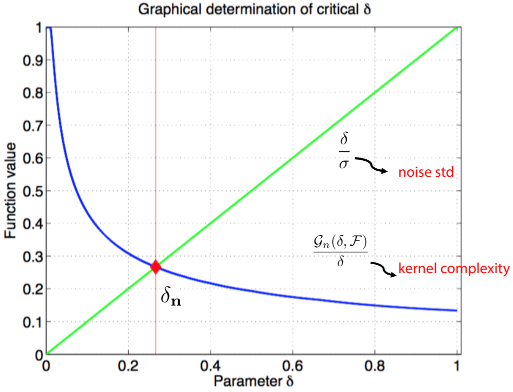
Statistical error is determined by fixed point equation:

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \min\left\{1, \frac{\mu_i}{\delta^2}\right\}} = \frac{\delta}{\sigma},$$

where μ_i are the eigenvalues of the kernel operator, and σ is the noise level.

Fixed point equation

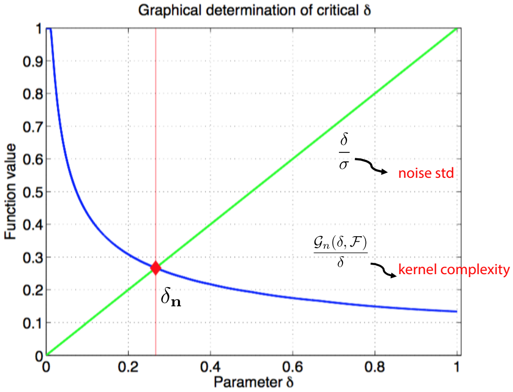
- Stopping rule T depends on critical radius δ_n



$$\frac{\mathcal{G}_n(\delta, \mathcal{F})}{\delta} = \frac{\delta}{\sigma}$$

*van de Geer'00, Bartlett'02, Koltchinskii'07, Raskutti et al.'13

Fixed point equation

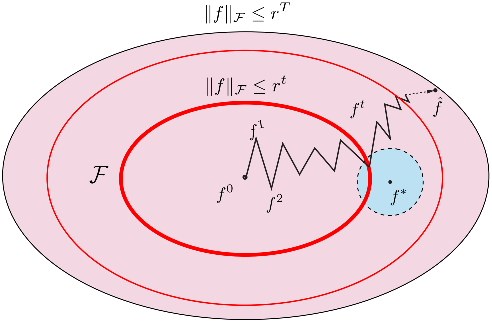


$$\frac{\mathcal{G}_n(\delta, \mathcal{F})}{\delta} = \frac{\delta}{\sigma}$$

- penalized estimator $\equiv$ early-stopped estimator

*van de Geer'00, Bartlett'02, Koltchinskii'07, Raskutti et al.'13

Geometric intuition in boosting analysis



- Boosted sequence $\{f^t\}_{t=1}^{\infty}$ takes a particular path
- Effective function classes $\mathcal{F}^t$ explored at iteration t increases

Minimax optimality

Our early stopped estimator:

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \;\lesssim\; \delta_n^2$$

Theorem (W*, Yang* & Wainwright '17)

Given any kernel class $\mathcal{F}$, and i.i.d. samples $\{y_i\}_{i=1}^n$ from a class of generalized linear model with some function f^* then

$$\inf_{\widehat{f}} \boxed{\sup_{\|f^*\|_{\mathcal{H}} \leq 1}} \;\mathbb{E}\|\widehat{f} - f^*\|_n^2 \gtrsim \delta_n^2.$$

(Yang et al.'17)

Minimax optimality

Our early stopped estimator:

$$\mathcal{L}(\bar{f}^t) - \mathcal{L}(f^*) \lesssim \delta_n^2$$

Theorem (W*, Yang* & Wainwright '17)

Given any kernel class $\mathcal{F}$, and i.i.d. samples $\{y_i\}_{i=1}^n$ from a class of generalized linear model with some function f^* then

$$\boxed{\inf_{\widehat{f}}} \sup_{\|f^*\|_{\mathcal{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \gtrsim \delta_n^2.$$

(Yang et al.'17)

Running time v.s. kernel complexity

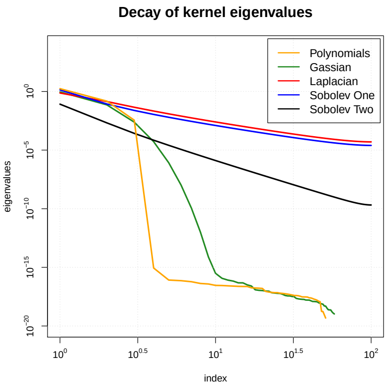
Table 1: Epochs to overfit (Laplacian)

Label	MNIST	SVHN	TIMIT
Original	4	8	3
Random	7	21	4

Table 2: Epochs to overfit (Gaussian)

Label	MNIST	SVHN	TIMIT
Original	20	46	7
Random	873	1066	22

Belkin et al.'18



Running time v.s. kernel complexity

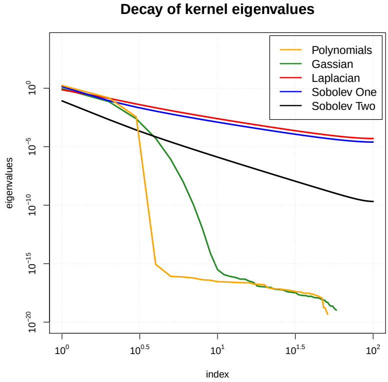
Table 1: Epochs to overfit (Laplacian)

Label	MNIST	SVHN	TIMIT
Original	4	8	3
Random	7	21	4

Table 2: Epochs to overfit (Gaussian)

Label	MNIST	SVHN	TIMIT
Original	20	46	7
Random	873	1066	22

Belkin et al.'18



Theoretically predicted running times to statistical precision:

Kernel	Laplacian	Gaussian
Time	$(\frac{n}{\sigma^2})^{2/3}$	$\frac{n}{\sigma^2}$

From kernels to neural networks

To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal
Department of Computer Science and Engineering
Ohio State University

Kernel Methods for Deep Learning

Youngmin Cho and Lawrence K. Saul
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404
{yoc002,saul}@cs.ucsd.edu

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot
École Polytechnique Fédérale de Lausanne
arthur.jacot@netopera.net

Franck Gabriel
Imperial College London
franckrgabriel@gmail.com

Clément Hongler
École Polytechnique Fédérale de Lausanne
clement.hongler@epfl.ch

Conclusion

- An effective way of early-stopping for boosting algorithms

Conclusion

- An effective way of early-stopping for boosting algorithms
- Connection between regularization through **penalization** and regularization through **early-stopping** over RKHS

Open questions

- Generalization ✓
  - $\bar{f}^t \to f^t$
  - kernel class $\to$ broader function classes

Open questions

- Generalization ✓
  - $\bar{f}^t \to f^t$
  - kernel class $\to$ broader function classes
- Boosting trees

Open questions

- Generalization ✓
  - $\bar{f}^t \to f^t$
  - kernel class $\to$ broader function classes
- Boosting trees
- Non-convex loss functions

References

- Y. Wei, F. Yang, and M. J. Wainwright. (2017) Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *NIPS'17* and arXiv (https://arxiv.org/abs/1707.01543)

Thanks! Questions?

Supplementary: RKHS

- Symmetric **kernel function** $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- RKHS is the closure of $f(\cdot) = \sum_{j \geq 1} \alpha_j \mathbb{K}(\cdot, x_j)$
- Reproducing relation

$$\langle f, \mathbb{K}(\cdot, x) \rangle_{\mathcal{F}} = f(x) \quad \text{for all } f \in \mathcal{F}$$

- Inner product

$$\langle f_1, f_2 \rangle_{\mathcal{F}} = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \alpha_i \beta_j \mathbb{K}(x_i, x_j)$$

for $f_1(\cdot) = \sum_{i=1}^{\ell_1} \alpha_i \mathbb{K}(\cdot, x_i)$ and $f_2(\cdot) = \sum_{j=1}^{\ell_2} \beta_j \mathbb{K}(\cdot, x_j)$

back

Supplementary: Proof idea

Key lemma 1

For any stepsize and any iteration t we have

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \le \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathcal{F}}^2 - \|\Delta^{t+1}\|_{\mathcal{F}}^2\Big\} + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle.$$

Key lemma 2

With high probability, we have

$$\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta\rangle \le 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathcal{F}} + \frac{m}{c}\|\Delta\|_n^2$$