Early Stopping for kernel boosting algorithms: A general analysis with localized complexities

Fanny Yang*, Yuting Wei*,
Martin Wainwright (*equal contribution)
UC Berkeley

NIPS Conference, December 2017

Problem setting

- Regression in functional space $\mathcal{F}$, with $(x_i, y_i) \sim \mathbb{P}$ for i = 1, … n

Problem setting

- Regression in functional space $\mathcal{F}$, with $(x_i, y_i) \sim \mathbb{P}$ for i = 1, … n

$$f^\star = \arg\min_{f \in \mathcal{F}} \underbrace{\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \phi(y_i, f(x_i))\Big]}_{\text{population loss } \mathcal{L}}$$

Problem setting

- Regression in functional space $\mathcal{F}$, with $(x_i, y_i) \sim \mathbb{P}$ for i = 1, … n

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))}_{\text{empirical loss } \mathcal{L}_n} \quad \text{vs.} \quad f^\star = \arg\min_{f \in \mathcal{F}} \underbrace{\mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))\Big]}_{\text{population loss } \mathcal{L}}$$

Problem setting

- Regression in functional space $\mathcal{F}$, with $(x_i, y_i) \sim \mathbb{P}$ for i = 1, … n

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))}_{\text{empirical loss } \mathcal{L}_n} \quad \text{vs.} \quad f^\star = \arg\min_{f \in \mathcal{F}} \underbrace{\mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))\Big]}_{\text{population loss } \mathcal{L}}$$

- Find $\hat{f}$ via functional gradient descent (~ boosting) on empirical loss

$$f^t = f^{t-1} - \alpha^t g^t \quad \text{with} \quad g^t \sim \frac{\partial \mathcal{L}_n}{\partial f}\bigg|_{f=f^t}$$

Problem setting

- Regression in functional space $\mathcal{F}$, with $(x_i, y_i) \sim \mathbb{P}$ for i = 1, … n

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))}_{\text{empirical loss } \mathcal{L}_n} \quad \text{vs.} \quad f^\star = \arg\min_{f \in \mathcal{F}} \underbrace{\mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))\Big]}_{\text{population loss } \mathcal{L}}$$

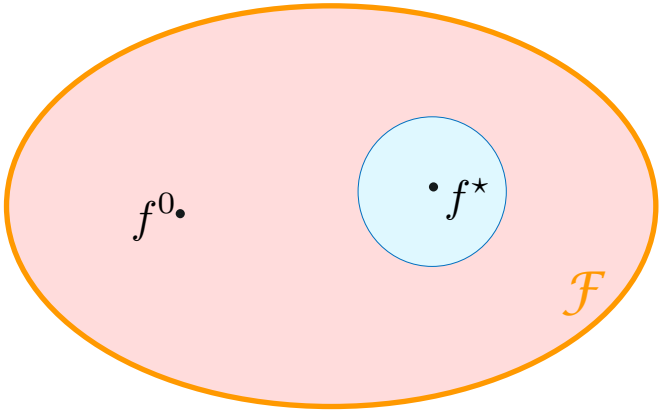
- Find $\hat{f}$ via functional gradient descent (~ boosting) on empirical loss

$$f^t = f^{t-1} - \alpha^t g^t \quad \text{with} \quad g^t \sim \left.\frac{\partial \mathcal{L}_n}{\partial f}\right|_{f=f^t}$$

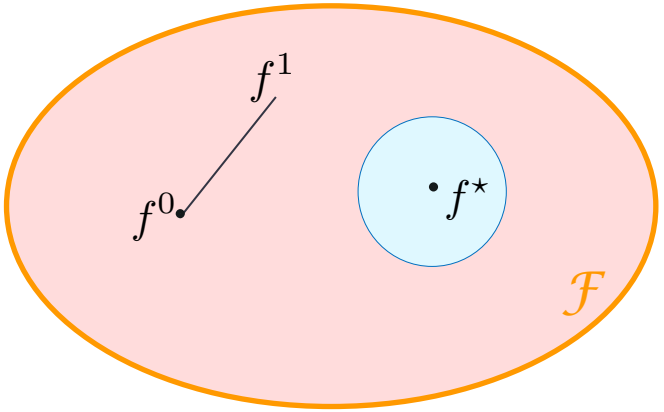
PROBLEM: Running until convergence to $\hat{f}$ might be overfitting!

$\rightarrow$ $\|\hat{f} - f^\star\|$ larger than optimal

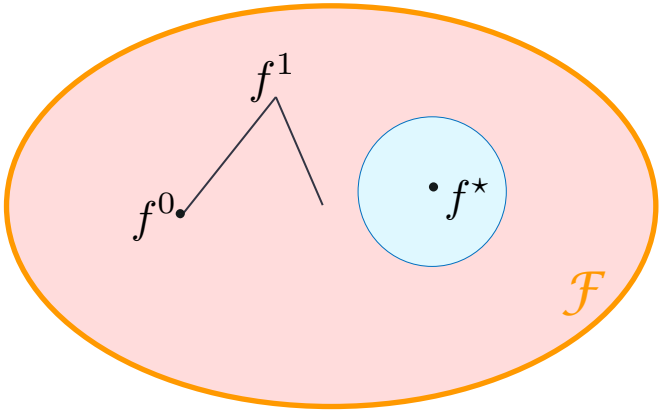
Overfitting and regularization



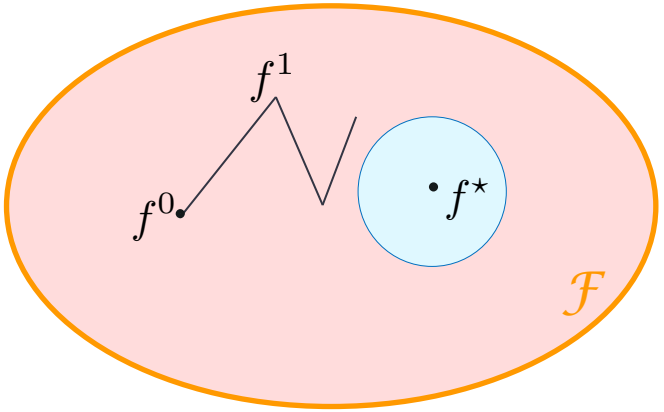
Overfitting and regularization



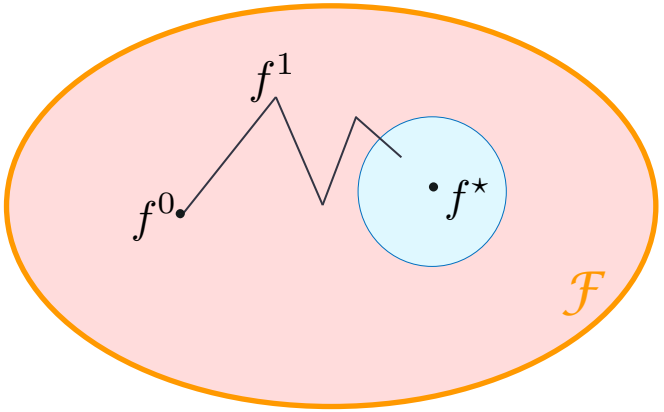
Overfitting and regularization



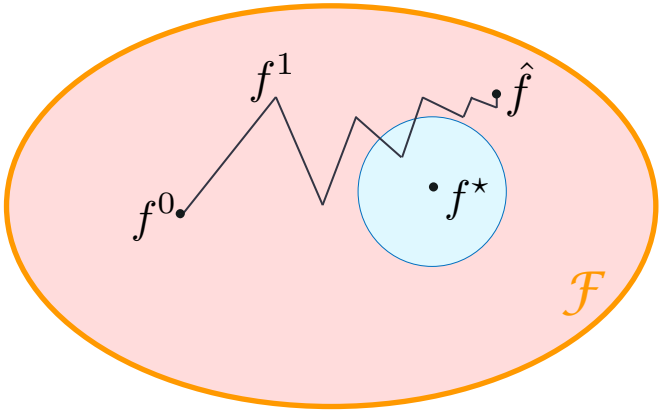
Overfitting and regularization



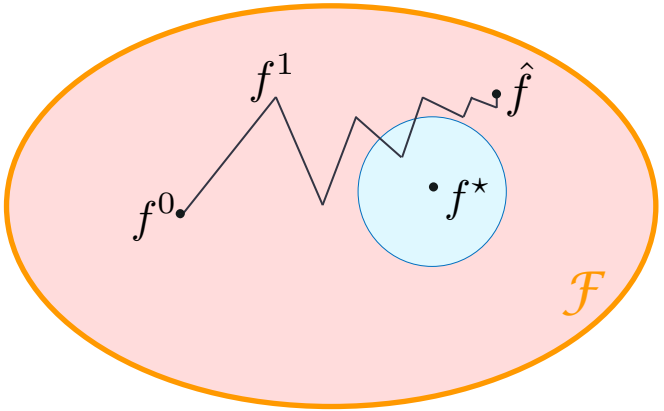
Overfitting and regularization

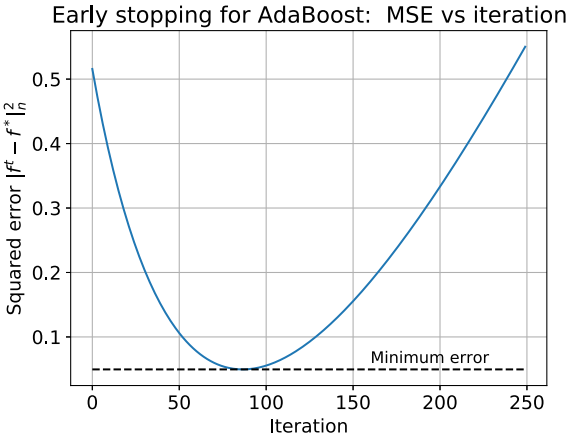


Overfitting and regularization

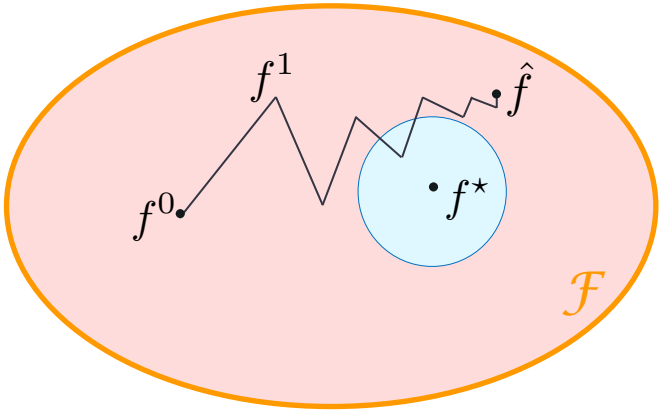


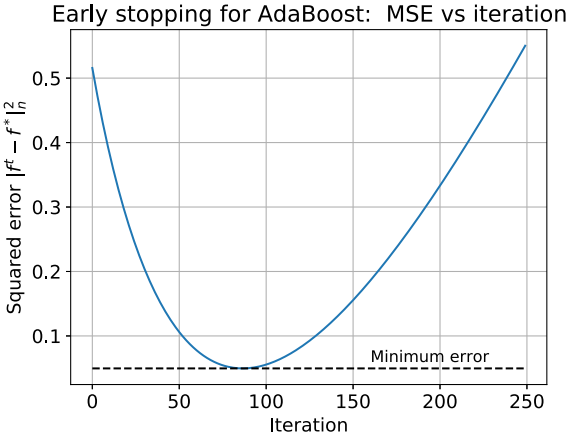
Overfitting and regularization





Overfitting and regularization

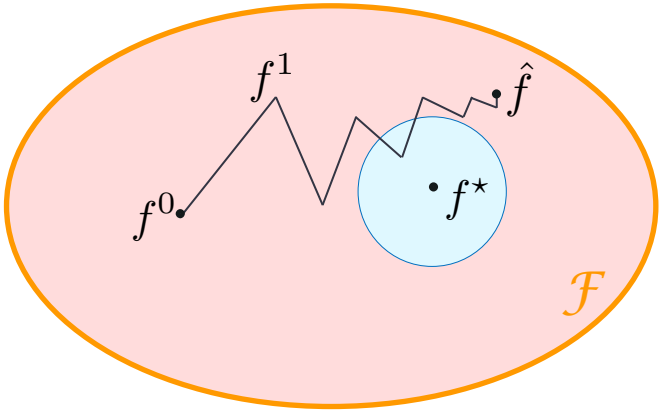


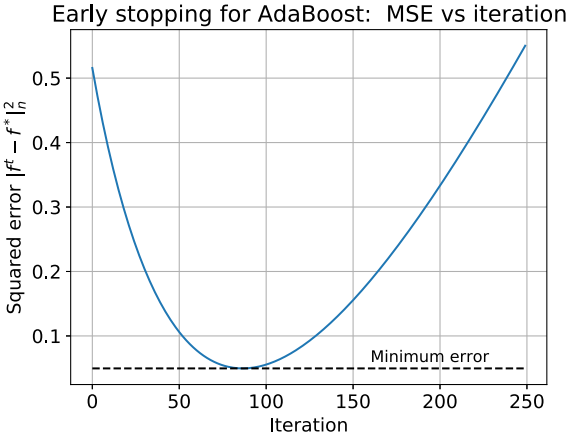


- Solve different problem: (penalized regularization)

$$f_{\text{pen}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i)) + \|f\|$$

Overfitting and regularization





- Solve different problem: (penalized regularization)

$$f_{\text{pen}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i)) + \|f\|$$

- Early stopping the iterates: (algorithmic regularization)

$$f_{\text{ES boost}} = f^{T(\mathcal{F},n)}$$

Previous work

Both types of regularization yield identical statistical rates for $\phi(y, f(x)) = \|y - f(x)\|^2$, i.e.

$$\|f_{\text{pen}} - f^\star\|_n \sim \|f_{\text{ES boost}} - f^\star\|_n$$

Previous work

Both types of regularization yield identical statistical rates for $\phi(y, f(x)) = \|y - f(x)\|^2$, i.e.

$$\|f_{\text{pen}} - f^\star\|_n \sim \|f_{\text{ES boost}} - f^\star\|_n$$

Our contribution

- Stat. rates for early stopping can be proven using key quantities in penalized regularization (localized complexities, critical radius)

Previous work

Both types of regularization yield identical statistical rates for $\phi(y, f(x)) = \|y - f(x)\|^2$, i.e.

$$\|f_{\text{pen}} - f^\star\|_n \sim \|f_{\text{ES boost}} - f^\star\|_n$$

Our contribution

- Stat. rates for early stopping can be proven using key quantities in penalized regularization (localized complexities, critical radius)
- This new technique extends guarantees to ϕ beyond least squares

Learn more at Poster #215

"Early Stopping for kernel boosting algorithms: A general analysis with localized complexities"

Yuting Wei*

Fanny Yang*

Martin Wainwright