COS324: Introduction to Machine Learning

Lecture $10\frac{1}{2}$: Generalization Revisited

Prof. Elad Hazan & Prof. Yoram Singer

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- Assume $\mathcal{Y} = \{-1, +1\}$ [binary classification]

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- Assume $\mathcal{Y} = \{-1, +1\}$ [binary classification]
- Probability label is $+1$ (-1 respectively)

$$\mathbb{P}[Y = +1|\mathbf{x}] = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x})} = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x}, -1) + \mathcal{D}(\mathbf{x}, +1)}$$

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- Assume $\mathcal{Y} = \{-1, +1\}$ [binary classification]
- Probability label is $+1$ (-1 respectively)

$$\mathbb{P}[Y = +1|\mathbf{x}] = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x})} = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x}, -1) + \mathcal{D}(\mathbf{x}, +1)}$$

- Given $\mathbf{x} \in \mathcal{X}$ we define a predictor

$$f^\star(x) = \text{sign}\left(\mathbb{P}[Y = +1|\mathbf{x}] - \frac{1}{2}\right)$$

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- Assume $\mathcal{Y} = \{-1, +1\}$ [binary classification]
- Probability label is $+1$ (-1 respectively)

$$\mathbb{P}[Y = +1|\mathbf{x}] = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x})} = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x}, -1) + \mathcal{D}(\mathbf{x}, +1)}$$

- Given $\mathbf{x} \in \mathcal{X}$ we define a predictor

$$f^\star(x) = \text{sign}\left(\mathbb{P}[Y = +1|\mathbf{x}] - \frac{1}{2}\right)$$

- $f^\star$ is a minimizer of $\mathcal{L}_{\mathcal{D}}^{0-1}$ (may not be efficiently computable)

Optimal Predictor

- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- Assume $\mathcal{Y} = \{-1, +1\}$ [binary classification]
- Probability label is $+1$ (-1 respectively)

$$\mathbb{P}[Y = +1|\mathbf{x}] = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x})} = \frac{\mathcal{D}(\mathbf{x}, +1)}{\mathcal{D}(\mathbf{x}, -1) + \mathcal{D}(\mathbf{x}, +1)}$$

- Given $\mathbf{x} \in \mathcal{X}$ we define a predictor

$$f^\star(x) = \text{sign}\left(\mathbb{P}[Y = +1|\mathbf{x}] - \frac{1}{2}\right)$$

- $f^\star$ is a minimizer of $\mathcal{L}_{\mathcal{D}}^{0-1}$ (may not be efficiently computable)
- How "far" is a learned classifier from $f^\star$?

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Gap between the learned and optimal predictors

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^{\star}) = ?$$

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Gap between the learned and optimal predictors

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^{\star}) = ?$$

- Let us define the following

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Gap between the learned and optimal predictors

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^{\star}) = ?$$

- Let us define the following
  - (Hypothetical) Best in hypothesis class w.r.t $\mathcal{D}$ & 0-1 loss

$$h^{\star} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_D^{0-1}(h)$$

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Gap between the learned and optimal predictors

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^{\star}) = ?$$

- Let us define the following
  - (Hypothetical) Best in hypothesis class w.r.t $\mathcal{D}$ & 0-1 loss

$$h^{\star} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_D^{0-1}(h)$$

  - Best in hypothesis class w.r.t S & 0-1 loss

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_S^{0-1}(h)$$

- Algorithm $\mathcal{A}$, surrogate loss $\ell(\cdot)$, sample S, learned predictor

$$\mathcal{A}(S) = \arg\widehat{\min}_{h \in \mathcal{H}} \mathcal{L}_S^{\ell}(h)$$

- Gap between the learned and optimal predictors

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^{\star}) = ?$$

- Let us define the following
  - (Hypothetical) Best in hypothesis class w.r.t $\mathcal{D}$ & 0-1 loss

$$h^{\star} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_D^{0-1}(h)$$

  - Best in hypothesis class w.r.t S & 0-1 loss

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_S^{0-1}(h)$$

  - Solution found by learning algorithm using ℓ

$$\hat{\hat{h}} \stackrel{\text{def}}{=} \mathcal{A}(S)$$

Error Decomposition

- ε_{app} approximation error w.r.t to hypothesis class $\mathcal{H}$

$$\varepsilon_{\text{app}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)$$

Error Decomposition

- ε_{app} approximation error w.r.t to hypothesis class $\mathcal{H}$

$$\varepsilon_{\text{app}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)$$

- ε_{est} estimation error due to finite sample S

$$\varepsilon_{\text{est}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h}) - \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star)$$

Error Decomposition

- ε_{app} approximation error w.r.t to hypothesis class $\mathcal{H}$

$$\varepsilon_{\text{app}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)$$

- ε_{est} estimation error due to finite sample S

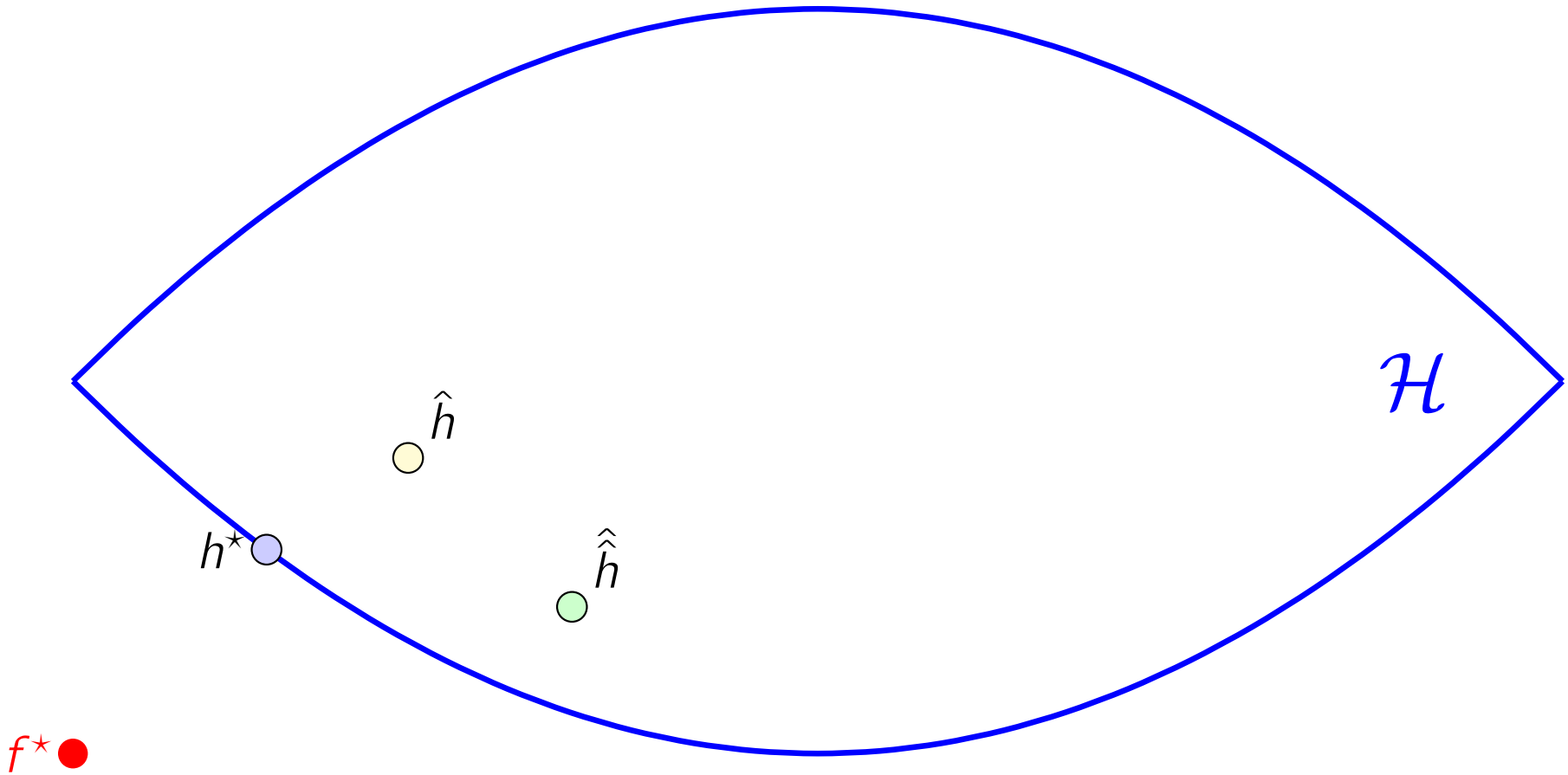
$$\varepsilon_{\text{est}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h}) - \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star)$$

- ε_{opt} optimization error

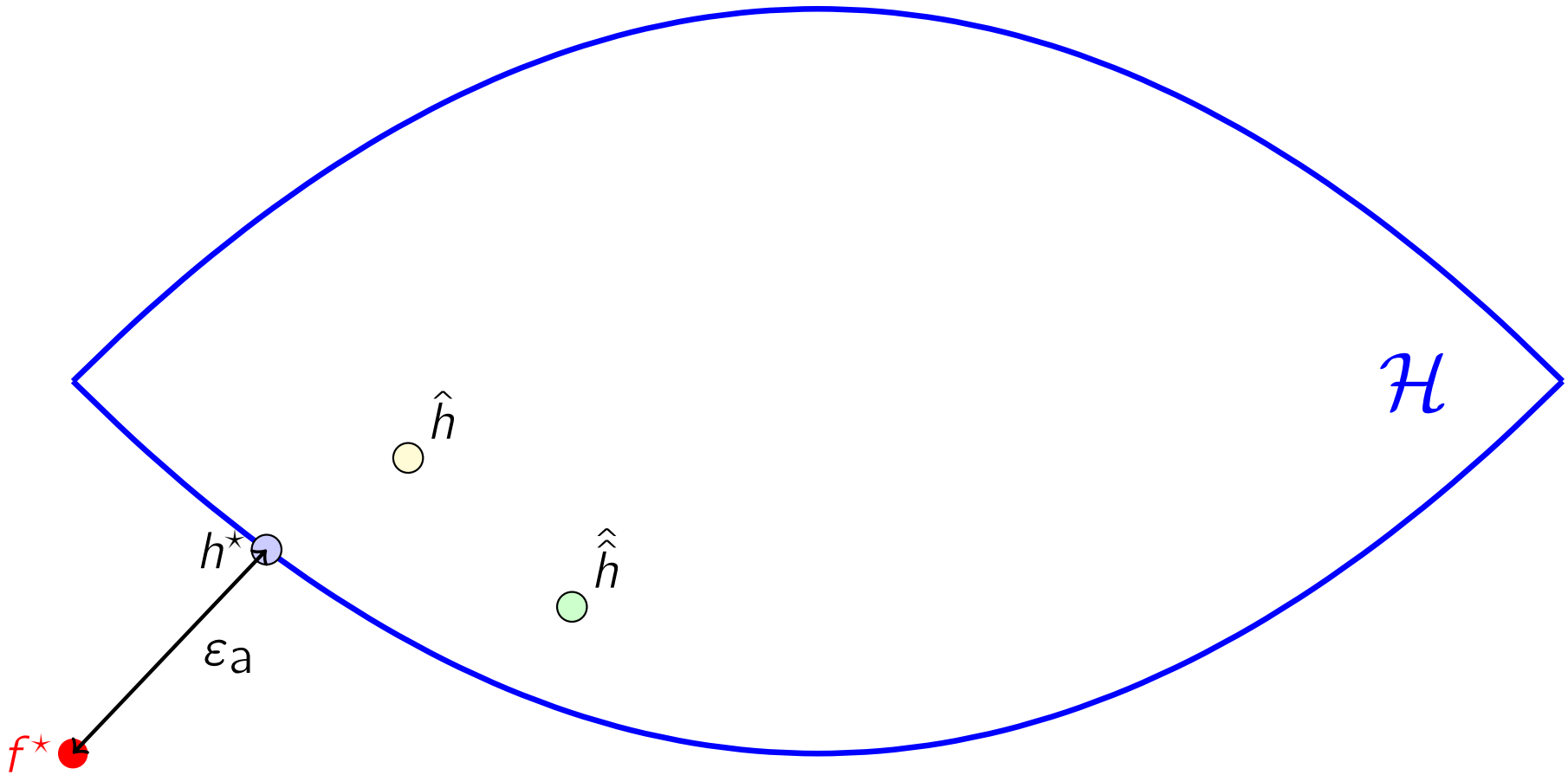
$$\varepsilon_{\text{opt}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{\hat{h}}) - \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h})$$

(surrogate loss ℓ, stochastic gradients, finite #iterations)

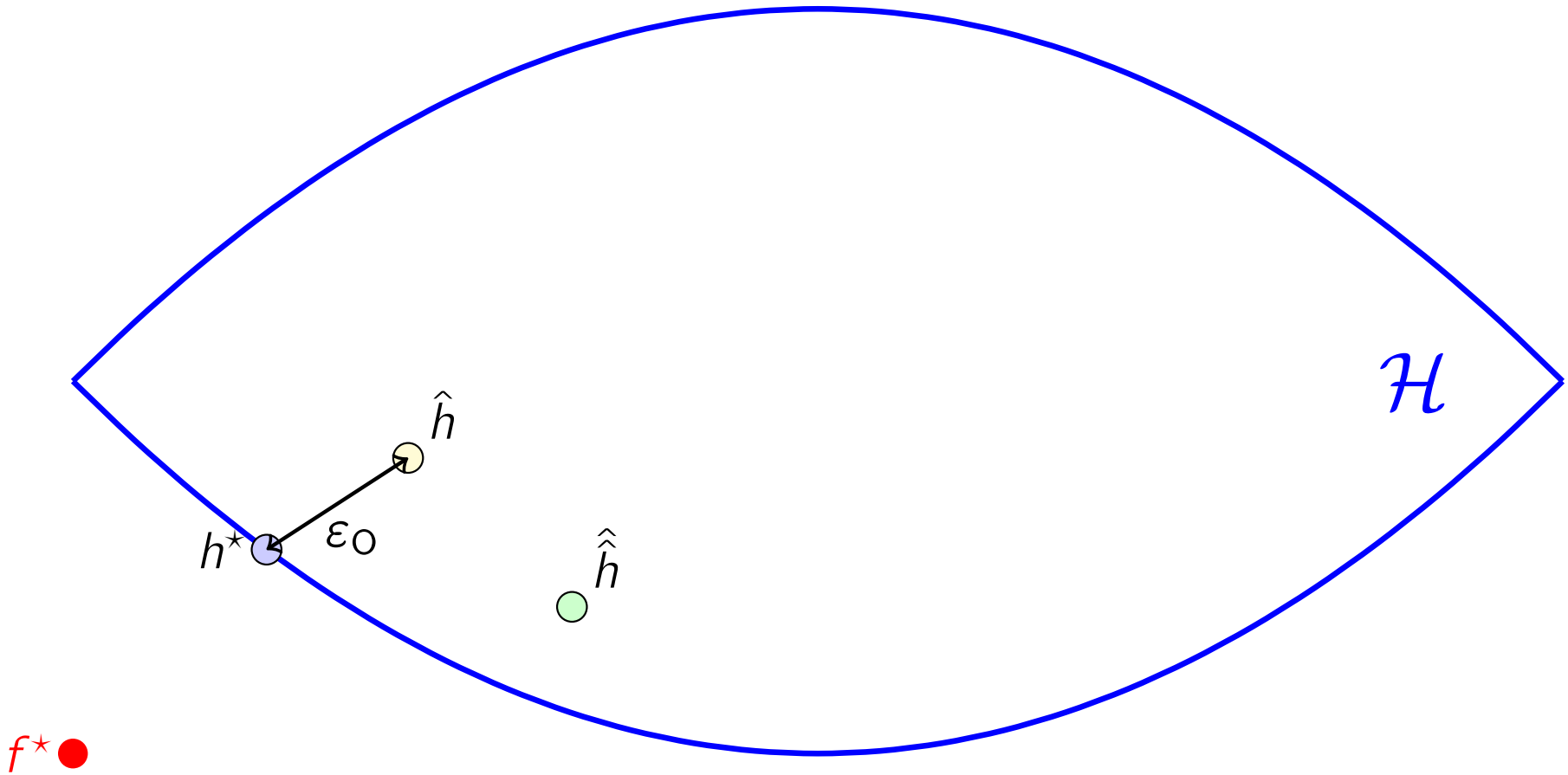
Generalization Revisited



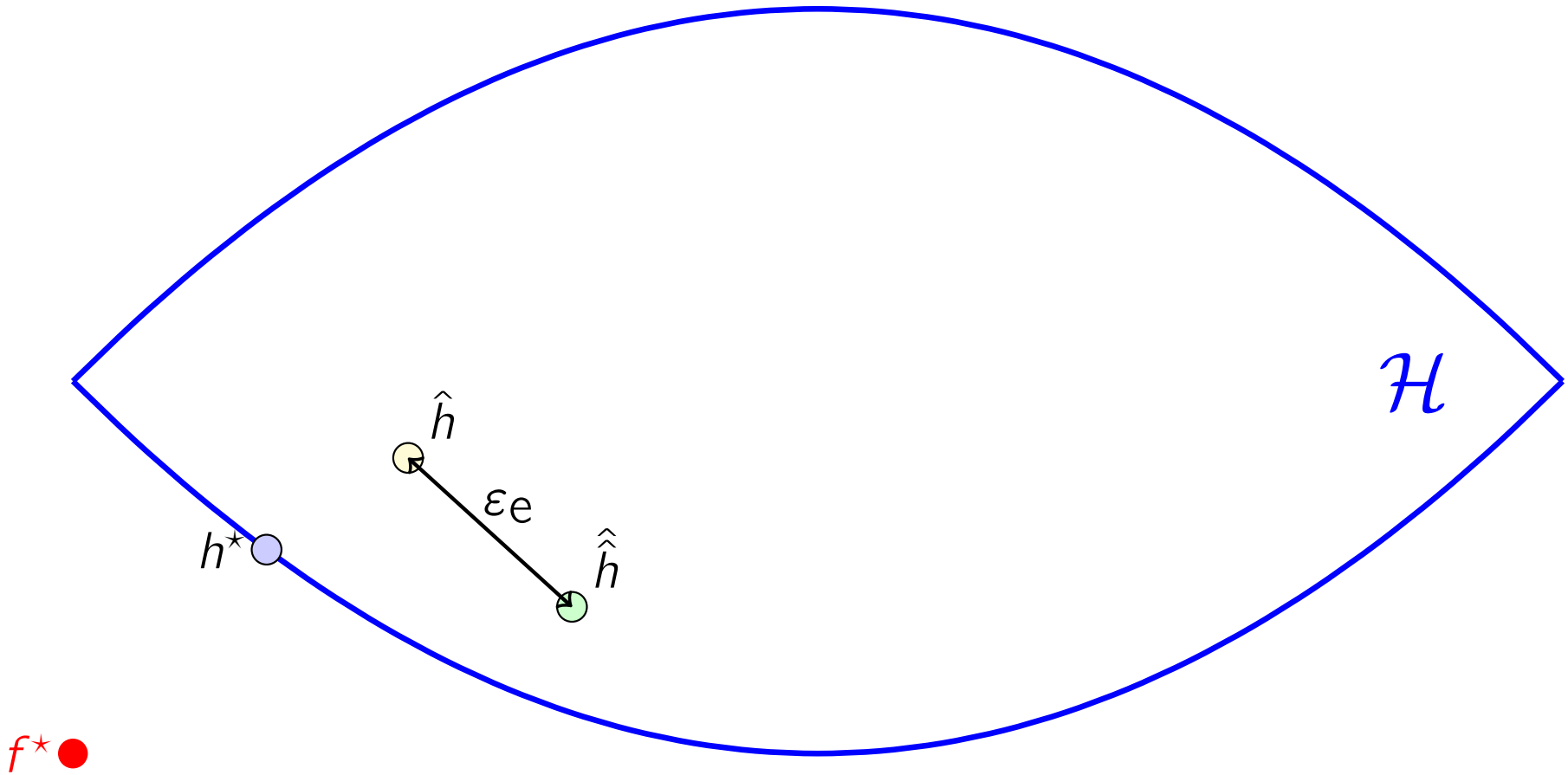
Generalization Revisited



Generalization Revisited



Generalization Revisited



Generalization Revisited

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star) \leq$$

$$\varepsilon_{\text{opt}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{\hat{h}}) - \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h})\right| \; +$$

$$\varepsilon_{\text{est}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h}) - \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star)\right| \; +$$

$$\varepsilon_{\text{app}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)\right|$$

Generalization Revisited

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star) \leq$$

$$\varepsilon_{\text{opt}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{\hat{h}}) - \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h})\right| \; +$$

$$\varepsilon_{\text{est}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h}) - \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star)\right| \; +$$

$$\varepsilon_{\text{app}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)\right|$$

- PAC Learning focuses on ε_{est} & complexity of $\mathcal{H}$

Generalization Revisited

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\mathcal{A}(S)) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star) \leq$$

$$\varepsilon_{\text{opt}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{\hat{h}}) - \mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h})\right| \; +$$

$$\varepsilon_{\text{est}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{h}) - \mathcal{L}_{\mathcal{D}}^{0-1}(h^\star)\right| \; +$$

$$\varepsilon_{\text{app}} \quad \Rightarrow \quad \left|\mathcal{L}_{\mathcal{D}}^{0-1}(h^\star) - \mathcal{L}_{\mathcal{D}}^{0-1}(f^\star)\right|$$

- PAC Learning focuses on ε_{est} & complexity of $\mathcal{H}$
- Stochastic optimization focuses on ε_{opt} & convexity of $\mathcal{H}$