Constrained vs. Penalized ERM

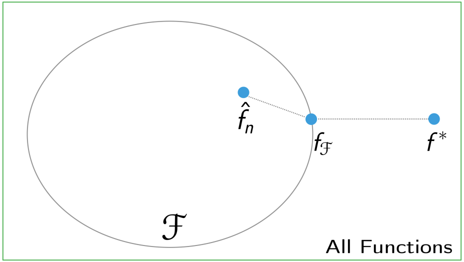
David S. Rosenberg

Bloomberg ML EDU

October 5, 2017

Regularization Paths in Function Space

Recall: Risk Decomposition Figure

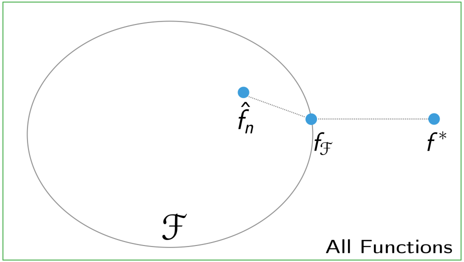


$$f^* = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Recall: Risk Decomposition Figure



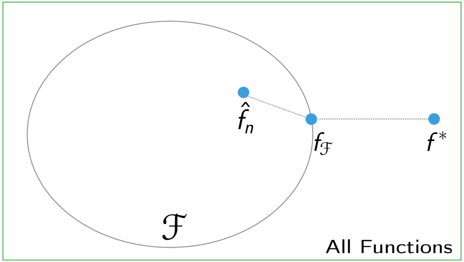
$$f^* = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= \; R(f_{\mathcal{F}}) - R(f^*)$

Recall: Risk Decomposition Figure



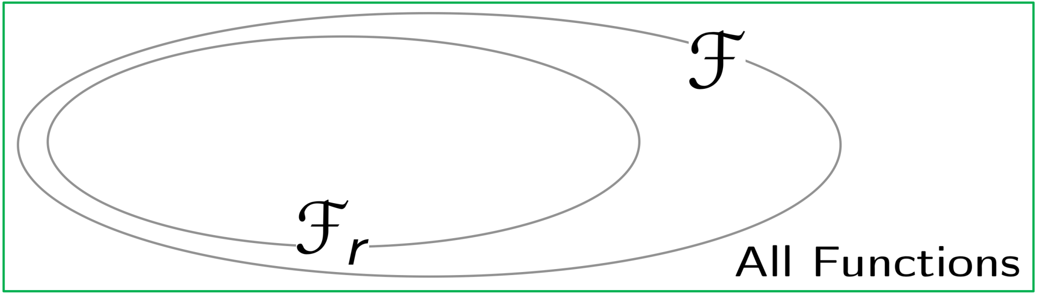
$$f^* = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} \mathbb{E}\ell(f(X), Y))$$

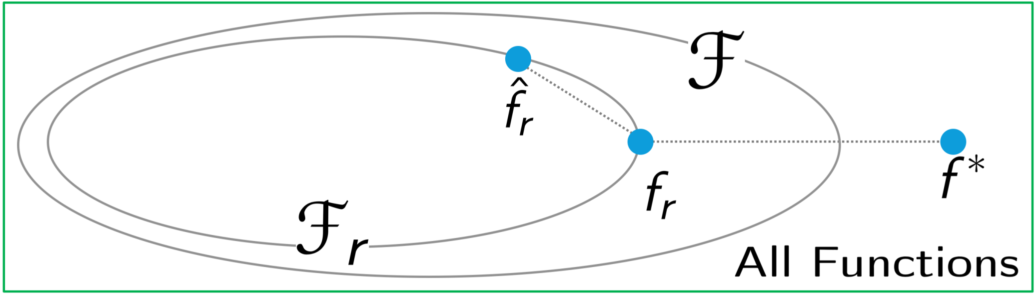
$$\hat{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= \ R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of $\hat{f}_n$ in $\mathcal{F}$) $= \ R(\hat{f}_n) - R(f_{\mathcal{F}})$

- Introduce **complexity-constrained hypothesis space**: $\mathcal{F}_r = \{f \in \mathcal{F} \mid \Omega(f) \leqslant r\}$

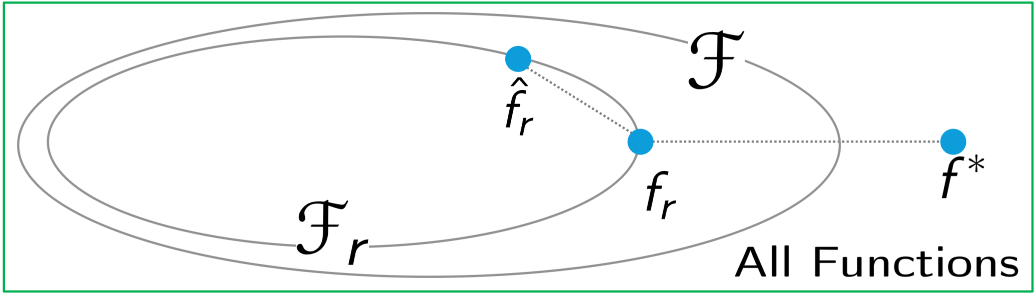


Risk Decomposition Figure: Complexity Constrained



- Revised notation:

$$\hat{f}_r = \underset{f \in \mathcal{F}_r}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \qquad f_r = \underset{f \in \mathcal{F}_r}{\arg\min} \mathbb{E}\ell(f(X), Y)) \qquad f^* = \underset{f}{\arg\min} \mathbb{E}\ell(f(X), Y)$$

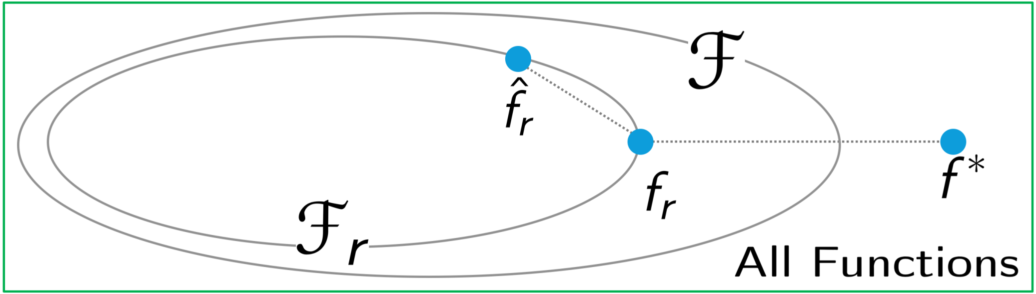


- Revised notation:

$$\hat{f}_r = \underset{f \in \mathcal{F}_r}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \qquad f_r = \underset{f \in \mathcal{F}_r}{\arg\min} \mathbb{E}\ell(f(X), Y)) \qquad f^* = \underset{f}{\arg\min} \mathbb{E}\ell(f(X), Y)$$

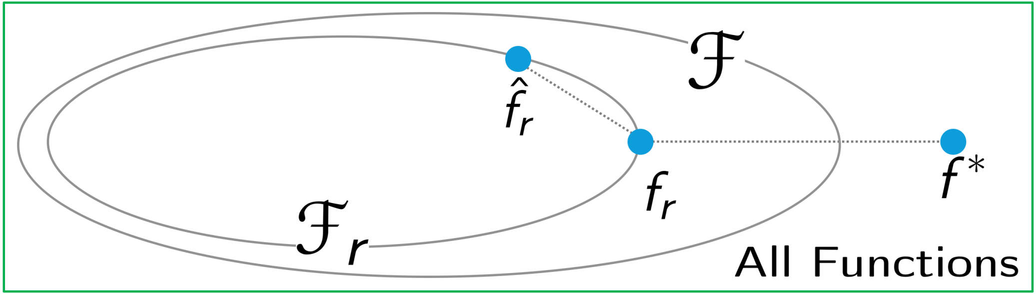
- This time we've put $\hat{f}_r$ on the boundary of $\mathcal{F}$ - why?

Risk Decomposition Figure: Complexity Constrained



- This time we've put $\hat{f}_r$ on the boundary of $\mathcal{F}$ - why?

Risk Decomposition Figure: Complexity Constrained



- This time we've put $\hat{f}_r$ on the boundary of $\mathcal{F}$ - why?
- **Typically**, $\hat{f}_r$ will have $\Omega(\hat{f}_r) = r$, since with more complexity we can usually fit the data better.

Risk Decomposition Figure: Complexity Constrained

- Consider complexity constraints $r = .001, .01, 1.0, 10, 1000$, corresponding to nested spaces:

$$\mathcal{F}_{0.001} \subset \mathcal{F}_{0.1} \subset \mathcal{F}_{1.0} \subset \mathcal{F}_{10} \subset \mathcal{F}_{1000}$$

Risk Decomposition Figure: Complexity Constrained

- Consider complexity constraints $r = .001, .01, 1.0, 10, 1000$, corresponding to nested spaces:

$$\mathcal{F}_{0.001} \subset \mathcal{F}_{0.1} \subset \mathcal{F}_{1.0} \subset \mathcal{F}_{10} \subset \mathcal{F}_{1000}$$

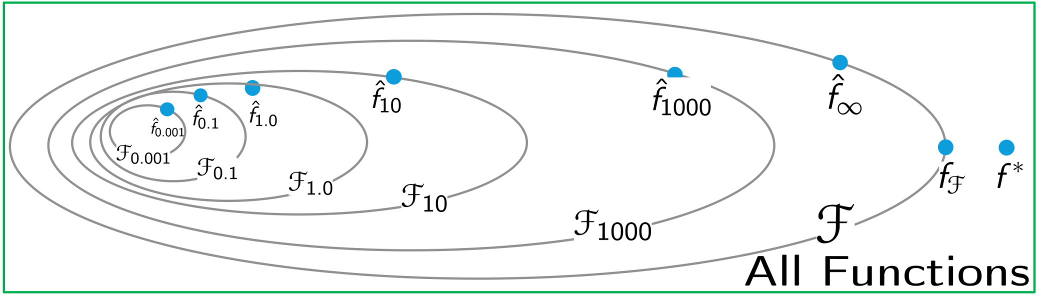
- We get corresponding sequence of ERMs: $\hat{f}_{0.001}, \hat{f}_{0.1}, \hat{f}_{1.0}, \hat{f}_{10}, \hat{f}_{1000}$

Risk Decomposition Figure: Complexity Constrained

- Consider complexity constraints $r = .001, .01, 1.0, 10, 1000$, corresponding to nested spaces:

$$\mathcal{F}_{0.001} \subset \mathcal{F}_{0.1} \subset \mathcal{F}_{1.0} \subset \mathcal{F}_{10} \subset \mathcal{F}_{1000}$$

- We get corresponding sequence of ERMs: $\hat{f}_{0.001}, \hat{f}_{0.1}, \hat{f}_{1.0}, \hat{f}_{10}, \hat{f}_{1000}$



- What if we found ERM's $\hat{f}_r$ for all $r \in [0, \infty]$?

Regularization Path for Constrained ERM

- What if we found ERM's $\hat{f}_r$ for all $r \in [0, \infty]$?
- Define the **regularization path** for constrained optimization in $\mathcal{F}$ with complexity Ω as

$$P_{\mathcal{F},\Omega}^{\text{constrained}} = \left\{ \hat{f}_r \mid r \in [0, \infty] \right\},$$

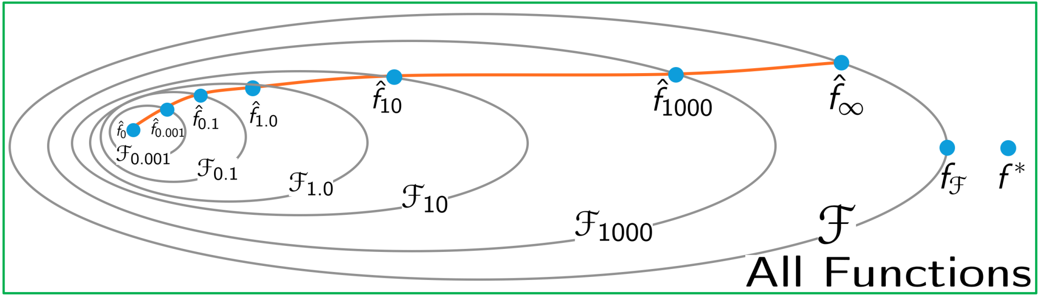
where $\hat{f}_r$ is the constrained ERM in $\mathcal{F}$ defined by $\hat{f}_r = \arg\min_{\{f \in \mathcal{F} \mid \Omega(f) \leqslant r\}} \hat{R}(f)$.

Regularization Path for Constrained ERM

- What if we found ERM's $\hat{f}_r$ for all $r \in [0, \infty]$?
- Define the **regularization path** for constrained optimization in $\mathcal{F}$ with complexity Ω as

$$P^{\text{constrained}}_{\mathcal{F},\Omega} = \left\{ \hat{f}_r \mid r \in [0, \infty] \right\},$$

where $\hat{f}_r$ is the constrained ERM in $\mathcal{F}$ defined by $\hat{f}_r = \arg\min_{\{f \in \mathcal{F} \mid \Omega(f) \leqslant r\}} \hat{R}(f)$.



Regularization Path for Penalized ERM

- Define the **regularization path** for penalized optimization in $\mathcal{F}$ with complexity Ω as

$$P_{\mathcal{F},\Omega}^{\text{penalized}} = \left\{ \hat{f}_\lambda \mid \lambda \in [0, \infty] \right\},$$

where $\hat{f}_r$ is the constrained ERM in $\mathcal{F}$ defined by $\hat{f}_r = \arg\min_{\{f \in \mathcal{F} | \Omega(f) \leqslant r\}} \hat{R}(f)$.

Regularization Path for Penalized ERM

- Define the **regularization path** for penalized optimization in $\mathcal{F}$ with complexity Ω as

$$P_{\mathcal{F},\Omega}^{\text{penalized}} = \left\{ \hat{f}_\lambda \mid \lambda \in [0, \infty] \right\},$$

  where $\hat{f}_r$ is the constrained ERM in $\mathcal{F}$ defined by $\hat{f}_r = \arg\min_{\{f \in \mathcal{F} \mid \Omega(f) \leqslant r\}} \hat{R}(f)$.
- For lasso, ridge, and many more, $P_{\mathcal{F},\Omega}^{\text{constrained}} = P_{\mathcal{F},\Omega}^{\text{penalized}}$.
  - Precise statement in homework.