Supervised Learning for Risk Functionals

- **Goal**: find a model in the hypothesis class that **performs well for a general risk functional**.
- Learner is given iid data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, hypothesis class $\mathcal{F}$, and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ for evaluating predictions

$$f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)]$$

Expectation
(Average Performance)

$$f^\star \in \arg\min_{f \in \mathcal{F}} \rho[\ell(f(X), Y)]$$

General Risk Functional
(Worst Performance, Variability, …)

Distributional Risk Estimation + Optimization

Let F_f be the CDF of losses $\ell(f(X), Y)$, and $\widehat{F}_f$ be the empirical CDF.

In order to estimate $f^\star$, solve:

$$\hat{f}^\star \in \arg\min_{f \in \mathcal{F}} \rho[\widehat{F}_f]$$

Risk Estimation / Optimization Method

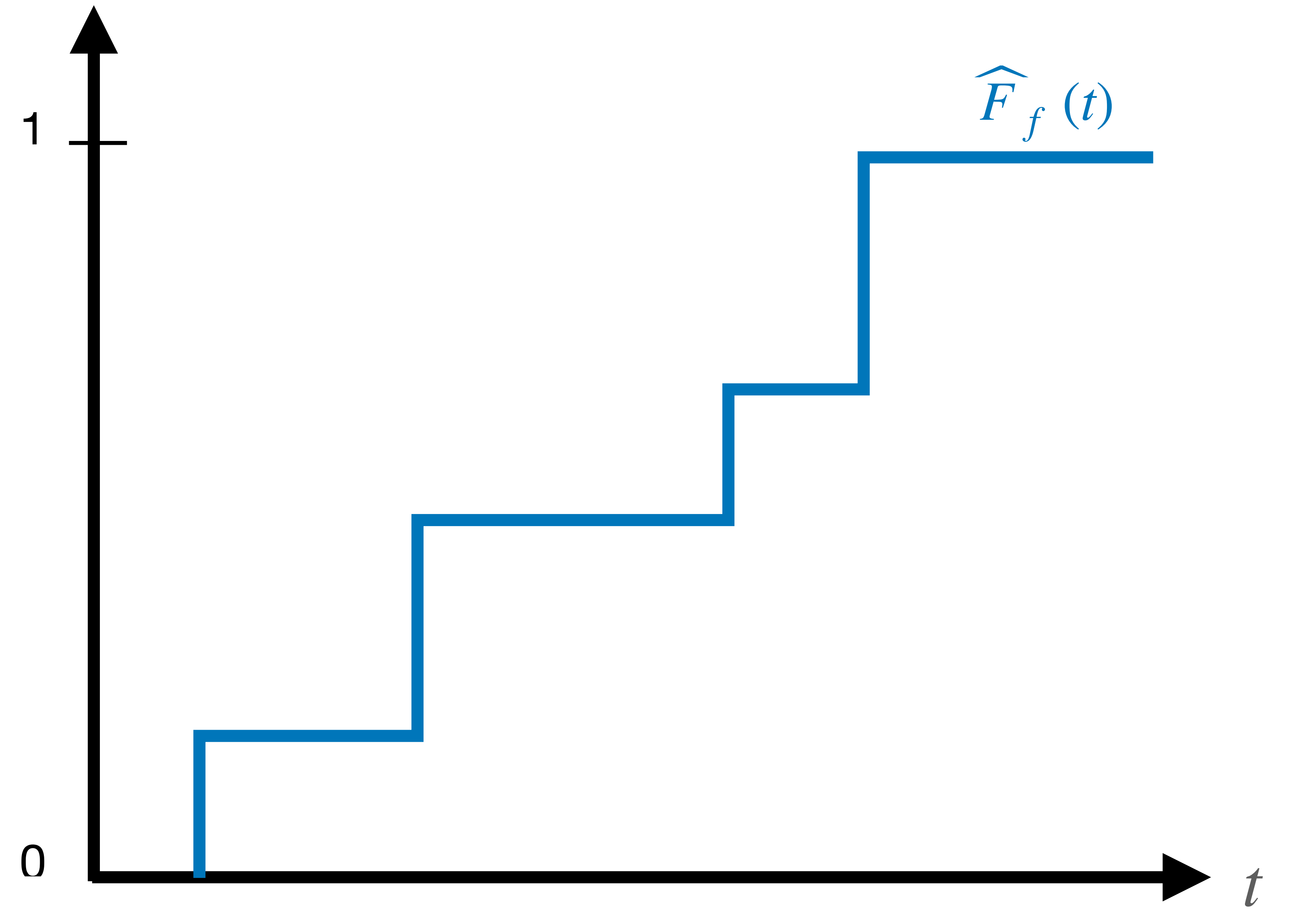
1. For hypothesis f, estimate the empirical loss CDF:

$$\widehat{F}_f(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\ell(f(\mathbf{X}_i), Y_i) \leq t\}$$

2. Evaluate risk functional ρ on the empirical CDF

$$\hat{\rho}_f = \rho(\widehat{F}_f)$$

3. Use $\hat{\rho}_f$ in gradient-based minimization method (or just to evaluate risk of f).



Risk Estimation / Optimization Method

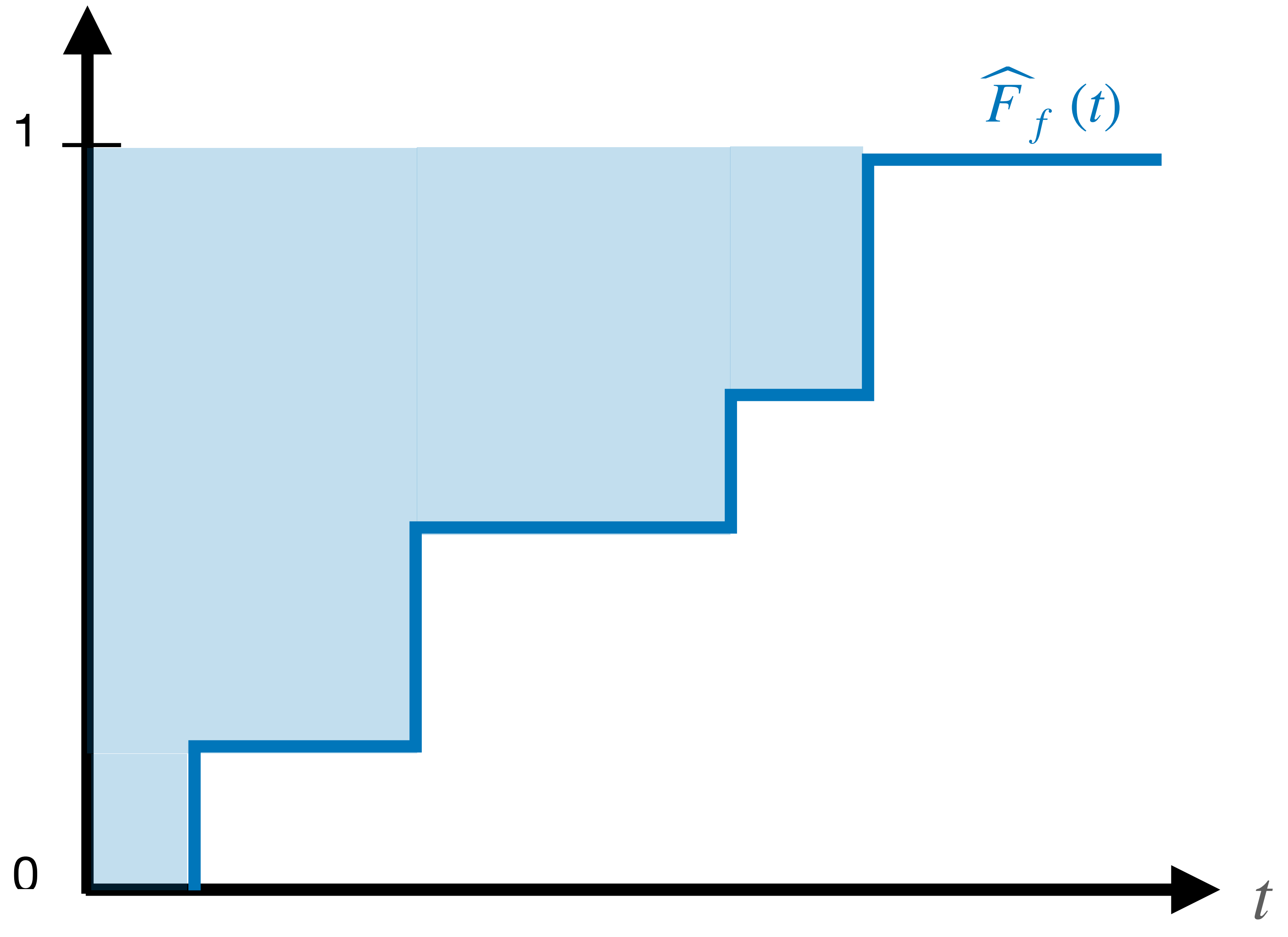
1. For hypothesis f, estimate the empirical loss CDF:

$$\widehat{F}_f(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\ell(f(\mathbf{X}_i), Y_i) \leq t\}$$

2. Evaluate risk functional ρ on the empirical CDF

$$\hat{\rho}_f = \rho(\widehat{F}_f)$$

3. Use $\hat{\rho}_f$ in gradient-based minimization method (or just to evaluate risk of f).



Expected value

Risk Estimation / Optimization Method

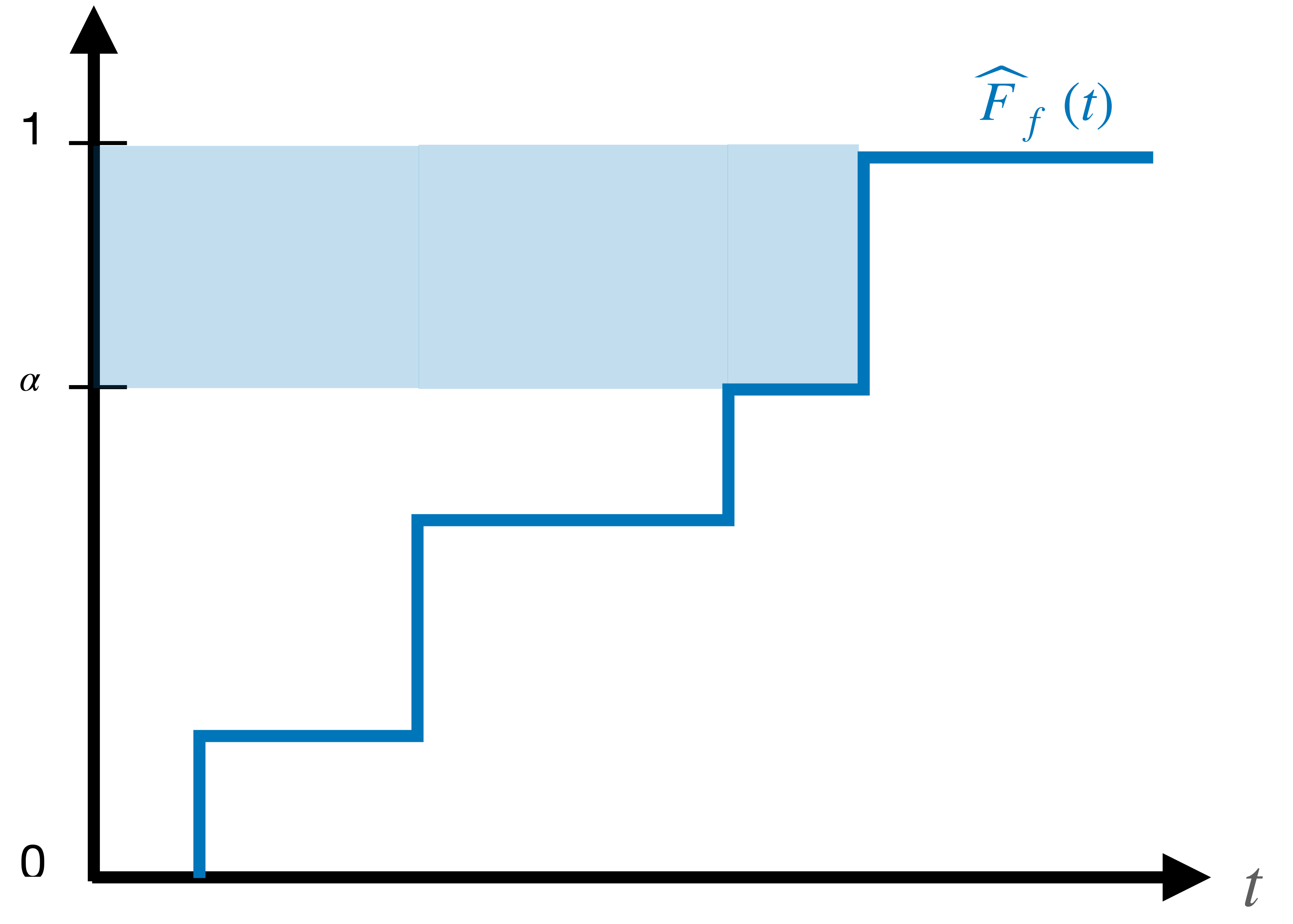
1. For hypothesis f, estimate the empirical loss CDF:

$$\widehat{F}_f(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\ell(f(\mathbf{X}_i), Y_i) \le t\}$$

2. Evaluate risk functional ρ on the empirical CDF

$$\hat{\rho}_f = \rho(\widehat{F}_f)$$

3. Use $\hat{\rho}_f$ in gradient-based minimization method (or just to evaluate risk of f).



Conditional Value-at-Risk (CVaRα)
"worst-case average"

Motivation for Distributional Approach

Excess risk is bounded by CDF estimation error:

$$\rho(F_{\hat{f}^\star}) - \rho(F_{f^\star}) \leq 2 \sup_{f \in \mathbb{F}} | \rho(\widehat{F}_f) - \rho(F_f) |$$

standard ERM inequality

$$\leq 2L \sup_{f \in \mathbb{F}} \| \widehat{F}_f - F_f \|_\infty$$

smoothness of ρ

Accurate empirical CDF estimation is sufficient for generalizable ERM.

Results

1. **Uniform convergence** of CDF (and risk) estimation
    - CDF estimation: $\sup_{f \in \mathcal{F}} \|F_f - \widehat{F}_f\|_\infty$
    - Risk estimation: $\sup_{f \in \mathcal{F}} |\rho_f - \widehat{\rho}_f|$ and $\sup_{\rho} \sup_{f \in \mathcal{F}} |\rho_f - \widehat{\rho}_f|$
2. **Convergence** of a gradient-based method for minimizing $\widehat{\rho}_f$

Supervised Learning with **General Risk Functionals**

Liu Leqi, Audrey Huang, Zachary Lipton, Kamyar Azizzadenesheli