Supervised Learning with General Risk Functionals

Liu Leqi[1], Audrey Huang[2], Zachary C. Lipton[1], Kamyar Azizzadenesheli[3]

[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign [3]Purdue University correspondence: leqil@cs.cmu.edu

Contribution

- **Motivation** Most supervised learning literatures center on learning risk-neutral models with low **expected** losses. However, real-world concerns often demand that we address training models under other **functionals of the loss distribution**:
  1. The desired model at test time is risk-sensitive for reasons including risk aversion, equitable allocations of benefits and harms, or alignment with human preferences.
  2. The desired model at test time is risk-neutral (i.e., with low expected losses) but the training objective is chosen as other functionals for reasons including distribution shifts, noisy labels, or imbalanced dataset.
- **Contribution** A general learning procedure along with guarantees for risk-sensitive supervised learning under general risk functionals:
  - **Uniform convergence for risk estimation** that holds simultaneously for all Hölder risks, yielding learning guarantees for empirical risk minimization for the broad class of distortion risks.
  - A **gradient-based method** for minimizing distortion risks that re-weights examples dynamically based on the empirical CDF of losses, and corresponding convergence guarantees.

Risk-sensitive Learning

- The learner is given iid data $Z_i = (\mathbf{X}_i, Y_i)$, a hypothesis class $\mathcal{F}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ for evaluating a prediction.
- The **goal** is to find a model in the hypothesis class that performs **well**. Traditionally, the models are evaluated in terms of their average performance which corresponds to the expectation functional. However, one may care about different functionals ρ of the loss distribution, e.g., the worst-case performance, the variability of the performance, etc.

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

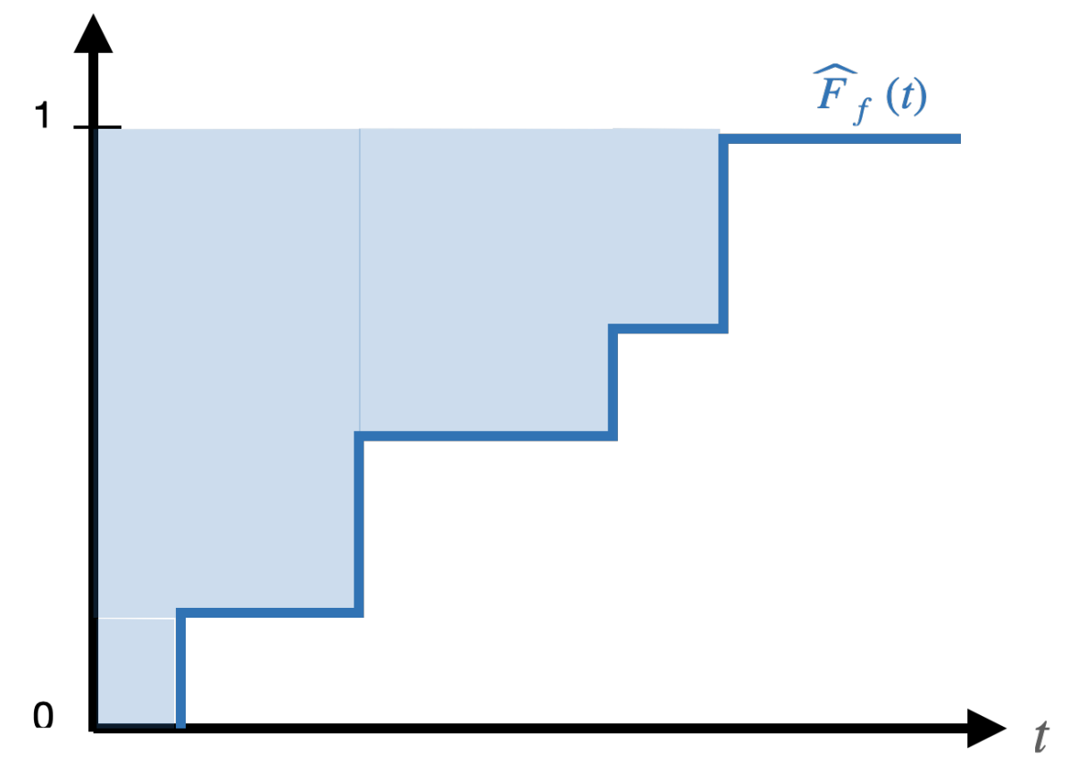
$$\downarrow$$

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

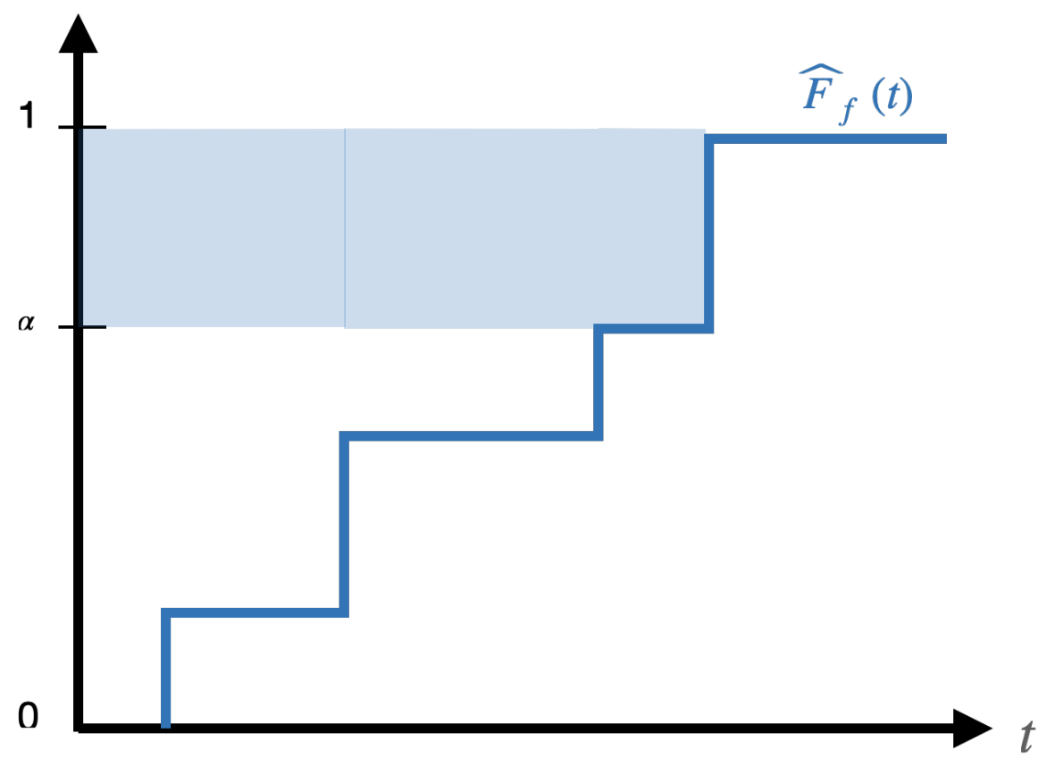
- We propose to consider the set of **Hölder risk functionals**. We provide a simplified definition here: A risk functional ρ is L-Hölder on a space of real-valued random variables $\mathcal{U}$ if there exist constants $p > 0$ and $L > 0$ such that for all $U, U' \in \mathcal{U}$ with CDF F_U and $F_{U'}$ respectively, the following holds:

$$|\rho(F_U) - \rho(F_{U'})| \leq L \|F_U, F_{U'}\|_\infty^p.$$

Examples: Expectation, mean-variance, conditional value at risks, entropic risks, and cumulative prospect theory risks are Hölder on the set of bounded random variables.



(a) Expected Value



(b) Conditional Value at Risk

General Learning Procedure

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

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

2. Evaluate the risk on the empirical CDF $\rho[\widehat{F}_f]$.
3. Minimizing the empirical risk: $\widehat{f}^\star \in \arg\min_{f\in\mathcal{F}} \ \rho(\widehat{F}_f)$.

Uniform Convergence of Risk Estimation

- Why do we study it? The **excess risk** of $\widehat{f}^\star$ is bounded by the following:

$$\begin{aligned}\rho(F_{\widehat{f}^\star}) - \rho(F_{f^\star}) &\leq 2\sup_{f\in\mathbb{F}} |\rho(\widehat{F}_f) - \rho(F_f)| && \text{(standard inequality)}\\ &\leq 2L \sup_{f\in\mathbb{F}} \|\widehat{F}_f - F_f\|_\infty && \text{(smoothness of } \rho)\end{aligned}$$

- The uniform convergence of risk estimation is reduced to the uniform convergence of CDF estimation.

Uniform Convergence of CDF Estimation

Theorem 1 *Given a hypothesis class $\mathcal{F}$, any loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, and n iid samples $\{Z_i\}_{i=1}^n$, we have that with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \sup_{r\in\mathbb{R}} |\widehat{F}(t; f) - F(t; f)| \leq 2\mathcal{R}(n, \mathcal{F}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}},$$

where $\mathcal{R}(n, \mathcal{F}) = \mathbb{E}_{\mathbb{P},\xi}\left[\sup_{f\in\mathcal{F}} \sup_{t\in\mathbb{R}} \frac{1}{n} \left|\Sigma_{i=1}^n \xi_i \mathbb{I}\{\ell(f(X_i), Y_i)) \leq t\}\right|\right]$ and ξ denotes a Rademacher random variable.

Corollary 1 *For a finite hypothesis class $\mathcal{F}$,*

$$\mathcal{R}(n, \mathcal{F}) \leq \sqrt{\frac{\log(4|\mathcal{F}|)}{2n}}.$$

For an infinite hypothesis class, we consider two approaches:

1. We propose a new notion of **permutation complexity** that captures the structure of CDF estimation (e.g., empirical CDFs are monotonic and bounded) and disentangle the complexity of $\mathcal{F}$ from the complexity of the indicator functions.
2. We use the traditional approach that directly work with a function class consisting of compositions of $f \in \mathcal{F}$ and the indicator functions.

Empirical Risk Minimization

We use a gradient-based method to minimize distortion risks (with Hölder distortion functions) a subset of Hölder risk functionals including the expectation and conditional value at risk:

$$\rho(F_f) = \textstyle\int_0^\infty g(1 - F_f(t))dt,$$

where the distortion function $g : [0, 1] \to [0, 1]$ is non-decreasing with $g(0) = 0$ and $g(1) = 1$.

- At iteration t,

$$\theta_{t+1} \leftarrow \theta_t - \eta \left(\nabla_\theta \rho(\widehat{F}_\theta) + w_t\right),$$

  where θ parameterizes f, η is the learning rate, $\nabla_\theta \rho(\widehat{F}_\theta)$ is the gradient and w_t is sampled from a d-dimensional Gaussian with mean 0 and variance $\frac{1}{d}$.
- If $\{\ell(f_\theta(x_i), y_i)\}_{i=1}^n$ are Lipschitz continuous and $\rho(\widehat{F}_\theta)$ is β-smooth in θ, when $\eta = \frac{1}{\beta\sqrt{T}}$, θ_t converges to a stationary point.

Experiment: Risk Assessment

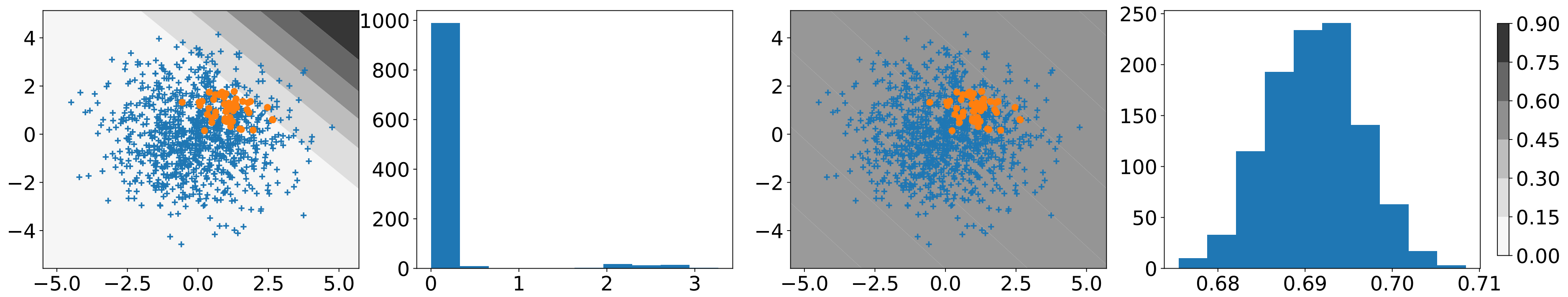
Setup. We perform risk assessments on pretrained Pytorch models for ImageNet classification. These models share similar accuracy (around 69%) on the validation set (Table 1).

	VGG-11	GoogLeNet	ShuffleNet	Inception	ResNet-18
Accuracy	69.022%	69.772%	69.356%	69.542%	69.756%
$\mathbb{E}[\ell_f]$	1.261	1.283	1.360	1.829	1.247
$\mathrm{CVaR}_{.05}(\ell_f)$	1.327	1.350	1.431	1.925	1.313
$\mathbb{E}[\ell_f] + 0.5\mathrm{Var}(\ell_f)$	5.215	4.376	6.718	14.416	5.353

Table: Risks for different ImageNet classification models evaluated on the validation set. $\ell_f(Z)$ is the cross-entropy loss for each model f. For simplicity, we omitted the arguments Z in the table. CVaR_α is the expected value of the top 100α percent losses. All results are rounded to 3 digits.

Experiment: Empirical Distortion Risk Minimization

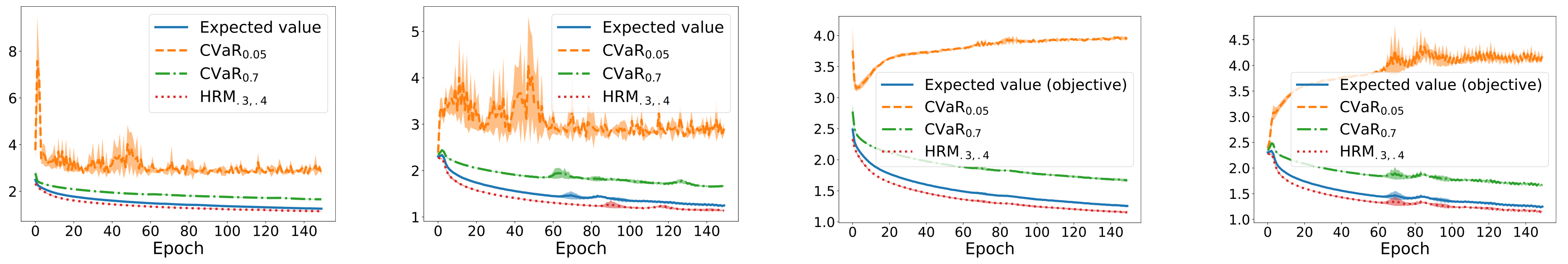
Toy Example The blue pluses and orange dots represent two classes, respectively. We have learned logistic regression models to minimize the expected loss and the $\mathrm{CVaR}_{.05}$ (expected value of the top 5% losses) through minimizing their empirical risks. The model learned under expected loss suffers high loss for a small subset of the covariates while the model learned under $\mathrm{CVaR}_{.05}$ have all losses concentrated around a small value. Indicated by the (uniform) grey color in the contour plot, the predictions (predicted probability of a covariate being labeled as 1) for the $\mathrm{CVaR}_{.05}$ model are around 0.5. In contrast, the predictions for the expected loss model spread across a wide range between 0 and 1.



(a) Prediction contour plot under expected value (b) Loss histogram under expected value (c) Prediction contour plot under $\mathrm{CVaR}_{.05}$ (d) Loss histogram under $\mathrm{CVaR}_{.05}$

Figure: Prediction (predicted probability of a covariate being labeled as 1) contours and loss histograms of two models learned under the expected loss and the $\mathrm{CVaR}_{.05}$ objective, respectively. The blue pluses and orange dots represent two classes. The loss distribution for the expected loss model has extremely high values for a small subset of the covariates.

CIFAR-10 We have trained VGG-16 models on CIFAR-10 through minimizing the empirical risks for expected loss, $\mathrm{CVaR}_{.05}$, $\mathrm{CVaR}_{.7}$ and $\mathrm{HRM}_{.3,.4}$. In general, the objective values are decreasing over the epochs during training and testing. In addition, we observe that minimizing the empirical risk for expected loss does not necessarily imply minimizing other risks, e.g., $\mathrm{CVaR}_{.05}$ (Figure 3(a)), suggesting the efficacy of our proposed optimization procedure for minimizing distortion risks.



(a) Training objective values (b) Testing objective values (c) Training risk evaluations (d) Testing risk evaluations

Figure: Performance of VGG-16 models trained under expected loss, $\mathrm{CVaR}_{.05}$, $\mathrm{CVaR}_{.7}$ and $\mathrm{HRM}_{.3,.4}$. In Figures 3(a) and 3(b), each model is trained and evaluated on the same objective. In Figures 3(c) and 3(d), we only train one model under the expected loss but report all four objectives of that model.

References

- Khim, J., Leqi, L., Prasad, A., & Ravikumar, P. (2020). Uniform convergence of rank-weighted learning. ICML.
- Lee, J., Park, S., & Shin, J. (2020). Learning bounds for risk-sensitive learning. NeurIPS.
- Huang, A., Leqi, L., Lipton, Z., & Azizzadenesheli, K. (2021). Off-policy risk assessment in contextual bandits. NeurIPS.
- Bhat, Sanjay P., & Prashanth LA. (2019). Concentration of risk measures: A Wasserstein distance approach. NeurIPS.