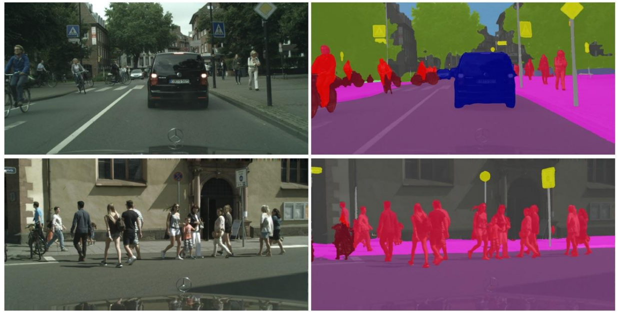
Learning Guarantees for Convex But Inconsistent Surrogates

Kirill Struminsky, Simon Lacoste-Julien, Anton Osokin

Structured Prediction

- Structured prediction = given x predict a structured output y
- Examples:
    - Image segmentation
    - Ranking
    - Handwriting recognition
- Key aspects:
    - Exponential number of labels y
    - Not all mistakes are equal - cost sensitive prediction
- We explore inconsistent structured prediction algorithms
    - Generalize the results from (Osokin17)
    - Theoretical guarantees for learning (optimization steps needed to minimize population risk)
    - More efficient than consistent algorithms

x y



Osokin A., Bach F., Lacoste-Julien S. On structured prediction theory with calibrated convex surrogate losses //Advances in Neural Information Processing Systems. – 2017. – C. 302-313.

Structured Prediction Setup

- Datapoints $(x, z) \in X \times Z$, output $y \in Y,\ (|Y|, |Z| < \infty)$
- Prediction with a score function $f(x) \in \mathbb{R}^k,\ k = |Y|$

$$y = \operatorname{pred} f(x) = \operatorname{argmax}_{\hat{y} \in Y} f_{\hat{y}}(x)$$

- Loss matrix $L \in \mathbb{R}^{|Y| \times |Z|}$ (e.g. Hamming distance between y and z)
- Choose f to minimize **population risk**

$$\mathcal{R}_L(f) := \mathbb{E}_{(x,z)\sim D} L(\operatorname{pred} f(x), z) \to \min_{f \in \mathcal{F}}$$

Learning With Surrogates

- Population risk can be hard to optimize

$$\mathcal{R}_L(f) := \mathbb{E}_{(x,z)\sim D} L(\text{pred}\, f(x), z) \to \min_{f \in \mathcal{F}}$$

- Workaround: use surrogate loss function

$$\mathcal{R}_\Phi(f) := \mathbb{E}_{(x,z)\sim D} \Phi(f(x), z) \to \min_{f \in \mathcal{F}}$$

- Examples:
  - Classification $\Phi_{\log}(f(x), y) := -f_y(x) + \log \sum_{\hat{y} \in Y} \exp f_{\hat{y}}(x)$
  - SSVM $\Phi_{SSVM}(f(x), z) := \max_{\hat{y} \in Y} (f_{\hat{y}}(x) + L(\hat{y}, z)) - f_{\hat{y}}(x)$

Consistency

- Connects actual loss and surrogate loss
- Surrogate is **consistent** if it has the same optimum as population risk
  - Cross-entropy defines a consistent surrogate for classification
- **Inconsistent** surrogates are still useful in practice
  - SSVM is inconsistent in some settings
- **Calibration function** is a tool for surrogate consistency analysis

Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization //Annals of Statistics. – 2004. – C. 56-85.
Zhang T. Statistical analysis of some multi-category large margin classification methods //Journal of Machine Learning Research. – 2004. – T. 5. – №. Oct. – C. 1225-1251.

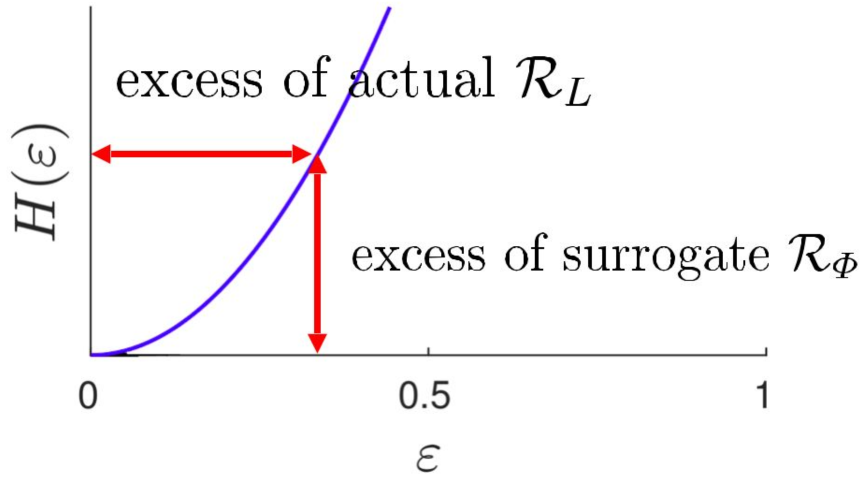
Calibration Function Intuition

- Calibration function connects **surrogate** optimization and **population risk** optimization

$$\mathcal{R}_\Phi(f) - \mathcal{R}^*_\Phi < H_{\Phi,L,\mathcal{F}}(\varepsilon) \Rightarrow \mathcal{R}_L(f) - \mathcal{R}^*_L < \varepsilon$$

- $H_{\Phi,L,F}(\varepsilon)$ Depends on
  - Surrogate Φ
  - Loss L
  - Score functions F
- Bigger values are better
- Consistent iff

$$H_{\Phi,L,F}(\varepsilon) > 0 \quad \text{for } \varepsilon > 0$$



Calibration Function Definition 1/3: Conditional Loss

- Conditional actual and surrogate risk

$$l(f, q) := \sum_{c=1}^{k} q_c L(\mathrm{pred}(f), c), \quad \phi(f, q) := \sum_{c=1}^{k} q_c \Phi(f, c)$$

For population loss we have

$$\mathcal{R}_L(\mathbf{f}) = \mathbb{E}_{x \sim D_X} l(f(x), P_{\mathcal{D}}(\cdot \mid x))$$
$$\mathcal{R}_\Phi(\mathbf{f}) = \mathbb{E}_{x \sim D_X} \phi(f(x), P_{\mathcal{D}}(\cdot \mid x))$$

Calibration Function Definition 2/3: Excess

- Conditional actual and surrogate risk

$$l(f,q) := \sum_{c=1}^{k} q_c L(\mathrm{pred}(f), c), \quad \phi(f,q) := \sum_{c=1}^{k} q_c \Phi(f,c)$$

- Excess

$$\delta l(f,q) := l(f,q) - \min_{\hat{f} \in \mathbb{R}^{|Y|}} l(\hat{f}, q)$$

$$\delta\phi(f,q) := \phi(f,q) - \min_{\hat{f} \in \mathcal{F}} \phi(\hat{f}, q)$$

- How far are $\mathcal{R}_L(\mathbf{f})$ and $\mathcal{R}_\Phi(\mathbf{f})$ from optimum?

$$\mathcal{R}_L(\mathbf{f}) - \mathcal{R}_L^* = \mathbb{E}_{x \sim D_X} \delta l(f(x), P_{\mathcal{D}}(\cdot \mid x))$$
$$\mathcal{R}_\Phi(\mathbf{f}) - \mathcal{R}_\Phi^* = \mathbb{E}_{x \sim D_X} \delta\phi(f(x), P_{\mathcal{D}}(\cdot \mid x))$$

Calibration Function Definition 3/3:

- Conditional actual and surrogate risk

$$l(f,q) := \sum_{c=1}^{k} q_c L(\mathrm{pred}(f), c), \quad \phi(f,q) := \sum_{c=1}^{k} q_c \Phi(f,c)$$

- Excess

$$\delta l(f,q) := l(f,q) - \min_{\hat{f} \in \mathbb{R}^{|Y|}} l(\hat{f}, q)$$

$$\delta\phi(f,q) := \phi(f,q) - \min_{\hat{f} \in \mathcal{F}} \phi(\hat{f}, q)$$

- The calibration function

$$\begin{aligned} H_{\Phi,L,\mathcal{F}}(\varepsilon) := \min_{f,q}\, & \delta\phi(f,q) \\ \text{s.t. } & \delta l(f,q) \geq \varepsilon, \\ & f \in \mathcal{F}, q \in \Delta_{|Y|} \end{aligned}$$

$$\boxed{H_{\Phi,L,\mathcal{F}}(\delta l(f,q)) \leq \delta\phi(f,q)}$$

Population Risk Guarantees

Assume $\mathcal{R}_\Phi(\mathbf{f}) - \mathcal{R}_\Phi^* \leq H_{\Phi,L,\mathcal{F}}(\varepsilon)$

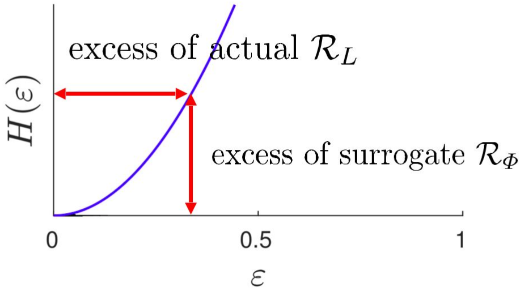
By the definition of calibration function we have $H_{\Phi,L,\mathcal{F}}(\delta l(f,q)) \leq \delta\phi(f,q)$

After averaging we obtain

$$\boxed{H_{\Phi,L,\mathcal{F}}\Big(\mathcal{R}_L(\mathbf{f}) - \mathcal{R}_L^*\Big) \leq} \mathbb{E}_{x\sim D_X} H_{\Phi,L,\mathcal{F}}\Big(\delta l(f(x), P(\cdot \mid x))\Big) \leq \mathcal{R}_\Phi(\mathbf{f}) - \mathcal{R}_\Phi^* \boxed{\leq H_{\Phi,L,\mathcal{F}}(\varepsilon)}$$

convexity

our assumptions



strict monotonicity

$$\mathcal{R}_L(\mathbf{f}) - \mathcal{R}_L^* \leq \varepsilon$$

Learning Guarantees Recipe

- Choose a framework to provide an upper bound for $\mathcal{R}_\Phi(\mathbf{f}) - \mathcal{R}^*_\Phi$

  “After **X** steps of SGD with high probability $\mathcal{R}_\Phi(\mathbf{f}) - \mathcal{R}^*_\Phi \leq$ **Y**“

- Use calibration function $H_{\Phi,L,\mathcal{F}}(\varepsilon)$ to replace the surrogate

  “After **X** steps of SGD with high probability $\mathcal{R}_L(\mathbf{f}) - \mathcal{R}^*_L \leq$ **Z**“

Nemirovski A. et al. Robust stochastic approximation approach to stochastic programming //SIAM Journal on optimization. – 2009. – T. 19. – №. 4. – C. 1574-1609.

Quadratic Surrogate

We consider

$$\Phi_{quad}(f(x), z) := \frac{1}{2|Y|} \sum_{\hat{y}=1}^{|Y|} (f_{\hat{y}}(x) + L(\hat{y}, z))^2$$

Intuition:

1. for (x, z) score function $f(x)$ predicts column of loss matrix $L(\cdot, z)$
2. predictor $\mathrm{pred}\, f(x) = \mathrm{argmax}_{\hat{y}}\, f_y(x)$ minimizes $L(y, z)$ over y

Pros: universal, consistent, easy to analyze

Cons: not common in practice

Ciliberto C., Rosasco L., Rudi A. A consistent regularization approach for structured prediction //Advances in neural information processing systems. – 2016. – C. 4412-4420.
Osokin A., Bach F., Lacoste-Julien S. On structured prediction theory with calibrated convex surrogate losses //Advances in Neural Information Processing Systems. – 2017. – C. 302-313.

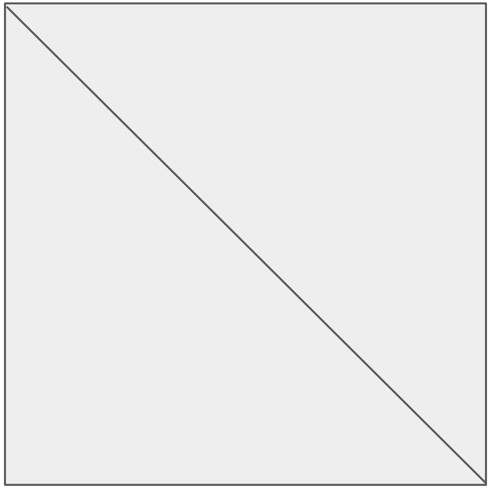
Calibration Function For 0-1 Loss

k- number of classes (exponential)

$$H_{\Phi_{quad}, L_{01}, \mathbb{R}^k}(\varepsilon) = \frac{\varepsilon^2}{4k}$$

The surrogate is consistent, but hard to learn

L



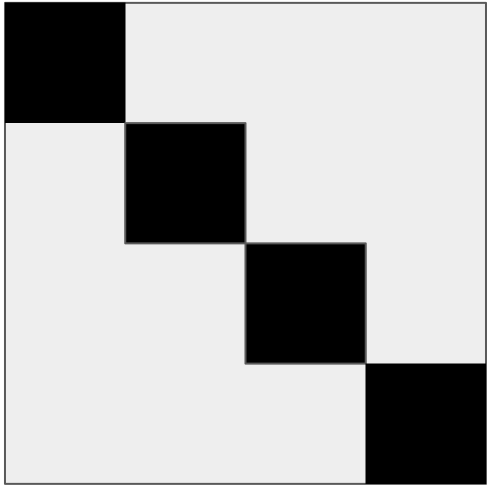
Calibration Function For Block 0-1 Loss

k - number of classes

s - number of blocks

$$H_{\Phi_{quad},L,\mathbb{R}^k}(\varepsilon) = \frac{\varepsilon^2}{4k}\frac{s+1}{s}$$

L



Almost no changes

Choice of Score Functions

Restrict score functions: $f(x) = F\theta(x)$

- $F \in \mathbb{R}^{|Y| \times r}$ is a fixed low-rank matrix
- $\theta : X \to \mathbb{R}^r$ is the function we learn

Φ_{quad} is consistent iff $\operatorname{span} L \subseteq \operatorname{span} F$

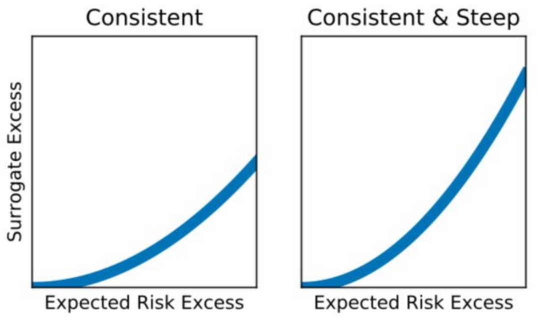
For block 0-1 loss the restrictions give $H_{\Phi_{quad},L,\mathcal{F}'}(\varepsilon) = \frac{\varepsilon^2}{4d}$

d- number of blocks

Choice of Score Functions

In general, if $\operatorname{span} L \subseteq \operatorname{span} F$ (i.e. the surrogate is consistent)

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) \geq \min_{y_1 \neq y_2} \frac{\varepsilon^2}{2k\ \pi_{y_1,y_2}}$$



For $\pi_{y_1,y_2} = \|\operatorname{Proj}_F(e_{y_1} - e_{y_2})\|_2^2$

If $F_1 \subset F_2$, then $H_{\Phi,L,\mathcal{F}_1}(\varepsilon) \geq H_{\Phi,L,\mathcal{F}_2}(\varepsilon)$ $\longrightarrow$ stronger guarantees for F_1

What happens if $\operatorname{span} F \subset \operatorname{span} L$?

Lower Bounds for the Inconsistent Case

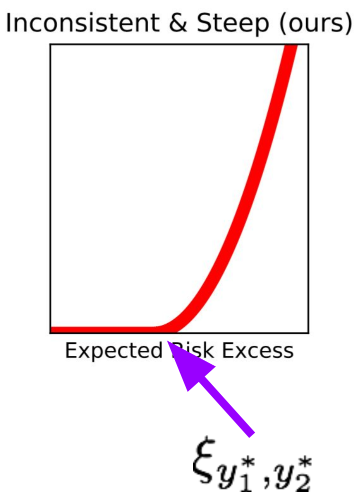
We have

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) \geq \min_{y_1 \neq y_2} \frac{(\varepsilon - \xi_{y_1,y_2})_+^2}{2k\ \pi_{y_1,y_2}}$$

(where $(x-a)_+^2 := (\min(0, x-a))^2$)

for $\quad \xi_{y_1,y_2} = \|L^T(I - \mathrm{Proj}_F)(e_{y_1} - e_{y_2})\|_\infty$

and $\pi_{y_1,y_2} = \|\,\mathrm{Proj}_F(e_{y_1} - e_{y_2})\|_2^2$

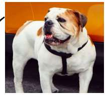
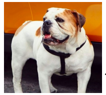
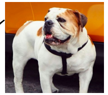


No consistency assumption $\mathrm{span}\, L \subseteq \mathrm{span}\, F$

Small $\boldsymbol{F} \rightarrow$ easier learning, but less guarantees (no consistency)

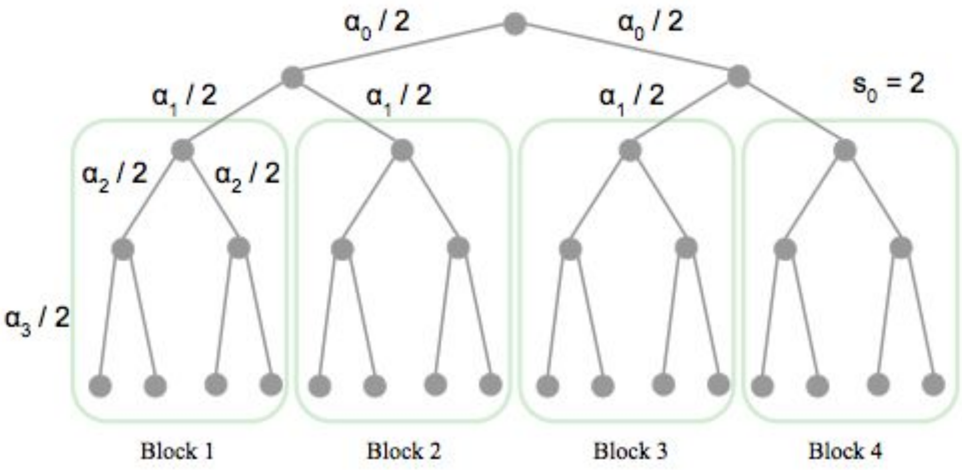
Tree-structured loss

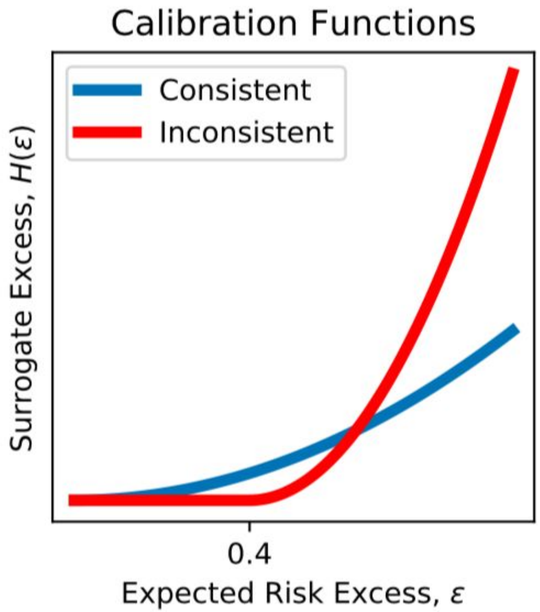
Classes form a hierarchy reflected in the loss matrix L

- $L($, $) = 0$
- $L($, $) = 0.5$
- $L($, $) = 1$

The example is a mixture of block-01 and 01 losses (full rank loss)

Tree-structured Loss





Labels = Tree leaves

L(i, j) - distance between leaves

Ranking with mAP

- Training labels - binary vectors (document relevance) $z \in \{0,1\}^r$
- Prediction - permutation of documents $\sigma \in S_n$
- Loss:

$$L(\sigma, z) = 1 - \frac{1}{|z|} \sum_{j=1}^{m} \frac{z_j}{\sigma(j)} \sum_{l=1}^{\sigma(j)} z_{\sigma^{-1}(l)}$$

Ramaswamy H. G., Agarwal S., Tewari A. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses //Advances in Neural Information Processing Systems. – 2013. – C. 1475-1483.

Consistent Surrogate for mAP

The loss can be rewritten as

$$L(\sigma, z) = 1 - \sum_{p=1}^{r} \sum_{q=1}^{p} \frac{1}{\max(\sigma(p), \sigma(q))} \frac{z_p z_q}{|z|}$$

For low-rank $F_{mAP} \in \mathbb{R}^{r! \times \frac{r(r+1)}{2}}$

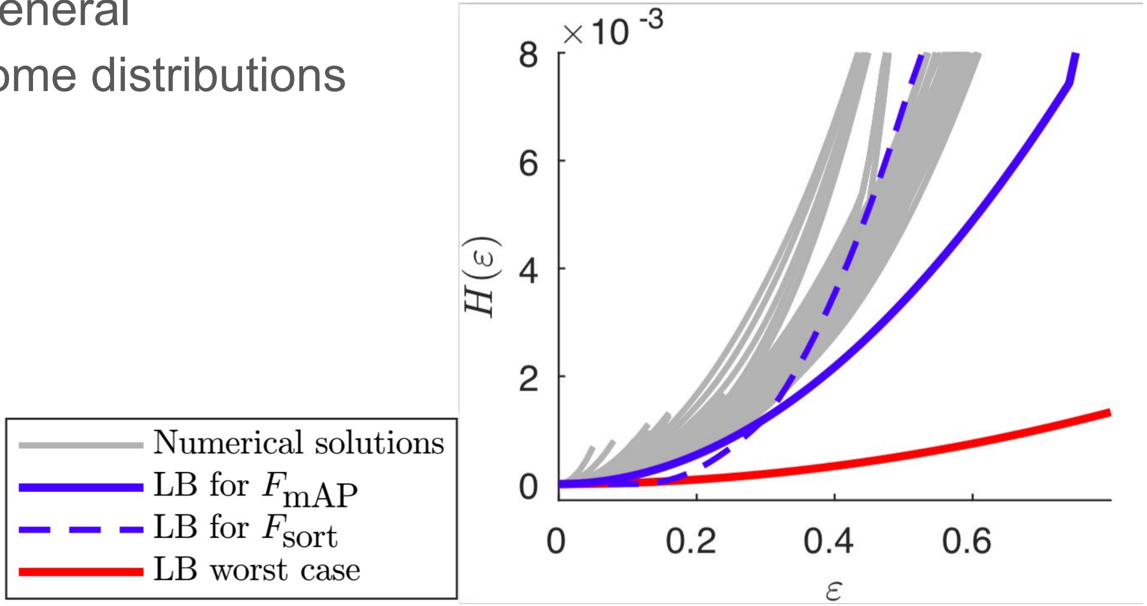
$$(F_{mAP})_{\sigma,pq} = \frac{1}{\max(\sigma(p), \sigma(q))}$$

$\operatorname{span} L \subseteq \operatorname{span} F$, i.e. consistent surrogate

Prediction $y = \operatorname{pred} f(x) = \operatorname{argmax}_{\hat{y}} f_{\hat{y}}(x)$ is a QAP

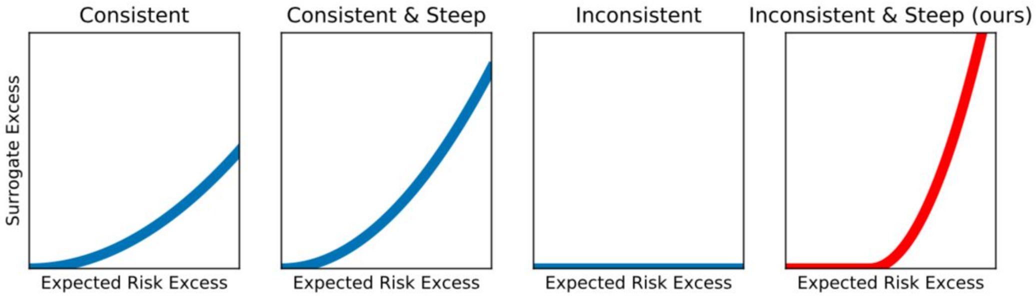
Inconsistent Surrogate for mAP

- Ramaswamy et. al showed that for prediction reduces to sorting $F_{sort} \in \mathbb{R}^{r! \times r}, (F_{sort})_{\sigma,p} = \frac{1}{\sigma(p)}$
- Inconsistent in general
- Consistent for some distributions



Conclusion

- Worst-case guarantees for learning with inconsistent surrogates
- Choice of score function matters
- Quantified the trade-off
  - Optimization complexity
  - Learning guarantees



Future directions:

- Beyond worst-case analysis? Data distribution assumptions
- Non-quadratic surrogates
- Explore more high-rank losses