On Structured Prediction Theory with Calibrated Convex Surrogate Losses

Rushab Munot

University of Pennsylvania

rushab@seas.upenn.edu

Instructor: Shivani Agarwal

Authors

Anton Osokin   Francis Bach   Simon Lacoste-Julien

October 30, 2018

Overview

- Structured Prediction - Obstacles
  - Exponential number of classes
  - Cost-Sensitive Task Loss
- Previous Work
  - Consistent but not efficient (non-convex)
  - Efficient (convex) but not consistent
- What we need
  - Consistency for surrogate losses **and**
  - Efficient algorithm

Overview - continued

- Convexity + Consistency $\not\Longrightarrow$ Efficient Algorithm
  - Exponential constant in Generalization Error bound (eg. Ciliberto et. al. 2010)
  - To have error ϵ on Task Loss
    1. Exponentially small error on surrogate loss
    2. Exponential number of iterations
- **Approach**
  - **Calibration function**: Connect actual excess risk and surrogate excess risk
  - **Exponential constants**: Constrain the score vector space
  - **Online SGD**

Recap - Calibrated Surrogates for 0-1 Loss

- Inner Risk: $L_l(\eta, \alpha) = E_{y \sim \eta}[l(y, \alpha)]$
- Bayesian Inner Risk: $H_l(\eta) = \inf_\alpha L_l(\eta, \alpha)$
- Inner Regret/Excess Risk: $\mathcal{R}_l(\eta, \alpha) = L_l(\eta, \alpha) - H_l(\eta)$

- **0-1 Calibrated Surrogate Loss** ($\psi : [0, 1] \to \mathbb{R}$):

$$\inf_{R_{0-1}(\eta, sign(f)) > 0} R_\psi(\eta, f) > 0 \ \ldots \ \forall \ \eta \in [0, 1]$$

- Bartlett et al (2006): Calibrated surrogate losses admit a regret transfer bound w.r.t. the 0-1 loss
- Calibrated Multi-class losses

Notation

- $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$ (structured)
  $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}$
- $|\mathcal{Y}| = k$ (finite, exponential in size of $\boldsymbol{y}$)
- Loss function $L \in \mathbb{R}^{k \times k}$ (non-negative)

- Score function: $\mathfrak{f} : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^k$
  $\text{pred}(\mathfrak{f}(x)) = \text{argmax}_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \mathfrak{f}_{\hat{\boldsymbol{y}}}(x)$
- $\mathfrak{F}_{\mathcal{F}}$ = Set of score functions
- Surrogate Loss $\Phi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ (continuous and bounded below)

Notation

- Generalization Error:

$$
\begin{aligned}
er_{\mathcal{D}}^{L} &= E_{(x,y)\sim\mathcal{D}}[L(\text{pred}(f(x)),\ y)] \\
&= E_{x\sim\mathcal{D}_{\mathcal{X}}}[l(\text{pred}(f(x)),\ P(\cdot|x))]
\end{aligned}
$$

$$
l(\boldsymbol{f}, \boldsymbol{q}) = \sum_{c=1}^{k} q_c L(\text{pred}(f), c) \ \ldots \text{ Inner Risk}
$$

- Surrogate Error:

$$
er_{\mathcal{D}}^{\Phi} = E_{(x,y)\sim\mathcal{D}}[\Phi(f(x),\ y)]
$$

$$
\phi(\boldsymbol{f}, \boldsymbol{q}) = \sum_{c=1}^{k} q_c \Phi(f, c) \ \ldots \text{ Surrogate Inner Risk}
$$

- R_l and R_ϕ denote the inner regret and surrogate inner regret

Loss functions under consideration

- 01 Loss
- Hamming Loss
- Block loss:
  Divide into b blocks of size s
  $k = bs$
- Mixed Loss:
  $L_{mixed,01,b}(\hat{y}, y) =$
  $\eta L_{01}(f, q)(\hat{y}, y) + (1 - \eta) L_b(\hat{y}, y)$

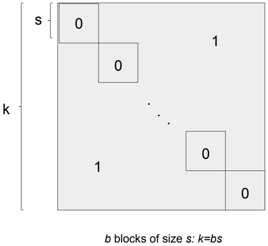


Figure: Block 0-1 Loss

Calibration Function

- Connects actual and surrogate inner regrets

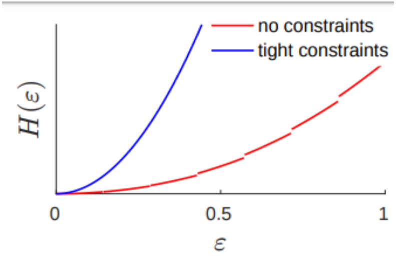
Definition (Calibration Function)

For a task loss L, surrogate loss Φ, and space of allowed score vectors $\mathcal{F}$,

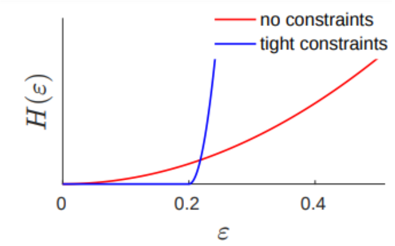
$$\mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon) = \inf_{\boldsymbol{f}\in\mathcal{F},\ \boldsymbol{q}\in\Delta_k} R_\phi(f, q)$$

$$\text{s.t. } R_l(\boldsymbol{f}, \boldsymbol{q}) > \epsilon$$

- Implications:
  1. $\mathcal{H}_{L,\Phi,\mathcal{F}}(R_l(f, q)) \leq R_\phi(f, q)$
  2. $\mathcal{H}_{L,\Phi,\mathcal{F}}$ is non-decreasing
  3. Larger $\mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon)$ is better



(a)
Surrogate Loss: Quadratic
Task Loss: Hamming
Constraints: $\mathcal{F} = \text{span}(L_{Ham,T})$



(b)
Surrogate Loss: Quadratic
Task Loss: Mixed $L_{01,b,\eta}$
Constraints: $\mathcal{F} = \text{span}(L_{01,b,\eta})$

Figure: Exponential rise in $\mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon)$ with constraints. $\mathcal{H}_{L,\Phi,\mathcal{F}}$ can be inconsistent/ not continuous

Consistency

Note that, $\forall\, \epsilon,\ \mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon) > 0 \implies$ consistency

Theorem

$\check{\mathcal{H}}_{L,\Phi,\mathcal{F}}$ is the lower convex envelope of $\mathcal{H}_{L,\Phi,\mathcal{F}}$. Then

$$\mathit{regret}_{\mathcal{D}}^{\Phi}[f] < \check{\mathcal{H}}_{L,\Phi,\mathcal{F}}(\epsilon) \implies \mathit{regret}_{\mathcal{D}}^{L} < \epsilon$$

Definition (η consistency)

Φ is η-consistent iff

1. $\mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon) > 0,\ \forall\, \epsilon > \eta$
2. $\mathcal{H}_{L,\Phi,\mathcal{F}}$ is finite for some $\hat{\epsilon} > 0$

Allows optimization up to a certain accuracy. Can be much faster (Mixed 01 block loss).

Constraints of $\mathcal{F}$

- **Approximation Error:**

$$er_{\mathcal{D}}^{L,\mathcal{F},*} - er_{\mathcal{D}}^{L,*} = \inf_{\mathfrak{f} \in \mathfrak{F}_{\mathcal{F}}} er_{\mathcal{D}}^{L}[\mathfrak{f}] - \inf_{\mathfrak{f}} er_{\mathcal{D}}^{L}[\mathfrak{f}]$$

- Given $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}$, $\boldsymbol{q} = P(\cdot|\boldsymbol{x}) \in \mathbb{R}^k$ and $L \in \mathbb{R}^{k \times k}$
  - Expected loss if you predict $y = c$ is $(L\boldsymbol{q})_c$
  - Observe that $\boldsymbol{f} = -L\boldsymbol{q}$ is optimal
- $\text{span}(L) \subseteq \mathcal{F} \implies 0$ approximation error
- Can restrict $\mathcal{F} = \text{span}(L)$
- Low Rank L may get rid of exponential constants

Quadratic Surrogate Loss

Definition

$\boldsymbol{f} \in \mathcal{F}, \; \boldsymbol{y} \in \mathcal{Y}$

$$\Phi_{quad}(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{2k} ||\boldsymbol{f} + L(:, \boldsymbol{y})||^2 = \frac{1}{2k} \sum_{c=1}^{k} \Big(\boldsymbol{f}_c + L(c, \boldsymbol{y}) \Big)^2$$

- Let $F \in \mathbb{R}^{k \times r}$ (e.g. F is the basis of $span(L)$)
- Constrain $\mathcal{F} = \text{span}(F) = \{F\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^r\}$
- $\dim(\mathcal{F}) = \text{rank}(F) \leq r << k$
- Surrogate inner regret: $R_\phi(\boldsymbol{f} = F\boldsymbol{\theta}) = \frac{1}{2k} ||F\boldsymbol{\theta} + L\boldsymbol{q}||^2$ (Convex in $\boldsymbol{f}$ and $\boldsymbol{q}$)

Lower Bound on $\mathcal{H}_{\Phi_{quad},L,\mathcal{F}}$

- i = Label with lowest expected loss
  j = Label with highest predicted score

$$\mathcal{H}_{ij}(\epsilon) = \inf_{i,j \in pred(\mathcal{F})} R_\phi(f, q)$$

$$\begin{aligned}
&\text{s.t.} \\
&(Lq)_i \le (Lq)_j - \epsilon \\
&(Lq)_i \le (Lq)_c \ \forall\ c \in pred(\mathcal{F}) \\
&f_j \ge f_c \ \forall\ c \in pred(\mathcal{F}) \\
&\boldsymbol{f} \in \mathcal{F} \\
&\boldsymbol{q} \in \Delta_k
\end{aligned}$$

- The union of feasibility sets is $\mathcal{F} \times \Delta_k$
- $\mathcal{H}_{L,\Phi,\mathcal{F}}(\epsilon) = \min_{i,j \in pred(\mathcal{F});\ i \neq j} \mathcal{H}_{ij}(\epsilon)$

Lower Bound on $\mathcal{H}_{\Phi_{quad},L,\mathcal{F}}$

- For the quadratic surrogate, this translates to

$$\mathcal{H}_{ij}(\epsilon) = \inf_{\theta,\mathbf{q}} ||F\boldsymbol{\theta} + L\boldsymbol{q}||^2$$
$$(\mathbf{e_i} - \mathbf{e_j})^T Lq = -\epsilon$$
$$(\mathbf{e_i} - \mathbf{e_j})^T \boldsymbol{\theta} \le 0$$

Can prove that

$$\mathcal{H}_{\Phi_{quad},L,\mathcal{F}} \ge \frac{\epsilon^2}{2k}\left(\frac{1}{\max_{i\neq j}||P_{\mathcal{F}}(\mathbf{e_i} - \mathbf{e_j})||_2^2}\right) \ge \frac{\epsilon^2}{4k}$$

$P_{\mathcal{F}}$: Operator for projection onto $\mathcal{F}$
$\mathbf{e_c}$: c^{th} standard basis vector in $\mathbb{R}^k$

Task Loss: 0-1

- $\mathcal{H}_{L_{01},\Phi_{quad},\mathcal{F}}(\epsilon) \geq \frac{\epsilon^2}{4k}$
- This bound is tight
- Note that $\mathcal{F} = \text{span}(L_{01}) = \mathbb{R}^k$
  Hence the projection is always $e_i - e_j$
- k is exponential in the dimension of $\boldsymbol{y}$
- 0-1 Loss has poor guarantees as expected

Block 0-1 Loss

- $\mathcal{F} = \text{span}(L_{block,b})$
- $rank(L) = b$
- $||P_{\mathcal{F}}(\mathbf{e_i} - \mathbf{e_j})||_2^2 = \frac{2}{s} \ldots i \neq j,\ else\ 0$
- $\mathcal{H}_{L_{block,T},\Phi_{quad},\mathcal{F}}(\epsilon) \geq \frac{\epsilon^2}{4b}$ (not tight)

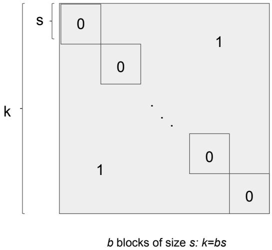


Figure: Block 0-1 Loss

Hamming Loss

Hamming Loss: Fraction of wrong labels

- T = Sequence Length
- $\mathcal{F} = \text{span}(L_{Ham,T})$
- $rank(L) = T + 1$
- $\mathcal{H}_{L_{Ham,T},\Phi_{quad},\mathcal{F}}(\epsilon) \geq \frac{\epsilon^2}{8T}$ (tight)

Upper Bound on $\mathcal{H}_{\Phi_{quad},L,\mathcal{F}}$

- Turns out we cannot give good guarantees if $\mathcal{F}$ is not constrained

Theorem

For a Loss Matrix L which is a pseudometric, and unconstrained $\mathcal{F}$ (i.e. $\mathcal{F} = \mathbb{R}^k$),

$$\mathcal{H}_{\Phi_{quad},L,\mathcal{F}}(\epsilon) < \frac{\epsilon^2}{2k}$$

- This is why we need to constrain the score vector space.
- Also this bound is the reason why previous algorithms were not efficient.

Learning Guarantees

- Online (kernel) Projected Averaged SGD
- $F \in \mathbb{R}^{k \times r}$ $\psi : \mathcal{X} \to \mathbb{R}^d$ Feature Map; $W \in \mathbb{R}^{r \times d}$
- $\mathfrak{f}(x) = FW\psi(x)$
- SGD Update for $(\boldsymbol{x}^n, \boldsymbol{y}^n) \sim \mathcal{D}$:

$$W^{(n)} = P_D\left[W^{(n-1)} - \gamma^{(n)} F^T \nabla\Phi\psi(\boldsymbol{x}^n)^T\right]$$

P_D is the projection on the ball of radius D wrt the Hilbert Schmidt norm. γ is the step size. (No need to calculate the feature map - kernel trick)

Learning Guarantees

- **Assumption:** The surrogate regret $er_{\mathcal{D}}^{\Phi}$ has a global minimum $\boldsymbol{f}^*$ over the function class $\mathfrak{F}_{\mathcal{F}} = \{\boldsymbol{f} \mid \boldsymbol{f}(\boldsymbol{x}) = FW\psi(\boldsymbol{x})\}$.
  Not required - but analysis becomes complicated!
- If we have
  1. $\Phi(f, y)$ is bounded below and convex wrt f for all y
  2. $||F^T \nabla \Phi \psi(\boldsymbol{x}^n)^T||_{HS}^2 \leq M^2$
  3. $||W^*||_{HS} < D$

  then with $\gamma = \frac{2D}{M\sqrt{N}}$, we have

$$\boldsymbol{E}\left[\mathrm{er}_{\Phi}\left[\bar{\boldsymbol{f}}^{(N)}\right]\right] - \mathrm{er}^{\Phi,\mathcal{F},*} \leq \frac{2DM}{\sqrt{N}}$$

$\bar{f}^{(N)} = \frac{1}{N}\sum_{i=1}^{N} f^{(N)}; \quad f^{(n)} = FW^{(n)}\psi(x)$

Learning Guarantees

- Under the same assumptions, for any $\epsilon > 0$, we have $\boldsymbol{E}\left[\mathrm{er}_L\left[\bar{\boldsymbol{f}}^{(N)}\right]\right] - er^{L,\mathcal{F},*} < \epsilon$ if

$$N > \frac{4D^2M^2}{\check{\mathcal{H}}^2_{L,\Phi,\mathcal{F}}(\epsilon)}$$

- 01 Loss $DM = O(k)$
- Hamming Loss $DM = O(\log_2 k^3$
- Block 01 Loss $DM = O(b)$

Thank You