On Structured Prediction Theory with Calibrated Convex Surrogate Losses

Anton Osokin — NRU HSE, Moscow, Russia
Francis Bach — INRIA/ENS, Paris, France
Simon Lacoste-Julien — MILA and DIRO, Université de Montréal, Canada

Summary

We focus on **theoretical aspects** of structured prediction and provide some insights how to build methods with guarantees. We want **consistency guarantees** and **learning (optimization) complexity**

Contributions:
- Compute tight bounds on **calibration function** to relate convex surrogate risk to true risk
- Use SGD analysis to get learning complexity
- Monitor carefully exponential constants (vs. prior work)
- Formalize intuition that learning is easier for some losses: structured losses vs. 0-1 loss

Structured Prediction

Structured prediction = ML for predicting structured objects.

Examples

Handwriting recognition (OCR)

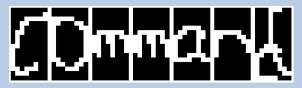
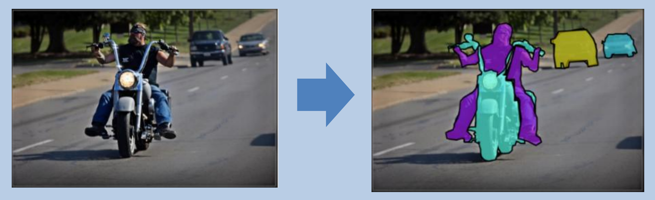
 → command

Image segmentation



Key differences from binary classification:
- Exponential number of classes
- Cost-sensitive prediction – not all mistakes are equal

Structured prediction setup

- Given input $\boldsymbol{x}$ =  predict $\boldsymbol{y}$ = command
- k = **number of labels** (exponential in sequence length)
- **Loss** $L \in \mathbb{R}^{k \times k}$ (e.g., $L(\hat{\boldsymbol{y}}, \boldsymbol{y})$ is the Hamming distance)
- Model = **score function** $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^k$
- **Prediction**: $\underset{\boldsymbol{y}=1,\dots,k}{\operatorname{argmax}} \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{x})$ need inference (e.g., Viterbi)
- Non-parametric: scores defined by universal kernels on $\boldsymbol{x}$
- An optimal predictor: $\boldsymbol{f}^*(\boldsymbol{x}) = -L\,\boldsymbol{q}_{\boldsymbol{y}}$ with $\boldsymbol{q}_{\boldsymbol{y}} = P(\boldsymbol{y} \mid \boldsymbol{x})$
- Loss matrix rank (connected with complexity):
  - 0-1 loss: $\dim(\operatorname{span}(L)) = k$
  - Hamming loss: $\dim(\operatorname{span}(L)) \approx \log_2(k)$

Learning in structured prediction

Learning = minimize the **population risk**

$$\mathcal{R}_L(\boldsymbol{f}) := \mathbf{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\, L\Big(\operatorname{argmax}(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{y}\Big)$$

⚠ Non-convex => no guarantees!

Instead, minimize the **surrogate risk**

$$\mathcal{R}_\Phi(\boldsymbol{f}) := \mathbf{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\, \Phi(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x}))$$

Convex => optimization guarantees!

Examples: structured SSVM, conditional likelihood

Theory for Structured Prediction

Comparison with prior work:

- **PAC-Bayes bounds** (McAllester, 2007) (McAllester&Keshet, 2011) (London et al., 2016)
  - Consistent, but not convex
  - No provable optimization guarantees
- **Rademacher complexity bounds** (Cortes et al., 2016)
  - Convex, but not consistent
  - No provable optimization guarantees
- **Input-output kernel regression** (Ciliberto et al., 2016) (Brouard et al., 2016)
  - Convex, consistent
  - Exponential constants in sample complexity bounds
- **This work**
  - Consistency
  - Convex: efficient optimization
  - No exponential constants

  } advantages

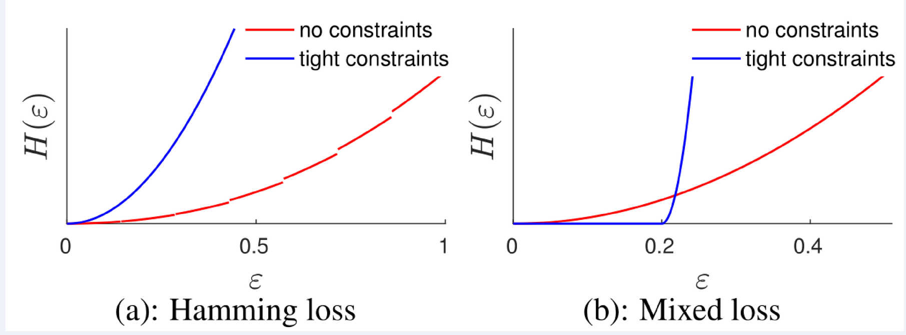
Calibration Functions

Calibration function connects the actual and surrogate risks

$$H(\text{excess of actual } \mathcal{R}_L) \le \text{excess of surrogate } \mathcal{R}_\Phi$$

Calibration functions can characterize **consistency** ($H(\varepsilon) > 0,\ \varepsilon > 0$)

Constraints ($\boldsymbol{f} = F\boldsymbol{\theta}$) on the set of scores influence H.
- Tight constraints increase H
- Can break consistency
- Good choice: $\operatorname{span}(L)$



(a): Hamming loss (b): Mixed loss

Optimization Accuracy

Calibration functions are not sufficient because
- scale is arbitrary defined
- no connection to the actual optimization
- no notion of sample complexity

Online SGD convergence rate: $\mathbf{E}[\mathcal{R}_\Phi(\bar{\boldsymbol{f}}^{(N)})] - \mathcal{R}^*_{\Phi,\mathcal{F}} \le \frac{2DM}{\sqrt{N}}$

(Upper bound on the expected squared norm on the stochastic gradient → M; Upper bound on the solution norm → D; averaged iterate → $\bar{\boldsymbol{f}}^{(N)}$)

⚠ We need structure of F and L to run SGD efficiently

In expectation, online SGD needs $N^* := \frac{4D^2M^2}{H^2(\varepsilon)}$ iterations to have $\mathbf{E}[\mathcal{R}_L(\bar{\mathrm{f}}^{(N)})] < \mathcal{R}^*_{L,\mathcal{F}} + \varepsilon$

(calibration function → $H^2(\varepsilon)$)

Analysis for Quadratic Surrogate

- Computing calibration functions is difficult in general.
- Compute only for a special "quadratic" surrogate

$$\Phi_{\text{quad}}(\boldsymbol{f}, \boldsymbol{y}) := \tfrac{1}{2k}\|\boldsymbol{f} + L(:, \boldsymbol{y})\|_2^2 = \tfrac{1}{2k}\sum_{c=1}^{k}(f_c + L(c, \boldsymbol{y}))^2, \quad \boldsymbol{f} = F\boldsymbol{\theta}$$

- Hardness result: **upper bound** (pseudo-metric losses, no constraints)

$H(\varepsilon) \le \frac{\varepsilon^2}{2k}$ ⚠ Exponentially small!

- Easiness result: **lower bound** for all losses: if there are good constraints then the calibration function is not small

$H(\varepsilon) \ge \frac{\varepsilon^2}{2k \max_{i \ne j} \|P_{\mathcal{F}}\Delta_{ij}\|_2^2} \ge \frac{\varepsilon^2}{4k}$ Can be large!

Depends on projections on span(F) of "bad direction" $\Delta_{ij} = \mathbf{e}_i - \mathbf{e}_j \in \mathbb{R}^k$

- Some **exact values**: (with scores in $\operatorname{span}(L)$)
  - 0-1 loss $H(\varepsilon) = \frac{\varepsilon^2}{4k}$ ⚠ Exponentially small!
  - Hamming loss (T variables) $H(\varepsilon) = \frac{\varepsilon^2}{8T}$ Large!
  - Block 0-1 loss (b blocks) $H(\varepsilon) = \frac{\varepsilon^2}{4b}$ Large!

- choice of constraints **can break consistency** (for small ε), but **make learning much faster**

Example: mixed loss $L_{01,b,\eta} := \eta L_{01} + (1-\eta)L_{01,b}$, scores in $\operatorname{span}(L_{01,b})$

$$H(\varepsilon) = \begin{cases} \frac{O(1)}{4b}(\varepsilon - \frac{\eta}{2})^2, & \frac{\eta}{2} \le \varepsilon \le 1, \\ 0, & 0 \le \varepsilon \le \frac{\eta}{2} \end{cases}$$

Large! ⚠ Non-consistent!

- Computing the SGD constants:
  - 0-1 loss $DM = O(k)$ ⚠ Exponentially large!
  - Hamming loss (T variables) $DM = O(\log_2^3 k)$ Small!
  - Block 0-1 loss (b blocks) $DM = O(b)$ Small!

References

(McAllester, 2007) McAllester, D. Generalization bounds and consistency for structured labeling. In Predicting Structured Data. MIT Press, 2007.

(McAllester&Keshet, 2011) McAllester, D. and Keshet, J. Generalization bounds and consistency for latent structural probit and ramp loss. In NIPS, 2011.

(London et al., 2016) London, B., Huang, B. and Getoor, L. Stability and generalization in structured prediction. Journal of Machine Learning Research (JMLR), 17(222):1–52, 2016.

(Cortes et al., 2016) Cortes, C., Kuznetsov, V., Mohri, M. and Yang, S.. Structured prediction theory based on factor graph complexity. In NIPS, 2016.

(Ciliberto et al., 2016) Ciliberto, C., Rudi, A. and Rosasco, L. A consistent regularization approach for structured prediction. In NIPS, 2016.

(Brouard et al., 2016) Brouard, C., Szafranski, M. and d'Alché-Buc, F. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. Journal of Machine Learning Research (JMLR), 17(176):1–48, 2016.