On Structured Prediction Theory with Calibrated Convex Surrogate Losses

Anton Osokin Francis Bach Simon Lacoste-Julien

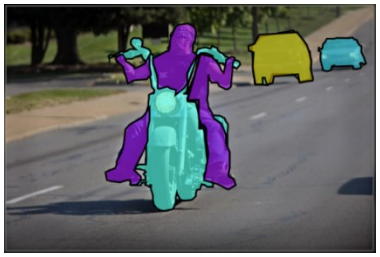
08.09.2017

Structured prediction

- Data is often very structured
  - sequences, tables, images, video, etc.
- Structured prediction = output is structured
- Examples:

Input data | Output labels

- Image segmentation

- Handwriting recognition

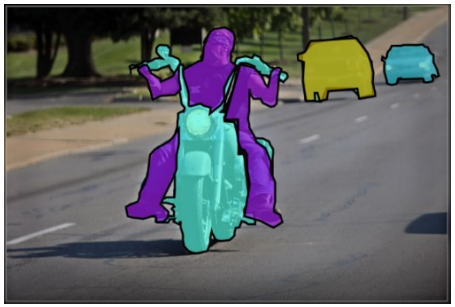

command

Inference and learning

- Inference (prediction)

Input data

+ Model =

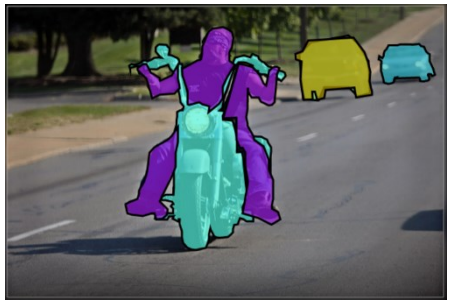


Output labels

- Learning

Input data

+



Output labels

= Model

Training data

We want a method that works!

and has guarantees!

Number of iterations to get ε-accuracy on the test set

Conditional likelihood (e.g., CRFs)

Input ***x***, output ***y***, loss *L*; set of possible ***y*** is exponential!

Training – maximize conditional likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$$

Prediction – minimize expected loss

$$\operatorname{pred}(\boldsymbol{x}) = \underset{\hat{\boldsymbol{y}}}{\operatorname{argmin}} \sum_{\boldsymbol{y}} L(\hat{\boldsymbol{y}}, \boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$$

Problems:

- hard to connect to prediction error
- training does not know about the loss
- the whole distribution is hard to get

Empirical risk minimization

Input $\boldsymbol{x}$, output $\boldsymbol{y}$, loss L; set of possible $\boldsymbol{y}$ is exponential!
We want to minimize the population risk

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

Here $\boldsymbol{f}$ gives a score for each label
Usually we cannot optimize the risk, so we use a surrogate

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

Examples:

$$\Phi_{\mathrm{SSVM}}(\boldsymbol{f}, \boldsymbol{y}) := \max_{\hat{\boldsymbol{y}}\in\mathcal{Y}} \big(f_{\hat{\boldsymbol{y}}} + L(\hat{\boldsymbol{y}}, \boldsymbol{y})\big) - f_{\boldsymbol{y}}$$

$$\Phi_{\log}(\boldsymbol{f}, \boldsymbol{y}) := -f_{\boldsymbol{y}} + \log \sum_{\hat{\boldsymbol{y}}\in\mathcal{Y}} \exp f_{\hat{\boldsymbol{y}}}$$

Where is structure?

Input ***x***, output ***y***, loss *L*; set of possible ***y*** is exponential!
We want to minimize the population risk

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

Here ***f*** gives a score for each label
Usually we cannot optimize the risk, so we use a surrogate

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

1) Loss *L* is structured (Hamming, block 0-1)

2) Scores are structured: $\boldsymbol{f} \in \mathcal{F}$ (for example, separable)

Empirical risk minimization

Input $\boldsymbol{x}$, output $\boldsymbol{y}$, loss L; set of possible $\boldsymbol{y}$ is exponential!
We want to minimize the population risk

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

Here $\boldsymbol{f}$ gives a score for each label
Usually we cannot optimize the risk, so we use a surrogate

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

Issues to be dealt with:

- Surrogate Φ instead of the actual loss L — Consistency
- Finite dataset
- Optimization accuracy

} Online SGD

Consistency for binary classification

Input ***x***, output $y \in \{-1, 1\}$, loss $L(\hat{y}, y) = [\hat{y} \neq y]$
We want to minimize the population risk

$$\mathcal{R}_L(f) := \mathbf{E}_{(x,y)\sim\mathcal{D}} \, [yf(x) > 0]$$

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

Theorem (Bartlett et al., 2006)
Assume Φ is convex, differentiable at 0 and $\Phi'(0) < 0$.
Then Φ is classification-calibrated (consistent).

Examples: exponential, hinge, logistic, truncated quadratic

Calibration function

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

With expected and excess loss/surrogate defined as

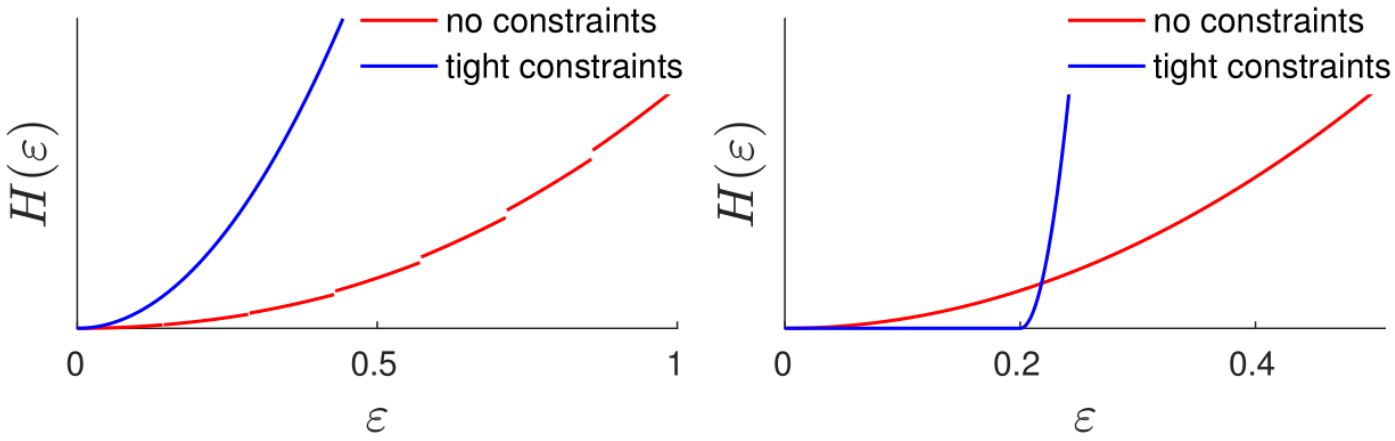
$$\ell(\boldsymbol{q},\boldsymbol{f}) = \sum_{c=1}^{k} q_c L(\text{argmax}(\boldsymbol{f}), c), \quad \phi(\boldsymbol{q},\boldsymbol{f}) = \sum_{c=1}^{k} q_c \Phi(c,\boldsymbol{f})$$
$$\delta\phi(\boldsymbol{q},\boldsymbol{f}) = \phi(\boldsymbol{q},\boldsymbol{f}) - \min_{\hat{\boldsymbol{f}}\in\mathcal{F}} \phi(\boldsymbol{q},\hat{\boldsymbol{f}})$$
$$\delta\ell(\boldsymbol{q},\boldsymbol{f}) = \ell(\boldsymbol{q},\boldsymbol{f}) - \min_{\hat{\boldsymbol{f}}\in\mathcal{F}} \ell(\boldsymbol{q},\hat{\boldsymbol{f}})$$

(Zhang, 2004) and others

Calibration function

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$



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

(Zhang, 2004) and others

Calibration function and consistency

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

If we optimize up to $\mathcal{R}_\Phi(\boldsymbol{f}) < \mathcal{R}_\Phi^* + H_{\Phi,L,\mathcal{F}}(\varepsilon)$

We get $\mathcal{R}_L(\boldsymbol{f}) < \mathcal{R}_L^* + \varepsilon$

If $H_{\Phi,L,\mathcal{F}}(\varepsilon) > 0,\ \varepsilon > 0$ the surrogate is consistent.

(Zhang, 2004) and others

If consistent, what can go wrong?

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

If we optimize up to $\mathcal{R}_\Phi(\boldsymbol{f}) < \mathcal{R}_\Phi^* + H_{\Phi,L,\mathcal{F}}(\varepsilon)$

We get $\mathcal{R}_L(\boldsymbol{f}) < \mathcal{R}_L^* + \varepsilon$

If $H_{\Phi,L,\mathcal{F}}(\varepsilon) > 0,\ \varepsilon > 0$ the surrogate is consistent.

If consistent, what can go wrong?

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

Required accuracy $H_{\Phi,L,\mathcal{F}}(\varepsilon)$ can be exponentially small!

We will never reach it: finite dataset and runtime

Easy-to-analyze surrogate

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

Simple surrogate, consistent for any given loss:

$$\Phi_{\text{quad}}(\boldsymbol{y},\boldsymbol{f}) = \tfrac{1}{2k}\|\boldsymbol{f} + L(:,\boldsymbol{y})\|_2^2 \qquad \text{where} \quad \mathfrak{f}(\boldsymbol{y}) = F\boldsymbol{\theta}(\boldsymbol{x})$$

01-loss

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

k – number of classes

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

If *k* is big then H is very small, i.e. we need crazy accuracy!

Block 01-loss

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

k – number of classes, *s* – block size

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

The loss is highly structured, but *H* is still very small!

What is missing?

Block 01-loss + constraints on the scores

Calibration function connects the surrogate and the loss.

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) = \min_{\boldsymbol{f},\boldsymbol{q}} \delta\phi(\boldsymbol{q},\boldsymbol{f})$$
$$\text{s.t. } \delta\ell(\boldsymbol{q},\boldsymbol{f}) \geq \varepsilon,$$
$$\boldsymbol{f} \in \mathcal{F},\ \boldsymbol{q} \in \Delta_k,$$

k – number of classes, *s* – block size, *d* – number of blocks

Calibration function: $H_{\Phi_{\text{quad}},L,\mathcal{L}}(\varepsilon) = \frac{\varepsilon^2}{4d}$ *d* is small!

d represents complexity of the loss

Hamming loss

Hamming loss on binary sequences:

$$H_{\Phi_{\text{quad}},\text{Ham},\mathbb{R}^k} = \frac{\varepsilon^2}{4k} \text{ for small } \varepsilon$$ very small!

With constraints on the scores:

$$H_{\Phi_{\text{quad}},\text{Ham},\mathcal{F}} = \frac{\varepsilon^2}{8T}$$ much better!

T – length of the sequence

Lower bound for any loss

Theorem 1. *For any loss L, its quadratic surrogate $\Phi_{\boldsymbol{\theta}}$, and a score subspace $\mathcal{F}$ containing the column space of L, the calibration function can be lower bounded:*

$$H_{\Phi_{\boldsymbol{\theta}},L,\mathcal{F}}(\varepsilon) \geq \frac{\varepsilon^2}{2k} \min_{i,j=1,\ldots,k} \frac{1}{\|P_{\mathcal{F}}\delta_{ij}\|_2^2} \geq \frac{\varepsilon^2}{4k}$$

where $P_{\mathcal{F}}$ is the projection of subspace $\mathcal{F}$ of allowed scores,
$\delta_{ij} = \delta_i - \delta_j \in \mathbb{R}^k$ defines the error direction.

The bound is tight in some cases!
(Block 01 loss, Hamming loss with appropriate constraints)

The bound suggests that good losses are low-rank.
Subspace should not be aligned with δ_{ij}

OK, say calibration function is large. Are we done?

No, scale of calibration function is defined arbitrarily.

We have to connect to optimization and trace constants.

Minimizing the surrogate risk

Consider minimizing

$$\mathcal{R}_{\Phi}(\mathfrak{f}) = \mathbf{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}\ \tfrac{1}{2k}\|\mathfrak{f}(\boldsymbol{x}) + L(:,\boldsymbol{y})\|_2^2$$

Where we assume linear dependency on features $\psi(\boldsymbol{x}) \in \mathbb{R}^d$

$$\mathfrak{f}(\boldsymbol{x}) = FW\psi(\boldsymbol{x}), \quad W \in \mathbb{R}^{r\times d}$$

Stochastic gradient:

$$\nabla_W \mathcal{R}_{\Phi}^{(n)} = \tfrac{1}{k}F^{\mathsf{T}}(FW\psi(\boldsymbol{x}) - L(:,\boldsymbol{y}))\psi(\boldsymbol{x})^{\mathsf{T}}$$

Now we just need a suitable algorithm with convergence rate.

Normalizing by SGD

Projected averaged SGD with constant step size

$$W^{(n)} := P_D\left[W^{(n-1)} - \gamma \nabla_W \mathcal{R}_\Phi^{(n)}\right], \;\; \gamma = \frac{2D}{M\sqrt{N}}$$

D – bounds the distance to optimum

M^2 – bounds the expectation of squared L2-norm of the gradient

(Nemirovski et al.) Convergence rate of averaged iterates

$$\mathbf{E}[\mathcal{R}_\Phi(\bar{\mathfrak{f}}^{(N)})] - \mathcal{R}_\Phi^* \leq \frac{2DM}{\sqrt{N}}, \qquad \bar{\mathfrak{f}}^{(N)} := \frac{1}{N}\sum_{n=1}^{N} F W^{(n)} \psi(\boldsymbol{x})^\mathsf{T}$$

Combining this with calibration functions?

Big assumption: well-specified optimization

There exists a global minimum of $\mathcal{R}_\Phi(\mathfrak{f})$
w.r.t. all measurable function **f(x)**
that belongs to the function class $\mathfrak{f}(\boldsymbol{x}) = FW\psi(\boldsymbol{x})$

This assumption can be relaxed if we use a universal kernel
(that can approximate any function)
and if we are very careful with optimization

Normalizing by SGD

Projected averaged SGD with constant step size

$$W^{(n)} := P_D\left[W^{(n-1)} - \gamma \nabla_W \mathcal{R}_\Phi^{(n)}\right], \quad \gamma = \frac{2D}{M\sqrt{N}}$$

D – bounds the distance to optimum

M^2 – bounds the expectation of squared L2-norm of the gradient

(Nemirovski et al.) Convergence rate of averaged iterates

$$\mathbf{E}[\mathcal{R}_\Phi(\bar{\mathfrak{f}}^{(N)})] - \mathcal{R}_\Phi^* \leq \frac{2DM}{\sqrt{N}}, \qquad \bar{\mathfrak{f}}^{(N)} := \frac{1}{N}\sum_{n=1}^{N} F W^{(n)} \psi(\boldsymbol{x})^\mathsf{T}$$

Combining this with calibration functions, we need

$$T \geq \frac{4D^2M^2}{H^2_{\Phi,L,\mathcal{F}}(\varepsilon)} \quad \text{iterations to reach} \quad \mathcal{R}_L(\tilde{\boldsymbol{f}}^{(T)}) < \mathcal{R}_L^* + \varepsilon,$$

Normalizing by SGD: quadratic

We need $T \geq \frac{4D^2M^2}{H^2_{\Phi,L,\mathcal{F}}(\varepsilon)}$ iterations to reach $\mathcal{R}_L(\tilde{\boldsymbol{f}}^{(T)}) < \mathcal{R}^*_L + \varepsilon$

D – bounds the distance to optimum
M^2 – bounds the expectation of squared L2-norm of the gradient

For quadratic surrogate

$$DM = L^2_{\max}\xi(\kappa(F)\sqrt{r}RQ_{\max}), \quad \xi(z) = z^2 + z,$$

$\kappa(F)$ – the condition number of *F*
r – rank of *F*
R – upper bound on $\|\psi(\boldsymbol{x})\|_2$
$Q_{\max}$ – upper bound on the sum of the norms of marginal probs

Normalizing by SGD: quadratic

We need $T \geq \frac{4D^2M^2}{H^2_{\Phi,L,\mathcal{F}}(\varepsilon)}$ iterations to reach $\mathcal{R}_L(\tilde{\boldsymbol{f}}^{(T)}) < \mathcal{R}^*_L + \varepsilon$

D – bounds the distance to optimum
M^2 – bounds the expectation of squared L2-norm of the gradient

For quadratic surrogate

$$DM = L^2_{\max}\xi(\kappa(F)\sqrt{r}RQ_{\max}), \quad \xi(z) = z^2 + z,$$

0-1 loss: $DM = O(k)$

Block 0-1 loss: $DM = O(b)$

Hamming loss: $DM = O(\log_2^3 k)$

Conclusion

- What are good surrogates?
- Consistency is not enough
- Calibration functions can be computed
- Scale can be chosen by connection to optimization
- Need to be careful with constants

Future work:

- Bounding calibration functions and constants for more losses
- Generalize analysis to other surrogates
- Make this practical!