Today's topics:
- Consistency
- CRF objective
- Variance reduced SGD

**Consistency & calibration function**

Need to relate $S(w)$ to $L(w)$ — "calibration function" \[ \text{[Steinwart]} \]

Relationship is usually very complicated;
- Current results: fake mainly of non-parametric setting
- Where dependence on $x$ is evacuated \[ \big(?\big) \]
- We suppose that $S(x; y; w)$ can be arbitrary for any $x$
  (i.e., $w$ is no-dum)

\[ \text{[Can use a universal kernel to formalize this and}\]
\[ \text{it include } x. \]

* Binary classification: Bartlett et al. characterized a whole family
  of consistent surrogate losses

For multiclass: \[ \text{[Lee et al. 2004, M. C.
  A. 2006, M. A. 2007] showed that multiclass hinge loss is not consistent}
\]
\[ \text{for 0-1 loss}\]
\[ \text{when no majority class (i.e., } p(y|x) \leq \frac{1}{2} \forall y) \]

2 aspects of structured prediction which give a much richer theory than for binary classification for consistency:

1) $p(y|x)$ "noise model" is much richer
2) $l(y, y')$ much richer

Proposed a fix with a surrogate loss that has $\frac{1}{4}$ instead of max $\frac{3}{2}$
[Osoin et al. 2017] — looked at this empirical aspect of optimization, in the simplest setup.

**Calibration Function** for structural cost $\ell$, surrogate loss $\tilde{q}$, and set $W$

$$H_\ell,\tilde{q}(\varepsilon) \triangleq \inf_{w \in W} \sup_{q \in \Delta[1]} \left\{ \tilde{q}(w) - \min_{w'} \tilde{q}(w') \right\}$$

$$\tilde{q}(w) \leq \mathbb{E}_{q \sim (y_i \mid x_i \mid z_i)} \left[ l(y_i, h(x)) \right]$$

Smallest optimization surrogate regret possible (over all dist. $q$)

Consequence (Thm. 3)

$$\forall q, \tilde{q}(w) \leq \tilde{q}^* + H(c) \Rightarrow \mathbb{L}_q(w) \leq \mathbb{L}_{q^*}(w) + H(c)$$

Complex envelope of $H(c)$ is $H^*$

If $H$ is invertible

$$\Rightarrow \tilde{q}(w) \leq \mathbb{L}_q(w) \leq H^{-1} (\tilde{q}(w) - \tilde{q}^*)$$

Sample complexity:

Can link learning with optimization using SGD. While running SGD with $q^{(i)} = q$

We can optimize population $\tilde{q}$ instead of $q$.

Thus, from SGD convergence on $p$, we can see how many iterations needed.
Thus from SGD convergence on $L_q$, can see how many iterations needed
\[ = \frac{4}{\varepsilon} \text{ of samples} \]

\[ \Rightarrow \text{true regret} \quad H(\varepsilon) \]

\[ \Rightarrow \text{translates to } \varepsilon \text{-true regret} \]

Q.E.D. constant step-size SGD projected on a ball of radius $D$ gives $E[L(q_{w^*}) - L_q] \leq \frac{DM}{\varepsilon^2}$ after $\varepsilon$ iterations

(thm 4): to get $E[L(q_{w^*}) - L_q] \leq \varepsilon$

\[ \text{need at most } \varepsilon \geq \left( \frac{DM}{H(\varepsilon)} \right)^2 \]

in our paper, consider the consistent convex surrogate

\[ L(y,w) = \frac{1}{\|y\|} \leq \frac{1}{\|y\|} \left( \langle w, y \rangle + l(y, \hat{y}) \right)^2 \]

\[ \frac{|L(y,w)|}{\|y\|} \leq \frac{1}{|\|y\||} \left( \langle w, y \rangle + l(y, \hat{y}) \right)^2 + \text{const.} \]

\[ \frac{L(y,w)}{\|y\|} \]

so $w^*$ is s.t. $sw^*(\hat{y}) = -Lq(\hat{y})$

\[ \text{Maximize } \Rightarrow \min \text{ cardinal risk} \quad \text{and thus } \text{consistent} \]

$S$ is consistent for any $l$ and $q$ as long as $sw \geq \frac{1}{g} l(y, *); y \in \mathbb{R}^d$

\[ \Rightarrow \text{in paper, we show that if no constraint on } sw(\cdot) \]

\[ \text{then } H(\varepsilon) \leq \frac{\varepsilon}{\varepsilon^2} \text{ for any } \varepsilon \]
main pointer covered today:
- Anton Osokin, Francis Bach, Simon Lacoste-Julien
  On Structured Prediction Theory with Calibrated Convex Surrogate Losses
  https://arxiv.org/abs/1703.02403
- other pointers:
  - canonical paper which presented consistency analysis for binary classification:
    Bartlett, Peter L., Jordan, Michael I., and McAuliffe, Jon D
    Convexity, classification, and risk bound
  - paper which showed that multiclass SVM is not consistent (for the 0-1 loss) and proposed a consistent alternative:
    Lee, Yoonkyung, Lin, Yi, and Wahba, Grace.
    Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data.
    see also the McAllester 2007 paper:

\[
\begin{align*}
\text{need exponential accuracy} &\Rightarrow \text{exponential sample complexity} \\
\text{also, for 0-1 loss, to be consistent, we basically need no constraint on } \mathbf{w}(x). \\
\text{but for logit loss, if small constraint that } S(y) = \frac{e^y}{e^y + e^{-y}} \text{ } \\
\text{over 1 binary variables, then } H(x) = \frac{e^2}{e^2 + 1} \text{ not too big } \Rightarrow \text{can learn?}
\end{align*}
\]

- Moral here:
  - some losses are harder to learn than others (0-1 difficult in general)
  - have linked computation to statistical performance in consistency framework
    \(L(Y, \text{convex surrogate loss})\)
    \(L(Y, \text{kernel stuff})\)
    but lost dependence on \(x\) (need to use SGD with RKHS to fall in details)
    \(\Rightarrow\) different approach than gen. error bound; but gives insights
  - still need more theory?
and interestingly, this recent paper shows that the multiclass SVM *is* consistent for a loss on 3 classes with an "abstain" notion: Ramaswamy, Harish G. and Agarwal, Shivani. *Convex calibration dimension for multiclass loss matrices.* *JMLR*, 17(14):1–45, 2016.

- see also the extensive related work section of the arxiv 2017 paper Osokin et al.