Today:

* PAC-Bayes
* Proof of PAC-Bayes
* Review of surrogate loss

**PAC-Bayes:**

Ockham's razor implies \( \hat{L}(w) \) with \( L_p(w) \)

uniformly over all \( w \in W \rightarrow \) but countable

using complexity \( W \) \( \propto \) "prior"

**PAC-Bayes:**

generalize this to

- arbitrary \( W \)
- general \( f(y,y') \in [0,1] \)

concretely: switch to a randomized predictor

i.e. instead of learning \( \hat{w} \), predicting \( y = h_\hat{w}(x) \)

consider a distribution over \( W \)

predict: first \( w \sim q(w) \); \( y = h_w(x) \)

\[ \hat{L}(W) \rightarrow \mathbb{E}_{q} \left[ L(w) \right] \text{ as the generalization error for } \hat{w} \]

i.e. \( \mathbb{E}_{(y,y')} \mathbb{P}_{q(w)} \mathbb{E}_{w \sim q} L(y, h_w(x)) \)

empirical version

\[ \hat{L}(W) \rightarrow \mathbb{E}_{q} \left[ L(w) \right] \rightarrow \text{ an structured prediction will yield risk surrogate loss} \]

Optimize over \( q \) to get \( \hat{w} \)

**PAC-Bayes Theorem** [McAllester, 1997, 2003]

(\( (y,y') \in [0,1] \)) for any fixed prior \( p \) over \( W \)

and any distribution \( q \) over \( X \times X \)

then with prob \( \geq 1 - \delta \) over \( D_n \cup P \times X \)

it holds that \( \forall \hat{w} \) \( \mathbb{P} \) over \( W \)

\[ \mathbb{E}_{q} \left[ L_p(w) \right] \leq \mathbb{E}_{q} \left[ L_n(w) \right] + \sqrt{\frac{\mathbb{E}_{Q} \left[ L_n(w) \right]}{2m} \mathbb{K}(q || p) \ln \frac{en}{\delta}} \]

\( m \) new complexity term

Note: if \( W \) is countable, \( \mathbb{P} \) \( Q \) \( \propto \) \( \frac{1}{2} \) \( \|w\| \)
McAllester 2001 use Catoni’s PAC-Bayes (thm. version)

\[
\forall q, \quad \text{Eq} (\mathbb{E} (L(w))) \leq \frac{1}{2} \text{Eq} \left[ \mathbb{L} (w) \right] + \delta_n \left[ KL(q || \pi^*) + \ln \frac{1}{\delta_n} \right]
\]

if we use \( \pi = N(0, I) \)

\[
q_0 = N(w, I) \quad \text{motivates \( \hat{w}_n \) (used)}
\]

Then, I in paper:

get \( \delta_n \) slow enough so that \( \delta_n \ln n \to 0 \)

then \( \text{Sprodit} (\hat{w}_n) \overset{a.s.}{\to} L^* = \min_{w \in W} L_p (w) \)

McAllester calls this ‘consistency’

[LaCroix-Julien unpublished result:]

\( L(w) \) is def.

\( \text{Sprodit} (\hat{w}_n) \overset{a.s.}{\to} L^* \implies L(w) \overset{a.s.}{\to} L^* = L(w) \)

proof idea: use Catoni’s PAC-Bayes bound

with prob \( \geq 1 - \delta_n \)

\[
\text{Sprodit} (\hat{w}_n) \leq \left( \frac{1}{2 - \delta_n} \right) \left[ \text{Sprodit} (\hat{w}_n) + \frac{\delta_n}{2} ||w^*||^2 + \frac{\ln \frac{1}{\delta_n}}{2 \delta_n} \right]
\]

\( \leq \text{Sprodit} (\hat{w}^*_n) + \frac{\delta_n}{2} ||w^*||^2 \)

\( \leq \text{Sprodit} (\hat{w}^*_n) + \sqrt{\ln \frac{1}{\delta_n}} \quad \text{Using Cauchy bound for} \ w^*_n \text{ with prob} \ \geq 1 - \frac{1}{n^2} \)

\( \implies \lim_{n \to \infty} \text{Sprodit} (\hat{w}_n) = \text{Sprodit} (\hat{w}^*_n) \quad \text{(with prob} \ 1) \)

\( \exists \ w \) also use \( \lim_{n \to \infty} \text{Sprodit} (\hat{w}_n) \leq L(w) \)

\( \implies \lim_{n \to \infty} \text{Sprodit} (\hat{w}_n) = L(w) \quad \text{[see paper for details]} \)

\( \text{Problem: } \text{Sprodit} (x, y, w) \text{ is non-convex in } w \implies \text{no optimization guarantees} \)

\( \text{Now: convex surrogates on score } \quad \text{def: } \quad s(y) \triangleq s(x, y, w) \text{ i.e. } z \hat{=} w \text{ is implicit} \)

\( \text{convex surrogates mentioned earlier} \)
Now: convex surrogates on score

\[ s(y) = \min_{y \neq y^*} s(y) \]

\[ \text{Score} \left( y \right) = \frac{1}{p} \log \left( \frac{1}{p} \exp(p s(y)) \right) - s(y) \left[ -\log p s(y) \right] \]

\[ \frac{1}{p} \log \left( \frac{1}{p} \exp(p s(y)) \right) \]

```
\text{Note: Slack rescaling more robust when have small } l(y, y^*) \text{ small.}
```

```
\text{but more computationally costly}
```

What theoretical properties could look at?

a) Generalization enforces [next class]

b) Consistency properties & calibration feat. [next class]

\[ \text{Relationship between } L(w) \text{ & } g(w) \]

Why structured score functions?

\[ s(z, y) = \sum_{C \in \mathcal{C}} \text{Sc}(z, y, C) \]

- Similarity to graphical models
- Statistical efficiency: less # of parameters (simpler score feat. Sc)
  \[ \rightarrow \text{easier to learn} \]
  \[ \text{(generalization guarantees)} \]
  \[ \text{(see Cortes et al. NeurIPS 2016)} \]
  \[ \text{next class} \]

2) Computational \[ \text{compute agmax } s(y) \]