Lecture 4:
MDL and PAC-Bayes
MDL and Universal Learning

- **Theorem:** For any prior $p(h)$, $\sum_h p(h) \leq 1$ (e.g. $p(h) = 2^{-|d(h)|}$) for a prefix-ambiguous $d(h)$, and any source distribution $\mathcal{D}$, if there exists $h^*$ with $L(h^*) = 0$, then w.p. $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$:

$$L(MDL_p(S)) \leq \sqrt{-\log p(h^*) + \log 2/\delta} = \sqrt{|d(h^*)| + \log 2/\delta}$$

- Sample complexity: $m = O\left(\frac{|d(h^*)|}{\epsilon^2}\right)$ (more careful analysis: $O\left(\frac{|d(h^*)|}{\epsilon}\right)$)

- Can learn any countable class!
  - Class of all computable functions, with $p(h) = 2^{-\min_{c(\sigma)=h} |\sigma|}$.
  - Class enumerable with $n: \mathcal{H} \to \mathbb{N}$ with $p(h) = 2^{-n(h)}$

- But $\text{VCdim}(\text{all computable functions}) = \infty$!

- Why no contradiction to Fundamental Theorem?
  - PAC Learning: Sample complexity $m(\epsilon, \delta)$ is uniform for all $h \in \mathcal{H}$. Depends only on class $\mathcal{H}$, NOT on specific $h^*$
  - MDL: Sample complexity $m(\epsilon, \delta, h)$ depends on $h$. 

Uniform and Non-Uniform Learnability

- **Definition**: A hypothesis class \( \mathcal{H} \) is **agnostically PAC-Learnable** if there exists a learning rule \( A \) such that \( \forall \epsilon, \delta > 0, \exists m(\epsilon, \delta), \forall \mathcal{D}, \forall h \in \mathcal{H}, \forall S \sim \mathcal{D}^m(\epsilon, \delta), \)
  \[ L^\mathcal{D}(A(S)) \leq L^\mathcal{D}(h) + \epsilon \]

- **Definition**: A hypothesis class \( \mathcal{H} \) is **non-uniformly learnable** if there exists a learning rule \( A \) such that \( \forall \epsilon, \delta > 0, \forall h \in \mathcal{H}, \exists m(\epsilon, \delta, h), \forall \mathcal{D}, \forall S \sim \mathcal{D}^m(\epsilon, \delta, h), \)
  \[ L^\mathcal{D}(A(S)) \leq L^\mathcal{D}(h) + \epsilon \]

- **Theorem**: \( \forall \mathcal{D} \), if there exists \( h \) with \( L(h) = 0 \), then \( \forall S \sim \mathcal{D}^m \)
  \[ L\left(\text{MDL}_p(S)\right) \leq \sqrt{-\log p(h) + \log 2/\delta} \]

  Compete also with \( h \) s.t. \( L(h) > 0 \)?
Allowing Errors: From MDL to SRM

\[ L(h) \leq L_S(h) + \sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}} \]

Minimized by ERM

Minimized by MDL

• Structural Risk Minimization:

\[ SRM_p(S) = \arg \min_h L_S(h) + \sqrt{\frac{-\log p(h)}{2m}} \]

fit the data

match the prior / simple / short description

• Theorem: For any prior \( p(h) \), \( \sum_h p(h) \leq 1 \), and any source distribution \( \mathcal{D} \), w.p. \( \geq 1 - \delta \) over \( S \sim \mathcal{D}^m \):

\[ L \left( SRM_p(S) \right) \leq \inf_h \left( L(h) + 2 \sqrt{\frac{-\log p(h) + \log 2/\delta}{m}} \right) \]
Non-Uniform Learning: Beyond Cardinality

- MDL still essentially based on cardinality ("how many hypothesis are simpler than me") and ignores relationship between predictors.

- Generalizes the cardinality bound: Using $p(h) = \frac{1}{|\mathcal{H}|}$ we get

$$m(\epsilon, \delta, h) = m(\epsilon, \delta) = \frac{\log|\mathcal{H}| + \log 2/\delta}{\epsilon^2}$$

- Can we treat continuous classes (e.g. linear predictors)? Move from cardinality to "growth function"?

- E.g.:
  - $\mathcal{H} = \left\{ \text{sign} \left( f(\phi(x)) \right) \mid f: \mathbb{R}^d \to \mathbb{R} \text{ is a polynomial} \right\}, \phi: \mathcal{X} \to \mathbb{R}^d$
  - $\text{VCdim}(\mathcal{H}) = \infty$
  - $\mathcal{H}$ is uncountable, and there is no distribution with $\forall h \in \mathcal{H} \ p(h) > 0$
  - But what if we bias toward lower order polynomials?

- **Answer 1: prior over hypothesis classes**
  - Write $\mathcal{H} = \bigcup \mathcal{H}_r$ (e.g. $\mathcal{H}_r =$degree-$r$ polynomials)
  - Use prior $p(H_r)$ over hypothesis classes
Prior Over Hypothesis Classes

- VC bound: \( \forall_r \mathbb{P} \left[ \forall_{h \in \mathcal{H}_r} L(h) \leq L_S(h) + O \left( \sqrt{\frac{\text{VCdim}(\mathcal{H}_r) + \log^1/\delta_r}{m}} \right) \right] \geq 1 - \delta_r \)

- Setting \( \delta_r = \rho(\mathcal{H}_r) \cdot \delta \) and taking a union bound,

\[
\forall_{S \sim \mathcal{D}^m} \forall_{\mathcal{H}_r} \forall_{h \in \mathcal{H}_r} L(h) \leq L_S(h) + O \left( \sqrt{\frac{\text{VCdim}(\mathcal{H}_r) - \log \rho(\mathcal{H}_r) + \log^1/\delta}{m}} \right)
\]

- Structural Risk Minimization over hypothesis classes:

\[
\text{SRM}_p(S) = \arg \min_{h \in \mathcal{H}_r} L_S(h) + C \sqrt{\frac{-\log \rho(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r)}{m}}
\]

- Theorem: w.p. \( \geq 1 - \delta \),

\[
L_D \left( \text{SRM}_p(S) \right) \leq \min_{\mathcal{H}_r, h \in \mathcal{H}_r} L_D(h) + O \left( \sqrt{\frac{-\log \rho(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r) + \log^1/\delta}{m}} \right)
\]
Structural Risk Minimization

• Theorem: For a prior \( p(\mathcal{H}_r) \) with \( \sum_{\mathcal{H}_r} p(\mathcal{H}_r) \leq 1 \) and any \( \mathcal{D}, \forall S \sim \mathcal{D}^m \),

\[
L_D \left( \text{SRM}_p(S) \right) \leq \min_{\mathcal{H}_r, h \in \mathcal{H}_r} L_D(h) + O \left( \sqrt{-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r) + \log \frac{1}{\delta}} \right)
\]

• For \( \mathcal{H}_i = \{h_i\} \):
  • \( \text{VCdim}(\mathcal{H}_r) = 0 \)
  • Reduces to “standard” SRM with a prior over hypothesis
Structural Risk Minimization

• Theorem: For a prior $p(\mathcal{H}_r)$ with $\sum_{\mathcal{H}_r} p(\mathcal{H}_r) \leq 1$ and any $\mathcal{D}$, $\forall S \sim \mathcal{D}$, $\forall S \sim \mathcal{D}$, $m$,

$$L_D \left( \text{SRM}_p(S) \right) \leq \min_{\mathcal{H}_r, h \in \mathcal{H}_r} L_D(h) + O \left( \sqrt{-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r) + \log \frac{1}{\delta}} \right)$$

• For $\mathcal{H}_i = \{h_i\}$:
  • $\text{VCdim}(\mathcal{H}_r) = 0$
  • Reduces to “standard” SRM with a prior over hypothesis

• For $p(\mathcal{H}_r) = 1$
  • Reduces to ERM over a finite-VC class
Structural Risk Minimization

• Theorem: For a prior \( p(\mathcal{H}_r) \) with \( \sum_{\mathcal{H}_r} p(\mathcal{H}_r) \leq 1 \) and any \( \mathcal{D}, \forall \mathcal{S} \sim \mathcal{D}^m \),

\[
L_{\mathcal{D}} \left( \text{SRM}_{p}(S) \right) \leq \min_{\mathcal{H}_r,h \in \mathcal{H}_r} L_{\mathcal{D}}(h) + O \left( \sqrt{-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r) + \log \frac{1}{\delta}} \right)
\]

• For \( \mathcal{H}_i = \{h_i\} \):
  • \( \text{VCdim}(\mathcal{H}_r) = 0 \)
  • Reduces to “standard” SRM with a prior over hypothesis
• For \( p(\mathcal{H}_r) = 1 \)
  • Reduces to ERM over a finite-VC class
• More general. Eg for polynomials over \( \phi(x) \in \mathbb{R}^d \) with \( p(\text{degree } r) = 2^{-r} \),

\[
m(\epsilon, \delta, h) = O \left( \frac{\text{degree}(h) + (d + 1)^{\text{degree}(h)} + \log \frac{1}{\delta}}{\epsilon^2} \right)
\]

• Allows non-uniform learning of a countable union of finite-VC classes
Uniform and Non-Uniform Learnability

• **Definition:** A hypothesis class \( \mathcal{H} \) is **agnostically PAC-Learnable** if there exists a learning rule \( A \) such that \( \forall \epsilon, \delta > 0, \exists m(\epsilon, \delta), \forall \mathcal{D}, \forall h, \forall_S S \sim_D m(\epsilon, \delta), \)
  \[ L_D(A(S)) \leq L_D(h) + \epsilon \]

• **Definition:** A hypothesis class \( \mathcal{H} \) is **non-uniformly learnable** if there exists a learning rule \( A \) such that \( \forall \epsilon, \delta > 0, \forall h, \exists m(\epsilon, \delta, h), \forall \mathcal{D}, \forall_S S \sim_D m(\epsilon, \delta, h), \)
  \[ L_D(A(S)) \leq L_D(h) + \epsilon \]

• **Theorem:** A hypothesis class \( \mathcal{H} \) is non-uniformly learnable **if and only if** it is a countable union of finite VC class \( (\mathcal{H} = \bigcup_{i \in \mathbb{N}} \mathcal{H}_i, \text{VCdim}(\mathcal{H}_i) < \infty) \)

• **Definition:** A hypothesis class \( \mathcal{H} \) is **“consistently learnable”** if there exists a learning rule \( A \) such that \( \forall \epsilon, \delta > 0, \forall h, \forall \mathcal{D}, \exists m(\epsilon, \delta, h, \mathcal{D}), \forall_S S \sim_D m(\epsilon, \delta, h, \mathcal{D}), \)
  \[ L_D(A(S)) \leq L_D(h) + \epsilon \]
Consistency

- \( \mathcal{X} \) countable (e.g. \( \mathcal{X} = \{0,1\}^* \), all strings or “sentences”)
- \( \mathcal{H} = \{\pm 1\}^\mathcal{X} \) (all possible functions over strings)
- \( \mathcal{H} \) is uncountable, it is not a countable union of finite VC classes, and is thus not non-uniformly learnable

- Claim: \( \mathcal{H} \) is “consistently learnable” using
  \[
  \text{ERM}_{\mathcal{H}}(S)(x) = \text{MAJORITY}(y_i \text{ s.t. } (x, y_i) \in S)
  \]

- Proof sketch: for any \( \mathcal{D} \),
  - Sort \( \mathcal{X} \) by decreasing probability. The tail has diminishing probability and thus for any \( \epsilon \), there exists some prefix \( \mathcal{X}' \) of the sort s.t. the tail \( \mathcal{X} \setminus \mathcal{X}' \) has probability mass \( \leq \epsilon/2 \).
  - We’ll give up on the tail. \( \mathcal{X}' \) is finite, and so \( \{\pm 1\}^{\mathcal{H}} \) is also finite.
- Why only “consistently learnable”? 
  - Size of \( \mathcal{X}' \) required to capture \( 1 - \epsilon/2 \) of mass depends on \( \mathcal{D} \).
Uniform and Non-Uniform Learnability

- **Definition:** A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall h$, $\forall \delta S_{\sim \mathcal{D}} m(\epsilon, \delta)$,
$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

- (Agnostically) PAC-Learnable iff $\text{VCdim}(\mathcal{H}) < \infty$

- **Definition:** A hypothesis class $\mathcal{H}$ is **non-uniformly learnable** if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\forall h$, $\exists m(\epsilon, \delta, h)$, $\forall \mathcal{D}$, $\forall \delta S_{\sim \mathcal{D}} m(\epsilon, \delta, h)$,
$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

- Non-uniformly learnable iff $\mathcal{H}$ is a countable union of finite VC classes

- **Definition:** A hypothesis class $\mathcal{H}$ is **“consistently learnable”** if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\forall h \forall \mathcal{D}$, $\exists m(\epsilon, \delta, h, \mathcal{D})$, $\forall \delta S_{\sim \mathcal{D}} m(\epsilon, \delta, h, \mathcal{D})$,
$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$
SRM In Practice

\[ SRM_p(S) = \arg \min_{h \in \mathcal{H}_r} L_S(h) + C \sqrt{\frac{-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r)}{m}} \]

- Bound is loose anyway. Better to view as bi-criteria optimization:
  \[ \arg \min L_S(h) \quad \text{and} \quad (-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r)) \]
  E.g. serialize as
  \[ \arg \min L_S(h) + \lambda (-\log p(\mathcal{H}_r) + \text{VCdim}(\mathcal{H}_r)) \]

- Typically \(-\log p(\mathcal{H}_r), \text{VCdim}(\mathcal{H}_r)\) monotone in “complexity” \(r\)
  \[ \arg \min L_S(h) \quad \text{and} \quad r(h) \]
  where
  \[ r(h) = \min r \quad \text{s.t.} \quad h \in \mathcal{H}_r. \]
SRM as a Bi-Criteria Problem

\[
\text{arg min } L_S(h) \text{ and } r(h)
\]

Regularization Path = \{ \text{arg min } L_S(h) + \lambda \cdot r(h) \mid 0 \leq \lambda \leq \infty \}

Select \( \lambda \) using a validation set—exact bound not needed
Non-Uniform Learning: Beyond Cardinality

• MDL still essentially based on cardinality ("how many hypothesis are simpler then me") and ignores relationship between predictors.

• Can we treat continuous classes (e.g. linear predictors)? Move from cardinality? Take into account that many predictors are similar?

• **Answer 1:** prior $p(\mathcal{H})$ over hypothesis class

• **Answer 2:** PAC-Bayes Theory
  • Prior distribution $P$ (not necessarily discrete) over $\mathcal{H}$
PAC-Bayes

• Until now (MDL, SRM) we used a discrete “prior” (discrete “distribution” \( p(h) \) over hypothesis, or discrete “distribution” \( p(\mathcal{H}_r) \) over hypothesis classes)

• Instead: encode inductive bias as distribution \( P \) over hypothesis

• Use randomized (averaged) predictor \( h_Q \), where for each prediction chooses \( h \sim Q \) and predicts \( h(x) \)
  • \( h_Q(x) = y \) w. p. \( P_{h \sim Q}(h(x) = y) \)
  • \( L_D(h_Q) = \mathbb{E}_{(x,y) \sim D} \mathbb{E}_{h \sim Q} [[h(x) \neq y]] = \mathbb{E}_{h \sim Q}[L_D(h)] \)

• **Theorem**: for any distribution \( P \) over hypothesis and any \( D, \forall_{S \sim D^m} \)
  \[
  |L_D(h_Q) - L_S(h_Q)| \leq \sqrt{KL(Q||P) + \log \frac{2m}{\delta}} \frac{1}{2(m - 1)}
  \]
### KL-Divergence

\[ KL(Q||P) = \mathbb{E}_{h \sim Q} \left[ \log \frac{dQ}{dP} \right] \]

\[ = \sum_h q(h) \log \frac{q(h)}{p(h)} \quad \text{for discrete dist with pmf } p, q \]

\[ = \int f_Q(h) \log \frac{f_Q(h)}{f_P(h)} dh \quad \text{for continuous distributions} \]

- Measures how much \( Q \) deviates from \( P \)
- \( KL(Q||P) \geq 0 \), and \( KL(Q||P) = 0 \) if and only if \( Q = P \)
- If \( Q(A) > 0 \) while \( P(A) = 0 \), \( KL(Q||P) = \infty \) (other direction is allowed)
- \( KL(H_1||H_0) \) = information per sample for rejecting \( H_0 \) when \( H_1 \) is true
- \( KL(Q||\text{Unif}(n)) = \log n - H(Q) \)
- \( I(X,Y) = KL(P(X,Y)||P(X)P(Y)) \)

Solomon Kullback

Richard Leibler
PAC-Bayes

• For any distribution $P$ over hypothesis and any $D$, $\forall_{\delta \sim D} m$

$$|L_D(h_Q) - L_S(h_Q)| \leq \sqrt{\frac{KL(Q||P) + \log 2m/\delta}{2(m-1)}}$$

• Can only use hypothesis in the support of $P$ (otherwise $KL(Q||P) = \infty$)

• For a finite $\mathcal{H}$ with $P = \text{Unif}(\mathcal{H})$
  • Consider $Q = \text{point mass on } h$
  • $KL(Q||P) = \log |\mathcal{H}|$
  • Generalizes cardinality bound (up to $\log m$)

• More generally, for a discrete $P$ and $Q = \text{point mass on } h$
  • $KL(Q||P) = \sum q(h) \log \frac{q(h)}{p(h)} = \frac{1}{p(h)}$
  • Generalizes MDL/SRM (up to $\log m$)

• For continuous $P$ (eg over linear predictors or polynomials)
  • For $Q=\text{point-mass}$ (or any discrete), $KL(Q||P) = \infty$
  • Take $h_Q$ as average over similar hypothesis (eg with same behavior on $S$)
PAC-Bayes

\[ L_D(h_Q) \leq L_S(h_Q) + \sqrt{\frac{KL(Q||P) + \log \frac{2m}{\delta}}{2(m-1)}} \]

- What learning rule does the PAC-Bayes bound suggest?

\[ Q_\lambda = \arg\min_Q L_S(h_Q) + \lambda \cdot KL(Q||P) \]

- **Theorem:**

\[ q_\lambda(h) \propto p(h)e^{-\beta L_S(h)} \]

for some "inverse temperature" \( \beta \)

- As \( \lambda \to \infty \) we ignore the data, corresponding to infinite temperature, \( \beta \to 0 \)

- As \( \lambda \to 0 \) we insist on minimizing \( L_S(h_Q) \), corresponding to zero temperature, \( \beta \to \infty \), and the prediction becomes ERM (or rather, a distribution over the ERM hypothesis in the support of \( P \))
PAC-Bayes vs Bayes

Bayesian approach:

- Assume $h \sim \mathcal{P}$,
- $x_1, \ldots, x_m$ iid from some $p(x)$
- $y_1, \ldots, y_m$ independent conditioned on $h$, with
  
  $$y_i | x_i, h = \begin{cases} 
  h(x_i), & \text{w. p. } 1 - \nu \\
  -h(x_i), & \text{w. p. } \nu
  \end{cases}$$

Bayesian posterior:

$$p(h|S) = \frac{p(h)p(S|h)}{p(S)} \propto p(h)p(S|h)$$

$$= p(h) \prod_i p(x_i)p(y_i|x_i)$$

$$\propto p(h) \prod_i \left(\frac{\nu}{1-\nu}\right)^{[h(x_i) \neq y_i]} = p(h) \left(\frac{\nu}{1-\nu}\right)^{\sum_i [h(x_i) \neq y_i]}$$

$$= p(h)e^{-\beta L_S(h)}$$

where $\beta = m \log \frac{1-\nu}{\nu}$