Rate-distortion Bounds on Bayes Risk in Supervised Learning

Matthew Nokleby, Wayne State University
Ahmad Beirami, MIT/Duke
Robert Calderbank, Duke University
Supervised Learning

- Set of labeled training data:

\[ X^n = (X_1, \ldots, X_n), \quad X_i \in \mathcal{X}, |\mathcal{X}| \leq |\mathcal{N}| \]

\[ Y^n = (Y_1, \ldots, Y_1), \quad Y_i \in \mathcal{Y} = \{1, \ldots, c\} \]

- Data drawn i.i.d. from unknown distribution:

\[ (X_i, Y_i) \sim p(x, y) \]

- Choose classifier \( h : \mathcal{X} \rightarrow \mathcal{Y} \) to optimize empirical risk:

\[ h^* = \arg\min_h \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)) \]

- When do we have enough training samples?
Generalization Error: PAC Bounds

• Probably Approximately Correct (PAC) framework:
  • restrict attention to a set of classifiers $\mathcal{H}$
  • choose the classifier that best fits the training set
• Bound performance via the VC dimension of $\mathcal{H}$

**Theorem [Vapnik]:** For any distribution, and with high probability over the training set, the classifier error probability is bounded by

$$\text{error probability} \leq \text{empirical error} + O \left( \sqrt{\frac{\text{VC dim}}{n}} \right)$$

• Leads to structural risk minimization (SRM):
  • Construct a hierarchy of families: $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$
  • Choose the family that minimizes the VC bound
Generalization Error: PAC Bounds

- PAC bounds are minimax over the data distribution
- Pessimistic in practice, especially when:
  - the training set is small(-ish)
  - there are more than two classes
- Practitioners usually ignore these bounds, resort to cross-validation

This talk:
- Framework for characterizing supervised learning performance for parametric distribution families
- Lower bounds on Bayes risk, averaged over distributions
- Has a rate-distortion interpretation
Bayes Risk over Parametric Families

• Distribution comes from a parametric family:
  \[(X_i, Y_i) \sim p(x, y|\theta), \ \theta \in \Theta \subset \mathbb{R}^b\]

• If distribution known, use MAP classifier
  \[h^*(x) = \arg \max_y p(y|x, \theta)\]

• Specific distribution unknown, learning machine knows parametric family as well as a prior over the family:
  \[\theta \sim p(\theta)\]
L₁ Bayes Risk

- Learning machine approximates MAP classifier from training
- Estimate the posterior distribution for every $x \in \mathcal{X}$

\[ W(y|x) \triangleq p(y|x, \theta) \]

\[ \hat{W}(y|x) = f(X^n, Y^n) \]

- Measure learning performance by the L₁ error:

\[ L_1(W, \hat{W}) = E_X \left[ \sum_{y=1}^{c} |W(y|X) - \hat{W}(y|X)| \right] \]

\[ W(y|X) - \hat{W}(y|X) \]

\[ \sum_{y=1}^{c} |\cdot| \]

\[ E_X[\cdot] \]
Rate-distortion Analysis

- Posterior function is a random “matrix” (Random source)
- Approximate to L1 risk tolerance (Distortion constraint)
- Using finitely many training samples (Finite-rate encoding)

**Lemma:** There exists a learning rule with Bayes risk less than $\epsilon$ only if
\[
I(X^n, Y^n; \theta) \geq \min_{p(\hat{W}|W)} \min_{E[L_1(W, \hat{W})]} I(W; \hat{W}) \\
\leq \epsilon
\]

- Can we find an expression of the form $h(W) - h(W - \hat{W})$?
- Challenge: Posterior is redundant under a parametric model
Interpolation Dimension

- A parametric model has limited degrees of freedom
- Posterior determined by finitely many entries

\[ S = \{ (x_1, y_1), (x_2, y_2), \ldots \} \]

**Definition:** A finite set \( S \) is an **interpolation set** for \( W(y|x) \) if:

1. \( h(W(S)) \) exists and is finite
2. \( h(W(T)) \) is negative infinity for any proper superset \( T \) of \( S \)
Interpolation Dimension

- A parametric model has limited degrees of freedom
- Posterior determined by finitely many entries

**Definition:** The cardinality of the smallest interpolation set $S$ is called the **interpolation dimension**, and is denoted by $d$.
Main Result: Lower Bounds on $L_1$ Risk

**Theorem:** Let $e(x,y) = E[|W(y|x) - \hat{W}(y|x)|]$. There exists a learning rule with $L_1(W,\hat{W}) < \epsilon$ only if

$$I(X^n, Y^n; \theta) \geq \sup_{S} \inf_{e(x,y)} h(W(S)) - \sum_{(x_i,y_i) \in S} (\log(1/e(x_i,y_i)) - 1)$$

**Proof “guts”:** Entropy of the error $W-\hat{W}$, subject to the average $L_1$ constraint, is maximized by a Laplace distribution.

Balances three quantities:

- **Mutual information:** bits provided by the training samples
- **Entropy of posterior:** bits needed to describe $W$
- **Bayes $L_1$ risk:** slack bits granted by the risk tolerance
Main Result: Lower Bounds on $L_1$ Risk

Corollary: Suppose the following conditions hold:

1. There exists a set of interpolation sets $S$ whose union is $X \times Y$
2. Each pair $(x,y)$ appears equally many times in the sets $S$
3. Each set $S$ has cardinality $d$

Then, there exists a learning rule with $L_1(W,\hat{W}) < \epsilon$ only if

$$I(X^n, Y^n; \theta) \geq \sup_S h(W(S)) - d \left( \log \left( \frac{c - 1}{\epsilon} \right) - 1 \right)$$

Proof: Conditions imply that constant error is optimum

- No need to perform max-min analysis
- Other expressions follow for slightly different conditions
Theorem: Let $I(\theta)$ be the Fisher information matrix of the data distribution $p(x,y;\theta)$. Under mild regularity conditions

$$I(X^n, Y^n; \theta) = \frac{1}{2} \log \left( \frac{n}{2\pi e} \right) + \frac{1}{2} E[\log |I(\theta)|] + h(\theta) + o(1)$$

Proof: Application of [Clarke & Barron, 1990] to $I(X^n, Y^n; \theta)$

Need to compute several quantities:

- Fisher information matrix (easy)
- Entropy over parametric family (easy)
- Entropy of interpolated posterior (hard)
Example: Discrete Distribution

- Trivial signal space: \( \mathcal{X} = \{0\} \)
- Equivalent to estimating \( c \)-dimensional discrete distribution

- Parametric family: all \( c \)-ary distributions
- Symmetric Dirichlet prior: \( p(\theta) = \text{Dir}(\alpha) \)

- Interpolation dimension: \( d = c-1 \)
- Can compute all the quantities in closed form:

\[
L_1(W, \hat{W}) \geq \sqrt{\frac{2\pi e}{n}} \frac{c}{2} \exp \left( \frac{c}{2(c-1)} (\psi(\alpha) - \psi(c\alpha)) - \frac{1}{2(c-1)} - 1 \right) (1 + o(1)) = O(\sqrt{c/n})
\]

\( \lambda \in \mathcal{X} \)

\( x \in \mathcal{X} \)
Example: Discrete Distribution

\[
L_1(W, \hat{W}) \geq \sqrt{\frac{2\pi e c}{n}} \cdot \frac{c}{2} \exp \left( \frac{c}{2(c-1)} \left( \psi(\alpha) - \psi(c\alpha) \right) - \frac{1}{2(c-1)} - 1 \right) (1 + o(1))
\]

\[= O(\sqrt{c/n})\]

- Compare to minimax bound of [Kamath et al., 2015]
- \(c = 100\)

\[
L_1(W, \hat{W}) \geq \sqrt{\frac{2\pi e c}{n}} \cdot \frac{c}{2} \exp \left( \frac{c}{2(c-1)} \left( \psi(\alpha) - \psi(c\alpha) \right) - \frac{1}{2(c-1)} - 1 \right) (1 + o(1))
\]

\[= O(\sqrt{c/n})\]

- Easily extends to arbitrary discrete classification problems
Example: Gaussian Classifier

- Gaussian classes with different means, Gaussian prior:
  \[ p(x|y) = \mathcal{N}(\theta_y, \sigma^2 I), \quad p(y) = 1/c \quad p(\theta) = \mathcal{N}(0, 1/m I) \]

- Posterior entropy difficult to compute analytically

- Compute bounds numerically for \( m=50, \sigma^2 = 0.1 \)

Figures 8-10, we evaluate the bounds numerically. We again choose \( d=50, \sigma^2 = 0.1 \), and we take \( M = \{5, 10, 15\} \). We plot the bounds and empirical Bayes risk and the empirical average of \( L(w, w_{MAP}) \), again over 2500 samples of \( \theta \), the training set, and the test point, and again using the MAP estimate of \( \theta \) from \( \mathbb{Z}_n \) to compute the empirical quantities. We do not plot a PAC bound because the VC dimension does not apply to non-binary cases.

We observe a few phenomena. First, the gap between the \( \ell_1 \) and \( \ell_2 \) risk grows with increasing \( M \), as predicted. Second, the bounds give rather accurate predictions of the classification error. Finally, the empirical \( \ell_1 \) and \( \ell_2 \) errors are closer together than predicted by theory, although the gap does grow in \( M \). This is because the estimation error of the posterior \( W(y|x; \theta) \) tends, empirically, to be "peaky" in \( y \), concentrating around a single value, leading to similar errors regardless or norm. The bounds on \( \tilde{R}_1 \) and \( \tilde{R}_1 \), on the other hand, optimistically suppose the error is evenly spread among the classes, in which case the norms differ substantially. The extent to which this result generalizes beyond this scenario is a topic for further investigation.
Conclusions

Summary:
• Rate-distortion framework for bounding learning performance
• Pros: predictive for multiple classes, small training sets
• Cons: difficult to evaluate bounds analytically

Future work:
• Evaluate performance of complex data models (DNNs/RBMs)
• Numerical/approximation techniques for entropy computation
• “Structural” Bayes risk minimization