Using Region Tests to Evaluate PAC Bounds

Thomas Walker

Imperial College London

thomas.walker21@imperial.ac.uk

August 29, 2023

Notation and Definitions

- Feature space $\mathcal{X}$, a label space $\mathcal{Y}$ to form data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ on which unknown distribution $\mathcal{D}$ is defined.
- Training data $S = \{(x_i, y_i)\}_{i=1}^{m} \overset{\text{i.i.d}}{\sim} \mathcal{D}^m$.
- Parameter space $\mathcal{W}$ indexing a hypothesis set $\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$.
    - The $h_{\mathbf{w}}$ are neural networks, with $\mathbf{w}$ being a vector of weights and biases.
- Loss function, $l : \mathcal{Y} \times \mathcal{Y} \to [0, C]$ quantifies performance of a hypothesis.

Notations and Definitions

Definition

The risk of a hypothesis is $R(\mathbf{w}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left(l(h(x), y)\right)$ and its empirical risk is $\hat{R}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} l(h_{\mathbf{w}}(x_i), y_i)$.

Note that $\mathbb{E}_{S\sim\mathcal{D}^m}\left(\hat{R}(\mathbf{w})\right) = R(\mathbf{w})$.

Remarks

- We don't know $R(\mathbf{w})$.
- We train for low $\hat{R}(\mathbf{w})$.
- The generalization gap is $R(\mathbf{w}) - \hat{R}(\mathbf{w})$.

Goal

Bound the generalization gap with high probability.

Bounds[1]

Uniform Convergence Bounds

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left(\sup_{\mathbf{w}\in\mathcal{W}}\left|R(\mathbf{w})-\hat{R}(\mathbf{w})\right|\leq\epsilon\left(\frac{1}{\delta},\frac{1}{m},\mathcal{W}\right)\right)\geq 1-\delta.$$

Algorithmic-Dependent Bounds

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left(\left|R\left(A(S)\right)-\hat{R}\left(A(S)\right)\right|\leq\epsilon\left(\frac{1}{\delta},\frac{1}{m},A\right)\right)\geq 1-\delta.$$

With equivalent expectation bounds.

[1] Viallard, Germain, Habrard, and Morvant 2021.

Assumption

Assumption

For a parameter $\mathbf{w}$ we can guarantee that $h_{\mathbf{w}}$ performs as expected on a region $\Delta \subset \mathcal{Z}$.

- For the 0-1 error this means $l_{\Delta}(\mathbf{w}) = 0$.

Questions

- How can we leverage this information to update our PAC bounds?
- How do these updates compare to increasing the size of the training data?

Leveraging the Assumption

We obtain information about the shape of $\mathcal{D}$ in the region Δ. Suppose we have a value for

$$p_\Delta = \mathbb{P}_{z \sim \mathcal{D}}(z \in \Delta) = \int_{z \in \Delta} \mathcal{D}(z) dz.$$

There are two potential improvements we can make to a PAC bound.

1. Tighten the bound, or
2. Improve the confidence with which the bound holds.

PAC Bound[2]

Theorem (PAC-Bound)

For a fixed $\mathbf{w} \in \mathcal{W}$, *let* $\delta \in (0,1)$ *then it follows that*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + C\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}\right) \geq 1-\delta.$$

Approach

1. Rework the proof of the theorem with our added assumption.
2. Condition the probability with our added assumption.

[2] Alquier 2023.

Improving Bounds

Theorem

For $\mathbf{w} \in \mathcal{W}$ *and* $\delta \in (0,1)$ *we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(R(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + CB(m, p_\Delta, \delta) \middle| l_\Delta(\mathbf{w}) = 0 \right) \geq 1 - \delta$$

for

$$B(m, p_\Delta, \delta) = \sqrt{\frac{\log\left(\frac{(1-p_\Delta)+\sqrt{(1-p_\Delta)^2+4\delta^{\frac{1}{m}} p_\Delta}}{2\delta^{\frac{1}{m}}} \right)}{2}}.$$

Remark

- With $p_\Delta = 0$ we recover Theorem PAC-Bound.
- With $p_\Delta = 1$ we note that $B(m, p_\Delta, \delta) > 0$.

Improving Confidence

Theorem

For $\mathbf{w} \in \mathcal{W}$ *and* $\delta \in (0,1)$ *we have that*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + C\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}\middle| l_\Delta(\mathbf{w}) = 0\right) \geq 1 - \left(\sum_{k=1}^{m}\binom{m}{k}\delta_k p_\Delta^{m-k}(1-p_\Delta)^k\right)$$

where

$$\delta_k = \frac{1}{\left(\frac{1}{\delta}\right)^{\frac{m^2}{k^2}}}.$$

Remark

- With $p_\Delta = 0$ we recover Theorem PAC-Bound.
- With $p_\Delta = 1$ we get full confidence in our bound.

PAC-Bayes Framework

Bayesian Machine Learning

1. A prior distribution π is defined on the parameter space.
2. A learning algorithm forms the updated posterior distribution ρ from the training data.
3. Infer a parameter from the posterior distribution to define a learned network.

Added Assumption

A subset of the parameter space, $\Omega \subset \mathcal{W}$, such that for $\mathbf{w} \in \Omega$ we have that $l_\Delta(\Omega) = 0$.

Conditioned PAC-Bayes Bound

Theorem

For all $\lambda > 0$, for all $\rho \in \mathcal{M}(\mathcal{W})$ and $\delta \in (0,1)$, conditioned on the fact that $l_\Delta(\Omega)$

$$R(\rho) \leq \hat{R}(\rho) + \frac{\log\left(B(\lambda, m, p_\Delta, p_\Omega)\right) + \mathrm{KL}(\rho, \pi) + \log\left(\frac{1}{\delta}\right)}{\lambda},$$

holds with probability greater than $1 - \delta$ over sampled training sets S where

$$B(\lambda, m, p_\Delta, p_\Omega) = p_\Omega\left(p_\Delta + (1 - p_\Delta)\exp\left(\frac{\lambda^2 C^2}{8m^2}\right)\right)^m + (1 - p_\Omega)\exp\left(\frac{\lambda^2 C^2}{8m}\right).$$

The original theorem was taken from Catoni 2009.

Approximating p_Δ

Using an independent random sample S_A we can form a confidence interval for p_Δ.

1. Let Z_i be random variable that $z_i \in S_A$ is in Δ.
   1.1 $Z_i \sim \mathrm{Bern}(p_\Delta)$.
2. Define the estimator $\hat{p}_\Delta$.
3. Construct $1-\alpha$ one-sided Clopper-Pearson (exact) confidence interval

$$[q_B(\alpha, m_A\hat{p}_\Delta, m_A - m_A\hat{p}_\Delta + 1), 1].$$

Update our result accordingly

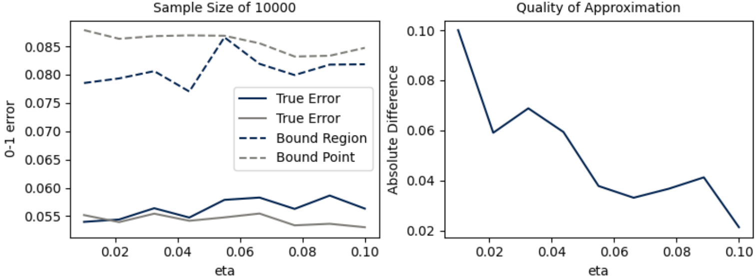
$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big(R(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + B\big(q_B(\alpha, m_A\hat{p}_\Delta, m_A - m_A\hat{p}_\Delta + 1)\big)\Big) \geq 1 - (\delta + \alpha(1-\delta)).$$

Experiment Details

- Define discrete underlying distribution.
- Sample m points randomly.
  - $m_A = \eta m$ points to approximate p_Δ,
  - $m_E = \zeta(1-\eta)m$ points to determine empirical error, and
  - $m_T = (1-\zeta)(1-\eta)m$ points to train the network.

1. Train with cross-entropy loss.
2. Determine correctly classified points of the underlying distribution, $\mathcal{C}$.
3. Sample $\mathcal{C}$ to determine Δ.
4. Approximate Δ using the determined segment.
5. Evaluate empirical 0-1 error on the m_E points.
6. Evaluate bound.

Results


Sample Size of 10000
0-1 error
0.085
0.080
0.075
0.070
0.065
0.060
0.055
0.02
0.04
0.06
0.08
0.10
eta
True Error
True Error
Bound Region
Bound Point
Quality of Approximation
Absolute Difference
0.10
0.08
0.06
0.04
0.02
0.02
0.04
0.06
0.08
0.10
eta

Let

$$\mathcal{D}_\Delta(z) = \begin{cases} \frac{\mathcal{D}(z)}{p_\Delta} & z \in \Delta \\ 0 & \text{otherwise,} \end{cases} \quad \mathcal{D}_{\Delta'}(z) = \begin{cases} \frac{\mathcal{D}(z)}{1-p_\Delta} & z \in \Delta' \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$R(\mathbf{w}) = p_\Delta R_\Delta(\mathbf{w}) + (1 - p_\Delta) R_{\Delta'}(\mathbf{w}). \tag{1}$$

for

$$R_\Delta(\mathbf{w}) = \mathbb{E}_{z\sim\mathcal{D}_\Delta}(l_z(\mathbf{w})), \text{ and } R_{\Delta'}(\mathbf{w}) = \mathbb{E}_{z\sim\mathcal{D}_{\Delta'}}(l_z(\mathbf{w})).$$

Proposition

With notation as above we have that,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big((1 - p_\Delta)R_{\Delta'}(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + B(\delta, m) - p_\Delta R_\Delta(\mathbf{w})\Big) \geq 1 - \delta,$$

for all $\mathbf{w} \in \mathcal{W}$ and $\delta \in (0, 1)$.

Experiment Details

1. Obtain a sample of size m from our data space according to a discrete underlying distribution.
2. Partition the data set according to some parameter ξ.
   2.1 Use ξm data points to determine the region Δ.
      - $\eta\xi m$ points to approximate p_Δ.
      - $(1-\eta)\xi m$ points to train a network to determine the region Δ.

   2.2 $(1-\xi)m$ points to evaluate our bound.
      - $(1-\zeta)(1-\xi)m$ points to train the model.
      - $\zeta(1-\xi)m$ points to evaluate the empirical errors for the bound.

Results

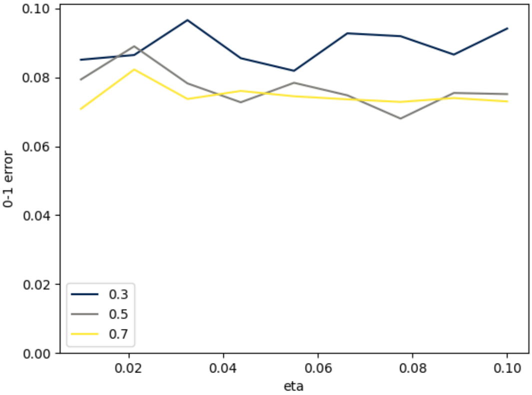


Figure: Plot of the value $\hat{R}(\mathbf{w}) + B(\delta, \zeta(1-\xi)m) - p_L R_\Delta(\mathbf{w})$ for $\zeta = 0.3$, and $\xi \in \{0.3, 0.5, 0.7\}$.

Summary

Conclusions

- Bounds can be updated not only by increasing training data size but also by using regional certificates of model performance.
- Updating bounds with this information can break the uniformity of results.
- Improvements in bounds through conditioning on regional certificates of neural network performance to are not significant.

Future Work

- Understand how this could work with other techniques for optimizing PAC bounds, such as data-informed priors, and compression bounds.
- Investigate whether informed sampling is effective.

References

Catoni, Olivier (Jan. 2009). "A PAC-Bayesian approach to adaptive classification". In.

Viallard, Paul, Pascal Germain, Amaury Habrard, and Emilie Morvant (2021). *A General Framework for the Disintegration of PAC-Bayesian Bounds*.

Alquier, Pierre (2023). *User-friendly introduction to PAC-Bayes bounds*.