Seeking information-theoretic bounds that explain generalization

Giuseppe Durisi

Chalmers, Sweden

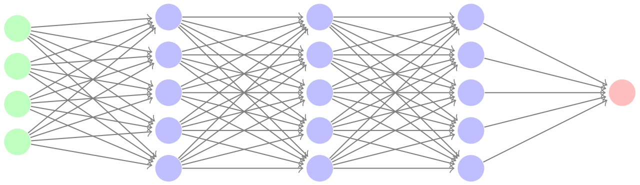
Information Theory and Tapas Workshop

Jan., 2023

CHALMERS

Joint work with Fredrik Hellström

Generalization performance of deep neural networks

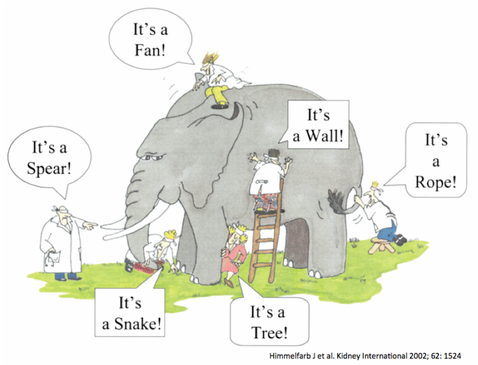


- State of the art in many fields

One of many mysteries

Why do DNN generalize despite being largely overparameterized?

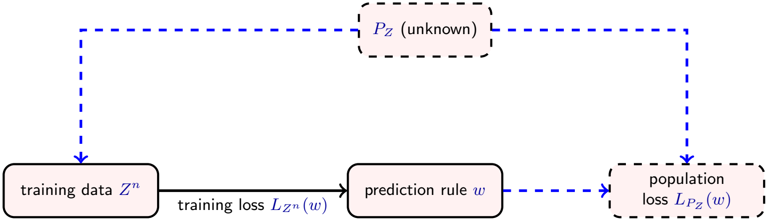
A complex problem that can be tackled from many angles. . .



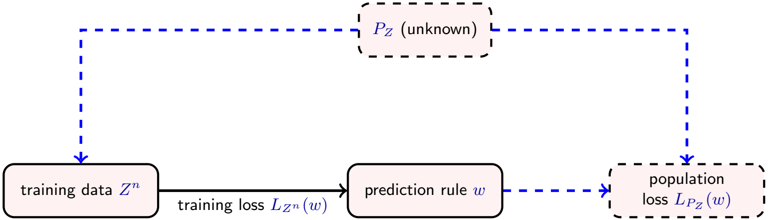
This talk

- Focus on information theoretic bounds
- Tutorial overview + recent results
- Numerically tight bounds but the question remains open

Supervised-learning setup

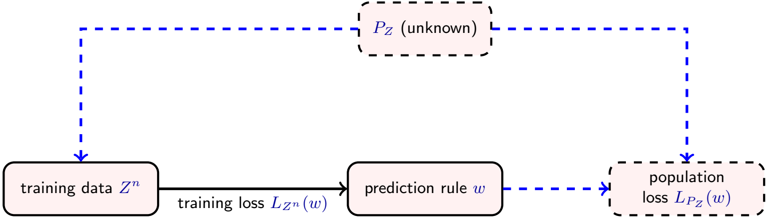


Supervised-learning setup



- $z = (x, y)$; x instance; y: label, $w(x)$: prediction; example: $x =$ 🚲, $y =$ bicycle, $w($🚲$) =$ car
- $\ell(\cdot, \cdot)$: nonnegative loss function; $\ell(w(x), y) \triangleq \ell(w; z)$
- $Z^n = [Z_1, \dots, Z_n]$: i.i.d. $\sim P_Z$ training data
- $L_{Z^n}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; Z_i)$: training loss; $L_{P_Z}(w) = \mathbb{E}_{P_Z}[\ell(w; Z)]$: population loss

Supervised-learning setup



- $z = (x, y)$; x instance; y: label, $w(x)$: prediction; example: $x =$ 🚲, $y =$ bicycle, $w($🚲$) =$ car
- $\ell(\cdot, \cdot)$: nonnegative loss function; $\ell(w(x), y) \triangleq \ell(w; z)$
- $Z^n = [Z_1, \ldots, Z_n]$: i.i.d. $\sim P_Z$ training data
- $L_{Z^n}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; Z_i)$: training loss; $L_{P_Z}(w) = \mathbb{E}_{P_Z}[\ell(w; Z)]$: population loss

Generalization problem: Under which conditions is $L_{P_Z}(w)$ close to $L_{Z^n}(w)$?

Probably approximately correct (PAC) learnability

- $\mathcal{W}$: set of prediction rules (hypothesis class)
- $c(\mathcal{W})$: "complexity" of $\mathcal{W}$

PAC bound [Vapnik & Chervonenkis, Valiant]

For all P_Z, with probability $1-\delta$ over the training set, we have that

$$L_{P_Z}(w) \leq L_{Z^n}(w) + \underbrace{\sqrt{\frac{c(\mathcal{W}) + \log 1/\delta}{2n}}}_{\text{penalty term}}$$

uniformly over the $w \in \mathcal{W}$

Probably approximately correct (PAC) learnability

- $\mathcal{W}$: set of prediction rules (hypothesis class)
- $c(\mathcal{W})$: "complexity" of $\mathcal{W}$

PAC bound [Vapnik & Chervonenkis, Valiant]

For all P_Z, with probability $1-\delta$ over the training set, we have that

$$L_{P_Z}(w) \leq L_{Z^n}(w) + \underbrace{\sqrt{\frac{c(\mathcal{W}) + \log 1/\delta}{2n}}}_{\text{penalty term}}$$

uniformly over the $w \in \mathcal{W}$

A vacuous bound

- CIFAR-10, convolutional neural network with $c(\mathcal{W}) \approx 10^7$
- Classification using 0–1 loss
- $n \approx 10^4$ suffices for good empirical performance but PAC bound is ≥ 1

Seeking nonvacuous bounds: the PAC-Bayes approach

PAC bounds for DNN

- Vacuous because the complexity term depends on the entire class $\mathcal{W}$
- Seek instead bounds with complexity term that depends on the prediction rule

Seeking nonvacuous bounds: the PAC-Bayes approach

PAC bounds for DNN

- Vacuous because the complexity term depends on the entire class $\mathcal{W}$
- Seek instead bounds with complexity term that depends on the prediction rule

PAC-Bayes approach

- Originally proposed in [McAllester, '98–'99 & Shawe-Taylor & Williamson, '98]
- Prediction rule modeled as Markov kernel (posterior) $P_{W \mid Z^n}$
- Prior Q_W is also available, used to embed *a priori* knowledge, or impose structure on prediction
- Objective: establish high-probability bounds on the average (over posterior) generalization gap

$$\mathbb{E}_{P_{W \mid Z^n}}[L_{P_Z}(W) - L_{Z^n}(W)]$$

- Available results scattered in many publication venues (outside IT)
- See [Alquier, arXiv 2021] for a recent primer on PAC-Bayes

Some PAC-Bayes bounds (bounded $\ell(\cdot,\cdot)$)

McAllester "square-root" bound [McAllester, 1999]

For a given Q_W the following bound holds with prob. $1-\delta$ w.r.t. P_{Z^n}

$$\mathbb{E}_{P_{W\,|\,Z^n}}[L_{P_Z}(W)] \le \mathbb{E}_{P_{W\,|\,Z^n}}[L_{Z^n}(W)] + \underbrace{\sqrt{\frac{1}{2(n-1)}\left[D(P_{W\,|\,Z^n}\,||\,Q_W) + \log\frac{\sqrt{n}}{\delta}\right]}}_{\text{penalty term}}$$

uniformly over all posterior distributions $P_{W\,|\,Z^n}$

Some PAC-Bayes bounds (bounded $\ell(\cdot,\cdot)$)

McAllester "square-root" bound [McAllester, 1999]

For a given Q_W the following bound holds with prob. $1-\delta$ w.r.t. P_{Z^n}

$$\mathbb{E}_{P_{W \mid Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W \mid Z^n}}[L_{Z^n}(W)] + \underbrace{\sqrt{\frac{1}{2(n-1)}\left[D(P_{W \mid Z^n} \,||\, Q_W) + \log\frac{\sqrt{n}}{\delta}\right]}}_{\text{penalty term}}$$

uniformly over all posterior distributions $P_{W \mid Z^n}$

Catoni "linear" bound [Catoni, 2007]

For a given Q_W and for a given $\beta > 0$, the following bound holds with prob. $1-\delta$ w.r.t. P_{Z^n}

$$\mathbb{E}_{P_{W \mid Z^n}}[L_{P_Z}(W)] \leq \frac{1}{1-e^{-\beta}}\left(\beta\, \mathbb{E}_{P_{W \mid Z^n}}[L_{Z^n}(W)] + \frac{D(P_{W \mid Z^n} \,||\, Q_W) + \log(1/\delta)}{n}\right)$$

uniformly over all posterior distributions $P_{W \mid Z^n}$

A 3-step proof template [Rivasplata et al., NeurIPS, 2020]

Step 1: concentration bound

- Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w), L_{Z^n}(w)\right)}\right] \leq \beta_n$$

where β_n does not depend on w

- Consequence:

$$\mathbb{E}_{Q_W}\left[\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right]\right] = \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}_{Q_W}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right]\right] \leq \beta_n$$

A 3-step proof template

Step 2: change of measure via Donsker-Varadhan

$$\log \mathbb{E}_{Q_W}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right] = \sup_{P_{W\,|\,Z^n}} \left\{\mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W)\right\}$$

Consequence: exponential inequality

$$\mathbb{E}_{P_{Z^n}}\left[e^{\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log \beta_n}\right] \leq 1$$

A 3-step proof template

Step 3: Chernoff bound

$$P_{Z^n}\left[\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n} \,||\, Q_W) - \log \beta_n > \log \frac{1}{\delta}\right] \le \delta$$

A 3-step proof template

Step 3: Chernoff bound

$$P_{Z^n}\left[\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log\beta_n > \log\frac{1}{\delta}\right] \le \delta$$

To conclude the proof

- Take complement
- Depending on the choice of $f(\cdot,\cdot)$, use Jensen's inequality

Examples of functions $f(\cdot,\cdot)$

McAllester "square-root" bound

$$\mathbb{E}_{P_{W\,|\,Z^n}}[L_{P_Z}(W)] \le \mathbb{E}_{P_{W\,|\,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{1}{2(n-1)}\left[D(P_{W\,|\,Z^n}\,||\,Q_W) + \log\frac{\sqrt{n}}{\delta}\right]}$$

Step 1: concentration bound

$$\mathbb{E}_{P_{Z^n}}\left[e^{2\frac{n-1}{n}\left(L_{P_Z}(w)-L_{Z^n}(w)\right)^2}\right] \le n$$

Catoni "linear" bound

$$\mathbb{E}_{P_{W\,|\,Z^n}}[L_{P_Z}(W)] \le \frac{1}{1-e^{-\beta}}\left(\beta\,\mathbb{E}_{P_{W\,|\,Z^n}}[L_{Z^n}(W)] + \frac{D(P_{W\,|\,Z^n}\,||\,Q_W)+\log(1/\delta)}{n}\right)$$

Step 1: concentration bound

$$\mathbb{E}_{Z^n}\left[e^{nd_\gamma\left(L_{P_Z}(w)\|L_{Z^n}(w)\right)}\right] \le 1, \text{ with } d_\gamma(p\|q) = \gamma p - \log(1-q+qe^{\gamma})$$

PAC-Bayes bounds and DNN

Catoni "linear" bound

For a given Q_W and for a given $\beta > 0$, the following bound holds with prob. $1-\delta$ w.r.t. P_{Z^n}

$$\mathbb{E}_{P_{W\,|\,Z^n}}[L_{P_Z}(W)] \leq \frac{1}{1-e^{-\beta}}\left(\beta\,\mathbb{E}_{P_{W\,|\,Z^n}}[L_{Z^n}(W)] + \frac{D(P_{W\,|\,Z^n}\,||\,Q_W) + \log(1/\delta)}{n}\right)$$

uniformly over all posterior distributions $P_{W\,|\,Z^n}$

- PAC-Bayes bounds can be optimized to find a good posterior $P_{W\,|\,Z^n}$
- Applied in many fields to obtain numerical certificates for randomized prediction rules
- DNN: Naïve application of PAC-Bayes yields vacuous bounds
- Solution: data-dependent prior

Data-dependent prior

- Split training data as $Z^n = [Z_{\mathrm{p}}^m, \quad Z_{\mathrm{t}}^{n-m}]$
- Let the prior depend on $Z_{\mathrm{p}}^m \Rightarrow$ data-dependent prior $Q_{W \mid Z_{\mathrm{p}}^m}$
- Use Z_{t}^{n-m} to evaluate the training error in the bound
- This approach yields some of the numerically tightest bounds known for randomized DNN

Catoni linear bound with data-dependent prior [Dziugaite et al., AISTATS, 2021]

For a given given $\beta > 0$, the following bound holds with prob. $1-\delta$ w.r.t. P_{Z^n}

$$\mathbb{E}_{P_{W \mid Z^n}}[L_{P_Z}(W)] \leq \frac{1}{1-e^{-\beta}}\left(\beta\, \mathbb{E}_{P_{W \mid Z^n}}\left[L_{Z_{\mathrm{t}}^{n-m}}(W)\right] + \frac{D(P_{W \mid Z^n} \,||\, Q_{W \mid Z_{\mathrm{p}}^m}) + \log(1/\delta)}{n-m}\right)$$

Proof: just modify step-1 in our proof template

Step 1: concentration bound

- Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z_{\mathrm{t}}^{n-m}}}\left[e^{f\left(L_{P_Z}(w), L_{Z_{\mathrm{t}}^{n-m}}(w)\right)}\right] \leq \beta_{n-m}$$

where β_{n-m} does not depend on w

- Consequence:

$$\mathbb{E}_{Q_{W|Z_{\mathrm{p}}^m} P_{Z_{\mathrm{p}}^m}}\left[\mathbb{E}_{P_{Z_{\mathrm{t}}^{n-m}}}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right]\right] = \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}_{Q_{W\,|\,Z_{\mathrm{p}}^m}}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right]\right] \leq \beta_n$$

Concluding the proof

Donsker-Varadhan to change measure from $Q_{W|Z_{\mathrm{p}}^m}$ to $P_{W\,|\,Z^n}$ and the proceed as before

Generalization bounds in the information-theory literature

- [T. Zhang, IT, 2006]: exponential inequalities, optimization of posterior distribution
- [Xu & Raginsky, NeurIPS, 2017]: average (rather than high-probability) generalization bound

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{1}{2n} I(W;Z^n)}$$

- Observation: $I(W;Z^n) = D(P_{W\,|\,Z^n} \,||\, P_W \,|\, P_{Z^n}) \leq D(P_{W\,|\,Z^n} \,||\, Q_W \,|\, P_{Z^n})$
- P_W: oracle prior

Almost identical 3-step proof template

Step 1: Concentration of measure (unchanged)

- Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w),L_{Z^n}(w)\right)}\right] \leq \beta_n$$

  where β_n does not depend on w

- Consequence:

$$\mathbb{E}_{Q_W}\left[\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(W),L_{Z^n}(W)\right)}\right]\right] = \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}_{Q_W}\left[e^{f\left(L_{P_Z}(W),L_{Z^n}(W)\right)}\right]\right] \leq \beta_n$$

An almost identical 3-step proof template

Step 2: change of measure via Donsker-Varadhan (unchanged)

$$\log \mathbb{E}_{Q_W}\left[e^{f\left(L_{P_Z}(W), L_{Z^n}(W)\right)}\right] = \sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W)$$

Consequence: exponential inequality

$$\mathbb{E}_{P_{Z^n}}\left[e^{\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log \beta_n}\right] \leq 1$$

An almost identical 3-step proof template

Step 3: Jensen's inequality (instead of Chernoff)

$$e^{\mathbb{E}_{P_{Z^n}}\left[\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log \beta_n\right]} \leq 1$$

An almost identical 3-step proof template

Step 3: Jensen's inequality (instead of Chernoff)

$$e^{\mathbb{E}_{P_{Z^n}}\left[\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log\beta_n\right]} \leq 1$$

As a consequence

$$\mathbb{E}_{P_{W,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W\,|\,P_{Z^n}) - \log\beta_n \leq 0$$

Depending on the choice of $f(\cdot,\cdot)$, use Jensen's inequality.

An almost identical 3-step proof template

Step 3: Jensen's inequality (instead of Chernoff)

$$e^{\mathbb{E}_{P_{Z^n}}\left[\sup_{P_{W\,|\,Z^n}} \mathbb{E}_{P_{W\,|\,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W) - \log\beta_n\right]} \leq 1$$

As a consequence

$$\mathbb{E}_{P_{W,Z^n}}\left[f\left(L_{P_Z}(W), L_{Z^n}(W)\right)\right] - D(P_{W\,|\,Z^n}\,||\,Q_W\,|\,P_{Z^n}) - \log\beta_n \leq 0$$

Depending on the choice of $f(\cdot,\cdot)$, use Jensen's inequality.

Choice of $f(\cdot,\cdot)$ in [Xu & Raginsky, NeurIPS, 2017]

$$f\left(L_{P_Z}(W), L_{Z^n}(W)\right) = \lambda\left(L_{P_Z}(W) - L_{Z^n}(W)\right)$$

Then optimization performed on λ

Implication

- We can leverage PAC-Bayes results to obtain a variety of average bounds
- Example: linear bound (a la Catoni)

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \frac{1}{1-e^{-\beta}}\left(\beta\, \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \frac{D(P_{W\,|\,Z^n}\,||\,Q_W\,|\,P_{Z^n})}{n}\right)$$

- But actually more can be done that has no correspondence in the PAC-Bayes literature

Samplewise bounds

Mutual information bound [Xu & Raginskiy, NeurIPS, 2017]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{1}{2n} I(W;Z^n)}$$

Samplewise bounds

Mutual information bound [Xu & Raginskiy, NeurIPS, 2017]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{1}{2n} I(W; Z^n)}$$

Individual-sample mutual information bound [Bu, Zou, Veeravalli, JSAIT, 2020]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2} I(W; Z_i)}}_{\leq \sqrt{\frac{1}{2n} I(W; Z^n)}}$$

Samplewise bounds

Mutual information bound [Xu & Raginskiy, NeurIPS, 2017]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{1}{2n} I(W;Z^n)}$$

Individual-sample mutual information bound [Bu, Zou, Veeravalli, JSAIT, 2020]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2} I(W;Z_i)}}_{\leq \sqrt{\frac{1}{2n} I(W;Z^n)}}$$

It tightens the MI bound and extends its applicability

The 3-step proof template still applies

Replace

Step 1: Concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w),L_{Z^n}(w)\right)}\right] \leq \beta_n$$

where β_n does not depend on w

The 3-step proof template still applies

Replace

Step 1: Concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w), L_{Z^n}(w)\right)}\right] \leq \beta_n$$

where β_n does not depend on w

with

Step 1b: samplewise concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that for all $i = 1, \dots, n$

$$\mathbb{E}_{P_{Z_i}}\left[e^{f\left(L_{P_Z}(w), \ell(w; Z_i)\right)}\right] \leq \beta$$

where β does not depend on w and i

The 3-step proof template still applies

Concluding the proof

- Step 2 and 3 result in

$$\mathbb{E}_{P_{W,Z_i}}\left[f\left(L_{P_Z}(W), \ell(W; Z_i)\right)\right] - D(P_{W\,|\,Z_i} \,||\, Q_W \,|\, P_{Z_i}) - \log \beta \leq 0$$

- Sum over i and use Jensen

The 3-step proof template still applies

Concluding the proof

- Step 2 and 3 result in

$$\mathbb{E}_{P_{W,Z_i}}\left[f\left(L_{P_Z}(W), \ell(W;Z_i)\right)\right] - D(P_{W\,|\,Z_i}\,||\,Q_W\,|\,P_{Z_i}) - \log\beta \leq 0$$

- Sum over i and use Jensen

Implication

- We can leverage PAC-Bayes results to obtain a variety of average, samplewise bounds
- On the contrary, PAC-Bayes samplewise bounds are generally vacuous [Harutyunyan, ITW, 2022]

Average bounds and conditional mutual information

Problem

- Average and PAC-Bayes bounds reviewed so far apply only to randomized prediction rules
- Easy to construct prediction rules with finite complexity in the PAC sense, but infinite $I(W; Z^n)$ or $D(P_{W \mid Z^n} \,||\, Q_W)$

The supersample approach [Steinke & Zakynthinou, COLT, 2020]

$$\underbrace{\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}}_{S^n} \Longrightarrow \underbrace{\begin{bmatrix} Z_1 & Z_{n+1} \\ Z_2 & Z_{n+2} \\ \vdots & \vdots \\ Z_n & Z_{2n} \end{bmatrix}}_{\mathbf{Z}} \Longrightarrow \underbrace{\begin{bmatrix} Z_1 \\ Z_{n+2} \\ \vdots \\ Z_n \end{bmatrix}}_{\mathbf{Z}(S^n)} \xrightarrow{P_{W \mid \mathbf{Z}(S^n)}} \boxed{W}$$

Conditional mutual information (CMI) bounds

[Steinke & Zakynthinou, COLT, 2020]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \sqrt{\frac{2}{n} I(W; S^n \,|\, \mathbf{Z})}$$

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq 2\,\mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \frac{3}{n} I(W; S^n \,|\, \mathbf{Z})$$

Advantages

- $I(W; S^n \,|\, \mathbf{Z})$ always bounded
- bounds applicable to fixed (deterministic) prediction rule

The 3-step proof template still applies (and tightens the bound)

Replace

Step 1: Concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w), L_{Z^n}(w)\right)}\right] \leq \beta_n$$

where β_n does not depend on w

The 3-step proof template still applies (and tightens the bound)

Replace

Step 1: Concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that

$$\mathbb{E}_{P_{Z^n}}\left[e^{f\left(L_{P_Z}(w),L_{Z^n}(w)\right)}\right] \le \beta_n$$

where β_n does not depend on w

with

Step 1c: Samplewise CMI concentration bound

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that for all $i = 1,\dots,n$

$$\mathbb{E}_{P_{S_i}}\left[e^{f\left(\ell\left(w;Z_{i,\bar{S}_i}\right),\ell\left(w;Z_{i,S_i}\right)\right)}\right] \le \beta$$

where β does not depend on w and i and $\mathbf{Z}$; then average w.r.t. $Q_{W\,|\,\mathbf{Z}}$

To conclude the proof

- Use Donsker-Varadhan to change the measure from $Q_{W\,|\,\mathbf{z}}$ to $P_{W\,|\,\mathbf{z},S_i}$ and apply Jensen
- Take expectation w.r.t to $\mathbf{Z}$
- Nonsamplewise concentration bound $+$ Chernoff $\Rightarrow$ PAC-Bayes CMI bounds

Examples of more general CMI bounds

Disintegrated, samplewise CMI bounds [Haghifam et al., NeurIPS, 2020]

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \le \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \mathbb{E}_{P_{\mathbf{Z}}}\left[\frac{1}{n}\sum_{i=1}^{n}\sqrt{2D(P_{W\,|\,\mathbf{Z},S_i}\,||\,Q_{W\,|\,\mathbf{Z}})}\right]$$

PAC-Bayes bounds for random subset setting [Hellström & Durisi, ICML-WS, 2021]

With probability at least $1-\delta$ with respect to $P_{\mathbf{Z},S^n}$,

$$\underbrace{\mathbb{E}_{P_{W\,|\,\mathbf{Z},S^n}}\left[L_{\mathbf{Z}(\bar{S}^n)}\right]}_{\text{text error}} \le \mathbb{E}_{P_{W\,|\,\mathbf{Z},S^n}}\left[L_{\mathbf{Z}(S^n)}\right] + \sqrt{\frac{2}{n-1}\left(D(P_{W\,|\,\mathbf{Z},S^n}\,||\,Q_{W\,|\,\mathbf{Z}}) + \log\frac{\sqrt{n}}{\delta}\right)}$$

$$\mathbb{E}_{P_{W\,|\,\mathbf{Z},S^n}}\left[L_{\mathbf{Z}(\bar{S}^n)}\right] \le 2\,\mathbb{E}_{P_{W\,|\,\mathbf{Z},S^n}}\left[L_{\mathbf{Z}(S^n)}\right] + \frac{3D(P_{W\,|\,\mathbf{Z},S^n}\,||\,Q_{W\,|\,\mathbf{Z}}) + \log(1/\delta)}{n}$$

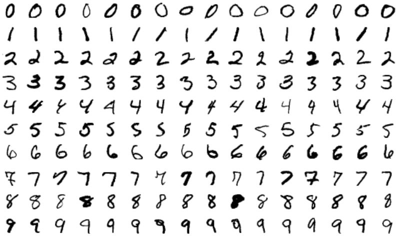
It gives automatically data-dependent prior; recovers state of the art bounds for randomized DNN

Numerical experiments for PAC-Bayes CMI bound

LeNet-5

Convolutional layer, 20 units, 5×5 size, linear activation, 1×1 stride, valid padding
Max pooling layer, 2×2 size, 2×2 stride
Convolutional layer, 50 units, 5×5 size, linear activation, 1×1 stride, valid padding
Max pooling layer, 2×2 size, 2×2 stride
Flattening layer
Fully connected layer, 500 units, ReLU activation
Fully connected layer, 10 units, $\mathrm{softmax}$ activation

MNIST dataset



Choice of posterior and prior distributions

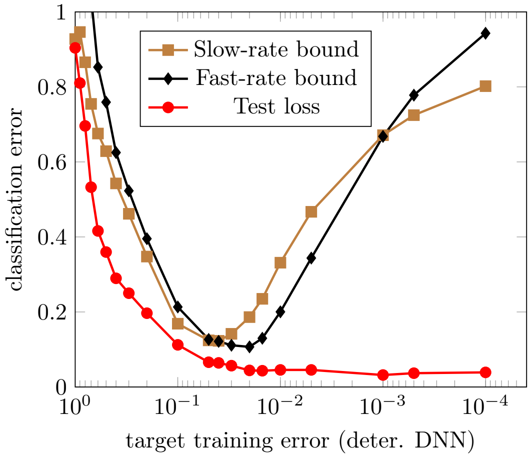
Posterior distribution $P_{W \mid \mathbf{Z}(S^n)}$

- Randomly generate S^n and determine $\mathbf{Z}(S^n)$
- Use SGD to find the weights μ_1 of the DNN
- Set posterior as $\mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})$, with σ_1^2 largest variance for which deterministic DNN has training error similar to stochastic DNN

Prior distribution $P_{W \mid \mathbf{Z}}$

- Evaluate (via Monte-Carlo) average μ_2 of the weight vectors of neural networks trained via SGD on $\mathbf{Z}(S^n)$ averaged over S^n
- Set prior as $\mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I})$ with σ_2^2 chosen as before

Classification error for SGD with momentum (random DNN)



- Slow-rate: square-root bound
- Fast-rate: linear bound
- The bounds are not vacuous
- Significant loss in accuracy for low training error (similar to [Dziugaite et al., AISTAT, 2021])

Evaluated conditional mutual information (eCMI) bounds

- The generalization performance depends on W indirectly through $\ell(W;Z)$
- Seek bounds where the information-theory metrics in the complexity term depend on $\ell(W;Z)$ rather than W
- First bounds of this kind appeared in [Steinke & Zakynthinou, COLT, 2020] and [Harutyunyan et al., NeurIPS, 2021] (fCMI)

General eCMI average and PAC-Bayes bounds

A family of both average, and PAC-Bayes eCMI bounds obtained using the 3-step proof template [Hellström, Durisi, NeurIPS, 2022]

Example: square-root, sample-wise, eCMI bound

$$\mathbb{E}_{P_{W,Z^n}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W,Z^n}}[L_{Z^n}(W)] + \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I\Big(\underbrace{\ell(W(\mathbf{Z}(S^n)); Z_{i1}), \ell(W(\mathbf{Z}(S^n)); Z_{i2})}_{\text{loss on train and test sample on } i\text{th row}}; S_i \,|\, \mathbf{Z}\Big)}$$

- Can be computed for deterministic DNN
- Can be evaluated efficiently for the case of 0-1 loss
- It requires the numerical estimation of a mutual information between Bernoulli random variables
- Expressiveness: can be used to recover classical PAC bounds

Key modification in proof template

Step 1c as in CMI, but with a different final averaging

Prove for a suitably chosen convex function $f(\cdot,\cdot)$ that for all $i = 1, \ldots, n$

$$\mathbb{E}_{P_{S_i}}\left[e^{f\left(\ell\left(w;Z_{i,\bar{S}_i}\right),\ell\left(w;Z_{i,S_i}\right)\right)}\right] \leq \beta$$

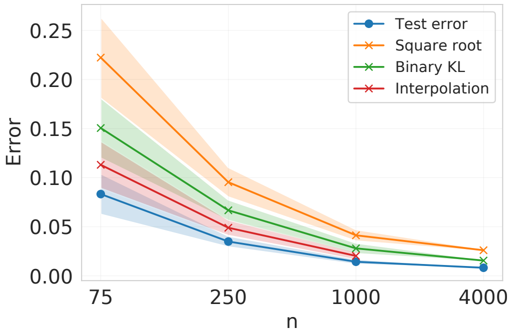
where β does not depend on w and i and $\mathbf{Z}$; then average w.r.t. $P_{\ell(W;Z_{i1}),\ell(W;Z_{i2})\,|\,\mathbf{Z}}$

Concluding the proof

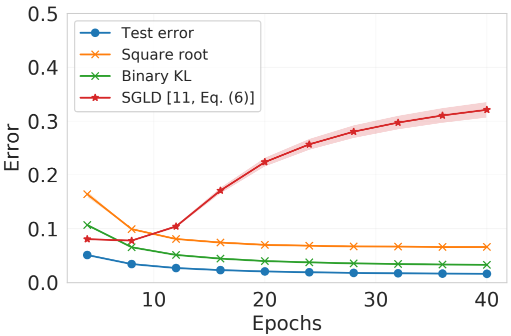
- Donsker-Varadhan to change measure from $P_{\ell(W;Z_{i1}),\ell(W;Z_{i2})\,|\,\mathbf{Z}}$ to $P_{\ell(W;Z_{i1}),\ell(W;Z_{i2})\,|\,S_i,\mathbf{Z}}$
- Then Jensen as usual

Numerical results, binarized version of MNIST

Deterministic DNN trained with SGD



(Randomized) DNN trained with SGLD



Conclusions

Take home message

Information-theoretic bounds that are numerically tight for neural networks and expressive enough to recover classical PAC bounds

We have not explained generalization (yet)

- Can we obtain tight bounds that can be evaluated analytically rather than numerically?
- Can the bound provide principled guidelines for DNN design and algorithm improvements?