Information Complexity and Generalization Bounds

Pradeep Kr. Banerjee
MPI MiS
pradeep@mis.mpg.de

Joint work with Guido Montúfar (UCLA and MPI MiS)

Foundations Reading Group Seminar
DeepMind, London
14th June, 2021

European Research Council
Established by the European Commission

Max-Planck-Institut für Mathematik in den Naturwissenschaften

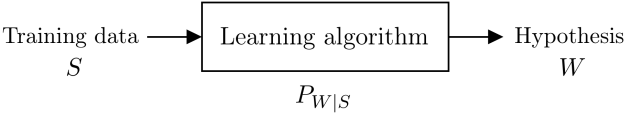
UCLA

One Bound to rule them all, One Bound to find them,
One Bound to bring them all, and in the darkness bind them.

– J.R.R. Tolkein (roughly)

Blum and Langford, 2003 "This quote is intended to describe the motivation for this line of work rather than our current state."

Formulation of the learning problem



- Ingredients
  - Example domain $\mathcal{Z}$
  - Hypothesis space $\mathcal{W}$
  - Loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$
- Learning algorithm $P_{W|S}$
  - Input: training data $S = (Z_1, \ldots, Z_n), \quad Z_i \overset{\text{i.i.d.}}{\sim} \mu$
  - Output: hypothesis $W \in \mathcal{W}$
- Population risk of a hypothesis $w \in \mathcal{W}$ w.r.t. μ

$$L_\mu(w) \triangleq \mathbb{E}_\mu[\ell(w, Z)]$$

- Goal: Output a hypothesis W based on S such that $L_\mu(W)$ is suitably small either in expectation or with high probability under any μ

Generalization error and mutual information

- Empirical risk $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$
- Population risk $L_\mu(w) = \mathbb{E}_{S' \sim \mu^{\otimes n}}[L_{S'}(w)]$, where $S' = (Z'_1, \ldots, Z'_n)$ is an i.i.d. sample
- **Objective**: Control the generalization error $\mathrm{g}(W, S) \triangleq L_\mu(W) - L_S(W)$, both in expectation and with high probability.
- Fitting-overfitting tradeoff

$$\mathbb{E}[L_\mu(W)] = \mathbb{E}[L_S(W)] + \mathbb{E}[L_\mu(W)] - \mathbb{E}[L_S(W)] = \mathbb{E}[L_S(W)] + \mathbb{E}[\mathrm{g}(W, S)]$$

- Expected generalization error

$$\mathbb{E}_{SW}[\mathrm{g}(W, S)] = \mathbb{E}_{P_S \otimes P_W}[L_S(W)] - \mathbb{E}_{P_{SW}}[L_S(W)], \qquad P_{SW} = \mu^{\otimes n} \otimes P_{W|S}$$

controlled by the mutual information $I(S; W)$ (Russo & Zou, 2016; Xu & Raginsky, 2017).

PAC-Bayesian inequalities

- Control the generalization error with high probability over a random draw of a sample S.
- McAllester (1999): Under bounded loss $\ell \in [0,1]$, for every $\delta \in (0,1)$, distribution μ on $\mathcal{Z}$, and fixed prior distribution Q over $\mathcal{W}$, we have for all posterior distributions $P \ll Q$ over $\mathcal{W}$, even such that depend on S,

$$\Pr_{S\sim\mu^{\otimes n}}\left(\mathbb{E}_P[\mathrm{g}(W,S)] \leq \sqrt{\frac{D(P\|Q) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}\right) \geq 1-\delta.$$

- For a fixed posterior P, $\mathbb{E}_S[D(P\|Q)]$ is minimized by the oracle prior,

$$Q^\star = \mathbb{E}_{S\sim\mu^{\otimes n}}[P_{W|S}(\cdot|S)].$$

- $\mathbb{E}_S[D(P\|Q^\star)] = I(S;W)$.
  - For any Q s.t. $D(P_W\|Q) < \infty$, $I(S;W) = D(P_{W|S}\|Q|P_S) - D(P_W\|Q)$, where $D(P_{W|S}\|Q|P_S) = \int_{\mathcal{Z}^n} D(P_{W|S=s}\|Q)\mu^{\otimes n}(\mathrm{d}s)$.

Outline

1. A whirlwind tour of information stability
2. One bound to rule'em all
3. PAC-Bayes-CMI and differentially private priors
4. Information complexity minimization and "flat" minima

A whirlwind tour of information stability

Generalization and stability

- Let $S \sim \mu^{\otimes n}$, $S' \sim \mu^{\otimes n}$ be two independent training samples
- Replace-one operation: Run $P_{W|S}$ after replacing Z_i with Z_i' for each $i \in [n]$

$$\left.\begin{aligned} S &= (Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_n) \xrightarrow{P_{W|S}} W \\ S^{(i)} &= (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n) \xrightarrow{P_{W|S}} W^{(i)} \end{aligned}\right\} \quad (W, S, Z_i') \stackrel{\mathsf{d}}{=} (W^{(i)}, S^{(i)}, Z_i)$$

- Population risk of $P_{W|S}$ is the empirical risk evaluated on a fresh independent sample S'

$$\mathbb{E}_{S,W}[L_\mu(W)] = \mathbb{E}_{S,S'}\mathbb{E}_W\left[\frac{1}{n}\sum_{i=1}^n \ell(W, Z_i')\right]$$

$$\mathbb{E}_{S,W}[L_S(W)] = \mathbb{E}_{S'}\mathbb{E}_{S,W}\left[\frac{1}{n}\sum_{i=1}^n \ell(W, Z_i)\right] = \mathbb{E}_{S,S'}\mathbb{E}_W\left[\frac{1}{n}\sum_{i=1}^n \ell(W^{(i)}, Z_i')\right]$$

- Expected generalization error measures stability of $P_{W|S}$ w.r.t. local perturbations in S

$$\Delta \triangleq \mathbb{E}_{S,W}[L_\mu(W) - L_S(W)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,S'}\mathbb{E}_W\left[\ell(W, Z_i') - \ell(W^{(i)}, Z_i')\right]$$

In expectation, generalization equals stability

- $P_{W|S}$ is stable on-average (w.r.t. to the replace-one operation) if

$$s_n(P_{W|S}) \triangleq \sup_{\mu} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{S,S'} \mathbb{E}_W \left[\ell(W, Z_i') - \ell(W^{(i)}, Z_i') \right] \right| \xrightarrow{n \to \infty} 0.$$

- $P_{W|S}$ generalizes on-average if

$$g_n(P_{W|S}) \triangleq \sup_{\mu} \left| \mathbb{E}_{S,W} [L_\mu(W) - L_S(W)] \right| \xrightarrow{n \to \infty} 0.$$

Lemma (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010)

For any learning algorithm $P_{W|S}$, $g_n(P_{W|S}) = s_n(P_{W|S})$. In particular, $P_{W|S}$ generalizes on-average if and only if it is stable on-average.

Distributional stability and differential privacy

Definition (Dwork and Roth, 2014)

For any $\epsilon > 0$, $P_{W|S}$ *is* ϵ*-differentially private if, for any two datasets* s, $s' \in \mathcal{Z}^n$ *with*

$$d_{\mathrm{H}}(s, s') \triangleq \sum_{i=1}^{n} \mathbb{1}_{\{z_i \neq z_i'\}} \leq 1,$$

and for any measurable set $\mathcal{O} \subseteq \mathcal{W}$,

$$P_{W|S=s}(\mathcal{O}) \leq e^{\epsilon} P_{W|S=s'}(\mathcal{O}).$$

Definition (Dwork et al., 2015)

Let X *and* Y *be random variables in arbitrary measurable spaces, and let* X' *be independent of* Y *and equal in distribution to* X*. For* $\alpha \geq 0$*, the* α*-approximate max-information* $I_{\infty}^{\alpha}(X; Y)$ *is the least value of* k *such that for all events* $\mathcal{O} \subseteq \mathcal{Z}^n \times \mathcal{W}$,

$$\Pr\left((X, Y) \in \mathcal{O}\right) \leq e^{k} \cdot \Pr\left((X', Y) \in \mathcal{O}\right) + \alpha.$$

Stability in max-information

- Max-information of an algorithm: $P_{W|S}$ has α-approximate max-information of k, denoted as $I^{\alpha}_{\infty,\mu}(P_{W|S}, n) \leq k$, if for every distribution μ over $\mathcal{Z}$, $I^{\alpha}_{\infty}(S;W) \leq k$.

Proposition

Let $S' \perp\!\!\!\perp W$ be an independent sample with the same distribution as S. If for some $\alpha \geq 0$, $I^{\alpha}_{\infty}(S;W) = k$, then for any event $\mathcal{O} \subseteq \mathcal{Z}^n \times \mathcal{W}$,

$$\Pr((S,W) \in \mathcal{O}) \leq e^k \cdot \Pr((S',W) \in \mathcal{O}) + \alpha.$$

Proposition (Dwork et al. 2015)

If $P_{W|S}$ is an ϵ-differentially private algorithm, then $I_{\infty,\mu}(P_{W|S}, n) \leq n\epsilon$, and

$$I^{\alpha}_{\infty,\mu}(P_{W|S}, n) \leq \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{2}{\alpha}}, \quad \textit{for any } \alpha > 0.$$

- Since $I_{\infty}(S;W) \geq I(S;W)$, stability in max-information $\implies$ stability in MI for any μ.

Comparing different notions of stability

Pure differential privacy

⇓

Stability in max-information

⇓

Stability in mutual information
for any μ

Different approaches to generalization

- Uniform convergence and VC dimension: Property of the hypothesis class

$$\mathbb{E}_{S\sim\mu^{\otimes n}}\left[\sup_{w\in\mathcal{W}}|L_\mu(w)-L_S(w)|\right]\leq\frac{\mathsf{C}}{\sqrt{n}}$$

  where C is some distribution-independent measure of complexity

- Distributional stability: Property of the algorithm
  - Differential privacy, TV-stability, KL-stability, Average leave-one-out KL-stability, etc.

- Uniform stability: Property of the loss and algorithm

- Mutual information stability: Property of the input and the algorithm
  - Limitation: Not sensitive to low-probability failures
    - e.g., compare sample complexities $\Omega\left(\frac{\mathsf{VCdim}(\mathcal{F})+\ln 1/\delta}{\varepsilon^2}\right)$ and $\Omega\left(\frac{I(S;W)}{\varepsilon^2\delta}\right)$

One bound to rule'em all

The information exponential inequality

For any $\beta > 0$, define the annealed expectation

$$M_\beta(w) = -\beta^{-1}\Lambda_{-\ell(w,Z)}(\beta) = -\beta^{-1}\ln \mathbb{E}_\mu[e^{-\beta\ell(w,Z)}].$$

Lemma (Zhang, 2006)

For any real-valued loss ℓ, fixed prior Q over $\mathcal{W}$, and any posterior distribution $P \ll Q$ over $\mathcal{W}$ that depends on an i.i.d. training sample S,

$$\mathbb{E}_S \exp\left\{n\beta\mathbb{E}_P\left[M_\beta(W) - L_S(W)\right] - D(P\|Q)\right\} \leq 1.$$

$M_\beta(w)$ acts as a surrogate for $L_\mu(w)$:

- By Jensen's inequality: $M_\beta(w) \leq L_\mu(w)$
- Bounds in the opposite direction under different assumptions on the loss
  - e.g., if $\ell(w, Z)$ is σ-sub-Gaussian under μ for every $w \in \mathcal{W}$, then $L_\mu(w) \leq M_\beta(w) + \frac{\beta}{2}\sigma^2$

Theorem

Suppose that there exist a convex function $\psi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisfying* $\psi(0) = \psi'(0) = 0$*, such that*

$$\sup_{w \in \mathcal{W}} \beta\Big(L_\mu(w) - M_\beta(w)\Big) \leq \psi(\beta), \quad \beta > 0.$$

Then, for any $\beta > 0$*,* $\delta \in (0,1)$*, and fixed prior distribution* Q *over* $\mathcal{W}$*,*

$$\Pr_{S \sim \mu^{\otimes n}} \left(\forall P \quad \mathbb{E}_P[g(W,S)] \leq \frac{1}{n\beta}\left[D(P\|Q) + \ln\frac{1}{\delta}\right] + \frac{\psi(\beta)}{\beta} \right) \geq 1 - \delta.$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \psi^{*-1}\left(\frac{D(P\|Q|P_S)}{n}\right),$$

where ψ^{*-1} *is the inverse of the Legendre dual of* ψ*.*

Properties of the cumulant generating function

Cumulant generating function of a random variable X:

$$\Lambda_X(\beta) = \ln \mathbb{E}[e^{\beta X}], \quad \beta > 0$$

Properties of $\Lambda_X(\beta)$ for $\beta > 0$:

- $\Lambda_X(\beta)$ is infinitely differentiable and convex in β
- $\frac{1}{\beta}\Lambda_X(\beta)$ is an increasing function of β
- $\mathbb{E}[X] \leq \frac{1}{\beta}\Lambda_X(\beta) \leq \Lambda'_X(\beta)$
- If $a \leq X \leq b$ a.s., then $a \leq \Lambda'_X \leq b$

Examples of Λ_X for concrete random variables:

- Bernoulli X: $\frac{1}{\beta}\Lambda_X(\beta) = \frac{1}{\beta}\ln\left(1 - (1 - e^{\beta})\mathbb{E}[X]\right)$
- σ-sub-Gaussian X: $\frac{1}{\beta}\Lambda_X(\beta) \leq \mathbb{E}[X] + \frac{\beta\sigma^2}{2}$

Legendre dual of a smooth convex function and its inverse

Lemma (Boucheron et al, 2013)

Let ψ *be a convex and continuously differentiable function defined on the interval* $[0, b)$*, where* $0 < b \leq \infty$*. Assume that* $\psi(0) = \psi'(0) = 0$*.*

Then, the Legendre dual of ψ*,*

$$\psi^*(t) \triangleq \sup_{\beta \in [0,b)} \{\beta t - \psi(\beta)\},$$

is a nonnegative convex and nondecreasing function on $[0, \infty)$ *with* $\psi^*(0) = 0$*.*

Moreover, its inverse $\psi^{*-1}(y) \triangleq \inf\{t \geq 0 : \psi^*(t) > y\}$ *is concave, and can be written as*

$$\psi^{*-1}(y) = \inf_{\beta \in (0,b)} \frac{y + \psi(\beta)}{\beta}.$$

Bounded mutual information implies generalization

- σ-sub-Gaussian loss: $\psi(\beta) = \frac{\beta^2\sigma^2}{2}$ for $\beta > 0$, and $\psi^{*-1}(y) = \sqrt{2\sigma^2 y}$
- (σ, c)-sub-gamma loss: $\psi(\beta) = \frac{\beta^2\sigma^2}{2(1-c\beta)}$ for $\beta \in \left(0, \frac{1}{c}\right)$, and $\psi^{*-1}(y) = \sqrt{2\sigma^2 y} + cy$

Corollary (Recovers Xu-Raginsky bound)

If $\ell(w, Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then

$$\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \sqrt{\frac{2\sigma^2}{n} I(S;W)}.$$

Corollary

If $\ell(w, Z)$ is (σ, c)-sub-gamma under μ for all $w \in \mathcal{W}$, then

$$\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \sqrt{\frac{2\sigma^2}{n} I(S;W)} + c\,\frac{I(S;W)}{n}.$$

The Gibbs algorithm and ERM

- Idea: Stabilize ERM by controlling the input-output mutual information $I(S; W)$.
- Xu-Raginsky: Given a prior Q over $\mathcal{W}$, the unique solution to the optimization problem

$$\arg\inf_{P_{W|S}} \left(\mathbb{E}[L_S(W)] + \frac{1}{\beta} D(P_{W|S} \| Q | P_S) \right)$$

  is the Gibbs algorithm, which satisfies

$$P^{\star}_{W|S=s}(\mathrm{d}w) = \frac{e^{-\beta L_s(w)} Q(\mathrm{d}w)}{\mathbb{E}_Q[e^{-\beta L_s(W')}]}, \qquad \text{for each } s \in \mathcal{Z}^n.$$

- In the zero temperature limit ($\beta \to \infty$), the Gibbs algorithm recovers ERM. For $\beta = 0$, the posterior reduces to the prior.
- When $\ell \in [0, 1]$, the Gibbs algorithm is $(2\beta/n)$-differentially private.

One-shot channel simulation and mutual information

- "Single-draw" bound (Xu & Raginsky, 2017; Bassily et al., 2018):

$$\text{Under bounded loss } \ell \in [0,1], \quad \Pr_{S,W}\left(|\mathrm{g}(W,S)| > \epsilon\right) = O\left(\frac{I(S;W)}{n\epsilon^2}\right).$$

- One-shot channel simulation (Harsha et al, 2010): Find the minimum amount of communication over a noiseless channel needed to simulate one use of $P_{W|S}$.
    - Alice and Bob has access to unlimited common randomness
    - Alice observes a sample $s \in \mathcal{Z}^n$ drawn according to P_S
    - Alice sends a message M to Bob via a noiseless channel

Q: What is the minimum $\mathbb{E}[L(M)]$ s.t. Bob can output a $w \in \mathcal{W}$ that is distributed according to $P_{W|S=s}$?

$$\text{A}: \quad \mathbb{E}[L(M)] \approx I(S;W)$$

Recovering classical PAC-Bayesian bounds

Corollary

[...] *with probability of at least* $1-\delta$ *over the choice of* $S \sim \mu^{\otimes n}$, *for all* $P \ll Q$ *over* $\mathcal{W}$:

- *The Catoni (2007) bound under* $\{0,1\}$*-valued loss*:

$$\mathbb{E}_P[L_\mu(W)] \leq \Phi_\beta^{-1}\left\{ \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta} D(P\|Q) + \frac{1}{n\beta}\ln\frac{1}{\delta} \right\}, \text{ where } \Phi_\beta^{-1}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}.$$

- *The McAllester (2013) "linear PAC-Bayes bound" under* $[0,1]$*-valued loss*:

$$\mathbb{E}_P[L_\mu(W)] \leq \frac{1}{1-\frac{\beta}{2}}\left[\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta} D(P\|Q) + \frac{1}{n\beta}\ln\frac{1}{\delta} \right], \quad \beta < 2.$$

- *The Germain et al. (2016) bound under* (σ, c)*-sub-gamma loss*:

$$\mathbb{E}_P[L_\mu(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta} D(P\|Q) + \frac{1}{n\beta}\ln\frac{1}{\delta} + \frac{\beta\sigma^2}{2(1-c\beta)}.$$

Recovering Catoni's bound

- [. . .] w.p. at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all $P \ll Q$ over $\mathcal{W}$

$$\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right).$$

- For $\ell \in \{0,1\}$,

$$M_\beta(w) = \Phi_\beta(L_\mu(w)) \triangleq -\beta^{-1}\ln\left(1-(1-e^{-\beta})L_\mu(w)\right), \quad \beta > 0.$$

- $\Phi_\beta : (0,1) \mapsto (0,1)$ is convex, increasing with inverse $\Phi_\beta^{-1}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}$

$$\mathbb{E}_P[L_\mu(W)] \leq \Phi_\beta^{-1}\left\{\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right)\right\}.$$

Compare with the more common PAC-Bayes derivation

- Since $Z_i \overset{\text{i.i.d.}}{\sim} \mu$, for any $w \in \mathcal{W}$ and $\beta > 0$,

$$e^{-n\beta M_\beta(w)} = \mathbb{E}_{S' \sim \mu^{\otimes n}} \left[e^{-n\beta L_{S'}(w)} \right]$$

$$\Pr_{S \sim \mu^{\otimes n}} \left(\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta} \left[D(P \| Q) + \ln \frac{1}{\delta} + \underbrace{\ln \mathbb{E}_Q \mathbb{E}_{S' \sim \mu^{\otimes n}} e^{n\beta \left(M_\beta(W) - L_{S'}(W) \right)}}_{= 0} \right] \right) \geq 1 - \delta$$

$$\Pr_{S \sim \mu^{\otimes n}} \left(\mathbb{E}_P[L_\mu(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta} \left[D(P \| Q) + \ln \frac{1}{\delta} + \ln \mathbb{E}_Q \mathbb{E}_{S' \sim \mu^{\otimes n}} e^{n\beta \left(L_\mu(W) - L_{S'}(W) \right)} \right] \right) \geq 1 - \delta$$

Optimizing β

Proposition

If $\ell(w, Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then for any constants $\alpha > 1$ and $v > 0$, with probability of at least $1 - \delta$,

$$\mathbb{E}_P[\mathrm{g}(W, S)] \leq \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln \frac{\log_\alpha \sqrt{n} + K}{\delta}\right) + \frac{\beta\sigma^2}{2}, \quad \forall \beta \in (0, v],$$

where $K = \max\left\{\log_\alpha\left(\frac{v\sigma}{\sqrt{2\alpha}}\right), 0\right\} + e$.

- Choice of β balances the first and second terms. Optimal order would be for $1/\sqrt{n}$.
- β *cannot* be optimized for "free". Overlooked in Hellström and Durisi, 2020a; 2020b.
- Under $[0, 1]$-valued loss, Maurer (2004) gave a version of the McAllester (2013) linear PAC-Bayes bound that is uniform in β at the cost of a $O\left(\frac{\ln \sqrt{n}}{n}\right)$ term.

PAC-Bayes-CMI and differentially private priors

The conditional mutual information (CMI) bound

Steinke & Zakynthinou (2020)

- Draw an i.i.d. "supersample" $\widetilde{Z} \in \mathcal{Z}^{2n}$

$$\begin{bmatrix} \widetilde{Z}_{1,0} & \widetilde{Z}_{2,0} & \cdots & \widetilde{Z}_{n,0} \\ \widetilde{Z}_{1,1} & \widetilde{Z}_{2,1} & \cdots & \widetilde{Z}_{n,1} \end{bmatrix} \begin{matrix} U_i = 0 \\ U_i = 1 \end{matrix} \qquad S \triangleq \widetilde{Z}_U = \left(\widetilde{Z}_{1,U_1}, \ldots, \widetilde{Z}_{n,U_n}\right)$$

- Randomly partition $\widetilde{Z}$ into input samples $S \triangleq \widetilde{Z}_U$ and "ghost" samples $G \triangleq \widetilde{Z}_{\overline{U}}$. "Selector" U (n uniform bits) specifies the partition independently of $\widetilde{Z}$ and the randomness of the algorithm. $\overline{U}$ is a vector obtained by inverting the bits of U.
- Run algorithm on input $S = \widetilde{Z}_U$ mapping it to a random element W of $\mathcal{W}$.
- After observing the output, how well can one distinguish the true inputs from their ghosts?

$$\mathsf{CMI}_\mu\left(P_{W|S}\right) \triangleq I(W;U|\widetilde{Z}) \qquad \mathsf{CMI}_\mu(P_{W|S}) \leq n \log 2$$

- When the loss is bounded in $[0,1]$, $\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \sqrt{\frac{2}{n} \cdot \mathsf{CMI}_\mu\left(P_{W|S}\right)}$.

Unconditional vs. conditional mutual information

“How much information does the output reveal
about the input?”

vs.

“How much information does the output reveal
about a randomly chosen subset of the supersample?”

Courtesy: Macaroni Penguin Art Print Smallwoods Studios

Sketch a suspect

vs.

Courtesy: speedbump.com

Recognize a suspect from a lineup

A hypothesis testing interpretation of CMI

- Suppose that we observe the output W and wish to identify S given access to $\widetilde{Z}$.
- For any estimator $\widehat{U} = \phi(W, \widetilde{Z})$ of U,

$$\inf_{\phi} \Pr\left(\phi(W, \widetilde{Z}) \neq U\right) \geq 1 - \frac{I(W; U|\widetilde{Z}) + \log 2}{n \log 2}.$$

- $I(W; U|\widetilde{Z})$ upper-bounds the probability of successfully identifying U from $\widehat{U}$.
- Mutual information decomposition and the CMI

$$W - \widetilde{Z}U - S \text{ and } W - S - \widetilde{Z}U \implies I(S; W) = I(\widetilde{Z}U; W) = I(W; \widetilde{Z}) + I(W; U|\widetilde{Z}).$$

PAC-Bayes-CMI bound

- Ghost sample $G \triangleq \widetilde{Z}_{\overline{U}}$ is independent of W

$$\mathrm{g}(W, \widetilde{Z}, U) \triangleq L_G(W) - L_S(W) \qquad \text{where} \begin{cases} L_G(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell\big(w, (\widetilde{Z}_{\overline{U}})_i\big) \\ L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell\big(w, (\widetilde{Z}_{U})_i\big) \end{cases}$$

- Prior $Q \equiv Q_{W|\widetilde{Z}=\widetilde{z}}$ and posterior $P \equiv P_{W|\widetilde{Z}=\widetilde{z}, U=u}$

Proposition

Under bounded loss $\ell \in [0,1]$, for every $\beta > 0$, $\delta \in (0,1)$,

$$\mathbb{E}_P[\mathrm{g}(W, \widetilde{Z}, U)] \leq \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\beta}{2}$$

with probability of at least $1-\delta$ over a draw of $\widetilde{Z}, U$. Moreover,

$$\mathbb{E}_{W,\widetilde{Z},U}[\mathrm{g}(W, \widetilde{Z}_U)] \leq \sqrt{\frac{2}{n}\cdot D(P\|Q|P_{\widetilde{Z},U})}.$$

Differentially private data-dependent priors

- A PAC-Bayes prior *cannot* depend on S but *can* depend on μ. However, our access to μ is only through S.
- Learn a prior using S in a differentially private fashion. Can then treat the prior "as if" it is independent of S.

Proposition

Let $\mathcal{K}(\mathcal{S}, \mathcal{W})$ denote the set of Markov kernels from $\mathcal{S}$ to $\mathcal{W}$. Let $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{W})$ be an ϵ-differentially private algorithm. Then with probability of at least $1 - \delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions P over $\mathcal{W}$,

$$\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P \| Q^0(S)) + \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}\right)$$

- The bound is valid for *any* loss and similar in spirit to a result by Dziugaite & Roy (2018), who gave a bound for the $[0,1]$-valued loss.

Information complexity minimization and “flat” minima

Information Complexity Minimization (ICM)

- Recipe: Given any prior Q, minimize the Information Complexity (IC) w.r.t. the posterior

$$\mathbb{E}_P[L_s(W)] + \frac{1}{\beta} D(P \| Q).$$

- The minimizing distribution is the Gibbs distribution $P^\star(w) \propto e^{-\beta L_s(w)} Q(w)$ and

$$\mathbb{E}_{P^\star}[L_s(W)] + \frac{1}{\beta} D(P^\star \| Q) = \underbrace{-\frac{1}{\beta} \ln \mathbb{E}_Q[e^{-\beta L_s(W)}]}_{\text{Optimal IC}}.$$

- Optimal IC and "flat" minima: For $Q = \mathcal{N}\big(w, (\beta\gamma)^{-1}\mathbb{I}_k\big)$,

$$-\frac{1}{\beta} \ln \int_{w' \in \mathbb{R}^k} e^{-\beta\left[L_s(w') + \frac{\gamma}{2}\|w - w'\|^2\right]} \mathrm{d}w'$$

measures the log-volume of low-loss parameter configurations around w.
- Entropy-SGD (Chaudhari et al., 2017): Minimize the Optimal IC w.r.t. Q.

PAC-Bayes-SGD

Langford & Caruana (2002), Dziugaite & Roy (2018)

- $\mathcal{G}$: Set of all Gaussian posteriors of the form $P = \mathcal{N}(w_P, \mathrm{diag}(\gamma))$.
- Prior $Q = \mathcal{N}(w_0, \lambda \mathbb{I}_k)$ centered at a non-trainable random initialization, w_0.

Proposition

Under bounded loss $\ell \in [0,1]$, for any $\delta, \delta' \in (0,1)$, fixed $\alpha > 1$, $c \in (0,1)$, $b \in \mathbb{N}$, and $m, n \in \mathbb{N}$, with probability of at least $1 - \delta - \delta'$ over a draw of $S \sim \mu^{\otimes n}$ and $W \sim P^{\otimes m}$,

$$\mathbb{E}_P[L_\mu(f_W)] \;\leq\; \inf_{P \in \mathcal{G},\ \beta > 1,\ \lambda \in (0,c)} \Phi_\beta^{-1} \left\{ \hat{L}_S(f_W) + \frac{\alpha}{n\beta} D(P \| Q) + R(\lambda, \beta; \delta, \delta') \right\},$$

where $R \triangleq \frac{\alpha}{n\beta} \left(\ln \left(\frac{\ln \alpha^2 \beta n}{\ln \alpha} \right)^2 + \ln \left(\frac{\pi^2 b^2}{6\delta} \left(\ln \frac{c}{\lambda} \right)^2 \right) \right) + \sqrt{\frac{1}{2m} \ln \frac{2}{\delta'}}$, *and* $\Phi_\beta^{-1}(x) = \frac{1 - e^{-\beta x}}{1 - e^{-\beta}}$.

- For large n, m, optimization is dominated by the IC term.

A PAC-Bayes bound using loss curvature information

Laplace approximation of the Gibbs posterior given a fixed prior $Q = \mathcal{N}(w_Q, \lambda^{-1}\mathbb{I}_k)$

- Quadratic approximation of loss around a local minimizer w_P,

$$\widetilde{L}_S(w) = \frac{1}{2}(w - w_P)^\top H(w - w_P), \quad H = \nabla^2 L_S(w)|_{w=w_P}$$

- Optimal posterior

$$P = \mathcal{N}(w_P, H_\lambda^{-1}), \text{ where } H_\lambda \triangleq (n\beta H + \lambda \mathbb{I}_k)\,.$$

- $\lambda > 0$ is sufficiently large so that H_λ is positive definite

Proposition

Let $\{\lambda_i\}_{i=1}^k$ be the eigenvalues of H_λ and suppose that $\lambda_i \geq \lambda > 0$ for all i. Then with probability of at least $1 - \delta$ over a draw of the sample S,

$$\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\ln\frac{1}{\delta} + \frac{1}{n\beta}\left(\frac{\lambda}{2}\|w_Q - w_P\|^2 + \frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\right).$$

Occam factor and flat minima

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\ln\frac{1}{\delta} + \frac{1}{n\beta}\left(\frac{\lambda}{2}\|w_Q - w_P\|^2 + \frac{1}{2}\sum_{i=1}^{k}\ln\frac{\lambda_i}{\lambda}\right).$$

- Negative of the log-ratio term

$$-\frac{1}{2}\sum_{i=1}^{k}\ln\frac{\lambda_i}{\lambda} = \ln\sqrt{\det\frac{\lambda}{H_\lambda}}$$

  is the logarithm of the Occam factor (Mackay, 1992).
- Occam factor: Fraction of the prior parameter space consistent with the training data.
- Log-Occam factor: Entropy of a Gaussian posterior with scaled covariance $\lambda(H_\lambda)^{-1}$.
  - Information we gain about the model's parameters after seeing the data
- Minimizing the bound w.r.t. the posterior leads to solutions with higher entropy and hence wider minima.

Conclusion and future work

Summary

- Unified treatment of PAC-Bayes and IT-based generalization bounds
- New bounds
  - PAC-Bayes-CMI bound
  - PAC-Bayes bound for data dependent priors and unbounded losses
  - PAC-Bayes bound motivated by an Occam factor argument in relation to flat minima
- Examples of ICM for learning with neural networks: Entropy- and PAC-Bayes- SGD

Future scope

- Bounds we studied embody the dictum "bounded information implies learning"
- Does learning imply bounded information? No!
  - Results due to Bassily et al. (2018); Nachum & Yehudayoff (2019) for IT-based framework
  - Result due to Livni & Moran (2020) in a similar vein for PAC-Bayesian framework
- Identify the common structural properties of these negative results

References I

Olivier Bousquet and André Elisseeff.
Stability and generalization.
Journal of Machine Learning Research, 2(Mar):499–526, 2002.

Avrim Blum and John Langford.
PAC-MDL bounds.
In *Learning theory and kernel machines*, pages 344–357. Springer, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.
Concentration inequalities: A nonasymptotic theory of independence.
Oxford University Press, 2013.

Pradeep Kr Banerjee and Guido Montúfar.
Information complexity and generalization bounds.
arXiv preprint arXiv:2105.01747, 2021.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff.
Learners that use little information.
In *International Conference on Algorithmic Learning Theory (ALT)*, pages 25–55, 2018.

Olivier Catoni.
PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, volume 56.
Institute of Mathematical Statistics, 2007.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina.
Entropy-SGD: Biasing gradient descent into wide valleys.
In *International Conference on Learning Representations*, 2017.

References II

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth.
Generalization in adaptive data analysis and holdout reuse.
In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

Cynthia Dwork and Aaron Roth.
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.

Gintare Karolina Dziugaite and Daniel M. Roy.
Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data.
In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy.
Data-dependent PAC-Bayes priors via differential privacy.
In *Advances in Neural Information Processing Systems*, pages 8430–8441, 2018.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien.
PAC-Bayesian theory meets Bayesian inference.
In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

Fredrik Hellström and Giuseppe Durisi.
Generalization bounds via information density and conditional information density.
IEEE Journal on Selected Areas in Information Theory, pages 824–839, 2020.

Fredrik Hellström and Giuseppe Durisi.
Generalization error bounds via mth central moments of the information density.
In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 2741–2746. IEEE, 2020.

References III

Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan.
The communication complexity of correlation.
IEEE Transactions on Information Theory, 56(1):438–449, 2009.

Roi Livni and Shay Moran.
A limitation of the PAC-Bayes framework.
In *Advances in Neural Information Processing Systems*, volume 33, 2020.

David J. C. MacKay.
A practical Bayesian framework for backpropagation networks.
Neural computation, 4(3):448–472, 1992.

Andreas Maurer.
A note on the PAC Bayesian theorem.
arXiv preprint cs/0411099, 2004.

David A. McAllester.
Some PAC-Bayesian theorems.
Machine Learning, 37(3):355–363, 1999.

David A. McAllester.
A PAC-Bayesian tutorial with a dropout bound.
arXiv preprint arXiv:1307.2118, 2013.

Ido Nachum and Amir Yehudayoff.
Average-case information complexity of learning.
In *International Conference on Algorithmic Learning Theory (ALT)*, pages 633–646, 2019.

References IV

Daniel Russo and James Zou.
Controlling bias in adaptive data analysis using information theory.
In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240, 2016.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan.
Learnability, stability and uniform convergence.
The Journal of Machine Learning Research, 11:2635–2670, 2010.

Thomas Steinke and Lydia Zakynthinou.
Reasoning about generalization via conditional mutual information.
In *Conference On Learning Theory*, pages 3437–3452, 2020.

Aolin Xu and Maxim Raginsky.
Information-theoretic analysis of generalization capability of learning algorithms.
In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

Tong Zhang.
Information-theoretic upper and lower bounds for statistical estimation.
IEEE Transactions on Information Theory, 52(4):1307–1321, 2006.