Data-dependent PAC-Bayes priors via differential privacy

Speaker: Yi-Shan Wu

Institute of Information Science

Academia Sinica

Taiwan

May 25, 2018

Notations

- observe $\mathcal{S} \sim \mathcal{D}^m$: m i.i.d. samples from $\mathcal{D}$
- parameter of NN : $\mathbf{w} \in \mathbb{R}^p$
- $L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(\mathbf{w}, z)\right]$
- $L_{\mathcal{D}}(\mathcal{Q}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[L_{\mathcal{D}}(\mathbf{w})\right] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(\mathbf{w}, z)\right]\right]$
- $\hat{L}_{\mathcal{S}}(\mathcal{Q}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\ell(\mathbf{w}, z_i)\right]$
- generalization error : $L_{\mathcal{D}}(\mathcal{Q}) - \hat{L}_{\mathcal{S}}(\mathcal{Q})$

PAC-Bayes bounds

Theorem 7 (PAC-Bayes; Maurer 2004, Thm. 5) *Under bounded loss $\ell \in [0,1]$, for every $\delta > 0$, $m \in \mathbb{N}$, distribution $\mathcal{D}$ on Z, and distribution P on $\mathbb{R}^p$,*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\Big((\forall Q)\ \mathrm{KL}(\hat{L}_S(Q)||L_{\mathcal{D}}(Q)) \leq \frac{\mathrm{KL}(Q||P) + \log 2\sqrt{m}}{m} + \frac{1}{m}\log\frac{1}{\delta}\Big) \geq 1 - \delta. \qquad (2)$$

General PAC-Bayes bound, where the prior $\mathcal{P}$ is chosen before observing samples.

Typically the KL term is the dominant quantity in the bound and analysis is constrained by the need to choose $\mathcal{Q}$ such that $KL(\mathcal{Q}\|\mathcal{P})$ is not large. If $\mathbf{w}^*$ receives low probability from $\mathcal{P}$, ($\mathcal{P}$ is chosen badly), the bound is obvious bad.

Prior results by Lever[2013]

Catoni [2007] started studying the problem on finding bounds for data-distribution-dependent prior. Lever [2013] studied the randomized classifier with $\mathcal{Q}(\mathcal{S}) \propto exp(-\tau \hat{L}_{\mathcal{S}})$, $\mathcal{P}(\mathcal{S}) \propto exp(-\tau L_{\mathcal{D}})$. They are able to bound $KL(\mathcal{Q}\|\mathcal{P})$ independent of $\mathcal{D}$, yielding

Theorem 1 (Lever, Laviolette, and Shawe-Taylor 2013) *For $S \in Z^m$, let $Q(S) = P_{\exp(-\tau \tilde{L}_S)}$ be a Gibbs posterior with respect to some measure P on $\mathbb{R}^p$ and some bounded (surrogate) risk $\tilde{L}_S$. Under bounded loss $\ell \in [0,1]$, for every $\delta > 0$, $m \in \mathbb{N}$, distribution $\mathcal{D}$ on Z,*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\mathrm{KL}(\hat{L}_S(Q(S))||L_{\mathcal{D}}(Q(S))) \leq \frac{1}{m}\left(\tau\sqrt{\frac{2}{m}\log\frac{2\sqrt{m}}{\delta}} + \frac{\tau^2}{2m} + \log\frac{2\sqrt{m}}{\delta}\right)\right) \geq 1-\delta. \quad (1)$$

However, the bound can be still loose, since it allows classifier overfits on some distribution.

data-dependent PAC-Bayes bound (Goal)

Trivial but bad approach to consider several $\mathcal{P}$s : by union bound. Fix a distribution $\mathcal{D}$ on $\mathcal{Z}$,

$$R(\epsilon) = \{\mathcal{S} \in \mathcal{Z}^m : (\exists \mathcal{Q}) KL(\hat{L}_{\mathcal{S}}(\mathcal{Q}) \| L_{\mathcal{D}}(\mathcal{Q})) > \epsilon\} - -\text{before}$$

$$R(\epsilon_p) = \{\mathcal{S} \in \mathcal{Z}^m : (\exists \mathcal{Q}) KL(\hat{L}_{\mathcal{S}}(\mathcal{Q}) \| L_{\mathcal{D}}(\mathcal{Q})) > \epsilon_p\} - -\text{now}$$

If we union M kinds of $\mathcal{P}$, then the probability becomes $1 - M\delta$.

Theorem 8 *Under bounded loss $\ell \in [0,1]$, for every $\delta > 0$, $m \in \mathbb{N}$, distribution $\mathcal{D}$ on Z, and ϵ-differentially private data-dependent prior $\mathscr{P} \colon Z^m \rightsquigarrow \mathcal{M}_1(\mathbb{R}^p)$,*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\Big((\forall Q)\ \mathrm{KL}(\hat{L}_S(Q) || L_{\mathcal{D}}(Q)) \leq \frac{\mathrm{KL}(Q || \mathscr{P}(S)) + \ln 2\sqrt{m}}{m} + \max\Big\{\frac{2}{m}\ln\frac{3}{\delta},\ 2\epsilon^2\Big\}\Big) \geq 1 - \delta.$$

differential privacy

Since now our goal is to obtain a "data-dependent prior $\mathcal{P}$, bad $\mathcal{S}$ may lead to bad prior $\mathcal{P}$, and bad prior lead to bad $R(\epsilon_p)$. However, If $\mathcal{P}_{\mathcal{S}}$ is generated by some "stable" mechanism, then maybe we don't have to union too much of them.

differential private algorithm

Definition 3 *A randomized algorithm* $\mathscr{A} : Z^m \rightsquigarrow T$ *is* (ϵ, δ)-differentially private *if, for all pairs* $S, S' \in Z^m$ *that differ at only one coordinate, and all measurable subsets* $B \subseteq T$*, we have* $\mathbb{P}\{\mathscr{A}(S) \in B\} \leq e^{\epsilon}\, \mathbb{P}\{\mathscr{A}(S') \in B\} + \delta$.

Under such intuition, we have the following theorems,

differential privacy

Theorem 6 (Dwork et al. 2015b, Thm. 11) *Let $m \in \mathbb{N}$, let $\mathscr{A} : Z^m \rightsquigarrow T$, let $\mathcal{D}$ be a distribution over Z, let $\beta \in (0,1)$, and, for each $t \in T$, fix a set $R(t) \subseteq Z^m$ such that $\mathbb{P}_{S\sim\mathcal{D}^m}\{S \in R(t)\} \leq \beta$. If $\mathscr{A}$ is ϵ-differentially private for $\epsilon \leq \sqrt{\ln(1/\beta)/(2m)}$, then $\mathbb{P}_{S\sim\mathcal{D}^m}\{S \in R(\mathscr{A}(S))\} \leq 3\sqrt{\beta}$.*

(for each $\mathcal{P} \in \mathcal{M}_1(\mathbb{R}^p)$, fix a set $R(\mathcal{P}) = \{\mathcal{S} \in \mathcal{Z}^m : \text{conditions}\}$ such that $\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\{\mathcal{S} \in R(\mathcal{P}) \leq \beta\}$)
That is, with a stable algorithm, the probability only enlarge to $3\sqrt{\beta}$.

Theorem 8 *Under bounded loss $\ell \in [0,1]$, for every $\delta > 0$, $m \in \mathbb{N}$, distribution $\mathcal{D}$ on Z, and ϵ-differentially private data-dependent prior $\mathscr{P} : Z^m \rightsquigarrow \mathcal{M}_1(\mathbb{R}^p)$,*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big((\forall Q)\ \mathrm{KL}(\hat{L}_S(Q)||L_{\mathcal{D}}(Q)) \leq \frac{\mathrm{KL}(Q||\mathscr{P}(S)) + \ln 2\sqrt{m}}{m} + \max\{\frac{2}{m}\ln\frac{3}{\delta},\ 2\epsilon^2\}\Big) \geq 1-\delta.$$

If we have probability $\geq 1-\beta$ in Theorem 7, then $\delta = 3\sqrt{\beta}$

Ideally

For convenience, choose $\mathcal{P}_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \Sigma)$, where $\mathbf{w}_0 \in \mathbb{R}^p$ is chosen by private mechanism and approximately minimzing $\hat{L}_{\mathcal{S}}$.
(Ideal prior : easy to calculate & data dependent)

Theorem 10 *Let $\tau > 0$ and assume loss is bounded $\ell \in [0,1]$. One sample from the Gibbs posterior $P_{\exp(-\tau \hat{L}_S)}$ is $\frac{2\tau}{m}$-differentially private.*

However,

- Gibbs distribution is intractable.
- We can only approximate it by producing a sequence of samples that converges in distribution to the target Gibbs distribution. Weak convergence, however, does not yield privacy guarantees.

Weak convergence suffice

Theorem 16 (Weak convergence suffices) *Let $\tau > 0$ and let $\Sigma \in \mathbb{R}^{p\times p}$ be positive definite. For every $\mathbf{w} \in \mathbb{R}^p$ and $S \in Z^m$, let $P_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \Sigma)$, $Q^S_{\mathbf{w}} = (P_{\mathbf{w}})_{\exp(-\tau \tilde{L}_S)}$, where $\tilde{L}_S(\cdot)$ is bounded (surrogate) risk. Then, for every $\epsilon' > 0$ and $\delta, \delta' \in (0,1)$, with probability at least $1-\delta-\delta'$ over $S \sim \mathcal{D}^m$ and a sequence $\mathbf{w}_1, \mathbf{w}_2, \ldots$ that, conditional on S, converges in distribution a.s. to an ϵ-differentially private mean $\mathbf{w}^*(S)$, there exists $N \in \mathbb{N}$, such that, for all $n > N$,*

$$\mathrm{KL}(\hat{L}_S(Q^S_{\mathbf{w}_n})||L_{\mathcal{D}}(Q^S_{\mathbf{w}_n})) \leq \frac{\mathrm{KL}(Q^S_{\mathbf{w}_n}||P_{\mathbf{w}_n}) + \ln 2\sqrt{m}}{m} + \max\{\frac{2}{m}\ln\frac{3}{\delta},\ \epsilon^2\} + \epsilon'.$$

Here, the sequence of $\mathbf{w}_i$ are obtained by SGLD that SGLD converges weakly to Gibbs distribution.

Some prior works

Lemma 15

Using bounded cross-entropy. Define $\mathcal{P}_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \Sigma)$ and let $G, G_1, G_2, \cdots \in \mathcal{M}_1(\mathbb{R}^p)$, where $G_n \to G$ weakly. Then $\forall \epsilon, \delta, \exists N \in \mathbb{N}$ such that $\forall n > N$, w.h.p. over $(\mathbf{w}_n, \mathbf{w}_n^*) \sim (G_n, G)$ that

$$KL(\hat{L}_{\mathcal{S}}(\mathcal{Q}_{\mathbf{w}_n}) \| L_{\mathcal{D}}(\mathcal{Q}_{\mathbf{w}_n})) \leq KL(\hat{L}_{\mathcal{S}}(\mathcal{Q}_{\mathbf{w}_n^*}) \| L_{\mathcal{D}}(\mathcal{Q}_{\mathbf{w}_n^*})) + \epsilon$$

$$KL(\mathcal{Q}_{\mathbf{w}_n^*} \| \mathcal{P}_{\mathbf{w}_n^*}) \leq KL(\mathcal{Q}_{\mathbf{w}_n} \| \mathcal{P}_{\mathbf{w}_n}) + \epsilon$$

1. Let $(\mathbf{w}_n, \mathbf{w}_n^*) \sim (G_n, G)$. There exists a sequence of random variables $\mathbf{w}^*, \mathbf{w}_1, \mathbf{w}_2, \cdots$ such that $\mathbf{w}_n \sim G_n$ for all $n \in \mathbb{N}$, $\mathbf{w}^* \sim G$ and $\mathbf{w}_n \sim \mathbf{w}^*$ a.s.
2. If $\mathcal{P}_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \Sigma)$, then $\mathcal{P}_{\mathbf{w}_n} \to \mathcal{P}$ and KL() have these inequalities