Information-theoretic analysis of generalization capability of learning algorithms

EE6143

K Nithin Varma

Indian Institute of Technology Madras

EE18B052

Learning: Data-Driven Stochastic Optimization

- Goal: stochastic optimization

$$\text{minimize} \quad L_\mu(w) := \mathbf{E}_\mu[\ell(w, Z)] = \int_{\mathrm{Z}} \ell(w, z)\mu(\mathrm{d}z) \tag{1}$$

  where:
  - w is an element of the hypothesis space W
  - Z is a random element of the instance space Z
  - $\mu := \mathcal{L}(Z)$ is unknown
  - $\ell : \mathrm{W} \times \mathrm{Z} \to \mathbb{R}_+$is the loss function
- $L_\mu(w)$ is the population loss of the hypothesis w w.r.t. μ
- Data-driven optimization: μ is unknown, but we have access to training data

$$\boldsymbol{Z} = (Z_1, \ldots, Z_n)\,, \quad Z_i \overset{\text{i.i.d.}}{\sim} \mu \tag{2}$$

Learning Algorithms and their performance

- Given: training data $\boldsymbol{Z} \sim \mu^{\otimes n}$
- Learning algorithm: a stochastic transformation (channel) from training data to hypotheses:
$$\boldsymbol{Z} \xrightarrow{P_{W|Z}} W \tag{3}$$
where W is a random element of the hypothesis space W
- Goal of learning (broadly speaking): design $P_{W|Z}$, such that the out-of-sample loss
$$L_\mu(W) = \int_{\mathrm{Z}} \ell(W, z)\mu(\mathrm{d}z) \tag{4}$$
is suitably small (either in expectation or with high probability)
- Caution!! $L_\mu(W)$ is a random variable

Empirical Loss and generalization Error

- The data-generating distribution μ is unknown; how do we evaluate the quality of the learned hypothesis W ?
- Empirical loss of a fixed hypothesis $w \in \mathrm{W}$:

$$L_Z(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(w, Z_i\right) \tag{5}$$

- unbiased estimate of $L_\mu(w), \mathbf{E}\left[L_{\boldsymbol{Z}}(w)\right] = L_\mu(w)$

- Empirical loss of W (a.k.a. resubstitution estimate)

$$L_{\boldsymbol{Z}}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(W, Z_i\right) \tag{6}$$

can be computed from the available information $(\boldsymbol{Z}, W)$, but is a biased estimate: $\mathbf{E}\left[L_{\boldsymbol{Z}}(W)\right] \neq \mathbf{E}\left[L_\mu(W)\right]$

Generalization error:

$$\operatorname{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) := \mathbf{E}\left[L_\mu(W) - L_{\boldsymbol{Z}}(W)\right] \tag{7}$$

What Does $\mathrm{gen}\left(\mu, P_{W|Z}\right)$ Tell Us?

Suppose there exists an optimal hypothesis $w_{\mathrm{opt}} \in \mathrm{W}$:

$$L_\mu\left(w_{\mathrm{opt}}\right) = \min_{w \in \mathrm{W}} L_\mu(w) \tag{8}$$

Let us analyze the expected excess risk of $P_{W|Z}$ w.r.t. μ :

$$\begin{aligned} \mathrm{ex}\left(\mu, P_{W|\boldsymbol{Z}}\right) &:= \mathbf{E}\left[L_\mu(W)\right] - L_\mu\left(w_{\mathrm{opt}}\right) \\ &= \mathbf{E}\left[L_\mu(W) - L_{\boldsymbol{Z}}(W)\right] + \mathbf{E}\left[L_{\boldsymbol{Z}}(W)\right] - L_\mu\left(w_{\mathrm{opt}}\right) \\ &= \mathbf{E}\left[L_\mu(W) - L_{\boldsymbol{Z}}(W)\right] + \mathbf{E}\left[L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}\left(w_{\mathrm{opt}}\right)\right] \\ &= \mathrm{gen}\left(\mu, P_{W|Z}\right) + \mathbf{E}\left[L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}\left(w_{\mathrm{opt}}\right)\right] \end{aligned} \tag{9}$$

Thus, $\mathrm{ex}\left(\mu, P_{W|\boldsymbol{Z}}\right)$ will be small if:

- $\mathrm{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right)$ is small (i.e., learning algo generalizes well on average)
- the empirical risks of W and w_{opt} are close on average

Uniform Convergence and Generalization

We can always bound the generalization error as

$$\mathrm{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) \leq \mathbf{E}\left[\sup_{w \in \mathrm{W}} \left|L_{\boldsymbol{Z}}(w) - L_{\mu}(w)\right|\right]$$

... but this bound:

- relies on restricting the complexity of the hypothesis space;
- ignores the details of the interaction between the data $\boldsymbol{Z}$ and the algo. output W;
- in particular, may be too conservative if the algo. does not explore the entire W due to fixed computational budget.

Learning does not require uniform convergence: One can construct examples of (ℓ, W), where uniform convergence does not hold (the upper bound does not converge to 0 as $n \to \infty$), yet learning still takes place (Shalev-Shwartz et al., 2010)

A Decoupling Estimate

Proposition:

Let U and V be two jointly distributed random objects, and let $f(U,V)$ be a real-valued function such that

$$\sup_u \log \mathbf{E}\left[e^{\lambda(f(u,V)-\mathbf{E}[f(u,V)])}\right] \leq \psi_+(\lambda), \quad \lambda > 0$$
$$\sup_u \log \mathbf{E}\left[e^{\lambda(f(u,V)-\mathbf{E}[f(u,V)])}\right] \leq \psi_-(-\lambda), \quad \lambda < 0$$

where ψ_+, ψ_- are convex and $\psi_\pm(0) = \psi'_\pm(0) = 0$. Then

$$\mathbf{E}[f(U,V) - f(\bar{U},\bar{V})] \leq \psi_+^{*-1}(I(U;V))$$
$$\mathbf{E}[f(\bar{U},\bar{V}) - f(U,V)] \leq \psi_-^{*-1}(I(U;V))$$

where:

- $P_{\bar{U},\bar{V}} = P_U \otimes P_V$
- $\psi_\pm^{*-1}$ is the inverse of the Legendre dual $\psi_\pm^*$

Proof

1. Donsker-Varadhan duality: for any $\lambda > 0$,

$$\begin{aligned} D\left(P_{V|U=u} \| P_V\right) &\geq \lambda \mathbf{E}[f(u,V) \mid U=u] - \log \mathbf{E}\left[e^{\lambda f(u,V)}\right] \\ &\geq \lambda(\mathbf{E}[f(u,V) \mid U=u] - \mathbf{E}[f(u,V)]) - \psi_+(\lambda) \end{aligned}$$

2. Rearrange and optimize:

$$\begin{aligned} \mathbf{E}[f(u,V) \mid U=u] - \mathbf{E}[f(u,V)] &\leq \inf_{\lambda>0} \frac{D\left(P_{V|U=u} \| P_V\right) + \psi_+(\lambda)}{\lambda} \\ &= \psi_+^{*-1}\left(D\left(P_{V|U=u} \| P_V\right)\right) \end{aligned}$$

(see, e.g., the book of Boucheron-Lugosi-Massart)

3. Average w.r.t. $U \sim P_U$:

$$\begin{aligned} \mathbf{E}[f(U,V)] - \mathbf{E}[f(\bar{U},\bar{V})] &\leq \int P_U(\,\mathrm{d}u)\left[\psi_+^{*-1}\left(D\left(P_{V|U=u}\right) \| P_V\right)\right] \\ &\leq \psi_+^{*-1}(I(U;V)), \end{aligned}$$

where we have used the fact that ψ_+^{*-1} is concave (since ψ_+^* is convex).

4. The case with $\lambda < 0$ is analogous.

Bounding gen $\left(\mu, P_{W|Z}\right)$ via Mutual Information

Theorem:

Suppose that there exist convex functions $\psi_\pm : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\psi_\pm(0) = \psi'_\pm(0) = 0$, such that

$$\sup_{w \in \mathrm{W}} \mathbf{E}\left[e^{\pm\lambda(\ell(w,Z)) - \mathbf{E}[\ell(w,Z)]}\right] \leq \psi_\pm(\pm\lambda), \quad \lambda > 0.$$

Then, for any learning algorithm $P_{W|Z}$ such that $I(W : \boldsymbol{Z}) < \infty$,

$$\psi_+^{*-1}\left(\frac{1}{n} I(W; \boldsymbol{Z})\right) \leq \text{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) \leq \psi_-^{*-1}\left(\frac{1}{n} I(W; \boldsymbol{Z})\right).$$

Remarks:

1. The subgaussian case is due to Xu-Raginsky (2017); related results by Russo-Zou (2016).
2. The general case was analyzed by Jiao-Han-Weissman (2018); Bu-Zou-Veeravalli (2019).

Proof

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

$$\begin{aligned}&\log \mathbf{E}\left[\exp\left\{\pm\lambda\left(L_{\boldsymbol{Z}}(w) - L_\mu(w)\right)\right\}\right] \\ &= n \log \mathbf{E}\left[\exp\left\{\frac{\pm\lambda}{n}(\ell(w,Z) - \mathbf{E}[\ell(w,Z)])\right\}\right] \leq n\psi_\pm(\pm\lambda/n)\end{aligned}$$

2. Now take $U = W, V = \boldsymbol{Z}, \ell(U,V) = L_{\boldsymbol{Z}}(W)$:

$$\mathbf{E}[f(U,V)] = \mathbf{E}\left[L_{\boldsymbol{Z}}(W)\right], \quad \mathbf{E}[f(\bar{U},\bar{V})] = \mathbf{E}\left[L_\mu(W)\right].$$

Apply the Decoupling Estimate to get

$$\begin{aligned}\text{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) &\leq \inf_{\lambda>0} \frac{I(W;\boldsymbol{Z}) + n\psi_-(\lambda/n)}{\lambda} \\ &= \inf_{\lambda>0} \frac{\frac{1}{n}I(W;\boldsymbol{Z}) + \psi_-(\lambda)}{\lambda} \\ &= \psi_-^{*-1}\left(\frac{1}{n}I(W;\boldsymbol{Z})\right)\end{aligned}$$

3. The lower bound is similar.

Subgaussian Case

- When $\ell(w, Z)$ is σ^2-subgaussian for every $w \in W$, we can take

$$\psi_{\pm}(t) = \frac{t^2\sigma^2}{2}, \quad \forall t \in \mathbb{R}$$
$$\psi_{\pm}^{*-1}(r) = \inf_{\lambda > 0} \frac{r + \lambda^2\sigma^2/2}{\lambda} = \sqrt{2r\sigma^2}$$

- Under the above assumption, for any learning algo. $P_{W|\boldsymbol{Z}}$ we have

$$\left|\text{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right)\right| \leq \sqrt{\frac{2\sigma^2}{n} I(W; \boldsymbol{Z})}$$

Tighter Bound via Individual-Sample Mutual Info

Theorem (Bu-Zou-Veeravalli)

Suppose $\ell(w, Z)$ is σ^2-subgaussian for each $w \in W$. Then for any learning algo. $P_{W|Z}$ we have

$$\left|\text{gen}\left(\mu, P_{W|Z}\right)\right| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma^2 I\left(W; Z_i\right)}$$

This bound is tighter than the $Xu - Raginsky$ bound:

$$\begin{aligned}
\sqrt{I(W; \boldsymbol{Z})} &= \sqrt{\sum_{i=1}^{n} I\left(W, Z^{i-1}; Z_i\right)} && \text{(chain rule, independence)} \\
&\geq \sqrt{\sum_{i=1}^{n} I\left(W; Z_i\right)} && \text{(data processing)} \\
&\geq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{I\left(W; Z_i\right)} && \text{(Jensen)}
\end{aligned}$$

Proof

1. Decompose

$$
\begin{aligned}
\operatorname{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) &= \mathbf{E}\left[L_\mu(W) - L_{\boldsymbol{Z}}(W)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[L_\mu(W) - \ell\left(W, Z_i\right)\right]
\end{aligned}
$$

2. Apply the Decoupling Estimate to each term in the sum: take $U = W, V = Z_i, f(U, V) = \ell\left(W, Z_i\right)$, then

$$
\left|\mathbf{E}\left[L_\mu(W) - \ell(W, Z)\right]\right| \leq \sqrt{2\sigma^2 I\left(W; Z_i\right)}
$$

3. Triangle inequality:

$$
\begin{aligned}
\left|\operatorname{gen}\left(\mu, P_{W|Z}\right)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[L_\mu(W) - \ell\left(W, Z_i\right)\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left|\mathbf{E}\left[L_\mu(W) - \ell\left(W, Z_i\right)\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\sigma^2 I\left(W; Z_i\right)}
\end{aligned}
$$

A Concentration Inequality for $|L_Z(W) - L_\mu(W)|$

- So far, we have been concerned with

$$\text{gen}\left(\mu, P_{W|\boldsymbol{Z}}\right) = \mathbf{E}\left[L_\mu(W) - L_{\boldsymbol{Z}}(W)\right]$$

- What about $\mathbf{P}\left[|L_\mu(W) - L_{\boldsymbol{Z}}(W)| > \varepsilon\right]$?
- Let's consider an extreme (and boring) case: W independent $\boldsymbol{Z}$ – the learning algoithm just ignores the data.
  Then, assuming $\ell(w, Z)$ is σ^2-subgaussian for all w,

$$\mathbf{P}\left[|L_{\boldsymbol{Z}}(W) - L_\mu(W)| > \varepsilon\right] \leq 2e^{-\frac{n\varepsilon^2}{2\sigma^2}}, \quad \forall \varepsilon > 0$$

  - that is, given $\varepsilon > 0$ and $0 < \delta \leq 1$, a sample size $n = \Omega\left(\frac{2\sigma^2}{\varepsilon^2} \log \frac{2}{\delta}\right)$ suffices to guarantee

$$|L_{\boldsymbol{Z}}(W) - L_\mu(W)| \leq \varepsilon \quad \text{with prob. at least } 1 - \delta.$$

- What happens if $I(W; \boldsymbol{Z})$ is suitably 'small?'

A Concentration Inequality for $|L_Z(W) - L_\mu(W)|$

Theorem (Xu-Raginsky)

Suppose $\ell(w, Z)$ is σ^2-subgaussian for all $w \in \mathrm{W}$. Let $P_{W|Z}$ be a learning algo. with $I(W; \boldsymbol{Z}) < \infty$. Let $\varepsilon > 0$ and $0 < \delta \leq 1$ be given . Then, provided

$$n \geq \frac{8\sigma^2}{\varepsilon^2}\left(\frac{I(W; \boldsymbol{Z})}{\delta} + \log\frac{2}{\delta}\right),$$

we will have

$$\mathbf{P}\left[|L_{\boldsymbol{Z}}(W) - L_\mu(W)| > \varepsilon\right] \leq \delta$$

Remarks:

1. The proof uses the monitor technique of Bassily et al.: run the algo. on m independent datasets $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$, then select the output with the largest value of $\left|L_\mu\left(W_j\right) - L_{\boldsymbol{Z}_j}\left(W_j\right)\right|$; the resulting 'super-algo.' has bounded mutual information.
2. The theorem does not give a 'high-probability' bound, due to $\frac{1}{\delta}$ additive term. Bassily et al. obtain such a bound assuming differential privacy and $0 \leq \ell \leq 1$.

Concentration using f-Divergence

Theorem (Esposito-Gastpar-Issa)

- Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probabiltiy spaces.
- Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $f(1) = 0$, and assume f is non-decreasing on $[0, +\infty)$.
- Suppose also that f is such that for every $y \in \mathbb{R}^+$the set $\{t \geq 0 : f(t) > y\}$ is non-empty, i.e. the generalized inverse, defined as $f^{-1}(y) = \inf\{t \geq 0 : f(t) > y\}$, exists.
- Let $f^*(t) = \sup_{\lambda \geq 0} \lambda t - f(\lambda)$ be the Fenchel-Legendre dual of $f(t)$. Given an event $E \in \mathcal{F}$, we have that:

$$\mathcal{P}(E) \leq \mathcal{Q}(E) \cdot f^{-1}\left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + (\mathcal{Q}(E^c))\, f^*(0)}{\mathcal{Q}(E)}\right)$$

Proof

$\forall \lambda > 0$:

$$\begin{aligned}
\mathcal{P}(E) &= \mathbb{E}_{\mathcal{P}}\left[\mathbb{I}_E\right] \\
&= \mathbb{E}_{\mathcal{Q}}\left[\mathbb{I}_E \frac{d\mathcal{P}}{d\mathcal{Q}}\right] \\
&\overset{(a)}{\leq} \frac{1}{\lambda}\mathbb{E}_{\mathcal{Q}}\left[f^*\left(\lambda \mathbb{I}_E\right) + f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] \\
&= \frac{D_f(\mathcal{P}\|\mathcal{Q}) + \mathbb{E}_{\mathcal{Q}}\left[f^*\left(\lambda \mathbb{I}_E\right)\right]}{\lambda} \\
&\overset{(b)}{\leq} \frac{D_f(\mathcal{P}\|\mathcal{Q}) + f^*(\lambda)\mathcal{Q}(E) + f^*(0)\left(\mathcal{Q}\left(E^c\right)\right)}{\lambda},
\end{aligned}$$

where (a) follows from Young's inequality and where f^* is the Legendre-Fenchel dual of f and (b) follows as, being $\mathbb{I}_E \in [0, 1]$ and we can write:

$$\begin{aligned}
f^*\left(\lambda \mathbb{I}_E\right) &= f^*\left(\lambda\left(\mathbb{I}_E + \left(1 - \mathbb{I}_E\right)0\right)\right) \\
&\leq \mathbb{I}_E f^*(\lambda) + \left(1 - \mathbb{I}_E\right) f^*(0)
\end{aligned}$$

Proof cont.

We can now minimize above over all $\lambda > 0$:

$$\begin{aligned}
\mathcal{P}(E) &\leq \inf_{\lambda>0} \left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + f^*(\lambda)\mathcal{Q}(E) + \left(\mathcal{Q}\left(E^c\right)\right) f^*(0)}{\lambda} \right) \\
&= \mathcal{Q}(E) \cdot \inf_{\lambda>0} \frac{\frac{D_f(\mathcal{P}\|\mathcal{Q}) + (\mathcal{Q}(E^c)) f^*(0)}{\mathcal{Q}(E)} + f^*(\lambda)}{\lambda} \\
&\overset{(c)}{=} \mathcal{Q}(E) \cdot f^{-1} \left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + \left(\mathcal{Q}\left(E^c\right)\right) f^*(0)}{\mathcal{Q}(E)} \right),
\end{aligned}$$

with (c) following from [book of Boucheron-Lugosi-Massart] as seen earlier.

Corollary 1 (Esposito-Gastpar-Issa)

Corlollary 1

Let X, Y be two random variables. Let $(\Omega, \mathcal{F}, \mathcal{P}_{XY}), (\Omega, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces where $\mathcal{F} = \sigma(X, Y)$ (i.e., the σ-algebra generated by (X, Y)). Let f be a convex function satisfying the assumptions of Theorem 1. Given an event $E \in \mathcal{F}$, we have that:

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot f^{-1}\left(\frac{I_f(X;Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))\, f^*(0)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right)$$

Corollary 3 (Esposito-Gastpar-Issa)

Corlollary 3

Corollary 3. Let $f(t) = t^2 - 1$, we have that $I_f(X;Y) = \chi^2(X,Y)$. Let $E \subseteq \mathcal{X} \times \mathcal{Y}$ we have that

$$\mathcal{P}_{XY}(E) \leq \sqrt{(\chi^2(X,Y)+1)\,\mathcal{P}_X\mathcal{P}_Y(E)}$$

Proof. We have that $f^*(t) = t^2/4 + 1$ and thus $f^*(0) = 1$. We also have that $f^{-1}(t) = \sqrt{t+1}$. Applying Corollary 1 we have that:

$$\begin{aligned}\mathcal{P}_{XY}(E) &\leq \mathcal{P}_X\mathcal{P}_Y(E)\sqrt{\frac{\chi^2(X,Y)+(1-\mathcal{P}_X\mathcal{P}_Y(E))}{\mathcal{P}_X\mathcal{P}_Y(E)}+1} \\ &= \sqrt{(\chi^2(X,Y)+1)\,\mathcal{P}_X\mathcal{P}_Y(E)}\end{aligned}$$

Corollary 4 (Esposito-Gastpar-Issa)

Corlollary 4

Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function such that $\ell(h, Z)$ is a σ^2-sub-Gaussian random variable [2], for some σ and for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$ Then,

$$\mathbb{P}(E) \leq \sqrt{2} \exp\left(\frac{1}{2}\left(\log\left(\chi^2(S, \mathcal{A}(S)) + 1\right) - n\frac{\eta^2}{2\sigma^2}\right)\right)$$

Corollary 6 (Esposito-Gastpar-Issa)

Corlollary 6

Let $E \in \mathcal{F}$ and let $\mathcal{P}_{XY}(E) \geq \mathcal{P}_X\mathcal{P}_Y(E)$, we have that:

$$\mathcal{P}_{XY}(E) - \mathcal{P}_X\mathcal{P}_Y(E) \leq H^2(X;Y) + 2H(X;Y)\sqrt{\mathcal{P}_X\mathcal{P}_Y(E)}$$

where $H^2(X;Y)$ denotes $H^2\left(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y\right)$

Rényi information bounds (Modak-Asnani-Prabhakaran)

Definition 1. The Rényi divergence of positive order $\alpha \neq 1$ between distributions $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ is given by [24],

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^\alpha {q_i}^{1-\alpha}$$

Lemma 1. Suppose P and Q are probability measures defined on $\mathcal{X}$ and g is a measurable function such that $e^{(\alpha-1)g} \in L^1(P)$ and $e^{\alpha g} \in L^1(Q)$. Then, for $\alpha \in \mathbb{R} \backslash \{0, 1\}$,

$$D_\alpha(P\|Q) \geq \frac{\alpha}{\alpha - 1} \log \mathbb{E}_P \left[e^{(\alpha-1)g(X)} \right] - \log \mathbb{E}_Q \left[e^{\alpha g(X)} \right]$$

Theorem 1. Suppose the loss function $\ell(w, Z)$ is σ subgaussian under P_{Z_i} and $P_{Z_i|W=w}$ for all w in the hypothesis set $\mathcal{W}$ and for each $i = 1, 2, \dots$, n. Then, for $\alpha \in (0, 1)$,

$$\left|\text{gen}\left(\mu, P_{W|S}\right)\right| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_W \left[\sqrt{2\sigma^2 \frac{D_\alpha\left(P_{Z_i|W} \| P_{Z_i}\right)}{\alpha}} \right]$$

Concentration using Rényi information

Lemma 3. Consider the measure space $(\mathcal{X}, \mathcal{F})$. Let P and Q be two probability measures on this space such that $P \ll Q$. Let $E \in \mathcal{F}$. Then, for $\alpha > 1$,

$$P(E) \leq \exp\left[\frac{\alpha - 1}{\alpha}\left(D_\alpha(P\|Q) + \log Q(E)\right)\right]$$

Proof:

$$\begin{aligned}
P(E) &= \mathbb{E}_P\left[\mathbf{1}_E\right] \\
&\overset{(a)}{=} \mathbb{E}_Q\left[\mathbf{1}_E\left(\frac{dP}{dQ}\right)\right] \\
&\overset{(b)}{\leq} \left(\mathbb{E}_Q\left[\mathbf{1}_E^{\frac{\alpha}{\alpha-1}}\right]\right)^{\frac{\alpha-1}{\alpha}}\left(\mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}} \\
&\overset{(c)}{=} (Q(E))^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} D_\alpha(P\|Q)\right) \\
&= \exp\left[\frac{\alpha - 1}{\alpha}\left(D_\alpha(P\|Q) + \log Q(E)\right)\right]
\end{aligned}$$

where we used $P \ll Q$ to write $(a), (b)$ follows from Holder's inequality with conjugates $\frac{\alpha}{\alpha-1}$ and α, and (c) is by the definition of Rényi divergence.

Concentration using Rényi information

Theorem 2. Let $E = \{(s, w) : |L_\mu(w) - L_S(w)| \geq \epsilon\}$ and P_{SW} be the joint distribution on $\mathcal{S} \times \mathcal{W}$. Suppose Q_{SW} is a measure such that $P_{SW} \ll Q_{SW}$. Then, for $\alpha > 1$ we have the following bound

$$P_{SW}(E) \leq \exp\left[\frac{\alpha - 1}{\alpha}\left(D_\alpha\left(P_{SW} \| Q_{SW}\right) + \log Q_{SW}(E)\right)\right]$$

Proof. Follows directly from Lemma 3.

My work

$\forall \lambda > 0$:

$$
\begin{aligned}
\mathcal{P}(E) &= \mathbb{E}_{\mathcal{P}}\left[\mathbb{I}_E\right] \\
&= \mathbb{E}_{\mathcal{Q}}\left[\mathbb{I}_E \frac{d\mathcal{P}}{d\mathcal{Q}}\right] \\
&\overset{(a)}{\leq} \frac{1}{\lambda}\mathbb{E}_{\mathcal{Q}}\left[f^*\left(\lambda \mathbb{I}_E\right) + f\left(\mathbb{I}_E \frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] \\
&= \frac{\mathbb{E}_{\mathcal{Q}}\left[f\left(\mathbb{I}_E \frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] + \mathbb{E}_{\mathcal{Q}}\left[f^*\left(\lambda \mathbb{I}_E\right)\right]}{\lambda} \\
&\overset{(b)}{\leq} \frac{\mathbb{E}_{\mathcal{Q}}\left[\mathbb{I}_E f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] + f(0)\left(\mathcal{Q}\left(E^c\right)\right) + f^*(\lambda)\mathcal{Q}(E) + f^*(0)\left(\mathcal{Q}\left(E^c\right)\right)}{\lambda},
\end{aligned}
$$

where (a) follows from Young's inequality and where f^* is the Legendre-Fenchel dual of f and (b) follows as, being $\mathbb{I}_E \in [0,1]$ and by convexity of $f()$ and $f^*()$

Cont.

We can now minimize above over all $\lambda > 0$:

$$\mathcal{P}(E) \leq \inf_{\lambda>0} \left(\frac{\mathbb{E}_{\mathcal{Q}} \left[\mathbb{I}_E f \left(\frac{d\mathcal{P}}{d\mathcal{Q}} \right) \right] + f(0) \left(\mathcal{Q} \left(E^c \right) \right) + f^*(\lambda) \mathcal{Q}(E) + f^*(0) \left(\mathcal{Q} \left(E^c \right) \right)}{\lambda} \right)$$

$$= \mathcal{Q}(E) \cdot \inf_{\lambda>0} \frac{\frac{\mathbb{E}_{\mathcal{Q}} \left[\mathbb{I}_E f \left(\frac{d\mathcal{P}}{d\mathcal{Q}} \right) \right] + f(0)(\mathcal{Q}(E^c)) + f^*(0)(\mathcal{Q}(E^c))}{\mathcal{Q}(E)} + f^*(\lambda)}{\lambda}$$

$$\overset{(c)}{=} \mathcal{Q}(E) \cdot f^{-1} \left(\frac{\mathbb{E}_{\mathcal{Q}} \left[\mathbb{I}_E f \left(\frac{d\mathcal{P}}{d\mathcal{Q}} \right) \right] + f(0) \left(\mathcal{Q} \left(E^c \right) \right) + f^*(0) \left(\mathcal{Q} \left(E^c \right) \right)}{\mathcal{Q}(E)} \right),$$

with (c) following from [book of Boucheron-Lugosi-Massart] as seen earlier.

Cont.

$$\mathbb{E}_{\mathcal{Q}}\left[\mathbb{I}_E f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] \overset{(d)}{\leq} \left(\mathbb{E}_Q\left[\mathbf{1}_E^{\frac{\alpha}{\alpha-1}}\right]\right)^{\frac{\alpha-1}{\alpha}} \left(\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}}$$

$$= (Q(E))^{\frac{\alpha-1}{\alpha}} \left(\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}}$$

where (d) follows from Holder's inequality with conjugates $\frac{\alpha}{\alpha-1}$ and α,

$\mathcal{P}(E)$

$$\leq \mathcal{Q}(E) \cdot f^{-1}\left(\frac{(Q(E))^{\frac{\alpha-1}{\alpha}}\left(\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}} + f(0)\left(\mathcal{Q}\left(E^c\right)\right) + f^*(0)\left(\mathcal{Q}\left(E^c\right)\right)}{\mathcal{Q}(E)}\right)$$

Cont.

if we choose $f(t) = t^2 - 1$, we have $I_f(X, Y) = \chi^2(X, Y)$ and :

- $f^{-1}(t) = \sqrt{t+1}$
- $f^*(t) = \frac{t^2}{4} + 1$

Plugging in we have:

$$\mathcal{P}(E) \leq \mathcal{Q}(E) \cdot \sqrt{\left(\frac{(Q(E))^{\frac{\alpha-1}{\alpha}} \left(\mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right)^{\alpha} \right] \right)^{\frac{1}{\alpha}})}{\mathcal{Q}(E)} \right) + 1}$$

$$= \sqrt{\left((Q(E))^{\frac{\alpha-1}{\alpha}} \left(\mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right)^{\alpha} \right] \right)^{\frac{1}{\alpha}}) + \mathcal{Q}(E) \right) \mathcal{Q}(E)}$$

cont.

Esposito-Gastpar-Issa Bound

$$\mathcal{P}(E) \leq \sqrt{(\chi^2(X,Y)+1)\,\mathcal{Q}(E)}$$

Our Bound

$$\mathcal{P}(E) \leq \sqrt{\left((Q(E))^{\frac{\alpha-1}{\alpha}}\left(\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}}) + \mathcal{Q}(E)\right)\mathcal{Q}(E)}$$

if $\alpha \to 1$ we have:

$$\mathcal{P}(E) \leq \sqrt{(\chi^2(X,Y)+\mathcal{Q}(E))\,\mathcal{Q}(E)}$$

By optimising α we might get even better bound.

References

[1] Maxim Raginsky *"Information, Concentration, and Learning slides"*

[2] Aolin Xu, Maxim Raginsky *"Information-theoretic analysis of generalization capability of learning algorithms"*

[3] Yuheng Bu, Shaofeng Zou, Venugopal V. Veeravalli *"Tightening Mutual Information-Based Bounds on Generalization Error"*

[4] Amedeo Roberto Esposito, Michael Gastpar, Ibrahim Issa *"Robust Generalization via f Mutual Information"*

[5] Eeshan Modak, Himanshu Asnani, Vinod M. Prabhakaran*"Renyi Divergence Based Bounds on Generalization Error"*