Rate-Distortion Analysis of Minimum Excess Risk in Bayesian Learning

Hassan Hafez-Kolahi[1] Behrad Moniri[2] Shohreh Kasaei[1] Mahdieh Soleymani Baghshah[1]

[1] Department of Computer Engineering, Sharif University of Technology [2] Department of Electrical Engineering, Sharif University of Technology

ICML International Conference On Machine Learning

Problem Formulation

Data Generation Model:

$$P_{W,Z^n,Z} = P_W \otimes \prod_{i=1}^{n} P_{Z_i|W} \otimes P_{Z|W}$$

$$\forall\, i \in [n],\quad P_{Z_i|W} = P_{Z|W}$$

Bayes Risk of predicting Y given U:

$$R_\ell(Y|U) = \inf_{\psi:\mathcal{U}\to\mathcal{Y}} \mathbb{E}[\ell(Y,\psi(U))] \rightsquigarrow \psi^*_{Y|U}(u)$$

Minimum Excess Risk (MER):

$$\mathrm{MER}^n_\ell = R_\ell(Y|Z^n,X) - R_\ell(Y|W,X)$$

Related Literature

Theorem (Xu & Raginsky 2020). *Consider an arbitrary non-negative bounded function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,b]$. We have*

$$\mathrm{MER}^n_\ell \le \sqrt{\frac{b^2}{2} I(Y;W|Z^n,X)} \le \sqrt{\frac{b^2}{2n} I(W;Z^n)}.$$

Remark 1: Under mild conditions, $I(Y;W|Z^n) = O(1/n)$ as $n\to\infty$ giving $\mathrm{MER}^n_\ell = O(\sqrt{1/n})$.

Remark 2: The lower bound was left as an open problem in [Xu & Raginsky 2020]

Remark 3: We showed that no lower bound of the form $\alpha\sqrt{I(Y;W|Z^n,X)}$ exists.

Droping the Square Root

For bounded random variables, the upper bound can be improved to $O(1/n)$ if the loss is quadratic or the problem is realizable.

Lemma. *Consider random variables Y, U, and V forming Markov chain $Y - U - V$ and an arbitrary bounded function $\ell : \mathcal{Y}\times\mathcal{Y}\to[0,b]$. We have*

$$R_\ell(Y|V) \le 2R_\ell(Y|U) + 3bI(Y;U|V).$$

Rate-Distortion View

The Markov chain $W \to Z^n \to \hat{h}(\cdot)$ holds.

Goal: Observe Z^n and find $\hat{h}(.)$ which performs well compared to the case where W is known.

Not a standard rate-distortion problem!

Lower Bound:
If asked to use $R = I(W;Z^n)$ nats to represent W by a variable Ξ in a way that it is possible to recover a good $\hat{h}$, is it a good idea to set $\Xi = Z^n$?

Upper Bound:
Is it possible to have $I(Z^n;\hat{h}) = I(W;Z^n)$ and still achieve the optimal $\hat{h}(.)$.

R/D Optimization

Define the distortion function as $h^*_w(x)$, i.e.

$$d(w,\hat{h}) = \mathbb{E}^w_{XY}[\ell(Y,\hat{h}(X)) - \ell(Y,h^*_w(X))].$$

We have $\mathbb{E}_{WZ^n}[d(W,\psi^*_{Y|Z^nX}(Z^n,\cdot))] = \mathrm{MER}^n_\ell$.

The (Constrained) Rate-Distortion Function:

$$D_n(R) = \inf_{P^{Z^n}_{\hat{h}}} \mathbb{E}\big[d(W,\hat{h})\big] \quad \text{s.t.} \quad I(W;\hat{h}) \le R.$$

Theorem. *For a given training set size n, for all rates $R \ge I(W;Z^n)$, we have $D_n(R) = \mathrm{MER}^n_\ell$.*

Upper Bound

Add the constraint that $I(Z^n;\hat{h}) \le R$:

$$D^U_n(R) = \inf_{P^{Z^n}_{\hat{h}}} \mathbb{E}[d(W,\hat{h})] \text{ s.t. } I(Z^n;\hat{h}) \le R.$$

We have $\forall R,\ \forall n;\ D_n(R) \le D^U_n(R)$.

Theorem. *For any bounded loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,b]$, and for all $n \ge 1$, we have*

$$D^U_n(I(W;Z^n)) \le \sqrt{\frac{b^2}{2} I(W;Y|Z^n,X)}.$$

Lower Bound

Remove the constraint that $\hat{h}$ is generated only using the samples Z^n:

$$D^L(R) = \inf_{P^W_{\hat{h}}} \mathbb{E}[d(W,\hat{h})], \text{ s.t. } I(W;\hat{h}) \le R.$$

The feasible set is enlarged; hence

$$\forall R,\ \forall n;\quad D^L(R) \le D_n(R).$$

Comparing Bounds

Theorem. *For any bounded loss $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,b]$, we have*

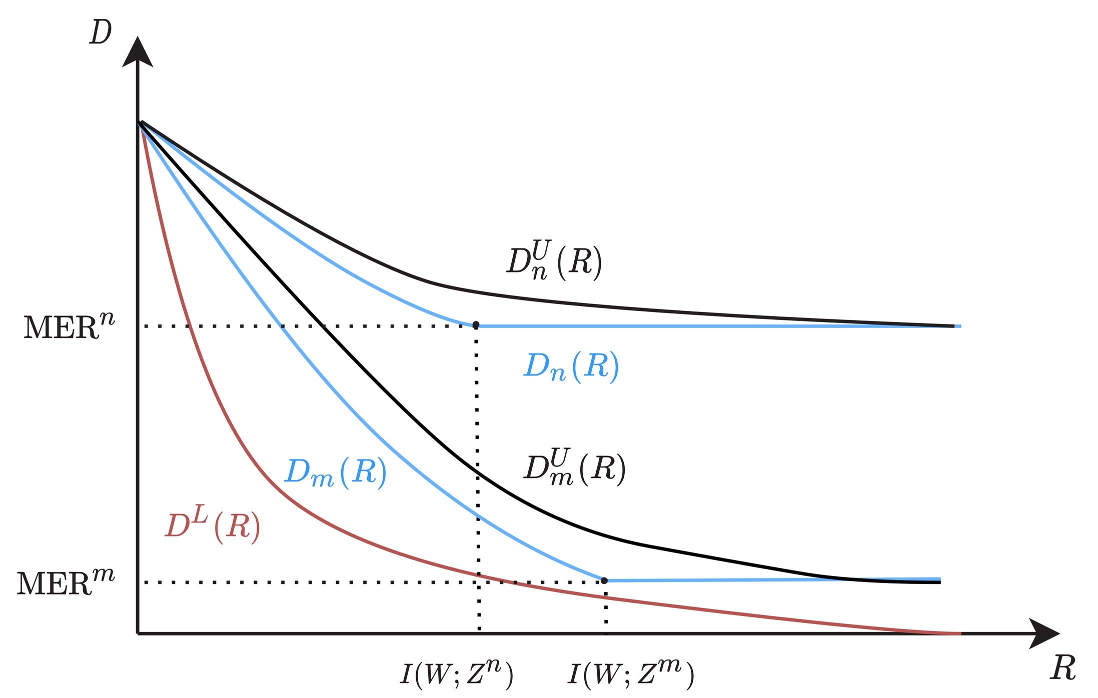
$$D^U_n(R) - D^L(R) \le \sqrt{\frac{b^2}{2} I(W;\hat{h}_R|Z^n)},$$

where $P_{W,\hat{h}_R Z^n} = P_W \otimes P^{*W}_{\hat{h}_R} \otimes P^W_{Z^n}$ and $P^{*W}_{\hat{h}_R}$ is a solution for $D^L(R)$.

As $n\to\infty$, if the posterior is concentrated to the true realization, it is reasonable to expect that $I(W;\hat{h}_R|Z^n) \to 0$ and all of the rate-distortion functions converge.

Theorem. *Suppose the distortion $d(W,\hat{h})$ can be represented as a distance $d'(h^*_W,\hat{h})$. Let W and W' be two samples independently generated from $P^{Z^n}_W$. If we have $\lim_{n\to\infty}\mathbb{E}[d'(h^*_W,h^*_{W'})] = 0$, then*

$$\forall R \ge 0;\ D^L(R) = \lim_{n\to\infty} D_n(R) = \lim_{n\to\infty} D^U_n(R).$$



Lower Bound on MER

For quadratic loss we have

$$d(w,\hat{h}) = ||\psi_{Y|W,X}(w,\cdot) - \hat{h}(\cdot)||_{L^2(P_X)}.$$

Theorem. *Let $\mathcal{W}$ be a p-dimensional compact and convex subspace of $\mathbb{R}^p$. Under some mild conditions (see Section 6 of the paper), as $n\to\infty$ we have*

$$\mathrm{MER}^n_\ell \ge \frac{p\pi}{n\big(V_p\,\Gamma(1+\frac{p}{2})\big)^{\frac{2}{p}}} \exp\left(\frac{-\mathbb{E}\log|J^W_Z(W)|}{p}\right).$$

The bounds are tight for the cases where the upper rate of $O(1/n)$ holds.

Application in Linear Models

Under the conditions that:

- P_W is supported on a compact & convex subset of $\mathbb{R}^p$.
- $Y = W^\top X + \sigma\nu$, where $W \sim P_W$, $X \sim \mathcal{N}(0,\Sigma_X)$, and $\nu \sim \mathcal{N}(0,1)$.
- Variables W, X, and ν are independent.
- The matrix Σ_X is full-rank.

We have $\mathrm{MER}^n_\ell = \Omega(p/n)$.

Similar results can be derived for Neural Tangent Kernels $f(\cdot,w) = f(\cdot,w_0) + \Phi^\top_{w_0}(\cdot)(w - w_0)$.

Future Work

One of the limitations of the current work, is that our result requires some technical conditions for the $\Omega(p/n)$ to be guaranteed. Analyzing lower rates under more general conditions, for example non-parametric problems, is an interesting direction for future studies.

References

[Xu & Raginsky 2020] Aolin Xu and Maxim Raginsky. Minimum Excess Risk in Bayesian Learning, ArXiv, 2020.