Integral Probability Metrics PAC-Bayes Bounds

Ron Amit, Baruch Epstein, Shay Moran, Ron Meir

Technion – Israel Institute of Technology

NeurIPS 2022

European Research Council

TECHNION
Israel Institute
of Technology

Problem description

- Samples $z \in \mathcal{Z}$
- Data-set of m samples $S = \{z_1, ..., z_m\}$
- Statistical model $z \underset{i.i.d}{\sim} \mathcal{D}$ $\mathcal{D}$ unknown distribution
- Hypothesis $h \in \mathcal{H}$ Learning algorithm $\mathcal{A} : \mathcal{Z}^m \rightarrow \mathcal{H}$
- Loss function $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$

Problem description

- The generalization gap:

$$\Delta_S(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h,z) - \frac{1}{m}\sum_{j=1}^{m} \ell(h,z_j) \quad h \in \mathcal{H}$$

Expected risk (out-of-sample)

Empirical risk (in-sample)

- We want a high probability bound on $\Delta_S(h)$

Uniform Convergence (UC)

- Uniform convergence bound

$$\Delta_S(h) \leq u(m,\delta), \forall h \in \mathcal{H} \quad \text{w.p.} \geq 1-\delta$$

UC bound $u(m,\delta) \underset{m\to\infty}{\to} 0$

e.g., $u(m,\delta) = c\sqrt{\frac{\mathrm{VC}(\mathcal{H}) + \ln(1/\delta)}{m}}$

- independent of the algorithm and data
- Usually very loose for large classes

PAC-Bayes (PB) Bounds

- Posterior: $Q \in \mathcal{M}(\mathcal{H})$ Distributions over $\mathcal{H}$

  $$\mathcal{A} : \mathcal{Z}^m \to \mathcal{M}(\mathcal{H})$$

Non-Bayesian setting!

- Generalization gap:

$$\Delta_S(Q) = \underbrace{\mathbb{E}_{h \sim Q} \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z)}_{\text{Expected risk}} - \underbrace{\mathbb{E}_{h \sim Q} \frac{1}{m} \sum_{j=1}^{m} \ell(h, z_j)}_{\text{Empirical risk}}$$

PAC-Bayes (PB) Bounds

- For any data-independent $P \in \mathcal{M}(\mathcal{H})$ "prior"
- Classical KL-PB bound

$$\Delta_S(Q) \leq \sqrt{\frac{1}{2(m-1)}\left(D_{KL}(Q||P) + \log\frac{m}{\delta}\right)} \quad \text{w.p. } \geq 1-\delta$$

for all $Q \in \mathcal{M}(\mathcal{H})$

KL-divergence $$D_{KL}(Q||P) \triangleq \mathop{\mathbb{E}}_{h\sim Q} \log\frac{Q(h)}{P(h)}$$

Requires $\text{supp}(Q) \subset \text{supp}(P)$

New Family of PB bounds

- New family of Integral Probability Metrics (IPMs) based PB bounds
- KL-divergence is replaced by
  - Total-Variation (TV) distance
  - Wasserstein distance
- No requirement for the distributions supports (Works even for Dirac delta measures!)

Total-Variation (TV) PB Bound

- Given a UC bound

$$\Delta_S(h) \le u(m,\delta), \forall h \in \mathcal{H} \qquad \text{w.p. } \ge 1-\delta$$

- we get a TV-PB bound:

$$\Delta_S(Q) \le \sqrt{u^2(m, \delta/2) D_{\mathrm{TV}}(Q,P) + \frac{\ln(2m/\delta)}{2(m-1)}} \qquad \text{w.p. } \ge 1-\delta$$

$$\le \tilde{O}\left(u \cdot \underbrace{\sqrt{D_{\mathrm{TV}}(Q,P)}}_{\le 1}\right)$$

Origin UC bound

algorithm and data-dependent factor

Wasserstein-PB Bound

- [1] Given $\Delta_S^2(\cdot)$ is K-Lipschitz with $K = K(m, \delta)$ with probability $\geq 1 - \delta$

▶ we get a Wasserstein-PB (WPB):

$$\Delta_S(Q) \leq \sqrt{\underbrace{K(m, \delta/2)}W_1(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}$$

Finite class with G-Lipschitz loss $\tilde{O}\left(\frac{G \log|\mathcal{H}|}{m}\right)$

- Linear regression in $\mathbb{R}^d$ $\tilde{O}\left(\frac{d}{m}\right)$

Linear Regression Example

UC bound $\tilde{O}\left(\sqrt{\frac{d}{m}}\right)$

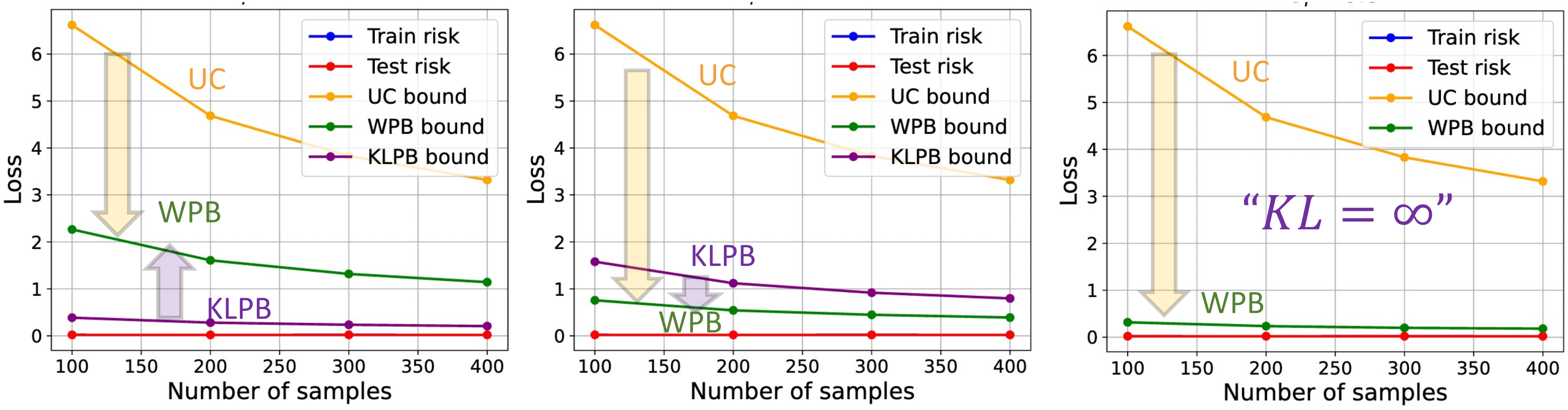
WPB bound $\tilde{O}\left(\sqrt{W_1(Q,P)\frac{d}{m}}\right)$

KLPB bound $\tilde{O}\left(\sqrt{\frac{\mathrm{KL}(Q \parallel P)}{m}}\right)$

wide prior

narrow prior

Dirac's deltas



(a) $\sigma_{\mathbf{P}} = 10^{-2}, \sigma_{\mathbf{Q}} = 10^{-3}$

(b) $\sigma_{\mathbf{P}} = 10^{-4}, \sigma_{\mathbf{Q}} = 10^{-3}$

(c) $\sigma_{\mathbf{P}} = \sigma_{\mathbf{Q}} = 0$