A Unified Framework for Generalization Bounds via Change of Measure Inequalities

Yanxiao Liu* [1] Yijun Fan* [2] Deniz Gündüz [1]

[1]Imperial College London [2]The Chinese University of Hong Kong

Generalization Error Bounds

Central challenge in machine learning: quantify the "generalization" guarantee of learning algorithms

- If an algorithm performs well on training data, will it also perform well on new samples?

Problem Setting:

- **Stochastic learning algorithm** $P_{W|S}$ as a probabilistic mapping from a training dataset $S := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ to hypothesis $W \in \mathcal{W}$.
- **Loss function** $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is σ-sub-Gaussian.
- Generalization error is

$$\mathrm{gen}(S, W) := \mathbf{E}_{P_Z}\left[\ell(Z, W)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, W),$$

i.e., the gap between the expected and empirical loss.

Existing Bounds:

Average bound via mutual information:

$$\mathbf{E}\left[\mathrm{gen}(S, W)\right] \leq \sqrt{\frac{2\sigma^2}{n} I(S; W)}.$$

High-probability bound via maximal leakage:

$$\mathbf{P}\left(\left|\mathrm{gen}(S, W)\right| \geq \eta\right) \leq 2 \cdot \exp\left(\mathcal{L}(S \to W) - \frac{n\eta^2}{2\sigma^2}\right).$$

We consider the high-probability case.

Change of Measure Inequalities

A class of change of measure inequalities: for probability measures P, Q on measurable $(\mathcal{X}, \mathcal{F})$ such that $P \ll Q$, for measurable set $E \in \mathcal{F}$, we consider

$$P(E) \leq \xi\big(Q(E), \mathrm{d}P/\mathrm{d}Q\big)$$

where $\xi(\cdot)$ is some functional.

It can be converted to generalization error bounds by taking

$$\begin{aligned} P &:= P_{SW}, \\ Q &:= P_W P_S, \\ E &= \{W, S : |\mathrm{gen}(S, W)| \geq \epsilon\}. \end{aligned}$$

For example, the strong converse lemma

$$P(E) \leq \gamma Q(E) + P\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\right), \quad \forall \gamma \in \mathbb{R}.$$

has been used for generalization bounds [1].

Information Measure

Consider $f : [0, \infty) \to \mathbb{R}$ be convex and $f(0) = \lim_{t \downarrow 0} f(t)$. For distributions $P \ll Q$, define f-divergence

$$D_f(P\|Q) := \int f\left(\mathrm{d}P/\mathrm{d}Q\right) \mathrm{d}Q.$$

- By $f(t) = t \log t$ it recovers KL divergence.
- By $f(t) = t^2 - 1$ it recovers χ^2-divergence $\chi^2(P\|Q)$.
- By $f(t) = \frac{t^\beta - 1}{\beta - 1}$ it recovers power divergence $\mathcal{H}_\beta(P\|Q)$.
- By $f(t) = (1 - \sqrt{t})^2$ it recovers Hellinger squared distance $H(P; Q)$.
- By $f(t) = [t - \gamma]_+$ it recovers E_γ-divergence $E_\gamma(P\|Q)$.

Data Processing Inequality

For a Markov kernel T from $\mathcal{X}$ to $\mathcal{Y}$ and probability measures P, Q on $\mathcal{X}$, we define $T \circ P$ as the pushforward measure: $(T \circ P)(B) = \int_{\mathcal{X}} T(B|x) P(\mathrm{d}x)$, $B \in \mathcal{F}_{\mathcal{Y}}$ and define $T \circ Q$ similarly. Then

$$D_f(T \circ P \| T \circ Q) \leq D_f(P\|Q).$$

Choose $T = \mathbb{1}_E$ and denote $p := P(E)$, $q := Q(E)$, we observe

$$\begin{aligned} D_f(P\|Q) &\geq D_f(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) \\ &= D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) \\ &= q f\left(\frac{p}{q}\right) + (1 - q) f\left(\frac{1 - p}{1 - q}\right). \end{aligned}$$

Machinery proof: simply specify f and calculate the last term.

E_γ-divergence

Fix probability measures P, Q on $\mathcal{X}$ such that $P \ll Q$. For measurable E,

$$P(E) \leq \gamma Q(E) + E_\gamma(P\|Q), \quad \forall \gamma \in \mathbb{R}. \tag{1}$$

Proof Sketch:

E_γ is an f-divergence with $f(t) = [t - \gamma]_+$, then for any $\gamma \in \mathbb{R}$,

$$\begin{aligned} E_\gamma(P\|Q) &\geq E_\gamma(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) \\ &= q \cdot [p/q - \gamma]_+ + (1 - q) \cdot [(1 - p)/(1 - q) - \gamma]_+ . \end{aligned}$$

Since $(1-q)[\cdot]_+ \geq 0$, it follows that $E_\gamma(P\|Q) \geq q(p/q - \gamma)$, and rearrangement yields the result.

This result (1) is strictly stronger than the strong converse lemma.

More examples

For probability measures P, Q on $\mathcal{X}$ such that $P \ll Q$, for measurable E,

$$P(E) \leq Q(E) + \sqrt{Q(E)(1 - Q(E))\chi^2(P\|Q)}, \tag{2}$$

$$P(E) \leq \big(\mathrm{KL}(P\|Q) + \log(1 + Q(E)(e^c - 1))\big)/c, \quad c > 0, \tag{3}$$

$$2\left(1 - \sqrt{P(E)Q(E)} - \sqrt{(1 - P(E))(1 - Q(E))}\right) \leq H^2(P; Q), \tag{4}$$

$$P(E)^\beta Q(E)^{1-\beta} + (1 - P(E))^\beta (1 - Q(E))^{1-\beta} \leq 1 + (\beta - 1)\mathcal{H}_\beta(P\|Q). \tag{5}$$

These results are stronger or equivalent to results in [2].

Related Works

We recover [3, Theorem 3] by

$$\begin{aligned} f\big(p/q\big) &\leq \big(D_f(P\|Q) - (1 - q) f\big((1 - p)/(1 - q)\big)\big)/q \\ &\leq \big(D_f(P\|Q) + (1 - q) f^*(0)\big)/q, \end{aligned}$$

since $f^*(0) = \sup_{t \geq 0}(-f(t)) = -\inf_{t \geq 0} f(t) \geq -f\big((1 - p)/(1 - q)\big)$.

There is a concurrent work [4] for PAC-Bayesian bounds; we can derive PAC-Bayesian bounds tighter than some of their results.

References

[1] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 3, pp. 824–839, 2020.

[2] A. Picard-Weibel and B. Guedj, "On change of measure inequalities for f-divergences," arXiv preprint arXiv:2202.05568, 2022.

[3] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," IEEE Transactions on Information Theory, vol. 67, no. 8, pp. 4986–5004, 2021.

[4] M. Guan, F. Farokhi, and J. Zhu, "A DPI-PAC-Bayesian framework for generalization bounds," in IEEE Information Theory Workshop (ITW) 2025, Australia, 2025, pp. 1–6.

yliu25@ic.ac.uk

https://arxiv.org/pdf/2602.07999