Chained Generalisation Bounds

E. Clerico, A. Shidani, G. Deligiannidis, A. Doucet

University of Oxford, Department of Statistics

COLT 2022

Setting

- **Loss Function:** $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$
- **Learning Algorithm:**

$$s = \{x_1, \ldots, x_m\} \sim \mathbb{P}_S \mapsto W \sim \mathbb{P}_{W|S=s}$$

- **Population Loss:** $\mathscr{L}_{\mathcal{X}}(w) = \mathbb{E}_{\mathbb{P}_X}[\ell(w, X)]$
- **Empirical Loss:** $\mathscr{L}_s(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, x_i)$
- **Generalisation Gap:** $g_s(w) = \mathscr{L}_{\mathcal{X}}(w) - \mathscr{L}_s(w)$

GOAL: Control the expected generalisation gap $\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W,S}}[g_S(W)]$.

Examples of Generalisation Bounds

Proposition (Standard MI bound - Russo & Zou, 2019)

Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. *If* $\ell(w, X)$ *is* ξ*-SG,* $\forall w \in \mathcal{W}$*, then* $|\mathcal{G}| \leq \xi\sqrt{\frac{2I(W;S)}{m}}$.

Proposition (Standard Wasserstein bound - Lopez & Jog, 2018)

Suppose that $d_{\mathcal{X}}$ *and* $d_{\mathcal{S}}$ *are related by* $d_{\mathcal{S}}(s, s')^2 = \sum_{i=1}^m d_{\mathcal{X}}(x_i, x_i')^2$. *If,* $\forall w \in \mathcal{W}$, $x \mapsto \ell(w, x)$ *is* ξ*-Lipschitz on* $\mathcal{X}$*, then* $|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W})]$.

- In both bounds uniform (in $\mathcal{W}$) regularity (in $\mathcal{X}$) of the loss.
- Key idea: $\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]$.

General framework

Definition ($\mathfrak{D}$-regularity)

Let $\mathfrak{D}$ be a measurable map $\mathscr{P} \times \mathscr{P} \to [0, +\infty]$. Fix $\mu \in \mathscr{P}$ and $\xi \geq 0$. We say that $f : \mathcal{Z} \to \mathbb{R}$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$, with respect to μ, if $f \in L^1(\mu)$ and, for every $\nu \in \mathscr{P}$ such that $\mathrm{Supp}(\nu) \subseteq \mathrm{Supp}(\mu)$ and $f \in L^1(\nu)$,

$$|\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)]| \leq \xi \, \mathfrak{D}(\mu, \nu) \, .$$

We say that $F : \mathcal{Z} \to \mathbb{R}^q$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ (wrt μ) if $z \mapsto v \cdot F(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi \|v\|)$ (wrt μ), for all $v \in \mathbb{R}^q$.

Theorem (General unchained bound)

Assume that $s \mapsto \mathscr{L}_s(w)$ *has regularity* $\mathcal{R}_{\mathfrak{D}}(\xi)$ *wrt* $\mathbb{P}_S$, $\forall w \in \mathcal{W}$. *Then we have*

$$|\mathcal{G}| = |\mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]| \leq \xi\, \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})]\,,$$

where $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] = \int_{\mathcal{W}} \mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W=w})\, \mathrm{d}\mathbb{P}_W(w)$.

- If $f(Z)$ is ξ-SG for $Z \sim \mu \in \mathscr{P}$:
  f has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ where $\mathfrak{D} : (\mu, \nu) \mapsto \sqrt{2\mathrm{KL}(\nu\|\mu)}$.
- If f is ξ-Lipschitz on $\mathcal{Z}$:
  f has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ where $\mathfrak{D} : (\mu, \nu) \mapsto \mathfrak{W}(\mu, \nu)$.

$\Longrightarrow$ unchained MI and Wasserstein bounds!

Chained Bounds

The chained bounds are multiscale generalisation bounds that leverage the dependencies between different hypotheses by mean of the chaining technique.

Definition (ε-Nets)

Given $\varepsilon > 0$, we define an ε-projection on $\mathcal{W}$ as a measurable mapping $\pi : \mathcal{W} \to \mathcal{W}$ such that $\pi(\mathcal{W})$ has finitely many elements and, for all $w \in \mathcal{W}$, $\|\pi(w) - w\| \leq \varepsilon$. The image $\pi(\mathcal{W})$ is called an ε-net on $\mathcal{W}$.

Definition (Refining Sequences of Nets)

Consider a positive, vanishing, decreasing sequence $\{\varepsilon_k\}_{n\in\mathbb{N}}$ and assume that $\exists w_0 \in \mathcal{W}$ such that $\|w - w_0\| \leq \varepsilon_0$, $\forall w \in \mathcal{W}$. We call $\{\pi_k(\mathcal{W})\}_{n\in\mathbb{N}}$ an $\{\varepsilon_k\}$-refining sequence of nets if $\pi_0(\mathcal{W}) = \{w_0\}$ and, for all $k \geq 1$, we have that π_k is a ε_k-projection and $\pi_{k-1} \circ \pi_k = \pi_{k-1}$.

Proposition (CMI bound - Asadi et al., 2018)

Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ and $\mathcal{W}$ be a compact set, with an $\{\varepsilon_k\}$-refining sequence of nets $\{\mathcal{W}_k\}$ defined on it. Suppose that $w \mapsto \ell(w, x)$ is continuous, for $\mathbb{P}_X$-almost every x, and that $\{\ell(w, X)\}_{w \in \mathcal{W}}$ is a ξ-SG stochastic process. Then we have

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2 I(W_k; S)}\,.$$

The bounds based on chaining, such as the CMI bound, do not fit naturally in the framework presented so far.

Key idea: The regularity of $x \mapsto \ell(w, x)$ is transferred onto $x \mapsto (\ell(w, x) - \ell(w', x))$.

If ℓ is regular enough we can focus on the gradient $\nabla_w \ell(w, x)$ instead.

Assumptions (♣)

- *The set* $\mathcal{W} \subset \mathbb{R}^d$ *is convex, compact, and with non-empty interior.*
- *The function* $w \mapsto \ell(w, x)$ *is of class* C^1 *on* $\mathcal{W}$, $\mathbb{P}_X$*-a.s.*
- $\sup_{\mathcal{W}\times\mathcal{X}} |\ell(w,x)| < +\infty$ *and* $\sup_{\mathcal{W}\times\mathcal{X}} \|\nabla_w \ell(w,x)\| < +\infty$.

Theorem

Assume ♣ *and that* $s \mapsto \nabla_w \mathscr{L}_s(w)$ *has regularity* $\mathcal{R}_{\mathfrak{D}}(\xi)$ *wrt* $\mathbb{P}_S$, $\forall w \in \mathcal{W}$. *Then*

$$|\mathcal{G}| = |\mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})],$$

where $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] = \int_{\mathcal{W}} \mathfrak{D}\left(\mathbb{P}_S, \mathbb{P}_{S|W\in\pi_k^{-1}(w)}\right) \mathrm{d}\mathbb{P}_W(w)$.

Key idea: $\mathscr{L}_s(w) = \mathscr{L}_s(w_0) + \sum_{k\geq 1}(\mathscr{L}_s(w_k) - \mathscr{L}_s(w_{k-1}))$.

- Recovering the CMI bound:

Proposition (Variant of the CMI bound)

Under ♣, if $\nabla_w \ell(w, X)$ is ξ-SG $\forall w$:

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2I(W_k; S)}\,.$$

- Finding new chained bounds:

Proposition (Chained Wasserstein bound)

Under ♣, if $x \mapsto \nabla_w \ell(w, x)$ is ξ-Lipschitz $\forall w$:

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W} \left[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W_k}) \right].$$

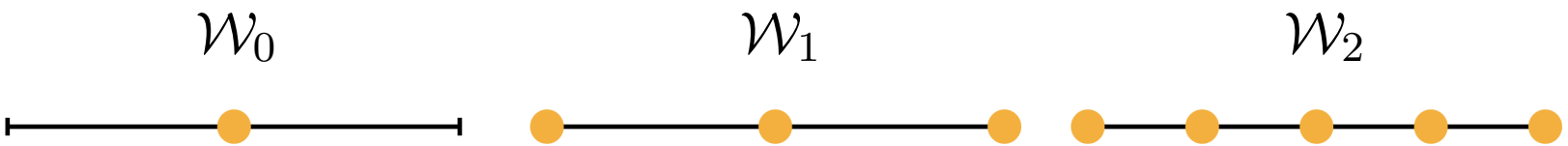
To chain or not to chain?

- The requirements for the chained bounds are somewhat **stronger**: whenever we derive a chained bound in our framework, we can always state an unchained counterpart.
- However, conditioning on W_k instead of W can often be helpful: $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \leq \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})]$ if $\mathfrak{D}(\mathbb{P}_S, \cdot)$ is convex.
- If $\mathbb{P}_W$ is very **concentrated on a tiny region** of $\mathcal{W}$, S is almost independent of W_k up to a small scale and the chained result tends to be the tightest.

Toy Example

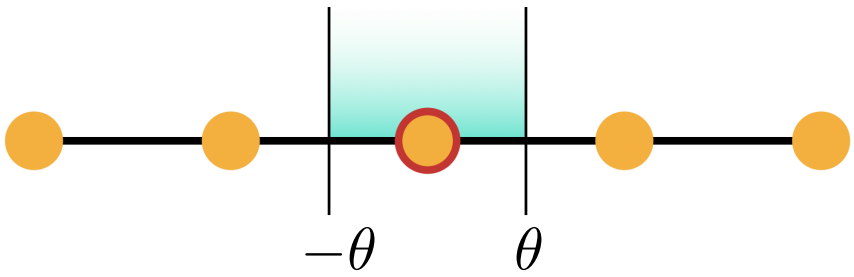
$$\mathcal{W} = [-1, 1] \qquad \mathcal{X} = [-1, 1] \qquad \varepsilon_k = 2^{-k}$$

$$\mathcal{W}_k = \{2^{1-k} j \ : \ j \in [-2^{k-1} : 2^{k-1}]\}, \text{ where } [a : b] = [a, b] \cap \mathbb{Z}$$



$$\theta = 2^{-k^\star} \qquad X \sim \text{Unif}(-\theta, \theta) \qquad \ell(w, x) = \tfrac{1}{2}(w - x)^2 \qquad \mathbb{P}_{W|X=x} = \delta_x$$

$X \perp\!\!\!\perp W_k$ for $k \leq k^\star$:



The first $k^\star$ terms in the chained sum are null, and $\mathcal{B}_{\mathsf{CWass}}/\mathcal{B}_{\mathsf{Wass}} \leq 2\theta$:

$$\begin{aligned}\mathcal{B}_{\mathsf{CWass}} &= \sum_{k>0} 2^{1-k}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k}] = \sum_{k>k^\star} 2^{1-k}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] \\ &\leq \sum_{k>k^\star} 2^{1-k}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W})] = 2\theta\mathcal{B}_{\mathsf{Wass}}\,.\end{aligned}$$

Exact results:

$$|\mathcal{G}| = \frac{1}{3}\,\theta^2 \simeq 0.33\,\theta^2\,; \qquad \mathcal{B}_{\mathsf{Wass}} = \frac{2}{3}\,\theta \simeq 0.67\,\theta\,;$$

$$\mathcal{B}_{\mathsf{CWass}} = \frac{247}{105}\,\theta^2 \simeq 2.35\,\theta^2; \qquad \mathcal{B}_{\mathsf{CMI}} \simeq 3.50\,\theta\,.$$

Conclusion

- We introduced a general framework allowing us to derive several generalisation bounds.
- Under suitable regularity conditions we established a duality between chained and unchained generalisation bounds.
- The same technical machinery can apply to broader settings, *e.g.* chained PAC-Bayes bound.