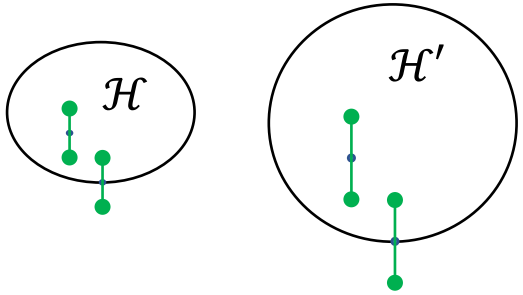
PAC-Bayesian Analysis

Yevgeny Seldin

PAC-Bayesian Analysis



- Selection → bias

- PAC-Bayesian analysis
  - Randomized classifiers → active avoidance of selection → reduced bias
  - The idea: instead of committing to a particular classifier, return a distribution over classifiers (avoid commitment)
  - For example: if two classifiers have the same empirical error, do not select among them, but return a 50/50 distribution
    - Stays at the same level of approximation error, but reduces the estimation error
  - Can be applied to uncountably infinite $\mathcal{H}$

Randomized Classifiers

- Let ρ be a distribution on $\mathcal{H}$

- Randomized classification:
  1. Sample $h \sim \rho(h)$
  2. Observe X
  3. Return $h(X)$

- ρ is a *randomized classifier* / *Gibbs classifier*

- Expected error: $\mathbb{E}_{h\sim\rho(h)}[L(h)]$
- Empirical error: $\mathbb{E}_{h\sim\rho(h)}\left[\hat{L}(h,S)\right]$

PAC-Bayes-kl inequality

- Theorem: For any "prior" distribution π on $\mathcal{H}$ that is independent of S

$$\mathbb{P}\left(\exists\rho\colon \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle|\middle|\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

- In other words, with probability at least $1-\delta$, for all ρ simultaneously

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle|\middle|\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n}$$

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

Understanding the bound

- Refined Pinsker's relaxation of kl:

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$

- Pick ρ that optimizes the trade-off between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathrm{KL}(\rho||\pi)$
  - $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ - assign high weight on h with small $\hat{L}(h,S)$
  - $\mathrm{KL}(\rho||\pi)$ - stay close to π in the KL sense
    - $\mathrm{KL}(\rho||\pi) = \sum_h \rho(h) \ln\frac{\rho(h)}{\pi(h)} = \sum_h \rho(h)\ln\frac{1}{\pi(h)} + \sum_h \rho(h)\ln\rho(h) = \sum_h \rho(h)\ln\frac{1}{\pi(h)} - \mathrm{H}(\rho)$
    - Distribute $\rho(h)$ uniformly among h with similar $\hat{L}(h,S)$ and $\pi(h)$ (maximize $\mathrm{H}(\rho)$)
  - Extreme case: if $\rho = \pi$, then $\mathrm{KL}(\rho||\pi) = 0$. No selection, no penalty!
  - Fast rates: small $\hat{L}(h,S)$ allows more aggressive deviation from π

PAC-Bayes vs. VC vs. VC+Occam

- PAC-Bayes (Pinsker's relaxation)

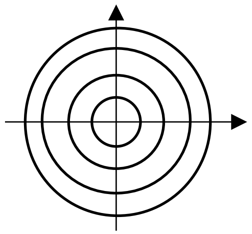
$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho[\hat{L}(h,S)] + \sqrt{\frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}$$

- VC (Empirical Risk Minimization – ERM)

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{8\ln\frac{2\left((2n)^{d_{\mathrm{VC}}(\mathcal{H})}+1\right)}{\delta}}{n}}$$

- VC+Occam (Structural Risk Minimization – SRM)

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{8\ln\frac{2\left((2n)^{d_{\mathrm{VC}}(\mathcal{H}_{j(h)})}+1\right)}{\pi(j)\delta}}{n}}$$



Special case: SVMs

$$L_{\mathrm{FAT}}(h_{w,b}) \leq \hat{L}_{\mathrm{FAT}}(h_{w,b},S) + \sqrt{\frac{8\ln\frac{2\left((2n)^{\lceil ||w||^2\rceil+1}+1\right)\lceil ||w||^2\rceil\left(\lceil ||w||^2\rceil+1\right)}{\delta}}{n}}$$

- PAC-Bayes
  - Complexity $\pi(h)$ defined for each h individually
  - Penalization by the actual selection $\mathrm{KL}(\rho||\pi)$
    - No selection, no penalty
    - $\rho = \pi \quad \Rightarrow \quad KL(\rho||\pi) = 0$
  - Possibility to incorporate prior knowledge via $\pi(h)$
- VC
  - Complexity $d_{VC}(\mathcal{H})$ defined for all $\mathcal{H}$
  - "Pre-paid" penalization $d_{VC}(\mathcal{H})$
    - Same cost irrespective of selection
  - No possibility to incorporate prior knowledge
- VC+Occam: in-between
  - Occam provides partial adaptivity across subsets of $\mathcal{H}$
  - $d_{VC}(\mathcal{H}_j)$ dominates the complexity
  - $\pi(j)$ allows to incorporate structural prior knowledge, but only limited data-dependent prior knowledge

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] || \mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

PAC-Bayes vs. Bayesian learning

- PAC-Bayesian bounds
  - $\pi(h)$ is an auxiliary construction in the proof (same as in Occam); the bounds always hold
  - High-probability guarantees on the distance between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathbb{E}_\rho[L(h)]$
  - Depend on the loss function $\ell(h(X),Y)$

- Bayesian learning
  - $\pi(h)$ is a prior "belief". The Bayes rule provides a way to update it to posterior "belief" given evidence (data)
  - No guarantees
  - Built for the log-loss

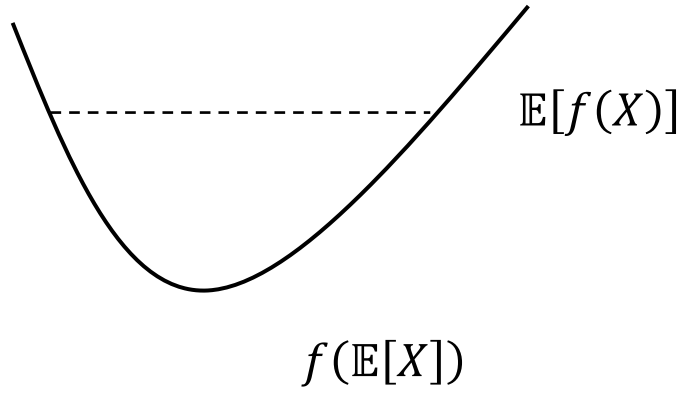
A proof of the PAC-Bayes-kl inequality

- Theorem: For any “prior” distribution π on $\mathcal{H}$ that is independent of S

$$\mathbb{P}\left(\exists\rho: \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle|\middle|\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

Proof tools – Jensen's inequality

- Jensen's inequality:
  For convex f: $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$



- Example:
  - $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$

Main proof tool – Change of measure inequality

- Lemma (Change of measure inequality):

  For any f, ρ, and π: $\mathbb{E}_\rho[f(h)] \leq KL(\rho||\pi) + \ln \mathbb{E}_\pi\left[e^{f(h)}\right]$

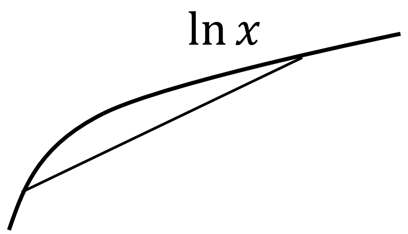
- Proof:

$$\mathbb{E}_\rho[f(h)] = \mathbb{E}_\rho\left[\ln\left(\frac{\rho(h)}{\pi(h)} e^{f(h)} \frac{\pi(h)}{\rho(h)}\right)\right]$$

$$= \mathrm{KL}(\rho||\pi) + \mathbb{E}_\rho\left[\ln\left(e^{f(h)} \frac{\pi(h)}{\rho(h)}\right)\right]$$

$$\underbrace{\leq}_{\text{Jensen}} \mathrm{KL}(\rho||\pi) + \ln \mathbb{E}_\rho\left[e^{f(h)} \frac{\pi(h)}{\rho(h)}\right]$$

$$= \mathrm{KL}(\rho||\pi) + \ln \mathbb{E}_\pi\left[e^{f(h)}\right]$$



- We obtain a deterministic relation between $\mathbb{E}_\rho[f(h)]$ and $\mathbb{E}_\pi\left[e^{f(h)}\right]$, no probabilities involved.

$$\mathbb{P}\left(\exists\rho\colon \mathrm{kl}\big(\mathbb{E}_\rho[\hat{L}(h,S)]||\mathbb{E}_\rho[L(h)]\big) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

Proof of PAC-Bayes-kl

Markov:

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon} = \delta$$

$$\mathbb{P}\left(X \leq \frac{1}{\delta}\mathbb{E}[X]\right) \geq 1 - \delta$$

- Change of measure inequality

  $\mathbb{E}_\rho[f(h)] \leq \mathrm{KL}(\rho||\pi) + \ln \mathbb{E}_\pi\left[e^{f(h)}\right]$

  Deterministic inequality relating all ρ to a single π

- PAC-Bayes Lemma

  $\mathbb{E}_\rho[f(h,S)] \leq \mathrm{KL}(\rho||\pi) + \ln \overbrace{\mathbb{E}_\pi\left[e^{f(h,S)}\right]}$ (deterministically $\forall\rho$)

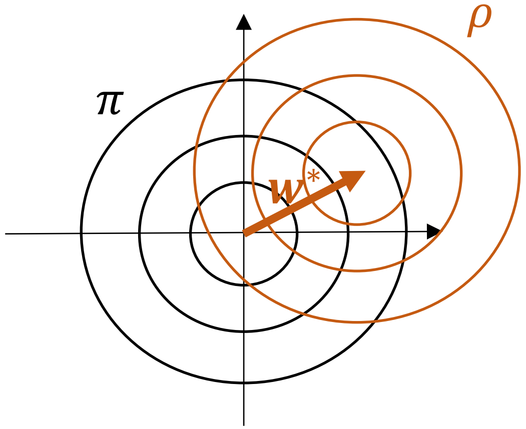
  $\underbrace{\leq}_{\substack{\text{Markov}\\ \text{w.p.}\geq 1-\delta}} \mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_S\left[\mathbb{E}_\pi\left[e^{f(h,S)}\right]\right]$ (single application of Markov to $\mathbb{E}_\pi\left[e^{f(h,S)}\right]$)

  $\underbrace{=}_{\pi \text{ independent of } S} \mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]$

- Choice of $f(h,S) = n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)$
  - kl-Lemma: $\mathbb{E}_S\left[e^{f(h,S)}\right] \leq 2\sqrt{n}$
  - $n\mathrm{kl}\big(\mathbb{E}_\rho[\hat{L}(h,S)]||\mathbb{E}_\rho[L(h)]\big) \underbrace{\leq}_{\substack{\text{convexity}\\ \text{of kl}}} \mathbb{E}_\rho\left[n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)\right] \underbrace{\leq}_{\substack{w.p.\geq 1-\delta\\ \text{PAC}-\text{Bayes lemma}}} \mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}$

Working with the bound

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \le \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$



- Select π. Example: $\pi(h_w) = \mathcal{N}(0, I)$
- Select ρ. Example: $\rho(h_w) = \mathcal{N}(w^*, I)$
- Calculate $\mathrm{KL}(\rho||\pi)$. In the example: $\mathrm{KL}(\rho||\pi) = ||w^*||^2$
- Calculate $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$
  - The challenging part

Modularity of the bound

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

- Different choices of $f(h,S)$ give divergence measures between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathbb{E}_\rho[L(h)]$
  - PAC-Bayes-kl: $f(h,S) = n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)$
    - kl-Lemma: $\mathbb{E}_S\left[e^{n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)}\right] \leq 2\sqrt{n}$
  - PAC-Bayes-Hoeffding: $f(h,S) = n\lambda\left(L(h) - \hat{L}(h,S)\right)$
    - Hoeffding's Lemma: $\mathbb{E}_S\left[e^{n\lambda\left(L(h)-\hat{L}(h,S)\right)}\right] \leq e^{\frac{n\lambda^2}{8}}$
- Different choices of ρ and π give different regularizations.
  - Gaussian prior and posterior $\rightarrow$ regularization by $||w||^2 = \sum_{i=1}^d w_i^2$
  - Laplacian prior and posterior $\rightarrow$ regularization by $||w||_1 = \sum_{i=1}^d |w_i|$

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

Minimization of the bound

- Relaxation: PAC-Bayes-λ (based on refined Pinsker's inequality)

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}$$

- For a fixed ρ convex in λ, for a fixed λ convex in ρ
- Apply alternating minimization
  - $\rho_\lambda^*(h) = \frac{\pi(h)e^{-n\lambda\hat{L}(h,S)}}{\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h,S)}\right]}$ (Gibbs distribution)
  - Holds for any $\mathcal{H}$, but computationally tractable only for finite $\mathcal{H}$
  - $\lambda_\rho^* = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{\mathrm{KL}(\rho||\pi)}+1}+1} \in (0,1]$
- The bound is *not* jointly convex in ρ and λ
  - Convergence to a *local* minimum, but in many practical cases still global

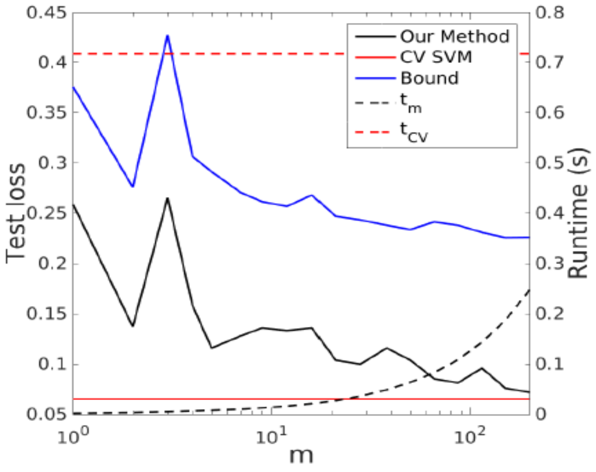
Construction of an interesting finite $\mathcal{H}$

- Train m classifiers on subsamples of size r
- Validate each classifier on the corresponding remaining $n - r$ samples
- Same as in cross-validation or bagging (e.g., random forests)
- Let $\hat{L}^{\mathrm{val}}(h, S)$ be the corresponding validation losses
- Adapted bound:

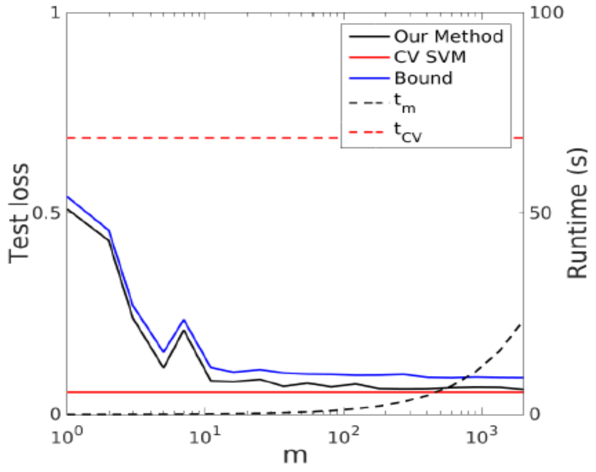
$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho\left[\hat{L}^{\mathrm{val}}(h,S)\right]}{1 - \frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho||\pi) + \ln \frac{2\sqrt{n-r}}{\delta}}{\lambda\left(1 - \frac{\lambda}{2}\right)(n-r)}$$

- Can be used to provide generalization bounds for random forests
- In the case of kernel SVMs provides computational speed-up when r is small ("inverse cross-validation") and $n > mr$
  - Kn^{2+} vs. $m(r^{2+} + r(n-r) + \mathrm{BoundOptimization})$

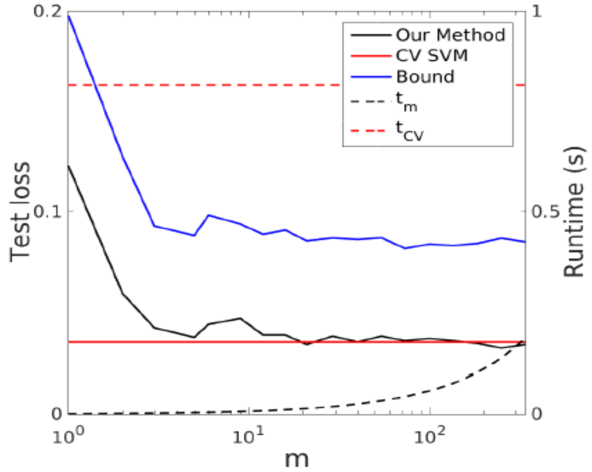
Empirical Evaluation – ensembles of small SVMs



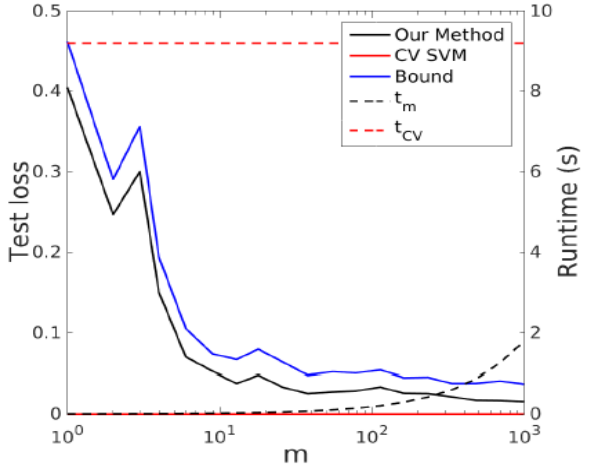
(a) Ionosphere dataset. $n = 200$, $r = d + 1 = 35$.



(b) Waveform dataset. $n = 2000$, $r = d + 1 = 41$.



(c) Breast cancer dataset. $n = 340$, $r = d + 1 = 11$.



(d) AvsB dataset. $n = 1000$, $r = d + 1 = 17$.

Summary

- PAC-Bayes-kl bound: with probability at least $1-\delta$, for all ρ

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

- Refined Pinsker's relaxation for intuition

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$

- Refined Pinsker's relaxation for alternating minimization

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}$$