PAC-Bayes under potentially heavy tails

Matthew J. Holland (Osaka University)

Background

PAC-Bayes theory is powerful, but relies on strong assumptions on the data.

Under **bounded** losses:

$$G_\rho \leq \widehat{G}_\rho + \sqrt{\frac{\boldsymbol{K}(\rho;\nu) + \log(2\sqrt{n}\delta^{-1})}{2n}}$$

(McAllester, 2003)

Under potentially **heavy-tailed** losses:

$$G_\rho \leq \widehat{G}_\rho + \left(\frac{\mathbf{E}_\nu\,|\widehat{R}-R|^q}{\delta}\right)^{\frac{1}{q}} \left(\int_{\mathcal{H}} \left(\frac{d\rho}{d\nu}\right)^p d\nu\right)^{\frac{1}{p}}$$

(Alquier and Guedj, 2018)

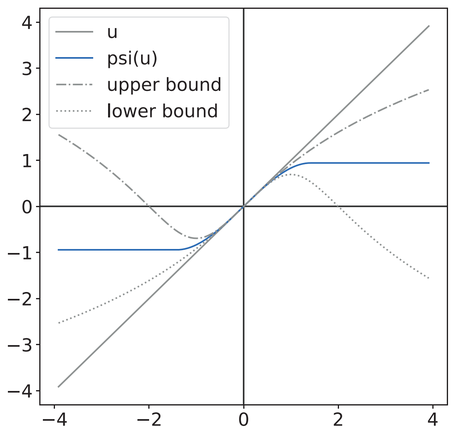
Goal: extend original framework to heavy-tailed case with near-optimality.

New estimator

Simple soft truncation:

$$\widehat{x} := \frac{s}{n}\sum_{i=1}^n \psi\left(\frac{x_i}{s}\right)$$

$$\psi(u) := \begin{cases} u - u^3/6, & -\sqrt{2} \leq u \leq \sqrt{2} \\ 2\sqrt{2}/3, & u > \sqrt{2} \\ -2\sqrt{2}/3, & u < -\sqrt{2} \end{cases}$$



Call on "PAC-Bayesian inequality" machinery.

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n f(x_i,\epsilon_i)\right) \leq \int \log \mathbf{E}_\mu \exp(f(x,\epsilon))\,d\rho(\epsilon) + \frac{\boldsymbol{K}(\rho;\nu)+\log(\delta^{-1})}{n}$$

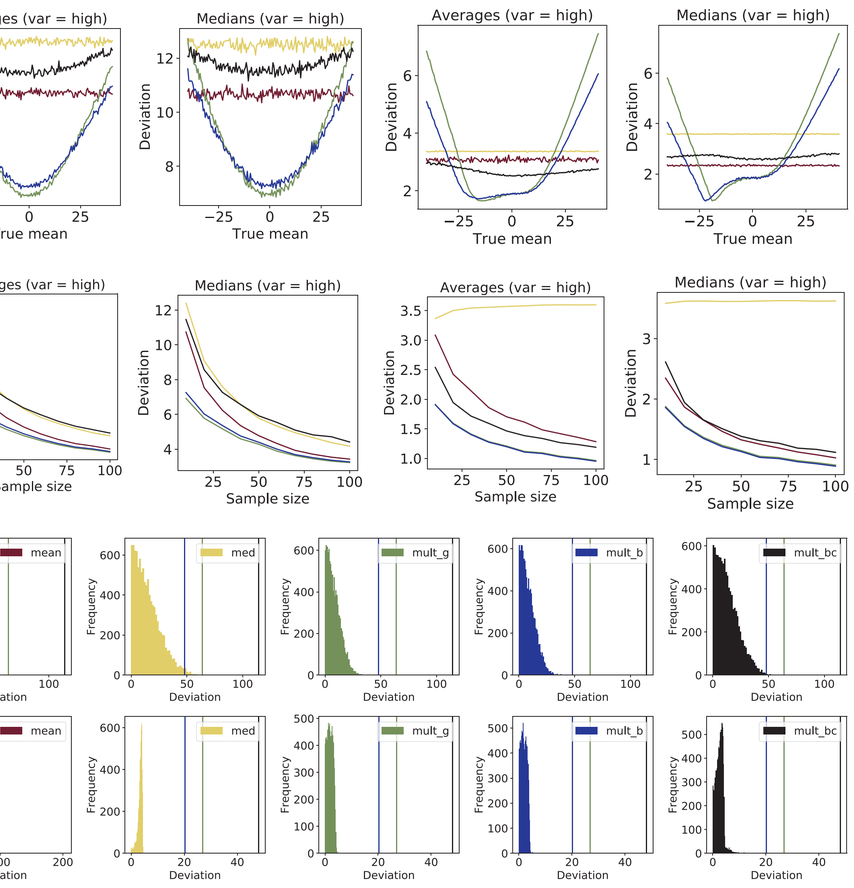
(cf. Catoni and Giulini, 2017)

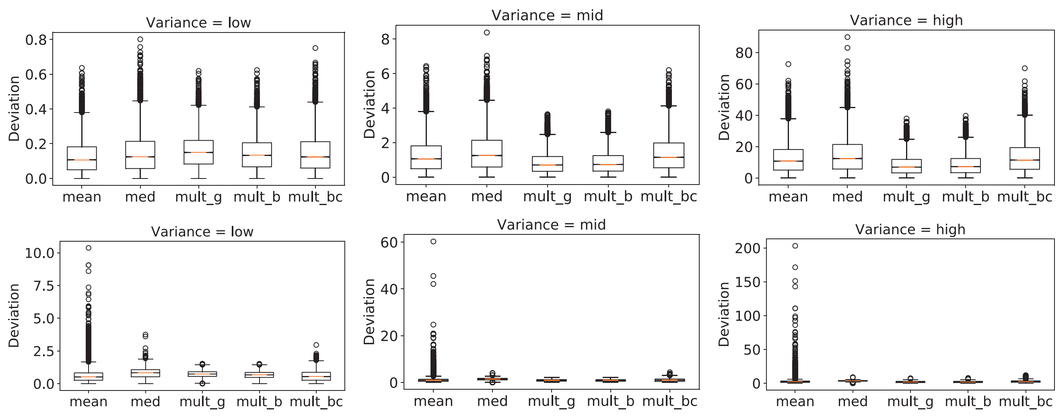
Next, make simple link with **"Bernoulli noise."**

$$\widehat{x} = \frac{1}{\theta}\,\mathbf{E}\left(\frac{s}{n}\sum_{i=1}^n \psi\left(\frac{x_i\,\epsilon_i}{s}\right)\right)$$

Lets us establish bounds using 2nd moments.

$$|\widehat{x} - \mathbf{E}_\mu\, x| \leq \sqrt{\frac{2\,\mathbf{E}_\mu\, x^2 \log(\delta^{-1})}{n}}$$





Applications to PAC-Bayes theory

Strategy: use robust estimators via PAC-Bayesian inequalities to robustify PAC-Bayesian theory.

Risk estimator: $\widehat{R}_\psi(h) := \frac{s}{n}\sum_{i=1}^n \psi\left(\frac{l(h;\boldsymbol{z}_i)}{s}\right)$

Pre-theorem (finite model):

$$R(h) \leq \widehat{R}_\psi(h) + \sqrt{\frac{2m_2(h)\,(\log(1/\nu(h)) + \log(1/\delta))}{n}}.$$

Proof:

Get high-prob bounds, pointwise in *h*.

$$R(h) \leq \frac{s}{n}\sum_{i=1}^n \psi\left(\frac{l(h;\boldsymbol{z}_i)}{s}\right) + \sqrt{\frac{2m_2(h)\log(\delta^{-1})}{n}}$$

Replace δ with $\nu(h)\delta$ and set error as

$$\varepsilon^*(h) := \sqrt{\frac{2m_2(h)\,(\log(1/\nu(h)) + \log(1/\delta))}{n}},$$

Finally take a union bound, showing

$$\mathbf{P}\left\{\exists\, h \in \mathcal{H} \text{ s.t. } R(h) - \widehat{R}_\psi(h) > \varepsilon^*(h)\right\} \leq 2\delta.$$

Theorem (infinite model):

$$G_\rho \leq \widehat{G}_{\rho,\psi} + \frac{1}{\sqrt{n}}\left(\boldsymbol{K}(\rho;\nu) + \frac{\log(8\pi M_2\delta^{-2})}{2} + M_2 + \nu_n^*(\mathcal{H}) - 1\right) + O\left(\frac{1}{n}\right)$$

KL div. of prior/posterior; 2nd moment bound; log-confidence term; Prior-dependent $\frac{\mathbf{E}_\nu \exp(\sqrt{n}(R-\widehat{R}_\psi))}{\mathbf{E}_\nu \exp(R-\widehat{R}_\psi)}$

minimize

Robust optimal Gibbs posterior

$$\left(\frac{d\widehat{\rho}}{d\nu}\right)(h) = \frac{\exp\left(-\sqrt{n}\widehat{R}_\psi(h)\right)}{\mathbf{E}_\nu \exp\left(-\sqrt{n}\widehat{R}_\psi\right)}.$$

Bounds established under **weak assumptions, but too sensitive to prior**. New machinery is desirable.