PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain†, Francis Bach†, Alexandre Lacoste‡, Simon Lacoste-Julien†

† INRIA Paris - École Normale Supérieure ‡ Google

Spoiler: Under the negative log-likelihood loss function, the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood.

PAC-Bayesian Theory

The PAC-Bayesian theory claims to provide "PAC guarantees to Bayesian algorithms" (McAllester, 1999). However, it is mostly used as a *frequentist* method.

Under a frequentist assumption...

The training set (X,Y) contains n ***i.i.d.* samples** from a data distribution $\mathcal{D}$.

...PAC-Bayes provides Probably Approximately Correct bounds...

With probability at least "$1-\delta$", the loss of predictor f is less than "ε",

$$\Pr_{X,Y\sim\mathcal{D}^n}\Big(\mathcal{L}_{\mathcal{D}}(f) \leq \varepsilon(\widehat{\mathcal{L}}_{X,Y}(f), n, \delta, \ldots)\Big) \geq 1-\delta\,.$$

...to Bayesian-like (averaged) predictors.

Given a prior π and a posterior $\hat\rho$ over a class of predictors $\mathcal{F}$,

$$\Pr_{X,Y\sim\mathcal{D}^n}\Big(\mathop{\mathbf{E}}_{f\sim\hat\rho}\mathcal{L}_{\mathcal{D}}(f) \leq \varepsilon\big(\mathop{\mathbf{E}}_{f\sim\hat\rho}\widehat{\mathcal{L}}_{X,Y}(f), n, \delta, \mathrm{KL}(\pi\|\hat\rho), \ldots\big)\Big) \geq 1-\delta\,.$$

where $\mathrm{KL}(\pi\|\hat\rho)$ is the **Kullback-Leibler divergence** between π and $\hat\rho$.

Two appealing aspects of PAC-Bayesian guarantees:

1. Data-driven generalization bounds computed on the training sample (*i.e.*, they do not rely on a testing sample) ;
2. Uniformly valid for all posteriors $\hat\rho$ over predictors class $\mathcal{F}$ (can be used as model selection criteria or optimized by a learning algorithm).

PAC-Bayesian Theorem for Bounded Losses

Given a loss function $\ell(f,x,y)\in[a,b]$, a predictor $f\in\mathcal{F}$, a data distribution $\mathcal{D}$, and a sample $(X,Y)=\{(x_i,y_i)\}_{i=1}^n\sim\mathcal{D}^n$,

$$\mathcal{L}_{\mathcal{D}}(f) \;\coloneqq\; \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}\ell(f,x,y)\,;\qquad \widehat{\mathcal{L}}_{X,Y}(f) \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n}\ell(f,x_i,y_i)\,.$$

Theorem ($^{\text{adapted}}_{\text{from}}$ Catoni, 2007). With probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,

$$\forall\hat\rho\text{ on }\mathcal{F}:\quad \mathop{\mathbf{E}}_{f\sim\hat\rho}\mathcal{L}_{\mathcal{D}}(f) \leq a + \frac{b-a}{1-e^{a-b}}\left[1-e^{\,a-\mathop{\mathbf{E}}_{f\sim\hat\rho}\widehat{\mathcal{L}}_{X,Y}(f)-\frac{1}{n}\left(\mathrm{KL}(\hat\rho\|\pi)+\ln\frac{1}{\delta}\right)}\right].$$

The bound suggests minimizing the following trade-off:

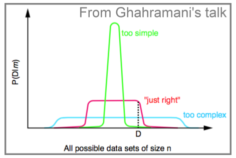
$$n\mathop{\mathbf{E}}_{f\sim\hat\rho}\widehat{\mathcal{L}}_{X,Y}(f) + \mathrm{KL}(\hat\rho\|\pi)\,.$$

Bayesian Model Selection

Bayesian Rule.
Consider a parameter set Θ.

For all $\theta\in\Theta$:

$$p(\theta|X,Y) = \frac{p(\theta)\,p(Y|X,\theta)}{p(Y|X)}$$



- $p(\theta|X,Y)$ is the posterior for each $\theta\in\Theta$ (similar to $\hat\rho$ over $\mathcal{F}$)
- $p(\theta)$ is the prior for each $\theta\in\Theta$ (similar to π over $\mathcal{F}$)
- $p(Y|X,\theta)$ is the *likelihood* of the parameter θ given the sample X,Y.
- $p(Y|X)$ is the marginal likelihood of Θ.

Bridging Bayes and PAC-Bayes

Negative log-likelihood loss function
Given a Bayesian likelihood $p(Y|X,\theta)$, let

$$\ell_{\mathrm{nll}}(\theta,x,y) \;=\; \ln\frac{1}{p(y|x,\theta)}\,.$$

The PAC-Bayesian and Bayesian posteriors align:

$$\underbrace{\hat\rho^*(\theta) = \frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}_{X,Y}^{\ell_{\mathrm{nll}}}(\theta)}}{Z_{X,Y}}}_{\text{PAC-Bayesian posterior}} = \underbrace{\frac{p(\theta)\,p(X,Y|\theta)}{p(Y|X)}}_{\text{Bayesian posterior}} = p(\theta|X,Y)\,.$$

The normalization constant *is* to the Bayesian *marginal likelihood*:

$$Z_{X,Y} \;=\; p(Y|X) \;=\; \int_{\Theta}\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}_{X,Y}^{\ell_{\mathrm{nll}}}(\theta)}d\theta\,.$$

Moreover,

$$-\ln Z_{X,Y} \;=\; n\mathop{\mathbf{E}}_{\theta\sim\hat\rho^*}\widehat{\mathcal{L}}_{X,Y}^{\ell_{\mathrm{nll}}}(\theta) + \mathrm{KL}(\hat\rho^*\|\pi)\,.$$

Thus, the following gives a PAC-Bayesian result based on the marginal likelihood $Z_{X,Y}$ of the optimal posterior $\hat\rho^*$.

Corollary. If $\ell_{\mathrm{nll}}(\cdot)\in[a,b]$, with probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,

$$\mathop{\mathbf{E}}_{\theta\sim\hat\rho^*}\mathcal{L}_{\mathcal{D}}^{\ell_{\mathrm{nll}}}(\theta) \;\leq\; a + \tfrac{b-a}{1-e^{a-b}}\left[1-e^{a}\sqrt[n]{Z_{X,Y}\,\delta}\right].$$

PAC-Bayesian Theorem for Unbounded Losses

Theorem (Alquier, Ridgway, Chopin, 2015). Let $\lambda>0$.
With probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,

$$\forall\hat\rho\text{ on }\mathcal{F}:\quad \mathop{\mathbf{E}}_{f\sim\hat\rho}\mathcal{L}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat\rho}\widehat{\mathcal{L}}_{X,Y}(f) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat\rho\|\pi) + \ln\frac{1}{\delta} + \Psi_{\ell,\pi,\mathcal{D}}(\lambda,n)\right],$$

$$\text{where}\quad \Psi_{\ell,\pi,\mathcal{D}}(\lambda,n) \;=\; \ln\mathop{\mathbf{E}}_{f\sim\pi}\mathop{\mathbf{E}}_{X',Y'\sim\mathcal{D}^n}\exp\left[\lambda\left(\mathcal{L}_{\mathcal{D}}(f)-\widehat{\mathcal{L}}_{X',Y'}(f)\right)\right].$$

Sub-gamma losses. The loss function ℓ is sub-gamma with a variance factor s^2 and scale parameter c, under a prior π and a data distribution $\mathcal{D}$, if it can be described by a sub-gamma random variable $V=\mathcal{L}_{\mathcal{D}}(f)-\ell(f,x,y)$, *i.e.*, its moment generating function is upper bounded by

$$\ln\mathbf{E}\,e^{\lambda V} = \ln\mathop{\mathbf{E}}_{f\sim\pi}\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}\exp\left[\lambda\left(\mathcal{L}_{\mathcal{D}}(f)-\ell(f,x,y)\right)\right] \;\leq\; \frac{\lambda^2 s^2}{2(1-c\lambda)}\,,\qquad \forall\lambda\in(0,\tfrac{1}{c})\,.$$

Corollary. If the loss is sub-gamma with variance factor s^2 and scale $c<1$, we have, With probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,

$$\forall\hat\rho\text{ on }\mathcal{F}:\quad \mathop{\mathbf{E}}_{f\sim\hat\rho}\mathcal{L}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat\rho}\widehat{\mathcal{L}}_{X,Y}(f) + \tfrac{1}{n}\left[\mathrm{KL}(\hat\rho\|\pi)+\ln\tfrac{1}{\delta}\right] + \tfrac{1}{2(1-c)}\,s^2\,.$$

As a special case, with $\ell\coloneqq\ell_{\mathrm{nll}}$ and $\hat\rho\coloneqq\hat\rho^*$, we have

$$\mathop{\mathbf{E}}_{\theta\sim\hat\rho^*}\mathcal{L}_{\mathcal{D}}^{\ell_{\mathrm{nll}}}(\theta) \;\leq\; \tfrac{s^2}{2(1-c)} - \tfrac{1}{n}\ln\left(Z_{X,Y}\,\delta\right).$$

Analysis of Model Selection

Consider a discrete set of L models $\{\mathcal{M}_i\}_{i=1}^L$ with parameters $\{\Theta_i\}_{i=1}^L$.

(PAC-)Bayesian Rule. For each model, the optimal posterior is

$$\hat\rho_i^*(\theta) \;=\; p(\theta|X,Y,\mathcal{M}_i) \;=\; \frac{p(\theta|\mathcal{M}_i)\,p(Y|X,\theta,\mathcal{M}_i)}{p(Y|X,\mathcal{M}_i)}\,.$$

$$p(Y|X,\mathcal{M}_i) \;=\; \int_{\Theta_i}p(\theta|\mathcal{M}_i)\,p(Y|X,\theta,\mathcal{M}_i)\,d\theta = Z_{X,Y,i}\ \text{is the \textit{model evidence}.}$$

Corollary. With probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,

$$\forall i\in\{1,\ldots,L\}\;:\qquad \mathop{\mathbf{E}}_{\theta\sim\hat\rho_i^*}\mathcal{L}_{\mathcal{D}}^{\ell_{\mathrm{nll}}}(\theta) \;\leq\; \tfrac{1}{2(1-c)}\,s^2 - \tfrac{1}{n}\ln\left(Z_{X,Y,i}\,\tfrac{\delta}{L}\right).$$

Provide a new interpretation of the Bayesian Occam's razor criteria!
To improve bounds, perform model averaging ($\Rightarrow$ *hierarchical Bayes*.)

Experiments with Bayesian Linear Regression

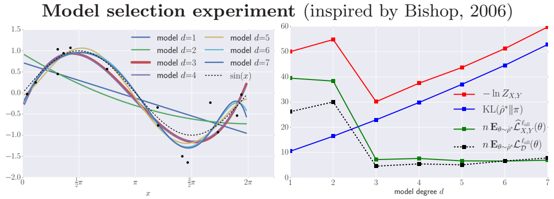
We consider a mapping function $\boldsymbol{\phi}:\mathbb{R}\to\mathbb{R}^d$, model parameters $\theta\coloneqq\mathbf{w}\in\mathbb{R}^d$, and noise σ.
Under the likelihood $p(y|x,\mathbf{w})=\mathcal{N}(y|\mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x}),\sigma^2)$, the negative log-likelihood loss function is

$$\ell_{\mathrm{nll}}(\mathbf{w},x,y) \;=\; -\ln p(y|x,\mathbf{w}) \;=\; \tfrac{1}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2\sigma^2}(y-\mathbf{w}\cdot\boldsymbol{\phi}(x))^2$$

For the Gaussian prior $p(\mathbf{w}|\sigma_\pi) = \mathcal{N}(\mathbf{0},\sigma_\pi\mathbf{I})$. the *optimal posterior* is given by $p(\mathbf{w}|\sigma,\sigma_\pi)=\mathcal{N}(\mathbf{w}\,|\,\widehat{\mathbf{w}},A^{-1})$, The negative log **marginal likelihood** is

$$-\ln Z_{X,Y} = \underbrace{n\,\widehat{\mathcal{L}}_{X,Y}^{\ell_{\mathrm{nll}}}(\widehat{\mathbf{w}}) + \tfrac{1}{2\sigma^2}\,\mathrm{tr}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}A^{-1})}_{n\,\mathbf{E}_{\mathbf{w}\sim\hat\rho^*}\widehat{\mathcal{L}}_{X,Y}^{\ell_{\mathrm{nll}}}(\mathbf{w})} + \underbrace{\tfrac{1}{2\sigma_\pi^2}\,\mathrm{tr}(A^{-1}) - \tfrac{d}{2} + \tfrac{1}{2\sigma_\pi^2}\|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2}\log|A| + d\ln\sigma_\pi}_{\mathrm{KL}\left(\mathcal{N}(\widehat{\mathbf{w}},A^{-1})\,\|\,\mathcal{N}(\mathbf{0},\sigma_\pi^2\mathbf{I})\right)}\,.$$

Model selection experiment (inspired by Bishop, 2006)



Comparison of bound values