On the properties of variational approximations of Gibbs posteriors

James Ridgway
Joint work with: P. Alquier, N. Chopin

University of Bristol

March 2016

The problem at hand

Aim: We want to learn from a given sample without any asumption on the likelihood This makes sense in particular when the design follows a complex generative model (i.e. images, text etc.)
To this extend we define the framework as follows:

Statistical Learning model (classification)

- A collection of labeled random variables $(Y_1, X_2), (Y_2, X_2), \dots$ where $(Y_i, X_i) \in \{-1, 1\} \times \mathcal{X}$ in this talk we suppose $(X_i, Y_i) \overset{\text{iid}}{\sim} \mathbb{P}$
- A collection prediction function $\{f_\theta, \theta \in \Theta\}$

$$f_\theta : \mathcal{X} \mapsto \{-1, 1\}$$

In this talk we can assume a linear model $f_\theta(x) = 2\mathbb{1}_{x\theta > 0} - 1$

The problem at hand

Statistical Learning (continued)

A loss function

$$\ell : \{-1,1\} \times \{-1,1\} \mapsto \mathbb{R}_+$$

to wich we associate. Example: $\ell(y, f_\theta(x)) = \mathbb{1}_{y \neq f_\theta(x)}$ the 0-1 loss.

- A theoretical risk $R(\theta) := \mathbb{E}\ell(Y, f_\theta(X))$
- A emprical risk $R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$

The final goal is to find a minimizer to $R(\theta)$.

The PAC solution to the problem

Define the 0-1 loss $\ell(y, y') := \mathbb{1}_{y \neq y'}$

Theorem (Vapnik [2000])

Suppose the above model with a 0-1 loss, and the linear classifier, $\Theta = \mathbb{R}^d$ and

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} R_n(\theta)$$

then $\forall \epsilon > 0$ with probability at least $1 - \epsilon$,

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 4\sqrt{\frac{(d+1)\log(n+1) + \log 2}{n}} + \sqrt{\frac{\log(2/\epsilon)}{2n}}$$

A Bayesian solution

PAC-Bayesian bounds

"PAC-Bayesian learning methods combine the informative priors of Bayesian methods with distribution-free PAC guarantees [...] The Bayesian approach has the advantage of using arbitrary knowledge in the form of a prior " McAllester [1998]

Define a prior measure $\pi \in \mathcal{M}_1^+(\Theta)$ the set of probability measures on Θ
We are going to use a Gibbs posterior with the risk as negative energy,

$$\pi_\lambda(\mathrm{d}\theta|\mathcal{D}) := \frac{1}{Z_\pi} e^{-\lambda R_n(\theta)} \pi(\mathrm{d}\theta)$$

and where $Z_\pi := \int_\Theta e^{-\lambda R_n(\theta)} \pi(\mathrm{d}\theta)$, and $\mathcal{D} := \{(Y_1, X_1), \cdots, (Y_n, X_n)\}$

PAC-Bayesian bounds in practice (1)

Aim: Show oracle inequalities and empirical bounds under the above framework. Consider the following modification of proposition 5.2 in Catoni [2004].

PAC-Bayesian oracle inequality

For a $0-1$ loss, for any $\epsilon > 0$ with probability at least $1-\epsilon$,

$$\int R \mathrm{d}\pi_\lambda(\mathrm{d}\theta|X) \leq \mathcal{B}_\lambda(\mathcal{M}_1^+)$$

$$:= \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int R \mathrm{d}\rho + \frac{\lambda}{n} + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}$$

PAC-Bayesian bounds in practice (2)

Gibbs measures in practice

- We would like similar results for computable estimators.
- Past implementation of the results rely mostly on MCMC
  - RJMCMC in Alquier and Biau [2013],
  - Unajusted Langevin in Dalalyan and Tsybakov [2008].
- For some non-asymptotic studies of properties of MCMC see Dalalyan [2014], Durmus and Moulines [2015] and others.

Goal

Ultimately we want to find polynomial time algorithm in the dimension (i.e. an algorithm that stops after a number of given steps that is a polynomial of the dimension).

Crash course in variational approximation

Instead let's have a look at Variational Bayes

Minimizing the KL divergence

Define

$$\rho_\lambda^{vb} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_\lambda(.|\mathcal{D}))$$

where

- $\mathcal{K}$ denotes the KL divergence $\mathcal{K}(\mu, \nu) = \int \mu(\mathrm{d}x) \log \frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}$ if $\nu >> \mu$, ∞ otherwise.
- $\mathcal{F}$ is a family of probability measures.

The choice of the of the family $\mathcal{F}$ will strongly influence the quality of the approximation. Two examples,

- $\mathcal{F}^\Phi = \left\{\Phi_{m,\Sigma}, m \in \mathbb{R}^d, \Sigma \in \mathcal{S}^d\right\}$ the set of Gaussian measures.
- $\mathcal{F}^{mf} = \left\{\rho \in \mathcal{M}_+^1(\Theta) \text{ s.t. } \rho(\mathrm{d}\theta) = \prod_{i \in J} \rho_i(\mathrm{d}\theta_i)\right\}$ the set of factorizable measures on a set of indices J.

Main result

Aim: Find a PAC-Bayesian bound for the Gaussian approximation.

Theorem

Using the 0-1 *loss, for any* $\varepsilon > 0$, *with probability at least* $1-\varepsilon$ *we have*

$$\int R \mathrm{d}\rho_\lambda^{vb} \leq \mathcal{B}_\lambda(\mathcal{F}) := \inf_{\rho\in\mathcal{F}} \left\{ \int R \mathrm{d}\rho + \frac{\lambda}{n} + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}.$$

Moreover,

$$\mathcal{B}_\lambda(\mathcal{F}) = \mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) + \frac{2}{\lambda}\inf_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{\frac{\lambda}{2}}), \text{ where } \pi_\lambda(\mathrm{d}\theta) \propto e^{-\lambda R(\theta)}\pi(\mathrm{d}\theta)$$

Application: 0-1 loss

Let's take the special case

- $\ell(y, y') = \mathbb{1}_{y \neq y'}$,
- with prior $\pi(\mathrm{d}\theta) = \mathcal{N}(0, \vartheta I)$
- Using the linear classifier $f_\theta(x) = 2\mathbb{1}_{\theta x > 0} - 1$

Corollary

Assume that the VB approximation is done on $\mathcal{F}^\Phi$, Take $\lambda = \sqrt{nd}$ and $\vartheta = \frac{1}{\sqrt{d}}$. Under some necessary assumption, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously

$$\left.\begin{matrix} \int R(\theta)\pi_\lambda(\mathrm{d}\theta|X) \\ \int R(\theta)\mathrm{d}\rho_\lambda^{vb}(\theta) \end{matrix}\right\} \leq \inf_{\theta\in\Theta} R(\theta) + \mathcal{O}\left(\sqrt{\frac{d}{n}}\log(n)\right) + \frac{2}{\sqrt{nd}}\log\frac{2}{\epsilon}$$

Application: 0-1 loss

We end up solving the following optimization problem:

$$(\hat{m}, \hat{\Sigma}) \in \arg \min_{m,\Sigma \in \mathbb{R}^d \times \mathcal{S}^+} \mathcal{L}_{\lambda,\vartheta}(m, \Sigma),$$

where $\mathcal{L}_{\lambda,\vartheta}(m, \Sigma) = -\frac{\lambda}{n} \sum_{i=1}^{n} \Phi \left(-Y_i \frac{X_i m}{\sqrt{X_i \Sigma X_i^t}} \right) - \frac{m^T m}{2\vartheta} + \frac{1}{2} \left(\log|\Sigma| - \frac{1}{\vartheta} \mathrm{tr} \Sigma \right)$

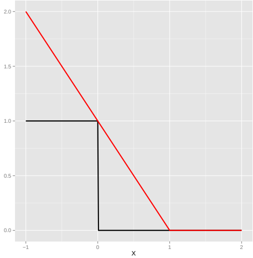
Optimizing the bound in practice

- The previous results tels us that $\int R(\theta) \mathrm{d}\Phi_{\hat{m},\hat{\Sigma}}(\theta)$ will converge to the oracle risk at a quantifiable rate.
- However optimizing $\mathcal{L}_{\lambda,\vartheta}$ is difficult (impossible ?) in practice.
- The target is nonconvex and in general multimodal

Application: Hinge loss

The usual way to deal with this is to use a convex upper bound on the loss.

- Hinge loss: $max(0, 1 - yf_\theta(x))$
- Prediction function: $f_\theta(x) = x^t\theta$
- with prior $\pi(\mathrm{d}\theta) = \mathcal{N}(0, \vartheta I)$



Hinge loss

- We now have a convex loss that will lead to convex optimization procedures
- We can use theoretical results from the optimization community to bound the numerical error

VB in practice

To minimize the KL divergence over $\mathcal{F}_1 = \left\{\Phi_{m,\sigma I}, m \in \mathbb{R}^d, \sigma \in \mathbb{R}_+\right\}$
On needs to minimize the following objective,

$$\mathcal{L}(\mathbf{m},\sigma) = -\frac{\lambda}{n}\left\{\sum_{i=1}^{n}(1-\Gamma_i\mathbf{m})\,\Phi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right) + \sum_{i=1}^{n}\sigma\|\Gamma_i\|\varphi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right)\right\} - \frac{\|\mathbf{m}\|_2^2}{2\vartheta} + \frac{d}{2}\left(\log\sigma^2 - \frac{\vartheta}{\sigma^2}\right).$$

The optimal mean and variance are given by

$$(m^\star,\sigma^\star) = \arg\min_{\mathbf{m}\in\mathbb{R}^d,\sigma>0}\mathcal{L}(\mathbf{m},\sigma).$$

Define $\rho^{vb}_{\lambda,k}$ the approximation formed of the mean and variance $(\mathbf{m}_k,\sigma_k)$ given by the k-th iterate of a gradient descent.

$$(\mathbf{m}_{k+1},\sigma_{k+1}) = (\mathbf{m}_k,\sigma_k) - \alpha\nabla\mathcal{L}(\mathbf{m}_k,\sigma_k)$$

Oracle bound for the Hinge loss

Using results from convex optimization [Nesterov, 2004] we can bound the risk integrated with respect to the approximation obtain after a fixed number of iteration of the solver.

Theorem

Assume that the VB approximation is done on $\mathcal{F}^{\Phi}$. Denote by $\rho^{vb}_{\lambda,k}(\mathrm{d}\theta)$ the VB approximated measure after the kth iteration of an optimal convex solver using the hinge loss. Take $\lambda = \sqrt{nd}$ and $\vartheta = \frac{1}{\sqrt{d}}$ then under the correct hypotheses with probability $1-\epsilon$

$$\int R^H \mathrm{d}\rho^{vb}_{\lambda,k} \leq \inf_{\theta\in\Theta} R^H + \frac{LM}{\sqrt{1+k}} + \mathcal{O}\left(\sqrt{\frac{d}{n}\log\frac{n}{d}}\right) + 2\frac{c_x}{\sqrt{nd}}\log\frac{2}{\epsilon}$$

where L is the Lipschitz coefficient on a ball of radius M of the objective function maximized in VB.

Numerical Application

Dataset	Covariates	Full cov. ($\mathcal{F}_3$)	SMC	VB Hinge	SVM linear
Pima	7	21.3	22.3	19.5	21.6
German Credit	60	33.6	32.0	26.2	33.2
DNA	180	23.6	23.6	4.2	5.1
SPECTF	22	06.9	08.5	10.1	21.4
Glass	10	19.6	23.3	2.8	6.5
Indian	11	25.5	26.2	25.5	25.3
Breast	10	1.1	1.1	0.5	1.7

Table : Comparison of misclassification rates (%).

Other losses

- We can get similar results for:
  - Ranking using AUC risk (application of stochastic variational Bayes).
  - Matrix completion (application with family $\mathcal{F}^{mf}$)

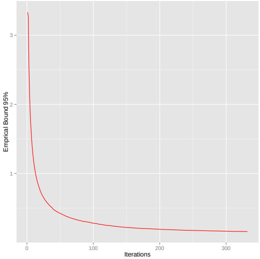


Figure : Error bound at each iteration, stochastic descent, Adult datasets.

Stochastic VB with fixed temperature $\lambda = 1000$, batch size of 50. The adult dataset has $n = 32556$ observation and $n_+ n_- = 193,829,520$ possible pairs. The convergence is obtained in order of seconds. The bounds are the empirical bounds obtained for a probability of 95%.

Closing remarks

- Development of R package (PACVB) to perform Hinge loss VB and a Hinge version of bipartite ranking. Available on the CRAN repository.
- Other question are still open
  - Can we do better than cross-validation for the choice of λ?
  - Online learning ?

Thank you for your attention!

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(1):243–280, 2013.

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. *arXiv preprint arXiv:1306.6430*, 2013.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

N. Chopin and J. Ridgway. Leave pima indians alone: binary regression as a benchmark for bayesian computation. *arXiv:1506.08640*, 2015.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave distribution. *arXiv preprint arXiv:1412.7392*, 2014.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted langevin algorithm. *arXiv:1507.05021*, 2015.

P.D. Grünwald and T. van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *arXiv preprint*, 2014.

D.A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, New York, 1998.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

V. Vapnik. *The nature of statistical learning*. Springer, 2000.