# 9.520: Class 11

# Bayesian Interpretations of Regularization

*Tomaso Poggio*

# Plan

- Bayesian interpretation of Regularization
- Bayesian interpretation of the regularizer
- Bayesian interpretation of quadratic loss
- Bayesian interpretation of SVM loss

# Bayesian Interpretation of RN, SVM, and BPD in Regression

Consider

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_K^2$$

We will show that there is a Bayesian interpretation of RN in which the data term – that is the term with the loss function – is a model of the noise and the stabilizer is a prior on the hypothesis space of functions $f$.

# Definitions

1. $D_\ell = \{(\mathbf{x}_i, y_i)\}$ for $i = 1, \cdots, \ell$ is the set of training examples

2. $\mathcal{P}[f|D_\ell]$ is the conditional probability of the function $f$ given the examples $g$.

3. $\mathcal{P}[D_\ell|f]$ is the conditional probability of $g$ given $f$, i.e. a model of the noise.

4. $\mathcal{P}[f]$ is the *a priori* probability of the random field $f$.

# Posterior Probability

The posterior distribution $\mathcal{P}[f|g]$ can be computed by applying Bayes rule:

$$\mathcal{P}[f|D_\ell] = \frac{\mathcal{P}[D_\ell|f]\ \mathcal{P}[f]}{P(D_\ell)}.$$

If the noise is normally distributed with variance $\sigma$, then the probability $\mathcal{P}[D_\ell|f]$ is

$$\mathcal{P}[D_\ell|f] = \frac{1}{Z_L} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathrm{x}_i))^2}$$

where $Z_L$ is a normalization constant.

# Posterior Probability

Informally (we will make it precise later), if

$$\mathcal{P}[f] = \frac{1}{Z_r} e^{-\|f\|_K^2}$$

where $Z_r$ is another normalization constant, then

$$\mathcal{P}[f|D_\ell] = \frac{1}{Z_D Z_L Z_r} e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \|f\|_K^2\right)}$$

# MAP Estimate

One of the several possible estimates of $f$ from $\mathcal{P}[f|D_\ell]$ is the so called MAP estimate, that is

$$\max \mathcal{P}[f|D_\ell] = \min \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2 + 2\sigma^2 \|f\|_K^2 \ .$$

which is the same as the regularization functional if

$$\lambda = 2\sigma^2/\ell.$$

# Bayesian Interpretation of the Data Term (quadratic loss)

As we just showed, the quadratic loss (the standard RN case) corresponds in the Bayesian interpretation to assuming that the data $y_i$ are affected by additive independent Gaussian noise processes, i.e. $y_i = f(x_i) + \epsilon_i$ with $E[\epsilon_j \epsilon_j] = 2\delta_{i,j}$

$$P(\mathbf{y}|f) \propto \text{exp}(-\sum(y_i - f(x_i))^2)$$

## Bayesian Interpretation of the Stabilizer

The stabilizer $\|f\|_K^2$ is the same for RN and SVM. Let us consider the corresponding prior in a Bayesian interpretation within the framework of RKHS:

$$P(f) = \frac{1}{Z_r} \text{exp}(-\|f\|_K^2) \propto \text{exp}(-\mathbf{c}^T \mathbf{K} \mathbf{c}).$$

The most likely hypotheses are the ones with small RKHS norm.

## Bayesian Interpretation of RN and SVM.

- For SVM the prior is the same Gaussian prior, but the noise model is different and is NOT Gaussian additive as in RN.

- Thus also for SVM (regression) the prior $P(f)$ gives a probability measure to $f$ in terms of the the norm in the RKHS defined by $K$.

# Why a Bayesian Interpretation can be Misleading

Minimization of functionals such as $H_{RN}(f)$ and $H_{SVM}(f)$ can be interpreted as corresponding to the MAP estimate of the posterior probability of $f$ given the data, for certain models of the noise and for a specific Gaussian prior on the space of functions $f$.

Notice that a Bayesian interpretation of this type is *inconsistent* with Structural Risk Minimization and more generally with Vapnik's analysis of the learning problem. Let us see why (Vapnik).

# Why a Bayesian Interpretation can be Misleading

Consider regularization (including SVM). The Bayesian interpretation with a MAP estimates leads to

$$\min H[f] = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \frac{1}{\ell} 2\sigma^2 \|f\|_K^2 \ .$$

Regularization (in general and as implied by VC theory) corresponds to

$$\min H_{RN}[f] = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_K^2 \ .$$

where $\lambda$ is found by solving the Ivanov problem

$$\min \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2$$

subject to

$$\|f\|_K^2 \leq A$$

# Why a Bayesian Interpretation can be Misleading

The parameter $\lambda$ in regularization and SVM is a function of the data (through the SRM principle) and in particular is $\lambda(\ell)$. In the Bayes interpretation $\tilde{\lambda}$ depends on the data as $\frac{2\sigma^2}{\ell}$: notice that $\sigma$ has to be part of the prior and therefore has to be independent of the size $\ell$ of the training data. It seems unlikely that $\lambda$ could simply depend on $\frac{1}{\ell}$ as the Bayesian interpretation requires for consistency. For instance note that in the statistical interpretation of classical regularization (Ivanov, Tikhonov, Arsenin) the asymptotic dependence of $\lambda$ on $\ell$ is different from the one dictated by the Bayesian interpretation. In fact (Vapnik, 1995, 1998)

$$\lim_{\ell \to \infty} \lambda(\ell) = 0$$

$$\lim_{\ell \to \infty} \ell\lambda(\ell) = \infty$$

implying a dependence of the type $\lambda(\ell) = O(log\ell/\ell)$. A similar dependence is probably implied by results of Cucker and Smale, 2002. Notice that this is a sufficient and not a necessary condition. Here an interesting question (a project?): which $\lambda$ dependence does stability imply?

# Bayesian Interpretation of the Data Term (nonquadratic loss)

To find the Bayesian interpretation of the SVM loss, we now assume a more general form of noise. We assume that the data are affected by additive independent noise sampled form a continuous mixture of Gaussian distributions with variance $\beta$ and mean $\mu$ according to

$$P(\mathbf{y}|f) \propto \exp\left(-\int_0^{\infty} d\beta \int_{-\infty}^{\infty} d\mu \sqrt{\beta} e^{-\beta(y-f(x)-\mu)^2} P(\beta,\mu)\right),$$

The previous case of quadratic loss corresponds to

$$P(\beta,\mu) = \delta\left(\beta - \frac{1}{2\sigma^2}\right)\delta(\mu).$$

# Bayesian Interpretation of the Data Term (absolute loss)

To find $P(\beta, \mu)$ that yields a given loss function $V(\gamma)$ we have to solve

$$V(\gamma) = -\log \int_0^\infty d\beta \int_{-\infty}^\infty d\mu \sqrt{\beta} e^{-\beta(\gamma-\mu)^2} P(\beta, \mu),$$

where $\gamma = y - f(x)$.

For the absolute loss function $V(\gamma) = |\gamma|$. Then

$$P(\beta, \mu) = \beta^{-2} e^{-\frac{1}{4\beta}} \, \delta(\mu).$$

For unbiased noise distributions the above derivation can be obtained via the inverse Laplace transform.

# Bayesian Interpretation of the Data Term (SVM loss)

Consider now the case of the SVM loss function $V_\epsilon(\gamma) = \max\{|\gamma| - \epsilon, 0\}$. To solve for $P_\epsilon(\beta, \mu)$ we assume independence

$$P_\epsilon(\beta, \mu) = P(\beta) P_\epsilon(\mu).$$

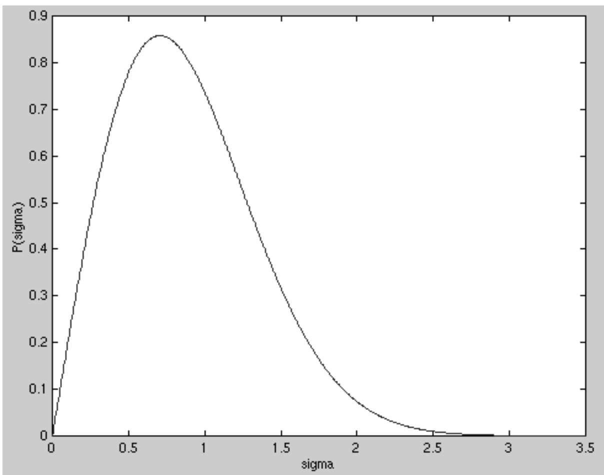
Solving

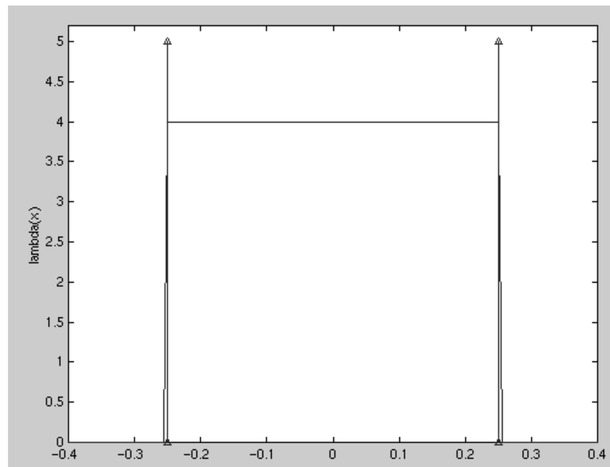
$$V_\epsilon(\gamma) = -\log \int_0^\infty d\beta \int_{-\infty}^\infty d\mu \sqrt{\beta} e^{-\beta(\gamma-\mu)^2} P(\beta) P_\epsilon(\mu)$$

results in

$$\begin{aligned} P(\beta) &= \beta^{-2} e^{-\frac{1}{4\beta}} \, , \\ P_\epsilon(\mu) &= \frac{1}{2(\epsilon+1)} \Big( \chi_{[-\epsilon,\epsilon]}(\mu) + \delta(\mu - \epsilon) + \delta(\mu + \epsilon) \Big) \, . \end{aligned}$$

# Bayesian Interpretation of the Data Term (SVM)



# Bayesian Interpretation of the Data Term (SVM loss and absolute loss)

Note $\lim_{\epsilon \to 0} V_\epsilon = |\gamma|$

So

$$P_0(\mu) = \frac{1}{2}\Big(\chi_{[-0,0]}(\mu) + \delta(\mu) + \delta(\mu)\Big) = \delta(\mu)$$

and

$$P(\beta, \mu) = \beta^{-2} e^{-\frac{1}{4\beta}}\ \delta(\mu),$$

as is the case for absolute loss.