

Chapter 14

Probabilistic methods

Chapter summary

- Probabilistic models can lead to intuitive algorithmic formulations but sometimes misleading interpretations. In particular, maximum a posteriori (MAP) estimation does *not* work best when the parameters are generated from the prior distribution. Minimum mean-square estimation (MMSE) is preferable (for the square loss).
- Generative models (such as linear discriminant analysis) that explicitly try to model the input data with simple models can lead to biased but efficient estimators in large dimensions compared to their discriminative counterparts (such as logistic regression).
- Bayesian inference can be used naturally for model selection using the marginal likelihood, both for model selection among a finite number of choices or with Gaussian processes.
- PAC-Bayesian analysis: aggregating estimators provide natural statistically efficient estimators with an elegant link with Bayesian inference.

In this chapter, we first consider probabilistic modeling interpretations of several learning methods, focusing primarily on identifying losses and priors with log-densities but drawing clear distinctions between what this analogy brings and what it does not. We then show how Bayesian inference naturally leads to model selection criteria and end the chapter with a description of PAC-Bayes analysis.

14.1 From empirical risks to log-likelihoods

Many methods in machine learning may be given a probabilistic interpretation through maximum likelihood or “maximum a posteriori” (MAP) estimation. For example, con-

sider the regularized empirical risk as:

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i)) + \frac{\lambda}{n} \Omega(\theta),$$

multiply by $-n$ and take the exponential to get:

$$\begin{aligned} \exp(-n\hat{\mathcal{R}}(\theta)) &= \exp\left(-\sum_{i=1}^n \ell(y_i, f_{\theta}(x_i)) - \lambda\Omega(\theta)\right) \\ &= \prod_{i=1}^n \exp[-\ell(y_i, f_{\theta}(x_i))] \cdot \exp[-\lambda\Omega(\theta)]. \end{aligned} \quad (14.1)$$

We can give a probabilistic interpretation by considering a *likelihood*, that is, a density (with respect to a well-defined base measure),

$$p(y_i|x_i, \theta) \propto \exp[-\ell(y_i, f_{\theta}(x_i))],$$


and a *prior* density


$$p(\theta) \propto \exp[-\lambda\Omega(\theta)],$$

so that we have:

$$\exp(-n\hat{\mathcal{R}}(\theta)) \propto \prod_{i=1}^n p(y_i|x_i, \theta) \cdot p(\theta),$$

which is precisely the (conditional) likelihood for the model where θ is a parameter and where, given θ , all pairs (x_i, y_i) are independent and identically distributed.

 Overloading of notations for probability densities, where the symbol p is used for all random variables.

 Note the difference between conditional likelihood and likelihood.



There is more to probabilistic interpretation than simply taking the exponential, e.g., generative models, Bayesian inference for hyperparameter learning (as done in later sections), dealing with missing data through the expectation-maximization algorithm, etc.



We only scratch the surface here, and from a learning theory point of view. See [Murphy \(2012\)](#); [Bishop \(2006\)](#) for many more details.

In this section, we primarily focus on the formulation in Eq. (14.1) and now look at specific examples for data likelihoods and priors.

14.1.1 Conditional likelihoods

For logistic regression where $\mathcal{Y} \in \{-1, 1\}$, we can interpret the loss as the conditional log-likelihood of the model where

$$\mathbb{P}(y_i = 1|x_i) = \frac{1}{1 + \exp(-f_\theta(x_i))},$$

which can be put in a compact way as $p(y_i|x_i) = \frac{1}{1+\exp(-y_i f_\theta(x_i))} = \sigma(y_i f_\theta(x_i))$.



To apply logistic regression, there is no need to assume that the model is well-specified, that is, there exists a θ_* so that the data are actually generated from the model above. For the non-parametric analysis, this is often assumed.

For least-squares regression, we can interpret the loss $\frac{1}{2}(y_i - f_\theta(x_i))^2$ as a Gaussian model with mean $f_\theta(x_i)$ and variance 1. We can also estimate a more general variance parameter that is uniform across all x (homoscedastic regression) or depends on x (heteroscedastic regression).



No need to have Gaussian noise! Simply zero mean and bounded variance are enough for the analysis.

Exercise 14.1 Show that the negative log-density of the Gaussian distribution with mean μ and variance σ^2 , that is, $-\log p(y|\mu, \sigma) = \frac{1}{2\sigma^2}(x-\mu)^2 + \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\sigma^2$ is not convex in (μ, σ^2) , but is jointly convex in $(\mu/\sigma^2, \sigma^{-2})$.

14.1.2 Classical priors

We can interpret classical regularizers that we have already encountered in previous chapters. For the squared ℓ_2 -norm with $\Omega(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, this corresponds to a Gaussian distribution with mean zero and covariance matrix $\lambda^{-1}I$.

For the ℓ_1 -norm with $\Omega(\theta) = \lambda\|\theta\|_1$, this is the so-called Laplace (or double exponential) prior:

$$p(\theta) = \prod_{j=1}^d \frac{\lambda}{2} \exp(-\lambda|\theta_j|).$$

Exercise 14.2 Show that the variance of a Laplace-distributed random variable is equal to $\frac{2}{\lambda^2}$.

The interactions between regularization terms and priors can go both ways, and we can consider other classical priors. One that will be useful later in the Bayesian setting is the multivariate Student distribution (often used marginally for independent components):

$$p(\theta) \propto \left(\beta + \frac{1}{2}\|\theta\|_2^2\right)^{-\alpha-d/2},$$

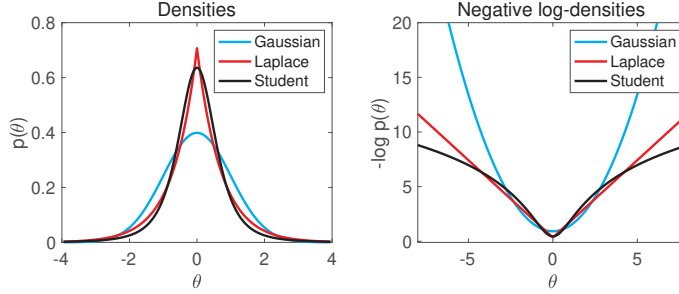


Figure 14.1: Classical priors in one dimension, all normalized to zero mean and unit variance: (left) densities, (right) negative log-densities.

leading to the regularizer $(\alpha + d/2) \log(\beta + \frac{1}{2}\|\theta\|_2^2)$, which is not convex in θ . This will be used within sparse priors in the next section.

Exercise 14.3 (♦) We consider a random vector θ which is Gaussian with mean zero and covariance matrix ηI , with $1/\eta$ being distributed as a Gamma random variable with parameters α and β , that is, η with density $p(\eta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (1/\eta)^{\alpha+1} \exp(-\beta/\eta)$. Show that the marginal density of θ is the Student distribution with density $p(\theta) = c(\beta + \frac{1}{2}\|\theta\|_2^2)^{-\alpha-d/2}$, with $c = \frac{1}{(2\pi)^{d/2}} \frac{\beta^\alpha \Gamma(\alpha+d/2)}{\Gamma(\alpha)}$, and that $\mathbb{E}[\theta\theta^\top] = \frac{\beta}{\alpha-1} I$ if $\alpha > 1$.



This can be misleading as even when the target function is sampled from the prior, MAP estimation may not work. See Section 14.1.4.

The expression of regularizers as log-densities may lead to the impression that MAP estimation is particularly well suited when (1) the conditional model is well-specified, that is, there exists θ_* such that $p(y|x)$ is indeed proportional to $\exp(-\ell(y, f_{\theta_*}))$, and (2) the optimal θ_* is sampled from the prior distribution proportional to $\exp(-\lambda\Omega(\theta))$. As we now explain, this is not the case *at all*.

14.1.3 Sparse priors

As shown in the next section, the Laplace prior is not a good prior for sparse data. We consider the following ones instead. For each one-dimensional component, we consider:

- Generalized Gaussians: $p(\theta) = \frac{\alpha}{2} \frac{\lambda^{1/\alpha}}{\Gamma(1/\alpha)} \exp(-\lambda|\theta|^\alpha)$, with variance $\lambda^{-2/\alpha} \frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}$.
- Student: $p(\theta) = \frac{1}{(2\pi)^{1/2}} \frac{\beta^\alpha \Gamma(\alpha+1/2)}{\Gamma(\alpha)} 1(\beta + \frac{1}{2}\theta^2)^{-\alpha-1/2}$, with variance $\frac{\beta}{\alpha-1}$ if $\alpha > 1$.
- Mixture of two Gaussians: $p(\theta) = \alpha \mathcal{N}(\theta|0, \sigma_0^2) + (1-\alpha) \mathcal{N}(\theta|0, \tau^2)$, with variance $\alpha\sigma_0^2 + (1-\alpha)\tau^2$.

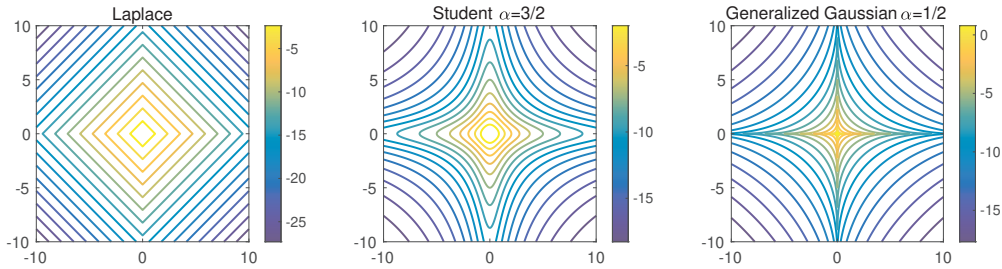


Figure 14.2: Sparse priors in two dimensions.

It turns out that all of these examples happen to be “scale mixtures of Gaussians”, that is, they can be seen as the (potentially continuous) mixtures of Gaussian distributions with zero mean but different variances:

$$p(\theta) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{1}{2}\frac{\theta^2}{\eta}} dq(\eta),$$

where q is a probability measure on \mathbb{R}_+ . For the third example, this is straightforward, with q being a weighted sum of two Diracs at σ_0^2 and τ^2 . For the Laplace distribution (generalized Gaussians with $\alpha = 1$), one can check by direct integration that we can take q to be an exponential distribution, that is, with density $q(\eta) = \frac{\lambda^2}{2} \exp(-\eta\lambda^2/2)$, while for the Student distribution, q has an inverse Gamma distribution, with density $q(\eta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \eta^{-\alpha-1} e^{-\beta/\eta}$ (see Exercise 14.3).

As we show in Section 14.3.2, this hierarchical model can be used with marginal likelihood maximization, leading to reweighted least-squares algorithms that are close to the “ η -trick” from Section 8.3.1, and thus provides a Bayesian interpretation.

Exercise 14.4 A density $p(\theta)$ on \mathbb{R} is said *super-Gaussian* if $\log p(\theta)$ is convex in θ^2 and non-increasing. Show that scale mixtures of Gaussians are super-Gaussian.¹

14.1.4 On the relationship between MAP and MMSE (♦)

In this section, following Gribonval (2011), we consider a very simple conditional model of the form

$$y = \theta + \varepsilon, \tag{14.2}$$

where ε is normal with zero mean and covariance matrix $\sigma^2 I$, assuming σ^2 is known. We have prior knowledge on θ in the form of a prior density $q(\theta)$. Given the observation of y , our goal is to obtain an estimator of θ with the most favorable properties, which we define here as the minimum squared error (this estimator will be generalized in Section 14.3).

That is, given an estimator $\hat{\theta}(y)$, we consider the criterion:

$$\mathcal{J}(\hat{\theta}) = \int_{\mathbb{R}^d} q(\theta) \|\theta - \hat{\theta}(y)\|_2^2 d\theta.$$

¹The converse is not true, see Palmer et al. (2005).

As shown in Section 2.2.3, the optimal estimator (i.e., function) $\hat{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the *a posteriori mean*, that is

$$\hat{\theta}_{\text{MMSE}}(y) = \mathbb{E}[\theta|y],$$

assuming that θ is sampled according to $q(\theta)$ and y follows the model in Eq. (14.2). Here, MMSE stands for “minimum mean-square error”. We now want to compare it with the maximum a posteriori parameter

$$\hat{\theta}_{\text{MAP}}(y) \in \arg \max_{\theta \in \mathbb{R}^d} p(\theta|y) = \arg \max_{\theta \in \mathbb{R}^d} q(\theta)p(y|\theta).$$

Gaussian prior. When q is a Gaussian distribution with mean zero and covariance matrix $\tau^2 I$, then (θ, y) is a Gaussian vector and from conditioning results presented in Section 1.1.3, we have

$$\hat{\theta}_{\text{MMSE}}(y) = \mathbb{E}[\theta|y] = \frac{\tau^2}{\tau^2 + \sigma^2} y,$$

while the MAP estimate is also equal to $\frac{\tau^2}{\tau^2 + \sigma^2} y$ because, for Gaussians, the mean and the mode are the same, but, as we will show later, Gaussian priors are the only ones for which these two are equal.

Simple expression of the MMSE. We denote by $p(y)$ the density of y , that is,

$$\begin{aligned} p(y) &= \int_{\mathbb{R}^d} p(y, \theta) d\theta = \int_{\mathbb{R}^d} p(\theta) p(y|\theta) d\theta \\ &= \int_{\mathbb{R}^d} q(\theta) \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|\theta - y\|_2^2\right) d\theta, \end{aligned}$$

using the expression of the Gaussian density. We can now express the a posteriori mean as, introducing the gradient of the Gaussian density:

$$\begin{aligned} \hat{\theta}_{\text{MMSE}}(y) &= \mathbb{E}[\theta|y] = \int_{\mathbb{R}^d} \frac{p(\theta, y)}{p(y)} \theta d\theta \\ &= y + \sigma^2 \int_{\mathbb{R}^d} \frac{p(y|\theta)p(\theta)}{p(y)} \frac{1}{\sigma^2} (\theta - y) d\theta \\ &= y + \frac{\sigma^2}{p(y)} \int_{\mathbb{R}^d} q(\theta) \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|\theta - y\|_2^2\right) \frac{1}{\sigma^2} (\theta - y) d\theta \\ &= y - \frac{1}{(2\pi\sigma^2)^{d/2}} \frac{\sigma^2}{p(y)} \int_{\mathbb{R}^d} q(\theta) \frac{\partial}{\partial \theta} \left[\exp\left(-\frac{1}{2\sigma^2} \|\theta - y\|_2^2\right) \right] d\theta. \end{aligned}$$

Thus, using integration by parts, we get:

$$\begin{aligned} \hat{\theta}_{\text{MMSE}}(y) &= y + \frac{1}{(2\pi\sigma^2)^{d/2}} \frac{\sigma^2}{p(y)} \int_{\mathbb{R}^d} q'(\theta) \exp\left(-\frac{1}{2\sigma^2} \|\theta - y\|_2^2\right) d\theta \\ &= y + \frac{1}{(2\pi\sigma^2)^{d/2}} \frac{\sigma^2}{p(y)} \int_{\mathbb{R}^d} q'(y - \eta) \exp\left(-\frac{1}{2\sigma^2} \|\eta\|_2^2\right) d\eta \\ &= y + \frac{\sigma^2}{p(y)} p'(y) = y + \sigma^2 \frac{d}{dy} (\log p(y)). \end{aligned} \tag{14.3}$$

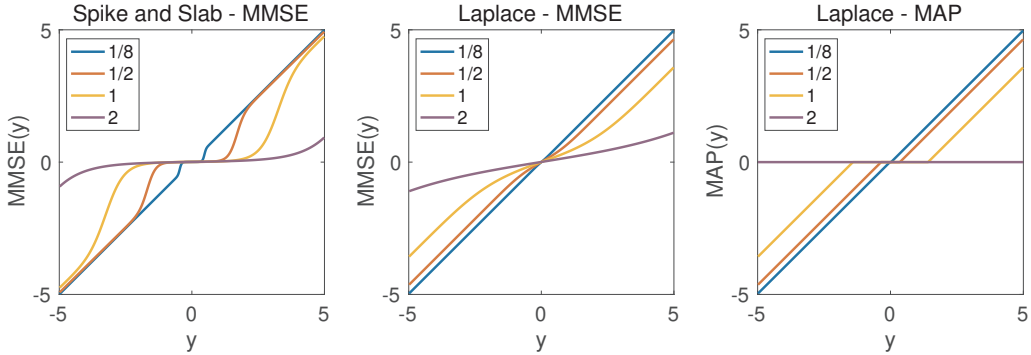


Figure 14.3: Comparison of MMSE and MAP for the spike-and-slab and Laplace priors: MMSE for the spike-and-slab prior (left), MMSE for the Laplace prior (middle), MAP for the Laplace prior (right).

We thus get an explicit expression of the minimum mean square error estimate. Note that for a Gaussian prior, y is (marginally) normally distributed; hence, the gradient of $\log p(y)$ is a linear function, and the MMSE is affine in y if and only if the prior is Gaussian.

Exercise 14.5 (♦) Show that the posterior covariance matrix can be expressed as $\text{var}(\theta|y) = \sigma^2 I + \sigma^4 \frac{d^2}{dy dy^\top} (\log p(y))$.

Expression of the MAP estimate. If $q(\theta) = \exp(-h(\theta))$, then the MAP estimate is

$$\hat{\theta}_{\text{MAP}}(y) \in \arg \max_{\theta \in \mathbb{R}^d} \frac{1}{2\sigma^2} \|\theta - y\|_2^2 + h(\theta),$$

with optimality condition, for differentiable h , $\theta - y - \sigma^2 \frac{d}{d\theta} (\log q(\theta)) = 0$; thus we have:

$$\hat{\theta}_{\text{MAP}}(y) = y + \sigma^2 \frac{d}{dy} (\log q) [\hat{\theta}_{\text{MAP}}(y)]. \quad (14.4)$$

Exercise 14.6 (♦♦) We denote by $f(y) = -\log p(y)$. Show that the MMSE estimator $\hat{\theta}_{\text{MMSE}}(y) = y - \sigma^2 f'(y)$ defined in Eq. (14.3) is the MAP estimator for the negative log-prior g that satisfies $g(\hat{\theta}_{\text{MMSE}}(y)) = f(y) - \frac{\sigma^2}{2} \|f'(y)\|_2^2$ for all $y \in \mathbb{R}^d$.

Differences between MMSE and MAP. Given the expressions in Eq. (14.3) and Eq. (14.4), we can now study how the two estimators differ for the various sparse priors that we have described above, where we consider the one-dimensional case for simplicity (which extends to independent marginal priors in the multi-dimensional case) (see plots in Figure 14.3):

- **Spike-and-slab:** this is the model essentially used in the analysis of the Lasso in Chapter 8, for which MAP with the Laplace prior (that is, the Lasso) is shown to have favorable properties. We consider the prior, which is the mixture of a Dirac at zero (with weight α) and a Gaussian with mean zero and variance τ^2 . The variance is then equal to $(1 - \alpha)\tau^2$, and $p(y)$ is the mixture of two Gaussian distributions, centered in zero, with variances σ^2 and $\sigma^2 + \tau^2$.

Exercise 14.7 Show that the marginal density $p(y)$ for the spike-and-slab prior is equal to $p(y) = \alpha \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) + (1 - \alpha) \frac{1}{(2\pi(\sigma^2 + \tau^2))^{1/2}} \exp\left(-\frac{y^2}{2(\sigma^2 + \tau^2)}\right)$. Provide an expression of $\hat{\theta}_{\text{MMSE}}(y)$ and of $\hat{\theta}_{\text{MAP}}(y)$.

- **Laplace:** this is the model for which the MAP estimation leads to the Lasso method. For $q(\theta) = \frac{\lambda}{2} \exp(-\lambda|\theta|)$, the variance is equal to $2/\lambda^2$. We can compute the MMSE by explicitly computing $p(y)$ by integrating separately over positive and negative numbers (see the exercise below). We see in Figure 14.3 that the MMSE is very far from the soft-thresholding operator from Section 8.3.1. In other words, the Lasso is not adapted to signals that are sampled from the Laplace distribution but rather to signals sampled from the spike-and-slab prior.

Exercise 14.8 Show that the marginal density $p(y)$ for the Laplace prior can be expressed using the Gauss error function $\text{erf}(\alpha) = \frac{2}{\sqrt{\pi}} \int_0^\alpha \exp(-t^2) dt$, as: $p(y) = \frac{\lambda}{4} \exp\left(\frac{\lambda^2 \sigma^2}{2} - \lambda y\right) \left[1 - \text{erf}\left(\frac{\lambda \sigma - \frac{y}{\sigma}}{\sqrt{2}}\right)\right] + \frac{\lambda}{4} \exp\left(\frac{\lambda^2 \sigma^2}{2} + \lambda y\right) \left[1 - \text{erf}\left(\frac{\lambda \sigma + \frac{y}{\sigma}}{\sqrt{2}}\right)\right]$. Provide an expression of $\hat{\theta}_{\text{MMSE}}(y)$ and of $\hat{\theta}_{\text{MAP}}(y)$.

Exercise 14.9 When q is a Gaussian distribution with mean zero and covariance matrix C , provide an expression of the MMSE and MAP estimates.

Exercise 14.10 (♦) Provide a closed-form expression for the marginal density $p(y)$ for the Student prior.

14.2 Discriminative vs. generative models

We consider a traditional supervised learning set-up, with $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The goal is for any $x \in \mathcal{X}$ to obtain a good conditional predictive model of y given x , that is, to get a good model for $p(y|x)$.

We can first directly model $p(y|x)$ with a parameterized conditional model (like done for least-squares or logistic regression). This will be called the *discriminative* approach.

We can also consider a joint density $p(x, y)$, and obtain $p(y|x) = \frac{p(x, y)}{p(x)} \propto p(x, y)$ using Bayes rule. Most often (in particular for classification problems), the joint model is obtained by modeling y and $x|y$, that is, the conditional model of the inputs given the outputs, with a particularly simple model. This will be called the *generative* approach.

14.2.1 Linear discriminant analysis and softmax regression

We consider a generative model with Gaussian class-conditional densities with a common covariance matrix, with $x \in \mathbb{R}^d$ and $y \in \{1, \dots, k\}$:

$$\begin{aligned} y &\sim \text{multinomial}(\pi) \\ x|y = i &\sim \text{Gaussian}(\mu_i, \Sigma). \end{aligned}$$

We can then compute the distribution of y given x as (removing all parts that are independent of i):

$$\begin{aligned} \mathbb{P}(y = i|x) &\propto \mathbb{P}(y = i, x) = \pi_i \exp \left[-\frac{1}{2}(x - \mu_i)^\top \Sigma^{-1}(x - \mu_i) \right] \\ &\propto \pi_i \exp \left[-\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i \right] \exp(\mu_i^\top \Sigma^{-1}x). \end{aligned}$$

This implies that, defining the softmax function $\text{softmax} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ through $\text{softmax}(v)_j = \frac{e^{v_j}}{e^{v_1} + \dots + e^{v_k}}$:

$$\mathbb{P}(y = i|x) = \text{softmax}[(\mu_i^\top \Sigma^{-1}x + \log \pi_i - \frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i)_i] = \text{softmax}[(w_i^\top x + b_i)_i],$$

that is, the conditional model is the softmax function of a linear model, which is precisely the definition of softmax regression from Section 13.1.1, with $w_i = \Sigma^{-1}\mu_i$, and $b_i = \log \pi_i - \frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i$. The availability of a generative model will lead to alternative parameter estimation algorithms (see below). Note that (a) for $k = 2$, we recover logistic regression and that (b) we can apply the softmax regression model for any set of k prediction functions f_1, \dots, f_k beyond affine functions.

Note, finally, that the common covariance matrix is often restricted to be diagonal.

Maximum likelihood estimation. Given observations $(x_1, y_1), \dots, (x_n, y_n)$, the parameters of the model above can be estimated naturally by maximum likelihood. It turns out that for the particular case of multinomial and Gaussian random variables, this is equivalent to computing empirical moments (proof left as an exercise), that is, the estimator of π is the vector of empirical proportions of each class. In contrast, the estimator of each mean μ_i is the empirical mean of observations with class i , and the joint covariance is a weighted combination of the empirical covariances of each class.

Exercise 14.11 (Quadratic discriminant analysis) Assume that the class-conditional covariance matrices are different for each class. Show that the conditional model is still a softmax function, but now of “affine + quadratic” functions of x .

14.2.2 Naive Bayes

We consider discrete data, that is $x \in \{1, \dots, m\}^d$ and $y \in \{1, \dots, k\}$, and the following generative model

$$\begin{aligned} y &\sim \text{multinomial}(\pi) \\ x|y = i &\sim \prod_{j=1}^d \text{multinomial}(x_j|\theta_{ji}), \end{aligned}$$

where $\pi \in \mathbb{R}^k$, and each θ_{ji} is in the simplex in \mathbb{R}^m . In other words, given y , the d components x_1, \dots, x_d are independent.

Using the usual “one-hot” encoding of discrete distribution, we see each x_j in \mathbb{R}^m as one of the canonical basis vectors so that the probability of $x_j|y = i$ is equal to $\prod_{a=1}^m \theta_{jia}^{x_{ja}}$. We can then compute

$$\begin{aligned}\mathbb{P}(y = i|x) &\propto \mathbb{P}(y = i, x) = \prod_{i=1}^k \pi_i^{y_i} \prod_{i=1}^k \prod_{j=1}^d \prod_{a=1}^m \theta_{jia}^{x_{ja} y_i} \\ \log \mathbb{P}(y = i|x) &\propto \sum_{i=1}^n y_i \left(\log \pi_i + \sum_{j=1}^d \sum_{a=1}^m (\log \theta_{jia}) x_{ja} \right).\end{aligned}$$

Like for linear discriminant analysis, we thus also get a softmax model $\text{softmax}[(w_i^\top x + b_i)_i]$, with $b_i = \log \pi_i$, and w_i with components $\log \theta_{jia}$. Also, like for linear discriminant analysis, we can obtain maximum likelihood estimates for each parameter of multinomial variables using empirical proportions.

14.2.3 Maximum likelihood estimations

As shown above, for linear discriminant analysis and naive Bayes, we obtain conditional models corresponding to softmax regression, for which we can use optimization algorithms to get the relevant parameters (this is the discriminative approach followed in this book).

However, we can also use the generative models to estimate parameters in closed form. For example, for linear discriminant analysis, the maximum likelihood estimates for the class proportions are the empirical class proportions $\hat{\pi}_i$, the means are the empirical means, and $\hat{\Sigma} = \sum_{i=1}^k \hat{\pi}_i \hat{\Sigma}_i$, which allows us to compute \hat{w}_i and \hat{b}_i , through the formula above, instead of having to solve a convex problem. The key question is: which one is better?

Discriminative vs. generative learning. When making an even simpler assumption of Σ diagonal, we can study the potential benefits of the discriminative and the generative set-up, following [Ng and Jordan \(2001\)](#): the generative approach has a stronger bias but potentially a lower variance.

For both linear discriminant analysis (LDA²) in Section 14.2.1 and Naive Bayes in Section 14.2.2, if we use the conditional log-likelihood as a criterion, the discriminative approaches in the population case optimize the correct criterion directly, and thus must lead to better or equal performance. However, in the unregularized case, to approach the population case for logistic regression, we need a number of samples proportional to d (e.g., by considering our bounds on Rademacher complexities in Section 4.5 with data with equal variance in all directions). For LDA or Naive Bayes, we need to estimate d separate quantities simultaneously, and when using concentration inequalities and the

²Not to be mixed with Latent Dirichlet Allocation ([Blei et al., 2003](#)), which is a generative model for collections of text documents.

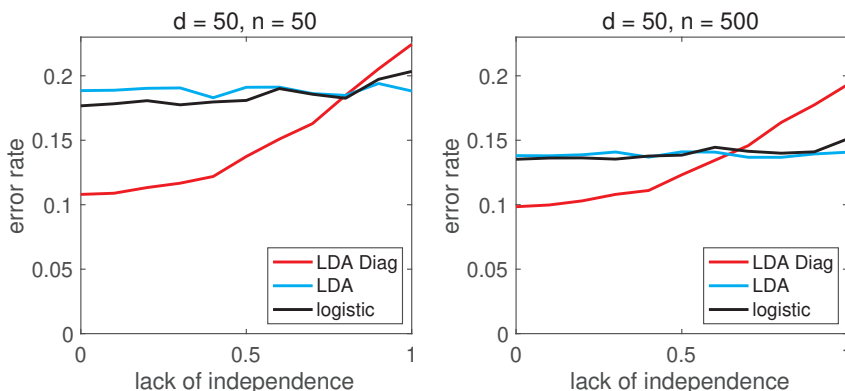


Figure 14.4: Comparison of LDA with full covariance matrix, LDA with diagonal covariance matrix, and logistic regression, on a well-specified binary classification problem (Gaussian class-conditional densities with same covariance matrix), with independent components and non-independent components (with a smooth transition, which is linear in the matrix logarithm). For independent components (left parts of the plots), linear discriminant analysis with the independence assumptions leads to better performance.

union bound, we should expect to have n larger than a constant times $\log d$ to attain the population performance. We thus get a larger bias with generative approaches but significantly less variability. See the experiments in Figure 14.4, more details by [Ng and Jordan \(2001\)](#), and a similar approach to variable selection in regression ([Fan and Lv, 2008](#)).

14.3 Bayesian inference

For simplicity, in this section, we consider random observations $z \in \mathcal{Z}$ that could be the traditional pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in supervised learning, but we note that Bayesian inference applies much more generally. See more details by [Robert \(2007\)](#).

We assume that we have a set of probability distributions over z , with densities with respect to some base measure, which are parameterized by some vector $\theta \in \Theta$ (a subset of a vector space), and which we denote $p(z|\theta)$, and refer to as the *likelihood function*. We assume some *prior distribution* with density $q(\theta)$ with respect to the Lebesgue measure. In the Bayesian methodology, we assume that θ is sampled once from the prior distribution and that we obtain *i.i.d.* observations $z_1, \dots, z_n \in \mathcal{Z}$ sampled from $p(z|\theta)$.

By independence and identical distributions, the overall joint distribution of the data and θ is

$$p(z_1, \dots, z_n, \theta) = q(\theta) \prod_{i=1}^n p(z_i|\theta).$$

We can then obtain the *posterior distribution* of θ given the data (z_1, \dots, z_n) , which is

proportional to $p(z_1, \dots, z_n, \theta)$, and with density:

$$p(\theta|z_1, \dots, z_n) = \frac{q(\theta) \prod_{i=1}^n p(z_i|\theta)}{\int_{\Theta} q(\eta) \prod_{i=1}^n p(z_i|\eta) d\eta}.$$

As already noted, the mode of the posterior distribution is the “maximum a posteriori” (MAP) estimate, which is rarely used within Bayesian inference (see some reasons in Section 14.1.4). Other estimates or estimation procedures are preferred, all using the posterior distribution as the main source. Thus, being able to characterize this posterior distribution is the computational tool (see below).

Posterior mean. A good summary of the posterior distribution is the posterior mean $\int_{\Theta} \theta p(\theta|z_1, \dots, z_n) d\theta$ and is traditionally associated with parameter estimation with the square loss. This was called the MMSE in Section 14.1.4.

Bayesian model averaging. Given the multiple models characterized by the posterior distribution, we can consider performing inference on unseen data through the mixture distribution

$$\int_{\Theta} p(z|\theta) p(\theta|z_1, \dots, z_n) d\theta.$$

Thus, overall, Bayesian inference naturally leads to parameter estimation procedures that can be studied both from a computational perspective (see Section 14.3.1) and a statistical perspective, as part of the “PAC-Bayes” framework described in Section 14.4. But it can also be used for model selection, as described in Section 14.3.2.

14.3.1 Computational handling of posterior distributions

This section gives only a brief account of algorithms to characterize posterior distributions. See many more details by [Gelman et al. \(1995\)](#); [Robert \(2007\)](#).

Conjugate priors. In rare instances, the posterior distribution has a simple form. Two classic examples are the Gaussian prior on the mean parameter of a Gaussian variable and the Dirichlet prior on the parameters of a multinomial distribution.

Gaussian approximation (Laplace method). When the number of observations gets large, then, the integral defining the normalizing factor of the posterior distribution can be written as:

$$\int_{\Theta} q(\eta) \prod_{i=1}^n p(z_i|\eta) d\eta = \int_{\Theta} \exp \left[n \times \left(\frac{1}{n} \log q(\eta) + \frac{1}{n} \sum_{i=1}^n \log p(z_i|\eta) \right) \right],$$

and thus as $\int_{\Theta} \exp(nh(\theta)) d\theta$ for a certain function h . The Laplace method is a traditional approximation technique for approximating integrals of that form when the function h

has a global maximum within the interior of Θ .³ This maximizer is exactly the MAP estimate $\hat{\theta}_{\text{MAP}}$, and the approximation is exactly equivalent to modeling the posterior density as a Gaussian with mean $\hat{\theta}_{\text{MAP}}$ and covariance matrix $\frac{1}{n}h''(\hat{\theta}_{\text{MAP}})^{-1}$.

Sampling. Obtaining independent samples from the posterior distribution is often enough for inference purposes, and many algorithms exist, such as Markov chain Monte Carlo methods (Robert and Casella, 2005).

Variational inference. An alternative to sampling is to approximate the posterior distribution by a family of simple tractable distributions that are made to fit the posterior as closely as possible. See Blei et al. (2017) and references therein.

14.3.2 Model selection through marginal likelihood

Probabilistic models are often naturally defined hierarchically, with prior distributions that have themselves parameters (which we can call hyperparameters), themselves with their own prior distribution (often called hyperprior distribution). For example, using the above notations, the prior distribution is $q(\theta|\lambda)$ with a hyperprior $r(\lambda)$, with often a data distribution that depends on both θ and λ .

While we could still treat λ as a random variable on which Bayesian inference is performed, it is common to perform maximum-likelihood estimation on λ , or more generally, maximum a posteriori estimation. This is sometimes called “type II maximum likelihood” or “empirical Bayes”. This leads to a form of hyperparameter selection for λ . More precisely, we maximize

$$\begin{aligned} p(\lambda|z_1, \dots, z_n) &\propto p(\lambda, z_1, \dots, z_n) = \int_{\Theta} p(\lambda, \theta, z_1, \dots, z_n) d\theta \\ &\propto r(\lambda) \int_{\Theta} \prod_{i=1}^n p(z_i|\theta, \lambda) q(\theta|\lambda) d\theta. \end{aligned}$$

The quantity $\int_{\Theta} \prod_{i=1}^n p(z_i|\theta) q(\theta|\lambda) d\theta$ is referred to as the *marginal likelihood*, and its maximization is a generic tool for hyperparameter selection, with many applications. We present briefly two of them below.

Selection among finitely many models. A classical application of marginal likelihood maximization is to consider m different models, that is, m different distribution $p_j(z|\theta_j)$, with potentially parameters $\theta_j \in \Theta_j$ living in different spaces, with prior distribution $q_j(\theta_j)$. With a uniform distribution on the models, model selection is performed by maximizing with respect to $j \in \{1, \dots, m\}$:

$$\int_{\Theta_j} \prod_{i=1}^n p_j(z_i|\theta_j) q_j(\theta_j) d\theta_j.$$

³See <https://francisbach.com/laplace-method/> for details.

Let's consider the Gaussian approximation obtained from the Laplace approximation. One can show that we obtain penalized maximum log-likelihood with a penalty equal to $\frac{d_j}{2} \log n$, where d_j is the dimension of Θ_j , leading to the Bayesian information criterion (BIC) (see [Robert, 2007](#), Chapter 7). See the discussion in Section 8.2.2.

Sparsity with automatic relevance determination. As mentioned in Section 14.1.3, we consider a prior distribution $q(\theta|\eta)$ which is Gaussian with mean zero and covariance matrix ηI . Maximizing the penalized marginal likelihood ends up being similar to the “ η -trick” from Section 8.3.1. Indeed, when we consider regression with Gaussian noise, that is, when y given θ is normal with mean $\Phi\theta$ and covariance matrix $\sigma^2 I$, then y given η is Gaussian with mean $\Phi \text{Diag}(\eta)\Phi^\top + \sigma^2 I$, and thus we can compute the log-likelihood in closed form, which leads to a natural (non-convex) cost function to estimate η .

Gaussian processes. The example above may be extended to kernel methods presented in Chapter 7. Indeed, it is possible to define a probabilistic model of random function from a set \mathcal{X} to \mathbb{R} such that the marginal distribution of $f(x_1), \dots, f(x_n)$ is Gaussian with mean zero and covariance matrix $K \in \mathbb{R}^{n \times n}$ where $K_{ij} = k(x_i, x_j)$, where k is a positive definite kernel function. This allows us to combine Bayesian inference with non-parametric kernel learning. See more details by [Rasmussen and Williams \(2006\)](#).

14.4 PAC-Bayesian analysis

In this section, we briefly review a generic framework to obtain generalization guarantees for randomized or averaged predictors like those from Bayesian inference. For more details, see [Alquier \(2021\)](#) and the many references therein.

14.4.1 Set-up

We consider the classical supervised learning framework that we have been following throughout the book, namely, with n pairs of i.i.d. observations (x_i, y_i) from a distribution p on $\mathcal{X} \times \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$. We assume that we have a family of prediction functions $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$, parameterized by $\theta \in \Theta$ (which is a subset of a vector space equipped with the Lebesgue measure).

We consider predictors that are not based on selecting a single $\theta \in \Theta$, but a probability distribution ρ over θ . Given that probability distribution, we can consider:

- (a) a randomized predictor f_θ , where θ is sampled from ρ . Then the generalization performance will be considered with this extra randomness (on top of the randomness of the training data),
- (b) the posterior mean $x \mapsto \int_\Theta f_\theta(x) d\rho(\theta)$ which is a function from \mathcal{X} to \mathbb{R} and then only the randomness of the training data need to be considered. Note that in this situation, the final prediction function is not in the set of all f_θ , $\theta \in \Theta$, and is often called an “aggregated predictor”.

The generalization bounds that will be presented will be valid for *all* potential probability distributions ρ , including ones that depend on the data, which implies that we can then optimize the bounds over the distribution, leading to a candidate which is very close to the Bayesian posterior distribution (but with an added temperature). Like in Bayesian inference, we consider a fixed probability distribution q on Θ , which we will refer to as the prior.

We use the notation $\mathcal{R}(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$ for the expected risk (a deterministic function of θ), and $\widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_\theta(x_i))$ for the empirical risk (which is a random function of θ with expectation \mathcal{R}).

14.4.2 Uniformly bounded loss functions

We assume that almost surely, for all $\theta \in \Theta$, we have: $\ell(y, f_\theta(x)) \in [0, \ell_\infty]$ (for example, with the 0-1 loss for binary classification, or with bounded predictors for regression). Following the exposition of [Alquier \(2021\)](#); [Catoni \(2003\)](#), in the proof of Hoeffding's inequality in Section 1.2.1, we saw that for all $\theta \in \Theta$ and $s \in \mathbb{R}_+$, we have:

$$\mathbb{E}[\exp(s(\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta)))] \leq \exp\left(\frac{s^2 \ell_\infty^2}{8n}\right).$$

Integrating over θ , we get

$$\int_{\Theta} \mathbb{E}[\exp(s(\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta)))] dq(\theta) \leq \exp\left(\frac{s^2 \ell_\infty^2}{8n}\right).$$

We now use the variational formulation of the log-partition function (also known as the Donsker-Varadhan formula), with $h(\theta) = s(\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta))$.

$$\log \int_{\Theta} \exp(h(\theta)) dq(\theta) = \sup_{\rho \in \mathcal{P}(\Theta)} \int_{\Theta} h(\theta) d\rho(\theta) - D(\rho \| q),$$

with $\mathcal{P}(\Theta)$ the set of probability distribution on Θ and $D(\rho \| q)$ the Kullback-Leibler divergence between ρ and q , defined as (see also Section 15.1.3):

$$D(\rho \| q) = \int_{\Theta} \log\left(\frac{d\rho}{dq}(\theta)\right) d\rho(\theta).$$

This leads to

$$\mathbb{E}\left[\exp\left(\sup_{\rho \in \mathcal{P}(\Theta)} \int_{\Theta} s(\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta)) d\rho(\theta) - D(\rho \| q)\right)\right] \leq \exp\left(\frac{s^2 \ell_\infty^2}{8n}\right). \quad (14.5)$$

Thus, using Chernoff bound,⁴ we obtain that with probability greater than $1 - \delta$,

$$\sup_{\rho \in \mathcal{P}(\Theta)} \int_{\Theta} s(\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta)) d\rho(\theta) - D(\rho \| q) \leq \frac{s^2 \ell_\infty^2}{8n} + \log \frac{1}{\delta},$$

⁴See https://en.wikipedia.org/wiki/Chernoff_bound.

or, in other words, for all $\rho \in \mathcal{P}(\theta)$,

$$\int_{\Theta} \mathcal{R}(\theta) d\rho(\theta) \leq \int_{\Theta} \widehat{\mathcal{R}}(\theta) d\rho(\theta) + \frac{1}{s} D(\rho \| q) + \frac{1}{s} \log \frac{1}{\delta} + \frac{s\ell_{\infty}^2}{8n}.$$

We thus get a bound on the average generalization error based on the average empirical error. The bound can be empirically computed for any ρ and minimized, with the optimal distribution being proportional to $\exp(-s\widehat{\mathcal{R}}(\theta))dq(\theta)$, which is often called the Gibbs posterior distribution. With $s = n$, this is exactly the Bayesian posterior distribution. Denoting $\hat{\rho}_s$ this distribution, we get with probability greater than $1 - \delta$, that

$$\int_{\Theta} \mathcal{R}(\theta) d\hat{\rho}_s(\theta) \leq \inf_{\rho \in \mathcal{P}(\Theta)} \int_{\Theta} \widehat{\mathcal{R}}(\theta) d\rho(\theta) + \frac{1}{s} D(\rho \| q) + \frac{1}{s} \log \frac{1}{\delta} + \frac{s\ell_{\infty}^2}{8n}.$$

Beyond integrated risks. For convex loss functions, by Jensen's inequality, the risk of the posterior mean $x \mapsto \int_{\Theta} f_{\theta}(x) d\rho(\theta)$ is less than the integrated risk, so the bound applies.

Moreover, by applying Jensen's inequality to Eq. (14.5), we can get a bound in expectation as for all $\rho \in \mathcal{P}(\theta)$ (again ρ may depend on the data):

$$\mathbb{E} \left[\int_{\Theta} \mathcal{R}(\theta) d\rho(\theta) \right] \leq \mathbb{E} \left[\int_{\Theta} \widehat{\mathcal{R}}(\theta) d\rho(\theta) + \frac{1}{s} D(\rho \| q) + \frac{s\ell_{\infty}^2}{8n} \right].$$

Moreover, for the Gibbs posterior, by applying Jensen's inequality, we get:

$$\mathbb{E} \left[\int_{\Theta} \mathcal{R}(\theta) d\hat{\rho}_s(\theta) \right] \leq \inf_{\rho \in \mathcal{P}(\Theta)} \int_{\Theta} \mathcal{R}(\theta) d\rho(\theta) + \frac{1}{s} D(\rho \| q) + \frac{s\ell_{\infty}^2}{8n}. \quad (14.6)$$

Finite set of models. We consider m prediction functions $\hat{f}_1, \dots, \hat{f}_n$. By considering all Diracs in Eq. (14.6), we get that

$$\mathbb{E} \left[\int_{\Theta} \mathcal{R}(\theta) d\hat{\rho}_s(\theta) \right] \leq \inf_{\theta \in \Theta} \mathcal{R}(\theta) + \frac{1}{s} \log \frac{1}{q(\theta)} + \frac{s\ell_{\infty}^2}{8n}.$$

With $q(\theta) = 1/m$ and optimizing over s , we get the usual $\ell_{\infty} \sqrt{\frac{\log m}{n}}$ like we obtained for empirical risk minimization in Section 4.4.3.

Lipschitz-continuous losses, linear predictions, and Gaussian priors. See the tutorial from [Alquier \(2021\)](#) to recover rates similar to ones that can be obtained with Rademacher complexities in Chapter 4.

Application to sparse regression. PAC-Bayesian analysis can be considered in many settings, including the sparse linear regression problems as dealt with in Chapter 8. For example, [Alquier and Lounici \(2011\)](#); [Rigollet and Tsybakov \(2011\)](#) consider the combination of all least-squares predictors with supports restricted to a set $A \subset \{1, \dots, d\}$ for all such sets A . The combination is performed with exponential weights, and the estimator is shown to exhibit the same performance as the ℓ_0 -penalty from Section 8.2.2, but now requires sampling as an estimation algorithm instead of combinatorial optimization.

14.5 Conclusion

Probabilistic modeling is an important part of machine learning. In this chapter, we simply highlighted some topics related to learning theory, namely (1) the link between prior models and predictive performance, where we showed that maximum a-priori estimation may not correctly leverage the knowledge of the prior distribution, namely (2) the use of generative models to obtain alternatives to discriminative estimators, and (3) the link between Bayesian inference.