

## Machine Learning Worksheet 3

### Parameter Inference

---

## 1 Optimizing Likelihoods: Monotonic Transforms

Usually one considers the *log likelihood*,  $\log p(x_1, \dots, x_n \mid \theta)$ . The next problems justify this.

In the lecture, we encountered the likelihood maximization problem

$$\arg \max_{\theta \in [0,1]} \theta^t (1 - \theta)^h,$$

where  $t$  and  $h$  denoted the number of tails and heads in a sequence of coin tosses, respectively.

**Problem 1:** Compute the first and second derivative of this likelihood w.r.t.  $\theta$ . Then compute first and second derivative of the log likelihood  $\log \theta^t (1 - \theta)^h$ .

**Problem 2:** Show that every local maximum of  $\log f(\theta)$  is also a local maximum of the differentiable, positive function  $f(\theta)$ . Considering this and the previous exercise, what is your conclusion?

## 2 Properties of MLE and MAP

**Problem 3:** Show that  $\theta_{\text{MLE}}$  can be interpreted as a special case of  $\theta_{\text{MAP}}$  in the sense that there always exists a prior  $p(\theta)$  such that  $\theta_{\text{MLE}} = \theta_{\text{MAP}}$ .

**Problem 4:** Consider a Bernoulli random variable  $X$  and suppose we have observed  $m$  occurrences of  $X = 1$  and  $l$  occurrences of  $X = 0$  in a sequence of  $N = m + l$  Bernoulli experiments. We are only interested in the number of occurrences of  $X = 1$ —we will model this with a Binomial distribution with parameter  $\theta$ . A prior distribution for  $\theta$  is given by the Beta distribution with parameters  $a, b$ . Show that the posterior *mean* value  $E[\theta \mid \mathcal{D}]$  (not the MAP estimate) of  $\theta$  lies between the prior mean of  $\theta$  and the maximum likelihood estimate for  $\theta$ .

To do this, show that the posterior mean can be written as  $\lambda$  times the prior mean plus  $(1 - \lambda)$  times the maximum likelihood estimate, with  $0 \leq \lambda \leq 1$ . This illustrates the concept of the posterior mean being a compromise between the prior distribution and the maximum likelihood solution.

The probability mass function of the Binomial distribution for some  $m \in \{0, 1, \dots, N\}$  is

$$p(x = m \mid N, \theta) = \binom{N}{m} \theta^m (1 - \theta)^{N-m}.$$

*Hint:* Identify the posterior distribution. You may then look up the mean rather than computing it.

### 3 Poisson Distribution

**Problem 5:** Let  $X$  be Poisson distributed. Again, for  $n$  i.i.d. samples from  $X$ , determine the maximum likelihood estimate for  $\lambda$ .

In class we also talked about avoiding overfitting of parameters via *prior* information. Compute the posterior distribution over  $\lambda$ , assuming a  $\text{Gamma}(\alpha, \beta)$  prior for it. Compute the MAP for  $\lambda$  under this prior. Show your work.