Solutions to Homework Set Three

ECE 271A
Electrical and Computer Engineering
University of California San Diego
Nuno Vasconcelos

1.

a) To minimize $f(\theta) = ||\mathbf{z} - \mathbf{\Phi}\theta||^2$ we compute the gradient

$$\nabla_\theta f = -2\mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\theta)$$

and set it to zero, which leads to the standard least squares solution

$$\theta^* = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{z}.$$

The matrix $\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T$ is referred to as the *pseudo-inverse* of $\mathbf{\Phi}$. To check that we have a minimum we compute the Hessian

$$\nabla^2_\theta f = 2\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^T = 2\mathbf{\Phi}^T\mathbf{\Phi}$$

and use the theorem, which we already mentioned in the last homework set, that $\mathbf{A}$ is positive definite if and only if there is a non-singular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}^T\mathbf{R}$. (See e.g. *Linear Algebra and its Applications* by G. Strang, Harcourt Brace Jovanovic, 1988.) Since this is exactly the case of our Hessian, we only have to prove that $\mathbf{\Phi}$ is non-singular. We cannot really do this without knowing what the sample points x_i are but, assuming that all the x_i are different, this will hold.

b) Since, when x is known, $f(x,\theta)$ is a deterministic function of θ, and $\epsilon \sim \mathcal{N}(0,\sigma^2)$ it follows that

$$P_{Z|X}(z|x;\theta) = \mathcal{G}(x, f(x,\theta), \sigma^2)$$

Given an iid sample $\mathcal{D} = \{\mathcal{D}_x, \mathcal{D}_y\}$, the ML estimate of θ is then

$$\begin{aligned}
\theta^* &= \arg\max_\theta \sum_i \log P_{Z|X}(z_i|x_i;\theta) \\
&= \arg\max_\theta \sum_i -\frac{(z_i - f(x_i,\theta))^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) \\
&= \arg\min_\theta \sum_i (z_i - f(x_i,\theta))^2 \\
&= \arg\min_\theta \sum_i (z_i - \phi_i\theta))^2 \qquad (1)
\end{aligned}$$

where $\phi_i = (1, x_i, \ldots, x_i^K)$, i.e. the i^{th} row of $\mathbf{\Phi}$. It follows that the ML solution is the one that minimizes $||\mathbf{z} - \mathbf{\Phi}\theta||^2$ and therefore the same as the least squares solution.

c) The only difference is that now we have a different σ_i^2 for each x_i. Hence,

$$\begin{aligned}
\theta^* &= \arg\max_\theta \sum_i \log P_{Z|X}(z_i|x_i;\theta) \\
&= \arg\max_\theta \sum_i -\frac{(z_i - f(x_i,\theta))^2}{2\sigma_i^2} - \frac{n}{2}\log(2\pi\sigma_i^2)
\end{aligned}$$

$$
\begin{aligned}
&= \arg\min_{\theta} \sum_i \frac{(z_i - f(x_i,\theta))^2}{\sigma_i^2} \\
&= \arg\min_{\theta} \sum_i \frac{(z_i - \phi_i\theta)^2}{\sigma_i^2} \\
&= \arg\min_{\theta} (\mathbf{z} - \mathbf{\Phi}\theta)^T \Sigma^{-1} (\mathbf{z} - \mathbf{\Phi}\theta)
\end{aligned}
$$

where $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ is the covariance matrix of $(\epsilon_1, \ldots, \epsilon_n)$, i.e. the diagonal matrix of variances σ_i^2. This is usually referred to as the *weighted least squares problem* with *weighting matrix* $\mathbf{W} = \mathbf{\Sigma}^{-1}$. Note that the statistical formulation gives this matrix an interpretation that makes a lot of sense: the contribution of each measurement (x_i, z_i) to the cost function is weighted by the inverse variance of the noise ϵ_i associated with that measurement. To obtain the ML solution, denoting $g(\theta) = (\mathbf{z} - \mathbf{\Phi}\theta)^T \Sigma^{-1} (\mathbf{z} - \mathbf{\Phi}\theta)$, we compute the gradient

$$\nabla_\theta g = -2\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \mathbf{\Phi}\theta)$$

and set it to zero, which leads to

$$\theta^* = \left(\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{z}.$$

Note that this can be written as

$$\theta^* = \left(\mathbf{\Phi'}^T \mathbf{\Phi'}\right)^{-1} \mathbf{\Phi'}^T \mathbf{z'}.$$

where $\mathbf{\Phi'} = \mathbf{S\Phi}$, $\mathbf{z'} = \mathbf{Sz}$, and $\mathbf{S} = diag(1/\sigma_1, \ldots, 1/\sigma_n)$. Compared to the solution in **a**) this corresponds to weighing each ϕ_i and z_i by $1/\sigma_i$, i.e. the inverse of the standard deviation of the corresponding measurement error ϵ_i. Hence, the statistical formulation of weighted least squares leads to a solution where measurements in which we have a lot of confidence (small σ_i) are weighted very heavily, while those that are very noisy (large σ_i^2) receive small weight. It should be clear that this is a significant advantage over the *deterministic* formulation of least squares, where there is no principled way of setting up the weights or to capture the intuition that they should reflect how much confidence we have in each measurement.

We still have to check that we have a minimum and, for this, we compute the Hessian

$$\nabla_\theta^2 g = 2\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} = 2(\mathbf{S\Phi})^T \mathbf{S\Phi}$$

Since $\mathbf{S}$ is non-singular we, once again, have a positive definite $\nabla_\theta^2 g$, as long as the x_i are all different. Hence, we have a minimum.

d) We have answered most of this question in **c)**. First, we have seen that the solution is

$$\theta^* = \left(\mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{W} \mathbf{z}.$$

Second, we have seen that $\mathbf{W}$ is the inverse of the covariance matrix $\mathbf{\Sigma}$ of the noise process ϵ. A fully symmetric $\mathbf{W}$ only means that *the errors are no longer independent*, i.e. the covariance σ_{ij} between ϵ_i and ϵ_j is no longer zero. This means that we have a model of the form

$$z_i = f(x_i, \theta) + \epsilon_i$$

where the random process $(\epsilon_1, \ldots, \epsilon_n)$ is no longer iid, but Gaussian with zero mean and covariance $\mathbf{\Sigma} = \mathbf{W}^{-1}$.

e) The easiest way to answer this question is to work backwards fro equation (1), replacing the L_2 norm with the L_1. It should not be difficult to convince yourself that this is equivalent to adopting a likelihood function of the form

$$P_{Z|X}(z|x;\theta) = \frac{1}{Z} e^{-\frac{|z-f(x,\theta)|}{\sigma^2}}$$

which corresponds to a probabilistic model of the form

$$z = f(x,\theta) + \epsilon$$

with *Laplacian* noise, i.e.

$$P_\epsilon(t) = \frac{1}{Z} e^{-\frac{|t|}{\sigma^2}}.$$

We can also compute the normalizing constant Z

$$Z = \int_{-\infty}^{\infty} e^{-\frac{|t|}{\sigma^2}} dt = 2\int_0^{\infty} e^{-\frac{t}{\sigma^2}} dt = 2\sigma^2$$

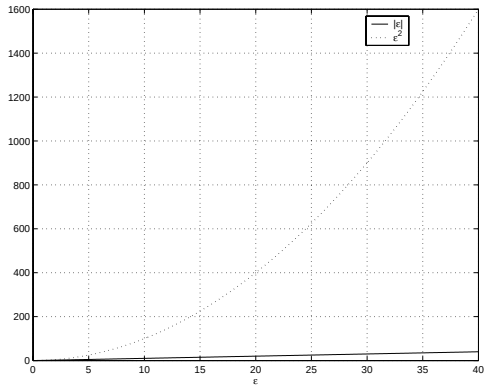
to obtain

$$P_\epsilon(t) = \frac{1}{2\sigma^2} e^{-\frac{|t|}{\sigma^2}}.$$

The conclusion that this model is more robust to outliers can be obtained in two ways. First, the L_1 norm is more robust to outliers because it penalizes less large errors. Consider the error $\epsilon = z - f(x,\theta)$ and the two norms

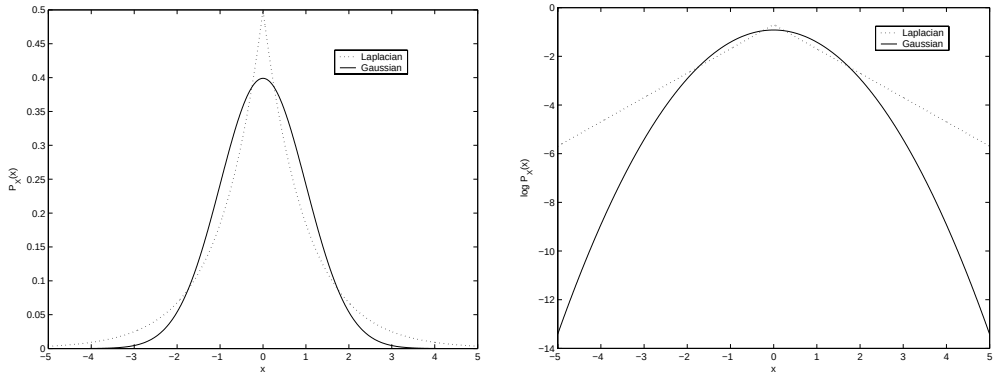
$$L_1 = \sum_i |\epsilon_i|$$
$$L_2 = \sum_i \epsilon_i^2.$$

As can be seen from the plot in the figure below the L^2 norm is much larger for large values of ϵ. Hence, outliers will contribute significantly more to the total cost than when the L^1 norm is used and, since we are trying to minimize the overall cost, the solution will be significantly more affected by them.



The second way to conclude that the formulation based on the L_1 norm is more robust is to consider the associated probabilistic model. Noting that outliers are rare events of large amplitude, the ability of a given model to explain outliers is a function of how much probability mass is contained in its tails. In particular, an *heavy-tailed* distribution (more probability mass in the tails) will explain outliers better than a distribution that is not heavy-tailed. It is indeed the case that the Laplacian is more heavy-tailed

than the Gaussian. You can convince yourself of this by plotting the two distributions. Notice that, since both functions are exponentially decreasing with the distance from the mean, it is usually difficult to see anything about the tails in the plot of the pdf. A better strategy is to look at the plot of the log of the pdf, which makes the differences more salient. Both are shown on the figure below, for the Gaussian and Laplacian distributions when $\sigma^2 = 1$. On the left we show the pdf plots, on the right their log. As you can see, the Gaussian decays much faster (its log is a quadratic function of the distance from the mean) than the Laplacian (log is a linear function of this distance). For example, for x five standard deviations away from the mean, the difference between the two functions is about 8 dbs, i.e. the Laplace probability is about $3,000$ larger. This implies that Laplacian noise explains outliers much better than Gaussian noise. Hence, in the presence of outliers there will be less of a mismatch under the Laplacian model and the ML solution is therefore better than that achievable with the Gaussian.



2.

a) Problem 3.5.17 in DHS

i) Since the samples are independent

$$\begin{aligned} P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) &= \prod_{k=1}^{n} P_{\mathbf{X}|\theta}(\mathbf{x}_k|\theta) && (2)\\ &= \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{(\mathbf{x}_k)_i}(1-\theta_i)^{1-(\mathbf{x}_k)_i} && (3)\\ &= \prod_{i=1}^{d} \theta_i^{\left(\sum_{k=1}^{n}\mathbf{x}_k\right)_i}(1-\theta_i)^{\left(\sum_{k=1}^{n}1-\mathbf{x}_k\right)_i} && (4)\\ &= \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} && (5) \end{aligned}$$

ii) To compute the posterior we apply Bayes rule

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= \frac{P_{\mathbf{T}|\theta}(\mathcal{D}|\theta)P_\theta(\theta)}{\int P_{\mathbf{T}|\theta}(\mathcal{D}|\theta)P_\theta(\theta)d\theta}\\ &= \frac{P_{\mathbf{T}|\theta}(\mathcal{D}|\theta)}{\int_0^1 P_{\mathbf{T}|\theta}(\mathcal{D}|\theta)d\theta}\\ &= \prod_{i=1}^{d}\frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i} \end{aligned}$$

iii) For $d=1$, $n=1$, $s_1\in\{0,1\}$ and

$$P_{\theta|\mathbf{T}}(\theta|s_1) = 2\theta_1^{s_1}(1-\theta_1)^{1-s_1}.$$

The plots of the two densities are shown in Figure 1. Notice that the observation of a $'0'$ turns the uniform prior into a posterior with a lot more probability for small values of θ (which is the probability of $'1'$). On the other hand, the observation of a $'1'$ turns the uniform prior into a posterior with a lot more probability for large values of θ. As usual for Bayesian inference, this is intuitive.

iv)

$$\begin{aligned} P_{\mathbf{X}|\mathbf{T}}(\mathbf{x}|\mathcal{D}) &= \int_0^1 P_{\mathbf{X}|\theta}(\mathbf{x}|\theta)P_{\theta|\mathbf{T}}(\theta|\mathcal{D})d\theta\\ &= \int_0^1 \prod_{i=1}^{d}\theta_i^{x_i}(1-\theta_i)^{1-x_i}\prod_{i=1}^{d}\frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}d\theta\\ &= \int_0^1 \prod_{i=1}^{d}\frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i+x_i}(1-\theta_i)^{n-s_i-x_i+1}d\theta\\ &= \prod_{i=1}^{d}\frac{(n+1)!}{s_i!(n-s_i)!}\int_0^1 \theta_i^{s_i+x_i}(1-\theta_i)^{n-s_i-x_i+1}d\theta_i \end{aligned}$$

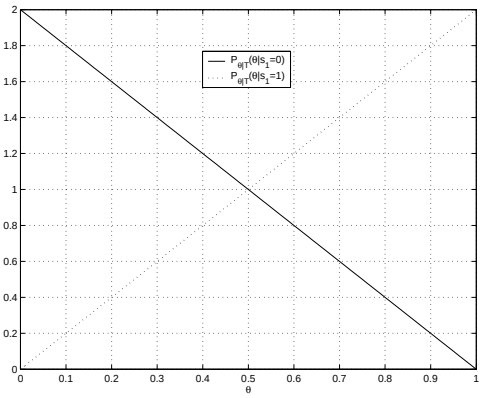


Figure 1: Posteriors for θ.

$$
\begin{aligned}
&= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \frac{(s_i+x_i)!(n-s_i-x_i+1)!}{(n+2)!} \\
&= \prod_{i=1}^{d} \frac{(s_i+x_i)^{x_i}(n-s_i+1)^{1-x_i}}{(n+2)} \\
&= \prod_{i=1}^{d} \left(\frac{s_i+x_i}{n+2}\right)^{x_i} \left(\frac{n-s_i+1}{n+2}\right)^{1-x_i} \\
&= \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i} \left(1-\frac{s_i+1}{n+2}\right)^{1-x_i}
\end{aligned}
$$

v) The effective estimate for θ is

$$\hat{\theta}_i = \frac{s_i+1}{n+2}, \;\; i \in 1, \ldots, d$$

b) The ML solution is

$$
\begin{aligned}
\theta_{ML} &= \arg\max_{\theta} \; \mathbf{l}(\theta) \\
&= \arg\max_{\theta} \; \log P_{\mathbf{X}}(\mathcal{D};\theta) \\
&= \arg\max_{\theta} \; \sum_{i=1}^{d} s_i \log\theta_i + (n-s_i)\log(1-\theta_i).
\end{aligned}
$$

Note that we should impose the constraints $0 \leq \theta_i \leq 1$. However, as we will see below, the maximum of the unconstrained problem is inside the constraint region. Hence, we can get away without doing this. It therefore suffices to compute the gradient

$$\frac{\partial \mathbf{l}}{\partial \theta_i} = \frac{s_i}{\theta_i} - \frac{n-s_i}{1-\theta_i}$$

and set to zero, from which we obtain

$$\theta_i^* = \frac{s_i}{n}.$$

Since $s_i \leq n$ this is indeed inside the constraint region. However, we still need to check that we have a maximum. For this we compute the Hessian

$$\begin{aligned} \frac{\partial^2 \mathbf{l}}{\partial \theta_i \partial \theta_j} &= 0 \\ \frac{\partial^2 \mathbf{l}}{\partial \theta_i^2} &= -\frac{s_i}{\theta_i^2} - \frac{n - s_i}{(1-\theta_i)^2}. \end{aligned}$$

This is a diagonal matrix will negative entries in the main diagonal, hence clearly negative definite for all values of θ_i. It follows that we have a maximum. Furthermore the function $\mathbf{l}(\theta)$ is concave and this maximum is global. Hence, we do not need to check boundary constraints.

The MAP solution is

$$\begin{aligned} \theta_{MAP} &= \arg\max_{\theta} \ \mathbf{p}(\theta) \\ &= \arg\max_{\theta} \ \log P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) + \log P_{\theta}(\theta) \end{aligned}$$

Notice that, since $P_\theta(\theta)$ is uniform, the second term does not depend on θ and can be dropped. Hence, the MAP solution is the same as the ML solution. This result is always true when we have a uniform prior, in which case there is no point in favoring MAP over ML.

3.

a) The MAP estimate is

$$\begin{aligned}
\mathbf{\Pi}_{MAP} &= \arg\max_{\pi_1,\ldots,\pi_N} P_{\mathbf{C}|\mathbf{\Pi}}(c_1,\ldots,c_N|\pi_1,\ldots,\pi_N)P_{\mathbf{\Pi}}(\pi_1,\ldots,\pi_N) \\
&= \arg\max_{\pi_1,\ldots,\pi_N} Z\prod_{j=1}^{N}\pi_j^{c_j}\prod_{j=1}^{N}\pi_j^{u_j-1} \\
&= \arg\max_{\pi_1,\ldots,\pi_N} \prod_{j=1}^{N}\pi_j^{c_j+u_j-1}
\end{aligned}$$

where Z is a normalizing constant that does not depend on the π_i. Notice that this maximization is exactly the one that we did in the previous assignment to compute the ML estimate. In fact the MAP estimate is identical to the ML estimate for the case where we have $n+\sum_j u_j - N$ observations from X and the number of the times that the observed value of X is k is $c_j + u_j - 1$. Hence, we can simply recycle last week's solution, and conclude that

$$\pi_i{}^* = \frac{c_j+u_j-1}{n+\sum_j u_j - N} = \frac{c_j+u_j-1}{\sum_j(c_j+u_j-1)}$$

b) We have already answered most of this in **a)**. The solution is equivalent to the ML solution of an equivalent problem with a larger number of trials. In particular we have $\sum_j c_j$ observations with relative counts c_j (real observations) and $\sum_j(u_j-1)$ observations with relative counts $u_j - 1$ (virtual observations). The prior allows us to force the estimate to go in a direction that we believe to be more plausible. For example, if we set $\sum_j(u_j-1) >> \sum_j c_j$ and set one of the u_js very close to $\sum_j u_j$ we are basically forcing π_j to go to 1. Conversely, if we set $\sum_j(u_j-1) >> \sum_j c_j$ and set all the u_js to the same value, we force the π_js to be identical, i.e. we have a uniform distribution. But these are extreme examples of prior "manipulation". In practice the Dirichlet prior can be very useful as a way to deal with the *empty bin* problem.

Whenever one estimates histograms in high dimensions, it is very common to end up with a lot of bins that have zero counts just because we have a limited amount of training data. This creates all sorts of problems. For example, if one needs to compute the ratio of two histograms (e.g. to compute the likelihood ratio in a binary Bayes decision rule) an empty bin in the denominator leads to a ratio of ∞. This is bad, not only numerically, but because it can lead to a very poor decision boundary. That is, the ratio goes to ∞ not because that is the case for the true distributions, but simply because there was not enough data in the training sample and the bins were left empty. By enabling us to add a number of virtual samples, the Bayesian solution allows us to eliminate the problem. One simple solution is to add a uniform prior (i.e. set the u_j all equal) and make $\sum_j(u_j-1) << \sum_j c_j$. In this way, we eliminate the empty bins but, at the same time, things do not change much for the bins that have enough data. This procedure is called *regularization*, i.e. we regularize the histogram to reduce overfitting to the training sample. In this case overfitting means that, just because there was no observation $x = k$ in the training set, x will never take value k, a pretty strong statement that is likely to be wrong.

c) We start by noting that

$$\lim_{n\to\infty}\frac{u_j-1}{n+\sum_j u_j - N} = 0.$$

Hence

$$\lim_{n\to\infty}\pi_{j\,MAP} = \lim_{n\to\infty}\frac{c_j}{n+\sum_j(u_j-1)}$$

and

$$\lim_{n\to\infty} \frac{\pi_{j\,ML}}{\pi_{j\,MAP}} = \lim_{n\to\infty} \frac{n + \sum_j u_j - N}{n} = 1.$$

That is, as the number of observations increases, the MAP estimate converges to the ML estimate and therefore becomes independent of the prior itself. This is a recurrent theme in Bayesian estimation. Whenever the size of the observed data is very large, the likelihood function always dominates. It also makes sense: if the amount of evidence presented by the data is overwhelming there is no point in paying much attention to the prior. The prior is important only when there is not enough data to obtain reliable estimates by ML, as in the regularization example above.

d) We can now go back to problem **3)** and note that the equivalent estimate that we obtained also fits the framework of "Bayesian equal to adding virtual samples". Note that, in the case of **3**, we obtained an estimate which is equivalent to that of ML when obtained with two extra observations, one equal to $'0'$ and the other to $'1'$. This is consistent with the fact that we used a uniform (non-informative) prior: we added an equal amount of virtual samples of each type. Once again, if $n >> 2$, this is basically irrelevant except for the variables x_i that received zero counts due to the empty bin problem. Note that, for such variables, changing the estimate from s_i/n to $s_i + 1/(n+2)$ changes the π_i from 0 to $1/(n+2)$. That is, we regularized the probability estimates.

The fact that we obtained a qualitatively similar result with the two approaches ("added virtual samples") tells us that the the Bayesian framework is very robust. Note that, while in the case of the non-informative prior we are averaging over all likelihood functions, in the MAP case we are picking one single model. It is remarkable that such different approaches lead to the same qualitative outcome. And if these two extremes lead to the same result (qualitatively) it is likely that all others (i.e. all other priors) will also do. This is a recurring property of Bayesian parameter estimation!