CS5487 Problem Set 3

Bayesian Parameter Estimation

Antoni Chan
Department of Computer Science
City University of Hong Kong

Misc. Math

Problem 3.1 Gamma function

The gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du. \tag{3.1}$$

Use integration by parts to prove the relation $\Gamma(x+1) = x\Gamma(x)$. Also show that $\Gamma(1) = 1$. Hence, $\Gamma(x+1) = x!$ when x is a non-negative integer, and we can think of the gamma function as an extension of factorial to positive real numbers.

.....

MAP Estimation

Problem 3.2 MAP for the exponential density

Let x have an exponential density,

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

Assume that the prior distribution of θ is also exponential,

$$p(\theta) = \begin{cases} \lambda e^{-\lambda \theta}, & \theta \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

(a) Given a set of i.i.d. samples $\{x_1, ..., x_N\}$, show that the MAP estimate for θ is

$$\hat{\theta}_{MAP} = \frac{1}{\bar{x} + \frac{\lambda}{N}}, \tag{3.4}$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$.

Now suppose we change the parameterization of the density to $\alpha = 1/\theta$,

$$p(x|\alpha) = \begin{cases} \frac{1}{\alpha} e^{-\frac{x}{\alpha}}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

We might naively assume that the MAP estimator for α is $\hat{\alpha} = \bar{x} + \frac{\lambda}{N}$. *We will show that this is not true!*

(b) First, we need to obtain the prior distribution of α that is equivalent to (3.3). This is not as simple as plugging in $\theta = 1/\alpha$, since θ is a random variable (we have to preserve the normalization of the pdf). Recall the following rule to obtain the distribution of a transformation of a random variable. If $\theta \sim p_\theta(\theta)$ and $\alpha = g(\theta)$ is a transformation of the r.v. θ, then $p(\alpha) = \left|\frac{\partial \alpha}{\partial \theta}\right|^{-1} p_\theta(g^{-1}(\alpha))$. Use this rule to obtain the prior of α,

See file r.v. transformation

$$p(\alpha) = \begin{cases} \frac{\lambda}{\alpha^2} e^{-\frac{\lambda}{\alpha}}, & \lambda \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

(c) Show that the MAP estimate for α is

$$\hat{\alpha}_{MAP} = \frac{N\bar{x} + \lambda}{N + 2}. \tag{3.7}$$

Note: this demonstrates that the MAP estimator is not invariant to all transformations!

.....

Problem 3.3 Invariance of MAP to linear transformations

Prove that MAP estimators are invariant to *linear* transformations.

(a) Let $p(x|\theta)$ be the distribution with scalar parameter θ, which has prior distribution $p(\theta)$. Let the new parameter be $\lambda = a\theta + b$, which is a linear transformation of θ (and $a \neq 0$). Show that $\hat{\lambda}_{MAP} = a\hat{\theta}_{MAP} + b$.

(b) Extend this result to vector parameters, i.e. show that for $\lambda = A\theta + b$, where $\theta \in \mathbb{R}^d$ and A is invertible, the MAP estimator is $\hat{\lambda}_{MAP} = A\hat{\theta}_{MAP} + b$.

.....

Bayesian Parameter Estimation

Problem 3.4 Bayesian estimation for the Gaussian mean

Let x be a r.v. with a Gaussian distribution, with mean μ and known variance σ^2, and place a Gaussian prior on the mean,

$$p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2), \qquad p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2). \tag{3.8}$$

Show that the posterior distribution of μ, given samples $\mathcal{D} = \{x_1, \ldots, x_n\}$, is

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu|\hat{\mu}_n, \hat{\sigma}_n^2), \tag{3.9}$$

$$\hat{\mu}_n = \frac{n\sigma_0^2 \hat{\mu}_{ML} + \sigma^2 \mu_0}{\sigma^2 + n\sigma_0^2}, \qquad \hat{\sigma}_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}, \tag{3.10}$$

where $\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^n x_i$.

Note: the prior and posterior of the mean are both Gaussian distributions. In other words, conditioning on the data yields an updated posterior, which "reproduces" from the original prior (it has the same form). This is an example of a *conjugate prior*. See Problem 3.9 for more about conjugate priors.

.....

Problem 3.5 Bayesian estimation for the precision of a Gaussian

In this problem we will look at estimating the precision of a Gaussian. (The precision is the inverse of the variance). Let x be a r.v. with a Gaussian distribution,

$$p(x|\lambda) = \mathcal{N}(x|\mu, \lambda^{-1}), \tag{3.11}$$

where λ is the precision and μ is the known mean.

A Gamma distribution is a distribution over positive real numbers ($\lambda > 0$),

$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}, \tag{3.12}$$

where $a > 0$ is called these shape parameter, and $b > 0$ is the inverse scale parameter (also called rate parameter). $\Gamma(x)$ is the gamma function (See Problem 3.1, above). The mean and variance of the Gamma distribution are

$$\mathbb{E}[\lambda] = \frac{a}{b}, \qquad \text{var}(\lambda) = \frac{a}{b^2}. \tag{3.13}$$

(a) Let the prior of λ be a Gamma distribution, $p(\lambda) = \text{Gam}(\lambda|a_0, b_0)$. Given a dataset $\mathcal{D} = \{x_1, \cdots, x_n\}$, show that the posterior of λ is also a Gamma distribution,

$$p(\lambda|\mathcal{D}) = \text{Gam}(\lambda|\hat{a}_n, \hat{b}_n), \tag{3.14}$$

$$\hat{a}_n = a_0 + \frac{n}{2}, \qquad \hat{b}_n = b_0 + \frac{n}{2}\sigma^2_{ML}, \tag{3.15}$$

where $\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is the ML estimate of the variance. What is the interpretation of the posterior mean $\mathbb{E}[\lambda|\mathcal{D}]$ in terms of "virtual" samples to an equivalent ML estimate?

(b) Using the same Gamma prior, show that the MAP estimate of λ is

$$\hat{\lambda}_{MAP} = \frac{2(a_0 - 1) + n}{2b_0 + n\sigma^2_{ML}}. \tag{3.16}$$

How does the MAP estimate behave when n is small and when $n \to \infty$?

(c) Now assume a noninformative (and improper) prior, $p(\lambda) \propto \frac{1}{\lambda}$. Show that the posterior of λ is

$$p(\lambda|\mathcal{D}) = \text{Gam}(\lambda|\tilde{a}_n, \tilde{b}_n), \tag{3.17}$$

$$\tilde{a}_n = \frac{n}{2}, \qquad \tilde{b}_n = \frac{n}{2}\sigma^2_{ML}, \tag{3.18}$$

and that the MAP estimate is

$$\tilde{\lambda}_{MAP} = \frac{n - 2}{n\sigma^2_{ML}}. \tag{3.19}$$

(d) Discuss the relationship between these estimators, including similarities and differences. How are the estimators related? What are the intuitive interpretations of the posterior mean in terms of adding "virtual" samples to an equivalent ML estimate? What is the regularization effect?

(e) So far we have only looked at the estimate of the precision parameter λ. Now let's consider the predictive distribution of a novel x. Show that the predictive distribution

$$p(x|\mathcal{D}) = \int p(x|\lambda)p(\lambda|\mathcal{D})d\lambda, \tag{3.20}$$

where $p(\lambda|\mathcal{D}) = \text{Gam}(\lambda|a, b)$ and $p(x|\lambda) = \mathcal{N}(x|\mu, \lambda^{-1})$, can be written as a Student's t distribution,

$$p(x|\mathcal{D}) = \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)\sqrt{2\pi b}}\left(1+\frac{1}{2b}(x-\mu)^2\right)^{-(a+\frac{1}{2})}. \tag{3.21}$$

(Hint: use Problem 3.1). The Student's t-distribution is a "fat" version of the Gaussian with longer tails, and hence can represent outliers better. The mean and variance are given by (for $a > 1$),

$$\mathbb{E}[x|\mathcal{D}] = \mu, \qquad \text{var}(x|\mathcal{D}) = \frac{b}{a-1}, \tag{3.22}$$

Calculate the variance of the predictive distribution using the values of $\{\hat{a}_n, \hat{b}_n\}$ from the posterior distribution in (a). What is the interpretation in terms of "virtual" samples added to an ML estimate? Plot the predictive distribution for a few values of $\{a_0, b_0, n, \sigma^2_{ML}\}$. How does the distribution behave when a_0 is large or when n is large?

.....

Problem 3.6 Bayesian estimation for a multivariate Gaussian

Let x be a vector r.v. with a multivariate Gaussian distribution of dimension d, and place a prior on the vector mean μ (assume that Σ is known),

$$p(x|\mu) = \mathcal{N}(x|\mu, \Sigma), \qquad p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0). \tag{3.23}$$

(a) Given a set of i.i.d. samples $\mathcal{D} = \{x_1, \cdots, x_n\}$, show that the posterior distribution of μ is

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu|\hat{\mu}_n, \hat{\Sigma}_n), \tag{3.24}$$

$$\hat{\mu}_n = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\hat{\mu}_{ML} + \frac{1}{n}\Sigma(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\mu_0, \tag{3.25}$$

$$\hat{\Sigma}_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}, \tag{3.26}$$

where $\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^n x_i$. Hint: note that $p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu)$, then examine the exponent and rewrite it into a quadratic form by completing the square.

(b) Show that (3.25) can be rewritten as

$$\hat{\mu}_n = A\hat{\mu}_{ML} + (I - A)\mu_0, \qquad A = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}. \tag{3.27}$$

What is the intuitive interpretation of this Bayes estimator $\hat{\mu}_n$?

(c) Show that the predictive distribution is

$$p(x_*|\mathcal{D}) = \mathcal{N}(x_*|\hat{\mu}_n, \hat{\Sigma}_n + \Sigma). \tag{3.28}$$

Hint: use Problem 1.9.

.....

Problem 3.7 Bayesian estimation for a Bernoulli distribution

In this problem we will consider Bayesian estimation and prediction for a Bernoulli r.v. Let x be a r.v. with a Bernoulli distribution,

$$p(x|\pi) = \pi^x(1-\pi)^{1-x}, \tag{3.29}$$

where $\pi = P(x = 1)$ is the parameter.

(a) Let $\mathcal{D} = \{x_1, \cdots, x_n\}$ be a set of samples, show that

$$p(\mathcal{D}|\pi) = \pi^s(1-\pi)^{n-s}, \tag{3.30}$$

where $s = \sum_{i=1}^n x_i$ is the sum of the samples (sufficient statistic).

(b) Assume a uniform prior over π. Use the identity

$$\int_0^1 \pi^m(1-\pi)^n d\pi = \frac{m!n!}{(m+n+1)!}, \tag{3.31}$$

to show that

$$p(\pi|\mathcal{D}) = \frac{(n+1)!}{s!(n-s)!}\pi^s(1-\pi)^{n-s}. \tag{3.32}$$

Plot this density for $n = 1$ for each value of s.

(c) Given the posterior in (3.32), show that the predictive distribution is

$$p(x|\mathcal{D}) = \left(\frac{s+1}{n+2}\right)^x \left(1 - \frac{s+1}{n+2}\right)^{1-x}. \tag{3.33}$$

What is the effective Bayesian estimate of π? What is the intuitive explanation in terms of "virtual" samples added to an equivalent MLE estimate?

(d) What is the ML estimate for π? What is the MAP estimate for π using the uniform prior? Do you see any advantage in favoring one of the estimates in favor of the others? How does that relate to the uniform prior that was assumed for π?

(e) Consider two other priors on π,

$$p_1(\pi) = 2\pi, \tag{3.34}$$
$$p_0(\pi) = 2 - 2\pi, \tag{3.35}$$

and $0 \leq \pi \leq 1$. The prior $p_1(\pi)$ favors a r.v. biased to 1, while p_0 favors a r.v. biased to 0. Calculate the MAP estimates using these two priors. What is the effective estimate of π using these Bayesian estimates? What is the intuitive explanation in terms of "virtual" samples?

.....

Problem 3.8 Bayesian estimation for a multinomial distribution

Consider Problem 2.7 in the previous problem set. Let x be a r.v. such that $p(x = j) = \pi_j, j \in \{1, \ldots, K\}$, N independent observations from x, a random vector $c = [c_1, \ldots, c_K]^T$ where c_j is the number of times that the observed value is j (i.e. c is the histogram of the sample of observations). We have seen that, c has multinomial distribution

$$p(c|\pi) = \frac{N!}{\prod_{i=1}^{K} c_i!} \prod_{j=1}^{K} \pi_j^{c_j}, \tag{3.36}$$

where $\pi = [\pi_1, \cdots, \pi_K]^T$ is the probability of each bin. In this problem we are going to compute MAP estimates for this model. Notice that the parameters π are probabilities and, therefore, not every prior will be acceptable here (since $\pi_j > 0$ and $\sum_j \pi_j = 1$ for the prior to be valid). One distribution over vectors π that satisfies this constraint is the Dirichlet distribution

$$p(\pi) = \frac{\Gamma(\sum_{j=1}^{K} u_j)}{\prod_{j=1}^{K} \Gamma(u_j)} \prod_{j=1}^{K} \pi_j^{u_j - 1}$$

where the u_j are a set of *hyperparameters* (parameters of the prior) and $\Gamma(x)$ is the gamma function (see Problem 3.1).

(a) Derive the MAP estimator for the parameters π using the Dirichlet prior.

(b) Compare this estimator with the ML estimator derived in Problem 2.7. What is the use of this prior equivalent to, in terms of the ML solution?

(c) What is the effect of the prior as the number of samples n increases? Does this make intuitive sense?

.....

Problem 3.9 Exponential family and conjugate priors

In this problem we explore the exponential family and conjugate priors. The *exponential family* is the family of densities of the form

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T u(x)}, \tag{3.37}$$

where θ is the parameter, and $f(x)$, $g(\theta)$, $\phi(\theta)$, and $u(x)$ are fixed functions for a particular probability distribution. $g(\theta)$ is the normalization constant (also called the partition function),

$$g(\theta) = \left(\int f(x)e^{\phi(\theta)^T u(x)} dx \right)^{-1}. \tag{3.38}$$

Given a likelihood $p(x|\theta)$ and a prior $p(\theta)$, when the posterior distribution $p(\theta|\mathcal{D})$ is of the same type as the prior distribution (e.g., both are Gaussians), then we call this prior distribution a *conjugate prior* to likelihood $p(x|\theta)$. In other words, conditioning on the data yields an updated posterior, which "reproduces" from the original prior (it has the same form). We have already seen an example of a conjugate prior in Problem 3.4, where a Gaussian prior on the mean of a Gaussian likelihood yields a Gaussian posterior of the mean given the data.

In this problem, we will consider the conjugate priors for other likelihood densities.

(a) Show that, for a density in the exponential family, the likelihood of data $\mathcal{D} = \{x_1, \ldots, x_n\}$ is

$$p(\mathcal{D}|\theta) \propto \prod_{i=1}^{n} f(x_i) \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} u(x_i)\right\}. \tag{3.39}$$

What is the normalization constant? Denote $s = \sum_{i=1}^{n} u(x_i)$ as *the sufficient statistic*.

(b) It has been shown that, apart from certain irregular cases, the exponential family is the only family of distributions for which there is a conjugate prior. Show that

$$p(\theta|\eta, \nu) = \frac{g(\theta)^\eta e^{\phi(\theta)^T \nu}}{\int g(\theta)^\eta e^{\phi(\theta)^T \nu} d\theta} \tag{3.40}$$

is a conjugate prior for the exponential family, where $\{\eta, \nu\}$ are the parameters of the prior distribution. Compute the posterior distribution $p(\theta|\mathcal{D})$. Using the definition of the sufficient statistic s, compare the posterior with the prior density. What is the result of "propagating" the prior through the likelihood function?

(c) Consider the table below. For each row **i)** show that the likelihood function on the right column belongs to the exponential family, **ii)** show that the prior on the left column is a conjugate prior for the likelihood function on the right column, **iii)** compute the posterior $p(\theta|\mathcal{D})$, and **iv)** interpret the meaning of the sufficient statistic and the "propagation" discussed in **b)**.

Likelihood	$p(\mathcal{D}\|\theta)$	Prior	$p(\theta)$
Bernoulli	$\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$	Beta	$p(\theta\|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
Poisson	$\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$	Gamma	$p(\theta\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$
Exponential	$\prod_{i=1}^{n} \theta e^{-\theta x_i}$	Gamma	$p(\theta\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$
Normal ($\theta = 1/\sigma^2$)	$\prod_{i=1}^{n} \sqrt{\frac{\theta}{2\pi}} \exp\{-\frac{\theta}{2}(x_i-\mu)^2\}$	Gamma	$p(\theta\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$

Table 1: In the case of the normal distribution, μ is assumed known, the parameter is the precision $\theta = 1/\sigma^2$.

Note: use the general form of $p(\theta|\mathcal{D})$ you derived in (b)!

(d) Repeat the steps of **c)** for the distributions of problem 3.8, i.e. the multinomial as the likelihood function and the Dirichlet as the prior.

.....

Bayesian Regression

Problem 3.10 Bayesian regression with Gaussian prior

In the last problem set, we showed that various forms of linear regression by the method of least squares are really just particular cases of ML estimation under the model

$$y = \Phi^T \theta + \epsilon \tag{3.41}$$

where $\theta = [\theta_1, \ldots, \theta_D]^T$ is the parameter vector, $y = [y_1, \ldots, y_n]^T$ is the vector of outputs, $\{x_1, \cdots, x_n\}$ are the set of corresponding inputs, $\phi(x_i)$ is a feature transformation, with

$$\Phi = \left[\phi(x_1), \cdots, \phi(x_n)\right] \tag{3.42}$$

and $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$ is a normal random process $\epsilon \sim \mathcal{N}(0, \Sigma)$, with some covariance matrix Σ.

It seems only natural to consider the Bayesian extension of this model. For this, we simply extend the model considering a Gaussian prior

$$p(\theta) = \mathcal{N}(\theta|0, \Gamma),$$

where Γ is the covariance matrix. We will first derive a general result (for generic covariance matrices Σ and Γ), and then show how it relates to other methods.

(a) Given a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, show that the posterior distribution is

$$p(\theta|\mathcal{D}) = \mathcal{N}(\theta|\hat{\mu}_\theta, \hat{\Sigma}_\theta), \tag{3.43}$$

$$\hat{\mu}_\theta = (\Gamma^{-1} + \Phi\Sigma^{-1}\Phi^T)^{-1}\Phi\Sigma^{-1}y, \tag{3.44}$$

$$\hat{\Sigma}_\theta = (\Gamma^{-1} + \Phi\Sigma^{-1}\Phi^T)^{-1}, \tag{3.45}$$

where $\hat{\mu}_\theta$ is the posterior mean and $\hat{\Sigma}_\theta$ is the posterior covariance. Do not assume any specific form of the covariance matrices Σ and Γ. Hint: complete the square (Problem 1.10).

(b) Consider the MAP estimate,

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_\theta p(\theta|\mathcal{D}). \tag{3.46}$$

How does it differ from the least squares estimate and the weighted-least squares estimate? What is the role of the terms that were not present before? Is there any advantage in setting them to anything other than zero?

(c) Consider the case where we assume the prior covariance and observation noise are both i.i.d., $\Gamma = \alpha I$ and $\Sigma = \sigma^2 I$. Show that the MAP estimate under these assumptions is

$$\hat{\theta} = (\Phi\Phi^T + \lambda I)^{-1}\Phi y, \tag{3.47}$$

for some $\lambda \geq 0$. Show that (3.47) is also the solution to the *regularized least-squares problem*,

$$\hat{\theta} = \operatorname*{argmin}_\theta \left\|y - \Phi^T\theta\right\|^2 + \lambda \left\|\theta\right\|^2. \tag{3.48}$$

In various fields, this formulation is also called *ridge regression*, *Tikhonov regression*, *shrinkage* (as in shrinking the weights in the parameter vector), or *weight decay*.

Now let's consider the fully Bayesian version of regression, with the same assumptions as in (c), i.e. $\Gamma = \alpha I$ and $\Sigma = \sigma^2 I$. This formulation is the linear version of *Gaussian process regression*.

(d) What happens to the mean and covariance of the posterior $p(\theta|\mathcal{D})$ for different values of α and σ^2 (e.g., $\alpha = 0$, $\alpha \to \infty$, $\sigma^2 = 0$)?

(e) Given a novel input x_*, show that the predictive distribution of $f_* = f(x_*, \theta)$ is

$$p(f_*|x_*, \mathcal{D}) = \mathcal{N}(f_*|\hat{\mu}_*, \hat{\sigma}_*^2), \tag{3.49}$$

$$\hat{\mu}_* = \phi(x_*)^T \hat{\mu}_\theta, \tag{3.50}$$

$$\hat{\sigma}_*^2 = \phi(x_*)^T \hat{\Sigma}_\theta \phi(x_*). \tag{3.51}$$

(Hint: see Problem 1.1). Assuming the same observation noise σ^2 as the training set, show that the predictive distribution of y_* is

$$p(y_*|x_*, \mathcal{D}) = \int p(y_*|x_*, \theta) p(\theta|\mathcal{D}) d\theta = \mathcal{N}(y_*|\hat{\mu}_*, \sigma^2 + \hat{\sigma}_*^2). \tag{3.52}$$

Hint: note that $p(y_*|x_*, \theta)$ only depends on θ through $f_* = \phi(x_*)^T \theta$. Hence, we can rewrite the integral over θ with an integral over f_*, while replacing $p(\theta|\mathcal{D})$ with $p(f_*|\mathcal{D})$.

This is the linear version of Gaussian process regression. We will see how to derive the nonlinear (kernel) version in a later problem set.

.....

Problem 3.11 Estimating hyperparameters with maximum marginal likelihood

In the Bayesian linear regression from Problem 3.10, there are two *hyperparameters*, α and σ^2, that parameterize the prior and likelihood covariance matrices, $\Gamma = \alpha I$ and $\Sigma = \sigma^2 I$. The observation likelihood (of the vector of observations) and the prior are

$$p(y|\theta, X, \sigma^2) = \mathcal{N}(y|\Phi^T \theta, \sigma^2 I), \tag{3.53}$$

$$p(\theta|\alpha) = \mathcal{N}(\theta|0, \alpha I), \tag{3.54}$$

where the dependence on the hyperparameters is now made explicit.

One way to estimate these hyperparameters is to *maximize the marginal likelihood* of the training data,

$$\{\hat{\alpha}, \hat{\sigma}^2\} = \underset{\alpha, \sigma^2}{\operatorname{argmax}}\, p(y|X, \sigma^2, \alpha) = \underset{\alpha, \sigma^2}{\operatorname{argmax}} \log p(y|X, \sigma^2, \alpha) \tag{3.55}$$

where the *marginal likelihood* (also called the *evidence*) is the likelihood of the training data, averaged over all possible parameter values θ,

$$p(y|X, \sigma^2, \alpha) = \int p(y|\theta, X, \sigma^2) p(\theta|\alpha) d\theta. \tag{3.56}$$

In some sense, (3.55) is selecting the hyperparameters that can best explain the data for all plausible values of the parameter θ (according to the prior).

(a) Show that the marginal log-likelihood for the above model for Bayesian linear regression is

$$\log p(y|X, \sigma^2, \alpha) = \underbrace{-\frac{1}{2\sigma^2} \left\| y - \Phi^T \hat{\mu}_\theta \right\|^2}_{0} + \frac{1}{2\alpha} \left\| \hat{\mu}_\theta \right\|^2 - \frac{1}{2} \log \left| \frac{1}{\alpha} I + \frac{1}{\sigma^2} \Phi \Phi^T \right| - \frac{n}{2} \log \sigma^2 - \frac{D}{2} \log \alpha - \frac{n}{2} \log(2\pi). \tag{3.57}$$

Hint: complete the square.

(b) What is the effect of each term in (3.57), when changing the hyperparameters α and σ^2? In particular, which term is the data-fitting penalty? Which are complexity penalties?

Estimating hyperparameters by maximizing the marginal likelihood is also known as *evidence approximation*, *type-II maximum likelihood*, *generalized maximum likelihood*, or *empirical Bayes*.

.....

Problem 3.12 L1-regularized least-squares (LASSO)

In this problem we will consider a different form of regularized least-squares, which uses the L1-norm to regularize the weights (parameter vector). This is called LASSO (for "least absolute shrinkage and selection operator"), and is a widely studied regression problem. The problem setup is the same as regression problems before (e.g., the previous problem), except now we include a regularization term on the weights based on the L1-norm,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \left\| y - \Phi^T \theta \right\|^2 + \lambda \left\| \theta \right\|_1 , \tag{3.58}$$

where $\|\theta\|_1 = \sum_{i=1}^{D} |\theta_i|$ is the L1-norm, and λ is a hyperparameter that controls its influence. The effect of the L1-norm is to force some of the weights to zero, leading to a parameter vector that is *sparse* (has few non-zero entries). For example, this could be useful for automatically selecting a subset of features from $\phi(x_i)$, or to control the complexity of a polynomial function (by forcing some weights to 0).

(a) Rewrite (3.58) as an MAP estimation problem. What is the prior distribution assumed by LASSO? Plot the Gaussian prior (from the previous problem) and the LASSO prior. How does this explain why LASSO prefers the weights to be close to zero?

There is no closed-form solution to (3.58), and hence an iterative method is needed to perform the optimization. Next, we will rewrite (3.58) into an equivalent *quadratic programming* (QP) problem, which can be plugged into a standard solver (e.g., `quadprog` in MATLAB).

(b) First, we rewrite θ as a difference between two vectors with positive entries.

$$\theta = \theta^+ - \theta^-, \tag{3.59}$$

$$\theta^+ \geq 0, \quad \theta^- \geq 0. \tag{3.60}$$

The original optimization problem (3.58) can now be rewritten as

$$\hat{\theta} = \underset{\theta^+, \theta^-}{\operatorname{argmin}} \frac{1}{2} \left\| y - \Phi^T (\theta^+ - \theta^-) \right\|^2 + \lambda \sum_i |\theta_i^+ - \theta_i^-|, \tag{3.61}$$

$$\text{s.t. } \theta^+ \geq 0, \quad \theta^- \geq 0.$$

Using a bit of optimization theory "magic", we can rewrite (3.61) as

$$\hat{\theta} = \underset{\theta^+, \theta^-}{\operatorname{argmin}} \frac{1}{2} \left\| y - \Phi^T (\theta^+ - \theta^-) \right\|^2 + \lambda \sum_i (\theta_i^+ + \theta_i^-), \tag{3.62}$$

$$\text{s.t. } \theta^+ \geq 0, \quad \theta^- \geq 0.$$

Why is the optimization problem in (3.62) equivalent to that of (3.61)? Hint: at the optimum, what can we say about the values of the pair $\{\theta_i^+, \theta_i^-\}$?

(c) Finally, define $\mathbf{x} = \begin{bmatrix} \theta^+ \\ \theta^- \end{bmatrix}$. Show that (3.62) can be rewritten in the standard form of a quadratic program,

$$\begin{aligned} &\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} \\ &\text{s.t. } \mathbf{x} \geq 0. \end{aligned} \tag{3.63}$$

where

$$\mathbf{H} = \begin{bmatrix} \Phi\Phi^T & -\Phi\Phi^T \\ -\Phi\Phi^T & \Phi\Phi^T \end{bmatrix}, \qquad \mathbf{f} = \lambda\mathbf{1} - \begin{bmatrix} \Phi y \\ -\Phi y \end{bmatrix}, \tag{3.64}$$

and $\mathbf{1}$ is the vector of ones. Now we can use a standard QP solver!

Note: there are many customized algorithms for estimating the weights of LASSO, but this is perhaps the easiest to implement since we can use `quadprog` in MATLAB.

.....

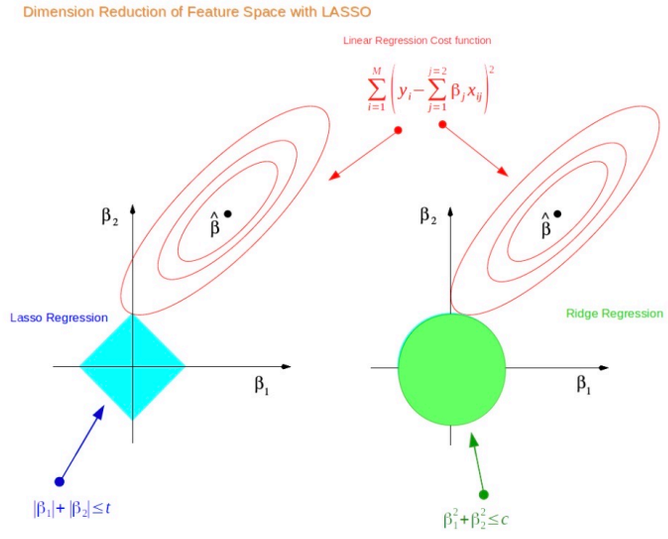

Figure 3: Why LASSO can reduce dimension of feature space? Example on 2D feature space. Modified from the plot used in 'The Elements of Statistical Learning' by Author.

For a two dimensional feature space, the constraint regions (see supplement 1 and 2) are plotted for Lasso and Ridge regression with cyan and green colours. The elliptical contours are the cost function of linear regression (eq. 1.2). Now if we have relaxed conditions on the coefficients, then the constrained regions can get bigger and eventually they will hit the centre of the ellipse. This is the case when Ridge and Lasso regression resembles linear regression results. Otherwise, **both methods determine coefficients by finding the first point where the elliptical contours hit the region of constraints. The diamond (Lasso) has corners on the axes, unlike the disk, and whenever the elliptical region hits such point, one of the features completely vanishes!** For higher dimensional feature space there can be many solutions on the axis with Lasso regression and thus we get only the important features selected.