1 Inference, Learning and Optimization

One of the most fundamental problems in statistics, signal processing, and machine learning, is the *inference* problem, where we wish to construct an estimate of some random quantity of interest $\boldsymbol{\gamma}$, given observations of a related random variable $\boldsymbol{h}$. The quantity of interest $\boldsymbol{\gamma}$ may take on continuous or discrete values, referred to as the "dependent variable", "state of nature", "class" or "label" depending on the application. We will most commonly refer to it as the label. We will refer to the observed random variable $\boldsymbol{h}$ generally as the feature, although in some applications it is known as "regressor" or "predictor".

1.1 BAYSIAN INFERENCE

If we are provided with the conditional distribution of $f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h)$ along with a single realization of the feature $\boldsymbol{h}$, it is quite natural to estimate $\boldsymbol{\gamma}$ as the most likely outcome given $\boldsymbol{h}$. We can express this formally as:

$$\gamma^\star \triangleq \arg\max_{\gamma\in\Gamma} f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h) \tag{1.1}$$

We note that it is more common in the literature on estimation to employ hat-notation to refer to estimates of a random quantity based on data. In this sense, we could have opted to denote the optimal solution to (1.1) as $\widehat{\gamma}$, rather than $\gamma^\star$. Nevertheless, as we transition from inference to learning and optimization, we will increasingly interpret estimates as solutions to generic optimization problems. In this context, it will be more appropriate to employ *-notation to denote an optimal solution. To preserve consistency throughout this textbook, we opt to use *-notation from the onset, with the understanding that within the Baysian framework (1.1) the optimal solution $\gamma^\star$ carries the interpretation of an estimate.

Example 1.1 (**Rainy day**) Suppose one steps outside in the morning and observes that the grass is wet. The question then arises whether it had rained overnight. The fact that the grass is wet (the observation $\boldsymbol{h}$) naturally contains information about the unobserved weather overnight (the state of nature $\boldsymbol{\gamma}$).

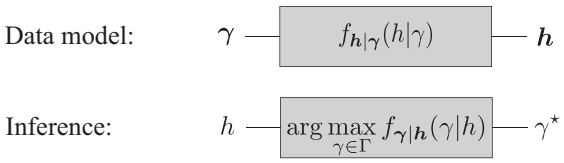

Figure 1.1 A general model underlying most inference problems.

This is formalized through the conditional likelihood:

$$\Pr(\boldsymbol{h} = \text{wet grass}|\boldsymbol{\gamma} = \text{rainy night}) = 0.9 \tag{1.2}$$
$$\Pr(\boldsymbol{h} = \text{dry grass}|\boldsymbol{\gamma} = \text{rainy night}) = 0.1 \tag{1.3}$$
$$\Pr(\boldsymbol{h} = \text{wet grass}|\boldsymbol{\gamma} = \text{dry night}) = 0.05 \tag{1.4}$$
$$\Pr(\boldsymbol{h} = \text{dry grass}|\boldsymbol{\gamma} = \text{dry night}) = 0.95 \tag{1.5}$$

The conditional distribution (or likelihood) above captures the fact that the state of the grass oberserved in the morning is highly correlated with the weather overnight. Suppose we live in an area of even distribution between rainy and dry nights, implying a non-informative prior:

$$\Pr(\boldsymbol{\gamma} = \text{rainy night}) = \Pr(\boldsymbol{\gamma} = \text{dry night}) = 0.5 \tag{1.6}$$

For the marginal probabilities we hence find:

$$\begin{aligned}\Pr(\boldsymbol{h} = \text{wet}) &= \Pr(\boldsymbol{h} = \text{wet grass}|\boldsymbol{\gamma} = \text{rainy night}) \cdot \Pr(\boldsymbol{\gamma} = \text{rainy night}) \\ &\quad + \Pr(\boldsymbol{h} = \text{wet grass}|\boldsymbol{\gamma} = \text{dry night}) \cdot \Pr(\boldsymbol{\gamma} = \text{dry night}) \\ &= 0.9 \cdot 0.5 + 0.05 \cdot 0.5 = 0.45 + 0.025 = 0.475 \end{aligned} \tag{1.7}$$
$$\begin{aligned}\Pr(\boldsymbol{h} = \text{dry}) &= \Pr(\boldsymbol{h} = \text{dry grass}|\boldsymbol{\gamma} = \text{rainy night}) \cdot \Pr(\boldsymbol{\gamma} = \text{rainy night}) \\ &\quad + \Pr(\boldsymbol{h} = \text{dry grass}|\boldsymbol{\gamma} = \text{dry night}) \cdot \Pr(\boldsymbol{\gamma} = \text{dry night}) \\ &= 0.1 \cdot 0.5 + 0.95 \cdot 0.5 = 0.05 + 0.475 = 0.525 \end{aligned} \tag{1.8}$$

Bayes' rule then allows us to find the desired conditional probabilities via:

$$\Pr(\boldsymbol{\gamma} = \gamma|\boldsymbol{h} = h) = \frac{\Pr(\boldsymbol{h} = h|\boldsymbol{\gamma} = \gamma) \cdot \Pr(\boldsymbol{\gamma} = \gamma)}{\Pr(\boldsymbol{h} = h)} \tag{1.9}$$

We have:

$$\Pr(\boldsymbol{\gamma} = \text{rainy night}|\boldsymbol{h} = \text{wet grass}) = \frac{0.9 \cdot 0.5}{0.475} \approx 0.947 \tag{1.10}$$
$$\Pr(\boldsymbol{\gamma} = \text{rainy night}|\boldsymbol{h} = \text{dry grass}) = \frac{0.1 \cdot 0.5}{0.525} \approx 0.095 \tag{1.11}$$
$$\Pr(\boldsymbol{\gamma} = \text{dry night}|\boldsymbol{h} = \text{wet grass}) = \frac{0.05 \cdot 0.5}{0.475} \approx 0.053 \tag{1.12}$$
$$\Pr(\boldsymbol{\gamma} = \text{dry night}|\boldsymbol{h} = \text{dry grass}) = \frac{0.95 \cdot 0.5}{0.525} \approx 0.905 \tag{1.13}$$

We can interpret these numbers as follows. If we step outside, and observe wet grass, it is significantly more likely that it rained overnight, than it didn't. A

reasonable conclusion given the observation of wet grass is then that it rained overnight. Similarly, if we observe dry grass, it is significantly more likely that it did not rain overnight, making this a reasonable conclusion.

Example 1.2 (**Predicting the Output of a Linear Channel**) Let us consider a signal $\boldsymbol{h} \in \mathbb{R}^M$, sent over a linear channel parametrized by $w^o \in \mathbb{R}^M$, resulting in a measurement $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = \boldsymbol{h}^\mathsf{T} w^o + \boldsymbol{v} \tag{1.14}$$

Here, $\boldsymbol{v}$ denotes measurement noise, that we may assume to be normally distributed $\boldsymbol{v} \sim \mathcal{N}(0, \sigma_v^2)$ with variance $\sigma_v^2 > 0$. Suppose we would like to predict the most likely measurement $\boldsymbol{\gamma}$ for a given signal $\boldsymbol{h}$. We can then formulate the conditional pdf:

$$f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{1}{2}\frac{(\gamma - h^\mathsf{T} w^o)^2}{\sigma_v^2}} \tag{1.15}$$

It then follows directly that:

$$\gamma^\star = \arg\max_{\gamma \in \Gamma} f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h) = h^\mathsf{T} w^o \tag{1.16}$$

The above examples have illustrated how the Baysian framework (1.1) can be used to formalize inference problems given data, conditional likelihoods and prior information. In its current form, however, the framework has two important limitations.

- First, relation (1.1) and the following examples provided a rule to map a single observation h to an estimate $\gamma^\star$ of γ. In most inference and learning settings, we are provided with multiple realizations $\{h_n\}_{n=1}^N$ of the random variable $\boldsymbol{h}$, and seek to reconcile multiple realizations in generating $\gamma^\star$. We show in Section 1.1.1 how (1.1) can be generalized to such settings.
- Perhaps more importantly, the resulting estimation rules are driven in large part on the conditional likelihoods (e.g., $f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h)$ when estimating $\boldsymbol{\gamma}$ from $\boldsymbol{h}$). This essentially requires full knowledge of the statistical model relating the label $\boldsymbol{\gamma}$ to its feature $\boldsymbol{h}$. For instance, in the case of the linear channel treated in Example 1.2, this translates into knowledge of the channel taps contained in w^o, along with the distribution of the noise term $\boldsymbol{v}$. In practice, we do not expect to have access to this knowledge, and need to estimate relevant conditional distributions directly from data. We refer to the process of estimating models from data as "learning" and develop it from (1.1) in 1.2.

1.1.1 Inference using a Batch of Features

Let us now consider a setting where, for a single observation of the random variable $\boldsymbol{\gamma}$, we observe multiple related random variables $\{\boldsymbol{h}_n\}_{n=1}^N$ through their realizations $\{h_n\}_{n=1}^N$. Although this condition can be relaxed significantly, we will assume for the time being that the observations $\{\boldsymbol{h}_n\}_{n=1}^N$ are distributed identically and independently after conditioning on the state $\boldsymbol{\gamma}$. Formally, this translates to:

$$f_{\{\boldsymbol{h}_n\}_{n=1}^N|\boldsymbol{\gamma}}\left(\{h_n\}_{n=1}^N|\gamma\right) \overset{(a)}{=} \prod_{n=1}^N f_{\boldsymbol{h}_n|\boldsymbol{\gamma}}\left(h_n|\gamma\right) \overset{(b)}{=} \prod_{n=1}^N f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right) \tag{1.17}$$

where (a) holds by conditional independence and (b) holds by identical distribution of the random variables $\boldsymbol{h}_n$. We can then adjust the framwork (1.1) to include multiple observations as follows:

$$\begin{aligned}
\gamma^\star &\triangleq \arg\max_{\gamma\in\Gamma} f_{\boldsymbol{\gamma}|\{\boldsymbol{h}_n\}_{n=1}^N}\left(\gamma|\,\{h_n\}_{n=1}^N\right) \\
&= \arg\max_{\gamma\in\Gamma} \frac{f_{\{\boldsymbol{h}_n\}_{n=1}^N|\boldsymbol{\gamma}}\left(\{h_n\}_{n=1}^N|\gamma\right)\cdot f_{\boldsymbol{\gamma}}(\gamma)}{f_{\{\boldsymbol{h}_n\}_{n=1}^N}\left(\{h_n\}_{n=1}^N\right)} \\
&= \arg\max_{\gamma\in\Gamma} f_{\{\boldsymbol{h}_n\}_{n=1}^N|\boldsymbol{\gamma}}\left(\{h_n\}_{n=1}^N|\gamma\right)\cdot f_{\boldsymbol{\gamma}}(\gamma) \\
&= \arg\max_{\gamma\in\Gamma} \left(\prod_{n=1}^N f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right)\right)\cdot f_{\boldsymbol{\gamma}}(\gamma)
\end{aligned} \tag{1.18}$$

As we saw in Example 1.2, there are situations where optimizing the *logarithm* of the posterior distribution is more tractable. The fact that the logarithmic funcition is monotonically increasing, ensures that:

$$\begin{aligned}
\gamma^\star &= \arg\max_{\gamma\in\Gamma} \log\left[\left(\prod_{n=1}^N f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right)\right)\cdot f_{\boldsymbol{\gamma}}(\gamma)\right] \\
&= \arg\max_{\gamma\in\Gamma} \left\{\sum_{n=1}^N \log f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right) + \log f_{\boldsymbol{\gamma}}(\gamma)\right\} \\
&= \arg\min_{\gamma\in\Gamma} \left\{-\sum_{n=1}^N \log f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right) - \log f_{\boldsymbol{\gamma}}(\gamma)\right\}
\end{aligned} \tag{1.19}$$

We can normalize by the sample size N, since the minimizing argument $\gamma^\star$ is unaffected by scaling:

$$\boxed{\gamma^\star = \arg\min_{\gamma\in\Gamma} \left\{-\frac{1}{N}\sum_{n=1}^N \log f_{\boldsymbol{h}|\boldsymbol{\gamma}}\left(h_n|\gamma\right) - \frac{1}{N}\log f_{\boldsymbol{\gamma}}(\gamma)\right\}} \tag{1.20}$$

Example 1.3 (**Inverting a Linear Channel**) We consider a binary class variable $\boldsymbol{\gamma} \in \{+1, -1\}$. Suppose that the observed feature $\boldsymbol{h}$ is normally distributed with mean that depends on the class variable $\boldsymbol{\gamma}$. Specifically, we model $\boldsymbol{h}_n \sim \mathcal{N}\left(w^o, \sigma_v^2 I_M\right)$ when $\gamma = +1$ and $\boldsymbol{h}_n \sim \mathcal{N}\left(-w^o, \sigma_v^2 I_M\right)$ when $\gamma = -1$. This relation can be summarized through the linear model:

$$\boldsymbol{h} = \boldsymbol{\gamma} w^o + \boldsymbol{v} \tag{1.21}$$

with $\boldsymbol{v} \sim \mathcal{N}(0, \sigma_v^2)$. In this sense, we may view the present example as the inverse problem to Example 1.2. The linear relationship translates to the conditional distribution:

$$f_{\boldsymbol{h}|\boldsymbol{\gamma}}(h|\gamma) = \frac{1}{\sqrt{(2\pi)^M \sigma_v^M}} e^{-\frac{1}{2}\frac{\|h-\gamma w^o\|^2}{\sigma_v^2}} \tag{1.22}$$

Applying (1.20), and assuming an equal prior $f_{\boldsymbol{\gamma}}(+1) = f_{\boldsymbol{\gamma}}(-1) = \frac{1}{2}$, we obtain:

$$\begin{aligned}
\gamma^\star &= \arg\min_{\gamma \in \pm 1} \frac{1}{N} \sum_{n=1}^{N} \|h_n - \gamma w^o\|^2 \\
&= \arg\min_{\gamma \in \pm 1} \frac{1}{N} \sum_{n=1}^{N} \left\{ \|h_n\|^2 - 2\gamma h_n^{\mathsf{T}} w^o + \|\gamma w^o\|^2 \right\} \\
&= \arg\min_{\gamma \in \pm 1} \left\{ \frac{1}{N} \sum_{n=1}^{N} \|h_n\|^2 - 2\gamma \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o + \|\gamma w^o\|^2 \right\} \\
&\overset{(a)}{=} \arg\min_{\gamma \in \pm 1} \left\{ -2\gamma \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o + \gamma^2 \|w^o\|^2 \right\} \\
&\overset{(b)}{=} \arg\min_{\gamma \in \pm 1} \left\{ -2\gamma \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o \right\} \\
&\overset{(c)}{=} \arg\max_{\gamma \in \pm 1} \left\{ 2\gamma \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o \right\} \\
&\overset{(d)}{=} \operatorname{sign} \left\{ \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o \right\}
\end{aligned} \tag{1.23}$$

In this sequence of reformulations, (a) follows from the fact that $\frac{1}{N}\sum_{n=1}^{N} \|h_n\|^2$ is independent of γ, (b) uses the fact that $\boldsymbol{\gamma}^2 = 1$ with probability one and (c) reverses minimization and maximization. Step (d) uses the fact that

$$\gamma \left(\frac{1}{N} \sum_{n=1}^{N} h_n \right)^{\mathsf{T}} w^o \geq 0 \tag{1.24}$$

if, and only if, $\gamma = \text{sign}\left\{\left(\frac{1}{N}\sum_{n=1}^{N} h_n\right)^{\mathsf{T}} w^o\right\}$.

1.2 FROM INFERENCE TO LEARNING

Our previous discussion, culminating in (1.20), has uncovered how to perform optimal inference of a random quantity of interest $\boldsymbol{\gamma}$ given observations $\{\boldsymbol{h}_n\}_{n=1}^N$, models $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ and prior information $f_{\boldsymbol{\gamma}}(\gamma)$. In many practical situations, we are not provided with prior knowledge about the model, and instead need to estimate $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ from data before performing inference. We refer to the process of estimating $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ from data as *learning*. As we will see in this section, the learning of models can be formalized using a Baysian framework analogous to (1.1). To this end, we will assume that the conditional likelihood $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ is parametrized by a set of learnable parameters $w \in \mathbb{R}^{M_w}$, denoted by $f_{\boldsymbol{h}|\boldsymbol{\gamma},\boldsymbol{w}}(\cdot|\gamma, w)$. Under this parametrization, learning the conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{h}$ is equivalent to learning the parameters $\boldsymbol{w}$ that parametrize $f_{\boldsymbol{h}|\boldsymbol{\gamma},\boldsymbol{w}}(\cdot|\gamma, w)$.

To formulate a procedure for learning $\boldsymbol{w}$ from pairs $\{\boldsymbol{h}, \boldsymbol{\gamma}\}$ we mirror the argument in Section 1.1.1. Suppose we collect labeled data in the form of pairs $\{h_n, \gamma_n\}_{n=1}^N$. If we suppose that feature-label pairs $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}$ are identically and independently distributed for a given set of parameters $\boldsymbol{w}$, we can factorize:

$$\begin{aligned} f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right) &\overset{(a)}{=} \prod_{n=1}^{N} f_{\boldsymbol{h}_n,\boldsymbol{\gamma}_n|\boldsymbol{w}}(h_n,\gamma_n|w) \\ &\overset{(b)}{=} \prod_{n=1}^{N} f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h_n,\gamma_n|w) \end{aligned} \tag{1.25}$$

Again, (a) holds by conditional independence and (b) holds by identical distribution of the pairs of random variables $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}$. We can then define the optimal estimate of the weight vector w as:

$$\begin{aligned} w^\star &\triangleq \arg\max_{w\in\Omega} f_{\boldsymbol{w}|\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N}\left(w|\,\{h_n,\gamma_n\}_{n=1}^N\right) \\ &= \arg\max_{w\in\Omega} \frac{f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right)\cdot f_{\boldsymbol{w}}(w)}{f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N}\left(\{h_n,\gamma_n\}_{n=1}^N\right)} \\ &= \arg\max_{w\in\Omega} f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right)\cdot f_{\boldsymbol{w}}(w) \\ &= \arg\max_{w\in\Omega}\left(\prod_{n=1}^{N} f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h_n,\gamma_n|w)\right)\cdot f_{\boldsymbol{w}}(w) \end{aligned} \tag{1.26}$$

Following the same argument that led to (1.20), we arrive at:

$$\boxed{w^\star = \arg\min_{w\in\Omega} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_n,\gamma_n|w\right) - \frac{1}{N}\log f_{\boldsymbol{w}}(w) \right\}} \tag{1.27}$$

Example 1.4 (**Ridge Regression for Channel Estimation**) Let us consider again the linear channel from Example 1.2:

$$\boldsymbol{\gamma} = \boldsymbol{h}^{\mathsf{T}}\boldsymbol{w} + \boldsymbol{v} \tag{1.28}$$

We assume that the parameters $\boldsymbol{w}$ are independent of all other random variables, impose a normal prior of the form $\boldsymbol{w} \sim \mathcal{N}(0, \sigma_w^2 I_M)$. We continue to assume $\boldsymbol{v} \sim \mathcal{N}(0,\sigma_v^2)$ and $\boldsymbol{h} \sim \mathcal{N}(0,\sigma_h^2)$. It then follows that:

$$f_{\boldsymbol{w}}(w) = \frac{1}{\sqrt{(2\pi)^M \sigma_w^M}} e^{-\frac{1}{2}\frac{\|w\|^2}{\sigma_w^2}} \tag{1.29}$$

$$\begin{aligned} f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h,\gamma|w) &= f_{\boldsymbol{\gamma}|\boldsymbol{h},\boldsymbol{w}}(\gamma|h,w)\cdot f_{\boldsymbol{h}|\boldsymbol{w}}(h|w) \\ &\overset{(a)}{=} f_{\boldsymbol{\gamma}|\boldsymbol{h},\boldsymbol{w}}(\gamma|h,w)\cdot f_{\boldsymbol{h}}(h) \\ &= \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{1}{2}\frac{(\gamma-h^{\mathsf{T}}w)^2}{\sigma_v^2}} \cdot \frac{1}{\sqrt{(2\pi)^{M_h}\sigma_h^{2M_h}}} e^{-\frac{1}{2}\frac{\|h\|^2}{\sigma_h^2}} \end{aligned} \tag{1.30}$$

Plugging these relations into (1.27), and removing constant terms independent of w, we obtain:

$$w^\star = \arg\min_{w\in\Omega} \left\{ \frac{1}{2N}\sum_{n=1}^{N} \left(\gamma_n - h_n^{\mathsf{T}}w\right)^2 + \frac{1}{2N}\frac{\sigma_v^2}{\sigma_w^2}\|w\|^2 \right\} \tag{1.31}$$

1.2.1 Asymptotic Behavior

It is instructive to examine the asymptotic behavior of the optimal estimate (1.27) as the sample size N grows. We have for all w such that $f_{\boldsymbol{w}}(w) > 0$:

$$\begin{aligned} &\lim_{N\to\infty} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_n,\gamma_n|w\right) - \frac{1}{N}\log f_{\boldsymbol{w}}(w) \right\} \\ &= \lim_{N\to\infty} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_n,\gamma_n|w\right) \right\} - \lim_{N\to\infty} \left\{ \frac{1}{N}\log f_{\boldsymbol{w}}(w) \right\} \\ &\overset{(a)}{=} \lim_{N\to\infty} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_n,\gamma_n|w\right) \right\} \\ &\overset{(b)}{=} -\mathbb{E}_{\boldsymbol{h},\boldsymbol{\gamma}} \left\{ \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(\boldsymbol{h},\boldsymbol{\gamma}|w\right) \right\} \end{aligned} \tag{1.32}$$

Here, (a) follows since $f_{\boldsymbol{w}}(w) > 0 \Longrightarrow |\log f_{\boldsymbol{w}}(w)| < \infty$ and hence:

$$\lim_{N\to\infty} \left\{ \frac{1}{N} \log f_{\boldsymbol{w}}(w) \right\} = 0 \tag{1.33}$$

Step (b) follows from the law of large numbers under suitable regularity conditions on the distributions of the involved random variables. The precise nature and rate of convergence as well as limiting distribution of the estimator can be quantified more finely. To avoid a digression, we refer the reader to the appendix for a more detailed discussion on asymptotic properties of Bayesian estimates. For the purpose of our discussion, the key take-away from (1.32) is that it motivates the introduction of the loss:

$$Q(w; \boldsymbol{h}, \boldsymbol{\gamma}) \triangleq -\log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(\boldsymbol{h}, \boldsymbol{\gamma}|w\right) \tag{1.34}$$

as well as the risk:

$$J(w) \triangleq \mathbb{E}_{\boldsymbol{h},\boldsymbol{\gamma}} Q(w; \boldsymbol{h}, \boldsymbol{\gamma}) = -\mathbb{E}_{\boldsymbol{h},\boldsymbol{\gamma}} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(\boldsymbol{h}, \boldsymbol{\gamma}|w\right) \tag{1.35}$$

We may then define the asymptotically optimal model:

$$w^o \triangleq \arg\min_{w\in\Omega} J(w) = \arg\min_{w\in\Omega} \left\{ -\mathbb{E}_{\boldsymbol{h},\boldsymbol{\gamma}} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(\boldsymbol{h}, \boldsymbol{\gamma}|w\right) \right\} \tag{1.36}$$

Example 1.5 (**Least Mean-Square (LMS) Estimation**) Considering the linear model from Examples 1.4, we find in the limit as $N \to \infty$ from (1.31), we obtain the negative log-likelihood as:

$$\begin{aligned}
&\log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h, \gamma|w\right) \\
\overset{(1.30)}{=} &-\log\left(\frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{1}{2}\frac{(\gamma - h^{\mathsf{T}}w)^2}{\sigma_v^2}} \cdot \frac{1}{\sqrt{(2\pi)^{M_h}\sigma_h^{2M_h}}} e^{-\frac{1}{2}\frac{\|h\|^2}{\sigma_h^2}} \right) \\
= &-\log\left(e^{-\frac{1}{2}\frac{(\gamma - h^{\mathsf{T}}w)^2}{\sigma_v^2}} \right) \underbrace{-\log\left(\frac{1}{\sqrt{2\pi\sigma_v^2}} \cdot \frac{1}{\sqrt{(2\pi)^{M_h}\sigma_h^{2M_h}}} e^{-\frac{1}{2}\frac{\|h\|^2}{\sigma_h^2}} \right)}_{\triangleq C} \\
= &\frac{1}{2\sigma_v^2}\left\|\gamma - h^{\mathsf{T}}w\right\|^2 + C
\end{aligned} \tag{1.37}$$

Since $\frac{1}{\sigma_v^2}$ and C are independent of w, we conclude after shifting and scaling:

$$w^o = \arg\min_{w} \frac{1}{2}\mathbb{E}\|\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}}w\|^2 \tag{1.38}$$

Since the objective function in the case of the least mean-square estimation problem takes the form of a quadratic, we can develop a closed-form expression

for w^o. Indeed, after differentiating, we find:

$$\nabla_w \mathbb{E}\|\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w\|^2 = -\underbrace{\mathbb{E}\boldsymbol{\gamma}\boldsymbol{h}}_{\triangleq r_{\gamma h}} + \underbrace{\mathbb{E}\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}}_{\triangleq R_h} w = -r_{\gamma h} + R_h w \tag{1.39}$$

At w^o, this derivative must be equal to zero, and hence after rearranging:

$$w^o = R_h^{-1} r_{\gamma h} \tag{1.40}$$

The perceptive reader will notice that we chose the notation w^o to denote the optimal solution to the least-mean square error cost (1.38), while in case of a deterministic linear channel in Example 1.2, w^o corresponds to the *true* channel weights in (1.14). This choice is deliberate. Indeed, if we condition on $\boldsymbol{w} = w^o$, we have almost surely:

$$\boldsymbol{\gamma} = \boldsymbol{h}^{\mathsf{T}} w^o + \boldsymbol{v} \tag{1.41}$$

If we multiply by $\boldsymbol{h}$ from the left and take expectations, we find:

$$r_{\gamma h} = R_h w^o \tag{1.42}$$

Rearranging again yields $w^o = R_h^{-1} r_{\gamma h}$, and we can conclude that in the case of a linear channel, the optimal solution w^o of the least-squares cost (1.38) corresponds to the true channel.

1.3 CASE STUDY: DATA FUSION

In the previous sections we introduced the Baysian framework as a useful methodology to reason about optimal inference and learning. We concluded that in many cases, the answer boils down to solving an optimization problem of the form (1.27) or (1.36), which depends the data and the associated log likelihoods. We will now proceed to illustrate how this same methodology can be applied to develop optimal techniques for fusing multiple data sources.

To this end, we consider a collection of K agents. Each agent has sensing capabilities, and collects a local data set $\{\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_{k,n}\}_{n=1}^N$. For simplicity, we assume that the sample size N is common to all agents. Each pair of samples is sampled independently and identically from the distribution:

$$f_{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}|\boldsymbol{w}_k}(h,\gamma|w) = f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h,\gamma|w) \tag{1.43}$$

Now suppose different agents are completely disconnected, with no way to exchange information. In that case, it is reasonable to formulate independent local learning problems:

$$w_k^\star \triangleq \arg\max_{w\in\Omega} f_{\boldsymbol{w}|\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N}\left(w|\,\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right) \tag{1.44}$$

Following the argument laid out in Sec. 1.2 we arrive at the equivalent formulation:

$$\begin{aligned} w_k^\star &= \arg\min_{w\in\Omega} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_{k,n},\gamma_{k,n}|w\right) - \frac{1}{N}\log f_{\boldsymbol{w}}(w) \right\} \\ &= \arg\min_{w\in\Omega} J_k(w) + R(w) \end{aligned} \tag{1.45}$$

where we defined:

$$J_k(w) \triangleq -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_{k,n},\gamma_{k,n}|w\right) \tag{1.46}$$

$$R(w) \triangleq -\frac{1}{N}\log f_{\boldsymbol{w}}(w) \tag{1.47}$$

Through the subscript k we emphasize here that because the local samples $\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N$ are distinct, this implies that $J_k(\cdot)$ as well as the resulting optimal models w_k^o are distinct as well. While we can claim local optimality of each model w_k^o in the sense of (1.45), this model is making use only of the data available to agent k.

As an alternative to (1.45), we can instead define the *global* inference problem:

$$w^\star \triangleq \arg\max_{w\in\Omega} f_{\boldsymbol{w}|\left\{\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N\right\}_{k=1}^K}\left(w|\left\{\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right\}_{k=1}^K\right) \tag{1.48}$$

As long as the samples $\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N$ are also mutually independent between agents after conditioning on $\boldsymbol{w}$, we can factorize the conditional likelihoods:

$$\begin{aligned} w^\star &\triangleq \arg\max_{w\in\Omega} f_{\boldsymbol{w}|\left\{\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N\right\}_{k=1}^K}\left(w|\left\{\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right\}_{k=1}^K\right) \\ &= \arg\max_{w\in\Omega} \frac{f_{\left\{\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N\right\}_{k=1}^K|\boldsymbol{w}}\left(\left\{\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right\}_{k=1}^K|w\right)\cdot f_{\boldsymbol{w}}(w)}{f_{\left\{\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N\right\}_{k=1}^K}\left(\left\{\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right\}_{k=1}^K\right)} \\ &= \arg\max_{w\in\Omega} f_{\left\{\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^N\right\}_{k=1}^K|\boldsymbol{w}}\left(\left\{\{h_{k,n},\gamma_{k,n}\}_{n=1}^N\right\}_{k=1}^K|w\right)\cdot f_{\boldsymbol{w}}(w) \\ &= \arg\max_{w\in\Omega} \left(\prod_{k=1}^K f_{\left\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\right\}_{n=1}^N|\boldsymbol{w}}\left(\{h_{k,n},\gamma_{k,n}\}_{n=1}^N|w\right)\right)\cdot f_{\boldsymbol{w}}(w) \end{aligned} \tag{1.49}$$

After taking logarithms, normalizing and making the substitutions (1.46)–(1.47), we find:

$$\begin{aligned} w^\star &= \arg\min_{w} \left\{ -\frac{1}{KN}\sum_{k=1}^{K}\sum_{n=1}^{N} \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}\left(h_{k,n},\gamma_{k,n}|w\right) - \frac{1}{KN}\log f_{\boldsymbol{w}}(w) \right\} \\ &= \arg\min_{w} \frac{1}{K}\sum_{k=1}^{K} J_k(w) + \frac{1}{K}R(w) \end{aligned} \tag{1.50}$$

We refer to (1.50) as the *aggregate* optimization problem of the collection of

agents, since it is obtained by aggregating data $x_{k,n}$ from across the network. The formulation provides a recipe for optimal learning in a collection of agents.

1.3.1 Example: Least-Squares Fusion

Let us specialize our discussion to the least-squares problem of Example 1.4. For the local models we then have:

$$\begin{aligned} w_k^\star &= \arg\min_{w\in\Omega} \left\{ \frac{1}{2N}\sum_{n=1}^{N} \left(\gamma_{k,n} - h_{k,n}^{\mathsf{T}} w\right)^2 + \frac{1}{2N}\frac{\sigma_v^2}{\sigma_w^2}\|w\|^2 \right\} \\ &= \arg\min_{w\in\Omega} \left\{ \frac{1}{2N}\sum_{n=1}^{N} \left(\gamma_{k,n} - h_{k,n}^{\mathsf{T}} w\right)^2 + \frac{\rho}{2}\|w\|^2 \right\} \end{aligned} \tag{1.51}$$

where we defined $\rho = \frac{\sigma_v^2}{\sigma_w^2}$ for convenience. Since (1.51) is a quadratic, we can determine a closed-form expression of $w_k^\star$ by differentiating:

$$-\frac{1}{N}\sum_{n=1}^{N}\gamma_{k,n}h_{k,n} + \left(\frac{1}{N}\sum_{n=1}^{N}h_{k,n}h_{k,n}^{\mathsf{T}}\right) w_k^\star + \frac{\rho}{N}w_k^\star = 0 \tag{1.52}$$

or after inverting:

$$\begin{aligned} w_k^\star &= \left(\sum_{n=1}^{N}h_{k,n}h_{k,n}^{\mathsf{T}} + \rho I\right)^{-1}\left(\sum_{n=1}^{N}\gamma_{k,n}h_{k,n}\right) \\ &= (H_k + \rho I)^{-1} d_k \end{aligned} \tag{1.53}$$

where we defined for compactness:

$$H_k = \sum_{n=1}^{N} h_{k,n}h_{k,n}^{\mathsf{T}} \tag{1.54}$$

$$d_k = \sum_{n=1}^{N}\gamma_{k,n}h_{k,n} \tag{1.55}$$

For the aggregate problem (1.50) we have analogously:

$$\begin{aligned} w^\star &= \left(\sum_{k=1}^{K}\sum_{n=1}^{N}h_{k,n}h_{k,n}^{\mathsf{T}} + \rho I\right)^{-1}\left(\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_{k,n}h_{k,n}\right) \\ &= (H + \rho I)^{-1} d \end{aligned} \tag{1.56}$$

where we defined:

$$H = \sum_{k=1}^{K}\sum_{n=1}^{N} h_{k,n}h_{k,n}^{\mathsf{T}} \tag{1.57}$$

$$d = \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_{k,n}h_{k,n} \tag{1.58}$$

These relationships show how the measurements $h_{k,n}$ ought to be processed to compute an optimal estimate of the weight vector w^o. As it stands, however, evaluating (1.56) requires *central aggregation* of the raw data $h_{k,n}, \gamma_{k,n}$ across the entire network. A natural question is then whether we can learn $w^\star$ by sharing only a processed and compressed version of the raw data. In the case of a least-squares problem, as we show in this case study, this is rather straightforward. For general, and more complex learning problems, this will turn out to be substantially more involved, and we will learn the relevant tools in future chapters. Returning to the quadratic problem, note that:

$$(H + \rho I)w^\star = d = \sum_{k=1}^{K} d_k = \sum_{k=1}^{K} (H_k + \rho I)w_k^\star \tag{1.59}$$

It then follows that $w^\star$ can be evaluated equivalently as:

$$w^\star = \left(\sum_{k=1}^{K} H_k + \rho I\right)^{-1} \left(\sum_{k=1}^{K} (H_k + \rho I)w_k^\star\right) \tag{1.60}$$

In other words, the globally optimal estimate $w^\star$ is fully determined by the locally optimal estimates $w_k^\star$ and the local matrices H_k. This means that rather than communicate and collect raw data $h_{k,n}$ at fusion center, and subsequently computing $w^\star$ via (1.56), we can instead locally compute H_k and $w_k^\star$, only transmit those processed quantities, and subsequently find $w^\star$ via (1.60). The first approach requires the communication of $K(M+1)N$ scalar quantities, where K is the number of agents, M is the dimension of the feature vector, and N is the local sample size. The second approach requires K times the exchange of a matrix of size M^2 and a vector of size M, hence a cost of $K(M+1)M$. Sample sizes are generally significantly larger than the dimension, and hence we benefit from reduced communication cost. We have developed our first distributed algorithm! We will further examine its properties in the problems at the end of this chapter.

1.4 PROBLEMS

1.1 In this problem we implement the least-squares fusion technique in (1.60) in code to verify empirically verify the benefit of data fusion. You are free to choose any unspecified parameters.

(a) Generate local data sets $\{h_{k,n}, \gamma_{k,n}\}_{n=1}^{N}$ following the statistical model of Example 1.4.

(b) Compute $w_k^\star$ for each agent along with the globally optimal model $w^\star$.

(c) Generate a new independent test set $\left\{\widetilde{h}_n, \widetilde{\gamma}_n\right\}_{n=1}^{\widetilde{N}}$ and evaluate prediction perfor-

mance of local models:

$$\frac{1}{\widetilde{N}}\sum_{n=1}^{\widetilde{N}}(\widetilde{\gamma}_n - \widetilde{h}_n^{\mathsf{T}} w_k^{\star})^2 \tag{1.61}$$

as well as of the performance of the global model:

$$\frac{1}{\widetilde{N}}\sum_{n=1}^{\widetilde{N}}(\widetilde{\gamma}_n - \widetilde{h}_n^{\mathsf{T}} w^{\star})^2 \tag{1.62}$$

Compare the performance for varying choices of N and K. What do you conclude?

1.2 We consider again the setting of Section 1.3, but generalize to the setting where local data sets $\{\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_{k,n}\}_{n=1}^{N_k}$ have differing sizes N_k. Adjust the derivation leading up to (1.60) as necessary to find an optimal fusion protocol in this generalized setting. How do the local data sizes N_k affect the fusion?