LECTURE 1: FOUNDATIONS IN DETERMINISTIC AND STOCHASTIC OPTIMIZATION

STEFAN VLASKI AND ALI H. SAYED

ABSTRACT. These notes cover foundational material in statistical learning theory as well as deterministic and stochastic optimization with a focus on single agent learners. It is expected that the reader will have had some exposure to concepts such as maximum likelihood and maximum a posteriori estimation, gradient descent as well as its stochastic variants, and as a result the exposition is kept brief. Nevertheless, we find it useful to collect this information here as it will form the foundation for the development of multi-agent learning algorithms in future lectures.

1. BAYESIAN INFERENCE

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 $\boldsymbol{\gamma} \in \mathbb{R}^{M_{\boldsymbol{\gamma}}}$, given observations of a related random variable $\boldsymbol{h} \in \mathbb{R}^{M_{\boldsymbol{h}}}$. The quantity of interest $\boldsymbol{\gamma}$ may take on continuous or discrete values, and is 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 "observation".

If we are provided with the conditional distribution $f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h)$ 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:

$$\gamma^\star \triangleq \arg\max_{\gamma \in \Gamma} f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h) \tag{1}$$

The conditional distribution $f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|h)$ is frequently referred to as the *posterior distribution* of $\boldsymbol{\gamma}$, making $\gamma^\star$ the *maximum a posteriori (MAP) estimate* of $\boldsymbol{\gamma}$ given $\boldsymbol{h}$. We note that it is common in the literature on estimation theory to the 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) by $\widehat{\gamma}$, 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 appropriate to employ the *-notation to refer to an optimal solution. To preserve consistency throughout this text, we opt to use the *-notation from the onset, with the understanding that within the Bayesian framework (1)

the optimal solution $\gamma^\star$ carries the interpretation of an estimate. Finally, we note that (1) is formulated for a particular realization h of the random variable $\boldsymbol{h}$. As we will see, most MAP estimators are derived for a particular realization of the feature vector, resulting in a deterministic estimate $\gamma^\star$ as a function of the realization h. We can also regard the MAP solution as a random variable, denoted in boldface notation $\boldsymbol{\gamma}^\star$, when it is viewed as a function of the random observation $\boldsymbol{h}$:

$$\boldsymbol{\gamma}^\star \triangleq \arg\max_{\gamma\in\Gamma} f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|\boldsymbol{h}) \tag{2}$$

where the posterior distribution $f_{\boldsymbol{\gamma}|\boldsymbol{h}}(\gamma|\boldsymbol{h})$ is now a function of the random variable $\boldsymbol{h}$, rather than its realization h.

Remark 1.1 (**Bold font indicates random variables**). We note that we employ bold font to denote a random variables, for example $\boldsymbol{h}$, while regular font denotes its realization or a deterministic quanity, as in h. As a general rule, we use lowercase fonts to denote vectors, while uppercase letters denote matrices. In this way, a random matrix would be denoted by $\boldsymbol{H}$, while its realization would correspond to H.

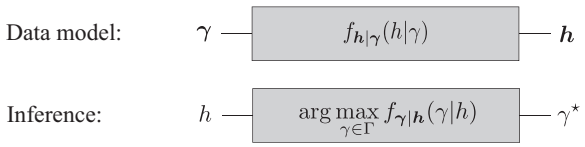

FIGURE 1. A general model underlying most inference problems. The top row illustrates the generative model for the observations $\boldsymbol{h}$, while the bottom row illustrates the Bayesian inference formulation for recovering the label variable.

1.1. **From Inference to Learning.** Our previous discussion leading to (**??**) has uncovered how to perform MAP 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 relating the label $\boldsymbol{\gamma}$ with the feature $\boldsymbol{h}$. Instead, we need to estimate $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ from data before performing inference. We refer to the process of estimating statistical properties of data needed for inference, such as $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$, as *learning*. As we will see in this section, the learning of models can be formalized using a Bayesian MAP framework analogous to (1). To this end, we will assume that the conditional likelihood $f_{\boldsymbol{h}|\boldsymbol{\gamma}}(\cdot|\gamma)$ is parameterized by a learnable parameter $w \in \mathbb{R}^{M_w}$ and write instead $f_{\boldsymbol{h}|\boldsymbol{\gamma},\boldsymbol{w}}(\cdot|\gamma, w)$. Under this parameterization, learning the conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{h}$ is equivalent to learning the parameter $\boldsymbol{w}$ that parameterizes $f_{\boldsymbol{h}|\boldsymbol{\gamma},\boldsymbol{w}}(\cdot|\gamma, w)$.

To formulate a procedure for learning $\boldsymbol{w}$ from pairs $\{\boldsymbol{h}, \boldsymbol{\gamma}\}$ we mirror the argument in Section **??**. Suppose the model $\boldsymbol{w}$ is sampled once, yielding the

realization w^o. The *training data* $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}_{n=1}^N$ is sampled N times, where each pair $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}$ is sampled from $f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h, \gamma|w^o)$. If we suppose that feature-label pairs $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}$ are identically and independently distributed after conditioning on the parameterization $\boldsymbol{w}$, we can factorize:

$$\begin{aligned} f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right) &\overset{(a)}{=} \prod_{n=1}^N f_{\boldsymbol{h}_n,\boldsymbol{\gamma}_n|\boldsymbol{w}}(h_n,\gamma_n|w) \\ &\overset{(b)}{=} \prod_{n=1}^N f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h_n,\gamma_n|w) \end{aligned} \tag{3}$$

Step (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 MAP estimate of the weight vector w as:

$$\begin{aligned} w^\star &\triangleq \arg\max_{w\in\mathbb{R}^{M_w}} f_{\boldsymbol{w}|\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N}\left(w|\,\{h_n,\gamma_n\}_{n=1}^N\right) \\ &= \arg\max_{w\in\mathbb{R}^{M_w}} \frac{f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right)\times f_{\boldsymbol{w}}(w)}{f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N}\left(\{h_n,\gamma_n\}_{n=1}^N\right)} \\ &= \arg\max_{w\in\mathbb{R}^{M_w}} f_{\{\boldsymbol{h}_n,\boldsymbol{\gamma}_n\}_{n=1}^N|\boldsymbol{w}}\left(\{h_n,\gamma_n\}_{n=1}^N|w\right)\times f_{\boldsymbol{w}}(w) \\ &= \arg\max_{w\in\mathbb{R}^{M_w}} \left(\prod_{n=1}^N f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h_n,\gamma_n|w)\right)\times f_{\boldsymbol{w}}(w) \end{aligned} \tag{4}$$

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

$$w^\star = \arg\min_{w\in\mathbb{R}^M}\left\{-\frac{1}{N}\sum_{n=1}^N \log f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h_n,\gamma_n|w) - \frac{1}{N}\log f_{\boldsymbol{w}}(w)\right\} \tag{5}$$

Remark 1.2 (**Distinction between the true model and the MAP estimate**). Observe that we make a deliberate distinction between the *true* model w^o, which parameterizes the distribution $f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h, \gamma|w^o)$ from which the samples $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}_{n=1}^N$ are generated, and the MAP *estimate* $w^\star$, which maximizes the posterior distribution of $\boldsymbol{w}$ after observing the samples $\{\boldsymbol{h}_n, \boldsymbol{\gamma}_n\}_{n=1}^N$. In general, there will be a difference between the MAP model $w^\star$ and the true model w^o. The expectation is that as the size of the data set N grows, the MAP model $w^\star$ will become increasingly accurate and approach w^o as $N \to \infty$. We will verify that this is the case further below. □

Note that in order to be able to *learn* an optimal parameterization $w^\star$ according to (5), we need to be able to collect a batch of feature-label pairs $\{h_n, \gamma_n\}_{n=1}^N$. In the context of machine learning this step is frequently referred to as *training*, and the labeled data $\{h_n, \gamma_n\}_{n=1}^N$ is referred to as the *training data*. Once $w^\star$ is determined, we can use the learned parameterization $w^\star$ on an unlabeled feature vector h^{test}, to compute an approximate

MAP estimate for its label:

$$\gamma^{\text{test}\star} \triangleq \arg\max_{\gamma\in\Gamma} f_{\boldsymbol{\gamma}|\boldsymbol{h},\boldsymbol{w}}(\gamma|h^{\text{test}}, w^\star) \tag{6}$$

We emphasize that (6) is only an *approximate* MAP estimate for the true label γ^{test}, since it is generated using an estimate of the posterior distribution $f_{\boldsymbol{\gamma}|\boldsymbol{h},\boldsymbol{w}}(\gamma|h^{\text{test}}, w^\star)$ using the learned parameterization $w^\star$. In general, and in particular for finite sample sizes N of the training data, the learned parameterization $w^\star$ will be different from the true parameterization w^o that actually generated the data. The difference between predictions made using the true model w^o and the learned model $w^\star$ is known as a *generalization error*.

Remark 1.3 (**Compact notation for feature-label pairs**). As is evident from (5), during learning, we are provided with feature-label pairs $\{\boldsymbol{h}, \boldsymbol{\gamma}\}$ or their realizations $\{h_n, \gamma_n\}_{n=1}^N$. To simplify the notation, we will collect features $\boldsymbol{h} \in \mathbb{R}^{M_{\boldsymbol{h}}}$ and labels $\boldsymbol{\gamma} \in \mathbb{R}^{M_{\boldsymbol{\gamma}}}$ into a single augmented data vector $\boldsymbol{x} \in \mathbb{R}^{M_{\boldsymbol{h}}+M_{\boldsymbol{\gamma}}}$, such that:

$$\boldsymbol{x} \triangleq \operatorname{col}\{\boldsymbol{h}, \boldsymbol{\gamma}\} = \begin{pmatrix} \boldsymbol{h} \\ \boldsymbol{\gamma} \end{pmatrix} \tag{7}$$

Similarly, we will collect realizations into $x_n \triangleq \operatorname{col}\{h_n, \gamma_n\}$. In this manner, we can write more compactly:

$$f_{\boldsymbol{h},\boldsymbol{\gamma}|\boldsymbol{w}}(h, \gamma|w) = f_{\boldsymbol{x}|\boldsymbol{w}}(x|w) \tag{8}$$

The MAP learning problem (5) then becomes:

$$w^\star = \arg\min_{w\in\mathbb{R}^M} \left\{ -\frac{1}{N}\sum_{n=1}^N \log f_{\boldsymbol{x}|\boldsymbol{w}}(x_n|w) - \frac{1}{N}\log f_{\boldsymbol{w}}(w) \right\} \tag{9}$$