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.