

## Problem Set 5 — Solutions (Stochastic Gradient Descent)

**Exercise 37.** Let  $Y$  be a random variable over a finite probability space  $(\Omega, \text{prob})$  where  $\text{prob} : 2^\Omega \rightarrow [0, 1]$ ; this avoids subtleties in defining conditional probabilities and expectations; and it covers the random variables occurring in SGD, since in each step, we are randomly choosing among a finite set of  $n$  indices. Furthermore, let  $B \subseteq \Omega$  be an event.

For nonempty  $B$ , the conditional expectation of  $Y$  given  $B$  is the number

$$\mathbb{E}[Y|B] := \sum_{y \in Y(\Omega)} y \cdot \text{prob}(Y = y|B).$$

where  $Y = y$  is shorthand for the event  $\{\omega \in \Omega : Y(\omega) = y\}$ .

Finally, for two events  $A$  and  $B \neq \emptyset$ , the conditional probability  $\text{prob}[A|B]$  is defined as

$$\text{prob}(A|B) := \frac{\text{prob}(A \cap B)}{\text{prob}(B)}.$$

If  $B = \emptyset$ ,  $\mathbb{E}[Y|B]$  can be defined arbitrarily.

Prove the following statements.

(i) Alternative definition of conditional expectation:

$$\text{prob}(B) \cdot \mathbb{E}[Y|B] = \sum_{\omega \in B} Y(\omega) \text{prob}(\omega).$$

(ii) Partition Theorem: Let  $B_1, \dots, B_m$  be a partition of  $\Omega$ . Then

$$\mathbb{E}[Y] = \sum_{i=1}^m \mathbb{E}[Y|B_i] \text{prob}(B_i).$$

(iii) Linearity of conditional expectation: For random variables  $Y_1, \dots, Y_m$  over  $(\Omega, \text{prob})$  and real numbers  $\lambda_1, \dots, \lambda_m$ , and if  $B \neq \emptyset$ ,

$$\sum_{i=1}^m \lambda_i \mathbb{E}[Y_i|B] = \mathbb{E}\left[\sum_{i=1}^m \lambda_i Y_i | B\right].$$

**Solution:** (i) By definition, we have

$$\begin{aligned} \text{prob}(B) \cdot \mathbb{E}[Y|B] &= \sum_{y \in Y(\Omega)} y \cdot \text{prob}[\{Y = y\} \cap B] \\ &= \sum_{y \in Y(\Omega)} y \sum_{\omega \in \Omega: Y(\omega)=y, \omega \in B} \text{prob}(\omega) \\ &= \sum_{y \in Y(\Omega)} \sum_{\omega \in \Omega: Y(\omega)=y, \omega \in B} Y(\omega) \text{prob}(\omega) \\ &= \sum_{\omega \in \Omega: \omega \in B} Y(\omega) \text{prob}(\omega). \end{aligned}$$

Part(ii) is an immediate consequence—just sum up (i) for all the  $B_i$ 's.

For (iii), we use (i) to compute

$$\begin{aligned}\text{prob}(B) \cdot \sum_{i=1}^m \lambda_i \mathbb{E}[Y_i|B] &= \sum_{i=1}^m \lambda_i \cdot \text{prob}(B) \cdot \mathbb{E}[Y_i|B] \\ &= \sum_{i=1}^m \lambda_i \sum_{\omega \in B} Y_i(\omega) \text{prob}(\omega) \\ &= \sum_{\omega \in B} \sum_{i=1}^m \lambda_i Y_i(\omega) \text{prob}(\omega) \\ &= \sum_{\omega \in B} \mathbb{E}\left[\sum_{i=1}^m \lambda_i Y_i|B\right] \\ &= \text{prob}(B) \cdot \mathbb{E}\left[\sum_{i=1}^m \lambda_i Y_i|B\right].\end{aligned}$$

The desired statement follows after dividing by  $\text{prob}(B) > 0$ .

## Practical Implementation of SGD

Follow the Python notebook provided here:

[github.com/epfml/OptML\\_course/tree/master/labs/ex05/](https://github.com/epfml/OptML_course/tree/master/labs/ex05/)