Labs

Optimization for Machine Learning Spring 2025

EPFL

School of Computer and Communication Sciences
Nicolas Flammarion

github.com/epfml/OptML_course

Problem Set 5 — Solutions (Stochastic Gradient Descent)

Exercise 37. Let Y be a random variable over a finite probability space $(\Omega, \operatorname{prob})$ where $\operatorname{prob}: 2^{\Omega} \to [0,1]$; 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 nonemepty B, the conditional expectation of Y given B is the number

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

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 $\operatorname{prob}[A|B]$ is defined as

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

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

Prove the following statements.

(i) Alternative definition of conditional expectation:

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

(ii) Partition Theorem: Let B_1, \ldots, B_m be a partition of Ω . Then

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

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

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

Solution: (i) By definition, we have

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

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

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

$$\operatorname{prob}(B) \cdot \sum_{i=1}^{m} \lambda_{i} \mathbb{E}[Y_{i} | B] = \sum_{i=1}^{m} \lambda_{i} \cdot \operatorname{prob}(B) \cdot \mathbb{E}[Y_{i} | B]$$

$$= \sum_{i=1}^{m} \lambda_{i} \sum_{\omega \in B} Y_{i}(\omega) \operatorname{prob}(\omega)$$

$$= \sum_{\omega \in B} \sum_{i=1}^{m} \lambda_{i} Y_{i}(\omega) \operatorname{prob}(\omega)$$

$$= \sum_{\omega \in B} \mathbb{E}[\sum_{i=1}^{m} \lambda_{i} Y_{i} | B]$$

$$= \operatorname{prob}(B) \cdot \mathbb{E}[\sum_{i=1}^{m} \lambda_{i} Y_{i} | B].$$

The desired statement follows after dividing by prob(B) > 0.

Practical Implementation of SGD

Follow the Python notebook provided here:

 $github.com/epfml/OptML_course/tree/master/labs/ex05/$