Labs
Optimization for Machine Learning
Spring 2022

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 6 — Solutions (Stochastic Gradient Descent)

Exercise 37. *Let Y be a random variable over a finite probability space (Ω, prob) where $\mathrm{prob} : 2^\Omega \to [0,1]$; 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*

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

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

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

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

Prove the following statements.

(i) Alternative definition of conditional expectation:

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

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

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

(iii) Linearity of conditional expectation: For random variables $Y_1, \ldots, Y_m$ over (Ω, 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

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

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

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

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

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

Practical Implementation of SGD

Follow the Python notebook provided here:

colab.research.google.com/github/epfml/OptML_course/blob/master/labs/ex06/template/Lab 6 - Stochastic Gradient Descent.ipynb