

random experiment

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

A random experiment is a physical or computational process that produces an outcome ω from a known set Ω of possibilities, the sample space, without the outcome being known in advance. Probability theory represents the process by a probability space $(\Omega, \Sigma, \mathbb{P})$, where the σ -algebra Σ collects the events and the probability measure $\mathbb{P}(\cdot)$ assigns each event a probability; this representation is a probabilistic model of the process, not a property of it. A measurement of the outcome is represented by a random variable (RV), a measurable function $x : \Omega \rightarrow \mathcal{X}$, and the measured value $x(\omega)$ is a realization. Repeating the experiment m times is itself a random experiment on the product sample space, and the measurements of the runs are modeled as independent and identically distributed (i.i.d.) RVs. In machine learning (ML), the collection of a training set, the draw of a subset of it by stochastic gradient descent (SGD), and the noise added for differential privacy (DP) are random experiments.

Definition

Rolling a die, reading the maximum temperature of a day at the weather station Krems an der Donau, or drawing a handful of data points from a training set for one iteration of stochastic gradient descent (SGD): each of these is a random experiment. Its outcome is not known before the process is carried out, while the set of possible outcomes is. Formally, a random experiment is a physical or computational process that produces an outcome ω from a set Ω of possibilities, the sample space of the experiment (Bertsekas and Tsitsiklis, 2008, Sect. 1.2; Billingsley, 1986). Its defining characteristic is that the outcome is uncertain, and a key conceptual property is that it can be repeated under identical conditions.

Probability theory does not study the process itself but a mathematical representation of it, a probability space $(\Omega, \Sigma, \mathbb{P})$ (Bertsekas and Tsitsiklis, 2008, Sect. 1.2; Billingsley, 1986). The sample space Ω collects the possible outcomes. The σ -algebra Σ collects the events, those subsets $\mathcal{A} \subseteq \Omega$ to which a probability is assigned. The probability measure $\mathbb{P}(\cdot)$ assigns to each event $\mathcal{A} \in \Sigma$ a number $\mathbb{P}(\mathcal{A}) \in [0, 1]$ such that $\mathbb{P}(\Omega) = 1$ and the probability of a countable union of pairwise disjoint events is the sum of their probabilities. Choosing Ω , Σ and $\mathbb{P}(\cdot)$ is a modeling step: the probability measure is a property

of the probabilistic model, not of the process. For a die, the sample space is $\{1, \dots, 6\}$ and the uniform probability measure is the model of a fair die; whether a given die is fair is a question about the process, answered by the relative frequencies of its outcomes in many rolls.

What is observed of an outcome is a measurement, a function $x : \Omega \rightarrow \mathcal{X}$ of the outcome. Such a measurable function is a random variable (RV), and the value $x(\omega)$ it takes for the actual outcome ω is a realization of it (Billingsley, 1986; Gray, 2009). The temperature read at 13:00 is an RV on the weather of a day, and the number read on one day is a realization. The probability measure on Ω induces the probability distribution of the RV, $\mathbb{P}^{(x)}(\mathcal{B}) = \mathbb{P}(x \in \mathcal{B})$ for every measurable set $\mathcal{B} \subseteq \mathcal{X}$. Machine learning (ML) methods work with the RVs and their probability distributions; the sample space behind them rarely needs to be written down. Fig. 1 depicts the two layers, the process and its mathematical representation.

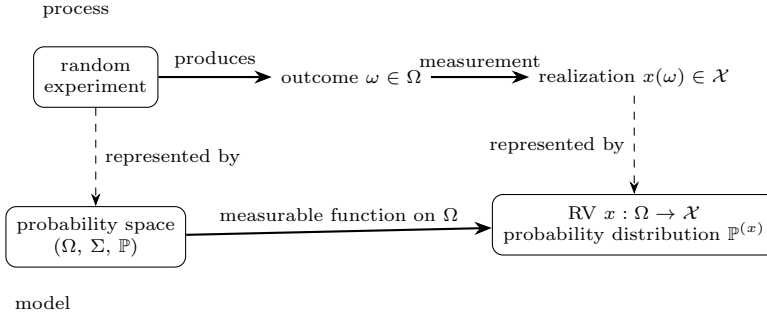


Figure 1: A random experiment and its mathematical representation. The process produces an outcome ω , of which a measurement $x(\omega)$ is observed. Probability theory represents the process by a probability space $(\Omega, \Sigma, \mathbb{P})$ and the measurement by an RV x , a measurable function on Ω whose probability distribution $\mathbb{P}^{(x)}$ is induced by $\mathbb{P}(\cdot)$

Repeating a random experiment m times defines a new random experiment whose outcomes are sequences $(\omega^{(1)}, \dots, \omega^{(m)})$ of outcomes of the original one (Fig. 2). Its sample space is the product $\Omega \times \dots \times \Omega$, and repetition under identical conditions, with no run influencing another, is represented by the product probability measure: the measurements $x^{(1)}, \dots, x^{(m)}$ of the runs are then independent and identically distributed (i.i.d.) RVs (Bertsekas and Tsitsiklis, 2008, Sect. 1.5). While the outcome of a single run is uncertain, the long-run behavior of repeated runs becomes predictable: the law of large numbers and the central limit theorem (CLT) make this precise.

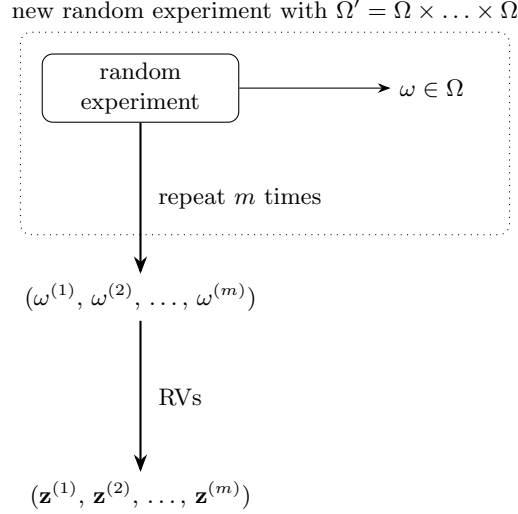


Figure 2: Repeating a random experiment m times yields another random experiment whose outcomes are sequences $(\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(m)}) \in \Omega \times \dots \times \Omega$. One random experiment arising in many ML applications is the gathering of a training set $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$, whose data points are RVs on the repeated runs

Random experiments arise in ML in three places:

- Data collection: the data points of a training set are modeled as RVs, i.e., as functions of the outcome $\omega \in \Omega$ of a random experiment, and the independent and identically distributed assumption (i.i.d. assumption) models their collection as the repetition of one experiment. This is what makes the risk of a hypothesis, an expectation under $\mathbb{P}^{(\mathbf{z})}$, a meaningful target for empirical risk minimization (ERM).
- SGD runs a random experiment at each iteration, the draw of a subset of the training set; the bootstrap draws resamples of the training set.
- Privacy protection methods run a random experiment to perturb the outputs of an ML method, which is what differential privacy (DP) requires.

Cross References

outcome, sample space, event, RV, realization, probability, probability space, probability measure, probability distribution, probabilistic model, i.i.d., law of large numbers, CLT

References

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.
2. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.
3. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.