# Tutorial

# Stochastic gradient descent (SGD) and variants: Evolution and recent trends

**Presenter:**
Paul A. Rodriguez
Electrical Department
Pontifical Catholic University of Peru

## 1. Tutorial overview

Gradient descent (GD) is a well-known first order optimization method, which uses the gradient of the loss function, along with a step-size (or learning rate), to iteratively update the solution. When the loss (cost) function is dependent on datasets with large cardinality, such in cases typically associated with deep learning (DL), GD becomes impractical.

In this scenario, stochastic GD (SGD), which uses a noisy gradient approximation (computed over a random fraction of the dataset), has become crucial. There exits several variants/improvements over the "vanilla" SGD, such SGD+momentum, Adagrad, RMSprop, Adadelta, Adam, Nadam, AdaBelief, etc., which are usually given as black-boxes by most of DL's libraries (TensorFlow, PyTorch, MXNet, etc.)

The primary objective of this tutorial is to open such black-boxes by explaining their "evolutionary path", in which each SGD variant may be understood as a set of add-on features over the vanilla SGD. Furthermore, since the hyper-parameters associated with each SGD variant do directly influence their performance, they will also be assessed from a theoretical and computational point of view.

## 2. Learning Outcome

- Learn about differences between gradient descent (GD) and stochastic GD.
- Learn about SGD variants, as well as on their improvements over the vanilla SGD.
- Observed (computational point of view) the influence of SGD variant's hyper-parameters and learn about their theoretical interpretation.
- Build confidence to take on his/her own project related to machine learning / deep learning.

## 3. Topics (3 hrs.)

- Gradient descent (GD) and stochastic GD.
- Accelerated GD (AGD).
    - Adaptive step-sizes.

    - Momentum.
    - Nesterov acceleration.
    - Anderson acceleration.
- SGD variants (NOTE: other algorithms may be included later).
    - AdaGrad.
    - AdaDelta.
    - RMSprop.
    - Adam.
    - Nadam.
    - AdaBelief
    - e-Adam
- SGD variant's parameters
    - Computational examples.
    - Theoretical interpretation.
- SGD in praxis
    - Multiclass classification
    - Batch size and learning rate scheduling.

**4. Target audience, and the expected prerequisite technical knowledge**

The targeted audiences are senior-year undergraduate, postgraduate (specially first year), as well as industry signal processing engineers and practitioners, with some background in python, random signal processing, (basic) optimization and linear algebra.

**5. Supporting course resources, software, tools and readings**

- Lecture notes from the slides presented in the tutorial.
- Python Jupyter notebooks (TensorFlow and PyTorch based) for after-the-lecture hands-on practice.
- References to papers used in the tutorial.

**6. Pre-reading:**


1. J Watt, R. Borhani, A. Katsaggelos, “First Order Optimization Techniques”, in “Machine Learning Refined: Foundations, Algorithms, and Applications” (2nd Ed.), 2020, (Ch. 34). Cambridge University Press.
2. N. Vishnoi, “Gradient Descent”, in “Algorithms for Convex Optimization”, 2021, (Ch. 6).

Cambridge University Press.

3. S. Shalev-Shwartz, S. Ben-David, “Stochastic Gradient Descent”, in “Understanding Machine Learning: From Theory to Algorithms”, 2014 (Ch. 14). Cambridge University Press.

## 7. Presenters’ contact information and short biography

| Presenter | Short Biography |
|---|---|
| **Paul Rodriguez**<br><br>prodrig@pucp.edu.pe<br>+511 626 2000 ext 4681 | Paul Rodriguez received the BSc degree in electrical engineering from the ``Pontificia Universidad Católica del Perú' {PUCP}), Lima, Peru, in 1997, and the MSc and PhD degrees in electrical engineering from the University of New Mexico, U.S., in 2003 and 2005 respectively. He spent two years (2005-2007) as a postdoctoral researcher at Los Alamos National Laboratory, and is currently a Full Professor with the Department of Electrical Engineering at PUCP.<br><br>His research interests include AM-FM models, parallel algorithms, adaptive signal decompositions, and optimization algorithms for inverse problems in signal and image processing such Total Variation, Basis Pursuit, principal component pursuit (a.k.a. robust PCA), convolutional sparse representations, extreme learning machines, etc. |