

# Stochastic Optimization: [ICML 2010](#) Tutorial

[Nathan Srebro](#) and [Ambuj Tewari](#)

## Topic Overview

Stochastic Optimization played an important role in Machine Learning in the past, and is lately again playing an increasingly important role, both as a conceptual framework and as a computational tool. This is not at all surprising.

First, the standard Statistical Learning setups, e.g. the (agnostic) PAC Framework and Vapnik's General Learning Setting can be viewed as special cases of Stochastic Optimization, where one wishes to minimize the generalization error based on stochastic estimates. Viewed as such, familiar Statistical Learning Theory guarantees can be seen as special cases of results in Stochastic Optimization, and familiar learning rules as special cases of Stochastic Optimization methods. Lower bounds for oracle models in Stochastic Optimization can also sometimes be translated to learning lower bounds.

Second, even when training is phrased as a global optimization of some deterministic objective involving the training data, Stochastic Optimization approaches are often the correct choice for optimizing the training objective. The advantages of stochastic approaches in such cases have recently been explored both empirically and theoretically \cite{BoLe03,Leon,MDLW}, and are gaining in popularity.

The goal of this tutorial is both to explore the historical, conceptual and theoretical connections (and differences) between Stochastic Optimization and Statistical Learning, and to introduce the Machine Learning community to the latest research and techniques coming out of the Stochastic Optimization community, exploring how these can be used in learning problems.

## Target Audience

The tutorial is aimed at Machine Learning practitioners and researchers working on learning algorithms, wanting to obtain a better familiarity with Stochastic optimization and its techniques; at researchers working on Online Learning and Statistical Learning Theory wanting to see how ideas from Stochastic Optimization can benefit them, and wanting to familiarize themselves with the latest results coming from Stochastic Optimization; and broadly at researchers interested in understanding the connections between Stochastic Optimization, Online learning and Statistical Learning Theory.

We will not expect any background in Stochastic Optimization, nor will we assume any specific knowledge in Statistical Learning. The conceptual and theoretical parts of the tutorial will be self contained, although a general familiarity with basic concepts in Machine Learning (generalization error, training set, learning rule, etc) will of course be assumed. In describing how specific Stochastic Optimization methods can be applied to learning problems, we will be assuming some familiarity with popular learning methods, such as SVMs,  $L_1$  regularized learning (e.g. the LASSO), multi-task learning with matrix regularization, etc. Although we will not assume any specific background, we will make comments and point out connections aimed at attendees familiar with typical Statistical Learning guarantees, with Online Learning and with regret analysis. We expect that attendees not previously familiar with these topics too will be able to follow along, and enjoy, appreciate and learn from the tutorial.

## Content

- Stochastic Optimization: Basic setup; Statistical Learning as a Stochastic Optimization problem
- A bit of history in different fields: Statistics, Signal Processing, Optimization, Machine Learning
- Two approaches

- SAA (sample average approximation) approach in Stochastic Optimization/ERM (Empirical Risk Minimization) in Machine Learning
- SA (stochastic approximation) approach in Stochastic Optimization/Online Learning approach in Machine Learning
- Hybrid approach: Using SA/Online techniques to minimize an empirical objective.
- Online Convex Optimization (OCO) and regret
  - Reducing Stochastic Convex Optimization to OCO
  - Online to batch conversion via concentration inequalities
- Stochastic Gradient Descent (SGD) for convex Lipschitz functions
  - Basic guarantees
  - Generalization to Mirror Descent
- SGD for strongly convex objectives
- Accelerated methods for smooth objectives
  - Deterministic case -- how are accelerated rates possible?
  - Recent extensions to the stochastic setting; discussion of the variance in the stochastic (sub)gradient estimates
- Lower bounds for the deterministic and stochastic (first-order) oracle models
- Stochastic Coordinate Descent
- Examples from Machine Learning: PEGASOS, SMIDAS, SCD for dual of SVM, SCD for  $L_1$  regularization, gradient descent for trace norm regularization.

## Slides

Slides presented at the tutorial are [here](#).

## References

- Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In Proceedings of the 21st Annual Conference on Learning Theory, pages 414–424. Omnipress, 2008.
- Leon Bottou. Online algorithms and stochastic approximations. In David Saad, editor, Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK, 1998.
- Leon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In NIPS, 2007.
- L'eon Bottou and Yann LeCun. Large scale online learning. In NIPS, 2003.
- N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. IEEE Transactions on Information Theory, 50(9):2050-2057, September 2004.
- Chonghai Hu, James Kwok, and WeiKe Pan. Accelerated gradient methods for stochastic optimization and online learning. In NIPS, 2009.
- Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In Advances in Neural Information Processing Systems 21, pages 801-808. MIT Press, 2009.
- G. Lan. An optimal method for stochastic composite optimization, 2008. submitted.
- Guanghui Lan. Convex Optimization Under Inexact First-order Information. PhD thesis, Georgia Institute of Technology, 2009.
- Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and Bregman iterative methods for matrix rank minimization. Mathematical Programming Series A, 2009. submitted.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 19(4):1574-1609, 2009.
- A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. Nauka Publishers, Moscow, 1978.
- Yu. Nesterov. A method for solving a convex programming problem with convergence rate  $1/k^2$ . Soviet Mathematics Doklady, 27(2):372-376, 1983.
- Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-GrAdient SOLver for SVM. In ICML, pages 807-814, 2007.
- Shai Shalev-Shwartz and Nathan Srebro. SVM optimization: inverse dependence on training set size. In ICML, pages 928-935, 2008.

- Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for  $l_1$  regularized loss minimization. In Proceedings of the 26th International Conference on Machine Learning, pages 929-936. ACM Press, 2009.
- Alexander Shapiro, Darinko Dentcheva, and Andrzej Ruszczyński. Lectures on Stochastic Programming: Modeling and Theory. SIAM, 2009.
- P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. SIAM Journal on Optimization, 2008. submitted.