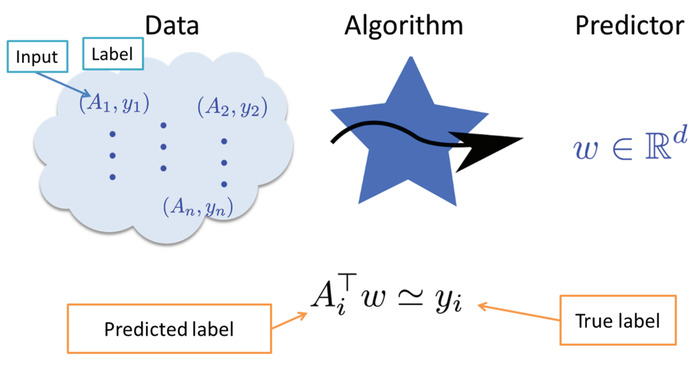
Randomized Methods in Large-Scale Optimization

Dr. Z. Qu, Department of Mathematics

In the field of machine learning, the goal of supervised learning is to infer a discriminant function which maps objects to labels on the basis of a set of labeled training examples (data). Objects (input) are represented via vectors desc ibing their characteristics (features). Labels (output) can either be discrete (classification problem) or continuous (regression problem). The inferred function should predict the correct output value for any valid input object.

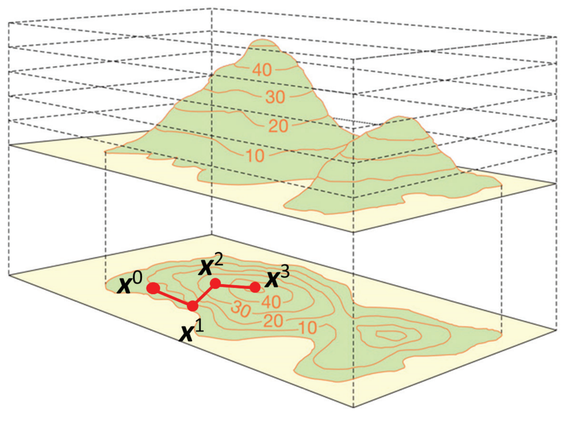


The performance of the inferred function (predictor) is measured through the average loss over the training examples, each of which is a pair of vector and label. A learning algorithm seeks to find a discriminant function minimizing the average loss, which in various types of models leads to the following empirical risk minimization (ERM) problem:

$$\min_{w\in\mathbb{R}^d} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi_i(A_i^\top w)}_{\text{empirical risk}} + \underbrace{g(w)}_{\text{regularization}}$$

Many key machine learning optimization problems are supported by this framework, including the problem of training linear support vector machines (SVM), logistic regression, ridge regression, LASSO and multiclass SVM. Real-world applications can be found in recommender system, image classification, power management, web search, fraud-detection, spam filtering, handwriting recognition and computational finance.

With the advent of "big data" era it becomes increasingly common in the machine learning context that both the number of training examples and the number of features can be huge (billions or more). In that case, even the simplest full-dimensional vector operations (such as a full gradient evaluation) may be too expensive and so algorithms relying on full gradient information become inefficient.



To tackle problems of huge size, stochastic or randomized optimization techniques have become very popular in the past few years for their cheap computational cost per iteration, enabling them to efficiently find an approximate solution. We focus on the development of randomized algorithms for large-scale optimization problems arising in the context of statistical learning and big data analysis.

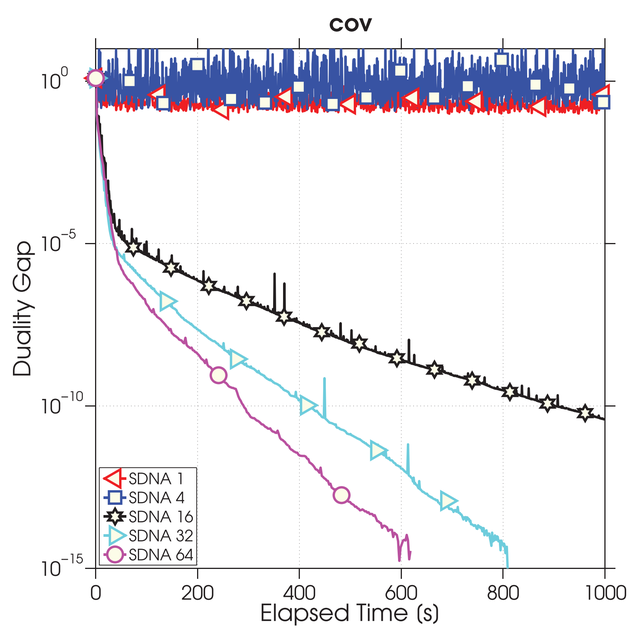


Figure 1 [1]

Newton method achieves local quadratic convergence rate by making use of the curvature information contained in the Hessian matrix of the objective function. It is clear that in a large-scale setting, second order methods like quasi-Newton methods are more prohibitive than first-order full gradient methods. A new research direction consists in incorporating local curvature information into the update of randomly selected coordinates. Figure 1 demonstrates how we can accelerate the existing randomized first-order method for solving the ERM problem by incorporating local curvature information hidden in the principal submatrices of the Hessian matrix. It is clear from the figure that the algorithm converges significantly faster by increasing the size of the submatrix (from 1 to 64). This encouraging numerical result still lacks strong theoretical support and we are currently working in this direction. Another promising new research direction includes the use of adaptive sampling probabilities in randomized methods. Almost all of the existing randomized type methods fix beforehand a sampling distribution and then select iteratively the subset of features or samples to update in an i.i.d. fashion. In [2] we propose an adaptive sampling probabilities by employing the duality theory and provide a theoretical guarantee. The introduction of adaptive sampling results in a striking improvement in computational time.(see Figure 2).

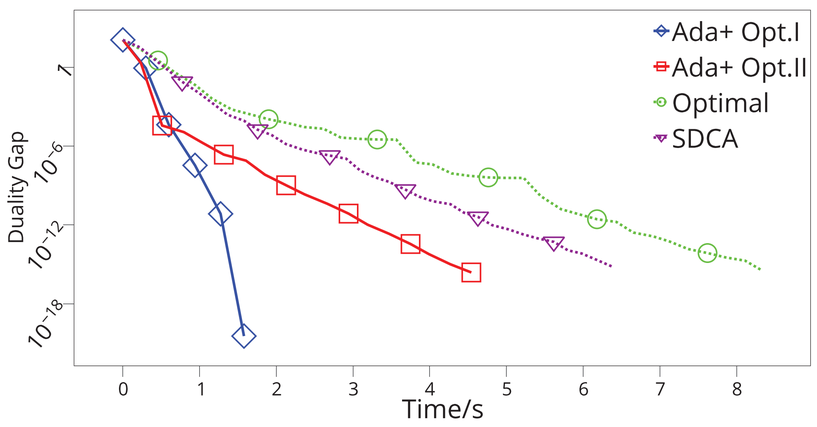


Figure 2 [2]

Further Reading

[1] Z. Qu, P. Richtarik, M. Takac & O. Fercoq. SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization. *ICML* 2016.
[2] D. Csiba, Z. Qu & P. Richtarik. Stochastic Dual Coordinate Ascent with Adaptive Probabilities. *ICML* 2015.
[3] Z. Qu, P. Richtarik & T. Zhang. Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling. *NIPS 2015.*