Neural Networks 84 (2016) 28–38

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Piece-wise quadratic approximations of arbitrary error functions for fast and robust machine learning

A.N. Gorban [a,*], E.M. Mirkes [a], A. Zinovyev [b]

[a] Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK
[b] Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, F-75005, Paris, France

HIGHLIGHTS

- The quadratic error functionals demonstrate many weaknesses for complex data.
- The back side of the non-quadratic error functionals is cost for optimization.
- New algorithms use Piece-wise Quadratic potentials of SubQuadratic growth (PQSQ).
- PQSQ-based algorithms are as fast as the fast heuristic methods but more accurate.
- PQSQ-based algorithms are computationally efficient for regularized sparse regression.

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 10 August 2016
Accepted 19 August 2016
Available online 30 August 2016

Keywords:
Data approximation
Nonquadratic potential
Principal components
Clustering
Regularized regression
Sparse regression

ABSTRACT

Most of machine learning approaches have stemmed from the application of minimizing the mean squared distance principle, based on the computationally efficient quadratic optimization methods. However, when faced with high-dimensional and noisy data, the quadratic error functionals demonstrated many weaknesses including high sensitivity to contaminating factors and dimensionality curse. Therefore, a lot of recent applications in machine learning exploited properties of non-quadratic error functionals based on L_1 norm or even sub-linear potentials corresponding to quasinorms L_p ($0 < p < 1$). The back side of these approaches is increase in computational cost for optimization. Till so far, no approaches have been suggested to deal with *arbitrary* error functionals, in a flexible and computationally efficient framework. In this paper, we develop a theory and basic universal data approximation algorithms (k-means, principal components, principal manifolds and graphs, regularized and sparse regression), based on piece-wise quadratic error potentials of subquadratic growth (PQSQ potentials). We develop a new and universal framework to minimize *arbitrary sub-quadratic error potentials* using an algorithm with guaranteed fast convergence to the local or global error minimum. The theory of PQSQ potentials is based on the notion of the cone of minorant functions, and represents a natural approximation formalism based on the application of min-plus algebra. The approach can be applied in most of existing machine learning methods, including methods of data approximation and regularized and sparse regression, leading to the improvement in the computational cost/accuracy trade-off. We demonstrate that on synthetic and real-life datasets PQSQ-based machine learning methods achieve orders of magnitude faster computational performance than the corresponding state-of-the-art methods, having similar or better approximation accuracy.

1. Introduction

Modern machine learning and artificial intelligence methods are revolutionizing many fields of science today, such as medicine, biology, engineering, high-energy physics and sociology, where large amounts of data have been collected due to the emergence of new high-throughput computerized technologies. Historically and methodologically speaking, many machine learning algorithms have been based on minimizing the mean squared error potential, which can be explained by tractable properties of normal distribution and existence of computationally efficient methods for quadratic optimization. However, most of the real-life datasets

* Corresponding author.
E-mail addresses: ag153@le.ac.uk (A.N. Gorban), em322@le.ac.uk (E.M. Mirkes), Andrei.Zinovyev@curie.fr (A. Zinovyev).

http://dx.doi.org/10.1016/j.neunet.2016.08.007
0893-6080/

are characterized by strong noise, long-tailed distributions, presence of contaminating factors, large dimensions. Using quadratic potentials can be drastically compromised by all these circumstances: therefore, a lot of practical and theoretical efforts have been made in order to exploit the properties of non-quadratic error potentials which can be more appropriate in certain contexts. For example, methods of regularized and sparse regression such as lasso and elastic net based on the properties of L_1 metrics (Tibshirani, 1996; Zou & Hastie, 2005) found numerous applications in bioinformatics (Barillot, Calzone, Hupe, Vert, & Zinovyev, 2012), and L_1 norm-based methods of dimension reduction are of great use in automated image analysis (Wright et al., 2010). Not surprisingly, these approaches come with drastically increased computational cost, for example, connected with applying linear programming optimization techniques which are substantially more expensive compared to mean squared error-based methods.

In practical applications of machine learning, it would be very attractive to be able to deal with *arbitrary error potentials*, including those based on L_1 or fractional quasinorms L_p ($0 < p < 1$), in a computationally efficient and scalable way. There is a need in developing methods allowing to tune the *computational cost/accuracy of optimization* trade-off accordingly to various contexts.

In this paper, we suggest such a universal framework able to deal with a large family of error potentials. We exploit the fact that finding a minimum of a piece-wise quadratic function, or, in other words, a function which is the *minorant of a set of quadratic functionals*, can be almost as computationally efficient as optimizing the standard quadratic potential. Therefore, if a given arbitrary potential (such as L_1-based or fractional quasinorm-based) can be approximated by a piece-wise quadratic function, this should lead to relatively efficient and simple optimization algorithms. It appears that only potentials of quadratic or subquadratic growth are possible in this approach: however, these are the most useful ones in data analysis. We introduce a rich family of piece-wise quadratic potentials of subquadratic growth (PQSQ-potentials), suggest general approach for their optimization and prove convergence of a simple iterative algorithm in the most general case. We focus on the most used methods of data dimension reduction and regularized regression: however, potential applications of the approach can be much wider.

Data dimension reduction by constructing explicit low-dimensional approximators of a finite set of vectors is one of the most fundamental approach in data analysis. Starting from the classical data approximators such as k-means (Lloyd, 1957) and linear principal components (PCA) (Pearson, 1901), multiple generalizations have been suggested in the last decades (self-organizing maps, principal curves, principal manifolds, principal graphs, principal trees, etc.) (Gorban, Kegl, Wunsch, & Zinovyev, 2008; Gorban & Zinovyev, 2009) in order to make the data approximators more flexible and suitable for complex data structures.

We solve the problem of approximating a finite set of vectors $\vec{x}_i \in R^m$, $i = 1, \ldots, N$ (dataset) by a simpler object L embedded into the data space, such that for each point $\vec{x}_i$ an approximation error $err(\vec{x}_i, L)$ function can be defined. We assume this function in the form

$$err(\vec{x}_i, L) = \min_{y \in L} \sum_k u(x_i^k - y^k), \tag{1}$$

where the upper $k = 1, \ldots, m$ stands for the coordinate index, and $u(x)$ is a monotonously growing symmetrical function, which we will be calling the error potential. By data approximation we mean that the embedment of L in the data space minimizes the error

$$\sum_i err(\vec{x}_i, L) \to \min .$$

Note that our definition of error function is coordinate-wise (it is a sum of error potential over all coordinates).

The simplest form of the error potential is quadratic $u(x) = x^2$, which leads to the most known data approximators: mean point (L is a point), principal points (L is a set of points) (Flury, 1990), principal components (L is a line or a hyperplane) (Pearson, 1901). In more advanced cases, L can possess some regular properties leading to principal curves (L is a smooth line or spline) (Hastie, 1984), principal manifolds (L is a smooth low-dimensional surface) and principal graphs (eg., L is a pluri-harmonic graph embedment) (Gorban, Sumner, & Zinovyev, 2007; Gorban & Zinovyev, 2009).

There exist multiple advantages of using quadratic potential $u(x)$, because it leads to the most computationally efficient algorithms usually based on the splitting schema, a variant of expectation–minimization approach (Gorban & Zinovyev, 2009). For example, k-means algorithm solves the problem of finding the set of principal points and the standard iterative Singular Value Decomposition finds principal components. However, quadratic potential is known to be sensitive to outliers in the dataset. Also, purely quadratic potentials can suffer from the curse of dimensionality, not being able to robustly discriminate 'close' and 'distant' point neighbors in a high-dimensional space (Aggarwal, Hinneburg, & Keim, 2001).

There exist several widely used ideas for increasing approximator's robustness in presence of strong noise in data such as: (1) using medians instead of mean values, (2) substituting quadratic norm by L_1 norm (e.g. Ding, Zhou, He, & Zha, 2006 and Hauberg, Feragen, & Black, 2014), (3) outliers exclusion or fixed weighting or iterative reweighting during optimizing the data approximators (e.g. Allende, Rogel, Moreno, & Salas, 2004; Kohonen, 2001 and Xu & Yuille, 1995), and (4) regularizing the PCA vectors by L_1 norm (Candès, Li, Ma, & Wright, 2011; Jolliffe, Trendafilov, & Uddin, 2003; Zou, Hastie, & Tibshirani, 2006). In some works, it was suggested to utilize 'trimming' averages, e.g. in the context of the k-means clustering or some generalizations of PCA (Cuesta-Albertos, Gordaliza, & Matrán, 1997; Hauberg et al., 2014). In the context of regression, iterative reweighting is exploited to mimic the properties of L_1 norm (Lu, Lin, & Yan, 2015). Several algorithms for constructing PCA with L_1 norm have been suggested (Brooks, Dulá, & Boone, 2013; Ke & Kanade, 2005; Kwak, 2008) and systematically benchmarked (Brooks & Jot, 2012; Park & Klabjan, 2014). Some authors go even beyond linear metrics and suggest that fractional quasinorms L_p ($0 < p < 1$) can be more appropriate in high-dimensional data approximation (Aggarwal et al., 2001).

However, most of the suggested approaches exploiting properties of non-quadratic metrics either represent useful but still arbitrary heuristics or are not sufficiently scalable. The standard approach for minimizing L_1-based norm consists in solving a linear programming task. Despite existence of many efficient linear programming optimizer implementations, by their nature these computations are much slower than the iterative methods used in the standard SVD algorithm or k-means.

In this paper, we provide implementations of the standard data approximators (mean point, k-means, principal components) using a PQSQ potential. As an other application of PQSQ-based framework in machine learning, we develop PQSQ-based regularized and sparse regression (imitating the properties of lasso and elastic net).

2. Piecewise quadratic potential of subquadratic growth (PQSQ)

2.1. Definition of the PQSQ potential

Let us split all non-negative numbers $x \in R_{\geq 0}$ into $p + 1$ non-intersecting intervals $R_0 = [0; r_1), R_1 = [r_1; r_2), \ldots, R_k = [r_k; r_{k+1}), \ldots, R_p = [r_p; \infty)$, for a set of thresholds $r_1 < r_2 < \cdots < r_p$. For convenience, let us denote $r_0 = 0, r_{p+1} = \infty$. Piecewise quadratic potential is a continuous monotonously