RC25183 (W1107-016) July 6, 2011
Mathematics

IBM Research Report

Convexization

Dimitri Kanevsky
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

CONVEXIZATION

Dimitri Kanevsky

Abstract

Maximization of many functions is done efficiently via a recursive process that involve optimization of auxiliary concave functions at intermediate steps. In the paper we describe a process that allows constructively build concave auxiliary functions. This process can be applied to optimization of exponential families and to recently introduced $\mathcal{A}$-functions.

1 Introduction

The goal of this paper is to introduce a general convexization process for arbitrary functions to assist maximization of non-concave functions. There are various efficient methods in general to maximize concave functions. It is naturally in optimization of non-concave functions to use concave lower bound functions in intermediate steps. Since many processes in general are governed by non concave objective functions this make maximization process based on them difficult. Standard regularization of these processes while improves optimization accuracy performance it slows them significantly. In our paper we suggest a different convexization recursion process that is based on a novel way to transform locally any objective function into auxiliary concave functions. The recursion process is then applied to this auxiliary functions and update is found. This update process is used to create a new concave auxiliary function and so on. The rest of the paper is structured as follows. In Section 2 we introduce a notion of auxiliary function and describe a process how to build them. In Section 3 we extend this method to stochastic functions.

2 Auxiliary functions

2.1 Definition of auxiliary functions

Let $f(x) : \mathcal{U} \subset \mathbb{R}^n \to \mathbb{R}$ be a real valued differentiable function in an open subset $\mathcal{U}$. Let $\mathbf{Q}_f = \mathbf{Q}_f(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be twice differentiable in $x \in \mathcal{U}$ for each $y \in \mathcal{U}$. We define $\mathbf{Q}_f$ as an auxiliary function for f in $\mathcal{U}$ if the following properties hold.

1. $\mathbf{Q}_f(x, y)$ is a strictly concave function of x for any $y \in \mathcal{U}$ with its (unique) maximum point belonging to $\mathcal{U}$ (recall that twice differentiable function is strictly concave or convex over some domain if its Hessian function is positive or negative definite in the domain, respectively).
2. Hyperplanes tangent to manifolds defined by $z = g_y(x) = \mathbf{Q}_f(x, y)$ and $z = f(x)$ at any $x = y \in \mathcal{U}$ are parallel to each other, i.e.

$$\nabla_x \mathbf{Q}_f(x, y)|_{x=y} = \nabla_x f(x) \tag{1}$$

3. For any $x \in \mathcal{U}$ $f(x) = \mathbf{Q}_f(x, x)$

4. For any $x, y \in \mathcal{U}$ $f(x) \geq \mathbf{Q}_f(x, y)$

In an optimization process via an $\mathcal{Q}$-function it is usually assumed that finding an optimum of an $\mathcal{Q}$-function is "easier" than finding a (local) optimum of the original function f. Naturally, a desired outcome is for the equation $\nabla_x \mathbb{Q}_f(x, y) = 0$ to have a closed form solution.

2.2 Optimization process for auxiliary functions

The optimization recursion via an auxiliary function can be described as follows (where we use EM style).

E-step Given x^t construct $\mathbb{Q}_f(x, x^t)$

M-step Find

$$x^{t+1} = \arg\max_{x \in \mathcal{U}} \mathbb{Q}_f(x, x^t) \tag{2}$$

For updates (2) we have $f(x^{t+1}) = \mathbb{Q}_f(x^{t+1}, x^{t+1}) \geq \mathbb{Q}_f(x^t, x^{t+1}) \geq \mathbb{Q}_f(x^t, x^t) = f(x^t)$. This means that iterative update rules have a "growth" property (i.e. the value of the original function increases for the new parameters values). This is illustrated in the following plot.

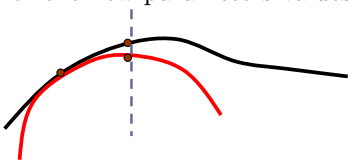

Fig. 1. Auxiliary function

In this figure the upper curve denotes the plot of the objective function $f : x \to \mathbb{R}$ and the curve in red, i.e. the convex lower curve, represents an auxiliary function.

Let call a point $x \in \mathcal{U}$ critical if $\nabla_x f(x) = 0$. We can prove the following convergence statement.

Proposition 1 *Let $\mathbf{Q}_f$ be an auxiliary function for f in $\mathcal{U}$ and let $\mathcal{S} = \{x^t, t = 1, 2, ...\}$. Then all limit points of $\mathcal{S}$ that lie in $\mathcal{U}$ are critical points. Assume in addition that f has a local maximum at some limit point of the sequence $\mathcal{S}$ in $\mathcal{U}$ and that f is strictly concave in some open neighborhood of this point. Then there exists only one critical point of $\mathcal{S}$ in $\mathcal{U}$.*

Proof
Let us define the following map

$$T : \{x \in \mathcal{U} \to \tilde{x} = \arg\max_y \mathbf{Q}_f(y, x)\} \tag{3}$$

Let us first prove that limit points of $\mathcal{S}$ are fixed points of T. Indeed, let a be a limit point of $T^{n_i}(x), i = 1, 2, ...$ Then $b = T(a)$ is a limit point for the sequence $T^{n_i+1}(x)$. If $b \neq a$ then we have

$$f(T^{n_i}(x)) \leq f(T^{n_i+1}(x)) \leq f(T^{n_{i+1}}(x)) \tag{4}$$

I.e. $F(a) \leq F(b) \leq F(a)$. This implies $a = b$.
Next, one can see that fixed points of T that belong $\mathcal{U}$ are critical points. Indeed, if $T(a) = a \in \mathcal{U}$ then $a = \arg\max_x \mathbf{Q}_f(x, a)$ and $\nabla_x \mathbf{Q}_f(x, a)|_{x=a} = \nabla_a f(a) = 0$.
The final statement of the proposition follows from the fact that if f is strictly concave in some open neighborhood of a critical point $a \in \mathcal{U}$ than for sufficiently close to a points x^t we have $x^{t+1} = \arg\max_x \mathbf{Q}_f(x, x^t)$ is close to a.

2.3 Auxiliary functions for convex functions

In this section we describe how to construct auxiliary functions (that are strictly concave) for strictly convex functions.
Assume that $f(x)$ is strictly convex in $\mathcal{U}$. Then for any point $x \in \mathcal{U}$ we can construct a family of auxiliary functions as follows.
Let us consider the following family of functions.

$$\mathbf{Q}_f(y, x; \lambda) = -\lambda f\left(-\frac{y}{\lambda} + x\left(1 + \frac{1}{\lambda}\right)\right) + f(x) + \lambda f(x) \tag{5}$$

These family functions (5) obey properties 1-3 for any $\lambda > 0$ in the definition of auxiliary function. The family of functions (5) are obtained via subsequent applications of the following three transformations.

Reflection along x-axis

$$\mathbf{H}_f(y, x) = -f(y) + 2f(x) \tag{6}$$

Reflection along y-axis

$$\mathbf{G}_f(y, x) = \mathbf{H}_f(-y + 2x, x) + 2\mathbf{H}_f(x, x) \tag{7}$$

Scaling

$$\mathbf{Q}_f(y, x; \lambda) = \lambda \mathbf{G}_f\left(\frac{y}{\lambda} + x\left(1 - \frac{1}{\lambda}\right), x\right) + (1 - \lambda)\mathbf{G}_f(x, x) \tag{8}$$

Various properties of these transformation will be described in forthcoming papers. Specifically, one can easily to see that when a scaling factor λ grows, the function $\mathbf{Q}_f(y, x; \lambda)$ becomes larger and its extremum is moving father from x. Another useful fact is: if $f(x)$ is convex then for any $\lambda > 0, x, y \in \mathcal{U}$ one has $\mathbf{Q}_f(y, x; \lambda) \leq f(x)$

In general, for an arbitrary function $f(x)$ one can construct auxiliary functions $\mathbf{Q}_f(y, x; \lambda)$ locally (with different λ in neighborhoods for different points x).

This method of building auxiliary functions can be extended to more general class of functions as shown in next sections.

2.4 Objective function that is sum of convex and concave functions

Assume that

$$f(x) = g(x) + h(x) \tag{9}$$

where $h(x)$ is strictly convex in $\mathcal{U}$. Then we can define an auxiliary function for $f(x)$ as following

$$Q_f(y,x) = Q_g(y,x) + Q_h(y,x,\lambda) \tag{10}$$

where $Q_g(y,x)$ is some auxiliary function associated with g (for example it coincides with $g(x)$ if $g(x)$ is strictly concave). In practical applications some function $Q_h(y,x,\lambda)$ may be concave but not strictly concave. In this case one can add a small regularized penalty to it to make it strictly concave.

2.4.1 Exponential families

The important example of convex functions is an exponential family. We define an exponential family as any family of densities on $\mathbb{R}^D$, parameterized by θ, that can be written as $\xi(x,\theta) = \frac{\exp\{\theta^T\phi(x)\}}{Z(\theta)}$ where x is a D-dimensional base observation. The function $\phi : \mathbb{R}^D \to \mathbb{R}^d$ characterizes the exponential family. $Z(\theta) = \int_{\Xi} \exp\{\theta^T\phi(x)\}dx$ is the partition function, that provides the normalization necessary for $\xi(x,\theta)$. The function $\log\xi(x,\theta)$ is convex and it is strictly convex if $Var[\phi(x)] \neq 0$ ([3]). Some objective functions of exponential densities (e.g. in energy-based models) can be optimized via a recursion procedure that at each recursion require optimization of weighed sum of exponential densities, i.e., a sum of convex and concave functions.

2.5 $\mathcal{A} - functions$

$\mathcal{A}$-functions were introduced in ([2]). An $\mathcal{A}$-function $\mathbb{A}_f$ for f is defined exactly as the auxiliary function $\mathbf{Q}_f$ except that in the condition 1. one need to require that $\mathcal{A}$-function is either strictly concave or strictly convex. A growth transformation in ([2]) was formulated such that the next step in the parameter update that increases $f(x)$ is obtained as a linear combination of the current parameter values with the value that optimizes the $\mathcal{A}$-function, i.e. $\nabla_x \mathbb{A}_f(x,y)|_{x=\tilde{x}} = 0$. If an $\mathbb{A}_f$ is convex then in our novel approach one can apply a transformation (5) to create an auxiliary function. Otherwise, if this $\mathcal{A}$-function is concave already, one can just apply a scaling transformation (8) to construct an auxiliary function.

3 Stochastic Auxiliary functions

3.1 Stochastic convex learning (summary)

Let us consider a convex minimization problem as follows. We assume that our goal is to find some parameter vector $x \in \mathcal{U}$ such that sum of convex functions $f^i \to \mathbb{R}$ takes on the smallest value possible. Order of the functions f_i can be chosen in response to our previous choice of x or the functions f^i can be drawn from some distribution. It is our goal to find a sequence of x^i that cumulative value (empirical loss or risk in machine learning terminology) of $f^i(x^i)$ is minimized. Let average cumulative loss is defined as

$$f^*(x) = \frac{1}{T}\sum_{t=1}^{T} f^t(x) \tag{11}$$

and

$$x^* = \arg\min_{x \in \mathcal{U}} f^*(x) \tag{12}$$

We assume that $x^* \in \mathcal{U}$ exists and satisfies $||x^*|| < R$. Minimizing the function (11) can be done using batch gradient descent

$$x^{t+1} = x^t - \gamma_t \nabla_x f^*(x^t) = x^t - \gamma_t \frac{1}{T} \sum_{i=1}^{T} \nabla_x f^i(x^t) \tag{13}$$

The properties of this algorithm are well known. When learning rates of γ_t are small enough the algorithm converges to a local minimum of $f^*(x)$.Each iteration of batch gradient descent involves however a burdening computation of the average of the gradient of the gradients of the function $f^*(x)$ over an entire training set. Significant resources must be allocated in order to store the large enough training set and compute this average. The elementary online gradient descent algorithm is obtained by dropping the averaging operation in the batch descent algorithm. Instead of averaging the gradient of the function f^* over the complete training set each operation of the online gradient descent consist of choosing a function f^t at random (as corresponding to a random training example) and and updating the parameter x^t according to the formula

$$x^{t+1} = x^t - \gamma_t \nabla_x f^t(x^t t) \tag{14}$$

It is well known that general convexity condition

$$\forall \epsilon > 0 \ , \qquad \inf_{(x-x^*)^2 > \epsilon} (x - x^*) \nabla_x f(x) > o \tag{15}$$

and condition

$$\sum \gamma_t^2 < \infty \tag{16}$$

and also the condition that $\nabla_x^2 f(x)$ behaves quadratically with the final convergence region are sufficient conditions for almost sure convergence of the stochastic gradient descent (14) to the optimum x^* (see, for example, [4]).

3.2 Auxiliary stochastic functions

Assume now that functions $f_i(x)$ are non-concave and we need to solve the maximization problem

$$\max \sum f^i(x) \tag{17}$$

Assume also that $Q^i(x, y)$ are auxiliary functions for $f^i(x)$ at x. In this case one can consider the following optimization process.
Let

$$Q^*(x, y) = \sum Q^i(x, y) \tag{18}$$

Then $Q^*(x, y)$ is an auxiliary function for $f^*(x) = \sum f^i(x)$. For $t = 1, 2, ...$ we can optimize $Q^*(x^t, y)$ using stochastic descent methods and find x^{t+1}. This induces the optimization process for $f^*(x)$ via the auxiliary function $Q^*(x, y)$.

References

[1] Yann LeCunn et al., A tutorial on Energy-Based Learning, http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf

[2] D. Kanevsky, D. Nahamoo, T. Sainath and B. Ramabhadran and ,"A-Functions: A Generalization of Extended Baum-Welch Transformations o Convex Optimization", in Proc. ICASSP, 2011

[3] Goel, V., Olsen, P., Acoustic Modeling Using Exponential Families, 2009, October, Proc. Interspeech

[4] Leon Bottou, "Stochastic learning", Advanced Lectures on Machine Learning, 146-168, Edited by Olivier Bousquet and Ulrike von Luxburg, Lecture Notes in Artificial Intelligence, LNAI 3176, Springer Verlag, Berlin, 2004.