q-Gaussian based Smoothed Functional Algorithm for Stochastic Optimization

Debarghya Ghoshdastidar, Ambedkar Dukkipati and Shalabh Bhatnagar
Department of Computer Science and Automation
Indian Institute of Science
Bangalore – 560012
Email: {gdebarghya@ee,ambedkar@csa,shalabh@csa}.iisc.ernet.in

Abstract—The q-Gaussian distribution results from maximizing certain generalizations of Shannon entropy under some constraints. The importance of q-Gaussian distributions stems from the fact that they exhibit power-law behavior, and also generalize Gaussian distributions. In this paper, we propose a Smoothed Functional (SF) scheme for gradient estimation using q-Gaussian distribution, and also propose an algorithm for optimization based on the above scheme. Convergence results of the algorithm are presented. Performance of the proposed algorithm is shown by simulation results on a queuing model.

I. INTRODUCTION

Stochastic optimization algorithms play an important role in optimization problems involving objective functions that cannot be computed analytically. These schemes are extensively used in discrete event systems, such as queuing systems, for obtaining optimal or near-optimal performance measures.

Gradient descent algorithms are used for stochastic optimization by estimating the gradient of average cost in the long run. Methods for gradient estimation by random perturbation of parameters have been proposed in [1]. The Smoothed Functional scheme, described in [2], approximates the gradient of expected cost by its convolution with a multivariate normal distribution. Based on all the above schemes, two-timescale stochastic approximation algorithms have been presented in [3], which simultaneously perform cost averaging and parameter updation using different step-size schedules.

In this paper, we propose a new SF method based on q-Gaussian distribution, which is a generalization of the normal distribution. We show that q-Gaussian satisfies all the conditions for smoothing kernels proposed by Rubinstein [4]. We illustrate a method for gradient estimation using q-Gaussian. We also present a two-timescale algorithm for stochastic optimization using q-Gaussian based SF, and show the convergence of the proposed algorithm.

The rest of the paper is organized as follows. The framework for the optimization problem and some of the preliminaries are presented in Section II. The gradient estimates using q-Gaussian SFs have been derived in Section III. Section IV presents the proposed algorithm. Numerical experiments comparing our algorithm with a previous algorithm is presented in Section V. An outline of convergence analysis of our algorithm is discussed in Section VI. Finally, Section VII provides the concluding remarks.

II. BACKGROUND AND PRELIMINARIES

A. q-Gaussian distribution

Most of the distributions, like normal, uniform, exponential etc., can be obtained by maximizing Shannon entropy functional defined as $H(p) = \int X p(x) \ln p(x) dx$, where $p(.)$ is a pdf defined on the sample space $X$. Other entropy functions have also been proposed as generalized information measures. One of the most popular among them is Tsallis entropy [5] defined as

$$H_q(p) = \frac{1 - \int X p(x)^q dx}{q - 1}, \quad q \in \mathbb{R}. \quad (1)$$

Its consistency with the discrete case is shown in [6], and the Shannon-Khinchin axioms for the discrete case in such a generalized setting is established in [7]. This entropy function produces Shannon entropy as $q \rightarrow 1$. Corresponding to this generalized measure, $q$-expectation of a function $f(.)$ can be defined as

$$\langle f(x) \rangle_q := \frac{\int X f(x)p(x)^q dx}{\int X p(x)^q dx}. \quad (2)$$

Maximizing Tsallis entropy under the following constraints:

$$\langle x \rangle_q = \mu \quad \text{and} \quad \langle x^2 \rangle_q = \beta^2, \quad (3)$$

results in q-Gaussian distribution [8], which is of the following form:

$$G_{q,\beta}(x) = \frac{1}{\beta K_q} \left(1 - \frac{(1 - q)(x - \mu)^2}{(3 - q)\beta^2}\right)^{-\frac{1}{q-1}}, \quad (4)$$

where $y_{+} = \max(y, 0)$ is called Tsallis cut-off condition, and $K_q$ is the normalizing constant, which depends on the value of $q$. The function defined in (4) is not integrable for $q < 3$, and hence, q-Gaussian is a probability density function only for $q \geq 3$.

Multivariate form of the q-Gaussian distribution [9] is defined in the following way:

$$G_{q,\beta}(X) = \frac{1}{\beta^N K_{q,N}} \left(1 - \frac{(1 - q)\|X\|^2}{(3 - q)\beta^2}\right)^{-\frac{1}{q-1}}, \quad (5)$$

where $K_{q,N}$ is the normalizing constant. The multivariate normal distribution is a special case of (5) as $q \rightarrow 1$. 

B. Problem Framework

Let \( \{Y_n\}_{n \in \mathbb{N}} \subset \mathbb{R}^d \) be a parameterized Markov process, depending on a tunable parameter \( \theta \in C \), where \( C \) is a compact and convex subset of \( \mathbb{R}^N \). Let \( \mathcal{F}_n(x, dy) \) denote the transition kernel of \( \{Y_n\} \) when the operative parameter is \( \theta \in C \). Let \( h : \mathbb{R}^d \to \mathbb{R}^+ \bigcup \{0\} \) be a Lipschitz continuous cost function associated with the process.

**Assumption I.** The process \( \{Y_n\} \) is ergodic for any given \( \theta \) as the operative parameter, i.e.,

\[
\frac{1}{L} \sum_{m=0}^{L-1} h(Y_m) \to \nu_0 \text{ as } L \to \infty,
\]

where \( \nu_0 \) is the stationary distribution of \( \{Y_n\} \).

Our objective is to minimize the long-run average cost

\[
J(\theta) = \lim_{L \to \infty} \frac{1}{L} \sum_{m=0}^{L-1} h(Y_m) = \int h(x) \nu_0(dx) \quad (6)
\]

by choosing an appropriate \( \theta \in C \). The existence of the above limit is given by Assumption [1]. In addition, we assume that the average cost \( J(\theta) \) satisfies the following condition:

**Assumption II.** \( J(\theta) \) is continuously differentiable with respect to any \( \theta \in C \).

We also assume the existence of a stochastic Lyapunov function through the following assumption:

**Assumption III.** Let \( \{\theta(n)\} \) be a sequence of random parameters, obtained using an iterative scheme, controlling the process \( \{Y_n\} \), and \( \mathcal{F}_n = \sigma(\theta(m), Y_m, m \leq n) \), \( n \geq 0 \) denote the sequence of associated \( \sigma \)-fields. There exists \( \varepsilon_0 > 0 \), \( K \subset \mathbb{R}^d \) compact, and a continuous \( \mathbb{R}^d \)-valued function \( V \), with \( \lim_{\|x\| \to \infty} V(x) = \infty \), such that under any non-anticipative \( \theta(n) \),

(i) \( \sup_{\eta} \mathbb{E}[V(Y_n^2)] < \infty \) and

(ii) \( \mathbb{E}[V(Y_{n+1}) | \mathcal{F}_n] \leq V(Y_n) - \varepsilon_0 \), when \( Y_n \notin K \), \( n \geq 0 \).

Assumption [1] is a technical requirement, whereas Assumption [III] is used to show the stability of the scheme. Assumption [III] will not be required, for instance, if the single-stage cost function \( h \) is bounded in addition.

C. Smoothed Functionals

Given any function \( f : C \mapsto \mathbb{R} \), its smoothed functional is defined as

\[
S_{\beta}[f(\theta)] = \int_{-\infty}^{\infty} G_{\beta}(\eta) f(\theta - \eta) d\eta = \int_{-\infty}^{\infty} G_{\beta}(\theta - \eta) f(\eta) d\eta,
\]

where \( G_{\beta} : \mathbb{R}^N \to \mathbb{R} \) is a kernel function.

The idea behind using smoothed functionals is that if \( f(\theta) \) is not well-behaved, i.e., it has a fluctuating character, then \( S_{\beta}[f(\theta)] \) is better-behaved. This ensures that any optimization algorithm with objective function \( f(\theta) \) does not get stuck at any local minimum, but converges to the global minimum. The parameter \( \beta \) controls the degree of smoothness. Rubinstein [4] shows that the SF algorithm achieves these properties if the kernel function satisfies the following sufficient conditions:

(P1) \( G_{\beta}(\eta) = \frac{1}{\beta^N} G_q \left( \frac{\eta}{\beta} \right) \),

where \( G_q \left( \frac{\eta}{\beta} \right) = G_q \left( \frac{\eta^{(1)}}{\beta}, \frac{\eta^{(2)}}{\beta}, \ldots, \frac{\eta^{(N)}}{\beta} \right) \).

(P2) \( G_{\beta}(\eta) \) is piecewise differentiable in \( \eta \).

(P3) \( G_{\beta}(\eta) \) is a probability distribution function, i.e., \( S_{\beta}[f(\theta)] = \mathbb{E}_{G_{\beta}(\eta)}[f(\theta - \eta)] \).

(P4) \( \lim_{\beta \to \infty} G_{\beta}(\eta) = \delta(\eta) \),

where \( \delta(\eta) \) is the Dirac delta function.

(P5) \( \lim_{\beta \to 0} S_{\beta}[f(\theta)] = f(\theta) \).

The normal distribution satisfies the above conditions, and has been used as a kernel by Katkovnik [2].

Based on (7), a form of gradient estimator has been derived in [3] which is given by:

\[
\nabla_{\theta} J(\theta) \approx \frac{1}{\beta M^2} \sum_{n=0}^{M-1} \sum_{m=0}^{L-1} \eta(n) h(Y_m) \quad (8)
\]

for large \( M \), \( L \) and small \( \beta \). The process \( \{Y_m\} \) is governed by parameter \( (\theta(n) + \beta \eta(n)) \), where \( \eta(n) \in C \subset \mathbb{R}^N \) is obtained through an iterative scheme. \( \eta(n) \) is a \( N \)-dimensional vector composed of i.i.d. \( N(0,1) \)-distributed random variables.

III. \( q \)-GAUSSIAN FOR SMOOTHED FUNCTIONALS

**Proposition 3.1.** The \( q \)-Gaussian distribution satisfies the kernel properties (P1) – (P5) for all \( q < 3 \), \( q \neq 1 \).

**Proof:**

(P1) From (5), it is evident that \( G_{q,\beta}(\eta) = \frac{1}{\beta^N} G_q \left( \frac{\eta}{\beta} \right) \).

(P2) For \( 1 < q < 3 \), \( 1 - \frac{(1-q)}{(3-q)\beta^2} \|\eta\|^2 > 0 \), for all \( \eta \in \mathbb{R}^N \). Hence, \( G_{q,\beta}(\eta) = \frac{1}{\beta^N K_{q,N}(1 - \frac{(1-q)}{(3-q)\beta^2} \|\eta\|^2)^{\frac{1}{q-2}}} \).

Thus, \( \nabla_{\eta} G_{q,\beta}(\eta) = -\frac{2\eta}{(3-q)\beta^2 \left( 1 - \frac{(1-q)}{(3-q)\beta^2} \|\eta\|^2 \right)^{\frac{2}{q-2}}} \). \hspace{1cm} (9)

For \( q < 1 \), when \( \|\eta\|^2 < \frac{(3-q)\beta^2}{(1-q)} \), we have

\[
\left( 1 - \frac{(1-q)}{(3-q)\beta^2} \|\eta\|^2 \right) > 0.
\]

So, (9) holds. On the other hand, when \( \|\eta\|^2 \geq \frac{(3-q)\beta^2}{(1-q)} \), we have

\[
\frac{1}{\beta^N K_{q,N}} \left( 1 - \frac{(1-q)}{(3-q)\beta^2} \|\eta\|^2 \right) \leq 0,
\]

which implies

\[
G_{q,\beta}(\eta) = 0 \quad \text{and} \quad \nabla_{\eta} G_{q,\beta}(\eta) = 0.
\]

Thus, \( G_{q,\beta}(\eta) \) is differentiable for \( q > 1 \), and piecewise differentiable for \( q < 1 \).

(P3) \( G_{q,\beta}(\eta) \) is a distribution for \( q < 3 \) and hence,

\[
S_{\beta}[f(\theta)] = \mathbb{E}_{G_{q,\beta}(\eta)}[f(\theta - \eta)].
\]
(P4) As $\beta \to 0$, $G_{q,\beta}(0) = \frac{1}{\beta^N K_{q,N}} \to \infty$. But, we have
\[ \int_{\mathbb{R}^N} G_{q,\beta}(\eta) d\eta = 1 \text{ for } q < 3. \] So, $\lim_{\beta \to 0} G_{q,\beta}(\eta) = \delta(\eta)$.

(P5) It follows from dominated convergence theorem that
\[ \lim_{\beta \to 0} S_{q,\beta}[f(\theta)] = \lim_{\beta \to 0} \int_{\mathbb{R}^N} G_{q,\beta}(\eta) f(\theta - \eta) d\eta \]
\[ = \int_{\mathbb{R}^N} \delta(\eta) f(\theta - \eta) d\eta = f(\theta). \]

Our objective is to estimate $\nabla_{\theta} J(\theta)$ using the SF approach. The existence of $\nabla_{\theta} J(\theta)$ is due to Assumption [II]. Now,
\[ \nabla_{\theta} J(\theta) = \left[ \nabla_{\theta}^{(1)} J(\theta) \quad \nabla_{\theta}^{(2)} J(\theta) \ldots \nabla_{\theta}^{(N)} J(\theta) \right]^T. \]

Let us define, $\Omega_q = \left\{ \eta \in \mathbb{R}^N : ||\eta||^2 < \frac{(3-q)\beta^2}{1-q} \right\}$ for $q < 1$, and $\Omega_q = \mathbb{R}^N$ for $1 < q < 3$. It is evident that $\Omega_q$ is the support set for the $q$-Gaussian distribution with $q$-variance $\beta^2$. Define the SF for gradient of average cost as,
\[ D_{q,\beta}[J(\theta)] = \left[ S_{q,\beta}[\nabla_{\theta}^{(1)} J(\theta)] \ldots S_{q,\beta}[\nabla_{\theta}^{(N)} J(\theta)] \right]^T \]
\[ = \int_{\mathbb{R}^N} G_{q,\beta}(\theta - \eta) \nabla_{\eta} J(\eta) d\eta \]

It follows from integration by parts and the definition of $\Omega_q$,
\[ D_{q,\beta}[J(\theta)] = \int_{\Omega_q} \nabla_{\eta} G_{q,\beta}(\eta) J(\theta - \eta) d\eta \]

Substituting $\bar{\eta} = -\frac{2}{\beta}$, we have
\[ D_{q,\beta}[J(\theta)] = \int_{\Omega_q} 2 \frac{\bar{\eta} J(\theta + \bar{\eta})}{(3-q)\beta} G_q(\eta) d\eta \]
\[ = \frac{2}{\beta^2(3-q)} \mathbb{E}_{G_q(\eta)} \left[ \frac{\bar{\eta} J(\theta + \bar{\eta})}{1 - \frac{1-q}{3-q} ||\bar{\eta}||^2} \right]. \quad (10) \]

We first state the following lemma which will be required to prove the result in Proposition [3.3].

**Lemma 3.2.** Let $f : \mathbb{R}^N \mapsto \mathbb{R}$ be a function defined over a standard $q$-Gaussian distributed random variable $X \in \mathbb{R}^N$.
\[ \text{i.e., } \langle X \rangle_q = 0 \text{ and } \langle X X^T \rangle_q = I_{N \times N}, \]

then,
\[ \langle f(X) \rangle_q = \frac{1}{\Lambda_q} \mathbb{E}_{G_q(X)} \left[ \frac{f(X)}{1 - \frac{1-q}{3-q} ||X||^2} \right], \]

where $\Lambda_q = [(K_{q,N})^{q-1} \int_{\mathbb{R}^N} G_q(x)^q dx]$, $K_{q,N}$ being the normalizing constant for $N$-variate $q$-Gaussian.

**Proof:** From [2]
\[ \langle f(X) \rangle_q = \frac{1}{\Lambda_q} \int_{\mathbb{R}^N} G_q(x)^q dX \]
\[ = \frac{1}{\Lambda_q} \int_{\mathbb{R}^N} f(x) \left( 1 - \frac{(1-q)||x||^2}{3-q} \right)^{\frac{1-q}{2}} dX \]
\[ = \frac{1}{\Lambda_q} \int_{\Omega_q} \frac{f(x)}{1 - \frac{1-q}{3-q} ||x||^2} G_q(x) dX \]
\[ = \frac{1}{\Lambda_q} \mathbb{E}_{G_q(X)} \left[ \frac{f(X)}{1 - \frac{1-q}{3-q} ||X||^2} \right]. \]

**Proposition 3.3.** For a given $q < 3$, $q \neq 1$, as $\beta \to 0$, SF for the gradient converges to a scaled version of the gradient,
\[ \text{i.e., } \left| D_{q,\beta}[J(\theta)] - \frac{2\Lambda_q}{(3-q)} \nabla_{\theta} J(\theta) \right| \to 0 \text{ as } \beta \to 0. \]

**Proof:** For small $\beta$, using Taylor series expansion,
\[ J(\theta + \bar{\eta}) = J(\theta) + \bar{\eta} \nabla_{\theta} J(\theta) + \frac{1}{2} \beta \bar{\eta}^T \nabla_{\theta}^2 J(\theta) \bar{\eta} + o(\beta^2) \]
By Lemma 3.2
\[ D_{q,\beta}[J(\theta)] = \frac{2\Lambda_q}{\beta(3-q)} \left( \langle \bar{\eta} J(\theta + \bar{\eta}) \rangle_q \right) \]
\[ = \frac{2\Lambda_q}{\beta(3-q)} \left[ \langle \bar{\eta} J(\theta) \rangle_q + \beta \langle \bar{\eta}^T \nabla_{\theta} J(\theta) \rangle_q \right] + \frac{1}{2} \beta \bar{\eta}^T \nabla_{\theta}^2 J(\theta) \bar{\eta} + o(\beta) \]
\[ = \frac{2\Lambda_q}{(3-q)} \left[ \nabla_{\theta} J(\theta) + \beta \left( \langle \bar{\eta}^T \nabla_{\theta}^2 J(\theta) \bar{\eta} \rangle_q + o(\beta) \right) \right] \]
Thus, $D_{q,\beta}[J(\theta)] \to \left( \frac{2\Lambda_q}{3-q} \nabla_{\theta} J(\theta) \right)$ as $\beta \to 0$. **■**

As a consequence of the Proposition 3.3 for large $M$ and small $\beta$, the form of gradient estimate suggested by (10) is
\[ \nabla_{\theta}[J(\theta)] \approx \frac{1}{\Lambda_q \beta M} \sum_{n=0}^{M-1} \frac{\bar{\eta}(n)J(\theta(n) + \bar{\eta}(n))}{1 - \frac{1-q}{3-q} ||\bar{\eta}(n)||^2} \]

Using an approximation of (10), for large $L$, we can write the above equation as
\[ \nabla_{\theta}[J(\theta)] \approx \frac{1}{\Lambda_q \beta M} \sum_{n=0}^{M-1} \sum_{m=0}^{L-1} \frac{\bar{\eta}(n)h(Y_m)}{1 - \frac{1-q}{3-q} ||\bar{\eta}(n)||^2} \]
where $\{Y_m\}$ is governed by parameter $(\theta(n) + \bar{\eta}(n))$.

However, since $\Lambda_q > 0$, $\Lambda_q$ need not be explicitly determined as estimating $[\Lambda_q \nabla_{\theta} J(\theta)]$ instead of $\nabla_{\theta} J(\theta)$ does not affect the gradient descent approach. As a special case, for $q = 1$, we have $\Lambda_q = 1$ from definition. Hence, we obtain the same form as in (8).
IV. Proposed Algorithms

In this section, we propose a two-timescale algorithm corresponding to the estimate obtained in [12].

The $q$-Gaussian distributed parameters ($\eta$) have been generated in the algorithm using the method proposed in [10]. Let $\{a(n)\}, \{b(n)\}$ be two step-size sequences satisfying

\[
\text{Assumption IV.}\; a(n) = o(b(n)), \quad \sum_{n=0}^{\infty} a(n) = \sum_{n=0}^{\infty} b(n) = \infty,
\]

and

\[
\sum_{n=0}^{\infty} a(n)^2, \sum_{n=0}^{\infty} b(n)^2 < \infty.
\]

For $\theta = (\theta^{(1)}, \ldots, \theta^{(N)})^T \in \mathbb{R}^N$, let $\Gamma(\theta) = (\Gamma^{(1)}(\theta), \ldots, \Gamma^{(N)}(\theta))^T$ represent the projection of $\theta$ onto the set $C$. $\{Z^{(i)}(n), i = 1, \ldots, N\}_{n \in \mathbb{N}}$ are quantities used to estimate $\{A_q\}$ via the recursions below.

The $q$-SF Algorithm
1: Fix $M, L, q$ and $\beta$.
2: Set $Z^{(i)}(0) = 0, i = 1, \ldots, N$.
3: Fix parameter vector $\theta(0) = (\theta^{(1)}(0), \ldots, \theta^{(N)}(0))^T$.
4: for $n = 0$ to $M - 1$ do
5: \hspace{1em} Generate i.i.d. standard $q$-Gaussian distributed random variables $\eta^{(1)}(n), \ldots, \eta^{(N)}(n)$ and set $\eta(n) := (\eta^{(1)}(n), \ldots, \eta^{(N)}(n))^T$.
6: \hspace{1em} for $m = 0$ to $L - 1$ do
7: \hspace{2em} Generate the simulation $Y_{nL+m}$ governed with parameter $(\theta(n) + \beta \eta(n))$.
8: \hspace{2em} for $i = 1$ to $N$ do
9: \hspace{3em} $Z^{(i)}(nL + m + 1) = (1 - b(n))Z^{(i)}(nL + m) + b(n) \left[ \frac{\eta^{(i)}(n)\eta(Y_{nL+m})}{\beta(1 + \frac{1}{\beta}||\eta(n)||^2)} \right]$.
10: \hspace{2em} end for
11: \hspace{1em} end for
12: \hspace{1em} for $i = 1$ to $N$ do
13: \hspace{2em} $\theta^{(i)}(n + 1) = \Gamma(\theta^{(i)}(n) - a(n)Z^{(i)}(nL))$.
14: \hspace{2em} end for
15: Set $\theta(n + 1) := (\theta^{(1)}(n + 1), \ldots, \theta^{(N)}(n + 1))^T$.
16: end for
17: Output $\theta(M) := (\theta^{(1)}(M), \ldots, \theta^{(N)}(M))^T$ as the final parameter vector.

V. Numerical Experiment

A. Numerical Setting

We consider a two-node network of $M/G/1$ queues with feedback. This setting here is somewhat similar to that considered in [3]. Nodes 1 and 2 are fed with independent Poisson external arrival processes with rates $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$, respectively. After departing from Node-1, customers enter Node-2. Once the service at Node-2 is completed, a customer either leaves the system with probability $p = 0.4$ or joins Node-1. The service time processes of the two nodes, $\{S_1^{(1)}(\theta_1)\}_{n \geq 1}$ and $\{S_2^{(2)}(\theta_2)\}_{n \geq 1}$, respectively, are defined as

\[
S_i^{(j)}(\theta_i) = U_i(n)(1 + \frac{||\theta_i(n) - \theta_i||^2}{R_i})
\]

for $i = 1, 2, \quad n \geq 1,$

where $R_1 = 10$ and $R_2 = 20$ are constants. Here, $U_1(n)$ and $U_2(n)$ are independent samples drawn from the uniform distribution on $(0,1)$. The service time of each node depends on the $N_i$-dimensional tunable parameter vector $\theta_i$, whose individual components lie in a certain interval $[\theta_i^{(j)}(\min), \theta_i^{(j)}(\max)]$, $j = 1, 2, \ldots, N_i, \ i = 1, 2$. $\theta_i(n)$ represents the $n$th update of the parameter vector at Node-$i$, and $\theta_i$ represents the target parameter vector.

The cost function is chosen to be the sum of the two queue lengths at any instant. For the cost to be minimum, $S_i^{(j)}(\theta_i)$ should be minimum, and hence, we should have $\theta_i(n) = \theta_i$, $i = 1, 2$. We denote $\theta = (\theta_1^{(1)}, \ldots, \theta_1^{(N_1)}, \theta_2^{(1)}, \ldots, \theta_2^{(N_2)})^T \in \mathbb{R}^N$, and $\theta = (\theta_1^{(1)}, \ldots, \theta_1^{(N_1)}, \theta_2^{(1)}, \ldots, \theta_2^{(N_2)})^T \in \mathbb{R}^N$, where $N = N_1 + N_2$.

In order to compare the performance of various algorithms, we consider the Euclidean distance between $\theta(n)$ and $\theta$, as our performance measure.

For the simulations, we use the following values of parameters:
1) $N_1 = N_2 = 2$,
2) $\theta_i^{(j)}(\min) = 0, \theta_i^{(j)}(\max) = 5$ for $j = 1, \ldots, N_i, \ i = 1, 2$, which implies $C = [0, 5]^N$.
3) $\theta^{(j)}(0) = 5, \theta^{(j)}(1) = 1$ for $j = 1, 2, \ldots, N$,
4) $M = 10000, \; L = 100$,
5) $a(n) = 1/n, b(n) = 1/n^{2/3}$.

B. Simulation Results

The simulations are performed by varying the parameters $q$ and $\beta$. For each case, the results are averaged over 20 independent trials. We compare the performance of our algorithm with the SF algorithm proposed in [3], which uses Gaussian smoothing. Table I presents the performance comparison for different values of $q$ and $\beta$.

![Fig. 1: Convergence behavior of the algorithm for $\beta = 0.25$.](image-url)

It can be observed from the results that for very low values of $\beta$, our algorithm performs better than Gaussian-SF algorithm for $q > 1$. However, for relatively high values of $\beta$, better performance can be obtained with $q < 1$. It can also be seen that for higher values of $q$, the algorithm tends to become unstable for high $\beta$. As per the observations, the best value of $q$ is 0.9, which performs better than Gaussian in 63% cases. In fact, $q = 0.9$ also gives the best performance (least distance) among all values of $q$ in most of the cases (50%).
TABLE I: Performance (mean distance from optimum).

VI. SCHEME OF CONVERGENCE ANALYSIS

Here, we give a sketch of the proof of convergence of the proposed algorithm. We just state the important results. The proofs will be given in a longer version of the paper.

Let $F(l) = \sigma(\beta(i)k), \tilde{\eta}(i)k, \tilde{\eta}(i)k, k \geq l, i = 1, \ldots, N$, $l \geq 1$ denote the $\sigma$-fields generated by the above mentioned quantities, where $\tilde{\eta}(i)k = \tilde{\eta}(i)k(n)$ and $\tilde{\eta}(i)k = \eta(i)k(n)$ for $i = 1, \ldots, N, nL \leq k < (n + 1)L$. Define $\{\tilde{b}(n)\}_{n \geq 0}$ such that $\tilde{b}(n) = b([\frac{n}{L}])$, where $[x]$ is the integer part of $x$. Thus, $\sum_{n=0}^{\infty} \tilde{b}(n) = \infty$, $\sum_{n=0}^{\infty} \tilde{b}(n)^2 < \infty$ and $\tilde{b}(n) = o(b(n))$.

Using above notation, we can rewrite Step 9 of our algorithm, with $p = nL + m$, in the following way:

$$Z^{(i)}(p + 1) = (1 - \tilde{b}(p))Z^{(i)}(p) +$$

$$\tilde{b}(p) \left[ \tilde{\eta}^{(i)}(p)h(Y_p) \right] \left[ \beta \left( 1 - \frac{1-q}{(3-q)} \right) ||\tilde{\eta}(p)||^2 \right]$$

(14)

Define the sequences $\{M^{(i)}(p)\}_{p \geq 1}, i = 1, \ldots, N$ as

$$M^{(i)}(p) = \sum_{k=1}^{p} \tilde{b}(k) \left[ \tilde{\eta}^{(i)}(k)h(Y_k) \right] \left[ \beta \left( 1 - \frac{1-q}{(3-q)} \right) ||\tilde{\eta}(n)||^2 \right]$$

$$E_G \left[ \tilde{\eta}^{(i)}(k)h(Y_k) \left[ \beta \left( 1 - \frac{1-q}{(3-q)} \right) ||\tilde{\eta}(k)||^2 \right] \right]$$

(15)

Lemma 6.1. The sequences $\{M^{(i)}(p), F(p)\}_{p \geq 1}, i = 1, 2, \ldots, N$ are almost surely convergent martingale sequences.

Consider the following ordinary differential equations:

$$\dot{\theta}(t) = 0,$$

(16)

$$\ddot{Z}(t) = D_{q, \beta}[J(\theta)] - Z(t).$$

(17)

Lemma 6.2. The sequence of updates $\{Z(p)\}$ is uniformly bounded with probability 1.

Lemma 6.3. For a given $q < 3$, $q \neq 1$, as $n \to \infty$,

$$\left\| Z(nL) - D_{q, \beta}[J(\theta(n))] \right\| \to 0$$

with probability 1.

The following corollary follows directly from Proposition 3.3 and Lemma 6.3 by triangle inequality.

Corollary 6.4. Given a particular $q < 3$, with probability 1, as $n \to \infty$ and $\beta \to 0$,

$$\left\| Z(nL) - \frac{2\Lambda_n}{(3-q)} \nabla_{\theta} J(\theta) \right\| \to 0$$

Now, finally considering the ODE corresponding to the slowest timescale recursion:

$$\dot{\theta}(t) = \Gamma \left( -\frac{2\Lambda_n}{(3-q)} \nabla_{\theta} J(\theta(t)) \right),$$

(18)

where $\Gamma(f(x)) := \lim_{x \to 0} \left( \frac{\Gamma(x + e^{-f(x)}) - x}{x} \right)$ for any bounded, continuous function $f : \mathbb{R}^N \to \mathbb{R}^N$. The stable points of $\{18\}$ lie in the set $S = \{\theta \in C : \Gamma \left( -\frac{2\Lambda_n}{(3-q)} \nabla_{\theta} J(\theta(t)) \right) = 0\}$. Let for given $\delta > 0$, $S^\delta := \{\theta \in C : ||\theta - \theta_0|| < \delta, \theta_0 \in S\}$.

Theorem 6.5. Under Assumptions [7]–[10] given $q < 3, q \neq 1$ and $\delta > 0$, $\exists \beta_0 > 0$ such that for all $\beta \in (0, \beta_0)$, the sequence obtained using the $q$-SF algorithm converges to a point in $S^\delta$ with probability 1 as $n \to \infty$.

VII. CONCLUSION

$q$-Gaussians exhibit power law behavior, which gives a better control over smoothing of functions as compared to normal distribution. We have extended the Gaussian smoothed functional gradient estimation approach to $q$-Gaussians, and developed an optimization algorithm based on this. We have also presented results illustrating that for some values of $q$, our algorithm performs better than the SF algorithm in [3].

REFERENCES

[1] E. Kiefer and J. Woffowitz, “Stochastic estimation of a maximum regression function,” Annals of Mathematical Statistics, vol. 23, pp. 462–466, 1952.
[2] Y. Y. Kato, V. U. Kulchitsky, “Convergence of a class of random search algorithms,” Automation Remote Control, vol. 8, pp. 1321–1326, 1972.
[3] S. Bhatnagar and V. S. Borkar, “Multiscale chaotic SPSA and smoothed functional algorithms for simulation optimization,” Simulation, vol. 79, no. 9, pp. 568–580, 2003.
[4] R. Y. Rubinstein, Simulation and Monte-Carlo Method, John Wiley, New York, 1981.
[5] C. Tsallis, “Possible generalization of Boltzmann-Gibbs statistics,” Journal of Statistical Physics, vol. 52, no. 1-2, pp. 479–487, 1988.
[6] A. Dukkipati, S. Bhatnagar, and M. N. Murty, “On measure-theoretic aspects of nonextensive entropy functionals and corresponding maximum entropy prescriptions,” Physica A: Statistical Mechanics and its Applications, vol. 384, no. 2, pp. 758–774, 2007.
[7] H. Suyari, “Generalization of shannon-khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy,” IEEE Transactions on Information Theory, vol. 50, pp. 1783–1787, 2004.
[8] D. Prato and C. Tsallis, “Nonextensive foundation of Lévy distributions,” Physical Review E, vol. 60, no. 2, pp. 2398–2401, 1999.
[9] C. Vignati and A. Plastino, “Central limit theorem and deformed exponentials,” Journal of Physics A: Mathematical and Theoretical, vol. 20, no. 45, 2007.
[10] W. J. Thistleton, J. A. Marsh, K. Nelson, and C. Tsallis, “Generalized Box-Muller method for generating $q$-Gaussian random deviates,” IEEE Trans. Information Theory, vol. 53, no. 12, pp. 4805–4810, 2007.