Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

ON THE PERFORMANCE OF THE CROSS-ENTROPY METHOD

Jiaqiao Hu
Ping Hu

Department of Applied Mathematics and Statistics
State University of New York
Stony Brook, NY 11794, U.S.A.

ABSTRACT

We study the recently introduced Cross-Entropy (CE) method for optimization, an iterative random sampling approach that is based on sampling and updating an underlying distribution function over the set of feasible solutions. In particular, we propose a systematic approach to investigate the convergence and asymptotic convergence rate for the CE method through a novel connection with the well-known stochastic approximation procedures. Extensions of the approach to stochastic optimization will also be discussed.

1 INTRODUCTION

We consider the optimization problem

$$x^* \in \arg\max_{x \in \mathbb{X}} H(x), \quad \mathbb{X} \subseteq \Re^n, \tag{1}$$

where $\mathbb{X}$ is the solution space, which can be either continuous or discrete, and H is a bounded, deterministic, real-valued function. Note that in general, the objective function H itself may take the form of an expected value of a sample performance h, $H(x) = E[h(x, \psi)]$, where ψ is a random variable (possibly depending on x) representing the stochastic effects of the system, and only estimates of noisy function h are available. This stochastic problem will be discussed in Section 5.

Numerous algorithms have been proposed for optimization of such systems. In deterministic settings, these include simulated annealing (Kirkpatrick et al. 1983), genetic algorithms (Goldberg 1989), tabu search (Glover 1990), nested partitions (Shi and Ólafsson 2000a), and pure adaptive search (Zabinsky 2003). Approaches that have proven effective for solving problems in stochastic settings are stochastic approximation (SA) (Robbins and Monro 1951; Kushner and Clark 1978; Spall 1992; Kushner and Yin 1997), sample average approximation (Kleywegt et al. 2001), response surface methods (Barton and Meckesheimer 2006), the low-dispersion point sets method (Yakowitz et al. 2000), and many other discrete optimization algorithms based on random search, including stochastic ruler methods (Yan and Mukai 1992, Alrefaei and Andradòttir 2001), the random search methods (Andradòttir 1995, 2006), simulated annealing (Alrefaei and Andradòttir 1999), stochastic nested partitions (Shi and Òlafsson 2000b), and the COMPASS algorithm of Hong and Nelson (2006).

The Cross-Entropy (CE) method (Rubinstein and Kroese 2004), when viewed in an optimization context, is typical of a class of sampling-based algorithms known as model-based methods (Zolchin et al. 2004, Fu et al. 2006). The basic idea of CE is to work with a parameterized probability distribution on the solution space and randomly generate at each iteration a group of candidate solutions. These candidate solutions are then used to update the parameters associated with the distribution so that the future search will be biased toward the region containing high quality solutions. Ever since its introduction, the CE method has attracted a lot of attention from the optimization community due to its many successful applications to hard optimization problems (e.g., Allon et al. 2005, Chepuri and Homem De Mello 2005).

A primary focus of this paper is to combine the robust features of CE encountered in practice with rigorous theoretical convergence guarantees by exploiting its connections with the well-known stochastic approximation method. The underlying idea is to show that the CE method implicitly interprets a deterministic optimization problem (which might be a discrete combinatorial problem) in a continuous-parameter stochastic approximation framework in terms of the parameters of the

probability distribution rather than the original decision variables. In effect, the noise caused by Monte-Carlo random sampling in CE can be naturally transformed into the uncertainty in evaluating the objective function of the equivalent stochastic optimization problem. Note that our analysis differs from existing approaches on the convergence of CE, which have primarily focused on discrete optimization settings (Costa et al. 2007, Rubinstein and Kroese 2004). We restrict our discussion on the CE method, however the ideas contained in this paper can be used to study the capability and potential of some other model-based algorithms such as EDAs (Larrañaga and Lozano 2002) and MRAS (Hu et al. 2007) as well. For a more comprehensive development of the approach and detailed proofs, the reader is referred to Hu et al. (2009).

The rest of the paper is structured as follows. In Section 2, we review the idealized CE method and derive its equivalent gradient recursion form. In Section 3, we show almost sure convergence of the method in its Monte-Carlo version, followed by asymptotic convergence rate results in Section 4. Extensions of the approach to stochastic optimization settings are discussed in Section 5. In Section 6, we illustrative the performance of CE via four examples. Concluding remarks are given in Section 7.

2 THE IDEALIZED CE AND ITS ASSOCIATED GRADIENT ITERATION

The CE method was originally introduced in Rubinstein (1997) for estimating rare event probabilities in stochastic networks. Since then, the method together with its various extensions and adaptations have become useful tools for Monte Carlo simulation and multi-extremal nonlinear optimization. Let $\{f_\theta(\cdot), \theta \in \Theta\}$ be a specified parameterized distribution family, where Θ is the parameter space. The CE method for optimizing (1), in its idealized version, is summarized below:

Algorithm 1: Idealized CE Method

1. *Choose an initial pdf/pmf* $f_{\theta_0}(\cdot)$ *on* $\mathbb{X}$, $\theta_0 \in \Theta$. *Specify parameters* $\rho \in (0,1]$, $\varepsilon > 0$, *and a non-decreasing function* $S(\cdot) : \Re \to \Re^+$. *Set* $k = 0$.
2. *Calculate the* $(1-\rho)$*-quantile* γ_k *of* $H(X)$, *where* X *is a random vector taking values in* $\mathbb{X}$ *with distribution* f_{θ_k}.
3. *Compute the new parameter*

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} E_{\theta_k}\left[S(H(X))I(H(X),\gamma_k)\ln f_\theta(X)\right].$$

4. *If a stopping rule is satisfied, then terminate; otherwise set* $k = k+1$ *and go to Step* 2.

In Algorithm 1, $E_{\theta_k}[\cdot]$ is the expectation taken with respect to f_{θ_k}, the function $S(\cdot)$ is used to account for the cases where the objective function $H(x)$ is negative for some x, and

$$I(y,\gamma) := \begin{cases} 1 & \text{if } y \geq \gamma, \\ (y-\gamma+\varepsilon)/\varepsilon & \text{if } \gamma-\varepsilon < y < \gamma, \\ 0 & \text{if } y \leq \gamma-\varepsilon. \end{cases}$$

Note that for pure technical reasons, we have made a slight modification of the CE method by replacing the original indicator function with a "soft" threshold function $I(\cdot,\cdot)$ in the parameter updating step. Step 2 above is in the same spirit as the selection scheme employed in many population-based approaches such as genetic algorithms. The step is particularly useful in actual sampling-based version of the algorithm (see Section 3), and the motivation is to concentrate the computational effort on the top ρ-percent of the selected "elite" solutions.

An alternative interpretation of the CE method was given in Hu et al. (2007). It has been shown that there exists a sequence of intermediate distributions $\{g_k\}$ called the *reference* distributions implicit in CE, and Step 3 of Algorithm 1 is equivalent to minimizing the Kullback-Leibler (KL) divergence between g_{k+1} and f_θ, i.e.,

$$\theta_{k+1} = \arg\min_{\theta \in \Theta} \mathscr{D}(g_{k+1}, f_\theta) := \int_{\mathbb{X}} \ln \frac{g_{k+1}(x)}{f_\theta(x)} g_{k+1}(dx), \tag{2}$$

where

$$g_{k+1}(x) = \frac{S(H(x))I(H(x),\gamma_k)f_{\theta_k}(x)}{E_{\theta_k}[S(H(X))I(H(X),\gamma_k)]}. \tag{3}$$

We consider a slight generalization of the reference distributions given in (3)

$$g_{k+1}(x) = \alpha_k \frac{S(H(x))I(H(x),\gamma_k)f_{\theta_k}(x)}{E_{\theta_k}[S(H(X))I(H(X),\gamma_k)]} + (1-\alpha_k)f_{\theta_k}(x), \tag{4}$$

where $\alpha_k \in (0,1]$ is a smoothing parameter that ensures that the difference between the reference distribution g_{k+1} and f_{θ_k} is only incremental, so that the new distribution $f_{\theta_{k+1}}$ obtained via (2) does not deviate too much from the current sampling distribution f_{θ_k}.

When the parameterized family belongs to natural exponential families (NEFs), the optimization problem (2) can be solved analytically in closed form for arbitrary g_{k+1}, which makes the approach very convenient to implement in practice.

Definition: *A parameterized family* $\{f_\theta(\cdot), \theta \in \Theta \subseteq \Re^m\}$ *on* $\mathbb{X}$ *is called a natural exponential family if there exist functions* $\Gamma : \Re^n \to \Re^m$ *and* $K : \Re^m \to \Re$ *such that* $f_\theta(x) = \exp\left(\theta^T \Gamma(x) - K(\theta)\right)$*, where* $K(\theta) = \ln \int_{\mathbb{X}} \exp(\theta^T \Gamma(x)) \nu(dx)$*, and* ν *is the Lebesgue/discrete measure on* $\mathbb{X}$*.*

The function $K(\theta)$ plays an important role in the theory of natural exponential families. It is strictly convex with $\nabla K(\theta) = E_\theta[\Gamma(X)]$ and Hessian matrix $\mathrm{Cov}_\theta[\Gamma(X)]$. Therefore, the Jacobian of the mean vector function $m(\theta) := E_\theta[\Gamma(X)]$ is strictly positive definite and thus invertible. From the inverse function theorem, it follows that $m(\theta)$ is also invertible.

When g_{k+1} in (4) is used as the reference distribution in Algorithm 1, the following lemma states the explicit relationship between the two successive mean vectors in the idealized CE method.

Lemma 2.1. *If* f_θ *belongs to NEFs and the new parameters* θ_{k+1} *computed at step 3 of Algorithm 1 is an interior point of* Θ *for all k, then*

$$m(\theta_{k+1}) - m(\theta_k) = \alpha_k \nabla_\theta \ln E_\theta[S(H(X))I(H(X),\gamma_k)]|_{\theta=\theta_k} \quad \text{for all } k = 0,1,2,\ldots.$$

Proof. Follows from Lemma 2 in Hu et al. (2007). □

Lemma 2.1 brings out explicitly the updating direction of the mean vectors at each step of the CE method, which is in the direction of the gradient of the objective function for the maximization problem $\max_{\theta\in\Theta} \ln E_\theta[S(H(X))I(H(X),\gamma_k)]$. Note that the smoothing parameter sequence $\{\alpha_k\}$ turns out to be the gain sequence for the gradient iteration, so that the special case $\alpha_k = 1$ for all k corresponds to constant gain sequences. This implies that the original CE method can be viewed as gradient search methods on the parameter space with constant gain 1. This observation suggests that the CE method is essentially a gradient-based recursion for solving optimization problems on the parameter space (of the distribution model) with smooth differential structures. This key insight, which is new to CE, is crucial to understanding why the algorithm works well for hard optimization problems with little structure.

3 STRONG CONVERGENCE OF MONTE-CARLO VERSION OF CE

Algorithm 1 describes the idealized setting where quantile values and expectations can be evaluated exactly whenever needed. In practice, only a finite number of candidate solutions are generated at each iteration, expected values are replaced with their corresponding sample averages, and true quantiles are estimated by sample quantiles. This results in the following Monte-Carlo version of the CE method.

Algorithm 2: The Monte-Carlo Version of CE

1. *Choose an initial pdf/pmf* $f_{\widehat{\theta}_0}(\cdot)$ *on* $\mathbb{X}$*,* $\widehat{\theta}_0 \in \Theta$*. Specify the parameter* $\rho \in (0,1]$*, a small constant* $\varepsilon > 0$*, the gain sequence* $\{\alpha_k\}$*, and a non-decreasing function* $S(\cdot) : \Re \to \Re^+$*. Set* $k = 0$*.*
2. *Randomly sample* N_k *candidate solutions* $\Lambda_k = \{X_1, \ldots, X_{N_k}\}$ *from the distribution* $f_{\widehat{\theta}_k}$*.*
3. *Calculate the sample* $(1-\rho)$*-quantile* $\widehat{\gamma}_k = H_{(\lceil (1-\rho)N_k \rceil)}$*, where* $\lceil a \rceil$ *is the smallest integer greater than a, and* $H_{(i)}$ *is the ith order statistic of the sequence* $\{H(X_i), i = 1, \ldots, N_k\}$*.*
4. *Compute the new parameter*

$$\widehat{\theta}_{k+1} = \arg\max_{\theta\in\Theta} \mathscr{D}(\widehat{g}_{k+1}, f_\theta),$$

where $\widehat{g}_{k+1}(x) = \alpha_k \frac{S(H(x))I(H(x),\widehat{\gamma}_k)f_{\widehat{\theta}_k}(x)}{\frac{1}{N_k}\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)} + (1-\alpha_k) f_{\widehat{\theta}_k}(x)$, $x \in \Lambda_k$ *is the sample average approximation of (4) based on the sampled solutions in* Λ_k.

5. *If a stopping rule is satisfied, then terminate; otherwise set* $k = k+1$ *and go to Step* 2.

We make the following assumption on the parameter $\widehat{\theta}_{k+1}$ computed at step 4 of the algorithm:

Assumption A1. *The parameter* $\widehat{\theta}_{k+1}$ *computed at step 4 of Algorithm 2 is an interior point of* Θ *for all k.*

Lemma 3.1. *If A1 holds, then the mean vector* $m(\widehat{\theta}_{k+1})$ *in Algorithm 2 satisfies*

$$m(\widehat{\theta}_{k+1}) = \alpha_k \frac{\frac{1}{N_k}\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)\Gamma(x)}{\frac{1}{N_k}\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)} + (1-\alpha_k)\frac{1}{N_k}\sum_{x\in\Lambda_x}\Gamma(x), \quad \forall k. \tag{5}$$

Define terms

$$\begin{aligned}
L(m(\widehat{\theta}_k)) &= \frac{E_{\widehat{\theta}_k}[S(H(X))I(H(X),\gamma_k)\Gamma(X)]}{E_{\widehat{\theta}_k}[S(H(X))I(H(X),\gamma_k)]} - m(\widehat{\theta}_k) = \nabla_\theta \ln E_\theta[S(H(X))I(H(X),\gamma_k)]|_{\theta=\widehat{\theta}_k},\\
b_k(\widehat{\theta}_k) &= \frac{\frac{1}{N_k}\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)\Gamma(x)}{\frac{1}{N_k}\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)} - \frac{E_{\widehat{\theta}_k}[S(H(X))I(H(X),\gamma_k)\Gamma(X)]}{E_{\widehat{\theta}_k}[S(H(X))I(H(X),\gamma_k)]},\\
\xi_k(\widehat{\theta}_k) &= \frac{1-\alpha_k}{\alpha_k}\big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x) - m(\widehat{\theta}_k)\big),
\end{aligned}$$

where with a slight abuse of notation, we use γ_k to represent the true $(1-\rho)$-quantile of $H(X)$ under $f_{\widehat{\theta}_k}$.

We can rewrite (5) in the form of a generalized Robbins-Monro algorithm in terms of the true gradient L, the combined effect of bias and Monte-Carlo random sampling b_k, and an error term due to sample average approximation ξ_k.

$$m(\widehat{\theta}_{k+1}) = m(\widehat{\theta}_k) + \alpha_k\big[L(m(\widehat{\theta}_k)) + b_k(\widehat{\theta}_k) + \xi_k(\widehat{\theta}_k)\big]. \tag{6}$$

Before presenting the main convergence result of this section, we begin with some regularity conditions, some of which are standard in stochastic approximation literatures (cf. e.g., Kushner and Clark 1978; Kushner and Yin 1997).

Assumptions:

A2. *The gain* $\alpha_k > 0\ \forall k,\ \alpha_k \to 0$ *as* $k\to\infty$, *and* $\sum_{k=1}^{\infty}\alpha_k = \infty$. *The sample size* $N_k = O(k^\beta)$, *where* $\beta > 1$.

A3. *For given* $\rho \in (0,1)$ *and a distribution family* $\{f_\theta, \theta\in\Theta\}$, *the* $(1-\rho)$*-quantile of* $\{H(X), X\sim f_\theta(x)\}$ *is unique for all* $\theta\in\Theta$.

A4. *There exists a compact set* Π *so that the level sets* $\{x: H(x)\ge\gamma_k\}\cap\mathbb{X}\subseteq\Pi$ *and* $\{x: H(x)\ge\widehat{\gamma}_k\}\cap\mathbb{X}\subseteq\Pi$ *almost surely for k sufficiently large.*

A5. *Let* $\{\mathscr{F}_k\}$ *be the sequence of increasing* σ*-fields generated by* $\{\Lambda_0,\Lambda_1,\ldots\}$. *There exist constants* $\sigma^2>0$ *and* $\ell>0$ *such that* $\max_{1\le i\le m} Var_{\widehat{\theta}_k}[|\Gamma_i(X)|^{2+\ell}|\mathscr{F}_{k-1}] \le \sigma^2$ *for all* $\widehat{\theta}_k\in\Theta$ *and k, where* $\Gamma_i(X)$ *is the ith component of the vector* $\Gamma(X)$.

A6. (**Kushner and Clark 1978**) *Let* η^* *be a locally asymptotically stable (in the sense of Liapunov) solution to the ODE* $d\eta(t)/dt = L(\eta) = \nabla_\theta \ln E_\theta[S(H(X))I(H(X),\gamma(m^{-1}(\eta)))]|_{\theta=m^{-1}(\eta)}$ *with domain of attraction* $D(\eta^*)$, *where* $\gamma(m^{-1}(\eta))$ *is the true* $(1-\rho)$*-quantile of* $H(X)$ *under* $f_{m^{-1}(\eta)}$. $L(\eta)$ *is continuous in* η. *Moreover, there is a compact set* $A\subseteq D(\eta^*)$ *such that* $m(\widehat{\theta}_k)\in A$ *infinitely often for almost all sample paths generated by Algorithm 2.*

The lemma below establishes the strong consistency of the bias term $b_k(\widehat{\theta}_k)$ as $k\to\infty$. The proof is based on large deviations theory.

Lemma 3.2. *If assumptions A2 – A4 are satisfied, then*

$$b_k(\widehat{\theta}_k) \to 0 \quad as\ k \to \infty \quad w.p.1.$$

The noise term $\xi_k(\widehat{\theta}_k)$ has the following property.

Lemma 3.3. *If A2 and A5 hold, then for any $\varepsilon > 0$,*

$$\lim_{k\to\infty} P\Big(\sup_{l\geq k} \|\sum_{i=k}^{l} \alpha_i \xi_i\| \geq \varepsilon\Big) = 0.$$

The strong convergence of the sequence of parameters $\{\widehat{\theta}_k\}$ generated by Algorithm 2 follows from Lemma 3.2 and Lemma 3.3 and then by applying Lemma 2.2.1 and Theorem 2.3.1 in Kushner and Clark (1978).

Theorem 1. *Let $\eta_k = m(\widehat{\theta}_k)$. Assume A1 − A6 hold, then*

$$\eta_k \to \eta^* \ as\ k \to \infty\ w.p.1.$$

Thus, by the invertibility of $m(\cdot)$, the sequence of sampling distributions $\{f_{\widehat{\theta}_k}\}$ in Algorithm 2 converges to a limiting distribution.

4 ASYMPTOTIC NORMALITY

Let $J_L(\eta)$ be the Jacobian matrix of $L(\eta)$. Note that since $L(\eta)$ is the gradient of some underlying function $F(\eta)$, J_L turns out to be the Hessian of F. Recursion (6) is a gradient-based algorithm for solving the maximization problem $\max_\eta F(\eta)$. Consequently, by Assumption A6 and the strong convergence of the sequence $\{\eta_k\}$ (i.e., the sequence $\{\eta_k\}$ tracks the solution to the ODE $d\eta(t)/dt = L(\eta)$), it is reasonable to expect that η^* is a local maximizer of the objective function $F(\eta)$ in its neighborhood. The following assumption about J_L is natural.

Assumption B1. *$J_L(\eta)$ is continuous and symmetric negative definite in a small neighborhood of η^*.*

In addition, we need the following regularity condition on the distribution function of the objective function. Let f_k^H be the probability density/mass function of $H(X)$ when X is distributed according to $f_{\widehat{\theta}_k}$.

Assumption B2. (i) (*Continuous optimization*) For a given $\rho \in (0,1)$, there exist constants $\bar{\zeta} > 0$ and $\bar{\delta} > 0$ such that $f_k^H(\gamma) > \bar{\zeta},\ \forall \gamma \in (\gamma_k - \bar{\delta}, \gamma_k + \bar{\delta})$ almost surely for k sufficiently large. (ii) (*Discrete finite optimization*) For a given $\rho \in (0,1)$, there exists a constant $\bar{\zeta} > 0$ such that $P_{\widehat{\theta}_k}(H(X) \geq \gamma_k) \geq \rho + \bar{\zeta}$ and $P_{\widehat{\theta}_k}(H(X) > \gamma_k) \leq \rho - \bar{\zeta}$ almost surely for k sufficiently large.

We consider the standard gain sequence of the form $\alpha_k = c/k^\alpha$ for some constants $c > 0$ and $\alpha \in (0,1)$. By using a Taylor expansion of $L(\eta)$ in a small neighborhood of η^* and Assumption A6, it follows from (6) that the difference $\delta_k := \eta_k - \eta^*$ satisfies the recursion

$$\delta_{k+1} = \delta_k + ck^{-\alpha} J_L(\tilde{\eta}_k)\delta_k + ck^{-\alpha} b_k + ck^{-\alpha}\xi_k$$

where $\tilde{\eta}_k$ is on the line segment between η_k and η^*. In the notation of Fabian (1968), the above recursion can be recast in the form

$$\delta_{k+1} = \delta_k - k^{-\alpha}\Upsilon_k\delta_k + k^{-\frac{\alpha+\tau}{2}}\Phi_k V_k + k^{-\alpha-\frac{\tau}{2}} T_k,$$

where $\tau > 0$ is a constant,

$$\Upsilon_k = -cJ_L(\bar{\eta}_k), \qquad \Phi_k = c\mathbb{I}_{m\times m},$$
$$V_k = k^{-\frac{\alpha}{2}+\frac{\tau}{2}}\xi_k(\widehat{\theta}_k), \qquad T_k = ck^{\frac{\tau}{2}}b_k(\widehat{\theta}_k),$$

and $\mathbb{I}_{m\times m}$ denotes an m-by-m identity matrix.

The following is a strengthened version of Lemma 3.2, which indicates that the amplified bias term T_k vanishes to zero asymptotically.

Lemma 4.1. *For any constant $\tau > 0$, if A3, A4, and B2 hold and $\beta > 2\tau$, then*

$$T_k \to 0 \quad as\ k \to \infty \quad w.p.1.$$

Moreover, the amplified noise V_k satisfies the conditions below.

Lemma 4.2. *Let $\alpha_k = c/k^\alpha$, $\alpha \in (0,1)$, and $N_k = O(k^\beta)$, $\beta > 1$. If A1, A3 – A6 hold and $\beta \geq \alpha + \tau$, then there exists a matrix Σ such that $E_{\widehat{\theta}_k}[V_k|\mathscr{F}_{k-1}] = 0$ and $E_{\widehat{\theta}_k}[V_kV_k^T|\mathscr{F}_{k-1}] \to \Sigma$ as $k \to \infty$ w.p.1. Moreover, the sequence $\{V_k\}$ is uniformly square integrable in the sense that*

$$\lim_{k\to\infty} E\left[\mathscr{I}_{\{\|V_k\|^2 \geq rk^\alpha\}}\|V_k\|^2\right] = 0 \ \forall r > 0,$$

where $\mathscr{I}_{\{\cdot\}}$ is the indicator function.

The asymptotic normality result of CE follows directly from lemmas 4.1 and 4.2 above, and Theorem 2.2 in Fabian (1968).

Theorem 2. *Let $\alpha_k = c/k^\alpha$, $\alpha \in (1/2,1)$. If A1, A3–A6, B1, and B2 hold, $\tau \in (1-\alpha,\alpha)$, and $\beta \geq \alpha + \tau$, then*

$$k^{\frac{\tau}{2}}(\eta_k - \eta^*) \xrightarrow{dist} \mathscr{N}(0, QMQ^T),$$

where Q is an orthogonal matrix such that $Q^T(-J_L(\eta^*))Q = \Lambda$ with Λ being a diagonal matrix, and the (i,j)th entry of M is given by $M_{(i,j)} = (Q^T\Sigma Q)_{(i,j)}(\Lambda_{(i,i)} + \Lambda_{(j,j)})^{-1}$,

$$\Sigma := \begin{cases} Cov_{m^{-1}(\eta^*)}(\Gamma(X)) & \text{if } \beta = \alpha + \tau, \\ 0 & \text{if } \beta > \alpha + \tau. \end{cases}$$

5 EXTENSIONS TO SIMULATION OPTIMIZATION

For ease of exposition, we write (6) in the abstract form

$$\eta_{k+1} = \eta_k + \alpha_k\widehat{L}_k(\eta_k),$$

where $\widehat{L}_k$ represents an estimate for the true gradient L based on the sampled N_k solutions at each iteration.

The simulation optimization setting, where $h(x,\psi)$ is obtained in a simulation replication, requires an additional simulation allocation rule $\{M_k\}$, which allocates M_k simulation observations to each of the N_k candidate solutions generated at the kth iteration. Thus, in Algorithm 2, if the true performance at a sampled solution X is replaced by the sample average

$$\bar{H}_k(X) = \frac{1}{M_k}\sum_{j=1}^{M_k} h(X,\psi_j),$$

then an estimator of the true gradient $L(\eta)$ will take the form

$$\bar{L}_k(\eta) = \frac{N_k^{-1}\sum_{x\in\Lambda_k} S(\bar{H}_k(x))I(\bar{H}_k(x),\bar{\gamma}_k)\Gamma(x)}{N_k^{-1}\sum_{x\in\Lambda_k} S(\bar{H}_k(x))I(\bar{H}_k(x),\bar{\gamma}_k)} - \frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x),$$

where $\bar{\gamma}_k$ is the sample $(1-\rho)$-quantile of $\bar{H}_k(X)$ when X is distributed according to $f_{m^{-1}(\eta)}$. Consequently, the associated gradient iteration in the simulation setting can be expressed as

$$\bar{\eta}_{k+1} = \bar{\eta}_k + \alpha_k \widehat{L}_k(\bar{\eta}_k) + \alpha_k\big(\bar{L}_k(\bar{\eta}_k) - \widehat{L}_k(\bar{\eta}_k)\big), \tag{7}$$

where $\{\bar{\eta}_k\}$ is the sequence of mean vectors generated by Algorithm 2 when applied to simulation optimization. A large deviations approach similar to that of Hu et al. (2008) can be used to determine the conditions on the allocation rule $\{M_k\}$ under which $E[\bar{L}_k(\bar{\eta}_k) - \widehat{L}_k(\bar{\eta}_k)|\mathscr{F}_{k-1}] \to 0$, where $\{\mathscr{F}_k\}$ is some appropriate σ-field. This implies that the simulation noise effect $\bar{L}_k(\bar{\eta}_k) - \widehat{L}_k(\bar{\eta}_k)$ can in fact be treated as vanishing bias in gradient estimates. Consequently, the (possibly local) convergence and convergence rate analysis of (7) can be carried out along the same line as described previously.

6 NUMERICAL EXAMPLES

In all previous numerical studies, the CE algorithm was implemented using a smoothed parameter updating procedure (cf. e.g., Rubinstein and Kroese 2004, Hu et al. 2007):

$$\tilde{\theta}_{k+1} := \upsilon\widehat{\theta}'_{k+1} + (1-\upsilon)\tilde{\theta}_k, \quad \text{with } \tilde{\theta}_0 = \widehat{\theta}_0, \tag{8}$$

where $\upsilon \in (0,1]$ is a constant smoothing parameter, and $\widehat{\theta}'_{k+1}$ is the new parameter calculated at Step 4 of Algorithm 2 with $\alpha_k = 1\ \forall k$ in $\widehat{g}_{k+1}$. Such a smoothed parameter updating procedure is primarily used to prevent premature convergence of the algorithm and often works well in practice. Note that our proposed approach differs from this procedure in that the *reference distributions* rather than the *output parameters* are smoothed (cf. equation (4) and Step 4 in Algorithm 2). Because expectation is a linear operator, the solution to the optimization problem at Step 4 of Algorithm 2 can still be obtained in analytical form when f_θ belongs to NEFs. For example, in continuous optimization, if multivariate normal distributions $\mathscr{N}(\widehat{\mu}_k, \widehat{\Sigma}_k)$ are used in CE, the explicit parameter updating equations are given by

$$\widehat{\mu}_{k+1} = \alpha_k \frac{\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)x}{\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)} + (1-\alpha_k)\widehat{\mu}_k \quad \text{and}$$

$$\widehat{\Sigma}_{k+1} = \alpha_k \frac{\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)(x-\widehat{\mu}_{k+1})(x-\widehat{\mu}_{k+1})^T}{\sum_{x\in\Lambda_k} S(H(x))I(H(x),\widehat{\gamma}_k)} + (1-\alpha_k)\big(\widehat{\Sigma}_k + (\widehat{\mu}_k - \widehat{\mu}_{k+1})(\widehat{\mu}_k - \widehat{\mu}_{k+1})^T\big).$$

This is clearly different from the smoothed parameter updating procedure in that there is an extra term $(\widehat{\mu}_k - \widehat{\mu}_{k+1})(\widehat{\mu}_k - \widehat{\mu}_{k+1})^T$ in updating the covariance matrix.

To illustrate the potential influence on practice of such a parameter updating procedure, which was derived using gradient interpretation of CE, we consider four benchmark problems H_1 to H_4, where H_1 is highly multimodal with a huge number of local optima, whereas H_2 to H_4 are badly-scaled functions for which the original CE method has been reported to be not very successful (cf. e.g., Hu et al. 2007).

(1) Trigonometric function ($n = 20$)

$$H_1(x) = -1 + \sum_{i=1}^{n} 8\sin^2\left(7(x_i-0.9)^2\right) + 6\sin^2\left(14(x_i-0.9)^2\right) + (x_i-0.9)^2,$$

where $H_1(x^*) = -1$.

(2) Levy function ($n = 20$)

$$H_2(x) = -10\sin^2(\pi x_1) - (x_n - 1)^2 - 1 - \sum_{i=1}^{n-1} 100 x_i^2 (1 + \sin^2(\pi x_{i+1})),$$

where $H_2(x^*) = -1$.

(3) Powell badly scaled function ($n = 20$)

$$H_3(x) = -\sum_{i=1}^{n/4} \left[(10000 x_{2i+1} x_{2i+2} - 1)^2 \right] - \sum_{i=1}^{n/4} \left[(e^{-x_{2i-1}} + e^{-x_{2i}} - 1.0001)^2 \right] - 1,$$

where $H_3(x^*) = -1$.

(4) Pintér function ($n = 20$)

$$H_4(x) = -1 - \sum_{i=1}^{n} i x_i^2 - \sum_{i=1}^{n} 20 i \sin^2 \left(x_{i-1} \sin x_i - x_i + \sin x_{i+1} \right) - \sum_{i=1}^{n} i \log_{10} \left(1 + i (x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2 \right),$$

where $x_0 = x_n$, $x_{n+1} = x_1$, and $H_4(x^*) = -1$.

In Algorithm 2, we have used the multivariate normal distributions as the parameterized distribution family. The initial mean is uniformly selected from $[-50, 50]^n$ and the covariance matrix is initialized as an n-by-n diagonal matrix with diagonal entries equal to 100. The input parameters are $\rho = 0.1$, $\alpha_k = 2/(k+100)^{0.501}$, and $N_k = \max\{400, k^{1.01}\}$. The performance of Algorithm 2 was also compared with those of CE with smoothed parameter updating procedures, referred to as CE with dynamic parameter updating and CE with constant parameter updating. The difference between these two alternative approaches is that in the former case the smoothing parameter is iteration dependent with $\upsilon_k = \alpha_k$ for all k, whereas in the latter case we set $\upsilon = 0.2$. All other parameters in the two alternative algorithms are taken to be the same as Algorithm 2.

For each test function, we performed 100 independent replication runs of all three algorithms. The performance of the algorithms is shown in Figure 1, which plots the averaged objective function values at the estimated optimal solutions as a function of the number of function evaluations used. The figure clearly indicates convergence of Algorithm 2 on all four test functions as well as its superior performance over both versions of CE with smoothed parameter updating procedures. In particular, for functions H_2, H_3, and H_4, even with smoothed parameter updating, which was primarily designed to prevent premature convergence, the original CE may still frequently stagnate at solutions that are far from optimal. These examples support the proposed gradient interpretation of CE.

7 CONCLUSIONS

We have studied the convergence and convergence rate behavior of the recently proposed Cross-Entropy method by exploiting its connections with the well-known stochastic approximation procedure. Our analysis also provided some guidance on the implementation issues of CE. Preliminary numerical studies indicate that the proposed parameter updating procedure based on gradient interpretation of CE significantly outperforms the original CE with smoothed parameter updating. We hope that the ideas contained in this paper will open a new line of research and eventually lead to a systematic approach to explore the capability and potential of some other sampling-based algorithms such as EDAs and MRAS, while broadening the application scope of stochastic approximation by enabling its use for the analysis of a class of optimization tools for general non-differentiable problems.

ACKNOWLEDGMENTS

This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-07-1-0366, and by the National Science Foundation under Grant No. 0900332.

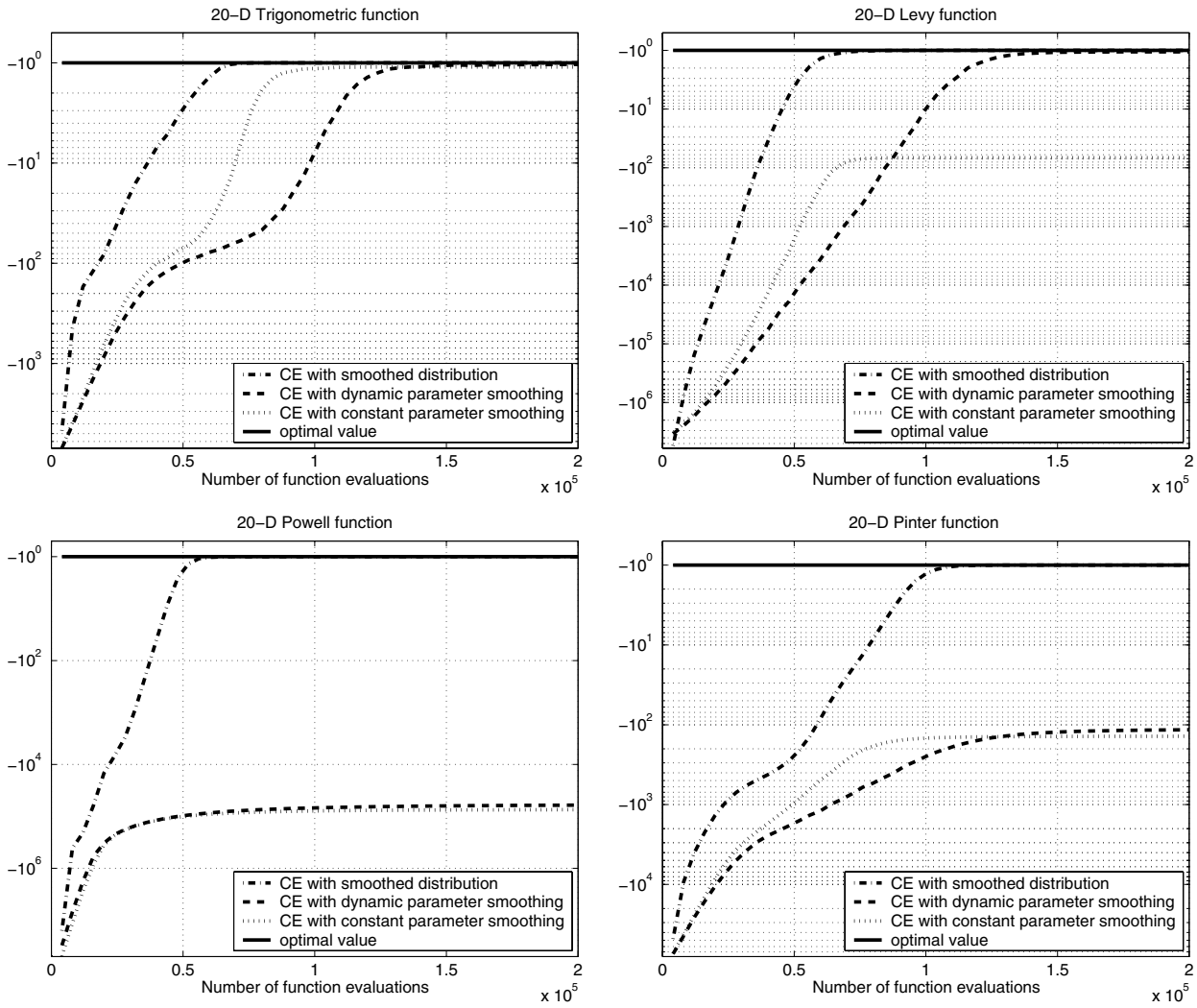

Figure 1: Average Performance of CE with smoothed reference distributions vs. smoothed parameter updating on (a) a 20-D Trigonometric function, (b) a 20-D Levy's function, (c) a 20-D Powell's function, and (d) a 20-D Pinter's function.

REFERENCES

Allon, G., D.P. Kroese, T. Raviv, and R.Y. Rubinstein. 2005. Application of the Cross-Entropy Method to the Buffer Allocation Problem in a Simulation-Based Environment. *Annals of Operations Research* 134:137-151.

Alrefaei, M. H., and S. Andradóttir. 1995. A modification of the stochastic ruler method for discrete stochastic optimization. *European Journal of Operational Research* 133(1):160-182.

Alrefaei, M. H., and S. Andradóttir. 1999. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Management Science* 45(5):748-764.

Andradóttir, S. 1995. A method for discrete stochastic optimization. *Management Science* 41(12):1946-1961.

Andradóttir, S. 2006. An Overview of Simulation Optimization with Random Search. Chapter 20 in *Handbooks in Operations Research and Management Science: Simulation*, ed. S.G. Henderson and B.L. Nelson, Elsevier, 617-632.

Barton, R. R., and M. Meckesheimer. 2006. Metamodel-Based Simulation Optimization. Chapter 18 in *Handbooks in Operations Research and Management Science: Simulation*, ed. S.G. Henderson and B.L. Nelson, Elsevier, 535-574.

Chepuri, K., and T. Homem De Mello. 2005. Solving the Vehicle Routing Problem with Stochastic Demands Using the Cross-Entropy Method. *Annals of Operations Research* 134:153-181.

Costa, A., O.D. Jones, and D. Kroese. 2007. Convergence Properties of the Cross-Entropy Method for Discrete Optimization. *Operations Research Letters* 35(5):573-580.

Fabian, V. 1968. On Asymptotic Normality in Stochastic Approximation. *The Annals of Mathematical Statistics* 39(4):1327-1332.

Fu, M. C., J. Hu, and S. I. Marcus. 2006. Model-based Randomized Methods for Global Optimization. *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems* 355-363.

Glover, F. W. 1990. Tabu Search: A Tutorial. *Interfaces* 20(4):74-94.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Kluwer Academic Publishers, Boston, MA.

Hong, L. J., and B.L. Nelson. 2006. Discrete Optimization via Simulation using COMPASS. *Operations Research* 54(1):115-129.

Hu, J., M. C. Fu, and S. I. Marcus. 2007. A Model Reference Adaptive Search Method for Global Optimization. *Operations Research* 55(3):549-568.

Hu, J., M. C. Fu, and S. I. Marcus. 2008. A Model Reference Adaptive Search Method for Stochastic Global Optimization. *Communications in Information and Systems* 8(3):245-276.

Hu, J., P., Hu., and H. S. Chang. 2009. A Stochastic Approximation Framework for a Class of Randomized Optimization Algorithms. Submitted for publication.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220(4598):671-680.

Kleywegt, A., A. Shapiro, and T. Homem-De-Mello. 2001. The Sample Average Approximation Method for Stochastic Discrete Optimization. *SIAM Journal on Optimization* 12(2):479-502.

Kushner, H. J., and D.S. Clark. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, New York, NY.

Kushner, H. J., and G.G. Yin. 1997. *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, NY.

Larrañaga, P., and J.A. Lozano (Eds.) 2002. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation.* Kluwer Academic Publisher, Boston, MA.

Robbins, H., and S. Monro. 1951. A Stochastic Approximation Method. *Annals of Mathematical Statistics* 22(3):400-407.

Rubinstein, R. Y. 1997. Optimization of Computer Simulation Models with Rare Events. *European Journal of Operational Research* 99(11):89-112.

Rubinstein, R. Y., and D.P. Kroese. 2004. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning.* Springer.

Shi, L., and S. Ólafsson. 2000a. Nested partitions method for global optimization. *Operations Research* 48(3):390-407.

Shi, L., and S. Ólafsson. 2000b. Nested Partitions Method for Stochastic Optimization. *Methodology and Computing in Applied Probability* 2(3):271-291.

Spall, J. C. 1992. Multivariate Stochastic Approximation using Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control* 37(3):332-341.

Yakowitz, S., P. L'Ecuyer, and F. Vázquez-abad. 2000. Global Stochastic Optimization with Low-Dispersion Point Sets. *Operations Research* 48(6):939-950.

Yan, D., and H. Mukai. 1992. Stochastic discrete optimization. *SIAM Journal on Control and Optimization* 30(3):594-612.

Zabinsky, Z. B. 2003. *Stochastic Adaptive Search for Global Optimization*. Kluwer Academic Publishers.

Zlochin, M., M. Birattari, N. Meuleau, and M. Dorigo. 2004. Model-based Search for Combinatorial Optimization: A Critical Survey. *Annals of Operations Research* 131:373-395.

AUTHOR BIOGRAPHIES

JIAQIAO HU is an Assistant Professor in the Department of Applied Mathematics and Statistics at the State University of New York, Stony Brook. He received a B.S. in automation from Shanghai Jiao Tong University, an M.S. in applied mathematics from the University of Maryland, Baltimore County, and a Ph.D. in electrical engineering from the University of Maryland, College Park. His research interests include Markov decision processes, applied probability, and simulation-based optimization. His e-mail address is <jqhu@ams.sunysb.edu>.

PING HU is a Ph.D. candidate in the Department of Applied Mathematics and Statistic at the State University of New York, Stony Brook. He received a B.S. degree in mathematics from Peking University, China in 2006. His research interests are in the areas of optimization and simulation. His e-mail address is <maycher0808@gmail.com>.