The Cross-Entropy Method

Guy Weichenberg

17 September 2003

1 Introduction

This report is a summary of the theory underlying the Cross-Entropy (CE) method, as discussed in the tutorial by de Boer, Kroese, Mannor and Rubinstein [1]. For a more thorough discussion of the method and its applications, please refer to the original tutorial and the references cited in the tutorial.

The CE method, pioneered by Rubinstein in 1997 as an adaptive algorithm for estimating probabilities of rare events, has been broadened as a generic and efficient tool for solving a myriad of NP-hard problems. Beyond its original purpose, the CE method has been employed in deterministic and stochastic combinatorial optimization problems (COPs) and continuous multi-extremal optimization problems.

This report is organized as follows. In Section 2, we discuss the fundamental theory of the CE method and specialize the method to rare-event simulation (RES) and COPs. In Section 3, we consider more sophisticated versions of the CE method, and briefly discuss convergence issues. We conclude the report in Section 4.

2 The Cross-Entropy method

Regardless of the application at hand, the crux of the CE method remains the same. Abstractly, the CE method is an iterative algorithm comprising the following two steps:

1. Generate random data samples using a set of dynamic parameters.
2. Update the parameters governing the random data generation using the data samples themselves with the objective of improving future data samples.

In the remainder of this section, we specialize the above general algorithm to RES and COPs. Refinements to the CE method and convergence are discussed in Section 3.

2.1 Rare-event simulation

The estimation of the probability of rare events often arises in assessing the performance of various engineering systems. If analytical or asymptotic characterizations of the system are unavailable, as is often the case, one must resort to simulation techniques. The simplest and most inefficient simulation technique is *Crude Monte Carlo* (CMC) simulation, where the system is simulated under normal operating parameters for a very long time. A more clever simulation technique is *Importance Sampling* (IS), where the system is simulated under a different (but related) set of parameters which render the occurrence of the rare-event of interest more likely. The difficultly with the IS technique, as we shall see, is obtaining an optimal (or near-optimal) alternative set of parameters under which we would like to simulate the system. The CE method, when used in the context of RES, acts as an *adaptive* IS simulation technique in that it iteratively refines estimates of the optimal set of alternative IS parameters.

We now discuss the theory behind the CE method when applied to RES. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector taking values in some space $\mathcal{X}$, let S be a real-valued function on $\mathcal{X}$, and let $f(\cdot; \mathbf{u})$ be a probability density function on $\mathcal{X}$ parameterized by $\mathbf{u}$. Suppose now that we are interested in the probability of occurrence of a rare event. Specifically, we are interested in the very small probability l that $S(\mathbf{x})$ is greater than or equal to a real number γ under $f(\cdot; \mathbf{u})$:

$$l = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

In a CMC approach to estimating l, we would simply draw a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from $f(\cdot; \mathbf{u})$ and compute:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}}$$

to arrive at an unbiased estimate of l. Clearly, for rare events, most of the terms in the above summation will be zero, thus requiring a very large value of N to obtain a meaningful estimate of l.

An IS approach to estimating l would proceed as follows. We first note that, using an *importance sampling density* g, we can represent l as:

$$l = \int I_{\{S(\mathbf{x}) \geq \gamma\}} \frac{f(\mathbf{x}; \mathbf{u})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \mathbb{E}_g I_{\{S(\mathbf{X}) \geq \gamma\}} \frac{f(\mathbf{X}; \mathbf{u})}{g(\mathbf{X})}.$$

Hence, an unbiased estimator of l is:

$$\widehat{l} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X_i}), \tag{1}$$

where $W(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})/g(\mathbf{x})$ is the *likelihood ratio* (LR), and $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. vectors drawn from g. The optimal importance sampling density is:

$$g^*(\mathbf{x}) = \frac{I_{\{S(\mathbf{x}) \geq \gamma\}} f(\mathbf{x}; \mathbf{u})}{l}$$

in that the resulting estimate (1) has zero variance and only one sample is required. As we mentioned earlier, the problem with the IS approach is obtaining g^*, as this density depends on the quantity l which we are attempting to estimate. In addition, it is often convenient to choose an importance sampling density of the same form as $f(\cdot;\mathbf{u})$. In this case, we are faced with the task of finding the density $f(\cdot;\mathbf{v})$ which is "closest" to g^* in some sense. A convenient measure of "closeness" for this task is the *Kullback-Leibler (K-L) distance*, which is defined as:

$$\mathcal{D}(g,h) = \mathbb{E}_g \ln \frac{g(\mathbf{X})}{h(\mathbf{X})} = \int g(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) d\mathbf{x}.$$

Minimizing the K-L distance between g^* and $f(\cdot;\mathbf{v})$ is equivalent to the following maximization problem:

$$\max_{\mathbf{v}} \ \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X})\geq\gamma\}} \ln f(\mathbf{X};\mathbf{v}),$$

which, using another importance sampling density $f(\cdot;\mathbf{w})$, can be rewritten as:

$$\max_{\mathbf{v}} \ \mathbb{E}_{\mathbf{w}} I_{\{S(\mathbf{X})\geq\gamma\}} W(\mathbf{X};\mathbf{u},\mathbf{w}) \ln f(\mathbf{X};\mathbf{v}). \tag{2}$$

where $W(\mathbf{x};\mathbf{u},\mathbf{w}) = f(\mathbf{x};\mathbf{u})/f(\mathbf{x};\mathbf{w})$ is the LR between $f(\cdot;\mathbf{u})$ and $f(\cdot;\mathbf{w})$ at $\mathbf{x}$. The value $\mathbf{v}^*$ which maximizes (2) can be estimated by solving the stochastic counterpart of (2):

$$\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} \ \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i)\geq\gamma\}} W(\mathbf{X}_i;\mathbf{u},\mathbf{w}) \ln f(\mathbf{X}_i;\mathbf{v}). \tag{3}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_N$ are iid vectors drawn from $f(\cdot;\mathbf{w})$. In instances where the function to be maximized in (3) is convex and differentiable with respect to $\mathbf{v}$, the solution may be analytically obtained by solving the following system of equations[1]:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i)\geq\gamma\}} W(\mathbf{X}_i;\mathbf{u},\mathbf{w}) \nabla \ln f(\mathbf{X}_i;\mathbf{v}) = \mathbf{0}. \tag{4}$$

Note, however, that the approaches implied by (3) and (4) to estimating $\mathbf{v}^*$ only yield meaningful results when not too many of the $I_{\{S(\mathbf{X}_i)\geq\gamma\}}$ terms are zero — that is, when the event of interest is not rare. For rare events, we must therefore resort to an alternative technique, such as an algorithm based on the CE method.

We approach the task of estimating $\mathbf{v}^*$ by following an iterative algorithm in which the two parameters $\mathbf{v}_t$ and γ_t associated with each iteration progressively approach $\mathbf{v}^*$ and γ, respectively.

[1] The convenience of the K-L distance measure is now apparent, since it lends to analytic solutions. On the other hand, alternative metrics, such as variance minimization, generally involve complicated numerical optimization techniques [1].

Specifically, we begin by assigning $\mathbf{v}_0 = \mathbf{u}$, generating N samples $\mathbf{X}_1, \dots, \mathbf{X}_N$ from the density $f(\cdot; \mathbf{v}_0)$, and assigning γ_1 to the value which makes the probability $l_1 = \mathbb{E}_{\mathbf{v}_0} I_{\{S(\mathbf{X}) \geq \gamma_1\}}$ approximately equal to ρ, a specified constant. We complete the first iteration by letting $\mathbf{v}_1$ be equal to the optimal vector parameter for estimating l_1. Continuing in this way, γ_t will increase until eventually reaching its destination value of γ. Simultaneously, $\mathbf{v}_t$ continually changes to reflect an estimate of the optimal vector parameter for estimating l_t, and it thus reaches its destination value of $\mathbf{v}^*$ at the termination of the algorithm. The following are the details of the algorithm.

Algorithm 1 (Probabilistic CE algorithm for RES [1])

1. *Define* $\widehat{\mathbf{v}}_o = \mathbf{u}$. *Set* $t = 1$.

2. *Generate samples* $\mathbf{X}_1, \dots, \mathbf{X}_N$ *from the density* $f(\cdot; \mathbf{v}_{t-1})$ *and compute the sample* $(1-\rho)$*-quantile of* $\widehat{\gamma}_t$ *of the performances — that is,* $\widehat{\gamma}_t = S_{(\lceil (1-\rho) \rceil)}$ *— provided* $\widehat{\gamma}_t$ *is less than* γ. *Otherwise, set* $\widehat{\gamma}_t = \gamma$.

3. *Use the same sample* $\mathbf{X}_1, \dots, \mathbf{X}_N$ *to solve the stochastic program:*

$$\widehat{\mathbf{v}}_t = \mathrm{argmax}_{\mathbf{v}} \ \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) \ln f(\mathbf{X}_i; \mathbf{v}).$$

4. *If* $\widehat{\gamma}_t < \gamma$, *set* $t = t + 1$ *and reiterate from Step 2. Otherwise, proceed with Step 5.*

5. *Estimate the rare-event probability* l *using the LR estimate:*

$$\widehat{l} = \frac{1}{N_1} \sum_{i=1}^{N_1} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_T),$$

where T *denotes the final number of iterations.*

The above algorithm is *probabilistic* in the sense that $\widehat{\gamma}_t$, $\widehat{\mathbf{v}}_t$ and $\widehat{l}$ are estimated quantities which depend probabilistically on the samples $\mathbf{X}_1, \dots, \mathbf{X}_N$. It is possible to define a *deterministic* version of the above algorithm by replacing sample means and sample quantiles by expectations and quantiles.

Algorithm 2 (Deterministic CE algorithm for RES [1])

1. *Define* $\mathbf{v}_o = \mathbf{u}$. *Set* $t = 1$.

2. *Calculate* γ_t *as:*

$$\gamma_t = \max_s \ \mathbb{P}_{\mathbf{v_{t-1}}}(S(\mathbf{X}) \geq s) \geq \rho,$$

provided this is less than γ; *otherwise, set* $\gamma_t = \gamma$.

3. *Calculate* $\mathbf{v}_t$ *as:*

$$\mathrm{argmax}_{\mathbf{v}} \ \mathbb{E}_{\mathbf{v}_{t-1}} I_{\{S(\mathbf{X}) \geq \gamma_t\}} W(\mathbf{X}; \mathbf{u}, \mathbf{v}_{t-1}) \ln f(\mathbf{X}; \mathbf{v}).$$

4. *If $\gamma_t = \gamma$, then stop; otherwise, set $t = t + 1$ and reiterate from Step 2.*

5. *Calculate the rare-event probability l as:*

$$l = \mathbb{E}_{\mathbf{v}_T} I_{\{S(\mathbf{X}) \geq \gamma\}} W(\mathbf{X}; \mathbf{u}, \mathbf{v}_T),$$

where T denotes the final number of iterations.

2.2 Combinatorial optimization

The CE method for COPs essentially involves recasting a deterministic COP into a probabilistic framework where the RES technique of Section 2.1 may be applied.

Let us assume that in our COP $\mathcal{X}$ denotes a finite set of states, and S denotes a real-valued performance function that we wish to maximize over $\mathcal{X}$. Let us denote the maximum value of S by γ^* and the state at which this maximum occurs by x^*. We recast this deterministic COP into a probabilistic framework as follows. Define $f(\cdot; \mathbf{v})$ to be a probability mass function on $\mathcal{X}$, parameterized by a real-valued parameter vector $\mathbf{v}$. For a certain parameter vector $\mathbf{u}$ and threshold γ we define $l(\gamma)$ as the probability that the performance function S exceeds γ:

$$l(\gamma) = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \sum_{\mathbf{x}} I_{\{S(\mathbf{x}) \geq \gamma\}} f(\mathbf{x}; \mathbf{u}) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

Now, in a typical COP, $|\mathcal{X}|$ is very large and $l(\gamma^*) = f(\mathbf{x}^*; \mathbf{u}) = 1/|\mathcal{X}|$ is consequently very small, rendering $l(\gamma^*)$ a *rare event*. Thus, we are now in a position to employ the algorithms of Section 2.1 to solve this problem. The following is the resulting procedure.

Algorithm 3 (Probabilistic CE algorithm for COPs [1])

1. *Define $\widehat{\mathbf{v}}_o = \mathbf{u}$. Set $t = 1$.*

2. *Generate samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from the density $f(\cdot; \mathbf{v}_{t-1})$ and compute the sample $(1-\rho)$-quantile of $\widehat{\gamma}_t$ of the performances — that is, $\widehat{\gamma}_t = S_{(\lceil(1-\rho)\rceil)}$ — provided $\widehat{\gamma}_t$ is less than γ. Otherwise, set $\widehat{\gamma}_t = \gamma$.*

3. *Use the same sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to solve the stochastic program:*

$$\widehat{\mathbf{v}}_t = \text{argmax}_{\mathbf{v}} \, \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} \ln f(\mathbf{X}_i; \mathbf{v}). \tag{5}$$

4. *If for some $t \geq d$:*

$$\widehat{\gamma}_t = \widehat{\gamma}_{t-1} = \cdots = \widehat{\gamma}_{t-d},$$

then stop; otherwise, set $t = t + 1$ and reiterate from Step 2.

Several remarks are now in order:

- A significant difference between Algorithm 1 and Algorithm 3 is the role of the initial parameter vector $\mathbf{u}$. In Algorithm 1, $\mathbf{u}$ uniquely characterizes the system under consideration, whereas in Algorithm 3, $\mathbf{u}$ is arbitrarily chosen to initialize the algorithm. The LR term W is thus implicitly set to unity in Algorithm 3 because it is meaningless.

- Rather than updating $\widehat{\mathbf{v}}_t$ through (5), it is often beneficial to use the following smoothing function:

  $$\widehat{\mathbf{v}}_t = \alpha \widehat{\mathbf{w}}_t + (1 - \alpha)\widehat{\mathbf{v}}_{t-1},$$

  where $\widehat{\mathbf{w}}_t$ is the vector derived via (5). Employing such a smoothing function prevents harmful recurrences of zeros and ones in the parameter vectors.

- As in the case of RES, a deterministic version of the above algorithm can be defined.

3 Enhancements and convergence

In this section, we consider two modifications to the basic CE method presented in Section 2. The first modification involves the use of alternative performance functions to the one inherently defined by the original problem. The second modification allows the CE method to be fully self-tuning.

3.1 Alternative performance functions

In our application of the CE method to RES and COPs in Section 2, we were required to estimate the rare-event probability:

$$l(\gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

We could modify the original CE algorithm by replacing the indicator function in the above expression with an alternative function $\varphi(s; \gamma)$ which, for a maximization problem, is increasing in s for each $\gamma \geq 0$ and decreasing in γ for each $s \geq 0$. For example, for functions $\varphi(s; \gamma)$ of the form:

$$\varphi(s; \gamma) = \psi(s) I_{\{s \geq \gamma\}},$$

the updating of $\widehat{\gamma}_t$ would remain unchanged and the updating of $\widehat{\mathbf{v}}_t$ would be given by:

$$\text{argmax}_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} \psi(S(\mathbf{X}_i) \ln f(\mathbf{X}_i; \mathbf{v}).$$

Empirical evidence shows that $\varphi = s$ speeds up the algorithm, while high-power polynomials should be avoided as they lead more easily to local minima.

3.2 Fully Adaptive CE algorithm

In this subsection, we discuss the Fully Adaptive CE (FACE) algorithm which is a fully automated version of Algorithm 3.

In FACE, parameters are updated using the best performing N^e samples, where N^e is a constant specified ahead of time. We define ρ_t as N^e/N_t. For each iteration, N_t is obtained using the following procedure. At each iteration, we would like:

$$S_{t,(N_t)} \geq S_{t-1,(N_{t-1})}, \tag{6}$$

which we attempt to satisfy by allowing N_t to take on a wide range of values from N^{min} to N^{max}. We begin by setting $N_t = N^{min}$ and incrementing N_t until (6) is satisfied. If $N_t = N^{max}$ and (6) is violated, then we update $\widehat{\gamma}_t$ and $\widehat{\mathbf{v}}_t$ as we did in Algorithm 3. If, however, samples of size N^{max} were generated for several iterations while violating (6), then FACE bails out declaring failure. At each iteration, we would also like:

$$\widehat{\gamma}_t > \widehat{\gamma}_{t-1}. \tag{7}$$

Collectively, (6) and (7) imply improvement in both the best and worst of the N^e samples in each iteration. Empirically, it was found that FACE speeds up convergence up to two times faster than the original CE algorithm with static N and ρ. The full FACE algorithm is detailed below.

Algorithm 4 (FACE Algorithm [1])

1. *At each iteration t, $t = 1, 2, \ldots$ take a sample of size N_t, ($N^{min} \leq N_t \leq N^{max}$) from $f(\cdot; \widehat{\mathbf{v}}_{t-1})$. Denote the corresponding ordered sample performances by $S_{t,(1)} \leq \cdots \leq S_{t,(N)}$.*

2. *If (6) and (7) hold, proceed with the updating Steps 2 and 3 of Algorithm 3 using the N_t samples in Step 1.*

3. *If (6) and (7) are violated, check whether:*

$$S_{t,(N_t)} = \cdots = S_{t,(N_t-d)} = S_{t-1,(N_{t-1})} \tag{8}$$

holds, for some integer d. If (8) holds, stop and deliver $S_{t,(N_t)}$ as an estimate of the optimal solution. Call such $S_{t,(N_t)}$ a reliable estimate of the optimal solution.

4. *If $N_t = N^{max}$ and (6), (7) and (8) are still violated, proceed with Steps 2 and 3 of Algorithm 3 using the N^{max} samples mentioned in Step 1 and go to Step 3.*

5. *If $N_t = N^{max}$ for several iterations in turn and (6), (7) and (8) are violated, stop and announce that FACE identified a "hard" problem. Call $S_{t,(N_t)}$ an unreliable estimate of the optimal solution.*

3.3 Convergence

The convergence of the CE method to an estimate of the optimal CE parameter in a finite number of iterations with a finite sample size was shown in [2] under certain assumptions. Specifically, it was assumed that the probability being estimated in the CE method did not vanish in a neighborhood of the optimal parameter $\mathbf{v}^*$. Furthermore, the algorithms for which convergence was proven were modified versions of Algorithms 1 and 2 in which ρ is adaptively decreased, and the sample size N is adaptively increased. In the same work, the authors showed that for sufficiently small ρ and sufficiently large N, the convergence to the optimal CE parameter is exponential in the sample size N.

4 Conclusion

In this report, we summarized the fundamental theoretical aspects of the CE method pioneered by Rubinstein. The basic method was specialized to simulation of rare events and combinatorial optimization. While these are the most popular applications of the CE method, the method has found use in other areas, such as machine learning and vector quantization and clustering [1].

While the CE method has been widely deployed to efficiently solve a wide range of difficult problems, such as the Max-Cut and Travelling Salesman problems, there still remains a great deal to be understood about the dynamics and convergence properties of the method. A better understanding of the method's dynamics which contribute to its resilience to local minima is desirable, in addition to bounds for rates of convergence of the method.

References

[1] P. T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, *A tutorial on the Cross-Entropy method*, (2003), Internet: http://wwwhome.cs.utwente.nl/

[2] T. Homem de Mello and R. Y. Rubinstein, *Rare event estimation for static models via Cross-Entropy and Importance Sampling*, (2003), submitted for publication.