BCMA-ES: A Bayesian approach to CMA-ES

Eric Benhamou
LAMSADE and A.I Square Connect
Paris Dauphine, France
eric.benhamou@aisquareconnect.com

David Saltiel
LISIC and A.I Square Connect
ULCO-LISIC, France
david.saltiel@aisquareconnect.com

Sebastien Verel
ULCO-LISIC
France
verel@univ-littoral.fr

Fabien Teytaud
ULCO-LISIC
France
fabien.teytaud@univ-littoral.fr

ABSTRACT

This paper introduces a novel theoretically sound approach for the celebrated CMA-ES algorithm. Assuming the parameters of the multi variate normal distribution for the minimum follow a conjugate prior distribution, we derive their optimal update at each iteration step. Not only provides this Bayesian framework a justification for the update of the CMA-ES algorithm but it also gives two new versions of CMA-ES either assuming normal-Wishart or normal-Inverse Wishart priors, depending whether we parametrize the likelihood by its covariance or precision matrix. We support our theoretical findings by numerical experiments that show fast convergence of these modified versions of CMA-ES.

CCS CONCEPTS

• **Mathematics of computing** → *Probability and statistics*;

KEYWORDS

CMA-ES, Bayesian, conjugate prior, normal-inverse-Wishart

ACM Reference Format:
Eric Benhamou, David Saltiel, Sebastien Verel, and Fabien Teytaud. 2019. BCMA-ES: A Bayesian approach to CMA-ES. In *Proceedings of A.I Square Working Paper*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The covariance matrix adaptation evolution strategy (CMA-ES) [12] is arguably one of the most powerful real-valued derivative-free optimization algorithms, finding many applications in machine learning. It is a state-of-the-art optimizer for continuous black-box functions as shown by the various benchmarks of the COmparing Continuous Optimisers INRIA platform for ill-posed functions. It has led to a large number of papers and articles and we refer the interested reader to [1, 2, 4–6, 10–12, 15, 21] and [25] to cite a few.

It has has been successfully applied in many unbiased performance comparisons and numerous real-world applications. In particular, in machine learning, it has been used for direct policy search in reinforcement learning and hyper-parameter tuning in supervised learning ([13, 14, 16]), and references therein, as well as hyperparameter optimization of deep neural networks [18]

In a nutshell, the (μ / λ) CMA-ES is an iterative black box optimization algorithm, that, in each of its iterations, samples λ candidate solutions from a multivariate normal distribution, evaluates these solutions (sequentially or in parallel) retains μ candidates and adjusts the sampling distribution used for the next iteration to give higher probability to good samples. Each iteration can be individually seen as taking an initial guess or *prior* for the multi variate parameters, namely the mean and the covariance, and after making an experiment by evaluating these sample points with the fit function updating the initial parameters accordingly.

Historically, the CMA-ES has been developed heuristically, mainly by conducting experimental research and validating intuitions empirically. Research was done without much focus on theoretical foundations because of the apparent complexity of this algorithm. It was only recently that [3, 8] and [21] made a breakthrough and provided a theoretical justification of CMA-ES updates thanks to information geometry. They proved that CMA-ES was performing a natural gradient descent in the Fisher information metric. These works provided nice explanation for the reasons of the performance of the CMA-ES because of strong invariance properties, good search directions, etc

There is however another way of explanation that has been so far ignored and could also bring nice insights about CMA-ES. It is Bayesian statistics theory. At the light of Bayesian statistics, CMA-ES can be seen as an iterative prior posterior update. But there is some real complexity due to tricky updates that may explain why this has always been ignored. First of all, in a regular Bayesian approach, all sample points are taken. This is not the case in the (μ/λ) CMA-ES as out of the λ generated paths, only the μ best are selected. The updating weights are also constant which is not consistent with Bayesian updates. But more importantly, the covariance matrix update is the core of the problem. It appeals important remarks. The update is done according to a weighted combination of a rank one matrix referred to $p_C p_C^T$ with parameter c_1 and a rank $min(\mu, n)$ matrix with parameter c_μ, whose details are given for instance in [9]. The two updates for the covariance matrix makes the Bayesian update interpretation challenging as these updates are done according to two paths: the isotropic and anisotropic evolution path. All this may explain why a Bayesian approach for interpreting and revisiting the CMA-ES algorithm have seemed a daunting task and not tackled before.

This is precisely the objective of this paper. Section 2 recalls various Bayesian concepts of updates for prior and posterior to highlight the analogy of an iterative Bayesian update. Section 3 presents in greater details the Bayesian approach of CMA-ES, with the corresponding family of derived algorithms, emphasizing the various design choices that can conduct to multiple algorithms. Section 4 provides numerical experiments and shows that Bayesian

A.I Square Working Paper, March 2019, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

adapted CMA-ES algorithms perform well on convex and non convex functions. We finally conclude about some possible extensions and further experiments.

However, the analogy with a successive Bayesian prior posterior update has been so far missing in the landscape of CMA-ES for multiple reasons. First of all, from a cultural point of view, the evolutionary and Bayesian community have always been quite different and not overlapping. Secondly, the CMA-ES was never formulated in terms of a prior and posterior update making its connection with Bayesian world non obvious. Thirdly, when looking in details at the parameters updates, the weighted combination between the global and local search makes the interpretation of a Bayesian posterior update non trivial. We will explain in this paper that the global search needs to be addressed with a special dilatation techniques that is not common in Bayesian wold.

2 FRAMEWORK

CMA-ES computes at each step an update of the mean and covariance of the distribution of the minimum. From a very general point of view this can be interpreted as a prior posterior update in Bayesian statistics.

2.1 Bayesian vs Frequentist probability theory

The justification of the Bayesian approach is discussed in [23]. In Bayesian probability theory, we assume a distribution on unknown parameters of a statistical model that can be characterized as a probabilization of uncertainty. This procedure leads to an axiomatic reduction from the notion of unknown to the notion of randomness but with probability. We do not know the value of the parameters for sure but we know specific values that these parameters can take with higher probabilities. This creates a prior distribution that is updated as we make some experiments as shown in [7, 19, 23]. In the Bayesian view, a probability is assigned to a hypothesis, whereas under frequentist inference, a hypothesis is typically tested without being assigned a probability. There are even some nice theoretical justification for it as presented in [17].

Definition 2.1. *(Infinite exchangeability). We say that $(x_1, x_2, \ldots)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(x_1, x_2, ..., x_n)$ is invariant to permutation of the indices. That is, for any permutation π,*

$$p(x_1, x_2, ..., x_n) = p(x_{\pi 1}, x_{\pi 2}, ..., x_{\pi n})$$

Equipped with this definition, the De Finetti's theorem as provided below states that exchangeable observations are conditionally independent relative to some latent variable.

Theorem 2.1. *(De Finetti, 1930s). A sequence of random variables $(x_1, x_2, \ldots)$ is infinitely exchangeable iff, for all n,*

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta),$$

for some measure P on θ.

This representation theorem 2.1 justifies the use of priors on parameters since for exchangeable data, there must exist a parameter θ, a likelihood $p(x|\theta)$ and a distribution π on θ. A proof of De Finetti theorem is for instance given in [24] (section 1.5).

Remark 2.1. *The De Finetti is trivially satisfied in case of i.i.d. sampling as the sequence is clearly exchangeable and that the joint probability is clearly given by the product of all the marginal distributions. However, the De Finetti goes far beyond as it proves that the infinite exchangeability is enough to prove that the joint distribution is the product of some marginal distribution for a given parameter θ. The sequence may not be independent neither identically distributed, which is a much stronger result!*

2.2 Conjugate priors

In Bayesian statistical inference, the probability distribution that expresses one's (subjective) beliefs about the distribution parameters before any evidence is taken into account is called *the prior* probability distribution, often simply called the prior. In CMA-ES, it is the distribution of the mean and covariance. We can then update our prior distribution with the data using Bayes' theorem to obtain a posterior distribution. The *posterior* distribution is a probability distribution that represents your updated beliefs about the parameters after having seen the data. The Bayes' theorem tells us *the fundamental rule* of Bayesian statistics, that is

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

The proportional sign indicates that one should compute the distribution up to a renormalization constant that enforces the distribution sums to one. This rule is simply a direct consequence of Baye's theorem. Mathematically, let us say that for a random variable X, its distribution p depends on a parameter θ that can be multi-dimensional. To emphasize the dependency of the distribution on the parameters, let us write this distribution as $p(x|\theta)$ and let us assume we have access to a prior distribution $\pi(\theta)$. Then the joint distribution of (θ, x) writes simply as

$$\phi(\theta, x) = p(x|\theta)\pi(\theta)$$

The marginal distribution of x is trivially given by marginalizing the joint distribution by θ as follows:

$$m(x) = \int \phi(\theta, x) d\theta = \int p(x|\theta)\pi(\theta) d\theta$$

The posterior of θ is obtained by Bayes's formula as

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta) d\theta} \propto p(x|\theta)\pi(\theta)$$

Computing a posterior is tricky and does not bring much value in general. A key concept in Bayesian statistics is conjugate priors that makes the computation really easy and is described at length below.

Definition 2.2. *A prior distribution $\pi(\theta)$ is said to be a conjugate prior if the posterior distribution*

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \tag{1}$$

remains in the same distribution family as the prior.

At this stage, it is relevant to introduce exponential family distributions as this higher level of abstraction that encompasses the multi variate normal trivially solves the issue of founding conjugate priors. This will be very helpful for inferring conjugate priors for the multi variate Gaussian used in CMA-ES.

Definition 2.3. *A distribution is said to belong to the exponential family if it can be written (in its canonical form) as:*

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp(\eta \cdot T(\mathbf{x}) - A(\eta)), \tag{2}$$

where η is the natural parameter, $T(\mathbf{x})$ is the sufficient statistic, $A(\eta)$ is log-partition function and $h(\mathbf{x})$ is the base measure. η and $T(\mathbf{x})$ may be vector-valued. Here $a \cdot b$ denotes the inner product of a and b. The log-partition function is defined by the integral

$$A(\eta) \triangleq \log \int_{\mathcal{X}} h(\mathbf{x}) \exp(\eta \cdot T(\mathbf{x})) \, \mathrm{d}x. \tag{3}$$

Also, $\eta \in \Omega = \{\eta \in \mathbb{R}^m | A(\theta) < +\infty\}$ where Ω is the natural parameter space. Moreover, Ω is a convex set and $A(\cdot)$ is a convex function on Ω.

Remark 2.2. *Not surprisingly, the normal distribution $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ belongs to the exponential family but with a different parametrisation. Its exponential family form is given by:*

$$\eta(\mu, \Sigma) = \begin{bmatrix} \Sigma^{-1}\mu \\ \text{vec}(\Sigma^{-1}) \end{bmatrix}, \qquad T(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \text{vec}(-\frac{1}{2}\mathbf{x}\mathbf{x}^{\mathrm{T}}) \end{bmatrix}, \tag{4a}$$

$$h(\mathbf{x}) = (2\pi)^{-\frac{d}{2}}, \qquad A(\eta(\mu, \Sigma)) = \frac{1}{2}\mu^{\mathrm{T}}\Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma|. \tag{4b}$$

where in equations (4a)*, the notation* $\text{vec}(\cdot)$ *means we have vectorized the matrix, stacking each column on top of each other and hence can equivalently write for a and b, two matrices, the trace result* $\text{Tr}(a^{\mathrm{T}}b)$ *as the scalar product of their vectorization* $\text{vec}(a) \cdot \text{vec}(b)$*. We can remark the canonical parameters are very different from traditional (also called moment) parameters. We can notice that changing slightly the sufficient statistic $T(x)$ leads to change the corresponding canonical parameters η.*

For an exponential family distribution, it is particularly easy to form conjugate prior.

Proposition 2.2. *If the observations have a density of the exponential family form $p(x|\theta, \lambda) = h(x)\exp\left(\eta(\theta, \lambda)^T T(x) - nA(\eta(\theta, \lambda))\right)$, with λ a set of hyper-parameters, then the prior with likelihood defined by $\pi(\theta) \propto \exp(\mu_1 \cdot \eta(\theta, \lambda) - \mu_0 A(\eta(\theta, \lambda)))$ with $\mu \triangleq (\mu_0, \mu_1)$ is a conjugate prior.*

The proof is given in appendix subsection 6.1. As we can vary the parameterisation of the likelihood, we can obtain multiple conjugate priors. Because of the conjugacy, if the initial parameters of the multi variate Gaussian follows the prior, the posterior is the true distribution given the information $\mathcal{X}$ and stay in the same family making the update of the parameters really easy. Said differently, with conjugate prior, we make the optimal update. And it is enlightening to see that as we get some information about the likelihood, our posterior distribution becomes more peak as shown in figure1.

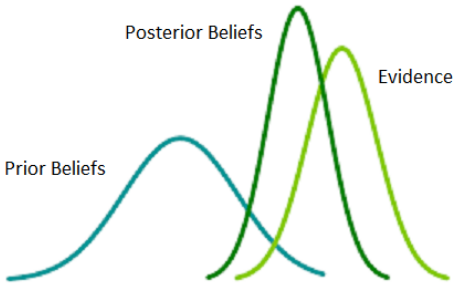

Figure 1: As we get more and more information using the likelihood, the posterior becomes more peak.

2.3 Optimal updates for NIW

The two natural conjugate priors for the Multi variate normal that updates both the mean and the covariance are the normal-inverse-Wishart if we want to update the mean and covariance of the Multi variate normal or the normal-Wishart if we are interested in updating the mean and the precision matrix (which is the inverse of the covariance matrix). In this paper, we will stick to the normal-inverse-Wishart to keep things simple. The Normal-inverse-Wishart distribution is parametrized by $\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi}, \nu$ and its distribution is given by

$$f(\boldsymbol{\mu}, \Sigma | \boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi}, \nu) = \mathcal{N}\left(\boldsymbol{\mu} \middle| \boldsymbol{\mu}_0, \frac{1}{\lambda}\Sigma\right) \mathcal{W}^{-1}(\Sigma | \boldsymbol{\Psi}, \nu)$$

where $\mathcal{W}^{-1}$ denotes the inverse Wishart distribution. The key theoretical guarantee of the BCMA-ES is to update the mean and covariance of our CMA-ES optimally as follows.

Proposition 2.3. *If our sampling density follows a d dimensional multivariate normal distribution $\sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ with unknown mean μ and covariance Σ and if its parameters are distributed according to a Normal-Inverse-Wishart $(\mu, \Sigma) \sim \text{NIW}(\mu_0, \kappa_0, \upsilon_0, \psi)$ and if we observe $\mathcal{X} = (x_1, .., x_n)$ samples, then the posterior is also a Normal-Inverse-Wishart with different parameters $\text{NIW}(\mu_0^\star, \kappa_0^\star, \upsilon_0^\star, \psi^\star)$ given by*

$$\begin{aligned} \mu_0^\star &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0 + n}, \\ \kappa_0^\star &= \kappa_0 + n, \\ \upsilon_0^\star &= \upsilon_0 + n \\ \psi^\star &= \psi + \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T \end{aligned} \tag{5}$$

with $\overline{x}$ the sample mean.

Remark 2.3. *This proposition is the cornerstone of the BCMA-ES. It provides the theoretical guarantee that the updates of the parameters in the algorithm are accurate and optimal under the assumption of the prior. In particular, this implies that any other formula for the update of the mean and variance and in particular the ones used in the mainstream CMA-ES assumes a different prior.*

Proof. A complete proof is given in the appendix section 6.2. □

3 BAYESIAN CMA-ES

3.1 Main assumptions

Our main assumptions are the followings :

- the parameters of the multi-variate Gaussian follow a conjugate prior distribution.
- the minimum of our objective function f follows a multivariate normal law.

3.2 Simulating the minimum

One of the main challenge is to simulate the likelihood to infer the posterior. The key question is really to use the additional information of the function value f for candidate points. At step t in our algorithm, we suppose multi variate Gaussian parameters μ and Σ follow a normal inverse Wishart denoted by $NIW(\mu_t, \kappa_t, \upsilon_t, \psi_t)$.

In full generality, we need to do a Monte Carlo of Monte Carlo as the parameters of our multi variate normal are themselves stochastic. However, we can simplify the problem and take their mean values. It is very effective in terms of computation and reduces Monte Carlo noise. For the normal inverse Wishart distribution, there exist closed form for these mean values given by:

$$\mathbb{E}_t[\mu] = \mu_t \tag{6}$$

and

$$\mathbb{E}_t[\Sigma] = \frac{\psi_t}{\upsilon_t - n - 1} \tag{7}$$

We simulate potential candidates $\mathcal{X} = \{X_i\} \sim \mathcal{N}(\mathbb{E}_t[\mu], \mathbb{E}_t[\Sigma])$ and evaluate them $f(X_i)$. If the distribution of the minimum was accurate, the minimum would concentrate around $\mathbb{E}_t[\mu]$ and be spread with a variance of $\mathbb{E}_t[\Sigma]$. When evaluating potential candidates, as our guess is not right, we do not get values centered around $\mathbb{E}_t[\mu]$ and spread with a variance of $\mathbb{E}_t[\Sigma]$. This comes from three things:

- Our assumed minimum is not right. We need to shift our normal to the right minimum!
- Our assumed variance is not right. We need to compute it on real data taken into additional information given by f.
- Last but not least, our Monte Carlo simulation adds some random noise.

For the last issue, we can correct any of our estimator by the Monte Carlo bias. This can be done using standard control variate as the simulated mean and variance are given: $\mathbb{E}_t[\mu]$ and $\mathbb{E}_t[\Sigma]$ respectively and we can compute for each of them the bias explicitly.

The first two issues are more complex. Let us tackle each issue one by one.

To recover the true minimum, we design two strategies.

- We design a strategy where we rebuild our normal distribution but using sorted information of our X's weighted by their normal density to ensure this is a true normal corrected from the Monte Carlo bias. We need to explicitly compute the weights. For each simulated point X_i, we compute it assumed density denoted by $d_i = \mathcal{N}(\mathbb{E}_t[\mu], \mathbb{E}_t[\Sigma])(X_i)$ where $\mathcal{N}(\mathbb{E}_t[\mu], \mathbb{E}_t[\Sigma])(.)$ denotes the p.d.f. of the multi-variate Gaussian.
  We divide these density by their sum to get weights $(w_i)_{i=1..k}$ that are positive and sum to one as follows. $w_j = d_j / \sum_{i=1}^k d_i$. Hence for k simulated points, we get $\{X_i, w_i\}_{i=1..k}$. We reorder jointly the uplets (points and density) in terms of their weights in decreasing order.
  To insist we take sorted value in decreasing order with respect to the weights $(w_i)_{i=1..k}$, we denote the order statistics $(i), w\downarrow$.
  This first sorting leads to k new uplets $\{X_{(i),w\downarrow}, w_{(i),w\downarrow}\}_{i=1..k}$. Using a *stable* sort (that keeps the order of the density), we sort jointly the uplets (points and weights) according to their objective function value (in increasing order this time) and get a k new uplets $\{X_{(i),f\uparrow}, w_{(i),w\downarrow}\}_{i=1..k}$. We can now compute the empirical mean $\overline{\mu}_t$ as follows:
$$\overline{\mu}_t = \underbrace{\sum_{i=1}^k w_{(i),w\downarrow} \cdot X_{(i),f\uparrow}}_{\text{MC mean for } X_{f\uparrow}} - \underbrace{\left(\sum_{i=1}^k w_i X_i - \overline{\mu}_t\right)}_{\text{MC bias for } X} \tag{8}$$
  The intuition of equation (8) is to compute in the left term the Monte Carlo mean using reordered points according to their objective value and correct our initial computation by the Monte Carlo bias computed as the right term, equal to the initial Monte Carlo mean minus the real mean. We call this strategy one.
- If we think for a minute about the strategy one, we get the intuition that when starting the minimization, it may not be optimal. This is because weights are proportional to $\exp\left\{\frac{1}{2}(X - \mathbb{E}_t[\mu])^T (\mathbb{E}_t[\Sigma])^{-1}(X - \mathbb{E}_t[\mu])\right\}$.
  When we start the algorithm, we use a large search space, hence a large covariance matrix $\overline{\Sigma}_t$ which leads to have weights which are quite similar. Hence even if we sort candidates by their fit, ranking them according to the value of f in increasing order, we will move our theoretical multi variate Gaussian little by little. A better solution is more to brutally move the center of our multi variate Gaussian to the best candidate seen so far, as follows:
$$\overline{\mu}_t = \arg\min_{X \in \mathcal{X}} f(X) \tag{9}$$
  We call this strategy two. Intuitively, strategy two should be best when starting the algorithm while strategy one would be better once we are close to the solution.

To recover the true variance, we can adapt what we did in stratey one as follows:

- $$\overline{\Sigma}_t = \underbrace{\sum_{i=1}^k w_{(i),w\downarrow} \cdot \left(X_{(i),f\uparrow} - \overline{X}_{(.),f\uparrow}\right)\left(X_{(i),f\uparrow} - \overline{X}_{(.),f\uparrow}\right)^T}_{\text{MC covariance for } X_{f\uparrow}} - \underbrace{\left(\sum_{i=1}^k w_i \cdot \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T - \overline{\Sigma}_t\right)}_{\text{MC covariance for simulated } X} \tag{10}$$
  where $\overline{X}_{(.),f\uparrow} = \sum_{i=1}^k w_{(i),w\downarrow} X_{(i),f\uparrow}$ and $\overline{X} = \sum_{i=1}^k w_i X_i$ are respectively the mean of the sorted and non sorted points.
- Again, we could design another strategy that takes part of the points but we leave this to further research.

Once we have the likelihood mean and variance using (9) and (10) or (8) and (10), we update the posterior law according to equation (5). This gives us the iterative conjugate prior parameters updates:
$$\begin{aligned}
\mu_{t+1} &= \frac{\kappa_t \mu_t + n\overline{\mu}_t}{\kappa_t + n}, \\
\kappa_{t+1} &= \kappa_t + n, \\
\upsilon_{t+1} &= \upsilon_t + n, \\
\psi_{t+1} &= \psi_t + \overline{\Sigma}_t + \frac{\kappa_t n}{\kappa_t + n}\left(\overline{\mu}_t - \mu_t\right)\left(\overline{\mu}_t - \mu_t\right)^T
\end{aligned} \tag{11}$$

The resulting algorithm is summarized in Algo 1.

Proposition 3.1. *Under the assumption of a NIW prior, the updates of the BCMA-ES parameters for the expected mean and variance write as a weighted combination of the prior expected mean and variance and the empirical mean and variance as follows*
$$\begin{aligned}
\mathbb{E}_{t+1}[\mu] &= \mathbb{E}_t[\mu] + w_t^{\mu}\left(\overline{\mu}_t - \mathbb{E}_t[\mu]\right), \\
\mathbb{E}_{t+1}[\Sigma] &= \underbrace{w_t^{\Sigma,1}}_{\text{discount factor}} \mathbb{E}_t[\Sigma] + w_t^{\Sigma,2} \underbrace{\left(\overline{\mu}_t - \mathbb{E}_t[\mu]\right)\left(\overline{\mu}_t - \mathbb{E}_t[\mu]\right)^T}_{\text{rank one matrix}} \\
&\quad + w_t^{\Sigma,3} \underbrace{\overline{\Sigma}_t}_{\text{rank (n-1) matrix}}
\end{aligned}$$
where
$$\begin{aligned}
w_t^{\mu} &= \frac{n}{\kappa_t + n}, \\
w_t^{\Sigma,1} &= \frac{\kappa_t n}{(\kappa_t + n)(\upsilon_t - 1)} \\
w_t^{\Sigma,2} &= \frac{\upsilon_t - n - 1}{\upsilon_t - 1}, \\
w_t^{\Sigma,3} &= \frac{1}{\upsilon_t - 1}
\end{aligned} \tag{12}$$

Remark 3.1. *The proposition above is quite fundamental. It justifies that under the assumption of NIW prior, the update is a weighted sum of previous expected mean and covariance. It is striking that it provides very similar formulae to the standard CMA ES update. Recall that these updates given for the mean m_t and covariance C_t can be written as follows:*

$$\begin{aligned} m_{t+1} &= m_t + \sum_{i=1}^{\mu} w_i \, (x_{i:\lambda} - m_t) \\ C_{t+1} &= \underbrace{(1 - c_1 - c_\mu + c_s)}_{\text{discount factor}} \, C_t + c_1 \underbrace{p_c p_c^T}_{\text{rank one matrix}} \\ &\quad + \, c_\mu \underbrace{\sum_{i=1}^{\mu} w_i \frac{x_{i:\lambda} - m_k}{\sigma_k} \left(\frac{x_{i:\lambda} - m_t}{\sigma_t} \right)^T}_{\text{rank } \min(\mu, n-1) \text{ matrix}} \end{aligned} \tag{13}$$

where the notations $m_t, w_i, x_{i:\lambda}, C_t, c_1, c_\mu, c_s$, etc… are given for instance in [26].

Proof. See 6.3 in the appendix section. □

Algorithm 1 Predict and Correct parameters at step t

Simulate candidate
Use mean values $\mathbb{E}_t[\mu] = \mu_t$ and $\overline{\Sigma}_t = \mathbb{E}[\Sigma] = \psi_t/(\upsilon_t - n - 1)$
Simulate k points $\mathcal{X} = \{X_i\} = 1..k \sim \mathcal{N}(\mathbb{E}_t[\mu], \overline{\Sigma}_t)$
Compute densities $(d_i)_{i..k} = (\mathcal{N}(\mathbb{E}_t[\mu], \overline{\Sigma}_t)(X_i))_{i..k} =$
Sort in decreasing order with respect to d to get $\{X_{(i),d\downarrow}, d_{(i),d\downarrow}\}_{i=1..k}$
Stable Sort in increasing order order with respect to $f(X_i)$ to get $\{X_{(i),f\uparrow}, d_{(i),d\downarrow}\}_{i=1..k}$

Correct $\mathbb{E}_t[\mu]$ and $\overline{\Sigma}_t$
Either Update $\mathbb{E}_t[\mu]$ and $\overline{\Sigma}_t$ using (9) and (10) **(strategy two)**
Or Update $\mathbb{E}_t[\mu]$ and $\overline{\Sigma}_t$ using (8) and (10) **(strategy one)**
Update $\mu_{n+1}, \kappa_{n+1}, \upsilon_{n+1}, \psi_{n+1}$ using (11)

3.3 Particularities of Bayesian CMA-ES

There are some subtleties that need to be emphasized.

- Although we assume a prior, we do not need to simulate the prior but can at each step use the expected value of the prior which means that we do not consume additional simulation compared to the standard CMA-ES.
- We need to tackle local minimum (we will give example of this in the numerical section) to avoid being trapped in a bowl! If we are in a local minimum, we need to inflate the variance to increase our search space. We do this whenever our algorithm does not manage to decrease. However, if after a while we do not get better result, we assume that this is indeed not a local minimum but rather a global minimum and start deflating the variance. This mechanism of inflation deflation ensures we can handle noisy functions like Rastrigin or Schwefel 1 or Schwefel 2 functions as defined in the section 4.

3.4 Differences with standard CMA-ES

Since we use a rigorous derivation of the posterior, we have the following features:

- the update of the covariance takes all points. This is different from λ/μ CMA-ES that uses only a subset of the point.
- by design, the update is optimal as we compute at each step the posterior.
- the contraction dilatation mechanism is an alternative to global local search path in standard CMA-ES.
- weights varies across iterations which is also a major difference between main CMA ES and Bayesian CMA ES. Weights are proportional to $\exp(\frac{1}{2} X^T \Sigma^{-1} X)$ sorted in decreasing order. Initially, when the variance is large,

3.5 Full algorithm

The complete Bayesian CMA ES algorithm is summarized in 2. It iterates until a stopping condition is met. We use multiple stopping conditions. We stop if we have not increase our best result for a given number of iterations. We stop if we have reached the maximum of our iterations. We stop if our variance norm is small. Additional stopping condition can be incorporated easily.

Algorithm 2 Bayesian update of CMA-ES parameters:

Initialization
Start with a prior distribution Π on μ and Σ
Set retrial to 0
Set f_{min} to max float
while stop criteria not satisfied **do**
  $X \sim \mathcal{N}(\mu, \Sigma)$
  update the parameters of the Gaussian thanks to the posterior law $\Pi(\mu, \Sigma|X)$ following details given in algorithm 1
  Handle dilatation contraction variance for local minima as explained in algorithm 3
  if DilateContractFunc($X, \overline{\Sigma}_t, X_{min}, f_{min}, \overline{\Sigma}_{t,min}$) == 1 **then**
    return best solution
  end if
end while
return best solution

Last but not least, we have a dilatation contraction mechanism for the variance to handle local minima with multiple level of contractions and dilatation that is given in function 3. The overall idea is first to dilate variance if we do not make any progress to increase the search space so that we are not trapped in a local minimum. Should this not succeed, it means that we are reaching something that looks like the global minimum and we progressively contract the variance. In our implemented algorithm, we take $L_1 = 5, L_2 = 20, L_3 = 30, L_4 = 40, L_5 = 50$ and the dilatation, contraction parameters given by $k_1 = 1.5, k_2 = 0.9, k_3 = 0.7, k_5 = 0.5$ We have also a restart at previous minimum level $L_* = L_2$.

4 NUMERICAL RESULTS

4.1 Functions examined

We have examined four functions to stress test our algorithm. They are listed in increasing order of complexity for our algorithm and correspond to different type of functions. They are all generalized function that can defined for any dimension n. For all, we present the corresponding equation for a variable $x = (x_1, x_2, .., x_n)$ of n dimension. Code is provided in supplementary materials. We have frozen seeds to have *reproducible results*.

Algorithm 3 Dilatation contraction variance for local minima:

Function DilateContractFunc($X, \overline{\Sigma}_t, X_{min}, f_{min}, \overline{\Sigma}_{t,min}$)
if $f(X) \leq f_{min}$ **then**
  Set $f_{min} = f(X)$
  Memorize current point and its variance:
    • $X_{min} = X$
    • $\overline{\Sigma}_{t,min} = \overline{\Sigma}_t$
  Set retrial = 0
else
  Set retrial += 1
  if retrial == L_* **then**
    Restart at previous best solution:
      • $X = X_{min}$
      • $\overline{\Sigma}_t = \overline{\Sigma}_{t,min}$
  end if
  if L_2 > retrial and retrial > L_1 **then**
    Dilate variance by k_1
  else if L_3 > retrial and retrial ≥ L_2 **then**
    Contract variance by k_2
  else if L_4 > retrial and retrial ≥ L_3 **then**
    Contract variance by k_3
  else if L_5 > retrial and retrial ≥ L_4 **then**
    Contract variance by k_4
  else
    return 1
  end if
  return 0
end if
End Function

4.1.1 Cone. The most simple function to optimize is the quadratic cone whose equation is given by (14) and represented in figure 2. It is also the standard Euclidean norm. It is obviously convex and is a good test of the performance of an optimization method.

$$f(x) = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \|x\|_2 \tag{14}$$

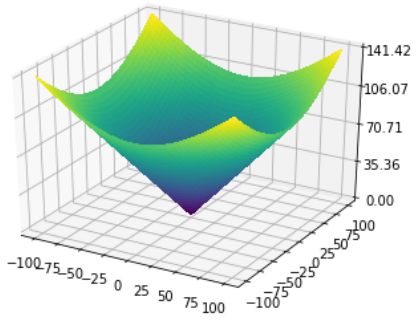

Figure 2: A simple convex function: the quadratic norm. Minimum in 0

4.1.2 Schwefel 2 function. A slightly more complicated function is the Schwefel 2 function whose equation is given by (15) and represented in figure 3. It is a piecewise linear function and validates the algorithm can cope with non convex function.

$$f(x) = \sum_{i=1}^{n} | x_i | + \prod_{i=1}^{n} | x_i | \tag{15}$$

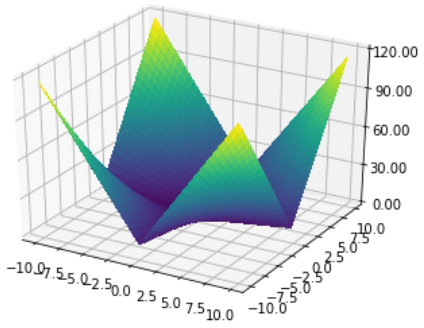

Figure 3: Schwefel 2 function: a simple piecewise linear function

4.1.3 Rastrigin. The Rastrigin function, first proposed by [22] and generalized by [20], is more difficult compared to the Cone and the Schwefel 2 function. Its equation is given by (16) and represented in figure 4. It is a non-convex function often used as a performance test problem for optimization algorithms. It is a typical example of non-linear multi modal function. Finding its minimum is considered a good stress test for an optimization algorithm, due to its large search space and its large number of local minima.

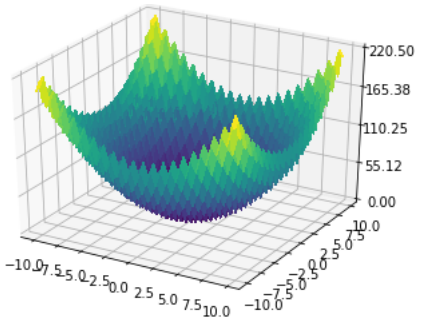

Figure 4: Rastrigin function: a non convex function multimodal and with a large number of local minima

$$f(x) = 10 \times n + \sum_{i=1}^{n} \left[x_i^2 - 10 \cos(2\pi x_i) \right] \tag{16}$$

4.1.4 Schwefel 1 function. The last function we tested is the Schwefel 1 function whose equation is given by (17) and represented in figure 5. It is sometimes only defined on $[-500, 500]^n$. The Schwefel 1 function shares similarities with the Rastrigin function. It is continuous, not convex, multi-modal and with a large number of local minima. The extra difficulty compared to the Rastrigin function, the local minima are more pronounced local bowl making the optimization even harder.

$$f(x) = 418.9829 \times n - \sum_{i=1}^{n} \left[x_i \sin(\sqrt{|x_i|}) \mathbb{1}_{|x_i|<500} + 500 \sin(\sqrt{500}) \mathbb{1}_{|x_i|\geq 500} \right] \quad (17)$$

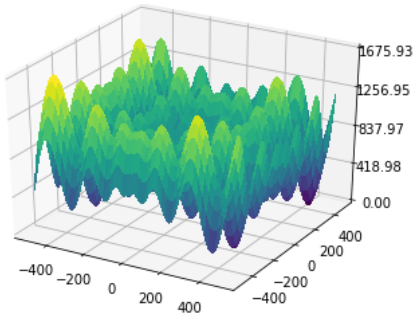

Figure 5: Schwefel 1 function: a non convex function multi-modal and with a large number of local pronounced bowls

4.2 Convergence

For each of the functions, we compared our method using strategy one entitled *B-CMA-ES S1*: update $\overline{\mu}_t$ and $\overline{\Sigma}_t$ using (8) and (10) plotted in *orange*, or strategy two *B-CMA-ES S2*: same update but using (9) and (10), plotted in *blue* and standard CMA-ES as provided by the opensource python package pycma plotted in *green*. We clearly see that strategies one and two are quite similar to standard CMA-ES. The convergence graphics that show the error compared to the minimum are represented:

- for the cone function by figure 6 (case of a convex function), with initial point (10, 10)
- for the Schwefel 2 function in figure 7 (case of piecewise linear function), with initial point (10, 10)
- for the Rastrigin function in figure 8 (case of a non convex function with multiple local minima), with initial point (10, 10)
- and for the Schwefel 1 function in figure 9 (case of a non convex function with multiple large bowl local minima), with initial point (400, 400)

The results are for one test run. In a forthcoming paper, we will benchmark them with more runs to validate the interest of this new method.

For the four functions, BCMAES achieves convergence similar to standard CMA-ES. The intuition of this good convergence is that shifting the multi variate mean by the best candidate seen so far is a good guess to update it at the next run (standard CMA-ES or B-CMA-ES S1).

5 CONCLUSION

In this paper, we have revisited the CMA-ES algorithm and provided a Bayesian version of it. Taking conjugate priors, we can find optimal update for the mean and covariance of the multi variate Normal. We have provided the corresponding algorithm that is a new version of CMA-ES. First numerical experiments show this new version is competitive to standard CMA-ES on traditional functions such as cone, Schwefel 1, Rastrigin and Schwefel 2. This faster convergence can be explained on a theoretical side from an optimal update of the prior (thanks to Bayesian update) and the use of the best candidate seen at each simulation to shift the mean of the multi-variate Gaussian likelihood. We envisage further works to benchmark our algorithm to other standard evolutionary algorithms, in particular to use the COCO platform to provide more meaningful tests and confirm the theoretical intuition of good performance of this new version of CMA-ES, and to test the importance of the prior choice.

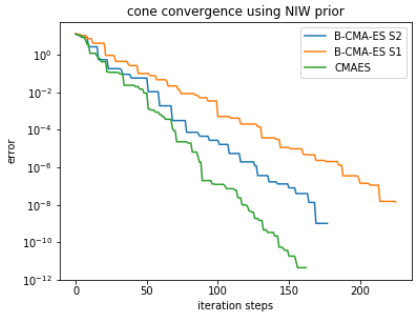

Figure 6: Convergence for the Cone function

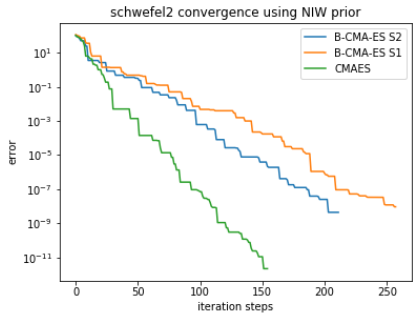

Figure 7: Convergence for the Schwefel 2 function

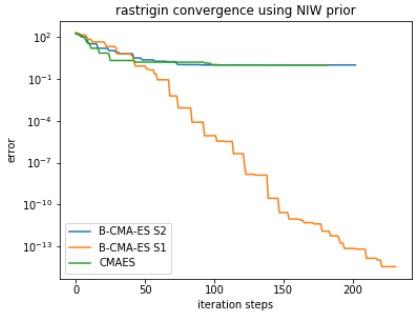

Figure 8: Convergence for the Rastrigin function

6 APPENDIX

6.1 Conjugate priors

Proof. Consider n independent and identically distributed (IID) measurements $\mathcal{X} \triangleq \{\mathbf{x}^j \in \mathbb{R}^d | 1 \leq j \leq n\}$ and assume that these variables have an exponential family density. The likelihood

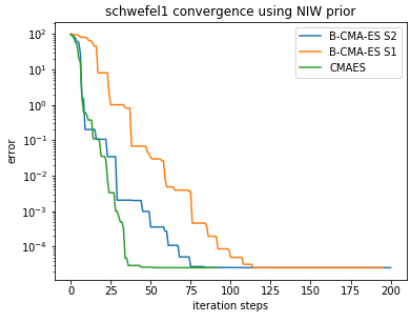

Figure 9: Convergence for the Schwefel 1 function

$p(\mathcal{X}|\theta,\lambda)$ writes simply as the product of each individual likelihood:

$$p(\mathcal{X}|\theta,\lambda)=\left(\prod_{j=1}^{n}h(\mathbf{x}^j)\right)\exp\left(\eta(\theta,\lambda)^T\sum_{j=1}^{n}T(x^j)-nA(\eta(\theta,\lambda))\right). \quad (18)$$

If we start with a prior $\pi(\theta)$ of the form $\pi(\theta)\propto\exp(\mathcal{F}(\theta))$ for some function $\mathcal{F}(\cdot)$, its posterior writes:

$$\begin{aligned}\pi(\theta|\mathcal{X}) &\propto p(\mathcal{X}|\theta)\exp(\mathcal{F}(\theta))\\ &\propto \exp\left(\eta(\theta,\lambda)\cdot\sum_{j=1}^{n}T(x^j)-nA(\eta(\theta,\lambda))+\mathcal{F}(\theta)\right). \quad (19)\end{aligned}$$

It is easy to check that the posterior (19) is in the same exponential family as the prior iff $\mathcal{F}(\cdot)$ is in the form:

$$\mathcal{F}(\theta)=\mu_1\cdot\eta(\theta,\lambda)-\mu_0 A(\eta(\theta,\lambda)) \quad (20)$$

for some $\mu\triangleq(\mu_0,\mu_1)$, such that:

$$p(\mathcal{X}|\theta,\lambda)\propto\exp\left(\left(\mu_1+\sum_{j=1}^{n}T(x^j)\right)^T\eta(\theta,\lambda)-(n+\mu_0)A(\eta(\theta,\lambda))\right). \quad (21)$$

Hence, the conjugate prior for the likelihood (18) is parametrized by μ and given by:

$$p(\mathcal{X}|\theta,\lambda)=\frac{1}{Z}\exp\left(\mu_1\cdot\eta(\theta,\lambda)-\mu_0 A(\eta(\theta,\lambda))\right), \quad (22)$$

where $Z=\int\exp\left(\mu_1\cdot\eta(\theta,\lambda)-\mu_0 A(\eta(\theta,\lambda))\right)\,dx$. □

6.2 Exact computation of the posterior update for the Normal inverse Wishart

To make our proof simple, we first start by the one dimensional case and show that in one dimension it is a normal inverse gamma. We then generalize to the multi dimensional case.

Lemma 6.1. *The probability density function of a Normal inverse gamma (denoted by NIG) random variable can be expressed as the product of a Normal and an Inverse gamma probability density functions.*

Proof. we suppose that $x|\mu,\sigma^2\sim\mathcal{N}(\mu_0,\sigma^2/v)$. We recall the following definition of conditional probability:

Definition 6.1. *Suppose that events A,B and C are defined on the same probability space, and the event B is such that $\mathbb{P}(B)>0$. We have the following expression:*
$\mathbb{P}(A\cap B|C)=\mathbb{P}(A|B,C)\mathbb{P}(B|C)$.

Applying 6.1, we have:

$$\begin{aligned}p\left(\mu,\sigma^2|\mu_0,v,\alpha,\beta\right)&=p\left(\mu|\sigma^2,\mu_0,v,\alpha,\beta\right)p\left(\sigma^2|\mu_0,v,\alpha,\beta\right)\\&=p\left(\mu|\sigma^2,\mu_0,v\right)p\left(\sigma^2|\alpha,\beta\right). \quad (23)\end{aligned}$$

Using the definition of the Normal inverse gamma law, we end the proof. □

Remark 6.1. *If $(x,\sigma^2)\sim NIG(\mu,\lambda,\alpha,\beta)$, the probability density function is the following:*

$$f(x,\sigma^2|\mu,\lambda,\alpha,\beta)=\frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{-\frac{2\beta+\lambda(x-\mu)^2}{2\sigma^2}\right\}. \quad (24)$$

Proposition 6.2. *The Normal Inverse Gamma NIG (μ_0,v,α,β) distribution is a conjugate prior of a normal distribution with unknown mean and variance.*

Proof. the posterior is proportional to the product of the prior and likelihood, then:

$$\begin{aligned}p\left(\mu,\sigma^2|X\right)\propto&\frac{\sqrt{v}}{\sqrt{2\pi}}\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{\frac{-v(\mu-\mu_0)^2}{2\sigma^2}\right\}\\&\times\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{\frac{-\beta}{\sigma^2}\right\}\\&\times\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}\right\}. \quad (25)\end{aligned}$$

Defining the empirical mean and variance as $\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{s}=\frac{1}{n}\sum_{i=1}^n(x_i-\bar{x})^2$, we obtain that $\sum_{i=1}^n(x_i-\mu)^2=n(\bar{s}+(\bar{x}-\mu)^2)$.

So, the conditional density writes:

$$\begin{aligned}p\left(\mu,\sigma^2|X\right)\propto&\sqrt{v}\left(\frac{1}{\sigma^2}\right)^{\alpha+n/2+3/2}\\&\times\exp\left\{-\frac{1}{\sigma^2}\left[\beta+\frac{1}{2}\left(v(\mu-\mu_0)^2+n\left(\bar{s}+(\bar{x}-\mu)^2\right)\right)\right]\right\}. \quad (26)\end{aligned}$$

Besides,

$$\begin{aligned}&v(\mu-\mu_0)^2+n\left(\bar{s}+(\bar{x}-\mu)^2\right)\\&=v\left(\mu^2-2\mu\mu_0+\mu_o^2\right)+n\bar{s}+n\left(\bar{x}^2-2\bar{x}\mu+\mu^2\right)\\&=\mu^2\left(v+n\right)-2\mu\left(v\mu_0+n\bar{x}\right)+v\mu_o^2+n\bar{s}+n\bar{x}^2. \quad (27)\end{aligned}$$

Denoting $a=v+n$ and $b=v\mu_0+n\bar{x}$, we have :

$$\begin{aligned}&\beta+\frac{1}{2}\left(v(\mu-\mu_0)^2+n\left(\bar{s}+(\bar{x}-\mu)^2\right)\right)\\&=\beta+\frac{1}{2}\left(a\mu^2-2b\mu+v\mu_o^2+n\bar{s}+n\bar{x}^2\right)\\&=\beta+\frac{1}{2}\left(a\left(\mu^2-\frac{2b}{a}\mu\right)+v\mu_o^2+n\bar{s}+n\bar{x}^2\right)\\&=\beta+\frac{1}{2}\left(a\left(\mu-\frac{b}{a}\right)^2-\frac{b^2}{a}+v\mu_o^2+n\bar{s}+n\bar{x}^2\right). \quad (28)\end{aligned}$$

So we can express the proportional expression of the posterior :

$$p\left(\mu,\sigma^2|X\right) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha^\star+3/2} \times \exp\left\{-\frac{2\beta^\star+\lambda^\star\left(\mu-\mu^\star\right)^2}{2\sigma^2}\right\},$$

with

- $\alpha^\star = \alpha + \frac{n}{2}$
- $\beta^\star = \beta + \frac{1}{2}\left(\sum_{i=1}^n (x_i-\overline{x})^2 + \frac{n\upsilon}{n+\upsilon}\frac{(\overline{x}-\mu_0)^2}{2}\right)$
- $\mu^\star = \frac{\upsilon\mu_0+n\overline{x}}{\upsilon+n}$
- $\lambda^\star = \upsilon + n$

We can identify the terms with the expression of the probability density function given in 6.1 to conclude that the posterior follows a NIG$(\mu^\star,\lambda^\star,\alpha^\star,\beta^\star)$. □

We are now ready to prove the following proposition:

Proposition 6.3. *The Normal Inverse Wishart (denoted by NIW) $(\mu_0,\kappa_0,\upsilon_0,\psi)$ distribution is a conjugate prior of a multivariate normal distribution with unknown mean and covariance.*

Proof. we use the fact that the probability density function of a Normal inverse Wishart random variable can be expressed as the product of a Normal and an Inverse Wishart probability density functions (we use the same reasoning that in 6.1). Besides, the posterior is proportional to the product of the prior and the likelihood.
We first express the probability density function of the multivariate Gaussian random variable in a proper way in order to use it when we write the posterior density function.

$$\begin{aligned} &\sum_{i=1}^n (x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \\ = \quad & n(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu) + \sum_{i=1}^n (x_i-\overline{x})^T\Sigma^{-1}(x_i-\overline{x}). \end{aligned} \tag{29}$$

We can inject the previous result and use the properties of the trace function to express the following probability density function of the multivariate Gaussian random variable of parameters μ and Σ. The density writes as:

$$\begin{aligned} &\frac{|\Sigma|^{-n/2}}{\sqrt{(2\pi)^{pn}}}\exp\left\{-\tfrac{n}{2}(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu)\right. \\ &\qquad\left.-\tfrac{1}{2}tr\left(\Sigma^{-1}\sum_{i=1}^n (x_i-\overline{x})(x_i-\overline{x})^T\right)\right\}. \end{aligned} \tag{30}$$

Hence, we can compute explicitly the posterior as follows:

$$\begin{aligned} p\left(\mu,\sigma^2|X\right) &\propto \frac{\sqrt{\kappa_0}}{\sqrt{(2\pi)^p|\Sigma|}}\exp\left\{-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\} \\ &\quad\times\frac{|\psi|^{\upsilon/2}}{2^{\upsilon p/2}\Gamma_p(\upsilon_0/2)}|\Sigma|^{-\frac{\upsilon_0+p+1}{2}}exp\left\{-\frac{1}{2}tr\left(\psi\Sigma^{-1}\right)\right\} \\ &\quad\times|\Sigma|^{-n/2}\exp\left\{-\frac{n}{2}(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu)\right. \\ &\quad\left.-\frac{1}{2}tr\left(\Sigma^{-1}\sum_{i=1}^n (x_i-\overline{x})(x_i-\overline{x})^T\right)\right\} \end{aligned} \tag{31}$$

$$\begin{aligned} &\propto |\Sigma|^{-\frac{\upsilon_0+p+2+n}{2}}\exp\left\{-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right. \\ &\quad-\frac{n}{2}(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu) \\ &\quad\left.-\frac{1}{2}tr\left(\Sigma^{-1}\left(\psi+\sum_{i=1}^n (x_i-\overline{x})(x_i-\overline{x})^T\right)\right)\right\}. \end{aligned} \tag{32}$$

We organize the terms and find the parameters of our Normal Inverse Wishart random variable NIW$(\mu_0^\star,\kappa_0^\star,\upsilon_0^\star,\psi^\star)$.

$$\begin{aligned} &\mu_0^\star = \frac{\kappa_0\mu_0+n\overline{x}}{\kappa_0+n}, \quad \kappa_0^\star = \kappa_0+n, \quad \upsilon_0^\star = \upsilon_0+n \\ &\psi^\star = \psi + \sum_{i=1}^n (x_i-\overline{x})(x_i-\overline{x})^T + \frac{\kappa_0 n}{\kappa_0+n}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T \end{aligned} \tag{33}$$

which are exactly the equations provided in (5). □

6.3 Weighted combination for the BCMA ES update

Proof.

$$\begin{aligned} \mathbb{E}_{t+1}[\mu] &= \mu_{t+1} \\ &= \frac{\kappa_t\mu_t+n\overline{\mu}_t}{\kappa_t+n} \\ &= \mathbb{E}_t[\mu] + w_t^\mu\left(\hat{\mu}-\mathbb{E}_t[\mu]\right) \end{aligned} \tag{34}$$

$$\begin{aligned} \mathbb{E}_{t+1}[\Sigma] &= \frac{\psi_{t+1}}{\upsilon_{t+1}-n-1} \\ &= \frac{1}{\upsilon_t-1}\psi_t + \frac{1}{\upsilon_t-1}\overline{\Sigma}_t \\ &\quad+\frac{\kappa_t n}{(\kappa_t+n)(\upsilon_t-1)}\left(\overline{\mu}_t-\mu_t\right)\left(\overline{\mu}_t-\mu_t\right)^T \\ &= \underbrace{w_t^{\Sigma,1}}_{\text{discount factor}}\mathbb{E}_t[\Sigma] + w_t^{\Sigma,2}\underbrace{(\hat{\mu}-\mathbb{E}_t[\mu])(\hat{\mu}-\mathbb{E}_t[\mu])^T}_{\text{rank one matrix}} \\ &\quad+ w_t^{\Sigma,3}\underbrace{\overline{\Sigma}_t}_{\text{rank (n-1) matrix}} \end{aligned}$$

$$\begin{aligned} \text{where} \quad w_t^\mu &= \frac{n}{\kappa_t+n}, \\ w_t^{\Sigma,1} &= \frac{\upsilon_t-n-1}{\upsilon_t-1}, \\ w_t^{\Sigma,2} &= \frac{\kappa_t n}{(\kappa_t+n)(\upsilon_t-1)} \end{aligned} \tag{35}$$

$\overline{\Sigma}_t$ is a covariance matrix of rank $n-1$ as we subtract the empirical mean (which removes one degree of freedom). The matrix $(\hat{\mu}-\mathbb{E}_t[\mu])(\hat{\mu}-\mathbb{E}_t[\mu])^T$ is of rank 1 as it is parametrized by the vector $\hat{\mu}$. □

REFERENCES

[1] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2015. Continuous Optimization and CMA-ES. *GECCO 2015, Madrid, Spain* 1 (2015), 313–344.

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2016. CMA-ES and Advanced Adaptation Mechanisms. *GECCO, Denver* 2016 (2016), 533–562.

[3] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. 2010. Bidirectional Relation between CMA Evolution Strategies and Natural Evolution Strategies. *PPSN* XI, 1 (2010), 154–163.

[4] Anne Auger and Nikolaus Hansen. 2009. Benchmarking the (1+1)-CMA-ES on the BBOB-2009 noisy testbed. *Companion Material* GECCO 2009 (2009), 2467–2472.

[5] Anne Auger and Nikolaus Hansen. 2012. Tutorial CMA-ES: evolution strategies and covariance matrix adaptation. *Companion Material Proceedings* 2012, 12 (2012), 827–848.

[6] Anne Auger, Marc Schoenauer, and Nicolas Vanhaecke. 2004. LS-CMA-ES: A Second-Order Algorithm for Covariance Matrix Adaptation. *PPSN VIII, 8th International Conference, Birmingham, UK, September 18-22, 2004, Proceedings* 2004, 2004 (2004), 182–191.

[7] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis* (2nd ed. ed.). Chapman and Hall/CRC, New York.

[8] T. Glasmachers, T. Schaul, S. Yi, D. Wierstra, and J.: Schmidhuber. 2010. Exponential natural evolution strategies. *In: Proceedings of Genetic and Evolutionary Computation Conference, pp* 2010, 2010 (2010), 393–400.

[9] Nikolaus Hansen. 2016. The CMA Evolution Strategy: A Tutorial, Preprint. arXiv:1604.00772

[10] Nikolaus Hansen and Anne Auger. 2011. CMA-ES: evolution strategies and covariance matrix adaptation. *GECCO 2011* 2011, 1 (2011), 991–1010.

[11] Nikolaus Hansen and Anne Auger. 2014. Evolution strategies and CMA-ES (covariance matrix adaptation). *GECCO Vancouver* 2014, 14 (2014), 513–534.

[12] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation* 9, 2 (2001), 159–195. https://doi.org/10.1162/106365601750190398

[13] Verena Heidrich-Meisner and Christian Igel. 2009. Neuroevolution strategies for episodic reinforcement learning. *J. Algorithms* 64, 4 (2009), 152–168.

[14] Christian Igel. 2010. Evolutionary Kernel Learning. In *Encyclopedia of Machine Learning and Data Mining*. Springer, New-York, 465–469.

[15] Christian Igel, Nikolaus Hansen, and Stefan Roth. 2007. Covariance Matrix Adaptation for Multi-objective Optimization. *Evol. Comput.* 15, 1 (March 2007), 1–28.

[16] Christian Igel, Verena Heidrich-Meisner, and Tobias Glasmachers. 2009. Shark. *Journal of Machine Learning Research* 9 (2009), 993–996.

[17] M. I. Jordan. 2010. Lecture notes: Justification for Bayes. https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture2.pdf

[18] Ilya Loshchilov and Frank Hutter. 2016. CMA-ES for Hyperparameter Optimization of Deep Neural Networks. *arXiv e-prints* 1604, Apr (April 2016), arXiv:1604.07269. arXiv:cs.NE/1604.07269

[19] Jean-Michel Marin and Christian P. Robert. 2007. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[20] H. Mühlenbein, M. Schomisch, and J. Born. 1991. The parallel genetic algorithm as function optimizer. *Parallel Comput.* 17, 6 (1991), 619–632.

[21] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. 2017. Information-geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *J. Mach. Learn. Res.* 18, 1 (Jan. 2017), 564–628.

[22] L. A. Rastrigin. 1974. *Systems of extremal control.* Mir, Moscow.

[23] C. P. Robert. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer, New York.

[24] M.J. Schervish. 1996. *Theory of Statistics.* Springer, New York. https://books.google.fr/books?id=F9A9af4It10C

[25] Konstantinos Varelas, Anne Auger, Dimo Brockhoff, Nikolaus Hansen, Ouassim Ait ElHara, Yann Semet, Rami Kassab, and Frédéric Barbaresco. 2018. A Comparative Study of Large-Scale Variants of CMA-ES. *PPSN XV - 15th International Conference, Coimbra, Portugal* 15, 2018 (2018), 3–15.

[26] Wikipedia. 2018. CMA-ES. https://en.wikipedia.org/wiki/CMA-ES