CMA-ES with Learning Rate Adaptation: Can CMA-ES with Default Population Size Solve Multimodal and Noisy Problems? Supplementary Material

A Derivation for Section 3.2

A.1 Derivation of Eq. (12)

This section presents the detailed derivation of Eq. (12). By ignoring $(1-\beta)^n$, $\mathcal{E}^{(t+n)}$ can be approximately calculated as follows:

$$\begin{aligned}
\mathcal{E}^{(t+n)} &= (1-\beta)\mathcal{E}^{(t+n-1)} + \beta\tilde{\Delta}^{(t+n-1)} \\
&= (1-\beta)\left\{(1-\beta)\mathcal{E}^{(t+n-2)} + \beta\tilde{\Delta}^{(t+n-2)}\right\} + \beta\tilde{\Delta}^{(t+n-1)} \\
&= \dots \\
&= (1-\beta)^n\mathcal{E}^{(t)} + \sum_{i=0}^{n-1}(1-\beta)^i\beta\tilde{\Delta}^{(t+n-1-i)} \\
&\approx \sum_{i=0}^{n-1}(1-\beta)^i\beta\tilde{\Delta}^{(t+n-1-i)}.
\end{aligned}$$

Here, we assume the $\tilde{\Delta}^{(\cdot)}$ are uncorrelated with each other; this corresponds to the scenario where η is sufficiently small. In this case, we can ignore the dependence of t, i.e., $\mathbb{E}[\tilde{\Delta}^{(t+n-1-i)}] =: \mathbb{E}[\tilde{\Delta}]$. Thus,

$$\mathbb{E}[\mathcal{E}^{(t+n)}] = \sum_{i=0}^{n-1}(1-\beta)^i\beta\mathbb{E}[\tilde{\Delta}].$$

Here,

$$\sum_{i=0}^{n-1}(1-\beta)^i = \frac{1\cdot\{1-(1-\beta)^n\}}{1-(1-\beta)} = \frac{1-(1-\beta)^n}{\beta}.$$

Then, by ignoring $(1-\beta)^n$, we can approximate $\mathbb{E}[\mathcal{E}^{(t+n)}]$ as follows:

$$\mathbb{E}[\mathcal{E}^{(t+n)}] = [1-(1-\beta)^n]\mathbb{E}[\tilde{\Delta}] \approx \mathbb{E}[\tilde{\Delta}].$$

Next, we consider the covariance $\mathrm{Cov}[\mathcal{E}^{(t+n)}]$:

$$\mathrm{Cov}[\mathcal{E}^{(t+n)}] = \mathbb{E}[\mathcal{E}^{(t+n)}(\mathcal{E}^{(t+n)})^\top] - \mathbb{E}[[\mathcal{E}^{(t+n)}]([\mathcal{E}^{(t+n)}])^\top.$$

First, we find the exact expression of $\mathcal{E}^{(t+n)}(\mathcal{E}^{(t+n)})^\top$:

$$\begin{aligned}
\mathcal{E}^{(t+n)}(\mathcal{E}^{(t+n)})^\top &= \beta^2\sum_{i=0}^{n-1}(1-\beta)^{2i}\tilde{\Delta}^{(t+n-1-i)}(\tilde{\Delta}^{(t+n-1-i)})^\top \\
&+ 2\beta^2\sum_{i,j=0:i\neq j}^{n-1}(1-\beta)^i(1-\beta)^j\tilde{\Delta}^{(t+n-1-i)}(\tilde{\Delta}^{(t+n-1-i)})^\top.
\end{aligned}$$

Note that, for $i,j \in \{0,\cdots n-1\} (i \neq j)$, $\mathbb{E}[\tilde{\Delta}^{(t+n-1-i)}(\tilde{\Delta}^{(t+n-1-j)})^\top] = \mathbb{E}[\tilde{\Delta}](\mathbb{E}[\tilde{\Delta}])^\top$, as we assume that they are uncorrelated. For $i \in \{0,\cdots n-1\}$, $\mathbb{E}[\tilde{\Delta}^{(t+n-1-i)}(\tilde{\Delta}^{(t+n-1-i)})^\top] = \mathbb{E}[\tilde{\Delta}](\mathbb{E}[\tilde{\Delta}])^\top + \mathrm{Cov}[\tilde{\Delta}]$. Thus,

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}^{(t+n)}(\mathcal{E}^{(t+n)})^\top] \\
&= \beta^2\sum_{i=0}^{n-1}(1-\beta)^{2i}\left(\mathbb{E}[\tilde{\Delta}](\mathbb{E}[\tilde{\Delta}])^\top + \mathrm{Cov}[\tilde{\Delta}]\right) \\
&\quad + 2\beta^2\sum_{i,j=0:i\neq j}^{n-1}(1-\beta)^i(1-\beta)^j\mathbb{E}[\tilde{\Delta}](\mathbb{E}[\tilde{\Delta}])^\top, \\
&= \mathbb{E}[\mathcal{E}^{(t+n)}](\mathbb{E}[\mathcal{E}^{(t+n)}])^\top + \beta^2\sum_{i=0}^{n-1}(1-\beta)^{2i}\,\mathrm{Cov}[\tilde{\Delta}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}[\mathcal{E}^{(t+n)}] &= \mathbb{E}[\mathcal{E}^{(t+n)}(\mathcal{E}^{(t+n)})^\top] - \mathbb{E}[[\mathcal{E}^{(t+n)}]([\mathcal{E}^{(t+n)}])^\top \\
&= \beta^2\sum_{i=0}^{n-1}(1-\beta)^{2i}\,\mathrm{Cov}[\tilde{\Delta}].
\end{aligned}$$

Here,

$$\sum_{i=0}^{n-1}(1-\beta)^{2i} = \frac{1-(1-\beta)^{2n}}{1-(1-\beta)^2} = \frac{1-(1-\beta)^{2n}}{\beta(2-\beta)}.$$

Thus, by ignoring $(1-\beta)^{2n}$, we can approximate $\mathrm{Cov}[\mathcal{E}^{(t+n)}]$ as follows:

$$\begin{aligned}
\mathrm{Cov}[\mathcal{E}^{(t+n)}] &= [1-(1-\beta)^{2n}]\frac{\beta}{2-\beta}\,\mathrm{Cov}[\tilde{\Delta}], \\
&\approx \frac{\beta}{2-\beta}\,\mathrm{Cov}[\tilde{\Delta}].
\end{aligned}$$

Therefore, $\mathcal{E}^{(t+n)}$ approximately follows the distribution

$$\mathcal{E}^{(t+n)} \sim \mathcal{D}\left(\mathbb{E}[\tilde{\Delta}], \frac{\beta}{2-\beta}\,\mathrm{Cov}[\tilde{\Delta}]\right).$$

This completes the derivation of Eq. (12).

A.2 Derivation of Estimates for $\|\mathbb{E}[\tilde{\Delta}]\|_2^2$

We organize the relation between $\mathcal{E}$ and $\tilde{\Delta}$ by the following equation:

$$\begin{aligned}
\mathbb{E}[\|\mathcal{E}\|_2^2] &= \mathbb{E}[\mathcal{E}]^\top I\mathbb{E}[\mathcal{E}] + \mathrm{Tr}(\mathrm{Cov}[\mathcal{E}]) \\
&\approx \|\mathbb{E}[\tilde{\Delta}]\|_2^2 + \mathrm{Tr}\left(\frac{\beta}{2-\beta}\,\mathrm{Cov}[\tilde{\Delta}]\right) \\
&= \|\mathbb{E}[\tilde{\Delta}]\|_2^2 + \frac{\beta}{2-\beta}\,\mathrm{Tr}(\mathrm{Cov}[\tilde{\Delta}]).
\end{aligned}$$

Now we apply the same arguments to $\mathcal{V}$ and obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{V}] &= [1-(1-\beta)^{t+1}]\mathbb{E}[\|\tilde{\Delta}\|_2^2] \\
&\approx \mathbb{E}[\|\tilde{\Delta}\|_2^2] = \|\mathbb{E}[\tilde{\Delta}]\|_2^2 + \mathrm{Tr}(\mathrm{Cov}[\tilde{\Delta}]).
\end{aligned}$$

By reorganizing these arguments, we obtain

$$\|\mathbb{E}[\tilde{\Delta}]\|_2^2 \approx \frac{2-\beta}{2-2\beta}\mathbb{E}[\|\mathcal{E}\|_2^2] - \frac{\beta}{2-2\beta}\mathbb{E}[\mathcal{V}].$$

This gives the rationale of the estimates $\frac{2-\beta}{2-2\beta}\|\mathcal{E}\|_2^2 - \frac{\beta}{2-2\beta}\mathcal{V}$ for $\|\mathbb{E}[\tilde{\Delta}]\|_2^2$.

B Additional Experiment Results

Figure 1 shows the success rate and SP1 results with respect to $\beta_\Sigma \in \{0.01, 0.02, \ldots, 0.05\}$ on the 30-D noiseless Sphere, Schaffer, and Rastrigin functions. Clearly, the performance was not significantly affected by β_Σ values within this range. However, similar to the case shown in Figure **??**, an excessively small β_Σ setting decelerated the convergence for the Rastrigin function.

Figures 2 and 3 show the success rate and SP1 values with respect to β_m and γ, respectively. The results show that the performance was relatively stable against these hyperparameters.

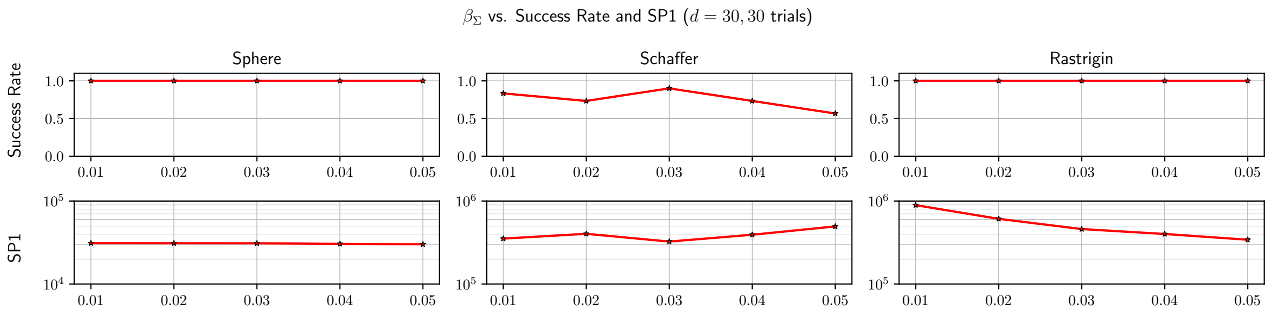

Figure 1: Success rate and SP1 versus hyperparameter $\beta_\Sigma \in \{0.01, 0.02, ..., 0.05\}$ on 30-D noiseless problems.

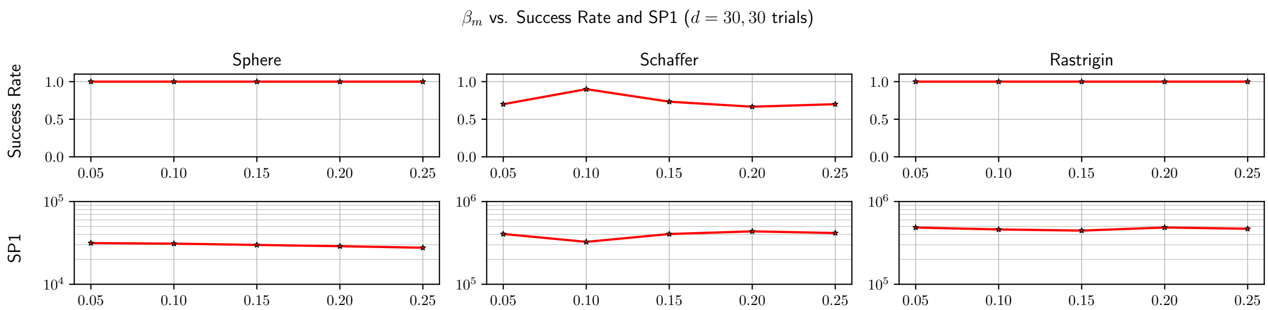

Figure 2: Success rate and SP1 versus hyperparameter β_m on 30-D noiseless problems.

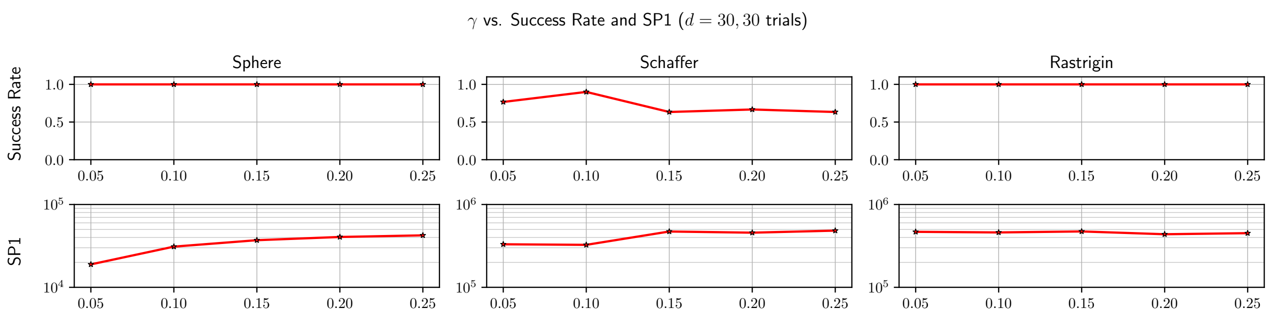

Figure 3: Success rate and SP1 versus hyperparameter γ on 30-D noiseless problems.