CMA-ES

Antoni Zajko

Warsaw University of Technology

2023

CMA-ES wysokopoziomowo

Cel **:** *Zminimalizować funkcję f*

1. Zainicjalizuj parametry:
   1.1 $\Sigma^{(0)} = I$
   1.2 $\mu^{(0)} \in \mathbb{R}^n$
   1.3 $\sigma^{(0)} \in \mathbb{R}$

2. Dopóki nie zostanie spełnione ustalone kryterium:
   2.1 Wygeneruj p punktów x_i tak, że : $x_i \sim \mathcal{N}(\mu^{(g)}, \sigma^{(g)}\Sigma^{(g)})$
   2.2 Zaktualizuj μ, Σ, σ na podstawie najlepszych punktów z x_i

Wizualizacja działania

https://blog.otoro.net/assets/20171031/rastrigin/cmaes.gif

Aktualizacja μ (wartości oczekiwanej rozkładu)

1. Wybierz $q \leqslant p$ "najlepszych" punktów spośród x
2. Posortuj te punkty od najlepszych do najgorszych
3. $\mu^{(g+1)} = \mu^{(g)} + c_\mu \sum_{i=1}^{q} w_i (x_i - \mu)'$

Aktualizacja μ (wartości oczekiwanej rozkładu)

- q jest hiperparametrem tej metody.
- Często $w_i = \frac{1}{q}$, chociaż niekoniecznie.
- Często się przyjmuje, że $\sum_{i=1}^{q} w_i = 1$, chociaż tak niekoniecznie musi być. Suma tych wag nie musi być nawet dodatnia.
- c_μ jest parametrem uczenia dla μ (jednym z wielu w tej metodzie).

Variance effective selection mass

$$q_{eff} = \left(\frac{||w||_1}{||w||_2} \right)^2 = \frac{1}{\sum_{i=1}^{q} w_i^2}$$

Dobór wag

1. Dla $w_i = \frac{1}{q}$, $q_{eff} = q$
2. Uważa się, że jeżeli $q_{eff} \approx \frac{q}{2}$, to dobór wag jest w porządku
3. Przykładowy dobór wag: $w_i \propto q - i + 1$, gdzie $q \approx \frac{p}{2}$

Adaptacja macierzy kowariancji

Estymacja macierzy kowariancji

$$\Sigma = \sum_{i=1}^{q} w_i (x_i - \mu)(x_i - \mu)^\top$$

Rank-q-update

$$\Sigma^{(g+1)} = (1 - c_q)\Sigma^{(g)} + c_q \sum_{i=1}^{q} w_i \left(\frac{x_i - \mu}{\sigma^{(g)}} \right) \left(\frac{x_i - \mu}{\sigma^{(g)}} \right)^\top$$

Gdzie c_q to jest learning rate. Ważny jest odpowiedni jego dobór. Za mały learning rate będzie powodował zbyt wolne uczenie, natomiast za duży spowoduje degeneracje macierzy kowariancji. Eksperymenty pokazują, że $c_q \approx \frac{q_{eff}}{n^2}$ jest optymalnym wyborem.

Ścieżka ewolucji

$$p_c^{(g)} = \begin{cases} 0, & g = 0 \\ (1 - c_c)p_c^{(g-1)} + \sqrt{c_c(2 - c_c)\mu_{eff}}\,\dfrac{m^{(g)} - m^{(g-1)}}{c_m \sigma^{(g)}}, & g > 0 \end{cases}$$

Gdzie c_c jest kolejnym learning rate.

Rank-One-Update

Wykorzystując ścieżkę ewolucji, można aktualizować macierz kowariancji:

$$\Sigma^{(g+1)} = (1 - c_1)\Sigma^{(g)} + c_1 p_c^{(g+1)} (p_c^{(g+1)})^\top$$

Gdzie c_1 jest kolejnym learning rate.

Finalny rezultat

$$\begin{aligned}\Sigma^{(g+1)} = &(1 - c_1 - c_q)\Sigma^{(g)} \\ &+ c_1 p_c^{(g+1)} (p_c^{(g+1)})^\top \\ &+ c_q \sum_{i=1}^{q} w_i \left(\frac{x_i - \mu^{(g)}}{\sigma^{(g)}} \right) \left(\frac{x_i - \mu^{(g)}}{\sigma^{(g)}} \right)^\top\end{aligned}$$

Zalety

- Dużo lepiej sobie radzi z optimami lokalnymi, niż tradycyjne algorytmy np. BFGS,
- Nie wymaga dużej populacji, dzięki odpowiedniemu dopasowaniu współczynników uczenia,
- Inwariancja ze względu na skalowanie, transformacje monotnoniczne, transformacje przestrzeni zachowujące kąty.

Wady

- Kosztowność obliczeniowa, ponieważ w każdej iteracji jest wymagane policzenie kilku macierzy kowariancji,
- inwariancja ze względu na skalowanie, transformacje monotnoniczne, transformacje przestrzenie zachowujące kąty,
- Nie nadaje się do problemów z dynamicznym zbiorem, po którym się optymalizuje tzn. takim, którego postać zależy od optymalizowanych parametrów,
- Nie działa ze zmiennymi kategorycznymi.

Zbieżność

Function	f_{stop}	init	n	CMA-ES	DE	RES	LOS
$f_{\text{Ackley}}(\boldsymbol{x})$	1e-3	$[-30, 30]^n$	20	**2667**	.	.	6.0e4
			30	**3701**	12481	1.1e5	9.3e4
			100	**11900**	36801	.	.
$f_{\text{Griewank}}(\boldsymbol{x})$	1e-3	$[-600, 600]^n$	20	**3111**	8691	.	.
			30	**4455**	11410 *	*8.5e-3/2e5*	.
			100	**12796**	31796	.	.
$f_{\text{Rastrigin}}(\boldsymbol{x})$	0.9	$[-5.12, 5.12]^n$	20	68586	**12971**	.	9.2e4
		DE: $[-600, 600]^n$	30	147416	**20150** *	1.0e5	2.3e5
			100	1010989	**73620**	.	.
$f_{\text{Rastrigin}}(\boldsymbol{Ax})$	0.9	$[-5.12, 5.12]^n$	30	**152000**	*171/1.25e6* *	.	.
			100	**1011556**	*944/1.25e6* *	.	.
$f_{\text{Schwefel}}(\boldsymbol{x})$	1e-3	$[-500, 500]^n$	5	43810	**2567** *	.	7.4e4
			10	240899	**5522** *	.	5.6e5

Bibliografia

1. Opis algorytmu
2. Benchmark algorytmu