Large-scale Surrogate Distribution-based CMA-ES

Nilo Schenke & Michele Sebag

March 8, 2015

1 Context

- Motivating application is to train a large scale neural net: the weight vector is in $\mathbb{R}^{300,000}$ or $\mathbb{R}^{400,000}$
- The point is to avoid building covariance matrix $\Sigma(D \times D)$.
- The idea: define problems in smaller dimension (a few hundred) based on random projections; use their distribution to emulate $\mathcal{N}(\mu, \Sigma_{D\times D})$
- Related: Hutter & Nando de Freitas, IJCAI 2013.
- Tool: Random projections, tutorial J.A. Tropp.

On va décomposer l'algorithme en deux étapes. La première (RP-CMA), on considère une seule projection aléatoire (évidemment, si on tombe mal, pas de bol). La seconde (ERP-CMA), on considère plusieurs projections aléatoires (on est bien mieux).

2 Algo RP-CMA, Random Projection CMA

Soit F la fonction à optimiser:

$$F : \mathbb{R}^D \mapsto \mathbb{R}$$

On va exécuter en parallèle deux CMA-ES. Le premier (le grand) est sur $\mathbb{R}^D$, défini par une distribution $\mathcal{N}_D = (\mu_D, \Sigma_D, \sigma_D)$ où Σ_D est une matrice de covariance diagonale.

Le second (le petit) est sur $\mathbb{R}^d$, défini par une distribution $\mathcal{N}_d = (\mu_d, \Sigma_d, \sigma_d)$.

2.1 Le grand

La seule chose qui change est dans la routine de génération des points.

Chaque échantillon $\mathbf{x}_i$ est obtenu :

- en générant K points $\mathbf{z}_j$ dans $\mathbb{R}^D$ selon $\mathcal{N}_D$;
- en retournant argmax $\{Pr(A\mathbf{z}_j|\mathcal{N}_d), j = 1 \dots K\}$: le point tel que $A\mathbf{z}_j$ soit le plus probable selon $\mathcal{N}_d$.

où K est un paramètre à définir et $\mathcal{N}_d$ est la distribution du petit CMA-ES, ci-dessous.

2.2 Le petit

Soit A une matrice aléatoire $d \times D$ (d de l'ordre de 50 ou 100).

On construit et on fait évoluer $\mathcal{N}_d$:

- à partir des points $A\mathbf{x}_i$ (dans $\mathbb{R}^d$) où $\mathbf{x}_i$ sont les points de $\mathbb{R}^D$ utilisés pour faire évoluer $\mathcal{N}_D$.

3 Algo ERP-CMA, Ensemble Random Projection - CMA

Comme une seule projection aléatoire peut être mal adaptée, on va en considérer M. M est un paramètre à définir - pas de valeur magique.

Chacune des RP$_i$ $i = 1 \dots M$ est associée à un poids w_i, initialement fixé à $1/M$.

On a donc maintenant le grand CMA-ES (comme dans RP-CMA) et M petits CMAs sur $\mathbb{R}^d$. On note $\mathcal{N}_i$ la distribution associée au petit CMA-ES de projection aléatoire A_i.

3.1 Génération des échantillons

Dans la routine de génération des points du grand CMA-ES, chaque échantillon $\mathbf{x}_i$ est maintenant obtenu :

- en tirant i dans $1 \dots M$ proportionnellement à w_i;
- en générant K points $\mathbf{z}_j$ dans $\mathbb{R}^D$ selon $\mathcal{N}_D$;
- en retournant argmax $\{Pr(A\mathbf{z}_j|\mathcal{N}_i), j = 1 \dots K\}$ (le point tel que $A\mathbf{z}_j$ soit le plus probable selon $\mathcal{N}_i$).

3.2 Mise à jour des poids w_i

Le poids w_i est mis à jour en considérant

$$w_i \propto Pr((\mathbf{x}, \mathbf{x}') \; s.t. \; Pr(A_i\mathbf{x}|\mathcal{N}_i) > Pr(A_i\mathbf{x}'|\mathcal{N}_i)|\mathcal{F}(\mathbf{x}) > \mathcal{F}(\mathbf{x}'))$$

la fraction des paires de points $(\mathbf{x}, \mathbf{x}')$ de $\mathbb{R}^D$ tels que $A_i\mathbf{x}$ est plus probable que $A_i\mathbf{x}'$ selon $\mathcal{N}_i$ conditionnellement au fait que $\mathcal{F}(\mathbf{x}) > \mathcal{F}(\mathbf{x}')$.

(Mann Whitney Wilcoxon).

Intuition: la projection i est d'autant meilleure que sa distribution ordonne les points comme $\mathcal{F}$.

Ici, on raffinera plus tard: par exemple

- Il faut tuer les RP avec un w trop mauvais; et en générer d'autres;
- Il faut aussi mesurer la diversité des RP entre elles.

4 Paramètres

- Dimension d: dépend de la fonction objectif
- K: force de la sélection d'un petit CMA (devrait dépendre de w_i - à voir).
- M: nombre de petits CMA.