

---

# Journal

---

Hugo Simon  
hugo.simon@telecom-paris.fr

## 1 Introduction

Let  $\mathcal{X} = \{x_1, \dots, x_n\}$  be a set of  $n$  datapoints. Let  $\Theta$  be a space of parameters, and  $\theta$  an element of  $\Theta$ . We consider cost functions of the form:

$$L(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$$

Let  $\mathcal{S} = \{x_{s_1}, \dots, x_{s_m}\}$  be a subset of  $\mathcal{X}$  (possibly with repetitions). To each element  $x \in \mathcal{S}$ , associate a weight  $\omega(x) \in \mathbb{R}^+$ . Define the estimated cost associated to the weighted subset  $\mathcal{S}$  as:

$$\hat{L}(\theta) = \sum_{x \in \mathcal{S}} \omega(x) f(x, \theta).$$

**Definition 1.1** (Coreset). Let  $\varepsilon \in ]0, 1[$ . The weighted subset  $\mathcal{S}$  is a  $\varepsilon$ -coreset for  $L$  if, for any parameter  $\theta$ , the estimated cost is equal to the exact cost up to a relative error:

$$\forall \theta \in \Theta \quad \left| \frac{\hat{L}(\theta)}{L(\theta)} - 1 \right| \leq \varepsilon$$

An important consequence of the coreset property is the following

$$(1 - \varepsilon)L(\theta^{\text{opt}}) \leq (1 - \varepsilon)L(\hat{\theta}^{\text{opt}}) \leq \hat{L}(\hat{\theta}^{\text{opt}}) \leq \hat{L}(\theta^{\text{opt}}) \leq (1 + \varepsilon)L(\theta^{\text{opt}})$$

## 2 Variance argument

### 2.1 Multinomial case

Multinomial case  $\mathcal{S} \sim \mathcal{M}(m, q)$  i.e.  $m$  independent categorical sampling where  $\mathbb{P}(x_i) = q(x_i)$

$$\text{Var}[\hat{L}(\theta)] = \frac{1}{m} \text{Var} \left[ \frac{f_{\theta}(x)}{q(x)} \right] = \frac{1}{m} \sum_{x \in \mathcal{X}} \frac{f_{\theta}(x)^2}{q(x)} - \frac{1}{m} L(\theta)^2$$

For any query  $\theta \in \Theta$ , the variance is minimized (reduced to 0) by

$$q_{\theta}(x) := \frac{f_{\theta}(x)}{\sum_{x' \in \mathcal{X}} f_{\theta}(x')}$$

### 2.2 DPP case

DPP case where  $\mathcal{S} \sim \mathcal{DPP}(K)$ ,  $\pi_i := K_{ii}$ . We have

$$\text{Var}[\hat{L}(\theta)] = \sum_{i,j} \mathbb{E}[\varepsilon_i \varepsilon_j] \frac{f_{\theta}(x_i) f_{\theta}(x_j)}{\pi_i \pi_j} - L(\theta)^2 \quad \text{with} \quad \mathbb{E}[\varepsilon_i \varepsilon_j] = \begin{cases} \det(K_{ij}) = \pi_i \pi_j - K_{ij}^2, & \text{if } i \neq j \\ \mathbb{E}[\varepsilon_i] = \pi_i, & \text{if } i = j \end{cases}$$

Introducing  $\Pi = \text{diag}(\pi)$  and  $\tilde{K} = \Pi^{-1} K^{\odot 2} \Pi^{-1}$ , we can rewrite

$$\mathbb{V}\text{ar}[\hat{L}(\theta)] = \sum_i \left( \frac{1}{\pi_i} - 1 \right) f_\theta(x_i)^2 - \sum_{i \neq j} \frac{K_{ij}^2}{\pi_i \pi_j} f_\theta(x_i) f_\theta(x_j) = f_\theta^\top (\Pi^{-1} - \tilde{K}) f_\theta$$

For a Bernoulli process where  $\mathbb{P}(x_i \in \mathcal{S}) = \pi_i$  independently,  $K = \Pi$  then  $\tilde{K} = I$ . The DPP variance beats uniformly the Bernoulli process variance if  $\tilde{K}$  dominates the identity i.e.

$$\forall f_\theta, \mathbb{V}\text{ar}[\hat{L}_K(\theta)] < \mathbb{V}\text{ar}[\hat{L}_\Pi(\theta)] \iff \tilde{K} \succ I$$

But  $\tilde{K}$  is a symmetric positive definite matrix and by Hadamard inequality  $\det(\tilde{K}) \leq \prod_i \tilde{K}_{ii} = 1$ . Therefore at least one of its eigenvalue is lower than 1, hence  $\tilde{K} \not\succ I$ .

### 3 Sensitivity

**Definition 3.1** (Sensitivity). The sensitivity  $\sigma_i$  of a datapoint  $x_i$  and the total sensitivity  $\mathfrak{S}$  of  $\mathcal{X}$  are

$$\begin{cases} \sigma_i = \sup_{\theta \in \Theta} q_\theta(x_i) = \sup_{\theta \in \Theta} \frac{f_\theta(x_i)}{L(\theta)} & \in [0, 1] \\ \mathfrak{S} = \sum_{i=1}^n \sigma_i \end{cases}$$

Let  $s$  be an upper bound on sensitivity  $\sigma$  i.e.  $\forall i, s_i \geq \sigma_i$ , and  $S := \sum_{i=1}^n s_i$ . Furthermore, let sample  $S \sim \mathcal{M}(m, s/S)$ , the multinomial sampling case. Define  $g_\theta(x_i) := \frac{q_\theta(x_i)}{s_i} \in [0, 1]$

By Hoeffding's inequality, we thus have for any  $\theta \in \Theta$  and  $\varepsilon' > 0$

$$\mathbb{P} \left[ \left| \mathbb{E}[g_\theta(x)] - \frac{1}{m} \sum_{x \in S} g_\theta(x) \right| > \varepsilon' \right] \leq 2 \exp(-2m\varepsilon'^2).$$

By definition,  $\mathbb{E}[g_\theta(x)] = \frac{1}{S}$  and  $\frac{1}{m} \sum_{x \in C} g_\theta(x) = \frac{\text{cost}(C, Q)}{S \text{cost}(\mathcal{X}, Q)}$ . As such, for any  $Q \in \mathcal{Q}$

$$\mathbb{P}[|\text{cost}(\mathcal{X}, Q) - \text{cost}(C, Q)| > \varepsilon' S \text{cost}(\mathcal{X}, Q)] \leq 2 \exp(-2m\varepsilon'^2)$$

Hence, the set  $C$  satisfies the coresnet property in (2.2) for any single query  $Q \in \mathcal{Q}$  and  $\varepsilon > 0$  with probability at least  $1 - \delta$ , if we choose

$$m \geq \frac{S^2}{2\varepsilon^2} \log \frac{2}{\delta}$$

### 4 SGD Paper