Journal

Hugo Simon
hugo.simon@telecom-paris.fr

1 Introduction

Let $\mathcal{X} = \{x_i \mid i \in [\![1, n]\!]\}$ be a set of n datapoints. Let Θ be a space of parameters, and θ an element of Θ. We consider cost functions of the form:

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

Let $\mathcal{S} = \{x_i \mid i \in [\![1, m]\!]\}$ be a submultiset (possibly with repetitions) of $\mathcal{X}$. To each element $x \in \mathcal{S}$, associate a weight $\omega(x) \in \mathbb{R}^+$. Define the estimated cost associated to the weighted submultiset $\mathcal{S}$ as:

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

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

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

An important consequence of the coreset property is the following

$$(1-\varepsilon) L\left(\theta^{\text{opt}}\right) \leq (1-\varepsilon) L\left(\hat{\theta}^{\text{opt}}\right) \leq \hat{L}\left(\hat{\theta}^{\text{opt}}\right) \leq \hat{L}\left(\theta^{\text{opt}}\right) \leq (1+\varepsilon) L\left(\theta^{\text{opt}}\right) \tag{2}$$

See Bachem et al. 2017.

2 Variance arguments

2.1 Multinomial case

In the multinomial case, we have $\mathcal{S} \sim \mathcal{M}(m, q)$ i.e. m i.i.d. categorical sampling with $\mathbb{P}(x_i) = q(x_i)$. Then an unbiased estimator of L is

$$\hat{L}_{\text{iid}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{m q(x_i)}$$

Its variance is

$$\mathbb{V}\text{ar}_{\text{iid}}(\theta) := \frac{1}{m} \mathbb{V}\text{ar}\left[\frac{f_\theta(x_i)}{q(x_i)}\right] = \frac{1}{m} \sum_{x \in \mathcal{X}} \frac{f_\theta(x)^2}{q(x)} - \frac{1}{m} L(\theta)^2 = f_\theta^\top (\frac{Q^{-1}}{m} - \frac{\boldsymbol{J}}{m}) f_\theta \tag{3}$$

where $Q = \text{diag}(q)$ and $\boldsymbol{J} = \boldsymbol{j}\boldsymbol{j}^\top$ the matrix full of ones.

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

$$q_\theta(x) := \frac{f_\theta(x)}{L(\theta)}$$

2.2 DPP case

In the DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K),\ \pi_i := K_{ii}$. Then an unbiased estimator of L is

$$\hat{L}_{\text{DPP}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{\pi_i}$$

Its variance can be computed using ε_i as the counting variable for x_i:

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

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

$$\mathbb{V}\text{ar}_{\text{DPP}}(\theta) = \sum_i \left(\frac{1}{\pi_i} - 1\right) f_\theta(x_i)^2 - \sum_{i \neq j} \frac{K_{ij}^2}{\pi_i \pi_j} f_\theta(x_i) f_\theta(x_j) = f_\theta^\top (\Pi^{-1} - \tilde{K}) f_\theta \tag{4}$$

For a Bernoulli process where $\mathbb{P}(x_i \in \mathcal{S}) = \pi_i$ independently, the DPP kernel reduces to its diagonal i.e. $K = \Pi$ then $\tilde{K} = I$. We denote its variance $\mathbb{V}\text{ar}_{\text{diag}}$.

2.3 m-DPP case

In the m-DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K) \mid |S| = m$, and the marginals $b_i := \mathbb{E}\left[\varepsilon_i\right]$ have an analytic form. Then an unbiased estimator of L is

$$\hat{L}_{\text{mDPP}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{b_i}$$

Note that we could also be interested in a biaised cost function such as the diversified risk introduced by Zhang et al. 2017

$$\tilde{L}(\theta) = \frac{1}{m} \mathbb{E}_{x \sim \text{mDPP}}[f_\theta(x)] = \frac{1}{m} \sum_{x_i \in \mathcal{X}} b_i f_\theta\left(x_i\right)$$

Then an unbiased estimator of $\tilde{L}$ is

$$\hat{\tilde{L}}_{\text{mDPP}}(\theta) = \frac{1}{m} \sum_{x_i \in \mathcal{S}} f_\theta(x_i)$$

We can switch between L and $\tilde{L}$, substituting $f_\theta(x_i)$ by $\frac{b_i f_\theta(x_i)}{m}$.

Returning to the estimation of L, we are interested in the variance of $\hat{L}_{\text{mDPP}}$ which is

$$\mathbb{V}\text{ar}_{\text{mDPP}}(\theta) = \sum_i \left(\frac{1}{b_i} - 1\right) f_\theta(x_i)^2 + \sum_{i \neq j} C_{ij} f_\theta(x_i) f_\theta(x_j) \tag{5}$$

where $C_{ij} = \frac{\mathbb{E}[(\varepsilon_i - b_i)(\varepsilon_j - b_j)]}{\mathbb{E}[\varepsilon_i]\mathbb{E}[\varepsilon_j]} = \frac{\mathbb{E}[\varepsilon_i \varepsilon_j]}{b_i b_j} - 1$

Observe that if the m-DPP kernel is reduced to its diagonal ($C_{ij} = 0$), we recover $\mathbb{V}\text{ar}_{\text{diag}}$, the variance of a Bernoulli process with same marginals ($\pi_i = b_i$), though here the number of elements sampled is fixed to m.

In order to benefit from some variance reduction, one should want $\forall i \neq j$, $C_{ij} f_\theta(x_i) f_\theta(x_j) < 0$ for a given m-DPP. Zhang et al. 2017 discuss that intuitively, if the m-DPP kernel rely on some similarity measure and that f is smooth for it, then 2 similar points should have both negative correlation ($C_{ij} < 0$) and their value have positive scalar product ($f_\theta(x_i) f_\theta(x_j) > 0$). Reversely, it is argued that 2 dissimilar points should have negative correlation, and their value show "no tendency to align" hinting $f_\theta(x_i) f_\theta(x_j) < 0$. We could more conservatively consider that the induced variance change, wether positive or negative, would in either case be small, as for DPP and m-DPP, 2 dissimilar points tend toward independance.

2.4 Variance comparaison

In order to compare processes with same marginals, we set $\Pi = mQ$. Then $\mathbb{V}\mathrm{ar}_{\text{iid}}$, $\mathbb{V}\mathrm{ar}_{\text{diag}}$ and $\mathbb{V}\mathrm{ar}_{\text{DPP}}$ are quadratic forms of f_θ associated with respective matrices

$$\begin{cases} \mathbb{V}\mathrm{ar}_{\text{iid}} \equiv \Pi^{-1} - \frac{\boldsymbol{J}}{m} \\ \mathbb{V}\mathrm{ar}_{\text{diag}} \equiv \Pi^{-1} - I \\ \mathbb{V}\mathrm{ar}_{\text{DPP}} \equiv \Pi^{-1} - \tilde{K} \end{cases}$$

2.4.1 DPP versus diag?

The DPP variance strictly beats uniformly the Bernoulli process variance if $\tilde{K}$ strictly dominates identity i.e.

$$\forall f_\theta,\ \mathbb{V}\mathrm{ar}_{\text{DPP}} < \mathbb{V}\mathrm{ar}_{\text{diag}} \iff \tilde{K} \succ I \tag{6}$$

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 $\tilde{K} \not\succ I$.

2.4.2 DPP versus i.i.d.?

The DPP variance strictly beats uniformly the multinomial variance if $\tilde{K}$ strictly dominates $\frac{\boldsymbol{J}}{m}$ i.e.

$$\forall f_\theta,\ \mathbb{V}\mathrm{ar}_{\text{DPP}} < \mathbb{V}\mathrm{ar}_{\text{iid}} \iff \tilde{K} \succ \frac{\boldsymbol{J}}{m} \tag{7}$$

K being positive of rank $r \in [\![0, n]\!]$, it exists $V = (V_i \mid i \in [\![1, n]\!]) \in \mathcal{M}_{r,n}$ such that $K = V^\top V$.

For any vector $v \in \mathbb{R}^r$, Copenhaver et al. 2013 define its diagram vector

$$\tilde{v} := \frac{1}{\sqrt{r-1}}((v_k^2 - v_l^2, \sqrt{2r} v_k v_l) \mid k < l)^\top \in \mathbb{R}^{r(r-1)}$$

concatenating all the differences of squares and products.

Then introducing $\tilde{V} = \left(\tilde{V}_i \mid i \in [\![1, n]\!]\right)$ allows us to rewrite $\tilde{K}_{ij} = \frac{\boldsymbol{J}}{r} + \frac{r-1}{r} \tilde{V}^\top \tilde{V}$. Therefore, for a projective DPP with rank $r = m$, we have $\tilde{K} - \frac{\boldsymbol{J}}{m} = \frac{m-1}{m} \tilde{V}^\top \tilde{V} \succeq 0$ ($\succ$ if $m > 1$). That is to say, for every multinomial sampling, we have a DPP which always beats it uniformly.

justify the use of a projective DPP. requires $r \geq m$ but we always have $m \leq r$, therefore $r = m$

3 State of the art

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

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

3.1 Main proof

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

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

$$\mathbb{P}\left[\left|\mathbb{E}\left[g_\theta(x)\right] - \frac{1}{m} \sum_{x \in \mathcal{S}} g_\theta(x)\right| > \varepsilon'\right] \leq 2 \exp\left(-2m\varepsilon'^2\right) \tag{8}$$

and by definition, $\mathbb{E}\left[g_\theta(x)\right] = \frac{1}{S}$ and $\frac{1}{m} \sum_{x \in \mathcal{S}} g_\theta(x) = \frac{\hat{L}_{\text{iid}}(\theta)}{S L(\theta)}$, thus

$$\mathbb{P}\left[|L(\theta) - \hat{L}_{\text{iid}}(\theta)| > \varepsilon' S L(\theta)\right] \leq 2 \exp\left(-2m\varepsilon'^2\right)$$

Hence, $\mathcal{S}$ satisfies the ε-coreset property 1.1 for any single query $\theta \in \Theta$ with probability at least $1 - \delta$, if we choose

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

3.2 Extension to all queries

See **Uniform guarantee for all queries** in Bachem et al. 2017. Introducing the pseudo-dimension d', it gives

$$m \geq \mathcal{O}(\frac{S^2}{2\varepsilon^2}(d' + \log\frac{2}{\delta})) \tag{9}$$

See **Theorem 5.5** of Braverman et al. 2016 for an improved bound (when f is positive?).

$$m \geq \mathcal{O}(\frac{S}{2\varepsilon^2}(d' \log S + \log\frac{2}{\delta})) \tag{10}$$

4 Improving concentration with DPP

Assume better variance with DPP, can we improve concentration?

- Can we use the $\sqrt{N^{1+\frac{1}{d}}}$ rate from the SGD paper?
- Concentration inequality for a sum of **dependant** variables?

Theorem 3.4. from Pemantle et al. 2011: Let $\mathbb{P}$ be a k-homogeneous probability measure on $\mathcal{B}_n$ satisfying the Stochastic Covering Property (SCP). Let f be a 1-Lipschitz function on $\mathcal{B}_n$. Then

$$\mathbb{P}(|f - \mathbb{E}f| \geq a) \leq 2\exp\left(-\frac{a^2}{8k}\right)$$

Bennett inequality: Let be $(X_i)_{i\in[\![1,n]\!]}$ independant and centered real-valued random variables, and $\sigma^2 = \frac{1}{n}\sum_i \mathbb{V}\mathrm{ar}[X_i]$, then for any $t > 0$

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

5 Discrete OPE

Can we bypass the Kernel Density Estimate (KDE) in SGD paper by using discrete OPE? See Gautschi 2004.

6 Holydays questions

- Variance for formula for k-DPP, in Zhang et al. 2017.
- How $\tilde{K}$ eigenspaces look like ? When $n \to \infty$?
  - How does it compare to Bardenet et al. 2020 ?
  - If f is given, can I find a K for which f is in "good" eigenspaces (eigenvalue ≥ 1).
- Defining discrete OPE, because discretized continuous OPE is probably not a DPP. See Gautschi Orthogonal Polynomials, 2004.
  - For making links with SGD paper Bardenet et al. 2021
  - Look at the limit e.g. for Jacobi ensembles.
- Take a Bernoulli and beat it with a DPP.
- Focus on metric we could have advantages on, e.g. look how variance decay with coreset size.
- Better with direct applications e.g. on k-means or linear regression

References

Bachem, Olivier et al. (2017). *Practical Coreset Constructions for Machine Learning*. DOI: 10.48550/ARXIV.1703.06476. URL: https://arxiv.org/abs/1703.06476.

Bardenet, Rémi et al. (2020). "Monte Carlo with Determinantal Point Processes". In: *Annals of Applied Probability*. URL: https://hal.archives-ouvertes.fr/hal-01311263.

Bardenet, Rémi et al. (2021). *Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD*. DOI: 10.48550/ARXIV.2112.06007. URL: https://arxiv.org/abs/2112.06007.

Braverman, Vladimir et al. (2016). *New Frameworks for Offline and Streaming Coreset Constructions*. DOI: 10.48550/ARXIV.1612.00889. URL: https://arxiv.org/abs/1612.00889.

Copenhaver, Martin S. et al. (2013). *Diagram vectors and Tight Frame Scaling in Finite Dimensions*. DOI: 10.48550/ARXIV.1303.1159. URL: https://arxiv.org/abs/1303.1159.

Gautschi, Walter (2004). *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Clarendon Press. ISBN: 0198506724. eprint: https://www.cs.purdue.edu/homes/wxg/OPmatlab.pdf.

Pemantle, Robin et al. (2011). *Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures*. DOI: 10.48550/ARXIV.1108.0687. URL: https://arxiv.org/abs/1108.0687.

Zhang, Cheng et al. (2017). *Determinantal Point Processes for Mini-Batch Diversification*. DOI: 10.48550/ARXIV.1705.00607. URL: https://arxiv.org/abs/1705.00607.