Data Mining: Exam

December 3rd, 2020

1 Problem 1: Reducing the variance of the Monte Carlo estimators [50 points]

Consider a random variable of the form:

$$X = f(\mathbf{g}^\top \mathbf{z}), \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^d$ is a fixed deterministic vector and $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$. Assume that f has a Taylor series expansion with positive coefficients. A standard unbiased Monte Carlo estimator of $\mathbb{E}[X]$ is of the form:

$$\widehat{X} = \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{g}_i^\top \mathbf{z}), \tag{2}$$

where $\mathbf{g}_1, ..., \mathbf{g}_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Propose an improved (still unbiased) version of that estimator that will rely on the same set of random projections $\mathbf{g}_1, ..., \mathbf{g}_m$ and which, under as weak additional assumptions about f as possible (maybe even no additional assumptions), is characterized by a strictly lower variance. Prove it and show that your estimator is still unbiased. **Note:** As a corollary, you might obtain an improved estimator of the softmax-kernel for the approximate attention.

2 Problem 2: SVM with dynamically changing labels [30 points]

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a dataset of size N and with labels assigned to its datapoints dynamically changing over time $t = 0, 1, 2, 3, ...T$. Denote by ω_t a vector defining the orientation of the optimal SVM-hyperplane at a given point t (you can assume that the labels are changing in such a way that at every given time t dataset $\mathcal{X}$ is linearly separable, thus in particular ω_t exists). Show that if $T \geq N$ then there exists $0 \leq t \leq N$ such that ω_t is a linear combination of ω_i for $i \in \{0, ..., N\} \backslash \{t\}$.

3 Problem 3: Gradient exploding/vanishing problems in neural network training [20 points]

Describe the problem of exploding/vanishing gradients in neural network training and explain why it affects deep neural network systems. Describe methods for handling this phenomenon.

4 Problem 4: ϵ-close estimators of kernels [40 points]

Let $\mathrm{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a kernel function. We say that its two estimators $\widehat{X}_1, \widehat{X}_2$ (in fixed two points: $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$) are ϵ-close if the following holds:

$$|\mathrm{MSE}(\widehat{X}_1) - \mathrm{MSE}(\widehat{X}_2)| \leq \epsilon. \tag{3}$$

Assume that the *total variation distance* between $\widehat{X}_1$ and $\widehat{X}_2$ is at most δ, i.e. for every measurable set $\mathcal{S}$ the following holds:

$$\|\mu_{\widehat{X}_1}(\mathcal{S}) - \mu_{\widehat{X}_2}(\mathcal{S})\| \leq \delta, \tag{4}$$

where μ_X stands for the probabilistic measure corresponding to random variable X. Assume furthermore that each $\widehat{X}_i$ is of the form $\widehat{X}_i = \frac{1}{m} f(\mathbf{G}_i \mathbf{v})^\top f(\mathbf{G}_i \mathbf{w})$, where $\mathbf{G}_i \in \mathbb{R}^{m \times d}$ and f is bounded (you can assume that $F = \sup_{x \in \mathbb{R}} |f(x)| < \infty$). Find the value of ϵ such that $\widehat{X}_1$ and $\widehat{X}_2$ are ϵ-close. **Note:** This analysis can be applied to show that estimators of Gaussian kernels relying on $\Theta(d \log(d))$ random Givens rotations are asymptotically characterized by strictly lower variance than their unstructured counterparts (even though you are not asked to show it here).