Exercise XIII, Sublinear Algorithms for Big Data Analysis

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students, and solve as many problems as you can. Problems marked (*) are more difficult, but also more rewarding. These problems have been taken from various sources on the Internet, too numerous to cite individually.

1 Show that a uniform sample of $O(1/(\epsilon^2\mu))$ points from a dataset P gives a $(1\pm\epsilon)$-approximation to $K(P,q)=\mu$ with high constant probability.

2 Show that a sample of $O(1/\epsilon^2)$ points where every point $p\in P$ is included independently with probability $\Omega(K(p,q)/(n\mu))$, appropriately rescaled, gives a $(1\pm\epsilon)$-approximation to $K(P,q)$ with high constant probability.

3 Suppose that you have a data structure that $(1\pm\epsilon)$-approximates $K(P,q)$, $|P|=n$, using space $S(1/\mu,1/\epsilon)$ and query time $Q(1/\mu,1/\epsilon)$ assuming that $K(P,q)\geq\mu$, and provides an unbiased estimate for all μ (i.e. not necessarily those that lower bound $K(P,q)$). Assuming that $S(1/\mu,1/\epsilon)$ and $Q(1/\mu,1/\epsilon)$ are increasing in μ, give a data structure that $(1\pm O(\epsilon))$-approximates $K(P,q)=\mu$ as long as $\mu\geq\mu^*$ for a universal lower bound μ^* using space $S(1/\mu^*,\epsilon)\text{poly}(\log 1/\mu^*)$ and query time $Q(1/\mu,\epsilon)\text{poly}(\log 1/\mu^*)$.