K-RANDOMIZATION

MAXIM ZHILYAEV AND DAVID ZEBER

Outline of the procedure

•

1. Theoretical setup

In the following we work with data in the form of bit vectors. A **bit vector** is a vector $v \in \{0,1\}^L$.

First we define the randomization procedure we will be applying.

Definition. The randomization procedure R with **lie probability** $0 < q < 1/2$ flips a bit with probability q, and leaves it as-is with probability $1-q$. In other words, for a bit $b \in \{0,1\}$,

$$R(b) = R(b; X) = (1-b) \cdot X + b \cdot (1-X) \quad \text{where } X \sim Ber(q).$$

When applied to a vector, each bit is randomized independently:

$$R(v) = R\big(v; (X_1, \dots, X_L)\big) = \big(R(v_1; X_1), \dots, R(v_L; X_L)\big) \quad \text{where } X \overset{\text{iid}}{\sim} Ber(q).$$

Remark. The randomization R reports the original bit value with probability $1-q > q$, and lies with probability q. This is equivalent to the randomized response procedure where the value is reported as-is with probability $1-f$, and with probability f the reported value is the outcome of the toss of a fair coin. In this case, $q = f/2$.

Remark. If $q = 1/2$, then $R(0) \overset{d}{=} R(1)$, and the reported value is "completely" randomly generated, i.e., independently of the original value.

Distribution of $R(v)$.

For a bit b, the randomization lies iff $R(b) \neq b$:

$$P[R(b) = s] = q^{\mathbf{1}_{\{b \neq s\}}} (1-q)^{\mathbf{1}_{\{b = s\}}}$$

Hence, for a bit vector v,

$$P[R(v) = s] = q^{\sum \mathbf{1}_{\{b_i \neq s_i\}}} (1-q)^{\sum \mathbf{1}_{\{b_i = s_i\}}} = q^{L - m(v,s)} (1-q)^{m(v,s)},$$

where $m(v,s) = |\{i : v_i = s_i\}|$. Note that this probability is maximized when $m(v,s) = L$ (the reported vector s is identical to the original vector v), and minimized when $m(v,s) = 0$. In other words, the most likely outcome of randomizing a bit vector is obtaining an identical vector.

For a collection T,

$$P[s \in R(T)] = 1 - P[s \notin R(T)] = 1 - \prod_{v \in T} P[R(v) \neq s] = 1 - \prod_{v \in T} \big[1 - q^{L-m(v,s)}(1-q)^{m(v,s)}\big].$$

2. Differential Privacy

Consider a collection T of bit vectors, and write $T_v = T \backslash \{v\}$. The randomization procedure R is ϵ-differentially private if

$$\log\left(\frac{P[R(T) \in S]}{P[R(T_v) \in S]}\right) \leq \epsilon$$

for any set of bit vectors S.