The AGM Query Bound

Scott Kovach
https://cutfree.net

December 31, 2023

1 Introduction

In this note we'll give some background on natural join queries and a self-contained proof of the "AGM" output size bound (Atserias et al., 2013).

We start with background on relations and join queries and state the goal: to bound the worst-case query size given size bounds on the input relations. At a high level, the proof is short and simple (see here), but it carefully relies on some facts about entropy. We then prove these facts. We'll loosely follow the *structured proof* style described by Lamport (2012) [1].

2 Background

This note is about a basic type of relational join query. The problem is to determine the set of tuples that satisfy a set of constraints, where each constraint is expressed as a relation. Formally, we have a *universe* set U and a set of *column attributes* A. When $\alpha \subseteq A$, U^α denotes the set of *tuples of shape* α. [2] That is, a tuple assigns a value of U to each value in α. A *relation* R *of shape* α is a subset of U^α. We use $\sigma(R)$ as notation for the shape of R.

Example: Suppose we maintain a database of wildlife observations. We might use the attributes

$$A = \{\text{user-id},\ \text{username},\ \text{profile-photo},\ \text{date},\ \text{image-data},\ \text{location}\}.$$

A *users* table would have shape {user-id, username, profile-photo}, while the primary table, *observations*, might have shape {user-id, date, image-data, location}.

[1] using his package `pf2` available at http://lamport.azurewebsites.net/latex/latex.html

[2] Often a *type* is assigned to each attribute, and tuples must assign an appropriate value to each. We work with one universe set U just to simplify some definitions. We think of U as the disjoint union of whatever types are necessary, and the main point of this note does not depend on the nature of U at all.

Example: An example tuple in R_{users}:

$$\left\{ \text{user-id} \mapsto 11,\ \text{username} \mapsto \texttt{"Emmy"},\ \text{profile-photo} \mapsto \right\}$$

Since $t \in U^\beta$ is a function, we can always *restrict* it to a smaller domain $\alpha \subseteq \beta$ to obtain a tuple over α, which we denote $t|_\alpha$. For instance if $t \in R_{obs}$ then $t|_{\{\text{date}\}}$ would be a tuple containing only the "date" field of t.

Definition [natural join query] A problem instance consists of a sequence of finite relations, $Q = [R_1, R_2, \ldots, R_k]$ (not necessarily of the same shape). Let $A := \bigcup_{R \in Q} \sigma(R)$ (we write $R \in Q$ if $R = R_i$ for some i). The *schema* of Q is the function $\sigma(Q) : \{1, 2, \ldots, k\} \to 2^A$ given by $\sigma(Q)(i) := \sigma(R_i)$. [3] The solution to Q is the relation $[\![Q]\!] \subseteq U^A$ defined by

$$[\![Q]\!] := \{t \in U^A \mid \forall R \in Q, t|_{\sigma(R)} \in R\}.$$

The purpose of this note is to prove an asymptotic upper bound on the size of this set given in terms of the sizes of each input relation. We can stipulate a family of problem instances by first fixing the schema $\sigma : \{1, 2, \ldots, k\} \to 2^A$. We can then characterize the size of a problem instance using a vector b of sizes, where $|Q| = b$ means $|R_i| = b_i$. We are looking for the worst case size,

$$\max_{\sigma(Q)=\sigma, |Q|=b} |[\![Q]\!]|,$$

which is the largest possible output size over all input relations with fixed schema and fixed size. We give a couple of examples for intuition before stating and proving the bound in Section 4.

Example: Suppose $R_1 \subseteq U^{\{a,b\}}$ and $R_2 \subseteq U^{\{b,c\}}$ and $|R_1| = |R_2| = n$. How big could the join of these relations be? First of all, we are interested in a certain subset of tuples from $U^{\{a,b,c\}}$, and each tuple consists of a value for each of a, b, and c. There are n tuples in R_1, so at most n distinct values can appear for a (and similarly for b and c). Thus n^3 is an upper bound. We can clearly do better: any tuple much correspond with one tuple of R_1 and one tuple of R_2, and there are at most $n \cdot n$ ways of choosing one from each, so n^2 is a better upper bound. Moreover, if we choose $R_1 := \{(1,1), (2,1), \ldots, (n,1)\}$ and $R_2 := \{(1,1), (1,2), \ldots, (1,n)\}$ then the join consists of n^2 tuples $\{(x,1,y) \mid x, y \in [n]\}$, so this upper bound is tight (we can conflate a tuple in $U^{\{a,b,c\}}$ with a tuple $(x,y,z) \in U^3$ as long as we are careful to keep track of which value is which, for instance by ordering our attributes).

[3] The schema is equivalently a subset of $A \times Q$, and some other sources refer to this as the *query hypergraph*.

Example: Suppose that in addition to R_1, R_2, we have $R_3 \subseteq U^{\{a,c\}}$ and $|R_3| = n$. We have not added any new attributes; the new relation can only *further constrain* the size of the output. Thus n^2 is still an upper bound. However, this is not the best upper bound (check to see why the example values for R_1 and R_2 that were just given can no longer generate an output of size n^2). For reasons we'll see below, in fact $n^{3/2}$ is an upper bound for this query, which is known as the *triangle query*.

Example: Suppose we have the same query, but we generalize the bounds: $|R_1| = b_1, |R_2| = b_2, |R_3| = b_3$. Then $(b_1 b_2 b_3)^{1/2}$ is an upper bound, but in some cases (for instance if $b_1 b_2 \ll b_3$) this isn't the *best* upper bound.

We'll see that the theorem we prove handles *all* such problems, but the answer it gives is less simple than other worst-case bounds you may have seen. The reason is that it includes an additional parameter that lets us choose "how much" of a given relation to include in the bound.

3 Entropy Definitions

We use capitals X, Y, Z for discrete random variables and x, y, z for values of the corresponding type. With some abuse of notation we write $[x]$ for the probability $P(X = x)$, $[x; Y]$ for the variable Y conditioned on $X = x$, and $[x; y]$ for the conditional probability $P([x; Y] = y) = P(Y = y \mid X = x) = [xy]/[x]$. Let $\mathrm{supp}(X)$ denote the support of a random variable.

$H(X)$ is the *entropy* of X:

$$H(X) := \sum_{x \in \mathrm{supp}(X)} [x] \log [x]^{-1}$$

$H(X; Y)$ is the *conditional entropy* of Y given X, which is defined to be the expected entropy of $[x; Y]$ over X:

$$H(X; Y) := \sum_{x \in \mathrm{supp}(X)} [x] \sum_{y \in \mathrm{supp}([x;Y])} [x; y] \log [x; y]^{-1}$$

4 AGM Bound

We reproduce the proof based on entropy from the appendix of Ngo et al. (2014). In Section 5, we prove the basic properties of entropy we use in the proof.

The key idea is pick a distribution over $[\![Q]\!]$, then bound the entropy of this distribution in terms of the entropy of various marginalizations corresponding to the shape of each $R \in Q$. Finally, we relate these entropy estimates to the size of each relation and the output.

Let $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \mid x \geq 0\}$. To state the bound, we need to introduce an extra "weight" parameter, $\lambda : Q \to \mathbb{R}_{\geq 0}$. Each relation has a shape $\sigma(R)$, which is the set of attributes over which it is defined. We number the attributes, so

$A \simeq \{1, 2, \ldots, n\}$, and when $i \in \sigma(R)$ we say that R *covers* i. The requirement on λ we impose is that each attribute is "sufficiently covered"; the sum of the weights of all relations covering i must be at least one:

$$1 \leq \sum_{R:\ i \in \sigma(R)} \lambda_R \tag{1}$$

Theorem [AGM bound] Given a query problem Q and $\lambda : Q \to \mathbb{R}_{\geq 0}$ satisfying condition (1),

$$|Q| \leq \prod_{R \in Q} |R|^{\lambda_R}.$$

Let X be the uniform random variable with support $[\![Q]\!] \subseteq U^A$. Then $X|_{\sigma(R)} = \{X_i\}_{i \in \sigma(R)}$ is the random variable formed by projecting X onto $\sigma(R)$.

⟨1⟩1. For any $\alpha \subseteq A$, $H(X|_\alpha) = \sum_{i \in \alpha} H(\{X_j\}_{j \in \alpha, j < i}; X_i)$

PROOF: Section 5.2.

⟨1⟩2. For any Z, if $\alpha \subseteq \beta$, then $H(X|_\beta; Z) \leq H(X|_\alpha; Z)$

⟨2⟩1. For arbitrary variables, $H(X, Y; Z) \leq H(X; Z)$

PROOF: Section 5.3.

PROOF: Repeated application of ⟨2⟩1.

⟨1⟩3. For all $R \in Q$, $H(X|_{\sigma(R)}) \leq \log |R|$

⟨2⟩1. For any random variable Z, $H(Z) \leq \log |\mathrm{supp}(Z)|$

PROOF: Section 5.1

⟨2⟩2. $\mathrm{supp}(X|_{\sigma(R)}) \subseteq R$.

PROOF: By definition of the natural join problem, and the assumption that $\mathrm{supp}(X) = [\![Q]\!]$.

PROOF: By ⟨2⟩1 and ⟨2⟩2

⟨1⟩4. $\log |[\![Q]\!]| \leq \sum_R \lambda_R \log |R|$

$$
\begin{aligned}
\textsc{Proof}: \ \log|[\![Q]\!]| &= H(X) && \text{Section 5.1}\\
&= \sum_i H(\{X_j\}_{j<i}; X_i) && \langle 1\rangle 1\\
&\leq \sum_i \left(\sum_{R:\ i\in\sigma(R)} \lambda_R \right) H(\{X_j\}_{j<i}; X_i) && \text{assumption}\\
&= \sum_R \lambda_R \sum_{i\in\sigma(R)} H(\{X_j\}_{j<i}; X_i) && \text{re-order sums}\\
&\leq \sum_R \lambda_R \sum_{i\in\sigma(R)} H(\{X_j\}_{j\in\sigma(R),j<i}; X_i) && \langle 1\rangle 2\\
&= \sum_R \lambda_R H(X|_{\sigma(R)}) && \langle 1\rangle 1\\
&\leq \sum_R \lambda_R \log|R| && \langle 1\rangle 3
\end{aligned}
$$

PROOF: By exponentiating both sides of $\langle 1\rangle 4$.

5 Entropy Lemmas

For this section we only take one fact for granted; log is concave, meaning that it satisfies Jensen's inequality:

$$\sum_i [i] \log y_i \leq \log \left(\sum_i [i] y_i \right)$$

for any random variable I with support $\{1, 2, \ldots, n\}$ and $\{y_i \in \mathbb{R}_{\geq 0} \mid 1 \leq i \leq n\}$.

5.1 Maximal Entropy

Theorem: If X is a random variable taking values in a finite set of size n, then $H(X) \leq \log n$, with equality if X follows the uniform distribution.

PROOF: using Jensen's inequality:

$$H(X) := \sum_x [x] \log [x]^{-1} \leq \log \left(\sum_x [x]/[x] \right) = \log n.$$

And when uniform,

$$\sum_x [x] \log [x]^{-1} = \sum_x \frac{1}{n} \log n = \log n.$$

5.2 Chain Rule

Theorem [Chain Rule]

$$H(X;Y) = H(XY) - H(X)$$

$$\begin{aligned} H(X;Y) &:= \sum_x [x] \sum_y [x;y] \log [x;y]^{-1} \\ &= \sum_{xy} [xy] \log([x]/[xy]) \\ &= \sum_{xy} [xy] \log [xy]^{-1} - \sum_{xy} [xy] \log [x]^{-1} \\ &= \sum_{xy} [xy] \log [xy]^{-1} - \sum_{x} [x] \log [x]^{-1} \\ &= H(XY) - H(X). \end{aligned}$$

Theorem [Chain Rule*]

$$H(X_1, X_2, \ldots, X_k) = \sum_{1 \le i \le k} H(X_1, \ldots, X_{i-1}; X_i)$$

$\langle 1 \rangle 1$. $H(\{\}) = 0$

PROOF: The empty variable is deterministic, so its entropy is zero.

PROOF:

$$\begin{aligned} \sum_{1 \le i \le k} H(X_1, \ldots, X_{i-1}; X_i) &= \sum_{1 \le i \le k} (H(X_1, \ldots, X_i) - H(X_1, \ldots, X_{i-1})) && \text{chain rule} \\ &= H(X_1, X_2, \ldots, X_k) - H(\{\}) && \text{sum telescopes} \\ &= H(X_1, X_2, \ldots, X_k) && \langle 1 \rangle 1 \end{aligned}$$

5.3 Conditioning

This section proves a basic fact about entropy: the conditional entropy $H(X;Y)$ can only increase if X is marginalized. This makes intuitive sense, and the proof contains nothing more than Jensen's inequality and some tricky manipulation of conditional probabilities. This proof follows Galvin (2014).

Theorem:

$$H(X, Y; Z) \le H(X; Z)$$

PROOF: The inequality step below is Jensen's inequality, which is applicable since $\sum_y [xz; y] = 1$. The equality steps are applications of the definition of

conditional probability.

$$\begin{aligned}
H(X,Y;Z) &:= \sum_{xy}[xy]\sum_{z}[xy;z]\log[xy;z]^{-1} \\
&= \sum_{xyz}[xyz]\log[xy;z]^{-1} \\
&= \sum_{xz}[xz]\sum_{y}[xz;y]\log[xy;z]^{-1} \\
&\le \sum_{xz}[xz]\log\left(\sum_{y}[xz;y]/[xy;z]\right) && \text{log is concave} \\
&= \sum_{xz}[xz]\log\left(\sum_{y}[xy]/[xz]\right) \\
&= \sum_{xz}[xz]\log([x]/[xz]) && \text{(marginalize } y\text{)} \\
&= \sum_{x}[x]\sum_{z}[x;z]\log[x;z]^{-1} \\
&=: H(X;Z).
\end{aligned}$$

References

Albert Atserias, Martin Grohe, and Dániel Marx. 2013. Size bounds and query plans for relational joins. *SIAM J. Comput.* 42, 4 (2013), 1737–1767.

David Galvin. 2014. Three tutorial lectures on entropy and counting. arXiv:1406.7872 [math.CO]

Leslie Lamport. 2012. How to write a 21 st century proof. *Journal of fixed point theory and applications* 11 (2012), 43–63.

Hung Q Ngo, Christopher Ré, and Atri Rudra. 2014. Skew strikes back: New developments in the theory of join algorithms. *ACM SIGMOD Record* 42, 4 (2014), 5–16. https://doi.org/10.1145/2590989.2590991