August 9, 2016

Estimating Runtime

The Bottom Line

Simplifying formulas

- An admissible polynomial of degree d is $\Theta(n^d)$
- When finding Θ, ignore terms of strictly lower asymptotic class

Finding model constants

- Growth Exponent $d \simeq \log_{10} T(10n_0)/T(n_0)$
- Growth Constant $A \simeq T(n_0)/M(n_0)$

where n_0 is a specific size for which we have data, T is actual runtime data, and M is the abstract model. The concrete modelling formula is then

$$T(n) \simeq A \times M(n).$$

1 Definitions, Notation, and Representations

Define a function F to be *admissable* iff it is defined on almost all of the non-negative integers and has positive real values on almost all of the domain. That is, there is an integer n_0 such that $F(n)$ is defined and $F(n) > 0$ for all integers $n > n_0$. Note that for polynomials, this condition requires that the leading coefficient is a positive number.

Note also we are introducing the terminology "*almost all* of the non-negative integers" to mean "for all integers n that are greater than some fixed value n_0" – in other words, the condition holds for all but a finite number of possible exceptions. This terminology is easier to talk about and allows us to stop mentioning "$n > n_0$", simplifying the wording of many statements.

Lemma 1. Suppose that F and G are admissible and the quotient $F(n)/G(n)$ tends to a positive constant c as n grows large - that is,

$$\lim_{n\to\infty} \frac{F(n)}{G(n)} = c > 0.$$

Then $F(n)$ and $G(n)$ are in the same Θ equivalence class.

Proof. We must show there are positive constants C_1 and C_2 such that $C_1 \times G(n) \leq F(n)$ and $F(n) \leq C_2 \times G(n)$ for almost all n. By definition of limit, we know that for any $\epsilon > 0$, $c - \epsilon \leq \frac{F(n)}{G(n)} \leq c + \epsilon$ for almost all n. Choosing $\epsilon = C/2$, $C_1 = c - \epsilon$, and $C_2 = c + \epsilon$ we have

$$C_1 = \frac{c}{2} = c - \epsilon \leq \frac{F(n)}{G(n)}$$

and

$$\frac{F(n)}{G(n)} \leq = c + \epsilon = 3\frac{c}{2} = C_2$$

for almost all n. Multiplying by $G(n)$ yields

$$C_1 G(n) \leq F(n) \leq C_2 G(n)$$

for almost all n, the desired result. □

2 Simplifying Formulas

Look at these two functions of n:

$$F(n) = a_0 n^d + a_1 n^{d-1} + \ldots + a_n$$
$$G(n) = n^d$$

(where we assume the leading coefficient $a_0 > 0$). $F(n)$ is the general form of an admissible polynomial of degree d, whereas $G(n)$ is the much simpler form of the highest power term.

Lemma 2. With the definitions above, $F(n)$ and $G(n)$ are in the same Θ equivalence class.

proof. First note the calculation

$$\begin{aligned}\frac{F(n)}{G(n)} &= \frac{a_0 n^d}{n^d} + \frac{a_1 n^{d-1}}{n^d} + \ldots + \frac{a_{d-1} n}{n^d} + \frac{a_d}{n^d} \\ &= a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots + \frac{a_{d-1}}{n^{d-1}} + \frac{a_d}{n^d}\end{aligned}$$

from which it is apparent that $\frac{F(n)}{G(n)}$ tends to a_0 as n becomes large. Since F and G are admissible, the result follows from Lemma 1. □

In the light of Lemma 2, n^d is our choice representative for the Θ class of any polynomial of degree d.

Along these same lines, suppose we have a function

$$H(n) = h(n) + \phi(n)$$

where $\phi(n)$ has the property that $\frac{\phi(n)}{h(n)} \to 0$ as $n \to \infty$.

Lemma 3. $H(n)$ and $h(n)$ are in the same Θ equivalence class.

Proof. Use the same idea as the proof of Lemma 2:

$$\begin{aligned}\frac{H(n)}{h(n)} &= \frac{h(n)}{h(n)} + \frac{\phi(n)}{h(n)} \\ &= 1 + \frac{\phi(n)}{h(n)} \\ &\to 1 + 0 \text{ as } n \to \infty \\ &= 1\end{aligned}$$

□

Lemmas 2 and 3 can be loosely paraphrased as follows:

(1) (Polynomial rule) A polynomial of degree d with positive leading coefficient is $\Theta(n^d)$.
(2) (Lower order cancellation rule) When finding the Θ class of a function, terms with strictly lower asymptotic order can be ignored.

Example applications of these two simplifying rules:

$$\begin{aligned} n(n+1)/2 &= \Theta(n^2) \\ n^2 + \log n &= \Theta(n^2) \\ n + \log n &= \Theta(n) \\ 5000n^2 + 2300\sqrt{n} &= \Theta(n^2)\end{aligned}$$

3 Estimating the growth exponent from Data

In many cases an exponent associated with the asymptotic class of an algorithm can be estimated from data. Start by assuming that the algorithm runtime has Θ class one of these forms (aka "abstract models"):

$$\begin{aligned} &An^d + B\phi(n) && \text{[Model 1]} \\ &An^d \log n + B\phi(n) && \text{[Model 2]}\end{aligned}$$

where $A > 0$ and $\phi(n)$ is dominated by the first term: $\frac{\phi(n)}{n^d} \to 0$ as $n \to \infty$ [Model 1] or $\frac{\phi(n)}{n^d \log n} \to 0$ as $n \to \infty$ [Model 2].

Lemma 4. The abstract models have Θ class as follows:

$$\begin{aligned} An^d + B\phi(n) &= \Theta(n^d) && \text{[Model 1]} \\ An^d \log n + B\phi(n) &= \Theta(n^d \log n) && \text{[Model 2]} \end{aligned}$$

The proof is a direct application of Lemma 3.

Thus we can "ignore" the second term when finding the exponent d in the models. In both cases we can find the exponent d from actual runtime data.

Example 1: Model 1 and insertion_sort

Assume that the asymptotic growth of an algorithm is modelled by $F(n) = An^d$ [Model 1]. and that we have data gathered from experimentation to evaluate F at size n and again at size $10n$:

$$\begin{aligned} F(10n) &= (10n)^d \\ &= n^d 10^d \\ &= 10^d F(n) \end{aligned}$$

which shows that raising the input size by one order of magnitute increases the runtime by d orders of magnitude. For instance, when $d = 2$ (the quadratic case), increasing the size of the input by one decimal place increases the runtime by two decimal places. Another way to phrase the result is as a ratio:

$$\frac{F(10n)}{F(n)} = \frac{(10n)^d}{n^d} = 10^d$$

which can be stated succintly as

$$d = \log_{10}(\frac{F(10n)}{F(n)}).$$

If we have actual timing data $T(n)$ for an algorithm modelled by F we can use the ratio to estimate d. Consider for example the insertion_sort algorithm, and use "comps", the number of data comparisons, as a measure of runtime. We know from theory that insertion_sort is modelled by F and we wish to know the exponent d. We have collected runtime data

$$\begin{aligned} T(1000) &= 244853 \\ T(10000) &= 24991950 \end{aligned}$$

The ratio $T(10000)/T(1000)$ is

$$\begin{aligned}\frac{T(10000)}{T(1000)} &= \frac{24991950}{244853}\\ &= 102.07\ldots\\ &\simeq 100\pm\\ &= 10^2\end{aligned}$$

yielding an estimate of $d = 2$, or quadratic runtime. Your eye might have noticed this in the data itself: $T(10000)$ is about 100 times $T(1000)$.

Example 2 - Model 2 and List::Sort

The somewhat more complex Model 2 works in the same way. Assume that the asymptotic growth of an algorithm is modelled by $G(n) = An^d \log n$ [Model d]. and that we have data gathered from experimentation to evaluate G at size n and again at size $10n$:

$$\begin{aligned}\frac{G(10n)}{G(n)} &= \frac{(10n)^d \log(10n)}{n^d \log n}\\ &= \frac{n^d 10^d \log(10n)}{n^d \log n}\\ &= \frac{10^d \log(10n)}{\log n}\\ &= 10^d(\frac{\log 10 + \log n}{\log n})\\ &= 10^d(\frac{1 + \log n}{\log n})\\ &= 10^d(1 + \frac{1}{\log n})\\ &\rightarrow 10^d\end{aligned}$$

because $\frac{1}{\log n} \rightarrow 0$ as $n \rightarrow \infty$. As in the pure exponential case, this concliusion can be stated in terms of logarithms:

$$d \simeq \log_{10}(\frac{F(10n)}{F(n)}).$$

Consider the bottom-up merge_sort specifically for linked lists, implemented as List::Sort. It is known from theory that the algorithm is modelled by G, and we have collected specific timing data as follows:

$$\begin{aligned} T(10000) &= 123674 \\ T(100000) &= 1566259 \end{aligned}$$

Then:

$$\begin{aligned} \frac{T(100000)}{T(10000)} &= \frac{1566259}{123674} \\ &= 11.66\ldots \\ &\simeq 10\pm \\ &= 10^1 \end{aligned}$$

predicting $d = 1$. Note here that the data will not likely be enough to discriminate between Models 1 and 2, so we must base that choice on other considerations.

4 Estimating the growth constant

We can refine an abstract model to a "concrete" version by finding the constant A such that $A \times Model(n)$ more accurately predicts runtime. The goal is to make timing data and the concrete model match as closely as possible:

$$T(n) \simeq A \times M(n) \quad \text{for all } n$$

At this point, we are assuming one of two "abstract" models for the runtime cost of an algorithm:

$$\begin{aligned} F(n) &= n^d \\ G(n) &= n^d \log n \end{aligned}$$

and further we have estimated a value for the (integer) exponent d. Given that, we want to calculate an estimate for the constant A such that $T(n) = A \times M(n)$ for either of our models M by solving one of the evaluated equations obtained from data for A:

$$A = \frac{T(n)}{M(n)}$$

where T is timing data and M is the growth model (F or G). In fact, we get different estimates for A for each known pair $(n, T(n))$ in our collected data - a classic over-constrained system. Ideally we would use a method such as least squares (linear regression) to optimize a value for A using all of the collected runtime data. A decent substitute would be to interpolate a value using the two data points we used to estimate the exponent. Here are those calculations using the two examples already given above.

Example 1 (continued)

We have this data for insertion_sort:

$$T(1000) = 244853$$
$$T(10000) = 24991950$$

The data points give estimates of A as

$$\begin{aligned} A &= \frac{T(1000)}{F(1000)} = \frac{244853}{1000^2} \\ &= 0.2485 \end{aligned}$$

$$\begin{aligned} A &= \frac{T(10000)}{F(10000)} = \frac{24991950}{10000^2} \\ &= 0.2499 \end{aligned}$$

It is reasonable to settle for $A = 0.25$ to complete our concrete model:

$$M(n) = 0.25 \times n^2 \qquad \text{Concrete Model for insertion\_sort}$$

This model can be used to estimate runtimes for values of n where actual data is lacking. Note that the choice of the quadratic abstract model is based on theory and known to be a correct abstract model for insertion_sort.

Example 2 (continued)

We have this data collected for List::Sort:

$$T(10000) = 123674$$
$$T(100000) = 1566259$$

The data points give estimates of A as

$$\begin{aligned} A &= \frac{T(10000)}{G(10000)} = \frac{123674}{10000 \log 10000} = \frac{123674}{10000 \times 4} \\ &= 3.09185 \end{aligned}$$

$$\begin{aligned} A &= \frac{T(100000)}{G(100000)} = \frac{1566259}{100000 \log 100000} = \frac{1566259}{100000 \times 5} \\ &= 3.132518 \end{aligned}$$

It is reasonable to settle for $A = 3.1$ to complete our concrete model:

$$M(n) = 3.1 \times n \log n \qquad \text{Concrete Model for List::Sort}$$

This model can be used to estimate runtimes for values of n where actual data is lacking. Note that the choice of the linear×log abstract model is based on theory and known to be a correct abstract model for List::Sort (a version of bottom-up merge_sort).

5 Cautions and Limitations

The reader was likely surprised that using the data as in Section 3 above is unable to distinguish between the pure power model F and the model G that is a power model multiplied by a logarithm. The reason at one level is simple: the quotients $G(10n)/G(n)$ and $F(10n)/F(n)$ differ by $10^d/\log n$. The numerator 10^d is a fixed number, whereas the denominator $\log n$ grows infinitely large with n (albeit rather slowly), so the difference gets ever smaller as n grows large. Given that data inevitably has some variation due to randomness, teasing out such a diminishingly fine distinction is problematic.

Another observation the reader likely made is that we used the base 10 logarithm instead of the more common base 2 logarithm. Any base could have been used. We chose base 10 because multiplying by 10 is a visually simple process - just move the decimal point - whereas if we used base 2 (and doubled our input size instead of multiplying it by 10) the results are similar, except it is less easy visually to recognize "approximately" $2n$ than "approximately" $10n$.

Different base logarithmic functions have the same Θ class, so when discussing Θ we are free to use any base log:

Lemma 5. $\log_a x = \log_a b \times \log_b x$

which tells us that $\log_2 n = \Theta(\log_{10} n)$, the first being a constant multiple of the second, that constant being $\log_2 10$.

Finally, and most important, we need to keep in mind that using the techniques of Section 3 are (1) only estimates - "estimate" being another word for "educated guess" - and (2) dependent on a choice of model. The choice of model may also be an educated guess, or it could be from theoretical considerations, or it could be a simplification from known theoretical constraints.

As in all of science, a model is an approximation of reality.