# Vectorise

If you've used R for any length of time, you've probably heard the admonishment to "vectorise your code". But what does that actually mean? Vectorising your code is not just about avoiding for loops, although that's often a step. Vectorising is about taking a "whole object" approach to a problem, thinking about vectors, not scalars. There are two key attributes of a vectorised function:

- It makes many problems simpler. Instead of having to think about the components of a vector, you only think about entire vectors.
- The loops in a vectorised function are written in C instead of R. Loops in C are much faster because they have much less overhead.

Functionals stressed the importance of vectorised code as a higher level abstraction. Vectorisation is also important for writing fast R code. This doesn't mean simply using `apply()` or `lapply()`, or even `Vectorise()`. Those functions improve the interface of a function, but don't fundamentally change performance. Using vectorisation for performance means finding the existing R function that is implemented in C and most closely applies to your problem.

Vectorised functions that apply to many common performance bottlenecks include:

- `rowSums()`, `colSums()`, `rowMeans()`, and `colMeans()`. These vectorised matrix functions will always be faster than using `apply()`. You can sometimes use these functions to build other vectorised functions.

  ```
  rowAny <- function(x) rowSums(x) > 0
  rowAll <- function(x) rowSums(x) == ncol(x)
  ```

- Vectorised subsetting can lead to big improvements in speed. Remember the techniques behind lookup tables (lookup tables) and matching and merging by hand (matching and merging by hand). Also remember that you can use subsetting assignment to replace multiple values in a single step. If `x` is a vector, matrix or data frame then `x[is.na(x)] <- 0` will replace all missing values with 0.
- If you're extracting or replacing values in scattered locations in a matrix or data frame, subset with an integer matrix. See matrix subsetting for more details.
- If you're converting continuous values to categorical make sure you know how to use `cut()` and `findInterval()`.
- Be aware of vectorised functions like `cumsum()` and `diff()`.

Matrix algebra is a general example of vectorisation. There loops are executed by highly tuned external libraries like BLAS. If you can figure out a way to use matrix algebra to solve your problem, you'll often get a very fast solution. The ability to solve problems with matrix algebra is a product of experience. While this skill is something you'll develop over time, a good place to start is to ask people with experience in your domain.

The downside of vectorisation is that it makes it harder to predict how operations will scale. The following example measures how long it takes to use character subsetting to lookup 1, 10, and 100 elements from a list. You might expect that looking up 10 elements would take 10x as long as looking up 1, and that looking up 100 elements would take 10x longer again. In fact, the following example shows that it only takes about 9 times longer to look up 100 elements than it does to look up 1.

```
lookup <- setNames(as.list(sample(100, 26)), letters)

x1 <- "j"
x10 <- sample(letters, 10)
x100 <- sample(letters, 100, replace = TRUE)

microbenchmark(
  lookup[x1],
  lookup[x10],
  lookup[x100]
)
#> Unit: nanoseconds
#>           expr    min     lq  mean median     uq    max neval
#>     lookup[x1]    960  1,100  1220  1,210  1,300  2,750   100
#>    lookup[x10]  3,040  3,160  3512  3,250  3,340 27,400   100
#>   lookup[x100] 10,600 10,900 11373 11,100 11,500 26,300   100
```

Vectorisation won’t solve every problem, and rather than torturing an existing algorithm into one that uses a vectorised approach, you’re often better off writing your own vectorised function in C++. You’ll learn how to do so in Rcpp.

## Do as little as possible

The easiest way to make a function faster is to let it do less work. One way to do that is use a function tailored to a more specific type of input or ouput, or a more specific problem. For example:

- `rowSums()`, `colSums()`, `rowMeans()`, and `colMeans()` are faster than equivalent invocations that use `apply()` because they are vectorised (the topic of the next section).
- `vapply()` is faster than `sapply()` because it pre-specifies the output type.
- If you want to see if a vector contains a single value, `any(x == 10)` is much faster than `10 %in% x`. This is because testing equality is simpler than testing inclusion in a set.

Having this knowledge at your fingertips requires knowing that alternative functions exist: you need to have a good vocabulary. Start with the basics, and expand your vocab by regularly reading R code. Good places to read code are the R-help mailing list and stackoverflow.

Some functions coerce their inputs into a specific type. If your input is not the right type, the function has to do extra work. Instead, look for a function that works with your data as it is, or consider changing the way you store your data. The most common example of this problem is using `apply()` on a data frame. `apply()` always turns its input into a matrix. Not only is this error prone (because a data frame is more general than a matrix), it is also slower.

Other functions will do less work if you give them more information about the problem. It's always worthwhile to carefully read the documentation and experiment with different arguments. Some examples that I've discovered in the past include:

- `read.csv()`: specify known column types with `colClasses`.
- `factor()`: specify known levels with `levels`.
- `cut()`: don't generate labels with `labels = FALSE` if you don't need them, or, even better, use `findInterval()` as mentioned in the "see also" section of the documentation.
- `unlist(x, use.names = FALSE)` is much faster than `unlist(x)`.
- `interaction()`: if you only need combinations that exist in the data, use `drop = TRUE`.

Sometimes you can make a function faster by avoiding method dispatch. As we saw in (Extreme dynamism), method dispatch in R can be costly. If you're calling a method in a tight loop, you can avoid some of the costs by doing the method lookup only once:

- For S3, you can do this by calling `generic.class()` instead of `generic()`.
- For S4, you can do this by using `findMethod()` to find the method, saving it to a variable, and then calling that function.

For example, calling `mean.default()` quite a bit faster than calling `mean()` for small vectors:

```
x <- runif(1e2)

microbenchmark(
  mean(x),
  mean.default(x)
)
#> Unit: microseconds
#>             expr  min   lq mean median   uq  max neval
#>          mean(x) 6.28 6.48 7.13   6.62 6.77 21.3   100
#>  mean.default(x) 2.22 2.40 2.94   2.50 2.58 23.1   100
```

This optimisation is a little risky. While `mean.default()` is almost twice as fast, it'll fail in surprising ways if `x` is not a numeric vector. You should only use it if you know for sure what `x` is.

Knowing that you're dealing with a specific type of input can be another way to write faster code. For example, `as.data.frame()` is quite slow because it coerces each element into a data frame and then `rbind()`s them together. If you have a named list with vectors of equal length, you can directly transform it into a data frame. In this case, if you're able to make strong assumptions about your input, you can write a method that's about 20x faster than the default.

```
quickdf <- function(l) {
  class(l) <- "data.frame"
  attr(l, "row.names") <- .set_row_names(length(l[[1]]))
  l
}

l <- lapply(1:26, function(i) runif(1e3))
names(l) <- letters

microbenchmark(
  quick_df      = quickdf(l),
  as.data.frame = as.data.frame(l)
)
#> Unit: microseconds
#>           expr     min    lq    mean  median      uq     max neval
#>       quick_df    21.1    23    27.7    26.4    30.5    90.2   100
#>  as.data.frame 2,080.0 2,130 2284.5 2,160.0 2,220.0 4,200.0   100
```

Again, note the trade-off. This method is fast because it's dangerous. If you give it bad inputs, you'll get a corrupt data frame:

```
quickdf(list(x = 1, y = 1:2))
#> Warning in format.data.frame(x, digits = digits, na.encode = FALSE):
#> corrupt data frame: columns will be truncated or padded with NAs
#>   x y
#> 1 1 1
```

To come up with this minimal method, I carefully read through and then rewrote the source code for `as.data.frame.list()` and `data.frame()`. I made many small changes, each time checking that I hadn't broken existing behaviour. After several hours work, I was able to isolate the minimal code shown above. This is a very useful technique. Most base R functions are written for flexibility and functionality, not performance. Thus, rewriting for your specific need can often yield substantial improvements. To do this, you'll need to read the source code. It can be complex and confusing, but don't give up!

The following example shows a progressive simplification of the `diff()` function if you only want computing differences between adjacent values. At each step, I replace one argument with a specific case, and then check to see that the function still works. The initial function is long and complicated, but by restricting the arguments I not only make it around twice as fast, I also make it easier to understand.

First, I take the code of `diff()` and reformat it to my style:

```r
diff1 <- function (x, lag = 1L, differences = 1L) {
  ismat <- is.matrix(x)
  xlen <- if (ismat) dim(x)[1L] else length(x)
  if (length(lag) > 1L || length(differences) > 1L ||
      lag < 1L || differences < 1L)
    stop("'lag' and 'differences' must be integers >= 1")

  if (lag * differences >= xlen) {
    return(x[0L])
  }

  r <- unclass(x)
  i1 <- -seq_len(lag)
  if (ismat) {
    for (i in seq_len(differences)) {
      r <- r[i1, , drop = FALSE] -
        r[-nrow(r):-(nrow(r) - lag + 1L), , drop = FALSE]
    }
  } else {
    for (i in seq_len(differences)) {
      r <- r[i1] - r[-length(r):-(length(r) - lag + 1L)]
    }
  }
  class(r) <- oldClass(x)
  r
}
```

Next, I assume vector input. This allows me to remove the `is.matrix()` test and the method that uses matrix subsetting.

```r
diff2 <- function (x, lag = 1L, differences = 1L) {
  xlen <- length(x)
  if (length(lag) > 1L || length(differences) > 1L ||
      lag < 1L || differences < 1L)
```

```
    stop("'lag' and 'differences' must be integers >= 1")

  if (lag * differences >= xlen) {
    return(x[0L])
  }

  i1 <- -seq_len(lag)
  for (i in seq_len(differences)) {
    x <- x[i1] - x[-length(x):-(length(x) - lag + 1L)]
  }
  x
}
diff2(cumsum(0:10))
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

I now assume that `difference = 1L`. This simplifies input checking and eliminates the for loop:

```
diff3 <- function (x, lag = 1L) {
  xlen <- length(x)
  if (length(lag) > 1L || lag < 1L)
    stop("'lag' must be integer >= 1")

  if (lag >= xlen) {
    return(x[0L])
  }

  i1 <- -seq_len(lag)
  x[i1] - x[-length(x):-(length(x) - lag + 1L)]
}
diff3(cumsum(0:10))
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

Finally I assume `lag = 1L`. This eliminates input checking and simplifies subsetting.

```
diff4 <- function (x) {
  xlen <- length(x)
  if (xlen <= 1) return(x[0L])

  x[-1] - x[-xlen]
}
diff4(cumsum(0:10))
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

Now `diff4()` is both considerably simpler and considerably faster than `diff1()`:

```
x <- runif(100)
microbenchmark(
  diff1(x),
  diff2(x),
  diff3(x),
  diff4(x)
)
#> Unit: microseconds
#>      expr   min    lq  mean median    uq  max neval
#>  diff1(x) 13.40 13.90 15.37  14.10 14.60 33.8   100
#>  diff2(x) 10.80 11.30 12.49  11.60 11.80 33.6   100
#>  diff3(x)  8.84  9.26 10.02   9.51  9.77 19.6   100
#>  diff4(x)  6.11  6.47  6.92   6.67  6.99 22.4   100
```

You'll be able to make `diff()` even faster for this special case once you've readRcpp.

A final example of doing less work is to use simpler data structures. For example, when working with rows from a data frame, it's often faster to work with row indices than data frames. For instance, if you wanted to compute a bootstrap estimate of the correlation between two columns in a data frame, there are two basic approaches: you can either work with the whole data frame or with the individual vectors. The following example shows that working with vectors is about twice as fast.

```
sample_rows <- function(df, i) sample.int(nrow(df), i,
  replace = TRUE)

# Generate a new data frame containing randomly selected rows
boot_cor1 <- function(df, i) {
  sub <- df[sample_rows(df, i), , drop = FALSE]
  cor(sub$x, sub$y)
}

# Generate new vectors from random rows
boot_cor2 <- function(df, i ) {
  idx <- sample_rows(df, i)
  cor(df$x[idx], df$y[idx])
}

df <- data.frame(x = runif(100), y = runif(100))
microbenchmark(
  boot_cor1(df, 10),
  boot_cor2(df, 10)
)
#> Unit: microseconds
#>               expr min  lq mean median  uq max neval
#>  boot_cor1(df, 10) 179 185  200    194 199 768   100
#>  boot_cor2(df, 10) 113 116  122    119 126 192   100
```

## Avoid copies

A pernicious source of slow R code is growing an object with a loop. Whenever you use `c()`, `append()`, `cbind()`, `rbind()`, or `paste()` to create a bigger object, R must first allocate space for the new object and then copy the old object to its new home. If you're repeating this many times, like in a for loop, this can be quite expensive. You've entered Circle 2 of the "R inferno".

Here's a little example that shows the problem. We first generate some random strings, and then combine them either iteratively with a loop using `collapse()`, or in a single pass using `paste()`. Note that the performance of `collapse()` gets relatively worse as the number of strings grows: combining 100 strings takes almost 30 times longer than combining 10 strings.

```
random_string <- function() {
  paste(sample(letters, 50, replace = TRUE), collapse = "")
}
strings10 <- replicate(10, random_string())
strings100 <- replicate(100, random_string())

collapse <- function(xs) {
  out <- ""
  for (x in xs) {
    out <- paste0(out, x)
  }
  out
}

microbenchmark(
  loop10  = collapse(strings10),
  loop100 = collapse(strings100),
  vec10   = paste(strings10, collapse = ""),
  vec100  = paste(strings100, collapse = "")
)
#> Unit: microseconds
#>     expr     min      lq   mean  median      uq     max neval
#>   loop10    49.5    51.0   54.8    52.0    58.4    89.7   100
#>  loop100 1,570.0 1,580.0 1649.2 1,590.0 1,670.0 2,180.0   100
#>    vec10    11.0    11.5   12.8    11.8    12.3    35.2   100
#>   vec100    79.6    80.2   83.5    80.6    84.7   104.0   100
```

Modifying an object in a loop, e.g., `x[i] <- y`, can also create a copy, depending on the class of `x`. Modification in place discusses this issue in more depth and gives you some tools to determine when you're making copies.

## Byte code compilation

R 2.13.0 introduced a byte code compiler which can increase the speed of some code. Using the compiler is an easy way to get improvements in speed. Even if it doesn't work well for your function, you won't have invested a lot of time in the effort. The following example shows the pure R version of `lapply()` from functionals. Compiling it gives a considerable speedup, although it's still not quite as fast as the C version provided by base R.

```
lapply2 <- function(x, f, ...) {
  out <- vector("list", length(x))
  for (i in seq_along(x)) {
    out[[i]] <- f(x[[i]], ...)
  }
  out
}

lapply2_c <- compiler::cmpfun(lapply2)

x <- list(1:10, letters, c(F, T), NULL)
microbenchmark(
  lapply2(x, is.null),
  lapply2_c(x, is.null),
  lapply(x, is.null)
)
#> Unit: microseconds
#>                   expr  min    lq  mean median    uq  max neval
#>    lapply2(x, is.null) 9.92 11.30 13.42  11.80 12.40 49.2   100
#>  lapply2_c(x, is.null) 4.40  5.18  5.79   5.53  5.83 15.1   100
#>     lapply(x, is.null) 4.58  5.36  6.22   5.70  6.16 34.0   100
```

Byte code compilation really helps here, but in most cases you're more likely to get a 5-10% improvement. All base R functions are byte code compiled by default.

## Case study: t-test

The following case study shows how to make t-tests faster using some of the techniques described above. It's based on an example in "Computing thousands of test statistics simultaneously in R" by Holger Schwender and Tina Müller. I thoroughly recommend reading the paper in full to see the same idea applied to other tests.

Imagine we have run 1000 experiments (rows), each of which collects data on 50 individuals (columns). The first 25 individuals in each experiment are assigned to group 1 and the rest to group 2. We'll first generate some random data to represent this problem:

```
m <- 1000
n <- 50
X <- matrix(rnorm(m * n, mean = 10, sd = 3), nrow = m)
grp <- rep(1:2, each = n / 2)
```

For data in this form, there are two ways to use `t.test()`. We can either use the formula interface or provide two vectors, one for each group. Timing reveals that the formula interface is considerably slower.

```
system.time(for(i in 1:m) t.test(X[i, ] ~ grp)$stat)
#>    user  system elapsed
#>     1.4     0.0     1.4
system.time(
  for(i in 1:m) t.test(X[i, grp == 1], X[i, grp == 2])$stat
)
#>    user  system elapsed
#>   0.348   0.000   0.354
```

Of course, a for loop computes, but doesn't save the values. We'll use `apply()` to do that. This adds a little overhead:

```
compT <- function(x, grp){
  t.test(x[grp == 1], x[grp == 2])$stat
}
system.time(t1 <- apply(X, 1, compT, grp = grp))
#>    user  system elapsed
#>   0.341   0.000   0.341
```

How can we make this faster? First, we could try doing less work. If you look at the source code of `stats:::t.test.default()`, you'll see that it does a lot more than just compute the t-statistic. It also computes the p-value and formats the output for printing. We can try to make our code faster by stripping out those pieces.

```
my_t <- function(x, grp) {
  t_stat <- function(x) {
    m <- mean(x)
    n <- length(x)
    var <- sum((x - m) ^ 2) / (n - 1)

    list(m = m, n = n, var = var)
  }

  g1 <- t_stat(x[grp == 1])
  g2 <- t_stat(x[grp == 2])

  se_total <- sqrt(g1$var / g1$n + g2$var / g2$n)
  (g1$m - g2$m) / se_total
}
system.time(t2 <- apply(X, 1, my_t, grp = grp))
#>    user  system elapsed
#>   0.049   0.000   0.049
stopifnot(all.equal(t1, t2))
```

This gives us about a 6x speed improvement.

Now that we have a fairly simple function, we can make it faster still by vectorising it. Instead of looping over the array outside the function, we will modify `t_stat()` to work with a matrix of values.
Thus, `mean()` becomes `rowMeans()`, `length()` becomes`ncol()`, and `sum()` becomes `rowSums()`. The rest of the code stays the same.

```
rowtstat <- function(X, grp){
  t_stat <- function(X) {
    m <- rowMeans(X)
    n <- ncol(X)
    var <- rowSums((X - m) ^ 2) / (n - 1)

    list(m = m, n = n, var = var)
  }

  g1 <- t_stat(X[, grp == 1])
  g2 <- t_stat(X[, grp == 2])

  se_total <- sqrt(g1$var / g1$n + g2$var / g2$n)
  (g1$m - g2$m) / se_total
}
system.time(t3 <- rowtstat(X, grp))
#>    user  system elapsed
#>   0.003   0.000   0.003
stopifnot(all.equal(t1, t3))
```

That's much faster! It's at least 40x faster than our previous effort, and around 1000x faster than where we started.

Finally, we could try byte code compilation. Here we'll need to use `microbenchmark()`instead of `system.time()` in order to get enough accuracy to see a difference:

```
rowtstat_bc <- compiler::cmpfun(rowtstat)

microbenchmark(
  rowtstat(X, grp),
  rowtstat_bc(X, grp),
  unit = "ms"
)
#> Unit: milliseconds
#>                 expr  min   lq mean median   uq  max neval
#>     rowtstat(X, grp) 2.58 2.62 2.89   3.01 3.04 3.83   100
#>  rowtstat_bc(X, grp) 2.56 2.60 2.90   2.99 3.05 4.90   100
```

In this example, byte code compilation doesn't help at all.