Optimal sequence alignments

Consider three words in different languages, for example Henry, Enrico and Heinrich. The alphabet used to construct these words is the Latin alphabet and just by looking at them we are convinced that the words look "the same". One way to make this rigorous is to "align" the words so that similar letters line up. For example,

H e - n r y - -
H e i n r i c h
- E - n r i c o

and then introduce a score function for which aligned letters gain a score $+1$, mismatched letters (like h and o in the last two words) get penalised by $-a$ and gaps (denoted by underscores) are penalised by $-b$.

An alignment is called optimal if it achieves the highest possible score from all possible alignments between the words and a high score indicates a higher probability that the words are "similar". This can be used for example when comparing DNA sequences of two species.

While the example above is deterministic, one can add randomness to this by creating two words over any finite alphabet with random letters, and then trying to find the optimal score and alignment, as well as the behaviour of these scores when the alphabet size k, costs a, b and size n change.

Key words: Sequence alignment, global alignment, optimality regions, multiple sequence alignments, algebraic statistics.

In a recent paper with Janosch Ortmann we studied the number of optimality regions when comparing two unequal sequences at the onset of the flat edge. The paper, titled *"Optimality Regions and Fluctuations for Bernoulli Last Passage Models"* is published under open access and it can be found here.

The sharper results in the article have to do with the number optimality regions in the discrete Hammersley process, where we obtain Tracy-Widom laws under different rescaling of the process.