Supplement to "Availability of Perfect Decomposition in Statistical Linkage Learning for Unitation-Based Function Concatenations"

Michal Prusik
Fac. of Pure and Applied Mathematics
Wroclaw Univ. of Science and Techn.
Wroclaw, Poland
michal.prusik@pwr.edu.pl

Bartosz Frej
Fac. of Pure and Applied Mathematics
Wroclaw Univ. of Science and Techn.
Wroclaw, Poland
bartosz.frej@pwr.edu.pl

Michal W. Przewozniczek
Dep. of Systems and Computer Networks
Wroclaw Univ. of Science and Techn.
Wroclaw, Poland
michal.przewozniczek@pwr.edu.pl

1 Pseudocode for the First Improvement Hill Climber

Pseudocode 1 First Improvement Hill Climber

```
1:  function RunForLevel(solution)
2:      optSolution ← solution
3:      geneOrder ← GenerateRandomGeneOrder(size(optSolution));
4:      repeat
5:          modified ← false;
6:          for each gene in geneOrder do
7:              fitness ← Fitness(optSolution);
8:              optSolution[gene] ← ¬optSolution[gene];
9:              fitnessNew ← Fitness(optSolution);
10:             if fitnessNew > fitness then
11:                 modified ← true;
12:             else
13:                 optSolution[gene] ← ¬optSolution[gene];
14:             end if
15:         end for
16:     until modified = true;
17:     return optSolution
18: end function
```

2 Calculations transfered from the main paper

For the convenience of the reader interested in mathematical details, we present below short calculations which we decided not to include in the main paper. We start from the lemmas stated at the end of the section 2 (Basic notions).

Denote by $\mathcal{S}$ the three-dimensional simplex (a tetrahedron) in $\mathbb{R}^4$, with four extremal points (vertices) having 1 at a single non-zero coordinate. Let $\widetilde{H}$ be the extension of the entropy function H from $\mathcal{S}$ to the set $\{(p_1,p_2,p_3,p_4) : p_1,...,p_4 \geqslant 0\}$ by $\widetilde{H}(p_1,p_2,p_3,p_4) = -\sum_{i=1}^{4} p_i \log p_i$. Let $\nabla\widetilde{H}$ denote the gradient of $\widetilde{H}$ and $\nabla H_{\mathcal{S}}$ be its orthogonal projection onto the hyperplane containing $\mathcal{S}$.

Lemma S-1 (Lemma 6 in the section Basic notions). *Let P and Q be probability vectors. If the dot product $(P-Q)\cdot\nabla\widetilde{H}(Q)$ is less than 0, then $H(P) \leqslant H(Q)$.*

Proof. If $H(P) > H(Q)$ then by concavity of the entropy function on $\mathcal{S}$

$$\widetilde{H}(Q+t(P-Q)) = \widetilde{H}((1-t)Q+tP) \geqslant (1-t)\widetilde{H}(Q)+t\widetilde{H}(P) > \widetilde{H}(Q)$$

for each $t \in (0,1)$. Hence $(P-Q)\cdot\nabla\widetilde{H}(Q)$ is nonnegative as the directional derivative of $\widetilde{H}$ along $(P-Q)$. □

Lemma S-2 (Lemma 8 in the section Basic notions). *For each $P = (t, \frac{1}{2}-t, \frac{1}{2}-t, t)$ with $t \in (\frac{1}{4}, \frac{1}{2})$ there is $k > 0$ such that $\nabla H_{\mathcal{S}}(P) = (-k,k,k,-k)$.*

Proof. The vector $(1,1,1,1)$ is normal to the hyperplane containing $\mathcal{S}$, hence for $P = (t, \frac{1}{2}-t, \frac{1}{2}-t, t)$ we have $\nabla H_{\mathcal{S}}(P) = \nabla\widetilde{H}(P) - (s,s,s,s)$, where s is suitably chosen. This yields the equation

$$\left(\nabla\widetilde{H}(P)-(s,s,s,s)\right)\cdot(1,1,1,1) = -2\log t - 2\log\left(\frac{1}{2}-t\right) - 4 - 4s = 0,$$

hence $s = -\frac{\log t + \log\left(\frac{1}{2}-t\right)}{2} - 1$ and, finally, $\nabla H_{\mathcal{S}}(P) = (-k,k,k,-k)$ for

$$k = \frac{\log t - \log\left(\frac{1}{2}-t\right)}{2}.$$

□

Finaly, a calculation cut out of the proof of Lemma 16—the map φ contracts distances between points in $\mathcal{T}$ by factor $\frac{1}{\sqrt{2}}$.

$$\begin{aligned}
&\|(p_1,p_2,p_2,1-p_1-2p_2)-(q_1,q_2,q_2,1-q_1-2q_2)\|^2 \\
&\quad = (p_1-q_1)^2 + 2(p_2-q_2)^2 + (p_1-q_1+2p_2-2q_2)^2 \\
&\quad = 2(p_1-q_1)^2 + 6(p_2-q_2)^2 + 4(p_1-q_1)(p_2-q_2) \\
&\quad = 2((p_1-q_1+p_2-q_2)^2 + 2(p_2-q_2)^2) \\
&\quad = 2\|(p_1+p_2, \sqrt{2}p_2)-(q_1+q_2,\sqrt{2}q_2)\|^2.
\end{aligned}$$

3 The distribution of the dependent genes

We present here the proof of the Theorem 14 from the section 5 of the paper.

Theorem S-3. *Let g have the monotonicity represented by*

$$MAX_g = (k_1,\ldots,k_N) \qquad \text{and} \qquad MIN_g = (l_0,l_1,\ldots,l_N).$$

Let (q_1, q_2, q_2, q_3) be the distribution of the dependent pairs in this model. Then we have

$$\begin{aligned} q_1 &= 2^{-k+1} \cdot \sum_{i=1}^{N} \frac{(k-k_i)(k-k_i-1)}{k(k-1)} \sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j} \\ &+ 2^{-k} \cdot \left(\sum_{j=0}^{l_0-1} \binom{k}{j} + \binom{k-1}{l_0-1} \right) \cdot \chi_{\{l_0>0\}}, \\ q_2 &= 2^{-k+1} \cdot \sum_{i=1}^{N} \frac{k_i(k-k_i)}{k(k-1)} \sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j}, \end{aligned}$$

where $\chi_{\{l_0>0\}} = 1$ for $l_0 > 0$ and $\chi_{\{l_0>0\}} = 0$ if $l_0 = 0$.

Proof. Take $x \in \{0,1\}^k$ and fix two genes x_v, x_w. We will start with computing q_2, since its formula will not depend on l_0. First, let us notice that the situation $x_v = 0$, $x_w = 1$ may appear after FIHC optimization only if the optimized solution has unitation equal to k_i for some $i \in \{1, \ldots, N\}$ (i.e., if we are at some maximum which is not the block of zeros and not the block of ones). Thus we have

$$\begin{aligned} q_2 &= \mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 01) \\ &= \sum_{i=1}^{N} \mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 01 | u(\mathcal{F}_\pi(x)) = k_i) \mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i). \end{aligned}$$

By similar arguments as in Section 4 in the main article, the events $\mathcal{F}_\pi(x) = y$ are equally probable for all $y \in \{0,1\}^k$ having unitation equal to k_i. Hence, to compute the probability $\mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 01 | u(\mathcal{F}_\pi(x)) = k_i)$ we just need to count the number of blocks y with unitation equal to k_i in which $y_{vw} = 01$ and divide it by the number of all such blocks y. Therefore we get

$$\mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 01 | u(\mathcal{F}_\pi(x)) = k_i) = \frac{\binom{k-2}{k_i-1}}{\binom{k}{k_i}} = \frac{k_i(k-k_i)}{k(k-1)}.$$

To compute $\mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i)$ let us first write it as

$$\mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i) = \sum_{\pi} \frac{1}{k!} \mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i | \pi).$$

For any permutation π we get $u(\mathcal{F}_\pi(x)) = k_i$ whenever one of the following holds:

- $u(x)$ is greater than l_{i-1} and smaller than l_i,
- $u(x) = l_{i-1}$ and $\pi x_1 = 0$,
- $u(x) = l_i$ and $\pi x_1 = 1$.

Therefore, we get

$$\begin{aligned} \mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i | \pi) &= \sum_{j=l_{i-1}+1}^{l_i-1} \mathbb{P}(u(x) = j | \pi) + \\ &+ \mathbb{P}(u(x) = l_{i-1}, \pi x_1 = 0 | \pi) + \mathbb{P}(u(x) = l_i, \pi x_1 = 1 | \pi) \\ &= \sum_{j=l_{i-1}+1}^{l_i-1} \mathbb{P}(u(x) = j) + \mathbb{P}(u(x) = l_{i-1},\, x_1 = 0) + \mathbb{P}(u(x) = l_i,\, x_1 = 1) \\ &= \sum_{j=l_{i-1}+1}^{l_i-1} \binom{k}{j} 2^{-k} + \binom{k-1}{l_{i-1}} 2^{-k} + \binom{k-1}{l_i-1} 2^{-k} \\ &= \left[\sum_{j=l_{i-1}+1}^{l_i-1} \left(\binom{k-1}{j-1} + \binom{k-1}{j} \right) + \binom{k-1}{l_{i-1}} + \binom{k-1}{l_i-1} \right] 2^{-k} \\ &= \left[\sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j} \right] 2 \cdot 2^{-k} = \left[\sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j} \right] 2^{-k+1}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i) &= \sum_{\pi} \left[\sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j} \right] 2^{-k+1} \cdot \frac{1}{k!} \\ &= \left[\sum_{j=l_{i-1}}^{l_i-1} \binom{k-1}{j} \right] 2^{-k+1}, \end{aligned}$$

which together with the previous calculations gives us the final formula for q_2.

For computing q_1 let us notice that apart from $u(\mathcal{F}_\pi(x)) = k_i$ we must include $u(\mathcal{F}_\pi(x)) = 0$ (the block of zeros), which guarantees that $\mathcal{F}_\pi(x)_{vw} = 00$. Thus

$$\begin{aligned} q_1 &= \mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 00) \\ &= \sum_{i=1}^{N} \mathbb{P}\big(\mathcal{F}_\pi(x)_{vw} = 00 | u(\mathcal{F}_\pi(x)) = k_i\big)\, \mathbb{P}\big(u(\mathcal{F}_\pi(x)) = k_i\big) + \\ &\quad + \mathbb{P}(u(\mathcal{F}_\pi(x)) = 0). \end{aligned}$$

The event $u(\mathcal{F}_\pi(x)) = 0$ has a nonzero probability if and only if $l_0 > 0$ (i.e., if we have a maximum at unitation 0). Then

$$\begin{aligned} &\mathbb{P}(u(\mathcal{F}_\pi(x)) = 0) = \\ &= \sum_{j=0}^{l_0-1} \mathbb{P}(u(x) = j) + \sum_{\pi} \frac{1}{k!} \mathbb{P}(u(x) = l_0, \pi x_1 = 1 | \pi) \\ &= 2^{-k} \cdot \sum_{j=0}^{l_0-1} \binom{k}{j} + 2^{-k} \cdot \sum_{\pi} \frac{1}{k!} \binom{k-1}{l_0-1} \\ &= 2^{-k} \cdot \left(\sum_{j=0}^{l_0-1} \binom{k}{j} + \binom{k-1}{l_0-1} \right). \end{aligned}$$

Using the same argumentation as for q_2 we get also

$$\mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 00 | u(\mathcal{F}_\pi(x)) = k_i) = \frac{\binom{k-2}{k_i}}{\binom{k}{k_i}} = \frac{(k-k_i)(k-k_i-1)}{k(k-1)}$$

for $k_i \neq k-1$. It is possible that $k_N = k-1$, however in this situation we have $\mathbb{P}(\mathcal{F}_\pi(x)_{vw} = 00 | u(\mathcal{F}_\pi(x)) = k-1) = 0$ (because we need at least two genes to be zero), so we can use the formula for this case, too. Computing $\mathbb{P}(u(\mathcal{F}_\pi(x)) = k_i)$ goes in the same way as previously. □