Formulation of the problem. One of the main algorithms for clustering \( n \) given \( d \)-dimensional points selects \( K \) “typical” values \( c_k \) and assignments \( k(i) \) for each \( i \) from 1 to \( n \) so as to minimize the sum \( \sum_i (x_i - c_{k(i)})^2 \). This minimization is usually done iteratively. First, we pick \( c_k \) and assign each point \( x_i \) to the cluster \( k \) whose representative \( c_k \) is the closest to \( x_i \). Then, we freeze \( k(i) \) and select new typical representatives \( c_k \) by minimizing the objective function. This leads to \( c_k \) being an average of all the points \( x_i \) assigned to the \( k \)-th cluster. Then, the procedure repeats again and again – until the process converges.

In practice, for some objects, we cannot definitely assign them to a single cluster. In such cases, it is reasonable to assign, to each object \( i \), degrees \( \mu_{ik} \) of belongs to different clusters \( k \), so that \( \sum_k \mu_{ik} = 1 \). In this case, it seems reasonable to take each term \((x_i - c_k)^2\) with the weight \( \mu_{ik} \), i.e., to find the values \( \mu_{ik} \) and \( c_k \) by minimizing the expression \( \sum_{i,k} \mu_{ik} \cdot (x_i - c_k)^2 \). However, this expression is linear in \( \mu_{ik} \), and the minimum of a linear function under linear constraints is always at a vertex, i.e., when one value \( \mu_{ik} \) is 1 and the rest are 0s. To come up with truly fuzzy clustering, with \( 0 < \mu_{ik} < 1 \), we need to replace the factor \( \mu_{ik} \) with a non-linear expression \( f(\mu_{ik}) \), i.e., to minimize \( \sum_{i,k} f(\mu_{ik}) \cdot (x_i - c_k)^2 \). In practice, the functions \( f(\mu) = \mu^p \) works the best. Why?

Our explanation. The weights \( \mu_{ik} \) are normalized so that their sum is 1. So, if we delete some clusters or add more clusters, we need to re-normalize these values. A usual way to do it is to multiply them by a normalization constant \( c \). It is therefore reasonable to require that the relative quality of different clustering ideas not change is we simply re-scale. This implies, e.g., that if \( f(\mu_1) \cdot v_1 = f(\mu_2) \cdot v_2 \), then after re-scaling \( \mu_i \to c \cdot \mu_i \), we should have \( f(c \cdot \mu_1) \cdot v_1 = f(c \cdot \mu_2) \cdot v_2 \). We show that this condition implies that \( f(\mu) = \mu^p \).

Indeed, \( \frac{f(c \cdot \mu_2)}{f(c \cdot \mu_1)} = \frac{v_1}{v_2} = \frac{f(\mu_2)}{f(\mu_1)} \), thus \( r \defeq \frac{f(c \cdot \mu_1)}{f(\mu_1)} = \frac{f(c \cdot \mu_2)}{f(\mu_2)} \) for all \( \mu_1 \) and \( \mu_2 \). Thus, the ratio \( r \) does not depend on \( \mu \): \( r = r(c) \), and \( f(c \cdot \mu) = r(c) \cdot f(\mu) \). It is known that the only continuous solutions of this functional equations are \( f(\mu) = C \cdot \mu^p \). Minimization is not affected if we divide the objective function by \( C \) and get \( f(\mu) = \mu^p \).