Handout on NPML Property of Kaplan-Meier

Assume we have the usual right-censored sample of survival data $(T_i\,\Delta_i,\ i = 1, \ldots, n)$, where $T_i = \min(X_i, C_i)$, $\Delta_i = I_{[XZ_i \le C_i]}$, with X_i and C_i independent, and parameter of interest equal to the marginal probability law of X_i. Among all such marginal probability laws it can be argued that the ones which have likelihood at least as large as any single competitor (the main idea of Weiss and Wolfowitz' 1956 'Nonparametric Maximum Likelihood') assign mass only to the set of observed death-times. Denote by $t_{(k)},\ k = 1, \ldots, D$ the ordered distinct death-times among T_i (i.e., the sorted unduplicated values T_i for which $\Delta_i = 1$), and let d_k be the number of observed deaths at time $T_i = t_{(k)}$. Let $p(t_{(k)})$ be the probability D-vector of masses which a discrete probability law assigns respectively to the death-times $t_{(k)},\ k = 1, \ldots, D$. Then also denote

$$Y_k \;=\; Y(t_{(k)}) \;=\; \sum_{i=1}^{n} I_{[T_i \ge t_{(k)}]}$$

$$S(t) \equiv \sum_{k: t_{(k)} \ge t} p(t_{(k)}) \quad , \qquad h_k \equiv \frac{p(t_{(k)})}{\sum_{j:j\ge k} p(t_{(j)})} \;=\; \frac{p(t_{(k)})}{S(t_{(k)})}$$

and note that the survival function $S(t)$ is left-continuous and given by the identity

$$S(t) \;=\; \prod_{k:\, t_k < t} (1 - h_k)$$

Now we calculate and re-express the log-likelihood in terms of the 'discrete hazard' intensity parameters h_k as follows:

$$\begin{aligned}
logLik(\underline{h}) \;&=\; \sum_{i=1}^{n} \Big(\Delta_i \,\log p(T_i) \;+\; (1-\Delta_i)\,\log S(T_i) \Big) \\
&=\; \sum_{i=1}^{n} \Big(\Delta_i \,\log \frac{p(T_i)}{S(T_i)} \;+\; \log S(T_i) \Big) \\
&=\; \sum_{k=1}^{D} d_k \,\log h_k \;+\; \sum_{i=1}^{n} \log \prod_{k:\, t_k < T_i} (1 - h_k) \\
&=\; \sum_{k=1}^{D} d_k \,\log h_k \;+\; \sum_{k=1}^{D} \sum_{i=1}^{n} I_{[t_k < T_i]} \log(1 - h_k)
\end{aligned}$$

$$= \sum_{k=1}^{D} (d_k \log h_k + (Y_k - d_k) \log(1 - h_k))$$

Finally, note that in this setup the discrete hazard values h_k are unrestricted strictly positive numbers, except that $h_D \equiv 1$ if $Y_D = d_D$. Therefore the previous *logLik* is maximized uniquely when

$$d_k/h_k = (Y_k - d_k)/(1 - h_k) \quad , k \le \begin{cases} D & \text{if } Y_D > d_D \\ D-1 & \text{otherwise} \end{cases}$$

But this means precisely that $h_k = d_k/Y_k$ for the same set of k values in the previous display, a condition which precisely specifies the discrete hazard as the Nelson-Aalen solution and the survival function $S(t)$ as the (left-continuous) Kaplan-Meier survival function estimator. **Thus these estimators can be interpreted as giving the Nonparametric Maximum Likelihood survival distribution.**