Sherlock et al. (2015b) (in a non-subsampling context) study the statistical efficiency for delayed acceptance random walk Metropolis and, moreover, an efficiency that also takes into account the computational efficiency for the case where the target is estimated (DA-PMMH in Section 4).

Christen and Fox (2005) note that, because the transition kernels of both the MH and delayed acceptance MH are derived from the same proposal q_1, and in addition $\alpha_2 \leq 1$, the delayed acceptance MH will be less statistically efficient than MH. The intuition is that under these conditions the chain clearly exhibits a more "sticky" behavior and an estimate based on these samples will have a larger asymptotic variance under DA-MH than MH. Notice that the closer α_2 is to 1, the more statistically efficient the delayed acceptance algorithm is, and when $\alpha_2 = 1$ it is equivalent to the standard MH which gives the upper bound of the possible statistical efficiency achieved by a DA-MH.

Result 1 in Payne and Mallick (2015) gives the alternative formulation (for state-independent approximations)

$$\alpha_2(\theta_c \to \theta_p) = \min\left\{1, \frac{\hat{p}_m(y|\theta_c, u)/p(y|\theta_c)}{\hat{p}_m(y|\theta_p, u)/p(y|\theta_p)}\right\}. \tag{3.3}$$

Let $l_k(\theta_c, \theta_p) = l_k(\theta_c) - l_k(\theta_p)$ and denote by $\hat{l}_m(\theta_c, \theta_p)$ the estimate of $l(\theta_c, \theta_p) = \sum_{k=1}^{n} l_k(\theta_c, \theta_p)$. Similarly to (2.3),

$$\hat{l}_m(\theta_c, \theta_p) = q(\theta_c, \theta_p) + \frac{1}{m}\sum_{i=1}^{m} \zeta_i, \quad \text{with } q(\theta_c, \theta_p) = \sum_{k=1}^{n} q_k(\theta_c, \theta_p), \tag{3.4}$$

where $q_k(\theta_c, \theta_p) = q_k(\theta_c) - q_k(\theta_p)$ and the ζ_i's are iid with

$$\Pr\left(\zeta_i = nD_k(\theta_c, \theta_p)\right) = \frac{1}{n}, \text{ with } D_k = \left(l_k(\theta_c, \theta_p) - q_k(\theta_c, \theta_p)\right) \quad \text{for } i = 1, \dots m.$$

We can also show that