Off-policy importance sampling (Precup, 2000) is a popular technique, and per-decision importance sampling is a popular variant. Thomas (2015) gives a great explanation of both, which I will try to compress into one page here, using a simpler setting.

Let our deterministic reward function be $R(S, A)$, and suppose we would like to estimate the expected value of discounted return $G = \sum_{t=1}^{T} u^{t-1} R(S_t, A_t)$ under a policy π using n length-T trajectories $(s_{i,t}, a_{i,t})_{i=1,\ldots,n,t=1,\ldots,T}$ generated by $\pi^{\mathscr{B}}$. Ordinarily, we would note

$$E_\pi[G] = E_\pi\left[\sum_{t=1}^{T} u^{t-1} R(S_t, A_t)\right] \tag{1}$$

$$= E_{\pi^{\mathscr{B}}}\left[\left[\prod_{k=1}^{T} \frac{\pi(A_k|S_k)}{\pi^{\mathscr{B}}(A_k|S_k)}\right] \sum_{t=1}^{T} u^{t-1} R(S_t, A_t)\right] \tag{2}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left[\left[\prod_{k=1}^{T} \frac{\pi(a_{i,k}|s_{i,k})}{\pi^{\mathscr{B}}(a_{i,k}|s_{i,k})}\right] \sum_{t=1}^{T} u^{t-1} R(s_{i,t}, a_{i,t})\right]. \tag{3}$$

Note that the ratio of policies is all that is left from the ratio of densities after cancellations.

In **per-decision importance sampling**, we just choose to bring the expectation into the sum before multiplying by the ratio of the densities.

$$\begin{aligned}
E_\pi[G] &= E_\pi\left[\sum_{t=1}^{T} u^{t-1} R(S_t, A_t)\right] \\
&= \sum_{t=1}^{T} u^{t-1} E_\pi\left[R(S_t, A_t)\right] \\
&= \sum_{t=1}^{T} u^{t-1} E_{\pi^{\mathscr{B}}}\left[\left[\prod_{k=1}^{t} \frac{\pi(A_k|S_k)}{\pi^{\mathscr{B}}(A_k|S_k)}\right] R(S_t, A_t)\right] \\
&\approx \sum_{t=1}^{T} u^{t-1} \frac{1}{n}\sum_{i=1}^{n}\left[\left[\prod_{k=1}^{t} \frac{\pi(a_{i,k}|s_{i,k})}{\pi^{\mathscr{B}}(a_{i,k}|s_{i,k})}\right] R(s_{i,t}, a_{i,t})\right]
\end{aligned}$$

Our ratio only goes only up until t instead of T because

$$\begin{aligned}
E[R(S_t, A_t)] &= E[E[R(S_t, A_t)|S_1, \ldots, S_T, A_1, \ldots, A_T]] \\
&= E[E[R(S_t, A_t)|S_1, \ldots, S_t, A_1, \ldots, A_t]].
\end{aligned}$$

Per-decision importance sampling is more stable than ordinary importance sampling, because the truncated ratio has lower variance.

References

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Philip S Thomas. *Safe reinforcement learning*. PhD thesis, 2015.