- Previously, we computed n-gram probabilities based on relative frequency

$$p(w_2|w_1) = \frac{\mathsf{count}(w_1, w_2)}{\mathsf{count}(w_1)}$$

- Good Turing smoothing adjusts counts c to expected counts c^*

$$\mathsf{count}^*(w_1, w_2) \leq \mathsf{count}(w_1, w_2)$$

- We use these expected counts for the prediction model (but 0^* remains 0)

$$\alpha(w_2|w_1) = \frac{\mathsf{count}^*(w_1, w_2)}{\mathsf{count}(w_1)}$$

- This leaves probability mass for the discounting function

$$d_2(w_1) = 1 - \sum_{w_2} \alpha(w_2|w_1)$$