A Implementation details

B Theoretical proofs for the sampling procedures

Let's refresh notation from Section 4.

Let $\mathcal{X}$ be a universe of possible outputs (e.g. relation instances), $\mathcal{Y} \subseteq \mathcal{X}$ be an unknown subset of this universe corresponding to the correct elements in $\mathcal{X}$ and $X_1, \ldots X_m \subseteq \mathcal{X}$ be known subsets that correspond to the predicted output from m systems, and $Y_1, \ldots, Y_m$ be the intersection of $X_1, \ldots, X_m$ with $\mathcal{Y}$. Furthermore, let $\hat{X}_i$ be a mulit-set of n_i independent samples drawn from X_i with the distribution p_i, $\hat{Y}_i$ be the intersection of these sets with $\mathcal{Y}$, and $\hat{Y}_0$ be a sample drawn from $\mathcal{Y}$ according to an unknown distribution $p'(x)$.

We would like to evaluate precision, π_i, and recall, r_i:

$$\pi_i \stackrel{\text{def}}{=} \mathbb{E}_{x \sim X_i}[f(x)] \qquad\qquad r_i \stackrel{\text{def}}{=} \mathbb{E}_{x \sim \mathcal{Y}}[g_i(x)],$$

In this section, we'll provide proofs that show that the joint estimators proposed in Section 4 are indeed unbiased, and we will characterize their variance.

B.1 Estimating precision

In Section 4, we proposed the following estimator for π_i:

$$\hat{\pi}_i \stackrel{\text{def}}{=} \sum_{j=1}^{m} \frac{w_{ij}}{n_j} \sum_{x \in \hat{X}_j} \frac{p_i(x) f(x)}{q_i(x)},$$

where $q_i(x) = \sum_{j=1}^{m} w_{ij} p_j(x)$ and $w_{ij} \geq 0$ are mixture parameters such that $\sum_{j=1}^{m} w_{ij} = 1$ and $q_i(x) > 0$ wherever $p_i(x) > 0$.

Theorem 1 (Statistical properties of $\hat{\pi}_i$). *$\hat{\pi}_i$ is an unbiased estimator of π_i and has a variance of:*

$$\operatorname{Var} \hat{\pi}_i = \sum_{j=1}^{m} \frac{w_j^2}{n_j} \mathbb{E}_{p_j} \left[\frac{p_i(x)^2 f(x)^2 - \pi_{ij} p_i(x) f(x) q_i(x)}{q_i(x)^2} \right],$$

where $\pi_{ij} \stackrel{\text{def}}{=} \mathbb{E}_{p_j} \left[\frac{p_i(x) f(x)}{q_i(x)} \right]$.

Proof. Let $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_m)$ which is drawn from the product distribution of $p_1 \times p_m$. By independence and the linearity of expectation,

$$\mathbb{E}_{\hat{X}} \left[\sum_{j=1}^{m} f(\hat{X}_j) \right] = \sum_{j=1}^{m} \mathbb{E}_{\hat{X}_j} [f(\hat{X}_j)].$$

First, let's show that $\hat{\pi}_i$ is unbiased:

$$\begin{aligned}
\mathbb{E}_{\hat{X}}[\hat{\pi}_i] &= \mathbb{E}_{\hat{X}}\left[\sum_{j=1}^{m}\frac{w_j}{n_j}\sum_{x\in\hat{X}_j}\frac{p_i(x)f(x)}{q_i(x)}\right] \\
&= \sum_{j=1}^{m}\frac{w_j}{n_j}\mathbb{E}_{\hat{X}_j}\left[\sum_{x\in\hat{X}_j}\frac{p_i(x)f(x)}{q_i(x)}\right] \\
&= \sum_{j=1}^{m}\frac{w_j}{n_j}n_j\mathbb{E}_{p_j}\left[\frac{p_i(x)f(x)}{q_i(x)}\right] \\
&= \sum_{j=1}^{m}w_j\sum_{x\in\mathcal{X}}p_j(x)\frac{p_i(x)f(x)}{q_i(x)} \\
&= \sum_{x\in\mathcal{X}}\sum_{j=1}^{m}w_jp_j(x)\frac{p_i(x)f(x)}{q_i(x)} \\
&= \sum_{x\in\mathcal{X}}q_i(x)\frac{p_i(x)f(x)}{q_i(x)} \\
&= \sum_{x\in\mathcal{X}}p_i(x)f(x) \\
&= \pi_i.
\end{aligned}$$

Now let's compute the variance.

$$\begin{aligned}
\operatorname{Var}\hat{\pi}_i &= \sum_{j=1}^{m}\frac{w_j^2}{n_j}\mathbb{E}_{p_j}\left[\frac{p_i(x)^2f(x)^2}{q_i(x)^2}\right] - \sum_{j=1}^{m}\frac{w_j^2}{n_j}\mathbb{E}_{p_j}\left[\frac{p_i(x)f(x)}{q_i(x)}\right]^2 \\
&= \sum_{j=1}^{m}\frac{w_j^2}{n_j}\mathbb{E}_{p_j}\left[\frac{p_i(x)^2f(x)^2}{q_i(x)^2} - \frac{\pi_{ij}p_i(x)f(x)}{q_i(x)}\right] \\
&= \sum_{j=1}^{m}\frac{w_j^2}{n_j}\mathbb{E}_{p_j}\left[\frac{p_i(x)^2f(x)^2 - \pi_{ij}p_i(x)f(x)q_i(x)}{q_i(x)^2}\right],
\end{aligned}$$

where $\pi_{ij} \stackrel{\text{def}}{=} \mathbb{E}_{p_j}\left[\frac{p_i(x)f(x)}{q_i(x)}\right]$. □

B.2 Estimating recall

In Section 4, we used the fact that the recall of system i, r_i, can be expressed as the recall of i within the pool, ν_i and the recall of the pool itself θ: $r_i = \theta\nu_i$:

$$\nu_i = \mathbb{E}_{x\sim\mathcal{Y}|Y}[g_i(x)] \qquad\qquad \theta = \mathbb{E}_{x\sim\mathcal{Y}}[g(x)],$$

where x is sampled under the distribution $p'(x \mid x \in Y)$ and $p'(x)$ respectively and $g(x) \stackrel{\text{def}}{=} \mathbb{I}[x \in \bigcup_{i=1}^{m} X_i] = \max_{j\in[1,m]} g_j(x)$ is the indicator function for x belonging to the pool.

Ideally, to estimate the pooled recall, ν_i, we need to take expectations with respect to $x \sim Y$. However, we only have samples drawn from individual X_i. To correct for this bias, we'll use a self-normalizing estimator for ν_i:

$$\hat{\nu}_i \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{m}\frac{w_j}{n_j}\sum_{x\in\hat{Y}_j}\frac{p_0(x)g_i(x)}{q(x)}}{\sum_{j=1}^{m}\frac{w_j}{n_j}\sum_{x\in\hat{Y}_j}\frac{p_0(x)}{q(x)}},$$

where $p'(x) \propto p_0(x)$, $q(x) = \sum_{j=1}^m w_j p_j(x)$ and $w_j \geq 0$ are mixture parameters such that $\sum_{j=1}^m w_j = 1$.

The pool recall θ can be estimated as follows:

$$\hat{\theta} \stackrel{\text{def}}{=} \sum_{x \in \hat{Y}_0} g(x),$$

where $g(x) \stackrel{\text{def}}{=} \mathbb{I}\left[x \in \bigcup_{i=1}^m X_i\right] = \max_{j \in [1,m]} g_j(x)$.

Finally, we proposed the following estimator for recall r_i:

$$\hat{r}_i \stackrel{\text{def}}{=} \hat{\theta}\hat{\nu}_i.$$

Let's start by showing that ν_i is unbiased.

Theorem 2 (Statistical properties of $\hat{\nu}_i$)**.** *$\hat{\nu}_i$ is a consistent estimator of ν_i.*

Proof. We have that $p'_Y(x) = \frac{w(x)}{Z_Y}$. While we do not know the value of Z_Y, we can divide both the numerator and denominator of $\hat{\nu}_i$ by this quantity:

$$\begin{aligned}
\hat{\nu}_i &= \frac{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p_0(x) g_i(x)}{Z_Y q(x)}}{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p_0(x)}{Z_Y q(x)}} \\
&= \frac{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x) g_i(x)}{q(x)}}{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x)}{q(x)}}.
\end{aligned}$$

As the number of samples $n_i \to \infty$,

$$\begin{aligned}
\mathbb{E}_X[\hat{\nu}_i] &= \mathbb{E}_X\left[\frac{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x) g_i(x)}{q(x)}}{\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x)}{q(x)}}\right] \\
&= \frac{\mathbb{E}_X\left[\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x) g_i(x)}{q(x)}\right]}{\mathbb{E}_X\left[\sum_{j=1}^m \frac{w_j}{n_j} \sum_{x \in \hat{Y}_j} \frac{p'_Y(x)}{q(x)}\right]}.
\end{aligned}$$

Following similar arguments as in the proof of Theorem 1, the numerator and denominator are unbiased estimators of $\mathbb{E}_{x \sim \mathcal{Y}|Y}[g_i(x)]$ and $\mathbb{E}_{x \sim \mathcal{Y}|Y}[1] = 1$ respectively. Thus,

$$\begin{aligned}
\mathbb{E}_X[\hat{\nu}_i] &= \mathbb{E}_{x \sim \mathcal{Y}|Y}[g_i(x)] \\
&= \nu_i.
\end{aligned}$$

$\hat{\nu}_i$ is an unbiased estimator of ν_i.

□

Finally, we turn to studying $\hat{r}$:

Theorem 3 (Statistical properties of $\hat{r}_i$)**.** *$\hat{r}_i$ is an unbiased estimator of r_i with variance*

$$\operatorname{Var} \hat{r}_i = \theta \operatorname{Var} \hat{\nu}_i + \nu_i \operatorname{Var} \hat{\theta} + \operatorname{Var} \hat{\theta} \operatorname{Var} \hat{\nu}_i.$$

Proof. First, let's show that $r_i = \theta\nu_i$:

$$\begin{aligned}
r_i &\stackrel{\text{def}}{=} \mathbb{E}_{x \sim \mathcal{Y}}[g_i(x)] \\
&= p'(Y_i) \\
&= p'(Y \wedge Y_i) \\
&= p'(Y) p'(Y_i \mid Y) \\
&= \mathbb{E}_{x \sim \mathcal{Y}}[g(x)] \mathbb{E}_{x \sim \mathcal{Y}|Y}[g_i(x)] \\
&= \theta\nu_i.
\end{aligned}$$

From Theorem 2, we have that $\hat{\nu}_i$ is an unbiased estimator of ν_i. It is evident that $\hat{\theta}$ is an unbiased estimator of θ. $\hat{\nu}_i$ and $\hat{\theta}$ are estimated using independent samples ($\hat{Y}$ and $\hat{Y}_0$ respectively), and hence

$$\begin{aligned}\mathbb{E}_{Y_0,Y}[\hat{r}] &= \mathbb{E}_{Y_0,Y}[\hat{\theta}\hat{\nu}_i] \\ &= \mathbb{E}_{Y_0}[\hat{\theta}]\mathbb{E}_{Y}[\hat{\nu}_i] \\ &= \theta\nu_i \\ &= \hat{r}.\end{aligned}$$

By Lemma 1,

$$\operatorname{Var}\hat{r}_i = \theta \operatorname{Var}\hat{\nu}_i + \nu_i \operatorname{Var}\hat{\theta} + \operatorname{Var}\hat{\theta}\operatorname{Var}\hat{\nu}_i.$$

□

B.3 Picking heuristic w_{ij}.

B.4 Picking optimal number of samples for a new system

In Section 4.3, we outlined a method to pick the optimal number of samples to draw and evaluate for a new system: we pick the minimum number of samples n_m required to evaluate system m within a target variance using a conservative estimate of the variance of $\hat{\pi}_m^{\text{(joint)}}$. In particular, we use the following estimate for variance using the result from Theorem 1:

$$\widehat{\operatorname{Var}}\hat{\pi}_m = \sum_{j=1}^{m-1} \frac{w_j^2}{n_j} \sum_{x \in \hat{X}_j} \frac{1}{n_j}\left[\frac{p_i(x)^2 f(x)^2 - \pi_{ij} p_i(x) f(x) q_i(x)}{q_i(x)^2}\right] + \frac{w_m^2}{n_m} \sum_{x \in X_m} p_m(x)\left[\frac{p_m(x)}{q(x)}\right]^2,$$

where the first $m-1$ terms are an empirical estimate of variance and the last term is an upper bound on the variance. We note that the actual output of each system, X_j, and the samples drawn from previous systems, $\hat{X}_j$, is known. Thus, the only variable in computing $\widehat{\operatorname{Var}}\hat{\pi}_m$ is n_m. Furthermore, $\widehat{\operatorname{Var}}\hat{\pi}_m$ is a monotonically decreasing in n_m, so we can easily solve for the minimum number of samples required to estimate $\hat{\pi}_m^{\text{(joint)}}$ within a confidence interval ϵ by using the bisection method (Burden and Faires, 1985).

C Basic probability lemmas

Lemma 1 (Mean and variance of the product of two random variables)**.** *Let x and y be two independent random variables with means μ_x and μ_y, and variances σ_x^2 and σ_y^2. Then, the estimator $z = xy$ has mean $\mu_x\mu_y$ and variance*

$$\sigma_z^2 = \sigma_x^2\sigma_y^2 + \mu_x^2\sigma_y^2 + \sigma_x^2\mu_y^2.$$

Proof. If x and y are independent, $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y]$. Thus $\mathbb{E}[z] = \mu_x\mu_y$.

The variance of z can be calculated as follows:

$$\begin{aligned}\operatorname{Var}(z) &= \mathbb{E}[z^2] - \mathbb{E}[z]^2 \\ &= \mathbb{E}[(xy)^2] - \mathbb{E}[xy]^2 \\ &= \mathbb{E}[x^2]\mathbb{E}[y^2] - \mathbb{E}[x]^2\mathbb{E}[y]^2 \\ &= (\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2) - \mu_x^2\mu_y^2 \\ &= \sigma_x^2\sigma_y^2 + \mu_x^2\sigma_y^2 + \sigma_x^2\mu_y^2 + \mu_x^2\mu_y^2 - \mu_x^2\mu_y^2 \\ &= \sigma_x^2\sigma_y^2 + \mu_x^2\sigma_y^2 + \sigma_x^2\mu_y^2.\end{aligned}$$

□

Lemma 2 (Mean and variance of the ratio of two random variables)**.** *Let x and y be two random variables such that y is strictly positive (i.e. $y > 0$) with means μ_x and μ_y, variances σ_x^2 and σ_y^2. Then, the first-order Taylor approximation of $z = x/y$ has mean μ_x/μ_y. Furthermore, if x and y are the mean of a n_x and n_y independent random variables, the approximation error of using the first-order approximation goes to 0 as $n_x, n_y \to \infty$.*

Proof. This is a standard result in statistics. For completeness, we provide a proof below.

Let $f(x, y) = \frac{x}{y}$. Even if x and y are independent, $\mathbb{E}[f(x, y)]$ is not necessarily equal to $f(\mathbb{E}[x], \mathbb{E}[y])$. However, taking a first-order Taylor expansion around (μ_x, μ_y), we get

$$\begin{aligned}
\mathbb{E}[f(x, y)] &\approx f(\mu_x, \mu_y) + f'_x(\mu_x, \mu_y)\mathbb{E}[x - \mu_x] + f'_y(\mu_x, \mu_y)\mathbb{E}[y - \mu_y] \\
&= \frac{\mu_x}{\mu_y}.
\end{aligned}$$

We note that if x and y are the sum of independent random variables, then by the central limit theorem all moments of x and y greater than 1 go to 0 as $n_x, n_y \to \infty$.

$\square$