Active Estimation of F-Measures (Online Appendix)

Christoph Sawade, Niels Landwehr, and Tobias Scheffer
University of Potsdam
Department of Computer Science
August-Bebel-Strasse 89, 14482 Potsdam, Germany
{sawade, landwehr, scheffer}@cs.uni-potsdam.de

A. Proof of Lemma 1

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be drawn according to $q(\mathbf{x})p(y|\mathbf{x})$. Let $\hat{G}^0_{n,q} = \sum_{i=1}^n v_i \ell_i w_i$ and $W_n = \sum_{i=1}^n v_i w_i$ with $v_i = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$, $w_i = w(\mathbf{x}_i, y_i, f_\theta)$ and $\ell_i = \ell(f_\theta(\mathbf{x}_i), y_i)$. Furthermore, we define $\mu_G = \mathbb{E}\left[\frac{1}{n}\hat{G}^0_{n,q}\right]$ and $\mu_W = \mathbb{E}\left[\frac{1}{n}W_n\right]$. Note that $G = \frac{\mathbb{E}\left[\frac{1}{n}\hat{G}^0_n\right]}{\mathbb{E}\left[\frac{1}{n}W_n\right]} = \frac{\mu_G}{\mu_W}$.

Starting with the definition of the bias (Equation 1), we derive

$$\left|Bias\left[\hat{G}_{n,q}\right]\right| = \left|\mathbb{E}\left[\hat{G}_{n,q}\right] - G\right| \tag{1}$$

$$= \left|\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)\sum_{j=0}^\infty \frac{\left(\frac{1}{n}W_n - \mu_W\right)^j}{\mu_W^{j+1}}(-1)^j\right] - G\right| \tag{2}$$

$$= \left|\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)\sum_{j=2}^\infty \frac{\left(\frac{1}{n}W_n - \mu_W\right)^j}{\mu_W^{j+1}}(-1)^j\right]\right.$$
$$\left. - \frac{\text{Cov}\left[\frac{1}{n}\hat{G}^0_{n,q}, \frac{1}{n}W_n - \mu_W\right] + \mathbb{E}\left[\frac{1}{n}\hat{G}^0_{n,q}\right]\mathbb{E}\left[\frac{1}{n}W_n - \mu_W\right]}{\mu_W^2}\right| \tag{3}$$

$$= \left|\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)\sum_{j=2}^\infty \frac{\left(\frac{1}{n}W_n - \mu_W\right)^j}{\mu_W^{j+1}}(-1)^j\right] - \frac{\text{Cov}\left[\hat{G}^0_{n,q}, W_n\right]}{n^2\mu_W^2}\right|$$

$$= \left|\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)\sum_{j=2}^\infty \frac{\left(\frac{1}{n}W_n - \mu_W\right)^j}{\mu_W^{j+1}}(-1)^j\right] - \frac{1}{n\mu_W^2}\text{Cov}\left[v_1\ell_1 w_1, v_1 w_1\right]\right| \tag{4}$$

$$\leq \sum_{j=2}^\infty \frac{1}{\mu_W^{j+1}}\sqrt{\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)^2\right]\mathbb{E}\left[\left(\frac{1}{n}W_n - \mu_W\right)^{2j}\right]} + \frac{1}{n}\cdot const \tag{5}$$

$$\leq \sum_{j=2}^\infty \frac{1}{\mu_W^{j+1}}\sqrt{const \cdot \frac{1}{n^j}\mathbb{E}\left[(v_1 w_1 - \mu_W)^{2j}\right]} + \frac{1}{n}\cdot const \tag{6}$$

$$= \sum_{j=2}^\infty const \cdot \frac{\sqrt{\mathbb{E}\left[(v_1 w_1 - \mu_W)^{2j}\right]}}{\mu_W\left(\mu_W\sqrt{n}\right)^j} + \frac{1}{n}\cdot const. \tag{7}$$

In Equation 2, we have expressed the quotient $\hat{G}_{n,q} = \frac{\frac{1}{n}\hat{G}^0_{n,q}}{\frac{1}{n}W_n}$ by a multivariate Taylor expansion around the expectations $\mathbb{E}[\frac{1}{n}\hat{G}^0_{n,q}]$ and $\mathbb{E}[\frac{1}{n}W_n]$. In Equation 3, G cancels out with the summand for $j = 0$. In Equation 4, we exploit that $\text{Cov}\left[v_i\ell_i w_i, v_j\ell_j\right] = 0$ for $i \neq j$. Equation 5 follows from the

Cauchy-Schwarz inequality for expectations. To derive Equation 6, we consider the $2j$-th derivative of the moment-generating function of $\frac{1}{n}W_n - \mu_W$. A straightforward calculation shows that

$$\mathbb{E}\left[\left(\frac{1}{n}W_n - \mu_W\right)^{2j}\right] \leq const \cdot \frac{1}{n^j}\,\mathbb{E}\left[(v_1 w_1 - \mu_W)^{2j}\right]$$

where $const$ denotes terms independent of n. The term $\mathbb{E}\left[\left(\frac{1}{n}\hat{G}^0_{n,q}\right)^2\right]$ from Equation 5 is also subsumed into $const$ as it is bounded independently of n. Finally, we note that for sufficiently large n, the series on the right-hand side of Equation 7 converges and is of order $\frac{1}{n}$. Thus, there exists a constant C such that for all $n \in \mathbb{N}$ it holds that

$$\left|Bias\left[\hat{G}_{n,q}\right]\right| \leq \frac{C}{n}.$$

$\square$

B. Comprehensive Empirical Results in the Text Classification Domain

Table 1 lists the true one-vs-rest precision, $F_{0.5}$-measure, recall, and accuracy of the actively trained model for the ten classes in the Reuters-21578 text classification domain.

Figure 1 shows the estimation error of $active_F$, $passive$, and $active_{err}$ over number of labeled data, for precision, $F_{0.5}$ and recall estimates and all ten classes in the text classification domain.

Table 1: Class ratios and model quality for active learned model for ten most frequent occurring topics in Reuters corpus

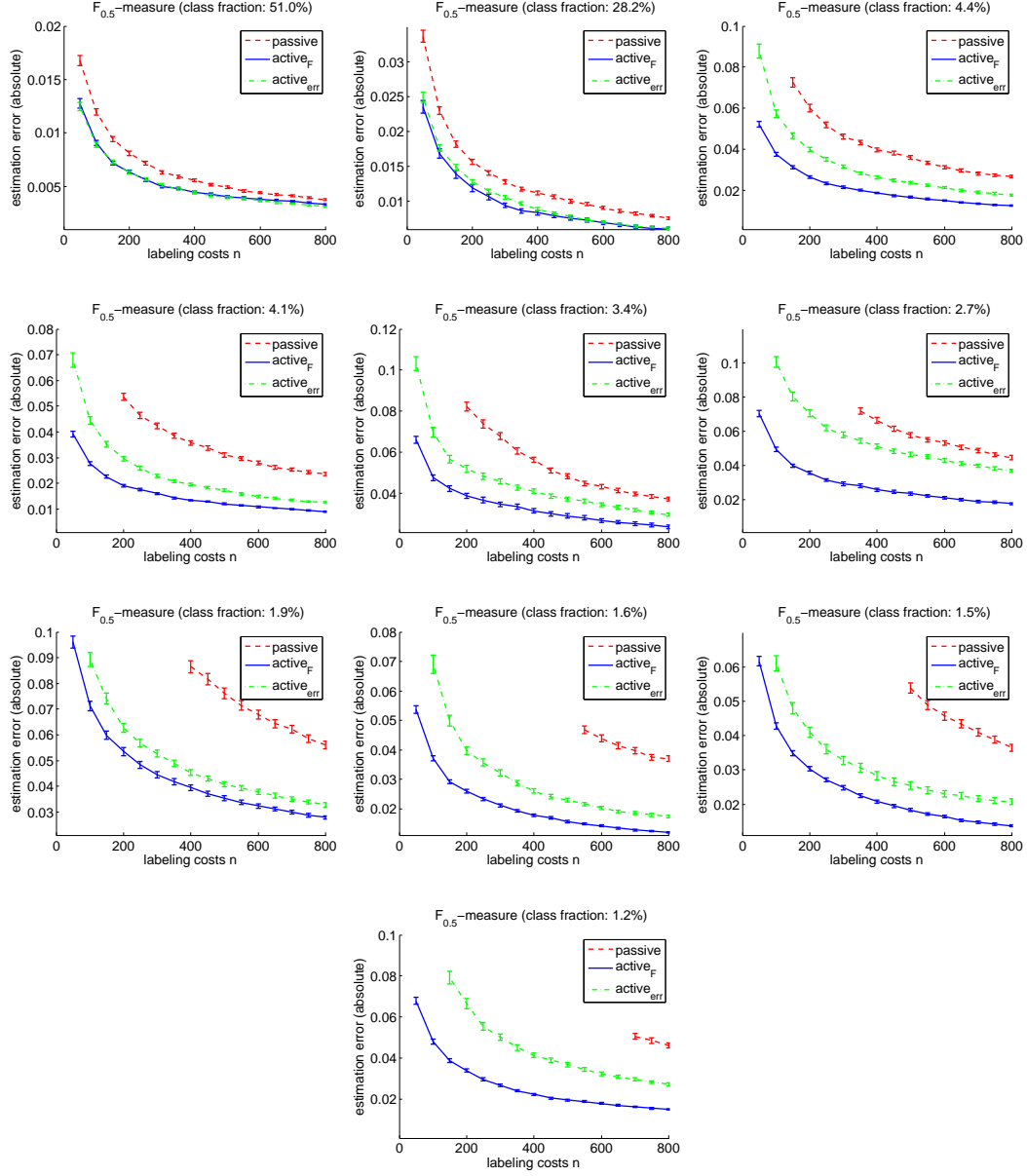| topic	earn	acq	crude	trade	money-fx	interest	ship	sugar	coffee	gold
class ratio	0.510	0.280	0.044	0.041	0.034	0.027	0.020	0.016	0.015	0.012
Acc	0.978	0.974	0.994	0.995	0.991	0.992	0.994	0.998	0.999	0.998
Prec	0.976	0.942	0.972	0.937	0.872	0.851	0.886	0.972	1.000	0.954
Rec	0.980	0.966	0.879	0.940	0.861	0.843	0.820	0.921	0.900	0.911
$F_{0.5}$	0.978	0.954	0.923	0.938	0.867	0.847	0.850	0.946	0.947	0.932

Figure 1: Text Classification: Estimation error over number of labeled data, for recall, $F_{0.5}$ and precision estimates. Classes (from left to right and top to bottom): "earn", "acq", "crude", "trade", "money-fx", "interest", "ship", "sugar", "coffee" and "gold". Error bars indicate the standard error.

C. Detailed Empirical Results in the Digit Recognition Domain

We consider a digit recognition domain in which training and testing distributions diverge because the data originate from different sources. To realize this scenario, a digit recognition model is trained on the USPS data set and evaluated on the MNIST data set. We use a version of MNIST prepared by Sam Roweis. MNIST images were rescaled from 28×28 to 16×16 pixels to match the resolution of USPS and the bounding box was recomputed. Images are represented by their 256 numeric pixel values. There are 10 classes, $11,000$ training and $70,000$ test instances. We train a single multi-class model using an RBF kernel. We evaluate one-versus-rest F-measures for each class, resulting in ten different evaluation tasks.

Table 2 lists the true one-vs-rest precision, $F_{0.5}$-measure, recall, and accuracy of the trained model on the ten different estimation problems corresponding to digits 0 to 9.

Figures 2 and 3 show the estimation error of $active_F$, $passive$, and $active_{err}$ over number of labeled data, for precision, $F_{0.5}$ and recall estimates and the ten different one-vs-rest estimation problems corresponding to digits 0 to 9 in the digit recognition domain.

Table 2: Class ratios and model quality of digit recognition model on MNIST

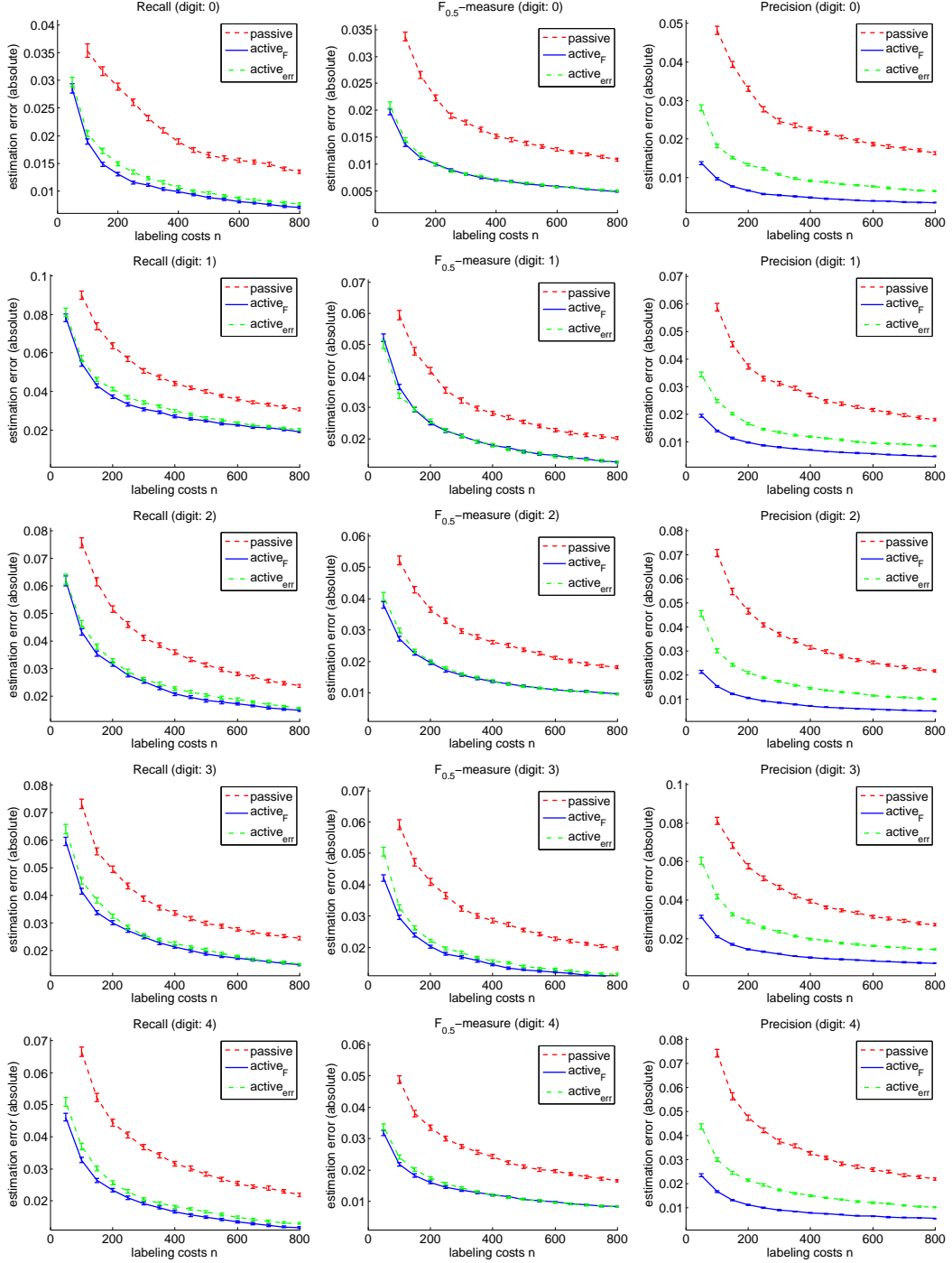| digit	0	1	2	3	4	5	6	7	8	9
class ratio	0.097	0.113	0.099	0.102	0.098	0.092	0.098	0.104	0.096	0.099
Acc	0.995	0.978	0.986	0.980	0.987	0.980	0.991	0.978	0.962	0.977
Prec	0.966	0.955	0.937	0.885	0.930	0.932	0.951	0.972	0.744	0.849
Rec	0.978	0.844	0.918	0.920	0.937	0.840	0.962	0.810	0.926	0.940
$F_{0.5}$	0.972	0.896	0.927	0.902	0.933	0.884	0.956	0.883	0.825	0.891

Figure 2: Digit Recognition: Estimation error over number of labeled data for recall, $F_{0.5}$ and precision estimates for digits 0 to 4 (from top to bottom). Error bars indicate the standard error.
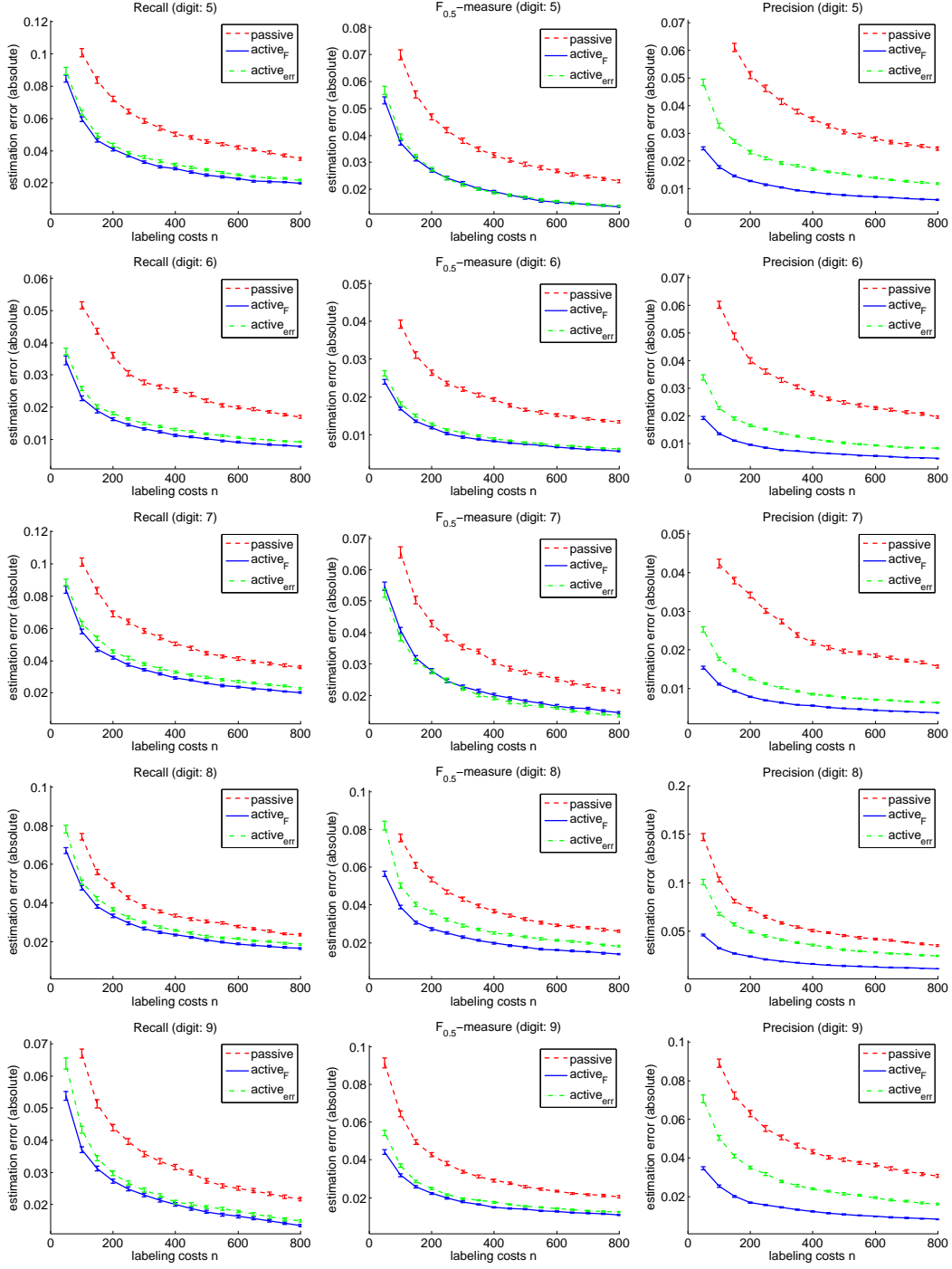
Figure 3: Digit Recognition: Estimation error over number of labeled data for recall, $F_{0.5}$ and precision estimates for digits 5 to 9 (from top to bottom). Error bars indicate the standard error.