Supplemental Material

9. Additional results

9.1. Alternate metrics

Here we report tabular results for several alternate metrics, providing additional points of comparison to supplement our main results of cumulative regret at step 100 (Tab. 1).

9.1.1. Variance between seeds

	Task	Random Sampling	Uncertainty	Active Testing	VMA	Model Selector	CODA (Ours)
DomainNet126	real→sketch	147.1± 38.5	197.7± 13.0	269.8± 119.8	119.2± 51.9	**88.8± 9.8**	101.2± 0.0
	real→painting	143.6± 37.1	167.9± 6.8	118.9± 68.8	139.9± 53.3	92.1± 1.6	**87.2± 2.8**
	real→clipart	237.7± 88.7	217.6± 23.7	**152.1± 40.0**	206.9± 100.7	153.2± 9.5	231.7± 0.0
	sketch→real	280.4± 164.4	252.7± 11.8	236.4± 160.5	273.4± 132.7	268.4± 7.5	**11.9± 0.0**
	sketch→painting	156.6± 72.6	181.9± 7.3	237.7± 151.7	157.1± 60.9	173.9± 7.1	**13.0± 0.0**
	sketch→clipart	228.2± 74.5	224.6± 43.2	162.0± 83.6	227.9± 85.6	**38.1± 9.2**	432.4± 1.9
	painting→real	364.8± 164.4	224.0± 7.2	215.6± 94.4	358.9± 193.6	293.4± 20.3	**2.4± 0.0**
	painting→sketch	179.3± 53.5	440.7± 32.9	202.5± 79.7	211.5± 73.3	209.8± 5.2	**72.3± 0.2**
	painting→clipart	222.7± 144.5	296.6± 6.1	251.4± 111.8	271.8± 116.9	73.2± 3.2	**43.1± 0.1**
	clipart→real	322.1± 127.8	177.1± 34.6	159.1± 55.6	306.4± 181.7	72.7± 15.3	**25.3± 0.0**
	clipart→sketch	247.2± 143.0	924.8± 22.7	282.6± 186.5	291.3± 163.9	532.7± 76.1	**51.3± 0.0**
	clipart→painting	147.0± 46.2	162.9± 13.3	222.6± 100.5	167.9± 100.2	131.7± 2.5	**122.2± 0.1**
WILDS	iwildcam	287.0± 65.1	380.4± 7.1	392.1± 194.1	440.6± 103.6	459.0± 56.0	**201.7± 0.0**
	camelyon	175.2± 81.7	311.6± 11.9	206.1± 121.1	**160.1± 76.8**	198.3± 90.6	288.7± 0.0
	fmow	189.7± 66.8	191.9± 8.4	153.0± 38.7	189.2± 62.6	211.7± 28.1	**70.0± 0.2**
	civilcomments	140.9± 75.5	**13.3± 2.6**	76.7± 119.0	50.1± 29.3	125.3± 59.6	318.6± 0.0
MSV	cifar10→4070	410.6± 231.1	629.7± 5.9	399.9± 207.5	476.5± 114.4	567.2± 23.9	**58.7± 0.0**
	cifar10→5592	346.2± 147.4	281.7± 21.3	154.9± 68.9	383.7± 141.9	90.9± 12.3	**74.1± 0.0**
	pacs	216.6± 86.8	**6.9± 5.4**	101.8± 55.9	116.5± 57.5	97.9± 6.0	57.9± 0.0
GLUE	cola	368.1± 80.6	**169.5± 27.1**	239.8± 162.4	317.8± 191.8	207.8± 61.1	2226.7± 0.0
	mnli	237.4± 132.2	80.8± 27.9	234.2± 105.2	312.8± 111.5	148.5± 114.4	**23.5± 0.0**
	qnli	222.7± 68.5	231.6± 76.2	246.6± 112.4	283.0± 78.8	185.7± 89.6	**120.4± 0.0**
	qqp	136.7± 20.9	388.9± 492.8	127.8± 16.5	169.2± 114.1	186.8± 137.5	**4.8± 0.0**
	rte	375.7± 184.6	390.3± 68.4	424.4± 244.4	674.7± 186.7	**243.6± 73.2**	283.8± 0.0
	sst2	219.9± 108.1	89.6± 39.6	174.9± 54.3	373.6± 128.6	202.0± 131.4	**51.7± 0.0**
	mrpc	318.2± 100.3	301.1± 64.4	235.2± 36.0	332.3± 156.0	173.1± 47.7	**49.0± 0.0**

Table 2. **Main results with variance between seeds: Cumulative regret at step 100.** Same as Tab. 1 but ± one standard deviation between runs with different random seeds (5 random seeds used). Standard deviation of 0.0 (as in many CODA runs) indicates method is not stochastic for that task, *e.g.* because of asymmetric priors resulting in deterministic selection.

9.1.2. Instantaneous regret

We report the average instantaneous regret of each method at steps 50 and 100 in Tab. 3 and Tab. 4, respectively. These results are more influenced by stochasticity than cumulative regret, *i.e.* instantaneous regret may fluctuate widely between steps (see Fig. 8).

9.1.3. Success rate

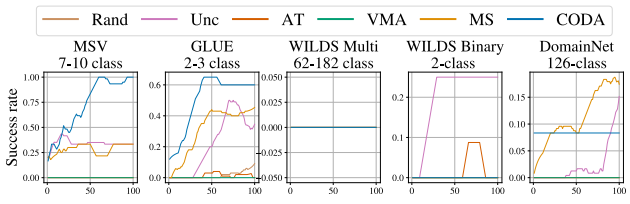

Figure 7. **"Success rate" of each model in selecting the *absolute best* model at each time step.** Mean over 5 random seeds. In all datasets, several methods have not yet selected the absolute best model by time step 100.

	Task	Random Sampling	Uncertainty	Active Testing	VMA	Model Selector	CODA (Ours)
DomainNet126	real→sketch	1.3± 0.6	2.0± 0.2	0.8± 0.6	1.3± 1.2	**0.6± 0.0**	1.0± 0.0
	real→painting	1.2± 0.4	1.2± 0.5	1.4± 1.4	1.5± 1.0	**0.9± 0.3**	1.2± 0.0
	real→clipart	1.6± 1.8	0.5± 0.6	1.0± 0.9	0.8± 0.8	**0.2± 0.1**	2.3± 0.0
	sketch→real	2.0± 3.0	0.4± 0.2	3.0± 3.0	2.6± 3.1	5.4± 0.0	**0.1± 0.0**
	sketch→painting	1.1± 1.3	0.9± 1.1	2.2± 2.3	0.9± 0.6	5.5± 0.0	**0.1± 0.0**
	sketch→clipart	1.8± 1.4	0.2± 0.2	1.6± 1.6	2.1± 0.8	**0.0± 0.0**	4.8± 0.0
	painting→real	4.1± 2.9	2.1± 0.7	1.3± 0.9	2.3± 1.1	4.6± 4.6	**0.0± 0.0**
	painting→sketch	1.5± 1.8	4.3± 0.0	2.3± 1.5	1.9± 1.5	1.8± 0.7	**0.9± 0.0**
	painting→clipart	1.9± 2.3	2.8± 0.0	3.2± 1.7	2.8± 1.5	0.7± 0.0	**0.4± 0.0**
	clipart→real	2.6± 1.7	0.4± 0.0	0.2± 0.4	2.8± 1.8	**0.1± 0.0**	0.3± 0.0
	clipart→sketch	3.0± 2.3	8.7± 2.9	0.8± 0.7	2.6± 2.0	5.4± 0.0	**0.4± 0.0**
	clipart→painting	1.6± 0.9	1.1± 0.0	2.6± 1.9	1.3± 1.5	**1.0± 0.1**	1.2± 0.0
WILDS	iwildcam	2.1± 0.9	3.0± 0.0	4.5± 2.3	6.1± 1.9	4.6± 1.3	**0.9± 0.0**
	camelyon	**0.5± 0.2**	3.5± 0.0	2.1± 1.7	0.9± 1.0	1.5± 0.8	0.6± 0.0
	fmow	1.4± 1.2	**0.6± 0.0**	1.2± 0.7	1.6± 1.0	1.0± 0.0	0.8± 0.0
	civilcomments	0.9± 1.1	**0.0± 0.0**	0.8± 1.5	0.3± 0.3	1.6± 2.7	3.3± 0.0
MSV	cifar10→4070	5.5± 4.1	7.7± 0.0	5.3± 3.4	4.7± 2.6	7.5± 0.4	**0.0± 0.0**
	cifar10→5592	3.1± 2.3	3.2± 0.0	0.2± 0.2	2.6± 1.7	**0.0± 0.0**	**0.0± 0.0**
	pacs	2.2± 1.9	**0.0± 0.0**	0.6± 0.8	0.8± 0.7	1.4± 0.0	**0.0± 0.0**
GLUE	cola	3.6± 1.3	2.0± 0.9	2.3± 1.9	2.5± 1.9	1.4± 1.0	**0.4± 0.0**
	mnli	2.6± 3.0	**0.1± 0.0**	1.0± 1.7	2.6± 1.6	**0.1± 0.1**	**0.1± 0.0**
	qnli	1.3± 1.3	0.9± 1.2	1.0± 1.4	1.7± 1.5	0.1± 0.2	**0.0± 0.0**
	qqp	1.4± 0.2	1.3± 2.8	1.1± 0.1	1.0± 0.6	0.3± 0.4	**0.0± 0.0**
	rte	3.2± 4.4	**0.0± 0.0**	3.7± 3.6	5.9± 3.9	**0.0± 0.0**	**0.0± 0.0**
	sst2	1.9± 2.5	0.2± 0.0	0.1± 0.1	3.1± 3.5	**0.0± 0.0**	**0.0± 0.0**
	mrpc	2.0± 2.4	1.6± 1.9	1.6± 1.2	2.8± 2.4	0.8± 0.5	**0.5± 0.0**

Table 3. **Average instantaneous regret at step 50.** Mean and standard deviation reported over 5 random seeds.

	Task	Random Sampling	Uncertainty	Active Testing	VMA	Model Selector	CODA (Ours)
DomainNet126	real→sketch	1.2± 1.1	**0.0± 0.0**	2.2± 2.2	1.0± 0.6	0.8± 0.5	1.0± 0.0
	real→painting	1.1± 0.6	0.6± 0.0	0.8± 0.2	0.9± 0.6	1.3± 0.3	**0.2± 0.0**
	real→clipart	0.6± 0.4	1.0± 0.0	0.6± 0.3	1.5± 2.0	**0.2± 0.0**	2.3± 0.0
	sketch→real	1.7± 2.8	0.5± 0.0	0.7± 1.1	0.2± 0.1	**0.1± 0.0**	**0.1± 0.0**
	sketch→painting	0.7± 0.7	**0.0± 0.0**	0.8± 0.8	0.8± 0.9	0.1± 0.1	0.1± 0.0
	sketch→clipart	0.9± 0.8	0.2± 0.3	0.9± 0.5	2.1± 2.0	**0.0± 0.0**	0.5± 0.0
	painting→real	3.3± 2.4	1.2± 0.0	1.1± 0.9	1.4± 0.8	1.2± 0.0	**0.0± 0.0**
	painting→sketch	0.6± 0.5	0.7± 0.0	0.7± 0.7	1.3± 1.9	1.7± 0.0	**0.5± 0.0**
	painting→clipart	2.0± 1.7	0.5± 0.2	2.3± 1.7	1.7± 1.6	**0.3± 0.0**	0.4± 0.0
	clipart→real	1.4± 1.6	0.6± 0.0	0.7± 0.8	0.7± 0.6	**0.1± 0.0**	0.3± 0.0
	clipart→sketch	1.6± 2.0	5.3± 2.1	1.0± 0.7	**0.6± 0.4**	**0.6± 0.0**	**0.6± 0.0**
	clipart→painting	**0.6± 0.5**	1.1± 0.0	1.4± 0.7	1.1± 0.3	1.0± 0.1	1.1± 0.0
WILDS	iwildcam	1.4± 1.7	3.0± 0.0	3.2± 3.0	2.8± 2.5	4.1± 1.6	**0.9± 0.0**
	camelyon	**0.5± 0.9**	3.3± 0.7	0.9± 1.5	1.4± 1.5	1.2± 0.9	0.6± 0.0
	fmow	1.4± 0.9	1.0± 0.8	0.7± 0.3	1.2± 0.5	1.5± 1.4	**0.4± 0.0**
	civilcomments	0.7± 1.2	**0.0± 0.0**	0.3± 0.4	0.2± 0.4	0.3± 0.2	2.8± 0.0
MSV	cifar10→4070	1.2± 1.3	5.7± 0.0	2.8± 2.9	3.2± 1.2	6.3± 1.9	**0.0± 0.0**
	cifar10→5592	1.0± 1.4	3.2± 0.0	0.3± 0.2	1.9± 1.6	**0.0± 0.0**	**0.0± 0.0**
	pacs	1.1± 0.7	**0.0± 0.0**	0.4± 0.6	0.6± 0.5	0.8± 0.5	**0.0± 0.0**
GLUE	cola	2.9± 1.8	**0.2± 0.5**	1.4± 1.2	2.8± 2.7	0.4± 0.0	56.0± 0.0
	mnli	0.3± 0.3	**0.1± 0.0**	0.4± 0.2	0.7± 0.8	0.2± 0.1	0.2± 0.0
	qnli	0.7± 1.1	**0.0± 0.0**	1.6± 2.1	1.5± 1.9	0.1± 0.2	0.4± 0.0
	qqp	1.0± 0.6	0.2± 0.3	1.0± 1.1	0.9± 0.7	0.7± 0.9	**0.0± 0.0**
	rte	1.2± 2.6	**0.0± 0.0**	0.8± 1.8	4.3± 4.2	0.0± 0.0	**0.0± 0.0**
	sst2	0.8± 0.7	0.1± 0.0	1.0± 0.9	2.0± 2.9	**0.0± 0.0**	**0.0± 0.0**
	mrpc	0.2± 0.2	0.2± 0.1	0.4± 0.5	2.6± 2.9	**0.0± 0.0**	0.5± 0.0

Table 4. **Average instantaneous regret at step 100.** Mean and standard deviation over 5 random seeds.

9.2. Unsupervised model selection results

Our method defines a prior over model performance that creates a strong starting point for active model selection. This can be used in isolation, without active label collection, to perform unsupervised model selection. We compare against five existing methods for model selection in unsupervised domain adaptation:

Source validation Model selection is performed using val-

	Task	Source val	DEV	Entropy	BNM	EnsV	CODA
DomainNet126	real → sketch	0.5	4.6	3.1	3.1	**0.4**	1.0
	real → clip	4.5	49.3	0.3	6.5	2.8	**2.3**
	real → paint	1.3	34.7	2.3	2.0	**0.2**	1.2
	sketch → real	4.7	6.3	9.1	8.5	4.5	**0.1**
	sketch → clip	2.4	13.4	5.9	6.4	**3.0**	4.8
	sketch → paint	3.8	3.5	3.2	3.2	1.7	**0.1**
	clip → real	1.5	6.3	5.6	5.6	0.5	**0.3**
	clip → sketch	2.8	4.9	5.1	4.9	2.2	**0.4**
	clip → paint	3.1	7.8	4.2	4.2	4.2	**1.2**
	paint → real	0.0	36.6	3.3	3.3	**0.1**	**0.1**
	paint → sketch	0.7	26.1	4.3	4.3	3.3	**0.9**
	paint → clip	2.9	16.2	6.3	6.3	1.1	**0.4**
WILDS	iwildcam	9.8	-	-	-	**0.9**	6.1
	camelyon	**4.3**	-	-	-	13.1	11.3
	fmow	0.8	-	-	-	0.9	**0.8**
	civilcomments	-	-	-	-	**3.8**	4.7
MSV	cifar10-low	-	-	-	-	**2.3**	**2.3**
	cifar10-high	-	-	-	-	**4.9**	**4.9**
	pacs	-	-	-	-	**0.4**	**0.4**
GLUE	cola	-	-	-	-	5.3	**5.0**
	mnli	-	-	-	-	**3.0**	**3.0**
	qnli	-	-	-	-	**3.3**	**3.3**
	qqp	-	-	-	-	**1.1**	**1.1**
	rte	-	-	-	-	**14.8**	**14.8**
	sst2	-	-	-	-	**3.6**	**3.6**
	mrcp	-	-	-	-	**1.0**	**1.0**

Table 5. **Unsupervised model selection results.** We report regret at step 0 (lower is better) for all methods on all tasks. Best method for each task is in bold. CODA matches or exceeds state-of-the-art performance on 20 out of 26 tasks. Note that because models/predictions for MSV and GLUE are black-box/one-hot (as in ModelSelector [45]), the only comparison we are able to make is to EnsV.

idation accuracy on "source" data, *i.e*. using a validation set from the same distribution as the training set.

Target entropy [43] Shannon entropy is computed on prediction scores on the test set. Model selection is performed by selecting the model with the lowest entopy, *i.e*. the model that is most confident in its test predictions.

Deep embedded validation [70] Computes a classification loss for each source validation sample, and weights each loss based on a computed probability that the sample "belongs to" the test domain. This probability comes from a separate domain classifier trained on source and test data.

Batch nuclear norm [11] Performs singular value decomposition on the prediction matrix. Model selection is performed by selecting the model with the minmium nuclear norm (*i.e*. minimum sum of singular values).

EnsV [20] All models in the hypothesis set are ensembled. To perform model selection, accuracy is estimated for each model with respect to the ensemble's predictions.

9.3. Unaggregated results on all tasks

In Fig. 8 we visualize regret and cumulative regret at every step for all baselines on all tasks.

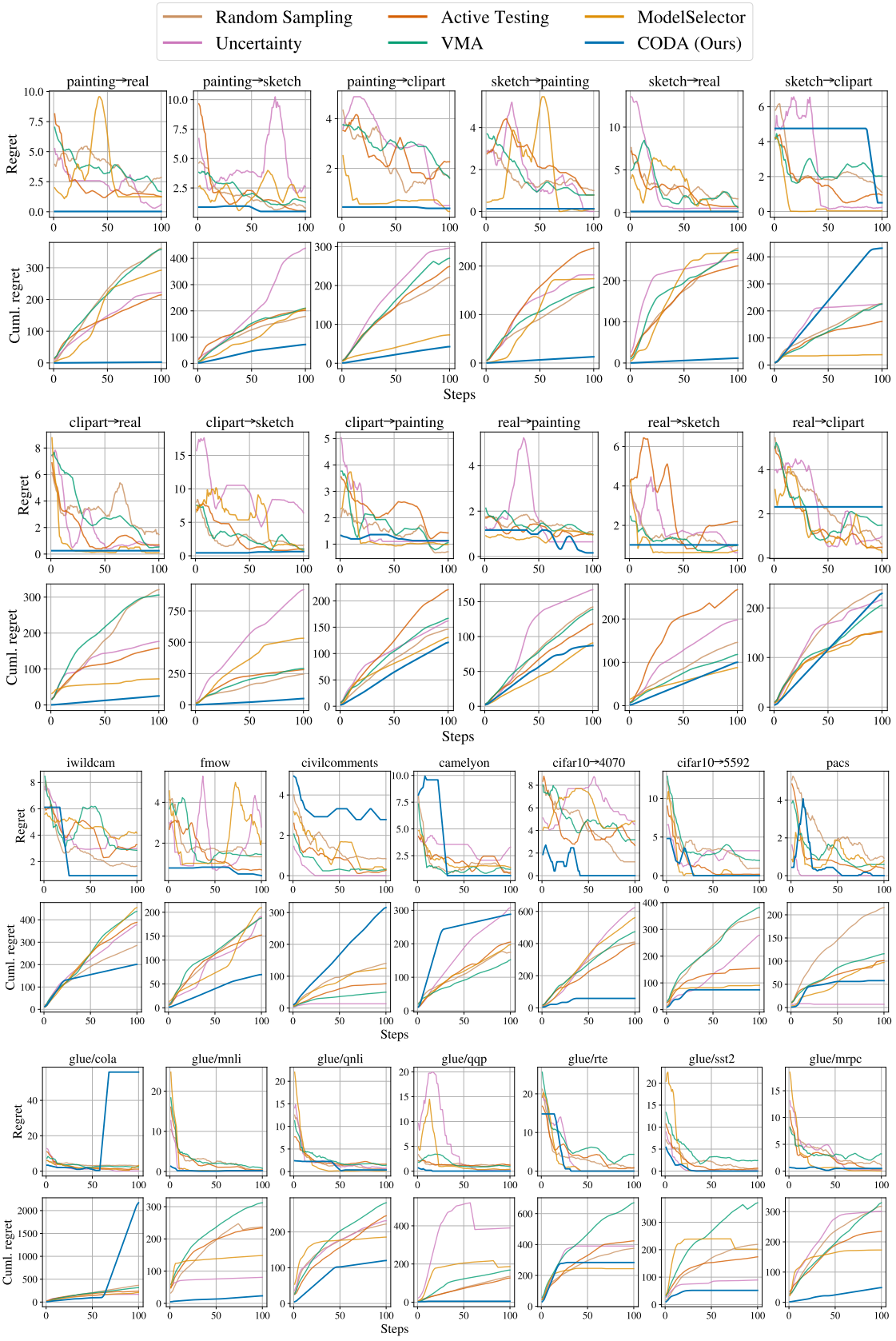

Figure 8. Results on all benchmarks.

10. Implementation details

10.1. Dawid-Skene data generating process

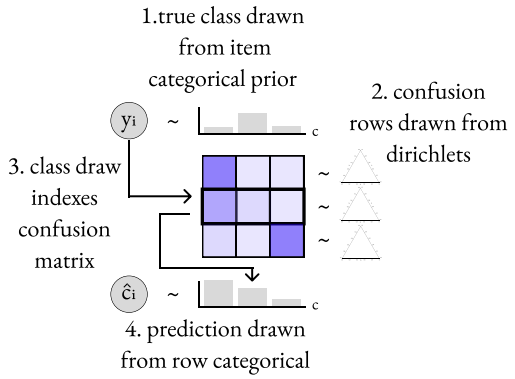

Figure 9. A visual depiction of the Dawid-Skene [14] data generating process that we adapt to active model selection. See Sec. 4.1 of the main paper for more details.

We visualize the data generating process of the Bayesian implementation of the Dawid-Skene model in Fig. 9, as described in Sec. 4.1. We repeat the text here for easy reference.

The data generating process proceeds as follows:

1. Each data point's true class label y_i is drawn randomly from per-data-point prior distributions over which class that data point could be, $y_i \sim \text{Cat}(\pi(x_i))$.
2. Each row of the classifier's confusion matrix is drawn randomly from per-row distributions, $M_{k,c,\cdot} \sim \theta_{k,c}$, where $\theta_{k,c}$ is the prior distribution over what the row of the confusion matrix could be. To accommodate Bayesian updates, we initialize each $\theta_{k,c}$ to be a Dirichlet prior.
3. The sampled true class indexes into the corresponding row of the classifier's confusion matrix, M_{k,y_i}.
4. The classifier's prediction for that data point is sampled from the distribution over that row's cells, $\hat{c}_{k,i} \sim \text{Cat}(M_{k,y_i})$.

10.2. Computing P_{Best}

We illustrate visually the computation of P_{Best} from Sec. 4.3 in Fig. 10 in the simplified case of two models. To compute the probability that Model 1 is best, we integrate over all possible accuracy values that Model 1 *could* have. For every possible accuracy, we compute the probability that Model 1 has that accuracy (defined by its PDF f), multiplied that the probability Model 2 has accuracy *less* than that value (defined by its CDF F).

11. Data and model details

We provide more details about the datasets and models in our benchmarking suite in Tab. 6, Tab. 7, and Tab. 8.

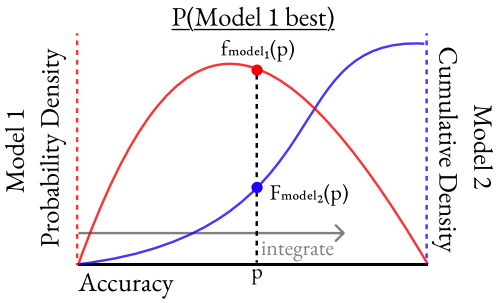

Figure 10. A visual depiction of the integration technique used to construct our distribution over which model is best, P_{Best}. See Sec. 10.2 of the supplemental and Sec. 4.3 of the main paper for more details.

DomainNet126 Task	Num. Classes	Num. Checkpoints	Test set size
sketch_painting	126	10 algs 20 epochs = 200	3086
sketch_real	126	10 algs 20 epochs = 200	20939
sketch_clipart	126	10 algs 20 epochs = 200	5611
clipart_painting	126	10 algs 20 epochs = 200	3086
clipart_sketch	126	10 algs 20 epochs = 200	7313
clipart_real	126	10 algs 20 epochs = 200	20939
painting_sketch	126	10 algs 20 epochs = 200	7313
painting_real	126	10 algs 20 epochs = 200	20939
painting_clipart	126	10 algs 20 epochs = 200	5611
real_painting	126	10 algs 20 epochs = 200	3086
real_sketch	126	10 algs 20 epochs = 200	7313
real_clipart	126	10 algs 20 epochs = 200	5611

Table 6. **Dataset details: DomainNet126.** For each transfer task, we first train a "source-only" model on the source domain. We then train 10 UDA models for each transfer task (source domain → target domain) using the Powerful Benchmarker codebase [43].

WILDS Task	Num. Classes	Num. Checkpoints		Test set size	Regret w/ val.
RxRx1	1139	4 algs 18 epochs	= 72	34,432	0.1
Amazon	5	4 algs 3 epochs	= 12	100,050	0.1
CivilComments	2	4 algs 5 epochs	= 20	133,782	N/A
fMoW	62	4 algs 60 epochs	= 240	22,108	0.8
iWildCam	182	4 algs 12 epochs	= 48	42,791	9.8
Camelyon17	2	4 algs 10 epochs	= 40	85,054	4.3

Table 7. **Dataset details: WILDS.** We show metrics for all classification datasets in WILDS where we could perform model selection. In our main experiments, we only use the benchmarks where near-perfect model selection can be performed trivially using the default in-distribution validation set. We train all models ourselves using the public code from Koh et al. [30].

MSV Dataset	Num. Classes	Num. Checkpoints	Test set size
CIFAR10-High	10	80	10000
CIFAR10-Low	10	80	10000
PACS	7	30	9991

GLUE Dataset	Num. Classes	Num. Checkpoints	Test set size
CoLA	2	109	1043
MNLI	3	82	9815
QNLI	2	90	5463
QQP	2	101	40430
RTE	2	87	277
SST2	2	97	872
MRPC	2	95	408

Table 8. **Dataset details: MSV and GLUE.** All model checkpoints sourced directly from Okanovic et al. [45].