A. Comparison of Strategies for Initializing EM Source Probabilities

Shift Estimator	Calibration Method	$\rho = 0.5$			$\rho = 0.9$		
		n=500	n=1000	n=1500	n=500	n=1000	n=1500
EM: source priors from preds	None	**2.0; 0.0**	**2.0; 0.0**	**2.2; 0.0**	**1.4; 0.0**	**1.5; 0.0**	**1.567; 0.0**
EM: source priors from labels	None	-3.4; 1.0	-3.6; 1.0	-3.367; 1.0	1.0; 1.0	1.0; 1.0	1.067; 1.0
EM: source priors from preds	TS	**2.2; 0.0**	**2.3; 0.0**	**2.667; 0.0**	**1.6; 0.0**	**2.0; 0.0**	**2.133; 0.0**
EM: source priors from labels	TS	-64.5; 1.0	-65.1; 1.0	-65.267; 1.0	-91.4; 1.0	-91.9; 1.0	-91.833; 1.0
EM: source priors from preds	NBVS	**3.5; 0.0**	**4.1; 0.0**	**4.233; 0.0**	**2.2; 0.0**	**2.4; 0.0**	**2.467; 0.0**
EM: source priors from labels	NBVS	-5.6; 1.0	-4.6; 1.0	-4.567; 1.0	0.4; 1.0	0.6; 1.0	0.7; 1.0
EM: source priors from preds	BCTS	**3.8; 0.0**	4.65; 0.0	4.633; 0.0	3.6; 0.0	3.6; 0.0	3.733; 0.0
EM: source priors from labels	BCTS	3.8; 1.0	4.65; 1.0	4.633; 1.0	3.6; 1.0	3.6; 1.0	3.733; 1.0
EM: source priors from preds	VS	3.8; 0.0	4.4; 0.0	4.633; 0.0	3.5; 0.0	3.6; 0.0	3.733; 0.0
EM: source priors from labels	VS	3.8; 1.0	4.4; 1.0	4.633; 1.0	3.5; 1.0	3.6; 1.0	3.733; 1.0

Table A.1. **The strategy for computing EM source priors heavily affects domain adaptation if probabilities retain systematic bias**. Value before the semicolon is the median improvement in %accuracy (across 100 trials) caused by applying domain adaptation to the predictions on a diabetic retinopathy prediction task. Value after the semicolon is the median rank of a particular method relative to the other method in the pair. Domain shift is induced by varying the proportion of "healthy" examples ρ; in the source distribution, $\rho = 0.73$. We see that calibration methods that lack class-specific bias parameters (i.e. no calibration, TS and NBVS) can hurt domain adaptation if source priors are initialized by averaging true labels rather than the predicted probabilities. A bold value in a pair is significantly better than the non-bold value according to a paired Wilcoxon test at $p \leq 0.01$. See **Sec. 4.1** for details on the experimental setup.

B. Calibration Quality Comparison

We find that bias-corrected versions of Temperature Scaling (namely Bias-Corrected Temperature Scaling and Vector Scaling) tend to yield the best Negative Log Likelihood on an unshifted test set, even if they do not always yield the best ECE. Results are shown in the tables below.

Calibration Method	NLL			ECE		
	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
None	0.299; 4.0	0.299; 4.0	0.299; 4.0	2.696; 4.0	2.696; 4.0	2.696; 4.0
TS	0.292; 3.0	0.292; 3.0	0.292; 3.0	**1.074; 1.0**	1.029; 2.0	0.996; 3.0
NBVS	0.277; 2.0	0.276; 2.0	0.276; 2.0	1.09; 1.5	0.997; 1.5	**0.932; 1.5**
BCTS	**0.274; 0.0**	**0.273; 1.0**	0.272; 1.0	**1.046; 1.0**	**0.963; 1.0**	**0.921; 1.0**
VS	0.276; 1.0	**0.274; 0.5**	**0.272; 0.0**	1.175; 2.5	1.022; 2.0	0.966; 1.0

Table B.1. **CIFAR10: NLL and ECE for different calibration methods**. Metrics were computed on a test set that had the same distribution as the validation set. Value before the semicolon is the median of the metric over all the runs. Value after the semicolon is the median rank of the method relative to other methods in the column. n indicates the number of examples used for calibratin. Bold values in a column are not significantly different from the best performing method in the column, as measured by a paired Wilcoxon test at $p \leq 0.01$. See **Sec. 4.1** for details on the experimental setup.

Calibration Method	NLL			ECE		
	n=7000	n=8500	n=10000	n=7000	n=8500	n=10000
None	1.727; 4.0	1.727; 4.0	1.727; 4.0	19.999; 4.0	19.999; 4.0	19.999; 4.0
TS	1.281; 3.0	1.281; 3.0	1.281; 3.0	3.149; 3.0	3.156; 3.0	3.184; 3.0
NBVS	1.236; 2.0	1.235; 2.0	1.234; 2.0	**2.291; 0.0**	**2.289; 0.0**	**2.344; 0.0**
BCTS	1.23; 1.0	1.228; 1.0	1.227; 1.0	2.91; 2.0	2.943; 2.0	2.918; 2.0
VS	**1.229; 0.0**	**1.225; 0.0**	**1.224; 0.0**	2.495; 1.0	2.501; 1.0	2.474; 1.0

Table B.2. **CIFAR100: NLL and ECE for different calibration methods**. Analogous to **Table B.1**.

Calibration Method	NLL			ECE		
	n=500	n=1000	n=1500	n=500	n=1000	n=1500
None	0.641; 4.0	0.64; 4.0	0.638; 4.0	8.738; 4.0	8.746; 4.0	8.745; 4.0
TS	0.57; 3.0	0.57; 3.0	0.568; 3.0	3.636; 3.0	3.749; 3.0	3.872; 3.0
NBVS	0.542; 2.0	0.54; 2.0	0.539; 2.0	**1.956; 0.0**	**2.012; 1.0**	**2.047; 1.0**
BCTS	**0.513; 0.0**	**0.511; 1.0**	0.509; 1.0	2.157; 1.0	**2.135; 1.0**	**2.08; 1.0**
VS	0.518; 1.0	**0.512; 0.0**	0.509; 0.0	2.187; 1.0	**2.006; 1.0**	**2.014; 1.0**

Table B.3. **Kaggle Diabetic Retinopathy Detection: NLL and ECE for different calibration methods**. Analogous to **Table B.1**.

C. CIFAR10 Supplementary Tables

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.75; 3.0	0.825; 3.0	0.813; 3.0	16.025; 4.0	15.837; 4.0	15.944; 4.0
EM	TS	0.775; 3.0	0.812; 3.0	0.862; 3.0	16.675; 3.0	16.412; 3.0	16.45; 3.0
EM	NBVS	**1.1; 1.0**	1.162; 2.0	1.169; 2.0	17.1; 2.0	17.062; 2.0	17.244; 2.0
EM	BCTS	**1.2; 1.0**	1.25; 1.0	1.194; 1.0	**17.3; 1.0**	17.025; 1.0	17.362; 1.0
EM	VS	**1.2; 1.0**	**1.275; 1.0**	**1.212; 1.0**	**17.125; 1.0**	**17.225; 1.0**	**17.419; 0.0**

Table C.1. **CIFAR10: Comparison of calibration methods when using EM adaptation to "tweak-one" shift, with Δ%accuracy as the metric.** Analogous to **Table 5**.

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.01049; 15.0	0.00763; 15.0	0.00795; 15.0	0.00283; 15.0	0.00202; 15.0	0.00147; 15.0	0.00191; 12.0	0.001; 10.0	0.0006; 14.0
EM	TS	0.00944; 15.0	0.00705; 15.0	0.00693; 15.0	0.00262; 14.0	0.00192; 15.0	0.00148; 15.0	0.00195; 13.0	0.00101; 13.0	0.00058; 14.0
EM	NBVS	0.0014; 2.0	0.00106; 2.0	0.00075; 4.0	0.00133; 2.0	0.00071; 2.0	0.00043; 3.0	**0.00161; 6.0**	**0.00079; 4.5**	**0.00047; 6.0**
EM	BCTS	**0.00037; 1.0**	**0.00034; 1.0**	**0.00022; 1.0**	**0.00126; 1.0**	0.00071; 2.0	**0.00043; 2.0**	**0.00161; 6.0**	**0.00077; 4.0**	**0.00047; 5.5**
EM	VS	0.00061; 1.0	**0.00031; 1.0**	**0.0002; 1.0**	**0.00118; 1.0**	**0.00067; 2.0**	**0.0004; 1.5**	**0.00167; 6.0**	**0.00076; 4.5**	**0.00045; 6.0**
BBSL-hard	None	0.00443; 13.0	0.00214; 13.0	0.00125; 13.0	0.00266; 14.0	0.00135; 14.0	0.00074; 13.0	0.00244; 15.0	0.00123; 15.0	0.00055; 13.0
BBSL-soft	None	0.00309; 9.0	0.00133; 9.0	0.00092; 9.0	0.00188; 9.0	0.00108; 9.0	0.00057; 7.5	0.00187; 9.0	0.0009; 9.0	0.00047; 9.0
BBSL-soft	TS	0.0031; 8.0	0.00132; 7.0	0.00091; 8.0	0.00191; 10.0	0.00102; 8.0	0.0006; 8.0	0.00179; 7.0	0.00095; 10.0	0.00048; 9.0
BBSL-soft	NBVS	0.0028; 9.0	0.00137; 8.0	0.00098; 8.0	0.00175; 8.0	0.00096; 7.0	0.00061; 7.0	0.00179; 7.0	0.00081; 7.0	0.00048; 7.0
BBSL-soft	BCTS	0.00288; 9.0	0.00126; 7.0	0.00104; 8.0	0.00171; 8.0	0.00102; 8.0	0.0006; 7.0	0.00176; 7.0	0.00083; 8.0	0.00048; 7.0
BBSL-soft	VS	0.00306; 9.0	0.00134; 9.0	0.00103; 8.0	0.00173; 8.0	0.00102; 7.0	0.00061; 7.0	0.00172; 7.0	0.00082; 7.0	0.00047; 7.0
RLLS-hard	None	0.00405; 13.0	0.00209; 13.0	0.00124; 12.0	0.00265; 13.0	0.00135; 13.0	0.00074; 13.0	0.00244; 15.0	0.00123; 15.0	0.00055; 13.0
RLLS-soft	None	0.00316; 9.0	0.00127; 8.5	0.0009; 8.0	0.00183; 10.0	0.00108; 10.0	0.00057; 8.0	0.00187; 8.5	0.0009; 9.0	0.00047; 9.0
RLLS-soft	TS	0.00312; 8.0	0.00125; 7.0	0.0009; 7.0	0.00186; 9.0	0.00102; 8.0	0.0006; 8.0	0.00179; 7.0	0.00095; 9.0	0.00048; 9.0
RLLS-soft	NBVS	0.00275; 8.0	0.00131; 8.0	0.00094; 7.5	0.00168; 7.0	0.00096; 7.0	0.00061; 7.5	0.00179; 7.0	0.00081; 7.0	0.00048; 7.0
RLLS-soft	BCTS	0.00284; 7.0	0.00124; 7.0	0.00093; 7.0	0.00169; 8.0	0.00102; 8.0	0.0006; 7.0	0.00176; 7.0	0.00083; 8.0	0.00048; 7.0
RLLS-soft	VS	0.00298; 7.5	0.00133; 8.0	0.00091; 8.0	0.00171; 7.0	0.00102; 7.0	0.00061; 7.0	0.00172; 7.0	0.00082; 7.0	0.00047; 7.0

Table C.2. **CIFAR10: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric (dirichlet shift).** Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. EM with BCTS or VS tends to achieve the best performance. See **Sec. 4.1** for details on the experimental setup.

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.00177; 13.0	0.00094; 13.0	0.00066; 15.0	0.04165; 15.0	0.03988; 15.0	0.03791; 15.0
EM	TS	0.00183; 13.0	0.00101; 14.0	0.00071; 15.0	0.04156; 16.0	0.04897; 16.0	0.04237; 16.0
EM	NBVS	**0.00139; 3.0**	**0.0006; 3.0**	0.0004; 4.0	0.00274; 2.0	0.00187; 2.0	0.00172; 2.0
EM	BCTS	**0.00132; 3.0**	**0.00058; 2.0**	**0.00036; 2.0**	**0.00125; 1.0**	**0.00083; 1.0**	**0.00072; 1.0**
EM	VS	**0.00139; 3.5**	**0.00057; 2.0**	**0.00038; 3.5**	**0.00165; 1.0**	**0.00083; 1.0**	**0.00066; 1.0**
BBSL-hard	None	0.00228; 15.0	0.001; 15.0	0.00054; 14.0	0.01524; 13.0	0.00834; 13.0	0.00473; 10.0
BBSL-soft	None	0.00148; 9.0	0.00067; 8.0	0.0004; 6.0	0.01356; 11.0	0.00779; 10.0	0.00488; 10.0
BBSL-soft	TS	0.00146; 7.0	0.00068; 8.0	0.0004; 6.0	0.01411; 10.0	0.00852; 11.0	0.00565; 12.0
BBSL-soft	NBVS	0.00153; 8.0	0.00073; 7.0	0.00043; 8.0	0.01187; 7.5	0.00636; 7.0	0.00346; 7.0
BBSL-soft	BCTS	0.00152; 8.0	0.00073; 8.0	0.00042; 7.0	0.01189; 8.0	0.00607; 7.0	0.00351; 6.0
BBSL-soft	VS	0.00159; 8.0	0.00075; 7.5	0.00043; 9.5	0.0121; 8.0	0.00607; 6.0	0.0033; 5.0
RLLS-hard	None	0.00228; 15.0	0.001; 14.0	0.00054; 13.0	0.01347; 13.0	0.00815; 13.0	0.00476; 10.0
RLLS-soft	None	0.00148; 9.0	0.00067; 8.0	0.0004; 6.0	0.01355; 10.0	0.00779; 11.0	0.00488; 11.0
RLLS-soft	TS	0.00146; 7.5	0.00068; 7.0	0.0004; 5.0	0.01368; 10.0	0.00832; 12.0	0.00565; 13.0
RLLS-soft	NBVS	0.00153; 8.0	0.00073; 7.0	0.00043; 7.0	0.01155; 8.0	0.00626; 7.0	0.00343; 8.0
RLLS-soft	BCTS	0.00152; 7.0	0.00073; 8.0	0.00042; 7.0	0.01178; 7.0	0.00601; 7.0	0.00337; 7.0
RLLS-soft	VS	0.00159; 7.0	0.00075; 7.0	0.00043; 9.0	0.01218; 7.0	0.00598; 6.0	0.00326; 6.0

Table C.3. **CIFAR10: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric ("tweak-one" shift).** Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. EM with BCTS or VS tends to achieve the best performance. See **Sec. 4.1** for details on the experimental setup.

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.00177; 2.0	0.00094; 2.0	0.00066; 4.0	0.04165; 4.0	0.03988; 4.0	0.03791; 4.0
BBSL-hard	None	0.00228; 3.0	0.001; 3.0	0.00054; 3.0	**0.01524; 2.0**	0.00834; 2.0	**0.00473; 1.5**
BBSL-soft	None	0.00148; 1.0	0.00067; 1.0	0.0004; 1.0	**0.01356; 1.0**	**0.00779; 1.0**	**0.00488; 2.0**
RLLS-hard	None	0.00228; 3.0	0.001; 3.0	0.00054; 3.0	**0.01347; 2.0**	0.00815; 2.0	**0.00476; 1.0**
RLLS-soft	None	**0.00148; 1.0**	**0.00067; 1.0**	0.0004; 1.0	**0.01355; 1.0**	0.00779; 1.0	0.00488; 2.0
EM	TS	0.00183; 2.0	0.00101; 2.0	0.00071; 2.0	0.04156; 2.0	0.04897; 2.0	0.04237; 2.0
BBSL-soft	TS	0.00146; 1.0	0.00068; 1.0	0.0004; 1.0	**0.01411; 1.0**	0.00852; 0.0	**0.00565; 0.0**
RLLS-soft	TS	**0.00146; 1.0**	**0.00068; 0.0**	**0.0004; 0.0**	**0.01368; 1.0**	0.00832; 1.0	0.00565; 1.0
EM	NBVS	**0.00139; 0.0**	**0.0006; 0.0**	**0.0004; 0.0**	**0.00274; 0.0**	**0.00187; 0.0**	**0.00172; 0.0**
BBSL-soft	NBVS	0.00153; 2.0	0.00073; 1.0	0.00043; 1.0	0.01187; 1.0	0.00636; 1.0	0.00346; 1.0
RLLS-soft	NBVS	0.00153; 1.0	0.00073; 1.0	0.00043; 1.0	0.01155; 2.0	0.00626; 2.0	0.00343; 2.0
EM	BCTS	**0.00132; 0.0**	**0.00058; 0.0**	**0.00036; 0.0**	**0.00125; 0.0**	**0.00083; 0.0**	**0.00072; 0.0**
BBSL-soft	BCTS	0.00152; 1.5	0.00073; 1.0	0.00042; 2.0	0.01189; 1.0	0.00607; 1.0	0.00351; 1.0
RLLS-soft	BCTS	0.00152; 1.0	0.00073; 1.0	0.00042; 1.0	0.01178; 2.0	0.00601; 2.0	0.00337; 2.0
EM	VS	**0.00139; 0.0**	**0.00057; 0.0**	**0.00038; 0.0**	**0.00165; 0.0**	**0.00083; 0.0**	**0.00066; 0.0**
BBSL-soft	VS	0.00159; 2.0	0.00075; 1.0	0.00043; 2.0	0.0121; 1.0	0.00607; 1.0	0.0033; 1.0
RLLS-soft	VS	0.00159; 1.0	0.00075; 1.0	0.00043; 1.0	0.01218; 2.0	0.00598; 2.0	0.00326; 2.0

Table C.4. **CIFAR10: Comparison of EM, BBSL and RLLS ("tweak-one" shift) using MSE as the metric.** Analogous to **Table 1**, but with tweak-one shift instead of dirichlet shift.

D. MNIST Tables

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.00438; 13.0	0.00294; 6.0	0.00227; 13.0	0.00339; 16.0	0.00256; 16.0	0.00207; 16.0	0.00217; 15.0	0.00133; 16.0	0.00093; 16.0
EM	TS	0.00322; 5.5	0.00195; 4.0	0.00189; 9.5	0.00203; 5.0	0.00102; 5.0	0.00084; 10.5	0.00158; 7.5	0.00081; 7.0	0.00052; 8.0
EM	NBVS	0.00144; 2.0	0.00104; 2.0	0.00085; 2.0	**0.00167; 3.0**	**0.00094; 2.0**	0.0007; 3.0	0.00165; 8.5	0.00081; 6.5	0.00051; 6.5
EM	BCTS	**0.00085; 1.0**	**0.00057; 1.0**	0.0004; 1.0	**0.00165; 2.0**	**0.00093; 2.0**	0.0007; 3.0	0.00157; 5.0	**0.0008; 4.0**	**0.00049; 4.0**
EM	VS	**0.00087; 1.0**	**0.00056; 1.0**	**0.00035; 1.0**	**0.00172; 2.0**	**0.00095; 2.0**	0.00073; 3.0	0.00156; 8.0	**0.0008; 7.0**	0.00054; 7.0
BBSL-hard	None	0.00491; 13.0	0.0036; 11.0	0.00244; 12.5	0.00276; 14.0	0.00167; 14.0	0.00125; 14.0	0.00208; 15.0	0.00106; 14.0	0.00073; 15.0
BBSL-soft	None	0.00459; 10.0	0.00324; 11.0	0.00213; 10.5	0.00227; 9.0	0.00137; 9.0	0.00085; 9.0	0.00174; 9.0	0.00083; 9.0	0.00052; 7.5
BBSL-soft	TS	0.00417; 9.0	0.00309; 9.0	0.00218; 8.0	0.00232; 7.0	0.00129; 6.0	0.00083; 7.0	0.00157; 5.0	**0.00079; 6.0**	**0.00049; 5.0**
BBSL-soft	NBVS	0.00428; 10.5	0.00306; 10.0	0.0022; 10.0	0.00234; 8.0	0.00129; 7.5	0.00082; 7.0	0.00171; 7.0	0.00085; 7.5	0.00052; 7.0
BBSL-soft	BCTS	0.00413; 10.0	0.00303; 10.0	0.00228; 9.0	0.00228; 8.0	0.0013; 8.0	0.00081; 6.0	0.00168; 6.0	0.00083; 6.5	**0.0005; 6.0**
BBSL-soft	VS	0.0039; 9.0	0.00322; 9.5	0.00221; 10.0	0.0022; 9.0	0.00136; 9.0	0.00085; 9.0	0.00163; 8.0	0.00084; 7.0	0.00054; 9.0
RLLS-hard	None	0.00472; 13.0	0.00347; 10.5	0.00238; 12.0	0.00271; 13.5	0.00167; 14.0	0.00125; 13.0	0.00208; 14.0	0.00106; 14.0	0.00073; 14.0
RLLS-soft	None	0.0044; 9.0	0.00314; 10.0	0.00211; 9.0	0.00226; 9.0	0.00137; 10.0	0.00085; 9.0	0.00174; 9.0	0.00083; 10.0	0.00052; 8.0
RLLS-soft	TS	0.00407; 8.0	0.00289; 8.0	0.00206; 8.0	0.00232; 7.0	0.00129; 7.0	0.00083; 7.0	0.00157; 5.0	**0.00079; 6.0**	**0.00049; 5.0**
RLLS-soft	NBVS	0.00409; 9.0	0.00284; 9.0	0.00211; 9.0	0.00232; 9.0	0.00129; 8.0	0.00082; 7.0	0.00171; 7.0	0.00085; 8.0	0.00052; 8.0
RLLS-soft	BCTS	0.00399; 8.0	0.00283; 8.0	0.00213; 8.0	0.00226; 7.5	0.0013; 8.0	0.00082; 6.0	0.00168; 6.5	0.00083; 6.0	**0.0005; 6.0**
RLLS-soft	VS	0.00385; 8.0	0.00282; 8.0	0.00212; 9.0	0.0022; 8.5	0.00136; 9.0	0.00085; 9.0	0.00163; 8.0	0.00084; 8.0	0.00054; 9.0

Table D.1. **MNIST: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric (dirichlet shift).** Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. EM with BCTS or VS tends to achieve the best performance, particularly for larger amounts of shift (corresponding to smaller α). See **Sec. 4.1** for details on the experimental setup.

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	0.00197; 14.0	0.00107; 14.0	0.00067; 14.0	0.00584; 11.0	0.00305; 5.0	0.00316; 12.0
EM	TS	0.00162; 7.0	0.00084; 5.0	0.00054; 3.0	0.00318; 3.0	0.00155; 3.0	0.00085; 3.0
EM	NBVS	0.00158; 5.0	0.00076; 3.5	0.00054; 2.0	0.00112; 2.0	0.00059; 2.0	0.00038; 2.0
EM	BCTS	**0.00153; 3.5**	**0.00076; 3.5**	**0.00052; 2.0**	**0.0006; 1.0**	**0.00037; 0.0**	**0.00028; 0.0**
EM	VS	0.00155; 5.0	**0.00076; 3.0**	**0.00052; 2.0**	0.00067; 1.0	0.00042; 1.0	0.00033; 1.0
BBSL-hard	None	0.00197; 15.0	0.00122; 16.0	0.00079; 16.0	0.00897; 15.0	0.00689; 16.0	0.00608; 16.0
BBSL-soft	None	0.00161; 10.0	0.00099; 10.0	0.00057; 11.0	0.00662; 12.0	0.00379; 6.0	0.00298; 5.0
BBSL-soft	TS	0.00161; 6.0	0.00091; 6.0	0.00054; 6.0	0.00659; 11.0	0.00395; 10.0	0.00339; 10.0
BBSL-soft	NBVS	0.00162; 9.0	0.00088; 9.0	0.00057; 9.0	0.00607; 10.0	0.00385; 9.0	0.00349; 9.0
BBSL-soft	BCTS	0.00158; 7.0	0.00087; 7.5	0.00055; 7.0	0.00623; 9.0	0.00383; 9.0	0.00338; 8.0
BBSL-soft	VS	0.00171; 9.5	0.00086; 9.0	0.00056; 11.0	0.00641; 10.0	0.00407; 11.0	0.00393; 13.0
RLLS-hard	None	0.00197; 15.0	0.00122; 15.0	0.00079; 15.0	0.00868; 14.0	0.00676; 15.0	0.00593; 15.0
RLLS-soft	None	0.00161; 9.0	0.00099; 9.0	0.00057; 10.0	0.00622; 10.0	0.00379; 6.0	0.00298; 5.0
RLLS-soft	TS	0.00161; 6.0	0.00091; 6.0	0.00054; 5.0	0.00627; 9.0	0.00393; 9.0	0.00339; 10.0
RLLS-soft	NBVS	0.00162; 8.0	0.00088; 8.0	0.00057; 8.0	0.006; 7.0	0.0038; 9.0	0.00349; 9.0
RLLS-soft	BCTS	0.00158; 6.0	0.00087; 7.0	0.00055; 6.0	0.00603; 7.0	0.00378; 8.0	0.00338; 8.0
RLLS-soft	VS	0.00171; 9.0	0.00086; 8.0	0.00056; 10.0	0.00626; 8.0	0.00405; 11.0	0.00393; 13.0

Table D.2. **MNIST: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric ("tweak-one" shift).** Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. EM with BCTS or VS tends to achieve the best performance. See **Sec. 4.1** for details on the experimental setup.

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	None	**0.00438; 2.0**	**0.00294; 2.0**	0.00227; 2.0	0.00339; 4.0	0.00256; 4.0	0.00207; 4.0	0.00217; 4.0	0.00133; 4.0	0.00093; 4.0
BBSL-hard	None	0.00491; 2.0	0.0036; 2.0	0.00244; 2.0	0.00276; 3.0	0.00167; 2.5	0.00125; 3.0	0.00208; 3.0	0.00106; 3.0	0.00073; 3.0
BBSL-soft	None	0.00459; 2.0	0.00324; 2.0	0.00213; 2.0	**0.00227; 1.0**	**0.00137; 1.0**	**0.00085; 1.0**	0.00174; 1.0	**0.00083; 1.0**	**0.00052; 1.0**
RLLS-hard	None	0.00472; 2.0	0.00347; 2.0	0.00238; 2.0	0.00271; 2.0	0.00167; 2.0	0.00125; 2.0	0.00208; 2.0	0.00106; 2.0	0.00073; 2.0
RLLS-soft	None	**0.0044; 1.0**	**0.00314; 2.0**	**0.00211; 1.0**	**0.00226; 1.0**	**0.00137; 1.0**	**0.00085; 1.0**	**0.00174; 1.0**	**0.00083; 1.0**	**0.00052; 0.5**
EM	TS	**0.00322; 0.0**	**0.00195; 0.0**	**0.00189; 1.0**	0.00203; 0.0	0.00102; 0.0	0.00084; 2.0	0.00158; 2.0	0.00081; 2.0	0.00052; 2.0
BBSL-soft	TS	0.00417; 1.0	0.00309; 1.0	0.00218; 1.0	0.00232; 1.0	0.00129; 1.0	0.00083; 1.0	0.00157; 1.0	0.00079; 1.0	0.00049; 1.0
RLLS-soft	TS	**0.00407; 1.0**	**0.00289; 1.0**	**0.00206; 1.0**	0.00232; 1.0	0.00129; 1.0	0.00083; 1.0	0.00157; 1.0	0.00079; 1.0	0.00049; 1.0
EM	NBVS	**0.00144; 0.0**	**0.00104; 0.0**	**0.00085; 0.0**	**0.00167; 0.0**	**0.00094; 0.0**	**0.0007; 0.0**	**0.00165; 1.0**	0.00081; 0.0	0.00051; 0.0
BBSL-soft	NBVS	0.00428; 1.5	0.00306; 1.0	0.0022; 1.0	0.00234; 1.0	0.00129; 1.0	0.00082; 1.0	0.00171; 1.0	0.00085; 1.0	0.00052; 1.0
RLLS-soft	NBVS	0.00409; 1.0	0.00284; 1.0	0.00211; 1.0	0.00232; 1.0	0.00129; 1.0	0.00082; 1.0	0.00171; 1.0	0.00085; 1.0	0.00052; 1.0
EM	BCTS	**0.00085; 0.0**	**0.00057; 0.0**	**0.0004; 0.0**	**0.00165; 0.0**	**0.00093; 0.0**	0.0007; 0.0	**0.00157; 0.0**	0.0008; 0.0	0.00049; 0.0
BBSL-soft	BCTS	0.00413; 2.0	0.00303; 2.0	0.00228; 2.0	0.00228; 1.0	0.0013; 1.0	0.00081; 1.0	0.00168; 1.0	0.00083; 1.0	0.0005; 1.0
RLLS-soft	BCTS	0.00399; 1.0	0.00283; 1.0	0.00213; 1.0	0.00226; 1.0	0.0013; 1.0	0.00082; 1.0	0.00168; 1.0	0.00083; 1.0	0.0005; 1.0
EM	VS	**0.00087; 0.0**	**0.00056; 0.0**	**0.00035; 0.0**	**0.00172; 0.0**	**0.00095; 0.0**	**0.00073; 0.0**	**0.00156; 0.0**	0.0008; 0.0	0.00054; 0.0
BBSL-soft	VS	0.0039; 2.0	0.00322; 2.0	0.00221; 2.0	0.0022; 1.0	0.00136; 1.0	0.00085; 1.0	0.00163; 1.0	0.00084; 1.0	0.00054; 1.0
RLLS-soft	VS	0.00385; 1.0	0.00282; 1.0	0.00212; 1.0	0.0022; 1.0	0.00136; 1.0	0.00085; 1.0	**0.00163; 1.0**	0.00084; 1.0	0.00054; 1.0

Table D.3. **MNIST: Comparison of EM, BBSL and RLLS (dirichlet shift).** Analogous to **Table 2**, but with dirichlet shift rather than tweak-one shift.

E. CIFAR100 Supplementary Tables

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10.0$		
		n=7000	n=8500	n=10000	n=7000	n=8500	n=10000	n=7000	n=8500	n=10000
EM	None	1.80113; 16.0	1.67187; 16.0	1.7157; 16.0	0.55795; 16.0	0.54112; 16.0	0.55955; 16.0	0.3896; 16.0	0.37585; 16.0	0.36337; 16.0
EM	TS	0.41127; 9.0	0.34963; 7.0	0.30451; 6.5	0.23803; 13.0	0.21715; 13.0	0.19781; 13.0	0.16408; 12.0	0.14477; 13.0	0.13852; 13.0
EM	NBVS	0.1637; 2.0	0.15904; 2.0	0.14348; 2.0	0.1042; 2.0	0.10523; 2.0	0.10777; 6.0	0.10122; 4.5	0.09874; 7.0	0.09729; 10.0
EM	BCTS	0.1101; 2.0	0.10824; 2.0	0.11636; 2.0	0.08838; 2.0	0.09102; 2.0	0.0876; 3.0	**0.09207; 2.0**	0.08707; 5.0	0.08781; 8.0
EM	VS	**0.09485; 1.0**	**0.08792; 1.0**	**0.0862; 1.0**	**0.07684; 1.0**	**0.07922; 1.0**	**0.07773; 2.0**	**0.09387; 2.0**	0.08796; 5.5	0.08981; 9.0
BBSL-hard	None	0.83095; 15.0	0.68099; 15.0	0.58656; 15.0	0.33542; 15.0	0.27998; 15.0	0.24659; 15.0	0.25694; 15.0	0.21736; 15.0	0.20637; 15.0
BBSL-soft	None	0.5279; 13.0	0.47521; 12.0	0.44263; 13.0	0.23637; 13.0	0.21324; 13.0	0.18605; 13.0	0.18842; 13.0	0.16108; 13.0	0.14018; 13.0
BBSL-soft	TS	0.40279; 9.0	0.34713; 9.0	0.33012; 9.0	0.18774; 9.5	0.15665; 9.0	0.12639; 9.0	0.13409; 8.5	0.10951; 7.0	0.09703; 7.0
BBSL-soft	NBVS	0.40856; 10.0	0.33122; 9.0	0.29594; 9.0	0.17545; 9.0	0.1409; 9.0	0.11336; 8.0	0.13601; 9.0	0.113; 8.0	0.09731; 7.0
BBSL-soft	BCTS	0.40887; 9.0	0.32629; 9.0	0.30435; 9.0	0.17923; 9.0	0.14756; 9.0	0.11536; 9.0	0.13698; 9.0	0.11267; 8.0	0.09748; 7.0
BBSL-soft	VS	0.43085; 9.0	0.3216; 8.0	0.30221; 9.0	0.16715; 8.0	0.13847; 8.0	0.11169; 7.0	0.13075; 8.0	0.10971; 7.0	0.09433; 6.0
RLLS-hard	None	0.4577; 12.0	0.38413; 12.0	0.37675; 12.0	0.20285; 12.0	0.16017; 12.0	0.14228; 12.0	0.16412; 12.0	0.13842; 13.0	0.12127; 12.0
RLLS-soft	None	0.33882; 8.0	0.2728; 8.0	0.2878; 8.0	0.13859; 8.0	0.12697; 7.0	0.10567; 8.0	0.12528; 8.0	0.11011; 8.5	0.10056; 8.0
RLLS-soft	TS	0.27236; 6.0	0.22426; 6.0	0.22697; 5.5	0.12255; 5.0	0.10639; 5.0	0.08221; 4.0	0.11395; 5.0	0.08803; 5.0	0.07868; 4.0
RLLS-soft	NBVS	0.2797; 7.0	0.21465; 5.0	0.23037; 5.0	0.11469; 4.0	0.09903; 4.0	0.08341; 4.0	0.11303; 5.0	0.08868; 4.0	0.07531; 4.0
RLLS-soft	BCTS	0.26594; 6.0	0.2305; 6.0	0.23604; 6.0	0.11498; 5.0	0.09755; 4.0	0.08309; 4.0	0.1085; 4.0	0.08672; 3.0	**0.07252; 3.0**
RLLS-soft	VS	0.28941; 6.0	0.21363; 6.0	0.22405; 6.0	0.11627; 5.0	0.09894; 4.0	0.08267; 3.0	0.11426; 4.0	0.08623; 4.0	0.07414; 3.0

Table E.1. **CIFAR100: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric (dirichlet shift).** Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. EM with VS tends to achieve the best performance, particularly for larger amounts of shift (corresponding to smaller α). See **Sec. 4.1** for details on the experimental setup.

F. Diabetic Retinopathy Supplementary Tables

Shift Estimator	Calibration Method	$\rho = 0.5$			$\rho = 0.9$		
		n=500	n=1000	n=1500	n=500	n=1000	n=1500
EM	None	**0.35073; 3.5**	0.30498; 4.0	0.19709; 3.0	0.0899; 10.0	0.0596; 12.0	0.05666; 13.0
EM	TS	**0.47048; 3.0**	**0.29683; 3.0**	0.20944; 2.0	0.08077; 11.0	0.05537; 12.0	0.05268; 13.0
EM	NBVS	**0.51515; 5.0**	0.33256; 4.0	0.24066; 4.0	0.09137; 12.0	0.08415; 14.0	0.08002; 15.0
EM	BCTS	**0.40245; 3.5**	**0.2475; 3.0**	**0.19461; 3.0**	**0.02267; 2.5**	**0.01354; 2.5**	**0.01043; 3.0**
EM	VS	0.53081; 5.0	**0.26442; 4.0**	0.21641; 4.0	**0.02388; 3.0**	**0.01307; 2.0**	**0.00953; 2.0**
BBSL-hard	None	1.83879; 15.0	1.4584; 15.0	0.81962; 15.0	0.31856; 16.0	0.10582; 15.0	0.0491; 13.5
BBSL-soft	None	1.4041; 11.0	0.71345; 10.5	0.43408; 10.0	0.10441; 12.0	0.04494; 11.0	0.02562; 10.0
BBSL-soft	TS	1.26587; 10.0	0.57077; 9.0	0.38372; 9.0	0.09552; 11.0	0.03859; 10.0	0.02165; 9.0
BBSL-soft	NBVS	1.62832; 12.0	0.70875; 10.5	0.45962; 11.0	0.09014; 10.0	0.03608; 10.0	0.01769; 8.0
BBSL-soft	BCTS	1.40016; 12.0	0.52628; 9.0	0.46372; 10.0	0.05524; 8.0	0.02529; 6.0	0.01419; 6.0
BBSL-soft	VS	1.36527; 11.0	0.50393; 10.0	0.48291; 10.0	0.04678; 7.0	0.02485; 6.0	0.01543; 6.0
RLLS-hard	None	1.07889; 10.0	1.08159; 14.0	0.74669; 13.5	0.05118; 7.0	0.02865; 7.0	0.02458; 8.0
RLLS-soft	None	0.91948; 8.0	0.63483; 9.0	0.41057; 7.0	0.03602; 6.0	0.0225; 6.0	0.02351; 7.0
RLLS-soft	TS	0.80812; 7.5	0.56293; 8.0	0.37624; 8.0	0.0352; 6.0	0.02457; 7.0	0.02054; 7.0
RLLS-soft	NBVS	0.77171; 7.0	0.56492; 7.0	0.40272; 9.0	0.04301; 7.0	0.03258; 8.0	0.02541; 8.0
RLLS-soft	BCTS	0.87398; 8.0	0.50668; 7.5	0.42255; 9.0	0.0384; 6.0	0.02826; 6.0	0.02245; 6.0
RLLS-soft	VS	0.82992; 7.0	0.47802; 7.0	0.42028; 8.5	0.04096; 6.0	0.02948; 7.0	0.022; 6.0

Table F.1. **Kaggle Diabetic Retinopathy Detection: Comparison of all calibration and domain adaptation methods, using MSE (Sec. 3.5) as the metric.** ρ represents the porportion of healthy examples in the sfhited domain; the source distribution has $\rho = 0.73$. Value before the semicolon is the median of the metric over all trials. Value after the semicolon is the median rank of the domain adaptation + calibration method combination relative to the other method combinations in the column. Bold values in a column are not significantly different from the best-performing method in the column as measured by a paired Wilcoxon test at $p < 0.01$. See **Sec. 4.1** for details on the experimental setup.

G. NLL Corresponds Better To Benefits In Label Shift Adaptation

To investigate whether NLL or ECE corresponded better to the benefits offered by a calibration method in the context of label shift adaptation, we adopted the following strategy: in a given experimental run, we identified the calibration method that provided the best NLL (or ECE) on the unshifted test set. We then looked at the performance of label shift adaptation using this calibration method. Note that the calibration method selected can differ from one run to the next. Across datasets, we observed that, by and large, selecting a calibration method according to the NLL produced better performance after domain adaptation as compared to selecting a calibration method according to ECE. Results are show in the tables below.

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	Best NLL	5.9; 0.0	5.775; 0.0	**5.75; 0.0**	2.15; 0.0	**2.075; 0.0**	**2.081; 0.0**	0.725; 0.0	**0.8; 0.0**	**0.838; 0.0**
EM	Best ECE	5.675; 1.0	5.85; 1.0	5.744; 1.0	2.15; 1.0	1.713; 1.0	1.775; 1.0	0.75; 1.0	0.188; 1.0	0.244; 1.0

Table G.1. **CIFAR10: NLL vs ECE, Δ%Accuracy, dirichlet shift.** Entry in "calibration method" column indicates how the calibration method for any given run was selected: either according to whether it produced the best NLL or whether it produced the best ECE, where NLL and ECE were calculated on the unshifted test set. Value before the semicolon is the median change in %accuracy relative to unadapted predictions. Value after the semicolon is the median rank of the given metric relative to the other metric in the pair. A bold value is significantly better than the non-bold value in the pair using a paired Wilcoxon test at $p \leq 0.01$. See **Sec. 4.1** for details on the experimental setup.

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	Best NLL	**1.25; 0.0**	**1.262; 0.0**	**1.225; 0.0**	17.3; 0.0	**17.025; 0.0**	**17.362; 0.0**
EM	Best ECE	1.05; 1.0	1.088; 1.0	1.137; 1.0	17.125; 1.0	16.412; 1.0	16.45; 1.0

Table G.2. **CIFAR10: NLL vs. ECE, metric: Δ%accuracy, "tweak-one" shift**. Analogous to **Table G.1**. The "tweak-one" shift strategy is explained in **Sec. 4.1**.

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	Best NLL	**0.04654; 0.0**	**0.03934; 0.0**	**0.02037; 0.0**	0.1262; 0.0	**0.07123; 0.0**	**0.04254; 0.0**	0.1609; 0.5	**0.07719; 0.0**	**0.04652; 0.0**
EM	Best ECE	0.14529; 1.0	0.08345; 1.0	0.04786; 1.0	0.118; 1.0	0.19194; 1.0	0.1479; 1.0	0.16718; 0.5	0.10077; 1.0	0.05845; 1.0
BBSL-soft	Best NLL	0.28846; 0.0	0.13062; 0.0	0.10355; 0.0	0.17072; 1.0	0.10215; 0.5	0.06013; 0.0	0.17633; 1.0	**0.0825; 0.0**	0.04782; 0.0
BBSL-soft	Best ECE	0.27553; 1.0	0.13608; 1.0	0.10031; 1.0	0.17342; 0.0	0.1021; 0.5	0.05966; 1.0	0.17173; 0.0	0.09483; 1.0	0.04756; 1.0
RLLS-soft	Best NLL	0.28416; 0.0	0.13019; 0.0	0.08999; 0.0	0.16909; 1.0	0.10211; 0.5	0.06013; 0.0	0.17633; 1.0	**0.0825; 0.0**	0.04782; 0.0
RLLS-soft	Best ECE	0.27631; 1.0	0.13053; 1.0	0.09304; 1.0	0.17087; 0.0	0.1021; 0.5	0.05977; 1.0	0.17173; 0.0	0.09483; 1.0	0.04756; 1.0

Table G.3. **CIFAR10: NLL vs. ECE, metric: MSE, dirichlet shift**. Analogous to **Table G.1**, but using MSE (**Sec. 3.5**) as the metric rather than change in %accuracy.

Shift Estimator	Calibration Method	$\rho = 0.01$			$\rho = 0.9$		
		n=2000	n=4000	n=8000	n=2000	n=4000	n=8000
EM	Best NLL	**0.13671; 0.0**	**0.05747; 0.0**	**0.03794; 0.0**	0.12478; 0.0	**0.08311; 0.0**	**0.07207; 0.0**
EM	Best ECE	0.14919; 1.0	0.06362; 1.0	0.04005; 1.0	0.16472; 1.0	4.89671; 1.0	4.23677; 1.0
BBSL-soft	Best NLL	0.15212; 0.0	0.07357; 1.0	0.04304; 1.0	1.18879; 1.0	**0.60651; 0.0**	**0.35089; 0.0**
BBSL-soft	Best ECE	0.14834; 1.0	**0.07226; 0.0**	**0.04296; 0.0**	1.21037; 0.0	0.85178; 1.0	0.5647; 1.0
RLLS-soft	Best NLL	0.15212; 0.0	0.07357; 1.0	0.04304; 1.0	1.17799; 1.0	**0.60084; 0.0**	**0.337; 0.0**
RLLS-soft	Best ECE	0.14834; 1.0	**0.07226; 0.0**	**0.04296; 0.0**	1.21754; 0.0	0.83205; 1.0	0.5647; 1.0

Table G.4. **CIFAR10: NLL vs ECE, metric: MSE, "tweak-one" shift.** Analogous to **Table G.1**.

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10.0$		
		n=7000	n=8500	n=10000	n=7000	n=8500	n=10000	n=7000	n=8500	n=10000
EM	Best NLL	**24.829; 0.0**	**24.518; 0.0**	24.51; 0.0	**21.3; 0.0**	**21.629; 0.0**	**21.93; 0.0**	**21.229; 0.0**	**21.306; 0.0**	**21.21; 0.0**
EM	Best ECE	24.079; 1.0	24.053; 1.0	24.575; 1.0	20.986; 1.0	21.382; 1.0	21.53; 1.0	21.05; 1.0	20.971; 1.0	21.06; 1.0

Table G.5. **CIFAR100: NLL vs ECE, metric: Δ%Accuracy, dirichlet shift.** Analogous to **Table G.1**

Shift Estimator	Calibration Method	$\alpha = 0.1$			$\alpha = 1.0$			$\alpha = 10.0$		
		n=7000	n=8500	n=10000	n=7000	n=8500	n=10000	n=7000	n=8500	n=10000
EM	Best NLL	**0.09485; 0.0**	**0.08792; 0.0**	**0.0862; 0.0**	**0.07684; 0.0**	**0.07922; 0.0**	**0.07773; 0.0**	**0.09307; 0.0**	**0.0868; 0.0**	**0.08981; 0.0**
EM	Best ECE	0.1637; 1.0	0.15904; 1.0	0.14348; 1.0	0.1042; 1.0	0.10523; 1.0	0.10777; 1.0	0.10042; 1.0	0.0966; 1.0	0.09683; 1.0
BBSL-soft	Best NLL	0.43085; 0.0	**0.3216; 0.0**	0.30221; 0.0	**0.16715; 0.0**	**0.13847; 0.0**	**0.11169; 0.0**	**0.13481; 0.0**	**0.11224; 0.0**	**0.09433; 0.0**
BBSL-soft	Best ECE	0.40856; 1.0	0.33122; 1.0	0.29594; 1.0	0.17545; 1.0	0.1409; 1.0	0.11336; 1.0	0.13601; 1.0	0.113; 1.0	0.09731; 1.0
RLLS-soft	Best NLL	0.28941; 0.5	0.21363; 1.0	0.22405; 1.0	0.11627; 0.0	0.09894; 1.0	0.08267; 0.0	0.11264; 0.0	0.08623; 0.0	**0.07414; 0.0**
RLLS-soft	Best ECE	0.2797; 0.5	0.21465; 0.0	0.23037; 0.0	0.11469; 1.0	0.09903; 0.0	0.08341; 1.0	0.11303; 1.0	0.08868; 1.0	0.07531; 1.0

Table G.6. **CIFAR100: NLL vs ECE, metric: MSE, dirichlet shift.** Analogous to **Table G.1**

Shift Estimator	Calibration Method	$\rho = 0.5$			$\rho = 0.9$		
		n=500	n=1000	n=1500	n=500	n=1000	n=1500
EM	Best NLL	**3.8; 0.0**	**4.5; 0.0**	**4.6; 0.0**	**3.6; 0.0**	3.6; 0.0	**3.733; 0.0**
EM	Best ECE	3.6; 1.0	4.4; 1.0	4.3; 1.0	2.2; 1.0	3.6; 1.0	2.467; 1.0

Table G.7. **KaggleDR: NLL vs ECE, metric: Δ%Accuracy.** Shift strategy modifies the proportion of healthy examples. Analogous to **Table G.1**

Shift Estimator	Calibration Method	$\rho = 0.5$			$\rho = 0.9$		
		n=500	n=1000	n=1500	n=500	n=1000	n=1500
EM	Best NLL	0.437; 0.0	**0.242; 0.0**	0.21; 0.0	**0.023; 0.0**	0.013; 0.0	**0.01; 0.0**
EM	Best ECE	0.485; 1.0	0.329; 1.0	0.225; 1.0	0.091; 1.0	0.013; 1.0	0.08; 1.0
BBSL-soft	Best NLL	1.4; 0.0	**0.554; 0.0**	0.486; 0.0	**0.055; 0.0**	0.025; 0.0	**0.015; 0.0**
BBSL-soft	Best ECE	1.498; 1.0	0.682; 1.0	0.416; 1.0	0.09; 1.0	0.025; 1.0	0.018; 1.0
RLLS-soft	Best NLL	0.907; 0.0	0.5; 0.0	0.404; 0.0	0.038; 0.0	0.029; 0.0	**0.022; 0.0**
RLLS-soft	Best ECE	**0.835; 1.0**	0.551; 1.0	0.403; 1.0	0.043; 1.0	0.029; 1.0	0.025; 1.0

Table G.8. **KaggleDR: NLL vs ECE, metric: MSE.** Shift strategy modifies the proportion of healthy examples. Analogous to **Table G.1**

H. Semi-supervised update

In some practical settings, we may be able to sample a portion of the test set and label them, for instance, by having domain experts provide labels for several cases. In this situation, the ideal thing to do would still be to leverage the remainder of the test data to estimate $q(y)$ in a semi-supervised fashion. The formula for the semi-supervised update would simply alter the M-step of EM to be:

$$q^{(s+1)}(y=i) = \frac{l_i + \sum_{k=1}^{U} q^{(s)}(y=i|k)}{U+L}$$

where l_i is the number of examples that ground-truth labels say are in class i, L is the total number of ground-truth labeled examples, and U is the total number of unlabeled examples.

I. Proof that EM Converges to a Stationary Point of the Likelihood Function

We slightly modify the derivation of Theorem 2 of Wu (1983), which states that (given certain conditions) continuity of the conditional expectation $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ in $q(\boldsymbol{\omega})$ and $q^{(s)}(\boldsymbol{\omega})$ implies that the sequence of parameter estimates produced by EM converges to a stationary point of the likelihood. The reason we do not directly invoke Theorem 2 of Wu (1983) is that its proof assumes that the conditional expectation $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ is defined at all points in the domain of the log-likelihood. In our application, $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ can be undefined when $q(\boldsymbol{\omega})$ is at the boundary of the domain - however, in spite of this, we observe that it is still possible to prove that EM converges to a stationary point of the log-likelihood.

The proof proceeds as follows: in **Sec. I.1**, we review the definition of the likelihood and recap key properties of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$. In **Sec. I.2**, we verify that certain conditions assumed in Wu (1983) hold for the case of domain adaptation to label shift. In **Sec. I.3**, we identify the domain of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ and prove differentiability over this domain. Finally, in **Sec. I.4**, we will invoke Theorem 1 of Wu (1983) to show that EM converges to a stationary point of the likelihood. Together with the concavity of the likelihood, this is sufficient to guarantee convergence to a global maximum.

I.1. Review of Likelihood and Key Properties of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$

We will begin by recapping certain definitions and expressions that will be useful later in the proof. Let ω_i denote membership in class i, and let $q(\omega_i)$ and $p(\omega_i)$ denote the target & source domain priors. Using the form of the log-likelihood derived in **Eqn. 3** in the main text, we have:

$$\begin{aligned} L(q(\boldsymbol{\omega})) &= \sum_k \log\left(\sum_i q(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) \\ &= \sum_k \log\left(\sum_i p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) \text{ (By the label shift assumption)} \end{aligned}$$

Let $q^{(s)}(\omega_i|\boldsymbol{x_k})$ denote the conditional probability that $\boldsymbol{x}_k$ originated from class i given the estimate $q^{(s)}(\boldsymbol{\omega})$ of the priors at

EM iteration s. The expression for $q^{(s)}(\omega_i|\boldsymbol{x_k})$ in terms of $q^{(s)}(\omega_i)$ can be found in equation A.4 of Saerens 2002:

$$q^{(s)}(\omega_i|\boldsymbol{x}_k) = \frac{\frac{q^{(s)}(\omega_i)}{p(\omega_i)}p(\omega_i|\boldsymbol{x}_k)}{\sum_{j=1}^{m}\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)} = \frac{q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)}{p(\omega_i)\sum_{j=1}^{m}\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)} \tag{9}$$

Following the standard derivation of EM, we can rewrite the log-likelihood as:

$$\begin{aligned}
L(q(\boldsymbol{\omega})) &= \sum_k \log\left(\sum_i \frac{q^{(s)}(\omega_i|\boldsymbol{x_k})}{q^{(s)}(\omega_i|\boldsymbol{x_k})}p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) \\
&= \sum_k \log\left(\mathbb{E}_{i\sim q^{(s)}(\omega_i|\boldsymbol{x_k})}\left[\frac{1}{q^{(s)}(\omega_i|\boldsymbol{x}_k)}p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right]\right) \\
&\geq \sum_k \mathbb{E}_{i\sim q^{(s)}(\omega_i|\boldsymbol{x_k})}\left[\log\left(\frac{1}{q^{(s)}(\omega_i|\boldsymbol{x_k})}p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right)\right] \text{ (By Jensen's inequality)} \\
&= \sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(\frac{1}{q^{(s)}(\omega_i|\boldsymbol{x_k})}p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) \\
&= \sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) - \sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right)
\end{aligned} \tag{10}$$

Let us define:

$$Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) := \sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right) \tag{11}$$

This agrees with the definition of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ as the conditional expectation of the complete data likelihood given $q^{(s)}(\boldsymbol{\omega})$; for a derivation, see equation A.5 in Saerens (2002).

Given our definition of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$, we can write:

$$\begin{aligned}
L(q(\boldsymbol{\omega})) &\geq Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - \sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right) \\
\sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right) &\geq Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - L(q(\boldsymbol{\omega}))
\end{aligned}$$

Let us define

$$H(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) := Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - L(q(\boldsymbol{\omega})) \tag{12}$$

We have:

$$\sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right) \geq H(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \tag{13}$$

When $q^{(s)}(\boldsymbol{\omega}) = q(\boldsymbol{\omega})$, the expectation $\mathbb{E}_{i\sim q^{(s)}(\omega_i|\boldsymbol{x_k})}\left[\frac{1}{q^{(s)}(\omega_i)}p(\boldsymbol{x}_k|\omega_i)q(\omega_i)\right]$ in (10) becomes $\mathbb{E}_{i\sim q^{(s)}(\omega_i|\boldsymbol{x_k})}\left[p(\boldsymbol{x}_k|\omega_i)\right]$, which is the expectation over a constant-valued random variable. Because Jensen's inequality holds with equality when the expectation is taken over a constant-valued random variable, we have:

$$\sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right) = H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \tag{14}$$

Because $\sum_k \sum_i q^{(s)}(\omega_i|\boldsymbol{x_k})\log\left(q^{(s)}(\omega_i|\boldsymbol{x_k})\right)$ does not depend on $q(\boldsymbol{\omega})$, this implies $H(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ is maximized when $q^{(s)}(\boldsymbol{\omega}) = q(\boldsymbol{\omega})$, i.e.

$$q^{(s)}(\boldsymbol{\omega}) = \arg\max_{q(\boldsymbol{\omega})} H(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \tag{15}$$

It follows that if we set $q^{(s+1)}(\boldsymbol{\omega})$ to be $\arg\max_{q(\boldsymbol{\omega})} Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$, we are guaranteed that $L(q^{(s+1)}) \geq L(q^{(s)})$:

$$\begin{aligned} L(q^{(s+1)}(\boldsymbol{\omega})) &= Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \\ &\geq Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \text{ (From the definition of } q^{(s+1)}(\boldsymbol{\omega})) \\ &\geq Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \text{ (From } \textbf{Eqn. 15}) \\ &= L(q^{(s)}(\boldsymbol{\omega})) \end{aligned} \tag{16}$$

The formula for the EM update rule $q^{(s+1)}(\boldsymbol{\omega}) = \arg\max_{q(\boldsymbol{\omega})} Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ can be found in Equation A.8 of Saerens (2002) and is reproduced below:

$$q^{(s+1)}(\omega_i) = \frac{1}{N} \sum_{k=1}^{N} q^{(s)}(\omega_i | \boldsymbol{x}_k) \tag{17}$$

where N denotes the number of examples from the target domain and $q^{(s)}(\omega_i|\boldsymbol{x}_k)$ is defined as in (9).

I.2. Verifying Conditions Assumed In Wu (1983)

Given our assumption that $p(\omega_i) \geq \epsilon \; \forall i$, we will show that the conditions assumed in Wu (1983) hold for domain adaptation to label shift. Let $\boldsymbol{\Omega}$ denote the domain of the log-likelihood function $L(q(\boldsymbol{\omega})) = C + \sum_k \log\left(\sum_i \frac{p(\omega_i|\boldsymbol{x}_k)}{p(\omega_i)} q(\omega_i)\right)$. The conditions are:

1. $\boldsymbol{\Omega}$ is a subset in d-dimensional Euclidean space $\mathbb{R}^d$
2. $\boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})} = \{q(\boldsymbol{\omega}) \in \boldsymbol{\Omega} : L(q(\boldsymbol{\omega})) \geq L(q_o(\boldsymbol{\omega}))\}$ is compact for any $L(q_o(\boldsymbol{\omega})) > -\infty$
3. $L(q(\boldsymbol{\omega}))$ is continuous in $\boldsymbol{\Omega}$ and differentiable in the interior of $\boldsymbol{\Omega}$
4. The EM starting point $q^{(0)}(\boldsymbol{\omega})$ satisfies $L(q^{(0)}(\boldsymbol{\omega})) > -\infty$
5. Each $q^{(s)}(\boldsymbol{\omega})$ (denoting the parameters at EM iteration s) lies in the interior of $\boldsymbol{\Omega}$

We verify each condition in turn.

I.2.1. $\boldsymbol{\Omega}$ IS A SUBSET IN d-DIMENSIONAL EUCLIDEAN SPACE $\mathbb{R}^d$

Proof: let d denote the number of classes. Given our assumption that $p(\omega_i) \geq \epsilon \; \forall i$, we observe that the log likelihood is defined so long as $\sum_{i=1}^{d} p(\omega_i|\boldsymbol{x}_k) q(\omega_i) > 0 \; \forall k$. Thus, the domain $\boldsymbol{\Omega}$ is $\{q(\boldsymbol{\omega}) : q(\omega_i) \geq 0 \; \forall i, \sum_{i=1}^{d} q(\omega_i) = 1, \sum_{i=1}^{d} p(\omega_i|\boldsymbol{x}_k) q(\omega_i) > 0 \; \forall k\}$, which is a subset of $\mathbb{R}^d$ (more specifically, a subset of the hyperplane defined by $\sum_{i=1}^{d} q(\omega_i) = 1$). ■

I.2.2. $\boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})} = \{q(\boldsymbol{\omega}) \in \boldsymbol{\Omega} : L(q(\boldsymbol{\omega})) \geq L(q_o(\boldsymbol{\omega}))\}$ IS COMPACT FOR ANY $L(q_o(\boldsymbol{\omega})) > -\infty$

Proof: for a subset of Euclidean space $\mathbb{R}^d$ to be compact, it must be both bounded and closed. The parameter space is bounded because $0 \leq q(\omega_i) \leq 1 \; \forall i$. To show closedness, we will use proof by contradiction. Assume that there exists some $q(\boldsymbol{\omega})$ s.t. $\boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})}$ is open and $L(q_o(\boldsymbol{\omega})) > -\infty$. We begin by noting that $L(q(\boldsymbol{\omega}))$ is finite everywhere in $\boldsymbol{\Omega}$ (by definition of $\boldsymbol{\Omega}$), and is also continuous everywhere in $\boldsymbol{\Omega}$ (proven in **Sec. I.2.3**). If $\boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})}$ were open, it would imply that $\exists x \notin \boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})}$ and a sequence $\{x'_n\} \in \boldsymbol{\Omega}_{q_o(\boldsymbol{\omega})}$ and $x'_n \to x$. From continuity of $L(q(\boldsymbol{\omega}))$ over $\boldsymbol{\Omega}$, we know that if $x \in \boldsymbol{\Omega}$, then $L(x'_n) \geq L(q_o(\boldsymbol{\omega})) \; \forall n$ as $x'_n \to x$ implies that $L(x) \geq L(q_o(\boldsymbol{\omega}))$. In other words, continuity necessitates that if

$x \in \mathbf{\Omega}$ and $x'_n \in \mathbf{\Omega}_{q_o(\boldsymbol{\omega})}$ $\forall n$, then $x \in \mathbf{\Omega}_{q_o(\boldsymbol{\omega})}$. Openness of $\mathbf{\Omega}_{q_o(\boldsymbol{\omega})}$ is therefore only possible if $\exists x, \{x'_n\}$ s.t. $x \notin \mathbf{\Omega}$ and $x'_n \in \mathbf{\Omega}_{q_o(\boldsymbol{\omega})}$ $\forall n$ - in other words, x must exist at an open boundary of $\mathbf{\Omega}$. However, the only open boundaries of $\mathbf{\Omega}$ are $\{q(\boldsymbol{\omega}) : \exists k \text{ s.t. } \sum_i p(\omega_i|\boldsymbol{x}_k)q(\omega_i) = 0\}$. We observe that $L(q(\boldsymbol{\omega})) \to -\infty$ as $\sum_i p(\omega_i|\boldsymbol{x}_k)q(\omega_i) \to 0$ for any k. The set $\mathbf{\Omega}_{q_o(\boldsymbol{\omega})}$ can therefore only be open if $L(q_o(\boldsymbol{\omega})) = -\infty$, leading to a contradiction. ■

I.2.3. $L(q(\boldsymbol{\omega}))$ IS CONTINUOUS IN $\mathbf{\Omega}$ AND DIFFERENTIABLE IN THE INTERIOR OF $\mathbf{\Omega}$

Proof: because differentiability everywhere in the domain implies continuity in the domain, it suffices to show that $L(q(\boldsymbol{\omega}))$ is differentiable at all points in the domain. The partial derivative of the log-likelihood L is:

$$\frac{\partial L(q(\boldsymbol{\omega}))}{\partial q(\omega_i)} = \sum_k \frac{p(\omega_i|\boldsymbol{x}_k)}{p(\omega_i)\left(\sum_j \frac{p(\omega_j|\boldsymbol{x}_k)}{p(\omega_j)} q(\omega_j)\right)} \tag{18}$$

Because we assumed $p(\omega_i) \geq \epsilon \; \forall i$, we note that the derivative is defined as long as $\sum_j p(\omega_j|\boldsymbol{x}_k)q(\omega_j) > 0 \; \forall k$. As the latter condition is a requirement for being in the domain $\mathbf{\Omega}$ (see **Sec. I.2.1**), we conclude that the log-likelihood L is both continuous and differentiable everywhere in $\mathbf{\Omega}$. ■

I.2.4. THE EM STARTING POINT $q^{(0)}(\boldsymbol{\omega})$ SATISFIES $L(q^{(0)}(\boldsymbol{\omega})) > -\infty$

Proof: we set $q^{(0)}$ to be equal to the source domain probabilities $p(\omega_i)$. Substituting $q(\omega_i) = p(\omega_i)$ into the expression for the log-likelihood gives:

$$\begin{aligned} L(q^{(0)}(\boldsymbol{\omega})) &= C + \sum_k \log\left(\sum_i \frac{p(\omega_i|\boldsymbol{x}_k)}{p(\omega_i)} p(\omega_i)\right) \\ &= C + \sum_k \log\left(\sum_i p(\omega_i|\boldsymbol{x}_k)\right) \\ &= C \end{aligned}$$

As a reminder, $C = \sum_k \log(p(\boldsymbol{x}_k))$ is a constant w.r.t. $q(\boldsymbol{\omega})$ and is therefore ignored when optimizing w.r.t. $q(\boldsymbol{\omega})$. ■

I.2.5. EACH $q^{(s)}(\boldsymbol{\omega})$ (DENOTING THE PARAMETERS AT EM ITERATION s) LIES IN THE INTERIOR OF $\mathbf{\Omega}$

We first state what it means to for $q^{(s)}(\boldsymbol{\omega})$ to lie in the interior of $\mathbf{\Omega}$. Because $\mathbf{\Omega}$ is a subset of the Euclidean hyperplane defined by $\sum_i q(\omega_i) = 1$, the interior of $\mathbf{\Omega}$ consists of $\{q(\boldsymbol{\omega}) : q(\omega_i) > 0 \; \forall i \text{ and } \sum_i q(\omega_i) = 1\}$. Similarly, the boundary of $\mathbf{\Omega}$ consists of $\{q(\boldsymbol{\omega}) : \exists i \text{ s.t. } q(\omega_i) = 0 \text{ and } \sum_i q(\omega_i) = 1\}$. We will use $\partial\mathbf{\Omega}$ to denote the boundary of $\mathbf{\Omega}$ and $(\mathbf{\Omega} \setminus \partial\mathbf{\Omega})$ to denote the interior.

We now consider the condition that $q^{(s)}(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega}) \; \forall s$. We first show that this condition is satisfied when $\sum_k p(\omega_i|\boldsymbol{x}_k) > 0 \; \forall i$.

Lemma: if $\sum_k p(\omega_i|\boldsymbol{x}_k) > 0 \; \forall i$ and $p(\omega_i) \geq \epsilon \; \forall i$, then $q^{(s)}(\omega_i) > 0 \; \forall i, s$.
Proof: Note that $\sum_k p(\omega_i|\boldsymbol{x}_k) > 0$ implies that $\exists k$ s.t. $p(\omega_i|\boldsymbol{x}_k) > 0$. We use proof by induction. The base case is satisfied because $q^{(0)}(\boldsymbol{\omega})$ is initialized to $p(\boldsymbol{\omega})$, which (by assumption) satisfies $p(\omega_i) \geq \epsilon \; \forall i$. We now observe that the product $q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)$ is the numerator of $q^{(s)}(\omega_i|\boldsymbol{x}_k) = \frac{q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)}{p(\omega_i)\sum_{j=1}^m q^{(s)}(\omega_j)p(\omega_j|\boldsymbol{x}_k)}$; thus, $q^{(s)}(\omega_i|\boldsymbol{x}_k) > 0$ if $q^{(s)}(\omega_i) > 0$ and $p(\omega_i|\boldsymbol{x}_k) > 0$. It therefore follows that $q^{(s+1)}(\omega_i) = \frac{1}{N}\sum_{k=1}^N q^{(s)}(\omega_i|\boldsymbol{x}_k) > 0$ if $q^{(s)}(\omega_i)$ and $\exists k$ s.t. $p(\omega_i|\boldsymbol{x}_k) > 0$. ■

We now consider the case where $\exists i'$ s.t. $\sum_k p(\omega_{i'}|\boldsymbol{x}_k) = 0$. In this case, the condition is not technically satisfied because $q^{(s)}(\omega_{i'}|\boldsymbol{x}_k) = 0$ when $p(\omega_{i'}|\boldsymbol{x}_k) = 0$, and thus $q^{(s+1)}(\omega_{i'}) = \frac{1}{N}\sum_{k=1}^N q^{(s)}(\omega_{i'}|\boldsymbol{x}_k)$ would be 0 for all $k+1 > 0$. However, the EM updates in this case would be equivalent to performing EM on a reduced problem where class i' is simply excluded from the EM optimization if $\sum_k p(\omega_{i'}|\boldsymbol{x}_k) = 0$. In the reduced problem, we would be guaranteed that $\sum_k p(\omega_i|\boldsymbol{x}_k) > 0 \; \forall i$, which (as we showed above) ensures that the EM parameters always lie in the interior of the parameter space.

I.3. Proving Differentiability of $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ over $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$

Given our assumption that $p(\omega_i) \geq \epsilon \; \forall i$, we will show that $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ is not only continuous but also differentiable over $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$. Because $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ is a composition of differentiable functions, to prove differentiability over $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$ it suffices to show that $Q(q(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ is defined over $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$.

Building on the expression for $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ from **Eqn. 11**, we have:

$$
\begin{aligned}
Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega})) &= \sum_{k=1}^{N}\sum_{i=1}^{m} q^{(s)}(\omega_i|\boldsymbol{x}_k)\left(\log q(\omega_i) + \log p(\boldsymbol{x}_k|\omega_i)\right) \\
&= \sum_{k=1}^{N}\sum_{i=1}^{m} q^{(s)}(\omega_i|\boldsymbol{x}_k)\left(\log q(\omega_i) + \log\left[\frac{p(\omega_i|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\omega_i)}\right]\right) \qquad (19) \\
&= \sum_{k=1}^{N}\sum_{i=1}^{m} \frac{q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)\left(\log q(\omega_i) + \log\left[\frac{p(\omega_i|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\omega_i)}\right]\right)}{p(\omega_i)\sum_{j=1}^{m}\left(\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)\right)} \qquad (20) \\
&= \sum_{k=1}^{N}\sum_{i=1}^{m} \frac{q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)\left(\log q(\omega_i) + \log p(\omega_i|\boldsymbol{x}_k) + \log p(\boldsymbol{x}_k) - \log p(\omega_i)\right)}{p(\omega_i)\sum_{j=1}^{m}\left(\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)\right)} \\
&= \sum_{k=1}^{N}\sum_{i=1}^{m} \frac{q^{(s)}(\omega_i)p(\omega_i|\boldsymbol{x}_k)\left(\log q(\omega_i) + \log p(\boldsymbol{x}_k) - \log p(\omega_i)\right) + q^{(s)}(\omega_i)\log\left[p(\omega_i|\boldsymbol{x}_k)^{p(\omega_i|\boldsymbol{x}_k)}\right]}{p(\omega_i)\sum_{j=1}^{m}\left(\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)\right)} \qquad (21)
\end{aligned}
$$

Where (19) comes from applying Bayes' rule to $p(\boldsymbol{x}_k|\omega_i)$, and (20) follows from substituting the definition of $q^{(s)}(\omega_i|\boldsymbol{x}_k)$ from **Eqn. 9**.

We make two observations: first, because $x^x > 0$ for all $x \geq 0$, the term $\log\left[p(\omega|\boldsymbol{x}_k)^{p(\omega|\boldsymbol{x}_k)}\right]$ is always defined. Second, because $\log p(\boldsymbol{x_k})$ does not depend on $q(\omega_i)$, it disappears when optimizing $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ w.r.t $q(\omega_i)$. If we assume, as before, that $p(\omega_i) \geq \epsilon \; \forall i$, we see that $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is defined if and only if $q(\omega_i) > 0 \; \forall i$ and $\sum_{j=1}^{m}\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)$ $\forall k$. The first condition holds true when $q(\boldsymbol{\omega}) \in (\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega})$, and the second condition holds true for all $q^{(s)}(\boldsymbol{\omega}) \in \boldsymbol{\Omega}$. Thus, we conclude that the domain of $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$. As stated earlier, because $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is composed of differentiable functions, it follows that $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is both continuous and differentiable over $(\boldsymbol{\Omega} \setminus \partial\boldsymbol{\Omega}) \bigtimes \boldsymbol{\Omega}$.

I.4. Invoking Theorem 1 of Wu (1983) to Prove EM Converges to a Stationary Point of the Likelihood

We now show that EM converges to a stationary point of the log-likelihood. We will closely follow the derivation of Theorem 2 in Wu (1983), but modifying it so as not to assume that $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is defined for all $q(\boldsymbol{\omega}) \in \boldsymbol{\Omega}$. Our proof will leverage Theorem 1 of Wu (1983), which is reproduced below for convenience:

Theorem 1 of Wu (1983): Let $\boldsymbol{\Omega}$ denote the domain of the likelihood function $L(\Phi)$, let $\mathscr{S}$ denote the set of stationary points in the interior of $\boldsymbol{\Omega}$, and let $\{\Phi_p\}$ denote a sequence of Generalized EM (GEM) parameter updates generated by $\Phi_{p+1} \in M(\Phi_p)$. If (i) M is a closed point-to-set map over the complement of $\mathscr{S}$ and (ii) $L(\phi_{p+1}) > L(\phi)$ for all $\phi_p \notin \mathscr{S}$. Then all the limit points of $\{\Phi_p\}$ are stationary points of L, and $L(\Phi_p)$ converges monotonically to $L^* = L(\Phi^*)$ for some $\Phi^* \in \mathscr{S}$. ■

To prove convergence to a stationary point, it suffices to show that both (i) and (ii) of Theorem 1 apply to the case of domain adaptation to label shift. We verify (i) and (ii) in turn below.

I.4.1. (i) of Theorem 1: M is a closed point-to-set map over the complement of $\mathscr{S}$

We first state what it means for M to be a closed point-to-set map. A map M from points of X to subsets of X is called a point-to-set map on X. In our case, M is a point-to-point map (a special case of a point-to-set map) given by the EM parameter update. A point-to-point map is said to be *closed* at x if $x_k \to x$, $x \in X$ and $y_k \to y$, $y_k = A(x_k)$ implies $y = A(x)$. For a point-to-point map, continuity implies closedness. Because the complement of $\mathscr{S}$ corresponds to the

domain $\mathbf{\Omega} \setminus \mathscr{S}$, if we show that M is continuous over the entirety of $\mathbf{\Omega}$, we would satisfy our conditions.

Referring to **Eqn. 17**, we see that M is defined as:

$$\begin{aligned} M(q^{(s)}(\boldsymbol{\omega}))_i &= \frac{1}{N}\sum_{k=1}^{N} q^{(s)}(\omega_i|\boldsymbol{x}_k) \\ &= \frac{1}{N}\sum_{k=1}^{N} \frac{\frac{q^{(s)}(\omega_i)}{p(\omega_i)}p(\omega_i|\boldsymbol{x}_k)}{\sum_{j=1}^{m}\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)} \end{aligned}$$

Because M is composed of differentiable functions, to show continuity of M over $\mathbf{\Omega}$ it suffices to show that $M(q^{(s)}(\boldsymbol{\omega}))$ is defined for all $q^{(s)}(\boldsymbol{\omega}) \in \mathbf{\Omega}$. We observe that, given our assumption of $p(\omega_i) \geq \epsilon \; \forall i$, the numerator $\frac{q^{(s)}(\omega_i)}{p(\omega_i)}$ is always defined. In order for the denominator $\sum_{j=1}^{m}\frac{q^{(s)}(\omega_j)}{p(\omega_j)}p(\omega_j|\boldsymbol{x}_k)$ to be defined $\forall k$, we additionally require that $\sum_j q^{(s)}(\omega_j)p(\omega_j|\boldsymbol{x}_k) > 0 \; \forall k$. The latter condition is satisfied if $q^{(s)}(\boldsymbol{\omega}) \in \mathbf{\Omega}$. Thus, we conclude that M is continuous over $\mathbf{\Omega}$, fulfilling (i) of Theorem 1 of Wu (1983). ■

I.4.2. (II) OF THEOREM 1: $L(\phi_{p+1}) > L(\phi)$ FOR ALL $\phi_p \notin \mathscr{S}$

We first show that $\frac{\partial H(q(\boldsymbol{\omega});q^{(s)}(\boldsymbol{\omega}))}{\partial q(\boldsymbol{\omega})} = 0$ when $q(\boldsymbol{\omega}) = q^{(s)}(\boldsymbol{\omega})$. For notational convenience, we will use $D^{10}H(a;b)$ and $D^{10}Q(a;b)$ to denote $\frac{\partial H(\Phi_1;\Phi_2)}{\partial \Phi_1}$ and $\frac{\partial Q(\Phi_1;\Phi_2)}{\partial \Phi_1}$ at $\Phi_1 = a$ and $\Phi_2 = b$, and will use $DL(a)$ to denote $\frac{\partial L(\Phi)}{\partial \Phi}$ at $\Phi = a$.

In **Sec. I.3**, we established that $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is differentiable for all $q(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega})$. Because $H(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega})) := Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega})) - L(q(\boldsymbol{\omega}))$, it follows that $H(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is also differentiable for all $q(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega})$. In **Sec. I.2.5**, we established that $q^{(s)}(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega}) \; \forall s$; thus, it follows that $H(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is differentiable in the neighborhood of $q^{(s)}(\boldsymbol{\omega})$. Because $q(\boldsymbol{\omega}) = q^{(s)}(\boldsymbol{\omega})$ maximizes $H(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ (**Eqn. 15**), we conclude that $D^{10}H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) = 0$.

We now show that $D^{10}H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) = 0$ and $q^{(s)}(\boldsymbol{\omega}) \notin \mathscr{S}$ implies $Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) > Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$. From the definition of $H(q^{(s)}(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$, we have $L(q^{(s)}(\boldsymbol{\omega})) = Q(q^{(s)}(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega})) - H(q^{(s)}(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$. If $D^{10}H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) = 0$ and $q^{(s)}(\boldsymbol{\omega}) \notin \mathscr{S}$, we have $DL(q^{(s)}(\boldsymbol{\omega})) = D^{10}Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \neq 0$. Because $q^{(s)}(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega})$ (**Sec. I.2.5**) and $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ is both defined and differentiable $\forall q(\boldsymbol{\omega}) \in (\mathbf{\Omega} \setminus \partial\mathbf{\Omega})$ (**Sec. I.3**), we conclude that if $D^{10}Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \neq 0$, then $q(\boldsymbol{\omega}) = q^{(s)}(\boldsymbol{\omega})$ does not maximize $Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$. Because $q^{(s+1)}(\boldsymbol{\omega}) = \arg\max_{q(\boldsymbol{\omega})} Q(q(\boldsymbol{\omega}), q^{(s)}(\boldsymbol{\omega}))$ by definition, we get $Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) > Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$ when $q^{(s)}(\boldsymbol{\omega}) \notin \mathscr{S}$.

Finally, we show that if $Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) > Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega}))$, then $L(q^{(s+1)}(\boldsymbol{\omega})) > L(q^{(s)}(\boldsymbol{\omega}))$. Analogous to the derivation of **Eqn. 16**, we have:

$$\begin{aligned} L(q^{(s+1)}(\boldsymbol{\omega})) &= Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \text{ (From the definition of } H) \\ &\quad \text{If } Q(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) > Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})), \text{ we get:} \\ L(q^{(s+1)}(\boldsymbol{\omega})) &> Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s+1)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \\ &\geq Q(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) - H(q^{(s)}(\boldsymbol{\omega}); q^{(s)}(\boldsymbol{\omega})) \text{ (From } \textbf{Eqn. 15}) \\ &= L(q^{(s)}(\boldsymbol{\omega})) \end{aligned}$$

Thus, we have shown that $L(q^{(s+1)}(\boldsymbol{\omega})) > L(q^{(s)}(\boldsymbol{\omega})) \; \forall q^{(s)}(\boldsymbol{\omega}) \notin \mathscr{S}$, fulfilling (ii) of Theorem 1 of Wu (1983).

■