Stabilizing EM-Based Quantification under Label Shift

Pablo González, Olaya Pérez-Mon, Juan José del Coz

Artificial Intelligence Center
University of Oviedo

LQ 2025 · September 15, 2025

Quantification and EMQ

- Quantification goal: estimate **class prevalences** in an *unlabelled* target sample.
- **Assumptions** for most quantification methods:
  - **Prior probability shift**: $P(X \mid Y) = Q(X \mid Y)$ while $P(Y) \neq Q(Y)$.
- Under these conditions, **EMQ**
  - coincides with the maximum-likelihood estimator,
  - is *Fisher-consistent*: perfect estimation as the sample size grows,
  - fast and a very strong baseline for quantification problems and for label shift adaptation.
- Yet, in practical problems, mild violations of the assumptions (e.g. mis-calibration of the underlying classifier) can lead to oscillations and degenerate estimates $\Rightarrow$ motivation for stabilization.

Baseline: EMQ

EMQ is an iterative algorithm that uses Bayes' rule to estimate the target class distribution $Q(c)$ by alternating between two steps:

- $Q(c)$: Prevalence of class c in the target set.
- $P(c)$: Prevalence of class c in the training set.
- $s_i(c)$: Score (e.g., soft prediction) assigned to class c by the classifier for instance i.
- $q_i^{(t)}(c)$: Estimated probability that instance i belongs to class c at iteration t.

E-step (Expectation)

Update instance-level posteriors:

$$q_i^{(t)}(c) = \frac{\frac{Q^{(t)}(c)}{P(c)}\, s_i(c)}{\sum_{j=1}^{K} \frac{Q^{(t)}(j)}{P(j)}\, s_i(j)}$$

M-step (Maximization)

Update class prevalences:

$$Q^{(t+1)}(c) = \frac{1}{N} \sum_{i=1}^{N} q_i^{(t)}(c)$$

Practical Limitations of EMQ

Although EMQ is a theoretically sound and widely adopted baseline, its practical performance is often mixed, especially in real-world settings:

- Highly sensitive to **mis-calibrated** posteriors.
- Tends to produce **degenerate estimates** when some classes are rare in the test distribution.
- In the **multiclass** setting, errors can propagate and compound across classes.
- **Empirical studies** in the literature report strong performance in some datasets, but also significant failures in others (even when a good calibrator is used).

Our Contributions

1. **Unified Framework**: Decompose EMQ into two modular components:
   - E-step: *Reweighted posterior Transformation* $\mathcal{C}$
   - M-step: *Prevalence Update* $\mathcal{U}$
2. Systematic taxonomy of existing & novel stabilisation heuristics (both for the E-step and the M-step).
3. Exhaustive evaluation on 20+ UCI datasets.

Unified EMQ Framework

Algorithm 1: Unified EMQ Framework

Input: Unlabeled instances $\mathcal{D}_T = \{x_1, \ldots, x_n\}$, classifier posteriors $s_i(c)$, posterior transformation $\mathcal{C}$, update function $\mathcal{U}$

Output: Estimated target class prevalences $Q(Y)$

Initialize prevalence estimate $Q^{(0)} \in \Delta^K$

repeat

  foreach *instance* $x_i \in \mathcal{D}_T$ **do**

    Compute reweighted posterior:

$$q_i^{(t)}(c) \leftarrow \frac{\frac{Q^{(t)}(c)}{P(c)} \cdot s_i(c)}{\sum_{j=1}^{K} \frac{Q^{(t)}(j)}{P(j)} \cdot s_i(j)}$$

    Apply reweighted posterior transformation:

$$\tilde{q}_i^{(t)} \leftarrow \mathcal{C}\big(q_i^{(t)}\big)$$

  end

  Compute raw prevalence update:

$$\hat{Q}^{(t+1)}(c) \leftarrow \frac{1}{N} \sum_{i=1}^{N} \tilde{q}_i^{(t)}(c)$$

  Update prevalence estimate:

$$Q^{(t+1)} \leftarrow \mathcal{U}\big(Q^{(t)},\ \hat{Q}^{(t+1)},\ \{\tilde{q}_i^{(t)}\}_{i=1}^{N}\big)$$

until *convergence*

return $Q^{(t+1)}$

E-Step Heuristics ($\mathcal{C}$)

- **Calibration (BCTS)**: Applies a post-processing transformation to classifier scores to improve reliability.
  Used only once before EM begins and helps align $s_i(c)$ with true posteriors (not part of the framework).
- **Posterior Smoothing**: $\tilde{q}_i^{(t)}(c) \leftarrow (q_i^{(t)}(c) + \varepsilon)/(1 + K \cdot \varepsilon)$
  Prevents zero probabilities and softens overconfident predictions, especially useful when some classes are rare or underrepresented.
- **Temperature Scaling**: $\tilde{q}_i^{(t)}(c) \leftarrow q_i^{(t)}(c)^{1/\tau}$ then renormalize
  Dynamically flattens (or sharpens) the score distribution during EMQ. Applies a power transformation to soften or sharpen posteriors during EMQ:
  - $\tau > 1$: flattens the distribution, reducing overconfidence.
  - $\tau < 1$: sharpens the distribution, making predictions more confident.
  Helps mitigate overconfident predictions and stabilizes updates.

M-Step Heuristics ($\mathcal{U}$)

- **MAP / Dirichlet Prior**: Add α-1 pseudo-counts to the prevalence update $Q^{(t+1)}(c) = \frac{\sum_{i=1}^{N} \tilde{q}_i^{(t)}(c)+\alpha-1}{N+K(\alpha-1)}$. When $\alpha > 1$, the prevalence updates are pulled towards the uniform distribution.
- **Damping**: Blend the new estimate with the previous one $Q^{(t+1)} \leftarrow (1-\lambda)Q^{(t)} + \lambda\hat{Q}^{(t+1)}$
  Slows down updates to avoid abrupt changes; effective for noisy or unstable posteriors.
- **Confidence Selection**: Use only the most confident instances to compute $Q^{(t+1)}$.
  Reduces the influence of uncertain predictions, filtering top-κ instances with highest max-score.
- **Entropy Regularization**: Encourages high-entropy (i.e., smoother) prevalence vectors:

$$\arg\max_{Q} \sum_{c=1}^{K} \hat{Q}(c) \log Q(c) + \eta H(Q)$$

  Acts as a regularizer to prevent overconfident or peaked estimates. η controls the strength of the penalty.

Experimental Setup

- **Datasets**: 24 multiclass UCI datasets, test bags = 1000 (per dataset), bag size = 500.
- **Metric**: Mean Absolute Error (MAE).
- **Hyper-parameters**: tuned on validation bags via *QuaPy* along with the classifier hyperparameters.

Heuristic	Hyperparameter range
Smooth	$\epsilon \in \{\mathbf{0}, 10^{-6}, 10^{-5}, 10^{-4}\}$
Temp	$\tau \in \{0.25, 0.5, 0.75, \mathbf{1.0}, 1.25, 1.5, 2.0, 3.0, 5.0\}$
Damp	$\lambda \in \{0.1, 0.2, \ldots, \mathbf{1.0}\}$
Ent	$\eta \in \{\mathbf{0.0}, 0.0001, 0.001\}$
MAP	$\alpha \in \{\mathbf{1.0}, 2.0, 5.0, 10.0\}$
Conf	$\kappa \in \{0.3, 0.5, 0.8, \mathbf{1.0}\}$

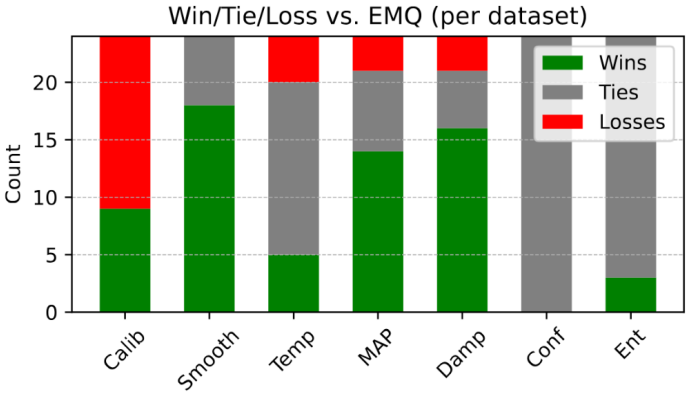
Overall Results

	Baseline		E-step mods		M-step mods			
	EMQ	Calib	Smooth	Temp	MAP	Damp	Conf	Ent
abalone	0.07050	0.07462	0.06753	0.05142	**0.05077**	0.05391	0.07050	0.05731
academic-success	**0.03074**	0.03184	0.03074	0.03074	0.04146	0.03074	0.03074	0.03074
chess	0.03563	0.03516	0.03441	0.03284	**0.02903**	0.03006	0.03563	0.03563
cmc	0.09456	0.11260	0.09403	0.09456	0.10482	**0.08968**	0.09456	0.09456
connect-4	0.03653	0.03489	0.03623	0.03653	0.03403	**0.03327**	0.03653	0.03653
digits	0.00271	**0.00270**	0.00271	0.00271	0.00271	0.00271	0.00271	0.00271
dry-bean	0.00425	**0.00409**	0.00423	0.00425	0.00425	0.00419	0.00425	0.00425
hand_digits	0.00286	0.00295	0.00286	0.00286	0.00286	**0.00285**	0.00286	0.00286
hcv	0.20780	0.30409	0.20619	0.16657	0.17029	0.19232	0.20780	**0.16233**
image_seg	0.00503	0.00529	0.00503	0.00547	0.00503	**0.00499**	0.00503	0.00503
isolet	0.00175	**0.00163**	0.00175	0.00174	0.00175	0.00168	0.00175	0.00175
letter	0.00545	0.00551	0.00539	0.00551	0.00519	**0.00511**	0.00545	0.00545
mhr	0.05940	0.06040	0.05873	0.05940	**0.05383**	0.05394	0.05940	0.05940
molecular	0.00847	**0.00788**	0.00846	0.00847	0.00832	0.00840	0.00847	0.00847
nursery	0.00579	0.00583	**0.00578**	0.00579	0.00579	0.00579	0.00579	0.00579
obesity	0.00619	0.00832	0.00617	0.00647	0.00614	**0.00603**	0.00619	0.00619
page_block	0.02440	**0.02061**	0.02431	0.02440	0.02313	0.02440	0.02440	0.02440
phishing	0.02902	0.02537	0.02878	0.02902	**0.02340**	0.02902	0.02902	0.02902
poker_hand	0.14458	0.17265	0.13097	0.08541	**0.08284**	0.15854	0.14458	0.08432
satellite	0.00954	0.00979	0.00949	0.00954	0.00930	**0.00920**	0.00954	0.00954
shuttle	0.06248	**0.04674**	0.06184	0.06248	0.06222	0.06248	0.06248	0.06248
waveform-v1	**0.00970**	0.01014	0.00970	0.01154	0.00985	0.00971	0.00970	0.00970
wine-quality	0.07639	0.07736	0.07043	0.07639	**0.06082**	0.06391	0.07639	0.07639
yeast	0.04386	0.04554	0.04363	0.04386	0.04386	**0.04336**	0.04386	0.04386
Mean	0.04073	0.04608	0.03956	0.03575	**0.03507**	0.03860	0.04073	0.03578

Bag-level: statistical analysis

Heuristic	Mean (MAE)	Std	% Wins	Wilcoxon p-value
Calib	0.00535	0.03005	0.45	1.0000
Smooth	-0.00118	0.00506	0.63	$<$ **0.0001**
Temp	-0.00499	0.02179	0.26	$<$ **0.0001**
MAP	-0.00566	0.02307	0.44	$<$ **0.0001**
Damp	-0.00214	0.01518	0.49	$<$ **0.0001**
Conf	0.00000	0.00000	0.0	–
Ent	-0.00496	0.02299	0.1	$<$ **0.0001**

Dataset level: Win-ties-losses charts



Combining heuristics

	EMQ	SmoothDamp	SmoothEnt	SmoothMAP	TempDamp	TempEnt	TempMAP
abalone	0.07050	0.05142	0.05142	**0.05077**	0.05362	0.05731	0.05078
academic-success	**0.03074**	0.03074	0.03074	0.04146	0.03074	0.03074	0.04129
chess	0.03563	0.03006	0.03284	0.02903	0.03065	0.03441	**0.02898**
cmc	0.09456	0.08968	0.09456	0.10482	**0.08955**	0.09403	0.10482
connect-4	0.03653	**0.03327**	0.03653	0.03403	0.03467	0.03623	0.03403
digits	0.00271	**0.00270**	0.00271	0.00271	0.00272	0.00271	0.00271
dry-bean	0.00425	**0.00419**	0.00425	0.00425	0.00423	0.00423	0.00423
hand_digits	0.00286	**0.00285**	0.00286	0.00286	**0.00285**	0.00286	0.00286
hcv	0.20780	0.19232	0.16657	0.17029	0.19104	**0.16233**	0.17029
image_seg	0.00503	0.00547	0.00547	0.00547	**0.00499**	0.00503	0.00503
isolet	0.00175	**0.00162**	0.00174	0.00174	0.00168	0.00175	0.00175
letter	0.00545	**0.00511**	0.00551	0.00519	**0.00511**	0.00539	0.00519
mhr	0.05940	0.05394	0.05940	0.05383	0.05382	0.05873	**0.05375**
molecular	0.00847	0.00840	0.00847	0.00908	0.00840	0.00846	**0.00832**
nursery	0.00579	0.00643	0.00579	0.00579	**0.00578**	**0.00578**	**0.00578**
obesity	0.00619	**0.00603**	0.00647	0.00614	**0.00603**	0.00617	0.00615
page_block	0.02440	0.02440	0.02440	0.02313	0.02431	0.02431	**0.02308**
phishing	0.02902	0.02902	0.02902	**0.02340**	0.02878	0.02878	**0.02340**
poker_hand	0.14458	0.08540	0.08382	**0.08284**	0.14292	0.08432	**0.08284**
satellite	0.00954	0.00920	0.00954	0.00930	**0.00919**	0.00949	0.00930
shuttle	0.06248	0.06248	0.06248	0.06222	**0.06184**	**0.06184**	0.06222
waveform-v1	**0.00970**	0.01122	0.01154	0.01076	0.00972	0.00970	0.00970
wine-quality	0.07639	0.06391	0.07639	0.06082	0.06352	0.07043	**0.06069**
yeast	0.04386	0.04336	0.04386	0.04386	**0.04334**	0.04363	0.04363
Mean	0.04073	0.03555	0.03568	0.03516	0.03790	0.03536	**0.03503**

Conclusions and Future work

- We presented a **unified framework** + extensive heuristic benchmarking.
- Simple heuristics can yield significant improvements over standard EMQ.
- Presented heuristics are lightweight, easy to implement and pretty safe to use, even though they should be considered based on the data.
- Future work
  - explore automatic selection of heuristics.
  - derive consistency guarantees (what properties should the functions $\mathcal{C}$ and $\mathcal{U}$ have so the resulting method is still Fisher consistent?).
  - explore heuristics for other types of shifts or situations where basic EMQ might fail.

Questions?