Robust Active Label Correction (Supplementary Material)

Jan Kremer
Department of Computer Science
University of Copenhagen

Fei Sha
Department of Computer Science
University of Southern California

Christian Igel
Department of Computer Science
University of Copenhagen

A ROBUST UNBIASED WEIGHTED UNCERTAINTY RE-LABELING (U-WURL)

If we know the noise rates ρ_{-1} and ρ_{+1}, we can define a loss ℓ_{u} on the noisy data as an unbiased estimator of the standard loss ℓ on the clean data,

$$\mathbb{E}_{\tilde{y}}\Big[\ell_{\mathrm{u}}(h(x),\tilde{y})\Big] = \ell(h(x),y)$$

for all x, y, and h. For binary classification, Natarajan et al. (2013) have shown that this holds for

$$\ell_{\mathrm{u}}(h(x),y) = \alpha_y \ell(h(x),y) - \beta_y \ell(h(x),-y) \ ,$$

with

$$\alpha_y := \frac{1-\rho_{-y}}{1-\rho_{-1}-\rho_{+1}} \ \text{ and } \ \beta_y := \frac{\rho_y}{1-\rho_{-1}-\rho_{+1}} \ .$$

Choosing the unbiased estimator ℓ_{u} for the noise-aware loss $\tilde{\ell}$ gives for logistic regression

$$\begin{aligned} L(\sigma) = &\sum_{(x,y)\in\mathcal{S}} \ell(\sigma(x),y) \\ &+ \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \Big(\alpha_{\tilde{y}}\ell(\sigma(x),\tilde{y}) - \beta_{\tilde{y}}\ell(\sigma(x),-\tilde{y})\Big) + \Omega(w) \end{aligned} \tag{1}$$

with gradient

$$\begin{aligned} g(\sigma) = &-\sum_{(x,y)\in\mathcal{S}} yx\sigma(-yx) \\ &-\sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \Big(\alpha_{\tilde{y}}\tilde{y}x\sigma(-\tilde{y}x) + \beta_{\tilde{y}}\tilde{y}x\sigma(\tilde{y}x)\Big) + \frac{\partial}{\partial w}\Omega(w) \ . \end{aligned} \tag{2}$$

Following our selection criterion, we then correct the example $(x^*,\tilde{y}^*)$ that maximizes

$$\begin{aligned} &\mathbb{E}_{y|x,\tilde{y}}\Big[\|g_j(\sigma) - g(\sigma)\|\Big] \\ &= \mathbb{E}_{y|x,\tilde{y}}\Big[\|x\|\Big| - y\sigma(-yx) + \tilde{y}\alpha_{\tilde{y}}\sigma(-\tilde{y}x) + \tilde{y}\beta_{\tilde{y}}\sigma(\tilde{y}x)\Big|\Big] \\ &= \|x\|\big(\alpha_{\tilde{y}} - p(Y=\tilde{y}|\tilde{Y}=\tilde{y},X=x)\big) \qquad (3) \\ &= \|x\|\Big(\frac{1-\rho_{-\tilde{y}}}{1-\rho_{-1}-\rho_{+1}} - \frac{(1-\rho_{\tilde{y}})\sigma(\tilde{y}x)}{\rho_{-\tilde{y}}\sigma(-\tilde{y}x) + (1-\rho_{\tilde{y}})\sigma(\tilde{y}x)}\Big) \end{aligned}$$

To arrive at Eq. (3) we use $\alpha_y - \beta_y = 1$ and $\sigma(-yx) = 1-\sigma(yx)$. Thus, we get the selection criterion

$$\begin{aligned} (x^*,\tilde{y}^*) &= \underset{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}}{\arg\max}\ \mathbb{E}_{y_j|x_j,\tilde{y}_j}\Big[\|g_j - g\|\Big] \\ &:= \underset{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}}{\arg\max}\ s_{\mathrm{U}}(x_j,\tilde{y}_j) \ , \end{aligned}$$

where

$$\begin{aligned} &s_{\mathrm{U}}(x_j,\tilde{y}_j) = \\ &\|x_j\|\Big(\frac{1-\rho_{-\tilde{y}_j}}{1-\rho_{-1}-\rho_{+1}} - \frac{(1-\rho_{\tilde{y}_j})\sigma(\tilde{y}_jx_j)}{\rho_{-\tilde{y}_j}\sigma(-\tilde{y}_jx_j) + (1-\rho_{\tilde{y}_j})\sigma(\tilde{y}_jx_j)}\Big) \ . \end{aligned}$$

This method has the advantage that the optimization of Eq. (1) is a convex problem in the case of the logistic loss (Natarajan et al., 2013).

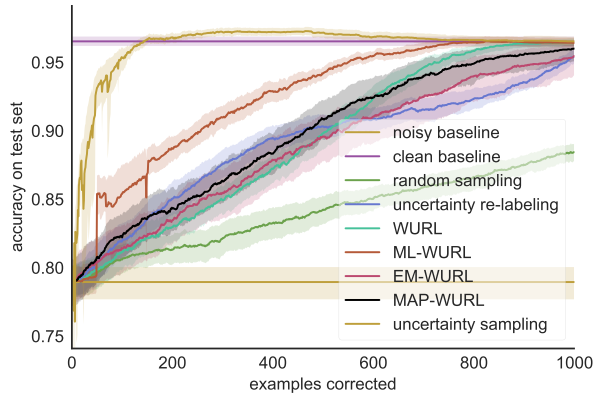
B ADDITIONAL EXPERIMENTS

We present results for the same datasets as in the main section plus several additional ones. We vary parameter settings to demonstrate that the overall performance of the proposed algorithms, relative to their competitors, is unaffected by specific choices. Furthermore, we show the performance of robust unbiased weighted uncertainty re-labeling (U-WURL).

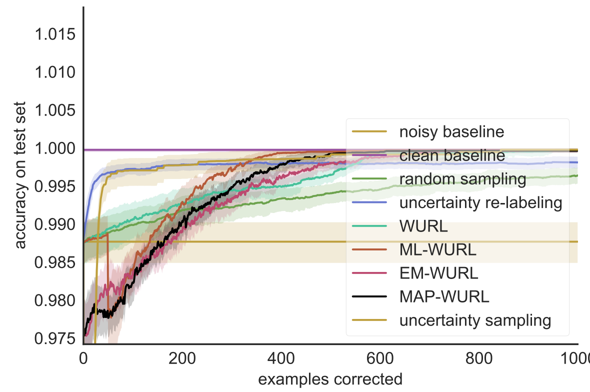
In general, the behavior of the algorithms matches the one shown on the datasets in the main section. There are only minor differences: In the 'mushrooms' experiments the noise rates are initially estimated incorrectly and the performance degrades, but the algorithms are able to quickly recover from the bad start. On the dataset 'ijcnn1' the performance of uncertainty re-labeling degrades quickly after the start, while the others perform as expected.

B.1 Additional datasets

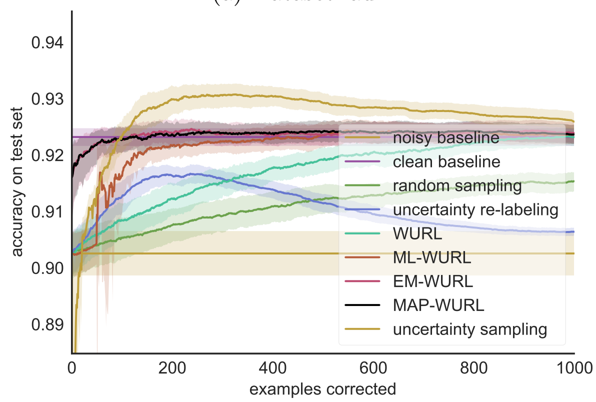
Figure 1 depicts some more results on additional datasets using the same parameter settings as in the main section. The algorithms showed the same behavior as with the parameter settings considered earlier.



(a) Dataset ‘ad’.



(b) Dataset ‘mushrooms’.



(c) Dataset ‘ijcnn1’.

Figure 1: Empirical results on the additional datasets with the parameters as in the main section: $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$, $\lambda = 1.0$, $n_{\text{burn-in}} = 50$, $\kappa = 0.5$.

B.2 Different choice of noise rate ρ_{-1}

We set the noise rates to values with smaller difference, that is, $\rho_{-1} = 0.2$ and $\rho_{+1} = 0.1$. The results are shown in Figure 2. It can be seen that the robust algorithms still perform well. It showed comparable performance as in the case of the larger difference in the noise rates.

B.3 Different choice of regularization λ

We set the regularization parameter to $\lambda = 10.0$, shown in Figure 3. The curves show a similar behavior as in the main section.

B.4 Different choice of burn-in sample size $n_{\textbf{burn-in}}$

We set the burn-in sample size to $n_{\text{burn-in}} = 0$, shown in Figure 4. We see that during the first few iterations ML-WURL fluctuated more than without initial uniform sampling for stabilization. However, it recovered quickly, converging to the clean baseline as fast as in the other settings.

B.5 Different choice of exploration parameter κ

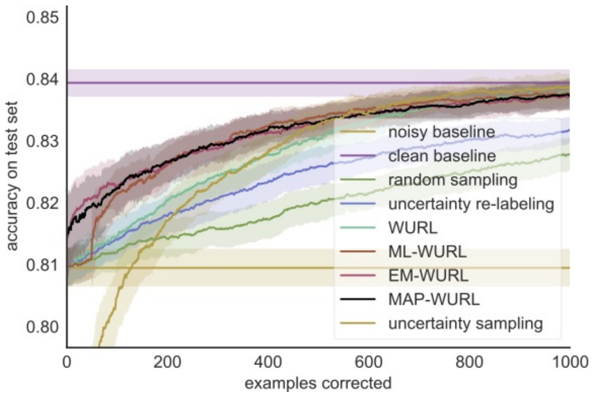
We set the exploration parameter $\kappa = 0.1$, shown in Figure 5. This lets the probability of picking an example uniformly at random decay more slowly with the number of re-labeled examples. We see that the algorithms behave more stably during the first few iterations.

B.6 Performance of the unbiased estimator

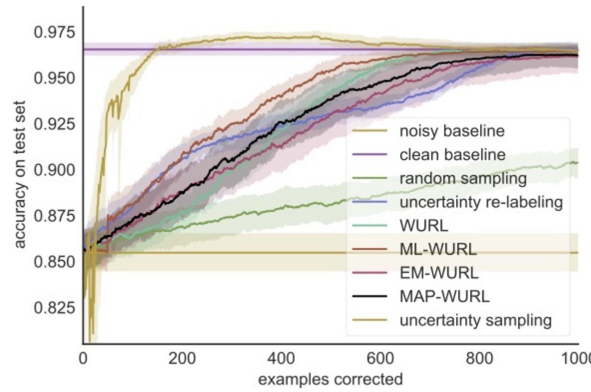
Due to its inferior performance we excluded U-WURL from the main section. Here, we include it for completeness, see Figure 6. Although on some datasets it performed better than random sampling, it performed always worse than the other active methods.

References

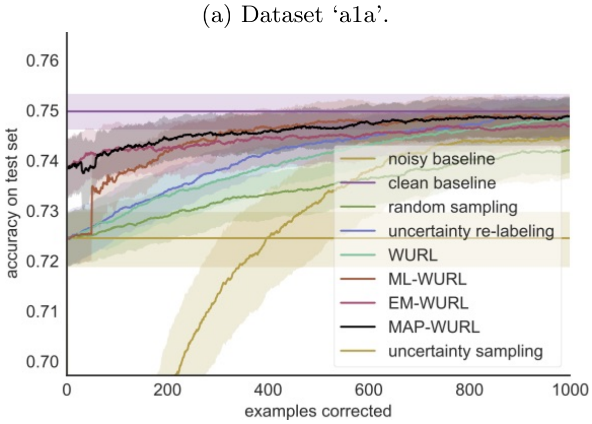
N. Natarajan, I. S. Dhillon, P. D. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.



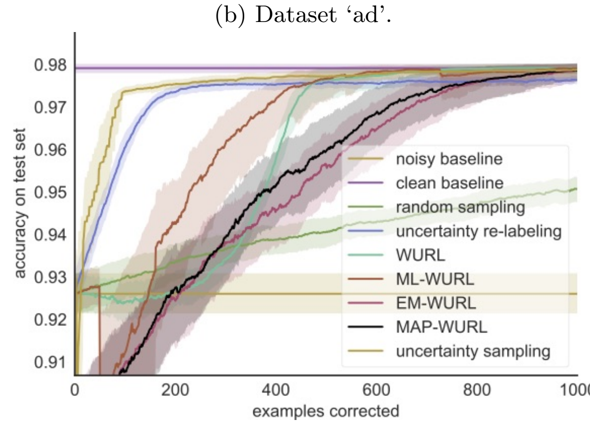
(a) Dataset 'a1a'.



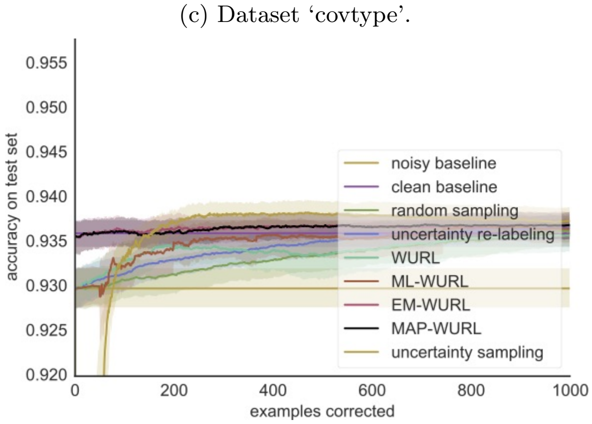
(b) Dataset 'ad'.



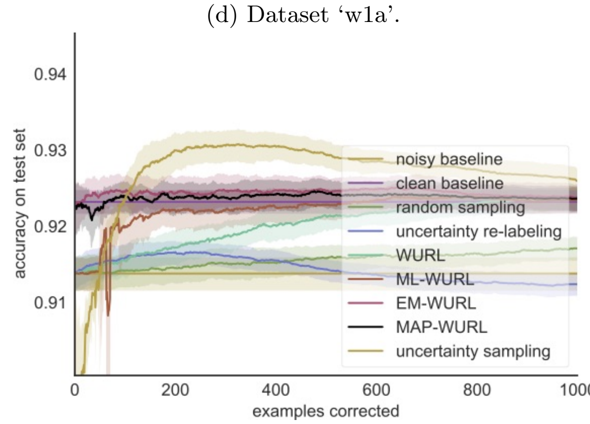
(c) Dataset 'covtype'.



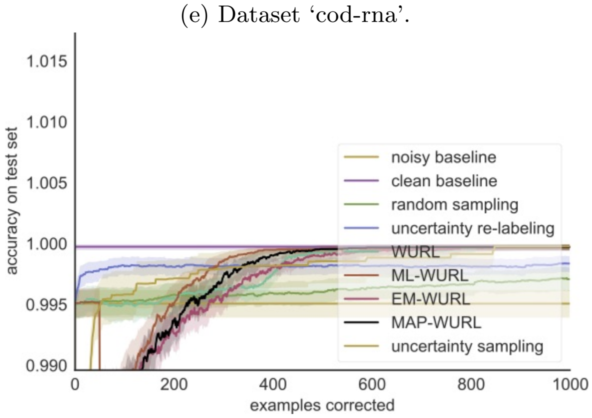
(d) Dataset 'w1a'.



(e) Dataset 'cod-rna'.

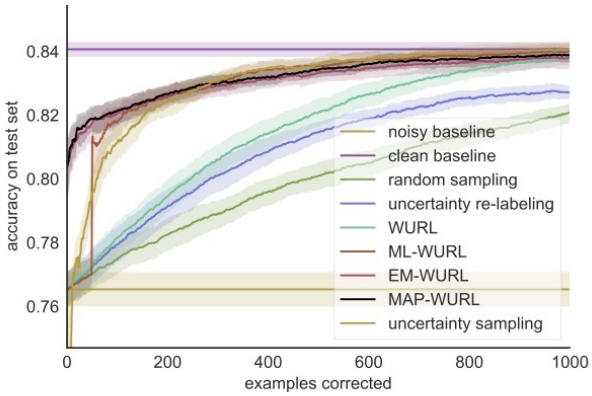


(f) Dataset 'ijcnn1'.

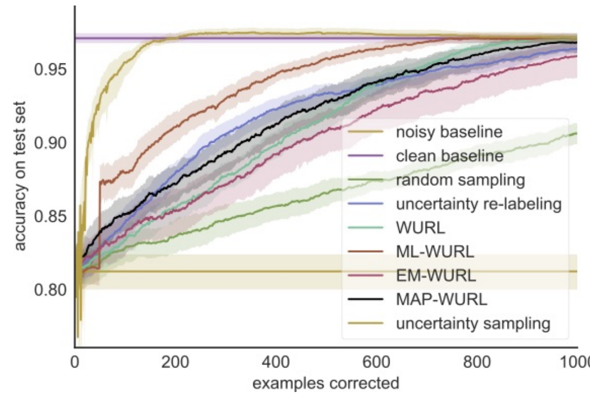


(g) Dataset 'mushrooms'.

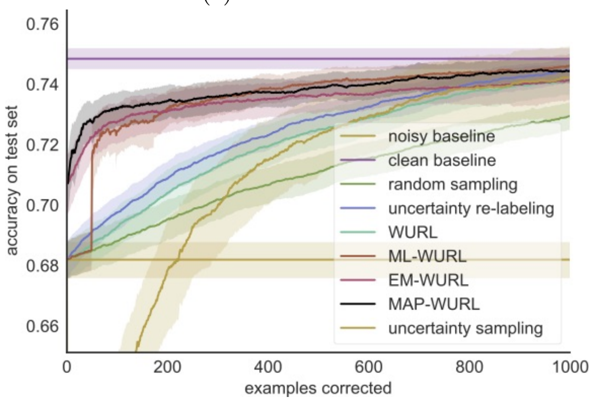
Figure 2: Setting noise rates $\rho_{-1} = 0.2$ and $\rho_{+1} = 0.1$. Empirical results on different benchmark datasets with parameters: $\rho_{-1} = 0.2$, $\rho_{+1} = 0.1$, $\lambda = 1.0$, $n_{\text{burn-in}} = 50$, $\kappa = 0.5$.



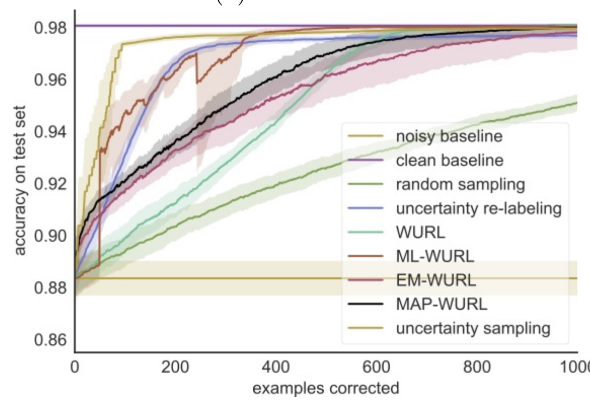
(a) Dataset ‘a1a’.



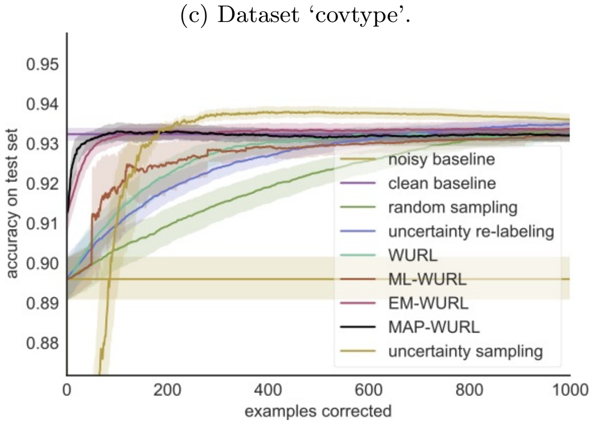
(b) Dataset ‘ad’.



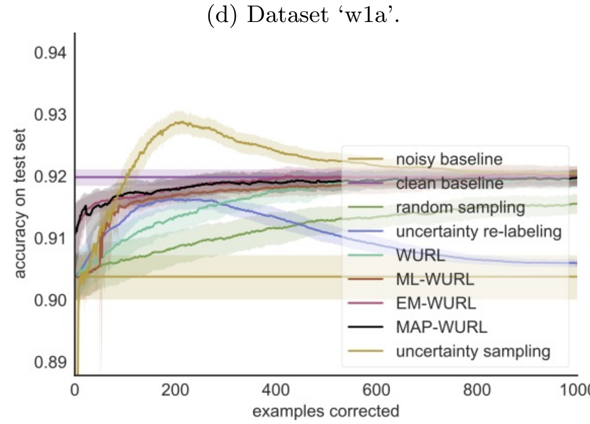
(c) Dataset ‘covtype’.



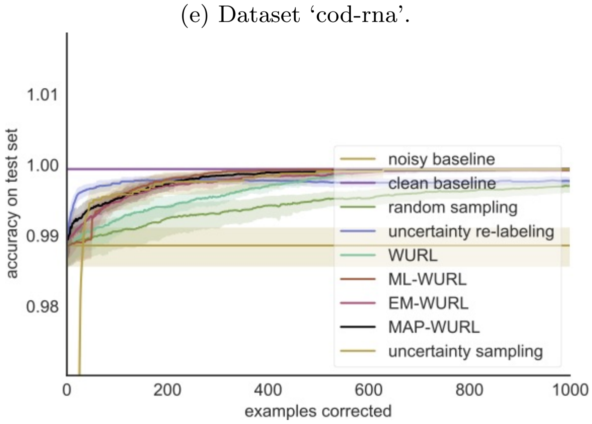
(d) Dataset ‘w1a’.



(e) Dataset ‘cod-rna’.

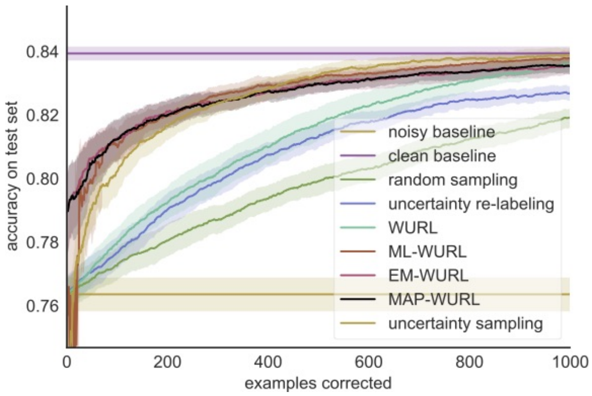


(f) Dataset ‘ijcnn1’.

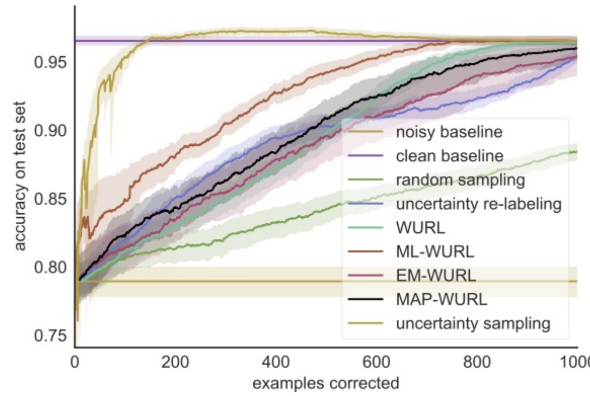


(g) Dataset ‘mushrooms’.

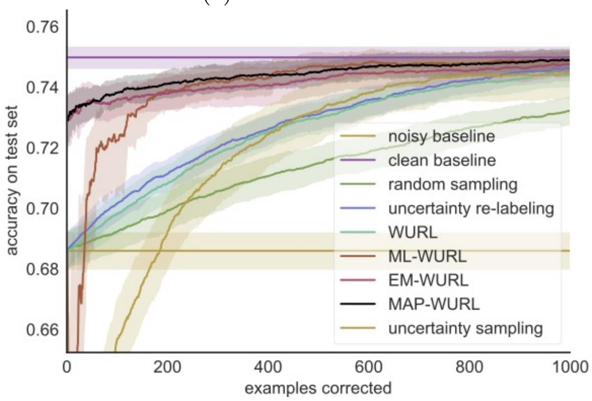
Figure 3: Setting regularization parameter $\lambda = 10.0$. Empirical results on different benchmark datasets with parameters: $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$, $\lambda = 10.0$, $n_{\text{burn-in}} = 50$, $\kappa = 0.5$.



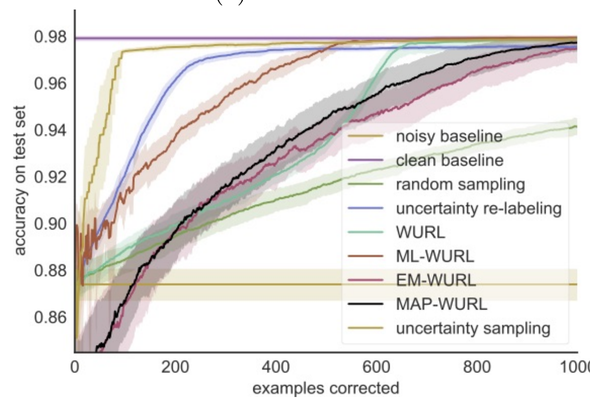
(a) Dataset ‘a1a’.



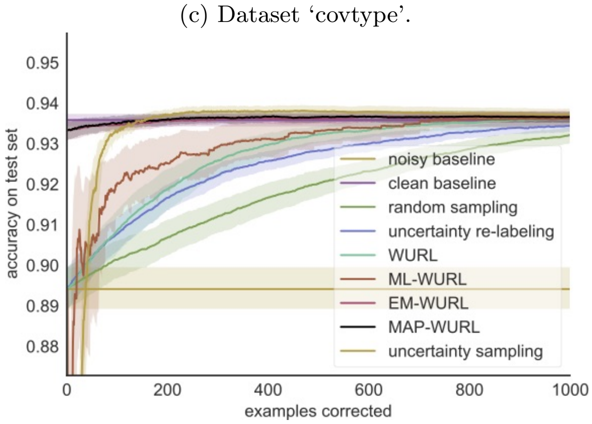
(b) Dataset ‘ad’.



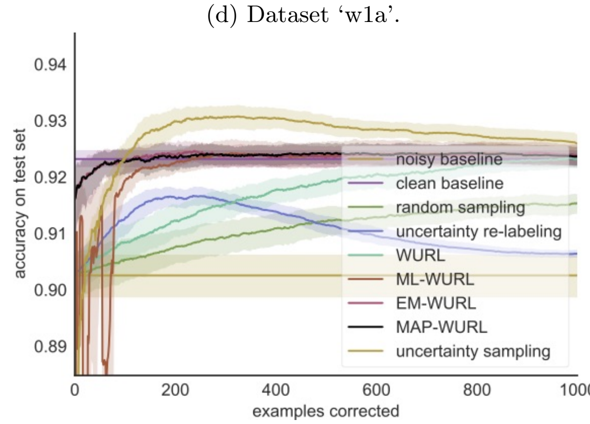
(c) Dataset ‘covtype’.



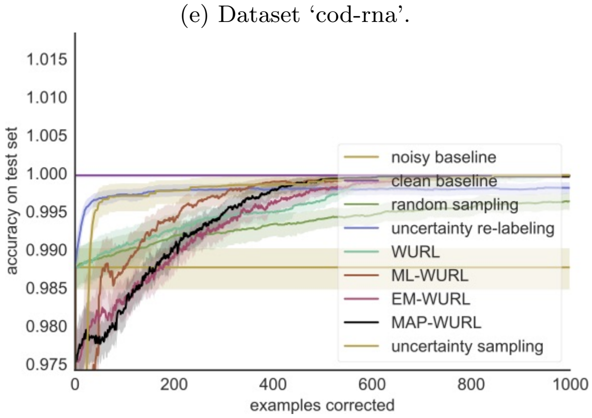
(d) Dataset ‘w1a’.



(e) Dataset ‘cod-rna’.

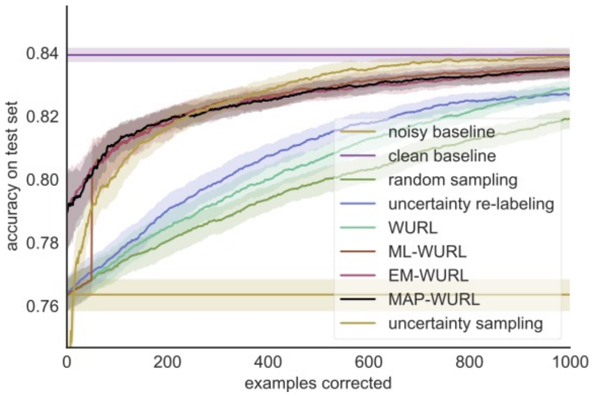


(f) Dataset ‘ijcnn1’.

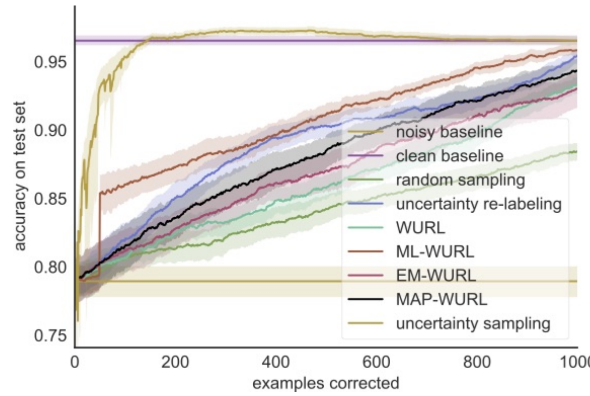


(g) Dataset ‘mushrooms’.

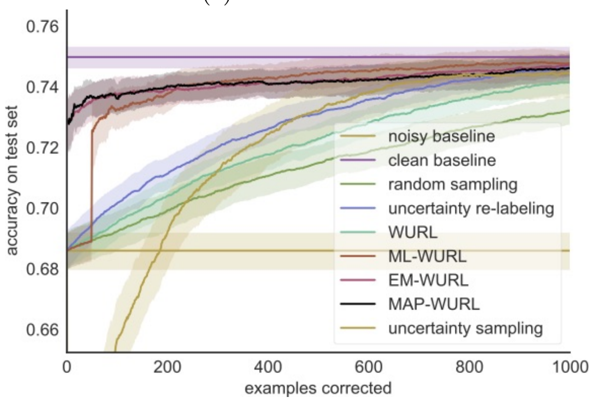
Figure 4: Setting burn-in sample size $n_{\text{burn-in}} = 0$. Empirical results on different benchmark datasets with parameters: $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$, $\lambda = 1.0$, $n_{\text{burn-in}} = 0$, $\kappa = 0.5$.



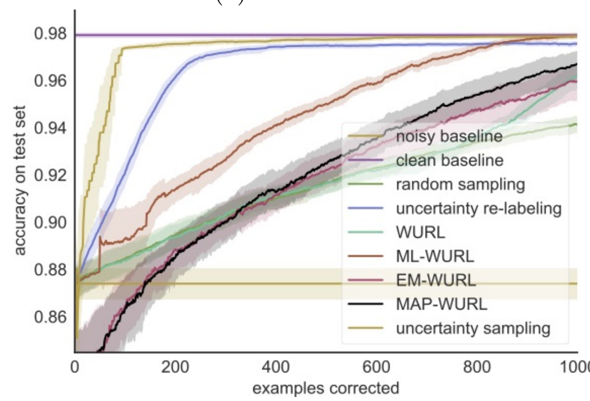
(a) Dataset ‘a1a’.



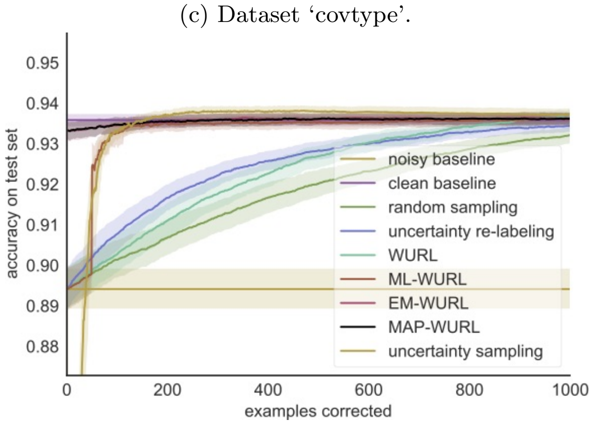
(b) Dataset ‘ad’.



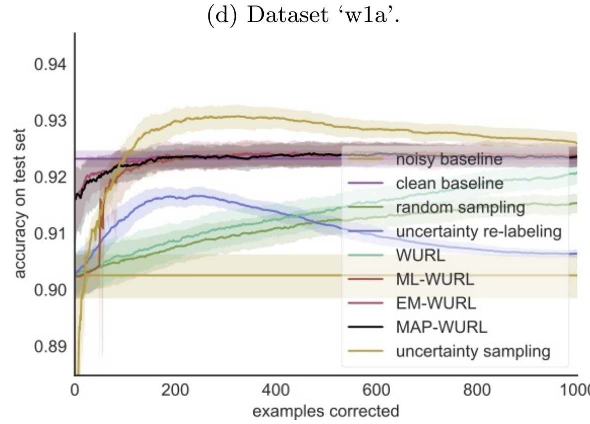
(c) Dataset ‘covtype’.



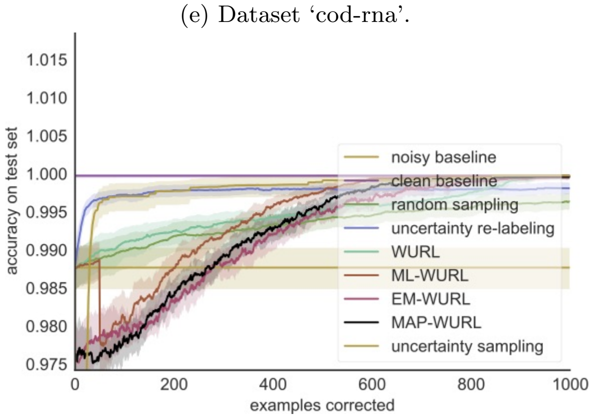
(d) Dataset ‘w1a’.



(e) Dataset ‘cod-rna’.

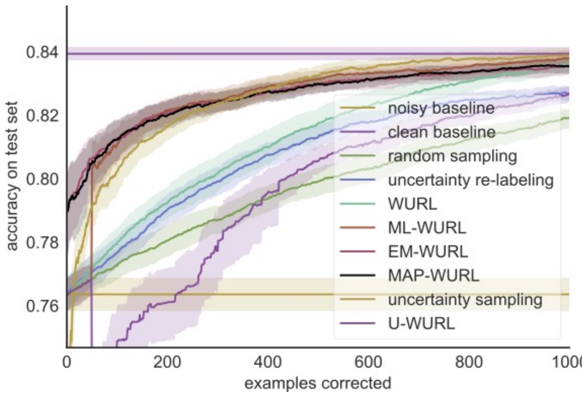


(f) Dataset ‘ijcnn1’.

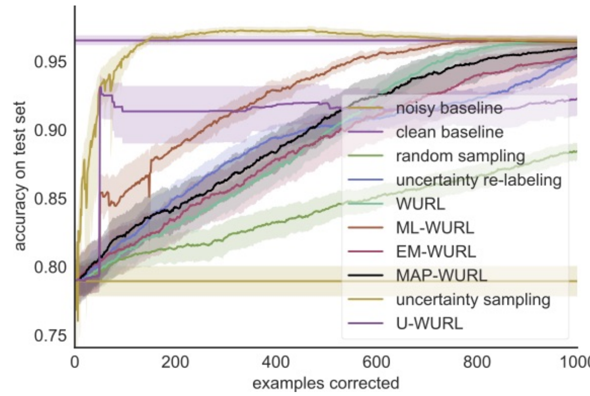


(g) Dataset ‘mushrooms’.

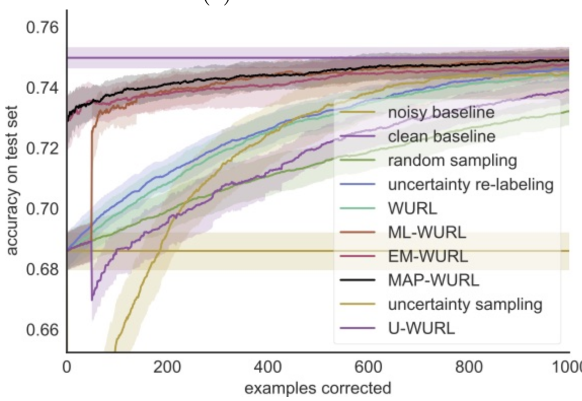
Figure 5: Setting the exploration parameter $\kappa = 0.1$. Empirical results on different benchmark datasets with parameters: $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$, $\lambda = 1.0$, $n_{\text{burn-in}} = 50$, $\kappa = 0.1$.



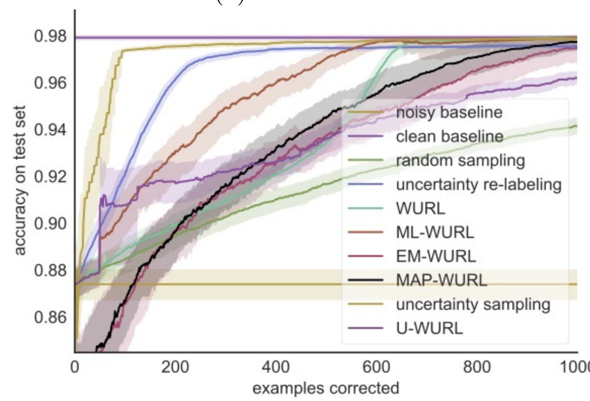
(a) Dataset 'a1a'.



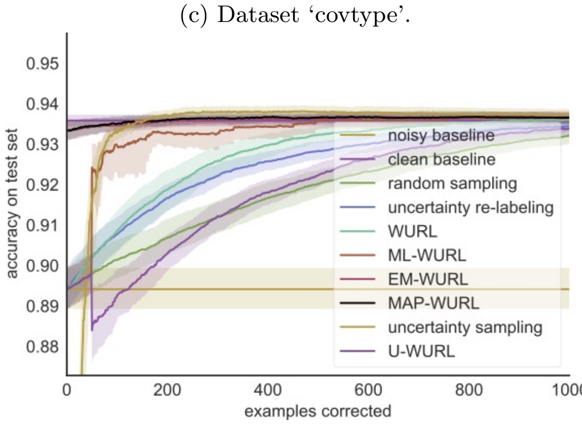
(b) Dataset 'ad'.



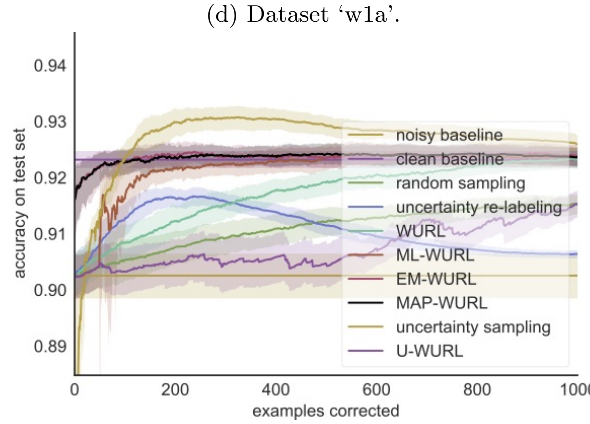
(c) Dataset 'covtype'.



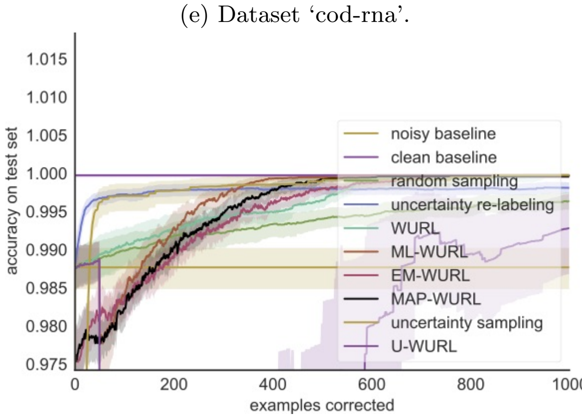
(d) Dataset 'w1a'.



(e) Dataset 'cod-rna'.



(f) Dataset 'ijcnn1'.



(g) Dataset 'mushrooms'.

Figure 6: Empirical results including the unbiased estimator with parameters: $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$, $\lambda = 1.0$, $n_{\text{burn-in}} = 50$, $\kappa = 0.5$.