Active Learning for Regression Tasks with Expected Model Output Changes – Supplementary Material

Christoph Käding[1]
christoph.kaeding@uni-jena.de
Erik Rodner[2]
erik.rodner@zeiss.com
Alexander Freytag[2]
alexander.freytag@zeiss.com
Oliver Mothes[1]
oliver.mothes@uni-jena.de
Björn Barz[1]
bjoern.barz@uni-jena.de
Joachim Denzler[1]
joachim.denzler@uni-jena.de

[1] Computer Vision Group
Friedrich Schiller University Jena
Jena, Germany

[2] Carl Zeiss AG
Jena, Germany

Abstract

The following document contains supplementary material for the paper *Active Learning of Regression Task with Expected Model Output Changes*. We provide additional information for five aspects: (i) we point towards important relationships between our EMOC criterion and other machine learning techniques (see Section S1), (ii) visualizations for the landmark prediction experiment presented in Section 4.2 of the main paper (see Section S2), (iii) illustrations of input images for our experiments shown in Section 4.4 of the main paper (see Section S3), (iv) additional experimental evaluations on more regression specific datasets (see Section S4), (v) and supporting video material for our results shown in Section S4.1 and Section 4.2 of the main paper (see Section S5).

S1 Relationships to EMOC

Our proposed EMOC criterion can be related to other machine learning techniques. In the following, we will show the connection to variance sampling and will comment on the relationship to marginalization in Gaussian processes. Furthermore, we will point towards differences in active learning for regression tasks and Bayesian optimization.

S1.1 Connection between EMOC and Variance Sampling

The main paper presents the theoretical derivation of the proposed EMOC criterion in Section 3. However, sampling of uncertain instances is one of the most popular selection

schemes in active learning (*e.g.,* [9]). Since we utilize GP models in this paper, the model uncertainty can directly be obtained as the predictive variance $\sigma_f^2(x')$. In the following, we derive the relation between the proposed EMOC criterion and uncertainty sampling in case of regression problems.

A possible approximation of the EMOC criterion is to estimate the output change of each candidate data point only on the sample itself. The approach originally proposed by [7] follows exactly this strategy. In this case, the expectation $\mathbb{E}_x$ would be ignored. Considering only x' in the calculation of the expected model output change leads to the following (compare Eq. 6 of the main paper):

$$\Delta f(x',x') \;=\; \left|\left|v'\right|\right|_P \cdot \mathbb{E}\left[||z||_P\right] \quad .$$

As noted before, the absolute Gaussian moment $\mathbb{E}\left[||z||_P\right]$ only depends on $\sigma_f(x')$ since $\tilde{\mu}(x')$ always equals 0. Therefore, ranking unlabeled samples corresponding to this moment is similar to the ranking based on $\sigma_f(x')$ directly. Furthermore, evaluating the v'-term only for the current sample x' leads to the following simplification:

$$\begin{aligned}
\left|\left|v'\right|\right|_P &= \left|\left| \frac{k'(x')^T}{\sigma_n^2+\sigma_f^2(x')} \begin{bmatrix} \left(K+\sigma_n^2 I\right)^{-1} k(x') \\ -1 \end{bmatrix} \right|\right|_P \\
&= \left|\left| \frac{k(x')^T \left(K+\sigma_n^2 I\right)^{-1} k(x') - \kappa(x',x')}{\sigma_n^2+\sigma_f^2(x')} \right|\right|_P \\
&= \left|\left| \frac{-\sigma_f^2(x')}{\sigma_n^2+\sigma_f^2(x')} \right|\right|_P \quad .
\end{aligned}$$

Since σ_n^2, $\sigma_f^2(x')$ and P are always positive values, ranking unlabeled samples relative to $\left|\left|v'\right|\right|_P$ is equivalent to ranking samples according to $\sigma_f(x')$.

Considering both facts leads to the following conclusion: for GP regression models, ranking using the predictive variance and using the proposed EMOC criterion is equivalent under the condition that the EMOC scores are calculated only on the candidate instances themselves. Since our *general* EMOC criterion is not relying on this approximation, it is able to overcome drawbacks of variance sampling by taking the empirical data distribution induced by all available samples into account.

S1.2 Comment on Marginalization in Gaussian Processes

While Gaussian processes are all about marginalization, it has to be pointed out that our EMOC type of marginalization is different from the one utilized in GPs. Whereas GP marginalizes over the likelihood of observations w.r.t. the prior function (*e.g.,* Eq. (2.28) of [14]), EMOC marginalizes the model output change (MOC) w.r.t. the continuous output y (Eq. (1) in the main paper). Nonetheless, our proposed EMOC criterion relies on some common fundamentals in integral calculation.

S1.3 Differences between Active Learning for Regression Tasks and Bayesian Optimization

Bayesian optimization [3] aims at the hyperparameter optimization of costly (unknown) objective functions with the use of surrogate functions. This optimization technique is closely

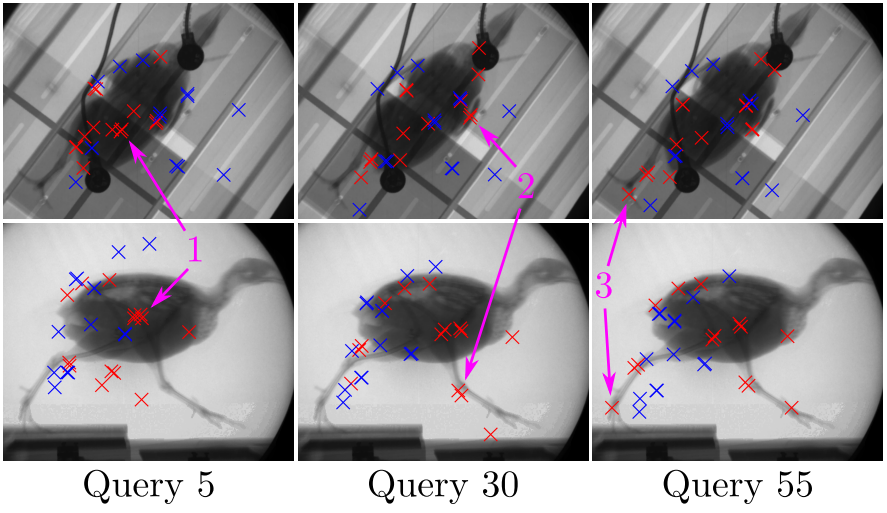

Figure S1: Qualitative results for both views of the X-ray data [] after different number of queries for EMOC (red) and random selection (blue). (1): Both regressors fail to localize landmarks after 5 queries. (2): EMOC enables the regressor to track several landmarks after only 30 queries. (3): The regressor learned with EMOC is able to predict all landmarks properly. Even difficult landmarks are estimated correctly after 55 queries. Images are cropped for better visualization.

related to active learning since it tries to improve the surrogate function (*i.e.,* an as precise as possible estimation of accuracy for different hyperparameter sets) by querying performance samples of hyperparameter constellations from the objective function. It is known that a set of criteria are applicable in both areas. However, an important prerequisite for Bayesian optimization is the availability of a non-changing objective function which can be evaluated and for which the maximum has to be found. Hence, even after multiple evaluations (*e.g.,* accuracy requests for hyperparameter sets), the objective function is still unchanged. In contrast, the objective function in active learning changes after every annotation request: since new annotations lead to an updated model, the mapping from samples to error reduction for this updated model is different from the mapping before the request. Additionally, in Bayesian optimization, the fact that the label given by the annotator is unknown is not taken into account since the objective function is independent of human interaction. Finally, while the aim of Bayesian optimization is to quickly find the maximum of the objective function, the goal of active learning is to discover the whole objective function as fast as possible.

S2 Visualizations of Landmark Prediction Experiment

Some qualitative results for the experiment described in Section 4.2 in the main paper are shown in Fig. S1. It can be seen that our approach correctly predicts almost all landmark positions with only 55 samples. In contrast to this, a regressor learned with randomly drawn data achieves poor results (compare Fig. S5). This shows how important the selection of influential samples is and how the annotation effort can be reduced. Additional video material is provided (see Section S5).

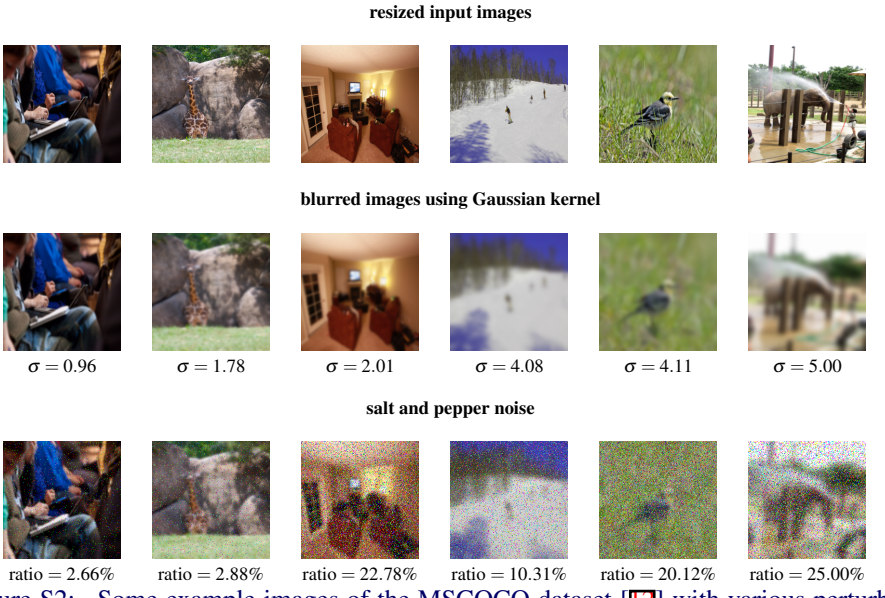

Figure S2: Some example images of the MSCOCO dataset [12] with various perturbance strengths.

S3 Visualizations of Image Quality Assessment Experiment

Section 4.4 of the main paper presents experiments for image quality assessment using the 2014 MSCOCO v1.0 dataset [12]. We show in Fig. S2 some example images with randomly applied perturbances. First, the image is resized (top row). Afterwards, the resized image is blurred using a Gaussian kernel (middle row) and perturbed with salt-and-pepper noise (bottom row). The resulting image is then fed to the CNN to generate features for our regression task.

S4 Additional Experiments

Besides the experimental evaluation presented in Section 4 of the main paper which focuses on computer vision applications, here we provide more general regression tasks. Additionally, due to the lack of space, we only showed the learning curve of a single evaluation scenario in Section 4 of the main paper. To compensate for this short-coming, we show a complete overview of all results in Table S1 while the remaining learning curves are gathered in Fig. S5. Explanations of the observable differences in accuracy can be found in the corresponding sections in the main paper as well as in Sections S4.1 and S4.2. A summary concerning all results is given in Section 4.5 of the main paper.

S4.1 Active Sensor Placement

Optimal sensor placement is a challenging problem in sensor networks [7]. The goal is to propose positions for sensors such that the recorded data contains most information about

	random	variance	EGAL	entropy	Di0.0/De1.0	Di0.5/De0.5	Di1.0/De0.0	mahalanobis	EMC	**EMOC**
AwA	63.78% (2)	69.69% (4)	86.15% (8)	71.51% (7)	87.27% (9)	70.93% (5)	71.23% (6)	67.77% (3)	100.00% (10)	**62.04% (1)**
ABL	59.51% (5)	49.28% (2)	77.17% (8)	77.85% (9)	100.00% (10)	68.85% (6)	69.54% (7)	55.24% (4)	50.92% (3)	**49.24% (1)**
C-Tai	87.00% (2)	91.89% (7)	95.67% (9)	100.00% (10)	94.10% (8)	90.47% (4)	90.58% (5)	91.19% (6)	88.64% (3)	**85.03% (1)**
yearbook	25.49% (3)	33.71% (6)	37.33% (7)	45.16% (9)	100.00% (10)	37.64% (8)	27.87% (4)	23.65% (2)	28.94% (5)	**22.81% (1)**
MSCOCO quality	60.99% (2)	97.01% (8)	61.89% (3)	100.00% (9)	66.27% (6)	64.74% (5)	96.82% (7)	- (10)	63.63% (4)	**60.61% (1)**
coastDat-1	6.89% (5)	5.02% (3)	54.19% (8)	100.00% (10)	80.95% (9)	36.73% (6)	4.99% (2)	44.27% (7)	6.27% (4)	**4.96% (1)**
UCI concrete	71.47% (5)	68.34% (4)	88.81% (7)	89.83% (8)	100.00% (10)	95.03% (9)	67.88% (2)	72.84% (6)	67.97% (3)	**66.92% (1)**
UCI housing	71.64% (6)	68.11% (3)	77.67% (7)	91.30% (9)	100.00% (10)	87.29% (8)	66.64% (2)	69.24% (4)	69.33% (5)	**64.66% (1)**
UCI redwine	91.23% (6)	90.52% (4)	93.28% (7)	99.27% (9)	100.00% (10)	98.84% (8)	90.20% (2)	90.21% (3)	90.52% (5)	**90.17% (1)**
UCI yacht	76.31% (5)	60.92% (3)	86.05% (7)	95.40% (9)	100.00% (10)	86.16% (8)	59.79% (2)	77.00% (6)	62.84% (4)	**57.90% (1)**
UCI whitewine	73.77% (6)	72.12% (4)	88.44% (7)	99.07% (9)	100.00% (10)	97.68% (8)	72.00% (3)	72.52% (5)	71.66% (2)	**71.56% (1)**
average rank	4.27	4.36	7.09	8.91	9.27	6.82	3.82	5.09	4.36	**1.00**

Table S1: Area under error curve in percent relative to the worst performing method on the same dataset (lower is better). Additionally, a ranking (lower is better) of all methods according to their area under error curve per dataset is given in brackets as well as an overall ranking at the bottom.

the problem to be solved. Examples for variables to be estimated are the composition of soil in agriculture [10] or the air contamination in cities [15]. Active learning can provide solutions to this field of research by interpreting the query process of selecting unlabeled data as identification of possible sensor positions. In the following, we use the coastDat-1 hindcast dataset [8] as basis for such an optimal selection process of points in space to measure certain environmental variables. The dataset consists of various marine climate variables measured at locations across the southern North Sea at different points in time between the years 1958 to 2007. Since no measurements are available at locations over land, we select the subset of the data between 53.9° N, 0° E and 56° N, 7.7° E at 1981-11-24, 02:00 AM. This results in 3,354 position-measurement pairs.

Experimental Setup We only consider the wave height for our experiments and the positions thereof. The positions are normalized to the range $[0,1]$. Hence, the regression problem consists of two-dimensional real-valued inputs and one-dimensional real-valued outputs. We use the geographical locations of four randomly chosen wave height measurements as initialization for $\mathfrak{L}$. The active learning techniques should then propose points on the grid where additional measurements shall be acquired, *i.e.,* new sensors should be placed to improve regression of wave heights. Evaluation is repeated ten times on random splits with 1,675 samples for the test set $\mathfrak{T}$ and 1,675 unlabeled samples for $\mathfrak{U}$. The GP model uses an RBF kernel.

Evaluation As the results in Table S1 and Fig. S5 show, our EMOC criterion performs best along with a few other baseline methods (*i.e.,* variance and diversity). The results clearly indicate how important the selection of data points is for labeling. Density sampling, for example, leads to a drastic deterioration after few queries while diversity sampling achieves nearly best performance. This difference in behavior might be explained as follows: Density sampling chooses positions in dense regions and broadcasts this knowledge over the entire map. If there are drastic wave height differences in those dense regions, the model predicts this all over the considered area. In contrast, diversity sampling selects positions maximally distant from the already known sensor placements and therefore builds a uniform sensor grid. This behavior leads to slightly worse results in comparison to our approach. The same arguments hold for variance sampling. Our method is able to outperform both baselines with small margin since it it able to explore the feature space more efficiently because it is not only relying on feature distances.

Visualization Fig. S3 shows visualizations of a sample experimental run. Here, we com-

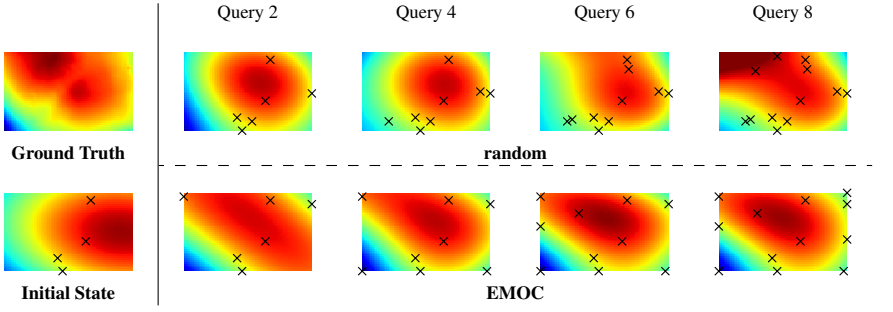

Figure S3: Qualitative results as heatmap for experiments on the coastDat-1 dataset [8]. Hotter colors encode higher waves. Crosses identify known positions.

pare the evolution of predictions for two GP regressors. One of them learns from points chosen at random and the other one is updated with positions queried according to the EMOC criterion. This visualization shows nicely that EMOC selects positions to define the expanse of the extremely high wave region in the center of the grid. In contrast to this, random selection is completely independent of the data which leads to poor results after 8 queries. This example shows that EMOC is able to increase the accuracy of regression even after little annotation effort.

S4.2 UCI Data

The UCI machine learning repository [11] is a widely known collection of different real world datasets from various problem domains and sizes. A usual task is to predict a descriptive real-valued attribute, for example the quality of wine, using real-valued features such as alcoholic strength or color intensity. Since the repository provides a great variety of different real world data, we follow [4, 13] and use the datasets `UCI concrete`, `UCI yacht`, `UCI housing`, `UCI redwine` and `UCI whitewine`.

Experimental Setup All of the used datasets consist of 8 to 13 dimensional inputs and a single real-valued output variable to predict. Each of those datasets contains between 308 and 4,898 samples. After selecting a certain number of initial samples for $\mathfrak{L}$, the remaining instances are split randomly in half to serve as unlabeled set $\mathfrak{U}$ and test set $\mathfrak{T}$. Results are averaged over multiple random initializations using an RBF kernel while we query all available samples in $\mathfrak{U}$. In detail, the settings are (reported as: feat. dim. / # inits / $||\mathfrak{L}||$ / $||\mathfrak{U}||$ and $||\mathfrak{T}||$): UCI concrete (8 / 10 / 30 / 500), UCI housing (13 / 10 / 5 / 250), UCI redwine (11 / 10 / 9 / 795), UCI yacht (6 / 10 / 7 / 150), UCI whitewine (11 / 3 / 8 / 2445). Following [4], feature dimensions for each experiment on the UCI datasets are normalized separately.

Evaluation Results for this experiment can be found in Table S1 as well as Fig. S5 and show similar behavior for all methods. Entropy and density sampling as well as the EGAL scheme do not seem to be a suitable choice for this data. The remaining active learning methods achieve good results on some and worse results on other UCI datasets. We attribute this to the small feature dimension of the data. Hence, it can be assumed that some feature dimensions are more important than others. For example, the *median value of owner-occupied homes* in the neighborhood may be more important for housing value than the *weighted distances to five Boston employment centres*. This could explain the AUC error rates of some

methods: distance based methods try to cover even unimportant dimensions, whereas the proposed EMOC criterion implicitly combines empirical density and model information and can identify important dimensions by estimating the output change since this information is encoded in the model itself.

S5 Qualitative Results Video

To support the presented results, we additionally provide two video clips with visualizations of the data selection and training process. A single video showing both clips can be found at youtu.be/2BxZkpdpZwg. See also Fig. S4 for a one-frame snapshot of both clips.

Clip 1 – Description In the first part of the video, results are shown for the anatomical landmark estimation (see Section 4.2 in the main paper). We overlay randomly selected frames from test videos with the predicted landmarks from two models. Jittering landmarks are due to retraining from increasing training set sizes over time. Since the prediction accuracy should increase with the available training data, the accuracy should be highest at the end of each mini-clip. Predictions from the baseline of passive (random) data selection are indicated in blue, whereas the model trained with active data selection with our EMOC criterion is shown in red.

Clip 1 – Observation Note that the red predictions match the anatomical landmarks of the actual bird earlier than the blue predictions and fluctuate less over time. Hence, our EMOC selection scheme is able to query informative samples right in the beginning of learning. Ultimately, this allows for training models which predict nearly correct landmark positions on important joints after few queries.

Clip 2 – Description The second part of the video visualizes the change of wave height estimation when new information becomes available (see Section S4.1). In the left column, we show the ground truth measurement of the wave and overlay the requested data points from our EMOC criterion (top) as well as random selection (bottom). Additionally, we show in the middle column the model predictions learned from currently queried data. The right column shows the absolute difference of the current model prediction and the ground truth data. Hence, overlayed crosses are identical in both clips of a row.

Clip 2 – Observation We observe that EMOC puts a strong focus on regions close to the currently estimated border of the wave. Hence, we conclude that the regression model is able to quickly adapt itself towards the correct regression after only a few queries with EMOC whereas random selection needs far more queries to achieve comparable low error rates (see most right column).

References

[1] Emanuel Andrada, Daniel Haase, Yefta Sutedja, John A. Nyakatura, Brandon M. Kilbourne, Joachim Denzler, Martin S. Fischer, and Reinhard Blickhan. Mixed gaits in small avian terrestrial locomotion. *Scientific Reports*, 2015.

[2] Paul Bodesheim, Erik Rodner, Alexander Freytag, and Joachim Denzler. Divergence-based one-class classification using gaussian processes. In *British Machine Vision Conference (BMVC)*, 2012.

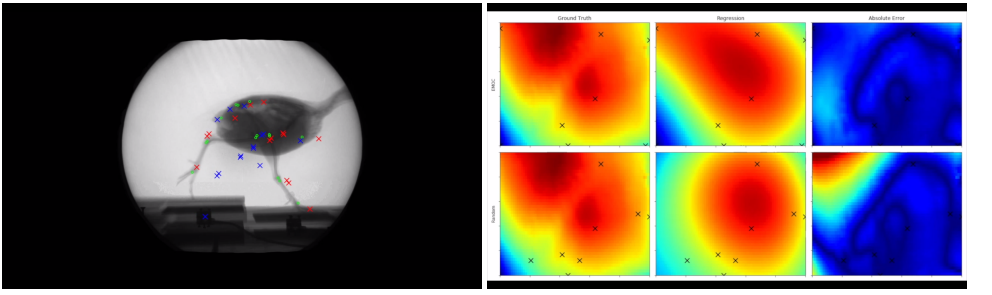

Figure S4: Snapshots taken from supporting video material.

[3] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[4] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *International Conference on Data Mining (ICDM)*, 2013.

[5] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition (GCPR)*, 2016.

[6] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *International Conference on Computer Vision Workshops (ICCV-WS)*, 2015.

[7] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *International Conference on Machine Learning (ICML)*, 2005.

[8] Zentrum für Material-und Küstenforschung GmbH Helmholtz-Zentrum Geesthacht. coastdat-1 waves north sea wave spectra hindcast (1948-2007), 2012.

[9] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 2010.

[10] Miroslav Kutílek, Donald R Nielsen, et al. *Soil hydrology: texbook for students of soil science, agriculture, forestry, geoecology, hydrology, geomorphology and other related disciplines.* Catena Verlag, 1994.

[11] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[13] Jack O'Neill, Sarah Jane Delany, and Brian MacNamee. Model-free and model-based active learning for regression. In *Advances in Computational Intelligence Systems (NIPS)*, 2017.

[14] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[15] S Zaromb and JR Stetter. Theoretical basis for identification and measurement of air contaminants using an array of sensors having partly overlapping selectivities. *Sensors and Actuators*, 1984.

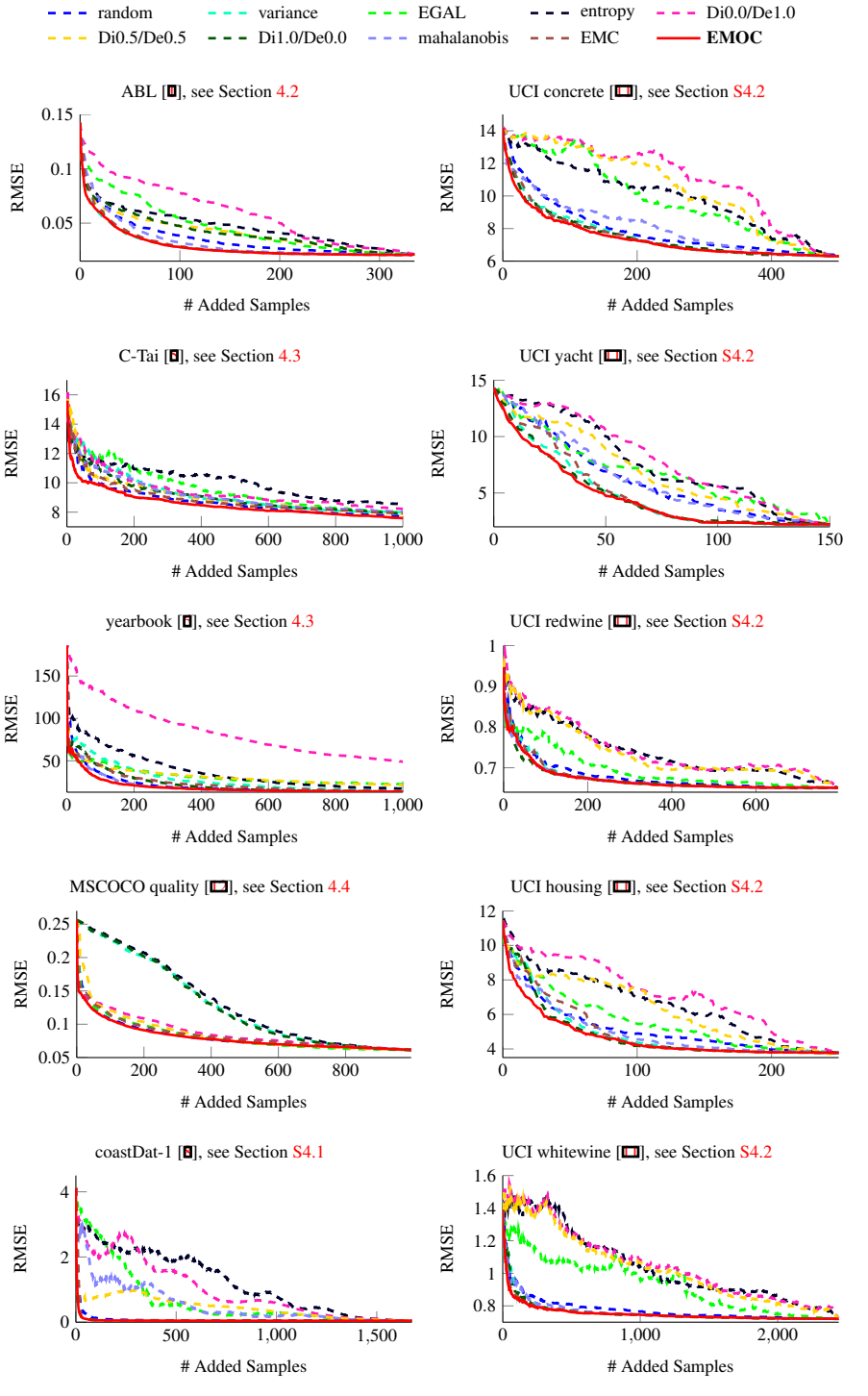

Figure S5: Experimental results as error curves on different datasets for all evaluated active learning methods (lower is better).