Large-scale Active Learning with Approximations of Expected Model Output Changes – Supplementary Material –

Christoph Käding[1,2], Alexander Freytag[1,2], Erik Rodner[1,2], Andrea Perino[3,4], and Joachim Denzler[1,2,3]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany
[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[4]Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Abstract. The following document contains supplementary derivations and empirical analyses for the paper "Large-scale Active Learning with Approximations of Expected Model Output Changes". We provide additional information for five aspects: (i) a detailed derivation of closed-form solutions required in Section 4 of the main submission (Section S1), (ii) an analysis of class distributions in MS-COCO (Section S2), (iii) further comparisons of our approximation techniques (Section S3), (iv) additional comparisons of computation times (Section S4), and (v) additional qualitative results on the biodiversity dataset (Section S5).

S1 Mathematical Derivation of LSR-EMOC

Derivation of the model update in Eq. (4) of the paper To derive efficient formulas for our LSR-EMOC approach, it is necessary to consider the actual change in the model parameters $\Delta\mathbf{w}_c$ itself which is defined as difference of the current model $\mathbf{w}_c$ and the updated model $\mathbf{w}'_c$. The model parameters are obtained by simple linear regression:

$$\mathbf{w}_c = \mathbf{C}_{\text{reg}}^{-1}\mathbf{X}\mathbf{y}_c \ , \tag{S1}$$

$$\mathbf{w}'_c = \mathbf{C}_{\text{reg}}'^{-1}\mathbf{X}'\mathbf{y}'_c \ , \tag{S2}$$

where $\mathbf{X}'$ and $\mathbf{y}'_c$ denote data matrix and binary label vector after the new example $\mathbf{x}'$ has been added with label y'_c. Without loss of generality, we can decompose the product of data matrix and label vector as follows:

$$\mathbf{X}'\mathbf{y}'_c = \mathbf{X}\mathbf{y}_c + \mathbf{x}'y'_c \ . \tag{S3}$$

This research was supported by grant DE 735/10-1 of the German Research Foundation (DFG).

Furthermore, the regularized covariance matrix[1] $\mathbf{C}_{\text{reg}}$ with respect to training data $\mathbf{X}$ is obtained by:

$$\mathbf{C}_{\text{reg}} = \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_n^2 \mathbf{I} \ . \tag{S4}$$

With some simple linear algebra, we can derive a well-known and efficient model update rule [6]:

$$\begin{aligned} \Delta\mathbf{w}_c \ &= \mathbf{w}'_c - \mathbf{w}_c && \text{(S5)} \\ &\overset{(S1),(S2)}{=} \mathbf{C}'^{-1}_{\text{reg}} \mathbf{X}' \mathbf{y}'_c - \mathbf{C}^{-1}_{\text{reg}} \mathbf{X} \mathbf{y}_c \ . && \text{(S6)} \end{aligned}$$

For the updated inverse matrix $\mathbf{C}'^{-1}_{\text{reg}}$ we can use the standard theorem of Sherman-Morrison [7] since $\mathbf{C}'_{\text{reg}}$ is a rank-one update of $\mathbf{C}_{\text{reg}}$:

$$\begin{aligned} \text{(S6)} \ &\overset{(5)}{=} \left(\mathbf{C}^{-1}_{\text{reg}} - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}}}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \right) \mathbf{X}' \mathbf{y}'_c - \mathbf{C}^{-1}_{\text{reg}} \mathbf{X} \mathbf{y}_c && \text{(S7)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \mathbf{X}' \mathbf{y}'_c - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}}}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \mathbf{X}' \mathbf{y}'_c - \mathbf{C}^{-1}_{\text{reg}} \mathbf{X} \mathbf{y}_c && \text{(S8)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \left(\mathbf{X} \mathbf{y}_c - \mathbf{X}' \mathbf{y}'_c \right) - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}}}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \mathbf{X}' \mathbf{y}'_c \ . && \text{(S9)} \end{aligned}$$

Applying the decomposition of the updated training data leads to the following simplification:

$$\begin{aligned} \text{(S9)} \ &\overset{(S3)}{=} \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \left(\mathbf{X} \mathbf{y}_c + \mathbf{x}' y'_c \right)}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} && \text{(S10)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{X} \mathbf{y}_c + \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \ . && \text{(S11)} \end{aligned}$$

Using the solution of model parameters in Eq. (S1) further shortens the equation:

$$\begin{aligned} \text{(S11)} \ &\overset{(S1)}{=} \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c - \frac{\mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \left(\mathbf{x}'^T \mathbf{w}_c + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c \right)}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} && \text{(S12)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \left(y'_c - \frac{\mathbf{x}'^T \mathbf{w}_c + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \right) && \text{(S13)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \left(\frac{y'_c \left(1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \right)}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} - \frac{\mathbf{x}'^T \mathbf{w}_c + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \right) && \text{(S14)} \\ &= \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' \left(\frac{y'_c + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c - \mathbf{x}'^T \mathbf{w}_c - \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}' y'_c}{1 + \mathbf{x}'^T \mathbf{C}^{-1}_{\text{reg}} \mathbf{x}'} \right) \ . && \text{(S15)} \end{aligned}$$

[1] Note that $\mathbf{C}_{\text{reg}}$ is actually a regularized matrix of the empirical non-central second order moments. Nonetheless we use the term regularized covariance matrix in the remainder.

Simplifying the given equation leads to the final model update:

$$\text{(S15)} \; = \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}' \left(\frac{y'_c - \mathbf{x}'^T \mathbf{w}_c}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \right) \quad . \tag{S16}$$

Efficient model output changes (Eq. (3) of the paper) The obtained efficient model update rule can now be used to derive our LSR-EMOC criterion for a given class y'. We start with the definition of the output change for linear models:

$$\begin{aligned} \Delta f_{mc}(\mathbf{x}', y') &= \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \mathcal{L}\left(f_{mc}(\mathbf{x}_j), f'_{mc}(\mathbf{x}_j)\right) && \text{(S17)} \\ &= \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \frac{1}{|C|} \sum_{c \in C} |f'_c(\mathbf{x}_j) - f_c(\mathbf{x}_j)| && \text{(S18)} \\ &= \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{w}_c'^T \mathbf{x}_j - \mathbf{w}_c^T \mathbf{x}_j \right| \quad . && \text{(S19)} \end{aligned}$$

The closed-form solution for $\Delta \mathbf{w}_c$ avoids the explicit computation of the updated model:

$$\text{(S19)} \overset{\text{(S5),(S16)}}{=} \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{x}_j^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}' \left(\frac{y'_c - \mathbf{x}'^T \mathbf{w}_c}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \right) \right| \quad . \tag{S20}$$

Some simplifications lead to the final output change for a given class y':

$$\begin{aligned} \text{(S20)} &= \frac{1}{|C|} \sum_{c \in C} \left| \frac{y'_c - \mathbf{x}'^T \mathbf{w}_c}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \right| \cdot \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \left| \mathbf{x}_j^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}' \right| && \text{(S21)} \\ &= \frac{1}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \cdot \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{w}_c^T \mathbf{x}' - y'_c \right| \\ &\quad \cdot \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \left| \mathbf{x}_j^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}' \right| \quad . && \text{(S22)} \end{aligned}$$

This formula allows for evaluating the EMOC criterion directly without explicitly estimating model parameters for the new example $(\mathbf{x}', y')$.

By using a proper class probability, we can obtain our proposed LSR-EMOC approach. Therefore, we have to consider every class in the label space $\mathcal{Y}$ as possible update and its corresponding model output change:

$$\Delta f_{mc}(\mathbf{x}') = \sum_{y' \in \mathcal{Y}} \left(p(y'|\mathbf{x}') \, \Delta f_{mc}(\mathbf{x}', y') \right) \quad . \tag{S23}$$

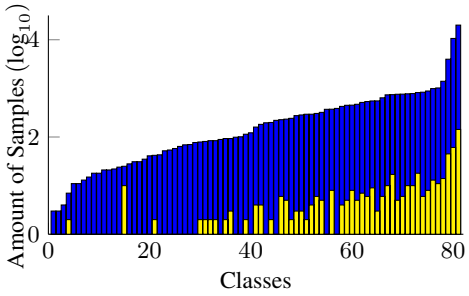

Fig. S1: Distribution of class frequencies in the full MS-COCO dataset including noise data (*blue*). Number of examples are shown in logarithmic scale. In addition, we show the distribution over examples drawn by our method (*yellow*).

Applying the derivation of the output change and simplifying the obtained equation leads to the proposed model output change criterion:

$$\text{(S23)} \overset{(S22)}{=} \sum_{y' \in \mathcal{Y}} \left(p\left(y'|\mathbf{x}'\right) \cdot \frac{1}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}'} \cdot \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{w}_c^T \mathbf{x}' - y'_c \right| \right.$$
$$\left. \cdot \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \left| \mathbf{x}_j^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}' \right| \right) \tag{S24}$$

$$= \frac{1}{1 + \mathbf{x}'^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}'} \cdot \sum_{y' \in \mathcal{Y}} \left(p\left(y'|\mathbf{x}'\right) \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{w}_c^T \mathbf{x}' - y'_c \right| \right)$$
$$\cdot \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \left| \mathbf{x}_j^T \mathbf{C}_{\text{reg}}^{-1} \mathbf{x}' \right| \quad . \tag{S25}$$

S2 Distribution of MS-COCO Dataset

As stated in the main paper, the MS-COCO dataset [5] is an imbalanced dataset with an heavy tailed class distribution (see blue bar plot in Fig. S1). Fig. S1 shows a sample class distribution after 500 queries with LSR-EMOC$^{\text{r-500}}$. It can be seen that our method selects examples of well presented classes. We further observe that the majority of the small classes are ignored since our selection criterion does not explicitly focus on rare class discovery.

S3 Comparison of Approximations of LSR-EMOC

Due to lack of space in the main submission, we presented only the comparison of our approximations for one of the chosen subsets of MS-COCO [5]. Here, we present further comparisons for the scenario with all data from MS-COCO (Fig. S2) as well as the class-balanced subset (Fig. S3) as used in Fig. 5 of the main paper. Note that we

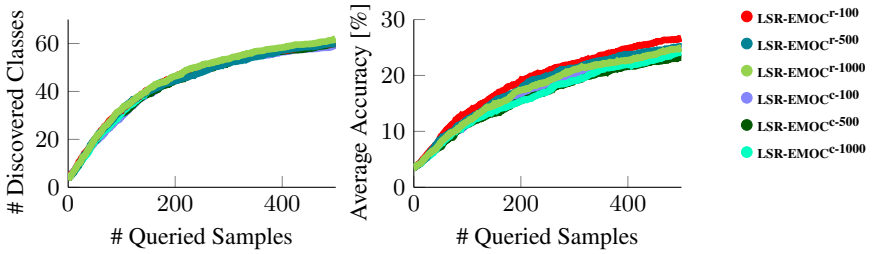

Fig. S2: Comparing approximations of LSR-EMOC on MS-COCO.

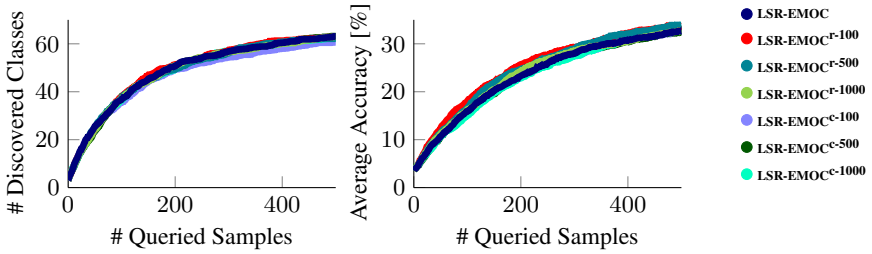

Fig. S3: Comparing approximations of LSR-EMOC on MS-COCO with balanced class distribution.

do not contain an evaluation for LSR-EMOC without approximations on the whole MS-COCO dataset since the required computation time was unaffordably high. As shown on smaller datasets, our presented approximations indeed reach comparable results compared to the original method. Further evidence for the computational efficiency of our proposed approximation techniques arises from the gained speedup of approximately 5.3 on the class-balanced subset of MS-COCO (LSR-EMOC $\approx$ 128.7s vs. LSR-EMOC$^{\text{r-100}}$ $\approx$ 24.4s for a whole query selection). A more detailed analysis regarding overall query times is presented in the next section.

S4 Runtime Comparisons

Besides the analyses presented in the main paper, we were further interested in a detailed runtime comparison of our evaluated methods. To this end, we follow the experiment in Section 5.1 on MS-COCO dataset (Fig. 4) and analyze the required computation times. We simulate a real world active learning experiment by assuming that a human annotator needs 10s to label an instance (this follows the setup presented in the supplementary material of [3]). In consequence, we can investigate the classification accuracy as a function of total time spent (including query selection and labeling). Results are shown in Fig. S4

As can be seen, our proposed method needs longer than the competitor methods to process all queries. This is indeed noticeable, since results are shown for a linear

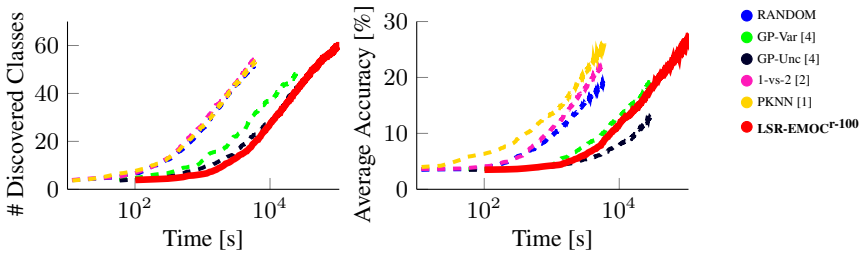

Fig. S4: Comparing runtimes considering 10s for labeling on MS-COCO.

model and the fasted of our approximation techniques. Hence, further speed-ups are hardly expectable. However, this drawback becomes less important if the annotation time increases. Besides this aspect, we have already seen in previous evaluations that all competitor methods are not able to reach the accuracy level of our method. Considering this, querying samples might be faster for other methods, but more labels are required to achieve comparable performances. In conclusion, which method to chose depends on whether the annotation cost or the selection time is the limiting factor in an application.

S5 Additional Qualitative Results of the Biodiversity Dataset

The main paper already contains qualitative results regarding queried images from the biodiversity dataset (see Section 5.2). In Fig. S5, we extend this qualitative analysis and present more selected examples for several methods. Shown is every 50$^{\text{th}}$ queried example of a single run for PKNN [1], random selection, and our proposed LSR-EMOC$^{\text{r-100}}$ algorithm. From the visual inspection of selected examples, we conclude that a mere random selection can not prevent from querying noisy data (*e.g.,* query 151 and query 351). Although PKNN does better in this aspect, it often selects redundant data (*e.g.,* query 201 and query 451 as well as query 101 and query 151). We hence conclude that it fails in learning a proper model for the biodiversity data. This incapability of learning a model for this data could explain the dropping performance in the presented experiment (see Section 5.2 of the main paper). In contrast to this, our proposed LSR-EMOC$^{\text{r-100}}$ algorithm even tries to refine decisions for some difficult cases (*e.g.,* query 251 versus query 451 – note the deer in the background).

References

1. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 762 –769 (2009)
2. Joshi, A., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2372 –2379 (2009)
3. Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4343–4352 (2015)

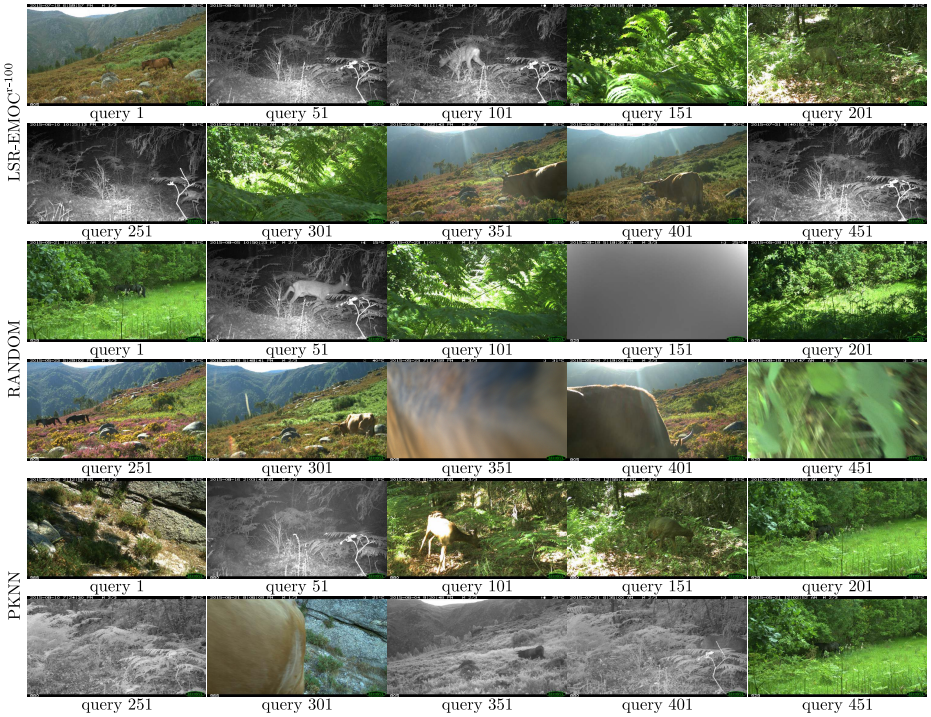

Fig. S5: Example queries from the biodiversity dataset.

4. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision (IJCV) 88, 169–188 (2010)
5. Lin, T., Maire, M., Belongie, S., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
6. Plackett, R.L.: Some theorems in least squares. Biometrika 37(1/2), pp. 149–157 (1950)
7. Press, W.H.: Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press (2007)