-Supplementary Material-
Labeling examples that matter: Relevance-Based Active Learning with Gaussian Processes

Alexander Freytag[1], Erik Rodner[1,2], Paul Bodesheim[1], and Joachim Denzler[1]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
http://www.inf-cv.uni-jena.de
[2]UC Berkeley ICSI & EECS, United States

Abstract. The following document gives additional details for the paper *Labeling examples that matter: Relevance-Based Active Learning with Gaussian Processes*. In particular, a detailed derivation of the weight vector $\bar{\boldsymbol{\alpha}}$ for Gaussian Process Regression after adding a new example is presented, learning curves of the conducted experiments are visualized, and a qualitative evaluation is given. The information given in this document is not necessary to understand the main paper.

S1 Update of Weight Vector $\boldsymbol{\alpha}$ for GP Regression

Theorem 1 (Closed form update of GP regression weights $\boldsymbol{\alpha}$). *Let y_* be the class label of a new example $\boldsymbol{x}^*$. Let further $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$ be the kernel matrix of all training examples seen so far, $\boldsymbol{k}_* = \kappa(\mathbf{X}, \boldsymbol{x}^*)$ is the vector of kernel values between the new example and all training examples, and $k_{**} = \kappa(\boldsymbol{x}^*, \boldsymbol{x}^*)$ is the self-similarity value of the new sample. In addition, we denote with $\boldsymbol{y}$ the vector of labels for all training samples, and σ_{n}^2 indicates the noise parameter for model regularization. Finally, let $\boldsymbol{\alpha} = \mathbf{K}^{-1}\boldsymbol{y}$ be the weight vector for GP regression. Then we can compute the weight vector $\bar{\boldsymbol{\alpha}}$ after adding $\boldsymbol{x}^*$ to the training set as follows:*

$$\bar{\boldsymbol{\alpha}} = \bar{\mathbf{K}}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \underbrace{\frac{1}{\sigma_{f_*}^2 + \sigma_{\mathsf{n}}^2}}_{\langle 1 \rangle} \underbrace{\begin{bmatrix} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{k}_* \\ -1 \end{bmatrix}}_{\langle 2 \rangle} \underbrace{\left(\boldsymbol{k}_*^T \boldsymbol{\alpha} - y_*\right)}_{\langle 3 \rangle} \quad \text{(T1)}$$

$$= \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \frac{\mu_* - y_*}{\sigma_{f_*}^2 + \sigma_{\mathsf{n}}^2} \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \\ -1 \end{pmatrix} .$$

The three factors can be interpreted as described in the main paper.

Proof.

$$
\begin{aligned}
\bar{\boldsymbol{\alpha}} &= \bar{\mathbf{K}}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{K} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & k_{**} + \sigma_\mathsf{n}^2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \\
\text{[2, Eq. (2.8.18),p. 117]} &= \begin{bmatrix} \left(\mathbf{K} - \frac{1}{k_{**}+\sigma_\mathsf{n}^2}\boldsymbol{k}_*\boldsymbol{k}_*^T\right)^{-1} & -\left(\mathbf{K} - \frac{1}{k_{**}+\sigma_\mathsf{n}^2}\boldsymbol{k}_*\boldsymbol{k}_*^T\right)^{-1}\frac{1}{k_{**}+\sigma_\mathsf{n}^2}\boldsymbol{k}_* \\ -\left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1}\boldsymbol{k}_*^T\mathbf{K}^{-1} & \left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \\
\text{[2, Eq. (2.8.19),p. 117]} &= \begin{bmatrix} \mathbf{K}^{-1} + \mathbf{K}^{-1}\boldsymbol{k}_*\left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1}\boldsymbol{k}_*^T\mathbf{K}^{-1} & -\left(\mathbf{K}^{-1} + \mathbf{K}^{-1}\boldsymbol{k}_*\left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1}\boldsymbol{k}_*^T\mathbf{K}^{-1}\right)\frac{1}{k_{**}+\sigma_\mathsf{n}^2}\boldsymbol{k}_* \\ -\left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1}\boldsymbol{k}_*^T\mathbf{K}^{-1} & \left(k_{**} + \sigma_\mathsf{n}^2 - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \\
\text{Definition of } \boldsymbol{\alpha} \text{ and GP variance} &= \begin{pmatrix} \boldsymbol{\alpha} + \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \left(\mathbf{K}^{-1} + \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\right)\frac{y_*}{k_{**}+\sigma_\mathsf{n}^2}\boldsymbol{k}_* \\ -\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} + \frac{y_*}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2} \end{pmatrix} \\
\text{Factor out } \boldsymbol{\alpha} &= \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_\mathsf{n}^2}\left(\mathbf{K}^{-1} + \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\right)\boldsymbol{k}_* \\ -\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} + \frac{y_*}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2} \end{pmatrix} \\
\text{Simplify denominator} &= \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_\mathsf{n}^2}\left(\mathbf{K}^{-1} + \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\right)\boldsymbol{k}_* \\ \frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
\text{Expand enumerator with } \boldsymbol{k}_* &= \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_\mathsf{n}^2}\left(\mathbf{K}^{-1}\boldsymbol{k}_* + \mathbf{K}^{-1}\boldsymbol{k}_*\frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_\mathsf{n}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&= (*)
\end{aligned}
$$

$$
\begin{aligned}
(*) &= \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_{\mathsf{n}}^2}\left(\mathbf{K}^{-1}\boldsymbol{k}_* + \mathbf{K}^{-1}\boldsymbol{k}_* \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Factor out } \mathbf{K}^{-1}\boldsymbol{k}_* \textit{ in enumerator}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \left(\frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_{\mathsf{n}}^2}\left(1 + \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*\right)\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Definition of GP variance}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \left(\frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_{\mathsf{n}}^2}\left(1 + \frac{k_{**}-\sigma_{f_*}^2}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\right)\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Expand 1 in enumerator}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \left(\frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_{\mathsf{n}}^2}\left(\frac{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2} + \frac{k_{**}-\sigma_{f_*}^2}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\right)\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Merge denominators}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \left(\frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{k_{**}+\sigma_{\mathsf{n}}^2}\left(\frac{k_{**}+\sigma_{\mathsf{n}}^2}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\right)\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Simplify enumerator}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \left(\frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\boldsymbol{k}_*^T\boldsymbol{\alpha} - \frac{y_*}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\right) \\ \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Factor out } \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \frac{1}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_*\left(\boldsymbol{k}_*^T\boldsymbol{\alpha} - y_*\right) \\ \left(y_* - \boldsymbol{k}_*^T\boldsymbol{\alpha}\right) \end{pmatrix} \\
&\boxed{\textit{Factor out } \boldsymbol{k}_*^T\boldsymbol{\alpha} - y_*} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \frac{\boldsymbol{k}_*^T\boldsymbol{\alpha} - y_*}{\sigma_{f_*}^2+\sigma_{\mathsf{n}}^2}\begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \\ -1 \end{pmatrix} \\
&\boxed{\textit{Definition of GP mean and variance}} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} + \frac{\mu_* - y_*}{\sigma_*^2}\begin{pmatrix} \mathbf{K}^{-1}\boldsymbol{k}_* \\ -1 \end{pmatrix} \qquad \square
\end{aligned}
$$

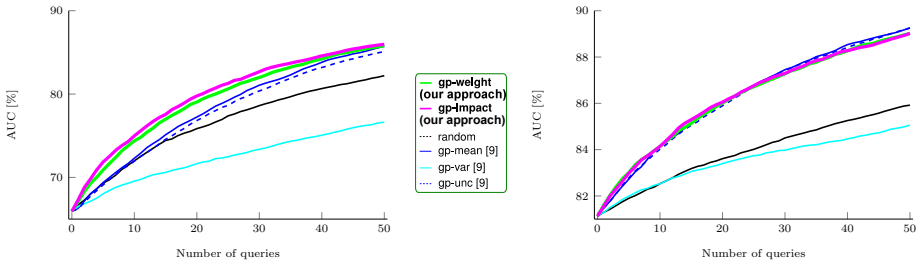

Fig. 1. Active Learning on **Caltech-256**. Initial number of samples per class is 1 and 5, respectively.

S2 Recognition Rates of Experimental Results

In the main paper, we presented active learning gains over random sampling for two popular image categorization datasets. Due to the lack of space, we moved the visualization of learning curves to this supplementary material document.

S2.1 Results on Caltech-256

Active learning curves for Caltech-256 are given in Fig. 1. The maximum accuracy when using all unlabeled data as additional training images is 93.81% AUC. As in the main document, results were obtained by averaging over the corresponding 1,000 binary settings. Hyperparameters were optimized by maximizing the marginal likelihood.

S2.2 Results on ImageNet

Active learning curves for ImageNet (2010 challenge) are given in Fig. 2. The maximum accuracy when using all unlabeled data as additional training images is 89.2% AUC. As in the main document, results were obtained by averaging over the corresponding 1,000 binary settings. Hyperparameters were optimized by maximizing the marginal likelihood.

S3 Queried Images of a Single Run on ImageNet

For a qualitative evaluation, we presented queried images of a single active learning setup in the main document. Due to the lack of space, we only included results of $\mathcal{Q}_{\text{impact}}$ and $\mathcal{Q}_{\sigma_*^2}$, and moved the remaining results to this document. The queried images of all strategies are given in Fig. 3.

References

2. Bernstein, D.S.: Matrix Mathematics. Princeton University Press, 2nd edn. (2009)
9. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. IJCV 88, 169–188 (2010)

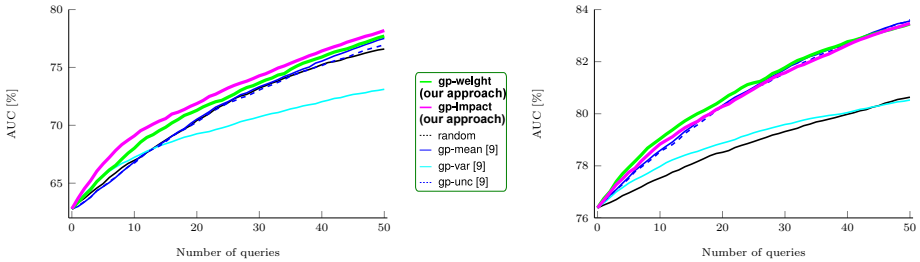

Fig. 2. Active Learning on **ImageNet**. Initial number of samples per class is 1 and 5, respectively.

Fig. 3. Queried images of GP-based active learning methods: (1) random queries, (2) $\mathcal{Q}_{\mu_*}$, (3) $\mathcal{Q}_{\sigma_*^2}$, and (4) $\mathcal{Q}_{\text{unc}}$ introduced by [9] as well as our strategies (5) $\mathcal{Q}_{\text{weight}}$ and (6) $\mathcal{Q}_{\text{impact}}$ (from top to bottom). Green and blue borders indicate images of positive and negative classes, respectively.