-Supplementary Material-
Large-Scale Gaussian Process Classification with Flexible Adaptive Histogram Kernels

Erik Rodner, Alexander Freytag, Paul Bodesheim, and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. The following document explains some details of the paper *Large-Scale Gaussian Process Classification with Flexible Adaptive Histogram Kernels* and gives additional background information for the last experiment performed. The information given in this document is not necessary to understand the paper.

S1 Details on Learning with Imbalanced Datasets

As shown by [1, p. 144], learning binary tasks with Gaussian process regression and classification can be related to the following optimization problem:

$$\underset{\boldsymbol{f}\in\mathbb{R}^n}{\text{minimize}} \quad -\sum_{i=1}^{n} \log p(y_i \mid f_i) + \frac{1}{2}\boldsymbol{f}^T\mathbf{K}^{-1}\boldsymbol{f} \quad . \tag{S1}$$

The vector $\boldsymbol{f} \in \mathbb{R}^n$ contains all values of the latent function f on the training data, *i.e.*, $\boldsymbol{f} = \mathbf{K}\boldsymbol{\alpha}$. For Gaussian process regression with a Gaussian noise model, as used in our paper (Eq. (1)), the optimization problem turns into:

$$\underset{\boldsymbol{f}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y}-\boldsymbol{f}\right|\right|^2 + \frac{1}{2}\boldsymbol{f}^T\mathbf{K}^{-1}\boldsymbol{f} \quad . \tag{S2}$$

The objective function can now be split into the quadratic error term and the regularization term. The noise variance controls the trade-off between those terms similar to the standard C parameter of SVM classifiers. Let us have a closer look on the error term in (S2):

$$\frac{1}{2\sigma^2}\left|\left|\boldsymbol{y}-\boldsymbol{f}\right|\right|^2 = \sum_{i=1}^{n}\left(\frac{1}{2\sigma^2}\right)(y_i - f_i)^2 \quad . \tag{S3}$$

As can be seen, each term is weighted equally with $w = \left(2\sigma^2\right)^{-1}$. For imbalanced training data with a large set of negatives but only a few positive examples, the optimization is biased towards the negative ones. This is also a common problem for SVM learning and the solution is to choose two different regularization parameters for each of the classes [2]. We choose the noise levels σ^2_{pos} and σ^2_{neg} for the n_{pos} positive and n_{neg} negative examples, respectively. If we require the sum of weights of the two classes to be equal:

$$\frac{1}{2\sigma^2_{\text{pos}}} n_{\text{pos}} = \frac{1}{2\sigma^2_{\text{neg}}} n_{\text{neg}} \tag{S4}$$

and to sum up to the original weight sum of the optimization problem (S2):

$$\frac{1}{2\sigma^2_{\text{pos}}} n_{\text{pos}} + \frac{1}{2\sigma^2_{\text{neg}}} n_{\text{neg}} = \frac{1}{2\sigma^2} n \enspace , \tag{S5}$$

we directly arrive at $\sigma^2_{\text{neg}} = 2\sigma^2\left(\frac{n_{\text{neg}}}{n}\right)$ and $\sigma^2_{\text{pos}} = 2\sigma^2\left(\frac{n_{\text{pos}}}{n}\right)$. This is similar in spirit to the adaptations of [3] for least-squares support vector machines to handle imbalanced datasets.

S2 Feature Relevance Experiment

For the synthetic experiments in Sect. 6.5, we used the same distributions as already done in [4]. The specific distributions with 500 samples per dimension and class are displayed in Fig. 1.

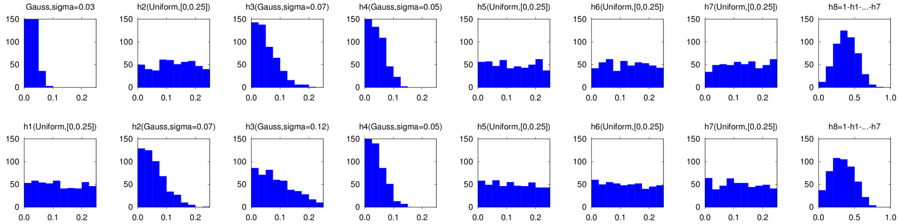

Fig. 1. Random distributions used in the synthetic feature relevance experiments (see Sect. 6.5). *Top row*: class 1, *bottom row*: class 2

S3 Details on the Log-Determinant Upper Bound

In our paper, we used the upper bound of the log-determinant of a positive definite matrix $\mathbf{D}$ as derived by Bai and Golub [5]:

$$\log\det(\mathbf{D}) \leq \left[\log\beta, \log\bar{t}\right] \begin{bmatrix} \beta & \bar{t} \\ \beta^2 & \bar{t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \doteq \mathrm{ub}(\beta, \mu_1, \mu_2) \tag{S6}$$

$$\text{with} \quad \bar{t} = \frac{\beta\mu_1 - \mu_2}{\beta n - \mu_1} \enspace . \tag{S7}$$

The bound itself depends on the maximum eigenvalue β, the trace μ_1, and the squared Frobenius norm μ_2 of the matrix $\mathbf{D}$. The only term that is not efficiently computable for HIK matrices is the Frobenius norm. Therefore, we propose in the paper to use a lower bound based on the sum of the M largest eigenvalues. In the following, we will prove that we obtain a valid upper bound of the log-determinant with the bound of [5] even when using a lower bound of the Frobenius norm. Our proofs are completely algebraic and do not require knowledge of the Gaussian quadrature techniques used in [5]. First of all, we show the validity of the modified upper bound for $\beta = 1$.

Lemma 1 (Monotonicity for $\beta = 1$). *Let $\tilde{\mu_2}$ with $0 < \tilde{\mu_2} \leq \mu_2$ be a lower bound of the squared Frobenius norm of a regular positive definite matrix $\mathbf{D}$, e.g., $\tilde{\mu_2} = \sum_{i=1}^{M} \lambda_i^2$ with $M < n$. Then the following holds for every positive definite matrix $\mathbf{D}$ with $\mu_1 = \operatorname{trace}(\mathbf{D})$ and $\beta = \lambda_1(\mathbf{D}) = 1$:*

$$\operatorname{ub}(1, \mu_1, \tilde{\mu_2}) \geq \operatorname{ub}(1, \mu_1, \mu_2) \ . \tag{S8}$$

Proof. First note that due to the conditions of the Lemma the following holds: $1 \leq \mu_2 < \mu_1 \leq n$ and $\bar{t} > 0$. Furthermore, the bound is only valid for $\beta \neq \bar{t}$, because otherwise the 2×2 matrix within the bound would be singular. We start by deriving the coefficients for μ_1 and μ_2. The first part of Eq. (S6) can be written as:

$$\begin{aligned}
\left[\log\beta, \log\bar{t}\right] \begin{bmatrix} \beta & \bar{t} \\ \beta^2 & \bar{t}^2 \end{bmatrix}^{-1} &= \left[\log\beta, \log\bar{t}\right] \left(\frac{1}{\beta\bar{t}^2 - \bar{t}\beta^2} \begin{bmatrix} \bar{t}^2 & -\bar{t} \\ -\beta^2 & \beta \end{bmatrix} \right) \\
&= \frac{1}{\beta\bar{t}^2 - \bar{t}\beta^2} \left[\bar{t}^2 \log\beta - \beta^2 \log\bar{t} \ , \ \ \beta\log\bar{t} - \bar{t}\log\beta \right] \\
&= \frac{1}{\bar{t} - \beta} \left[\frac{\log\beta}{\beta}\bar{t} - \frac{\log\bar{t}}{\bar{t}}\beta \ , \ \ \frac{\log\bar{t}}{\bar{t}} - \frac{\log\beta}{\beta} \right] \ .
\end{aligned} \tag{S9}$$

Therefore, we get the following short form of Eq. (S6) with $\beta = 1$:

$$\begin{aligned}
\operatorname{ub}(1, \mu_1, \mu_2) &= \frac{\log\bar{t}}{\bar{t}\,(\bar{t} - 1)} \left(\mu_2 - \mu_1\right) \\
\boxed{\textit{definition of } \bar{t}} &= \log\left(\frac{\mu_1 - \mu_2}{n - \mu_1}\right) \left(\frac{\mu_1 - \mu_2}{n - \mu_1} \left(\frac{\mu_1 - \mu_2}{n - \mu_1} - 1 \right) \right)^{-1} \cdot \left(\mu_2 - \mu_1\right) \\
\boxed{\textit{simplify}} &= \log\left(\frac{\mu_1 - \mu_2}{n - \mu_1}\right) \frac{(n - \mu_1)^2 (\mu_2 - \mu_1)}{(\mu_1 - \mu_2)(\mu_1 - \mu_2 - n + \mu_1)} \\
\boxed{\textit{cancel } \mu_1 - \mu_2} &= \log\left(\frac{\mu_1 - \mu_2}{n - \mu_1}\right) \frac{(n - \mu_1)^2}{n - 2\mu_1 + \mu_2} \ .
\end{aligned} \tag{S10}$$

Let $\tilde{\mu_2}$ with $0 < \tilde{\mu_2} \leq \mu_2$ be a lower bound of the squared Frobenius norm. If we replace μ_2 with $\tilde{\mu_2}$ in Eq. (S10), we notice that the log-term increases and the denominator of the second part decreases. This directly leads us to the validity of the Lemma. $\square$

The next Lemma shows that scaling the matrix $\mathbf{D}$ with $\gamma > 0$ leads to an additive constant in the bound, which is independent of μ_1 and μ_2. This constant is equivalent to the one occurring in $\log\det(\gamma\mathbf{D}) = \log\det(\mathbf{D}) + n\log\gamma$, therefore, the quality of the bound is invariant with respect to γ. Note that the squared Frobenius norm scales with γ^2 and $\bar{t}$ with γ.

Lemma 2 (Multiplicative scaling). *For all suitable parameters β, μ_1, and μ_2 of a positive definite matrix and every positive factor $\gamma > 0$, the following holds:*

$$\mathrm{ub}(\gamma\beta, \gamma\mu_1, \gamma^2\mu_2) = \mathrm{ub}(\beta, \mu_1, \mu_2) + n \cdot \log\gamma \ . \tag{S11}$$

Proof.

$$\begin{aligned}
\mathrm{ub}(\gamma\beta, \gamma\mu_1, \gamma^2\mu_2) &= \left[\log\gamma\beta,\, \log\gamma\bar{t}\right] \cdot \left(\begin{bmatrix}\gamma & 0\\ 0 & \gamma^2\end{bmatrix} \begin{bmatrix}\beta & \bar{t}\\ \beta^2 & \bar{t}^2\end{bmatrix} \right)^{-1} \begin{bmatrix}\gamma & 0\\ 0 & \gamma^2\end{bmatrix} \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
&= \left(\left[\log\beta,\, \log\bar{t}\right] + \left[\log\gamma,\, \log\gamma\right] \right) \cdot \begin{bmatrix}\beta & \bar{t}\\ \beta^2 & \bar{t}^2\end{bmatrix}^{-1} \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
\boxed{\textit{definition of } \mathrm{ub}} &= \mathrm{ub}(\beta, \mu_1, \mu_2) + \underbrace{\left[\log\gamma,\, \log\gamma\right] \cdot \begin{bmatrix}\beta & \bar{t}\\ \beta^2 & \bar{t}^2\end{bmatrix}^{-1} \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}}_{\doteq}\\
&= \mathrm{ub}(\beta, \mu_1, \mu_2) \quad + \quad \tilde{\mathrm{ub}}_\gamma(\beta, \mu_1, \mu_2) \ .
\end{aligned}$$

Now, we show that the second term equals to $n \cdot \log\gamma$ by using the definition of $\bar{t}$ and the calculation of the weights for μ_1 and μ_2 as done in the beginning of the proof of Lemma 1:

$$\begin{aligned}
\tilde{\mathrm{ub}}_\gamma(\beta, \mu_1, \mu_2) &= (\log\gamma)\left[1,\, 1\right] \cdot \begin{bmatrix}\beta & \bar{t}\\ \beta^2 & \bar{t}^2\end{bmatrix}^{-1} \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
\boxed{\textit{cf. proof of L1}} &= \frac{\log\gamma}{\bar{t} - \beta} \left[\frac{\bar{t}}{\beta} - \frac{\beta}{\bar{t}},\, \frac{1}{\bar{t}} - \frac{1}{\beta} \right] \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
&= \frac{\log\gamma}{(\bar{t} - \beta)\,\bar{t}\beta} \left[\bar{t}^2 - \beta^2,\, \beta - \bar{t}\right] \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
&= \frac{\log\gamma}{\bar{t}\beta} \left[\bar{t} + \beta,\, -1\right] \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
&= \frac{\log\gamma}{\bar{t}\beta} \left((\bar{t} + \beta)\,\mu_1 - \mu_2 \right)\\
\boxed{\textit{definition of } \bar{t}} &= (\log\gamma)\, \frac{\beta n - \mu_1}{\beta^2\mu_1 - \beta\mu_2} \left(\left(\frac{\beta\mu_1 - \mu_2 + \beta^2 n - \beta\mu_1}{\beta n - \mu_1} \right) \mu_1 - \mu_2 \right)\\
&= (\log\gamma)\, \frac{-\mu_1\mu_2 + \beta^2 n\mu_1 - \beta n\mu_2 + \mu_1\mu_2}{\beta^2\mu_1 - \beta\mu_2}\\
&= (\log\gamma)\, \frac{\beta^2 n\mu_1 - \beta n\mu_2}{\beta^2\mu_1 - \beta\mu_2}\\
&= n \cdot \log\gamma \ . \qquad \square
\end{aligned}$$

Theorem 1 (Upper bound with $\tilde{\mu}_2$). *For a given positive definite matrix* $\mathbf{D} \in \mathbb{R}^{n\times n}$ *with trace* μ_1 *and squared Frobenius norm* μ_2 *the following holds:*

$$\log\det(\mathbf{D}) \leq \mathrm{ub}(\beta,\mu_1,\mu_2) \leq \mathrm{ub}(\beta,\mu_1,\tilde{\mu}_2) \ , \tag{S12}$$

if $\tilde{\mu}_2$ *is a lower bound of* μ_2*.*

Proof. The first part of the inequality was proved by Bai and Golub [5] and the proof for the second part is straightforward by applying Lemma 2 with $\gamma = \frac{1}{\beta}$ followed by using Lemma 1:

$$\begin{aligned}
\mathrm{ub}(\beta,\mu_1,\mu_2) &= \mathrm{ub}\left(1,\frac{\mu_1}{\beta},\frac{\mu_2}{\beta^2}\right) - n\cdot\log\left(\frac{1}{\beta}\right) && \boxed{\textit{Lemma 2}}\\
&\leq \mathrm{ub}\left(1,\frac{\mu_1}{\beta},\frac{\tilde{\mu}_2}{\beta^2}\right) - n\cdot\log\left(\frac{1}{\beta}\right) && \boxed{\textit{Lemma 1}}\\
&= \mathrm{ub}(\beta,\mu_1,\tilde{\mu}_2) \ . && \boxed{\textit{Lemma 2}}
\end{aligned}$$

□

The upper bound of Bai and Golub [5] is also plotted in Fig. 2 for various values of μ_2, μ_1, and β. It can be seen that the value of the bound $\mathrm{ub}(\beta,\mu_1,\mu_2)$ is monotonically decreasing with respect to μ_2. Therefore, using a lower bound of μ_2 leads to an upper bound of $\mathrm{ub}(\beta,\mu_1,\mu_2)$.

References

1. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. The MIT Press (2006)
2. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on AI. (1999) 55–60
3. Tommasi, T., Caputo, B.: The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: Proceedings of the 2009 British Machine Vision Conference (BMVC'09). (2009)
4. Ablavsky, V., Sclaroff, S.: Learning parameterized histogram kernels on the simplex manifold for image and action classification. In: ICCV. (2011) 1473–1480
5. Bai, Z., Golub, G.: Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices. Annals of Num. Mathematics **4** (1997) 29–38

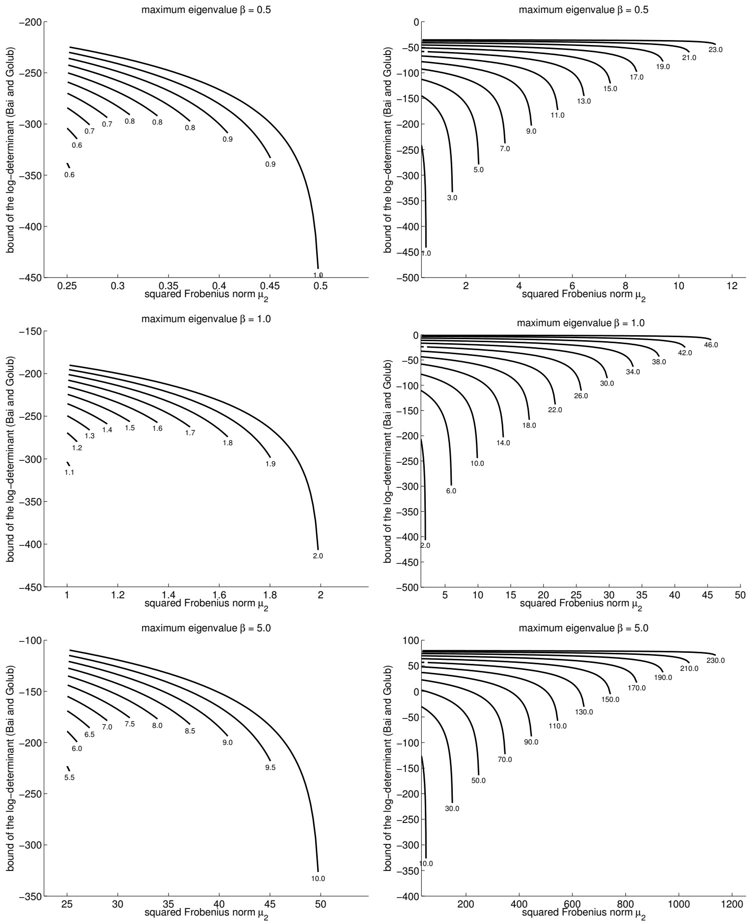

Fig. 2. Illustration of the fact that the upper bound of [5] is still an upper bound when using a lower bound of the squared Frobenius norm. The bound $\mathrm{ub}(\beta, \mu_1, \mu_2)$ is plotted in the valid range $\beta \leq \mu_2 \leq \min\{(\mu_1 - \beta)^2 + \beta^2, \beta\mu_1\}$ for $\beta = \{0.5, 1.0, 5.0\}$ and $n = 50$. It can be seen that ub is monotonically decreasing in μ_2.