On Information Gain and Regret Bounds in Gaussian Process Bandits: Supplementary Materials

1 Detailed Proofs

In this section, we provide detailed proofs for Theorem 3 and Corollary 1, as well as a comparison with the analysis of γ_T in Srinivas et al. (2010).

Proof of Theorem 3: We bound $I(\mathbf{y}_t; \hat{f}) = \frac{1}{2}\log\det(\mathbf{I}_t + \frac{1}{\tau}K_{\mathbf{X}_t,\mathbf{X}_t})$ for an arbitrary observation sequence $\mathbf{X}_t$. Recall k_P and k_O and the respective covariance matrices $K_{P,\mathbf{X}_t,\mathbf{X}_t} = [k_P(x_i,x_j)]_{i,j=1}^T$ and $K_{O,\mathbf{X}_t,\mathbf{X}_t} = [k_O(x_i,x_j)]_{i,j=1}^T$, where k_P corresponds to the D-dimensional projection in the RKHS of k, and k_O corresponds to the orthogonal element. Noticing $K_{\mathbf{X}_t,\mathbf{X}_t} = K_{P,\mathbf{X}_t,\mathbf{X}_t} + K_{O,\mathbf{X}_t,\mathbf{X}_t}$, we have

$$\begin{aligned} I(\mathbf{y}_t;\hat{f}) &= \frac{1}{2}\log\det(\mathbf{I}_t + \frac{1}{\tau}K_{\mathbf{X}_t,\mathbf{X}_t}) \\ &= \frac{1}{2}\log\det\left(\mathbf{I}_t + \frac{1}{\tau}(K_{P,\mathbf{X}_t,\mathbf{X}_t} + K_{O,\mathbf{X}_t,\mathbf{X}_t})\right) \\ &= \frac{1}{2}\log\det\left((\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})(\mathbf{I}_t + \frac{1}{\tau}(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}K_{O,\mathbf{X}_t,\mathbf{X}_t})\right) \\ &= \frac{1}{2}\log\det(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t}) + \frac{1}{2}\log\det\left(\mathbf{I}_t + \frac{1}{\tau}(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}K_{O,\mathbf{X}_t,\mathbf{X}_t}\right), \end{aligned} \tag{1}$$

where for the last line we used $\det(AB) = \det(A)\det(B)$ which holds for all two square matrices of the same dimensions. The equation (1) decouples the log det of the covariance matrix corresponding to k into that of k_P and a residual term depending on k_O. We now proceed to bounding the two terms on the right hand side of (1).

We can upper bound the first term on the right hand side of (1) using a bound on the log det of the Gram matrix in the D-dimensional feature space of k_P. Let us define $\mathbf{\Phi}_{t,D} = [\boldsymbol{\phi}_D(x_1), \boldsymbol{\phi}_D(x_2), \ldots, \boldsymbol{\phi}_D(x_t)]^\top$, a $t \times D$ matrix which stacks the feature vectors $\boldsymbol{\phi}_D^\top(x_s)$, $s = 1, \ldots, t$, at the observation points, as its rows. Notice that

$$K_{P,\mathbf{X}_t,\mathbf{X}_t} = \mathbf{\Phi}_{t,D}\Lambda_D\mathbf{\Phi}_{t,D}^\top.$$

Consider the Gram matrix

$$G_t = \Lambda_D^{\frac{1}{2}}\mathbf{\Phi}_{t,D}^\top\mathbf{\Phi}_{t,D}\Lambda_D^{\frac{1}{2}}.$$

By Weinstein–Aronszajn identity[1] (Pozrikidis, 2014)

$$\det(\mathbf{I}_D + \frac{1}{\tau}G_t) = \det(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t}). \tag{2}$$

We can prove the following lemma on the relation between the log det and the trace of a positive definite matrix.

Lemma 1. *For all positive definite matrices $P \in \mathbb{R}^{n\times n}$, we have*

$$\log\det(P) \leq n\log(tr(P)/n).$$

[1] That is a special case of matrix determinant lemma.

The proof is provided at the end of this section.

We next bound the trace of $\mathbf{I}_D + \frac{1}{\tau}G_t$. Notice that, for all $x \in \mathcal{X}$,

$$\begin{aligned}\|\boldsymbol{\phi}_D(x)\Lambda_D^{\frac{1}{2}}\|_2^2 &= \sum_{m=1}^{D}\lambda_m\phi_m^2(x)\\ &= k_P(x,x)\\ &\leq \bar{k}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{tr}(\mathbf{I}_D + \frac{1}{\tau}G_t) &= D + \frac{1}{\tau}\mathrm{tr}\left(\sum_{s=1}^{t}\Lambda^{\frac{1}{2}}\boldsymbol{\phi}_D(x_s)\boldsymbol{\phi}_D^\top(x_s)\Lambda^{\frac{1}{2}}\right)\\ &= D + \frac{1}{\tau}\sum_{s=1}^{t}\mathrm{tr}\left(\Lambda^{\frac{1}{2}}\boldsymbol{\phi}_D(x_s)\boldsymbol{\phi}_D^\top(x_s)\Lambda^{\frac{1}{2}}\right)\\ &= D + \frac{1}{\tau}\sum_{s=1}^{t}\mathrm{tr}\left(\Lambda^{\frac{1}{2}}\boldsymbol{\phi}_D^\top(x_s)\boldsymbol{\phi}_D(x_s)\Lambda^{\frac{1}{2}}\right)\\ &= D + \frac{1}{\tau}\sum_{s=1}^{t}\|\boldsymbol{\phi}_D(x_s)\Lambda^{\frac{1}{2}}\|_2^2\\ &\leq D + \frac{t\bar{k}}{\tau}.\end{aligned}$$

For the first line we expanded the Gram matrix, the second line holds by distributivity of trace over sum, and the third line is a result of $\mathrm{tr}(AA^\top) = \mathrm{tr}(A^\top A)$ which holds for any matrix A.

Using Lemma 1 and (2), we have

$$\begin{aligned}\log\det(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t}) &= \log\det(\mathbf{I}_D + \frac{1}{\tau}G_t)\\ &\leq D\log\left(\frac{\mathrm{tr}(\mathbf{I}_D + \frac{1}{\tau}G_t)}{D}\right)\\ &= D\log(1 + \frac{\bar{k}t}{\tau D}).\end{aligned} \tag{3}$$

To upper bound the second term on the right hand side of (1), we use $k_O(x,x') \leq \delta_D$. Notice that $(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}$ is a positive definite matrix whose largest eigenvalue is upper bounded by 1. For two positive definite matrices P_1, P_2 with the same dimensions, we have $\mathrm{tr}(P_1P_2) \leq \bar{\lambda}_{P_1}\mathrm{tr}(P_2)$ where $\bar{\lambda}_{P_1}$ is the largest eigenvalue of P_1 (cf. Fang et al. (1994)). Thus

$$\mathrm{tr}\left((\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}K_{O,\mathbf{X}_t,\mathbf{X}_t}\right) \leq \mathrm{tr}(K_{O,\mathbf{X}_t,\mathbf{X}_t}).$$

Since $\forall x, x' \in \mathcal{X},\ k_O(x,x') \leq \delta_D$, we have $\mathrm{tr}(K_{O,\mathbf{X}_t,\mathbf{X}_t}) \leq t\delta_D$. Therefore,

$$\mathrm{tr}\left(\mathbf{I}_t + \frac{1}{\tau}(\mathbf{I}_t + \frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}K_{O,\mathbf{X}_t,\mathbf{X}_t}\right) \leq t(1 + \frac{1}{\tau}\delta_D).$$

Using Lemma 1, we have

$$\begin{aligned}\log\det\left(\mathbf{I}_t+\frac{1}{\tau}(\mathbf{I}_t+\frac{1}{\tau}K_{P,\mathbf{X}_t,\mathbf{X}_t})^{-1}K_{O,\mathbf{X}_t,\mathbf{X}_t}\right) &\leq t\log\left(\frac{t(1+\frac{1}{\tau}\delta_D)}{t}\right)\\ &= t\log(1+\frac{1}{\tau}\delta_D)\\ &\leq \frac{t\delta_D}{\tau},\end{aligned} \tag{4}$$

where for the last line we used $\log(1+z)\leq z$ which holds for all $z\in\mathbb{R}$.

Putting (1), (3) and (4) together, we arrive at the following bound on the information gain.

$$I(\mathbf{y}_t;\hat{f}) \leq \frac{1}{2}D\log(1+\frac{\bar{k}t}{\tau D})+\frac{1}{2}\frac{t\delta_D}{\tau},$$

which holds for any arbitrary sequence $\mathbf{X}_t\subseteq\mathcal{X}$. Thus

$$\begin{aligned}\gamma_T &= \sup_{\mathbf{X}_T\subseteq\mathcal{X}} I(\mathbf{y}_T;\hat{f})\\ &\leq \frac{1}{2}D\log(1+\frac{\bar{k}T}{\tau D})+\frac{1}{2}\frac{T\delta_D}{\tau}.\end{aligned}$$

Proof of Lemma 1. Let $\{\kappa_m>0\}_{m=1}^n$ denote the eigenvalues of P. Using the inequality of arithmetic and geometric means

$$\prod_{m=1}^n\kappa_m\leq(\frac{1}{n}\sum_{m=1}^n\kappa_m)^n.$$

Thus,

$$\begin{aligned}\log\det(P) &= \log\left(\prod_{m=1}^n\kappa_m\right)\\ &\leq \log\left((\frac{1}{n}\sum_{m=1}^n\kappa_m)^n\right)\\ &= \log\left((\frac{\mathrm{tr}(P)}{n})^n\right)\\ &= n\log\left(\frac{\mathrm{tr}(P)}{n}\right).\end{aligned}$$

□

Comparison with the analysis of Srinivas et al. (2010): In comparison, Srinivas et al. (2010), in their analysis of the information gain, first showed that $I(\mathbf{y}_t,\hat{f})=\log\det(\mathbf{I}_t+K_{\mathbf{X}_t,\mathbf{X}_t})$ is a submodular function in $\mathbf{X}_t$. While finding the observation sequence that maximizes $I(\mathbf{y}_T,\hat{f})$ is NP-hard (Ko et al., 1995), Srinivas et al. (2010) used the properties of submodular functions to show that γ_T is within a constant factor of $\log\det(\mathbf{I}_T+K_{\tilde{\mathbf{X}}_T,\tilde{\mathbf{X}}_T})$ where $\tilde{\mathbf{X}}_T$ is a sequence of observation points that is selected, in a greedy fashion, to maximize $\mathcal{D}_T=\sum_{t=1}^T\sigma_{t-1}^2(x_t)$. Then, in order to bound $\log\det(\mathbf{I}_T+K_{\tilde{\mathbf{X}}_T,\tilde{\mathbf{X}}_T})$, they used the proximity of the eigenvalues of $K_{\tilde{\mathbf{X}}_T,\tilde{\mathbf{X}}_T}$ and those of the kernel k. In contrast, we directly work with the eigenvalues of k. The key idea in our analysis is the finite dimensional projection in the RKHS which allows us to bound the information gain for an arbitrary observation sequence, without having to handle the complexities of the greedy observation sequence and the eigenvalues of its covariance matrix. In a related work to the approach of Srinivas et al. (2010), Seeger et al. (2008) proved bounds

on $\mathbb{E}[I(\mathbf{y}_t, \hat{f})]$ where the expectation is taken with respect to a prior distribution on $\mathbf{X}_t$. Those bounds are not applicable to the sequential optimization problem due to the difference in the design of $\mathbf{X}_t$.

Proof of Corollary 1: Under the (C_p, β_p) polynomial eigendecay condition, the following bound on δ_D is straightforwardly derived from the decay rate of λ_m.

$$\begin{aligned}\delta_D &= \sum_{m=D+1}^{\infty} \lambda_m \psi^2 \\ &\leq \sum_{m=D+1}^{\infty} C_p m^{-\beta_p} \psi^2 \\ &\leq \int_{z=D}^{\infty} C_p z^{-\beta_p} \psi^2 dz \\ &= C_p D^{1-\beta_p} \psi^2.\end{aligned}$$

We select $D = \lceil (C_p \psi^2 T)^{\frac{1}{\beta_p}} \tau^{-\frac{1}{\beta_p}} \log^{-\frac{1}{\beta_p}}(1 + \frac{\bar{k}T}{\tau}) \rceil$ which is the smallest D ensuring $\frac{T\delta_D}{\tau} \leq D \log(1 + \frac{\bar{k}T}{\tau})$; thus, resulting in the lowest growth rate of γ_T based on Theorem 3, which implies $\gamma_T \leq D \log(1 + \frac{\bar{k}T}{\tau})$

$$\gamma_T \leq \left((C_p \psi^2 T)^{\frac{1}{\beta_p}} \tau^{-\frac{1}{\beta_p}} \log^{-\frac{1}{\beta_p}}(1 + \frac{\bar{k}T}{\tau}) + 1 \right) \log(1 + \frac{\bar{k}T}{\tau}).$$

Under the $(C_{e,1}, C_{e,2}, \beta_e)$ exponential eigendecay condition,

$$\begin{aligned}\delta_D &= \sum_{m=D+1}^{\infty} \lambda_m \psi^2 \\ &\leq \sum_{m=D+1}^{\infty} C_{e,1} \exp(-C_{e,2} m^{\beta_e}) \psi^2 \\ &\leq \int_{z=D}^{\infty} C_{e,1} \exp(-C_{e,2} z^{\beta_e}) \psi^2 dz.\end{aligned}$$

Now, consider two different cases of $\beta_e = 1$ and $\beta_e \neq 1$. When $\beta_e = 1$,

$$\begin{aligned}\int_{z=D}^{\infty} \exp(-C_{e,2} z^{\beta_e}) dz &= \int_{z=D}^{\infty} \exp(-C_{e,2} z) dz \\ &= \frac{1}{C_{e,2}} \exp(-C_{e,2} D).\end{aligned}$$

When $\beta_e \neq 1$, we have

$$\begin{aligned}\int_{z=D}^{\infty} \exp(-C_{e,2} z^{\beta_e}) dz &= \frac{1}{\beta_e} \int_{z=D^{\beta_e}}^{\infty} z^{\frac{1}{\beta_e}-1} \exp(-C_{e,2} z) dz \\ &= \frac{1}{\beta_e} \int_{z=D^{\beta_e}}^{\infty} z^{\frac{1}{\beta_e}-1} \exp(-C_{e,2} \frac{z}{2}) \exp(-C_{e,2} \frac{z}{2}) dz \\ &\leq \frac{1}{\beta_e} \int_{z=D^{\beta_e}}^{\infty} (\frac{2}{C_{e,2}} (\frac{1}{\beta_e} - 1))^{\frac{1}{\beta_e}-1} \exp(-(\frac{1}{\beta_e} - 1)) \exp(-C_{e,2} \frac{z}{2}) dz \\ &= \frac{2}{C_{e,2} \beta_e} (\frac{2}{C_{e,2}} (\frac{1}{\beta_e} - 1))^{\frac{1}{\beta_e}-1} \exp(-(\frac{1}{\beta_e} - 1)) \exp(-C_{e,2} \frac{D^{\beta_e}}{2}).\end{aligned}$$

The first equality is obtained by a change of parameter. The inequality holds since

$$\max_{z\in\mathbb{R}} z^{\frac{1}{\beta_e}-1}\exp(-C_{e,2}\frac{z}{2}) = (\frac{2}{C_{e,2}}(\frac{1}{\beta_e}-1))^{\frac{1}{\beta_e}-1}\exp(-(\frac{1}{\beta_e}-1)) \tag{5}$$

which can be verified using the standard method of equating the derivative of the left hand side to zero.

With a similar logic to the polynomial eigendecay case, when $\beta_e = 1$, we select

$$D = \lceil \frac{1}{C_{e,2}}\log(\frac{C_{e,1}\psi^2 T}{\tau C_{e,2}})\rceil.$$

When $\beta_e \neq 1$, we select

$$D = \left\lceil \left(\frac{2}{C_{e,2}}\left(\log(T) + \log(\frac{2C_{e,1}\psi^2}{\tau\beta_e C_{e,2}}) + (\frac{1}{\beta_e}-1)\left(\log(\frac{2}{C_{e,2}}(\frac{1}{\beta_e}-1))-1\right)\right)\right)^{\frac{1}{\beta_e}}\right\rceil.$$

Theorem 3 implies

$$\gamma_T \leq \left(\left(\frac{2}{C_{e,2}}(\log(T)+C_{\beta_e})\right)^{\frac{1}{\beta_e}}+1\right)\log(1+\frac{\bar{k}T}{\tau}),$$

$C_{\beta_e} = \log(\frac{C_{e,1}\psi^2}{\tau C_{e,2}})$ when $\beta_e = 1$, and $C_{\beta_e} = \log(\frac{2C_{e,1}\psi^2}{\tau\beta_e C_{e,2}}) + (\frac{1}{\beta_e}-1)\left(\log(\frac{2}{C_{e,2}}(\frac{1}{\beta_e}-1))-1\right)$, otherwise.

References

Y. Fang, K. A. Loparo, and X. Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

C. Pozrikidis. *An introduction to grids, graphs, and networks.* Oxford University Press, 2014.

M. W. Seeger, S. M. Kakade, and D. P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.