A Proof of Lemma 1

To prove Lemma 1, we will first upper bound the cumulative online Bayesian loss associated with OI-EGP, $\sum_{t=1}^{T}\ell_{t|t-1}$, relative to that incurred by any RF-based GP expert s, namely $\sum_{t=1}^{T}l^s_{t|t-1}$. Reorganizing (23), we have $\exp(-\ell_{t|t-1})/\exp(-l^s_{t|t-1})=w^s_{t-1}/w^s_t$, and after multiplying (23) from $t=1$ to T, it follows that

$$\frac{\exp(-\sum_{t=1}^{T}\ell_{t|t-1})}{\exp(-\sum_{t=1}^{T}l^s_{t|t-1})}=\frac{w^s_0}{w^s_T}=\frac{1}{Sw^s_T}$$

whose logarithm yields

$$\sum_{t=1}^{T}\ell_{t|t-1}=\sum_{t=1}^{T}l^s_{t|t-1}+\log S+\log w^s_T\overset{(a)}{\leq}\sum_{t=1}^{T}l^s_{t|t-1}+\log S \tag{41}$$

where (a) holds because $w^s_T\in[0,1]$. Thus, $\sum_{t=1}^{T}\ell_{t|t-1}-\sum_{t=1}^{T}l^s_{t|t-1}$ is upper bounded by $\log S$.

Next, we bound the difference between $\sum_{t=1}^{T}l^s_{t|t-1}$ and the loss incurred by any fixed strategy $\boldsymbol{\theta}^s_*$, for any expert $s\in\mathcal{S}$. To this end, we prove an intermediate lemma where we drop s for notational brevity; see also (Kakade and Ng, 2005). Upon defining the cumulative loss over T slots with a fixed strategy $\boldsymbol{\theta}$ as

$$\mathcal{L}_\theta:=-\log p(\mathbf{y}_T|\boldsymbol{\theta};\mathbf{X}_T)=\sum_{t=1}^{T}\mathcal{L}(\boldsymbol{\phi}^\top_{\mathbf{v}}(\mathbf{x}_t)\boldsymbol{\theta};y_t)$$

the expected cumulative loss over $q(\boldsymbol{\theta})$, a pdf of the fixed strategy $\boldsymbol{\theta}$, can be defined as

$$\bar{\mathcal{L}}_{q_\theta}:=\mathbb{E}_q[\mathcal{L}_\theta]=\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\mathcal{L}_\theta d\boldsymbol{\theta}\ .$$

Now we are ready to establish the following intermediate lemma.

Lemma 2: With $p(\boldsymbol{\theta})$ denoting the prior of $\boldsymbol{\theta}$ and KL the Kullback-Leibler divergence, it holds for any $q(\boldsymbol{\theta})$

$$\sum_{t=1}^{T}\ell_{t|t-1}\leq\bar{\mathcal{L}}_{q_\theta}+KL(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}))\ . \tag{42}$$

Proof: Based on Bayes rule, the following equality holds for the cumulative online Bayesian loss

$$\sum_{t=1}^{T}\ell_{t|t-1}=\sum_{t=1}^{T}-\log p(y_t|\mathbf{y}_{t-1};\mathbf{X}_t)=-\log p(\mathbf{y}_T;\mathbf{X}_T)$$

which after employing the definition of the KL divergence, leads to

$$\sum_{t=1}^{T}l_{t|t-1}-\bar{\mathcal{L}}_{q_\theta}=\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\log\frac{p(\mathbf{y}_T|\boldsymbol{\theta};\mathbf{X}_T)}{p(\mathbf{y}_T;\mathbf{X}_T)}d\boldsymbol{\theta}\ . \tag{43}$$

Further, since $p(\mathbf{y}_T,\boldsymbol{\theta};\mathbf{X}_T)=p(\mathbf{y}_T|\boldsymbol{\theta};\mathbf{X}_T)p(\boldsymbol{\theta})=p(\mathbf{y}_T;\mathbf{X}_T)p(\boldsymbol{\theta}|\mathbf{y}_T;\mathbf{X}_T)$, the RHS of (43) can be rewritten as

$$\begin{aligned}\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\log\frac{p(\mathbf{y}_T|\boldsymbol{\theta};\mathbf{X}_T)}{p(\mathbf{y}_T;\mathbf{X}_T)}d\boldsymbol{\theta}&=\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\log\frac{p(\boldsymbol{\theta}|\mathbf{y}_T;\mathbf{X}_T)}{p(\boldsymbol{\theta})}d\boldsymbol{\theta}\\&=\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}d\boldsymbol{\theta}-\int_{\boldsymbol{\theta}}q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y}_T;\mathbf{X}_T)}d\boldsymbol{\theta}\ \leq\ KL(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}))\end{aligned}$$

which completes the proof of Lemma 2.

To prove Lemma 1, we will use Lemma 2, and let $q(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta};\boldsymbol{\theta}_*,\xi^2\mathbf{I}_{2D})$, where ξ is a variational parameter we will tune later, and $p(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta};\mathbf{0},\sigma^2_\theta\mathbf{I}_{2D})$. It then follows that (42) becomes

$$\sum_{t=1}^{T}l_{t|t-1}-\bar{\mathcal{L}}_{q_\theta}\leq KL(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}))=2D\log\sigma_\theta+\frac{1}{2\sigma^2_\theta}\left(\|\boldsymbol{\theta}_*\|^2+2D\xi^2\right)-D-2D\log\xi\ . \tag{44}$$

Let $z_t = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_t)\boldsymbol{\theta}$ and $z_t^* = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_t)\boldsymbol{\theta}_*$. Taking the Taylor's expansion of $\mathcal{L}(z_t; y_t)$ around z_t^*, yields

$$\mathcal{L}(z_t; y_t) = \mathcal{L}(z_t^*; y_t) + \frac{d\mathcal{L}(z_t^*; y_t)}{dz_t}(z_t - z_t^*) + \frac{d^2}{dz_t^2}\mathcal{L}(h(z_t); y_t)\frac{(z_t - z_t^*)^2}{2} \tag{45}$$

where $h(z_t)$ is some function lying between z_t and z_t^*. Taking the expectation of (45) wrt $q(\boldsymbol{\theta})$, leads to

$$\begin{aligned}\mathbb{E}_q[\mathcal{L}(z_t; y_t)] &= \mathcal{L}(z_t^*; y_t) + \frac{d\mathcal{L}(z_t^*; y_t)}{dz_t} \times 0 + \mathbb{E}_q\left[\frac{d^2}{dz_t^2}\mathcal{L}(h(z_t); y_t)\frac{(z_t - z_t^*)^2}{2}\right] \\ &\overset{(a)}{\leq} \mathcal{L}(z_t^*; y_t) + c\mathbb{E}\left[\frac{(z_t - z_t^*)^2}{2}\right] \\ &\overset{(b)}{\leq} \mathcal{L}(\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_t)\boldsymbol{\theta}_*; y_t) + \frac{c\xi^2}{2}\end{aligned} \tag{46}$$

where (a) makes use of $\left|\frac{d^2}{dz^2}\mathcal{L}(z; y)\right| < c\ \forall z$ in (as1), and (b) relies on the bound $\|\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_t)\|^2 \leq 1$.

Summing (46) from $t = 1$ to T, we have

$$\bar{\mathcal{L}}_{q_\theta} \leq \mathcal{L}_{\theta_*} + \frac{Tc\xi^2}{2}\ . \tag{47}$$

Further leveraging (44), the following inequality holds

$$\sum_{t=1}^{T} l_{t|t-1} \leq \mathcal{L}_{\theta_*} + \frac{Tc\xi^2}{2} + 2D\log\sigma_\theta + \frac{1}{2\sigma_\theta^2}\left(\|\boldsymbol{\theta}_*\|^2 + 2D\xi^2\right) - D - 2D\log\xi \tag{48}$$

whose RHS is a convex function of ξ with minimal value taken at $\xi^2 = \frac{2D\sigma_\theta^2}{2D + Tc\sigma_\theta^2}$. Next, replacing the RHS with its mininal value, simplifies (48) to

$$\sum_{t=1}^{T} l_t^s \leq \sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\theta}_*^s; y_t\right) + \frac{1}{2\sigma_{\theta^s}^2}\|\boldsymbol{\theta}_*^s\|^2 + D\log\left(1 + \frac{Tc\sigma_{\theta^s}^2}{2D}\right) \tag{49}$$

for any expert $s \in \mathcal{S}$. With (49) and (41), Lemma 1 follows readily.

B Proof of Theorem 1

For a given shift-invariant standardized kernel $\bar{\kappa}^s$, the maximum point-wise error of the RF kernel approximant is uniformly bounded with probability at least $1 - 2^8(\frac{\sigma_s}{\epsilon})^2\exp\left(\frac{-D\epsilon^2}{4d+8}\right)$ by (Rahimi and Recht, 2008)

$$\sup_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}}\left|\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_i)\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_j) - \bar{\kappa}^s(\mathbf{x}_i, \mathbf{x}_j)\right| < \epsilon \tag{50}$$

where ϵ is a given constant, D is the number of spectral feature vectors, d is the dimension of $\mathbf{x}$, and $\sigma_s^2 := \mathbb{E}_{\pi_{\bar{\kappa}}^s}[\|\mathbf{v}^s\|^2]$ is the second-order moment of the RF vector $\mathbf{v}^s$.

The optimal function estimator in $\mathcal{H}^s$ incurred by κ^s is expressed as $\hat{f}^s(\mathbf{x}) := \sum_{t=1}^T \hat{\alpha}_t^s\kappa^s(\mathbf{x}, \mathbf{x}_t) = \sigma_{\theta^s}^2\sum_{t=1}^T \hat{\alpha}_t^s\bar{\kappa}^s(\mathbf{x}, \mathbf{x}_t)$; and its RF-based approximant is $\check{f}_*^s(\mathbf{x}) := \boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x})\boldsymbol{\theta}_*^s$ with $\boldsymbol{\theta}_*^s := \sigma_{\theta^s}^2\sum_{t=1}^T \hat{\alpha}_t^s\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_t)$. We then have that

$$\begin{aligned}\left|\sum_{t=1}^{T}\mathcal{L}\left(\check{f}_*^s(\mathbf{x}_t); y_t\right) - \sum_{t=1}^{T}\mathcal{L}\left(\hat{f}^s(\mathbf{x}_t); y_t\right)\right| &\overset{(a)}{\leq} \sum_{t=1}^{T}\left|\mathcal{L}\left(\check{f}_*^s(\mathbf{x}_t); y_t\right) - \mathcal{L}\left(\hat{f}^s(\mathbf{x}_t); y_t\right)\right| \\ &\overset{(b)}{\leq} \sum_{t=1}^{T} L\sigma_{\theta^s}^2\left|\sum_{t'=1}^{T}\hat{\alpha}_{t'}^s\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_{t'}) - \sum_{t'=1}^{T}\hat{\alpha}_{t'}^s\bar{\kappa}^s(\mathbf{x}_t, \mathbf{x}_{t'})\right| \\ &\overset{(c)}{\leq} \sum_{t=1}^{T} L\sigma_{\theta^s}^2\sum_{t'=1}^{T}|\hat{\alpha}_{t'}^s|\left|\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_{t'}) - \bar{\kappa}^s(\mathbf{x}_t, \mathbf{x}_{t'})\right|\end{aligned} \tag{51}$$

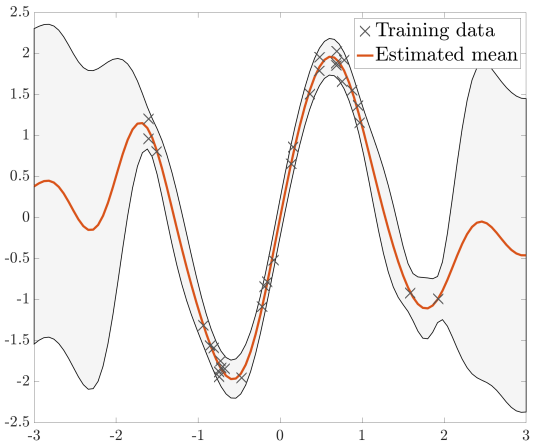

Figure 3: OI-EGP inference on a synthetic dataset. Shaded regions indicate 95% confidence intervals.

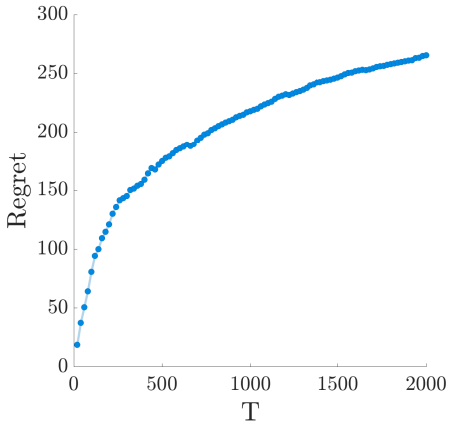

Figure 4: OI-EGP regret in (38) on the synthetic test.

where (a) follows from the triangle inequality; (b) makes use of (as2), which establishes the convexity and bounded derivative of $\mathcal{L}(z;y)$ wrt z, and (c) results from the Cauchy-Schwarz inequality. Combining with (50), we find

$$\left|\sum_{t=1}^{T}\mathcal{L}\left(\check{f}_*^s(\mathbf{x}_t);y_t\right)-\sum_{t=1}^{T}\mathcal{L}\left(\hat{f}^s(\mathbf{x}_t);y_t\right)\right|\leq\sum_{t=1}^{T}L\sigma_{\theta^s}^2\epsilon\sum_{t=1}^{T}|\hat{\alpha}_t^s|\leq\epsilon LTC,\ \text{w.h.p.} \tag{52}$$

where $C:=\max_{s\in\mathcal{S}}\sum_{t=1}^{T}\sigma_{\theta^s}^2|\hat{\alpha}_t^s|$. It thus holds that

$$\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x})\boldsymbol{\theta}_*^s;y_t\right)-\sum_{t=1}^{T}\mathcal{L}\left(\hat{f}^s(\mathbf{x}_t);y_t\right)\leq\epsilon LTC,\ \text{w.h.p.} \tag{53}$$

On the other hand, the uniform convergence bound in (50) and (as3) imply that

$$\sup_{\mathbf{x}_t,\mathbf{x}_{t'}\in\mathcal{X}}\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\phi}_{\mathbf{v}}^{s}(\mathbf{x}_{t'})\leq 1+\epsilon,\quad\text{w.h.p.} \tag{54}$$

which leads to

$$\|\boldsymbol{\theta}_*^s\|^2:=\left\|\sigma_{\theta^s}^2\sum_{t=1}^{T}\hat{\alpha}_t^s\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_t)\right\|^2=\sigma_{\theta^s}^4\sum_{t=1}^{T}\sum_{t'=1}^{T}\hat{\alpha}_t^s\hat{\alpha}_{t'}^s\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_{t'})\leq(1+\epsilon)C^2\ . \tag{55}$$

Hence, combining (55), (53) and Lemma 1, it follows for any $s\in\mathcal{S}$ that

$$\sum_{t=1}^{T}\ell_{t|t-1}-\sum_{t=1}^{T}\mathcal{L}(\hat{f}^s(\mathbf{x}_t);y_t)\leq\frac{(1+\epsilon)C^2}{2\sigma_{\theta^s}^2}+D\log\left(1+\frac{Tc\sigma_{\theta^s}^2}{2D}\right)+\log S+\epsilon LTC \tag{56}$$

thus completing the proof of Theorem 1 upon setting $s=s^*$.

Table 1: Statistics of the datasets

Datasets	Tom's hardware	SARCOS	Air Quality	Twitter
T	9725	44484	7322	98704
d	96	21	12	77

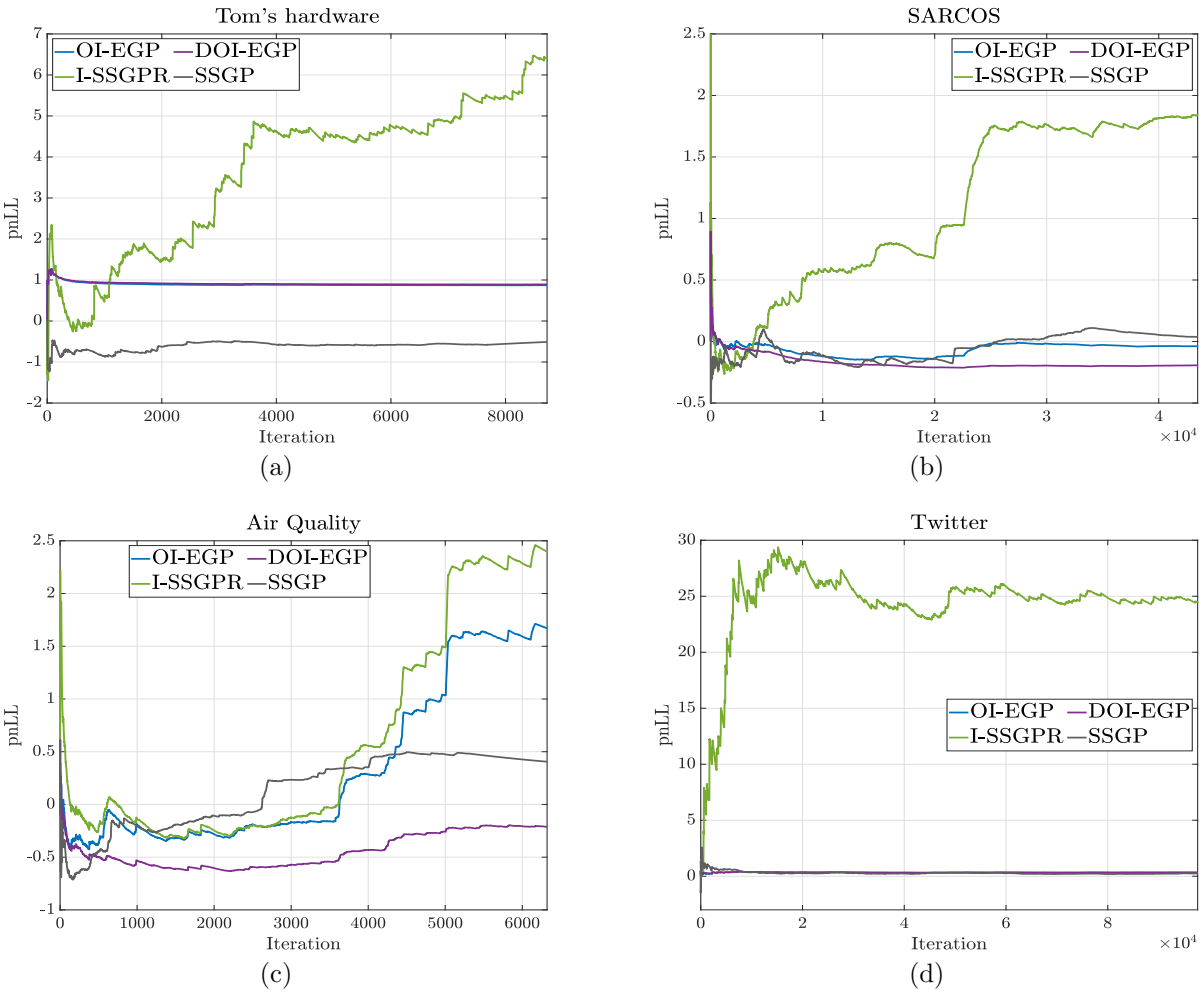

Figure 5: Predictive negative log-likelihood on (a) "Tom's hardware;" (b) SARCOS; (c) "Air quality"; and, (d) "Twitter" datasets.

C Additional real data tests

The statistics of the four datasets are summarized in Table 1. The nMSE performance and normalized running times of the competing approaches on the "Twitter" dataset are depicted in Figs. 6 and 7.

To further demonstrate (D)OI-EGP's uncertainty quantification performance, tests were conducted among GP-based approaches regarding the predictive negative log-likelihood (pnLL) as $\text{pnLL}_t := -\log p(y_t|\mathbf{y}_{t-1}; \mathbf{X}_t)$, which is computable from (20). As illustrated in Fig. 5, (D)OI-EGP always outperform I-SSGPR; while they outperform SSGP in SARCOS and air-quality datasets; they are comparable in the Twitter dataset; and perform inferior to SSGP on the Tom's hardware dataset, even though SSGP is two orders of magnitude slower than (D)OI-EGP.

D Synthetic tests

To assert the expected convergence characteristics, scalar input data $\{x_t\}_{t=1}^{30}$ were randomly drawn from a normal distribution, and outputs were generated as $y_t = \sin(2x_t) + \sin(3x_t) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, 0.01)$. The inferred mean function as well as (approximate) 95% confidence intervals are shown in Fig. 3. As expected, regions populated with training examples correspond to tighter confidence bands relative to unpopulated ones.

To validate the regret bound for OI-EGP (cf. (38)), datasets of increasing size T were generated from the aforementioned synthetic model, albeit with $x_t \sim \mathcal{N}(0, 100)$. It is evident from Fig. 4 that the regret can be

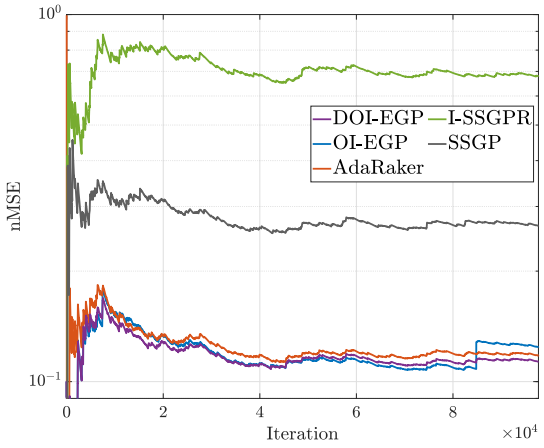

Figure 6: nMSE performance on the "Twitter" dataset. Notice the logarithmic scale.

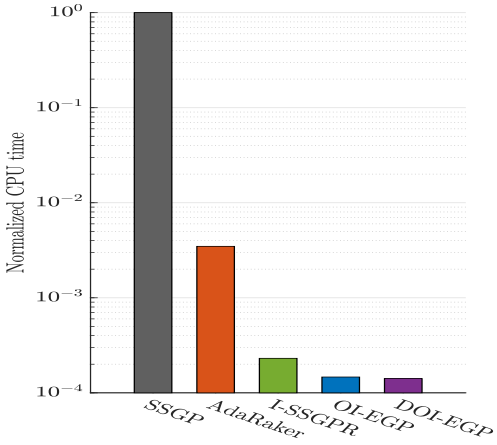

Figure 7: Normalized running times on the "Twitter" dataset. Notice the logarithmic scale.

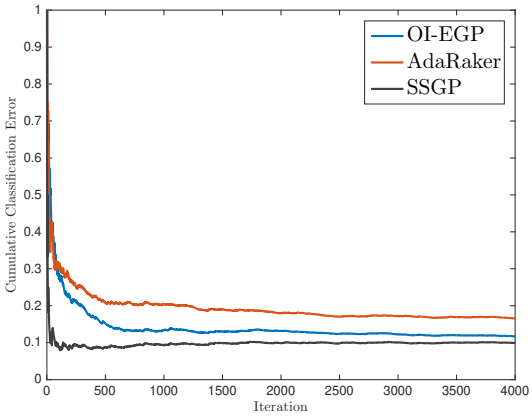

Figure 8: Classification error on the "Banana" dataset.

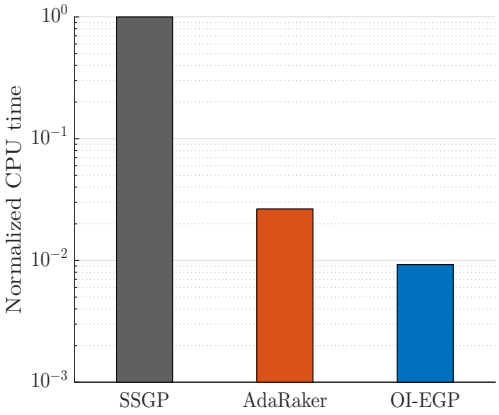

Figure 9: Running time on the "Banana" dataset. Notice the logarithmic scale.

upper bounded by $\mathcal{O}(\log T)$, as predicted by the theory.

E OI-EGP for Classification

Coupled with the logistic likelihood, our OI-EGP was tested for binary classification on the "Banana" dataset with 2-dimensional features ($d = 2$). The performance of OI-EGP was also compared with AdaRaker (Shen et al., 2019) and SSGP (Bui et al., 2017) in terms of classification error and running time. For OI-EGP and AdaRaker, the value of D was set to 15, and the kernel dictionary consisted of radial basis functions with lengthscales chosen from $\{10^{-2}, \ldots, 10^{2}\}$. As for SSGP, the ARD kernel was employed, the number of inducing points was 30, the batch size was chosen to be 200, and the first 1000 samples was used for model initialization. Targeting a tractable classification algorithm, each expert s in OI-EGP relies on Gaussian (Laplace) approximation of the posterior $p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t)$ (Rasmussen and Williams, 2006) to evaluate the integrals involved in (18)-(24) per slot t.

The cumulative classification error and running time of the three competing approaches are plotted in Figs. 8–9. While SSGP showcases the lowest classification error due to the more powerful ARD kernel, it is much more time-consuming than the other two RF-based alternatives, among which our novel OI-EGP outperforms AdaRaker in both classification error and running time.