Proofs and details of the paper Stratified Bayesian Optimization

1 Parameters of the posterior distribution of F

In this section we are going to calculate the posterior distribution of $F(\cdot,\cdot)$. We have placed a Gaussian process (GP) prior distribution over the function F:

$$F(\cdot,\cdot) \sim GP\left(\mu_0(\cdot,\cdot), \Sigma_0(\cdot,\cdot,\cdot,\cdot)\right)$$

where

$$\begin{aligned} \mu_0 : (x,w) &\rightarrow \mathbb{R}, \\ \Sigma_0 : (x,w,x',w') &\rightarrow \mathbb{R}, \end{aligned}$$

and Σ_0 is a positive semi-definite function. We choose Σ_0 such that closer arguments are more likely to correspond to similar values, i.e. $\Sigma_0(x,w,x',w')$ is a decreasing function of the distance between (x,w) and (x',w'). Specifically, we can use the squared exponential covariance function:

$$\Sigma_0\left(x, w^{(1)}, x', w'^{(1)}\right) = \sigma_0^2 \exp\left(-\sum_{k=1}^{n} \alpha_1^{(k)} \left[x_k - x'_k\right]^2 - \sum_{k=1}^{d_1} \alpha_2^{(k)} \left[\omega_k^{(1)} - \omega'_k(1)\right]^2\right)$$

where σ_0^2 is the common prior variance, and $\alpha_1^{(1)},\ldots,\alpha_1^{(n)},\alpha_2^{(1)},\ldots,\alpha_2^{(d_1)} \in \mathbb{R}_+$ are the length scales. These values are calculated using likelihood estimation from the observations of F.

First, observe that standard results from Gaussian process regression provide the following expressions for μ_n and Σ_n (the parameters of the posterior distribution of F).

$$\begin{aligned} \mu_n(x,w) &= \mu_0(x,w) \\ &\quad + \left[\Sigma_0(x,w,x_1,w_1) \ \cdots \ \Sigma_0(x,w,x_n,w_n)\right] A_n^{-1} \\ &\quad \times \begin{pmatrix} y_1 - \mu_0(x_1,w_1) \\ \vdots \\ y_n - \mu_0(x_n,w_n) \end{pmatrix} \\ \Sigma_n(x,w,x',w') &= \Sigma_0(x,w,x',w') \\ &\quad - \left[\Sigma_0(x,w,x_1,w_1) \ \cdots \ \Sigma_0(x,w,x_n,w_n)\right] A_n^{-1} \begin{pmatrix} \Sigma_0(x',w',x_1,w_1) \\ \vdots \\ \Sigma_0(x',w',x_n,w_n) \end{pmatrix} \end{aligned}$$

where

$$A_n = \begin{bmatrix} \Sigma_0(x_1,w_1,x_1,w_1) & \cdots & \Sigma_0(x_1,w_1,x_n,w_n) \\ \vdots & \ddots & \vdots \\ \Sigma_0(x_n,w_n,x_1,w_n) & \cdots & \Sigma_0(x_n,w_n,x_n,w_n) \end{bmatrix} + \operatorname{diag}\left(\sigma^2(x_1,w_1),\ldots,\sigma^2(x_n,w_n)\right).$$

2 Computation of the Value of Information

The following proposition that allows us to proof Lemma 1 in the paper.

Proposition 1. We have that

$$a_{n+1}(x) \mid \mathcal{F}_n, (x_{n+1}, w_{n+1}) \sim N\left(a_n(x), \sigma_n^2(x, x_{n+1}, w_{n+1})\right)$$

where

$$\sigma_n^2(x, x_{n+1}, w_{n+1}) = \mathrm{Var}_n[G(x)] - \mathbb{E}_n[\mathrm{Var}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}]$$

Proof.

$$a_{n+1}(x) = \mathbb{E}[\mu_{n+1}(x,w)] = \mathbb{E}[\mu_0(x,w)] + \begin{bmatrix} B(1) & \cdots & B(n+1) \end{bmatrix} A_{n+1}^{-1} \begin{pmatrix} y_1 - \mu_0(x_1, w_1) \\ \vdots \\ y_{n+1} - \mu_0(x_{n+1}, w_{n+1}) \end{pmatrix} \quad (1)$$

where

$$B(i) = \int \Sigma_0(x, w, x_i, w_i)\, dw$$

for $i = 1, \ldots, n+1$. Since y_{n+1} conditioned on $\mathcal{F}_n, x_{n+1}, w_{n+1}$ is normally distributed, then $a_{n+1}(x) \mid \mathcal{F}_n, x_{n+1}, w_{n+1}$ is also normally distributed. By tower property,

$$\begin{aligned} \mathbb{E}_n[a_{n+1}(x) \mid x_{n+1}, w_{n+1}] &= \mathbb{E}_n[\mathbb{E}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}] \\ &= \mathbb{E}_n[G(x)] \\ &= a_n(x) \end{aligned}$$

and

$$\begin{aligned} \sigma_n^2(x, x_{n+1}, w_{n+1}) &= \mathrm{Var}_n[\mathbb{E}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}] \\ &= \mathrm{Var}_n[G(x)] - \mathbb{E}_n[\mathrm{Var}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}]. \end{aligned}$$

Using the equation (1) and the previous proposition, we can get the following formula for a_n

$$a_{n+1} = a_n + \sigma_n(x, x_{n+1}, w_{n+1}) Z$$

where $Z \sim N(0,1)$, which is the Lemma 1 of the paper.

2.1 Computation of $\sigma_n(x, x_{n+1}, w_{n+1})$ and $a_n(x)$

In the following sections, we also give closed formulas for the equations when w follows a normal distribution and its components are indepent, specifically $w_i \sim N(\mu_i, \sigma_i^2)$ and the kernel is the squared exponential kernel.

Now, observe that

$$
\begin{aligned}
a_n(x) &= \mathbb{E}\left[\mu_n(x,w)\right] \\
&= \mathbb{E}\left[\mu_0(x,w)\right] \\
&\quad + \left[B(x,1) \ \ \cdots \ \ B(x,n)\right] A_n^{-1} \begin{pmatrix} y_1 - \mu_0(x_1,w) \\ \vdots \\ y_n - \mu_0(x_n,w) \end{pmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
B(x,i) &= \int \Sigma_0(x,w,x_i,w_i)\,dw \\
&= \sigma_0^2 \exp\left(-\sum_{k=1}^{n} \alpha_1^{(k)} \left[x_k - x_{ik}\right]^2\right) \prod_{k=1}^{d_1} \int \exp\left(-\alpha_2^{(k)} \left[w_k - w_{ik}\right]^2\right) dp(w_k)
\end{aligned}
$$

for $i = 1, \ldots, n$. In the particular case given at the beggining, we can compute $\int \exp\left(-\alpha_2^{(k)} \left[w_k - w_{ik}\right]^2\right) dp(w_k$ for any k and i:

$$
\begin{aligned}
&\int \exp\left(-\alpha_2^{(k)} \left[w_k - w_{ik}\right]^2\right) dp\left(w_k^{(1)}\right) \\
= \ & \frac{1}{\sqrt{2\pi}\sigma_k} \int \exp\left(-\alpha_2^{(k)} \left[z - w_{ik}\right]^2 - \frac{\left[z-\mu_k\right]^2}{2\sigma_k^2}\right) dz \\
= \ & \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\mu_k^2}{2\sigma_k^2} - \alpha_2^{(k)} \left(w_{ik}\right)^2 - \frac{\left(\frac{\mu_k}{\sigma_k^2} + 2\alpha_2^{(k)} w_{ik}\right)^2}{4\left(-\alpha_2^{(k)} - \frac{1}{2\sigma_k^2}\right)}\right) \\
& \times \int \exp\left(-\left(\alpha_2^{(k)} + \frac{1}{2\sigma_k^2}\right)\left[z - \frac{\frac{\mu_k}{\sigma_k^2} + 2\alpha_2^{(k)} w_{ik}}{2\left(b + \frac{1}{2\sigma_k^2}\right)}\right]^2\right) dz \\
= \ & \frac{1}{\sqrt{2}\sigma_k} \frac{1}{\sqrt{\alpha_2^{(k)} + \frac{1}{2\sigma_k^2}}} \exp\left(-\frac{\mu_k^2}{2\sigma_k^2} - \alpha_2^{(k)} \left(w_{ik}\right)^2 - \frac{\left(\frac{\mu_k}{\sigma_k^2} + 2\alpha_2^{(k)} w_{ik}\right)^2}{4\left(-\alpha_2^{(k)} - \frac{1}{2\sigma_k^2}\right)}\right)
\end{aligned}
$$

Now let's compute $\sigma_n^2(x, x_{n+1}, w_{n+1})$:

$$\begin{aligned}
& \sigma_n^2(x, x_{n+1}, w_{n+1}) \\
= \;& \mathrm{Var}_n[G(x)] - \mathbb{E}_n[\mathrm{Var}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}] \\
= \;& \mathrm{Var}_n[G(x) \mid x_{n+1}, w_{n+1}] - \mathrm{Var}_{n+1}[G(x) \mid x_{n+1}, w_{n+1}] \\
= \;& \int\int \Sigma_n(x, w, x, w') p(w) p(w') dwdw' \\
& - \int\int \Sigma_{n+1}(x, w, x, w') p(w) p(w') dw^{(1)} dw'^{(1)} \\
= \;& \int\int \Sigma_n(x, w, x_{n+1}, w_{n+1}) \frac{\Sigma_n(x, w', x_{n+1}, w_{n+1})}{\Sigma_n(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1})} p(w) p(w') dwdw' \\
= \;& \left[\frac{\int \Sigma_n(x, w, x_{n+1}, w_{n+1})}{\sqrt{\Sigma_n(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1})}} p(w) dw\right]^2 \\
= \;& \left[\frac{\int \Sigma_n(x, w, x_{n+1}, w_{n+1})}{\sqrt{\Sigma_n(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1})}} p(w) dw\right]^2 \\
= \;& \left[\frac{\left(B(x, n+1) - \begin{bmatrix} B(x,1) & \cdots & B(x,n)\end{bmatrix} A_n^{-1}\gamma\right)}{\sqrt{\left(\Sigma_0(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1}) - \gamma^T A_n^{-1}\gamma\right)}}\right]^2
\end{aligned}$$

where

$$\gamma = \begin{bmatrix} \Sigma_0(x_{n+1}, w_{n+1}, x_1, w_1) \\ \vdots \\ \Sigma_0(x_{n+1}, w_{n+1}, x_n, w_n) \end{bmatrix}.$$

2.2 Computation of ∇V_i

We have that

$$V_n(x, w) = \mathbb{E}_n[\mathrm{max}_{x'} a_{n+1}(x') \mid x_{n+1} = x, w_{n+1} = w] - \mathrm{max}_{x'} a_n(x')$$

where $a_n(x) := \mathbb{E}_n[\mathbb{E}[f(x, w, z)]] = \mathbb{E}_n[\mathbb{E}[F(x, w)]] = \mathbb{E}[\mu_n(x, w)]$. We need to discretize the domain of a_n and a_{n+1} to evaluate V_n.

By the previous part, conditioned on $\mathcal{F}_n, x_{n+1}, \omega_{n+1}^{(1)}$, we have that

$$\begin{aligned}
a_{n+1}(x) &= a_n(x) + \sqrt{\left(\mathrm{Var}_n[G(x)] - \mathbb{E}_n[\mathrm{Var}_{n+1}[G(x)] \mid x_{n+1}, w_{n+1}]\right)} Z_{n+1} \\
&= a_n(x) + \sigma_n(x, x_{n+1}, w_{n+1}) Z_{n+1}
\end{aligned}$$

where $Z_{n+1} \sim N(0, 1)$.

Then

$$\begin{aligned}
X^{KG}(\mathcal{F}_n) &= \arg\max_{x, \omega^{(1)}} \mathbb{E}[\mathrm{max}_{x'} a_n(x') + \tilde{\sigma}_n(x', x_{n+1}, w_{n+1}) Z_{n+1} \mid x_{n+1} = x, w_{n+1} = w] \\
& \quad - \mathrm{max}_{x'} a_n(x') \\
&= \arg\max_{x, \omega^{(1)}} h(a^n, \tilde{\sigma}_n(x, w))
\end{aligned}$$

where $a^n = (a_n(x_i))_{i=1}^M$,$\tilde{\sigma}_n(x, w) = (\tilde{\sigma}_n(x_i, x, w))_{i=1}^M$, $h : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ defined by $h(a, b) = \mathbb{E}[\mathrm{max}_i a_i + b_i Z] - \mathrm{max}_i a_i$, where a and b are any deterministic vectors, and$Z$ is a one-dimensional standard normal random variable.

Observe that h does not change its value if we reorder the components of the vectors a and b. Thus, we can suppose that $b_i \leq b_{i+1}$ for all i and $a_i \leq a_{i+1}$ if $b_i = b_{i+1}$. Using the Algorithm 1 in [], we can remove all those entries i for which $a_i + b_i z < \max_{k \neq i} a_k + b_k z$ for all z. Then, this algorithm gives us new vectors a' and b' such that

$$h(a,b) = \sum_{i=1}^{|a'|-1} \left(b'_{i+1} - b'_i\right) f\left(-|c_i|\right),$$

where

$$\begin{aligned} f(z) &:= \varphi(z) + z\Phi(z), \\ c_i &:= \frac{a'_{i+1} - a'_i}{b'_{i+1} - b'_i}, i = 1, \ldots, |a'| - 1 \end{aligned}$$

and φ, Φ are the standard normal cdf and pdf, respectively.

Now, let a' and b' be the vectors obtained when we apply the Algorithm 1 to the vectors $a^n, \tilde{\sigma}_n(x,w)$. If $|a'| = 1$, $V_n\left(x, \omega^{(1)}\right) = h\left(a^n, \tilde{\sigma}_n(x,w)\right) = 0$ and so $\nabla V_n(x,w) = 0$. On the other hand, if $|a'| > 1$,

$$\begin{aligned} \nabla V_n\left(x,\omega^{(1)}\right) &= \nabla h\left(a^n, \tilde{\sigma}_n(x,w)\right) \\ &= \sum_{i=1}^{|a'|-1} \left(b'_{i+1} - b'_i\right)\left(-\Phi\left(-|c_i|\right)\right)\nabla\left(|c_i|\right) - \left(\nabla b'_{i+1} - \nabla b'_i\right) f\left(-|c_i|\right) \\ &= \sum_{i=1}^{|a'|-1} \left(\nabla b'_{i+1} - \nabla b'_i\right)\left(-\Phi\left(-|c_i|\right)|c_i| - f\left(-|c_i|\right)\right) \\ &= \sum_{i=1}^{|a'|-1} \left(-\nabla b'_{i+1} + \nabla b'_i\right)\left(\varphi\left(|c_i|\right)\right). \end{aligned}$$

Then we only need to compute $\nabla b'_i$ for all i. Now, for the gaussian case and the squared exponential kernel,

$$\begin{aligned} \nabla \tilde{\sigma}_n(x, x_{n+1}, w) &= \nabla\left(\sqrt{\left(\mathrm{Var}_n\left[G(x)\right] - \mathbb{E}_n\left[\mathrm{Var}_{n+1}\left[G(x)\right] \mid x_{n+1}, w\right]\right)}\right) \\ &= \beta_1\left(\nabla B(x, n+1) - \nabla\left(\gamma^T\right) A_n^{-1} \begin{bmatrix} B(x,1) \\ \vdots \\ B(x,n) \end{bmatrix}\right) \quad (2) \\ &\quad -\frac{1}{2}\beta_1^3\beta_2\left[\nabla\Sigma_0\left(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1}\right) - 2\nabla\left(\gamma^T\right)A_n^{-1}\gamma\right] \quad (3) \end{aligned}$$

where

$$
\begin{aligned}
\beta_1 &= \left[\Sigma_0\left(x_{n+1}, w_{n+1}, x_{n+1}, w_{n+1}\right) - \gamma^T A_n^{-1} \gamma\right]^{-1/2} \\
\beta_2 &= B\left(x, n+1\right) - \left[B\left(x,1\right) \quad \cdots \quad B\left(x,n\right)\right] A_n^{-1}\gamma \\
\gamma &= \begin{bmatrix} \Sigma_0\left(x_{n+1}, w_{n+1}, x_1, w_1\right) \\ \vdots \\ \Sigma_0\left(x_{n+1}, w_{n+1}, x_n, w_1\right) \end{bmatrix} \\
\nabla\left(\gamma^T\right) &= \left[\nabla\Sigma_0\left(x_{n+1}, w_{n+1}, x_1, w_1\right) \cdots \nabla\Sigma_0\left(x_{n+1}, w_{n+1}, x_n, w_1\right)\right] \\
B\left(x,i\right) &= \sigma_0^2 \exp\left(-\sum_{k=1}^{n} \alpha_1^{(k)} \left[x_k - x_{ik}\right]^2\right) \\
& \quad \prod_{k=1}^{d_1} \frac{1}{\sqrt{2}\sigma_k} \frac{1}{\sqrt{\alpha_2^{(k)} + \frac{1}{2\sigma_k^2}}} \exp\left(-\frac{\mu_k^2}{2\sigma_k^2} - \alpha_2^{(k)}\left(w_{ik}\right)^2 - \frac{\left(\frac{\mu_k}{\sigma_k^2} + 2\alpha_2^{(k)} w_{ik}\right)^2}{4\left(-\alpha_2^{(k)} - \frac{1}{2\sigma_k^2}\right)}\right)
\end{aligned}
$$

Observe that we can compute (2) explicitly by plugging in

$$
\begin{aligned}
\nabla_{x_{n+1}}\Sigma_0\left(x_{n+1}, w_{n+1}, x_i, w_i\right) &= \begin{cases} 0, & i = n+1 \\ -2\alpha_1\left[x_{n+1} - x_i\right]\Sigma_0\left(x_{n+1}, w_{n+1}, x_i, w_i\right), & i < n+1 \end{cases} \\
\nabla_{w_{n+1}^{(1)}}\Sigma_0\left(x_{n+1}, w_{n+1}, x_i, w_i\right) &= \begin{cases} 0, & i = n+1 \\ -2\alpha_2\left[w_{n+1} - w_i^{(1)}\right]\Sigma_0\left(x_{n+1}, w_{n+1}, x_i, w_i\right), & i < n+1 \end{cases}
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla_{x_{n+1,i}} B\left(x, n+1\right) &= -2\alpha_1^{(j)}\left(x_j - x_{n+1,j}\right) B\left(x, n+1\right) \\
\nabla_{w_{n+1,k}} B\left(x, n+1\right) &= \sigma_0^2 \exp\left(-\sum_{i=1}^{n} \alpha_1^{(i)}\left[x_i - x_{n+1,i}\right]^2\right) \prod_{j \neq k} \int \exp\left(-\alpha_2^{(j)}\left[w_j - w_{n+1,j}\right]^2\right) dp\left(w_j\right) \\
& \quad \times \int \left(-2\alpha_2^{(k)}\left(w_k - w_{n+1,k}\right)\right) \exp\left(-\alpha_2^{(k)}\left[w_k - w_{n+1,k}\right]^2\right) dp\left(w_k\right)
\end{aligned}
$$