Bayesian Global Optimization

January 26, 2015

1 Introduction

Let $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ be a continuous function and $\left(\mathbb{R}^d, \mathcal{F}, P\right)$ be a probability space. We suppose that each evaluation has a cost. We denote the joint pdf of $\omega = \left(w^{(1)}, w^{(2)}\right) \in \mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with $d_1 \ll d_2$ by $p\left(w^{(1)}, w^{(2)}\right)$, which is assumed known. Specifically, we suppose that $p\left(w^{(1)}\right) = \prod_{i=1}^{d_1} p_i\left(w_i^{(1)}\right)$ where p_i is a normal distribution with parameters (μ_i, σ_i) for $i = 1, \ldots, d_1$. Our goal is to solve

$$\max_{x \in A \subset \mathbb{R}^n} \mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right] \tag{1}$$

for a given compact set A. We also suppose that $w^{(1)}$ has a much stronger effect on f than $w^{(2)}$, specifically we assume that

$$f\left(x, w^{(1)}, w^{(2)}\right) \mid x, w^{(1)} \sim N\left(F\left(x, w^{(1)}\right), \sigma^2\left(x, w^{(1)}\right)\right)$$

where $\sigma^2\left(x, w^{(1)}\right) := \mathrm{Var}\left(f\left(x, w^{(1)}, w^{(2)}\right) \mid w^{(1)}\right)$ and $F\left(x, w^{(1)}\right) := \mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right) \mid w^{(1)}\right]$. We suppose that $\sigma^2\left(x, w^{(1)}\right) < \infty$.

Consequently

$$\begin{aligned}
\max_{x \in A \subset \mathbb{R}^n} \mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right] &= \max_{x \in A \subset \mathbb{R}^n} \mathbb{E}\left[\mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right) \mid w^{(1)}\right]\right] \\
&= \max_{x \in A \subset \mathbb{R}^n} \mathbb{E}\left[F\left(x, w^{(1)}\right)\right]
\end{aligned}$$

We define the function $G(x) := \int F\left(x, w^{(1)}\right) dp\left(w^{(1)}\right)$.

2 Model

We place 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

$$\begin{aligned}
\mu_0 : \left(x, w^{(1)}\right) &\to \mathbb{R}, \\
\Sigma_0 : \left(x, w^{(1)}, x', w'^{(1)}\right) &\to \mathbb{R},
\end{aligned}$$

and Σ_0 is a positive semi-definite function. A typical choice of Σ_0 is the squared exponential function (see).

Let $y_n \approx F\left(x_n, w_n^{(1)}\right)$ be the observation at time n. Let $\mathcal{F}_n$ be the σ-algebra generated by $\left\{y_{1:n}, w_{1:n}^{(1)}, x_{1:n}\right\}$. At each time $n = 1, 2, \ldots, N$, our algorithm will choose some point $\left(x_n, w_n^{(1)}\right)$ based on $\mathcal{F}_{n-1}$, sample $w_{n,m}^{(2)} \sim p\left(w^{(2)} \mid w_n^{(1)}\right)$ for $m = 1, \ldots, M$ and observe $y_n = \frac{1}{M}\sum_{m=1}^{M} f\left(x_n, w_n^{(1)}, w_{n,m}^{(2)}\right)$ .The posterior distribution of F at time n is

$$F(\cdot,\cdot) \mid \mathcal{F}_n \sim GP\left(\mu_n(\cdot,\cdot), \Sigma_n(\cdot,\cdot,\cdot,\cdot)\right)$$

where μ_n and Σ_n can be computed using standard results from Bayesian linear regression. In fact, by the Kalman filter equations we have that

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

where

$$A_n = \begin{bmatrix} \Sigma_0\left(x_1, w_1^{(1)}, x_1, w_1^{(1)}\right) & \cdots & \Sigma_0\left(x_1, w_1^{(1)}, x_n, w_n^{(1)}\right) \\ \vdots & \ddots & \vdots \\ \Sigma_0\left(x_n, w_n^{(1)}, x_1, w_1^{(1)}\right) & \cdots & \Sigma_0\left(x_n, w_n^{(1)}, x_n, w_n^{(1)}\right) \end{bmatrix}.$$

Denote by $\mathbb{E}_n$ and Cov_n the expectation and covariance conditioned on $\mathcal{F}_n$, respectively. By Fubini's Theorem,

$$\begin{aligned}
\mathbb{E}_n\left[\mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right]\right] &= \mathbb{E}_n\left[\mathbb{E}\left[F\left(x, w^{(1)}\right)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_n\left[F\left(x, w^{(1)}\right)\right]\right] \\
&= \mathbb{E}\left[\mu_n\left(x, w^{(1)}\right)\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathrm{Cov}_n\left(\mathbb{E}\left[F\left(x', w'^{(1)}\right)\right], \mathbb{E}\left[F\left(x, w^{(1)}\right)\right]\right) \\
&= \int\int \Sigma_n\left(x, w^{(1)}, x', w'^{(1)}\right) p\left(w^{(1)}\right) p\left(w'^{(1)}\right) dw^{(1)} dw'^{(1)}
\end{aligned}$$

Then, if we were to stop after N evaluations of the simulator and choose the solution to (1) with the best estimated value, we would choose

$$x_N^* \in \arg\max_x \mathbb{E}_n\left[\mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right]\right] = \arg\max_x \mathbb{E}\left[\mu_n\left(x, w^{(1)}\right)\right]$$

This solution is Bayes-optimal when we are neutral with respect to the risk.

We now define a sequence of value of the information functions $(V_n)_n$ one for each time n. Let $V_n : \mathbb{R}^n \times \mathbb{R}^{d_1} \to \mathbb{R}$ defined by

$$V_n\left(x, \omega^{(1)}\right) = \mathbb{E}_n\left[\max_x a_{n+1}(x) \mid x_{n+1} = x, \omega^{(1)}_{n+1} = \omega^{(1)}\right] - \max_x a_n(x)$$

where $a_n(x) := \mathbb{E}_n\left[\mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right]\right] = \mathbb{E}_n\left[\mathbb{E}\left[F\left(x, w^{(1)}\right)\right]\right] = \mathbb{E}\left[\mu_n\left(x, w^{(1)}\right)\right]$.

The algorithm we present in §3 wants to evaluate the simulator at the point maximizing the value of the information. Thus, we seek to evaluate at time $n+1$

$$\left(x_{n+1}, \omega^{(1)}_{n+1}\right) \in \arg\max_{x,\omega} V_n(x, \omega).$$

To perform this computation, first we have to find the distribution of $a_{n+1}(x)$ conditioned on $\left(x_{n+1}, \omega^{(1)}_{n+1}\right)$ and $\mathcal{F}_n$ for any x. We perform these computations in section 4.

3 Algorithm

The following algorithm used the value functions to choose the points where the function is evaluated.

1. Evaluate F at a number randomly chosen. Fit a GP prior to F.
2. For $i \leftarrow 1$ to N do
   (a) If the stopping rule is met, go to Step 3; else go to Step 2b.
   (b) Update the distribution of a_i,V_i and ∇V_i.
   (c) Maximize $V_i(\cdot,\cdot)$ using multi-start gradient ascent. Let $\left(x_{i+1}, \omega^{(1)}_{i+1}\right)$ be the maximizer, and evaluate $\frac{1}{M}\sum_{m=1}^{M} f\left(x_{i+1}, w^{(1)}_{i+1}, w^{(2)}_{i+1,m}\right) \approx F\left(x_{i+1}, \omega^{(1)}_{i+1}\right)$ where $w^{(2)}_{i+1,m} \sim p\left(w^{(2)} \mid w^{(1)}_{i+1}\right)$
3. Return $x^* = \arg\max_x a_{N+1}(x) = \mathbb{E}\left[\mu_{N+1}\left(x, w^{(1)}\right)\right]$

4 Computations

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

$$\begin{aligned} \mu_0 : \left(x, w^{(1)}\right) &\to \mathbb{R}, \\ \Sigma_0 : \left(x, w^{(1)}, x', w'^{(1)}\right) &\to \mathbb{R}, \end{aligned}$$

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. $\Sigma_0\left(x, w^{(1)}, x', w'^{(1)}\right)$ is a decreasing function of the distance between $\left(x, w^{(1)}\right)$ and $\left(x', w'^{(1)}\right)$. Specifically, we use the squared exponential covariance function:

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

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.

The mean μ_0 is usually a linear regression function using basis functions. We are going to suppose that $\mu_0 \equiv b$ where b is a constant.

Lemma 1. We have that

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

where

$$\eta_n\left(x, x_{n+1}, \omega_{n+1}^{(1)}\right) = \mathrm{Var}_n\left[G(x)\right] - \mathbb{E}_n\left[\mathrm{Var}_{n+1}\left[G(x)\right] \mid x_{n+1}, \omega_{n+1}^{(1)}\right]$$

Proof.

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

where

$$B(i) = \int \Sigma_0\left(x, w^{(1)}, x_i, w_i^{(1)}\right) dw^{(1)}$$

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

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

and

$$\begin{aligned} \eta_n\left(x, x_{n+1}, \omega_{n+1}^{(1)}\right) &= \mathrm{Var}_n\left[\mathbb{E}_{n+1}\left[G(x)\right] \mid x_{n+1}, \omega_{n+1}^{(1)}\right] \\ &= \mathrm{Var}_n\left[G(x)\right] - \mathbb{E}_n\left[\mathrm{Var}_{n+1}\left[G(x)\right] \mid x_{n+1}, \omega_{n+1}^{(1)}\right] \end{aligned}$$

4.1 Computation of $\eta_n\left(x, x_{n+1}, \omega_{n+1}^{(1)}\right)$ and $a_n(x)$

$$\begin{aligned} a_n(x) &= \mathbb{E}\left[\mu_n\left(x, w^{(1)}\right)\right] \\ &= \mathbb{E}\left[\mu_0\left(x, w^{(1)}\right)\right] \\ &\quad + \left[B(x,1) \quad \cdots \quad B(x,n)\right] A_n^{-1} \begin{pmatrix} y_1 - \mu_0\left(x_1, w_1^{(1)}\right) \\ \vdots \\ y_n - \mu_0\left(x_n, w_n^{(1)}\right) \end{pmatrix} \end{aligned}$$

where

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

for $i = 1, \ldots, n$. We only need to compute $\int \exp\left(-\alpha_2^{(k)}\left[\omega_k^{(1)} - \omega_{ik}^{(1)}\right]^2\right) dp\left(w_k^{(1)}\right)$ for any k and i:

$$
\begin{aligned}
& \int \exp\left(-\alpha_2^{(k)}\left[\omega_k^{(1)} - \omega_{ik}^{(1)}\right]^2\right) dp\left(w_k^{(1)}\right) \\
= \; & \frac{1}{\sqrt{2\pi}\sigma_k} \int \exp\left(-\alpha_2^{(k)}\left[z - \omega_{ik}^{(1)}\right]^2 - \frac{\left[z-\mu_k\right]^2}{2\sigma_k^2}\right) dz \\
= \; & \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left[\left(\mu_k + 2\omega_{ik}^{(1)}\alpha_2^{(k)}\sigma_k^2\right)^2 \frac{\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-1}}{2\sigma_k^2} - \alpha_2^{(k)}\omega^{2}{}_{ik}^{(1)} - \frac{\mu_k^2}{2\sigma_k^2}\right] \\
& \times \sigma_k\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-.5} \int \exp\left(-\frac{u^2}{2}\right) du \\
= \; & \frac{1}{\sqrt{2\pi}}\exp\left[\left(\mu_k + 2\omega_{ik}^{(1)}\alpha_2^{(k)}\sigma_k^2\right)^2 \frac{\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-1}}{2\sigma_k^2} - \alpha_2^{(k)}\omega^{2}{}_{ik}^{(1)} - \frac{\mu_k^2}{2\sigma_k^2}\right]\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-.5}
\end{aligned}
$$

Now let's compute $\eta_n\left(x, x_{n+1}, \omega_{n+1}^{(1)}\right)$:

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

where

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

4.2 Computation of ∇V_i

We have that

$$V_n\left(x, \omega^{(1)}\right) = \mathbb{E}_n\left[\max_{x'} a_{n+1}(x') \mid x_{n+1} = x, \omega_{n+1}^{(1)} = \omega^{(1)}\right] - \max_{x'} a_n(x')$$

where $a_n(x) := \mathbb{E}_n\left[\mathbb{E}\left[f\left(x, w^{(1)}, w^{(2)}\right)\right]\right] = \mathbb{E}_n\left[\mathbb{E}\left[F\left(x, w^{(1)}\right)\right]\right] = \mathbb{E}\left[\mu_n\left(x, w^{(1)}\right)\right]$. We need to discretize the domain of a_n and a_{n+1} to evaluate V_n. We choose some positive integer N and discretize the domain via a mesh with N parts in each dimension, obtaining $M = N^n$ points.

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

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

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

Then

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

where $a^n = (a_n(x_i))_{i=1}^M$,$\tilde{\sigma}_n\left(x,\omega^{(1)}\right) = \left(\tilde{\sigma}_n\left(x_i, x, \omega^{(1)}\right)\right)_{i=1}^M$, $h : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ defined by $h(a,b) = \mathbb{E}\left[\max_i a_i + b_i Z\right] - \max_i a_i$, 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 $a^n, \tilde{\sigma}_n\left(x, \omega^{(1)}\right)$. If $|a'| = 1$, $V_n\left(x,\omega^{(1)}\right) = h\left(a^n, \tilde{\sigma}_n\left(x,\omega^{(1)}\right)\right) = 0$ and so $\nabla V_n\left(x,\omega^{(1)}\right) = 0$. On the other hand, if $|a'| > 1$,

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

Then we only need to compute $\nabla b'_i$ for all i. Now,

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

where

$$
\begin{aligned}
\beta_1 &= \left[\Sigma_0\left(x_{n+1}, w_{n+1}^{(1)}, x_{n+1}, \omega_{n+1}^{(1)}\right) - \gamma^T A_n^{-1}\gamma\right]^{-1/2} \\
\beta_2 &= B\left(x, n+1\right) - \left[B\left(x,1\right) \;\; \cdots \;\; B\left(x,n\right)\right] A_n^{-1}\gamma \\
\gamma &= \begin{bmatrix} \Sigma_0\left(x_{n+1}, w_{n+1}^{(1)}, x_1, w_1^{(1)}\right) \\ \vdots \\ \Sigma_0\left(x_{n+1}, w_{n+1}^{(1)}, x_n, w_n^{(1)}\right) \end{bmatrix} \\
\nabla\left(\gamma^T\right) &= \left[\nabla\Sigma_0\left(x_{n+1}, w_{n+1}^{(1)}, x_1, w_1^{(1)}\right) \cdots \Sigma_0\left(x_{n+1}, w_{n+1}^{(1)}, x_n, w_n^{(1)}\right)\right] \\
B\left(x,i\right) &= \sigma_0^2 \exp\left(-\sum_{k=1}^{n} \alpha_1^{(k)}\left[x_k - x_{ik}\right]^2\right) \\
&\quad \prod_{k=1}^{d_1} \frac{1}{\sqrt{2\pi}} \exp\left[\left(\mu_k + 2\omega_{ik}^{(1)}\alpha_2^{(k)}\sigma_k^2\right)^2 \frac{\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-1}}{2\sigma_k^2} - \alpha_2^{(k)}\omega^{2(1)}_{ik} - \frac{\mu_k^2}{2\sigma_k^2}\right]\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-.5}
\end{aligned}
$$

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

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

and

$$
\begin{aligned}
\nabla_{x_{n+1,i}} B\left(x, n+1\right) &= -2\alpha_1^{(i)}\left(x_{n+1,i} - x_i\right) B\left(x, n+1\right) \\
\nabla_{w_{n+1,k}^{(1)}} B\left(x, n+1\right) &= B\left(x, n+1\right)\left[2\left(\mu_k + 2\omega_{n+1,k}^{(1)}\alpha_2^{(k)}\sigma_k^2\right)\left(2\alpha_2^{(k)}\sigma_k^2+1\right)^{-1}\alpha_2^{(k)} - 2\alpha_2^{(k)}\omega_{n+1,k}^{(1)}\right]
\end{aligned}
$$