Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Bayesian Optimization Allowing for Common Random Numbers E-Companion

Michael Pearce
Complexity Science, Warwick University, Coventry, UK m.a.l.pearce@warwick.ac.uk

Matthias Poloczek
Amazon, San Francisco, USA*, matpol@amazon.com

Juergen Branke
Warwick Business School, Coventry, UK, juergen.branke@wbs.ac.uk

---

* The work was done while Matthias was affiliated with the University of Arizona in Tucson, AZ.

E-Companion: Proofs and Additional Experiments

EC.1. Proofs of Statements

EC.1.1. Estimating the Target

The data collected and the surrogate model are over the domain $X \times \mathbb{N}^+$ whereas the target of optimization is a function over X. In what follows, we show how to derive an estimate for the target. This result is an immediate consequence of the symmetry of the model across unobserved seeds proven in Lemma EC.1. As a result of this symmetry, when taking the limit of the sum over infinite seeds, unobserved seeds dominate this sum yielding a simple form of the GP posterior for the target. This result is consistent with other CRN and non-CRN methods that do not make the seed explicit but do incorporate off-diagonal noise covariance matrix.

LEMMA **??** *For any given kernel over the domain $X \times \mathbb{N}^+$ that is of the form $k_{\bar{\theta}}(x,x') + \delta_{ss'} k_\epsilon(x,x')$, and a dataset of n input-output triplets D^n, the posterior over the target is a Gaussian process given by*

$$\bar{\theta}(x)|D^n \sim GP(\mu^n_{\bar{\theta}}(x), k^n_{\bar{\theta}}(x,x')), \tag{EC.1}$$

$$\mu^n_{\bar{\theta}}(x) = \mu^n(x,s'), \tag{EC.2}$$

$$k^n_{\bar{\theta}}(x,x') = k^n(x,s',x',s''), \tag{EC.3}$$

where $s', s'' \in \mathbb{N}^+ \setminus S^n$ with $s' \neq s''$ any two unobserved unequal seeds.

Recall that the Gaussian process is defined over the domain $X \times \mathbb{N}^+$, with infinite seeds. The following result states that the Gaussian process model makes identical predictions for all the unobserved seeds.

LEMMA EC.1. *Let $\theta(x,s)$ be a realization of a Gaussian Process with $\mu^0(x,s) = 0$ and any positive semi-definite kernel of the form $k(x,s,x',s') = k_{\bar{\theta}}(x,x') + \delta_{ss'} k_\epsilon(x,x')$. For all $x \in X$, $s_{obs} \in S^n$, and unobserved seeds $s, s', s'' \in \mathbb{N}^+ \setminus S^n$, the posterior mean and kernel satisfy*

$$\mu^n(x,s) = \mu^n(x,s'), \tag{EC.4}$$

$$k^n(x, s_{obs}, x', s) = k^n(x, s_{obs}, x', s'), \tag{EC.5}$$

$$k^n(x, s, x', s') = k^n(x, s, x', s'') = k^n(x, s', x', s''). \tag{EC.6}$$

Proof Writing out the posterior mean in full from Equation **??** in the main paper,

$$\begin{aligned}\mu^n(x,s) &= k^0(x,s,\tilde{X}^n)K^{-1}Y^n \\ &= \begin{cases} \left(k_{\bar{\theta}}(x,X^n) + (\mathbb{1}^\intercal_{s=S^n} \circ k_\epsilon(x,X^n))\right)K^{-1}Y^n & s \in S^n \\ k_{\bar{\theta}}(x,X^n)K^{-1}Y^n & s \in \mathbb{N}^+ \setminus S^n \end{cases}\end{aligned}$$

where $a \circ b$ is element-wise product and $\mathbb{1}_{s=S^n} \in \{0,1\}^n$ is a binary masking column vector that is zero for all $s \in \mathbb{N}^+ \setminus S^n$. The proofs for Equations EC.5 and EC.6 follow similarly from Equation **??** in the main paper. □

We next prove the main lemma. Keep in mind that the target for the optimization is the infinite average over seeds, and the Gaussian process model makes identical predictions for unobserved seeds. The infinite average is dominated by unobserved seeds with identical predictions. Hence we may simply use the prediction of any one unobserved seed as a model for the infinite average/target.

Proof of Lemma **??** The target of optimization, $\bar{\theta}(x)$, is given by the average output over infinitely many seeds which may be written as the limit

$$\bar{\theta}(x) \;=\; \lim_{N_s\to\infty} \frac{1}{N_s}\sum_{s=1}^{N_s} \theta(x,s). \tag{EC.7}$$

Adopting the shorthand $\mathbb{E}_n[...] = \mathbb{E}[...|D^n]$, we first consider the posterior expected performance,

$$\mathbb{E}_n[\bar{\theta}(x)] = \mathbb{E}_n\left[\lim_{N_s\to\infty} \frac{1}{N_s}\sum_{s=1}^{N_s} \theta(x,s)\right] \tag{EC.8}$$

$$= \lim_{N_s\to\infty} \frac{1}{N_s}\sum_{s=1}^{N_s} \mathbb{E}_n\left[\theta(x,s)\right] \tag{EC.9}$$

$$= \lim_{N_s\to\infty} \frac{1}{N_s}\sum_{s=1}^{N_s} \mu^n(x,s). \tag{EC.10}$$

Let $n_s = \max\{S^n\}$ be the largest observed seed. The sum of posterior means can be split into sampled seeds $s \in \{1,...,n_s\}$ and unsampled seeds $s \in \{n_s+1,....,N_s\}$,

$$\mathbb{E}_n[\bar{\theta}(x)] = \lim_{N_s\to\infty} \frac{1}{N_s}\left(\sum_{s=1}^{n_s} \mu^n(x,s) + \sum_{s'=n_s+1}^{N_s} \mu^n(x,s')\right) \tag{EC.11}$$

$$= \lim_{N_s\to\infty} \frac{1}{N_s}\left(\sum_{s=1}^{n_s} \mu^n(x,s) + (N_s - n_s)\mu^n(x,n_s+1)\right) \tag{EC.12}$$

$$= \lim_{N_s\to\infty} \frac{1}{N_s}\left(\sum_{s=1}^{n_s} \mu^n(x,s) - n_s\mu^n(x,n_s+1)\right) + \mu^n(x,n_s+1) \tag{EC.13}$$

$$= \mu^n(x,n_s+1), \tag{EC.14}$$

where we have used Lemma EC.1 to simplify. Similarly for the covariance, writing each $\bar{\theta}(x)$ term as the limit of a sum over seeds,

$$\mathbb{E}_n\left[\left(\bar{\theta}(x) - \mathbb{E}_n[\bar{\theta}(x)]\right)\left(\bar{\theta}(x') - \mathbb{E}_n[\bar{\theta}(x')]\right)\right] \tag{EC.15}$$

$$= \mathbb{E}_n\left[\left(\lim_{N_s\to\infty} \frac{1}{N_s}\sum_{s=1}^{N_s} \theta(x,s) - \mu(x,s)\right)\left(\lim_{N_t\to\infty} \frac{1}{N_t}\sum_{s'=1}^{N_t} \theta(x',s') - \mu(x',s')\right)\right] \tag{EC.16}$$

$$= \lim_{N_s,N_t\to\infty} \frac{1}{N_sN_t}\sum_{s,s'=1}^{N_s,N_t} \mathbb{E}_n\left[\left(\theta(x,s) - \mu(x,s)\right)\left(\theta(x',s') - \mu(x',s')\right)\right] \tag{EC.17}$$

$$= \lim_{N_s,N_t\to\infty} \frac{1}{N_sN_t}\sum_{s,s'=1}^{N_s,N_t} k^n(x,s,x's'). \tag{EC.18}$$

The domain in the limit of the summation, $\mathbb{N}^+ \times \mathbb{N}^+$, is unaffected by setting $N_t = N_s$. The summation decomposes into four terms,

$$\begin{aligned}\sum_{s,s'=1}^{N_s} k^n(x,s,x',s') = &\underbrace{\sum_{s,s'=1}^{n_s} k^n(x,s,x',s')}_{\text{observed seeds full covariance}} + \underbrace{\sum_{s'=n_s+1}^{N_s}\sum_{s=1}^{n_s} k^n(x,s,x',s')}_{\text{observed-unobserved covariance}} \\ &+ \underbrace{\sum_{s=n_s+1}^{N_s} k^n(x,s,x',s)}_{\text{unobserved seeds variance}} + \underbrace{\sum_{n_s<s\neq s'\leq N_s} k^n(x,s,x',s')}_{\text{unobserved seeds covariance}} \\ = &\underbrace{\sum_{s,s'=1}^{n_s} k^n(x,s,x',s')}_{\text{constant with } N_s} + \underbrace{2(N_s-n_s)\sum_{s=1}^{n_s} k^n(x,s,x',s')}_{\text{linear with } N_s} \\ &+ \underbrace{(N_s-n_s)k^n(x,s',x',s')}_{\text{linear with } N_s} + \underbrace{(N_s-n_s)^2k^n(x,s',x',s'')}_{\text{quadratic with } N_s},\end{aligned}$$

where s' and s'' are two unequal unobserved seeds. Dividing the final equation by N_s^2 and taking the limit $N_s \to \infty$, only the final term remains. □

Given the assumed kernel with independent and identically distributed difference functions, the average of infinitely many seeds includes finite observed seeds and infinitely many identical

unobserved seeds. Unobserved seeds dominate the infinite average and the performance under any unobserved seed is an estimator for the objective function. Likewise the posterior covariance between infinite averages is the posterior covariance between any two unique unobserved seeds. Also note that the kernel of the Gaussian process prior for the objective evaluated at different seeds returns the prior kernel for the target $\bar{k}^0(x,x') = k^0(x,1,x',2) = k_{\bar{\theta}}(x,x')$ as desired.

EC.1.2. Proof of Theorem ??

We next show that, under certain assumptions on the target function, given an infinite sampling budget, $N \to \infty$, the $\mathrm{KG}^{\mathrm{CRN}}$ algorithm will discover the true optimum. We first restate the result.

THEOREM **??** *Let $x_r^N \in A$ be the point that KG^{CRN} recommends in iteration N. For each $p \in [0,1)$ there is a constant K_p such that with probability p*

$$\lim_{N\to\infty} \bar{\theta}(x_r^N) > \bar{\theta}(x^{OPT}) - K_p d.$$

We first prove properties of the $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ function and then consider the error due to discretization.

Lemma EC.2 ensures the GP model exists in the limit of infinite data. We then show that $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ is non-negative in Lemma EC.3 and that it is zero for sampled input pairs in Lemma EC.4. We then show that if a single x is sampled for infinitely many (not necessarily consecutive) seeds, again $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ tends to zero also for all unevaluated seeds in Lemma EC.5. Then in Lemma EC.6 we show the opposite direction, if $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ is zero, this implies that the peak of the target prediction will not change by sampling (x,s). This is extended in Lemma EC.7 that states that if for a new seed s, $\mathrm{KG}_n^{\mathrm{CRN}}(x,s) = 0$ for all X then no more samples will change the peak prediction of the target and the true peak is known when X is a discrete set.

The error due to discretization relies on the assumption of a differentiable GP kernel, such as Matérn, and using a Lipschitz continuity argument, the error may be bounded proving Theorem **??**. The following result simply states that the GP model exists in the limit of infinite data. First we define $V^n(x,x') = \mathbb{E}_n[\bar{\theta}(x)\bar{\theta}(x')]$.

LEMMA EC.2. *Let $x, x' \in X$. Then the limits of the series $(\bar{\mu}^n(x))_n$ and $(V^n(x,x'))_n$ exist and are denoted by $\bar{\mu}^\infty(x)$ and $V^\infty(x,x')$, respectively. Then we have*

$$\lim_{n\to\infty} \bar{\mu}^n(x) = \bar{\mu}^\infty(x) \tag{EC.19}$$

$$\lim_{n\to\infty} V^n(x,x') = V^\infty(x,x') \tag{EC.20}$$

almost surely.

Proof $\bar{\theta}(x)$ and $\bar{\theta}(x)\bar{\theta}(x')$ are integrable random variables for all $x, x' \in X$ by choice of $\bar{\theta}$. Proposition 2.7 in **?** states that any sequence of conditional expectations of an integrable random variable under an increasing filtration is uniformly integrable martingale. Thus, both sequences converge almost surely to their respective limit. □

The next result states the $\mathrm{KG}^{\mathrm{CRN}}(x,s)$ is non-negative for all input pairs.

LEMMA EC.3. *$KG_n^{CRN}(x,s) \geq 0$ holds for all $(x,s) \in X \times \mathbb{N}^+$.*

Proof Adopting the shorthand $x_r^n = \mathrm{argmax}_{x\in X}\mu^n(x,0)$, we may write $\max_x \mu^n(x,0) = \mu^n(x_r^n,0)$ and

$$\begin{aligned}
\mathrm{KG}_n^{\mathrm{CRN}}(x,s) &= \mathbb{E}\left[\max_{x'\in X}\{\mu^n(x',0) + \tilde{\sigma}^n(x',0;x,s)Z\} - \mu^n(x_r^n,0)\right] \\
&= \mathbb{E}\left[\max_{x'\in X}\{\mu^n(x',0) + \tilde{\sigma}^n(x',0;x,s)Z\} - \mu^n(x_r^n,0)\right] - \underbrace{\mathbb{E}[cZ]}_{=0} \\
&= \mathbb{E}\left[\max_{x'\in X}\{\mu^n(x',0) + (\tilde{\sigma}^n(x',0;x,s) - c)Z\} - \mu^n(x_r^n,0)\right]
\end{aligned}$$

where the expectation is over $Z \sim N(0,1)$ and c is an arbitrary constant. In particular, by setting $c = \tilde{\sigma}^n(x_r^n,0;x,s)$, the inner expression, when evaluated at $x_r^n \in X$, satisfies

$$\mu^n(x_r^n,0) + (\tilde{\sigma}^n(x_r^n,0;x,s) - \tilde{\sigma}^n(x_r^n,0;x,s))Z - \mu^n(x_r^n,0) = 0$$

for all $Z \in \mathbb{R}$ and

$$\max_{x'\in X}\{\mu(x',0) + (\tilde{\sigma}^n(x',0;x,s) - c)Z - \mu_0\} \;\geq\; \mu(x_r^n,0) + (\tilde{\sigma}^n(x_r^n,0;x,s) - c)Z - \mu^n(x_r^n,0) \;=\; 0$$

for all Z and $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ may be written as the expectation of a non-negative random variable.

□

The following result states that once an input pair (x,s) has been observed, $\theta(x,s)$ is known and $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ is zero. Combined with the result that $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ is non-negative, it follows that observed input pairs (x,s) are minima of the function $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$.

LEMMA EC.4. *Given deterministic simulation outputs, there is no improvement in re-sampling a sampled point.*

$$KG_n^{CRN}(x^i,s^i)=0$$

for all $(x^i,s^i)\in\tilde{X}^n$.

Proof The posterior covariance between the output at any point and the output at an observed point is zero, writing out the full matrix multiplication for the posterior kernel and simplifying yields

$$\begin{aligned}
k^n(x^i,s^i;x,s) &= k^0(x^i,s^i;x,s)-k^0(x^i,s^i;\tilde{X}^n)\left(k^0(\tilde{X}^n;\tilde{X}^n)\right)^{-1}k^0(\tilde{X}^n;x,s)\\
&= k^0(x^i,s^i;x,s)-\left[k^0(\tilde{X}^n;\tilde{X}^n)\right]_i\left(k^0(\tilde{X}^n;\tilde{X}^n)\right)^{-1}k^0(\tilde{X}^n;x,s)\\
&= k^0(x^i,s^i;x,s)-\mathbb{1}_i^{n\intercal}k^0(\tilde{X}^n;x,s)\\
&= k^0(x^i,s^i;x,s)-k^0(x^i,s^i;x,s)\\
&= 0
\end{aligned}$$

where $[\cdot]_i$ is the i^{th} row. The second line contains the i^{th} row of a matrix multiplied by its inverse returning the i^{th} row of the identity matrix denoted $\mathbb{1}_i^{n\intercal}$. Therefore $\tilde{\sigma}^n(x,s;x^i,s^i)=0$ for all (x,s) and $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)=0$. $\square$

Let ω denote an arbitrary sample path, $\omega=((x,s)^1,(x,s)^2,.....)$, determining an input pair for each query to the objective as $n\to\infty$. Lemmas EC.3 and EC.4 imply sampled point inputs are minima of $\mathrm{KG}^{\mathrm{CRN}}$ and recall that according to the algorithm, new samples are allocated to maxima $(x,s)^{n+1}=\mathrm{argmax}\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$. These facts together imply that no input (x,s) will be sampled more than once. We need only to consider sample paths ω where all sampled inputs pairs (x^i,s^i) are unique. Recall that we suppose a (finite) discretization of X, thus there must be an $x\in X$ that

is observed for an infinite number of seeds on ω as $n \to \infty$. We study the asymptotic behaviour $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ for $n \to \infty$ as a function of $\mu^n(x,0)$, $\tilde{\sigma}^n(x',0,x,s)$.

If s is a new seed and x has been observed for infinitely many seeds, the next result states that $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ tends to zero, there is less/no value in re-evaluating x for another new seed.

LEMMA EC.5. *If x is sampled for infinitely many (not necessarily consecutive) seeds, then $\tilde{\sigma}^\infty(x',0;x,s) = 0$ for all $x' \in X$ and all $s \in \mathbb{N}^+$ and $KG_\infty^{CRN}(x,s) = 0$ for all $s \in \mathbb{N}^+$ almost surely.*

Proof Setting $x^{n+1} = x$ and assuming (x^i, s^i) pairs are arranged such that s^{n+1} is always a new seed, the posterior variance reduces to zero

$$\begin{aligned}
\lim_{n\to\infty} |\tilde{\sigma}^n(x',0;x,s^{n+1})| &= \lim_{n\to\infty} \frac{|k^n(x',0,x,s^{n+1})|}{\sqrt{k^n(x,s^{n+1},x,s^{n+1})}} \\
&= \lim_{n\to\infty} \frac{\bar{k}^n(x',x)}{\sqrt{\bar{k}^n(x,x)+k_\epsilon(x,x)}} \\
&\leq \lim_{n\to\infty} \sqrt{\bar{k}^n(x',x')} \frac{\sqrt{\bar{k}^n(x,x)}}{\sqrt{\bar{k}^n(x,x)+k_\epsilon(x,x)}} \\
&= 0
\end{aligned}$$

where the final line is by noting that $\bar{k}^n(x,x) + k_\epsilon(x,x) > 0$ for all n and x. $\square$

The following result states that if there is no benefit of a new measurement for an input pair (x,s), then the change in the posterior mean, $\tilde{\sigma}^n(x',0;x,s)$ must be constant, i.e. the new sample at (x,s) will only have the effect of adding a constant to the prediction of the target, hence learning nothing about the peak of the target. The contrapositive is that for input points for which $\tilde{\sigma}^n(x',0;x,s)$ varies with x', $\mathrm{KG}^{\mathrm{CRN}}$ is strictly positive.

LEMMA EC.6. *Let (x,s) be an input pair for which $KG_n^{CRN}(x,s) = 0$. Then for all $x' \in X$*

$$\tilde{\sigma}^n(x',0;x,s) = c$$

where c is a constant.

Proof From Equation EC.21, $\mathrm{KG}_n^{\mathrm{CRN}}(x,s)$ can be written as the expectation of a non-negative random variable. Therefore the random variable itself must equate to zero almost surely implying

$$\begin{aligned}
\max_{x'\in X}\{\mu(x',0) + (\tilde{\sigma}(x',0;x,s) - c)Z - \mu^n(x_r^n,0)\} &= 0 \\
\max_{x'\in X}\{\mu(x',0) + (\tilde{\sigma}(x',0;x,s) - c)Z\} &= \max_{x''\in X}\{\mu(x'',0)\}
\end{aligned}$$

for all $Z \in \mathbb{R}$. This implies $\tilde{\sigma}(x', 0; x, s) = c$ for all $x' \in X$. $\square$

Note that in the case where $(x, s) \in \tilde{X}^n$ we have that $\tilde{\sigma}^n(x', 0; x, s) = 0$ for all $x' \in X$.

We next show that if there is no value in evaluating any input pair, then the optimizer of the target is known.

LEMMA EC.7. *Let $s \in \mathbb{N}^+ \setminus S^n$ be an unobserved seed, if $KG_n^{CRN}(x, s) = 0$ for all $x \in X$, then $argmax_x \mu^n(x, 0) = argmax_x \bar{\theta}(x)$*

Proof By Lemma EC.6, we have that $\bar{k}^n(x, x') = c$ for all $x, x' \in X$ and the covariance matrix $\bar{k}^n(X, X)$ is proportional to the all ones matrix. Hence $\bar{\theta}(x) - \mu^n(x, 0)$ is a normal random variable that is constant across all $x \in X$ and $\operatorname{argmax}_{x \in X} \mu(x, 0) = \operatorname{argmax}_{x \in X} \bar{\theta}(x)$ holds. $\square$

Lemmas EC.4, EC.5, consider evaluating $\mathrm{KG}^{\mathrm{CRN}}$ as the sampling budget increases in a specific way. More generally, recall that $\mathrm{KG}^{\mathrm{CRN}}$ picks $(x, s)^{n+1} \in \operatorname{argmax} \mathrm{KG}_n^{\mathrm{CRN}}(x, s)$ in each iteration n. Since $\theta(x, \cdot)$ is evaluated infinitely often (by choice of x), $\mathrm{KG}_n^{\mathrm{CRN}}(x, \cdot) \to 0$ for all $x \in A$ holds almost surely and by Lemma EC.7 the true optimizer is known.

Proof of Theorem 1 We give a bound on the loss due to discretization of a continuous search space. Suppose that $X \subset \mathbb{R}^d$ is compact and $A \subset X$ is a finite set of discretization points. Suppose that $\bar{\mu}^0(x) = 0$ for all x and that $k_{\bar{\theta}}(x, x')$ is a four times differentiable Matérn kernel, e.g., the popular squared exponential kernel. Moreover, suppose that $\bar{\theta}(x)$ is drawn from the prior, i.e., let $\bar{\theta}(x) \sim \mathrm{GP}(\bar{\mu}^0(x), k_{\bar{\theta}}(x, x'))$, then the sample $\bar{\theta}(x)$ from the set of functions is itself twice *continuously differentiable* in X with probability one (**?**). The extrema of $\frac{\delta}{\delta x_i} \bar{\theta}(x)$ over X are probabilistically bounded, since the derivative processes of $\bar{\theta}(x)$ are also Gaussian for our choice of $k_{\bar{\theta}}^0(x, x)$ (**?**). Let $x^{OPT} = \operatorname{argmax}_{x \in X} \bar{\theta}(x)$ and $d = \max_{x' \in X} \min_{x \in A} \operatorname{dist}(x, x')$ be the largest distance from any point in the continuous domain X to its nearest neighbor in A. We can compute for every $p \in [0, 1)$ a constant K_p such that $\bar{\theta}(x)$ is K_p Lipschitz continuous on X with probability at least p, thus there exists an $\bar{x} \in A$ with $\operatorname{dist}(\bar{x}, x^{OPT}) \leq d$ and

$$\bar{\theta}(\bar{x}) > \bar{\theta}(x^{OPT}) - K_p d$$

holds with probability p. Finally the point recommended by KG^{CRN} is the maximizer of $x_r^N \in \text{argmax}_{x\in A}\bar{\theta}(x)$ and therefore is not worse than $\bar{x}$

$$\begin{aligned}\lim_{N\to\infty} \bar{\theta}(x_r^N) &\geq \bar{\theta}(\bar{x}) \\ &\geq \bar{\theta}(x^{OPT}) - K_p d\end{aligned}$$

□

Thus, when applying the KG^{CRN} algorithm to a discretized search space, the true optimizer becomes known as the sampling budget increases without bound and if the underlying target function is continuous, the error is bounded simply due to Lipschitz continuity.

EC.1.3. Behavior in the Compound Spheric Case

We next provide proofs for Lemma **??** and Lemma **??** relating to the KG^{CRN} algorithm behaviour in the case of compound sphericity with full noise correlation. Recall this corresponds to the difference functions reducing to constant offsets and an algorithm may optimize one seed as a single seed is a deterministic function with the same optimizer as the target. This is essentially a best-case scenario for optimization with common random numbers. Lemma **??** of the main paper states that the difference $\mu^n(x,s) - \mu^n(x,s') = A_s - A_{s'}$ is constant for all x. Likewise the same relationship applies to $\tilde{\sigma}^n(x',s';x,s)$ that quantifies changes in the posterior mean and therefore must also maintain the symmetry over seeds s'. All results in this section assume $\theta(x,s)$ is a realization of a Gaussian process with the compound spheric kernel and full correlation $k_\epsilon(x,x') = \eta^2$.

The first result states that, when sampling a point (x,s), the update in the prediction for one seed differs from the update in prediction for another seed by an additive constant. Predictions for all seeds have the same shape/gradient and differ only by global constants.

LEMMA EC.8. *Let $x,x' \in X$, $s,s' \in \mathbb{N}^+$, then the difference in posterior mean updates satisfies*

$$\tilde{\sigma}^n(x',s';x,s) = \tilde{\sigma}^n(x',0;x,s) + h^n(s',x,s).$$

Proof

$$\begin{aligned}
\tilde{\sigma}^n(x',s';x,s) &= \frac{k^n(x',s';x,s)}{\sqrt{k^n(x,s,x,s)}} \\
&= \frac{1}{\sqrt{k^n(x,s,x,s)}}\left(k_{\bar{\theta}}(x',x)+\eta^2\delta_{ss'}-\left(k_{\bar{\theta}}(x',X^n)+\eta^2\mathbb{1}^\intercal_{s'=S^n}\right)K^{-1}\left(k_{\bar{\theta}}(X^n,x)+\eta^2\mathbb{1}_{s=S^n}\right)\right) \\
&= \tilde{\sigma}^n(x',0;x,s)+\underbrace{\frac{\eta^2\delta_{ss'}-\eta^2\mathbb{1}^\intercal_{s'=S^n}K^{-1}\left(k_{\bar{\theta}}(X^n,x)+\eta^2\mathbb{1}_{s=S^n}\right)}{\sqrt{k^n(x,s,x,s)}}}_{\text{independent of } x'} \\
&= \tilde{\sigma}^n(x',0;x,s)+h^n(s',x,s)
\end{aligned}$$

□

As a result of the symmetry over seeds it is possible to use any seed $s\in\mathbb{N}^+$ as the target of optimization formalized in the following Lemma.

LEMMA EC.9. *Let $x\in X$, $s,s'\in\mathbb{N}^+$, then*

$$KG_n^{CRN}(x,s)=\mathbb{E}[\max_{x'\in X}\mu^n(x',s')+\tilde{\sigma}^n(x',s';x,s)Z-\max_{x''\in X}\mu^n(x'',s')].$$

Proof

$$\begin{aligned}
\mathrm{KG}_n^{\mathrm{CRN}}(x,s) &= \mathbb{E}[\max_{x'\in X}\mu^n(x',0)+\tilde{\sigma}^n(x',0;x,s)Z-\max_{x''\in X}\mu^n(x'',0)] \\
&= \mathbb{E}[\max_{x'\in X}\mu^n(x',s')-A_{s'}+(\tilde{\sigma}^n(x',s';x,s)-h(s',x,s))Z-\max_{x''\in X}\mu^n(x'',s')-A_{s'}] \\
&= \mathbb{E}[\max_{x'\in X}\mu^n(x',s')+\tilde{\sigma}^n(x',s';x,s)Z-\max_{x''\in X}\mu^n(x'',s')]-h(s',x,s)\mathbb{E}[Z] \\
&= \mathbb{E}[\max_{x'\in X}\mu^n(x',s')+\tilde{\sigma}^n(x',s';x,s)Z-\max_{x''\in X}\mu^n(x'',s')]
\end{aligned}$$

We next prove Lemma **??** from the main paper: if there are finite solutions X and all have been evaluated on a common seed, then the value of sampling any solution on any seed is zero.

LEMMA **??** *Let $X=\{x_1,...,x_d\}$ and $\tilde{X}^n=\{(x_1,1),...,(x_d,1)\}$ then for all $(x,s)\in X\times\mathbb{N}^+$*

$$KG_n^{CRN}(x,s)=0$$

and the maximizer $\arg\max_x\bar{\theta}(x)$ *is known.*

Proof Lemma EC.9 shows that any seed can be used as the target of optimization. Therefore we may choose $s=1$ as the target. All x have been sampled for $s=1$ therefore $\tilde{\sigma}^n(x,1;x',s')=0$ for all $x\in X$ and $s'\in\mathbb{N}^+$. Hence

$$\begin{aligned}\mathrm{KG}^{\mathrm{CRN}}(x,s) &= \mathbb{E}[\max_{x'\in X}\mu^n(x',1)+0Z-\max_{x''\in X}\mu^n(x'',1)]\\ &= 0\end{aligned}$$

for all $x,s\in X\times\mathbb{N}^+$. By Lemma EC.7 the maximizer $\mathrm{argmax}_{x\in X}\bar{\theta}(x)$ is known (although it's underlying value, $\max\bar{\theta}(x)$, is not known). □

We next prove Lemma **??** from the main paper that $\mathrm{KG}^{\mathrm{CRN}}$, when evaluated as in **?**, never samples a new seed and reduces to Expected Improvement (EI) of **?**.

LEMMA **??** *Let $X\subset\mathbb{R}^d$ be a set of possible solutions, $\tilde{X}^n=\{(x^1,1),...,(x^n,1)\}$ be the set of sampled input pairs and $X^n=(x^1,...,x^n)$. Define*

$$KG_n^{CRN}(x,s;A)=\mathbb{E}\left[\max_{x'\in A\cup\{x\}}\mu^{n+1}(x',0)-\max_{x'\in A\cup\{x\}}\mu^n(x',0)\middle|D^n,(x,s)^{n+1}=(x,s)\right].$$

Then for all $x\in X$

$$KG_n^{CRN}(x,1;X^n)>KG_n^{CRN}(x,2;X^n)$$

and therefore $\max_x KG_n^{CRN}(x,1;X^n)>\max_x KG_n^{CRN}(x,2;X^n)$ and seed $s=2$ will never be evaluated. Further

$$KG_n^{CRN}(x,1;X^n)=\mathbb{E}\big[\max\{0,y^{n+1}-\max Y^n\}\big|D^n,x^{n+1}=x,s^{n+1}=1\big].$$

Proof By Lemma EC.9, we may set $s=1$ as the target of optimization. For all sampled points $i=1,...,n$, we have that $\tilde{\sigma}^n(x^i,1;x,s)=0$ and $\mu^n(x^i,1)=y^i$ therefore $\max\mu^n(\tilde{X}^n)=\max Y^n$. Define $\bar{Y}^n=\max Y^n$. The expression for Knowledge Gradient becomes

$$\begin{aligned}\mathrm{KG}_n^{\mathrm{CRN}}(x,s;X^n) &= \mathbb{E}\big[\max\{\bar{Y}^n,\mu^n(x,1)+\tilde{\sigma}^n(x,1;x,s)Z\}\big]-\max\{\bar{Y}^n,\mu^n(x,1)\}\\ &= \mathbb{E}\big[\max\{0,\mu^n(x,1)+\tilde{\sigma}^n(x,1;x,s)Z-\bar{Y}^n\}\big]\\ &= \Delta(x)\Phi\left(\frac{\Delta(x)}{|\tilde{\sigma}^n(x,1;x,s)|}\right)-|\tilde{\sigma}^n(x,1;x,s)|\phi\left(\frac{\Delta(x)}{|\tilde{\sigma}^n(x,1;x,s)|}\right)\\ &= f\big(\Delta(x),\ |\tilde{\sigma}^n(x,1;x,s)|\big)\end{aligned}$$

where $\Phi(\cdot), \phi(\cdot)$ are cumulative and density functions of the Gaussian distribution, $\Delta(x) = \mu^n(x,1) - \bar{Y}^n$ and $f(a,b)$ is the well known expected improvement acquisition function derived from the expectation of a truncated Gaussian random variable. Note that the function $f(a,b)$ is monotonically increasing in b, $\frac{d}{db}f(a,b) = \phi(-a/b) > 0$. Hence, to prove the lemma, it is sufficient to show $|\tilde{\sigma}^n(x,1;x,1)| > |\tilde{\sigma}^n(x,1;x,2)|$ for all $x \in X$. Firstly we may simplify $\tilde{\sigma}^n(x,1;x,1)$ as follows

$$\tilde{\sigma}^n(x,1;x,1) = k^n(x,1,x,1)/\sqrt{k^n(x,1,x,1)} \tag{EC.21}$$

$$= \sqrt{k^n(x,1,x,1)}. \tag{EC.22}$$

Substituting this into the inequality yields

$$\begin{aligned}
|\tilde{\sigma}^n(x,1;x,1)| &> |\tilde{\sigma}^n(x,1;x,2)| \\
\sqrt{k^n(x,1,x,1)} &> \frac{|k^n(x,1,x,2)|}{\sqrt{k^n(x,2,x,2)}} \\
1 &> \frac{|k^n(x,1,x,2)|}{\sqrt{k^n(x,2,x,2)k^n(x,1,x,1)}} \\
-1 &< \operatorname{corr}(\theta(x,1),\theta(x,2)|D^n) \leq 1
\end{aligned}$$

where the last line is true by the positive semi-definiteness of the kernel, the correlation between two random variables cannot be greater than one. The above result demonstrates that allocating samples according to $\mathrm{KG}^{\mathrm{CRN}}$ will always sample seed $s = 1$. The target is stochastic however the objective is deterministic and the new output $y^{n+1} \sim N(\mu^n(x,1), k^n(x,1,x,1))$. The acquisition function simplifies to

$$\begin{aligned}
\mathrm{KG}_n^{\mathrm{CRN}}(x,1;X^n) &= \mathbb{E}\big[\max\{0, \mu^n(x,1) + \sqrt{k^n(x,1,x,1)}Z - \bar{Y}^n\}\big] \\
&= \mathbb{E}\big[\max\{0, y^{n+1} - \bar{Y}^n\}\big|D^n, x^{n+1} = x, s^{n+1} = 1\big]
\end{aligned}$$

where the last line is exactly the EI acquisition criterion of **?**. □

EC.1.4. Suboptimality of $\mathrm{KG}^{\mathrm{PW}}$

Finally we prove that the Value of Information achieved by $\mathrm{KG}^{\mathrm{PW}}$ is less than that of $\mathrm{KG}^{\mathrm{CRN}}$, given the same set of observations and Gaussian process models.

LEMMA **??** *Let D^n be a dataset of observation triplets. For a Gaussian process with a kernel of the form $k_{\bar{\theta}}(x,x') + \delta_{ss'}k_{\epsilon}(x,x')$, the expected increase in value after two steps allocated according to KG^{CRN} is at least as big as two steps allocated according to KG^{PW},*

$$\begin{aligned}
&\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|(x,s)^{n+1},(x,s)^{n+2} \sim KG^{CRN}\right] \\
\geq\ &\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|(x,s)^{n+1},(x,s)^{n+2} \sim KG^{PW}\right]
\end{aligned}$$

Proof The suboptimality of one or two steps of the serial mode of $\mathrm{KG}^{\mathrm{PW}}$ is clear by noting it is constrained to a new seed, a subset of the same acquisition space considered by $\mathrm{KG}^{\mathrm{CRN}}$ as mentioned in Equation (**??**). We focus on the suboptimality of one step of the batch mode

$$\begin{aligned}
&\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|(x,s)^{n+1},(x,s)^{n+2} \sim \mathrm{KG}^{\mathrm{CRN}}\right] \\
=\ &\max_{(x,s)^{n+1}}\mathbb{E}_n\left[\max_{(x,s)^{n+2}}\mathbb{E}_{n+1}\left[\max_{x'}\mu^{n+2}(x',0)\middle|(x,s)^{n+2}\right] - \max_{x''}\mu^{n}(x'',0)\middle|(x,s)^{n+1}\right] \\
\geq\ &\max_{x^{n+1}}\mathbb{E}_n\left[\max_{x^{n+2}}\mathbb{E}_{n+1}\left[\max_{x'}\mu^{n+2}(x',0)\middle|x^{n+2}\right] - \max_{x''}\mu^{n}(x'',0)\middle|x^{n+1}, s^{n+1}=s^{n+2}=n+1\right] \quad \text{(EC.23)} \\
\geq\ &\max_{x^{n+1},x^{n+2}}\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|x^{n+1},x^{n+2},s^{n+1}=s^{n+2}=n+1\right] \quad \text{(EC.24)} \\
\geq\ &\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|(x^{n+1},x^{n+2}) = \arg\max \mathrm{KG}_n^{\mathrm{PW}}(x,x'), s^{n+1},s^{n+2}=n+1\right] \\
=\ &\mathbb{E}_n\left[\max_{x'}\mu^{n+2}(x',0) - \max_{x''}\mu^{n}(x'',0)\middle|(x,s)^{n+1},(x,s)^{n+2} \sim \mathrm{KG}^{\mathrm{PW}}\right]
\end{aligned}$$

where the first inequality is due to constraining the acquisition space to a new seed, the second is by Jensen's inequality and the convexity of the max operator implying sub-optimality due to batch pre-allocation, and the third inequality is due to the approximation with differences used in $\mathrm{KG}^{\mathrm{PW}}$ as pairs are not allocated to maximize the true batch value. □

Sequentially allocating two singles to the same new seed is guaranteed to have higher value per sample than a corresponding batch mode pre-allocating a pair to a single seed as shown by Equations (EC.23) and (EC.24). However the serial and batch mode of $\mathrm{KG}^{\mathrm{PW}}$ compute the value over different subsets of the full acquisition space and therefore the batch mode can return higher value per sample.

EC.2. Further Experimental Results

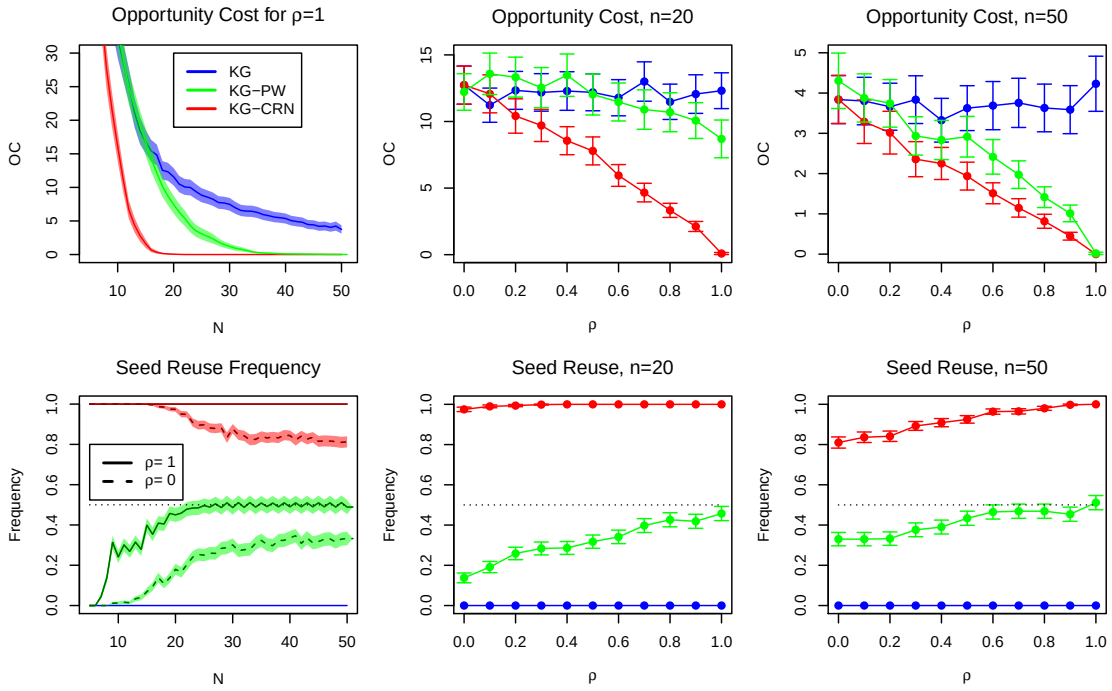

Figure EC.1 Results for synthetic data where the objective function was drawn from the algorithm's GP model with offsets and white noise ($\eta, \sigma_w \geq 0, \sigma_b = 0$) only. We let $\rho = \eta^2/(\eta^2+\sigma_w^2)$ and hold $\eta^2+\sigma_w^2 = 50^2$ constant. For low ρ, all algorithms perform similarly. As ρ increases, KG^{CRN} samples more old seeds and outperforms other methods. We observe that KG^{PW} samples singletons initially, hence its performance is comparable to KG in that phase. Later the performance of KG^{PW} improves over KG and when it samples correlated doubles later improving upon KG.

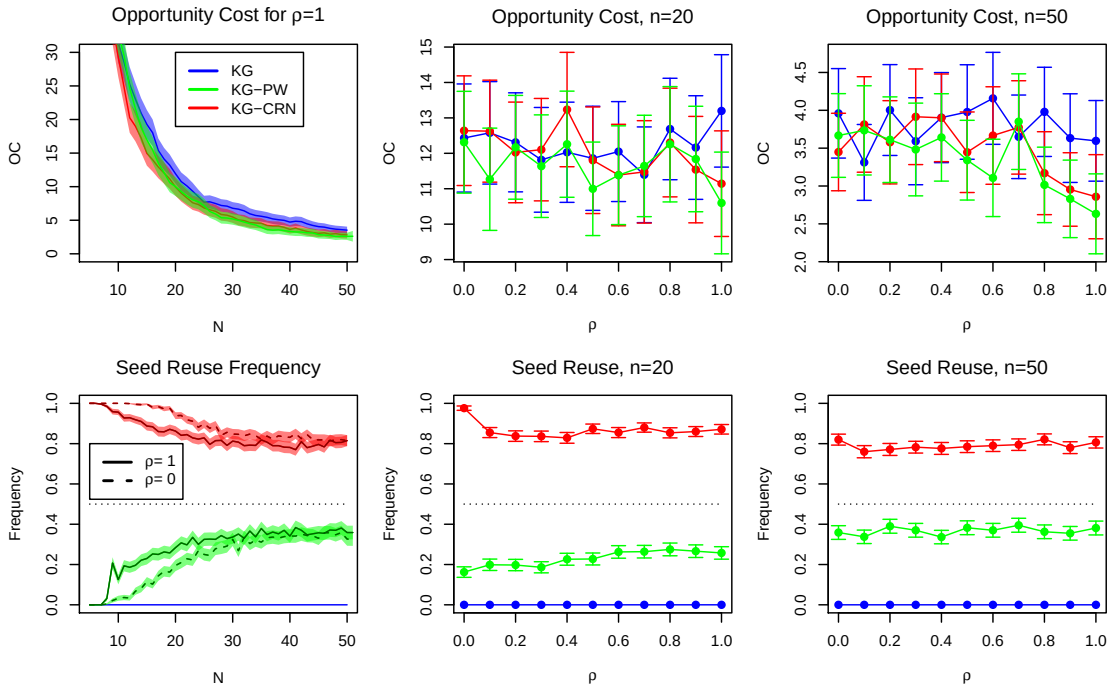

Figure EC.2 More results on GP synthetic data that was generated with $\eta^2 = 0$ and $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$, holding $\sigma_b^2 + \sigma_w^2 = 50^2$ constant. We do not observe a significant benefit from bias functions alone in the case of no offsets.

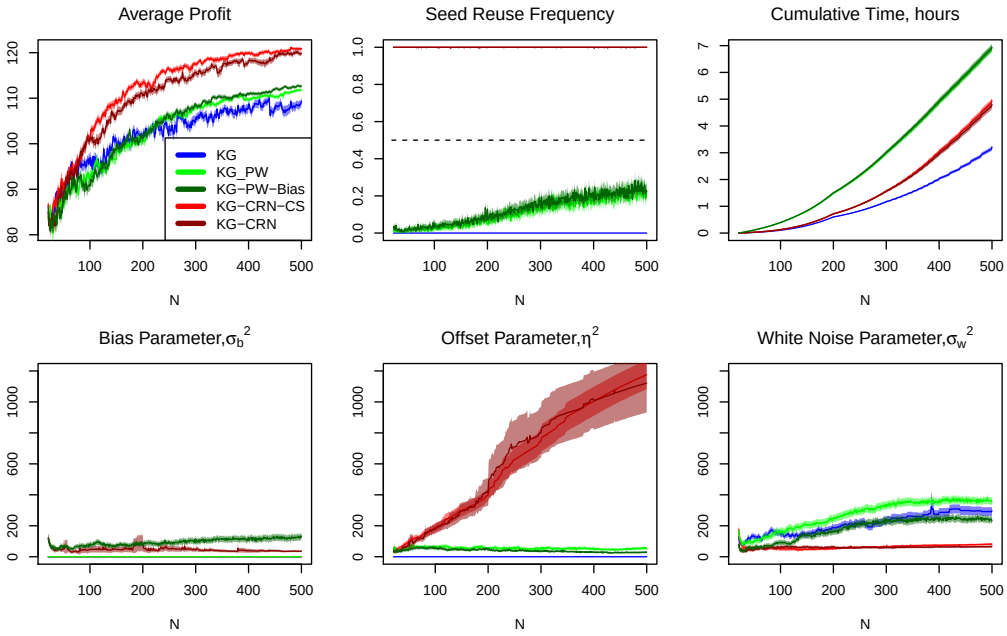

Figure EC.3 ATO results. The KG^{PW} algorithm samples singletons early on, and never learns a large offset parameter η^2. KG^{CRN} samples old seeds and eventually learns a large offset parameter and never samples any new seeds. KG has the smallest running time, followed by KG^{CRN} variants and the KG^{PW} variants.

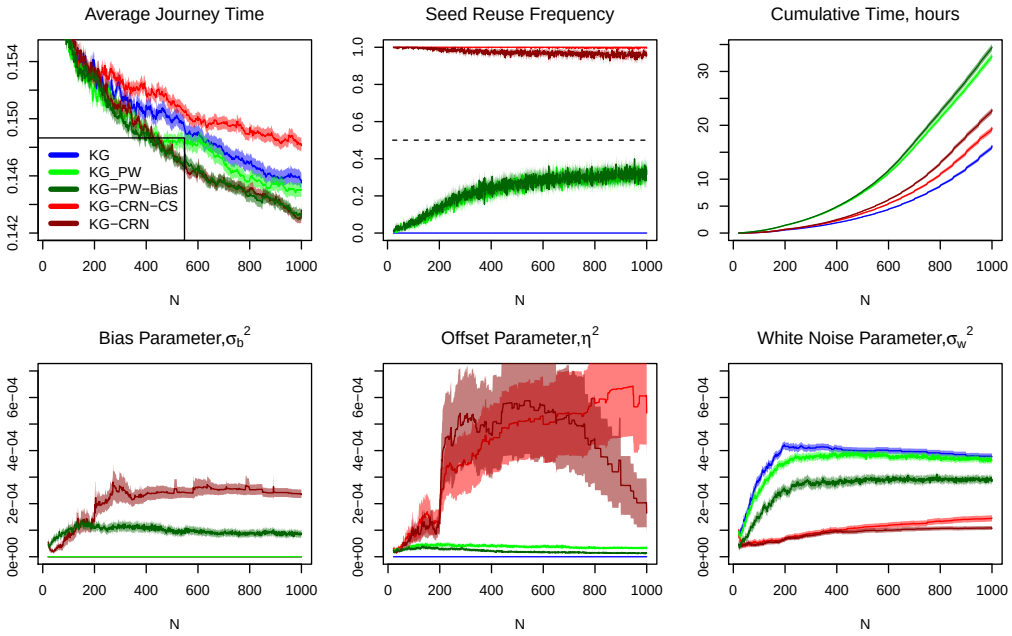

Figure EC.4 Ambulances in a square problem (AIS). The bias functions provide significant benefit to both KG^{CRN} and KG^{PW}. Excluding bias functions, $\text{KG}^{\text{CRN}} - CS$, leads to inefficient sampling of old seeds only. The KG^{CRN} variants learn larger offset parameters and require less computation time.

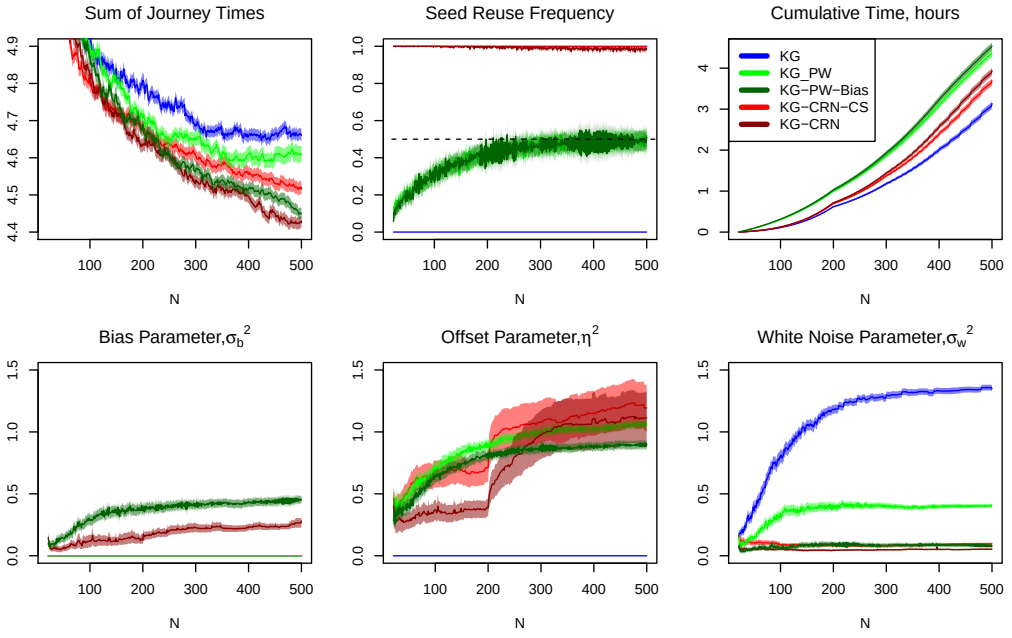

Figure EC.5 The AIS problem with the sum of journey times in a simulation as the objective. KG^{CRN} variants improve performance over KG, and bias functions improve performance over compound spheric variants. All methods reuse seeds to the extent that this is possible for them.

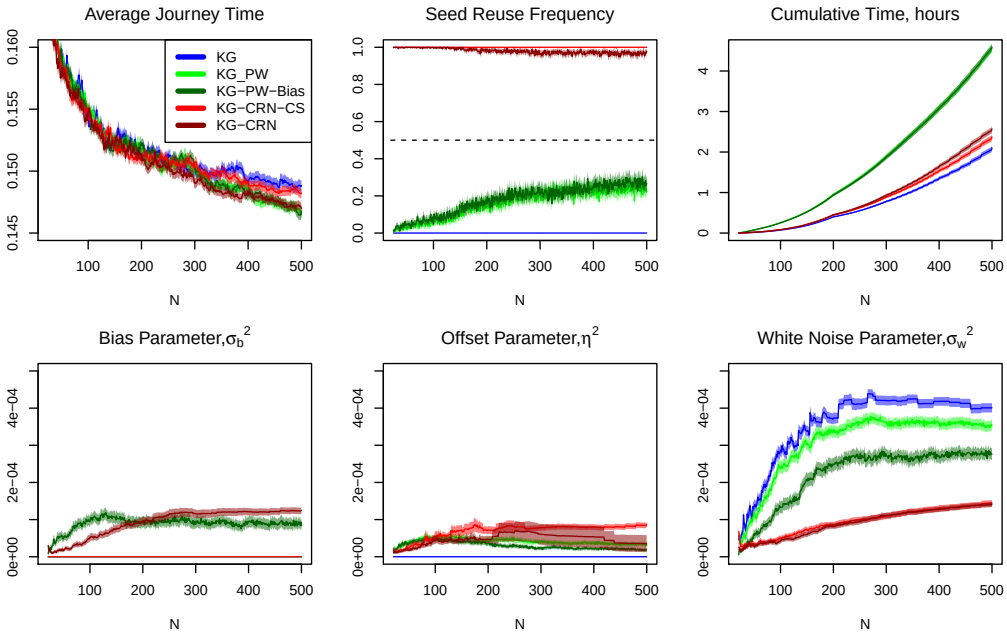

Figure EC.6 All algorithm variants perform similarly. The offset and bias parameters are considerably lower than the white noise parameter, suggesting there is little exploitable structure in the noise for this problem.

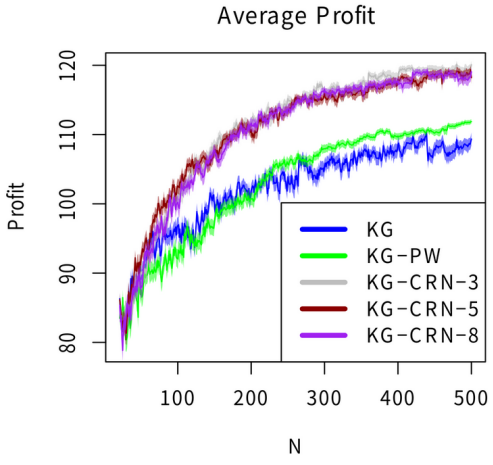

Figure EC.7 The ATO benchmark where the KG^{CRN} algorithm is initialized with 20 points spread over 3, 5, or 8 seeds. In all cases, the KG^{CRN} algorithm converges to exactly the same level of profit.

EC.3. Response Surfaces for ATO and Ambulances Problem

To understand the impact of the random seed on some real-world problems, Figure EC.8 depicts (empirically estimated) $\bar{\theta}(x)$ and $\theta(x,s)$ for each problem along a linear path within the higher dimensional solution space. For the ATO problem, the objective for each seed and the target have almost identical shape differing mostly by additive offsets, thus seed peaks align almost exactly with the target peak and CRN will be beneficial. For the Ambulance problem, each seed has an offset as well as more unique variation, the peaks of seeds are approximately similar to the target and CRN may be beneficial but less so than ATO.

To demonstrate the better fit of our proposed GP model, we also examined its predictive accuracy in a cross-validation analysis. For each of ATO, AIS, AIS-sum, AIS-1800s, using the sampled data from the KG^{CRN} optimization runs, X^n, S^n, Y^n, we fit both a standard GP to predict Y^n given X^n and our proposed CRN-GP to predict Y^n given X^n, S^n. Figure EC.9 depicts the average leave-one-out cross validation error for the 400 datasets, one from each experiment replication. We see that modeling the ATO and AIS-sum benefit greatly from the CRN-GP while AIS and AIS-1800s only show a marginal difference.

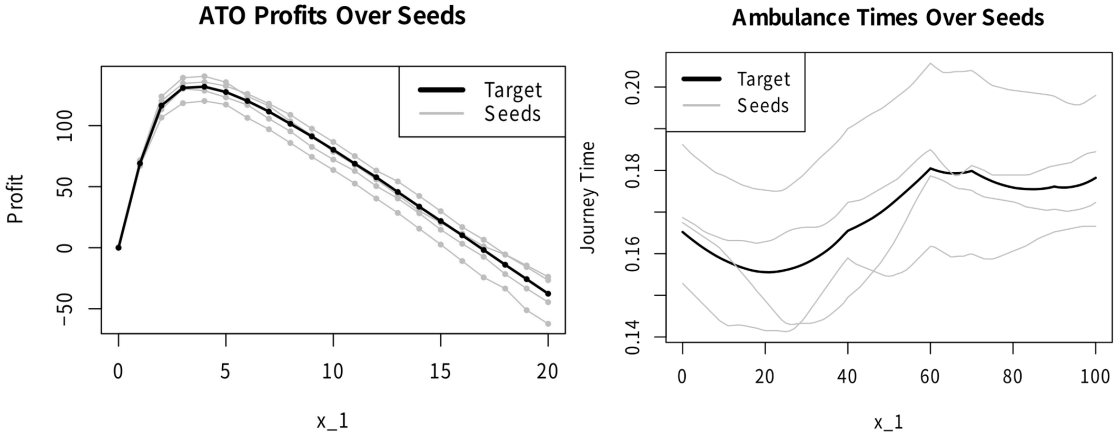

Figure EC.8 Left: ATO problem $x = (l, ..., l)$ for $l \in \{0, ..., 20\}$ and for 6 seeds of $\theta(x,s)$ (grey) and $\theta(x)$ (black). Each seed has shape and peak location almost identical to the target. Right: AIS where $x = (l, x_{2:6})$ with random fixed $x_{2:6}$. Each seed $\theta(x,s)$ has an offset and more seed specific variation from $\bar{\theta}(x)$.

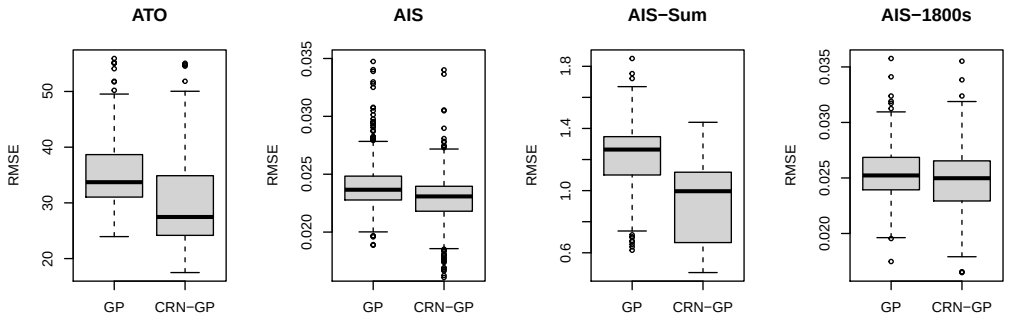

Figure EC.9 For ATO, and AIS-sum, the CRN-GP model performs significantly better and these problems also showed the most benefit from using CRN. For AIS and AIS-1800s, we observe no significant benefit from the CRN model.

EC.4. Algorithm Implementation Details

EC.4.1. Hyperparameter Learning

The hyperparameters of the GP prior are estimated by multi-start conjugate gradient ascent of the marginal likelihood (**?**). All parameters are lower bounded by zero. We set upper bounds for the length scale parameters to double the largest separation between points in each dimension. For all other parameters we set the upper bound to $1.5(\max Y^n - \min Y^n)^2$. We perform optimization in two ways, a full optimization and a finetuning optimization. For the full optimization, we evaluate the marginal likelihood at 1000 points that are randomly uniformly distributed. The best 20 points are used as starting points for 100 steps of conjugate gradient ascent each. This expensive search is used for each of the first 200 iterations of the algorithm, then at increasing intervals thereafter to save computation time. For all other iterations, we only fine tune by conducting 20 steps of gradient ascent using the current best hyperparameters as a starting point.

Recall that the likelihood has the following form:

$$\begin{aligned}\mathbb{P}[Y^n|\tilde{X}^n, L, \sigma_\theta^2, \eta^2, \sigma_b^2, \sigma_w^2] &= -\frac{1}{2}\left((Y^n-\bar{Y})^\intercal K^{-1}(Y^n-\bar{Y}) + \log(|K|) + n\log(2\pi)\right)\\ K_{ij} &= \sigma_\theta^2 \exp\left(-\frac{1}{2}(x^i-x^j)^\intercal L(x^i-x^j)\right)\\ &\quad + \mathbb{1}_{s^i=s^j}\left(\eta^2 + \sigma_b^2\exp\left(-\frac{1}{2}(x^i-x^j)^\intercal L(x^i-x^j)\right) + \mathbb{1}_{x^i=x^j}\sigma_w^2\right).\end{aligned}$$

Firstly, an independent noise model (IND) is fitted by clamping $\eta^2 = \sigma_b^2 = 0$ to yield

$$L^{IND}, {\sigma_{\tilde{\theta}}^2}^{IND}, {\sigma_w^2}^{IND} = \text{argmax}\mathbb{P}[Y^n|\tilde{X}^n, L, \sigma_{\tilde{\theta}}^2, \eta^2 = \sigma_b^2 = 0, \sigma_w^2]. \tag{EC.25}$$

Secondly, the noise parameters $\eta^2, \sigma_b^2, \sigma_w^2$ are optimized whilst keeping the total noise fixed $\eta^2 + \sigma_b^2 + \sigma_w^2 = {\sigma_w^2}^{IND}$ which is a two-dimensional optimization, we reparameterize as follows

$$\begin{aligned}
\eta^2(\alpha, \beta) &= \beta(1-\alpha){\sigma_w^2}^{IND} \\
\sigma_b^2(\alpha, \beta) &= (1-\beta)(1-\alpha){\sigma_w^2}^{IND} \\
\sigma_w^2(\beta) &= \alpha{\sigma_w^2}^{IND} \\
\alpha, \beta &= \text{argmax}_{[0,1]^2}\mathbb{P}[Y^n|\tilde{X}^n, L^{IND}, {\sigma_{\tilde{\theta}}^2}^{IND}, \eta^2(\alpha, \beta), \sigma_b^2(\alpha, \beta), \sigma_w^2(\beta)]
\end{aligned}$$

Thirdly, the final estimates of all hyperparameters are simultaneously fine-tuned by gradient ascent. This three-stage method guarantees that the found likelihood is greater than the equivalent non-CRN parameter estimates. Note that the second extra step of optimization is performed only over the unit square and is thus cheaper than learning all hyperparameters from scratch.

EC.4.2. Optimization of $\mathsf{KG}_n^{\mathsf{CRN}}(x, s)$

Derivatives of KG^{CRN} and KG^{PW}, when evaluated by discretization over X as we do, are easily (but tediously) derived and can be found in multiple previous works (**??**). Alternatively, any automatic differentiation package, (Autograd, TensorFlow, PyTorch) may be used as the mathematical operations are all common functions. We propose the following optimization procedure:

1. Evaluate $\text{KG}^{\text{CRN}}(x, s)$ across an initial Latin Hypercube design with 1000 points over the acquisition space $\tilde{X}_{acq} = X \times \{1, ..., \max S^n + 1\}$.

2. Use the top 20 initial points to initialize 100 steps of conjugate gradient ascent over X, holding the seed constant within each run.

3. For the largest (x, s) pair found, evaluate $\text{KG}^{\text{CRN}}(x, s)$ for the same x on all seeds $s \in \{1, ..., \max S^n + 1\}$

4. Perform 20 steps of gradient ascent to fine tune the x from the best seed.

When not using common random numbers, stages one and two use the same new seed and stages three and four are omitted.

Acknowledgments

The first author gratefully acknowledges funding through the UK Engineering and Physical Sciences Research Council (Grant no. EP/101358X/1).