Bayesian optimization with multiple steps look ahead: Notes and derivations

1 Conditional DPPs for step ahead locations

To approximate $\Lambda_n(\mathbf{x}_1)$ we compute $E_{p(\mathbf{y})}[\min(\mathbf{y}, \eta)]$ where $\mathbf{y} = \{y_1, \dots, y_n\}$ with $p(\mathbf{y}) \sim \mathcal{N}(\mathbf{y}; \mu, \Sigma)$ is the multivariate random vector that we obtain after evaluation the predictive distribution of the GP at locations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Point $\mathbf{x}_1$ is fixed (is the point where we evaluate Λ_n) and we consider $\mathbf{x}_2, \dots, \mathbf{x}_n$ to be a sample of the conditional k-DPP (for $k = n$) on $\mathbf{x}_1$ with kernel L (in principle, the one from the GP).

Let $\mathbf{L}$ be the kernel matrix corresponding to the evaluation of L on a finite set Ω of potential points (pre-uniformly sampled in the domain of interest, for instance). The distribution obtained by conditioning on having observed $\mathbf{x}_1 \in \Omega$ can be obtained as follows. Let $B \subset \Omega$ a non intersecting set with $\mathbf{x}_1$. We have that

$$p_L(\mathbf{x}_1 \cup B | \mathbf{x}_1 \subseteq Z) = \frac{p_L(Z = \mathbf{x}_1 \cup B)}{p(\mathbf{x}_1 \subseteq Z)} = \frac{det(\mathbf{L}_{\mathbf{x}_1 \cup B})}{det(\mathbf{L} - \mathbf{I}_{\bar{\mathbf{x}}_1})} \tag{1}$$

where $\mathbf{I}_{\bar{\mathbf{x}}_1}$ is the matrix with ones in the diagonal entries indexed by elements of $\Omega - \mathbf{x}_1$ and zeros elsewhere. This conditional distribution is again a DPP over subsets of $\Omega - \mathbf{x}_1$ (Borodin and Rains, 2005) with kernel

$$\mathbf{L}^{\mathbf{x}_1} = \left([(\mathbf{L} + \mathbf{I}_{\bar{\mathbf{x}}_1})^{-1}]_{\bar{\mathbf{x}}_1}\right)^{-1}.$$

$[\cdot]_{\bar{\mathbf{x}}_1}$ represents the restriction of the matrix to all rows and columns not indexed by $\mathbf{x}_1$. The previous inverses exist if and only if the probability of $\mathbf{x}_1$ appearing is nonzero, as is the case in our context. A second marginalization is later needed to generate samples of size k. Figure 1 shows an example of samples from a k-DPP and a conditional k-DPP with the same kernel.

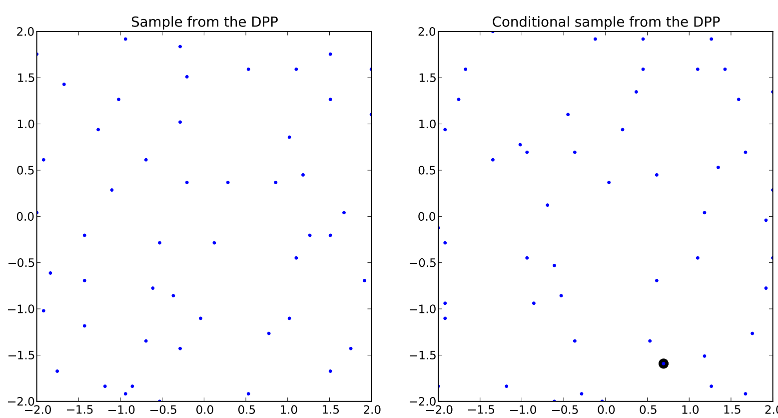

Figure 1: Left: sample from a k-DPP (k=50) for a SE kernel with length-scale 0.5. Right: sample from a k-DPP (k=50) conditional to x_1 (black dot) being in the selected set.

2 Computation of the Expected loss

Our goal is to compute $E_{p(\mathbf{y})}[\min(\mathbf{y},\eta)]$ for $\mathbf{y} = \{y_1,\dots,y_n\}$ and $p_0(\mathbf{y}) \sim \mathcal{N}(\mathbf{y};\mu,\Sigma)$. This approximates the n-steps ahead expected loss $\Lambda_n(\mathbf{x}_1)$. The goal of this section is to write $E_{p(\mathbf{y})}[\min(\mathbf{y},\eta)]$ in a way that it is suitable to be computed by Expectation Propagation. Next proposition will do the work (I think).

Proposition 1

$$E_{p(\boldsymbol{y})}[\min(\boldsymbol{y},\eta)] = \sum_{j=1}^{n}\int_{\mathbb{R}^n} y_j \prod_{i=1}^{n} t_{j,i}(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} + \eta\int_{\mathbb{R}^n}\prod_{i=1}^{n} h_i(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \tag{2}$$

where $h_i(\boldsymbol{y}) = \mathbb{I}\{y_i > \eta\}$ *and*

$$t_{j,i}(\boldsymbol{y}) = \begin{cases} \mathbb{I}\{y_j \leq \eta\} & \textit{if i=j} \\ \mathbb{I}\{0 \leq y_i - y_j\} & \textit{otherwise.} \end{cases}$$

Proof 1 *Denote by*

$$\begin{aligned} E_{p(\boldsymbol{y})}[\min(\boldsymbol{y},\eta)] &= \int_{\mathbb{R}^n} \min(\boldsymbol{y},\eta)\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \\ &= \int_{\mathbb{R}^n-(\eta,\infty)^n} \min(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} + \int_{(\eta,\infty)^n} \eta\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \end{aligned}$$

The first term can be written as follows:

$$\int_{\mathbb{R}^n-(\eta,\infty)^n} \min(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} = \sum_{j=1}^{n}\int_{P_j} y_j\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \tag{3}$$

where $P_j := \{\boldsymbol{y} \in \mathbb{R}^n - (\eta,\infty)^n : y_j \leq y_i,\ \forall i \neq j\}$. *We can do this because the regions* P_j *are disjoint and it holds that* $\cup_{j=1}^{n} P_j = \mathbb{R}^n - (\eta,\infty)^n$. *Also, note that the* $\min(\boldsymbol{y})$ *can be replaced within the integrals since within each* P_j *it holds that* $\min(\boldsymbol{y}) = y_j$. *Rewriting the integral in terms of indicator functions we have that*

$$\sum_{j=1}^{n}\int_{P_j} y_j\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} = \sum_{j=1}^{n}\int_{\mathbb{R}^n} y_j\prod_{i=1}^{n} t_{j,i}(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \tag{4}$$

where $t_{j,i}(y) = \mathbb{I}\{y_i \leq \eta\}$ *if* $j = i$ *and* $t_{j,i}(y) = \mathbb{I}\{y_j \leq y_i\}$ *otherwise.*

The second term can be written as

$$\int_{(\eta,\infty)^n} \eta\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} = \eta\int_{\mathbb{R}^n}\prod_{i=1}^{n} h_i(\boldsymbol{y})\mathcal{N}(\boldsymbol{y};\mu,\Sigma)d\boldsymbol{y} \tag{5}$$

where $h_i(\boldsymbol{y}) = \mathbb{I}\{y_i > \eta\}$. *Merge (4) and (5) to conclude the proof.*

All the elements in (2) can be rewritten in a way that can be computed using EP but the work in (Cunningham, 2011).

- The second term is a Gaussian probability on unbounded polyhedron in which the limits are aligned with the axis.
- The first term requires some more processing but it is still computable under the assumptions in (Cunningham, 2011). Let $\mathbf{w}_j$ the jth canonical vector. Then we have that

$$\begin{aligned} \int_{\mathbb{R}^n} y_j\prod_{i=1}^{n} t_{j,i}(\mathbf{y})\mathcal{N}(\mathbf{y};\mu,\Sigma)d\mathbf{y} &= \mathbf{w}^T\int_{\mathbb{R}^n} \mathbf{y}\prod_{i=1}^{n} t_{j,i}(\mathbf{y})\mathcal{N}(\mathbf{y};\mu,\Sigma)d\mathbf{y} \quad (6) \\ &= \mathbf{w}^T E[\mathbf{y}]z_j \quad (7) \end{aligned}$$

where the expectation is calculated over the normalized distribution over P_j, the one EP approximates with $q(\mathbf{y})$, and for z_j being the normalizing constant

$$z_j = \int_{\mathbb{R}^n} \prod_{i=1}^{n} t_{j,i}(\mathbf{y}) \mathcal{N}(\mathbf{y}; \mu, \Sigma) d\mathbf{y}$$

Because EP does moments matching, both the normalizing constant and the expectation are available.