GLASSES: Relieving The Myopia Of Bayesian Optimisation

Javier González[1,*] Michael Osborne[2] Neil D. Lawrence[1]

[1]University of Sheffield, [2]University of Oxford

Motivation and Summary

- We present GLASSES: *Global optimisation with Look-Ahead through Stochastic Simulation and Expected-loss Search.*
- GLASSES is a non-myopic loss for Bayesian Optimisation that permits the consideration of dozens of evaluations into the future.
- We show that the far-horizon planning thus enabled leads to substantive performance gains in empirical tests.
- Code available in the GPyOpt package (`https://github.com/SheffieldML/GPyOpt`).

Global optimisation problems

Let $f : \mathcal{X} \to \Re$ be well behaved function defined on a compact subset $\mathcal{X} \subseteq \Re^q$. Find

$$\mathbf{x}_M = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}).$$

f is a *black-box*: only evaluations of the type $y_i = f(\mathbf{x}_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are available.

Bayesian optimisation with a myopic expected loss

Figure: Two evaluations: if the first evaluation is made myopically, the second must be sub-optimal.

- $\mathcal{D}_0 = \{(\mathbf{x}_i, y_i)\}_{i=1}^N = (\mathbf{X}_0, \mathbf{y}_0)$: Available dataset.
- $p(f) = \mathcal{GP}(\mu; k)$: Gaussian process (GP) using $\mathcal{D}_0$.
- $\mathcal{I}_0$: conjunction of $\mathcal{D}_0$, the model parameters and the model likelihood type.
- $\eta = \min\{\mathbf{y}_0\}$: current best found value.
- One remaining evaluation before we need to report our inferred location of the minimum.

The *loss* of evaluating *f* this last time at $\mathbf{x}_*$ assuming it is returning y_* is

$$\lambda(y_*) \triangleq \begin{cases} y_*; & \text{if } y_* \leqslant \eta \\ \eta; & \text{if } y_* > \eta. \end{cases}$$

The *expectation of the loss*:

$$\Lambda_1(\mathbf{x}_*|\mathcal{I}_0) \triangleq \mathbb{E}[\min(y_*, \eta)] = \int \lambda(y_*) p(y_*|\mathbf{x}_*, \mathcal{I}_0) dy_* = \eta + (\mu - \eta)\Phi(\eta; \mu, \sigma^2) - \sigma^2\mathcal{N}(\eta, \mu, \sigma^2)$$

- We have abbreviated $\sigma^2(y_*|\mathcal{I}_0)$ as σ^2 and $\mu(y_*|\mathcal{I}_0)$ as μ.
- The subscript in Λ refers to the fact that we are considering one future evaluation.
- The next evaluation is located where $\Lambda_1(\mathbf{x}_*|\mathcal{I}_0)$ gives the minimum value [1].

$\Lambda_1(\mathbf{x}_*|\mathcal{I}_0)$ is myopic: doesn't take into account the number of remaining evaluations.

Ideal non-myopic expected loss

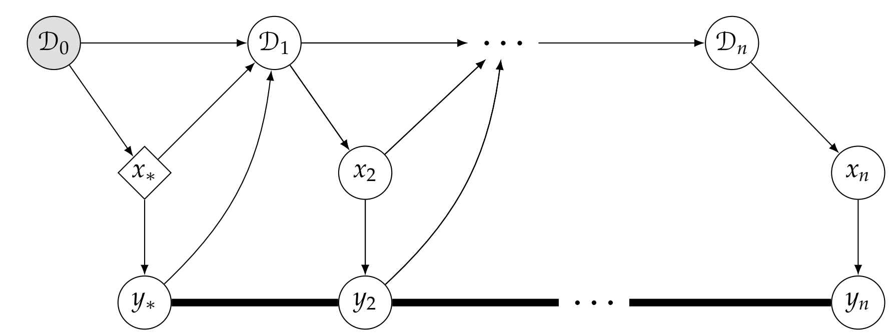

Figure: A Bayesian network describing the *n*-step lookahead problem.

The *ideal long-sight* loss is defined as:

$$\Lambda_n(\mathbf{x}_*|\mathcal{I}_0) = \int \lambda(y_n) \prod_{j=1}^{n} p(y_j|\mathbf{x}_j, \mathcal{I}_{j-1}) p(\mathbf{x}_j|\mathcal{I}_{j-1}) dy_* \dots dy_n d\mathbf{x}_2 \dots d\mathbf{x}_n$$

- $p(y_j|\mathbf{x}_j, \mathcal{I}_{j-1}) = \mathcal{N}\left(y_j; \mu(\mathbf{x}_j; \mathcal{I}_{j-1}), \sigma^2(\mathbf{x}_j|\mathcal{I}_{j-1})\right)$: predictive distribution of the GP at $\mathbf{x}_j$
- $p(\mathbf{x}_j|\mathcal{I}_{j-1}) = \delta\left(\mathbf{x}_j - \arg\min_{\mathbf{x}_*\in\mathcal{X}} \Lambda_{n-j+1}(\mathbf{x}_*|\mathcal{I}_{j-1})\right)$: optimisation step required to obtain $\mathbf{x}_j$.

The optimal long-sight loss is extremely expensive to compute

GLASSES

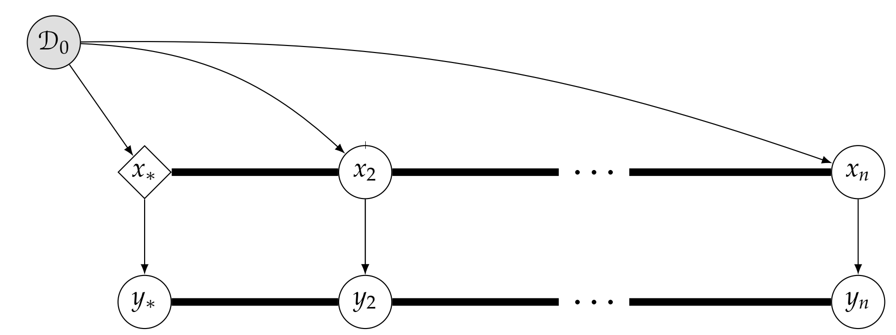

Figure: A Bayesian network describing our approximation to the *n*-step lookahead problem. Compare with the figure above: the sparser structure renders our approximation computationally tractable.

Take $\mathcal{F}_n(\mathbf{x}_*)$ is an oracle function able to predict the *n* future locations starting at $\mathbf{x}_*$:

$$\Lambda_n\left(\mathbf{x}_* \mid \mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*)\right) = \mathbb{E}[\min(\mathbf{y}, \eta)] = \eta \int_{\mathbb{R}^n} \prod_{i=1}^{n} h_i(\mathbf{y}) \mathcal{N}(\mathbf{y}; \mu, \Sigma) d\mathbf{y} + \sum_{j=1}^{n} \int_{\mathbb{R}^n} y_j \prod_{i=1}^{n} t_{j,i}(\mathbf{y}) \mathcal{N}(\mathbf{y}; \mu, \Sigma) d\mathbf{y},$$

where $h_i(\mathbf{y}) = \mathbb{I}\{y_i > \eta\}$ and $t_{j,i}(\mathbf{y}) = \mathbb{I}\{y_j \leqslant \eta\}$ if $i = j$ and $t_{j,i}(\mathbf{y}) = \mathbb{I}\{0 \leqslant y_i - y_j\}$ otherwise.

- $\Lambda_n\left(\mathbf{x}_* \mid \mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*)\right)$ can be computed with Expectation Propagation [2].
- A batch BO method [3] is used as a surrogate for $\mathcal{F}_n(\mathbf{x}_*)$.
- $\Lambda_n\left(\mathbf{x}_* \mid \mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*)\right)$ is optimised using a gradient-free method (DIRECT).

First non-myopic loss able to take into account dozens of future evaluations

Results

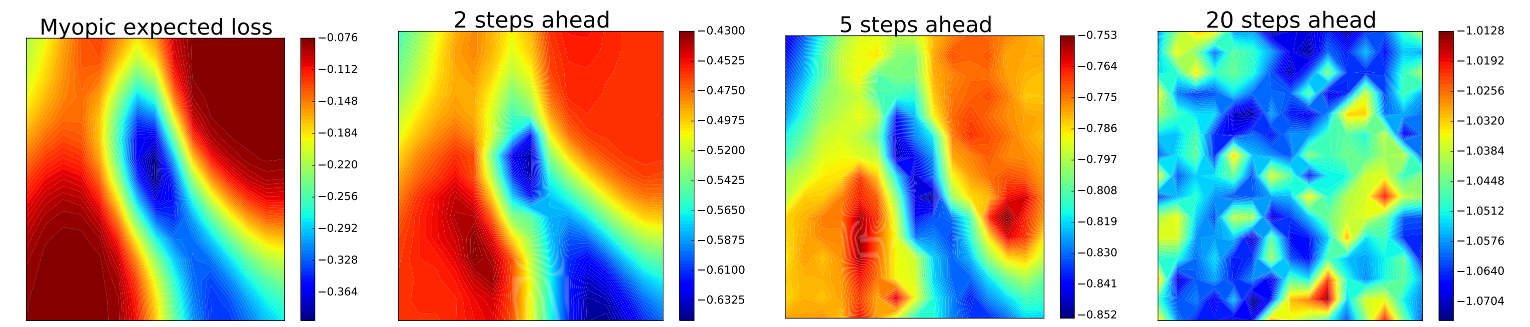

Figure: Expected loss for different number of steps ahead in an example with 10 data points and the Six-hump Camel function.

GLASSES is more explorative the more remaining evaluations are available

	MPI	GP-LCB	EL	EL-2	EL-3	EL-5	EL-10	GLASSES
SinCos	0.7147	0.6058	0.7645	*0.8656*	0.6027	0.4881	*0.8274*	***0.9000***
Cosines	0.8637	0.8704	0.8161	*0.8423*	*0.8118*	0.7946	0.7477	***0.8722***
Branin	0.9854	0.9616	**0.9900**	0.9856	0.9673	0.9824	0.9887	0.9811
Sixhumpcamel	0.8983	**0.9346**	0.9299	0.9115	0.9067	0.8970	0.9123	0.8880
Mccormick	**0.9514**	0.9326	0.9055	*0.9139*	*0.9189*	*0.9283*	*0.9389*	*0.9424*
Dropwave	0.7308	0.7413	0.7667	0.7237	0.7555	0.7293	0.6860	***0.7740***
Powers	0.2177	0.2167	0.2216	*0.2428*	*0.2372*	*0.2390*	*0.2339*	***0.3670***
Ackley-2	0.8230	**0.8975**	0.7333	0.6382	0.5864	0.6864	0.6293	0.7001
Ackley-5	0.1832	0.2082	0.5473	*0.6694*	0.3582	0.3744	***0.6700***	0.4348
Ackley-10	0.9893	0.9864	0.8178	*0.9900*	*0.9912*	***0.9916***	*0.8340*	*0.8567*
Alpine2-2	**0.8628**	0.8482	0.7902	0.7467	0.5988	0.6699	0.6393	0.7807
Alpine2-5	0.5221	0.6151	**0.7797**	0.6740	0.6431	0.6592	0.6747	0.7123

Table: Results for the average 'gap' measure (5 replicates) across different functions. EL-k: expect loss with *k* steps ahead. MPI: maximum probability of improvement. GP-LCB: lower confidence bound criterion.

GLASSES improves other myopic losses in practice

Conclusions and future work

- First non-myopic loss that allows taking into account dozens of future evaluations.
- The loss compares well with current myopic acquisitions.
- Challenge: making the optimisation of the loss more efficient.

References

1 Michael Osborne. Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature. PhD thesis, University of Oxford, 2010.

2 John P. Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation. arXiv:1111.6832 [stat], Nov 2011. arXiv: 1111.6832.

3 Javier González, Zhenwen Dai, Philipp Hennig, and Neil D Lawrence. Batch Bayesian optimization via local penalization. arXiv preprint arXiv:1505.08052, 2015.

Funded by the BRIC-BBSRC project No BB/K011197/1.

*j.h.gonzalez}@sheffield.ac.uk