Progressive-GLASSES?

Anonymous Author(s)
Affiliation
Address
email

Abstract

(Declaration of intent!) The vast majority of currently available Bayesian optimization (BO) methods used to tune the parameters of Machine Learning algorithms are cost-myopic: they try to make the best possible progress in the next function evaluation irrespective of the available cost budget to evaluate the objective. We present Progressive-GLASSES, *Progressive Global optimisation with Look-Ahead through Stochastic Simulation and Expected-loss Search*, the first BO approach able to make optimal non-myopic decisions according to a general cost function, possible unknown, given a budget restriction. We show the superior performance of Progressive-GLASSES in a variety of experiements in which a fixed time budget is used to tune the parameters of several algorithms.

1 Some notes and ideas

General selling point:

- Imagine you have to solve a machine learning task and you are training your favorite algorithm. But you have a problem: you only have one hour before your results have to be summited: *What about having a method able to find the best possible parameters of your algorithm within your limited time budget?*

Particular aspects of Progressive-GLASSES:

- Having a fixed time budget (rather than a fixed number of steps ahead) and different costs accross the domain of the function makes that the number of steps ahead is not fixed anymore: the sum of the costs of the steps ahead should also be smaller than the remaining time budget.
- Something interesting here is that the probability measure over the future steps should give measure zero to those sets of points (of whatever size) whose added cost is larger the available budget.
- The cost function of evaluating f can be fixed beforehand (we should start in this case, I guess) but a more realistic situation is to consider that we learn it as we run the optimization. we can use the log of a GP for this (perhaps use only the mean) and also to consider that the GP for $f(x)$ and $c(x)$ are coupled (and model it with a multi-output GP).

Issues when trying to use what we learned in the original GLASSES paper:

- The goal of this work is to generalize GLASSES to cases where there is a limitation in terms of a cost of evaluation rather than in terms of a number of evaluations. A specially interesting case of this is when the cost is the time to evaluate f.

- The main problem when trying to generalize GLASSES to this context is that, if you have a limited time budget, you need to be very efficient making the decisions. If not, you are limiting yourself a few number of evaluations of f.
- There are two bottlenecks on GLASSES that I am trying to solve: 1) the simulations of the steps ahead and 2) the computation of the loss. If we can speed up these two steps we will have something interesting.
- Regarding step 2: It is possible to write $E_{p(y)}[\min(y, \eta)]$ in terms of cumulated Gaussian distributions so we can avoid the EP step of GLASSES (and gradients would be available). The key to do this is in *Tallis, G 1961: The moment generating function of the truncated multi-normal distribution. J. Roy. Statist. Soc. Ser. B*. D Ginsbourger uses the main theorem of this paper to compute the multi-point expected improvement. The same type of arguments can be used in our context to obtain $E_{p(y)}[\min(y, \eta)]$.
- Regarding 1): I haven't thought yet so much about this point but perhaps we should give another try to the original DPP idea. If we could come up some efficient way of sampling from conditional dpps (of computing the MAP) this would be very elegant way of addressing this point.

Some citations to have in mind:

[3] [2] [6] [1] [5] [4]

2 Problem, notation, etc.

Let $f : \mathcal{X} \to \Re$ be well behaved function defined on a bounded $\mathcal{X} \subseteq \Re^q$. Our goal is to find

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

where we assume that f is a *black-box* from which only perturbed evaluations of the type $y_i = f(\mathbf{x}_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, are available. In this work we assume that every time we evaluate f at $\mathbf{x}_i$ we incur in a cost c_i, which is the output of a some smooth and differentiable function $c : \mathcal{X} \to \Re^+$. Especially interesting is the case in which the cost corresponds to the wall-clock time of evaluating f, but forms of cost are also valid here.

Our aim is to define an heuristic strategy able to make a series of evaluations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of f such that the the minimum of f is obtained as soon as possible while the condition $\sum_{i=1}^n c(\mathbf{x}_i) \leq C$, for some fixed cost budget $C \in \Re^+$, is satisfied.

- Gaussian process for f.
- $c(\mathbf{x})$ can be: fixed before hand, learned using another GP, that can be potentially coupled with f. We will assume that it is a deterministic function, even if it is learned from data (we use the posterior mean of a GP).
- Assume that N points have been gathered so far, having a dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, c_i, y_i)\}_{i=1}^N = (\mathbf{X}_0, \mathbf{c}_0, \mathbf{y}_0)$.
- $\eta = \min\{\mathbf{y}_0\}$ is the current best found value. We define the *loss per unit of cost* of evaluating f this last time at $\mathbf{x}_*$ assuming it is returning y_* as $\lambda^c(y_*) = \lambda(y_*)/c_*$ where

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

  and $c_* = c(\mathbf{x}_*)$, is the cost of evaluating f at $\mathbf{x}_*$. Its expectation is

  $$\mathbb{E}[\lambda^c(y_*)] = \int \lambda^c(y_*) p(y_*|\mathbf{x}_*, \mathcal{I}_0) \mathrm{d}y_* = \frac{1}{c_*}\mathbb{E}[\min(y_*, \eta)]$$

- Assuming we have an available budget C we can select the next (last) evaluation using the loss

  $$\Lambda^c(\mathbf{x}_*|\mathcal{I}_0) \triangleq \begin{cases} \mathbb{E}[\lambda^c(y_*)]; & \text{if} \quad c_* \leq C. \\ \infty; & \text{if} \quad c_* > C. \end{cases}$$

  And this is the myopic loss used in [6] (the expected improvement). Only points in $\mathcal{X}$ whose cost is lower that the current available budget are candidates.

- Assuming we have an available budget C to select the next (last) evaluation, the loss function is

$$\Lambda^c(\mathbf{x}_*|\mathcal{I}_0) \triangleq \begin{cases} \int \lambda(y_n) \prod_{j=1}^n p(y_j|\mathbf{x}_j, \mathcal{I}_{j-1}) p(\mathbf{x}_j|\mathcal{I}_{j-1}) \mathrm{d}y \mathrm{d}\mathbf{x}; & \text{if} \quad \sum c_j \leq C. \\ \infty; & \text{if} \quad \sum c_j > C. \end{cases}$$

- Assuming we have an available budget C to select the next (last) evaluation, the loss function is

$$\Lambda^c(\mathbf{x}_*|\mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*) \triangleq \begin{cases} \frac{1}{c_*}\mathbb{E}\big[\min(\mathbf{y}, \eta)\big]; & \text{if} \quad \sum c_j \leq C. \\ \infty; & \text{if} \quad \sum c_j > C. \end{cases}$$

  where in this case $\mathbf{y}$ is random Gaussian vector associated to the future

3 Some new results that we can potentially use

Proposition 1 *Denote by $\Phi_n(\boldsymbol{r}; \Sigma) = \mathbb{P}(\boldsymbol{z} \leq \boldsymbol{r})$ the c.d.f. of a general centered n-dimensional Gaussian vector $\boldsymbol{z}$ with covariance matrix Σ. Let $\boldsymbol{z} := (z_1, \ldots, z_n)^T$ be a Gaussian vector with mean $\mu \in \Re^n$ and $\Sigma \in \Re^{n \times n}$. It holds that (CHECK, THERE IS AN ERROR HERE!)*

$$\mathbb{E}[z_k|\boldsymbol{z} \leq 0] = \mu_k - \Phi_n(-\mu; \Sigma)^{-1} \sum_{i=1}^n \Sigma_{ik} \cdot \Phi_1(-\mu_i; \Sigma_{ii}) \cdot \Phi_{n-1}(0; \mu_{\neg i}, \Sigma_{\neg i})$$

where μ_i and σ_{ii} are respectively the ith entries of μ and Σ, $\mu_{\neg i}$ is the $(n-1)$ dimensional vector with jth element $\Sigma_{ij}\Sigma_{ii}^{-1}m_i - m_j, \forall i \neq j$ and $\Sigma_{\neg i}$ is the $(n-1)\times(n-1)$ matrix with qs-th elements $\Sigma_{qs} - \Sigma_{is}\Sigma_{ii}^{-1}\Sigma_{iq}$.

Proposition 2 *It holds that: (CHECK, THERE IS AN ERROR HERE!)*

$$\Lambda_n\big(\mathbf{x}_* \mid \mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*)\big) = \sum_{k=1}^n \left(\mu_k^k \cdot \Phi_n(-\mu^k; \Sigma^k) - \sum_{i=1}^n \Sigma_{ik}^k \cdot \Phi_1^k(-\mu_i^k; \Sigma_{ii}^k) \cdot \Phi_{n-1}(\mu_{\neg i}^k, \Sigma_{\neg i}^k) \right) + \eta$$

where for $k = 1, \ldots, n$ the elements of the vectors $\boldsymbol{z}^k := (\boldsymbol{z}_1^k, \ldots, \boldsymbol{z}_n^k)^T$, which are Gaussian with known mean $\mu^k \in \Re^n$ and covariance $\Sigma \in \Re^{n \times n}$, are defined as:

$$z_j^k = \begin{cases} y_k - \eta & \text{if } j = k \\ \\ y_k - y_j & \text{if } j \neq k \end{cases}$$

Remark 1 *The computation of $\Lambda_n\big(\mathbf{x}_* \mid \mathcal{I}_0, \mathcal{F}_n(\mathbf{x}_*)\big)$ requires n calls to $\Phi_n(\cdot)$ and n^2 calls to $\Phi_1(\cdot)$ and $\Phi_{n-1}(\cdot)$.*

4 Results

5 Conclusions

References

[1] Eric Brochu, Vlad M. Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[2] Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[3] Daniel James Lizotte. *Practical Bayesian Optimization*. PhD thesis, University of Alberta, 2008. AAINR46365.

[4] Roman Marchant, Fabio Ramos, and Scott Sanner. Sequential Bayesian optimisation for spatial-temporal monitoring. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2014.

[5] Michael A. Osborne, Roman Garnett, and Stephen J. Roberts. Gaussian processes for global optimization. In *3rd international conference on learning and intelligent optimization (LION3)*, pages 1–15, 2009.

[6] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. *Practical Bayesian optimization of machine learning algorithms*, page 2951–2959. 2012.

Supplementary materials for: 'XXX"

Authors here

S1 Proofs

Proof 1 *See (Tallis, 1961) and (Chevalier and Ginsbourger, 2012).*

Proof 2 *First, we write* $\mathbb{E}[\min(\mathbf{y},\eta)]$ *as the sum of the expectations of* n *marginal truncated Gaussians.*

$$\begin{aligned}
\mathbb{E}[\min(\mathbf{y},\eta)] &= \mathbb{E}[\min(\mathbf{y}-\eta,0)]+\eta \\
&= \mathbb{E}[\min(\mathbf{y}-\eta)]\cdot\sum_{k=1}^{n}\mathbb{I}\{\, y_k\leq\eta, y_k\leq y_j, \forall k\neq j\}+\eta \\
&= \sum_{k=1}^{n}\mathbb{E}[y_k-\eta|\; y_k\leq\eta, y_k\leq y_j, \forall k\neq j]\cdot\mathbb{I}\{y_k\leq\eta, y_k\leq y_j, \forall k\neq j\}+\eta \\
&= \sum_{k=1}^{n}\mathbb{E}[y_k-\eta|\; y_k-\eta\leq 0, y_k-y_j\leq 0, \forall k\neq j]\cdot\mathbb{I}\{y_k-\eta\leq 0, y_k-y_j\leq 0, \forall k\neq j\}+\eta \\
&= \sum_{k=1}^{n}\mathbb{E}[z_k^k|\boldsymbol{z}^k\leq 0]\cdot p(\boldsymbol{z}^k\leq 0)+\eta \qquad \text{(S.1)}
\end{aligned}$$

where for $k=1,\ldots,n$ *the elements of the vectors* $\boldsymbol{z}^k:=(z_1^k,\ldots,z_n^k)^T$ *are defined as:*

$$z_j^k=\begin{cases} y_k-\eta & \text{if } j=k \\ y_k-y_j & \text{if } j\neq k \end{cases}$$

Each element $\mathbb{E}[z_k^k|\boldsymbol{z}^k\leq 0]$ *can now be computed using Proposition 2. In particular, let* μ^k *and* Σ^k *be mean vector and covariance matrix associated to each* $\boldsymbol{z}^k$*. Then*

$$\mathbb{E}[z_k^k|\boldsymbol{z}^k\leq 0]=\mu_k^k-\Phi_n(-\mu^k;\Sigma^k)^{-1}\sum_{i=1}^{n}\Sigma_{ik}^k\cdot\Phi_1^k(-\mu_i^k;\Sigma_{ii}^k)\cdot\Phi_{n-1}(\mu_{\neg i}^k,\Sigma_{\neg i}^k)$$

and therefore

$$\begin{aligned}
\Lambda_n\big(\mathbf{x}_*\mid\mathcal{I}_0,\mathcal{F}_n(\mathbf{x}_*)\big) &= \sum_{k=1}^{n}\mathbb{E}[z_k^k|\boldsymbol{z}^k\leq 0]\cdot\Phi_n(-\mu^k;\Sigma^k)+\eta \\
&= \sum_{k=1}^{n}\left(\mu_k^k\cdot\Phi_n(-\mu^k;\Sigma^k)-\sum_{i=1}^{n}\Sigma_{ik}^k\cdot\Phi_1^k(-\mu_i^k;\Sigma_{ii}^k)\cdot\Phi_{n-1}(\mu_{\neg i}^k,\Sigma_{\neg i}^k)\right)+\eta
\end{aligned}$$