Exact Expected Regret Minimization partial-order derivatives for Bayesian Optimization with Gaussian Processes

Conor Clare

Ulster University, Shore Road, Newtownabbey, UK
clare-c@ulster.ac.uk

1 Gaussian Processes

GPs are multivariate, Gaussian distributions over functions f and offer a Bayesian, non-parametric solution to regression problems with non-linear $f(\mathbf{x})$ [6]. GPs have flexible covariance function (kernel) parameters for non-linear prediction, are simply-defined and are usually chosen as the surrogate for Bayesian optimization. A stochastic process $f(\mathbf{x})$ is Gaussian when observations jointly sampled have a multivariate Gaussian probability distribution. GPs are simply defined using two functions. The first is the mean function, $m(\mathbf{x})$, defining the expected value of a location, $\mathbf{x}$. The second is the kernel function $k(\mathbf{x}, \mathbf{x}')$, which calculates the covariance between two different locations $\mathbf{x}$ and $\mathbf{x}'$ [6]:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

The GP posterior mean is the vector $\hat{\mu}_{GP}(\mathbf{x}) = \mathbf{k}^T\mathbf{C^{-1}}y$, the GP posterior covariance is the matrix $\hat{\sigma}^2_{GP}(\mathbf{x}) = \kappa - \mathbf{k}^T\mathbf{C^{-1}}\mathbf{k}$ [6,8], with $\mathbf{C^{-1}}$ the (inverted) covariance defined by the kernel between the observed training inputs, $\mathbf{x}_k \in \mathcal{D}_N$. The inversion of $\mathbf{C}$, an $N \times N$ positive-definite, symmetrical matrix is the GP posterior's main computational expense and can be reduced by Cholesky decomposition for a faster and more numerically-stable result [6]. The covariance of the unobserved prediction inputs is κ, with $\mathbf{k}^T$ the (transposed) covariance of the kernel between the unobserved prediction inputs and observed training inputs. Finally, the GP posterior standard deviation $\hat{\sigma}_{GP}(\mathbf{x})$ is the square-root of the diagonal of $\hat{\sigma}^2_{GP}(\mathbf{x})$.

1.1 Kernels for Bayesian optimization

Common kernels widely-used in Bayesian optimization with GPs include the squared-exponential (SE) covariance function [6] and the Matérn class of covariance functions e.g. Matérn 3/2 and Matérn 5/2 [6]. The SE kernel uses the exponential function and is infinitely differentiable. Throughout this paper, we use $\mathbf{k}$ to denote a symmetric, SE kernel:

$$\mathbf{k} = k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}' - \mathbf{x})^2\right)$$

$\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}}$ is the Jacobian matrix of first-order partial derivatives for $\mathbf{k}$ (transposed) w.r.t to input $\mathbf{x}$:

$$\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}} = \left[(\mathbf{x}' - \mathbf{x}) \exp\left(-\frac{1}{2}(\mathbf{x}' - \mathbf{x})^2\right)\right]^T = [\mathbf{k}(\mathbf{x}' - \mathbf{x})]^T \tag{1}$$

$\frac{\partial^2 \mathbf{k}^T}{\partial \mathbf{x}^2}$ is the Hessian matrix of second-order partial derivatives for $\mathbf{k}$ (transposed) w.r.t to input $\mathbf{x}$:

$$\frac{\partial^2 \mathbf{k}^T}{\partial \mathbf{x}^2} = \left[\frac{\partial \mathbf{k}}{\partial \mathbf{x}}(\mathbf{x}' - \mathbf{x}) + \mathbf{k}(-1)\right]^T = \left[\mathbf{k}(\mathbf{x}' - \mathbf{x})^2 - \mathbf{k}\right]^T \tag{2}$$

1.2 Gaussian Processes: Posterior Partial Derivatives

Using Eq. 1 and [1,2,3,4,7], the first-order partial derivatives of the GP posterior mean $\hat{\mu}_{GP}(\mathbf{x})$ and the GP posterior covariance $\hat{\sigma}^2_{GP}(\mathbf{x})$ are respectively $\frac{\partial \hat{\mu}_{GP}}{\partial \mathbf{x}}(\mathbf{x})$ and $\frac{\partial \hat{\sigma}^2_{GP}}{\partial \mathbf{x}}(\mathbf{x})$:

$$\frac{\partial \hat{\mu}_{GP}}{\partial \mathbf{x}}(\mathbf{x}) = \frac{\partial \mathbf{k}^T}{\partial \mathbf{x}}\mathbf{C}^{-1}y \tag{3}$$

$$\frac{\partial \hat{\sigma}^2_{GP}}{\partial \mathbf{x}}(\mathbf{x}) = -\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}}\mathbf{C}^{-1}\mathbf{k} - \mathbf{k}^T\mathbf{C}^{-1}\frac{\partial \mathbf{k}}{\partial \mathbf{x}} = -2\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}}\mathbf{C}^{-1}\mathbf{k} \tag{4}$$

By differentiating Eq. 3 and Eq. 4 (using Eq. 1 and Eq. 2), the second-order partial derivatives of the GP posterior mean $\hat{\mu}_{GP}(\mathbf{x})$ and the GP posterior covariance $\hat{\sigma}^2_{GP}(\mathbf{x})$ are respectively $\frac{\partial^2 \hat{\mu}_{GP}}{\partial \mathbf{x}^2}(\mathbf{x})$ and $\frac{\partial^2 \hat{\sigma}^2_{GP}}{\partial \mathbf{x}^2}(\mathbf{x})$:

$$\frac{\partial^2 \hat{\mu}_{GP}}{\partial \mathbf{x}^2}(\mathbf{x}) = \frac{\partial^2 \mathbf{k}^T}{\partial \mathbf{x}^2}\mathbf{C}^{-1}y \tag{5}$$

$$\frac{\partial^2 \hat{\sigma}^2_{GP}}{\partial \mathbf{x}^2}(\mathbf{x}) = -2\left(\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}}\mathbf{C}^{-1}\frac{\partial \mathbf{k}}{\partial \mathbf{x}} + \frac{\partial^2 \mathbf{k}^T}{\partial \mathbf{x}^2}\mathbf{C}^{-1}\mathbf{k}\right) \tag{6}$$

1.3 Expected Regret Minimization with Gaussian Processes

Expected Regret Minimization is a new acquisition function for Bayesian optimisation with GPs [5]. Throughout this paper, we denote it as GP ERM and can define it as:

$$\mathrm{ERM}_{GP}(\mathbf{x}) = \hat{\sigma}_{GP}(\mathbf{x})\phi(z_{f^*}) + [f^* - \hat{\mu}_{GP}(\mathbf{x})]\Phi(z_{f^*}) \tag{7}$$

where: $z_{f^*} = \frac{f^* - \hat{\mu}_{GP}(\mathbf{x})}{\hat{\sigma}_{GP}(\mathbf{x})}$; with $\phi(z_{f^*})$ and $\Phi(z_{f^*})$ the respective probability density function (PDF) and cumulative distribution function (CDF) of a univariate, standard normal random variable, z_{f^*}. Prior knowledge of the best y-value is denoted f^*. Eq. 7 can be re-written as:

$$\mathrm{ERM}_{GP}(\mathbf{x}) = \hat{\sigma}_{GP}(\mathbf{x})[z_{f^*}\Phi(z_{f^*}) + \phi(z_{f^*})] \tag{8}$$

2 Deriving Exact GP ERM partial-order derivatives

2.1 The Exact GP ERM Jacobian: GP dERM

Using Eq. 8 and differentiation [1,2,3,4,7], the exact Jacobian matrix of first-order partial derivatives of GP ERM w.r.t. input $\mathbf{x}$ is $\frac{\partial \mathrm{ERM}_{GP}(\mathbf{x})}{\partial \mathbf{x}}$:

$$\frac{\partial \mathrm{ERM}_{GP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}}[z_{f^*}\Phi(z_{f^*}) + \phi(z_{f^*})] + \hat{\sigma}_{GP}(\mathbf{x})\Phi(z_{f^*})\frac{\partial z_{f^*}}{\partial \mathbf{x}} \tag{9}$$

where: $\frac{\partial z_{f^*}}{\partial \mathbf{x}} = \left(\frac{\partial \hat{\mu}_{GP}(\mathbf{x})}{\partial \mathbf{x}} - z_{f^*}\frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}}\right)/\hat{\sigma}_{GP}(\mathbf{x})$ and $\frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{2\hat{\sigma}_{GP}(\mathbf{x})}\frac{\partial \hat{\sigma}^2_{GP}(\mathbf{x})}{\partial \mathbf{x}}$, using Eq. 3 and Eq. 4 above.

2.2 The Exact GP ERM Hessian: GP $\mathbf{d^2}$ERM

This work's second knowledge contribution differentiates (using the product rule) Eq. 9 to derive $\frac{\partial^2 \mathrm{ERM}_{GP}(\mathbf{x})}{\partial \mathbf{x}^2}$, the exact Hessian matrix of second-order partial derivatives of GP ERM w.r.t. input $\mathbf{x}$:

$$\begin{aligned}\frac{\partial^2 \mathrm{ERM}_{GP}(\mathbf{x})}{\partial \mathbf{x}^2} = \frac{\partial^2 \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}^2}[z_{f^*}\Phi(z_{f^*}) + \phi(z_{f^*})] + 2\frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}}\Phi(z_{f^*})\frac{\partial z_{f^*}}{\partial \mathbf{x}} \\ + \hat{\sigma}_{GP}(\mathbf{x})\Phi(z_{f^*})\frac{\partial^2 z_{f^*}}{\partial \mathbf{x}^2} + \hat{\sigma}_{GP}(\mathbf{x})\phi(z_{f^*})\frac{\partial z_{f^*}}{\partial \mathbf{x}}\end{aligned} \tag{10}$$

where: $\frac{\partial^2 \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x^2}} = -\frac{1}{2\hat{\sigma}^2_{GP}(\mathbf{x})}\frac{\partial \hat{\sigma}^2_{GP}(\mathbf{x})}{\partial \mathbf{x}}\frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2\hat{\sigma}_{GP}(\mathbf{x})}\frac{\partial^2 \hat{\sigma}^2_{GP}}{\partial \mathbf{x}^2}(\mathbf{x})$ and $\frac{\partial^2 z_{f^*}}{\partial \mathbf{x}^2} = \left(\frac{\partial \hat{\mu}^2_{GP}(\mathbf{x})}{\partial \mathbf{x}^2} - z_{f^*}\frac{\partial^2 \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}^2} - 2\frac{\partial z_{f^*}}{\partial \mathbf{x}}\frac{\partial \hat{\sigma}_{GP}(\mathbf{x})}{\partial \mathbf{x}}\right)/\hat{\sigma}_{GP}(\mathbf{x})$, using Eq. 5 and Eq. 6 above.

References

1. Chandak, A., Dey, D., Mukhoty, B., Kar, P.: Epidemiologically and socio-economically optimal policies via Bayesian optimization. `https://doi.org/10.1007/s41403-020-00142-6` (2020)
2. Frean, M., Boyle, P.: Using Gaussian Processes to optimize expensive functions. In: AI 2008: Advances in Artificial Intelligence. pp. 258–267 (2008)
3. Klein, A., Falkner, S., Mansur, N., Hutter, F.: RoBO: A flexible and robust Bayesian optimization framework in Python. In: NIPS 2017 Bayesian Optimization Workshop (Dec 2017)
4. Marmin, S., Chevalier, C., Ginsbourger, D.: Differentiating the multi-point Expected Improvement for optimal batch design. In: Machine Learning, Optimization, and Big Data. pp. 37–48. Springer International Publishing (2015)
5. Nguyen, V., Osborne, M.A.: Knowing The What But Not The Where in Bayesian Optimization. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020) (2020)
6. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)

7. Roustant, O., Ginsbourger, D., Deville, Y.: DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software **51**(1), 1–55 (2012)
8. Tracey, B., Wolpert, D.: Upgrading from Gaussian Processes to Student's-t Processes. In: 2018 AIAA Non-Deterministic Approaches Conference (2018)