The Exact Expected Improvement Jacobian for Bayesian Optimization with Student's-t Processes

Conor Clare

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

1 Kernels for Bayesian optimization

Common kernels widely-used in Bayesian optimization with STPs include the squared-exponential (SE) covariance function [5] and the Matérn class of covariance functions e.g. Matérn 3/2 and Matérn 5/2 [5]. 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 $\mathbf{x}$:

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

2 Student's-t Processes: Partial Derivatives

Using Eq. 1 and [1,2,3,4,6], the first-order partial derivatives of the STP posterior mean $\hat{\mu}_{STP}(\mathbf{x})$ and the STP posterior covariance $\hat{\sigma}^2_{STP}(\mathbf{x})$ are respectively $\frac{\partial \hat{\mu}_{STP}}{\partial \mathbf{x}}(\mathbf{x})$ and $\frac{\partial \hat{\sigma}^2_{STP}}{\partial \mathbf{x}}(\mathbf{x})$:

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

$$\begin{aligned} \frac{\partial \hat{\sigma}^2_{STP}}{\partial \mathbf{x}}(\mathbf{x}) &= -\left(\frac{\nu + y^T \mathbf{C}^{-1} y + 2}{\nu + \mathcal{D}_N + 2}\right) \times \left(\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}} \mathbf{C^{-1}} \mathbf{k} - \mathbf{k}^T \mathbf{C}^{-1} \frac{\partial \mathbf{k}}{\partial \mathbf{x}}\right) \\ &= -2\left(\frac{\nu + y^T \mathbf{C}^{-1} y + 2}{\nu + \mathcal{D}_N + 2}\right) \times \left(\frac{\partial \mathbf{k}^T}{\partial \mathbf{x}} \mathbf{C^{-1}} \mathbf{k}\right) \end{aligned} \tag{3}$$

3 Expected Improvement with Student's-t Processes

Expected Improvement is a new acquisition function for Bayesian optimisation with STPs [7,8,9]. Throughout this paper, we denote it as STP EI and can define it as:

$$\mathrm{EI}_{STP}(\mathbf{x}) = \hat{\sigma}_{STP}(\mathbf{x}) \left(\frac{\nu}{\nu-1}\right) \left(1 + \frac{z_s^2}{\nu}\right) \phi(z_s) + [\hat{y} - \hat{\mu}_{STP}(\mathbf{x})] \Phi(z_s) \tag{4}$$

where: $z_s = \frac{\hat{y} - \hat{\mu}_{STP}(\mathbf{x})}{\hat{\sigma}_{STP}(\mathbf{x})}$; with $\phi(z_s)$ and $\Phi(z_s)$ the respective probability density function (PDF) and cumulative distribution function (CDF) of a univariate, standard Student's-t random variable, z_s. The best y-value sampled by Bayesian optimization is denoted $\hat{y}$. The exploration of Eq. 6 is modelled by $\hat{\sigma}_{STP}(\mathbf{x})\phi(z_s)$, with exploitation modelled by $[\hat{y} - \hat{\mu}_{STP}(\mathbf{x})]\Phi(z_s)$. Eq. 4 can be re-written as:

$$\mathrm{EI}_{STP}(\mathbf{x}) = \hat{\sigma}_{STP}(\mathbf{x})[\left(\frac{\nu + z_s^2}{\nu - 1}\right) \phi(z_s) + z_s \Phi(z_s)] \tag{5}$$

4 Deriving The Exact STP EI Jacobian: STP dEI

Our paper's first knowledge contribution differentiates (using the product rule) Eq. 5 to derive $\frac{\partial \mathrm{EI}_{STP}(\mathbf{x})}{\partial \mathbf{x}}$, the exact Jacobian matrix of first-order partial derivatives of STP EI w.r.t. input $\mathbf{x}$:

$$\begin{aligned} \frac{\partial \mathrm{EI}_{STP}(\mathbf{x})}{\partial \mathbf{x}} &= \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \left[\left(\frac{\nu + z_s^2}{\nu - 1}\right) \phi(z_s) + z_s \Phi(z_s) \right] + \\ \hat{\sigma}_{STP}(\mathbf{x}) \times &\left[\frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) \left(1 - \left(\frac{\nu + z_s^2}{\nu - 1}\right) + \frac{2}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}} \right) \right] \end{aligned} \tag{6}$$

where: $\frac{\partial z_s}{\partial \mathbf{x}} = \left(\frac{\partial \hat{\mu}_{STP}(\mathbf{x})}{\partial \mathbf{x}} - z_s \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \right) / \hat{\sigma}_{STP}(\mathbf{x})$ and $\frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{2\hat{\sigma}_{STP}(\mathbf{x})} \times \frac{\partial \hat{\sigma}^2_{STP}(\mathbf{x})}{\partial \mathbf{x}}$, using Eq. 2 and Eq. 3 above. A full derivation is shown in Appendix A.

A Derivation of STP dEI: exact STP EI gradients

Differentiating Eq. 5:

$$\begin{aligned} \frac{\partial \mathrm{EI}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} &\left[z_s \Phi(z_s) + \left(\frac{\nu + z_s^2}{\nu - 1}\right) \phi(z_s) \right] \\ &+ \hat{\sigma}_{STP}(\mathbf{x}) \frac{\partial [z_s \Phi(z_s) + \left(\frac{\nu + z_s^2}{\nu - 1}\right) \phi(z_s)]}{\partial \mathbf{x}} \end{aligned} \tag{7}$$

The second term in Eq. 7 can be written as:

$$= \hat{\sigma}_{STP}(\mathbf{x}) \frac{\partial [z_s \Phi(z_s) + \left(\frac{\nu}{\nu - 1}\right) \phi(z_s) + \left(\frac{z_s^2}{\nu - 1}\right) \phi(z_s)]}{\partial \mathbf{x}} \tag{8}$$

This means the derivative term in Eq. 8 can be re-written as:

$$= \frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \frac{\partial \Phi(z_s)}{\partial \mathbf{x}} + \left(\frac{\nu + z_s^2}{\nu - 1} \right) \frac{\partial \phi(z_s)}{\partial \mathbf{x}} + \frac{\phi(z_s)}{\nu - 1} \frac{\partial (z_s^2)}{\partial \mathbf{x}} \tag{9}$$

Using $\frac{\partial \phi(z_s)}{\partial \mathbf{x}} = -z_s \phi(z_s)$ and $\frac{\partial \Phi(z_s)}{\partial \mathbf{x}} = \phi(z_s)$, Eq. 9 becomes:

$$= \frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) - \left(\frac{\nu + z_s^2}{\nu - 1} \right) z_s \phi(z_s) + \frac{2 z_s \phi(z_s)}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}}$$

$$= \frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) \left(1 - \left(\frac{\nu + z_s^2}{\nu - 1} \right) + \frac{2}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}} \right) \tag{10}$$

Eq. 8 is now: $\hat{\sigma}_{STP}(\mathbf{x}) \times$ Eq. 10:

$$= \hat{\sigma}_{STP}(\mathbf{x}) \times \left[\frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) \left(1 - \left(\frac{\nu + z_s^2}{\nu - 1} \right) + \frac{2}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}} \right) \right] \tag{11}$$

As Eq. 7's second term equals Eq. 11, Eq. 7 can now be written as:

$$\frac{\partial \mathrm{EI}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \left[\left(\frac{\nu + z_s^2}{\nu - 1} \right) \phi(z_s) + z_s \Phi(z_s) \right] +$$
$$\hat{\sigma}_{STP}(\mathbf{x}) \times \left[\frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) \left(1 - \left(\frac{\nu + z_s^2}{\nu - 1} \right) + \frac{2}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}} \right) \right]$$

This matches Eq. 6 i.e.

$$\frac{\partial \mathrm{EI}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \left[\left(\frac{\nu + z_s^2}{\nu - 1} \right) \phi(z_s) + z_s \Phi(z_s) \right] +$$
$$\hat{\sigma}_{STP}(\mathbf{x}) \times \left[\frac{\partial z_s}{\partial \mathbf{x}} \Phi(z_s) + z_s \phi(z_s) \left(1 - \left(\frac{\nu + z_s^2}{\nu - 1} \right) + \frac{2}{\nu - 1} \frac{\partial z_s}{\partial \mathbf{x}} \right) \right]$$

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. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)

6. 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)
7. Shah, A., Wilson, A.G., Ghahramani, Z.: Bayesian optimization using Student-t Processes. NIPS Workshop on Bayesian Optimization (2013)
8. Shah, A., Wilson, A.G., Ghahramani, Z.: Student-t Processes as alternatives to Gaussian Processes. The Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2014 (2014)
9. Tracey, B., Wolpert, D.: Upgrading from Gaussian Processes to Student's-t Processes. In: 2018 AIAA Non-Deterministic Approaches Conference (2018)