The Exact Expected Regret Minimization 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 [7] and the Matérn class of covariance functions e.g. Matérn 3/2 and Matérn 5/2 [7]. 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 $\mathbf{x}$:

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

2 Student's-t Processes: Partial Derivatives

Using Eq. 1 and [1,3,4,5,8], 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 Regret Minimization with Student's-t Processes

Expected Regret Minimization [6] is a new acquisition function for Bayesian optimisation with STPs [2]. Throughout this paper, we denote it as STP ERM and can define it as:

$$\mathrm{ERM}_{STP}(\mathbf{x}) = \hat{\sigma}_{STP}(\mathbf{x}) \left(\frac{\nu}{\nu-1}\right) \left(1 + \frac{z_{s_{f^*}}^2}{\nu}\right) \phi(z_{s_{f^*}}) + [f^* - \hat{\mu}_{STP}(\mathbf{x})] \Phi(z_{s_{f^*}}) \tag{4}$$

where: $z_{s_{f^*}} = \frac{f^* - \hat{\mu}_{STP}(\mathbf{x})}{\hat{\sigma}_{STP}(\mathbf{x})}$; with $\phi(z_{s_{f^*}})$ and $\Phi(z_{s_{f^*}})$ the respective probability density function (PDF) and cumulative distribution function (CDF) of a univariate, standard Student's-t random variable, $z_{s_{f^*}}$. Prior knowledge of the best-known y-value is denoted f^*. Eq. 4 can be re-written as:

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

4 Deriving The Exact STP ERM Jacobian: STP dERM

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

$$\frac{\partial \mathrm{ERM}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \left[\left(\frac{\nu + z_{s_{f^*}}^2}{\nu - 1}\right) \phi(z_{s_{f^*}}) + z_{s_{f^*}} \Phi(z_{s_{f^*}})\right] +$$
$$\hat{\sigma}_{STP}(\mathbf{x}) \times \left[\frac{\partial z_{s_{f^*}}}{\partial \mathbf{x}} \Phi(z_{s_{f^*}}) + z_{s_{f^*}} \phi(z_{s_{f^*}}) \left(1 - \left(\frac{\nu + z_{s_{f^*}}^2}{\nu - 1}\right) + \frac{2}{\nu - 1} \frac{\partial z_{s_{f^*}}}{\partial \mathbf{x}}\right)\right] \tag{6}$$

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

A Derivation of STP dERM: exact STP ERM gradients

Differentiating Eq. 5:

$$\frac{\partial \mathrm{ERM}_{STP}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \hat{\sigma}_{STP}(\mathbf{x})}{\partial \mathbf{x}} \left[z_{s_{f^*}} \Phi(z_{s_{f^*}}) + \left(\frac{\nu + z_{s_{f^*}}^2}{\nu - 1}\right) \phi(z_{s_{f^*}})\right]$$
$$+ \hat{\sigma}_{STP}(\mathbf{x}) \frac{\partial [z_{s_{f^*}} \Phi(z_{s_{f^*}}) + \left(\frac{\nu + z_{s_{f^*}}^2}{\nu - 1}\right) \phi(z_{s_{f^*}})]}{\partial \mathbf{x}} \tag{7}$$

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

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

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

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

Using $\frac{\partial \phi(z_{s_{f^*}})}{\partial \mathbf{x}} = -z_{s_{f^*}} \phi(z_{s_{f^*}})$ and $\frac{\partial \Phi(z_{s_{f^*}})}{\partial \mathbf{x}} = \phi(z_{s_{f^*}})$, Eq. 9 becomes:

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

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

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

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

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

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

This matches Eq. 6 i.e.

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

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. Clare, C., Hawe, G., McClean, S.: Expected Regret Minimization for Bayesian optimization with Student's-t Processes. In: 3rd International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2020) (2020)
3. Frean, M., Boyle, P.: Using Gaussian Processes to optimize expensive functions. In: AI 2008: Advances in Artificial Intelligence. pp. 258–267 (2008)
4. 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)
5. 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)
6. 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)
7. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
8. 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)