Improved Convergence Rates for Sparse Approximation Methods in Kernel-Based Learning

Sattar Vakili[1], *Jonathan Scarlett* [2], *Da-shan Shiu*[1], *Alberto Bernacchia*[1]

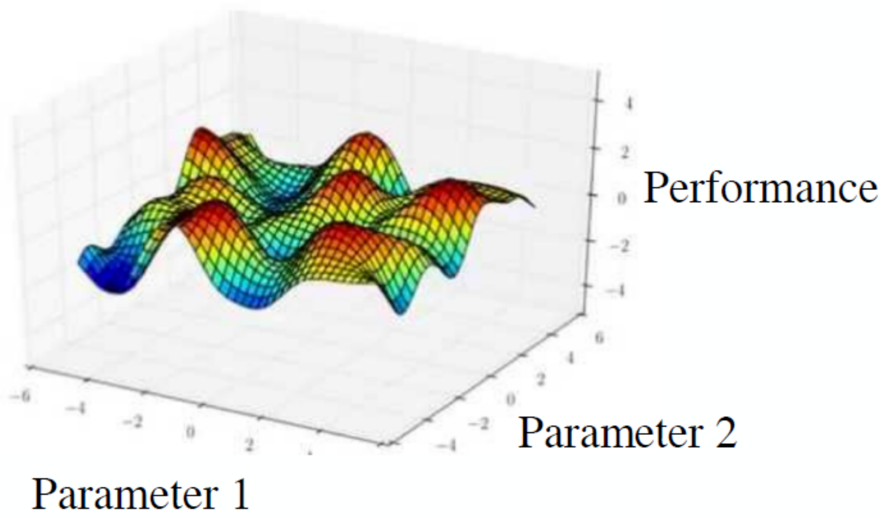
[1]*MediaTek Research*

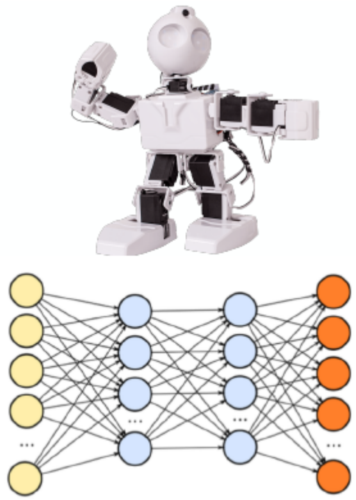
[2]*National University of Singapore*

Introduction

◇ Kernel-based regression is an elegant technique to extend linear models to very general nonlinear ones

◇ Applications: Parameter tuning for robotics/neural nets, molecular design, recommender systems, sensor networks, ...





Kernel Based Models

Provided a data set $\mathcal{D}_n = (X_n, Y_n)$:

Prediction: $$\mu_n(\cdot) = k_{X_n}^\top(\cdot)(\boldsymbol{\tau^2 I_n + K_{X_n,X_n}})^{-1}Y_n$$

Uncertainty: $$k_n(\cdot,\cdot') = k(\cdot,\cdot') - k_{X_n}^\top(\cdot)(\boldsymbol{\tau^2 I_n + K_{X_n,X_n}})^{-1}k_{X_n}(.')$$

◇ Nice closed form expressions

Kernel Based Models

Provided a data set $\mathcal{D}_n = (X_n, Y_n)$:

Prediction: $$\mu_n(\cdot) = k_{X_n}^\top(\cdot)(\boldsymbol{\tau^2 I_n + K_{X_n,X_n}})^{-1} Y_n$$

Uncertainty: $$k_n(\cdot, \cdot') = k(\cdot, \cdot') - k_{X_n}^\top(\cdot)(\boldsymbol{\tau^2 I_n + K_{X_n,X_n}})^{-1} k_{X_n}(.')$$

◇ Nice closed form expressions

◇ A high computational complexity of $\mathcal{O}(n^3)$

Sparse Approximation Methods

Choose a sparse set of inducing points $Z_m = [z_1, z_2, \ldots, z_m]^\top$

Approximate Prediction: $$\tilde{\mu}_n(\cdot) = \underbrace{k_{Z_m}^\top(\cdot)\boldsymbol{(\tau^2 k_{Z_m,Z_m} + k_{X_n,Z_m}^\top k_{X_n,Z_m})^{-1}} k_{X_n,Z_m}^\top}_{V_n^\top} Y_n$$

Approximate Uncertainty: $$\tilde{k}_n(\cdot,\cdot') = k(\cdot,\cdot') - k_{Z_m}^\top(\cdot)\boldsymbol{k_{Z_m,Z_m}^{-1}} k_{Z_m}(.') + k_{Z_m}^\top(\cdot)\boldsymbol{(k_{Z_m,Z_m} + \frac{1}{\tau^2} k_{X_n,Z_m}^\top k_{X_n,Z_m})^{-1}} k_{Z_m}(.')$$

◇ Reduces the computational complexity from $\mathcal{O}(\boldsymbol{n^3})$ to $\mathcal{O}(\boldsymbol{n}m^2 + m^3)$, where in practice $m << n$.

◇ Referred to as SVGP or Nyström method

Posterior Variance of the Approximate GP Model

Theorem: For the posterior variance of the approximate surrogate GP model, we have

$$\tilde{\sigma}_n^2(x) = \sup_{f:||f||_{\mathcal{H}_k}\leq 1}(f(x) - \tilde{f}(x))^2 + \sup_{g:\|g\|_{\mathcal{H}_q}\leq 1}(g(x) - V_n^\top(x)g_{X_n})^2$$
$$+\tau^2\|V_n(x)\|_{l^2}^2$$

- ◇ $V_n^\top(\cdot) = k_{Z_m}^\top(\cdot)\left(\tau^2 k_{Z_m,Z_m} + k_{X_n,Z_m}^\top k_{X_n,Z_m}\right)^{-1} k_{Z_m,X_n}$.
- ◇ Projection error from $\mathcal{H}_k$ to $\mathcal{H}_q$
- ◇ Prediction error from noise free observation within $\mathcal{H}_q$
- ◇ Effect of noise on prediction error

Inducing Points

◇ The theorem holds for any set of inducing points Z_m

◇ For $\tilde{\mu}_n$ to be a good approximation of f, however, the set of inducing points should be selected efficiently

◇ We observe that the effect of inducing points is concisely captured in spectral norm of error in kernel matrix $k_{X_n,X_n} - q_{X_n,X_n}$

◇ Let $\lambda_{\max}$ denote the maximum eigenvalue of $k_{X_n,X_n} - q_{X_n,X_n}$

◇ We next present our confidence interval

Confidence Interval

Theorem: We have the following, each with probability at least $1-\delta$, for a fixed $x \in \mathcal{X}$

$$f(x) - \tilde{\mu}_n(x) \leq \beta(\delta)\tilde{\sigma}_n(x)$$
$$f(x) - \tilde{\mu}_n(x) \geq -\beta(\delta)\tilde{\sigma}_n(x)$$

◇ $\beta(\delta) = \left(\left(2 + \frac{\sqrt{\lambda_{\max}}}{\tau}\right)C_k + \frac{R}{\tau}\sqrt{2\log\left(\frac{1}{\delta}\right)}\right),$

◇ $\|f\|_{\mathcal{H}_k} \leq C_k$

◇ Noise is R sub-Gaussian (can be relaxed)

Uniform Convergence of $\tilde{\mu}_n$ to f

◇ Collect a data $\mathcal{D}_n$ set that is well distributed over the entire domain

◇ $x_i = \arg\max_{x \in \mathcal{X}} \tilde{\sigma}_{i-1}(x)$

Theorem: We have with probability at least $1 - \delta$

$$\|f - \tilde{\mu}_n\|_{L^\infty} = \mathcal{O}\left(\sqrt{\frac{d\gamma_k(n)}{n} \log\left(\frac{n}{\delta}\right)}\right)$$

Comparison with Existing Work

◇ Celebrated work of Burt et al. (2019) bounds the KL divergence between variational and true distribution

◇ Their bounds hold in expectation over a prior distribution on f and dataset

◇ Nieman et al. (2021) proved similar bounds on the KL divergence of the approximate and exact (surrogate) GP posteriors, when f is a fixed function in the RKHS.

◇ Nieman et al. (2021) proved similar results to our theorem, their convergence is in expectation rather than uniformly

◇ Nieman et al. (2021) and we require $m = \tilde{\mathcal{O}}(n^{\frac{d}{2\nu+d}})$ inducing points, while Burt et al. (2019) require $m = \tilde{\mathcal{O}}(n^{\frac{2d}{2\nu-d}})$

Kernel-based Bandit

The performance is typically measured in terms of regret defined as

$$\mathcal{R}(N) = \sum_{n=1}^{N}(f(x^*) - f(x_n)),$$

Optimization Algorithm using Approximate Statistics

Sparse Batch Pure Exploration (S-BPE)

- ◇ An adaptation of BPE (Li and Scarlett, 2021) by replacing the exact GP statistics with approximate ones

- ◇ Proceeds with batches of observations

- ◇ During each batch: $x_j = \arg\max_{x \in \mathcal{X}_i} \tilde{\sigma}_{j-1,i}(x)$

- ◇ At the end of each batch, the points which are unlikely to be the maximizer are removed, using our confidence intervals

Regret Bound for S-BPE

Theorem: We have with probability at least $1 - \delta$

$$\mathcal{R}(N) = \tilde{\mathcal{O}}\left(\sqrt{N d \gamma_k(N) \log\left(\frac{N}{\delta}\right)}\right)$$

References

D. Burt, C. E. Rasmussen, and M. Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 862–871, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. *arXiv preprint arXiv:2110.07788*, 2021.

D. Nieman, B. Szabo, and H. van Zanten. Contraction rates for sparse variational approximations in Gaussian process regression. *arXiv preprint arXiv:2109.10755*, 2021.