Minimax Optimality of Bayesian Process Regression Techniques

Ryan Anderson
ID: 306076860

Daniel Chen
ID: 306072353

June 14, 2023

Abstract

This paper aims to summarize and clarify the results from (Yang & Tokdar, 2015) which derives the minimax L_2 risk for high dimensional non-parametric regression under generalized sparsity assumptions. Yang and Tokdar also show that the sum of selectively rescaled Gaussian processes can achieve minimax contraction rates up to $\log n$ terms when used as a prior for Gaussian process regression.

1 Introduction

One goal of high dimensional statistics is to understand the estimation of nonparametric regression models

$$Y = \mu + f(X_1, ..., X_p) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{1}$$

where minimal assumptions are made on f. In practical settings, the true relationship between Y and its predictors is highly nonlinear and involves interactions between the predictor variables. A known result is that if the only assumption made about target function f is that f is differentiable with smoothness $\alpha > 0$, then the associated minimax rate is $n^{-\alpha/(2\alpha+p)}$ (Stone, 1982). This rate is extremely slow when p is much larger than n such is in most modern high dimensional application. In order to make estimation practical, often sparsity assumptions are made:

$M1 : f$ can depend on all $X = (X1, ..., X_p),$ but X itself lies in a low dimensional manifold M^d.

$M2 : f$ depends on a subset of d predictors with $d \leq \min(n, p)$.

$M3 : f$ depends on $d \asymp \min(n^\gamma, p)$ variables for some $\gamma \in (0, 1)$ but admits an additive structure $f = \sum_{s=1}^{k} f_s,$ where each component function f_s depends on a small d_s number of predictors.

(Yang & Tokdar, 2015) propose $M3$ as a generalization of both the $M2$ assumption (i.e. $k = 1$) as well as the assumption that f is the sum of d univariate functions such as in the case of linear regression (i.e. $d_s = 1$). These sparsity assumptions on f allow the quality of estimation to be quantified through calculating the minimax risk defined as

$$r_n^2(\Sigma, Q, \mu, \sigma) - \inf_{\hat{f} \in A_n} \sup_{f \in \Sigma} E_{f,Q}[||\hat{f} - f||_Q^2] \tag{2}$$

where Σ is the class of functions f belongs to, and Q is some compactly supported probability measure on $\mathbb{R}^p$. A_n is the space of all measurable functions of data to $L_2(Q)$ where $L_2(Q)$ is the linear space of real valued functions on $\mathbb{R}^p$ equipped with the norm induced

by $\langle f, g\rangle_Q = \int f(x)g(x)Q(dx)$. The minimax rate is obtained by treating the minimax risk as a function of the number of observations n as n increases.

Under $M3$, each additive component f_s belongs to the class of centered functions that are α-Hölder smooth on $[0,1]^d$ and bound by λ in Banach space which we denote $\Sigma_s^p(\lambda_s, \alpha_s, d_s)$. The space of functions f belongs to is an additive convolution of multiple $\Sigma_s^p(\lambda_s, \alpha_s, d_s)$ denoted as $\Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d)$ where $\bar{d}$ is the maximum number of f_s that depend on each X_i.

One final assumption is required to achieve an upper and lower bound on the minimax rate.

> Assumption Q : Q admits a probability density function q on $[0,1]^p$ such that $\bar{q} := \sup_x q(x) < \infty$ and $\inf_{x \in [1/2-\Delta, 1/2+\Delta]^p} > 0$ for some $\Delta \in (0, 1/2]$.

The lower bound on q for some sub-hypercube ensures that X can't be reduced to a lower dimension without some loss of information such as in $M1$ assumption. This assumption allows a sharp upper and lower bound to be derived.

2 Bounding the Minimax Risk of Non-Parametric Estimation of Different Function Classes

The natural progression of research in minimax risk calculation gradually relaxes the set of assumptions placed on the target function f^* we are attempting to estimate.

As a starting point, consider the rather strong assumption that the domain of f^* is a manifold X in a low-dimensional space, leading to results from low-rank estimation. Problems like this can be solved with methods like SVD, dyadic decision trees, or nearest neighbor estimation. This is the result for the M1 class of functions defined above.

For the M2 class, we relax our assumptions by instead imposing sparsity on the input, allowing f to depend only on a subset of d predictors, with $d \leq \min(n, p)$. This is the realm of sparse linear regression, and so we can bend the distinction between parametric and nonparametric estimation a bit in order to follow the usual argument summarized in works like Wainwright (2019). With a convex relaxation of the ℓ_0 constraint and then by guaranteeing that the design matrix satisfies the restricted eigenvalue condition, we obtain the following:

Theorem 1 *Consider the Lagrangian lasso with a strictly positive regularization parameter* $\lambda_n \geq 2\|\frac{\mathbb{X}^t w}{n}\|_\infty$. *Then*

- *Any optimal solution* $\hat{\theta}$ *satisfies the bound*

$$\frac{\|\mathbb{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \leq 12\|\theta^*\|_1 \lambda_n$$

- *If* θ^* *is supported on a subset* S *of cardinality* d*, and the design matrix satisfies the* $(\kappa, 3)$*-RE condition over* S*, then any optimal solution satisfies the bound*

$$\frac{\|\mathbb{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{9}{\kappa} s\lambda_n^2$$

It turns out that one such possible regularization parameter is $\lambda_n = 2C\sigma(\sqrt{\frac{2\log(d)}{n}} + \delta)$ is valid, meaning that the minimax rate decays as $d\log(p)/n$, where d is the degree of

sparsity.

The preceding arguments introduce us to how we might approach estimation of the minimax risk for functions in M3. Raskutti, Wainwright, and Yu (2011) analyzes a restricted form of M3, one wherein all functions are univariate with equal sparsities. They find that for α-smooth functions, the minimax rate decays as

$$kn^{-2\alpha/(2\alpha+1)} + \frac{k\log(p)}{n}$$

This is a helpful form for understanding how the minimax rate evolves - for M3 functions, it is composed of the M2-like risk of estimating each of the functions and another term called the variable selection uncertainty. The latter of these can be analogized to the parametric lasso minimax rate obtained above.

Yang and Tokdar (2015) extend the result of Raskutti et al. (2011) to obtain the below result, which characterizes minimax rates for both M2 and M3.

Theorem 2 *Yang & Tokdar 3.1: Under Assumption Q, there exist* $N_0 \in \mathbb{N}, 0 < \underline{C} < 1 < \bar{C}$ *depending only on* $\bar{d}, \max_s d_s, \min_s \alpha_s, \max_s \alpha_s, \min_s \lambda_s, \max_s \lambda_s$ *such that for all* $n < N_0$,

$$\underline{C}\underline{\epsilon}_n^2 \leq r_n^2(\Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d), Q, \mu, \sigma) \leq \bar{C}\bar{\epsilon}_n^2$$

where

$$\underline{\epsilon}_n^2 = \sum_{s=1}^{k} \lambda_s^2 (\frac{\sqrt{n}\lambda_s}{\sigma})^{-4\alpha_s/(2\alpha_s+d_s)} + \frac{\sigma^2 \sum_s d_s}{n} \log \frac{p}{\sum_s d_s}$$

and

$$\bar{\epsilon}_n^2 = \sum_{s=1}^{k} \lambda_s^2 (\frac{\sqrt{n}\lambda_s}{\sigma})^{-4\alpha_s/(2\alpha_s+d_s)} + \frac{\sigma^2 \sum_s d_s}{n} \log \frac{p}{\min_s d_s}$$

As noted above, Theorem 3.1 characterizes both M2 and M3 minimax rates. If we let $k, \bar{d} = 1$ in the above, then we obtain

$$r_n^2(\Sigma_S^p(\lambda,\alpha,d), Q, \mu, \sigma) \asymp \lambda^2 (\frac{\lambda\sqrt{n}}{\sigma})^{-4\alpha/(2\alpha+d)} + \frac{\sigma^2 d}{n} \log(\frac{p}{d})$$

We can use the structure of Theorem 3.1 to do some rudimentary analysis of problems relating to the minimax risk of any given model. In the result just given for M2 functions, we need each of the terms to remain small to control the minimax rate. For the second term, the variable selection uncertainty, that roughly corresponds to having $\log(p) \asymp n^\beta, \beta \in (0,1)$. The first term is the risk of estimating one α-smooth function in d variables, and we can see that given a fixed α, this term is small as long as $d \approx o(n) \approx o(\log(\log(n)))$. This yields the very digestible intuition that meaningful learning in the M2 class is only possible when the number of really important predictors is much smaller than the total predictor count.

Moreover, we note that modeling a given problem as belonging to M3 allows us to have much worse variable structure and still get good error bounds.

Let each component f_s of $\hat{f}$ have the same dimension, smoothness, and magnitude, all of which not depending on k, n – this is a mild generalization of the conditions in Raskutti et al. (2011). Then our minimax rate is $r_n^2 \asymp kn^{-2\alpha/(2\alpha+d)} + kd\log(\frac{p}{d})$. This implies that so long as $k \approx o(\min(n^{-2\alpha/(2\alpha+d)}, \log(\frac{p}{d})))$ we are still well controlled. This means that the total number of predictors kd can be on the order of $\log(p)$, an exponential improvement over M2 modeling.

3 Adaptive Near Minimax Optimal Contraction Rate of Bayesian Additive Gaussian Process Regression

A Gaussian process in a stochastic process such that the value of the process at any finite subset of points on its domain is jointly Gaussian, or equivalently, any finite dimensional marginal of the stochastic process is multivariate normal. Another characterization is that a Gaussian process W is a random element of the supremum norm Banach space of continuous functions over some domain such that any linear functional of the sample is Gaussian. A Gaussian process is uniquely defined by its mean f and covariance Θ where Θ is a kernel that satisfies $\Theta(x, x') = E[(W_x - f(x))(W'_x - f(x')]$.

Standard Gaussian process regression involves defining a Gaussian process as a prior over smooth functions and using the observed data to obtain a posterior distribution. A popular convention is to let the mean of the prior be $f(x) = 0$ and let the covariance be $C^{SE}(x, x') = \exp(-||x - x'||^2)$. A known result for Gaussian process regression is that if the prior $W = GP(0, C^{SE})$ is rescaled such that the rescaled process $W^A = (W_{Ax} : x \in \mathcal{X})$ and A^p follows a gamma distribution then the resulting posterior distribution contracts to the true value f at the minimax rate of $n^{-\alpha/(2\alpha+p)}$ up to a factor of $\log n$ (van der Vaart & van Zanten, 2009).

(Yang & Tokdar, 2015) extend the rescaled Gaussian process method to a method called add-GP where the prior on f is defined by

$$\begin{aligned} &f = L_1 W_1^{A_1,B_1} + ... + L_K W_K^{A_K,B_K}; \quad K \sim \pi, \\ &W_s \overset{iid}{\sim} GP(0, C^{SE}), \quad L \overset{iid}{\sim} h, \\ &B_s \overset{iid}{\sim} [\bigotimes_{j=1}^{p} Be(1/p)]|_{|B| \le D_0}, \quad A_s^{|B_s|}|B_s \overset{iid}{\sim} Ga(a_1, a_2) \end{aligned} \tag{3}$$

where $W \in C(\mathbb{R}^p)$, π is a probability distribution over $\mathbb{N}$, h is a density on $(0, \infty)$, and a_1, a_2, and D_0 are positive hyperparameters. Obtaining a prior of Y also requires μ to be distributed as a Gaussian random variable, and σ to have positive density on $\mathbb{R}^+$.

The add-GP prior is analogous to a prior over sparse functions under $M3$ assumptions. B_s controls which predictors each additive component depends on, and the hyperparameter D_0 ensures that there exists some positive integer $\bar{d}$ such that at most $\bar{d}$ additive components can depend on any predictor.

(Yang & Tokdar, 2015) define the contraction rate of the posterior distribution as ϵ_n if

$$P_n(||\mu + f - \mu^* - f^*||_n + |\sigma - \sigma^*| \ge M\epsilon_n | \{(x^i, Y^i)\}_{i=1}^n) \to 0$$

as $n \to \infty$ for some constant M, and $||f||_n^2 = \frac{1}{n}\sum_{i=1}^n f(x^i)^2$.

Theorem 3 *Yang & Tokdar 4.1: Under assumption Q, for any $\mu^* \in \mathbb{R}, \sigma^* \in$ support(π_σ) and $f^* \in \Sigma_A^{\lambda^*,\alpha^*,\delta^*}$ and with $\max_s d_s^* \le D_0$ and $k < K_0$, the posterior contraction rate at $\theta^* = (\mu^*, f^*, \sigma^*)$ is of order $\epsilon_n(\log n)^{(1+D_0)/2}$ where*

$$\epsilon_n^2 = \sum_{s=1}^{k} \lambda_s^{*2} (\frac{\sqrt{n}\lambda_s^*}{\sigma^*})^{-4\alpha_s^*/(2\alpha_s^* + d_s^*)} (\log n)^{2q_s} + \frac{\sigma^{*2} \sum_s d_s^*}{n} \log p$$

with $q_s = (1 + d_s^*)/(2 + d_s^*/\alpha_s^*), 1 \le s \le k$, provided $K_0 \log p \le n\epsilon_n^2$.

The theorem follows from the existence of spaces of functions $\mathcal{F}_n$ such that the follow-

ing three conditions hold for prior probability P

$$\begin{aligned} P(||\mu + f - \mu^* - f^*||_n \leq \epsilon_n) &\geq \exp(-n\epsilon_n^2), \\ P(\mu + f \notin \mathcal{F}_n) &\leq \exp(-4n\epsilon_n^2), \\ \log N(\bar{\epsilon}_n, \mathcal{F}_n, ||\cdot||_\infty) &\leq n\epsilon_n^2, \end{aligned} \tag{4}$$

where $\hat{\epsilon}_n = \epsilon_n(\log(n)^{(1+D_0)/2})$. The first condition is satisfied by any sequence that asymptotically approaches the definition of ϵ_n as defined in theorem 4.1. The second and third conditions are satisfied when $n^{-\gamma_1} \leq \epsilon_n \leq n^{-\gamma_2}$ for some $\gamma_1 \leq \gamma_2 \leq 1/2$ and $K_0 \log p \leq n\epsilon_n^2$.

4 Discussion and Conclusions

Theorem 4.1 shows that the posterior of an add-GP prior is consistent for the fixed design regression problem which is a useful result. (van der Vaart & van Zanten, 2009) also demonstrates consistency of posteriors for transformed rescaled Gaussian process priors when applied to density estimation and classification problems where more information is known about target distribution f. We can also extend these application to add-GP priors. In the case of density estimation, we can achieve the optimal contraction rate if we let the target function f be equal to $\log(\pi)$ where π is the target density. The prior on π in this case is the exponential transformation of an add GP prior. For classification, if we transform the add GP process using a logistic or normal distribution density, the result is a prior over functions mapping $\mathcal{X} \to (0, 1)$.

Other non-Gaussian process methods of estimation, such as log spline models, are able to achieve an adaptive contraction rate that avoids the $\log n$ factor (Ghosal, Lember, & van der Vaart, 2008). Perhaps extensions of these methods under $M3$ assumptions would be able to achieve the $M3$ minimax rate.

References

Ghosal, S., Lember, J., & van der Vaart, A. (2008). Nonparametric Bayesian Model Selection and Averaging. *Electronic Journal of Statistics*, *2*. doi: 10.1214/07-EJS090

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). *Minimax-optimal rates for sparse additive models over kernel classes via convex programming*.

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Annals of Statistics*, *10*(4). doi: 10.1214/14-AOS1289

van der Vaart, A., & van Zanten, J. H. (2009). Adaptive Bayesian Estimation Using a Gaussian Random Field with Inverse Gamma Bandwidth. *Annals of Statistics*, *37*(5B). doi: 10.1214/08-AOS678

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (1st ed.). Cambridge University Press. Retrieved 2023-04-28, from `https://www.cambridge.org/core/product/identifier/9781108627771/type/book` doi: 10.1017/9781108627771

Yang, Y., & Tokdar, S. T. (2015). Minimax-Optimal Nonparametric Regression in High Dimensions. *Annals of Statistics*, *43*(2). doi: 10.1214/14-AOS1289