

Learning curves of Gaussian process regression with power-law priors and targets

<https://openreview.net/forum?id=KeI9E-gsoB>

Hui Jin¹

Pradeep Kr. Banerjee² and Guido Montúfar^{1,2}

¹Department of Mathematics
University of California, Los Angeles

²Max Planck Institute for
Mathematics in the Sciences

April 2022



UCLA

Main results

- For the realizable case, i.e., when the target function is within the span of the eigenfunctions with positive eigenvalues, we derive the power-law decay rate of the generalization error for Gaussian Process Regression (GPR) and Kernel Ridge Regression (KRR) depending on the eigenspectrum of the prior and the target.
- For the unrealizable case, we show that the generalization error converges to a constant.
- Our results are more general with more realistic assumptions than previous works [Cui et al., 2021, Bordelon et al., 2020].

Motivation

- Training of infinite neural network is equivalent to KRR and GPR in some circumstances [Lee et al., 2019, 2018].

Kernel Learning and Neural Networks

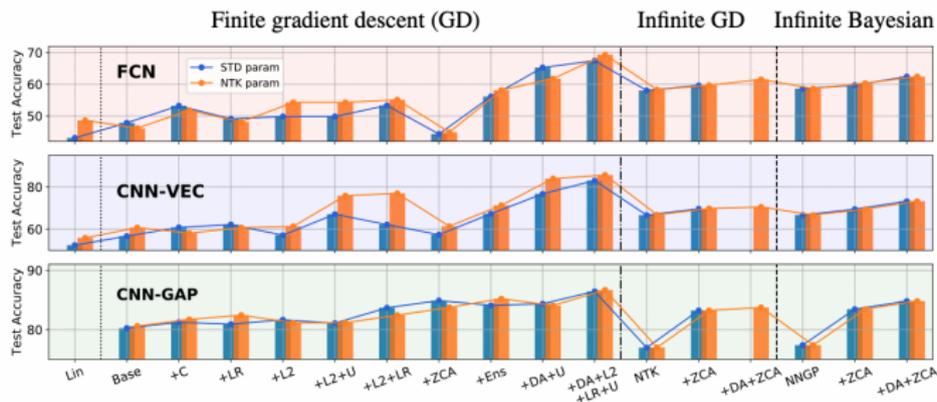


Figure: CIFAR-10 test accuracy for finite and infinite networks, from Lee et al. [2020].

Gaussian Process Regression (GPR)

- Goal: Learn a target function $f: \Omega \mapsto \mathbb{R}$.
- Training samples $D_n = \{(x_i, y_i)\}_{i=1}^n$ generated from an additive noise model $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{true}}^2)$ and $x_i \stackrel{\text{i.i.d.}}{\sim} X$ with pdf $\rho(x)$.
- The true distribution of (x_i, y_i) is $q(x, y) = \rho(x)q(y|x)$, where $q(y|x) = \mathcal{N}(y|f(x), \sigma_{\text{true}}^2)$.
- The *prior* distribution Π_0 over f is defined as a zero-mean GP with covariance function $k: \Omega \times \Omega \rightarrow \mathbb{R}$, i.e., $f \sim \mathcal{GP}(0, k)$.
- The *posterior* distribution over f given training data D_n is

$$d\Pi_n(f|D_n) = \frac{1}{Z(D_n)} \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma_{\text{model}}^2) d\Pi_0(f),$$

where $Z(D_n) = \int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma_{\text{model}}^2) d\Pi_0(f)$ is the *marginal likelihood* or *model evidence* and σ_{model} is the sample variance in GPR.

- In practice, we do not know the exact value of σ_{true} and our choice of σ_{model} can be different from σ_{true} .

Generalization Error

- The GP prior and the Gaussian noise assumption allows for *exact* Bayesian inference.
- The posterior is also a GP with mean and covariance

$$\begin{aligned}\bar{m}(x) &= K_{\mathbf{x}\mathbf{x}}(K_n + \sigma_{\text{model}}^2 I_n)^{-1} \mathbf{y}, \quad x \in \Omega \\ \bar{k}(x, x') &= k(x, x') - K_{\mathbf{x}\mathbf{x}}(K_n + \sigma_{\text{model}}^2 I_n)^{-1} K_{\mathbf{x}x'}, \quad x, x' \in \Omega,\end{aligned}$$

where $K_{\mathbf{x}\mathbf{x}} = K_{\mathbf{x}\mathbf{x}}^T = (k(x_1, x), \dots, k(x_n, x))^T$, $\mathbf{x} = (x_1, \dots, x_n)^T$, $\mathbf{y} = (y_1, \dots, y_n)^T$.

- The *excess mean squared error* is

$$\begin{aligned}M(D_n) &:= \mathbb{E}_{(x_{n+1}, y_{n+1})} (\bar{m}(x_{n+1}) - y_{n+1})^2 - \sigma_{\text{true}}^2 \\ &= \mathbb{E}_{x_{n+1}} (\bar{m}(x_{n+1}) - f(x_{n+1}))^2.\end{aligned}$$

- The *Bayesian generalization error* is KL divergence between true density and predictive density $p_n(y|x, D_n) = \int \mathcal{N}(y|f(x), \sigma_{\text{model}}^2) d\Pi_n(f|D_n)$,

$$G(D_n) := \int q(x, y) \log \frac{q(y|x)}{p_n(y|x, D_n)} dx dy.$$

Equivalence between GPR and Kernel Ridge Regression (KRR)

- The kernel ridge regression (KRR) estimator is the solution to the optimization problem

$$\hat{f} = \arg \min_{g \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^n (g(x_i) - y_i)^2 + \lambda \|g\|_{\mathcal{H}_k}^2.$$

- \mathcal{H}_k is chosen to be an RKHS corresponding to kernel function k .
- The solution is $\hat{f}(x) = K_{\mathbf{x}x}^T (K_n + n\lambda I_n)^{-1} \mathbf{y}$.
- The solution of KRR coincides with posterior mean function of GPR when $\sigma_{\text{model}}^2 = n\lambda$ [Kanagawa et al., 2018].

Spectrum of Kernel and Eigenexpansion of Target Function

- Consider the integral operator corresponding to kernel k :
 $L_k: L^2(\Omega, \rho) \mapsto L^2(\Omega, \rho); (L_k f)(x) = \int_{\Omega} k(x, s)f(s)d\rho(s).$
- Let $(\phi_p(x))_{p \geq 1}$ denote the eigenfunctions of L_k and $(\lambda_p)_{p \geq 1}$ the corresponding positive eigenvalues with $\lambda_1 \geq \lambda_2 \geq \dots > 0$.
- By Mercer's theorem, $k(x_1, x_2) = \sum_{p=1}^{\infty} \lambda_p \phi_p(x_1)\phi_p(x_2).$
- The target $f(x)$ can be decomposed into the orthonormal $(\phi_p(x))_{p \geq 1}$ and its orthogonal complement as

$$f(x) = \sum_{p=1}^{\infty} \mu_p \phi_p(x) + \mu_0 \phi_0(x) \in L^2(\Omega, \rho),$$

where $\mu = (\mu_0, \mu_1, \dots, \mu_p, \dots)^T$ are the coefficients of the decomposition, and $\phi_0(x)$ satisfies $\|\phi_0(x)\|_2 = 1$ and $\phi_0(x) \in \{\phi_p(x) : p \geq 1\}^{\perp}$.

- Let $\Lambda = \text{diag}\{0, \lambda_1, \dots, \lambda_p, \dots\}$ and $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots, \mu_p, \dots)^T$. We show that the generalization error mainly depends on Λ and $\boldsymbol{\mu}$.

Assumptions

- (*Capacity* condition) The eigenvalues $(\lambda_p)_{p \geq 1}$ follow a power law with $\alpha > 1$:

$$\underline{C}_\lambda p^{-\alpha} \leq \lambda_p \leq \overline{C}_\lambda p^{-\alpha}.$$

- (*Source* condition) The coefficients $(\mu_p)_{p \geq 1}$ of the decomposition of the target function follow a power law with $\beta > \frac{1}{2}$:

$$|\mu_p| \leq C_\mu p^{-\beta} \quad \text{and} \quad |\mu_{p_i}| \geq \underline{C}_\mu p_i^{-\beta}, \quad \forall i \geq 1,$$

where $\{p_i\}_{i \geq 1}$ is an increasing integer sequence such that $\sup_{i \geq 1} (p_{i+1} - p_i) < \infty$.

- The eigenfunctions $(\phi_p(x))_{p \geq 1}$ satisfy $\|\phi_p\|_\infty \leq C_\phi p^\tau$, $\forall p \geq 1$ with $\tau < \frac{\alpha-1}{2}$.

Assumptions

- Related to the effective dimension of the problem and the difficulty of learning the target function [Caponnetto and De Vito, 2007, Blanchard and Mücke, 2018].
- Velikanov and Yarotsky [2021]:
 - ▶ Derived the exact value of α when the kernel function has a homogeneous singularity on its diagonal, e.g., the arc-cosine kernel.
 - ▶ Gave examples of functions for which *source* condition is satisfied, such as functions that have a bounded support with smooth boundary and are smooth on the interior of this support, and derived the corresponding β .
- Ronen et al. [2019] showed that for inputs distributed uniformly on a hypersphere, the eigenfunctions of the arc-cosine kernel are spherical harmonics and the eigenvalues follow a power-law decay.

Main result I

Theorem (Asymptotics of excess mean squared error)

Assume $\sigma_{\text{model}}^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Then with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$, where $0 \leq q < \frac{[\alpha - (1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$, the excess mean squared error has the asymptotic:

$$\begin{aligned}\mathbb{E}_{\epsilon} M(D_n) &= (1 + o(1)) \left[\frac{\sigma_{\text{true}}^2}{\sigma_{\text{model}}^2} \left(\text{Tr} \left(I + \frac{n}{\sigma_{\text{model}}^2} \Lambda \right)^{-1} \Lambda - \|\Lambda^{1/2} \left(I + \frac{n}{\sigma_{\text{model}}^2} \Lambda \right)^{-1}\|_F^2 \right) \right. \\ &\quad \left. + \left\| \left(I + \frac{n}{\sigma_{\text{model}}^2} \Lambda \right)^{-1} \boldsymbol{\mu} \right\|_2^2 \right] \\ &= \Theta \left(\max \left\{ \sigma_{\text{true}}^2 n \frac{1-\alpha-t}{\alpha}, n \frac{(1-2\beta)(1-t)}{\alpha} \right\} \right)\end{aligned}$$

when $\mu_0 = 0$, and $\mathbb{E}_{\epsilon} M(D_n) = \mu_0^2 + o(1)$, when $\mu_0 > 0$.

- Exponent $\frac{1-\alpha-t}{\alpha}$ captures rate at which model suppresses noise.
- Exponent $\frac{(1-2\beta)(1-t)}{\alpha}$ captures rate at which model learns target function.

Related results

- Cui et al. [2021] derived similar asymptotics for KRR with Gaussian design, where $\Lambda_R^{1/2}(\phi_1(x), \dots, \phi_R(x))$ is assumed to follow a Gaussian distribution $\mathcal{N}(0, \Lambda_R)$.
 - ▶ Our assumption that the eigenfunctions are bounded by power functions is more general.
 - ▶ Our result is high probability result and is stronger than the expectation result of Cui et al. [2021].
- In the noiseless setting ($\sigma_{\text{true}} = 0$) with constant regularization ($t = 0$), Bordelon et al. [2020] showed that the mean squared error behaves as $\Theta(n^{\frac{1-2\beta}{\alpha}})$.
 - ▶ Our result is applicable to noisy data and non-constant regularization.
- By leveraging the equivalence between GPR and KRR, we can get the same statement for the generalization error of GPR.

Main result II

Theorem (Asymptotics of the Bayesian generalization error, $\mu_0 = 0$)

Assume that $\mu_0 = 0$ and $\sigma_{\text{model}}^2 = \sigma_{\text{true}}^2 = \sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Then with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$ where $0 \leq q < \frac{[\alpha - (1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$, the expectation of the Bayesian generalization error w.r.t. the noise ϵ has the asymptotic behavior:

$$\begin{aligned}\mathbb{E}_{\epsilon} G(D_n) &= \frac{1+o(1)}{2\sigma^2} \left(\text{Tr}(I + \frac{n}{\sigma^2} \Lambda)^{-1} \Lambda - \|\Lambda^{\frac{1}{2}}(I + \frac{n}{\sigma^2} \Lambda)^{-1}\|_F^2 + \|(I + \frac{n}{\sigma^2} \Lambda)^{-1} \boldsymbol{\mu}\|_2^2 \right) \\ &= \frac{1}{\sigma^2} \Theta(n^{\max\{\frac{(1-\alpha)(1-t)}{\alpha}, \frac{(1-2\beta)(1-t)}{\alpha}\}}).\end{aligned}$$

- Sollich and Halees [2002] obtained a corresponding result for the particular case when $f \sim \mathcal{GP}(0, k)$ and $t = 0$.

Main result III

Theorem (Asymptotics of the Bayesian generalization error, $\mu_0 > 0$)

Assume that $\mu_0 > 0$ and $\sigma_{\text{model}}^2 = \sigma_{\text{true}}^2 = \sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Then with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$, where $0 \leq q < \frac{[\alpha - (1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$, the expectation of the Bayesian generalization error w.r.t. the noise ϵ has the asymptotic behavior:

$$\mathbb{E}_{\epsilon} G(D_n) = \frac{1}{2\sigma^2} \mu_0^2 + o(1).$$

- In general, if $\mu_0 > 0$, the generalization error asymptotes to a constant.
- Hence, GPR can only learn functions within the span of eigenfunctions.

Experiments

Cho and Saul [2009] showed that arc-cosine kernel is the NNGP kernel of an infinitely wide shallow ReLU network. Let the input x be uniformly distributed on a unit circle, i.e., $\Omega = S^1$ and $\rho = \mathcal{U}(S^1)$.

	kernel function	α	activation function	bias
$k_{w/o \text{ bias}}^{(1)}$	$\frac{1}{\pi}(\sin \psi + (\pi - \psi) \cos \psi)$	4	$\max\{0, x\}$	no
$k_{w \text{ bias}}^{(1)}$	$\frac{1}{\pi}(\sin \bar{\psi} + (\pi - \bar{\psi}) \cos \bar{\psi})$	4	$\max\{0, x\}$	yes
$k_{w/o \text{ bias}}^{(2)}$	$\frac{1}{\pi}(3 \sin \psi \cos \psi + (\pi - \psi)(1 + 2 \cos^2 \psi))$	6	$(\max\{0, x\})^2$	no
$k_{w \text{ bias}}^{(2)}$	$\frac{1}{\pi}(3 \sin \bar{\psi} \cos \bar{\psi} + (\pi - \bar{\psi})(1 + 2 \cos^2 \bar{\psi}))$	6	$(\max\{0, x\})^2$	yes
$k_{w/o \text{ bias}}^{(0)}$	$\frac{1}{\pi}(\sin \psi + (\pi - \psi) \cos \psi)$	2	$\frac{1}{2}(1 + \text{sign}(x))$	no
$k_{w \text{ bias}}^{(0)}$	$\frac{1}{\pi}(\sin \bar{\psi} + (\pi - \psi) \cos \bar{\psi})$	2	$\frac{1}{2}(1 + \text{sign}(x))$	yes

Table: The different kernel functions of infinite shallow networks, their values of α , the corresponding neural network activation function. Here $\psi = \arccos(\langle x_1, x_2 \rangle)$ and $\bar{\psi} = \arccos(\frac{1}{2}(\langle x_1, x_2 \rangle + 1))$.

- Kernels corresponding to smoother activation function have faster decay rate of the eigenvalues.
- Networks with smoother activation function are better at compressing the noise, but less capable of fitting functions.

Experiments

	function value	β	μ_0	$\mathbb{E}_\epsilon G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{-3/4})$
f_2	θ^2	2	> 0	$\Theta(1)$
f_3	$(\theta - \pi/2)^2$	2	0	$\Theta(n^{-3/4})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{-1/4})$

Table: Target functions used in the experiments for the first order arc-cosine kernel without bias $k_{w/o \text{ bias}}^{(1)}$, their values of β and μ_0 , and theoretical rates for the Bayesian generalization error from our theorems.

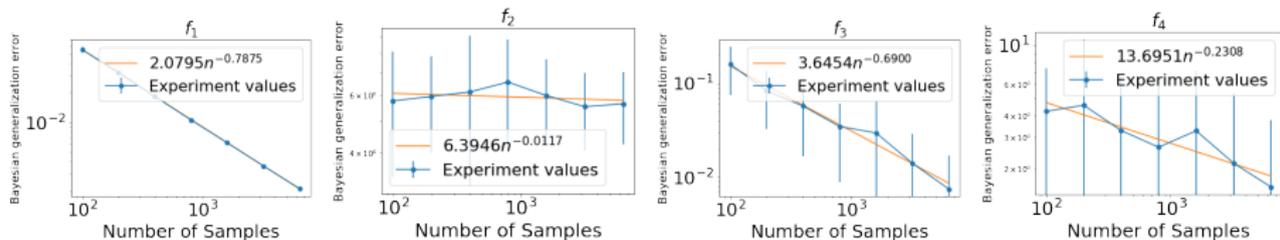


Figure: Bayesian generalization error for GPR with the kernel $k_{w/o \text{ bias}}^{(1)}$ and the target functions. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n .

Experiments

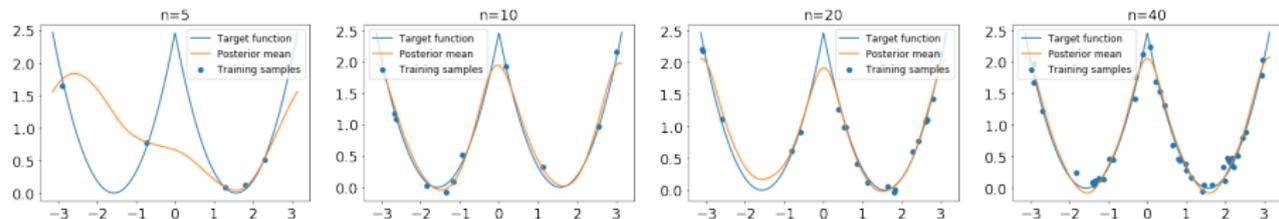


Figure: Experiment on the first order arc-cosine kernel without bias $k_{w/o \text{ bias}}^{(1)}$. Blue curve is the target function $f(\theta) = (|\theta| - \pi/2)^2$. Orange curve is the posterior mean and blue points are training samples.

Experiments

	function value	β	μ_0	$\mathbb{E}_\epsilon G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{-3/4})$
f_2	θ^2	2	0	$\Theta(n^{-3/4})$
f_3	$(\theta - \pi/2)^2$	2	0	$\Theta(n^{-3/4})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{-1/4})$

Table: Target functions used in the experiments for the first order arc-cosine kernel with bias, $k_w^{(1)}$, their values of β and μ_0 , and theoretical rates for the Bayesian generalization error from our theorems.

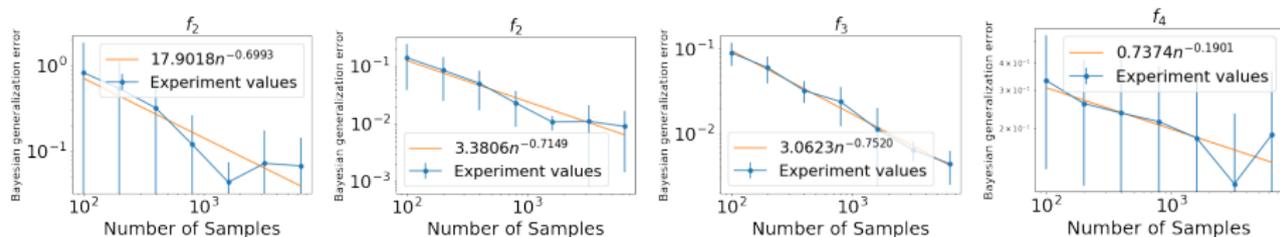


Figure: Bayesian generalization error for GPR with kernel $k_w^{(1)}$ and the target functions. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of $\log n$.

Experiments

	function value	β	μ_0	$\mathbb{E}_\epsilon G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{-5/6})$
f_2	$\text{sign}(\theta)$	1	0	$\Theta(n^{-1/6})$
f_3	$\pi/2 - \theta $	2	0	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	> 0	$\Theta(1)$

Table: Target functions used in the experiments for the second order arc-cosine kernel without bias, $k_{w/o \text{ bias}}^{(2)}$, their values of β and μ_0 , and theoretical rates for the Bayesian generalization error from our theorems.

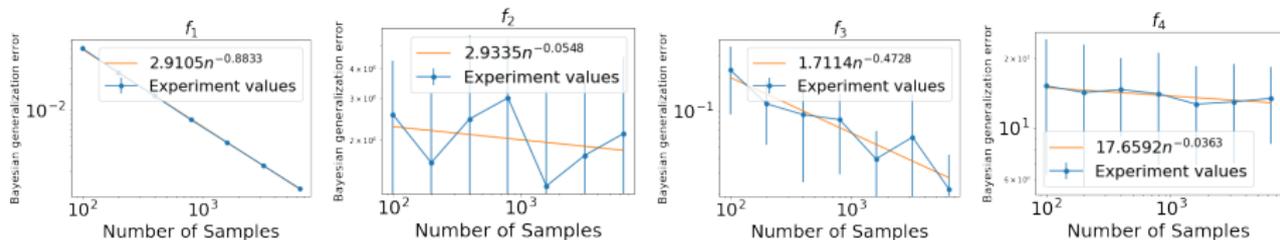


Figure: Bayesian generalization error for GPR with kernel $k_{w/o \text{ bias}}^{(2)}$ and the target functions.

Experiments

	function value	β	μ_0	$\mathbb{E}_\epsilon G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{-5/6})$
f_2	θ^2	2	0	$\Theta(n^{-1/2})$
f_3	$(\theta - \pi/2)^2$	2	0	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{-1/6})$

Table: Target functions used in the experiments for the second order arc-cosine kernel with bias, $k_w^{(2)} \text{bias}$, their values of β and μ_0 , and theoretical rates for the Bayesian generalization error from our theorems.

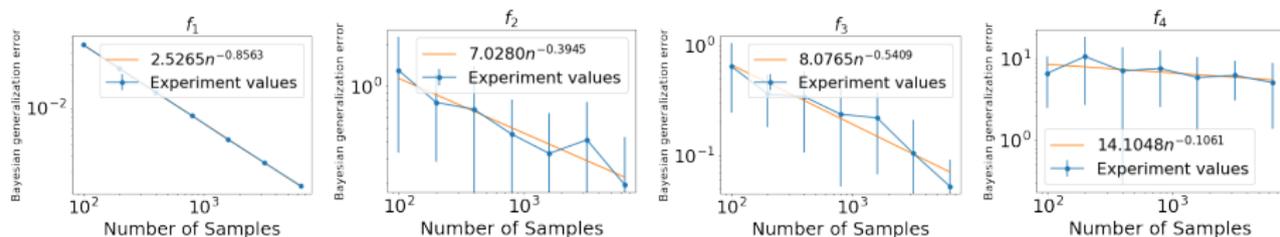


Figure: Bayesian generalization error for GPR with kernel $k_w^{(2)} \text{bias}$ and the target functions.

Conclusion and Future Work

- 1 Described the learning curves for GPR for the case that the kernel and target function follow a power law.
 - ▶ This setting is frequently encountered in kernel learning literatures.
 - ▶ The result can be applied to infinite neural networks.
- 2 In future work, it will be interesting to estimate the values of α and β for some specific settings.
 - ▶ The Neural Tangent Kernel (NTK) of deep fully-connected or convolutional neural networks.
 - ▶ Analyze the effect of data distribution.

References

- Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.
- Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1024–1034, 2020.
- Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 342–350, 2009.
- Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *arXiv preprint arXiv:2105.15004*, 2021.
- Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.
- Jaehoon Lee, Jascha Sohl-Dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.
- Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pages 8572–8583, 2019.
- Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *Advances in Neural Information Processing Systems*, 33:15156–15172, 2020.
- Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32:4761–4771, 2019.
- Peter Sollich and Anason Halees. Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.
- Maksim Velikanov and Dmitry Yarotsky. Universal scaling laws in the gradient descent training of neural networks. *arXiv preprint arXiv:2105.00507*, 2021.