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

Hui Jin

Department of Mathematics
University of California, Los Angeles

February 10, 2022

Joint work with Pradeep Banerjee and Guido Montúfar

To appear at ICLR 2022
https://openreview.net/forum?id=KeI9E-gsoB

Gaussian Process

- A *Gaussian Process* is a continuous stochastic process $fY_x : x \; 2 \; Xg$ where

$$Y_{x_1,\ldots,x_n} = (Y_{x_1}, \ldots, Y_{x_n})$$

  is a multivariate Gaussian random variable.
- Gaussian process $GP(m, k)$ is completely defined by its mean function $m(x)$ and covariance function $k(x, x^0)$.
- For every finite set of indices $x_1, \ldots, x_n$, we have $Y_{x_1,\ldots,x_n} \quad N(m_n, K_n)$, where

$$m_n = (m(x_1), \ldots, m(x_n))^T, \quad K_n = \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix}$$

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 $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{true}})$ and $x_i \overset{\text{i.i.d.}}{\sim} X$ with pdf $\rho(x)$.
- The true distribution of (x_i, y_i) is $q(x, y) = \rho(x)q(y|x)$, where $q(y|x) = N(y|f(x), \sigma^2_{\text{true}})$.
- The *prior* distribution Π_0 over f is defined as a zero-mean GP with covariance function $k : \Omega \times \Omega \to \mathbb{R}$, i.e., $f \sim GP(0, k)$.
- The *posterior* distribution over f given training data D_n is

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

where $Z(D_n) = \int \prod_{i=1}^n N(y_i|f(x_i), \sigma^2_{\text{model}}) d\Pi_0(f)$ 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

$$\bar{m}(x) = K_{x\mathbf{x}}(K_n + \sigma^2_{\text{model}} I_n)^{-1}\mathbf{y}, \; x \in \Omega$$
$$\bar{k}(x, x') = k(x, x') - K_{x\mathbf{x}}(K_n + \sigma^2_{\text{model}} I_n)^{-1} K_{\mathbf{x}x'}, \; x, x' \in \Omega,$$

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

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

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

- The *excess mean squared error* is

$$M(D_n) := \mathsf{E}_{(x_{n+1}, y_{n+1})}(\bar{m}(x_{n+1}) - y_{n+1})^2 - \sigma^2_{\text{true}}$$
$$= \mathsf{E}_{x_{n+1}}(\bar{m}(x_{n+1}) - f(x_{n+1}))^2.$$

Equivalence between GPR and Kernel Ridge Regression (KRR)

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

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

- 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 $\sigma^2_{\text{model}} = n\lambda$ [Kanagawa et al., 2018].

Kernel Learning and Neural Networks

- Training of infinite neural network is equivalent to kernel learning in some circumstances [Lee et al., 2019, 2018].
- Cho and Saul [2009] showed that arc-cosine kernel is the NNGP kernel of an infinitely wide shallow ReLU network with two inputs and no biases in the hidden layer.

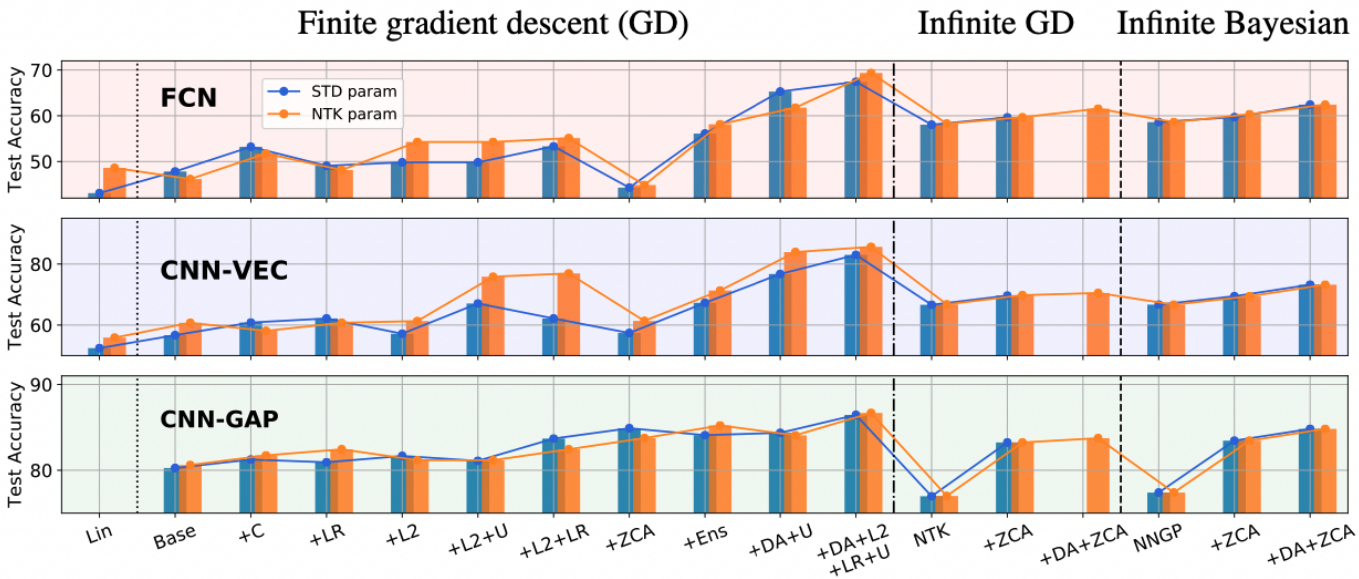

Figure: CIFAR-10 test accuracy for finite and infinite networks, from Lee et al. [2020].

Spectrum of Kernel and Eigenexpansion of Target Function

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

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

  where $\mu = (\mu_0, \mu_1, \ldots, \mu_p, \ldots)^T$ are the coefficients of the decomposition, and $\phi_0(x)$ satisfies $k\phi_0(x)k_2 = 1$ and $\phi_0(x) \; 2 \; f\phi_p(x) : p \quad 1g^?$.
- Let $\Lambda = \mathrm{diag} f0, \lambda_1, \ldots, \lambda_p, \ldots g$ and $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_p, \ldots)^T$. We show that the generalization error mainly depends on Λ and $\boldsymbol{\mu}$.

Assumptions

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

$$\underline{C_\lambda} p^{\ -\alpha} \quad \lambda_p \quad \overline{C_\lambda} p^{\ -\alpha}.$$

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

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

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

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

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

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

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

$$\mathsf{E}_\epsilon G(D_n) = \frac{1+o(1)}{2\sigma^2}\left(\mathrm{Tr}(I + \tfrac{n}{\sigma^2}\Lambda)^{\ 1}\Lambda \quad k\Lambda^{\frac{1}{2}}(I + \tfrac{n}{\sigma^2}\Lambda)^{\ 1}k_F^2 + k(I + \tfrac{n}{\sigma^2}\Lambda)^{\ 1}\mu k_2^2\right)$$
$$= \frac{1}{\sigma^2}\Theta(n^{\max\{\frac{(1 \quad \alpha)(1 \quad t)}{\alpha}, \frac{(1 \quad 2\beta)(1 \quad t)}{\alpha}\}}).$$

- The exponent $\frac{(1 \quad \alpha)(1 \quad t)}{\alpha}$ captures the rate at which the model suppresses the noise.
- The exponent $\frac{(1 \quad 2\beta)(1 \quad t)}{\alpha}$ captures the rate at which the model learns the target function.
- Sollich and Halees [2002] obtained a corresponding result for the particular case when $f \quad GP(0, k)$ and $t = 0$.

Main result

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

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

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

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

Main result

Theorem (Asymptotics of excess mean squared error)

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

$$\mathsf{E}_{\epsilon} M(D_n) = (1+o(1))\left[\frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{model}}}\left(\text{Tr}(I+\frac{n}{\sigma^2_{\text{model}}}\Lambda)^{\ 1}\Lambda \quad k\Lambda^{1/2}(I+\frac{n}{\sigma^2_{\text{model}}}\Lambda)^{\ 1}k_F^2\right)\right.$$
$$\left. + k(I+\frac{n}{\sigma^2_{\text{model}}}\Lambda)^{\ 1}\mu k_2^2\right]$$
$$= \Theta\left(\max\{\sigma^2_{\text{true}} n^{\frac{1 \quad \alpha \quad t}{\alpha}}, n^{\frac{(1 \quad 2\beta)(1 \quad t)}{\alpha}}\}\right)$$

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

- In this result σ^2_{model} and σ^2_{true} can be different.
- The asymptotic of excess mean squared error is the same as Bayesian generalization error when $\sigma^2_{\text{model}} = \sigma^2_{\text{true}}$.

Related results

- By leveraging the equivalence between GPR and KRR, we can get the same statement for the generalization error of KRR.

- Cui et al. [2021] derived similar asymptotics for KRR with Gaussian design, where $\Lambda_R^{1/2}(\phi_1(x), \ldots, \phi_R(x))$ is assumed to follow a Gaussian distribution $N(0, \Lambda_R)$.
  - 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 $\Theta(n^{\frac{1}{\alpha}\frac{2\beta}{\alpha}})$.
  - Our result is applicable to noisy data and non-constant regularization.

Experiments

Let the input x be uniformly distributed on a unit circle, i.e., $\Omega = S^1$ and $\rho = U(S^1)$.

	kernel function	α	activation function	bias
$k^{(1)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi - \psi)\cos\psi)$	4	$\max\{0, x\}$	no
$k^{(1)}_{\text{w bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi - \bar{\psi})\cos\bar{\psi})$	4	$\max\{0, x\}$	yes
$k^{(2)}_{\text{w/o bias}}$	$\frac{1}{\pi}(3\sin\psi\cos\psi + (\pi - \psi)(1 + 2\cos^2\psi))$	6	$(\max\{0, x\})^2$	no
$k^{(2)}_{\text{w bias}}$	$\frac{1}{\pi}(3\sin\bar{\psi}\cos\bar{\psi} + (\pi - \bar{\psi})(1 + 2\cos^2\bar{\psi}))$	6	$(\max\{0, x\})^2$	yes
$k^{(0)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi - \psi)\cos\psi)$	2	$\frac{1}{2}(1 + \text{sign}(x))$	no
$k^{(0)}_{\text{w bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi - \psi)\cos\bar{\psi})$	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 $\bar{\psi} = \arccos\left(\frac{1}{2}(\langle x_1, x_2\rangle + 1)\right)$.

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

Experiments

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

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

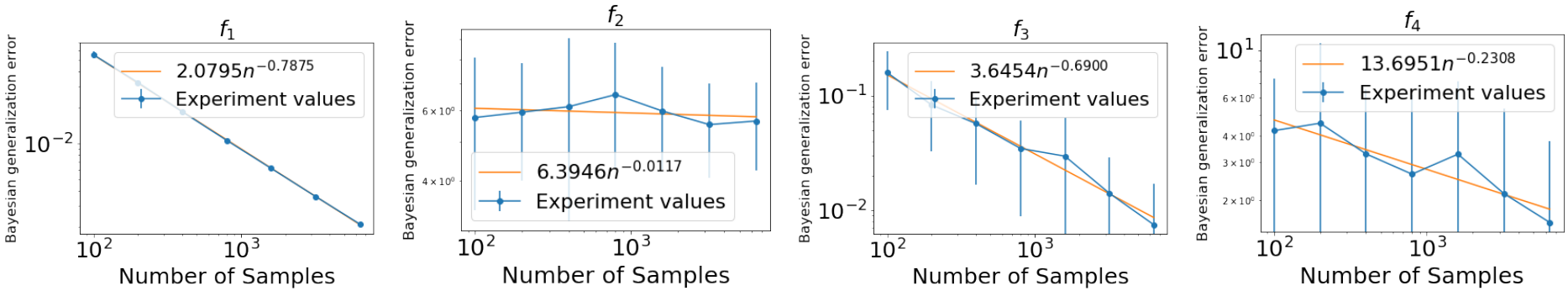

Figure: Bayesian generalization error for GPR with the kernel $k^{(1)}_{\text{w/o bias}}$ and the target functions. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

Experiments

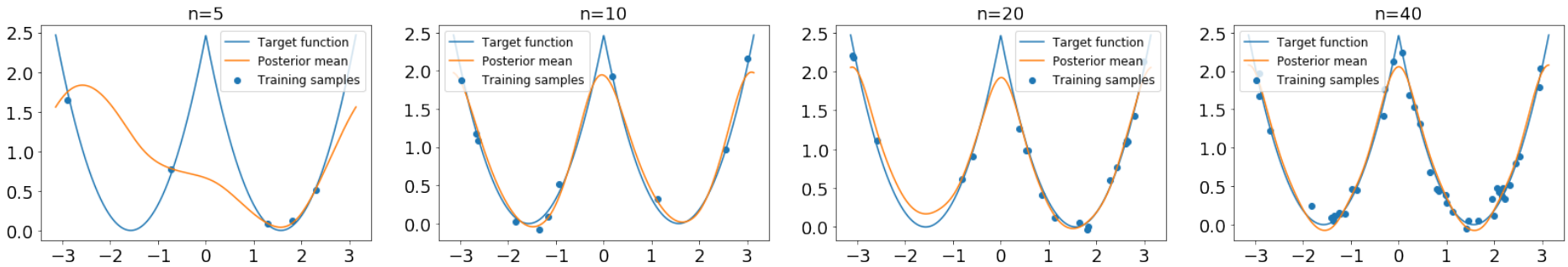

Figure: Experiment on the first order arc-cosine kernel without bias $k^{(1)}_{\text{w/o bias}}$. Blue curve is the target function $f(\theta) = (j\theta j \quad \pi/2)^2$. Orange curve is the posterior mean and blue points are training samples.

Experiments

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

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

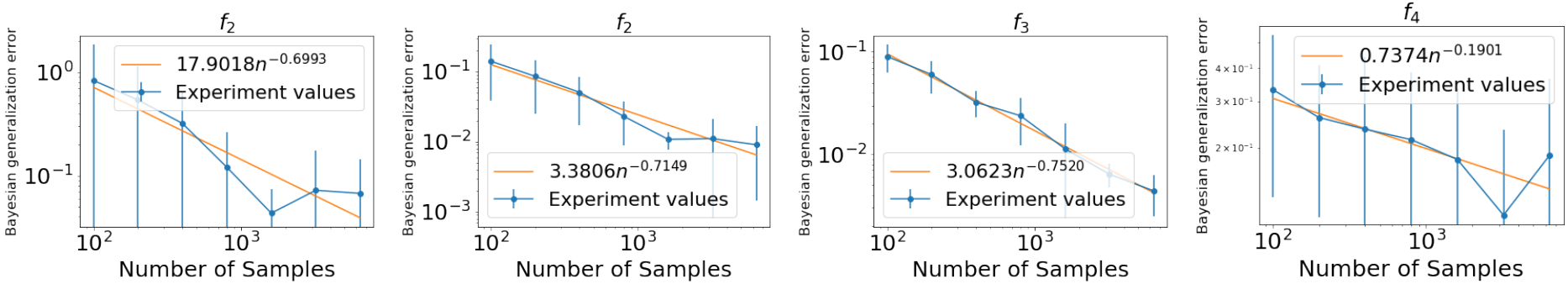

Figure: Bayesian generalization error for GPR with kernel $k^{(1)}_{\mathrm{w\ bias}}$ and the target functions. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

Experiments

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

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

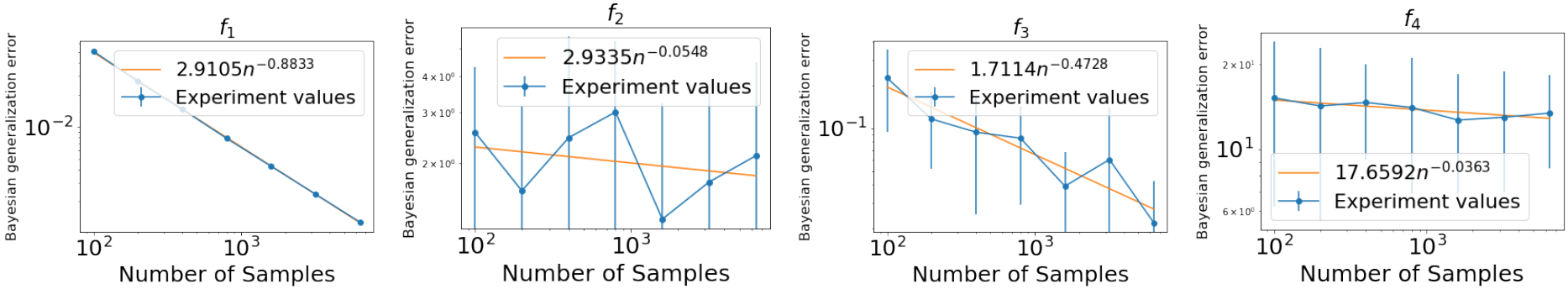

Figure: Bayesian generalization error for GPR with kernel $k^{(2)}_{\text{w/o bias}}$ and the target functions.

Experiments

	function value	β	μ_0	$\mathsf{E}_\epsilon G(D_n)$
f_1	$\cos 2\theta$	$+1$	0	$\Theta(n^{-5/6})$
f_2	θ^2	2	0	$\Theta(n^{-1/2})$
f_3	$(j\theta j - \pi/2)^2$	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, $k^{(2)}_{\mathrm{w\ bias}}$, their values of β and μ_0, and theoretical rates for the Bayesian generalization error from our theorems.

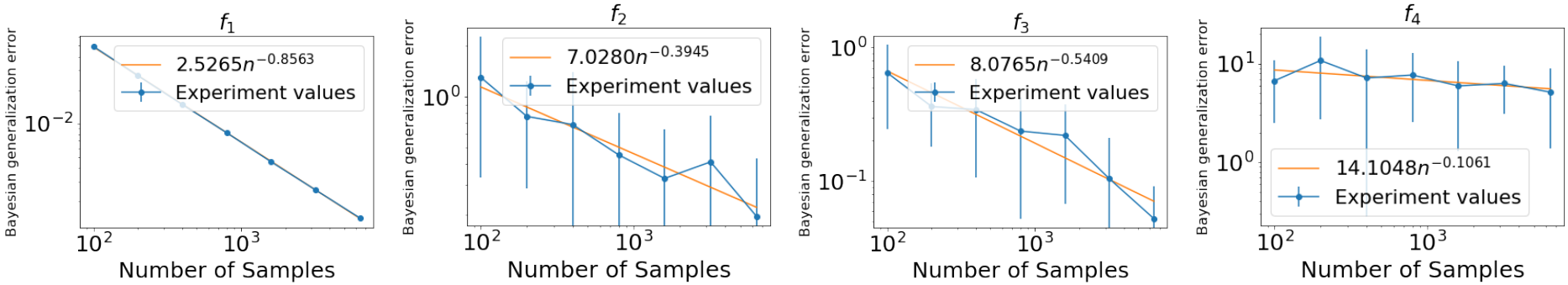

Figure: Bayesian generalization error for GPR with kernel $k^{(2)}_{\mathrm{w\ bias}}$ and the target functions.

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 speific 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.