A Technical Lemmas

Lemma 6. *Let $p_n(x) = \sum_{i=0}^{n} \frac{1}{i!} x^i$ be the degree-n Taylor polynomial of e^x around $x = 0$. Then for any $k \geq 1$ and any $x \in \mathbb{R}$, we have*

$$e^x \geq p_{2k-1}(x).$$

Proof. The proof is an induction on k. The base case $k = 1$ is trivial: $e^x \geq 1 + x$.

Consider the general case p_{2k+1}. Define $g(x) = e^x - p_{2k+1}(x)$. It is easy to see that $g'(x) = e^x - p_{2k}(x)$ and $g''(x) = e^x - p_{2k-1}(x)$. By the induction hypothesis, $g'' \geq 0$ and therefore g is convex. Thus, the minimum of g is given by its stationary points. It is easy to observe that $x = 0$ is indeed a stationary point. Thus, $\min_{x \in \mathbb{R}} g(x) = g(0) = 0$, which finishes the proof. □

Lemma 7. *Let $\sigma > 0$ and W be the principal branch of the Lambert W function. For any $m \in \mathbb{N}$, we have*

$$1 + W(-\frac{m}{e(m + \sigma^2)}) \leq \sqrt{\frac{2\sigma^2}{m + \sigma^2}}.$$

Proof. The problem is reduced to proving

$$1 + W(x) \leq \sqrt{2(1 + ex)}$$

for all $-\frac{1}{e} \leq x \leq 0$. In fact, the right hand side is exactly the first-order Taylor expansion of the left hand side at $x = -\frac{1}{e}$ [*e.g.*, 4].

Square both sides. It is equivalent to prove $(1 + W(x))^2 \leq 2(1 + ex)$ for all $x \in [-\frac{1}{e}, 0]$. Define

$$g(x) = (1 + W(x))^2 - 2(1 + ex).$$

The derivative of g in $(-\frac{1}{e}, 0)$ is

$$g'(x) = 2\bigg(\frac{W(x)}{x} - e\bigg).$$

By the definition of the Lambert W function, we have

$$ex = W(x)e^{1+W(x)} < W(x)$$

since $-1 < W(x) < 0$ for $x \in (-\frac{1}{e}, 0)$. Thus, $g'(x) < 0$ for $x \in (-\frac{1}{e}, 0)$ and g is a decreasing function in $[-\frac{1}{e}, 0]$. Observe that $g(-\frac{1}{e}) = 0$. Therefore $g(x) \leq 0$ for all $x \in [-\frac{1}{e}, 0]$, which completes the proof. □

Next, we present a lemma which states that the error function bounds the posterior covariance trace in each iteration of Algorithm 1.

Lemma 8. *In the t-th iteration of Algorithm 1, we have*

$$\operatorname{tr}\big(\nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t)\nabla^\top\big) \leq E_{d,k,\sigma}(b_t)$$

Proof. Without loss of generality, we assume $\mathbf{x}_t = \mathbf{0}$. Otherwise, shift the data $\mathcal{D}_t$ and $\mathbf{x}_t$ by $-\mathbf{x}_t$, which does not change the value of the left hand side becauce of stationarity of the kernel k. Let $\mathbf{Z} \in \mathbb{R}^{b_t \times d}$ be arbitrary candidates. Then, we have

$$\operatorname{tr}\big(\nabla k_{\mathcal{D}_{t-1} \cup \mathbf{Z}}(\mathbf{0}, \mathbf{0})\nabla^\top\big) \leq \operatorname{tr}\big(\nabla k(\mathbf{0}, \mathbf{0})\nabla^\top - \nabla k(\mathbf{0}, \mathbf{Z})(k(\mathbf{Z}, \mathbf{Z}) + \sigma^2\mathbf{I})^{-1}k(\mathbf{Z}, \mathbf{0})\nabla^\top\big).$$

Because the LHS conditions on both $\mathcal{D}_{t-1}$ and $\mathbf{Z}$ but the RHS only conditions on $\mathbf{Z}$. Now, we minimize $\mathbf{Z}$ on both sides.

On the one hand, the LHS becomes $\operatorname{tr}\big(\nabla k_{\mathcal{D}_t}(\mathbf{0}, \mathbf{0})\nabla^\top\big)$. This is because $\mathcal{D}_t$ is the union of $\mathcal{D}_{t-1}$ and the minimizer of the acquisition function $\alpha_{\mathrm{trace}}(\mathbf{0}, \mathbf{Z}) = \operatorname{tr}\big(\nabla k_{\mathcal{D}_{t-1} \cup \mathbf{Z}}(\mathbf{0}, \mathbf{0})\nabla^\top\big)$.

On the other hand, the RHS becomes $E_{d,k,\sigma}(b_t)$ by definition of the error function, which completes the proof.

Note that the edge case, where the minimizer of the acquisition function $\operatorname{argmin}_{\mathbf{Z}} \alpha_{\mathrm{trace}}(\mathbf{0}, \mathbf{Z})$ does not exist (*e.g.*, when $\sigma = 0$), can be handled by a careful limiting argument using the same idea. □

A.1 Lemmas for the RKHS Assumption

By Assumption 1, the kernel function k is four times continuously differentiable. The following lemma asserts the smoothness of $f \in \mathcal{H}$.

Lemma 9. *Suppose $f \in \mathcal{H}$ maps from a compact domain $\mathcal{X}$ to $\mathbb{R}$. Then f is L-smooth for some L.*

Proof. Since the kernel k is four times continuously differentiable, f is twice continuously differentiable. On a compact domain $\mathcal{X}$, the spectral norm of the Hessian $\|\nabla f(\mathbf{x})\|$ has a maximizer. Define $L = \max_{\mathbf{x}\in\mathcal{X}} \|\nabla f(\mathbf{x})\|$. Then f is L-smooth. □

Lemma 1. *For any $f \in \mathcal{H}$, any $\mathbf{x} \in \mathcal{X}$ and any $\mathcal{D}$, we have the following inequality*

$$\|\nabla f(\mathbf{x}) - \nabla \mu_{\mathcal{D}}(\mathbf{x})\|^2 \le \operatorname{tr}\big(\nabla k_{\mathcal{D}}(\mathbf{x},\mathbf{x})\nabla^\top\big)\|f\|_{\mathcal{H}}^2. \tag{4}$$

Proof. This is a simple corollary of a standard result in meshless scattered data approximation [*e.g.*, 28, Theorem 11.4]. The main idea is to express the estimation error as a linear functional, and then compute the operator norm of that linear functional.

Let $\lambda = \delta_{\mathbf{x}} D : \mathcal{H} \to \mathbb{R}$ be the composition of the evaluation operator and differential operator, *i.e.* $\lambda f = Df(\mathbf{x})$. Wendland [28, Theorem 11.4] provides a bound on $(\lambda f - \lambda\mu_{\mathcal{D}})^2$:

$$(\lambda f - \lambda\mu_{\mathcal{D}})^2 \le \lambda^{(1)}\lambda^{(2)} k_{\mathcal{D}}(\cdot,\cdot)\|f\|_{\mathcal{H}}^2,$$

where $\lambda^{(1)}$ applies λ to the first argument of $k_{\mathcal{D}}$ and $\lambda^{(2)}$ applies λ to the second argument of $k_{\mathcal{D}}$.

Pick the linear functional $\lambda : f \mapsto \frac{\partial}{\partial x_i} f(\mathbf{x})$ where $1 \le i \le d$. Then, the left hand side becomes $(\lambda f - \lambda\mu_{\mathcal{D}})^2 = (\frac{\partial}{\partial x_i} f(\mathbf{x}) - \frac{\partial}{\partial x_i}\mu_{\mathcal{D}}(\mathbf{x}))^2$. The right hand side $\lambda^{(1)}\lambda^{(2)} k_{\mathcal{D}}(\cdot,\cdot)$ is exactly the i-th diagonal entry of $\nabla k_{\mathcal{D}}(\mathbf{x},\mathbf{x})\nabla^\top$. For each i, use the above inequality, and then summing over all coordinates finishes the proof. □

A.2 Lemmas for Convergence on GP Sample Paths

In this section, we provide a few lemmas for the GP sample path $f \sim \mathcal{GP}(0,k)$. By Assumption 1, we have the follow lemma which asserts that f is smooth with high probability.

Lemma 3. *For $0 < \delta < 1$, there exists a constant $L > 0$ such that f is L-smooth w.p. at least $1-\delta$.*

Proof. The proof uses the Borell-TIS inequality. Let $f \sim \mathcal{GP}(0,k)$ be a Gaussian process. Provided that $\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$ is almost surely finite, the Borell-TIS inequality states that

$$\Pr(\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})| > u + \mathbb{E}\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|) \le \exp(-\frac{u^2}{2s^2}),$$

where $u > 0$ is an arbitrary positive constant and $s^2 = \sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}|f(\mathbf{x})|^2$. Namely, if the supremum of f is almost surely finite then the supremum of f is bounded with high probability.

Since k is four time continuously differentiable, the second-order derivative $\frac{\partial^2}{\partial x_i \partial x_j} f$ exists and is almost surely continuous. On a compact domain $\mathcal{X}$, the supremum $\sup_{\mathbf{x}\in\mathcal{X}} |\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})|$ is almost surely finite. By the Borell-TIS inequality, the expectation $\mathbb{E}\sup_{\mathbf{x}\in\mathcal{X}} |\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})|$ is finite and the supremum $\sup_{\mathbf{x}\in\mathcal{X}} |\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})|$ is bounded with high probability. Thus, the Frobenius norm of the Hessian $\nabla^2 f(\mathbf{x})$ is bounded with high probability by a union bound. Since the spectral norm of $\nabla f(\mathbf{x})$ can be bounded by its Frobenius norm, the spectral norm of the Hessian $\|\nabla f(\mathbf{x})\|$ is also bounded with high probability, which gives the smoothness constant. □

The next two lemmas bound the gradient estimation error with high probability.

Lemma 10. *Let $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ be a Gaussian vector. Then for any $t > 0$*

$$\Pr(\|\mathbf{u}\| > t) \le 2\exp\left(-\frac{t^2}{2\operatorname{tr}\mathbf{\Sigma}}\right).$$

Proof. This is a standard concentration inequality result, but we give a self-contained proof here for completeness. Denote the spectral decomposition $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. By Markov's inequality, we have

$$\begin{aligned}\Pr(\|\mathbf{u}\| > t) &= \Pr(e^{s\|\mathbf{u}\|} > e^{st}) \\ &\le e^{-st}\mathbb{E}e^{s\|\mathbf{u}\|},\end{aligned}$$

where $s > 0$ is an arbitrary positive constant. Let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a standard Gaussian variable. Then it is easy to see that $\|\mathbf{u}\| = \|\mathbf{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\|$. Then, replacing $\|\mathbf{u}\|$ with $\|\mathbf{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\|$ gives

$$\begin{aligned}\Pr(\|\mathbf{u}\| > t) &\le e^{-st}\mathbb{E}e^{s\|\mathbf{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\|} \\ &\le e^{-st}\mathbb{E}e^{s\|\mathbf{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\|_1} \\ &= e^{-st}\prod_{i=1}^{d}\mathbb{E}e^{s\sqrt{\lambda_i}|\epsilon_i|} \\ &\le 2e^{-st}\prod_{i=1}^{d}\mathbb{E}e^{s\sqrt{\lambda_i}\epsilon_i} \\ &= 2e^{-st+\frac{1}{2}s^2\sum_{i=1}^{d}\lambda_i},\end{aligned}$$

where the first line plugs in $\boldsymbol{\epsilon}$; the second line uses $\|\cdot\|_2 \le \|\cdot\|_1$; the third line is due to independence of ϵ_i; the forth line removes the absolute value resulting an extra factor of 2; the last lines uses the moment generating function of ϵ_i. Optimizing the bound over s gives the desired result

$$\Pr(\|\mathbf{u}\| > t) \le 2e^{-\frac{t^2}{2\sum_{i=1}^{d}\lambda_i}} = 2\exp(-\frac{t^2}{2\operatorname{tr}\mathbf{\Sigma}}).$$

□

Lemma 11. *For any $0 < \delta < 1$, let $C_t = 2\log(\frac{\pi^2 t^2}{6\delta})$. Then, the inequalities*

$$\|\nabla f(\mathbf{x}_t) - \nabla\mu_{\mathcal{D}_t}(\mathbf{x}_t)\|^2 \le C_t \operatorname{tr}\nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t)\nabla^\top$$

hold for any $t \ge 1$ with probability at least $1 - \delta$.

Proof. Since $\nabla f(\mathbf{x}_t) \sim \mathcal{N}(\nabla\mu_{\mathcal{D}_t}(\mathbf{x}_t), \nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t)\nabla^\top)$, applying Lemma 10 gives

$$\Pr(\|\nabla f(\mathbf{x}_t) - \nabla\mu_{\mathcal{D}_t}(\mathbf{x}_t)\|^2 \ge C_t \operatorname{tr}(\nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t)\nabla^\top)) \le 2\exp(-\frac{1}{2}C_t).$$

The particular choice of C_t makes the probability on the right hand side become $\frac{6\delta}{\pi^2 t^2}$. Using the union bound over all $t \ge 1$ and using the infinite sum $\sum_{t=1}^{\infty}\frac{1}{t^2} = \frac{\pi^2}{6}$ finishes the proof. □

We provide an important remark. The probability in Lemma 11 is taken over the randomness of f and the observation noise. On the other hand, the posterior mean gradient $\nabla\mu_{\mathcal{D}_t}$ is deterministic, since it is conditioned on the data $\mathcal{D}_t$.

B Bounds on the Error Function $E_{d,k,\sigma}$

This section is devoted to bounding the error function $E_{d,k,\sigma}(b)$ in terms of the batch size b. The results in this section immediately give a bound on the posterior covariance trace by Lemma 8.

Before diving into the proofs, we present some immediate corollaries of Assumption 1 on the kernel. Because k is stationary, the kernel can be written as $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x} - \mathbf{x}')$ for some positive-definite function ϕ. Observe that $\nabla k(\mathbf{x}, \mathbf{x}') = \nabla\phi(\mathbf{x} - \mathbf{x}')$ and $\nabla k(\mathbf{x}, \mathbf{x}')\nabla^\top = -\nabla^2\phi(\mathbf{x} - \mathbf{x}')$. Denote the first-order partial derivative $\partial_i\phi(\mathbf{x}) = \frac{\partial}{\partial x_i}\phi(\mathbf{x})$ and the second-order partial derivative $\partial_i^2\phi(\mathbf{x}) = \frac{\partial^2}{\partial x_i^2}\phi(\mathbf{x})$. It is easy to see that ϕ is an even function and $\nabla\phi$ is an odd function. In addition, $\mathbf{0}$ is a maximum of ϕ. Therefore, $\nabla\phi(\mathbf{0}) = \mathbf{0}$ and the Hessian $\nabla^2\phi(\mathbf{0})$ is negative semi-definite.

B.1 Noiseless Setting

The following is a bound for the error function $E_{d,k,0}$ for *arbitrary* kernels satisfying Assumption 1 in the noiseless setting $\sigma = 0$.

Lemma 2. *For $\sigma = 0$, the error function is bounded by $E_{d,k,0}(b) \le C \max\{0, 1 + d - b\}$, where $C = \max_{1 \le i \le d} \frac{\partial^2}{\partial x_i \partial x_i'} k(\mathbf{0}, \mathbf{0})$ is the maximum of the Hessian's diagonal entries at the origin.*

Proof. The bound holds trivially for $b = 0, 1$ and thus a proof is only needed for $b \ge 2$, which we split into two cases $2 \le b \le d + 1$ and $b > d + 1$.

We first focus on the case $2 \le b \le d + 1$. Let $\mathbf{z}_0 = \mathbf{0}$ and $\mathbf{z}_i = h\mathbf{e}_i$ where $i = 1, 2, \cdots b - 1$, where $\mathbf{e}_i$ is the i-th standard unit vector and $h > 0$ is a constant. Define $\mathbf{Z} = (\mathbf{z}_0 \quad \mathbf{z}_1 \quad \cdots \quad \mathbf{z}_{b-1})^\top$. By the definition of the error function, we have

$$\begin{aligned} E_{d,k,0}(b) &\le \operatorname{tr}\big(\nabla k(\mathbf{0}, \mathbf{0}) \nabla^\top - \nabla k(\mathbf{0}, \mathbf{Z}) k(\mathbf{Z}, \mathbf{Z})^{-1} k(\mathbf{Z}, \mathbf{0}) \nabla^\top\big) \\ &= \sum_{i=1}^{d} A_{ii} \\ &= C(1 + d - b) + \sum_{i=1}^{b-1} A_{ii}, \end{aligned}$$

where we define $\mathbf{A} = \nabla k(\mathbf{0}, \mathbf{0}) \nabla^\top - \nabla k(\mathbf{0}, \mathbf{Z}) k(\mathbf{Z}, \mathbf{Z})^{-1} k(\mathbf{Z}, \mathbf{0}) \nabla^\top$ and use the inequality $A_{ii} \le -\partial_i^2 \phi(\mathbf{0}) \le C$ for $b \le i \le d$.

Let us focus on the i-th diagonal entry A_{ii} where $1 \le i \le b - 1$. Then we have

$$\begin{aligned} A_{ii} &\le -\partial_i^2 \phi(\mathbf{0}) - (0 \quad \partial_i \phi(-h\mathbf{e}_i)) \begin{pmatrix} \phi(\mathbf{0}) & \phi(h\mathbf{e}_i) \\ \phi(h\mathbf{e}_i) & \phi(\mathbf{0}) \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \partial_i \phi(-h\mathbf{e}_i) \end{pmatrix} \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{1}{(\phi(\mathbf{0}))^2 - (\phi(h\mathbf{e}_i))^2} (0 \quad \partial_i \phi(-h\mathbf{e}_i)) \begin{pmatrix} \phi(\mathbf{0}) & -\phi(h\mathbf{e}_i) \\ -\phi(h\mathbf{e}_i) & \phi(\mathbf{0}) \end{pmatrix} \begin{pmatrix} 0 \\ \partial_i \phi(-h\mathbf{e}_i) \end{pmatrix} \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{\phi(\mathbf{0})(\partial_i \phi(h\mathbf{e}_i))^2}{(\phi(\mathbf{0}))^2 - (\phi(h\mathbf{e}_i))^2}, \end{aligned}$$

where the first line is because conditioning on the subset $\mathbf{z}_0$ and $\mathbf{z}_i$ does not make the posterior smaller. Now let $h \to 0$ and compute the limit by L'Hôpital's rule. We have

$$\begin{aligned} \lim_{h \to 0} A_{ii}(h) &= \lim_{h \to 0} -\partial_i^2 \phi(\mathbf{0}) - \frac{\phi(\mathbf{0})(\partial_i \phi(h\mathbf{e}_i))^2}{(\phi(\mathbf{0}))^2 - (\phi(h\mathbf{e}_i))^2} \\ &= \lim_{h \to 0} -\partial_i^2 \phi(\mathbf{0}) - \frac{\phi(\mathbf{0})}{\phi(\mathbf{0}) + \phi(h\mathbf{e}_i)} \cdot \frac{2 \partial_i \phi(h\mathbf{e}_i) \partial_i^2 \phi(h\mathbf{e}_i)}{-\partial_i \phi(h\mathbf{e}_i)} \\ &= 0. \end{aligned}$$

Thus letting $h \to 0$ gives the inequality $E_{d,k,\sigma}(b) \le C(1 + d - b)$ for $2 \le b \le d + 1$.

When $d > d + 1$, note that $E_{d,k,\sigma}(b)$ is an decreasing function in b and thus $E_{d,k,\sigma}(b) \le E_{d,k,\sigma}(d + 1) = 0$. Both cases can be bounded by the expression $C \max\{0, 1 + d - b\}$. □

B.2 Noisy Setting

This section proves bounds on the error function $E_{d,k,\sigma}$ for the RBF kernel and the $\nu = \frac{5}{2}$ Matérn kernel in the noisy setting. The lemmas in this section will implicitly use the assumption that $k(\mathbf{0}, \mathbf{0}) = 1$. This assumption is indeed satisfied by the RBF kernel and the Matérn kernel, which are of primary concern in this paper.

Before proving the bound on $E_{d,k,\sigma}$, we need one more technical lemma:

Lemma 12 (Central Differencing Designs)**.** *Consider the $2md$ points $\mathbf{Z} \in \mathbb{R}^{2md \times d}$ defined as*

$$\mathbf{z}_j^{(i)} = \begin{cases} -h\mathbf{e}_i, & j = 1, 2, \cdots m \\ h\mathbf{e}_i, & j = m + 1, m + 2 \cdots 2m, \end{cases}$$

where $1 \le i \le d$, $1 \le j \le 2m$ *and* $\mathbf{e}_i$ *is the* i*-th standard unit vector. Define*

$$\mathbf{A} = \nabla k(\mathbf{0}, \mathbf{0}) \nabla^\top - \nabla k(\mathbf{0}, \mathbf{Z})(k(\mathbf{Z}, \mathbf{Z}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{Z}, \mathbf{0}) \nabla^\top.$$

Then, we have

$$A_{ii} \le -\partial_i^2 \phi(\mathbf{0}) - \frac{2\beta_i^2}{(1-\alpha_i)+\gamma},$$

for all $1 \le i \le d$ *and thus*

$$\operatorname{tr}(\mathbf{A}) \le -\sum_{i=1}^{d} (\partial_i^2 \phi(\mathbf{0}) + \frac{2\beta_i^2}{(1-\alpha_i)+\gamma}),$$

where $\alpha_i = \phi(2h\mathbf{e}_i)$, $\beta_i = \partial_i \phi(-h\mathbf{e}_i)$ *and* $\gamma = \frac{\sigma^2}{m}$.

Proof. Note that $\mathbf{A}$ is the posterior covariance at the origin $\mathbf{0}$ conditioned on $\mathbf{Z}$. Denote $\mathbf{Z}^{(i)} = \begin{pmatrix} \mathbf{z}_1^{(i)} & \mathbf{z}_2^{(i)} & \cdots & \mathbf{z}_m^{(i)} & \mathbf{z}_{m+1}^{(i)} & \cdots & \mathbf{z}_{2m}^{(i)} \end{pmatrix}^\top$ the subset of $2m$ points that lie on the i-th axis. Then, the i-th diagonal entry A_{ii} can be bounded by

$$A_{ii} \le -\partial_i^2 \phi(\mathbf{0}) - \partial_i k(\mathbf{0}, \mathbf{Z}^{(i)})(k(\mathbf{Z}^{(i)}, \mathbf{Z}^{(i)}) + \sigma^2 \mathbf{I})^{-1} (\partial_i k(\mathbf{0}, \mathbf{Z}^{(i)}))^\top,$$

since conditioning on only a subset of points $\mathbf{Z}^{(i)}$ would not make the posterior variances smaller (*e.g.*, the posterior covariance is the Schur complement of a positive definite matrix). The remaining proof is dedicated to bounding the right hand side.

First, we need to compute the inverse of the $2m \times 2m$ kernel matrix

$$\begin{aligned} \widehat{\mathbf{K}} &= k(\mathbf{Z}^{(i)}, \mathbf{Z}^{(i)}) + \sigma^2 \mathbf{I} \\ &= \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \alpha_i \mathbf{1}\mathbf{1}^\top \\ \alpha_i \mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix} + \sigma^2 \mathbf{I}, \end{aligned}$$

where $\alpha_i = \phi(2h\mathbf{e}_i)$ is a nonnegative constant and $\mathbf{1}$ is a m dimensional vector (we drop the index i in the matrix $\widehat{\mathbf{K}}$ for notation simplicity). We compute the inverse analytically by forming its eigendecomposition

$$\widehat{\mathbf{K}} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top,$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_{2m})$ and $\mathbf{Q} = (\mathbf{q}_1 \quad \mathbf{q}_2 \quad \cdots \quad \mathbf{q}_{2m})$. Observe that:

$$\widehat{\mathbf{K}} = \begin{pmatrix} 1 & \alpha_i \\ \alpha_i & 1 \end{pmatrix} \otimes \mathbf{1}\mathbf{1}^\top + \sigma^2 \mathbf{I},$$

where $\otimes$ denotes the Kronecker product. Because the eigenvalues (vectors) of a Kronecker product equal the Kronecker product of the individual eigenvalues (vectors), and because adding a diagonal shift simply shifts the eigenvalues, the top two eigenvalues of $\widehat{\mathbf{K}}$ are $\lambda_1 = m(1+\alpha_i) + \sigma^2$ and $\lambda_2 = m(1-\alpha_i) + \sigma^2$. The remaining $2m-2$ eigenvalues are σ^2. The top two eigenvectors are

$$\mathbf{q}_1 = \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix}, \quad \mathbf{q}_2 = \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{1} \\ -\mathbf{1} \end{pmatrix},$$

Next, we cope with the term $\partial_i k(\mathbf{0}, \mathbf{Z}^{(i)})$, where the partial derivative is taken w.r.t. the first argument's i-th coordinate. Denote $\mathbf{v}^\top = \partial_i k(\mathbf{0}, \mathbf{Z}^{(i)})$. Then it is easy to see that

$$\mathbf{v} = \beta_i \begin{pmatrix} -\mathbf{1} \\ \mathbf{1} \end{pmatrix},$$

where $\beta_i = \partial_i \phi(-h\mathbf{e}_i)$. Note that $\mathbf{v}$ happens to be an eigenvector of $\widehat{\mathbf{K}}$ as well, because $\mathbf{v} \parallel \mathbf{q}_2$. As a result, $\mathbf{Q}^\top \mathbf{v}$ has a simple expression $\mathbf{Q}^\top \mathbf{v} = (0 \quad -\sqrt{2m}\beta_i \quad 0 \quad \cdots \quad 0)^\top$. Thus, we have

$$\begin{aligned} A_{ii} &\le -\partial_i^2 \phi(\mathbf{0}) - \mathbf{v}^\top \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^\top \mathbf{v} \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{2m\beta_i^2}{m(1-\alpha_i) + \sigma^2} \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{2\beta_i^2}{(1-\alpha_i)+\gamma}. \end{aligned}$$

Summing over the coordinates $1 \le i \le d$ finishes the proof. □

With Lemma 12, we are finally ready to present the bounds for the RBF kernel and the Matérn kernel.

Lemma 4 (RBF Kernel). *Let $k(\mathbf{x}_1, \mathbf{x}_2) = \exp\big(-\frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2\big)$ be the RBF kernel. We have*

$$E_{d,k,\sigma}(2md) \le d\bigg(1 + W\bigg(-\frac{m}{e(m+\sigma^2)}\bigg)\bigg) = \mathcal{O}(\sigma d m^{-\frac{1}{2}}),$$

where $m \in \mathbb{N}$ and W denotes the principal branch of the Lambert W function.

Proof. For any $\mathbf{Z} \in \mathbb{R}^{2md \times d}$, the following inequality holds by the definition of error function

$$E_{d,k,\sigma}(b) \le \operatorname{tr}\big(\nabla k(\mathbf{0},\mathbf{0})\nabla^\top - \nabla k(\mathbf{0},\mathbf{Z})(k(\mathbf{Z},\mathbf{Z}) + \sigma^2\mathbf{I})^{-1}k(\mathbf{Z},\mathbf{0})\nabla^\top\big).$$

Consider the following points $\mathbf{z}_j^{(i)}$ defined as

$$\mathbf{z}_j^{(i)} = \begin{cases} -h\mathbf{e}_i, & j = 1, 2, \cdots m \\ h\mathbf{e}_i, & j = m+1, m+2 \cdots 2m, \end{cases}$$

where $1 \le i \le d$. The total number of points is exactly $2md$. By Lemma 12, we have

$$E_{d,k,\sigma}(b) \le -\sum_{i=1}^{d}(\partial_i^2\phi(\mathbf{0}) + \frac{2\beta_i^2}{(1-\alpha_i)+\gamma}).$$

For the RBF kernel, the values of α_i and β_i are the same for each coordinate $1 \le i \le d$ since it is isotropic:

$$\alpha = \alpha_i = \phi(2h\mathbf{e}_i) = \exp(-2h^2), \quad \beta = \beta_i = \partial_i\phi(-h\mathbf{e}_i) = \exp\Big(-\frac{1}{2}h^2\Big)h.$$

Plugging the value of α, β, γ and $\partial_i^2\phi(\mathbf{0})$ into the bound on $E_{d,k,\sigma}$, we have

$$\begin{aligned} A_{ii} &\le 1 - \frac{2m\exp(-h^2)h^2}{m(1-\exp(-2h^2)) + \sigma^2} \\ &\le 1 - \frac{2m\exp(-2h^2)h^2}{m(1-\exp(-2h^2)) + \sigma^2} \end{aligned}$$

where the second inequality replaces $\exp(-h^2)$ with $\exp(-2h^2)$ in the numerator. Because the bound holds for arbitrary h, we can apply the transformation $h \mapsto \frac{1}{\sqrt{2}}h$, which gives the inequality

$$A_{ii} \le 1 - \frac{m\exp(-h^2)h^2}{m(1-\exp(-h^2)) + \sigma^2}.$$

Our goal is to bound A_{ii} in terms of m and σ^2. Therefore, we minimize the right hand side over h. Define $g(x) = 1 - \frac{me^{-x}x}{m(1-e^{-x})+\sigma^2}$, where $x \ge 0$. The derivative is given by:

$$g'(x) = \frac{m(m + (m+\sigma^2)e^x(x-1))}{(m(e^x-1)+\sigma^2e^x)^2}.$$

The unique stationary point is $x^* = 1 + W\big(-\frac{m}{e(m+\sigma^2)}\big)$, where W is the principal branch of the Lambert W function. It is easy to see the stationary point x^* is the global minimizer of $g(x)$ over $\mathbb{R}_{\ge 0}$. Plug x^* into the expression of g. Coincidentally, we have $g(x^*) = 1 + W\big(-\frac{m}{e(m+\sigma^2)}\big)$ as well — x^* is a fixed point of g.

In summary, we have shown $A_{ii} \le 1 + W\big(-\frac{m}{e(m+\sigma^2)}\big)$ for each coordinate i. Summing A_{ii} over all d coordinates proves the first inequality $E_{d,k,\sigma}(2md) \le d\Big(1 + W\big(-\frac{m}{e(m+\sigma^2)}\big)\Big)$. The second inequality is a direction implication of Lemma 7, which completes the proof. □

Lemma 5 (Matern Kernel). *Let $k(\cdot,\cdot)$ be the $\nu = 2.5$ Matérn kernel. Then, we have*

$$E_{d,k,\sigma}(2md) \lesssim \sigma d m^{-\frac{1}{2}} + \sigma^{\frac{3}{2}} d m^{-\frac{3}{4}} = \mathcal{O}(\sigma d m^{-\frac{1}{2}}).$$

Proof. The proof is similar to Lemma 4. The difference is that we need to upper bound $\partial_i^2\phi(\mathbf{0}) + \frac{2\beta_i^2}{(1-\alpha_i)+\gamma}$ by a rational function. Otherwise the expression is intractable to minimize.

The Matérn kernel with half integer ν can be written as a product of an exponential and a polynomial. In particular, for $\nu = \frac{5}{2}$, we have

$$\phi(\mathbf{x}) = (1 + \sqrt{5}\|\mathbf{x}\| + \frac{5}{3}\|\mathbf{x}\|^2)\exp(-\sqrt{5}\|\mathbf{x}\|).$$

When $\mathbf{x}$ is in the nonnegative orthant, the gradient is

$$\begin{aligned}\nabla\phi(\mathbf{x}) &= \exp(-\sqrt{5}\|\mathbf{x}\|)(\frac{\sqrt{5}}{\|\mathbf{x}\|}\mathbf{x} + \frac{10}{3}\mathbf{x}) - \frac{\sqrt{5}}{\|\mathbf{x}\|}\exp(-\sqrt{5}\|\mathbf{x}\|)(1 + \sqrt{5}\|\mathbf{x}\| + \frac{5}{3}\|\mathbf{x}\|^2)\mathbf{x} \\ &= -\frac{5}{3}\exp(-\sqrt{5}\|\mathbf{x}\|)(1 + \sqrt{5}\|\mathbf{x}\|)\mathbf{x}.\end{aligned}$$

Since the Matérn kernel is isotropic, the α_i and β_i as in Lemma 12 are the same across different coordinate i, and their values are

$$\begin{aligned}\alpha = \alpha_i &= \phi(2h\mathbf{e}_i) = \exp(-2\sqrt{5}h)(1 + 2\sqrt{5}h + \frac{20}{3}h^2), \\ \beta = \beta_i &= \partial_i\phi(-h\mathbf{e}_i) = -\partial_i\phi(h\mathbf{e}_i) = \frac{5}{3}\exp(-\sqrt{5}h)(1 + \sqrt{5}h)h,\end{aligned}$$

In addition, $-\partial_i^2\phi(\mathbf{0}) = \frac{5}{3}$. By Lemma 12, we have

$$\begin{aligned}A_{ii} &\le -\partial_i^2\phi(\mathbf{0}) - \frac{2\beta^2}{(1-\alpha)+\gamma} \\ &= \frac{5}{3} - \frac{10\exp(-2h)(1+h)^2h^2}{3(3 - \exp(-2h)(3 + 6h + 4h^2)) + 9\gamma}.\end{aligned}$$

Next, we approximate the exponential function $\exp(-2h)$ by its Taylor polynomials. By Lemma 6, use the inequality $\exp(-2h) \ge 1 - 2h$ for the numerator and the inequality $\exp(-2h) \ge 1 - 2h + 2h^2 - \frac{4}{3}h^3$ for the denominator. Applying these two inequalities gives

$$A_{ii} \le \frac{5}{3} - \frac{10(1-2h)(1+h)^2h^2}{2h^2(3 + 8h^3) + 9\gamma}.$$

Let $h = \gamma^{\frac{1}{4}}$ and thus $\gamma = h^4$. Then we have

$$\begin{aligned}A_{ii} &\le \frac{5}{3} - \frac{10(1-2h)(1+h)^2h^2}{2h^2(3 + 8h^3) + 9h^4} \\ &= \frac{5h^2(27 + 28h)}{3(6 + 9h^2 + 16h^3)} \\ &\le \frac{5}{18}h^2(27 + 28h) \\ &= \frac{15}{2}\gamma^{\frac{1}{2}} + \frac{70}{9}\gamma^{\frac{3}{4}} \\ &= \frac{15}{2}\sigma m^{-\frac{1}{2}} + \frac{70}{9}\sigma^{\frac{3}{2}}m^{-\frac{3}{4}}\end{aligned}$$

where the third line drops h^2 and h^3 in the denominator; the forth line plugs in the value $h = \gamma^{\frac{1}{4}}$ back and drops the constants. Summing over the coordinates $1 \le i \le d$ gives the first inequality:

$$E_{d,k,\sigma}(2md) \lesssim \sigma dm^{-\frac{1}{2}} + \sigma^{\frac{3}{2}}dm^{-\frac{3}{4}}$$

For large enough m, the bound is dominated by the first term $\sigma dm^{-\frac{1}{2}}$, which completes the proof. □

We end this section with a short summary. Lemma 4 and Lemma 5 happen to end up with the same rate $E_{d,k,s}(2md) = \mathcal{O}(\sigma dm^{-\frac{1}{2}})$. Replacing $2md$ with the batch size b, we have shown that $E_{d,k,s}(b) = \mathcal{O}(\sigma d^{\frac{3}{2}}b^{-\frac{1}{2}})$ for the RBF kernel and $\nu = 2.5$ Matérn kernel.

B.3 Discussion: Forward Differencing Designs

This section explores an alternative proof for the error function based on forward differencing designs, as opposed to the central differencing designs in Lemma 12. Similar to the previous section, we assume $k(\mathbf{0}, \mathbf{0}) = 1$, which is indeed satisfied by the RBF kernel and Matérn kernel.

Lemma 13. *Consider following $(d+1)m$ points $\mathbf{Z} \in \mathbb{R}^{(d+1)m \times d}$ defined as*

$$\begin{aligned} \mathbf{z}_j^{(0)} &= \mathbf{0}, && j = 1, 2, \cdots, m \\ \mathbf{z}_j^{(i)} &= h\mathbf{e}_i, \quad i \geq 1, && j = 1, 2, \cdots, m \end{aligned}$$

where $\mathbf{e}_i$ is the i-th standard unit vector. Define

$$\mathbf{A} = \nabla k(\mathbf{0}, \mathbf{0}) \nabla^\top - \nabla k(\mathbf{0}, \mathbf{Z})(k(\mathbf{Z}, \mathbf{Z}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{Z}, \mathbf{0}) \nabla^\top.$$

Then, we have

$$\operatorname{tr} \mathbf{A} \leq -\sum_{i=1}^{d} \left(\partial_i^2 \phi(\mathbf{0}) + \frac{(1+\gamma)\beta_i^2}{(1+\gamma)^2 - \alpha_i^2} \right)$$

where $\alpha_i = \phi(h\mathbf{e}_i)$, $\beta_i = \partial_i \phi(-h\mathbf{e}_i)$ and $\gamma = \frac{\sigma^2}{m}$.

Proof. The proof is similar to Lemma 12.

Denote $\mathbf{Z}^{(i)}$ be the subset of points consisting of $\mathbf{z}_j^{(i)}$ and $\mathbf{z}_j^{(0)}$, where $j = 1, 2, \cdots, m$. Notice that the i-th diagonal entry A_{ii} can be bounded by

$$A_{ii} \leq -\partial_i^2 \phi(\mathbf{0}) - \partial_i k(\mathbf{0}, \mathbf{Z}^{(i)})(k(\mathbf{Z}^{(i)}, \mathbf{Z}^{(i)}) + \sigma^2 \mathbf{I})^{-1} (\partial_i k(\mathbf{0}, \mathbf{Z}^{(i)}))^\top.$$

We need to invert the $2m \times 2m$ matrix

$$\begin{aligned} \widehat{\mathbf{K}} &= k(\mathbf{Z}^{(i)}, \mathbf{Z}^{(i)}) + \sigma^2 \mathbf{I} \\ &= \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \alpha_i \mathbf{1}^\top \\ \alpha_i \mathbf{1} & \mathbf{1}\mathbf{1}^\top \end{pmatrix} + \sigma^2 \mathbf{I}. \end{aligned}$$

Again, we resort to the eigendecomposition of $\widehat{\mathbf{K}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. The top two eigenvalues of $\widehat{\mathbf{K}}$ are $\lambda_1 = m(1+\alpha_i) + \sigma^2$ and $\lambda_2 = m(1-\alpha_i) + \sigma^2$. The remaining $2m-2$ eigenvalues are σ^2. The top two eigenvectors are

$$\mathbf{q}_1 = \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix}, \quad \mathbf{q}_2 = \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{1} \\ -\mathbf{1} \end{pmatrix}.$$

Denote $\mathbf{v}^\top = \partial_i k\big(\mathbf{0}, \mathbf{Z}^{(i)}\big)$. Note that $\mathbf{v}$ can be written as a linear combination of $\mathbf{q}_1$ and $\mathbf{q}_2$:

$$\mathbf{v} = \beta_i \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \end{pmatrix} = \frac{1}{2} \beta_i \sqrt{2m} (\mathbf{q}_1 - \mathbf{q}_2).$$

Then, straightforward calculation gives $\mathbf{Q}^\top \mathbf{v} = \frac{1}{2}\beta_i \sqrt{2m} (1 \quad -1 \quad 0 \quad \cdots \quad 0)^\top$.

Then, we have

$$\begin{aligned} A_{ii} &\leq -\partial_i^2 \phi(\mathbf{0}) - \mathbf{v}^\top \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^\top \mathbf{v} \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{1}{2} m \beta_i^2 \left(\frac{1}{m(1+\alpha_i) + \sigma^2} + \frac{1}{m(1-\alpha_i) + \sigma^2} \right) \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{1}{2} \beta_i^2 \left(\frac{1}{1+\alpha_i+\gamma} + \frac{1}{1-\alpha_i+\gamma} \right) \\ &= -\partial_i^2 \phi(\mathbf{0}) - \frac{\beta_i^2 (1+\gamma)}{(1+\gamma)^2 - \alpha_i^2}. \end{aligned}$$

Summing over all coordinates finishes the proof. □

Lemma 14. *Let $k(\mathbf{x}_1, \mathbf{x}_2) = \exp\big(-\frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2\big)$ be the RBF kernel. The forward differencing designs give a decay rate of $E_{d,k,\sigma}(b) = \mathcal{O}(\sigma d m^{-\frac{1}{2}})$.*

Proof. Define $\alpha = \phi(h\mathbf{e}_i)$ and $\beta = \partial_i \phi(-h\mathbf{e}_i)$. Plugging the values of α and β into the bounds in Lemma 13 yields

$$A_{ii} \le 1 - \frac{h^2 \exp(-h^2)(1+\gamma)}{(1+\gamma)^2 - \exp(-h^2)}.$$

The bound holds for arbitrary h. Applying the transformation $h \mapsto \sqrt{h}$ gives

$$\begin{aligned} A_{ii} &\le 1 - \frac{h \exp(-h)(1+\gamma)}{(1+\gamma)^2 - \exp(-h)} \\ &\le 1 - \frac{h(1-h)(1+\gamma)}{(1+\gamma)^2 - (1-h)} \\ &= 1 - \frac{h(1-h)(1+\gamma)}{\gamma^2 + 2\gamma + h} \\ &\le 1 - \frac{h(1-h)}{\gamma^2 + 2\gamma + h}, \end{aligned}$$

where the second line uses the inequality $\exp(-h) \ge 1 - h$ in the numerator and the denominator; the last line is because γ is nonnegative. Let $h = \gamma^{\frac{1}{2}}$ so that $\gamma = h^2$. Then we have

$$\begin{aligned} A_{ii} &\le 1 - \frac{h(1-h)}{h^4 + 2h^2 + h} \\ &= \frac{h^3 + 3h}{h^3 + 2h + 1} \\ &\le h^3 + 3h \\ &= \gamma^{\frac{3}{2}} + 3\gamma^{\frac{1}{2}} \\ &\lesssim \gamma^{\frac{1}{2}} \\ &= \sigma m^{-\frac{1}{2}} \end{aligned}$$

where the first line plugs in $\gamma = h^2$; the third line drops the $h^3 + 2h$ in the denominator; the forth lines plugs in $h = \gamma^{\frac{1}{2}}$; the fifth line is because $\gamma^{\frac{1}{2}}$ dominates the bound when m is large. Thus, we have shown $E_{d,k,\sigma}(dm + m) = \mathcal{O}(\sigma d m^{-\frac{1}{2}})$. □

The above lemma shows that forward differencing designs achieve the same asymptotic decay rate as the central differencing designs for the RBF kernel. Though, the leading constant in the big $\mathcal{O}$ notation is slightly larger.

C Convergence Proofs

The following is a useful lemma for biased gradient updates, which bounds the gradient norm via the cumulative bias. The proof is adapted from a lemma by Ajalloeian and Stich [1, Lemma 2].

Lemma 15. *Let f be L-smooth and bounded from below. Suppose the gradient oracle $\hat{\mathbf{g}}_t$ has bias bounded by ξ_t in the t-th iteration. Namely, we have $\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 \le \xi_t$ for all $t \ge 0$, where $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ is the ground truth gradient. Then the update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \hat{\mathbf{g}}_t$ with $\eta_t \le \frac{1}{L}$ produces a sequence $\{\mathbf{x}_t\}_{t=1}^{\infty}$ satisfying*

$$\min_{1 \le t \le T} \|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2(f(\mathbf{x}_1) - f^*)}{\sum_{t=1}^T \eta_t} + \frac{\sum_{i=1}^T \eta_t \xi_t}{\sum_{t=1}^T \eta_t}$$

Proof. By L-smoothness, we have

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.$$

Plugging in the update formula $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \hat{\mathbf{g}}_t$, we have

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \hat{\mathbf{g}}_t + \frac{1}{2} L \eta_t^2 \|\hat{\mathbf{g}}_t\|^2 \\
&\le f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \hat{\mathbf{g}}_t + \frac{1}{2} \eta_t \|\hat{\mathbf{g}}_t\|^2 \\
&\le f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \hat{\mathbf{g}}_t + \frac{1}{2} \eta_t \|\hat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)\|^2 \\
&\le f(\mathbf{x}_t) + \frac{1}{2} \eta_t (\|\hat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\|^2 - \|\nabla f(\mathbf{x}_t)\|^2) \\
&\le f(\mathbf{x}_t) - \frac{1}{2} \eta_t \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2} \eta_t \xi_t,
\end{aligned}$$

where the first inequality uses L-smoothness; the second inequality uses $\eta_t \le \frac{1}{L}$; the fourth inequality expands the squared Euclidean norm; the last inequality uses the definition of bias. Summing the inequalities for $t = 1, 2, \cdots, T$ and rearranging the terms, we have

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla f(\mathbf{x}_t)\|^2 &\le 2(f(\mathbf{x}_1) - f(\mathbf{x}_{T+1})) + \sum_{t=1}^{T} \eta_t \xi_t \\
&\le 2(f(\mathbf{x}_1) - f^*) + \sum_{t=1}^{T} \eta_t \xi_t.
\end{aligned}$$

Dividing both sides by $\sum_{t=1}^{T} \eta_t$ gives

$$\frac{\sum_{t=1}^{T} \eta_t \|\nabla f(\mathbf{x}_t)\|^2}{\sum_{t=1}^{T} \eta_t} \le \frac{2(f(\mathbf{x}_1) - f^*)}{\sum_{t=1}^{T} \eta_t} + \frac{\sum_{t=1}^{T} \eta_t \xi_t}{\sum_{t=1}^{T} \eta_t}.$$

The left hand side is a weighted average, which is greater than the minimum over $1 \le t \le T$, which completes the proof. □

Next, we prove a variant of Lemma 15 for the projected update $\mathbf{x}_{t+1} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t))$. For the ground truth gradient $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$, define the gradient mapping

$$G(\mathbf{x}_t) = \frac{1}{\eta_t}\big(\mathbf{x}_t - \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t)\big).$$

For the approximate gradient $\hat{\mathbf{g}}_t = \nabla \mu_{\mathcal{D}_t}(\mathbf{x}_t)$, define the gradient mapping

$$\widehat{G}(\mathbf{x}_t) = \frac{1}{\eta_t}\big(\mathbf{x}_t - \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \hat{\mathbf{g}}_t)\big).$$

In the following, we introduce two lemmas characterizing the projection operator $\mathrm{proj}_{\mathcal{X}}(\cdot)$.

Lemma 16 (*e.g.*, Lemma 3.1 of Bubeck et al. [2])**.** *Let $\mathcal{X}$ be convex and compact. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathbb{R}^d$. Then we have*

$$(\mathrm{proj}_{\mathcal{X}}(\mathbf{z}) - \mathbf{z})^\top (\mathbf{x} - \mathrm{proj}_{\mathcal{X}}(\mathbf{z})) \ge 0.$$

As a result, $\|\mathbf{x} - \mathbf{z}\|^2 \ge \|\mathrm{proj}_{\mathcal{X}}(\mathbf{z}) - \mathbf{z}\|^2 + \|\mathbf{x} - \mathrm{proj}_{\mathcal{X}}(\mathbf{z})\|^2$.

Lemma 17. *The following holds:*

1. $\|G(\mathbf{x}_t)\| \le \|\mathbf{g}_t\|$,
2. $\|G(\mathbf{x}_t) - \mathbf{g}\| \le \|\mathbf{g}_t\|$,
3. $\|\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\| \le \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|$,

4. $\mathbf{g}^\top G(\mathbf{x}_t) \geq \|G(\mathbf{x}_t)\|^2$.

Proof. (1-2): The first two inequalities are direct corollaries of Lemma 16.

(3): This is proved as follows:

$$\begin{aligned}\|\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\| &= \frac{1}{\eta_t}\|\mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) - \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \hat{\mathbf{g}}_t)\| \\ &\leq \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|,\end{aligned}$$

where the second line is because the projection operator is non-expansive.

(4): By Lemma 16, we have

$$-\eta_t G(\mathbf{x}_t)^\top (\eta_t G(\mathbf{x}_t) - \eta_t \mathbf{g}_t) \geq 0.$$

Rearranging the terms finishes the proof. □

Now we give a lemma proving biased gradient update with the projection operator. The proof is adapted from a lemma by Shu et al. [22].

Lemma 18. *Let f be L-smooth over a convex compact set $\mathcal{X}$. Moreover, assume the gradient norm $\|\nabla f(\mathbf{x})\|$ is bounded by L' on $\mathcal{X}$. Suppose the gradient oracle $\hat{\mathbf{g}}_t$ has bias bounded by ξ_t in the t-th iteration: $\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 \leq \xi_t$ for all $t \geq 0$, where $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$. Then the update $\mathbf{x}_{t+1} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \hat{\mathbf{g}}_t)$ with $\eta_t \leq \frac{1}{L}$ produces a sequence $\{\mathbf{x}_t\}_{t=1}^{\infty}$ satisfying*

$$\min_{1\leq t\leq T}\|G(\mathbf{x}_t)\|^2 \leq \frac{2(f(\mathbf{x}_1) - f^*)}{\sum_{t=1}^T \eta_t} + \frac{\sum_{i=1}^T \eta_t \xi_t}{\sum_{t=1}^T \eta_t} + \frac{L' \sum_{i=1}^T \eta_t \sqrt{\xi_t}}{\sum_{t=1}^T \eta_t},$$

where $G(\mathbf{x}_t) = \frac{1}{\eta_t}\big(\mathbf{x}_t - \mathrm{proj}_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t)\big)$ is the gradient mapping.

Proof. By L-smoothness, we have

$$\begin{aligned}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq \mathbf{g}_t^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{1}{2}L\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= -\eta_t \mathbf{g}_t^\top \widehat{G}(\mathbf{x}_t) + \frac{1}{2}L\eta_t^2 \|\widehat{G}(\mathbf{x}_t)\|^2 \\ &\leq -\eta_t \mathbf{g}_t^\top \widehat{G}(\mathbf{x}_t) + \frac{1}{2}\eta_t \|\widehat{G}(\mathbf{x}_t)\|^2,\end{aligned}$$

where the second line plugs in the update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \widehat{G}(\mathbf{x}_t)$ and the third line is due to $\eta_t \leq \frac{1}{L}$.

Now we analyze the two terms separately. For the first term, we have

$$\begin{aligned}-\eta_t \mathbf{g}_t^\top \widehat{G}(\mathbf{x}_t) &= -\eta_t \mathbf{g}_t^\top \big(\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big) - \eta_t \mathbf{g}_t^\top G(\mathbf{x}_t) \\ &\leq -\eta_t \mathbf{g}_t^\top \big(\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big) - \eta_t \big\|G(\mathbf{x}_t)\big\|^2,\end{aligned}$$

where the second inequality uses Lemma 17. For the second term, we have

$$\begin{aligned}\frac{1}{2}\eta_t \big\|\widehat{G}(\mathbf{x}_t)\big\|^2 &= \frac{1}{2}\eta_t \big\|\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t) + G(\mathbf{x}_t)\big\|^2 \\ &= \frac{1}{2}\eta_t \big\|\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big\|^2 + \eta_t \big(\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big)^\top G(\mathbf{x}_t) + \frac{1}{2}\eta_t \big\|G(\mathbf{x}_t)\big\|^2 \\ &\leq \frac{1}{2}\eta_t \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 + \eta_t \big(\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big)^\top G(\mathbf{x}_t) + \frac{1}{2}\eta_t \big\|G(\mathbf{x}_t)\big\|^2\end{aligned}$$

Summing them together, we have

$$\begin{aligned}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq \frac{1}{2}\eta_t \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 + \eta_t \big(\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big)^\top \big(G(\mathbf{x}_t) - \mathbf{g}_t\big) - \frac{1}{2}\eta_t \big\|G(\mathbf{x}_t)\big\|^2 \\ &\leq \frac{1}{2}\eta_t \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 + \eta_t \big\|\widehat{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\big\| \cdot \|\mathbf{g}_t\| - \frac{1}{2}\eta_t \big\|G(\mathbf{x}_t)\big\|^2 \\ &\leq \frac{1}{2}\eta_t \xi_t + \eta_t L' \sqrt{\xi_t} - \frac{1}{2}\eta_t \|G(\mathbf{x}_t)\|^2,\end{aligned}$$

where the second line uses Cauchy-Schwarz inequality.

A telescoping sum gives

$$\sum_{t=1}^{T} \eta_t \|G(\mathbf{x}_t)\|^2 \leq \sum_{t=1}^{T} \eta_t \xi_t + 2L' \sum_{t=1}^{T} \eta_t \sqrt{\xi_t} + 2(f(\mathbf{x}_1) - f(\mathbf{x}_{T+1})),$$

which results in

$$\min_{1 \leq t \leq T} \|G(\mathbf{x}_t)\|^2 \leq \frac{2(f(\mathbf{x}_1) - f^*)}{\sum_{t=1}^{T} \eta_t} + \frac{\sum_{i=1}^{T} \eta_t \xi_t}{\sum_{t=1}^{T} \eta_t} + \frac{L' \sum_{i=1}^{T} \eta_t \sqrt{\xi_t}}{\sum_{t=1}^{T} \eta_t}.$$

□

The rest of this section proves all theorems and their corollaries in the main paper.

Theorem 1. *Let $f \in \mathcal{H}$ whose smoothness constant is L. Running Algorithm 1 with constant batch size $b_t = b$ and step size $\eta_t = \frac{1}{L}$ for T iterations outputs a sequence satisfying*

$$\min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|^2 \leq \tfrac{1}{T}\big(2L(f(\mathbf{x}_1) - f^*)\big) + B^2 \cdot E_{d,k,0}(b). \tag{5}$$

Proof. By Lemma 1 and Assumption 3, we can bound the bias in the iteration t as

$$\|\nabla f(\mathbf{x}_t) - \nabla \mu_{\mathcal{D}_t}(\mathbf{x}_t)\|^2 \leq B^2 \operatorname{tr}\big(\nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t) \nabla^\top\big).$$

By Lemma 8, the trace in the RHS can be bounded by the error function $E_{d,k,\sigma}(b)$. Thus, the gradient bias is $B^2 E_{k,k,\sigma}(b)$. Applying Lemma 15 with $\eta_t = \frac{1}{L}$ and $\xi_t = B^2 E_{k,k,\sigma}(b)$ finishes the proof. □

Corollary 1. *Under the same assumptions of Theorem 1, using batch size $b_t = d + 1$, we have*

$$\min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|^2 \leq \tfrac{1}{T}\big(2L(f(\mathbf{x}_1) - f^*)\big).$$

Therefore, the total number of samples $n = \mathcal{O}(dT)$ and the squared gradient norm $\|\nabla f(\mathbf{x}_t)\|^2$ converges to zero at the rate $\mathcal{O}(d/n)$.

Proof. By Lemma 2, we have $E_{d,k,\sigma}(d+1) = 0$. Plugging it into Theorem 1 gives the rate in the iteration number T. To get the rate in samples n, note that $n = \sum_{t=1}^{T}(d+1) = (d+1)T$. Plugging $T = \frac{n}{d+1}$ into the rate finishes the proof. □

Theorem 2. *For $0 < \delta < 1$, suppose f is a GP sample whose smoothness constant is L w.p. at least $1 - \delta$. Algorithm 1 with batch size b_t and step size $\eta_t = \frac{1}{L}$ produces a sequence satisfying*

$$\min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|^2 \leq \tfrac{1}{T}\big(2L(f(\mathbf{x}_1) - f^*)\big) + \tfrac{1}{T} \textstyle\sum_{t=1}^{T} C_t E_{d,k,\sigma}(b_t) \tag{6}$$

with probability at least $1 - 2\delta$, where $C_t = 2\log\big((\pi^2/6)(t^2/\delta)\big)$.

Proof. By Lemma 11, we have

$$\|\nabla f(\mathbf{x}_t) - \nabla \mu_{\mathcal{D}_t}(\mathbf{x}_t)\|^2 \leq C_t \operatorname{tr}\big(\nabla k_{\mathcal{D}_t}(\mathbf{x}_t, \mathbf{x}_t) \nabla^\top\big)$$

with probability at least $1 - \delta$. The trace on the RHS can be further bounded by the error function $E_{d,k,\sigma}(b_t)$ by Lemma 8. Applying the union bound, with probability at least $1 - 2\delta$, the inequality

$$\|\nabla f(\mathbf{x}_t) - \nabla \mu_{\mathcal{D}_t}(\mathbf{x}_t)\|^2 \leq C_t E_{d,k,\sigma}(b_t)$$

holds for all $t \geq 0$ and f is L-smooth. Applying Lemma 15 with $\eta_t = \frac{1}{L}$ and $\xi = C_t E_{d,k,\sigma}(b_t)$ finishes the proof. □

Corollary 2. *Let $k(\cdot,\cdot)$ be either the RBF kernel or the $\nu = 2.5$ Matérn kernel. Under the same conditions as Theorem 2, if*

$$b_t = \begin{cases} d\log^2 t; \\ dt; \\ dt^2, \end{cases} \quad \textit{then} \quad \min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 = \begin{cases} \mathcal{O}(1/T) + \mathcal{O}(\sigma d); \\ \mathcal{O}\big(\sigma d T^{-\frac{1}{2}}\log T\big) = \mathcal{O}\big(\sigma d^{\frac{5}{4}} n^{-\frac{1}{4}}\log n\big); \\ \mathcal{O}\big(\sigma d T^{-1}\log^2 T\big) = \mathcal{O}\big(\sigma d^{\frac{4}{3}} n^{-\frac{1}{3}}\log^2 n\big), \end{cases}$$

with probability at least $1-2\delta$. Here, T is the total number of iterations and n is the total number of samples queried.

Proof. By Theorem 2, with probability at least $1-2\delta$, we have

$$\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 \le \tfrac{1}{T}\big(2L(f(\mathbf{x}_1) - f^*)\big) + \tfrac{1}{T}\textstyle\sum_{t=1}^T C_t E_{d,k,\sigma}(b_t)$$

The proof boils down to bounding the average cumulative bias. The full details is in Appendix D. □

Theorem 3. *Under the same conditions as Corollary 2, without Assumption 2, using the projected update rule* (7)*, Algorithm 1 obtains the following rates:*

$$\textit{if} \quad b_t = \begin{cases} dt; \\ dt^2, \end{cases} \quad \textit{then} \quad \min_{1\le t\le T}\|G(\mathbf{x}_t)\|^2 = \begin{cases} \mathcal{O}\big(\sigma d^{\frac{5}{4}} n^{-\frac{1}{4}}\log n + \sigma^{\frac{1}{2}} d^{\frac{5}{8}} n^{-\frac{1}{8}}\log n\big); \\ \mathcal{O}\big(\sigma d^{\frac{4}{3}} n^{-\frac{1}{3}}\log^2 n + \sigma^{\frac{1}{2}} d^{\frac{2}{3}} n^{-\frac{1}{6}}\log n\big), \end{cases}$$

with probability at least $1-2\delta$. Here, n is the total number of samples queried.

Proof. Since f is twice differentiable on a compact set $\mathcal{X}$, its gradient norm $\|\nabla f(\mathbf{x})\|$ attains a maximum. Thus, there exists a constant L' such that $L' \ge \|\nabla f(\mathbf{x})\|$ for all $\mathbf{x}\in\mathcal{X}$. By Lemma 18 and a similar argument in Theorem 2, with probability at least $1-2\delta$, we have

$$\min_{1\le t\le T}\|G(\mathbf{x}_t)\|^2 \le \tfrac{1}{T}\big(2L(f(\mathbf{x}_1) - f^*)\big) + \tfrac{1}{T}\textstyle\sum_{t=1}^T C_t^{(2)} E_{d,k,\sigma}(b_t) + \tfrac{1}{T}\sum_{t=1}^T C_t^{(2)}\sqrt{E_{d,k,\sigma}(b_t)},$$

where $C_t^{(1)}$ and $C_t^{(2)}$ are constants growing in $\mathcal{O}(\log t)$. By Lemmas 4 and 5, plug in the error function $E_{d,k,\sigma}(b) = \mathcal{O}(\sigma d^{\frac{3}{2}} b^{-\frac{1}{2}})$. The rest of the proof follows a similar argument in Corollary 2, as shown in Appendix D. □

Finally, we present a convergence result for GP sample path under noiseless assumption.

Theorem 4. *For $0 < \delta < 1$, suppose f is a GP sample whose smoothness constant is L with probability at least $1-\delta$. Assuming $\sigma = 0$, Algorithm 1 with batch size $b_t = d+1$ and step size $\eta_t = \frac{1}{L}$ produces a sequence satisfying*

$$\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L(f(\mathbf{x}_1) - f^*)}{T}$$

with probability at least $1-2\delta$.

Proof. By Theorem 2, we have

$$\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L(f(\mathbf{x}_1) - f^*)}{T} + \frac{1}{T}\sum_{t=1}^T C_t E_{d,k,\sigma}(d+1)$$

with probability at least $1-2\delta$. By Lemma 2, $E_{d,k,\sigma}(d+1) = 0$, the second term is essentially zero, which finishes the proof. □

D Optimizing the Batch Size

In this section, we optimize the batch size in Theorem 2 and give explicit convergence rates. The discussion in this section will give a proof for Corollary 2.

For the RBF kernel and $\nu = 2.5$ Matérn kernel, by Theorem 2, Lemma 4 and Lemma 5, we have shown the following bound

$$\begin{aligned}\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 &\lesssim \frac{1}{T}+\frac{1}{T}\sum_{t=1}^{T} E_{d,k,\sigma}(b_t)\log t\\ &\lesssim \frac{1}{T}+\frac{\sigma d^{\frac{3}{2}}}{T}\sum_{t=1}^{T} b_t^{-\frac{1}{2}}\log t.\end{aligned}$$

We discuss polynomially growing batch size $b_t = dt^a$, where $a > 0$. Then we have

$$\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 \lesssim \frac{1}{T}+\sigma d\cdot\frac{1}{T}\sum_{t=1}^{T} t^{-\frac{1}{2}a}\log t.$$

We discuss three cases: $0 < a < 2$, $a = 2$ and $a > 2$.

Case 1. When $0 \le a < 2$, the infinite sum $\sum_{t=1}^{T} t^{-\frac{1}{2}a}\log t$ diverges. Its growth speed is on the order of $\mathcal{O}(T^{1-\frac{1}{2}a}\log T)$. The total number of samples $n = \sum_{t=1}^{T} b_t = \mathcal{O}(dT^{a+1})$, and thus $T = \mathcal{O}(d^{-\frac{1}{a+1}} n^{\frac{1}{a+1}})$. Thus, the rate is

$$\begin{aligned}\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 &\lesssim T^{-1}+\sigma d\cdot T^{-\frac{1}{2}a}\log T\\ &\lesssim d^{\frac{1}{a+1}} n^{-\frac{1}{a+1}}+\sigma d^{\frac{3a+2}{2(a+1)}} n^{-\frac{a}{2(a+1)}}\log n\\ &\lesssim \sigma d^{\frac{3a+2}{2(a+1)}} n^{-\frac{a}{2(a+1)}}\log n,\end{aligned}$$

where the last inequality uses the fact that the second term dominates the rate when $0 < a < 2$.

Case 2. When $a = 2$, the infinite sum $\sum_{t=1}^{T} t^{-1}\log t$ diverges. Its growth speed is on the order of $\mathcal{O}(\log^2 T)$. The total number of samples is $n = d\sum_{t=1}^{T} t^2 = \mathcal{O}(dT^3)$, and thus $T = \mathcal{O}(d^{-\frac{1}{3}} n^{\frac{1}{3}})$. Then the rate is

$$\begin{aligned}\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 &\lesssim \frac{1}{T}+\frac{\sigma d\log^2 T}{T}\\ &\lesssim \sigma d T^{-1}\log^2 T\\ &\lesssim \sigma d^{\frac{4}{3}} n^{-\frac{1}{3}}\log^2 n.\end{aligned}$$

Case 3. When $a > 2$, the infinite sum $\sum_{t=1}^{T} t^{-\frac{1}{2}a}\log t$ converges to a constant when $T \to \infty$. The total number of samples $n = \mathcal{O}(dT^{a+1})$. Thus, the rate is

$$\begin{aligned}\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 &\lesssim T^{-1}+\sigma d T^{-1}\\ &\lesssim \sigma d^{\frac{a+2}{a+1}} n^{-\frac{1}{a+1}}.\end{aligned}$$

Note that $a = 2$ achieves the fastest rate $\mathcal{O}(n^{-\frac{1}{3}})$ in terms of samples n.

Now we discuss a batch size with logarithmic growth $b_t = d\log^2 t$. We have

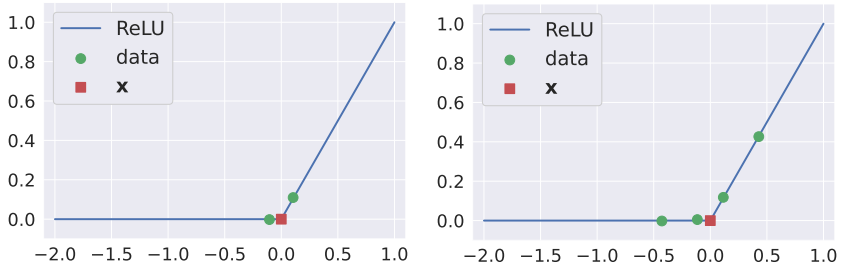
$$\begin{aligned}\min_{1\le t\le T}\|\nabla f(\mathbf{x}_t)\|^2 &\lesssim \frac{1}{T}+\frac{\sigma d}{T}\sum_{t=1}^{T}\mathcal{O}(1)\\ &\lesssim \frac{1}{T}+\sigma d.\end{aligned}$$

E Additional Experiments

This section presents additional experimental details and additional numerical simulations.

Details of Figure 2. In Figure 2a, we plot the error function starting from $b = 20$ to make sure $b \geq 2d$ so that Lemma 5 indeed applies. The decay rate $\mathcal{O}(\sigma d^{\frac{3}{2}} b^{-\frac{1}{2}})$ has a (leading) hidden constant of $\frac{15\sqrt{2}}{2}$ inside the big $\mathcal{O}$ notation (see the proof of Lemma 5), and thus the bounds plotted in Figure 2 are multiplied by this constant. Otherwise, the expression $\sigma d^{\frac{3}{2}} b^{-\frac{1}{2}}$ alone is not a valid upper bound.

ReLU. The ReLU function $\max\{0, x\}$ is non-differentiable at $x = 0$. Nevertheless, thanks to convexity the subdifferential at $x = 0$ is defined as $[0, 1]$. We estimate the "derivative" at $x = 0$ by minimizing the acquisition function. The estimate $\mu'_{\mathcal{D}}(0)$ and queries are plotted in Figure 5. The estimated derivative $\mu'_{\mathcal{D}}(0)$ is always in $[0, 1]$. Thus, the posterior mean gradient $\mu'_{\mathcal{D}}(0)$ produces a subgradient in this case.



(a) $n = 2$, $\mu'_{\mathcal{D}}(0) = 2.06 \times 10^{-5}$ (b) $n = 4$, $\mu'_{\mathcal{D}}(0) = 0.49$

Figure 5: Estimating the "derivative" of ReLU at $x = 0$ with noisy observations ($\sigma = 0.01$).