Geometric Rates of Convergence for Kernel-based Sampling Algorithms – Supplementary material

Rajiv Khanna[1] **Liam Hodgkinson**[1,2] **Michael W. Mahoney**[1,2]

[1]Department of Statistics, UC Berkeley
[2]International Computer Science Institute (ICSI), Berkeley

A APPENDIX

A.1 PROOF OF THEOREM 2

We begin by first proving Theorem 2, since the additional assumption of realizability makes it an easier read. For further ease of exposition, instead of directly working with $g(\cdot)$, we translate the function to remove any constants not dependent on the variable. We write,

$$l(\mathsf{S}) := \|\mu_\pi\|_k^2 - g(\mathsf{S}) = \mathbf{z}^\top \mathbf{K}^{-1}\mathbf{z}.$$

Some auxiliary Lemmas are proved later in this section. We use $Z(\mathsf{S}_j) := \sum_j w_j \phi(\mathbf{x}_j)$ Further, note that the Assumption 2, when applied for $h(\cdot)$, ensures that for any iterates considered in this proof we have that

$$\begin{aligned}
&-\frac{m_\omega}{2}\|Z(\mathsf{S}_i) - Z(\mathsf{S}_j)\|_k^2 \\
&\quad \geq l(\mathsf{S}_i) - l(\mathsf{S}_j) - \langle \nabla l(\mathsf{S}_i), Z(\mathsf{S}_i) - Z(\mathsf{S}_j)\rangle_k \\
&\quad \geq -\frac{M_\Omega}{2}\|Z(\mathsf{S}_i) - Z(\mathsf{S}_j)\|_k^2.
\end{aligned}$$

Proof. Say $(i-1)$ steps of the Algorithm 1 have been performed to select the set S. Let $\mathbf{w} \in \mathbb{R}^{(i-1)}$ be the corresponding weight vector. Let $h(\mathsf{S}, \mathbf{u}) := \|\mu_\pi\|_k^2 - \|\mu_\pi - \sum_j u_j \phi(\mathbf{x}_j)\|_k^2$, so that $l(\mathsf{S}) = \min_{\mathbf{u}} h(\mathsf{S}, \mathbf{u})$ (as per Lemma 1) . Set weight vector $\mathbf{u} \in \mathbb{R}^i$ as follows. For $j \in [0, i-1]$, $u_i = w_i$. Set $u_i = \alpha$, where α is an arbitrary scalar.

From weight optimality proved in Lemma 1,

$$l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) \geq h(\mathsf{S} \cup \{\mathbf{x}_i\}, \mathbf{u}) - l(\mathsf{S}),$$

for an arbitrary $\alpha \in \mathbb{R}$. From Assumption 2 (smoothness),

$$l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) \geq \alpha \langle \nabla l(\mathsf{S}), \phi(\mathbf{x}_i)\rangle_k - \alpha^2 \frac{M_\Omega}{2}.$$

Let γ_{S} be the optimum value of the solution of the inner LMO problem. Since $\mathbf{x}_i$ is the optimizing atom,

$$l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) \geq \alpha \gamma_{\mathsf{S}} - \alpha^2 \frac{M_\Omega}{2}.$$

Let $\mathsf{S}_\perp^\star$ be the set obtained by orthogonalizing $\mathsf{S}_r^\star$ with respect to S using the Gram-Schmidt procedure. Putting in $\alpha = \frac{\gamma_{\mathsf{S}}}{M_\Omega}$, we get,

$$\begin{aligned}
l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) &\geq \frac{1}{2M_\Omega}\gamma_{\mathsf{S}} && (1)\\
&\geq \frac{1}{2rM_\Omega}\sum_{\mathbf{x}_j \in \mathsf{S}_\perp^\star} \langle \phi(\mathbf{x}_j), \nabla l(\mathsf{S})\rangle_k^2 \\
&\geq \frac{m_\omega}{rM_\Omega}\left(l(\mathsf{S} \cup \mathsf{S}_\perp^\star) - l(\mathsf{S})\right) && (2)\\
&\geq \frac{m_\omega}{rM_\Omega}\left(l(\mathsf{S}_r^\star) - l(\mathsf{S})\right) \\
&= \frac{m_\omega}{rM_\Omega}\left(\|\mu_\pi\|_k^2 - l(\mathsf{S})\right).
\end{aligned}$$

The second inequality is true because $\gamma_{\mathsf{S}} = \langle \nabla l(\mathsf{S}), \mathbf{x}_i\rangle_k$ is the optimum value of the inner LMO problem in the i^{th} iteration. The third inequality follows from Lemma 2. The fourth inequality is true because of monotonicity of $l(\cdot)$, and the final equality is true because of Assumption 1 (realizability).

Let $C := \frac{m_\omega}{rM_\Omega}$. We have $l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) = g(\mathsf{S}) - g(\mathsf{S} \cup \{\mathbf{x}_i\}) \geq Cg(\mathsf{S}) \implies g(\mathsf{S} \cup \{\mathbf{x}_i\}) \leq (1-C)g(\mathsf{S})$. The result now follows.

□

A.2 PROOF OF THEOREM 1

Proof. We proceed as in the proof of Theorem 2, but by replacing $\mathsf{S}_r^\star$ with T_r. From (2),

$$l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S}) \geq \frac{m_\omega}{rM_\Omega}\left(l(\mathsf{T}_r) - l(\mathsf{S})\right).$$

Adding and subtracting $l(\mathsf{T}_r)$ on the LHS and rearranging,

$$l(\mathsf{T}_r) - l(\mathsf{S} \cup \{\mathbf{x}_i\}) \le (1 - \frac{m_\omega}{rM_\Omega})(l(\mathsf{T}_r) - l(\mathsf{S})).$$

Thus after k iterations,

$$l(\mathsf{T}_r) - l(\mathsf{S}_k) \le (1 - \frac{m_\omega}{rM_\Omega})^k \left(l(\mathsf{T}_r) - l(\emptyset)\right).$$

Rearranging,

$$\begin{aligned} l(\mathsf{S}_k) &\ge \left(1 - (1 - \frac{m_\omega}{rM_\Omega})^k\right) l(\mathsf{T}_r) \\ &\ge \left(1 - \exp(-\frac{km_\omega}{rM_\Omega})\right) l(\mathsf{T}_r). \end{aligned}$$

With $k = (r\frac{M_\Omega}{m_\omega}\log\frac{1}{\epsilon})$, we get,

$$l(\mathsf{S}_k) \ge (1-\epsilon) l(\mathsf{T}_r).$$

The result now follows. $\square$

A.3 AUXILIARY LEMMAS

The following Lemma proves that the weights w_i in $g(\mathsf{S})$ obtained using the posterior inference are an optimum choice that minimize the distance to μ_π in the RKHS over any set of weights [Khanna et al., 2019].

Lemma 1. *The residual $\mu_\pi - \sum_j w_j\phi(\mathbf{x}_j)$ is orthogonal to $\mathbf{x}_i \in \mathsf{S}\,\forall i$. In other words, for any set of samples S, $g(\mathsf{S}) = \min_{\mathbf{u}} \|\mu_\pi - \sum_i u_i\phi(\mathbf{x}_i)\|_k$.*

Proof. Recall that $w_i = \sum_j [\mathbf{K}^{-1}]_{ij}\mathbf{z}_j$, and $\mathbf{z}_i = \int k(\mathbf{x}, \mathbf{x}_i) d\pi(\mathbf{x})$. For an arbitrary index i,

$$\begin{aligned} &\langle \mu_\pi - \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_k \\ &\quad= \int k(\mathbf{x}, \mathbf{x}_i) d\pi(\mathbf{x}) - \langle \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_k \\ &\quad= \mathbf{z}_i - \langle \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_k \\ &\quad= \mathbf{z}_i - \sum_j w_j k(\mathbf{x}_j, \mathbf{x}_i) \\ &\quad= \mathbf{z}_i - \sum_j \sum_t [\mathbf{K}^{-1}]_{tj}\mathbf{z}_t k(\mathbf{x}_j, \mathbf{x}_i) \\ &\quad= \mathbf{z}_i - \sum_t \mathbf{z}_t \sum_j \mathbf{K}_{ji}[\mathbf{K}^{-1}]_{tj} \\ &\quad= \mathbf{z}_i - \mathbf{z}_i, \end{aligned}$$

where the last equality follows by noting that $\sum_j \mathbf{K}_{ji}[\mathbf{K}^{-1}]_{tj}$ is inner product of row i of $\mathbf{K}$ and row t of $\mathbf{K}^{-1}$, which is 1 if $t = i$ and 0 otherwise. This completes the proof. $\square$

Lemma 2. *For any set of chosen samples S_1, S_2, let $\mathcal{P}$ be the operator of projection onto span($\mathsf{S}_1 \cup \mathsf{S}_2$). Then, $l(\mathsf{S}_1 \cup \mathsf{S}_2) - l(\mathsf{S}_1) \le \frac{\mathcal{P}(\nabla l(\mathsf{S}_1))}{2m_\omega}$.*

Proof. Observe that

$$\begin{aligned} 0 &\le l(\mathsf{S}_1 \cup \mathsf{S}_2) - l(\mathsf{S}_1) \\ &\le \langle \nabla l(\mathsf{S}_1), Z(\mathsf{S}_1 \cup \mathsf{S}_2) - Z(\mathsf{S}_1)\rangle_k \\ &\quad - \frac{m_\omega}{2}\|Z(\mathsf{S}_1 \cup \mathsf{S}_2) - Z(\mathsf{S}_1)\|_k^2 \\ &\le \underset{X \in \mathsf{span}(\mathsf{S}_1 \cup \mathsf{S}_2)}{\arg\max} \langle \nabla l(\mathsf{S}_1), X - Z(\mathsf{S}_1)\rangle_k - \frac{m_\omega}{2}\|X - Z(\mathsf{S}_1)\|_k^2 \\ &= \arg\max_X \langle \mathcal{P}(\nabla l(\mathsf{S}_1)), X - Z(\mathsf{S}_1)\rangle_k - \frac{m_\omega}{2}\|X - Z(\mathsf{S}_1)\|_k^2. \end{aligned}$$

Solving the argmax problem on the RHS for X, we get the required result.

$\square$

A.4 PROOF OF THEOREM 3

We next present some notation and few lemmas that lead up to the main result of this section (Theorem 3). The domain of candidate atoms $\mathcal{X}$ is split into $\{\mathcal{X}_j, j \in [s]\}$ over s machines, where machine j runs WKH on $\mathcal{X}_j$. Let G_j be the k-sized solution returned by running Algorithm 1 on $\mathcal{X}_j$, i.e., $\mathsf{G}_j = \text{WKH}(\mathcal{X}_j, k)$. Note that each $\mathcal{X}_j$ induces a partition onto the optimal r-sized solution $\mathsf{S}_r^\star$ as follows ($r = 1$ for this theorem):

$$\begin{aligned} \mathsf{T}_j &:= \{x \in \mathsf{S}_1^\star : x \notin \text{WKH}(\mathcal{X}_j \cup x, k)\}, \\ \mathsf{T}_j^c &:= \{x \in \mathsf{S}_1^\star : x \in \text{WKH}(\mathcal{X}_j \cup x, k)\}. \end{aligned}$$

In other words, $\mathsf{T}_j = \mathsf{S}_1^\star$ if the j^{th} machine running WKH on $\mathcal{X}_j \cup \mathsf{S}_1^\star$ will not select it as among its output, and it is empty otherwise, since $\mathsf{S}_1^\star$ is a singleton. We re-use the definition of $l(\cdot)$ used in Appendix A.1.

Before moving to the proof of the main theorem, we prove two prerequisites. Recall G_j is the set of iterates selected by machine j. In this mini-result, we lower bound the expected improvement in the loss at the aggregator machine.

Lemma 3. *For the aggregator machine that runs* WKH *over $\cup_j \mathsf{G}_j$ (step 6 of Algorithm 2), we have, at selection of next sample point $\mathbf{x}_i$ after having selected S, $\exists$ machine j such that*

$$\mathbb{E}[l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S})] \ge \frac{m_\omega}{M_\Omega}\mathbb{E}\left(l(\mathsf{T}_j^c) - l(\mathsf{S})\right).$$

Proof. The expectation is over the random split of $\mathcal{X}$ into $\mathcal{X}_j$ for $j \in [s]$. We denote T_j^c to be the complement of T_j. Then, we have that

$$
\begin{aligned}
&\mathbb{E}[l(\mathsf{S} \cup \{\mathbf{x}_i\}) - l(\mathsf{S})] \\
&\quad \geq \mathbb{E}[\frac{1}{2M_\Omega}\gamma_{\mathsf{S}}] \\
&\quad \geq \frac{1}{2M_\Omega} \sum_{\mathbf{x} \in \mathsf{S}_1^\star} \mathbb{P}(\mathbf{x} \in \cup_j \mathsf{G}_j) \mathbb{E}\langle \phi(\mathbf{x}), \nabla l(\mathsf{S}) \rangle_k^2 \\
&\quad = \frac{1}{2sM_\Omega} \sum_{\mathbf{x} \in \mathsf{S}_1^\star} \left[\sum_{b=1}^{s} \mathbb{P}(\mathbf{x} \in \mathsf{T}_b^c) \right] \mathbb{E}\langle \phi(\mathbf{x}), \nabla l(\mathsf{S}) \rangle_k^2 \\
&\quad = \frac{1}{2sM_\Omega} \sum_{b=1}^{s} \sum_{\mathbf{x} \in \mathsf{T}_b^c} \mathbb{E}\langle \phi(\mathbf{x}), \nabla l(\mathsf{S}) \rangle_k^2 \\
&\quad \geq \frac{m_\omega}{sM_\Omega} \sum_{b=1}^{s} \mathbb{E}\left(l(\mathsf{S} \cup \mathsf{T}_b^c) - l(\mathsf{S}) \right) \\
&\quad \geq \frac{m_\omega}{sM_\Omega} \sum_{b=1}^{s} \mathbb{E}\left(l(\mathsf{T}_b^c) - l(\mathsf{S}) \right) \\
&\quad \geq \frac{m_\omega}{M_\Omega} \min_{b \in [s]} \mathbb{E}\left(l(\mathsf{T}_b^c) - l(\mathsf{S}) \right).
\end{aligned}
$$

The equality in step 3 above is because of Lemma 5. □

In the following lemma, we lower bound the greedy improvement in the loss on each machine.

Lemma 4. *For any individual worker machine j running local* WKH*, if S is the set of $(i-1)$ iterates already chosen, then at the selection of next sample point $\mathbf{x}_i$, $l(\mathsf{S} \cup \{\mathbf{x}_i\}) \geq (l(\mathsf{T}_j) - l(\mathsf{S}))$.*

Proof. We proceed as in proof of Theorem 2 in Appendix A.1. From (1), we have,

$$
\begin{aligned}
l(\mathsf{S} \cup \{\mathbf{x}\}) - l(\mathsf{S}) &\geq \frac{1}{2M_\Omega}\gamma_{\mathsf{S}} \\
&\geq \frac{1}{2M_\Omega} \sum_{\mathbf{x}_j \in \mathsf{T}_j} \langle \phi(\mathbf{x}_j), \nabla l(\mathsf{S}) \rangle_k^2 \\
&\geq \frac{m_\omega}{M_\Omega} \left(l(\mathsf{S} \cup \mathsf{T}_j) - l(\mathsf{S}) \right) \\
&\geq \frac{m_\omega}{M_\Omega} \left(l(\mathsf{T}_j) - l(\mathsf{S}) \right).
\end{aligned}
$$

□

We are now ready to prove Theorem 3.

Proof of Theorem 3. If, for a random split of $\mathcal{X}$, for any $j \in [s]$, $\mathsf{T}_j = \mathsf{S}_1^\star$, then the given rate follows from Lemma 4, after following the straightforward steps covered in proof of Theorem 2 for proving the rate from the given condition on $l(\cdot)$. On the other hand, if none of $j \in [s]$, $\mathsf{T}_j = \mathsf{S}_1^\star$, then $\forall j \in [s], \mathsf{T}_j = \emptyset \implies \mathsf{T}_j^c = \mathsf{S}_1^\star$. In this case, the given rate follows from Lemma 3. □

Finally, here is the statement and proof of an auxiliary lemma that was used above.

Lemma 5. *For any $x \in \mathcal{X}, \mathbb{P}(x \in \cup_j \mathsf{G}_j) = \frac{1}{s}\sum_j \mathbb{P}(x \in \mathsf{T}_j^c)$.*

Proof. We have

$$
\begin{aligned}
&\mathbb{P}(x \in \cup_j \mathsf{G}_j) \\
&\quad = \sum_j \mathbb{P}(x \in \mathcal{X}_j \cap x \in \text{WKH}(\mathcal{X}_j, k)) \\
&\quad = \sum_j \mathbb{P}(x \in \mathcal{X}_j)\mathbb{P}(x \in \text{WKH}(\mathcal{X}_j, k) | x \in \mathcal{X}_j) \\
&\quad = \sum_j \mathbb{P}(x \in \mathcal{X}_j)\mathbb{P}(x \in \mathsf{T}_j^c) \\
&\quad = \frac{1}{s}\mathbb{P}(x \in \mathsf{T}_j^c).
\end{aligned}
$$

□