Institut für Mathematik
Fakultät II – Mathematik und Naturwissenschaften
Technische Universität Berlin

Internship Report

Kernel herding as a Frank-Wolfe algorithm

Garam Kim
459187
Scientific Computing

Supervisors: Elias Wirth, Dr. Mathieu Besançon
Institution: Zuse Institute Berlin
Address: Takustraße 7, 14195 Berlin
Duration: 15.08.2022 - 03.11.2022 (212 hours, 12 weeks)

Contents

1 The Company

The Zuse Institute Berlin (ZIB) is a research institute dedicated to the advancement of mathematics and computer science. Established in 1984, the institute is named after Konrad Zuse, a renowned German computer scientist and inventor. ZIB conducts research in a wide range of areas such as optimization, simulation, visualization, and scientific computing. It provides mathematical and computational support to a variety of fields including physics, engineering, economics, and life sciences.

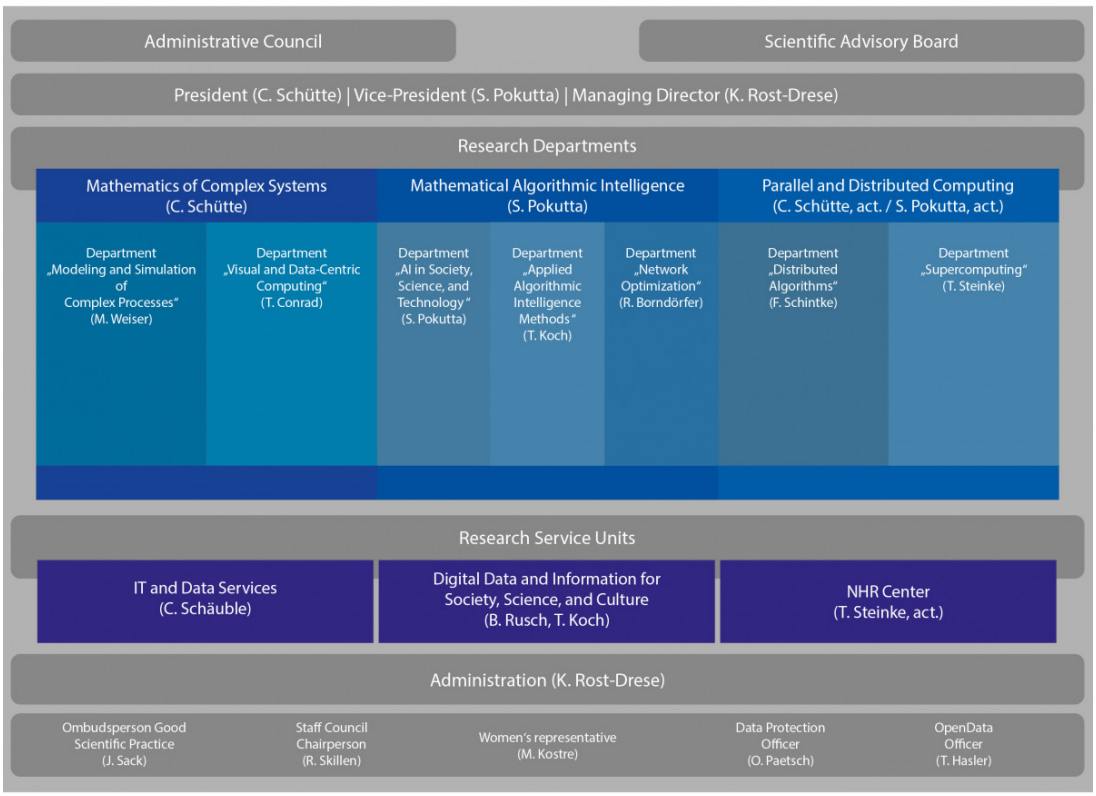


Figure 1: Organizational structure of Zuse Institute Berlin

One of the main goals of ZIB is to bridge the gap between mathematics and computer science by developing innovative algorithms and software tools for solving complex mathematical problems. This is achieved through close collaboration between mathematicians, computer scientists, and domain experts from various scientific fields.
ZIB is also involved in numerous interdisciplinary research projects, both nationally and internationally. Its researchers are frequently involved in collaborations with industry partners, and the institute actively seeks to transfer its research results to real-world applications. In addition to its research activities, ZIB provides a variety of services such as consulting, training, and workshops on topics related to mathematics and computer science.

1.1 Overview

The objective of the internship was solving open questions in Wirth et al. (2022), that is, proving a slower convergence rate of $\Omega(1/t)$ of the Frank-Wolfe algorithm (FW) with line search than FW with open-loop at a rate of $\mathcal{O}(1/t^2)$ in the infinite-dimensional setting. Working towards the goal, I understood kernel herding as well as FW algorithm and learned advanced `Julia` programming language. I begin with the finite kernel introduced in Bach et al. (2012), proving its convergence rate and implementing the code. All code is available on GitHub. Next, I studied the Matérn kernel, which is infinite-dimensional. The proof of the open question is almost in the and will be continued to my master's thesis topic.

In the rest of the report, I use "we" to describe mathematical description by convention, and "I" to give my personal experience.

2 Research Activities

During the internship, I was supervised by Elias Wirth and Dr. Mathieu Besançon, a graduate student and a post-doctor at the department "AI in society, science and technology (AISST)" of ZIB, respectively. We discussed the progress and I asked for theoretical as well as soft-skill questions.

2.1 Kernel Herding and Frank-Wolfe algorithm

In the first two weeks, I read and understood Kernel herding and the relation between the Frank-Wolfe algorithm and kernel herding, and learned `Julia` programming language.

Week 1: Understanding the topic of Kernel Herding, especially herding as a Frank-Wolfe algorithm, e.g., Bach et al. (2012); Chen et al. (2010); Wirth et al. (2022)

Week 2: Learning `Julia` programming language, toolbox for Frank-Wolfe which designed to solve optimization problem. Reading Julia tutorial

Frank-Wolfe algorithm

Frank-Wolfe algorithm is a popular first-order method for constrained convex optimization, i.e., solving

$$\min_{x \in \mathcal{C}} f(x),$$

where $\mathcal{C} \subset \mathbb{R}^d$ is a compact convex feasible set, and $f : \mathcal{C} \to \mathbb{R}$ is a convex and smooth function. These method, described in Algorithm 1, do not require projections on the feasible set $\mathcal{C}$, but access to a linear minimization oracle (LMO) which returns $\arg\min_{x \in \mathcal{C}} \langle c, x \rangle$.

Algorithm 1: Frank-Wolfe algorithms (FW)

Input: $x_0 \in \mathcal{C}$, convex function f

for $t = 0, 1, \cdots, T-1$ **do**

  $p_t \in \arg\min_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle$

  $x_{t+1} \leftarrow (1 - \eta_t) x_t + \eta_t p_t$

end

At each iterations, the minimum of $\langle \nabla f(x_t), p - x_t \rangle$ over $p \in \mathcal{C}$ is computed and the next iterate is taken as a convex combination of x_t and p_t, i.e., $x_{t+1} \leftarrow (1 - \eta_t) x_t + \eta_t p_t$. There are different choices of step-size, η_t as open loop step-size, $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \{1, 2\}$, line search $\eta_t \in \arg\min_{\eta \in [0,1]} f((1 - \eta) x_t + \eta p_t)$, or short-step stpe-size.

Kernel Herding

The herding algorithm is introduced by Welling (2009) that converts observed moments into a sequence of pseudo-sample. The algorithm generates samples deterministically, avoiding randomness, and is simpler computationally.
Let $\mathcal{X} \subset \mathbb{R}^d$ be an observation space, $\mathcal{H}$ a Reproducing Kernel Hilbert Space (RKHS) and $\Phi : \mathcal{X} \to \mathcal{H}$ a feature map. From Christmann and Steinwart (2008) Lemma 4.19, the feature map of RKHS is given by

$$\Phi(x) = k(\cdot, x),$$

for $x \in \mathcal{X}$. The positive definite kernel associated with Φ is denoted by $k : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, as $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$ for $x, y \in \mathcal{X}$. Let $\mathcal{M} \subset \mathcal{H}$ be the marginal polytope, the convex hull of all functions $\Phi(x)$ for $x \in \mathcal{X}$, i.e., $\mathcal{M} =: \mathrm{conv}(\{\Phi(x) \mid x \in \mathcal{X}\})$. Consider a fixed probability distribution $p(x)$ over $\mathcal{X}$ and denote μ the empirical moment vector on $\mathcal{X}$:

$$\mu = \mathbb{E}_{p(x)} \Phi(x) \in \mathcal{M}.$$

The herding algorithm generates pseudo-samples x_t and updates weight vector $w \in \mathcal{H}$ for given a distribution $p(x)$, as in Algorithm 2.

Algorithm 2: Kernel Herding

Input: $w_1 \in \mathcal{H}$

1 **for** $i = 1, 2, \cdots, t-1$ **do**

2 $\quad x_{i+1} \in \arg\max_{x \in \mathcal{X}} \langle w_i, \Phi(x) \rangle_{\mathcal{H}}$

3 $\quad w_{i+1} \leftarrow w_i + \mu_p - \Phi(x_{i+1})$

4 **end**

With an initialization of $w_1 \in \mathcal{H}$, Line 3 of algorithm 2 yields following expression:

$$w_t = w_1 + t\mu - \sum_{i=1}^{t} \Phi(x_i).$$

Assuming that $w_1 = \mu$, and the further restrictions, Chen et al. (2010) show that herding minimizes the squared error $\mathcal{E}_T^2$ defined as

$$\mathcal{E}_t^2 = \frac{1}{2} \| \mu - \frac{1}{t} \sum_{i=1}^{t} \Phi(x_i) \|^2. \tag{1}$$

Therefore, this algorithm constructs an associated empirical distribution $\hat{p}$ that is close to the true distribution p by generating pseudo-samples that greedily minimize this error at every iteration. Bach et al. (2012) suggests the learning perspective: approximate μ by considering n points $x_1, x_2, \ldots, x_n \in \mathcal{X}$ with positive weights $w_1, w_2, \ldots, w_n$ such that $\sum_{i=1}^{n} w_i = 1$, which defines $\hat{p}$ and $\hat{\mu}$ by

$$\hat{\mu} = \mathbb{E}_{\hat{p}(x)} \Phi(x) = \sum_{i=1}^{n} w_i \Phi(x_i) \in \mathcal{M}.$$

We thus get

$$\mathcal{E}^2 = \frac{1}{2} \| \mu - \hat{\mu} \|^2 = \frac{1}{2} \| \mu_p - \sum_{i=1}^{n} w_i \Phi(x_i) \|^2,$$

which is the generalization of (1). Hence, this finds a good estimate $\hat{\mu}$ of μ based on a weighted set of points from $\{\Phi(x), x \in \mathcal{X}\}$.

Another aspect of the herding algorithm is that the process remembers all previous samples and guides away from regions which have already been over-sampled. Therefore, kernel herding has a faster convergence rate of the error of functions in the Hilbert space in comparison to iid (independent and identical distribution) random sampling. That is, iid converges at a rate of $\mathcal{O}(1/\sqrt{t})$, kernel herding decreases it at a rate to $\mathcal{O}(1/t)$. To obtain the same order as i.i.d., the herding algorithm only requires $\sqrt{t}$ samples, see, e.g., Bach et al. (2012); Welling (2009); Chen et al. (2010).

Equivalence between Frank-Wolfe algorithms and kernel herding

The equivalence between kernel herding and FW is well-established. (Bach et al., 2012; Chen et al., 2010; Tsuji et al., 2021)

Consider the following optimization problem:

$$\min_{g \in \mathcal{M}} J(g) = \frac{1}{2}\|g - \mu\|^2. \tag{OPT-KH}$$

Following from the algorithm 1, FW to solve (OPT-KH) with step-size η_t uses the iterates

$$\begin{aligned} v_t &\in \arg\min_{g \in \mathcal{M}} \langle g_t - \mu, g \rangle \\ g_{t+1} &= (1 - \eta_t) g_t + \eta_t v_t. \end{aligned} \tag{2}$$

Since all extreme points of $\mathcal{M}$ are of the form $\Phi(x)$ for $x \in \mathcal{X}$ (Bach et al., 2012) and LMO always returns vertex of the feasible set, (2) returns an extreme point of $\mathcal{M}$, implying that $v_t = \Phi(x_t)$ for a certain $x_t \in \mathcal{H}$. By changing a variable $g_t = \mu - w_t/t$, the updates with step-size $\eta_t = \frac{1}{t+1}$ are exactly same as kernel herding. That is, i.e., we get $(t+1)g_{t+1} = tg_t + \Phi(x_t)$, and thus get uniform weights. General step-size $\eta_t \in [0, 1]$ leads to non-uniform weights.
Wirth et al. (2022) present the accelerated convergence rate of FW with the open loop step-size rule using the setting in Bach et al. (2012), and Wahba (1990), explaining that the equivalence of FW with kernel herding leads to a new convergence rate of $\mathcal{O}(1/t^2)$.

2.2 Finite-dimensional kernel

For the next four weeks, I began with a nite dimensional kernel introduced by Bach et al. (2012).

> **Week 3**: To implement a finite-dimensional kernel, I defined a suitable dot product and implemented it for the 1-dimensional case.
>
> **Week 4**: Successfully implement higher-dimensional kernel. Moreover, I made the test.jl to check its viability for every case.
>
> **Week 5, 6**: Establish mathematical description and show $\mathcal{O}(1/t^2)$ convergence rate.

2.2.1 Mathematical description

Let $\mathcal{X} = \{-1, 1\}^d$ be a feature space with dimensivon $d \in \mathbb{N}$, and

$$\mathcal{H} := \left\{ f \colon \mathcal{X} \to \mathbb{R} \mid f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}, \text{ and } \Phi(x) = (x, xx^T) \right\}, \tag{3}$$

a RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined by

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{d} f_i(x) g_i(x)$$

for $f, g \in \mathcal{H}$, where $f = [f_1, f_2, \ldots, f_d]$. For $x \in \mathcal{X}$, the feature map $\Phi(x) = (x, xx^T)$ is composed of x and of all of its pairwise products xx^T. Then the reproducing kernel of $\mathcal{H}$ is indeed

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \sum_{i=1}^{d} \Phi_i(x) \Phi_i(y) = \langle x, y \rangle_2, \tag{4}$$

for $x, y \in \mathcal{X}$. In particular, it holds that

$$k(x, y) = \big\langle k(z, x), k(z, y) \big\rangle_{\mathcal{H}}$$

for any $x, y, z \in \mathcal{X}$. We denote $\mathcal{M} \subset \mathcal{H}$ the marginal polytope defined by $\mathcal{M} = \text{conv}(\{\Phi(x) \mid x \in \mathcal{X}\}$. In this setting, we compute the expectation

$$\mu := \mathbb{E}_{p(x)} \Phi(x) = \sum_{i=1}^{2^d} p_i(x) \Phi_i(x) \in \mathcal{M}.$$

From the fact that LMO always returns an element of the form $\Phi(x) \in \mathcal{M}$ for $x \in \mathcal{X}$, the iterate g_t constructed with FW is of the form $\sum_{i=1}^{t} w_i \Phi(x_i)$. The associated empirical mean with corresponding empirical distribution $\hat{p}(x)$ is defined by

$$\hat{\mu} := \mathbb{E}_{\hat{p}(x)} \Phi(x) = \sum_{i=1}^{t} w_i \Phi(x_i) = g_t.$$

Note that in the kernel herding setting, a quadratic objective function gives that line search and short-step step-size rules coincide.

Theorem 1. Let $\mathcal{H}$ be the Hilbert space defined in (3), let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{H}$ be the kernel defined in (4). Suppose that there exists a ball of center x^* and radius $d > 0$ that is contained in $\mathcal{M}$. For the iterates of Algorithm 1 with line-search or short step size rule solving (OPT-KH), at iteration $t \in \mathbb{N}$, it holds that

$$J(g_t) - J(\mu) = \frac{1}{2}\|g_t - \mu\|_{\mathcal{H}}^2 \leq R^2 \exp\left(-\frac{d^2 t}{R^2}\right),$$

where R is a diameter of the marginal polytope $\mathcal{M}$.

Proof. For completeness, we repeat the proof from Beck and Teboulle (2004) and add some additional explanations. Let $g_{t+1} = g_t + \eta_t(v_t - g_t)$ be FW iterates and $\eta_t = \frac{\langle g_t - \mu, g_t - v_t \rangle_{\mathcal{H}}}{\|g_t - v_t\|_{\mathcal{H}}^2}$ be a line-search or short step step-size rule. Then, it holds that

$$\begin{aligned}
\|g_{t+1} - \mu\|_{\mathcal{H}}^2 &= \|g_t - \mu + \eta_t(v_t - g_t)\|_{\mathcal{H}}^2 \\
&= \|g_t - \mu\|_{\mathcal{H}}^2 + 2\eta_t \langle g_t - \mu, v_t - g_t \rangle_{\mathcal{H}} + \eta_t^2 \|v_t - g_t\|_{\mathcal{H}}^2 \\
&= \|g_t - \mu\|_{\mathcal{H}}^2 - 2\frac{\langle g_t - \mu, g_t - v_t \rangle_{\mathcal{H}}^2}{\|g_t - v_t\|_{\mathcal{H}}^2} + \frac{\langle g_t - \mu, g_t - v_t \rangle_{\mathcal{H}}^2}{\|g_t - v_t\|_{\mathcal{H}}^2} \\
&= \|g_t - \mu\|_{\mathcal{H}}^2 - \frac{\langle g_t - \mu, g_t - v_t \rangle_{\mathcal{H}}^2}{\|g_t - v_t\|_{\mathcal{H}}^2} \\
&= \frac{\|g_t - \mu\|_{\mathcal{H}}^2 \|g_t - v_t\|_{\mathcal{H}}^2 - \langle g_t - \mu, g_t - v_t \rangle_{\mathcal{H}}^2}{\|g_t - v_t\|_{\mathcal{H}}^2} \\
&\leq \frac{\|g_t - \mu\|_{\mathcal{H}}^2 \left(\|g_t - v_t\|_{\mathcal{H}}^2 - d^2\right)}{\|g_t - v_t\|_{\mathcal{H}}^2} \\
&= \|g_t - \mu\|_{\mathcal{H}}^2 \left(1 - \frac{d^2}{\|g_t - v_t\|_{\mathcal{H}}^2}\right) \\
&\leq \|g_t - \mu\|_{\mathcal{H}}^2 \left(1 - \frac{d^2}{R^2}\right),
\end{aligned}$$

where the sixth inequality holds due to Proposition 3.1 from Beck and Teboulle (2004) and the last inequality follows from diameter of the marginal polytope. Then, by Polyak (1987),

$$J(g_t) - J(\mu) = \frac{1}{2}\|g_t - \mu\|_{\mathcal{H}}^2 \leq R^2 \exp\left(-\frac{d^2 t}{R^2}\right)$$

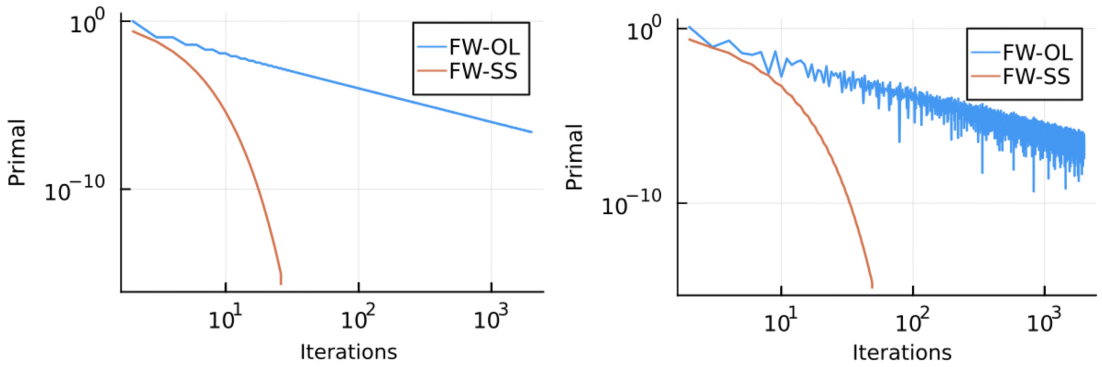
is satisfied for $t \in \mathbb{N}$.

□

For the version of open-loop step-size, Wirth et al. (2022) shows $\mathcal{O}(1/t^2)$ convergence when an optimum lies in the interior of the set. Moreover, experiments show the exact convergence for the finite kernel both line-search and open-loop step-size rules.

2.2.2 Experiments

For $d = 2$, we address (OPT-KH) with the setting defined in section 2.2.1. Both uniform and non-uniform distribution $p(x)$ are considered such that $\sum_{i=1}^{2^d} p_i(x) = 1$ by random number generator in `Julia`. The linear minimization oracle is implemented as a greedy search over the feature space, and the algorithms are run for 2000 iterations. The results of the experiments are plotted in log-log plots in Figure 2.



(a) uniform distribution (b) non-uniform distribution

Figure 2: Solving (OPT-KH) with FW with short step (FW-SS) and open loop step-size rules for the form $\eta_t = \frac{2}{t+2}$ (FW-OL) with uniform and non-uniform distributions, Figure 2a and 2b, respectively.

Results

I observe convergence rates of $\mathcal{O}(1/t^2)$ for FW with open loop step-size rule of the form $\eta_t = \frac{2}{t+2}$, and FW with line search appears to be converging linearly in both uniform and non-uniform cases, respectively.

2.3 Infinite-dimensional Kernel Herding

Week 7: Learn about the Matérn kernel and Bernoulli kernel which are the infinite-dimensional kernel.

Week 8, 9: Implement generalized Bernoulli kernel and Matérn kernel for the 1-dimensional case.

Week 10, 11: Attempt to prove the open question in Wirth et al. (2022), $\mathcal{O}(1/t^2)$ convergence rate of Frank-Wolfe algorithm with open-loop step-size for the Matérn kernel case. Since the assumption was too strong, instead, I decided to prove another open question, that in the kernel herding setting of Figure 3, FW with line search converges at a rate of $\Omega(1/t)$.

2.3.1 Matérn kernel

We focus on a specific kernel studied in Tsuji et al. (2022), that has an infinite dimensional feature space. Let $\mathcal{X} = [-1, 1]$ and

$$\mathcal{H} := \left\{ f \colon \mathcal{X} \to \mathbb{R} \mid f = \sum_{i=1}^{\infty} c_i k(\cdot, x_i) \text{ where } c_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}. \tag{5}$$

Kanagawa et al. (2018) showed that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) generated by the Matérn kernel

$$k(y, z) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|y - z\|_2}{\rho} \right)^{\nu} B_\nu \left(\sqrt{2\nu} \frac{\|y - z\|_2}{\rho} \right), \tag{6}$$

where $y, z \in \mathcal{X}$, Γ is the gamma function and B_ν is the modified Bessel function of the second kind and ρ and ν are positive parameters. If ν can be written as $\nu = m + \frac{1}{2}$ for non-negative integer m, then (6) has an explicit forms derived from Rasmussen (2004), giving

$$k(y,z) = \exp\left(-\sqrt{2\nu}\frac{\|y-z\|_2}{\rho}\right)\frac{\Gamma(m+1)}{\Gamma(2m+1)}\sum_{i=0}^{m}\frac{(m+i)!}{i!(m-i)!}\left(\sqrt{8\nu}\frac{\|y-z\|_2}{\rho}\right)^{m-i}.$$

In the following, we use the parameter $(\rho, \nu) = (\sqrt{3}, \frac{3}{2})$, which yields,

$$k(y,z) = (1 + \|y-z\|_2)\ \exp\left(-\|y-z\|_2\right). \tag{7}$$

Considering the uniform distribution, i.e., $p(y) := 1/2$ for all $y \in \mathcal{X}$, then for all $x \in \mathcal{X}$, it holds that

$$\mu(x) = \int_{\mathcal{Y}} k(x,y)p(y)dy = \frac{1}{2}\int_{\mathcal{X}} k(x,y)dy = \frac{1}{2}\left(e^{x-1}(x-3) - e^{-x-1}(x+3) + 4\right). \tag{8}$$

Hence, we have

$$\begin{aligned}
\|\mu\|_{\mathcal{H}}^2 &= \langle \mu(t), \mu(y)\rangle_{\mathcal{H}} \\
&= \left\langle \frac{1}{2}\int_{\mathcal{X}} k(x,t)dt, \frac{1}{2}\int_{\mathcal{X}} k(x,y)dy \right\rangle_{\mathcal{H}} \\
&= \frac{1}{4}\int_{\mathcal{X}}\int_{\mathcal{X}} k(t,y)dtdy \\
&= \frac{1}{2}\left(1 + \frac{5}{e^2}\right).
\end{aligned}$$

Let $a(x) = \sum_j w_j k(t_j, x)$ where $t_j \in \mathcal{X}$ such that $k(t_j, x)$ corresponds to a vertex of $\mathcal{M}$. Then

$$\begin{aligned}
\langle a(x), \mu(x)\rangle_{\mathcal{H}} &= \left\langle \sum_j w_j k(t_j,x), \frac{1}{2}\int_{\mathcal{X}} k(x,y)dy \right\rangle_{\mathcal{H}} \\
&= \frac{1}{2}\sum_j w_j \int_{\mathcal{X}} k(t_j,y)dy.
\end{aligned} \tag{9}$$

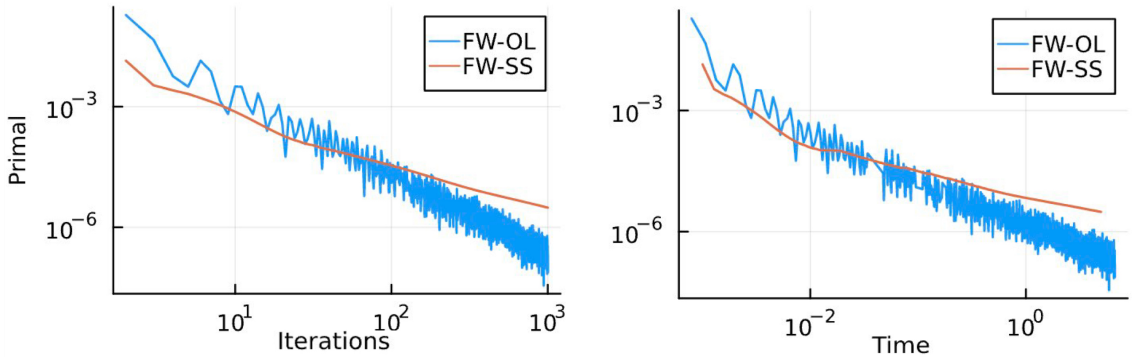
Wirth et al. (2022) prove $\mathcal{O}(1/t^2)$ convergence rate in Wahba Kernel with uniform distribution specifically, but the general proof of kernel herding problem and FW is still not fully understood. Hence, we implement experiments using the Matern Kernel and observe its convergence rate.

2.3.2 Experiments

Consider the infinite-dimensional kernel herding setting of Section 2.3.1 over $\mathcal{X} = [0,1]$. I address (OPT-KH) with FW with line search/short step (FW-SS) and open-loop step-size rules (FW-OL) of the form $\eta_t = \frac{2}{t+2}$. The linear minimization oracle is implemented as a greedy search over the feature space, and the algorithms are run for 1000 iterations. The results of the experiments are plotted in log-log plots in Figure 3.

Results

I observe convergence rates of $\mathcal{O}(1/t^2)$ for FW with open loop step-size rule of the form $\eta_t = \frac{2}{t+2}$, whereas FW with line search/short step step-size rules converges at a rate of $\mathcal{O}(1/t)$. This gives similar results of Wirth et al. (2022), which supports unexplained phenomenon in infinite-dimensional kernel herding.



(a) primal gap (b) computational times

Figure 3: Solving (OPT-KH) with FW with short step (FW-SS) and open loop step-size rules for the form $\eta_t = \frac{2}{t+2}$ (FW-OL) with uniform distributions and computational times, Figure 3a and 3b, respectively.

3 Conclusion and outlook

> **Week 12 and now**: Proving $\mathcal{O}(1/t)$ convergence rate of FW with line-search/short-step step-size rule. As the proof is almost in the end and the research is ongoing that will be published, I could not disclose the detail of the further proof but will in the master's thesis.

During the internship, I gained a deep understanding of the Frank-Wolfe (FW) algorithm and was able to derive its convergence rate by hand. Moreover, from the recent research that FW and kernel herding are equivalent, I learned possibilities of applications of FW and even it can improve the result. I was given meaningful responsibilities and was able to contribute to ongoing projects. ZIB provided a supportive environment, and I learned communication and received feedback on my work. Moreover, I am glad that I can continue researching this interesting topic, and the valuable experience helped me decide to pursue becoming a researcher in the optimization area.
Overall, my internship at ZIB was an excellent opportunity to gain practical experience and learn more about the field. I wish to express my gratitude to the ZIB, Elias Wirth, and Dr. Mathieu Besançon for their support.

References

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. 2012. URL https://arxiv.org/abs/1203.4523.

A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59:235–247, 01 2004.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, page 109–116, 2010.

A. Christmann and I. Steinwart. *Support Vector Machines*. 2008. ISBN 978-0-387-77241-7. doi: 10.1007/978-0-387-77242-4.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *Arxiv e-prints*, arXiv:1805.08845v1 [stat.ML], 2018.

B. Polyak. *Introduction to Optimization*. 1987.

C. E. Rasmussen. *Gaussian Processes in Machine Learning*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-28650-9_4. URL https://doi.org/10.1007/978-3-540-28650-9_4.

K. Tsuji, K. Tanaka, and S. Pokutta. Sparser kernel herding with pairwise conditional gradients without swap steps, 2021. URL https://arxiv.org/abs/2110.12650.

K. K. Tsuji, K. Tanaka, and S. Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128.

M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1121–1128. Association for Computing Machinery, 2009. doi: 10.1145/1553374.1553517.

E. Wirth, T. Kerdreux, and S. Pokutta. Acceleration of frank-wolfe algorithms with open loop step-sizes, 2022.